THE FUNDAMENTALS OF DATA INSIGHT

Patrick Boily[1,2,3], Jen Schellinck[2,4,5]

Abstract

In October 2012, the *Harvard Business Review* published an article calling data science the “sexiest job of the 21st century”, and comparing data scientists with the ubiquitous Wall Street “quants” of the ’80s and ’90s. Data science has since become the "it" career. In this chapter, we discuss important non-technical data science notions that are too often swept under the rug.

Keywords

Ethical guidelines, pipelines and workflows, data structures, cognitive biases, applications, basic data analysis methods.

Funding Acknowledgement

Parts of this chapter were funded by Carleton University’s Centre for Quantitative Analysis and Decision Support.

[1]Department of Mathematics and Statistics, University of Ottawa, Ottawa, Canada
[2]Data Action Lab, Ottawa, Canada
[3]Idlewyld Analytics and Consulting Services, Wakefield, Canada
[4]Sysabee, Ottawa, Canada
[5]Institute of Cognitive Science, Carleton University, Ottawa, Canada
Email: pboily@uottawa.ca

Contents

1. Introduction

> The Problem is Not New
>
> We have learned to fly the air like birds and swim the sea like fish, but we have not learned the simple art of living together as brothers.
>
> – M.L. King, Jr., Nobel Peace Prize Lecture , 1964

In October 2012, the *Harvard Business Review* published an article calling data science the “sexiest job of the 21st century”, and comparing data scientists with the ubiquitous “quants” of the ’90s: a data scientist is a ”hybrid of data hacker, analyst, communicator, and trusted adviser” [26].

Would-be data scientists are usually introduced to the field *via* machine learning algorithms and applications. Much could be said about those (and will be broached in future reports), but we would like, for the time being, to mention some important non-technical notions that are sometimes swept under the rug.

With that in mind, we discuss some of the fundamental ideas and concepts that underlie and drive forward the discipline of data science, as well as the contexts in which these concepts are typically applied. We also highlight issues related to the ethics of practical data science. We conclude the chapter by getting a bit more concrete and considering the analytical workflow of a typical data science project, the types of roles and responsibilities that generally arise during data science projects and some basics of how to think about data, as a prelude to more technical topics.

1.1 What is Data?

It is surprisingly difficult to give a clear-cut definition of **data** – we cannot even seem to agree on whether it should be used in the singluar or the plural:

"the data is ... " vs. "the data are ..."

From a strictly linguistic point of view, a *datum* (borrowed from Latin) is "a piece of information;" **data**, then, should mean "pieces of information." We can also think of it as a collection of "pieces of information", and we would then use *data* to represent the whole (being potentially greater than the sum of its parts) or simply the idealized concept.[1]

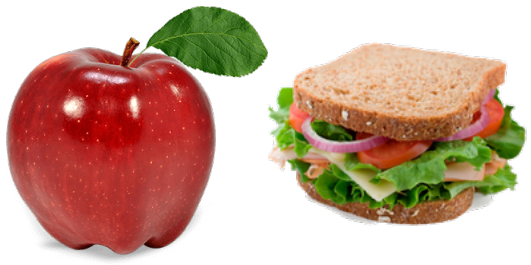
When it comes to actual data analysis, however, is the distinction really that important?

Is it even clear what data is, from the definition above, and where it comes from? Is the following data?

4,529 'red' 25.782 'Y'

To paraphrase Potter Stewart, while it may be hard to define what data is, "we know it when we see it." This position can strike some of you as unsatisfying; to overcome this objection, we will think of data simply as a collection of facts about **objects** and their **attributes**.

For instance, consider the apple and the sandwich below:



Let us say that they have the following attributes:

Object: apple

- **Shape:** spherical
- **Colour:** red
- **Function:** food
- **Location:** fridge
- **Owner:** Jen

Object: sandwich

- **Shape:** rectangle
- **Colour:** brown
- **Function:** food
- **Location:** office
- **Owner:** Pat

As long as we remember that a person or an object is not simply **the sum of its attributes**, this rough definition should not be too problematic.

Note, however, that there remains some ambiguity when it comes to **measuring** (and **recording**) the attributes.

We dare say that no one has ever beheld an apple quite like the one shown above: for starters, it is a 2-dimensional representation of a 3-dimensional object.

Additionally, while the overall shape of the sandwich is vaguely rectangular (as seen from above, say), it is not an exact rectangle. While no one would seriously dispute the shape attribute of the sandwich being recorded as "rectangle", a **measurement error** has occurred. For most analytical purposes, this error may not be significant, but it is impossible to dismiss it as such for all tasks.

More problematic might be the fact that the apple's shape attribute is given in terms of a volume, whereas the sandwich's is recorded as an a area; the measurement types are **incompatible**.

Similar remarks can be made about all the attributes – the function of an apple may be "food" from Jen's perspective, but from the point of view of an apple tree, that is emphatically not the case; the sandwich is definitely not uniformly "brown," and so on.

Furthermore, there are a number of potential attributes that are not even mentioned: size, weight, time, etc.

Measurement errors and incomplete lists are always part of the picture, but most people would recognize that the collection of attributes does provide a reasonable **description** of the objects. This is the pragmatic definition of data that we will use throughout.

1.2 From Objects and Attributes to Datasets

Raw data may exist in any format; we will reserve the term **dataset** to represent a collection fo data that could conceivably be fed into algorithms for analytical purposes.

Often, these appear in a **table** format, with rows and columns;[2] attributes are the **fields** (or columns) in such a dataset; objects are **instances** (or rows).

Objects are then described by their **feature vector** – the collection of attributes associated with value(s) of interest. The feature vector for a given observation is also know as the observation's **signature**.

For instance, the dataset of physical objects could contain the following items:

ID#	Shape	Colour	Function	Location	Owner
1	spherical	red	food	fridge	Jen
2	rectangle	brown	food	office	Pat
3	round	white	tell time	lounge	School
...	...	...	...	...	...

We will revisit this in more detail in Section 7.2.

[1] For what is worth, Jen prefers one, and Patrick the other.

[2] In practice, more complex **databases** are used.

1.3 Data in the News

We end this section with a sample of headlines and article titles showcasing the growing role of **data science** (DS), **machine learning** (ML), and a**rtificial/augmented intelligence** (AI) in different domains of society.

While these demonstrate some of the functionality/capabilities of DS/ML/AI technologies, it is important to remain aware that new technologies are always accompanied by emerging (and not always positive) social consequences.

- "Robots are better than doctors at diagnosing some cancers, major study finds" [29]
- "Deep-learning-assisted diagnosis for knee magnetic resonance imaging: Development and retrospective validation of MRNet" [10]
- "Google AI claims 99% accuracy in metastatic breast cancer detection" [8]
- "Data scientists find connections between birth month and health" [22]
- "Scientists using GPS tracking on endangered Dhole wild dogs" [50]
- "These AI-invented paint color names are so bad they're good" [63]
- "We tried teaching an AI to write Christmas movie plots. Hilarity ensued. Eventually." [34]
- "Math model determines who wrote Beatles' "In My Life": Lennon or McCartney?" [9]
- "Scientists use Instagram data to forecast top models at New York Fashion Week" [41]
- "How big data will solve your email problem " [38]
- "Artificial intelligence better than physicists at designing quantum science experiments" [69]
- "This researcher studied 400,000 knitters and discovered what turns a hobby into a business" [74]
- "Wait, have we really wiped out 60% of animals?" [79]
- "Amazon scraps secret AI recruiting tool that showed bias against women" [25]
- "Facebook documents seized by MPs investigating privacy breach" [7]
- "Firm led by Google veterans uses A.I. to 'nudge' workers toward happiness" [75]
- "At Netflix, who wins when it's Hollywood vs. the algorithm?" [62]
- "AlphaGo vanquishes world's top Go player, marking A.I.'s superiority over human mind" [43]
- "An AI-written novella almost won a literary prize" [49]
- "Elon Musk: Artificial intelligence may spark World War III" [51]
- "A.I. hype has peaked so what's next?" [65]

Opinions on the topic are varied – to some, DS/ML/AI provide examples of brilliant successes, while to others it is the dangerous failures that are at the forefront.

What do you think?

2. Analogue vs Digital

Humans have been collecting data for a long time. In the award-winning *Against the Grain: A Deep History of the Earliest States*, J.C. Scott argues that data collection was a major enabler of the modern nation-state (he also argues that this was not necessarily beneficial to humanity at large, but this is another matter altogether) [68].

For most of the history of data collection, humans were living in what might best be called the **analogue world** – a world where our understanding was grounded in a continuous experience of **physical reality**.

Nonetheless, even in the absence of computers, our data collection activities were, arguably, the first steps taken towards a different strategy for understanding and interacting with the world. Data, by its very nature, leads us to conceptualize the world in a way that is, in some sense, **more discrete than continuous**.

By translating our experiences and observations into numbers and categories, we re-conceptualize the world into one with sharper and more definable boundaries than our raw experience might otherwise suggest.

Fast-forward to the modern world and the culmination of this conceptual discretization strategy is clear to see in our adoption of the **digital computer**, which represents everything as a series of 1s and 0s.[3]

Somewhat surprisingly, this very minimalist representational strategy has been wildly successful at **representing our physical world**, arguably beyond our most ambitious dreams, and we find ourselves now at a point where what we might call the **digital world** is taking on a reality as pervasive and important as the physical one.

Clearly, this digital world is built on top of the physical world, but very importantly, the two do not operate under the same set of rules:

- in the physical world, the default is to **forget**; in the digital world, the default is to **remember**;
- in the physical world, the default is **private**; in the digital world, the default is **public**;
- in the physical world, copying is **hard**; in the digital world, copying is **easy**.

As a result of these different rules of operation, the digital is making things that were **once hidden, visible; once veiled, transparent**.

Considering data science in light of this new digital world, we might suggest that data scientists are, in essence, scientists of the **digital**, in much the same way that regular scientists are scientists of the **physical**: data scientists seek

[3] Or 'On' and 'Off', 'TRUE' and 'FALSE'.

to discover the **fundamental principles of data** and understand the ways in which these fundamental principles manifest themselves in different digital phenomena.

Ultimately, however, data and the digital world are **tied to the physical world**. Consequently, what is done with data has repercussions in the physical world; and it is crucial for analysts and consultants to have a solid grasp of the fundamentals and context of data work before leaping into the tools and techniques that drive it forward.

3. Conceptual Frameworks for Data Work

In simple terms, we use data to represent the world. But this is not the only strategy at our disposal: we might also (and in combination) describe the world using **language**, or represent it by building **physical models**.

The common thread is the more basic concept of **representation** – the idea that one object can stand in for another, and be used in its stead in order to indirectly engage with the object being represented.

Humans are representational animals *par excellence*; our use of representations becomes almost transparent to us, at times.

On some level, we do understand that "the map is not the territory", but we do not have to make much of an effort to use the map to navigate the territory. The transition from the **representation** to the **represented** is typically quite seamless.

This is arguably one of humanity's major strengths, but in the world of data science it can also act as an Achilles' heel, preventing analysts from working successfully with clients and project partners, and from appropriately **transferring analytical results** to the real world contexts that could benefit from them.

The best protection against these potential threats is the existence of a well thought out and explicitly described **conceptual framework**, by which we mean, in its broadest sense:

- a **specification** of which parts of the world are being represented;
- **how** they are represented;
- the **nature of the relationship** between the represented and the representing, and, coming out of this,
- **appropriate** and **rigorous strategies** for applying the results of the analysis that is carried out in this representational framework.

It would be possible to construct such a specification from scratch, in a piecemeal fashion, for each new project, but it is worth noting that there are some overarching **modeling frameworks** that are broadly applicable to many different phenomena, which can then be moulded to fit these more specific instances.

3.1 Three Modeling Strategies

We suggest that there are three main not mutually exclusive **modeling strategies** that can be used to guide the specification of a phenomenon or domain:

- **mathematical** modeling;
- **computer** modeling, and
- **systems** modeling.

We start with a description of the latter as it requires, in its simplest form, no special knowledge of techniques/concepts from mathematics or computer science.

Systems Modeling **General Systems Theory** was initially put forward by L. von Bertalanffy, a biologist, who felt that it should be possible to describe many **disparate** natural phenomena using a **common conceptual framework** – one which would be capable of describing many disparate phenomena, all as systems of interacting objects.

Although Bertalanffy himself presented abstracted, mathematical, descriptions of his general systems concepts, his broad strategy is relatively easily translated into a purely conceptual framework.

Within this framework, when presented with a novel domain or situation, we ask ourselves the following questions:

- which objects seem most relevant or involved in the system behaviours in which we are most interested?
- what are the properties of these objects?
- what are the behaviours (or actions) of these objects?
- what are the relationships between these objects?
- how do the relationships between objects influence their properties and behaviours?

As we find the answers to these questions about the system of interest, we start to develop a sense that we **understand the system** and its **relevant behaviours**.

By making this knowledge **explicit**, e.g. *via* diagrams and descriptions, and by sharing it amongst those with whom we are working, we can further develop a **consistent**, **shared understanding** of the system with which we are engaged.

If this activity is carried out prior to data collection, it can ensure that the **right data** is collected. If this activity is carried out after data collection, it can ensure that the process of **interpreting what the data represents** and how the latter should be used going forward is on solid footing.

Mathematical and Computer Modeling The other modeling approaches arguably come with their own general frameworks for interpreting and representing real-world phenomena and situations, separate from, but still compatible with, this systems perspective.

These disciplines have developed their own mathematical/digital (logical) worlds that are distinct from the tangible, physical world studied by chemists, biologists, and

so on; these frameworks can then be used to describe real-world phenomena by **drawing parallels** between the properties of objects in these different worlds and reasoning via these parallels.

Why these **constructed worlds** and the conceptual frameworks they provide are so effective at representing and describing the actual world, and thus allowing us to understand and manipulate it, is more of a philosophical question than a pragmatic one.

We will only note that they are **highly effective** at doing so, which provides the impetus and motivation to learn more about how these worlds operate, and how, in turn, they can provide data scientists with a means to engage with domains and systems through a powerful, rigorous and shared conceptual framework.

3.2 Information Gathering

The importance of achieving **contextual understanding** of a dataset cannot be over-emphasized. In the abstract we have suggested that this context can be gained by using conceptual frameworks. But more concretely, how does this understanding come about?

It can be reached through:

- **field trips**;
- interviews with **subject matter experts** (SMEs);
- **readings/viewings**;
- **data exploration** (even just **trying to obtain** or gain access to the data can prove a major pain),
- etc.

In general, clients or stakeholders are **not a uniform** entity – it is even conceivable that client data specialists and SMEs will **resent the involvement** of analysts (external and/or internal).

Thankfully, this stage of the process provides analysts and consultants the opportunity to show that every one is pulling in the same direction, by

- asking **meaningful** questions;
- taking an interest in the SMEs'/clients' experiences, and
- acknowledging everyone's ability to contribute.

A little tact goes a long way when it comes to information gathering.

Thinking in Systems Terms We have already noted that a **system** is made up of **objects** with **properties** that potentially change over time. Within the system we perceive **actions** and **evolving properties**, leading us to think in terms of **processes**.

To put it another way, in order to understand how various aspects of the world interact with one another, we need to **carve out chunks** corresponding to the aspects and define their boundaries.

Working with other intelligences requires this type of **shared understanding** of what is being studied.

Objects themselves have various properties. Natural processes generate (or destroy) objects, and may change the properties of these objects over time. We **observe**, **quantify**, and **record** particular values of these properties at particular points in time.

This process generates data points in our attempt to **capture the underlying reality** to some acceptable degree of **accuracy** and **error**, but it remains crucial for data analysts and data scientists to remember that **even the best system model only ever provides an approximation of the situation under analysis**; with some luck, experience, and foresight, these approximations might turn out to be **valid**.

Identifying Gaps in Knowledge A **gap in knowledge** is identified when we realize that what we thought we knew about a system proves **incomplete** (or blatantly false).

This can arise as the result of a certain naïveté vis-à-vis the situation being modeled, but it can also be emblematic of the nature of the project under consideration: with too many moving parts and grandiose objectives, there cannot help but be knowledge gaps.[4]

Knowledge gaps might occur **repeatedly**, at any moment in the process:

- data **cleaning**;
- data **consolidation**;
- data **analysis**;
- even during **communication of the results** (!).

When faced with such a gap, the best approach is to be flexible: **go back**, **ask questions**, and **modify the system representation** as often as is necessary. For obvious reasons, it is preferable to catch these gaps early on in the process.

Conceptual Models Consider the following situation: you are away on business and you forgot to hand in a very important (and urgently required) architectural drawing to your supervisor before leaving. Your office will send a gopher to pick it up in your living space. How would you explain to them, by phone, how to find the document?

If the gopher has previously been in your living space, if their living space is comparable to yours, or if your spouse is at home, the process may be able to be sped up considerably, but with somebody for whom the space is new (or someone with a visual impairment, say), it is easy to see how things could get complicated.

But time is of the essence – you and the gopher need to get the job done **correctly** as **quickly as possible**. What is your strategy?

[4] Note that it also happens with small, well-organized, and easily contained projects. It happens all the time, basically.

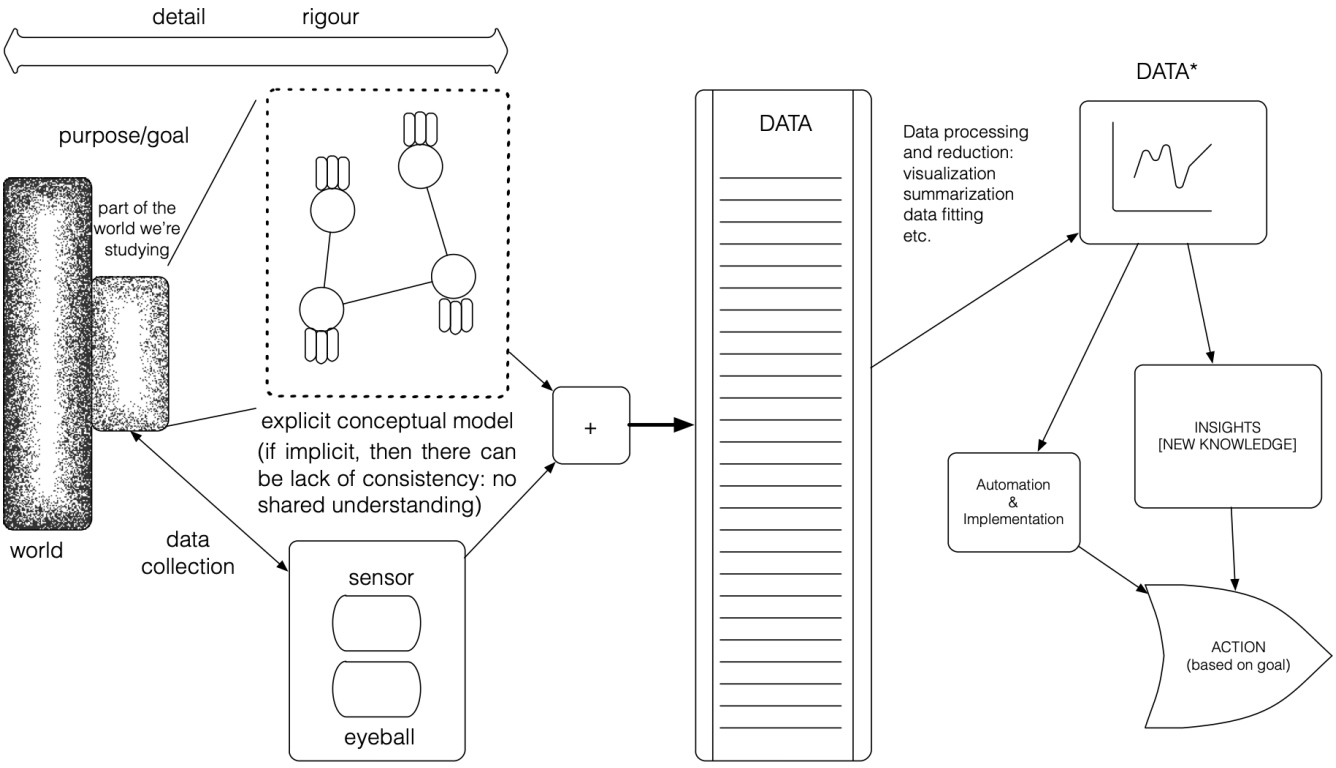


Figure 1. A schematic diagram of systems thinking as it applies to a general problem.

Conceptual models are built using methodical investigation tools:

- **diagrams**;
- structured **interviews**;
- structured **descriptions**,
- etc.

Data analysts and data scientists should beware **implicit conceptual models** – they go hand-in-hand with knowledge gaps.

In our opinion, it is preferable to err on the side of "too much conceptual modeling" than the alternative (although, at some point we have to remember that every modeling excercise is wrong[5] and that there is nothing wrong with building better models iteratively, over the bones of previously discarded simplistic models).

Roughly speaking, a **conceptual model** is a model that is not implemented as a scale-model or computer code, but one which exists only conceptually, often in the form of a diagram or verbal description of a system – boxes and arrows, mind maps, lists, definitions (see Figures 1 and 2).

Conceptual models do not necessarily attempt to capture specific behaviours, but they emphasize the **possible states** of the system: the focus is on object types, not on specific instances, with **abstraction** as the ultimate objective.

[5] "Every model is wrong; some models are useful." *George Box*.

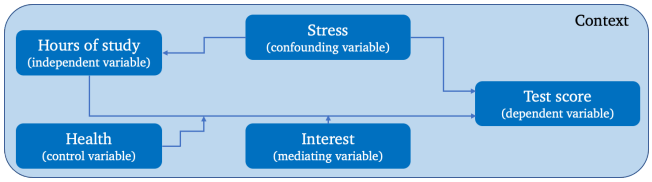
Conceptual modeling is not an exact science – it is more about making internal conceptual models **explicit** and **tangible**, and providing data analysis teams with the opportunity to **examine** and **explore** their ideas and assumptions. Attempts to formalize the concept include (see Figure 3):

- **Universal Modeling Language** (UML);
- **Entity Relationship Models** (ER), generally connected to relational databases.

In practice, we must first select a system for the task at hand, then generate a conceptual model that encompasses:

- **relevant** and **key objects** (abstract or concrete);
- **properties** of these objects, and their values;
- **relationships between objects** (part-whole, is-a, object-specific, one-to-many), and
- **relationships between properties** across instances of an object type.

A simplistic example describing a supposed relationship between a **presumed cause** (hours of study) and a **presumed effect** (test score) is shown below:



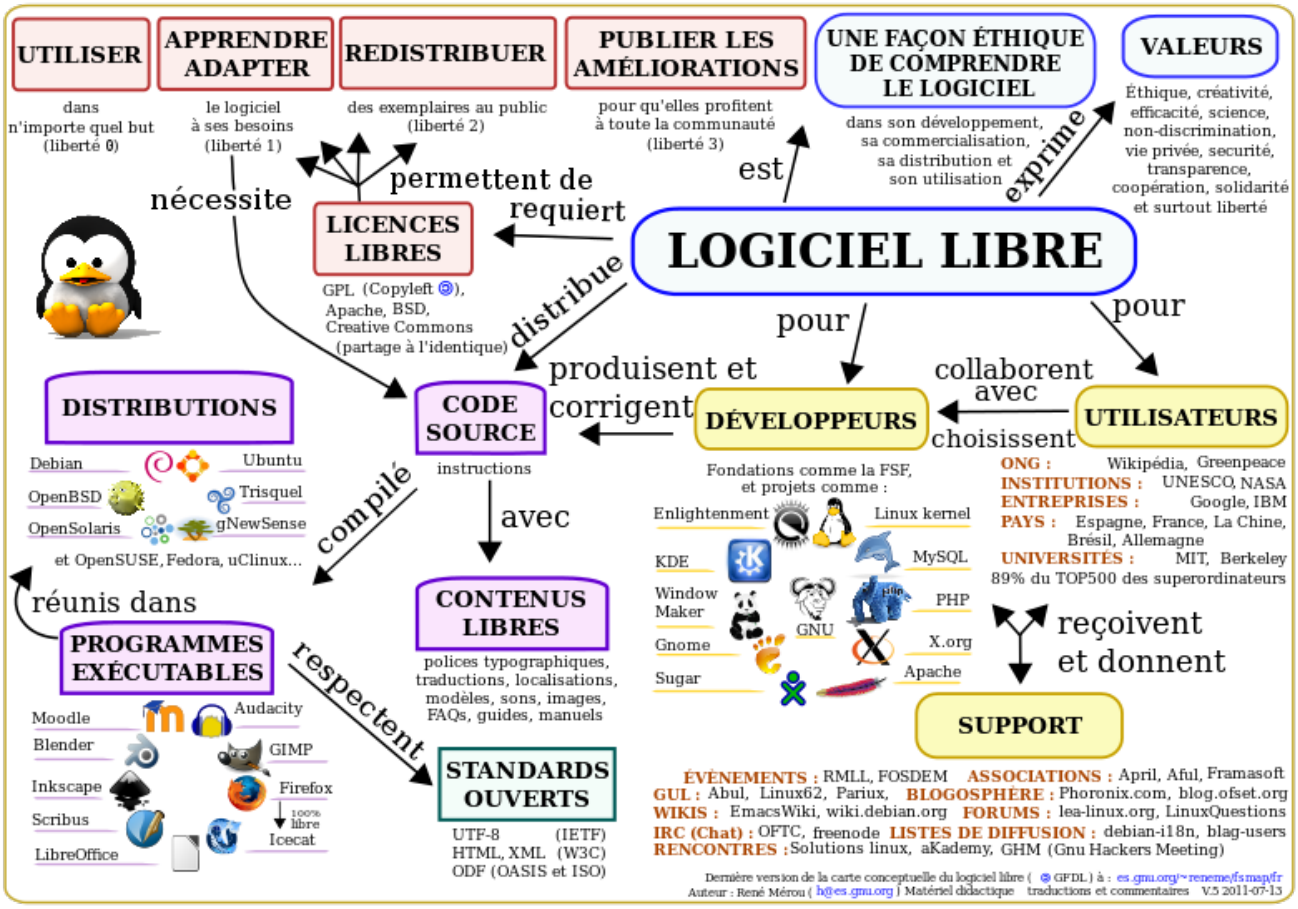


Figure 2. A conceptual model of the "free software" system (in French) [53].

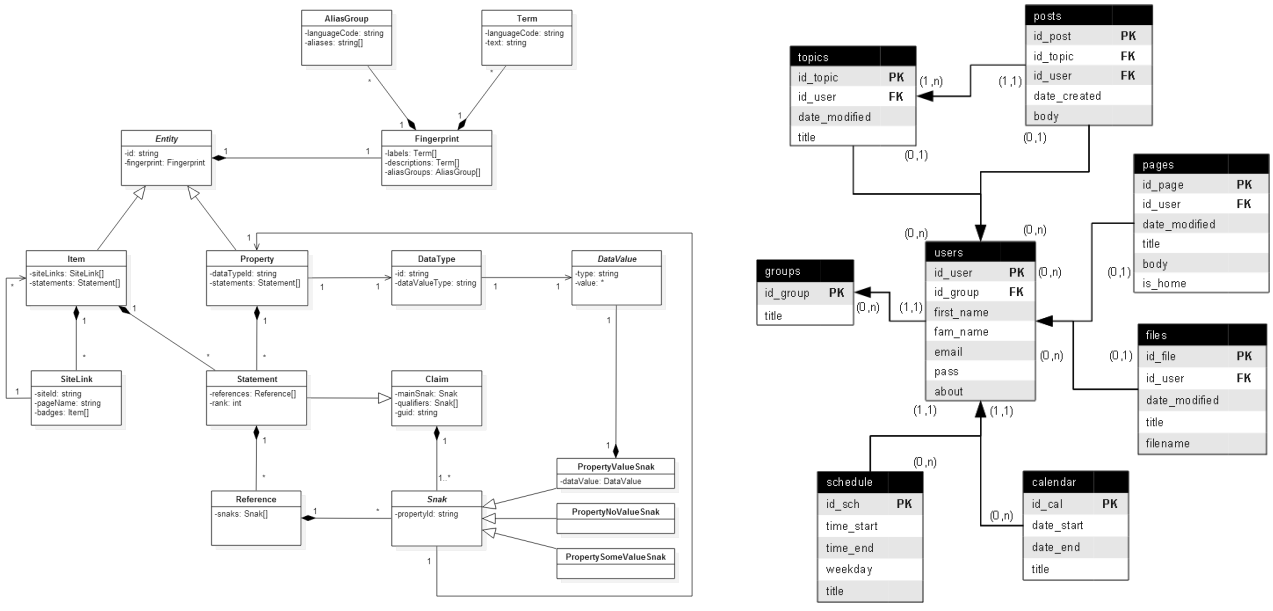


Figure 3. Examples of UML diagram (Wikibase Data Model, on the left [37]) and ER (on the right [78]) conceptual maps.

Relating the Data to the System From a pragmatic perspective, stakeholders and analysts alike need to know if the data which has been collected and analyzed will be useful to understand the system.

This question can best be answered if we understand:

- how the data is collected;
- the approximate nature of both data and system, and
- what the data represents (observations and features).

Is the **combination of system and data sufficient** to understand the aspects of the world under consideration? Again, this is difficult to answer in practice.

Contextual knowledge can help, but if the data, the system, and the world are **out of alignment**, any data insight drawn from mathematical, ontological, programmatical, or data models of the situation might ultimately prove useless.

3.3 Cognitive Biases

Adding to the challenge of building good conceptual models and using these to interpret the data is the fact that we are all vulnerable to a vast array of **cognitive biases**, which influence both how we construct our models and how we look for patterns in the data.

These biases are difficult to detect in the spur of the moment, but being aware of them, making a conscious effort to identify them, and setting up a clear and pre-defined set of thresholds and strategies for analysis will help reduce their negative impact.

Here is a sample of such biases (taken from [28, 48]).

Anchoring bias causes us to rely too heavily on the first piece of information we are given about a topic; in a salary negotiation, for instance, whoever makes the first offer establishes a range of reasonable possibilities in both parties' minds.

Availability heuristic describes our tendency to use information that comes to mind quickly and easily when making decisions about the future; someone might argue that climate change is a hoax because the weather in their neck of the woods has not (yet!) changed.

Choice-supporting bias causes us to view our actions in a positive light, even if they are flawed; we are more likely to sweep anomalous or odd results under the carpet when they arise from our own analyses.

Clustering illusion refers to our tendency to see patterns in random events; if a die has rolled five 3's in a row, we might conclude that the next throw is more (or less) likely to come up a 3 (gambling fallacy).

Confirmation bias describes our tendency to notice, focus on, and give greater credence to evidence that fits with our existing beliefs; gaffes made by politicians you oppose reinforces your dislike.

Conservation bias occurs when we favour prior evidence over new information; it might be difficult to accept that there is an association between factors X and Y if none had been found in the past.

Ostrich effect describes how people often avoid negative information, including feedback that could help them monitor their goal progress; a professor might chose to not consult their teaching evaluations, for whatever reason.

Outcome bias refers to our tendency to judge a decision on the outcome, rather than on why it was made; the fact that analysts gave Clinton an 80% chance of winning the 2016 U.S. Presidential Election does not mean that the forecasts were wrong.

Overconfidence causes us to take greater risks in our daily lives; experts are particularly prone to this, as they are more convinced that they are right.

Pro-innovation bias occurs when proponents of a technology overvalue its usefulness and undervalue its limitations; in the end, Big Data is not going to solve all of our problems.

Recency bias occurs when we favour new information over prior evidence; investors tend to view today's market as the "forever' market and make poor decisions as a result.

Salience Bias describes our tendency to focus on items or information that are more noteworthy while ignoring those that do not grab our attention; you might be more worried about dying in a plane crash than in a car crash, even though the latter occurs more frequently than the former.

Survivorship Bias is a cognitive shortcut that occurs when a visible successful subgroup is mistaken as an entire group, due to the failure subgroup not being visible; when trying to get the full data picture, it helps to know what observations did not make it into the dataset.

Zero-Risk Bias relates to our preference for absolute certainty; we tend to opt for situations where we can completely eliminate risk, seeking solace in the figure of 0%, over alternatives that may actually offer greater risk reduction.

Other biases impact our ability to make informed decisions:

bandwagon effect, base rate fallacy, bounded rationality, category size bias, commitment bias, Dunning-Kruger effect, framing effect, hot-hand fallacy, IKEA effect, illusion of explanatory depth, illusion of validity, illusory correlations, look elsewhere effect, optimism effect, planning fallacy, representative heuristic, response bias, selective perception, stereotyping, etc. [28, 48].

4. Ethics in the Data Science Context

Straight Talk From The Front Lines

> A lapse in ethics can be a conscious choice... but it can also be negligence.
>
> – R. Schutt, C. O'Neill [67]

In most empirical disciplines, **ethics** are brought up fairly early in the educational process and may end up playing a crucial role in researchers' activities.

At Memorial University of Newfoundland, for instance, "proposals for research in the social sciences, humanities, sciences, and engineering, including some health-related research in these areas," must receive approval from specific Ethics Research Boards .

This could, among other cases, apply to research and analysis involving [64]:

- living human subjects;
- human remains, cadavers, tissues, biological fluids, embryos or foetuses;
- a living individual in the public arena if s/he is to be interviewed and/or private papers accessed;
- secondary use of data – health records, employee records, student records, computer listings, banked tissue – if any form of identifier is involved and/or if private information pertaining to individuals is involved, and
- quality assurance studies and program evaluations which address a research question.

In our experience, data scientists and data analysts who come to the field by way of mathematics, statistics, computer science, economics, or engineering, however, are not as likely to have encountered ethical research boards or to have had **formal ethics training**.[6]

Furthermore, discussions on ethical matters are often tabled, perhaps understandably, in favour of pressing technical or administrative considerations (such as algorithm selection, data cleaning strategies, contractual issues, etc.) when faced with hard deadlines.

The problem, of course, is that the current deadline is eventually replaced by another deadline, and then by a new deadline, with the end result that the conversation may never take place.

It is to address this all-too-common scenario that we take the time to discuss ethics in the **data science context**; more information is available in [58].

[6] We are obviously not implying that these individuals have no ethical principles or are unethical; rather, that the opportunity to establish what these principles might be, in relation with their research, may never have presented itself.

4.1 The Need for Ethics

When large scale data collection first became possible, there was to some extent a 'Wild West' mentality to data collection and use. To borrow from the old English law principle, whatever was not prohibited (from a technological perspective) was allowed.

Now, however, **professional codes of conduct** are being devised for data scientists [1,17,73], outlining responsible ways to practice data science – ways that are legitimate rather than fraudulent, and ethical rather than unethical.[7]

Although this shifts some added responsibility onto data scientists, it also provides them with protection from clients or employers who would hire them to carry out data science in questionable ways – they can refuse on the grounds that it is against their professional code of conduct.

4.2 What Is/Are Ethics?

Broadly speaking, ethics refers to the study and definition of right and wrong conducts. Ethics may consider what is a right or a wrong action in general, or consider how broad ethical principles are appropriately applied in more specific circumstances.

And, as noted by R.W. Paul and L. Elder, ethics is not (necessarily) the same as social convention, religious beliefs, or laws [59]; that distinction is not always fully understood.

The following influential ethical theories are often used to frame the debate around ethical issues in the data science context:

- **Kant's golden rule:** do unto others as you would have them do unto you;
- **Consequentialism:** the end justifies the means;
- **Utilitarianism:** act in order to maximize positive effect;
- **Moral Rights:** act to maintain and protect the fundamental rights and privileges of the people affected by actions;
- **Justice:** distribute benefits and harm among stakeholders in a fair, equitable, or impartial way.

In general, it is important to remember that our planet's inhabitants subscribe to a wide variety of ethical codes, including also:

> Confucianism, Taoism, Buddhism, Shinto, Ubuntu, Te Ara Tika (Maori), First Nations Principles of OCAP, various aspects of Islamic ethics, etc.

It is not too difficult to imagine contexts in which either of these (or other ethical codes, or combinations thereof) would be better-suited – the challenge is to remember to **inquire**, and to **heed the answers**.

[7] This is not to say that ethical issues have miraculously disappeared – Volkswagen, Whole Foods Markets, General Motors, and Ashley Madison, to name but a few of the big data science and data analysis players, have all recently been implicated in ethical lapses [31]. More dubious examples can be found in [19,54].

4.3 Ethics and Data Science

How might these ethical theories apply to data analysis? The (former) University of Virginia's *Centre for Big Data Ethics, Law and Policy* suggested some specific examples of data science ethics questions [16]:

- who, if anyone, owns data?
- are there limits to how data can be used?
- are there value-biases built into certain analytics?
- are there categories that should never be used in analyzing personal data?
- should data be publicly available to all researchers?

The answers may depend on a number of factors, not least of which being who is actually providing them.

To give you an idea of some of the complexities, let us consider the first of those questions: who, if anyone, owns data?

In some sense, the **data analysts** who transform the data's potential into usable insights are only one of the links in the entire chain. Processing and analyzing the data would be impossible without raw data on which to work, so the **data collectors** also have a strong ownership claim to the data.

But collecting the data can be a costly endeavour, and it is easy to imagine how the **sponsors** or **employers** (who made the process economically viable in the first place) might feel that the data and its insights are rightfully theirs to dispose of as they wish.

In some instances, the **law** may chime in as well. One can easily include other players: in the final analysis, this simple question turns out to be far from easily answered.

This also highlights some of the features of the data analysis process, which we will discuss in Section 5: there is more to data analysis than *just* data analysis. The answer is not easily forthcoming, and may change from one case to another.

A similar challenge arises in regards to **open data**, where the "pro" and "anti" factions both have strong arguments (see [14, 55, 56], and [23] for a science-fictional treatment of the transparency-vs.-secrecy/security debate).

A general principle of data analysis is to **eschew the anecdotal** in favour of the **general** – from a purely analytical perspective, too narrow a focus on specific observations can end up obscuring the full picture (a vivid illustration can be found in [21]).

But data points are **not** solely marks on paper or electromagnetic bytes on the cloud. Decisions made on the basis of data science (in all manners of contexts, from security, to financial and marketing context, as well as policy) may **affect living beings in negative ways**. And it can not be ignored that outlying/marginal individuals and minority groups often suffer disproportionately at the hands of so-called evidence-based decisions [33, 44, 45].

4.4 Guiding Principles

Under the assumption that one is convinced of the importance of proceeding ethically, it could prove helpful to have a set of guiding principles to aid in these efforts.

In his seminal science fiction series about *positronic robots*, Isaac Asimov introduced the now-famous *Laws of Robotics*, which he believed would have to be built-in so that robots (and by extension, any tool used by human beings) could overcome humanity's *Frankeinstein*'s complex (the fear of mechanical beings) and help rather than hinder human social, scientific, cultural, and ecomomic activities [5]:

1. A robot may not injure a human being or, through inaction, allow a human being to come to harm.
2. A robot must obey the orders given to it by human beings, except where such orders would conflict with the 1st Law.
3. A robot must protect its own existence as long as such protection does not conflict with the 1st and 2nd Law.

Were they uniformly well-implemented and respected, the potential for story-telling would have been somewhat reduced; thankfully, Asimov found entertaining ways to break the Laws (and to resolve the resulting conflicts) which made the stories both enjoyable and insightful.

Interestingly enough, he realized over time that a Zeroth Law had to supersede the First in order for the increasingly complex and intelligent robots to succeed in their goals. Later on, other thinkers contributed a few others, filling in some of the holes (see Table 1).

We cannot speak for the validity of these laws for **robotics** (a term coined by Asimov, by the way), but we do find the entire set satisfyingly complete.

What does this have to do with data science? Various thinkers have discussed the existence and potential merits of different sets of Laws ([70]) – wouldn't it be useful if there were *Laws of Analytics*, **moral principles that could help us conduct data science ethically**?

Best Practices Such universal principles are unlikely to exist, but a number of best practices and guiding principles have been suggested.

"Do No Harm": data collected from an individual **should not be used to harm the individual**. This may be difficult to track in practice, as data scientists and analysts do not always participate in the ultimate decision process.

Informed Consent covers a wide variety of ethical issues, chief among them being that **individuals must agree to the collection and use** of their data, and that they must have a **real understanding of what they are consenting to**, and of **possible consequences** for them and others.

00. A robot may not harm sentience or, through inaction, allow sentience to come to harm.

0. A robot may not harm humanity, or, through inaction, allow humanity to come to harm, as long as this action/inaction does not conflict with the 00th Law.

1. A robot may not injure a human being or, through inaction, allow a human being to come to harm, as long as this does not conflict with the 00th or the 0th Law.

2. A robot must obey the orders given to it by human beings, except where such orders would conflict with the 00th, the 0th or the 1st Law.

3. A robot must protect its own existence as long as such protection does not conflict with the 00th, the 0th, the 1st or the 2nd Law.

4. A robot must reproduce, as long as such reproduction does not interfere with the 00th, the 0th, the 1st, the 2nd or the 3rd Law.

5. A robot must know it is a robot, unless such knowledge would contradict the 00th, the 0th, the 1st, the 2nd, the 3rd or the 4th Law.

Table 1. Asimov's (expanded) *Laws of Robotics*.

The Respect of "Privacy" is a dearly-held principle, but it is hard to adhere to it religiously with robots and spiders constantly trolling the net for personal data. In the *Transparent Society*, D. Brin (somewhat) controversially suggests that privacy and total transparency are closely linked [14]:

> "And yes, **transparency is also the trick to protecting privacy**, if we empower citizens to notice when neighbors [*sic*] infringe upon it. Isn't that how you enforce your own privacy in restaurants, where people leave each other alone, because those who stare or listen risk getting caught?'

Keeping Data Public is another aspect of data privacy, and a thornier issue – should some data be kept private? Most? All? It is fairly straightforward to imagine scenarios where adherence to the principle of public data could cause harm to individuals (revealing the source of a leak in a country without where the government routinely jails members of the opposition, say), contradicting the first principle against causing harm. But it is just as easy to imagine scenarios where keeping data private would have a similar effect.

Opt-in/Opt-out: informed consent requires the ability to **not consent**, i.e. to opt out. Non-active consent is not really consent.

Anonymize Data: identifying fields should be removed from the dataset **prior** to processing and analysis. Let any temptation to use personal information in an inappropriate manner be removed from the get-go, but be aware that this is easier said than done, from a technical perspective.

Let the Data Speak: absolutely no cherry-picking of your data. Use all of it in some way or another. Validate your analysis and make sure your results are repeatable.

4.5 The Good, the Bad, and the Ugly

Data projects could be classified as **good**, **bad** or **ugly**, either from a technical or from an ethical standpoint (or both).

We have identified instances in each of these classes (of course, our biases might show through):

- **good** projects increases knowledge, can help uncover hidden links, and so on: [6, 8–10, 15, 22, 24, 40, 42, 43, 47, 57, 60, 69, 74]
- **bad** projects, if not done properly, can lead to bad decisions, which can in turn decrease the public's confidence and potentially harm some individuals: [21, 41, 50, 62, 75]
- **ugly** projects are, flat out, unsavoury applications; they are poorly executed from a technical perspective, or put a lot of people at risk; these (and similar approaches/studies) should be avoided: [7, 25, 33, 44–46]

5. Analytics Workflow

An overriding component of the discussion so far has been the **importance of context**. And although the reader may be eager at this point to move into data analysis proper, there is one more context that should be considered first – the **project context**.

We have alluded to the idea that data science is much more than simply data analysis and this is apparent when we look at the typical steps involved in a data science project. Inevitably, data analysis pieces take place within this larger project context, as well as in the context of a larger **technical infrastructure** or **pre-existing system**.

5.1 The "Analytical" Method

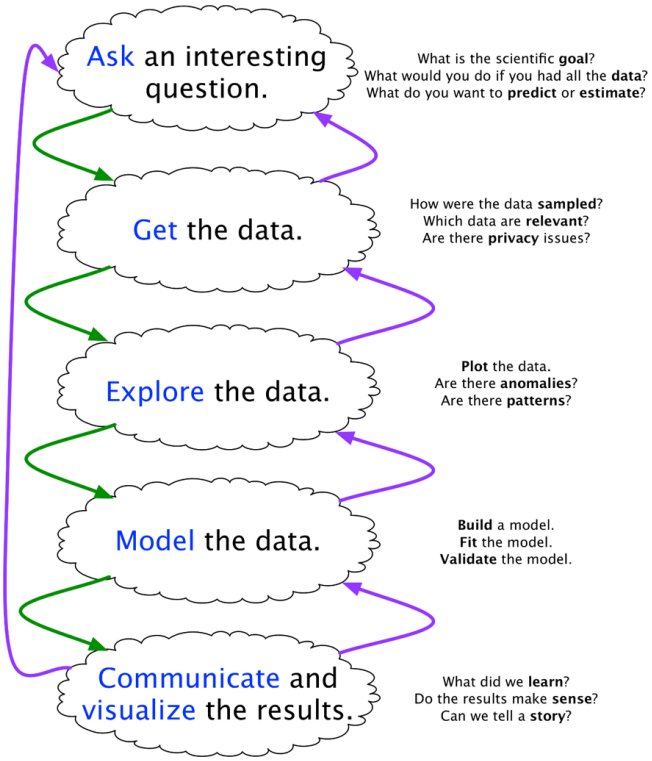
As with the **scientific method**, there is a "step-by-step" guide to data analysis:

1. statement of objective
2. data collection
3. data clean-up
4. data analysis/analytics
5. dissemination
6. documentation

Notice that **data analysis** only makes up a small segment of the entire flow.

In practice, the process often end up being a bit of a mess, with steps taken out of sequence, steps added-in, repetitions and re-takes (see Figure 4). And yet, it works on the whole (if done correctly).

J. Blitzstein and H. Pfister (who teach a well-rated data science course at Harvard) provide their own workflow diagram, but the similarities are easy to spot (see below).



The **Cross Industry Standard Process, Data Mining** is another framework, with projects consisting of 6 steps:

1. business understanding
2. data understanding
3. data preparation
4. modeling
5. evaluation
6. deployment

The process is iterative and interactive – the dependencies are highlighted in Figure 5.

In practice, the process is often corrupted by:

1. lack of clarity;
2. mindless rework;
3. blind hand-off to IT, and
4. failure to iterate.

CRISP-DM has a definite old-hat flavour (witness the use of the expression "data mining," which has become outdated), but it can be useful to check off its sub-components, if only for a sanity check.

Business Understanding:

- understanding the business goal
- assessing the situation
- translating the goal in a data analysis objective
- developing a project plan

Data Understanding:

- considering data requirements
- collecting and exploring data

Data Preparation:

- selection of appropriate data
- data integration and formatting
- data cleaning and processing

Modeling:

- selecting appropriate techniques
- splitting into training/testing sets
- exploring alternatives methods
- fine tuning model settings

Evaluation:

- evaluation of model in a business context
- model approval

Deployment:

- reporting findings
- planning the deployment
- deploying the model
- distributing and integrating the results
- developing a maintenance plan
- reviewing the project
- planning the next steps

All these approaches have a common core: data science projects are **iterative** and (often) **non-sequential**.

Helping the clients and/or stakeholders recognize this central truth will make it easier for analysts and consultants to **plan the data science process** and to obtain **actionable insights** for organizations and sponsors.

Another take-away is that there is a lot of real estate in the process before we can even start talking about modeling and analysis – **data analysis is not only about data analysis**.

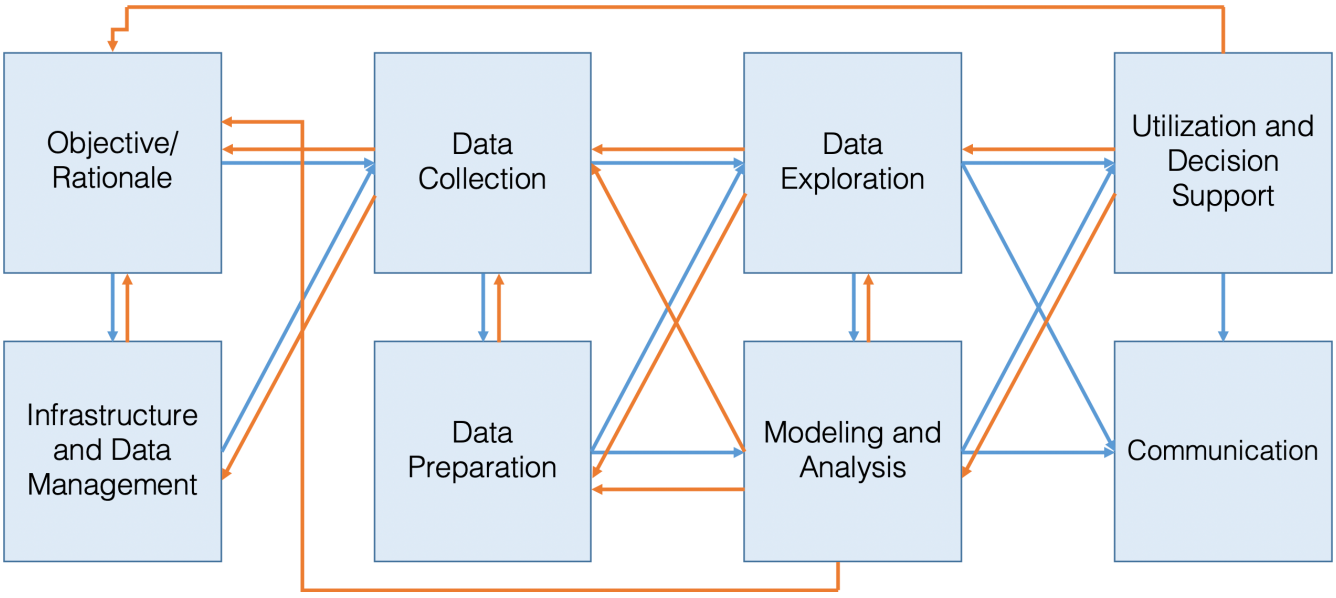


Figure 4. The reality of the analytic workflow – definitely not a linear process!

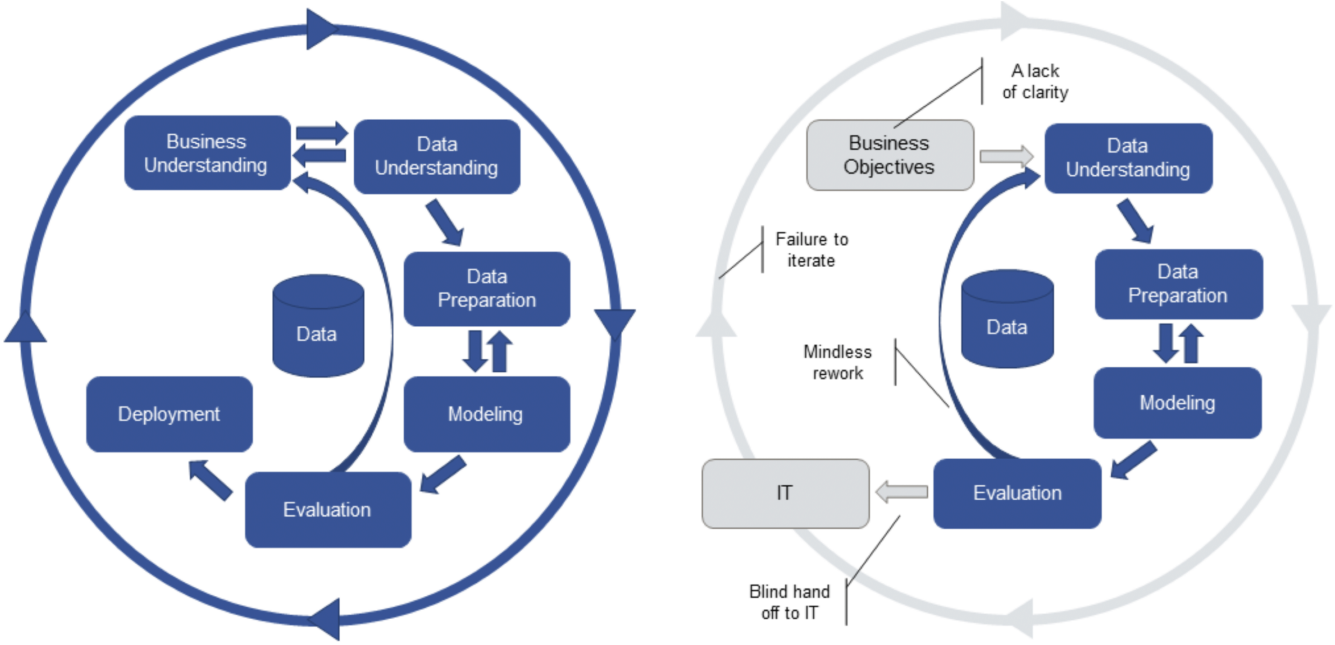


Figure 5. CRISP-DM in theory (left); corrupted CRISP-DM often found in practice (right) [72].

5.2 Data Collection, Storage, Processing, and Modeling

Data enters the pipeline by being **collected**. There are various possibilities:

- data may be collected in a **single pass**;
- it may be collected in **batches**, or
- it may be collected **continuously**.

This **mode of entry** may have an impact on the subsequent steps, including on how frequently models, metrics, and other outputs are **updated**.

Once it is collected, data must be **stored**. Choices related to storage (and **processing**) must reflect:

- how the data is collected (mode of entry);
- how much data there is to store and process (small vs. big), and
- the type of access and processing that will be required (how fast, how much, by whom).

Unfortunately, stored data may go **stale** (addresses no longer accurate, etc.); regular data audits are recommended.

Processing the data is required before it can be analyzed. This is discussed in detail in other reports/chapters/documents, but the key point is that **raw data** has to be converted into a format that is **amenable to analysis**, by

- identifying **invalid**, **unsound**, and **anomalous** entries;
- dealing with **missing values**;
- **transforming** the variables and the datasets so that they meet the requirements of the selected algorithms.

In contrast, the **analysis** step itself is almost anti-climactic – simply run the selected methods/algorithms on the processed data. The specifics of this procedure depend, of course, on the choice of method/algorithm.

We will not get into the details of how to make that choice[8], but data science teams should be familiar with a fair number of techniques and approaches:

- data cleaning
- descriptive statistics and correlation
- probability and inferential statistics
- regression analysis (linear and other variants)
- survey sampling
- bayesian analysis
- classification and supervised learning
- clustering and unsupervised learning
- anomaly detection and outlier analysis
- time series analysis and forecasting
- optimization
- high-dimensional data analysis
- stochastic modeling
- distributed computing
- etc.

These only represent a **small slice** of the analysis pie; you may not master them all (maybe not even a majority) at the moment, but that is one of the reasons why data science is a team sport (more on this in Section 6).

5.3 Model Assessment and Life After Analysis

Before applying the findings from a model or an analysis, one must first confirm that the model is reaching valid conclusions about the system of interest.

All analytical processes are, by their very nature, **reductive** – the raw data is eventually transformed into a small(er) **numerical outcome** (or summary) by various analytical methods, which we hope is still **related** to the system of interest (see Section 3).

Data science methodologies include an **assessment** (evaluation, validation) phase. This does not solely provide an analytical sanity check (i.e., are the results analytically compatible with the data?); it can be used to determine when the system and the data science process have stepped out of alignment.

[8]Truth be told, choosing wisely is probably the the most difficult aspect of a data science project.

Past successes can lead to reluctance to re-assess and re-evaluate a model. Even if the analytical approach has been vetted and has given useful answers in the past, it may not always do so.

At what point does one determine that the current data model is **out-of-date**? At what point does one determine that the current model is no longer **useful**? How long does it take a model to react to a **conceptual shift**?[9] This is another reason why regular audits are recommended – as long as the analysts remain in the picture, the only obstacle to performance evaluation might be the technical difficulty of conducting said evaluation.

When an analysis or model is 'released into the wild' or delivered to the client, it often take on a life of its own. When it inevitably ceases to be **current**, there may be little that (former) analysts can do to remedy the situation.

Consultants and analysts rarely have full (or even partial) control over **model dissemination**; consequently, results may be misappropriated, misunderstood, shelved, or failed to be updated, all without their knowledge. What can conscientious analysts do to prevent this?

Unfortunately, there is no easy answer, short of advocating for analysts and consultants to not solely focus on data analysis – data science projects afford an opportunity to **educate clients and stakeholders** as to the importance of these auxiliary concepts.

Finally, because of **analytic decay**, it is crucial not to view the last step in the analytical process as a **static dead end**, but rather as an invitation to return to the beginning of the process.

5.4 Automated Data Pipelines

In the **service delivery context**, the data analysis process is typically implemented as an **automated data pipeline**, to enable the analysis process to occur repeatedly and automatically.

Data pipelines are usually implemented in 9 components (5 stages and 4 transitions, see Figure 9, on p. 23):

1. data collection
2. data storage
3. data preparation
4. data analysis
5. data presentation

Each of these components must be **designed** and then **implemented**. Typically, at least one pass of the data analysis process has to be done **manually** before the implementation is completed. We will return to this topic in Section 7.

[9]How long does it take Netflix to figure out that you no longer like action movies and want to watch comedies instead, say? How long does it take Facebook to recognize that you and your spouse have separated and that you do not wish to see old pictures of them in your feed?

6. Roles and Responsibilities

> Straight Talk From The Front Lines
>
> To leverage Big Data efficiently, an organization needs business analysts, data scientists, and big data developers and engineers.
>
> – De Mauro, Greco, Grimaldi [27]

A data analyst or a data scientist (in the **singular**) is unlikely to get meaningful results – there are simply too many moving parts to any data project.

Successful projects require **team**s of highly-skilled individuals who understand the data, the context, and the challenges faced by their teammates. Our experience as consultants and data scientists has allowed us to identify the following roles.[10]

Project Managers / Team Leads have to understand the process to the point of being able to recognize whether what is being done makes sense, and to provide realistic estimates of the time and effort required to complete tasks. Team leads act as interpreters between the team and the clients/stakeholders, and advocate for the team.[11] They might not be involved with the day-to-day aspects of the projects but are responsible for the project deliverables.

Domain Experts / SMEs are, quite simply, authorities in a particular area or topic. Not "authority" in the sense that their word is law, but rather, in the sense that they have a comprehensive understanding of the context of the project, either from the client/stakeholder side, or from past experience. SMEs can guide the data science team through the unexpected complications that arise from the disconnect between data science team and the people "on-the-ground", so to speak.

Data Translators have a good grasp on the data and the data dictionary, and help SMEs transmit the underlying context to the data science team.

Data Engineers / Database Specialists work with clients and stakeholders to ensure that the data sources can be used down the line by the data science team. They may participate in the analyses, but do not necessarily specialize in esoteric methods and algorithms. Most data science activities require the transfer of some client data to the analysis team. In many instances, this can be as simple as sending a CSV file as an e-mail attachment. In other instances, there are numerous security and size issues.

Data Scientists are team members who work with the processed data to build sophisticated models that provide actionable insights. They have a sound understanding of algorithms and quantitative methods, and of how they can be applied to a variety of data scenarios. They typically have 2 or 3 areas of expertise and can be counted on to catch up on new material quickly.

Computer Engineers design and build computer systems and other similar devices. They are also involved in software development, which is frequently used to deploy data science solutions.

AI/ML QA/QC Specialists design testing plans for solutions that implement AI/ML models; in particular, they should help the data science team determine whether the models are able to learn.

Communication Specialists are team members who can communicate the actionable insights to managers, policy analysts, decision-makers and other stake holders. They participate in the analyses, but do not necessarily specialize in esoteric methods and algorithms. They should keep on top of popular accounts of quantitative results. They are often data translators, as well.

Data science projects can be downright **stressful**. In an academic environment, the pace is significantly looser, but

- deadlines still exist (exams, assignments, theses),
- work can pile up (multiple courses, TAs, etc.)

In the workplace, there are two major differences:

- a data science project can only receive 1 of 3 "grades": **A+** (exceeded expectations), **A-** (met expectation), or **F** (didn't meet expectations);
- while project quality is crucial, so is **timeliness** – missing a deadline is just as damaging as turning in uninspired or flawed work; perfect work delivered late may cost the client a sizeable amount of money.

Sound **project management** and **scheduling** can help alleviate some of the stress related to these issues. These are the purview of project managers and team leads, as is the maintenance of the quality of **team interactions**, which can make or break a project:

- treat colleagues/clients with respect AT ALL TIMES – that includes emails, Slack conversations, watercooler conversations, meetings, progress reports, etc.;
- keep interactions **cordial** and **friendly** – you do not have to like your teammates, but you are all pulling in the same direction;
- keep the team leader/team abreast of **developments** and **hurdles** – delays may affect the project management plan in a crucial manner (plus your colleagues might be able to offer suggestions), and
- respond to requests and emails in a timely manner (within reason, of course).

[10] Note that individuals can play more than one role on a team.
[11] They may also need to shield the team from clients/stakeholders.

Figure 6. A data science team in action, warts and all [Meko Deng, 2017].

7. Getting Insight From Data

With all of the appropriate context now in mind, we can finally turn to the main attraction, **data analysis** proper.

Let us start this section with a few definitions, in order to distinguish between some of the common categories of data analysis.

What is data analysis? We see **finding patterns in data** as being data analysis's main goal. Alternatively, we could describe the data analysis as **using data to**:

- answer specific questions;
- help in the decision-making process;
- create models of the data;
- describe or explain the situation or system under investigation;
- etc.

While some practitioners include other analytical-like activities, such as testing (scientific) hypotheses, or carrying out calculations on data, we view those as separate activities.

What is data science? One of the challenges of working in the data science field is that nearly all quantitative work can be described as data science (often to a ridiculous extent).

Our simple definition, to paraphrase T. Kwartler, is that data science is the collection of processes by which we extract **useful** and **actionable insights** from data. Robinson [66] further suggests that these insights usually come *via* **visualization** and (manual) **inferential analysis**.

The noted data scientist H. Mason thinks of the discipline as "the **working intersection** of statistics, engineering, computer science, domain expertise, and "hacking" [77])

What is machine learning? Starting in the 1940s, researchers began to take seriously the idea that machines could be taught to **learn**, **adapt** and **respond** to novel situations.

A wide variety of techniques, accompanied by a great deal of theoretical underpinning, were created in an effort to achieve this goal.

Machine learning is typically used in a second stage, to obtain "predictions" (or "advice"), while reducing the operator's analytical, inferential and decisional workload (although it is still present to some extent) [66].

What is artificial/augmented intelligence? The science fiction answer is that artificial intelligence is **non-human intelligence** that has been **engineered** rather than one that has evolved naturally. Practically speaking, this translates to "computers carrying out tasks that only humans can do".

A.I. attempts to remove the need for oversight, allowing for automatic "actions" to be taken by a completely unattended system.

These goals are laudable in an academic setting, but we believe that stakeholders (and humans, in general) should not seek to abdicate all of their agency in the decision-making process; as such, we follow the lead of various thinkers and suggest further splitting A.I. into "**general A.I.**" and "**augmented** intelligence".

7.1 Asking the Right Question

Definitions aside, however, data analysis, data science, machine learning, and artificial intelligence are about **asking questions** and **providing answers** to these questions.

We might ask various types of questions, depending on the situation. Our position is that, from a quantitative perspective, there are only really three types of questions:

- **analytics** questions
- **data science** questions, and
- **quantitative methods** questions.

Analytics questions could be something as simple as:

> how many clicks did a specific link on my website get?

Data science questions tend to be more complex – we might ask something along the lines of:

> if we know, historically, when or how often people click on links, can we predict how many people from Winnipeg will access a specific page on our website within the next three hours?

Whereas analytics-type questions are typically answered by **counting things**, data science-like questions are answered by using historical patterns to **make predictions**.

Quantitative methods questions might, in our view, be answered by making predictions but not necessarily based on historical data. We could build a model from **first principles** – the physics of the situation, as it were – to attempt to figure out what might happen.

For instance, if we thought there was a correlation between the temperature in Winnipeg and whether or not people click on the links in our website, then we might build a model that predicts "how many people from Winnipeg will access a page in the next week?", say, by trying to predict the weather instead.[12]

Analytics models do not usually predict or explain anything – they just **report** on the data, which is itself meant to represent the situation.

A data mining or a data science model tends to be **predictive**, but **not necessarily explanatory** – it shows the existence of connections, of correlations, of links, but without explaining why the connections exist.

In a quantitative method model, we may start by assuming that we know what the links are, what the connections are – which presumably means that we have an idea as to why these connections exist[13] – and then we try to **explore the consequences** of the existence of these connections and these links.

[12] Questions can also be asked in an **unsupervised** manner, see [4, 61], among others, and Section 7.5, briefly.

[13] Unless we're talking about quantum physics – nobody has any idea why things happen the way they do, down there.

We have a piece of advice for new data scientists and analysts, which may prove to be the single most important piece of advice they will receive in their quantitative career:

> **not every situation calls for analytics, data science, statistical analysis, quantitative methods, machine learning, or A.I.**

Take the time to identify instances where more is asked out of the data than what it can actually yield, and be prepared to warn stakeholders, as early as possible, when such a situation is encountered.

If we cannot ask the right questions of the data, of the client, of the situation, and so on, any associated project is doomed to fail from the very beginning.

Without questions to answer, analysts are wasting their time, running analyses for the sake of analysis – **the finish line cannot be reached if there is no finish line**. In order to help clients/stakeholders, analysts need:

- questions **to answer**,
- questions that **can be answered** by the types of methods and skills at their disposal, and
- answers that will be **recognized as answers**.

"How many clicks did this link get?" is a question that is easily answerable if we have a dataset of clicks, but it might not be a question that the client cares to see answered.

Data analysts and scientists often find themselves in a situation where they will ask the types of questions that can be answered with the **available data**, but the answers might not prove actually useful.

From a data science perspective, the right question is one that leads to **actionable insights**. And it might mean that new data has to be collected in order to answer it.

7.2 Structuring and Organizing Data

Let us resume the discussion started in Sections 1.1, 1.2.

Data Sources We cannot have insights from data without data. As with many of the points we have made, this may seem trivially obvious, but there are many aspects of **data acquisition**, **structuring**, and **organization** that have a sizable impact on what insights can be squeezed from data.

More specifically, there are a number of questions that can be considered:

- why do we collect data?
- what can we do with data?
- where does data come from?
- assuming we collect data so we can have a collection of data, what does "a collection" of data look like?
- how can we describe data?
- do we need to distinguish between data, information, knowledge?[14]

[14] According to the adage, "data is not information, information is not knowledge, knowledge is not understanding, understanding is not wisdom." (C.Stoll, attributed).

Historically, data has had three functions:

- **record keeping** – people/societal management (!);
- **science** – new general knowledge, and
- **intelligence** – business, military, police, social (?), domestic (?), personal (!)

Traditionally, each of these functions has

- used different **sources** of information;
- collected **different types of data**, and
- had **different data cultures** and **terminologies**.

Data science is an interdisciplinary field, it should come as no surprise that we may run into all of them on the same project (see Table 2).

Ultimately, data is generated from making observations about and taking measurements of the world. In the process of doing so, we are already imposing particular **conceptualizations** and **assumptions** on our raw experience.

More concretely, data comes from a variety of sources, including:

- records of activity,
- (scientific) observations,
- sensors and monitoring, and,
- more frequently lately, from computers themselves.

As discussed in Section 2, although data may be collected and recorded by handed, it is fast becoming a **mostly digital phenomenon**.

Computer science (and information science) has its own theoretical, **fundamental** viewpoint about data and information, operating over data in a fundamental sense – 1s and 0s that represent numbers, letters, etc. Pragmatically, the resulting data is now stored on computers, and is accessible through our world-wide computer network.

While data is necessarily a representation of **something else**, analysts should endeavour to remember that the data itself still has **physical properties**: it takes up physical space and requires energy to work with.

In keeping with this physical nature, data also has a shelf life – it ages over time. We use the phrase "**rotten data**" or "**decaying data**" in one of two senses:

- **literally**, as the data storage medium might decay, but also
- **metaphorically**, as when it no longer accurately represents the relevant objects and relationships (or even when those objects no longer exist in the same way) – compare with "analytical decay" (see Section 5.3).

Useful data must stay 'fresh' and 'current', and avoid going 'stale' – but that is both **context-** and **model-dependent**!

Before the Data The various data-using disciplines share some **core** (systems) **concepts** and elements, which should resonate with the systems modeling framework previously discussed in Section 3:

- all objects have **attributes**, whether concrete or abstract;
- for multiple objects, there are **relationships** between these objects/attributes, and
- all these elements evolve over time.

The **fundamental relationships** include:

- part–whole;
- is–a;
- is–a–type–of;
- cardinality (one-to-one, one-to-many, many-to-many),
- etc.,

while **object-specific relationships** include:

- ownership;
- social relationship;
- becomes;
- leads-to,
- etc.

Objects and Attributes We can examine concretely the ways in which objects have properties, relationships and behaviours, and how these are captured and turned into data through observations and measurements, *via* the apple and sandwich example of Section 1.1.

There, we **made observations** of an apple instance, **labeled the type of observation** we made, and **provided a value describing** the observation. We can further use these labels when observing other apple instances, and associate new values for these new apple instances.

Regarding the fundamental and object specified relationships, we might be able to see that:

- an apple is a type of fruit,
- a sandwich is part of a meal,
- this apple is owned by Jen,
- this sandwich becomes fuel,
- etc.

It is worth noting that while this all seems tediously obvious to adult humans, it is not so from the perspective of a toddler, or an artificial intelligence. Explicitly, "understanding" requires a basic grasp of:

- categories,
- instances,
- types of attributes,
- values of attributes, and
- which of these are important or relevant to a specific situation or in general terms.

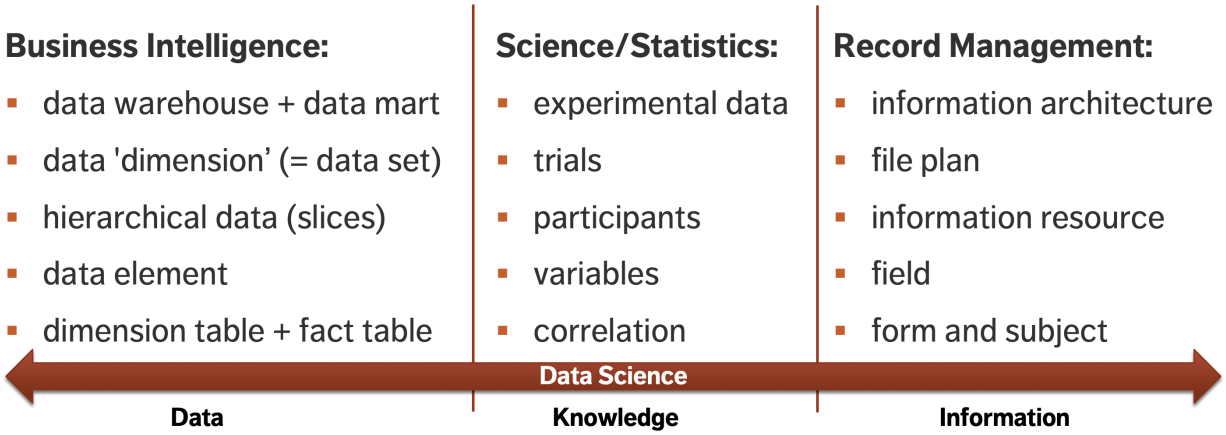


Table 2. Different data cultures and terms.

From Attributes to Datasets Were we to run around in an apple orchard, measuring and jotting down the height, width and colour of 83 different apples completely haphazardly on a piece of paper, the resulting data would be of limited value; although it would technically have been recorded, it would be lacking in **structure**.

We would not be able to tell which values were heights and which were widths, and which colours or which widths were associated with which heights, and *vice-versa*.

Structuring the data using **lists**, **tables**, or even **tree structures** allows analysts to **record** and **preserve** a number of important relationships:

- those between object types and instances, property, attribute types (sometimes also called fields, features or dimensions), and values,
- those between one attribute value and another value (i.e., both of these values are connected to this object instance),
- those between attribute types, in the case of hierarchical data, and
- those between the objects themselves (e.g., this car is owned by this person).

Tables, also called flat files, are likely the most familiar strategy for structuring data in order to preserve and indicate relationships. In the digital age, however, we are developing increasingly sophisticated strategies to store the **structure of relationships** in the data, and finding new ways to work with these increasingly complex relationship structures.

Formally, a **data model** is an abstract (logical) description of both the **dataset structure** and the **system**, constructed in terms that can be implemented in data management software.

In a sense, data models lie halfway between **conceptual models** and **database implementations**. The data proper relates to **instances**; the model to **object types**.

Ontologies provide an alternative representation of the system: simply put, they are **structured**, **machine-readable** collections of **facts** about a domain.[15]

In a sense, an ontology is an attempt to get closer to the level of detail of a full conceptual model, while keeping the whole machine-readable (see Figure 7 for an example).

Every time we move from a conceptual model to a specific type of model (a data model, a knowledge model), we lose some information. One way to preserve as much context as possible in these new models is to also provide rich **metadata** – data about the data!

Metadata is crucial when it comes to successfully working with and across datasets. Ontologies can also play a role here, but that is a topic for another day.

Typically data is stored in a **database**. A major motivator for some of the new developments in types of databases and other data storing strategies is the increasing availability of **unstructured** and (so-called) '**BLOB**' data.

Structured data is labeled, organized, and discrete, with a pre-defined and constrained form. With that definition, for instance, data that is collected *via* an e-form that only uses drop-down menus is structured.

Unstructured data, by comparison, is not organized, and does not appear in a specific pre-defined data structure – the classical example is text in a document. The text may have to subscribe to specific syntactic and semantic rules to be understandable, but in terms of storage (where spelling mistakes and meaning are irrelevant), it is highly unstructured since any data entry is likely to be completely different from another one in terms of length, etc.

The acronym "BLOB" stands for **B**inary **L**arge **O**bject data, such as images, audio files, or general multi-media files. Some of these files can be structured-like (all pictures taken from a single camera, say), but they are usually quite unstructured, especially in multi-media modes.

[15] We could facetiously describe ontologies as "data models on steroids."

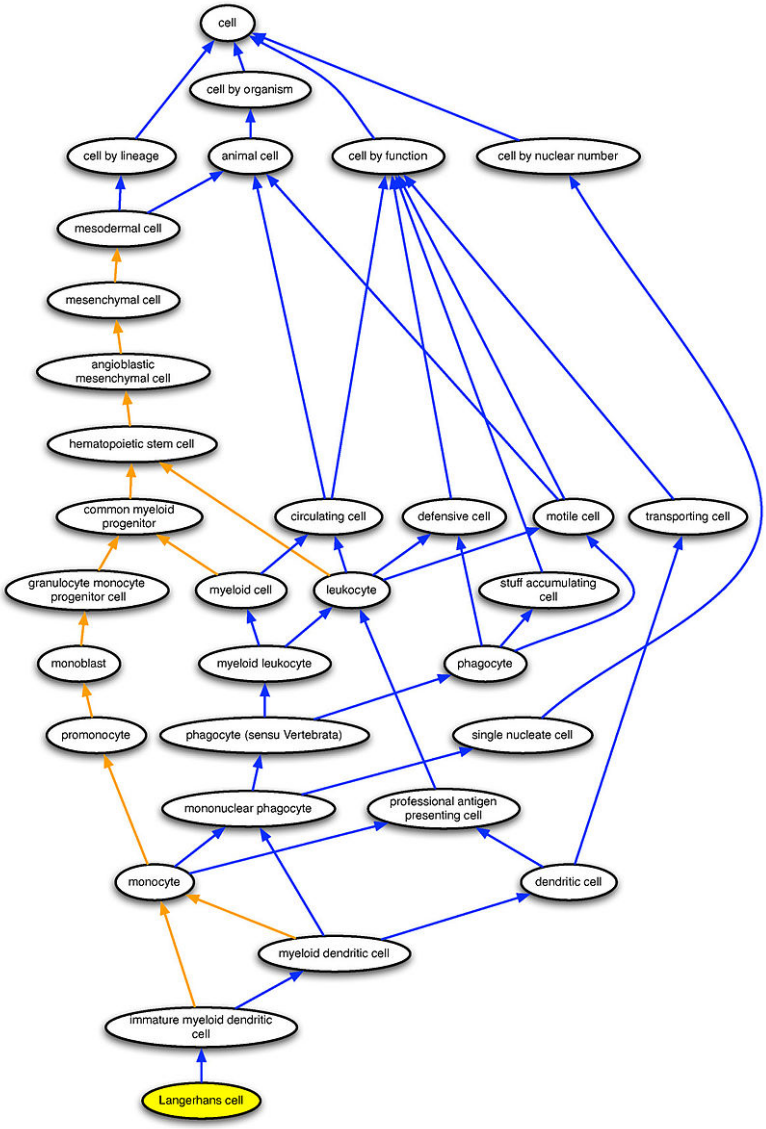


Figure 7. Representation of Langerhans cells in the *Cell Ontology* [52].

Not every type of database is well-suited to all data types.

Let us look at four currently popular database options in terms of fundamental **data and knowledge** modeling and structuring strategies:

- key-value pairs (e.g. JSON);
- triples (e.g. resource description framework – RDF));
- graph databases, and
- relational databases.

Key-Value Stores In these, all data is simply stored as a giant list of keys and values, where the 'key' is a name or a label (possibly of an object) and the 'value' is a value associated with this key.

Triple stores operate on the same principle, but data is stored according to 'subject – predicate – object'.

The following examples illustrate these concepts

1. The **apple type – apple colour** key-value store might contain
   - "Granny Smith – green" and
   - "Red Delicious[16] – red".
2. The **person – shoe size** key-value store might contain
   - "Jen Schellinck – women's size 7", and
   - "Colin Henein – men's size 10".
3. Other key-value stores: **word – definition**, **report name – report (document file)**, **url – webpage**.

[16] Now, there's a misnomer...

4. Triples stores just add a **verb** to the mix: **person – is – age** might contain

   - "Elowyn – is – 18",
   - "Llewellyn – is – 8", and
   - "Gwynneth – is – 4";

   while **object – is-colour – colour** might contain

   - "apple – is-colour – red" and
   - "apple – is-colour – green".

Both strategies results in a large amount of flexibility when it comes to the 'design' of the data storage, and not much needs to be known about the data structure prior to implementation. Additionally, missing values do not take any space in such stores.

In terms of their **implementation**, the devil is in the details; note that their extreme flexibility can also be a flaw [13], and it can be difficult to query them and find the data of interest.

In **graph databases**, the emphasis is placed on the relationships between different **types of objects**, rather than between an object and the properties of that object:

- the objects are represented by **nodes**;
- the relationships between these objects are represented by **edges**, and
- objects can have a relationship with other objects of the same type (such as "person is-a-sibling-of person").

They are fast and intuitive when using relation-based data, and might in fact be the only reasonable option to use in that case as traditional databases may slow to a crawl.

But they are probably too specialized for non relation-based data, and they are not yet widely supported.

In **relational databases**, data is stored in a series of tables. Broadly speaking, each table represents an object and some properties related to this object; special columns in tables connect object instances across table (the entity-relationship model diagram [ERD] of Figure 3 is an example of a relational database model).

For instance, a person lives in a house, which has a particular address. Sometimes that property of the house will be stored in the table that stores information about individuals; in other cases, it will make more sense to store information about the house in its own table.

The form of relational databases are driven by the **cardinality** of the relationships (one-to-one, one-to-many, or many-to-many). These concepts are illustrated in the cheat sheet of Figure 8.

Relational databases are widely supported and well understood, and they work well for many types of systems and use cases.

Note however, that it is difficult to modify them once they have been implemented and that, despite their name, they do not really handle relationships all that well.

We have said very little about keeping data in a single giant table (**spreadsheet**, **flatfile**), or multiple spreadsheets.

On the positive side, spreadsheets are very efficient when working with:

- **static data** (e.g., it is only collected once), or
- data about **one particular type of object** (e.g., scientific studies).

Most implementations of analytical algorithms require the data to be found in **one location** (such as an `R` data frame). Since the data will eventually need to be exported to a flatfile anyway, why not remove the middle step and work with spreadsheets in the first place?

The problem is that it is hard to manage **data integrity** with spreadsheets over the long term when data is collected (and processed) **continuously**. Furthermore, flatfiles are not ideal when working with systems involving many different types of objects and their relationships, and they are not optimzed for querying operations.

For small datasets or quick-and-dirty work, flatfiles are often a reasonable option, but analysts should look for alternatives when working on **large scale projets**.

We have said criminally little on the topic – be aware that projects have **sunk time and time again** when this aspect of the process has not been taken seriously.

Simply put, serious analyses cannot be conducted properly without the **right data infrastructure**.

Implementing a Model In order to **implement** the data/-knowledge model, data engineers and database specialists need access to **data storage** and **management software**.

Gaining this access might be challenging for individuals or small teams as the required software traditionally runs on **servers**.

A server allows multiple users to access the database **simultaneously**, from different client programs. The other side of the coin is that servers make it difficult to 'play' with the database.

User-friendly **embedded database software** (by opposition to client-server database engines) such as SQLite can help overcome some of these obstacles.

Data management software lets human agents interact easily with their data – in a nutshell, they are a **human–data interface**, through which

- data can be **added** to a data collection,
- subsets can be extracted from a data collection based on certain filters/criteria, and
- data can be deleted from (or edited in) a data collection.

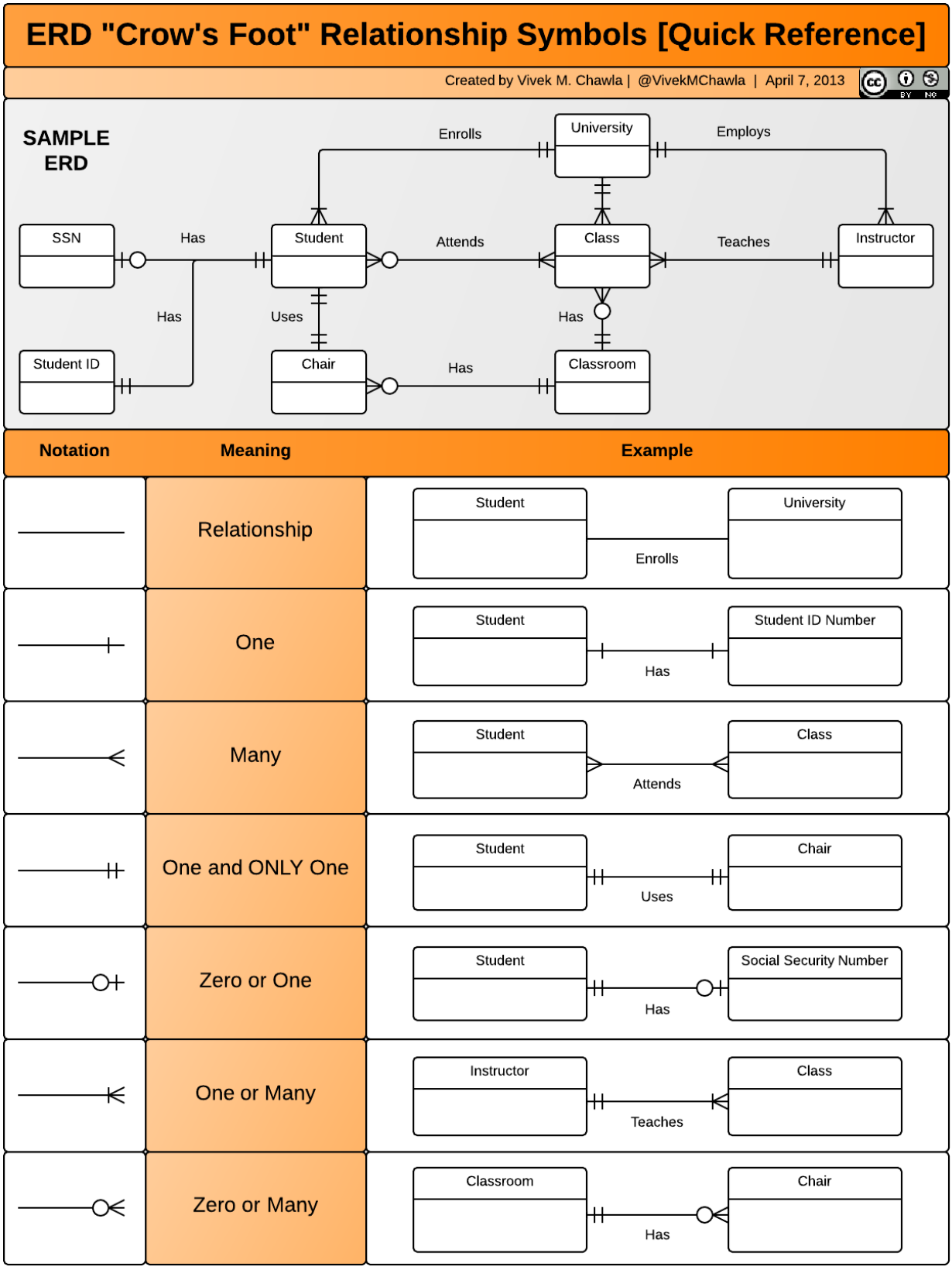


Figure 8. Entity-relationship model diagram "crow's foot" relationship symbols cheat sheet [18].

But *tempora mutantur, nos et mutamur in illis*[17] – whereas we used to speak of:

- databases and database management systems;
- data **warehouses** (data management system designed to enable **analytics**);
- data **marts** (used to retrieve client-facing data, usually oriented to a specific business line or team);
- **Structured Query Language** [SQL] (commonly-used programming language that helps manage (and perform operations on) relational databases),

we now speak of (see [30]):

- data **lakes** (centralized repository in which to store structured/unstructured data alike);
- data **pools** (a small collection of shared data that aspires to be a data lake, someday);
- data **swamps** (unstructured, ungoverned, and out of control data lake in which data is hard to find/use and is consumed out of context, due to a lack of process, standards and governance);
- database **graveyards** (where databases go to die?);
- data is stored in **non-traditional** data structures.[18]

Once a logical data model is complete, we need only:

1. **instantiate** it in the chosen software;
2. **load** the data, and
3. **query** the data.

Traditional relational databases use SQL; other types of databases either use **other query languages** (AQL, semantic engines, etc.) or rely on **bespoke (tailored) computer program**s (e.g. written in R, Python, etc.).

Once a data collection has been created, it must be **managed**, so that the data remains **accurate**, **precise**, **consistent**, and **complete**. **Databases decay**, after all; if a data lake turns into a data swamp, it will be difficult to squeeze usefulness out of it!

Data and Information Architectures There is no single correct structure for a given collection of data (or dataset).

Rather, consideration must be given to the **type of relationships** that exist in the data/system (and are thought to be important), the **types of analysis** that will be carried out, and **data engineering requirements** relating to the time and effort required to extract and work with the data.

The chosen structure, which stores and organizes the data, is called the **data architecture**; designing a specific architecture for a data collection is a necessary part of the data analysis process.

[17] "Times change, and we change with them." *C. Huberinus*

[18] Popular NoSQL database software include: ArangoDB, MongoDB, Redis, Amazon DynamoDB, OrientDB, Azure CosmosDB, Aerospike, etc.

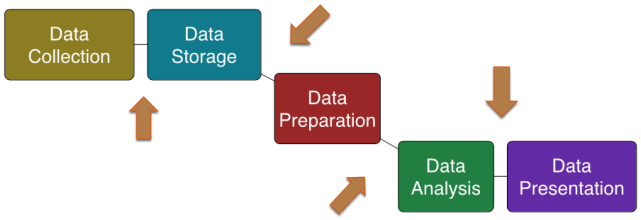


Figure 9. An implemented automated pipeline; note the transitions between the 5 stages.

The data architecture is typically embedded in the larger **data pipeline infrastructure** described in Section 5.

As another example, **automated data pipelines** in the **service delivery context** are usually implemented with 9 components (5 **stages**, and 4 **transitions**, as in Figure 9):

1. data collection
2. data storage
3. data preparation
4. data analysis
5. data presentation

Model validation could be added as a sixth stage, to combat model "drift".

By analogy, the **data storage** component, which houses the data and it's architecture, is the "heart" of the pipeline (the engine that makes the pipeline go), whereas the **data analysis** component is its "brain."[19]

Most analysts are familiar with mathematical and statistical models which are implemented in the data analysis component; **data models** tend to get constructed separately from the analytical models, at the data storage phase.

This separation can be problematic if the analytical model is not compatible with the data model. As an example, if an analyst needs a flatfile (with variables represented as columns) to feed into an algorithm implemented in R, say.

If the data comes from forms with various fields stored in a relational database, the discrepancy could create difficulties on the data preparation side of the process.

Building both the analytical model and the data model off of a **common conceptual model** might help the data science team avoid such quandaries.

In essence, the task is to structure and organize both data and knowledge so that it can be:

- stored in a useful manner;
- added to easily;
- usefully and efficiently extracted from that store (the "**extract-transform-load**" (ETL) paradigm), and
- operated over by humans and computers alike (programs, bots, A.I.) with minimal external modification.

[19] What does that make the other components?

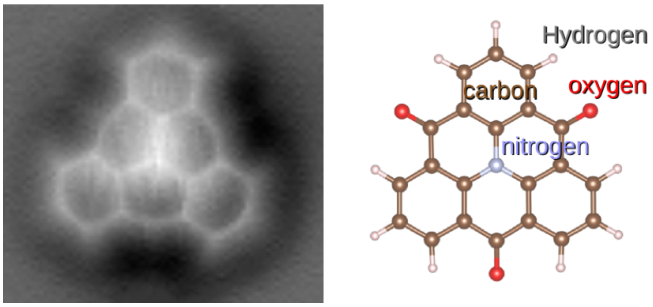


Figure 10. AFM image of 1,5,9-trioxo-13-azatriangulene (left) and its chemical structure model (right) [35].

7.3 Basic Data Analysis Techniques

Business Intelligence (BI) has evolved over the years:

1. we started to recognize that data could be used to **gain a competitive advantage** at the end of the 19th century;
2. the 1950s saw the advent of the first **business database** for decision support;
3. in the 1980s and 1990s, computers and data became increasingly available (**data warehouses**, **data mining**);
4. in the 2000s, the trend was to take business analytics out of the hands of data miners (and other specialists) and into the hands of **domain experts**.
5. Now, **big data** and specialized techniques have arrived on the scene, as have **data visualization**, **dashboards**, and **software-as-service**.

Historically, BI has been one of the streams contributing to modern day data science;

- **system of interest:** the commercial realm, specifically, the market of interest;
- **sources of data:** transaction data, financial data, sales data, organizational data;
- **goals:** provide awareness of competitors, consumers and internal activity and use this to support decision making;
- **culture and preferred techniques:** datamarts, key performance indicators, consumer behaviour, slicing and dicing, business 'facts'.

But no matter the realm in which we work, the ultimate goal remains the same: **obtaining actionable insight into the system of interest**. This can be achieved in a number of ways.

Traditionnally, analysts and consultants hope to do so by seeking:

- **patterns** – predictable, repeating regularities;
- **structure** – the organization of elements in a system, and
- **generalization** – the creation of general or abstract concepts from specific instances (see Figure 10).

The underlying analytical **hope** is to find patterns or structure in the data from which **actionable insight** arise.

While finding patterns and structure can be interesting in its own right (in fact, this is the ultimate reward for many scientists), in the data science context it is how these discoveries are used that trumps all.

Variable Types In the example of a conceptual model shown at the bottom of page 6, we have identified various types of variables; in an **experimental setting**, we typically encounter:

- **control/extraneous variables** – we do our best to keep these controlled and unchanging while other variables are changed;
- **independent variables** – we control their values as we suspect they influence the dependent variables;
- **dependent variables** – we do not control their values; they are generated in some way during the experiment, and presumed dependent on the other factors.

For instance, we could be interested in the **plant height** (dependent) given the **mean number of sunlight hours** (independent), given the **region of the country** in which each test site is located (control).

Data Types These variables need not be of the same **type**. In a typical dataset, we may encounter

- **numerical** data – integers or continuous numbers, such as 1, −7, 34.654, 0.000004, etc.
- **text** data – strings of text, which may be restricted to a certain number of characters, such as "Welcome to the park", "AAAAA", "345", "45.678", etc.
- **categorical** data – are variables with a fixed number of values, may be numeric or represented by strings, but for which there is no specific or inherent ordering, such as ('red','blue','green'), ('1','2','3'), etc.
- **ordinal** data – categorical data with an inherent ordering; unlike **integer** data, the spacing between values is not well-defined; (very cold, cold, tepid, warm, super hot)

We can transform categorical data into numeric data by generating **frequency counts** of the different values/levels of the categorical variable; regular analysis techniques could then be used on the now numeric variable.[20]

House colour	Frequency
red	40
blue	13
green	2

[20] Similar approach underlies most of modern text mining, natural language processing, and categorical anomaly detection. Information gets lost in the process, which explains why meaningful categorical analyses tend to stay fairly simple.

Year	Quarter	Count_Q
2012	1	34
2012	2	12
2012	3	52
2012	4	0
2013	1	21
2013	2	9
2013	3	112
2103	4	8

Year	Count_Y
2012	98
2013	150

Table 3. Nested data, with quarterly granularity.

Categorical data plays a special role in data analysis:

- in data science, categorical variables come with a **pre-defined** set of values;
- in experimental science, a **factor** is an independent variable with its levels being defined (it may also be viewed as a category of treatment)
- in business analytics, these are called **dimensions** (with members).

However they are labeled, these variable can be used to **subset** or **roll up/summarize** the data.

Hierarchical / Nested / Multilevel Data When a categorical variable has multiple levels of abstraction, new categorical variables can be created from these levels. We can view these levels as new categorical variables, in a sense. The 'new' categorical variable has pre-defined relationships with the more detailed level.

This is commonly the case with time and space variables – we can 'zoom' in or out, as needed, which allows us discuss the **granularity** of the data, i.e., the 'maximum zoom factor' of the data.

For instance, observations could be recorded hourly, and then further processed (mean value, total, etc.) at the daily level, the monthly level, the quarterly level, the yearly level, etc., as in Table 3.

Data Summarizing The **summary statistics** of variables can help analysts gain basic **univariate insights** into the dataset (and hopefully, into the system with which it is associated).

These data summaries do not typically provide the full picture and connections/links between different variables are often missed altogether. Still, they often give analysts a **reasonable sense** for the data, at least for **a first pass**.

Signal	Type
4.31	Blue
5.34	Orange
3.79	Blue
5.19	Blue
4.93	Green
5.76	Orange
3.25	Orange
7.12	Orange
2.85	Blue

Count	Signal avg	Signal stdev	Type mode
9	4.73	1.33	Blue/ Orange

Table 4. Artificial dataset, with roll-ups.

Common summary statistics include:

- **min** – smallest value taken by a variable
- **max** – largest value taken by a variable
- **median** – "middle" value taken by a variable
- **mean** – average value taken by a variable
- **mode** – most frequent value taken by a variable
- **# of obs** – number of observations for a variable
- **missing values** – # of missing observations for a variable
- **# of invalid entries** – number of invalid entries for a variable
- **unique values** – unique values taken by a variable
- **quartiles**, **deciles**, **centiles**
- **range**, **variance**, **standard deviation**
- **skew**, **kurtosis**
- **total**, **proportion**, etc.

We can also perform operations over subsets of the data – typically over its columns, in effect **compressing** or '**rolling up**' multiple data values into a single **representative value** (see Table 4 for 4 roll-up summaries).

Typical roll-up functions include the '`mean`', '`sum`', '`count`', and '`variance`', but these do not always give sensical outcomes: if the variable measures a proportion, say, the sum of that variable over all observations is a meaningless quantity, on its own.

We can apply the same roll-up function to many different columns, thus providing a **mapping** (list) of columns to values.

	Large	Medium	Small
Window	1	32	31
Door	14	11	0

Type	Count	Signal avg	Signal stdev
Blue	4	4.04	0.98
Green	1	4.93	N.A.
Orange	4	5.37	1.60

Table 5. Contingency table (top), pivot table (bottom).

Datasets can also be summarized *via* contingency and pivot tables. A **contingency table** is used to examine the relationship between two **categorical** variables – specifically the frequency of one variable relative to a second variable (this is also known as **cross-tabulation**).

A **pivot table**, on the other hand, is a table generated in a software application by applying operations (e.g. '`sum`', '`count`', '`mean`') to variables, possibly based on another (categorical) variable (see Table 5 for examples).[21]

Analysis Through Visualization Consider the broad definition of analysis as:

- identifying patterns or structure, and
- adding meaning to these patterns or structure by **interpreting** them in the context of the system,

There are two general options to achieve this:

1. use analytical methods of varying degrees of sophistication, and/or
2. **visualize** the data and use the brain's analytic (perceptual) power to reach meaningful conclusions about these patterns.

At this point, we will only list some simple visualization methods that are often used to reveal patterns:

- **scatter plots** are best suited for two numeric variables;
- **line charts**, for numeric variable and ordinal variable;
- **bar charts** for one categorical and one numeric, or multiple categorical/nested categorical data;
- **boxplots**, **histograms**, **bubble charts**, **small multiples**, etc.

An in-depth discussion of data visualization, as well as best practices and a more complete catalogues are provided in [12].

[21] Contingency tables are a special intance of pivot tables (where the roll-up function is '`count`').

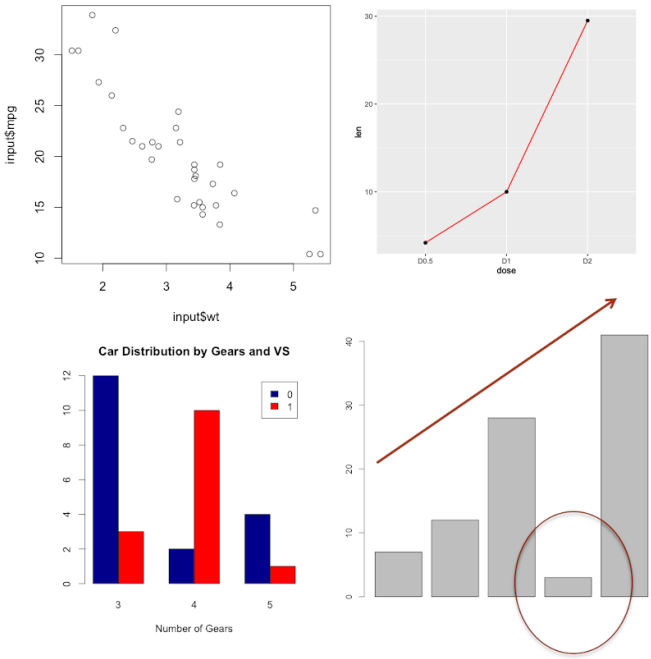


Figure 11. Analysis and pattern-reveal through visualization.

7.4 Statistical Analysis

The underlying reason for **statistical analysis** is to reach an **understanding of the data**.

In a first pass, a variable can be described along 2 dimensions: **centrality** and **spread** (**skew** and **kurtosis** are also sometimes used).

- **Centrality** measures include: **median**, **mean**, **mode** (less frequent);
- Spread (or **dispersion**) measures include: **standard deviation** (sd), **quartiles**, **inter-quartile range** (IQR), **range** (less frequent).

The median, range and the quartiles are easily calculated from an **ordered** list of the data.

Median The **median** of a quantitative variable with n observations is a value which splits the ordered data into 2 equal subsets – half the observations are below (or equal to) the median, and half are above (or equal to) it:

- if n is **odd**, then the median is the $\frac{n+1}{2}$—ordered observation;
- if n is **even**, then the median is any value between the $\frac{n}{2}$ and $\frac{n}{2}+1$ ordered observations (usually their average, but not necessarily so).

The procedure is simple: order the data and follow the even/odd rules to the letter.

As an example, imagine a quantitative variable with $n = 5$ observations, taking the values $4, 6, 1, 3, 7$.

Start by ordering the values: $1,3,4,6,7$. Since $n=5$ is odd, we use the $(n+1)/2=(5+1)/2=3$rd observation, which is 4.[22]

If instead the variable had $n=6$ observations, taking the values $4,6,1,3,7,23$, we again start by ordering the values: $1,3,4,6,7,23$. Since $n=6$ is even, we use any value between the $n/2=6/2=3$rd and the $n/2+1=6/2+1=4$th observations, say 5.2.[23]

Mean The **mean** of a sample is simply the arithmetic average of its observations. For observations $x_1,x_2,\ldots,x_n$, the sample mean is

$$\text{mean}(x_1,\ldots,x_n)=\overline{x}=\frac{x_1+x_2+\ldots+x_n}{n}=\frac{1}{n}\left(\sum_{i=1}^{n}x_i\right)$$

Using the same variable values as in the previous examples, we obtain

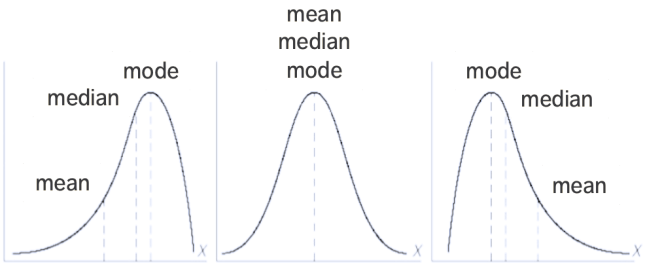
1. $\text{mean}(4,6,1,3,7)=\frac{4+6+1+3+7}{5}=\frac{21}{5}=4.2\approx 4$, which is $\text{median}(4,6,1,3,7)$.
2. $\text{mean}(1,3,4,6,7,23)=\frac{1+3+4+6+7+23}{6}=\frac{44}{6}\approx 7.3$, which is not as close to $\text{median}(1,3,4,6,7,23)=5$.

Mean or Median? We see that the median and the mean can differ. Which measure of centrality should be used to report on the data?

Historically, the mean has been favoured as it is supported, theoretically, by the **Central Limit Theorem**, which we will not be discussing here.

When the data distribution[24] is roughly **symmetric** then both values will be near one another.

But when the data distribution is **skewed**, then the mean is **pulled toward the long tail** and as a result gives a distorted view of the true centre.



Consequently, medians are generally used for house prices, incomes, etc., since that description is **robust** against outliers and incorrect readings, whereas the mean is not.

[22] There are 2 observations below 4 (1,3) and 2 above 4 (6,7).

[23] There are 3 observations below 5.2 (1,3,4) and 3 above 5.2 (6,7,23).

[24] The overall shape taken by the variable's values, taking into account repeated observations. We have left the definition vague by design to avoid getting lost in mathematical foundations, but consult any probability textbook for details.

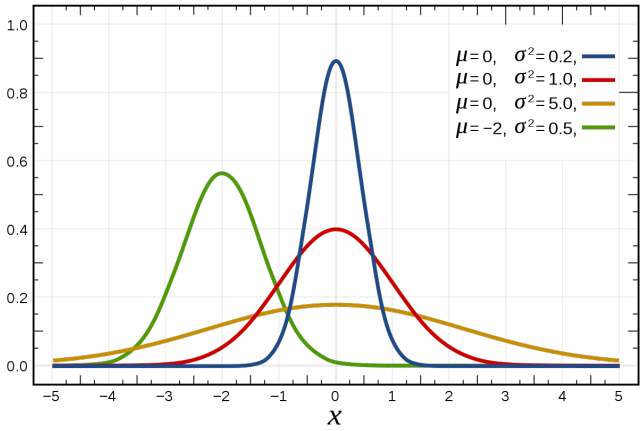


Figure 12. Spread σ of normal distributions [Wikipedia].

Standard Deviation The centrality measures provide an idea as to where the variable's values are "**massed**".

The spread of a distribution, on the other hand, provides an idea as to how disparate the variable observations can be.

The **standard deviation** (sd) is one measure of spread – the higher the value, the more disparate observations tend to be (see Figure 12). It is built from a **fancy average** of the n variable values $(x_1,\ldots,x_n)$. If their mean is μ, then

$$\text{sd}=\sqrt{\frac{(x_1-\mu)^2+\cdots+(x_1-\mu)^2}{n}}.$$

Using the same variable values as in the previous examples, we obtain

$$\text{sd}(4,6,1,3,7)=\sqrt{\frac{(4-4.2)^2+\cdots+(7-4.2)^2}{5}}\approx 2.14$$

and

$$\text{sd}(4,6,1,3,7,23)=\sqrt{\frac{(4-7.3)^2+\cdots+(23-7.3)^2}{6}}\approx 3.98.$$

The second of those is larger, and so we conclude that it has more spread (which is explained by the presence of an outlier, namely, the value 23).

The **variance** of a variable is another measure of dispersion; it is simply the square of the standard deviation.

Quantiles Another way to provide information about the spread of the data is with the help of **centiles**, **deciles**, or **quartiles**.

The **lower quartile** Q_1 of a column with n entries is a numerical value which splits the ordered data into 2 unequal subsets: 25% of the observations are **below** (or at) Q_1, and 75% of the observations are **above** (or at) Q_1.

Similarly, the **upper quartile** Q_3 splits the ordered data into 75% of the observations **below** (or at) Q_3, and 25% of the observations **above** (or at) Q_3.

The median can be interpreted as the **middle quartile** Q_2 of the data, the minimum as Q_0, and the maximum as Q_4; the vector $(Q_0, Q_1, Q_2, Q_3, Q_4)$ represents the **5-point summary** of the data – it is used to describe a variable **at a glance**.[25]

Centiles p_i, where $i = 0, \ldots, 100$, and **deciles** d_j, where $j = 0, \ldots, 10$, run through different **splitting percentages**

$$p_{25} = Q_1,\ p_{75} = Q_3,\ p_{50} = d_5 = Q_2, \text{etc.}$$

The procedure to obtain quantiles is simple:

1. **Sort** the n observations $\{x_1, x_2, \ldots, x_n\}$ in **increasing order** $y_1 \leq y_2 \leq \ldots \leq y_n$. The smallest y_1 has **rank** 1 and the largest y_n has **rank** n.
2. A value between the observations of ranks $\lfloor \frac{\ell n}{4} \rfloor$ and $\lfloor \frac{\ell n}{4} \rfloor + 1$ is a **quartile** Q_ℓ, $\ell = 1, 2, 3$ (if n is odd, use the formulation of the median for Q_2).[26]
3. A value between the observations of ranks $\lfloor \frac{jn}{10} \rfloor$ and $\lfloor \frac{jn}{10} \rfloor + 1$ is a **decile** d_j, $j = 1, \ldots, 9$ (if n is odd, use the formulation of the median for d_5).
4. A value between the observations of ranks $\lfloor \frac{in}{100} \rfloor$ and $\lfloor \frac{in}{100} \rfloor + 1$ is a **centile** p_i, $i = 1, \ldots, 99$ (if n is odd, use the formulation of the median for p_{50}).

In practice, we usually take the average of the observations of rank $\lfloor \frac{kn}{m} \rfloor$ and $\lfloor \frac{kn}{m} \rfloor + 1$ to obtain a unique $m-$**quantile of order** k for the data, where $k = 1, \ldots, m-1$.[27] Using the same variable values as in the previous examples, we obtain

$$Q_1(1,3,4,6,7) = \frac{y_{\lfloor 5/4 \rfloor} + y_{\lfloor 5/4 \rfloor + 1}}{2} = \frac{y_1 + y_2}{2} = \frac{1+3}{2} = 2$$

and

$$\begin{aligned} d_7(1,3,4,6,7,23) &= \frac{y_{\lfloor 7(6)/10 \rfloor} + y_{\lfloor 7(6)/10 \rfloor + 1}}{2} \\ &= \frac{y_4 + y_5}{2} = \frac{6+7}{2} = 6.5. \end{aligned}$$

Skewness When the data distribution is **symmetric**, then the median is equal to the mean, and Q_1 and Q_3 are **equidistant** from the median Q_2:

$$Q_3 - Q_2 \approx Q_2 - Q_1.$$

In general, when $Q_3 - Q_2 \gg Q_2 - Q_1$, the data distribution is said to be **skewed to the right**; when $Q_3 - Q_2 \ll Q_2 - Q_1$, the distribution is **skewed to the left**.[28]

[25] The 5-pt summary is associated with the **boxplot**.

[26] The **floor** function $\lfloor \cdot \rfloor$ returns the largest integer smaller than or equal to the input $\cdot$: $\lfloor 2.5 \rfloor = 2$, $\lfloor -2.5 \rfloor = -3$, $\lfloor 3 \rfloor = 3$.

[27] Remember that the quantiles of order 0 and m are the minimum and the maximum, respectively.

[28] In the illustrations on the previous page, at the bottom of the leftmost column of text, the first distribution is skewed to the right while the third is skewed to the left.

Other Measures In general, we can get a better understanding of a variable by looking at it through the lens of **multiple** descriptive measures.

Other, more exotic measures of **centrality** and **dispersion** are sometimes used:

- centrality: **mid-range** $(\frac{Q_0+Q_4}{2})$; **tri-mean** $(\frac{Q_1+2Q_2+Q_3}{4})$;
- dispersion: **range** $(Q_4 - Q_0)$, **inter-quartile range** $(Q_3 - Q_1)$.

In certain instances, specific values for the mean, variance, quantiles, skewness, etc. are interesting in their own right;[29] in other cases, the interest might arise from the description of the **overall shape of a dataset** that they provide, as this can suggest appropriate analytical and statistical models.

Correlation Consider the following $n = 20$ paired measurements (x_i, y_i) of hydrocarbon levels (x) and pure oxygen levels (y) in fuels:

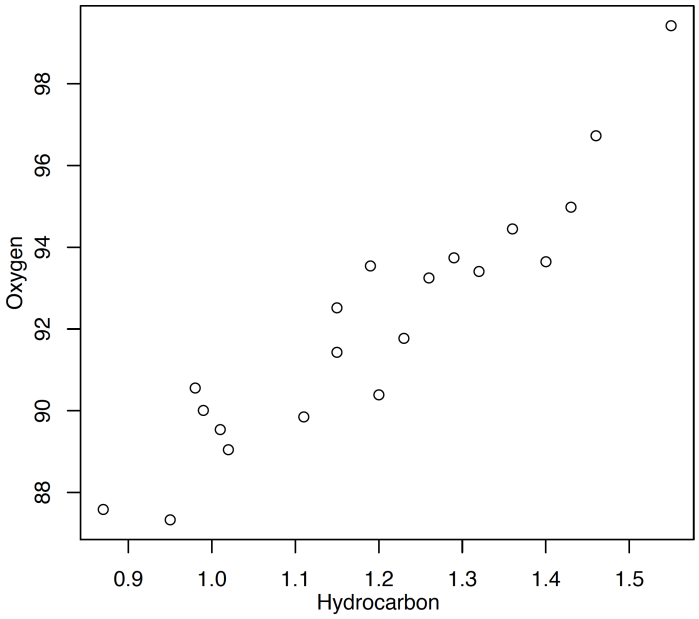
```
x:  0.99  1.02  1.15  1.29  1.46  1.36  0.87  1.23  1.55  1.40
y: 90.01 89.05 91.43 93.74 96.73 94.45 87.59 91.77 99.42 93.65

x:  1.19  1.15  0.98  1.01  1.11  1.20  1.26  1.32  1.43  0.95
y: 93.54 92.52 90.56 89.54 89.85 90.39 93.25 93.41 94.98 87.33
```

In such situations, the **goals** are often to:

- measure the **strength of association** between x and y, and
- **describe** the relationship between x and y.

A graphical display provides an initial description of the relationship.



It appears that the points lie around a **hidden line**!

[29] Polling/election scenarios, for instance.

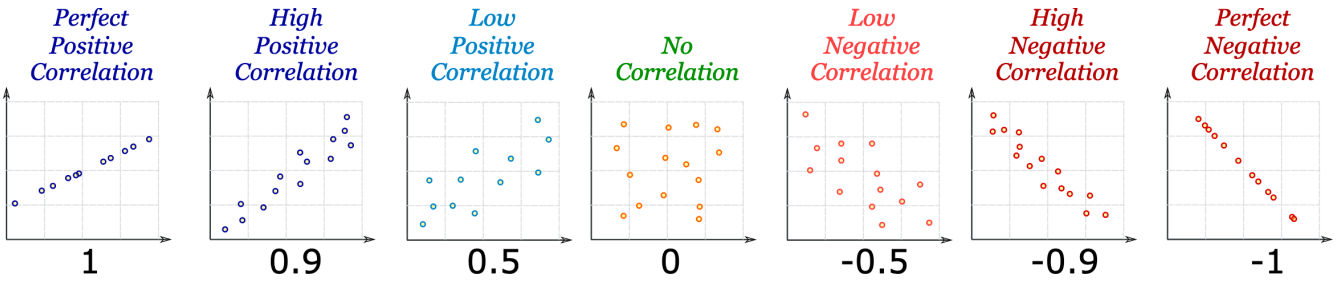


Figure 13. The gamut of correlation values (author unknown).

For paired data (x_i, y_i), $i = 1, \ldots, n$, let $\overline{x}, \overline{y}$ be the respective means of x, y; the **correlation coefficient** of x, y is

$$\rho_{XY} = \frac{\sum_{i=1}^n (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^n (x_i - \overline{x})^2 \sum_{i=1}^n (y_i - \overline{y})^2}} = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}}.$$

The coefficient ρ_{XY} is defined only if $S_{xx} \neq 0$ and $S_{yy} \neq 0$, i.e. neither x_i nor y_i are constant. The variables x and y are **uncorrelated** if $\rho_{XY} = 0$ (or very small, in practice), and **correlated** if $\rho_{XY} \neq 0$ (or $|\rho_{XY}|$ is "large", in practice).

For the hydrocarbon data, we have $S_{xy} \approx 10.18$, $S_{xx} \approx 0.68$, $S_{yy} \approx 173.38$, and $\rho_{XY} \approx \frac{10.18}{\sqrt{0.68 \cdot 173.38}} \approx 0.94$.

The coefficient of correlation has useful properties:

- ρ_{XY} is unaffected by changes of scale or origin. Adding constants to x does not change $x - \overline{x}$ and multiplying x and y by constants changes both the numerator and denominator equally;
- ρ_{XY} is symmetric in x and y (i.e. $\rho_{XY} = \rho_{YX}$) and $-1 \leq \rho_{XY} \leq 1$; if $\rho_{XY} = \pm 1$, then the observations (x_i, y_i) all lie on a straight line with a positive (negative) slope (see Figure 13);
- the sign of ρ_{XY} reflects the trend of the points;
- a high correlation coefficient value $|\rho_{XY}|$ does not necessarily imply a **causal relationship** between the two variables (but see Figure 15);
- note that x and y can have a very strong **non-linear** relationship without ρ_{XY} reflecting it (-0.12 on the left, 0.93 on the right, see Figure 14).

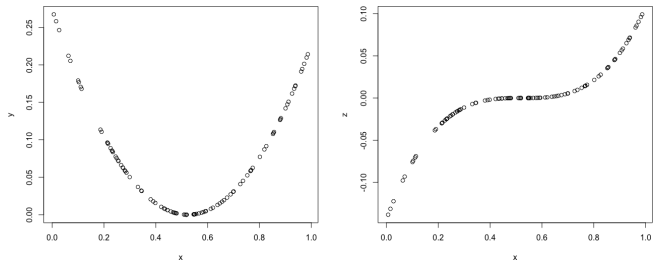


Figure 14. Non-linear relationships.

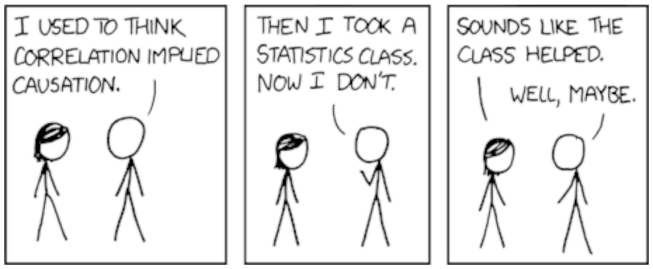


Figure 15. "Correlation doesn't imply causation, but it does waggle its eyebrows suggestively and gesture furtively while mouthing 'look over there'." xkcd.com, #552.

Linear Regression **Regression analysis** can be used to describe the relationship between a **predictor variable** (or regressor) X and a **response variable** Y. Assume that they are related through the model

$$Y = \beta_0 + \beta_1 X + \varepsilon,$$

where ε is a **random error** and β_0, β_1 are the **regression coefficients**.

It is assumed that the mean of the random error is 0,[30] and that the error's variance $\sigma_\varepsilon^2 = \sigma^2$ is constant. Then the model can be re-written as

$$E[Y|X] = \beta_0 + \beta_1 X.$$

Suppose that we have observations (x_i, y_i), $i = 1, \ldots, n$ so that

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, \ldots, n.$$

The aim is to find **estimators** b_0, b_1 of the unknown parameters β_0, β_1, in order to obtain the **estimated (fitted) regression line**

$$\hat{y}_i = b_0 + b_1 x_i$$

The **residual** or error in predicting y_i using $\hat{y}_i$ is thus

$$e_i = y_i - \hat{y}_i = y_i - b_0 - b_1 x_i, \quad i = 1, \ldots, n.$$

How do we find the estimators? One approach is to use the least square framework: find b_0, b_1 so that $\sum_{i=1}^n e_i^2$ is as small as possible.

[30] We usually denote the mean operation with E, so that $E[\epsilon] = 0$.

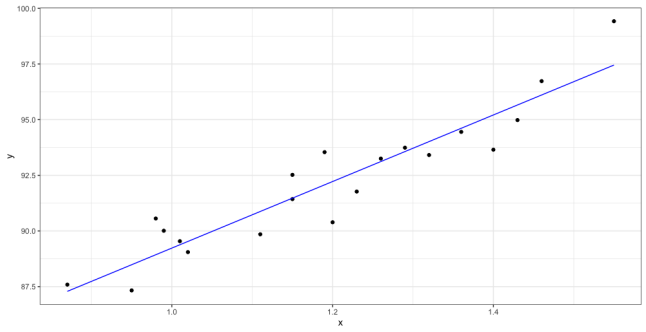


Figure 16. Fitted line for the fuels data: $\hat{y} = 74.28 + 14.95x$.

It is not too difficult to show that the **least square estimators** are

$$b_1 = \frac{S_{xy}}{S_{xx}} \quad \text{and} \quad b_0 = \overline{y} - b_1\overline{x}.$$

For the fuels data, we have already found that

$$S_{xy} \approx 10.18,\ S_{xx} \approx 0.68,\ \text{and}\ S_{yy} = 173.38.$$

Thus, $b_1 = \frac{10.18}{0.68} = 14.95$. Since

$$n = 20,\ \overline{x} = 1.20,\ \text{and}\ \overline{y} = 92.16,$$

we also have $b_0 = 92.16 - 20(1.20) = 74.28$.

Consequently, the **fitted regression line** is

$$\hat{y} = 74.28 + 14.95x.$$

The **exact values** of the estimators or the predictions may or may not be of interest; contextually, perhaps the main take-away is that as the hydrocarbon levels x increase, so do the oxygen levels y, and vice-versa.[31]

A whole slew of questions can be answered using the theoretical apparatus of regression analysis:

- How do we determine if the fitted line is a good model for the data?
- Can we estimate the variance σ^2?
- Can we predict the value of y given a specific x?
- Can we predict likely values of y given a specific x?
- Can we determine if the regression is significant?

The framework can also be extended to include **non-linear** models, **correlated variables**, **probability estimation**, and/or **multivariate** models; any book on statistical analysis contains at least one chapter or two on the topic (see [11, 39], for instance).

We will not pursue the topic further except to say that regression analysis is one of the arrows that every data scientist should have in their quiver.

[31] Remember, this does not necessarily mean that the relationship between the levels is **causal**.

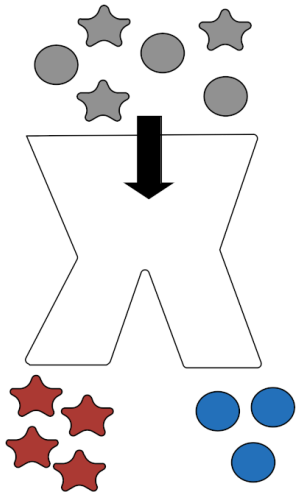


Figure 17. The trousers of classification.

7.5 Quantitative Methods

We provided a list of quantitative methods in Section 5.2; we finish this document by expanding on a few of them.

Classification and Supervised Learning Tasks Classification is one of the cornerstones of machine learning. Instead of trying to predict the numerical value of a response variable (as in regression), a **classifier** uses **historical data**[32] to identify general patterns that could lead to observations belonging to one of several **pre-defined categories**.

For instance, if a car insurance company only has resources to investigate up to 20% of all filed claims, it could be useful for them to predict:

- whether a claim is likely to be fraudulent?
- whether a customer is likely to commit fraud in the near future?
- whether an application for a policy is likely to result in a fraudulent claim?
- the amount by which a claim will be reduced if it is fraudulent?

Analysts and machine learners use a variety of different techniques to carry this process out (see Figure 17 and [2, 3, 36]), but the steps are always the same:

1. use **training data** to teach the classifier;
2. test/validate the classifier using **hold-out** data;
3. if it passes the test, use the classifier to classify **novel instances**.

Some classifiers (such as deep learning neural nets) are '**black boxes**': they might be very good at classification, but they are not **explainable**.

In some instances, that is an acceptable side effect of the process, in others, not so much – if an individual is refused refugee status, say, they might rightly want to know **why**.

[32] This training data usually consists of a **randomly** selected subset of the **labeled** (response) data.

Unsupervised Learning Techniques The hope of artificial intelligence is that intelligent behaviours will eventually be able to be **automated**. For the time being, however, that is still very much a work in progress.

But one of the challenges in that process is that not every intelligent behaviour arises from a supervised process.

Classification, for instance, is the prototypical supervised task: can we learn from historical/training examples? It seems like a decent approach to learning: evidence should drive the process.

There are limitations: it is difficult to make a **conceptual leap** solely on the basis of training data,[33] if only because the training data data might not be representative of the system, or because the learner target task is **too narrow**.

In **unsupervised** learning, we learn without examples, based solely on what is found in the data. There is no specific question to answer (in the classification sense), other than: what can we learn from the data? Typical unsupervised learning tasks include:

- **clustering** (novel categories);
- **association rules mining**;
- **recommender systems**, etc.

For instance, an online bookstore might want to make recommendations to customers concerning additional items to browse (and hopefully purchase) based on their buying patterns in prior transactions, the similarity between books, and the similarity between **customer segments**.

- But what are those patterns?
- How do we measure similarity?
- What are the customer segments?
- Can any of that information be used to create promotional bundles?

The lack of a specific target makes unsupervised learning much more **difficult** than supervised learning, as does the challenges of **validating the results**.

This contributes to the proliferation of clustering algorithms and cluster quality metrics [3, 4, 76].

Other Machine Learning Tasks Of course, this scratches but a **miniscule** part part of the machine learning ecosystem. Other common tasks include [61]:

- profiling and behaviour description;
- link prediction;
- data reduction;
- influence/causal modeling, etc.

to say nothing of more sophisticated learning frameworks (semi-supervised, reinforcement [71], deep learning [32], etc.).

[33] If our teaching experience is anything to go by...

Time Series Analysis and Process Monitoring Processes are often subject to **variability**:

- variability due the **cumulative effect** of many small, essentially unavoidable causes (a process that only operates with such **common causes** is said to be **in (statistical) control**;
- variability due to **special causes**, such as improperly adjusted machines, poorly trained operators, defective materials, etc. (the variability is typically much larger for special causes, and such processes are said to be **out of (statistical) control**.

The aim of **statistical process monitoring** (SPM) is to identify occurrence of special causes. This is often done *via* **time series analysis**.

Consider n observations $\{x_1, \ldots, x_n\}$ arising from some collection of processes. In practice, the index i is often a **time index** or a **location index**, i.e., the x_i are observed in **sequence** or in **regions**.[34]

The processes that generate the observations could change from one time/location to the next due to:

- **external factors** (war, pandemic, regime change, election results, etc.), or
- **internal factors** (policy changes, modification of manufacturing process, etc.).

The mean and standard deviation might not provide a useful summary of the situation.

To get a sense of what is going on with the data (and the associated system), it could prove preferable to **plot the data** in the **order that it has been collected** (or according to geographical regions, or both).

The horizontal coordinate represents:

- the **time of collection** t (order, day, week, quarter, year, etc.), or
- the **location** i (country, province, city, branch, etc.).

The vertical coordinate represents the observations of interest x_t or x_i (see Figure 19 for an example).

In process monitoring terms, we may be able to identify potential special causes by identifying **trend breaks**, **cycles discontinuities**, or **level shifts** in time series.

For instance, consider the three time series of Figure 18. Is any action necessary?

In the first example (left), there are occasional drops in sales from one year to the next, but the **upward trend** clear is clear. We see the importance of considering the full time series; if only the last two points are presented to stockholders, say, they might conclude that action is needed, whereas the whole series paints a more positive outlook.

[34] In the first situation, the observations form a **time series**.

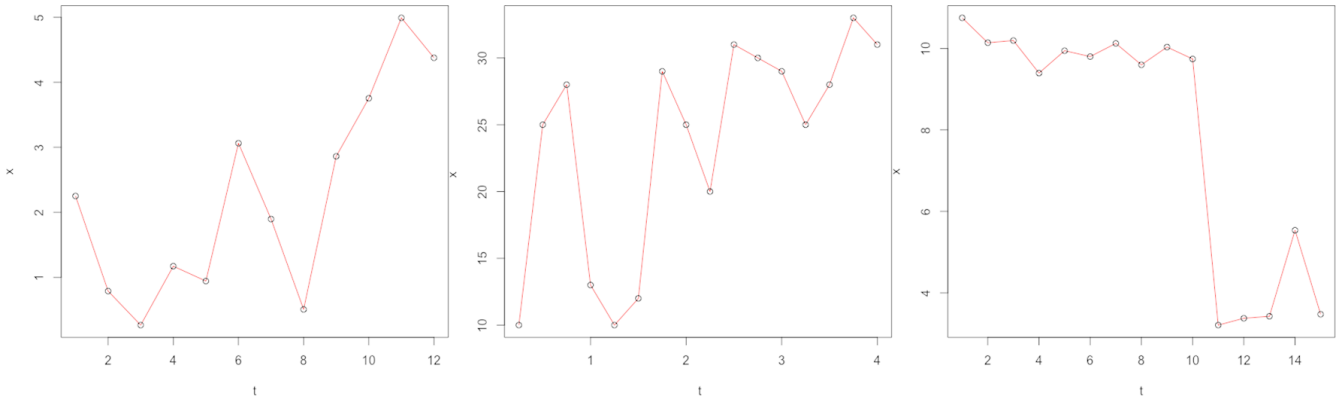


Figure 18. Sales (in 10,000$) for 3 different products – years (left), quarters (middle), weeks (right)

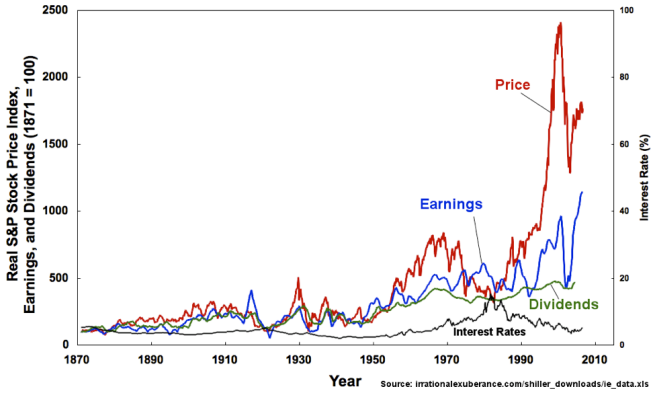


Figure 19. Real S&P stock price index (red), earnings (blue), and dividends (green), together with interest rates (black), from 1871 to 2009.

In the second case (middle), there is a **cyclic effect** with increases from Q1 to Q2 and from Q2 to Q3, but decreases from Q3 to Q4 and from Q4 to Q1. Overall, we also see an upward trend. The presence of regular patterns is a positive development.

Finally, in the last example (right), something clearly happened after the tenth week, causing a **trend level shift**. Whether it is due to internal or external factors depends on the context, which we do not have at our disposal, but some action certainly seems to be needed.

We might also be interested in using historical data to **forecast** the future behaviour of the variable.

This is similar to the familiar analysis goals of:

- **finding patterns** in the data, and
- **creating a (mathematical) model** that captures the essence of these patterns.

Time series patterns can be quite complex and must often be **broken down** into multiple component models (trend, seasonal, irregular, etc.).

Typically, this can be achieved with fancy analysis methods, but it is not a simple topic, in general. Thankfully, there are software libraries that can help.

Anomaly Detection The special points from process monitoring are anomalous in the sense that something unexpected happens there, something that changes the nature of the data pre- and post-break.

In a more general context, **anomalous observations** are those that are **atypical** or **unlikely**.

From an analytical perspective, anomaly detection can be approached using supervised, unsupervised, or conventional statistical methods.

The discipline is rich and vibrant (and the search for anomalies can end up being an arms race against the "bad guys"), but it is definitely one for which analysts should heed contextual understanding – blind analysis leads to blind alleys![35]

References

[1] ACM Code of Ethics and Professional Conduct. Association for Computing Machinery. Accessed: June 18, 2017.

[2] C. Aggarwal, editor. *Data Classification: Algorithms and Applications*. CRC Press, 2015.

[3] C. C. Aggarwal. *Data Mining: The Textbook*. Springer, Cham, 2015.

[4] C. C. Aggarwal and C. K. Reddy, editors. *Data Clustering: Algorithms and Applications*. CRC Press, 2014.

[5] I. Asimov. *Foundation Series*. Gnome Press, Spectra, Doubleday, 1942–1993.

[6] F. R. Bach and M. I. Jordan. Learning spectral clustering, with application to speech separation. *J. Mach. Learn. Res.*, 7:1963–2001, Dec. 2006.

[7] Facebook documents seized by MPs investigating privacy breach. BBC News, Nov 2018.

[8] BeauHD. Google AI Claims 99 Percent Accuracy In Metastatic Breast Cancer Detection. *Slashdot.com*, Oct 2018.

[35] A more thorough treatment is provided in [20].

[9] E. Betuel. Math Model Determines Who Wrote Beatles' "In My Life": Lennon or McCartney? *Inverse*, Jul 2018.

[10] N. Bien, P. Rajpurkar, R. L. Ball, J. Irvin, A. Park, E. Jones, M. Bereket, B. N. Patel, K. W. Yeom, K. Shpanskaya, S. Halabi, E. Zucker, G. Fanton, D. F. Amanatullah, C. F. Beaulieu, G. M. Riley, R. J. Stewart, F. G. Blankenberg, D. B. Larson, R. H. Jones, C. P. Langlotz, A. Y. Ng, and M. P. Lungren. Deep-learning-assisted diagnosis for knee magnetic resonance imaging: Development and retrospective validation of mrnet. *PLOS Medicine*, 15(11):1–19, 2018.

[11] P. Boily. MAT2377 - Probability and Statistics for Engineers . Course website.

[12] P. Boily, S. Davies, and J. Schellinck. *Practical Data Visualization*. Data Action Lab/Quadrangle, 2021.

[13] boot4life. What JSON structure to use for key-value pairs . StackOverflow, Jun 2016.

[14] D. Brin. *The Transparent Society: Will Technology Force Us to Choose Between Privacy and Freedom?* Perseus, 1998.

[15] S. E. Brossette, A. P. Sprague, J. M. Hardin, K. B. Waites, W. T. Jones, and S. A. Moser. Association Rules and Data Mining in Hospital Infection Control and Public Health Surveillance. *Journal of the American Medical Informatics Association*, 5(4):373–381, 07 1998.

[16] Centre for Big Data Ethics, Law, and Policy . Data Science Institute, University of Virginia. Accessed: June 18, 2017.

[17] Code of Ethics/Conducts . Certified Analytics Professional. Accessed: June 17, 2017.

[18] V. Chawla. ERD "Crow's Foot" Relationship Symbols Cheat Sheet , 2013.

[19] M. Chen. Is 'Big Data' Actually Reinforcing Social Inequalities? *The Nation*, Sep 2013.

[20] Y. Cissokho, S. Fadel, R. Millson, R. Pourhasan, and P. Boily. Anomaly Detection and Outlier Analysis . *Data Science Report Series*, 2020.

[21] N. Cohn. How One 19-Year-Old Illinois Man is Distorting National Polling Averages . *The Upshot* , 2016.

[22] Columbia University Irving Medical Center. Data Scientists Find Connections Between Birth Month and Health . *Newswire.com*, Jun 2015.

[23] J. Corey. *The Expanse*. Orbit Books, 2011–.

[24] J. Cranshaw, R. Schwartz, J. I. Hong, and N. M. Sadeh. The livehoods project: Utilizing social media to understand the dynamics of a city. In J. G. Breslin, N. B. Ellison, J. G. Shanahan, and Z. Tufekci, editors, *ICWSM*. The AAAI Press, 2012.

[25] J. Dastin. Amazon scraps secret AI recruiting tool that showed bias against women . *Reuters*, Oct 2018.

[26] T. Davenport and D. Patil. Data Scientist: The Sexiest Job of the 21st Century . *Harvard Business Review*, Oct 2012.

[27] A. De Mauro, M. Greco, and M. Grimaldi. A formal definition of big data based on its essential features. *Library Review*, 65(3):122–135, 2016.

[28] Cognitive Biases . The Decision Lab. Accessed: Sep 3, 2021.

[29] L. Donnelly. Robots are better than doctors at diagnosing some cancers, major study finds. *The Telegraph*, May 2018.

[30] N. Feldman. Data Lake or Data Swamp? , July 2015.

[31] K. Fung. The Ethics Conversation We're Not Having About Data . *Harvard Business Review*, Nov 2015.

[32] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT press Cambridge, 2016.

[33] A. Gumbus and F. Grodzinsky. Era of big data: danger of descrimination . *ACM SIGCAS Computers and Society*, 45(3):118–125, 2015.

[34] K. Hao. We tried teaching an AI to write Christmas movie plots. Hilarity ensued. Eventually . *MIT Technology Review*, Dec 2018.

[35] P. Hapala, M. Svec, O. Stetsovych, N. J. Van Der Heijden, M. Ondracek, J. Van Der Lit, P. Mutombo, I. Swart, and P. Jelinek. Mapping the electrostatic force field of single molecules from high-resolution scanning probe images. *Nature Communications*, 7(11560), 2016.

[36] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd ed.* Springer, 2008.

[37] Henning (WMDE). UML diagram of the Wikibase Data Model . Wikimedia.

[38] J. Hiner. How big data will solve your email problem . *ZDNet*, Oct 2013.

[39] R. Hogg and E. Tanis. *Probability andd Statistical Inference*. Pearson/Prentice Hall, 7 edition, 2006.

[40] K.-W. Hsu, N. Pathak, J. Srivastava, G. Tschida, and E. Bjorklund. *Data Mining Based Tax Audit Selection: A Case Study of a Pilot Project at the Minnesota Department of Revenue*, pages 221–245. Springer International Publishing, Cham, 2015.

[41] Indiana University. Scientists use Instagram data to forecast top models at New York Fashion Week . *Science Daily*, Sep 2015.

[42] A. Jensen, P. Moseley, T. Oprea, S. Ellesøe, R. Eriksson, H. Schmock, P. Jensen, L. Jensen, and S. Brunak. Temporal disease trajectories condensed from population-wide registry data covering 6.2 million patients. *Nature Communications*, 5, 2014.

[43] M. Jing. AlphaGo vanquishes world's top Go player, marking A.I.'s superiority over human mind . *South China Morning Post*, May 2017.

[44] I. Johnston. AI robots learning racism, sexism and other prejudices from humans, study finds . *The Independent*, Apr 2017.

[45] M. Judge. Facial-Recognition Technology Affects African Americans More Often . *The Root*, 2016.

[46] M. Kosinski and Y. Wang. Deep neural networks are more accurate than humans at detecting sexual orientation from facial images. *Journal of Personality and Social Psychology*, 114(2):246–257, Feb 2018.

[47] H. T. Kung and D. Vlah. A spectral clustering approach to validating sensors via their peers in distributed sensor networks. *Int. J. Sen. Netw.*, 8(3/4):202–208, Oct. 2010.

[48] S. L. Lee and D. Baer. 20 Cognitive Biases That Screw Up Your Decisions . *Business Insider*, Dec 2015.

[49] D. Lewis. An AI-Written Novella Almost Won a Literary Prize . *Smithsonian Magazine*, Mar 2016.

[50] Scientists Using GPS Tracking on Endangered Dhole Wild Dogs . Live View GPS, Oct 2018.

[51] E. Mack. Elon Musk: Artificial intelligence may spark World War III . *CNET*, Sep 2017.

[52] A. Masci, C. Arighi, A. Diehl, A. Lieberman, C. Mungall, R. Scheuermann, B. Smith, and L. Cowell. An improved ontological representation of dendritic cells as a paradigm for all cell types . *BMC Bioinformatics*, 2009.

[53] R. Mérou. Conceptual map of Free Software . Wikimedia, 2010.

[54] C. O'Neil. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy* . Crown, 2016.

[55] Open Data . Wikipedia. Accessed: June 19, 2017.

[56] Open Up Guide: Using Open Data to Combat Corruption . Open Data Charter, 2017. Accessed: June 20, 2017.

[57] V. U. Panchami and N. Radhika. A novel approach for predicting the length of hospital stay with dbscan and supervised classification algorithms. In *ICADIWT*, pages 207–212. IEEE, 2014.

[58] A Conversation with Julie Paquette: Ethics in Quantitative Contexts . Paquette, J. and Boily, P.

[59] R. Paul and L. Elder. *Understanding the Foundations of Ethical Reasoning* . Foundation for Critical Thinking, 2 edition, 2006.

[60] C. Plant, S. J. Teipel, A. Oswald, C. Böhm, T. Meindl, J. M. Miranda, A. L. W. Bokde, H. Hampel, and M. Ewers. Automated detection of brain atrophy patterns based on mri for the prediction of alzheimer's disease. *NeuroImage*, 50(1):162–174, 2010.

[61] F. Provost and T. Fawcett. *Data Science for Business*. O'Reilly, 2015.

[62] S. Ramachandran and J. Flint. At Netflix, who wins when it's Hollywood vs. the algorithm? *Wall Street Journal*, Nov 2018.

[63] S. Reichman. These ai-invented paint color names are so bad, they're good. *Curbed*, May 2017.

[64] Research integrity & ethics. Memorial University of Newfoundland.

[65] T. Rikert. A.I. hype has peaked so what's next? *TechCrunch*, Sep 2017.

[66] D. Robinson. What's the difference between data science, machine learning, and artificial intelligence? *Variance Explained*, Jan 2018.

[67] R. Schutt and C. O'Neill. *Doing Data Science: Straight Talk From the Front Line*. O'Reilly, 2013.

[68] J. C. Scott. *Against the Grain: A Deep History of the Earliest States*. Yale University Press, New Haven, 2017.

[69] B. Smith. Artificial intelligence better than physicists at designing quantum science experiments . *ABC Science*, Oct 2018.

[70] I. Stewart. The Fourth Law of Humanics . *Nature*, 535, 2016.

[71] R. Sutton and G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, 2018.

[72] J. Taylor. Four Problems in Using CRISP-DM and How To Fix Them . *KDnuggets.com*, 2017.

[73] Development of National Statistical Systems . United Nations, Statistics Division. Accessed: June 17, 2017.

[74] A. Van Dam. This researcher studied 400,000 knitters and discovered what turns a hobby into a business . *Washington Post*, Nov 2018.

[75] D. Wakabayashi. Firm led by Google veterans uses A.I. to 'nudge' workers toward happiness . *New York Times*, 12 2018.

[76] Wikipedia. Cluster Analysis Algorithms .

[77] D. Woods. bitly's Hilary Mason on "What is A Data Scientist?" . *Forbes*, Mar 2012.

[78] Wooptoo. Entity - Relationship Model . Wikimedia.

[79] E. Yong. Wait, have we really wiped out 60% of animals? *The Atlantic*, 10 2018.