Data Science Basics

DATA SCIENCE ESSENTIALS

2. Preliminaries

The Digital/Analog Data Dichotomy

Humans have been collecting data for a long time; J.C. Scott argues that data collection was a major enabler of the modern nation-state.

For most of the history of data collection, we have lived in the **analogue world** (understanding grounded in continuous experience of **physical reality**).

Our data collection activities were the first steps towards a different strategy for understanding and interacting with the world.

Data leads us to conceptualize the world in a way that is **more discrete than continuous**.

The Digital/Analog Data Dichotomy

Translating our experiences into numbers and categories, we create **sharper** and more definable boundaries than our raw experience might suggest.

This discretization strategy leads to the **digital computer** (series of 1s and 0s), which is surprisingly successful at representing our physical world: the **digital world** is taking on a reality as pervasive and important as the physical one.

This digital world is built on top of the physical world, but it **does not operate under the same set of rules**.

- in the physical world, the default is to **forget**; in the digital world, it 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**

The Digital/Analog Data Dichotomy

Digitization is making things that were **once hidden, visible; once veiled, transparent**.

Data scientists are scientists of the **digital world**. They seek to understand:

- the **fundamental principles of data**
- how these fundamental principles manifest themselves in different digital phenomena

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

What is Data?

It is difficult to give a clear-cut definition of **data** (is it singular or plural?).

Linguistically, a *datum* is “a piece of information”, so **data** means “pieces of information,” or **collection** of “pieces of information”.

Data represents the whole (potentially greater than the sum of its parts) or simply the idealized concept.

Is that clear?

What is Data?

Is the following data?

4,529 red 25.782 Y

Why? Why not? What, if anything is missing?

The Stewart approach: “we know it when we see it.”

Pragmatically, we think of data as a collection of facts about **objects** and their **attributes**.

Objects and 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

Remember: an object is not simply **the sum of its attributes**.

Objects and Attributes

Ambiguities when it comes to **measuring** (and **recording**) the attributes:

- apple picture is a 2-dimensional representation of a 3-dimensional object
- overall shape of the sandwich is vaguely rectangular, it is not exact (**measurement error**?)
- insignificant for most, but not necessarily all, analytical purposes
- apple's shape = volume, sandwich's shape = area (**incompatible measurements)**
- a number of potential attributes are not mentioned: size, weight, time, etc.
- are there other issues?

Measurement errors and incomplete lists are always part of the picture; is this collection of attributes providing a reasonable **description** of the objects?

[R. Magritte, La trahison des images (ceci n’est pas une pipe)]

From Objects and Attributes to Datasets

Raw data may exist in any format.

A **dataset** represents a collection of data that could conceivably be fed into algorithms for analytical purposes.

Datasets appear in a **table** format, with rows and columns; attributes are the **fields** (or columns, variables); objects are **instances** (or cases, rows, records).

Objects are described by their **feature vector** (observation's signature) – the collection of attributes associated with value(s) of interest.

From Objects and Attributes to Datasets

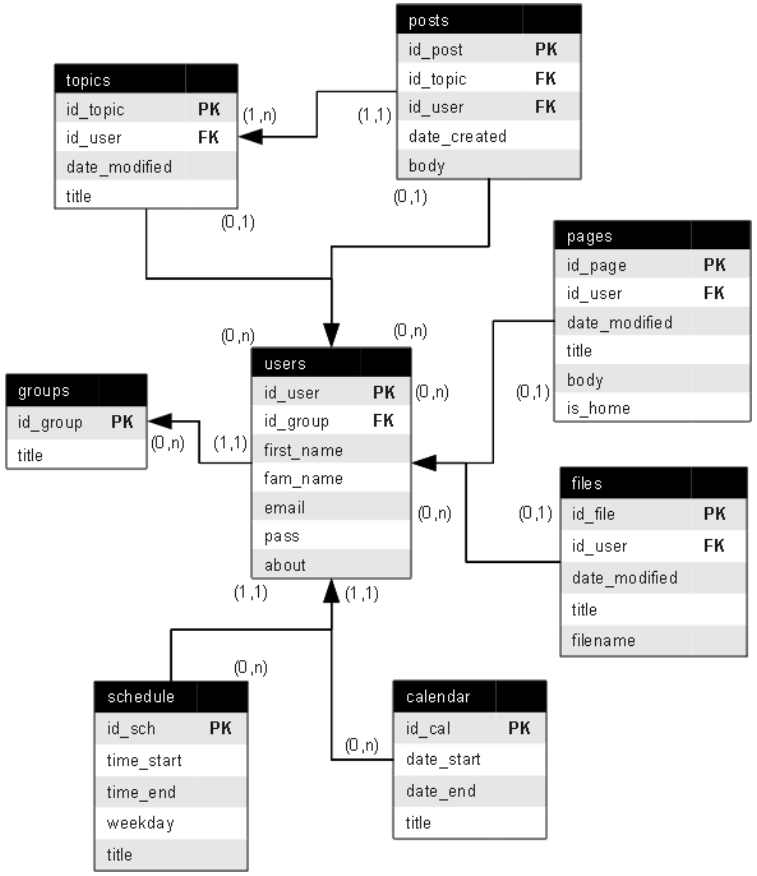
The dataset of physical objects could start with:

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
...	...	...	...	...	...

[Wooptoo, "Entity - relationship model." Wikimedia]

From Objects and Attributes to Data

In practice, more complex **databases** are used, for a variety of reasons that we briefly discuss at a later stage.



Data in the News

Here is a sample of headlines and article titles showcasing the growing role of **data science** (DS), **machine learning** (ML), and **artificial/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 social consequences (not always positive).

Data in the News

- “Robots are better than doctors at diagnosing some cancers, major study finds”
- “Deep-learning-assisted diagnosis for knee magnetic resonance imaging: Development and retrospective validation of MRNet”
- “Google AI claims 99% accuracy in metastatic breast cancer detection”
- “Data scientists find connections between birth month and health”
- “Scientists using GPS tracking on endangered Dhole wild dogs”
- “These AI-invented paint color names are so bad they’re good”
- “We tried teaching an AI to write Christmas movie plots. Hilarity ensued. Eventually.”
- “Math model determines who wrote Beatles’ "In My Life": Lennon or McCartney?”

Data in the News

- “Scientists use Instagram data to forecast top models at New York Fashion Week”
- “How big data will solve your email problem”
- “Artificial intelligence better than physicists at designing quantum science experiments”
- “This researcher studied 400,000 knitters and discovered what turns a hobby into a business”
- “Wait, have we really wiped out 60% of animals?”
- “Amazon scraps secret AI recruiting tool that showed bias against women”
- “Facebook documents seized by MPs investigating privacy breach”
- “Firm led by Google veterans uses A.I. to ‘nudge’ workers toward happiness”
- “At Netflix, who wins when it’s Hollywood vs.the algorithm?”

Data in the News

- "AlphaGo vanquishes world's top Go player, marking A.I.'s superiority over human mind"
- "An AI-written novella almost won a literary prize"
- "Elon Musk: Artificial intelligence may spark World War III"
- "A.I. hype has peaked so what's next?"

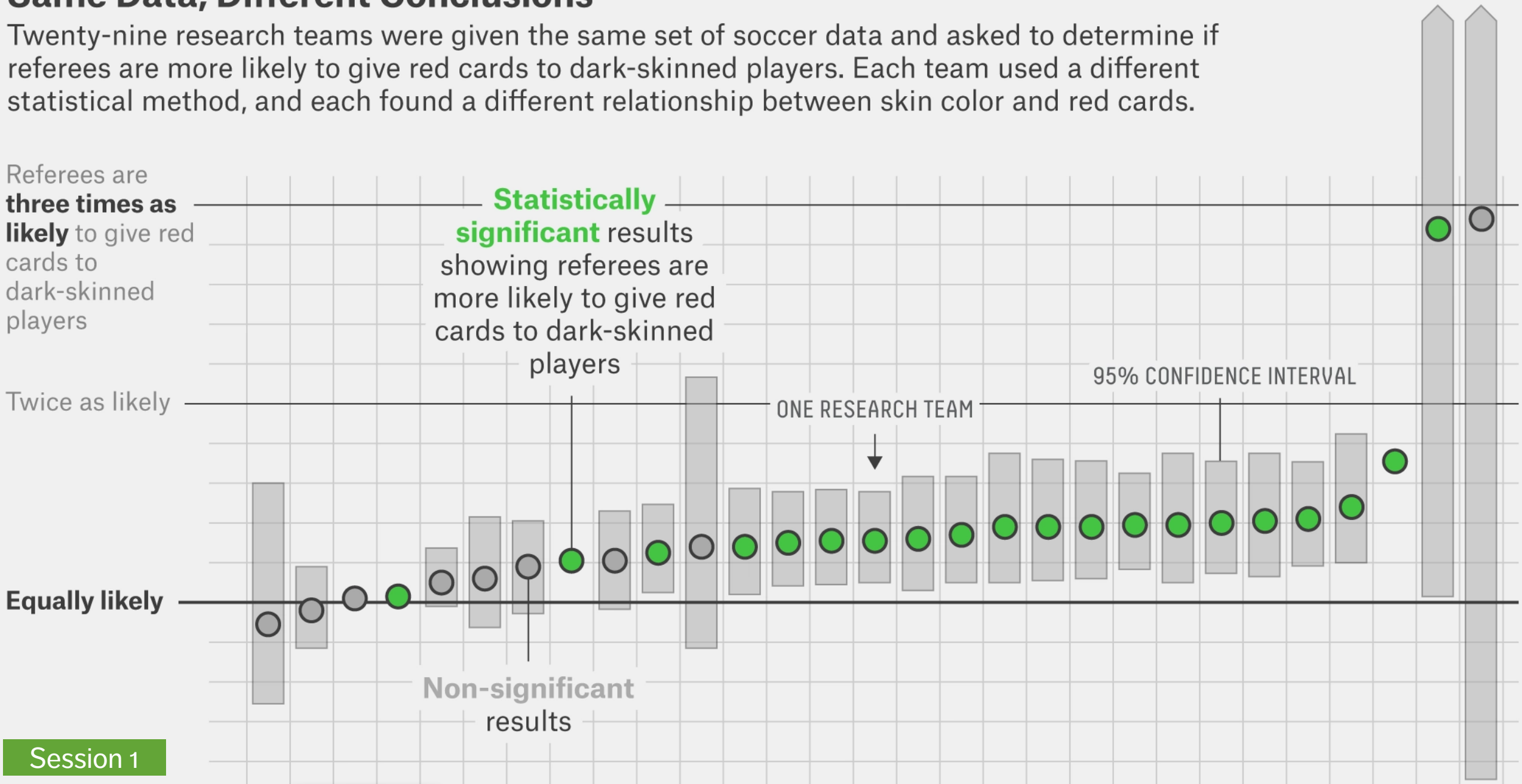
Opinions on the topic are varied – to some, DS/ML/AI provide examples of **brilliant successes**, while to others it is the **dangerous failures** that are at the forefront.

What do you think?

Are you a glass half-full or glass half-empty sort of person when it comes to data and applications?

Same Data, Different Conclusions

Twenty-nine research teams were given the same set of soccer data and asked to determine if referees are more likely to give red cards to dark-skinned players. Each team used a different statistical method, and each found a different relationship between skin color and red cards.



Session 1

From: Science isn't broken - It's just a hell of a lot harder than we give it credit for. **[Christie Aschwanden, 2015]**

Suggested Reading

Preliminaries

Data Understanding, Data Analysis, Data Science
Volume 2: Fundamentals of Data Insight

14. Data Science Basics

14.1 Introduction

- What is Data?
- From Objects and Attributes to Datasets
- Data in the News
- The Analog/Digital Data Dichotomy

Exercises

Preliminaries

1. Find examples of recent “Data in the News” stories. Were they successes or failures? What social consequences could emerge from the technologies described in the stories?

2. In what format is your organization’s data available? Are you able to access it easily? Is it updated regularly? Are there data dictionaries? Have you read them?

3. Conceptual Frameworks

Conceptual Frameworks

We use data to represent the world. But we also:

- describe the world using **language**
- represent it by building **physical models**

Common thread: **representation** (an object standing for another, being used in its stead in order to indirectly engage with the object being represented).

On one hand: “the map is not the territory”, but we do not need much effort to use the map to navigate the territory.

The transition from **representation** to **represented** can be seamless, which is risky: **it is easy to mistake the data/analytical results for the real world**.

Conceptual Frameworks

Best protection: thought out and explicitly described **conceptual framework**

- 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
- **appropriate** and **rigorous strategies** for applying the results of the analysis that is carried out in this representational framework

It could be built from scratch for each new project, but there are **modeling frameworks** that are broadly applicable to many different phenomena, which can be moulded to fit specific instances.

Three Modeling Strategies

There are 3 main (not mutually exclusive) **modeling strategies** that can be used to guide the specification of a phenomenon or domain:

- **mathematical** modeling
- **computer** modeling
- **systems** modeling

The first two have their own mathematical/digital (logical) worlds, distinct from the tangible, physical world studied by chemists, biologists, and so on:

- used to describe real-world phenomena by **drawing parallels** between the properties of objects in different worlds and reasoning via these parallels.

Three Modeling Strategies

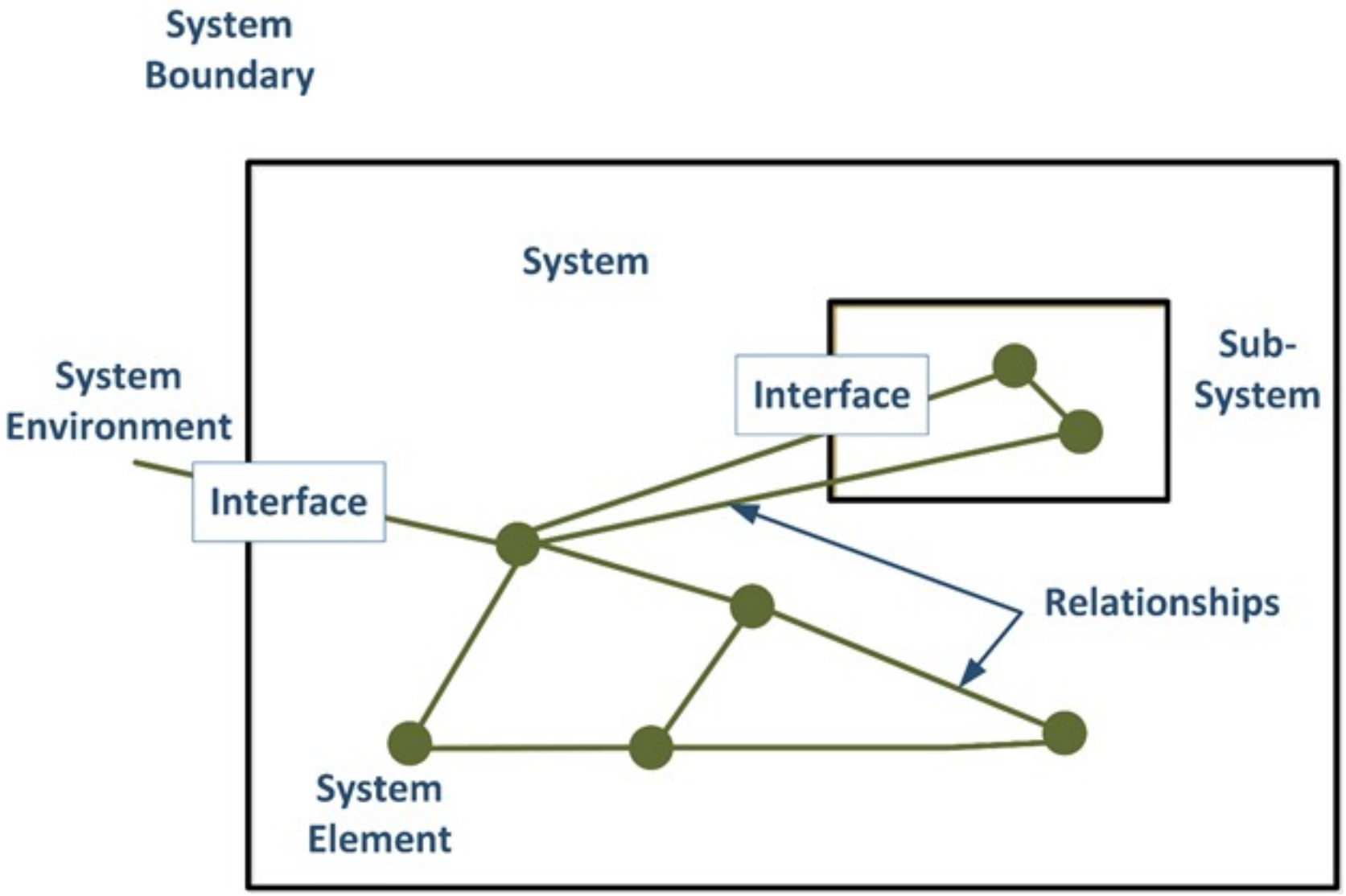
General Systems Theory describes **disparate** natural phenomena using a **common conceptual framework**, all as systems of interacting objects.

When presented with a new situation, we ask ourselves:

- which objects seem most relevant in the system behaviours of interest?
- 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?

Goals: **understand the system** and **relevant behaviours,** develop consistent **shared understanding**, inform **data collection, guide data interpretation**.

[Deepak, General Systems Theory, Wikimedia]



Information Gathering

Achieving **contextual understanding** of a dataset is crucial.

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.

Information Gathering

Clients or stakeholders are **not uniform** entities – client data specialists and SMEs may **resent the involvement** of analysts (external and/or internal).

Information gathering provides analysts the opportunity to show that everyone is pulling in the same direction, by:

- asking **meaningful** questions
- taking a **genuine interest** in the SMEs'/clients' experiences
- acknowledging everyone's ability to contribute

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

Thinking in Systems Terms

A **system** is made up of **objects** with **properties** that can change over time.

Within the system, there are **actions** and **evolving properties**, i.e., **processes**.

We understand how various aspects of the world interact with one another by **carving chunks** corresponding to the aspects and define their boundaries.

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

Objects themselves have various properties.

Thinking in Systems Terms

Natural processes generate/destroy objects, and change the properties of these objects over time.

We **observe**, **quantify**, and **record** values of these properties at particular points in time.

Observations are used to **capture the underlying reality** to an acceptable degree of **accuracy** and **error**, but … **even the best system model only ever provides an approximation of the situation under analysis.**

With luck, experience, foresight, these approximations might 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).

Causes:

- naïveté *vis-à-vis* the situation being modeled
- nature of the project under consideration

With **too many moving parts**, **unrealistic objectives**, **distance from pipeline**, knowledge gaps cannot be avoided (although they also occur with small, well-organized, easily contained projects).

Identifying Gaps in Knowledge

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 a knowledge gap, **be flexible**:

- **go back**
- **ask questions**
- **modify the system representation** as often as is necessary

It is preferable to catch these gaps early on in the process (obviously).

Conceptual Models

Conceptual models are built using methodical investigation tools:

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

Data scientists should beware **implicit conceptual models** (knowledge gaps).

It is preferable to err on the side of "too much conceptual modeling", but remember that "every model is wrong; some models are useful" [G.E. Box].

It is OK to build better models in an iterative manner.

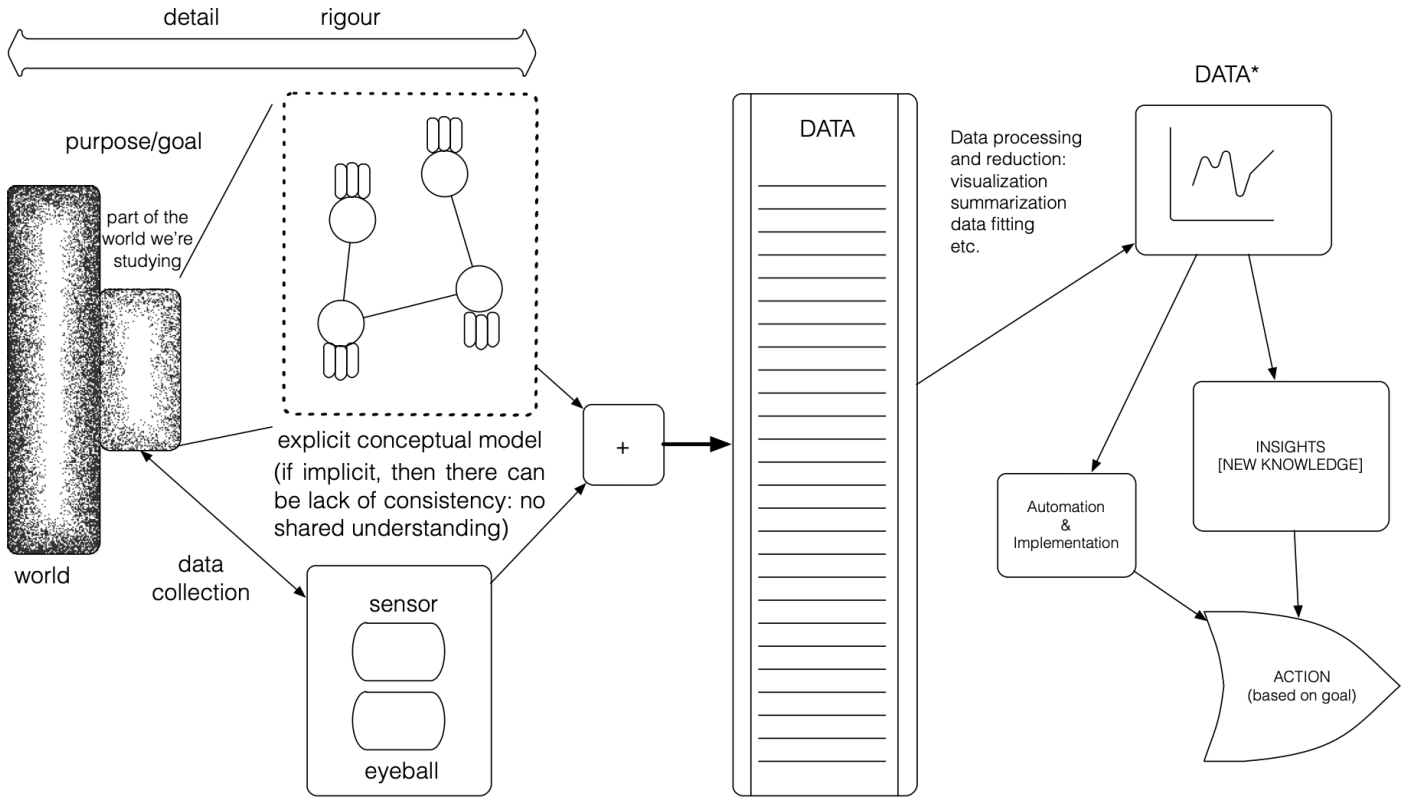
Conceptual Models

Conceptual model

- are not implemented as a scale-model or computer code
- exist only conceptually, often in the form of a diagram/verbal description of a system – boxes and arrows, mind maps, lists, definitions

Focus is on:

- **possible states** (not specific behaviour)
- object types, not specific instances; the goal is **abstraction**

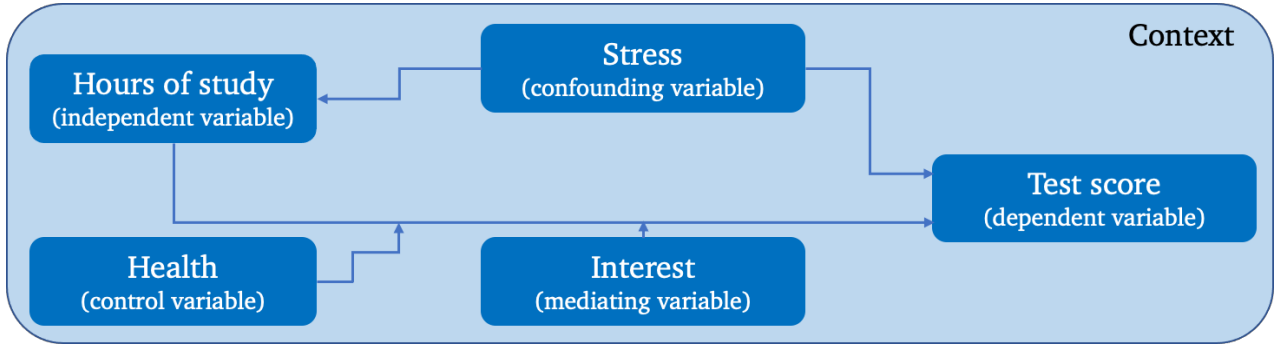


Conceptual Models

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).



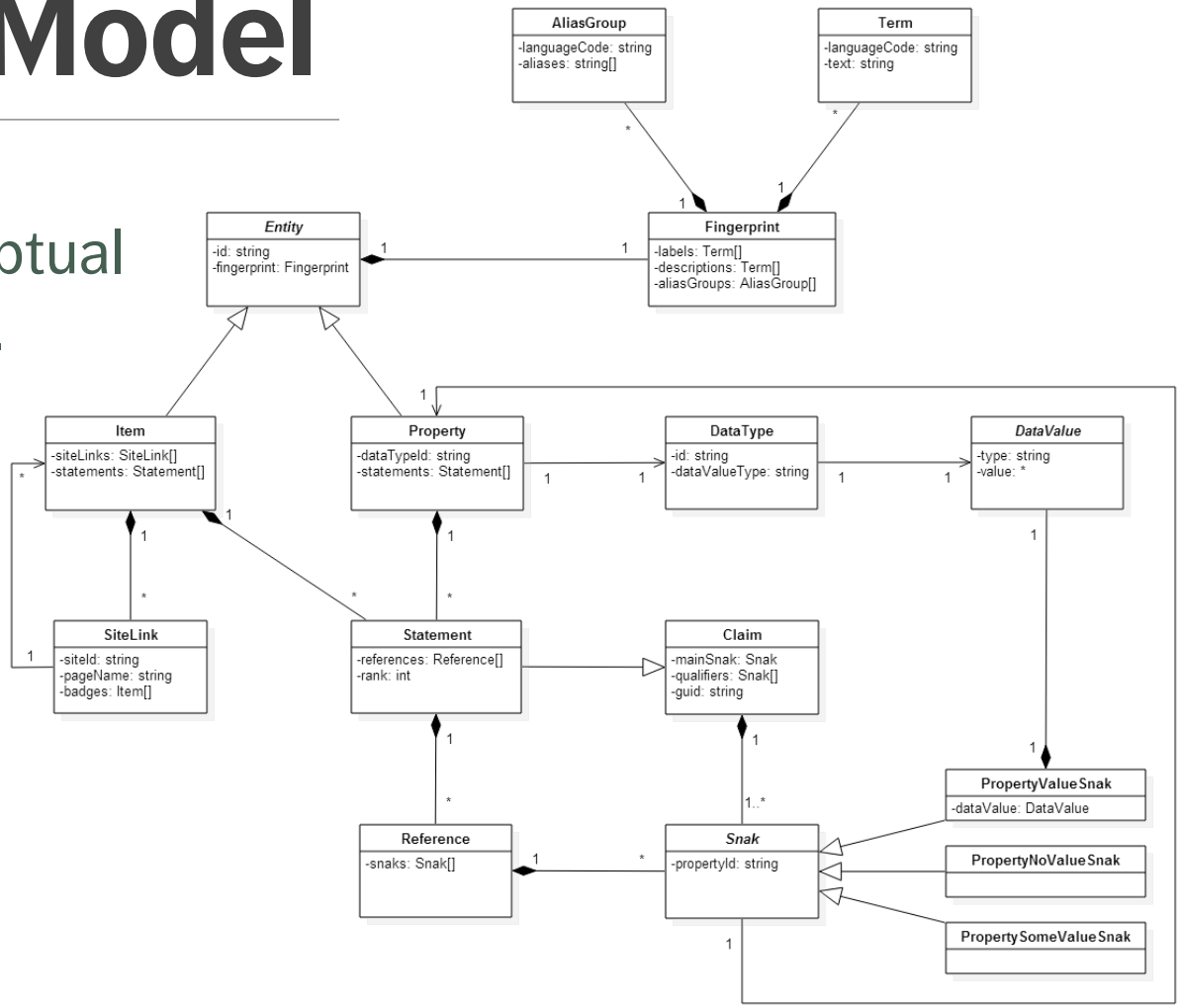
Formal Conceptual Model

Conceptual modeling turn implicit conceptual models into **explicit** and tangible models.

It provides opportunities to examine and explore ideas and assumptions.

Various efforts have been made to **formalize** conceptual modelling:

- UML (Universal Modelling Language)
- Entity Relationship (ER) Models



Relating Data to the System

Is the collected and analyzed data **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
- what the data represents (observations and features)

Is the **combination of system and data sufficient** to understand the situation under consideration? Difficult to answer in practice.

If the data, the system, and the real world are **out of alignment**, any data insight drawn from modeling and analysis might ultimately prove useless.

Cognitive Biases

Cognitive biases have an impact on how we construct models and look for patterns in the data:

- **Anchoring Bias** causes us to rely too heavily on the first piece of information we are given about a topic
- **Availability Heuristic** describes our tendency to use information that comes to mind quickly and easily when making decisions about the future
- **Bandwagon Effect** refers to our habit of adopting certain behaviours or beliefs because many others do the same
- **Choice-Supporting Bias** causes us to view our actions in a positive light, even if they are flawed
- **Clustering Illusion** refers to our tendency to see patterns in random events
- **Confirmation Bias** describes our tendency to notice, focus on, and give greater credence to evidence that fits with our existing beliefs
- **Conservation Bias** occurs when we favour prior evidence over new information
- **Ostrich Effect** describes how people often avoid negative information, including feedback that helps them monitor their goal progress

Cognitive Biases

- **Outcome Bias** refers to judging a decision on the outcome, rather than on why it was made
- **Overconfidence** causes us to take greater risks in our daily lives
- **Pro-innovation Bias** occurs when proponents of a technology overvalue its usefulness and undervalue its limitations
- **Recency Bias** occurs when we favour new information over prior evidence
- **Salience Bias** describes our tendency to focus on items or information that are more noteworthy while ignoring those that do not grab our attention
- **Survivorship Bias** is a cognitive shortcut that occurs when a visible successful subgroup is mistaken as an entire group
- **Zero-Risk Bias** relates to our preference for absolute certainty

Other biases:

- 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, response bias, selective perception, etc.

Suggested Reading

Conceptual Frameworks

Data Understanding, Data Analysis, Data Science
Volume 2: Fundamentals of Data Insight

14. Data Science Basics
    14.2 Conceptual Frameworks
    - Three Modeling Strategies
    - Information Gathering
    - Cognitive Biases

Exercises

Conceptual Frameworks

1. 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 an intern to pick it up in your living space. How would you explain to them, by phone, how to find the document? If the intern 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 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. Time is of the essence – you and the intern need to get the job done **correctly** as **quickly as possible**. What is your strategy?

2. Translate the cognitive biases to analytical contexts. What cognitive biases are you, your team, and your organization most susceptible to? Least?

4. Data Science Ethics

The Need for Ethics

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

Data scientists who come to the field by way of mathematics, statistics, computer science, economics, or engineering, however, are less likely to have encountered ethical research boards or **formal ethics training**.

Discussions on ethical matters are often tabled in favour of pressing technical or administrative considerations when faced with hard deadlines.

But the current deadline is replaced by another deadline, and then by another one, with the end result being that the conversation may never take place.

The Need for Ethics

When large-scale data collection first became possible, there was a 'Wild West' mentality to its use: **everything was allowed as long as it was feasible**.

Modern data science has **professional codes of conduct**

- outlining **responsible** ways to practice data science
- legitimate rather than fraudulent, ethical rather than unethical.

This shifts **added responsibility** to data scientists, but provides **protection** from clients/employers who want them to carry analysis in questionable ways.

The Need for Ethics

Recent focus on data ethics does not seem to have slowed breaches:

- Volkswagen
- Whole Foods Markets
- General Motors
- Cambridge Analytica
- Amazon
- Ashley Madison

What is/are Ethics?

Ethics refers to the study and definition of **right** and **wrong** conduct:

- in general
- applied in specific circumstances

Ethics is not (necessarily) the same as:

- social convention
- religious beliefs
- laws

What is/are Ethics?

In the West, ethical theories are used to frame debates around ethical issues:

- **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, impartial way

What is/are Ethics?

But humans subscribe to a wide variety of ethical codes/cultures, including:

- Confucianism
- Taoism
- Buddhism
- Shinto
- Ubuntu
- Te Ara Tika (Maori)
- etc.

It is easy to imagine contexts in which any of these would be better-suited to the task at hand –remember to **inquire** and to **heed the answers**.

Ethics and Data Science

How might these ethical theories apply to data analysis?

- 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 depend on a number of factors. To give you an idea of some of the complexities, let us the first question: *who, if anyone, owns data?*

Ethics and Data Science

Is it the **data analysts** who transform the data's potential into usable insights?

Is it the **data collectors** who have a copy and make the work possible?

Is it the **sponsors** or **employers** who made the process economically viable?

In some instances, the **law** may chime in as well. Anybody else?

This simple question is not easily answered; it's on a case-by-case basis.

Hidden truth: **there is more to data analysis than *just* data analysis**.

Ethics and Data Science

Similar challenge for **open data** (“pro” vs. “anti” both have strong arguments).

General principle of data analysis: eschew the **anecdotal** for the **general**. **Sound**, as focus on specific observations can obscure the full picture.

But data points are **not** just marks on paper or bytes on the cloud. Decisions made on the basis of data science may **affect living beings in negative ways**. It cannot be ignored that outlying individuals and minority groups often suffer disproportionately at the hands of so-called evidence-based decisions.

First Nations Principles of **OCAP** (Ownership, Control, Access, Possession).

Best Practices

"Do No Harm": data collected from an individual **should not be used to harm** the individual.

Informed Consent:

- Individuals must **agree to the collection and use** of their data
- Individuals must have a **real understanding of what they are consenting to**, and of **possible consequences** for them and others

Respect "Privacy": excessively hard to maintain in the age of constant trawling of the Internet for personal data.

Best Practices

Keep Data Public: data should be kept **public** (all? most? any?).

Opt-In/Opt-Out: Informed consent requires the ability to **opt out**.

Anonymize Data: removal of id fields from data prior to analysis.

“Let the Data Speak”:

- no cherry picking
- importance of validation (more on this later)
- correlation and causation (more on this later, too)
- repeatability

The Good, the Bad, and the Ugly

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

- **good** projects increase knowledge, can help uncover hidden links, etc., as harmlessly as possible
- **bad** projects can lead to bad decisions, which can in turn decrease the public's confidence and potentially harm some individuals
- **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

The Good, the Bad, and the Ugly

Good:

- P. A. B. Bien Nicholas AND Rajpurkar, "Deep-learning-assisted diagnosis for knee magnetic resonance imaging: Development and retrospective validation of MRNet," *PLOS Medicine*, vol. 15, no. 11, pp. 1–19, 2018, doi: 10.1371/journal.pmed.1002699.
- BeauHD, "Google AI claims 99 percent accuracy in metastatic breast cancer detection," *Slashdot.com*, Oct. 2018.
- Columbia University Irving Medical Center, "Data scientists find connections between birth month and health," *Newswire.com*, Jun. 2015.

The Good, the Bad, and the Ugly

Bad:

- Indiana University, "Scientists use Instagram data to forecast top models at New York Fashion Week," *Science Daily*, Sep. 2015.
- D. Wakabayashi, "Firm led by Google veterans uses A.I. to 'nudge' workers toward happiness," *New York Times*, Dec. 2018.
- N. Cohn, "How one 19-year-old illinois man is distorting national polling averages," *The Upshot*, 2016.

The Good, the Bad, and the Ugly

Ugly:

- J. Dastin, "Amazon scraps secret AI recruiting tool that showed bias against women," *Reuters*, Oct. 2018.
- I. Johnston, "AI robots learning racism, sexism and other prejudices from humans, study finds," *The Independent*, Apr. 2017.
- M. Judge, "Facial-recognition technology affects African-Americans more often," *The Root*, 2016.
- 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*, vol. 114, no. 2, pp. 246–257, Feb. 2018.

Suggested Reading

Data Science Ethics

Data Understanding, Data Analysis, Data Science
Volume 2: Fundamentals of Data Insight

14. Data Science Basics

14.3 Ethics in the Data Science Context

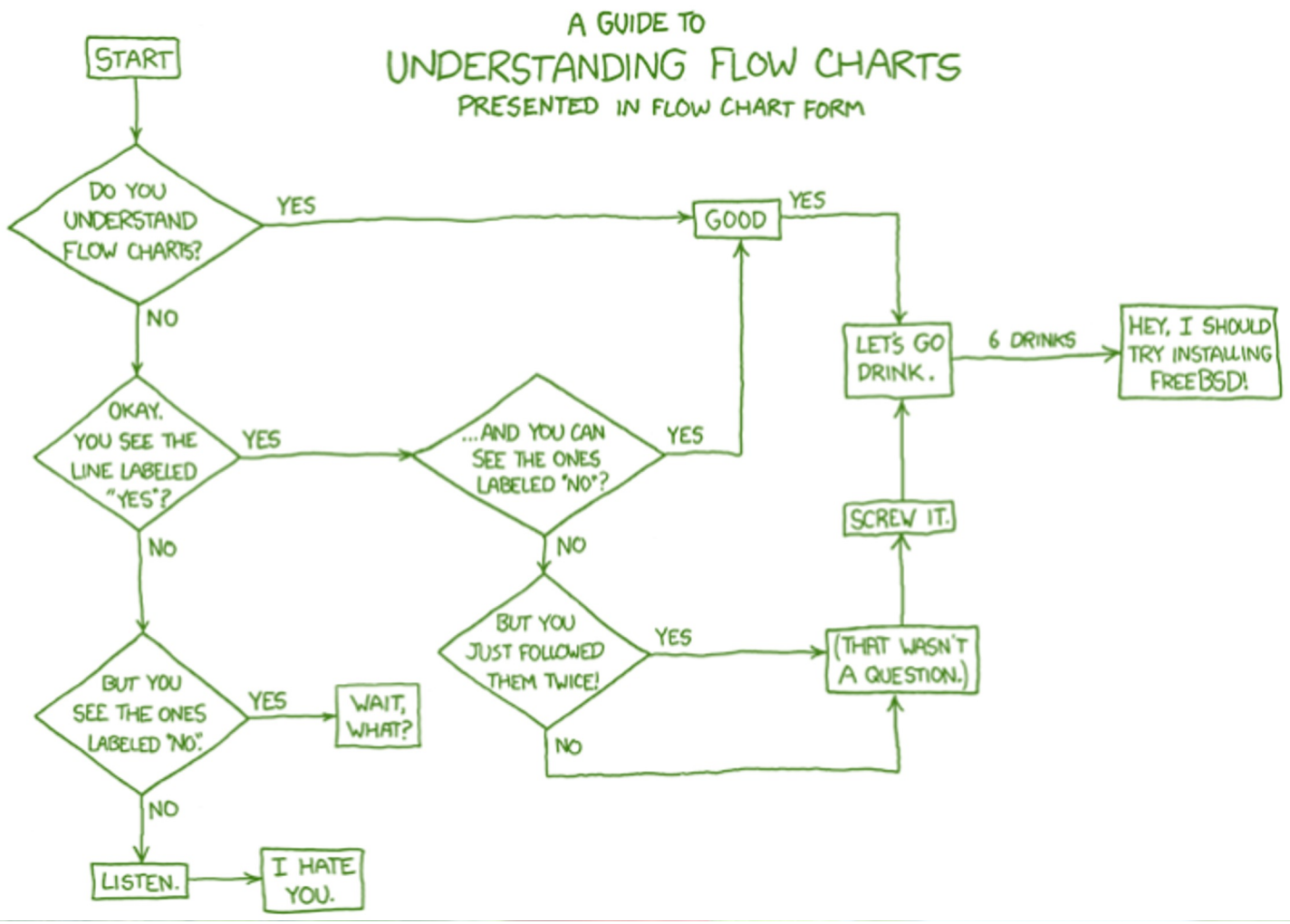
- The Need for Ethics
- What Is/Are Ethics?
- Ethics and Data Science
- Guiding Principles

Exercises

Data Science Ethics

1. Research the recent data ethics scandals involving Volkswagen, Amazon, Whole Foods Markets, Cambridge Analytica, Ashley Madison, General Motors, or any other organization. What transpired? Who was affected? What were the consequences to the general public, the organization, the data community? How could it have been avoided?

2. Establish a statement of ethics for your data work. Are there areas that you are unwilling to work on?



5. Analytics Workflows

Analytics Workflows

You are probably sick of **discussions about context** and would rather move to data analysis proper.

Very soon. One last thing, then: the **project context**.

Data science is more than just the analysis of data; this is apparent when we look at the typical steps involved in a **data science project.**

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**.

The "Analytical" Method

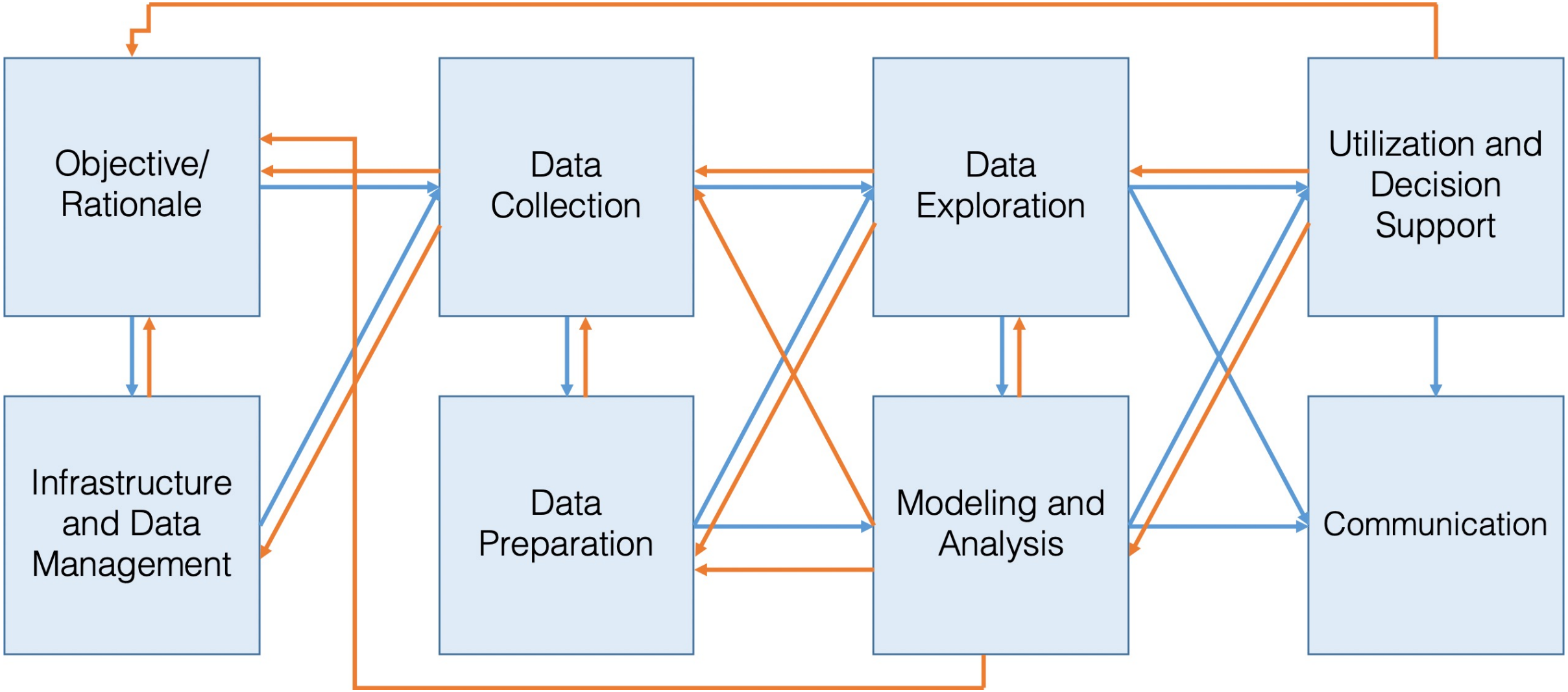
As with the **scientific method**, there is a "step-by-step" guide to data analysis:

- statement of objective
- data collection
- data clean-up
- data analysis/analytics
- dissemination
- documentation

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

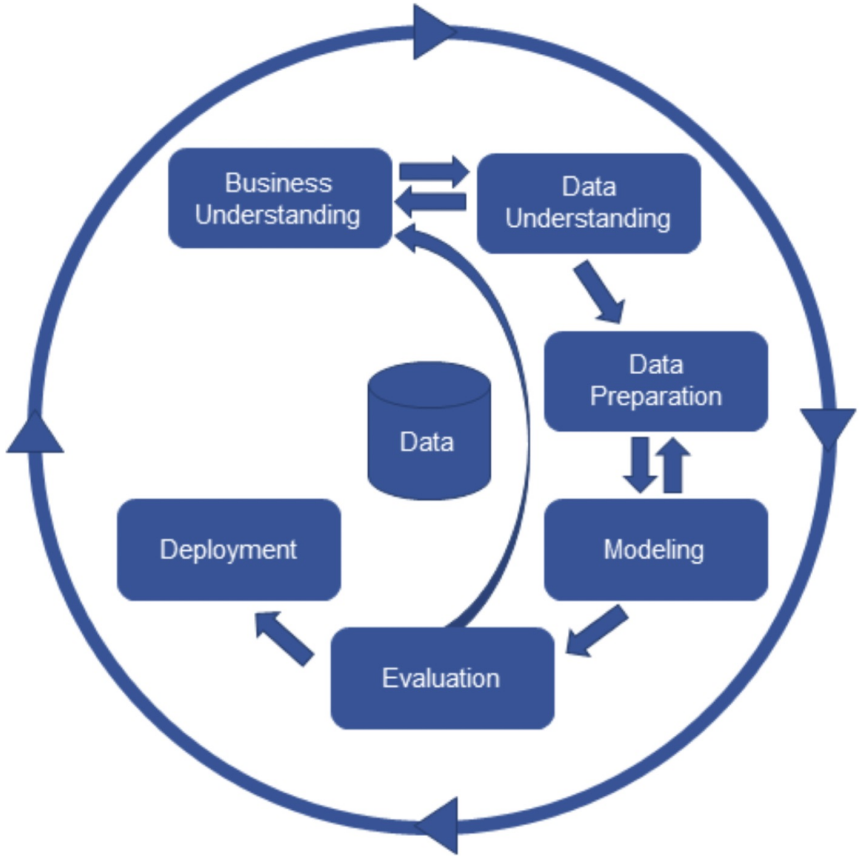
In practice, the process is quite often **messy**, with steps added in and taken out of the sequence, repetitions, re-takes, etc.

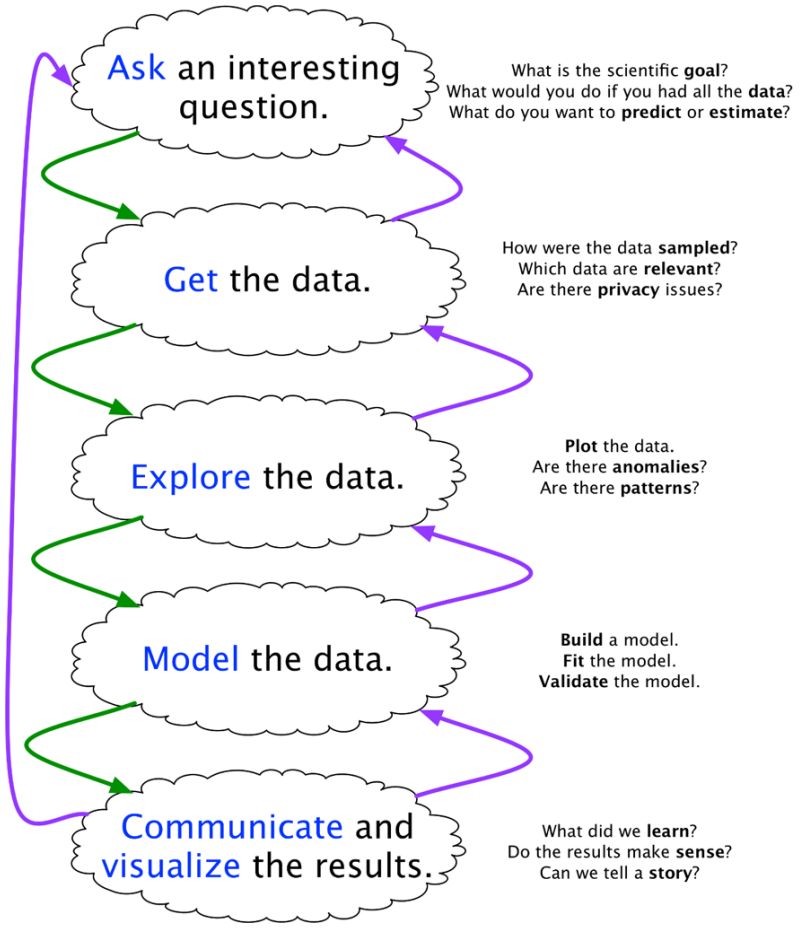
Surprisingly, it tends to work... when **conducted correctly**.


Objective/
Rationale
Data
Collection
Data
Exploration
Utilization and
Decision
Support
Infrastructure
and Data
Management
Data
Preparation
Modeling and
Analysis
Communication

[CRISP-DM framework (left); Blitzstein and Pfister's data science workflow (right)]

The "Analytical" Methods





The "Analytical" Methods

In practice, data analysis is often corrupted by:

- lack of clarity
- mindless rework
- blind hand-off to IT
- failure to iterate

All approaches have a common core

- data science projects are **iterative**
- (often) **non-sequential**.

Helping stakeholders recognize this **central truth** makes it easier for data scientists to:

- plan the **data science process**
- obtain **actionable insights**

Take-away: there is a lot to consider in advance of modeling and analysis

- **data analysis is not just about data analysis**.

Data Collection

Data enters the **data science pipeline** by being **collected**.

There are various ways to do this:

- data may be collected in a **single pass**
- it may be collected in **batches**
- it may be collected **continuously**

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

Data Storage

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**)
- the type of access and processing that will be required (**how fast**, **how much**, **by whom**)

Stored data may go **stale** (*figuratively* and *literally*); regular data audits are recommended.

Data Processing

The data must be **processed** before it can be analyzed.

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 so that they meet the requirements of the selected algorithms

The **analysis** itself is almost anti-climactic: simply run the selected methods or algorithms on the processed data.

Modeling

Data science teams should know:

- data cleaning
- descriptive statistics and correlation
- probability and inferential statistics
- regression analysis
- classification and supervised learning
- clustering and unsupervised learning
- anomaly detection and outlier analysis
- big data/high-dimensional data analysis
- stochastic modeling, etc.

These only represent a **small slice** of the analysis pie (see earlier slide).

No one analyst/data scientist could master all (or even a majority of them) at any moment, but that is one of the reasons why data science is a **team activity**.

Model Assessment and Life After Analysis

Before applying findings, we must first confirm that the model is reaching valid conclusions about the system of interest.

Analytical processes are **reductive:** raw data is transformed into a small(er) **numerical summaries**, which we hope is **related** to the system of interest.

Data science methodologies include an **assessment phase**

- analytical sanity check: is anything **out of alignment?**

Beware the **tyranny of past success:** even if the analytical approach has been vetted and has given useful answers in the past, it may not always do so.

[http://dailycatdrawings.tumblr.com; #309]

Real World

Model

Theory

Identification of details relevant to **description** and **translation** of real-world objects into model variables

Model Assessment and Life After Analysis

When an analysis or model is ‘released into the wild’, it often takes on a life of its own. When it inevitably ceases to be **current**, there may be little that data scientists can do to remedy the situation.

How do we determine if the current data model is:

- **out-of-date**?
- no longer **useful**?
- how long does it take a model to react to a **conceptual shift**?

Regular audits can be used to answer these questions.

Model Assessment and Life After Analysis

Data scientists rarely have full control over **model dissemination**.

- results may be misappropriated, misunderstood, shelved, or failed to be updated
- can conscientious analysts do anything to prevent this?

There is no easy answer: analysts should not only focus on the analysis, but also recognize opportunities that arises to **educate stakeholders** on the importance of these auxiliary concepts.

Due to **analytic decay**, the last step in the analytical process is not a **static dead end**, but an invitation to re-iterate to the beginning of the process.

Data Pipelines (First Pass)

In the **service delivery context**, the data analysis process is implemented as an **automated data pipeline** to enable automatic runs.

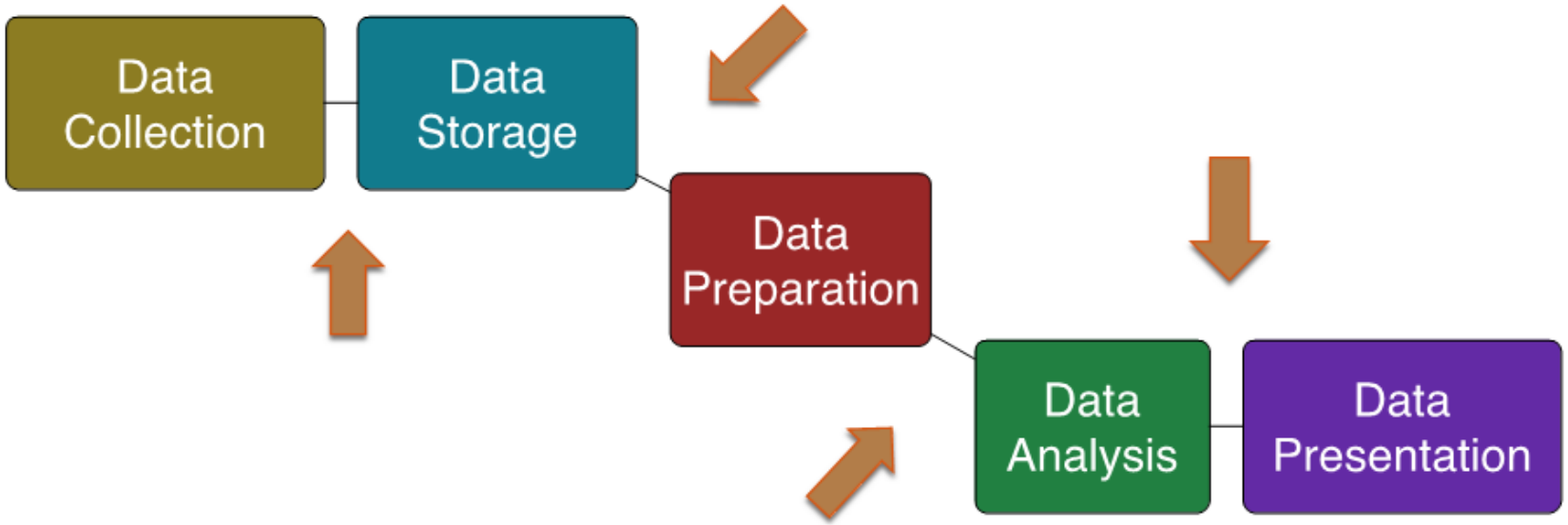
Data pipelines usually consist of 9 components (5 **stages** and 4 **transitions**):

- data collection
- data storage
- data preparation
- data analysis
- data presentation

Data Pipelines (First Pass)

Each components must be **designed** and then **implemented**.

Typically, at least one data analysis pass process has to be done **manually** before the implementation is completed.



Suggested Reading

Analytics Workflows

Data Understanding, Data Analysis, Data Science
Volume 2: Fundamentals of Data Insight

14. Data Science Basics

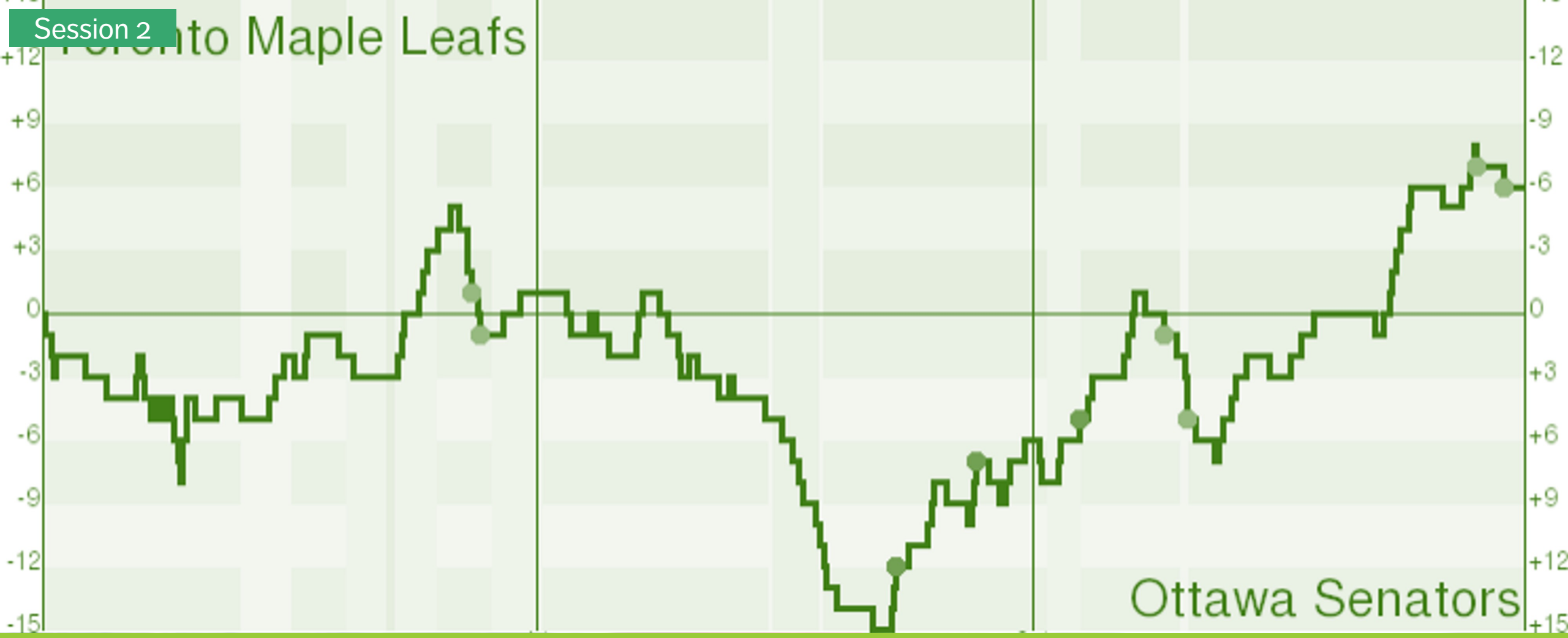
14.4 Analytics Workflows

- The “Analytical” Method
- Data Collection, Storage, Processing, and Modeling
- Model Assessment and Life After Analysis
- Automated Data Pipelines

Exercises

Analytics Workflows

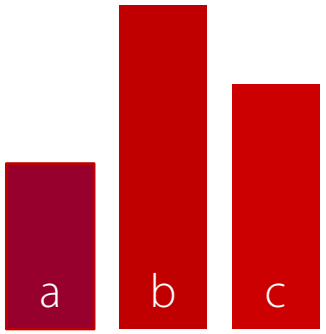
1. Install R / RStudio (Posit), and packages from the list the instructor will provide.
2. Test the installation with examples from the Programming Primer (sections 2 – 4) to make sure that the software performs as expected.



6. Getting Insight From Data

Analytics Modes

Descriptive



Show **what** happened

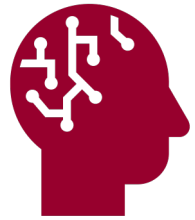
Diagnostic

Explain **why** something happened

Predictive

Guess **what will** happen

Prescriptive



Suggest **what should** happen

Low Value
Low Difficulty

High Value
High Difficulty

Asking the Right Questions

Data science is about asking and answering questions:

- **Analytics:** "How many clicks did this link get?"
- **Data Science:** "Based on this user's previous purchasing history, can I predict what links they will click on the next time they access the site?"

Data mining/science models are usually **predictive** (not **explanatory**): they show connections, but don't reveal why these exist.

Warning: not every situation calls for data science, artificial intelligence, machine learning, statistics, or analytics.

The Wrong Questions

Too often, analysts are asking the **wrong questions:**

- questions that are **too broad** or **too narrow**
- questions that **no amount of data could ever answer**
- questions for which **data cannot reasonably be obtained**

The **best-case scenario** is that stakeholders will recognize the answers as irrelevant.

The **worst-case scenario** is that they will erroneously implement policies or make decisions based on answers that have not been identified as misleading or useless.

Roadmap to Framing Questions

Understand the problem (opportunity vs problem)

What initial assumptions do I have about the situation?

How will the results be used?

What are the risks and/or benefits of answering this question?

What stakeholder questions might arise based on the answer(s)?

Do I have access to the data necessary to answering this question?

How will I measure my ‘success’ criteria?

Yes/No Trap

Examples of **bad** questions:

- Are our revenues **increasing** over time? **Has it** increased year-over-year?
- Are most of our customers from **this demographic**?
- **Does this project have** valuable ambitions to the broader department?
- **How great** is our hard-working customer success team?
- How often do you **triple check** your work?

Examples of **good** questions:

- What's the **distribution** of our revenues over the past three months?
- Where are our **top 5** high-spending cohorts from?
- What are the **different benefits** of pursuing this project?
- What are **three good** *and* **bad traits** of our customer success team?
- Do you **tend to** do quality assurance testing on your deliverables?

Question Audit Checklist

1. Did I avoid creating any yes/no questions?
2. Would anyone in my team/department understand the question irrespective of their backgrounds?
3. Does the question need more than one sentence to express?
4. Is the question 'balanced' - scope is not too broad that the question will never truly be answered, or too small that the resulting impact is minimal?
5. Is the question being skewed to what may be easier to answer for my/my team's particular skillset(s)?

Contingency/Pivot Tables

Contingency table: examines the relationship between two categorical variables via their relative (cross-tabulation).

Pivot table: a table generated by applying operations (sum, count, mean, etc.) to variables, possibly based on another (categorical) variable.

Contingency tables are special cases of pivot tables.

	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

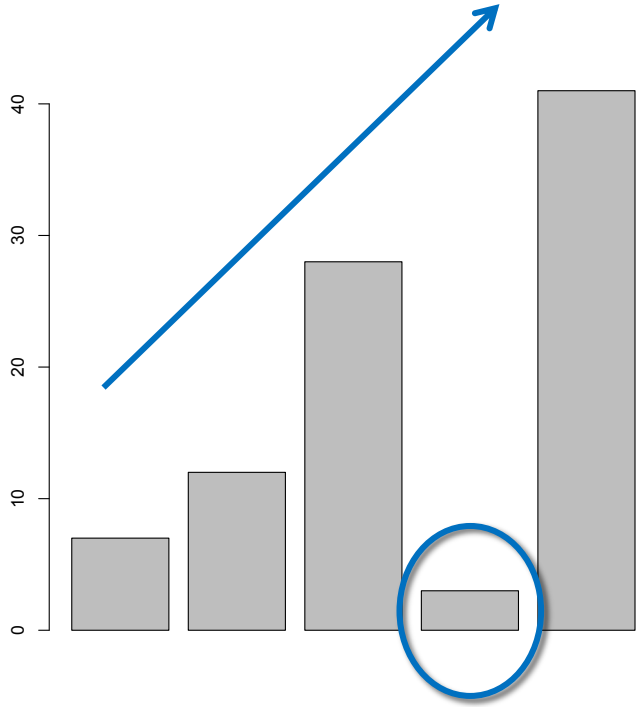
Analysis Through Visualization

Analysis (broad definition):

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

Option 1: use analytical methods to achieve this.

Option 2: visualize the data and use the brain's analytic power (perceptual) to reach meaningful conclusions about these patterns.



Numerical Summaries

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

Centrality measures include:

- median, mean, mode (less frequently)

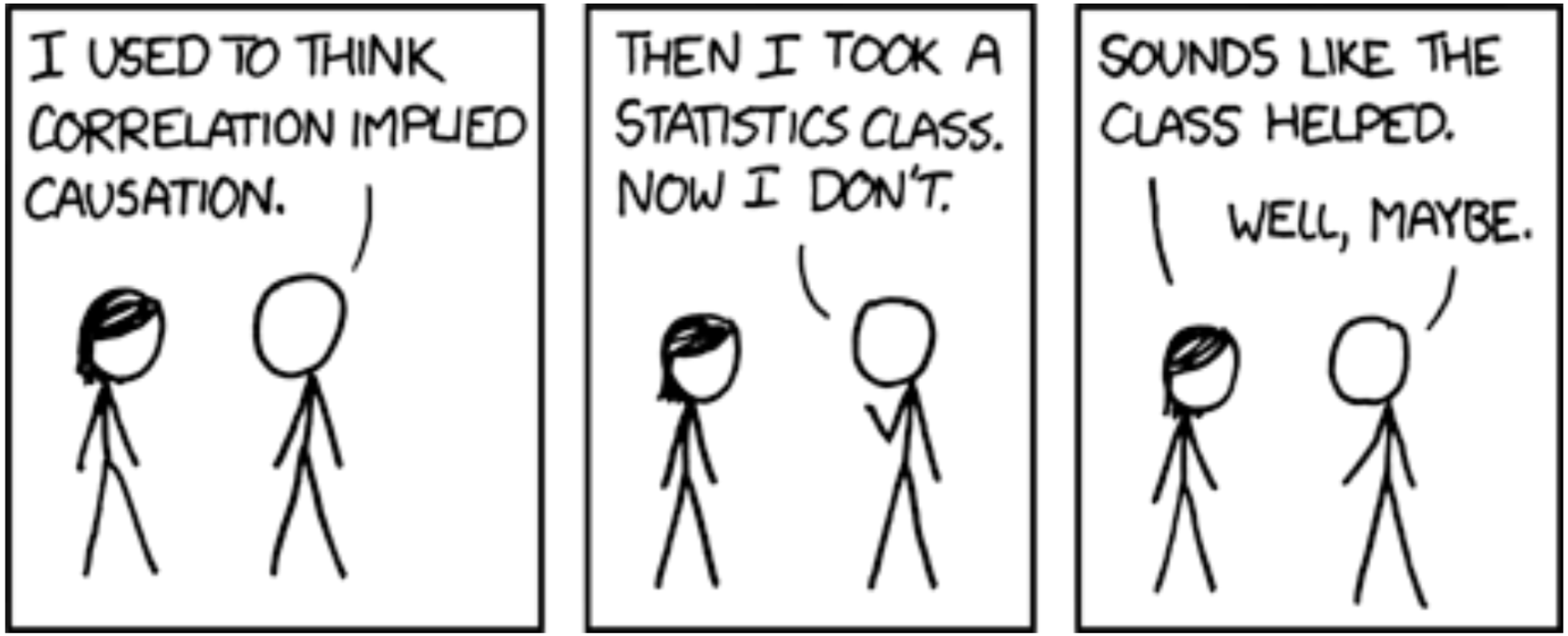
Spread (or dispersion) measures include:

- standard deviation (sd), variance, quartiles, inter-quartile range (IQR), range (less frequently)

The median, range and the quartiles are easily calculated from **ordered lists**.

[https://xkcd.com/552]

Correlation



Correlation doesn't imply causation, but it does waggle its eyebrows suggestively and gesture furtively while mouthing 'look over there'.

Linear Regression

The basic assumption of **linear regression** is that the dependent variable y can be approximated by a linear combination of the independent variables:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

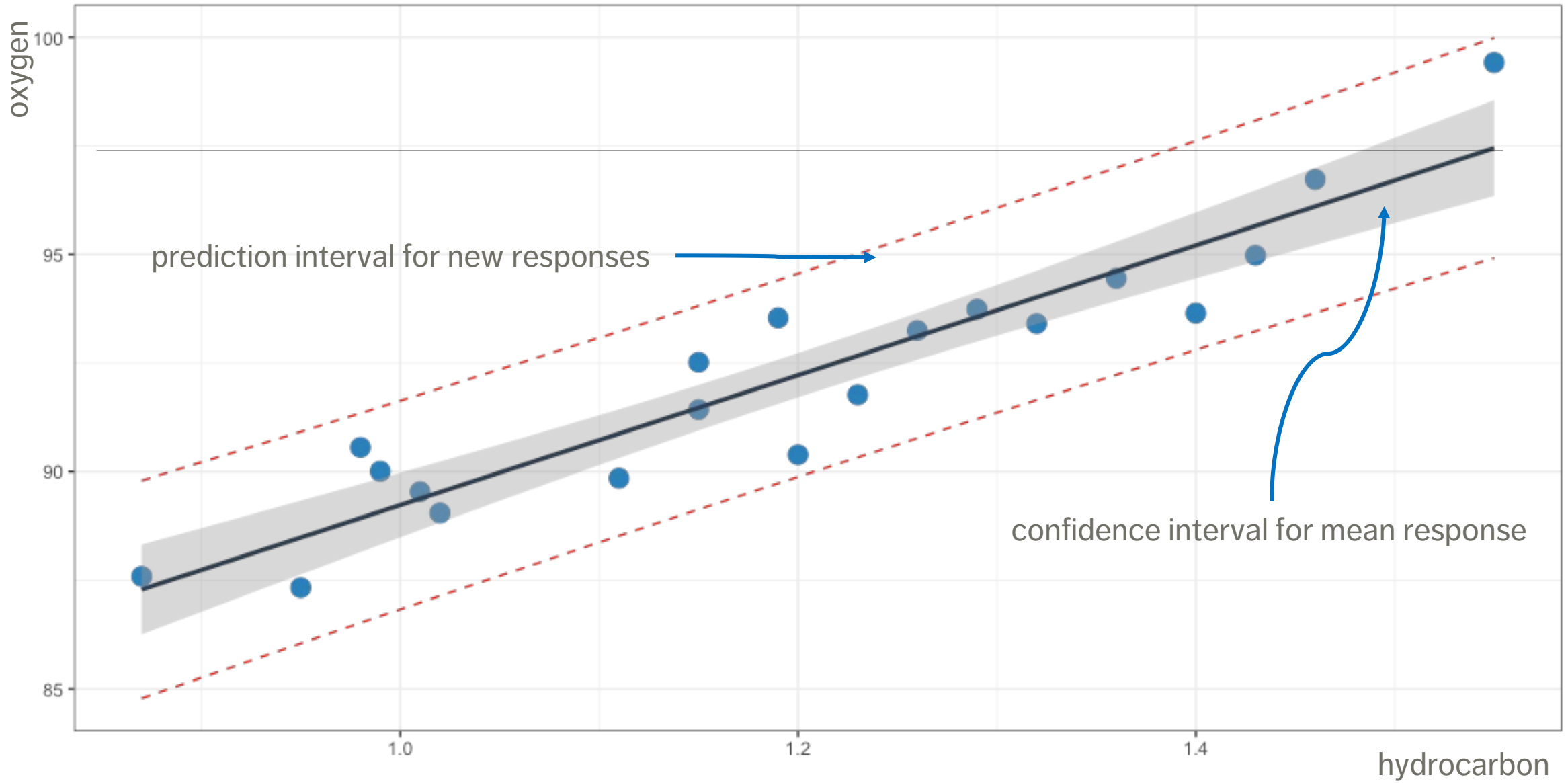
where $\boldsymbol{\beta} \in \mathbb{R}^p$ is to be determined based on the **training set**, and for which

$$E(\boldsymbol{\varepsilon}|\mathbf{X}) = \mathbf{0}, \qquad E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T|\mathbf{X}) = \sigma^2\mathbf{I}.$$

Typically, the errors are also assumed to be **normally distributed**:

$$\boldsymbol{\varepsilon}|\mathbf{X} \sim N(\mathbf{0}, \sigma^2\mathbf{I}).$$

$$\text{oxygen} = 14.95 \times \text{hydrocarbon} + 74.28$$



Machine Learning Tasks

Classification and **class probability estimation:** which clients are likely to be repeat customers?

Clustering: do customers form natural groups?

Association rule discovery: what books are commonly purchased together?

Others:

profiling and behaviour description; **link prediction**; **value estimation** (how much is a client likely to spend in a restaurant); **similarity matching** (which prospective clients are similar to a company's best clients?); **data reduction**; **influence/causal modeling**, etc.

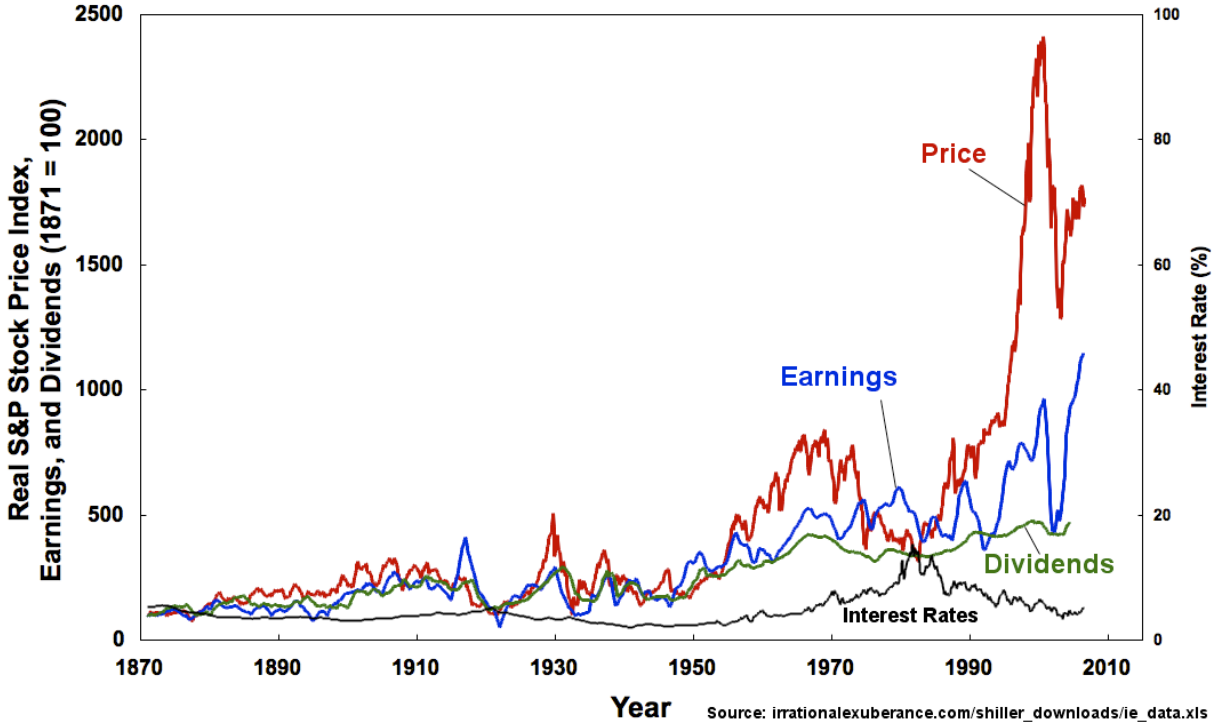
Time Series Analysis

A simple **time series:**

- has two variables: time + 2nd variable
- the second variable is *sequential*

What is the **pattern of behaviour** of this second variable over time? Relative to other variables?

Can we use this to **forecast the future behaviour** of the variable ?



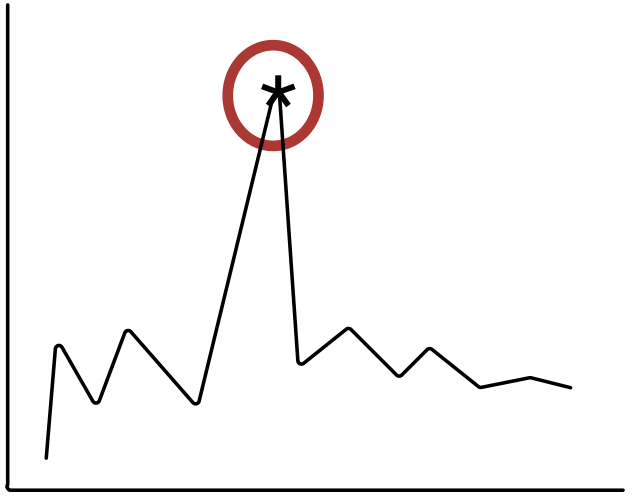
Anomaly Detection

Anomaly: an unexpected, unusual, atypical or statistically unlikely event

Wouldn't it be nice to have a data analysis pipeline that alerted you when things were out of the ordinary?

Many different analytic approaches to take!

- clustering
- classification
- ensemble techniques, etc.



Suggested Reading

Getting Insight From Data

Data Understanding, Data Analysis, Data Science

Volume 2: Fundamentals of Data Insight

14. Data Science Basics
    - 14.5 Getting Insight From Data
        - Asking the Right Questions
        - Basic Data Analysis Techniques
        - Common Statistical Procedures in R
        - Quantitative Methods

Volume 1: Prelude to Data Understanding

6. Probability and Applications
7. Introductory Statistical Analysis
8. Classical Regression Analysis
9. Times Series and Forecasting
10. Survey Sampling Methods
11. The Design of Experiments

Exercises

Getting Insight From Data

1. Do the exercise in Asking the Right Questions.
2. Recreate the examples of Common Statistical Procedures in R.
3. The file cities.txt contains population information about a country's cities. A city is classified as "small" if its population is below 75K, as "medium" if it falls between 75K and 1M, and as "large" otherwise. Locate and load the file into the workspace of your choice. How many cities are there? How many are there in each group? Display summary population statistics for the cities, both overall and by group.