Analytics for Decision-Making

Instructors:
Patrick Boily, Stephen Davies, Jen Schellinck, John Stroud

uOttawa

Institut de développement professionnel
Professional Development Institute

Analytics for Decision-Making

Instructors: Patrick Boily, Stephen Davies, Jen Schellinck, John Stroud

Slides: P. Boily (IACS, DAL, uOttawa), J. Schellinck (Sysabee, DAL, AI Guides), J. Stroud (AI Guides)
S. Davies (DAVHILL, DAL), B. Conway-Smith (Sysabee), M. Kashef ()

Patrick Boily

Bio

- Professor ['19 – now | Math/Stat | uOttawa]
- President, Consultant ['16 – now | Idlewyld Analytics]
- Manager, Senior Consultant ['12 – '19 | CQADS | Carleton]
- Public Service ['08 – '12 | CBSA | StatCan | TC | PWGSC]
- 60+ university courses; 250+ workshop days

Projects and Clients

- GAC; NWMO; CATSA; etc.
- 40+ analytics/data science projects

Specializations

- Data visualization; data cleaning (… unfortunately)
- Application of wide breadth of techniques to all kinds of data

Stephen Davies

Bio

- CEO [DAVHILL Group | GoC & Aviation]
- M. Sc. [Systems Engineering]
- Lecturer [Sprott School of Business, Carleton]
- Former executive in High Tech & Manufacturing
- 20+ years consulting and training

Projects and Clients

- All major GoC departments
- Pearson Airport, Porter, Sunwing, Flightstar, First Air, ICAO

Specializations

- Data visualization, business intelligence, data engineering
- Process transformation, risk management

Jen Schellinck

Bio

- CEO ['12 – now | Sysabee Inc.]
- Co-Founder ['18 – now | Data Action Lab]
- Co-Founder ['20 – now | AI Guides]
- Adjunct Professor ['09 – now | ICS, Carleton]
- Director of Modeling Methodologies ['09 – '12 | Cogniva]

Projects and Clients

- GAC; DFO; CATSA; CSPS Digital Academy, etc.

Specializations

- Unsupervised learning techniques
- Simulations

John Stroud

Bio

- AI Consultant [AIGuides]
- Former Vice President [CATSA | Strategy, HR, Law, Corporate]

Projects and Clients

- GoC
- Entrepreneurs: law, financial Services

Specializations

- AI as a Service
- Human factors
- Crowdsourcing
- Augmented Intelligence

Outline

Module 1
　　Decision-Making

Module 2
　　Reasoning, Evidence, Information, Data

Module 3
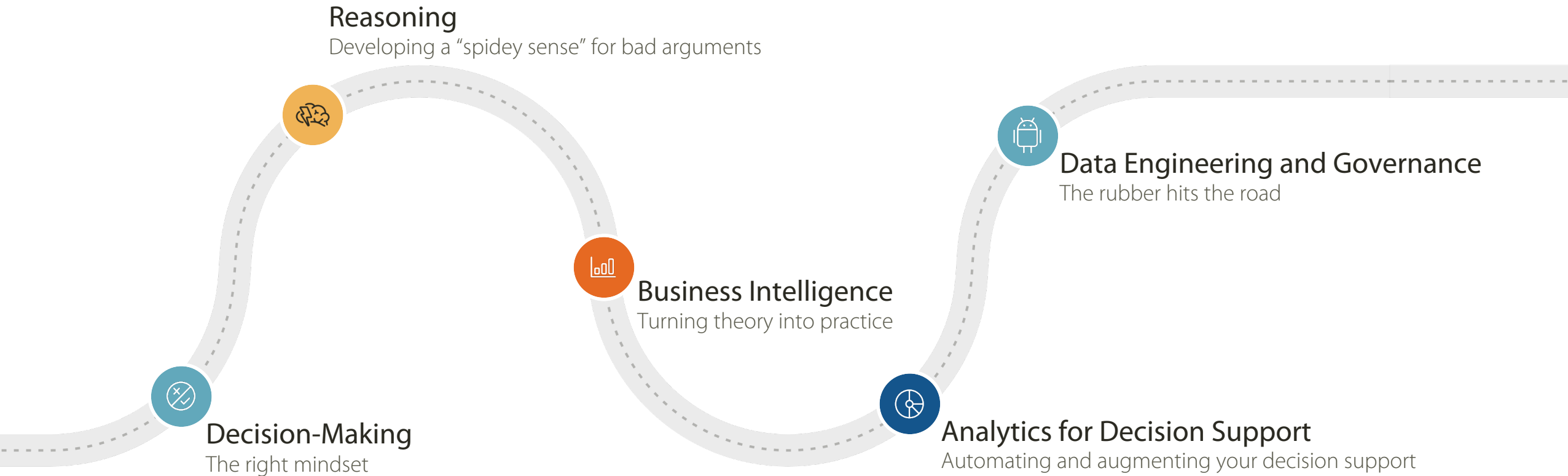　　People, Data Ethics, and the Law

Module 4
　　Business Intelligence and Analytics

Module 5
　　Analytics for Decision Support

Module 6
　　Data Engineering and Data Governance

Decisions Made in the Past Year

Should we renovate the roof this year or next year?

Should we travel to Nova Scotia in the midst of the pandemic?

Should I buy a car?

Should my son go to boarding school?

Analytics and Data Tools
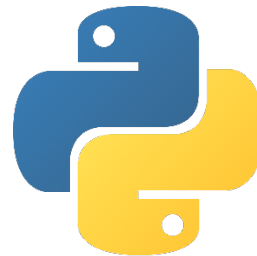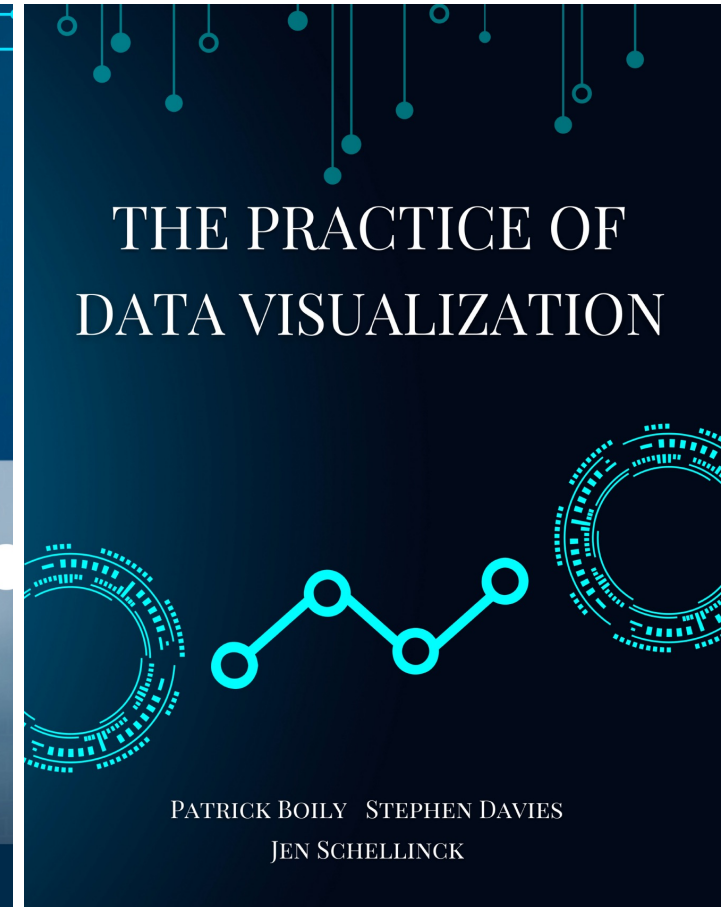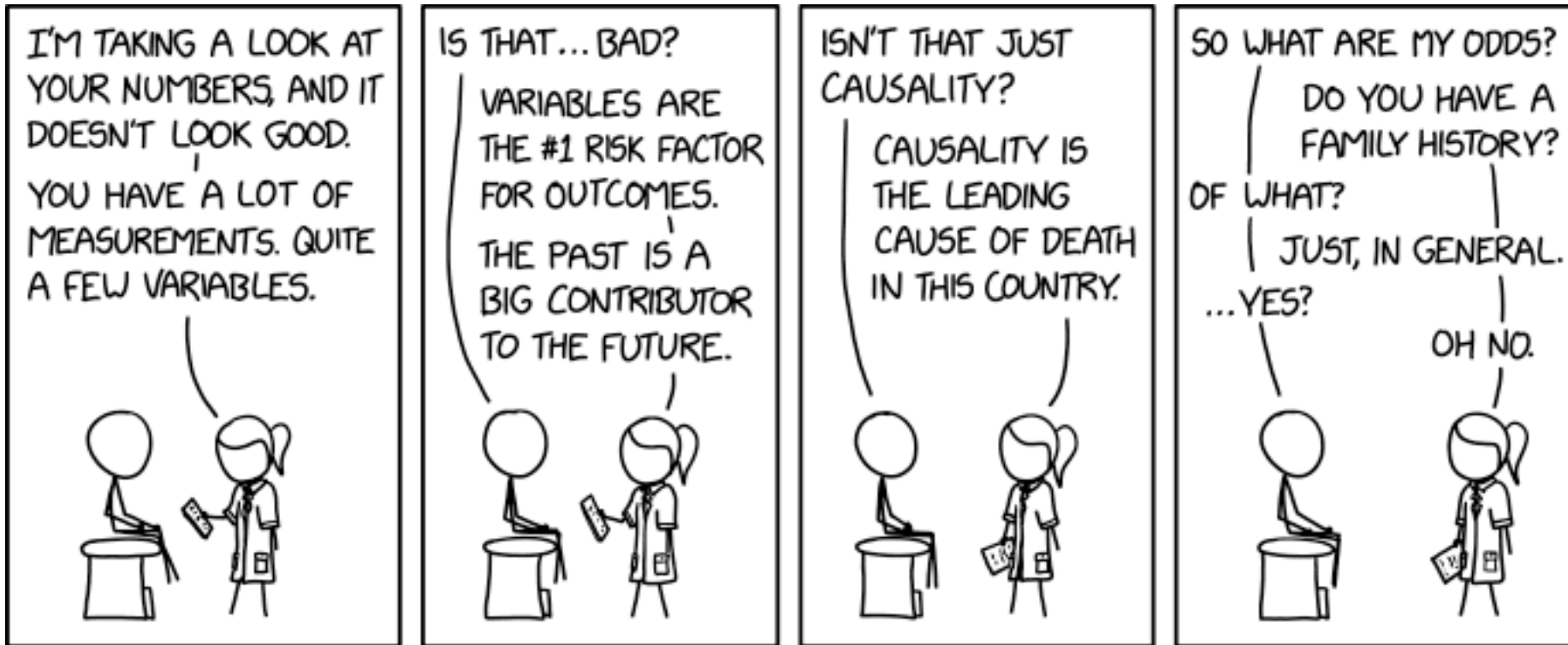
Suggested References

**Data Understanding,
Data Analysis, and
Data Science**

idlewyldanalytics.com

**The Practice of
Data Visualization**

pracicedataviz.com

Donate now to help us find a cure for causality.

No one should have to suffer through events because of other events.   [xkcd #2602]

Module 1
Decision-Making
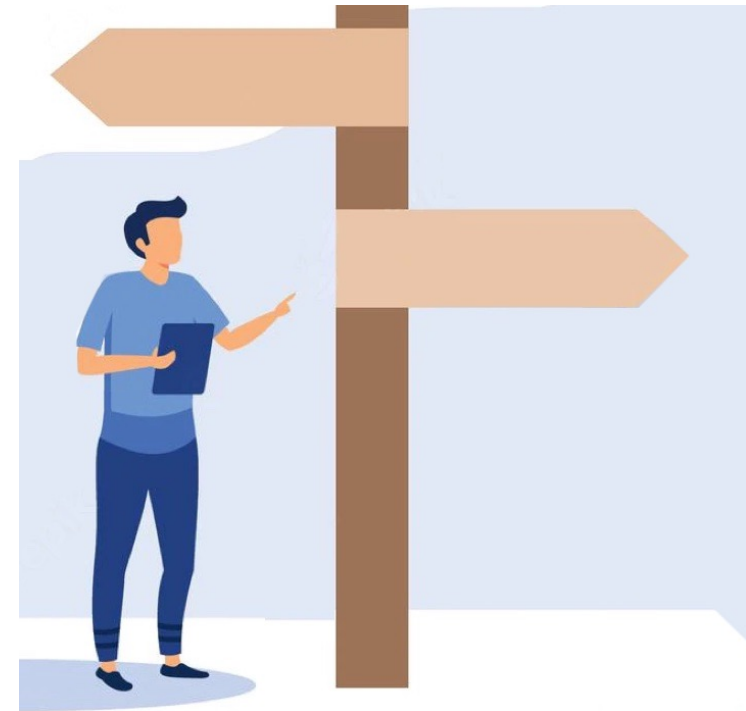
Life Turns on Two Things

Luck
(outside your control)

Decisions
(within your control)

The Game's on the Line

Situation

- 26 seconds left in the Super Bowl, trailing by 4 point, with the ball at 1 yard line.

- "Everyone" thinks you will hand off to your star running back.

Options

A: Run the ball (1 play). **Risk:** Fails to score and time runs out.

B: Throw then run if necessary (2 plays). **Risk:** 2% chance of interception.

The worst play call in NFL history will continue to haunt Seahawks in 2015

The worst play call in the history of the NFL will continue to haunt the Seattle Seahawks during the 2015 NFL season.

DON BANKS • JUL 20, 2015

Hillary Clinton's "Mistake"

Read the [Vanity Fair](#) article:

"You Could Fit All the Voters Who Cost Clinton the Election
in a Mid-Size Football Stadium"

Was it bad luck, or a mistake?

Buying the Rights to Star Wars

United Artists, Universal, and Disney passed.

20th Century Fox picked it up.

$775M for first movie, and $10B for the series.
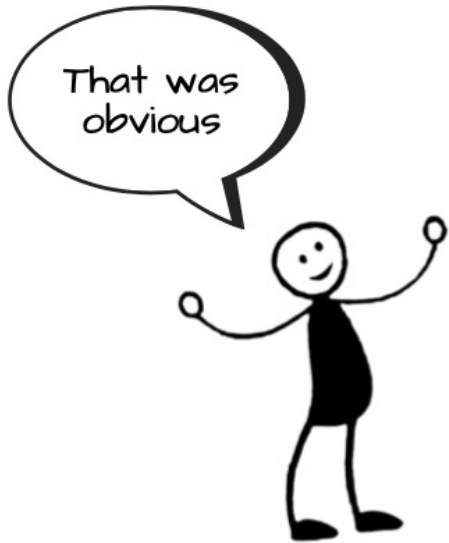
20th Century Fox: Smart or Lucky?

Resulting

Outcome Quality

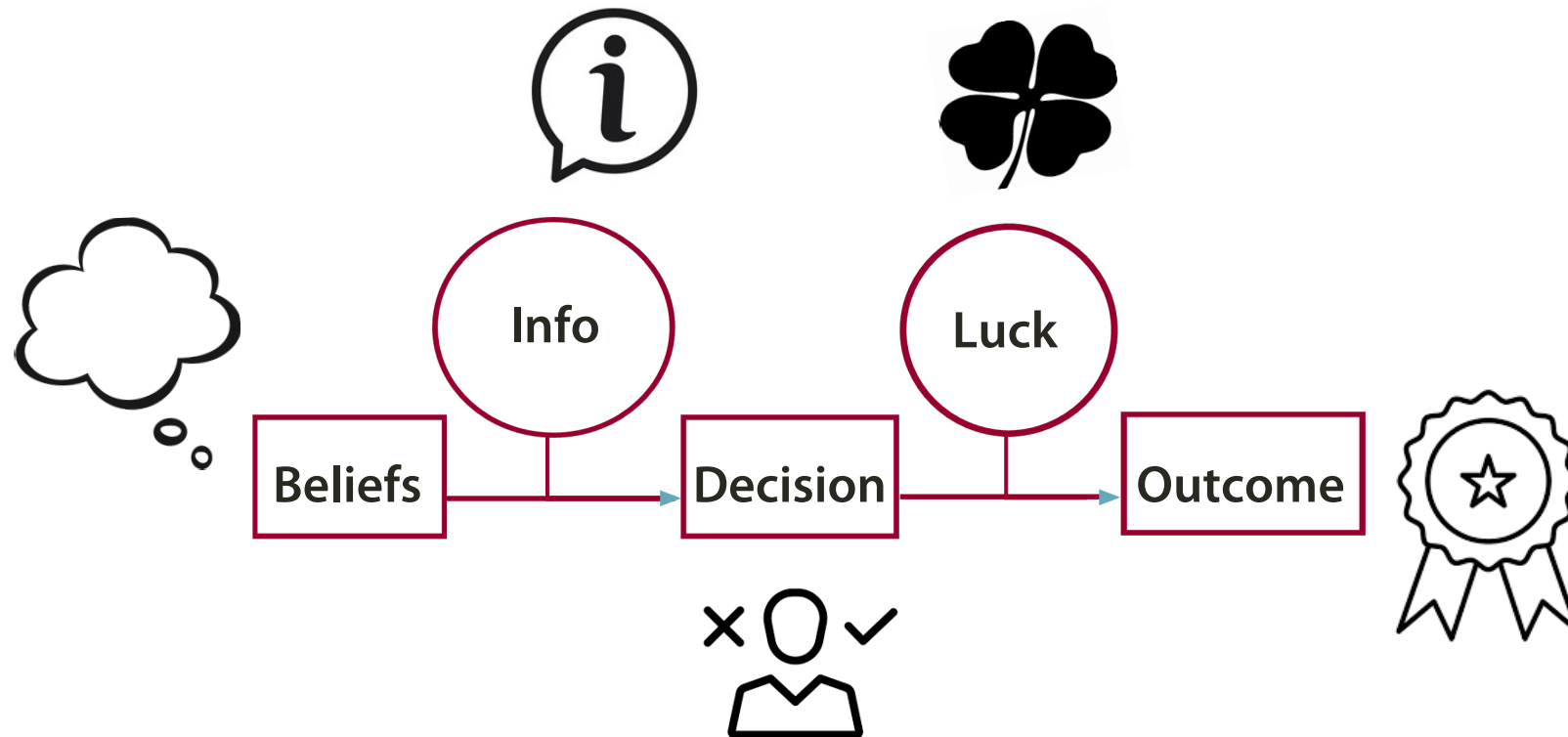|	Good	Bad
Good	Earned Reward	Bad Luck
Bad	Dumb Luck	Just Desserts

Decision Quality

Better Approach

Goals:

- can be used repeatedly, in the same way each time
- can be taught to someone else
- can be improved over time (in certain cases)
- allows for you to look back and examine whether you used it properly
- automated

Luck and Information

Personal Bias Types

Variable Bias

Confirmation Bias

Survivorship Bias

Planning Fallacy

The IKEA Effect

Progress Bias

Anchoring Bias

Availability Bias

The IKEA Effect

Giving too much credit to work we've done ourselves

Solution: avoid deciding or pushing a solution **just because** you're its architect

Variable Bias

Survivorship Bias

The IKEA Effect

Anchoring Bias

Confirmation Bias

Planning Fallacy

Progress Bias

Availability Bias

Anchoring Bias

Also known as "first impression bias" – jumping to conclusions

YOU ONLY GET ONE CHANCE TO MAKE A FIRST IMPRESSION

Solution: slowdown and recalibrate your thinking

Variable Bias

Survivorship Bias

The IKEA Effect

Anchoring Bias

Confirmation Bias

Planning Fallacy

Progress Bias

Availability Bias

Availability Bias

Weighing the first reasonable solution too highly

**Solution:** don't commit to the first solution **just to** avoid wasting time

Systems 1 & 2: Blind to Our Blindness

System 1: Automatic

- Generally works well, but lots of systemic biases
- Is a system for "jumping to conclusions"
- We are blind to our blindness about using System 1

System 2: Deliberate

- Requires attention
- "Paying attention" is apt because it costs one's deliberate effort
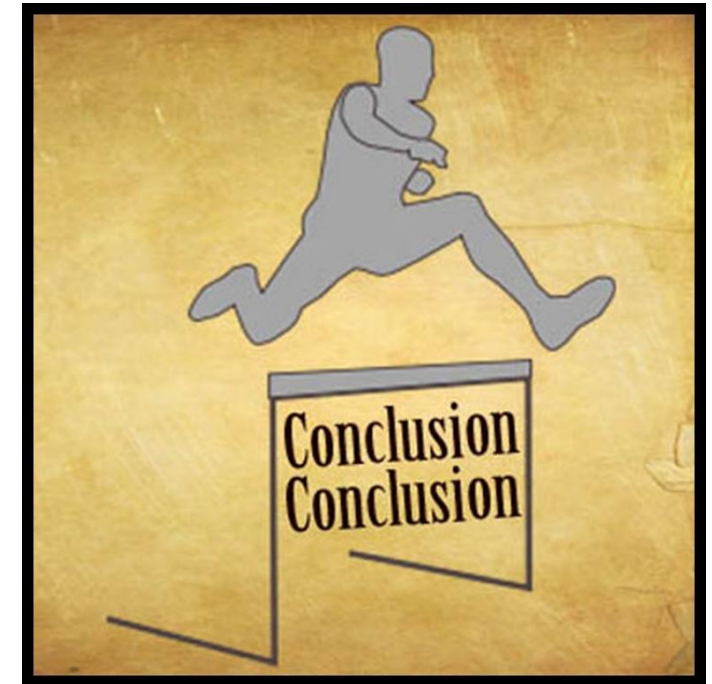- It requires the effortful application of logical rules
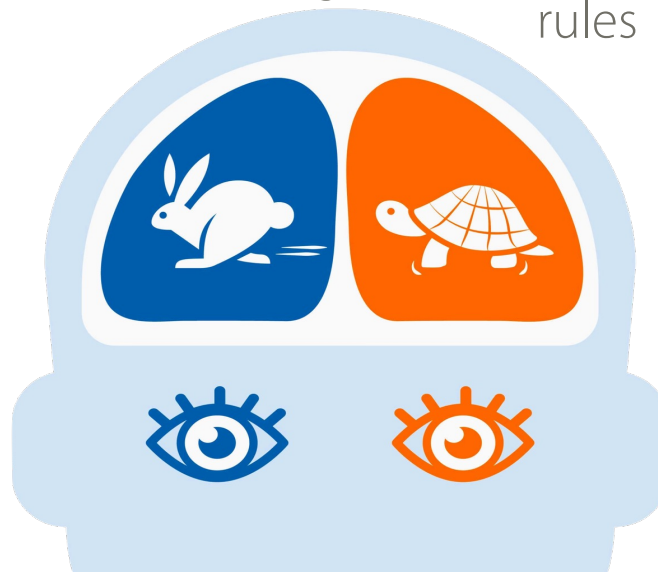
WYSIATI

What You See Is All There Is

- We (us and our bosses) jump to conclusions based on limited evidence.

Dunning-Kruger effect

- "You don't know what you don't know."
- Knowing little about something can make us overconfident about our knowledge and abilities.

Exercise: Medical Treatment Decision

A new disease is attacking your community. The good news is that treatments are emerging.

However, the effectiveness of the treatments vary. The hospital has elected to create two working groups (Team 1 and Team 2) to study the alternative treatments and return with a recommendation.

Study the evidence on treatments available and return with your recommendation.

Herd Mentality

"A person is smart. **People are dumb, panicky, dangerous animals and you know it.**"
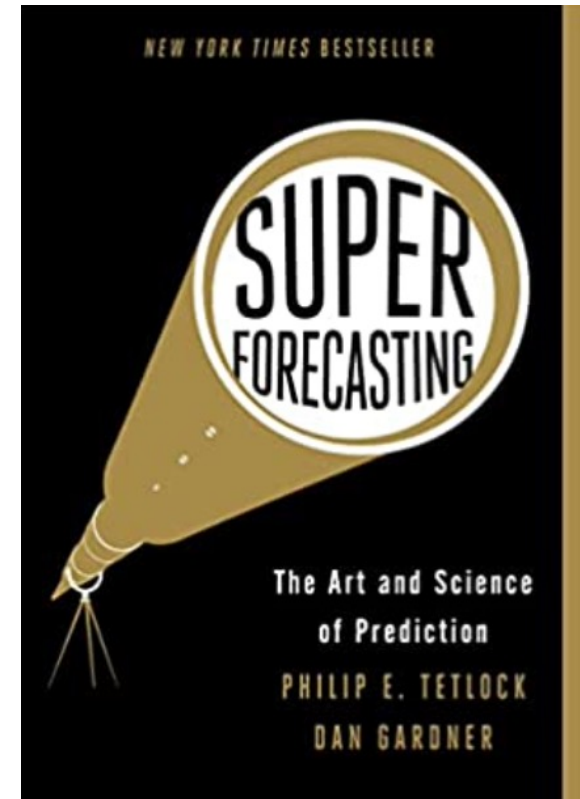
Wisdom of the Crowd

Risk Assessment Voting Heat Map

IARPA Tournament

Superforecaster Profile

No specific qualification or experience

Not WHO they are

What matters is HOW they decide
- curious
- self-critical
- humble
- open to new information

Be a Fox (not a Hedgehog)

Old Greek saying:
"A fox knows many things, but a hedgehog
knows one big thing"

Augmented Intelligence

Augmented Intelligence = Superforecasters + Algorithms

Instead of giving equal weight to all forecasters, **make 2 adjustments**

1.   Give extra weight to top forecasts

2.   Extremize the forecasts (i.e., push them closer to 0% or 100% probability)

For example, if the forecast is 70%, extremize it to 85%; if 30%, to 15%.

Ray Dalio: "Dot Collector" and A.I.

TIP: Avoid Echo Chambers at Work

1. Ask people for their views & Get More Data Points

2. Don't share your view when asking

3. Ask for views independently (especially before the boss speaks)

TIP: Pre-Mortem

A port-mortem is good for learning the causes of a bad outcome, with one tiny limitation: the patient is already dead.

Pre-mortem: imagine yourself at some time in the future, having **failed to achieve a goal**, and looking back at how you arrived at that destination – it is an autopsy **before** the patient dies.

Steps:

1. identify goal to achieve, or decision being considered
2. pick a timeline for achieving that goal.
3. imagine it is the day after the deadline and you are looking back from that date; give 5 reasons "within your control" and 5 reasons "outside of your control" for why things didn't work out

TIP: Backcasting

Backcasting: imagine yourself at some point in the future, having **succeeded in achieving a goal**, and looking back at how you arrived at that destination.

The process is like the one for pre-mortems (directions change,of course).

Benefits (pre-mortem/backcasting):

- get outside view
- wisdom of the crowd
- avoid group think
- eliminates the risk of people "not wanting to be the squeaky wheel"

TIP: Decision Exploration Table

	Pre-mortem	Backcasting
3 reasons **within** control/skill	1. 2. 3.	1. 2. 3.
3 reasons **outside** control/skill	1. 2. 3.	1. 2. 3.

Adapt To & Mitigate Bad Luck

After a premortem/backcasting:

1. modify your decision based on the new insights; increase the chance of good things happening and reduce the chance of bad things happening;

2. look for ways to mitigate the impact of bad luck; hedging (which reduces the impact of bad luck when it arises); there is a cost, and we hope never to have to use it (e.g., flood insurance).

The Rebel Alliance succeeded in stealing the Death Star's plans. The Empire did not undertake a pre-mortem: "any attack made by the Rebels against this station would be a useless gesture, no matter what technical data they have." The Rebels learned that a torpedo hitting a small external exhaust port could cause a fatal chain reaction.

Classic case of overconfidence.

Tips for Making Fast Personal Decisions

1. **Happiness test**
Will the decision have a significant impact on your happiness in 1 year?

2. **Worst-case scenario**
"What's the worst that can happen?"

3. **Closeness**
Would you rather a vacation next year in Rome or Paris? Is there a wrong choice?

4. **Only option**
"If this were the only option, would I be satisfied?"
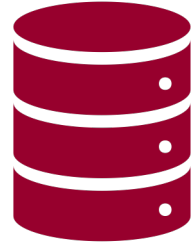
Module 2
Reasoning, Evidence, Information, Data

What does analysis look like in your organization?

What is Analysis in the GoC?

Drawing conclusions?

Gathering and presenting evidence (pivot tables)?

Providing options?

Providing opinions/hypotheses/beliefs/recommendations?

Pushing your agendas?

Typical Analysis Activities

Analysis is an activity done **to** something. We analyze the **situation** or the **problem:**

- Gathering facts and evidence
- Summarizing the facts
- Reviewing and evaluating facts
- Combining facts
- Generating new statements or hypotheses
- Breaking down concepts into simpler concepts
- Building up more complex concepts from simpler concepts
- Defining concepts
- Using reasoning to derive new facts
- Determining if statements are true (facts) or false
- Determining how confident we are about a statement being true or false

Common theme: facts!

Armchair Analysis

Critical Thinking

We want to show you why **critical thinking** (supported by analysis, reasoning, inference) is important.

Why adding rigour and methods is important

This is not a course on logic, BUT…

ultimately reasoning activities are all about getting at the (a?) **truth** – having enough true facts at your fingertips to keep you from making bad decisions.

How to be a Fox (and Not a Hedgehog)

We've already said that, as a decision maker, you need to be a fox. De facto, this means that analytics is your job

How can you explain to others why analytics – and by extension your job – matters when making decisions?

We will get down to specifics.
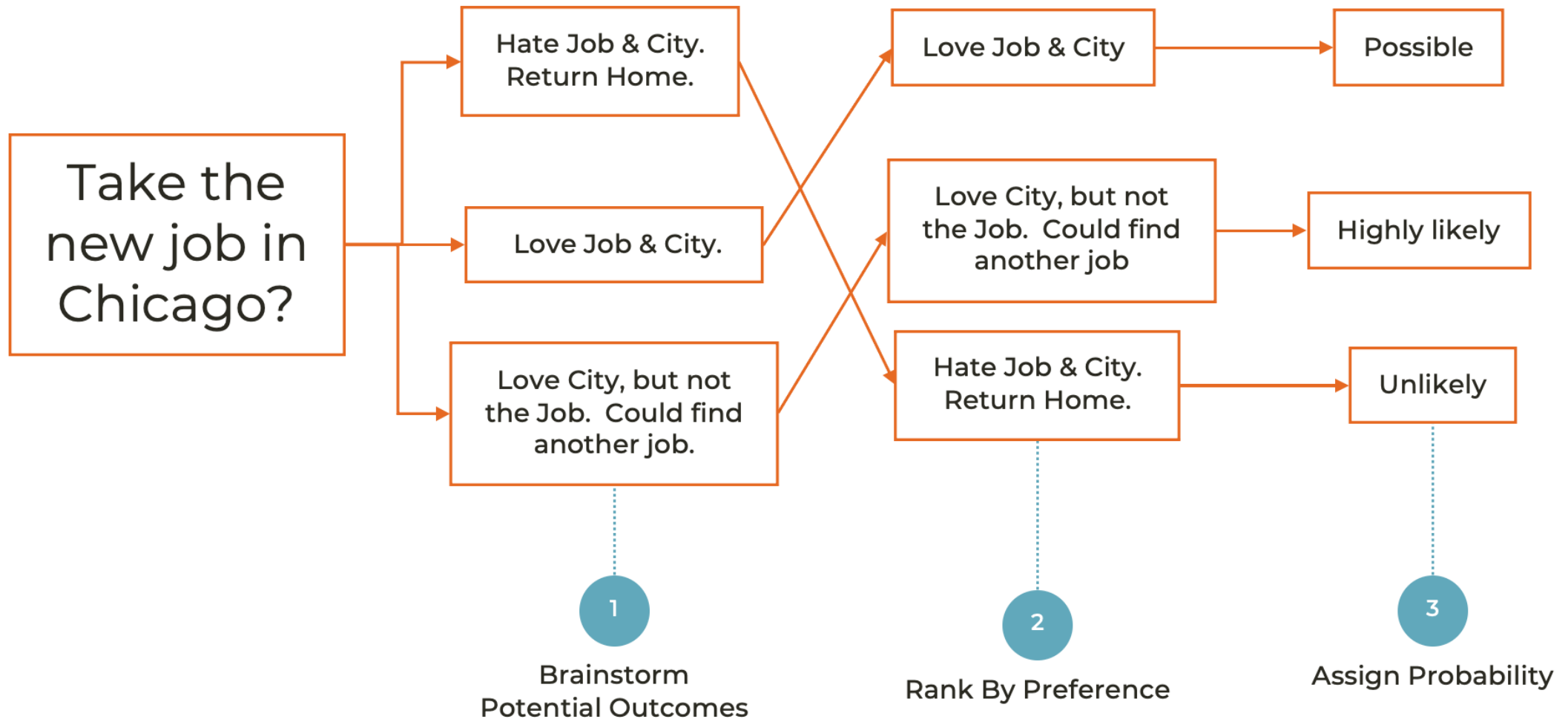
Some Effort Required

System 1 is easy

System 2 is hard

Is it worth it? Yes. Knowledge is a valuable commodity.

Exercise: Reasoning and Arguments

Are these strong? If not, what are their flaws? How could you improve them?

a) COVID vaccinations lead to increased hospitalizations since half of the hospitalizations were vaccinated.

b) Turning the Large Hadron Collider on was a mistake because either it destroys the Earth, or it does not; that's a 50% chance – way too risky.

c) We know that the Earth is not flat because none of the other planets we know are flat.

d) You should not vote in the next election because one vote rarely ever makes a difference.

e) The solution to reduce congestion is to reduce the number of lanes because with fewer lanes, people will seek alternative modes of transportation.

f) Airport security measures are proportionate to the risk because it's ok to wait a few hours if it means that my plane won't be hijacked.

Common Denominator of Weak Arguments

Major common denominators coming out of this exercise:

1. Not enough true or accurate information – need more facts!

2. Need more System 2
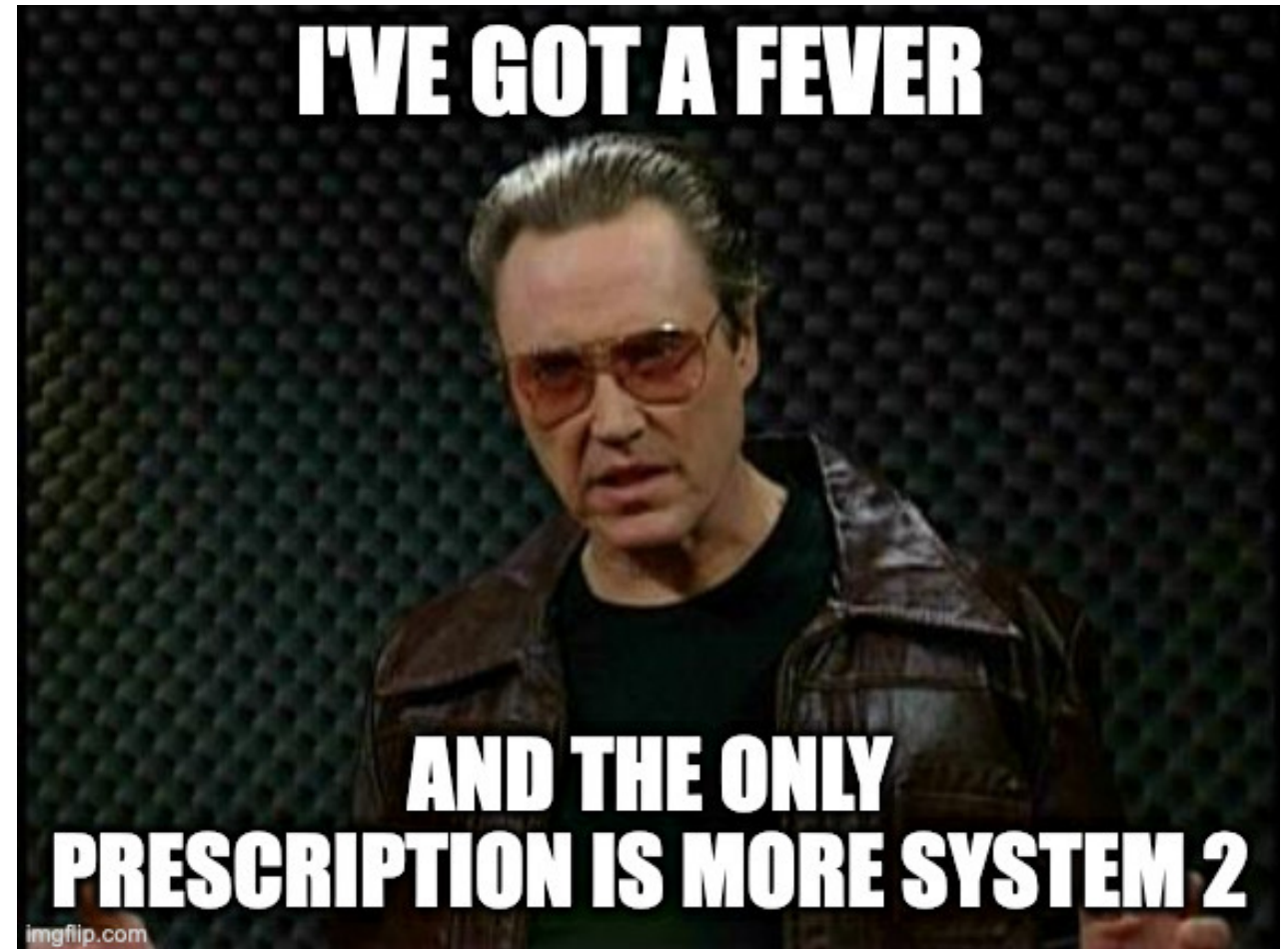
Is Reasoning Worth the Effort?

Using **more rigor** requires **more effort** (system 2 is more work than system 1).

But there are consequences to NOT using analysis techniques are:

- unable to distinguish between what is true and what is not.
- get things wrong, which will lead to waste, etc.

If our beliefs don't match up with the world, we make bad decisions.

System Options (Revisited)

System 1: Automatic decisions

System 2: Scientific method [controlled environment]

But also…

System 2: Everything else (i.e., your job)

We need to use all reasoning types…

with emphasis on what is **plausibly** true.

Science: Focus On Generalizing
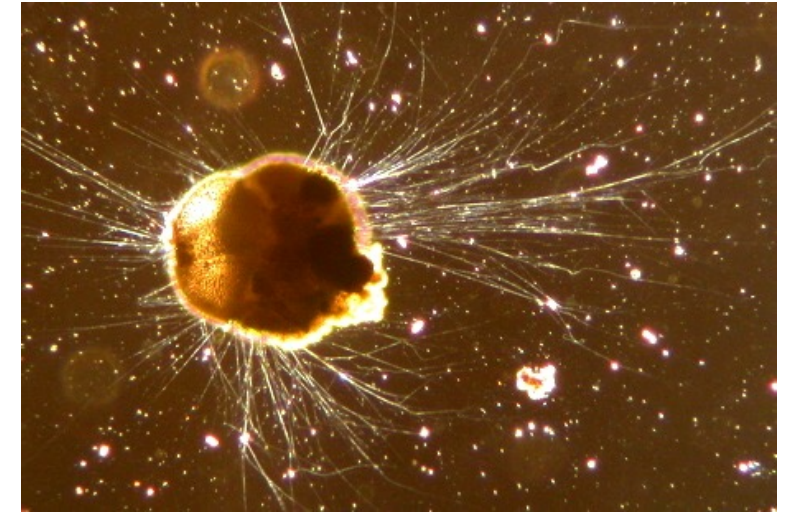
Science: Specific Experimental Context

Scientific data analysis techniques are sometimes only relevant:

- in a **very specific experimental context**
- on **certain types of data**

Now that data is so much more prevalent and usable, we need to **grow and adapt** these techniques

We might need to break out of the 'science mindset' (**without sacrificing rigour in the process**)

Science: Variables Types

In an experimental setting, we usually work with:

- **control/extraneous variables:** we do our best to keep these controlled and unchanging while other variables are changed;
- **independent variables:** we control the values of the variable, and we suspect that they influence the dependent variables;
- **dependent variables**: we don't control the values, they are generated during the experiment, and presumably are dependent on everything else.

How do these translate over to other datasets?
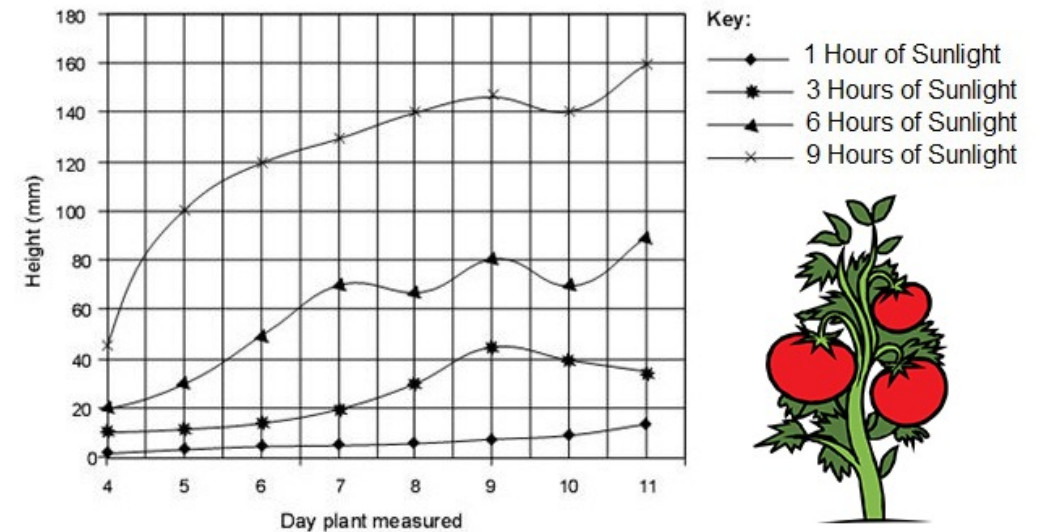
Same Data, Different Conclusions

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.

Referees are **three times as likely** to give red cards to dark-skinned players

Statistically significant results showing referees are more likely to give red cards to dark-skinned players

Twice as likely

ONE RESEARCH TEAM

95% CONFIDENCE INTERVAL

Equally likely

Non-significant results
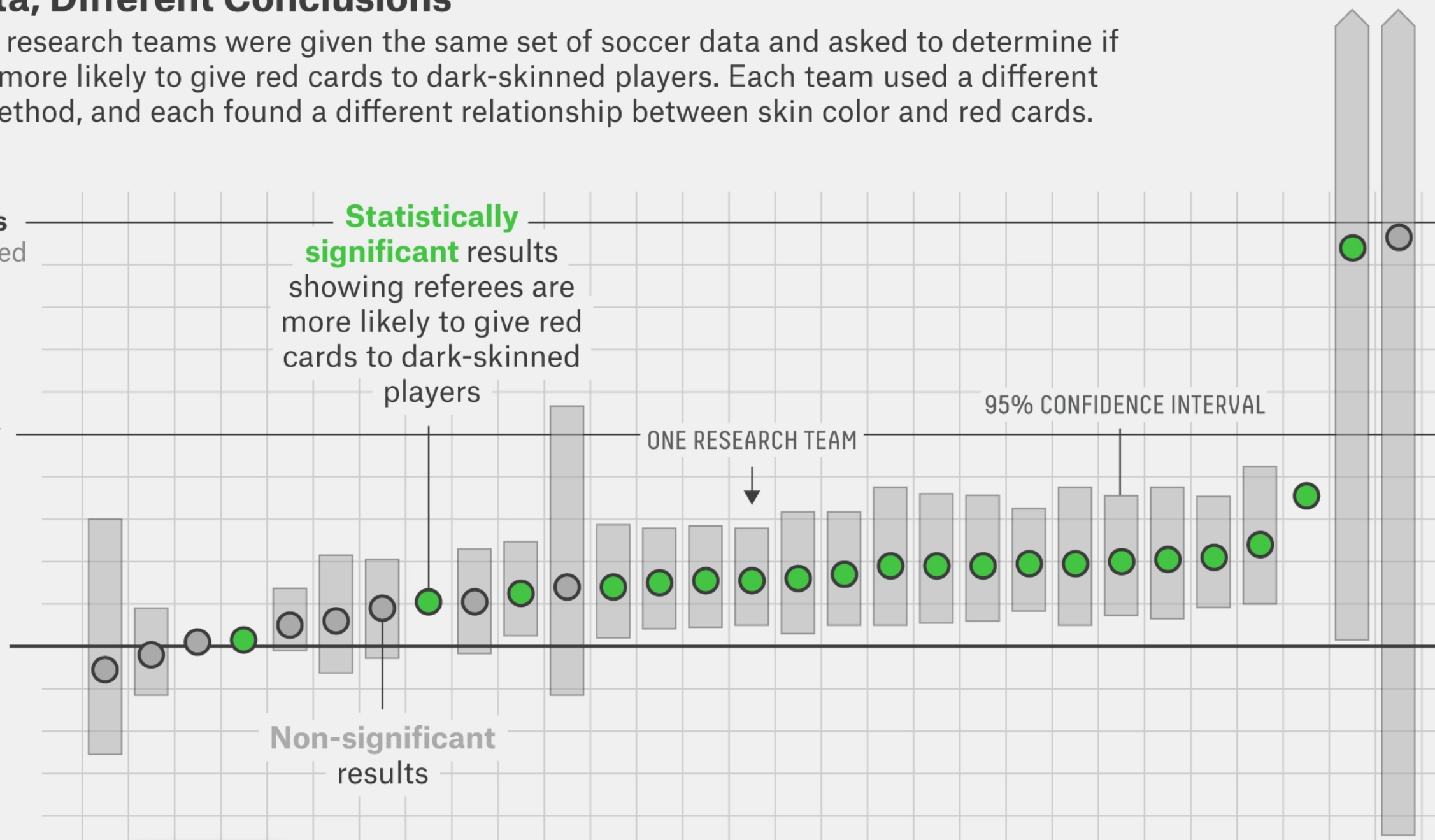
Analytics: Immediate and Focused

Analytics: Analysis Paralysis

We may need to make a decision with less than complete information. What is the risk of not deciding vs. the risk of making a less-than-perfect decision?

Analysis paralysis is caused by overthinking a situation and worrying about the outcome at the expense of decision-making. It is perfectionism, taken to an extreme.

"It doesn't matter in which direction you choose to move when under a mortar attack, just so long as you move. Decisions are never final for the simple fact that change is never absolute. Rather, change is ongoing. To stay competitive and progress at the rate of change requires adaptive decisions that can be iterated and improved upon on the fly." [Jeff Boss, Forbes]

Analytics: Avoiding Analysis Paralysis

1. Recognize it

2. Prioritize the decisions

3. Take a break

4. Ask for advice

5. Make small, quick decisions

6. Set a deadline

7. Understand your goals

8. Limit your information intake

Source: Indeed

What Is Your Analysis Goal?

Do you want to:

- Carry out actions based on what is in your data (maybe not analysis?).
- gain a deeper understanding of something **specific** (specific individuals? A specific group?).
- come to some **general** conclusions that extend beyond the specific.

Local vs. Global

Here vs. Everywhere

Past/Present vs. Future

Situational Awareness vs. Contingency Planning

Typical Sequence of Reasoning

1. Start with premises: knowledge/assumed true beliefs

2. Carry out reasoning

3. Reach conclusions: new knowledge/potentially true beliefs

This approach can also be used to generate a **logical argument**

You Already Do This, Informally

Suppose I pause at the top of a set of stairs with an armful of stuff. What argument might be playing out **unconsciously**...?

- IF I have too many things in my hands, THEN I can't hold on to the railing going down the stairs
- IF I don't hold onto the railing, THEN I might stumble
- IF I stumble, tTHEN I might drop my stuff to stop myself falling down the stairs
- IF I drop my stuff, THEN some of it might break
- IF my stuff breaks, THEN then I'll be sad

CONCLUSION: I currently have too many things in my hands

You Already Do This, Informally

How do we act on such a conclusion?

- Because I have too many things in my hands, I might drop them on the stairs and break some of them
- This would make me sad :(
- Instead, I could choose to make two trips so I can hold on to the railing
- if I make two trips instead of one, this doesn't mean I won't drop something and break it, but it does increase my confidence that I won't drop something

Decision and Action: I will split the load into two parts and make two trips

or …

Nah, that's not likely to happen; I'll tough it out and make one trip.

Reasoning Vulnerabilities

Conspiracy theories **mindset:** individuals jumping to invalid conclusions because they cannot reason and/or recognize bad evidence.

Is it **plausible** that there are microchips in the COVID vaccine? How would you gauge the degree of plausibility?

Thought exercise: you are given a stable of deductive logicians and a stable of debaters to help you make decisions. Which would you chose? Is any of them of use to you?

Formal Rigorous Reasoning

Mathematicians and philosophers developed **formal methods** to bring **rigour** to informal reasoning – we can think of these as reasoning tools.

Using these tools increase our chances that we will **end up with true statements**, in which we can feel confident (if not always 100% so).

Without rigor, we can succumb to **biased reasoning**, which prevents us from reaching either **true conclusions** or **justified conclusions**.

Formal Reasoning Techniques

INDUCTIVE, PLAUSIBLE, DEDUCTIVE, ABDUCTIVE, ANALOGICAL REASONING

FURTHER SPECIALIZED TECHNIQUES: SCIENTIFIC METHOD, STATISTICAL REASONING, MATHEMATICAL AND COMPUTER MODELLING

EVIDENCE-BASED ANALYSIS, WHICH MAY BE MORE OR LESS TECHNICAL

Formal Reasoning Techniques

Reasoning strategies:

- deducing new facts from existing facts (**deductive** reasoning)
- generalizing from examples (**inductive** reasoning)
- reasoning to the best explanation (**abductive** reasoning)
- using analogies and models (**analogical** reasoning)

These last three techniques are examples of plausible reasoning – you are not guaranteed to reach the truth, but you are increasing your level of certainty.

Plausible Reasoning
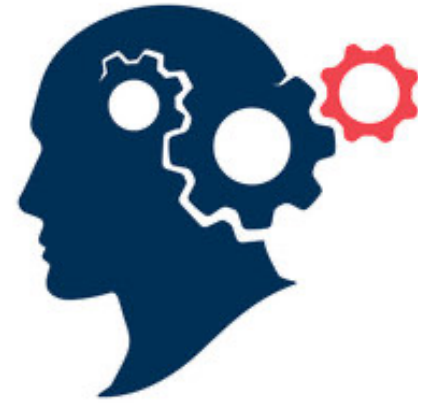
Consider the following scenario [Jaynes, 2003]:

- you are walking down a deserted street at night
- you hear a security alarm, look across the street, and see a store with a broken window
- someone wearing a mask crawls out of the broken window with a bag full of smart phones

What would your first system 1 conclusion be?

What would your system 2 conclusion be?

Plausible Reasoning

Say you concluded that the person crawling out of the store is stealing merchandise from the store. How do you come to that conclusion? It **cannot** come from a logical deduction based on evidence. Indeed,

- the person crawling out of the store **could** have been its owner who,
- upon returning from a costume party, realized that they had misplaced their keys
- just as a passing truck was throwing a brick in the store window,
- triggering the security alarm, after which
- the owner then went into the store to retrieve items before they could be stolen,
- which is when you happened unto the scene.

The original conclusion is not **deductive**, but it is at least **plausible**.

Deductive vs. Plausible Reasoning

Plausible reasoning:

If A is true, then B is more plausible
B is true

A is more plausible

If "the person is a thief" (A is true), you would not be surprised to "see them crawling out of the store with a bag of phones" (B is plausible).

You do "see them crawling out of the store with a bag of phones" (B is true).

Thus, you would not be surprised if "the person were a thief" (A is more plausible).

Exercise: Reasoning Strategies

Consider the following items found in a briefing note:

- The last 7 times pipelines were constructed in caribou territories, populations decreased in the territory

- Biologists created a map showing the caribou migration paths; based on this map, we conclude that placing the pipeline over the territory will not interfere with caribou migration.

- Pipelines have not affected geese populations; as they and caribous are both social animals, the pipeline will not affect the caribou population.

- Biologists have shown that caribous are not scared of large objects; if caribous are not scared, their breeding habits will not be affected; as pipelines are large objects, constructing this pipeline will not affect the breeding habits of the caribous on the territory.

Identify the reasoning strategies being used in each of these arguments. What would you conclude about the pipeline construction and the effect it might have on caribous, using plausible reasoning? What additional information would you need/want, before drawing a conclusion?

Exercise: Plausibility

In Tom Stoppard's 1966 play Rosencrantz and Guildenstern are Dead, the main characters bet on coin flips. Rosencrantz wins by flipping heads 92 times in a row.

This result is of course not impossible, but is it plausible? If this happened to you, what would you conclude?

Reasoning Machines

Another advantage of formalizing reasoning is that it allows us to **automate it**, by programming it into computers.

We can reframe reasoning as a process that **takes inputs** (premises/observations/ evidence) and **produces outputs** (conclusions).

By automating this process, we can get machines to carry out reasoning for us.

The result could be more **reliable**, **dependable**, **consistent**.

BUT: **garbage in, garbage out**! Weak premises in, possibly false conclusion out!

Rise of Analysis?

Statistics and other formal techniques (late 1800s): techniques to draw rigorous conclusions from (scientific) data to support research

Information Age (1990s): automated, electronic data collection

1970s

now

late 1800s

1990s

Computer Age (1970s): computers move into the public sphere

Where are we now?
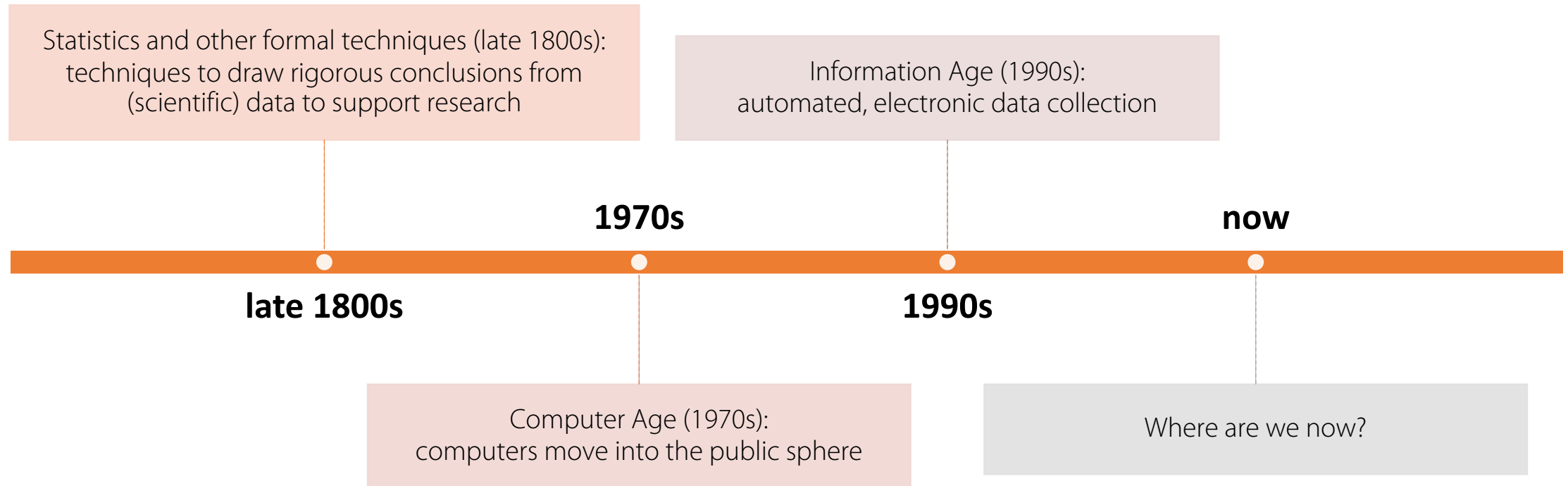
Analysis in the Pre-Digital Age vs. Analysis in the Digital Age

Then:

- only people could carry out the activity of analysis and the components of an analysis process

Now:

- we can distill the essence of an analysis process into an algorithm
- we can automate analytical activities and its supporting process
- we have analysis machines

Then:

- a given analysis of a situation was typically a one-time, one-off activity
- a single person might carry out 'an analysis' and then move on

Now:

- we can expect that we will probably want to repeat variations of the same analysis over and over on new data that is streaming in on a regular basis
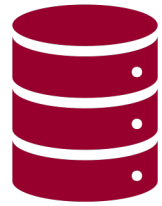
Module 3
People, Data Ethics, and the Law

Data Analysis: Why Me?

In general, even those who are not analysts must be able to talk to the analysts!

- **data engineers:** kitchen designers, who need to know what the …

- **data analysts:** cooks will be doing, who in turn need to know what …

- **data translators:** customers want to eat, who might work with the …

- **data presenters:** menu designers/waiters, who ensure the customers know what they are getting, all under the guidance of …

- **data project lead:** chef/sous-chef, who keep the kitchen stocked and manage all the moving pieces.

Data Teams and Decision-Making

What **team** are you on?

Does the information and data at **your disposal** allow you to make **timely** and **effective** decisions?

How often do you need the **support** of the data team?

To what extent does their service allow you to make timely and effective decisions?

What are **significant barriers** when trying to use data in your decision-making? What are some other **challenges** of data-driven decision-making?

Source: Google

Data Collection Data Storage Data Preparation Data Analysis Data Presentation
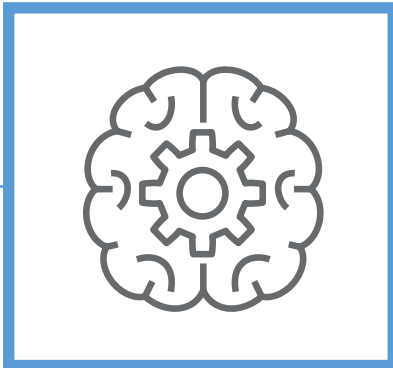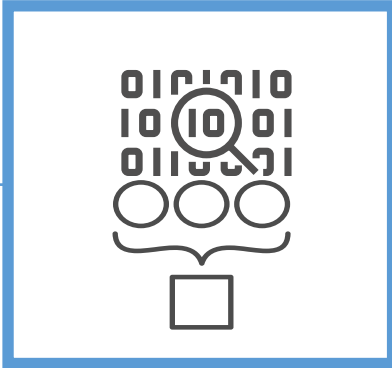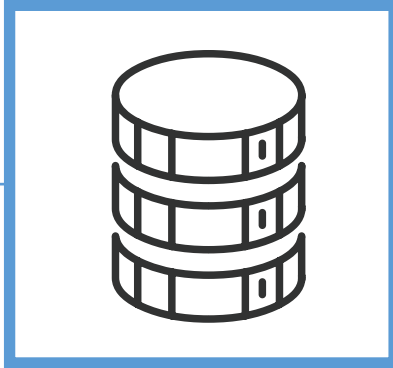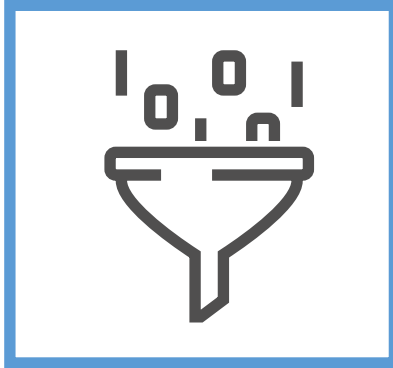
Data Analysis as a Team Sport

Data Collection Data Storage Data Preparation Data Analysis Data Presentation
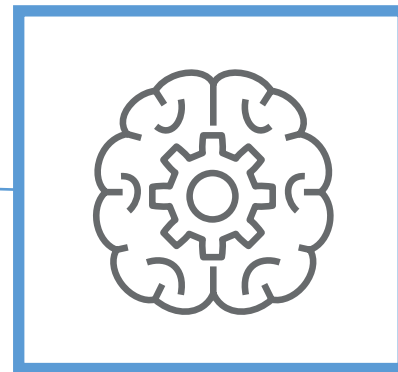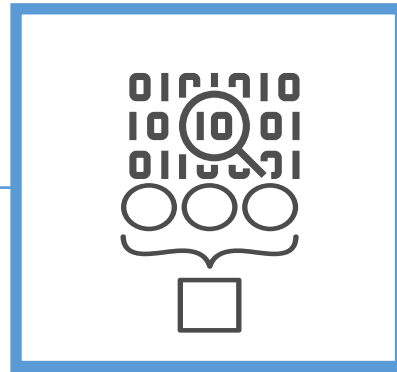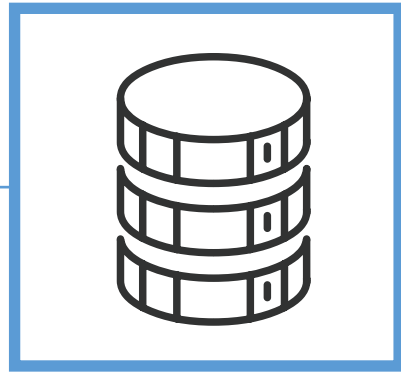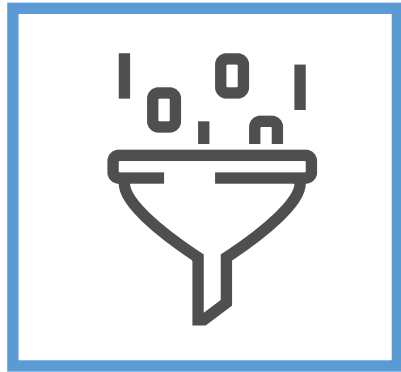
What roles support this pipeline?

Data Collection   Data Storage   Data Preparation   Data Analysis   Data Presentation
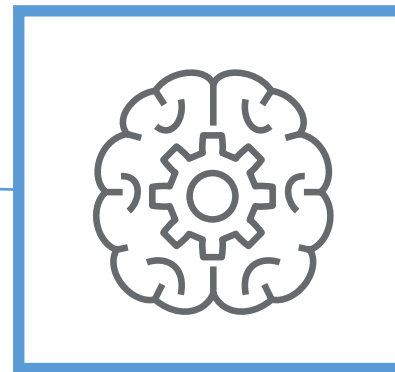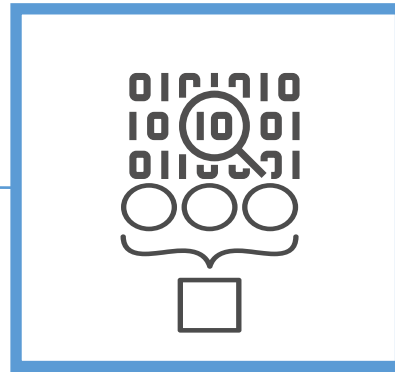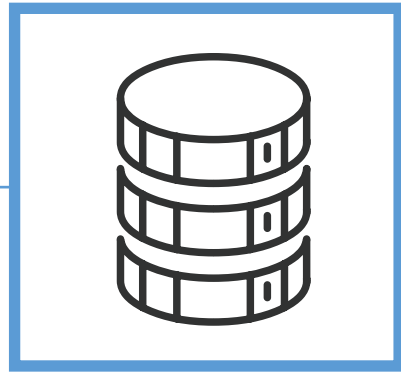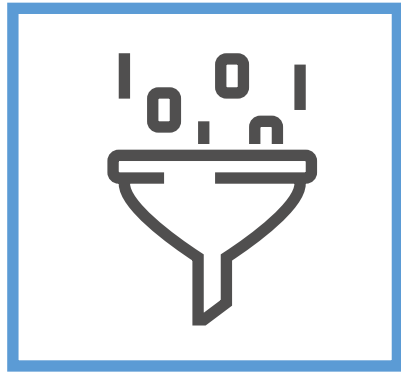
Ratio of other team members to analysts is ~8 : 1

Data Collection | Data Storage | Data Preparation | Data Analysis | Data Presentation
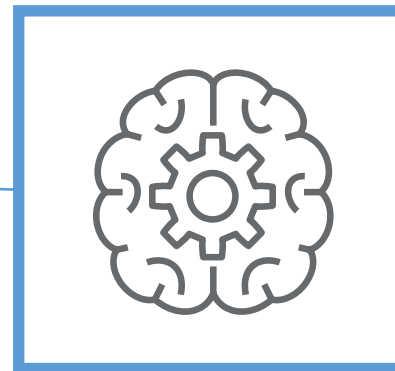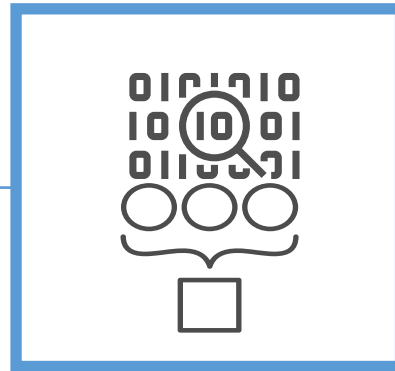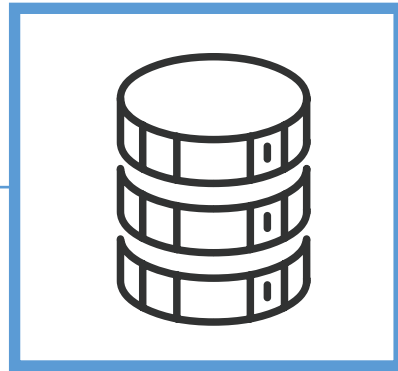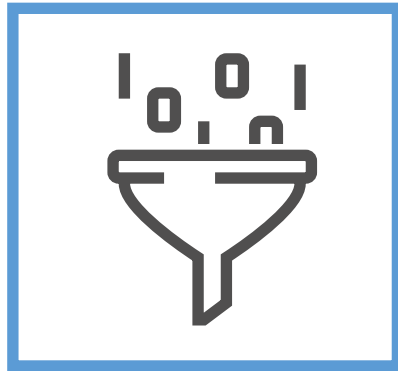
Ratio of analysis to other activities is ~1 : 8

Data Team Roles and Tasks (I)

Data Engineering: data infrastructure design and implementation – IT and DevOps heavy

Data Collection: design of data collection strategies and implementation of data collection tools

Data Architect/Manager: data storage and data architecture design and implementation

Data Preparation (in collaboration with **data collection**): data managers and data analysts process the data into an analysis-ready state; this pivotal task can be automated, but must be audited regularly

Data Analysis: analysts determine how to extract actionable insights from data and information; they design and implement algorithms to automate the analysis

Data Team Roles and Tasks (II)

Data Pipeline UX Expert: interface design, user experience

Data Communication: data visualization, data presentation

Subject Matter Experts: know much about the situation; understand what is important, what data could provide insight, how to interpret/apply the analysis results

Business or Organization Strategy Experts: hold the picture and know where the organization wants to head; share this information with the team

Project Lead: keep everyone on track and working together

Data Translators: know how the different pieces of the pipeline work at a high level, know a useful amount about the subject matter; connect people and help them talk to each other

Specialist:
in-depth knowledge of a topic

Generalist:
broad knowledge of a topic

Amateur:
passion project

Professional:
paid and expected to perform at a certain level, with a certain level of skill (obligations)

Let's think of desktop data analysis as "**semi-pro**"

A team can still be built in an amateur or semi-pro situation

It might be less of a formal arrangement, but the different roles can all still come into play

Some Useful Analogies

	Medicine	Cooking	The World of Cars
Amateur	Everyone	Home Cook	Car Owner
	First Aider		Car Hobbyist
Semi-Pro	Paramedic	Bake-Sale Folks	Semi-Pro Racer
			Gas Station Mechanic
Professional	Hospital Director	Restaurant Owner	Garage Mechanic
	Nurse	Chef	
	Doctor: GP, Specialist	Pastry Chef	Body Shop Specialist

Where Do Your Team Members Fit In?

Preliminary questions:

1. What part of the pipeline is most appealing to you?
2. Do you prefer designing or implementing what someone else has designed? Or both?
3. Are you a generalist (big picture) who likes to know a little bit of everything, or a specialist (detail-oriented person)?
4. Can you currently write computer programs/scripts (or do you want to be able to do so)?
5. Do you have a math or statistics (STEM) background?
6. Do you like working with IT technologies?
7. Do you like to facilitate communication between different members of a team
8. Do you have a deep knowledge of your organization's operations or subject matter
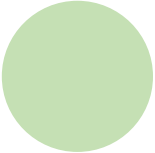9. Do you have a deep knowledge of organizational goals? Do you like strategy?
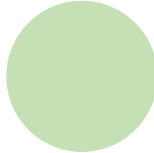
Generalists vs. Specialist: You Can't Do It All! (Emma)

Generalist: multi-purpose, can communicate across lanes

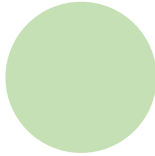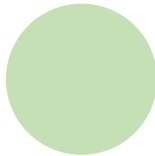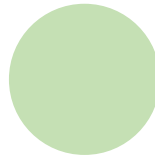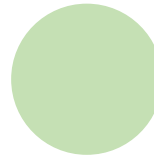|	Data Collect	Data Store	Data Prep	Analysis	Presentation	Subject Matter	Organization
Amateur	implementation	implementation			user experience	initially	initially
Semi-Pro	questionnaire design	data architecture (design)		statistics	design implementation	post start of project	post start of project
Professional	architecture for machine learning		unstructured, semi-structured	machine learning / unsupervised learning, SNA			

Generalists vs. Specialist: You Can't Do It All! (Jesse)

G

S

Hybrid: focus on a tight range of data tasks (other expertise in non-data fields)

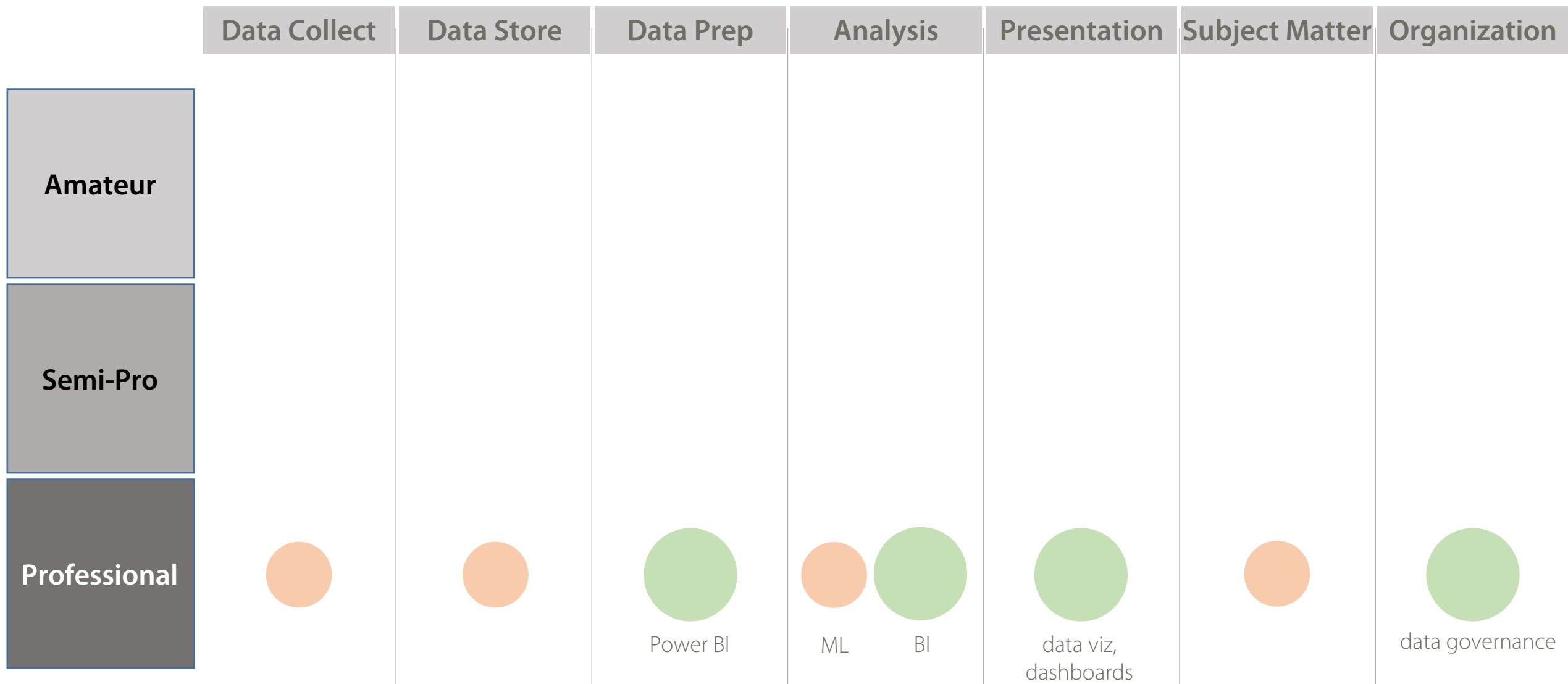|	Data Collect	Data Store	Data Prep	Analysis	Presentation	Subject Matter	Organization
Amateur							
Semi-Pro					storytelling	A.I. as a service, human factors	
Professional				needs analysis		legal matters and GoC	management

Generalists vs. Specialist: You Can't Do It All! (Ibrahim)

Specialist: focused, can communicate to certain audiences

	Data Collect	Data Store	Data Prep	Analysis	Presentation	Subject Matter	Organization
Amateur							
Semi-Pro							
Professional	survey sampling, web scraping		data wrangling, data cleaning	EDA, statistics, OR, ML	data visualization		
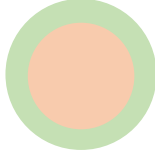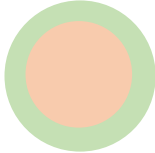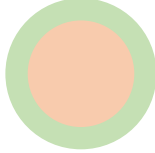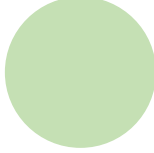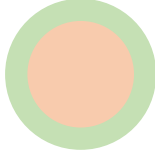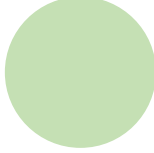
Generalists vs. Specialist: You Can't Do It All! (Li)

G

S

Hybrid: focus on wide range of data tasks

	Data Collect	Data Store	Data Prep	Analysis	Presentation	Subject Matter	Organization
Amateur							
Semi-Pro							
Professional	●	●	● Power BI	● ML ● BI	● data viz, dashboards	●	● data governance

Exercise: Where Do You (or Your Team) Fit?

	Data Collect	Data Store	Data Prep	Analysis	Presentation	Subject Matter	Organization
Amateur							
Semi-Pro							
Professional							

G   S

Experience vs. Fresh Perspective

Starting from scratch (**red team**) / buying ready-made solutions/people (**green team**)

Fresh Perspectives	**Biased Viewpoints of Data**
No Analytic Anchors	**Susceptible to 'Been there Done That'**
Natural Anomaly Detection	**Can Overlook Areas via Assumptions**

Can triple-check your assumptions and methodologies by employing the **red team** to challenge the perceived conclusions

Time to pay-off might be long

Source: The Head Game

Differentiating: IT and Data People

IT

Servers/cloud

Connection to data

Access and authorization

DS

Code environments

Code repositories

Project hierarchies

Open source packages

Production models and analysis

Potential Criteria for Hiring Data 'People'

Curious

Coachable

Communicative

Eager

Knowledgeable

Analytical

What are Ethics?

Broadly speaking, ethics refers to the **study** and **definition** of **right** and **wrong** conducts.

We all have a personal ethical system, don't we?

- be honest
- be fair
- be objective
- be responsible
- be compassionate?
- etc.
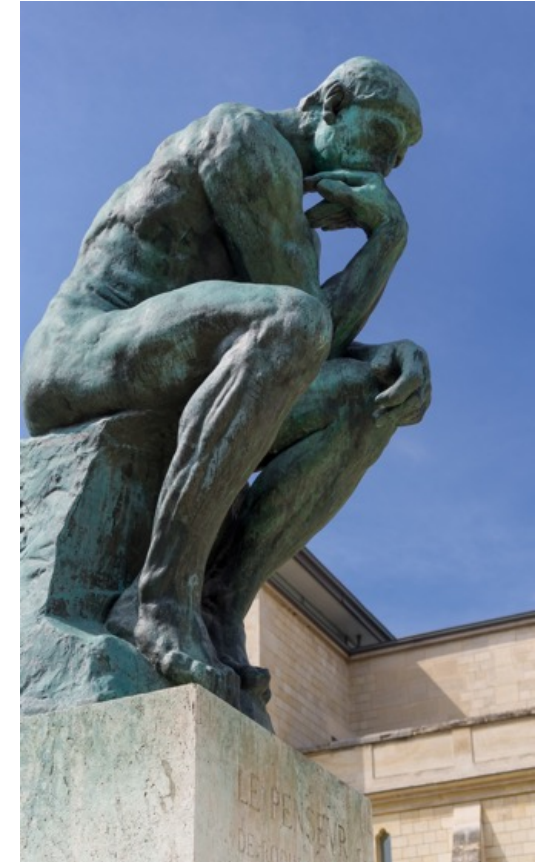
Upcoming Exercise:
Veil of Ignorance

All people are biased by their situations, so how can people agree on the rules to govern how the world should work?

We should imagine we sit behind a veil of ignorance that keeps us from knowing who we are and from identifying with our personal circumstances.

Case Study

Your company is always looking for the most talented people, especially for technical positions.

Corporate policy supports diversity and inclusion.

The hiring process is time-consuming, and you are concerned about personal biases of panel members influencing the decisions.

With the help of an outstanding AI team, you automate this process.

The AI-assisted processes finds a high percentage talented people, who fit into the organizational culture, and who like their jobs (low turnover).

Case Study: Amazon Hiring A.I.

More likely to get hired if your name was Jared and you played lacrosse.

A.I. was behaving in a biased manner, not recommending women be hired.

Amazon was not confident they could remove the bias or identify future biased behaviours, so they opted to discontinue the project.

Let's discuss this (and the other case studies).

Amazon ditched AI recruiting tool that favored men for technical jobs

Specialists had been building computer programs since 2014 to review résumés in an effort to automate the search process

▲ Amazon's automated hiring tool was found to be inadequate after penalizing the résumés of female candidates.
Photograph: Brian Snyder/Reuters

The Threat

In her book about data power, Dr. Cathy O'Neil presents several cautionary examples and tales.

"A computer program could speed through thousands of résumés […] and sort them into neat lists […]. This not only saved time but also was marketed as fair and objective. After all, it didn't involve prejudiced humans digging through reams of paper, just machines processing cold numbers. […]

The math-powered applications driving the data economy were based on choices made by fallible human beings. Some of these choices were no doubt made with the best intentions. Nevertheless, many of these models and algorithms encoded human prejudice, misunderstanding and bias into the software systems that increasingly managed our lives."

WEAPONS OF MATH DESTRUCTION

HOW BIG DATA INCREASES INEQUALITY AND THREATENS DEMOCRACY

CATHY O'NEIL

'Wise, fierce and desperately necessary'
JORDAN ELLENBERG
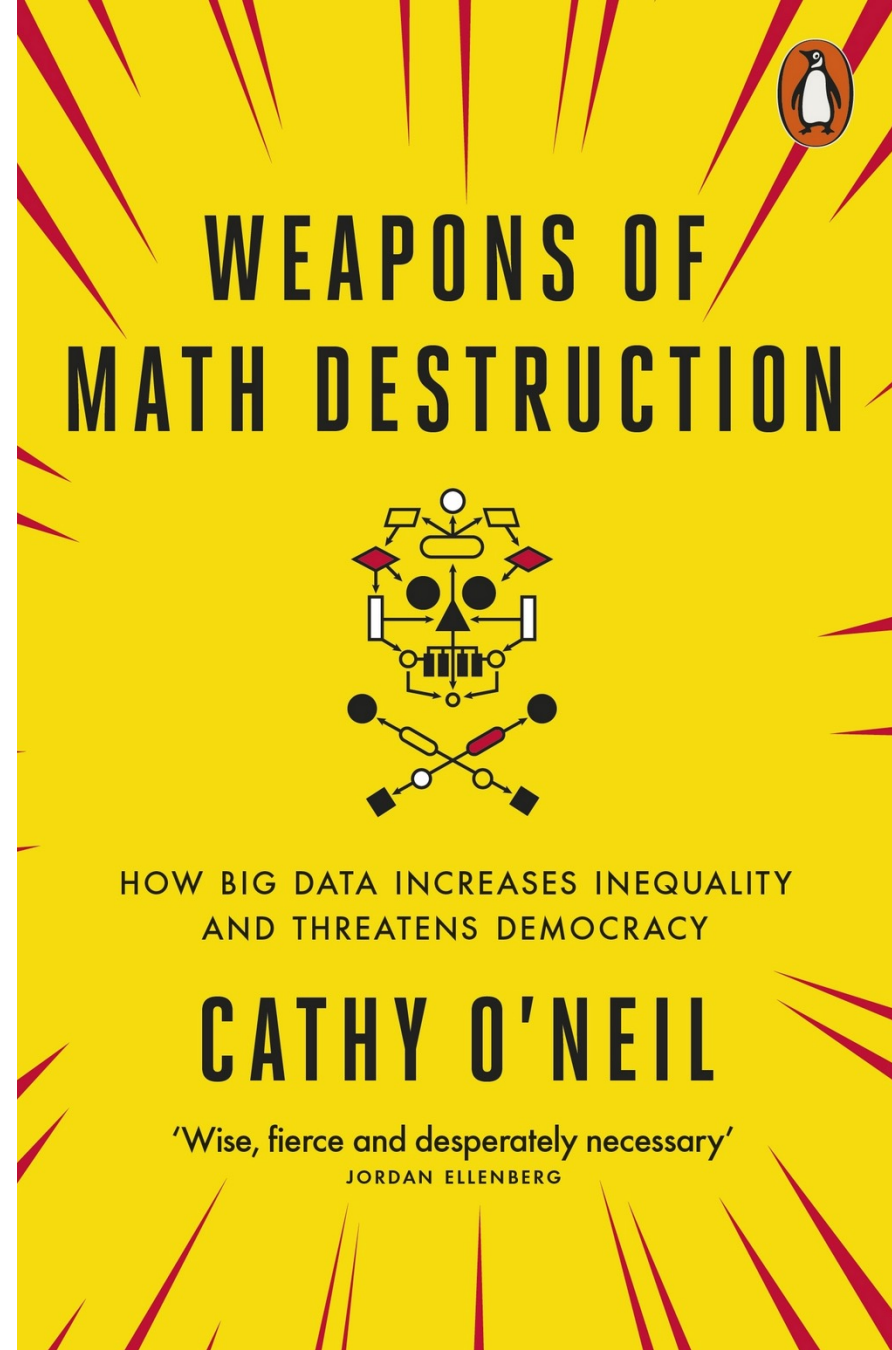
Ethics in the Data Context

Data ethics questions:

- 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 not be used in analyzing personal data?

- Should some data be publicly available to all researchers?

Are there lessons to be learned from the First Nations Principles of OCAP®?

- **ownership**, **control**, **access**, **possession**.

Exercise: Veil of Ignorance

Assume the following:

- a bank has an obligation to increase shareholder value
- it is considering using a new A.I.-driven decision-making process for loan applications, with the goals being to save staff from having to complete tedious tasks, and to reduce the defaulting rate
- the bank has access to proprietary and public information about loan applicants

Not knowing what position in society you might hold (e.g., whether you would work for the bank, or be applying for a loan), do you have questions about this proposal?

What ethical principles would come into play?

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.

Does anything else come to mind?

Emerging Legal Trends

Canada

- GoC: use <u>Algorithmic Impact Assessment</u> prior to the production of any Automated Decision System
- Privacy Commissioner (Personal Information Protection and Electronic Documents Act) Proposed amendment:
  - Defines automated decision systems any tech that assists or replaces the judgment of humans.
  - Need to give people an explanation of the prediction/recommendation, and how their personal info was used.

Europe

- General Data Protection Regulation (GDPR)
  - Article 22: not subject to a decision based solely on automated processing (with exceptions)
  - Article 15: if subject to such a decision, have right to meaningful information about the logic involved.

Legal Questions

Profiling:

- are you using personal data to draw inferences that are unfair, unethical or discriminatory

Surveillance:

- are people being placed in a perpetual line-up?

Liability:

- are you liable for what an A.I. does?

Case Study: IRCC Programs

Immigration, Refugees and Citizenship Canada (IRCC) has been quite pro-active at looking for ways to improve their services through judicious use of automation and A.I.

A recent pilot project used automation and machine learning techniques to fast track some applications.

Planning highlights

Client experience, automation and experimentation

In recent years, the Department has been working to advance its network of online tools to support both clients and staff and improve services, particularly as they relate to application processing times. These efforts will continue in 2018–2019 across key business lines as the

BOTS AT THE GATE

A HUMAN RIGHTS ANALYSIS OF
AUTOMATED DECISION-MAKING IN CANADA'S
IMMIGRATION AND REFUGEE SYSTEM

In the end, it's not (just) about the analysis!

In a professional setting, analysis **will not be happening just for the sake of analysis**.

In an applied setting, analysis supports business goals.

- **cooking analogy:** in the cooking world, the customer is royalty!
- **car hobby analogy:** some people work on cars for fun, and some people work on data for fun… but in the end, it's about the owner/driver of the car and what they need their car for.

Don't lose sight of your end goals.

Who knows what the end goals are? Spoilers: **you might not!**

Module 4
Business Intelligence and Analytics

What Do We Mean by "Intelligence"

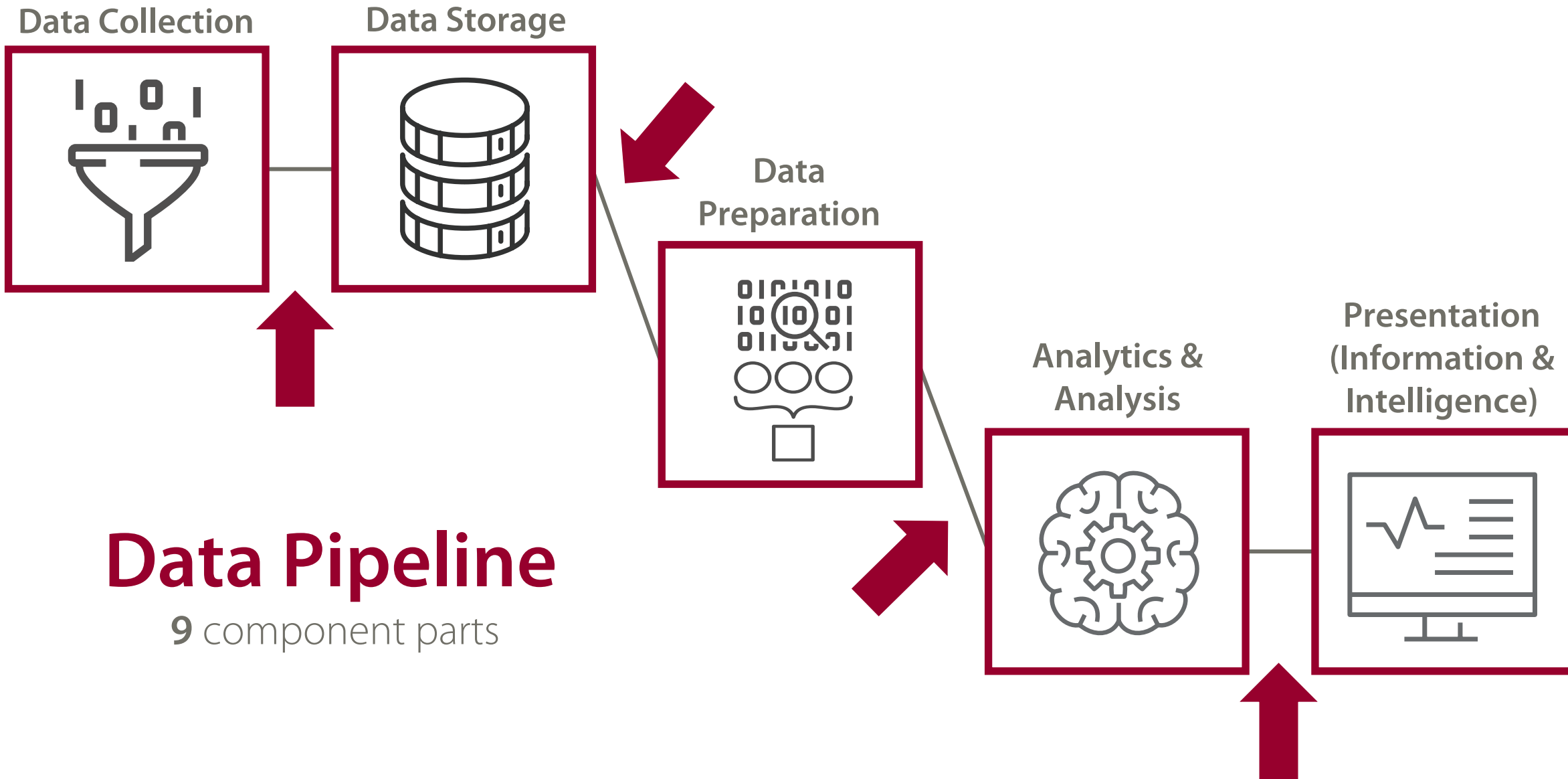| Data Analysis	Data Analytics	Business Intelligence
Typically seen as subcomponent of Data Analytics	Scope broader than Data Analysis	Uses the outputs of analytics and analysis
Various methods used to provide historical view of data	Includes methods to include future predictions	Uses past view and future predictions as situational awareness to make decisions
Process of observing data then arranging / visualizing to extract useful information	Includes various methods that are useful in the management of data	Takes analysis and analytics methods in a business context

What Do We Mean by "Analytics"

We often use analytics and analysis **interchangeably**, but typically (not always) analysis is seen as a **subcomponent** of analytics

We often think of them as methods that give us **useful information**

But analytics has come to represent a **broad range of tools** for processing data, covering data warehousing, enterprise information management, enterprise performance management, and governance

Some "methods" provide information/intelligence; other "methods" transform, clean, manipulate, and store the data **(data pipeline)**

Data Collection

Data Storage

Data Preparation

Analytics & Analysis

Presentation (Information & Intelligence)

Data Pipeline

9 component parts

Data Analytics

We often see analytics talked about in terms of "**modes**"

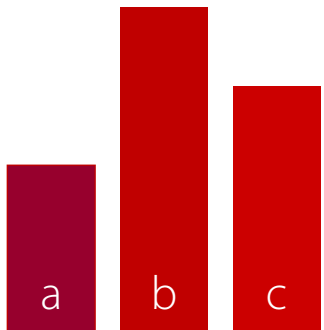Some modes can be more "**valuable**" than others depending on context

We might **never be able to achieve certain modes** (e.g., Predictive) because of issues with our data

These modes are effectively **technology-dependent**, the higher value modes requiring more sophisticated tools and methods

Analytics Modes

Analytics can be broken down into four core **key modes:**
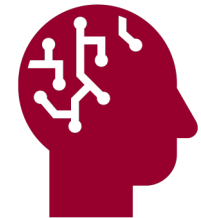
Descriptive	**Diagnostic**	**Predictive**	**Prescriptive**
Show **what** happened	Explain **why** something happened	Guess **what will** happen	Suggest **what should** happen

Low Value
Low Difficulty

⟵─────────────────────⟶

High Value
High Difficulty

Analytics Modes

Descriptive: Budget report providing a snapshot at a particular time

Diagnostic: Why do I have less in my budget than I originally thought?

Predictive: How much will I have in my budget next quarter?

Prescriptive: How should I change my spend profile to avoid deficit?

Descriptive

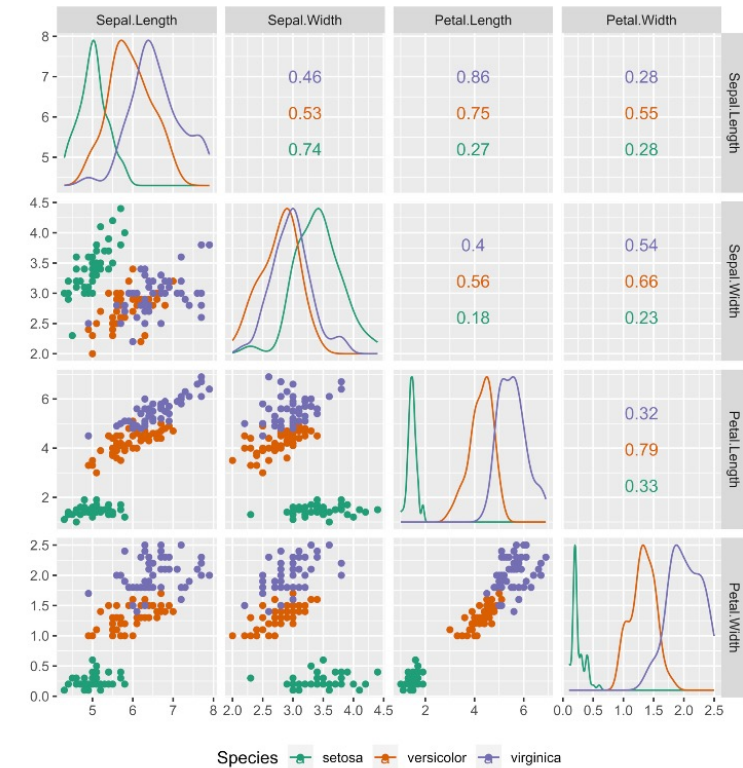Looks at data statistically to show what happened **in the past**

Helps a business understand its performance by **providing context** to help stakeholders interpret information

Typically, in the form of **graphs**, **charts**, **reports**, **dashboards**

Intelligence is usually **lagging** (data is typically not current)

Intelligence is usually **focused on a subset** or **single dataset**

Datasets very often **not appropriate/large enough** to infer statistical significance or to perform reliable tests

Diagnostic

Provides deeper analysis to answer: **why did this happen**?

Very often uses **formalized root cause analysis tools** (Ishikawas, 5 Why, Taproot, Causal Tree, etc.)

"Difficult" to automate and prone to "human" variability

Intelligence often **conflates correlation with causality**

If data sets are appropriate, AI/ML can **provide potentially significant value** (e.g., decision trees)

Predictive

Applies various methods (stats, machine learning, etc.) to **model and extract patterns** from past data

Applies models and patterns to **current data** to predict **what might happen next**

Intelligence can suffer from **model drift**, poor choice of training data for modeling, and many other issues (including dataset size)

Very popular, but **easy to mess up**
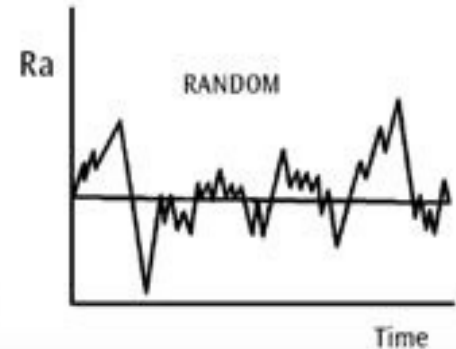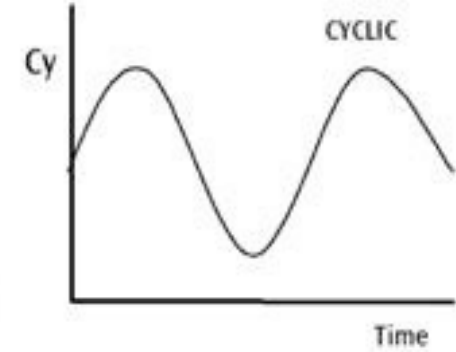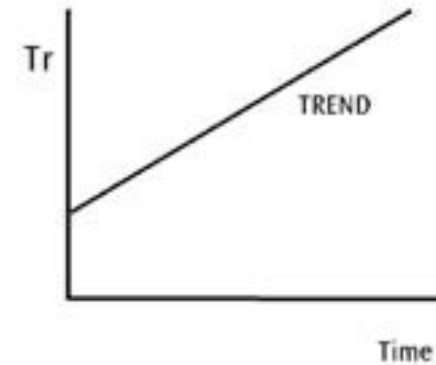
Prescriptive

Takes advantage of approaches that can suggest **various courses of action**

Intelligence is very often **probabilistic** and thus **potentially subjective/hedged** (65% chance of rain, so maybe take an umbrella?)

Intelligence often used in **augmented decision-making framework**

Holy grail of analytics

Prisoner 2

	Cooperate	Defect
Cooperate	3, 3	0, 5
Defect	5, 0	1, 1

Prisoner 1

What is Business Intelligence?

Many definitions, most of which are at best inconsistent and at worst contradictory

We should compare to other forms of intelligence gathering, for example:

- battlefield/combat
- diagnostic (medical)
- disaster response
- sports

What is Business Intelligence?

In all cases, the **actor** (Commander, Doctor, Coordinator, Player, etc.) requires **accurate situational awareness** to make informed and timely decisions

In all cases, the **quality** of the situational awareness has improved with technology

Consequently, the term **Business Intelligence** has evolved to represent a range of technologies that **support decision-makers** within a "business"

The fundamental core of situational awareness is still the key focus regardless of technology

What is Business Intelligence?

Category	Sports Example
Analysis	Home team won 60% past games again same opponent
Analytics	Bookmakers predicting win at 3:1 odds for home team
Experience	All team members > 2 years major leagues
Environment	Playing at home
Priorities	Pre-season game, result does not count in league
Risks	85% chance of rain

History of Business Intelligence

Late 1800s: people started to recognize that they could use data to gain a competitive advantage

1950s: advent of the first business database for decision support

1980s -1990s: computers and data becoming increasingly available - data warehouses, data mining – still very technical and specialized

2000s: trying to take business analytics out of the hands of data miners and other specialists and more into the hands of domain experts

Now: big data, specialized techniques, dashboards, software as a service

1865

1950s

1980-90s

2000s

Today

BI Process
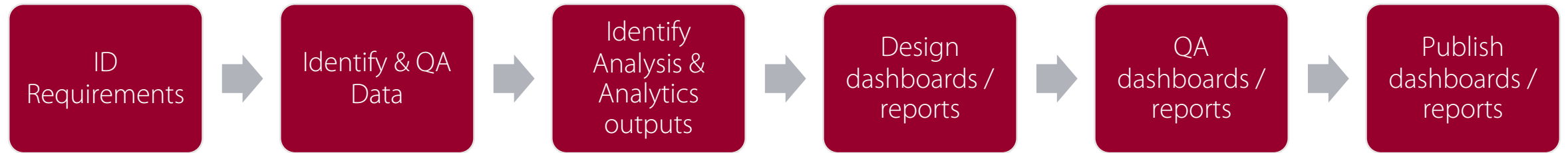
Getting Better at Forecasting

We have seen from our **superforecasting** examples that people can get better at the prediction (forecasting) aspect of business intelligence

What are some human-centered strategies or techniques for improving this skill?

Two possibilities:
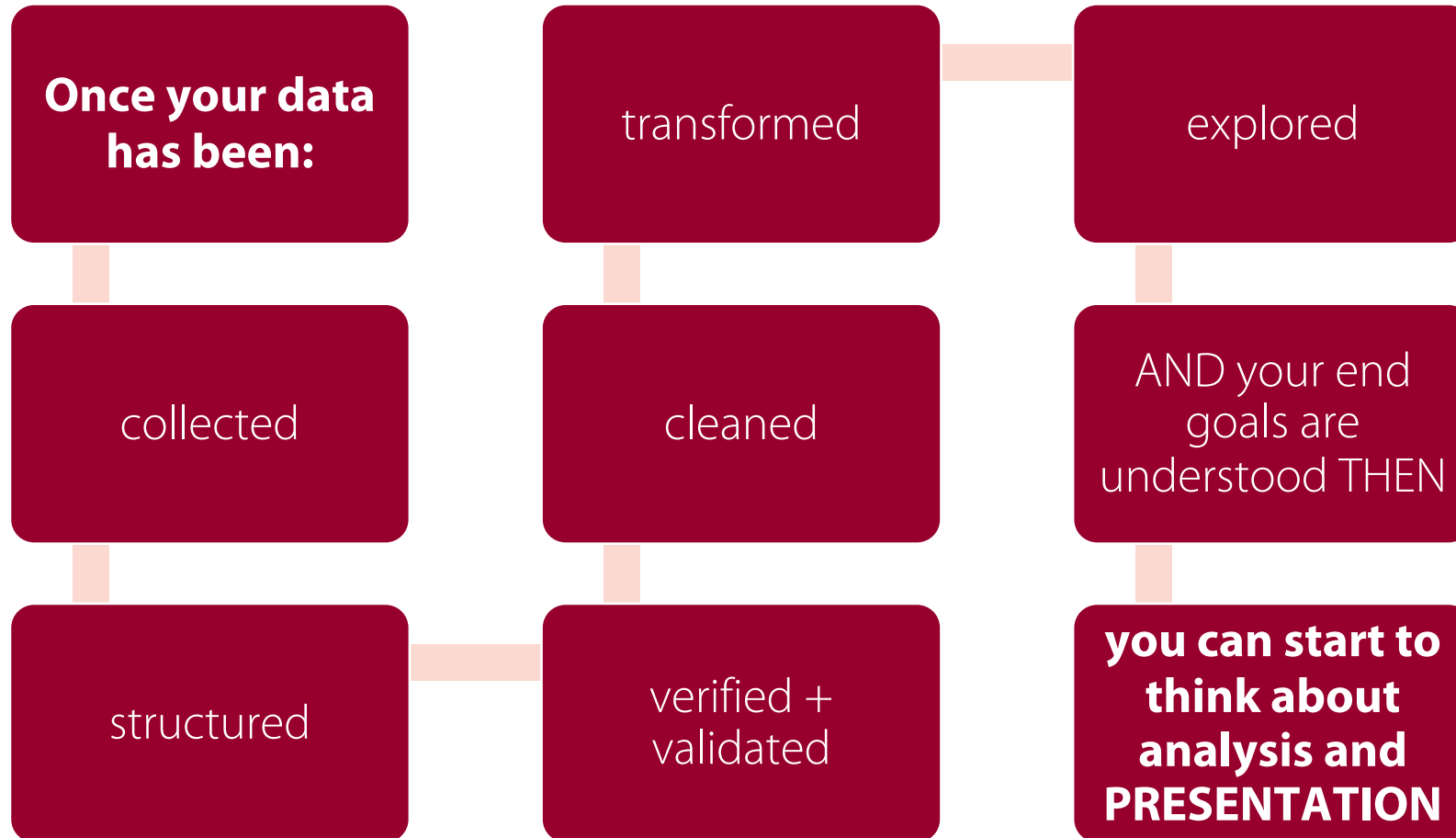- **pre-mortem**
- **backcasting**

What Happens After Analytics?

An analytics result is only as amazing as how well it is communicated.

We cannot overemphasize this enough.

Are We There Yet?

Once your data has been:

transformed

explored

collected

cleaned

AND your end goals are understood THEN

structured

verified + validated

you can start to think about analysis and PRESENTATION

Persuasion

You've made a data/evidence-driven decision: how do you convince people it was the right one?

We've talked about description, explanation, prediction, prescription.

Without **persuasion**, it could prove to be a waste of time.

In a business context, it's not enough to know (or suspect) that something is or will be the case, you need to **communicate** and **convince other people**.
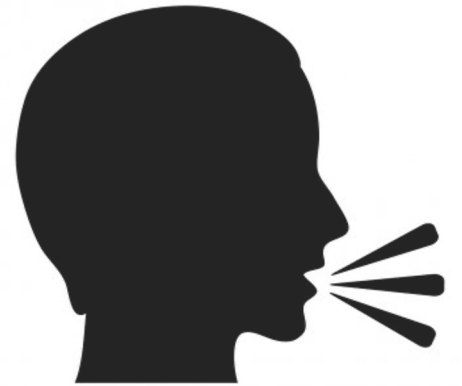
Story Spine: PIXAR

Once upon a time there was _
Every day, _.
One day _.
Because of that, _.
Because of that, _.
Until finally _.

The Story Spine

Exercise: construct a story spine relating to a successful decision made by your organization.

Is this all there is to data storytelling?

Storytelling Risks

"Open any newspaper, watch any TV news show, and you find experts who forecast what's coming. Some are cautious. Most are bold and confident. A handful claim to be Olympian visionaries able to see decades into the future. With few exceptions, they are not in front the camera because they possess any skill at forecasting.

Accuracy is seldom even mentioned. Old forecasts, old news – soon forgotten. The one undeniable talent that talking heads have is their skill at **telling a compelling story with conviction**, and that is enough. Many have become wealthy peddling forecasting of untested value to corporate executives, government officials and ordinary people who would never think of swallowing medicine of unknown efficacy and safety but who routinely pay for forecasts that are as dubious as elixirs sold from the back of a wagon." [Tetlock]
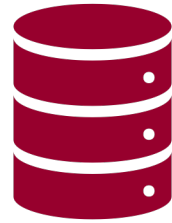
Module 5
Analytics for Decision Support

Alternative – Data Maturity Model

| Data Maturity Level	Description	Likely Decision-Making Criteria
0	No data	Instinct
1	Not enough data	Manual extractions
2	Too much data	Adopting BI platform
3	Analysis	Multiple platforms
4	Learning	Beginning to understand predictive
5	Acting	Predictive/prescriptive

High Level Concepts and Methods

Statistics

Artificial Intelligence

Machine Learning
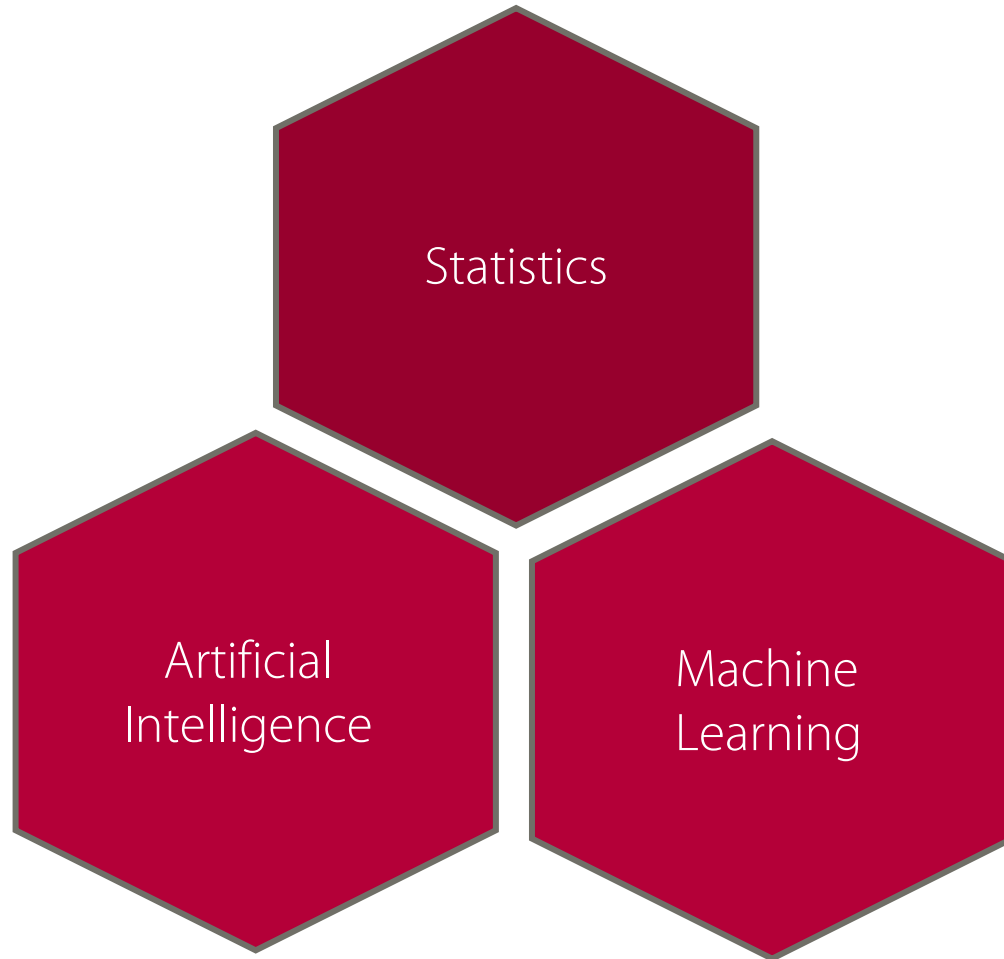
Association Rules Mining, Clustering

Regression and Classification

Time Series Analysis

Anomaly Detection and Outlier Analysis

Text Mining and Sentiment Analysis

Natural Language Processing

Reinforcement Learning

Recommender Systems

(Social) Network Analysis

Monte Carlo Methods and OR

Data Streams, etc.

Crawl
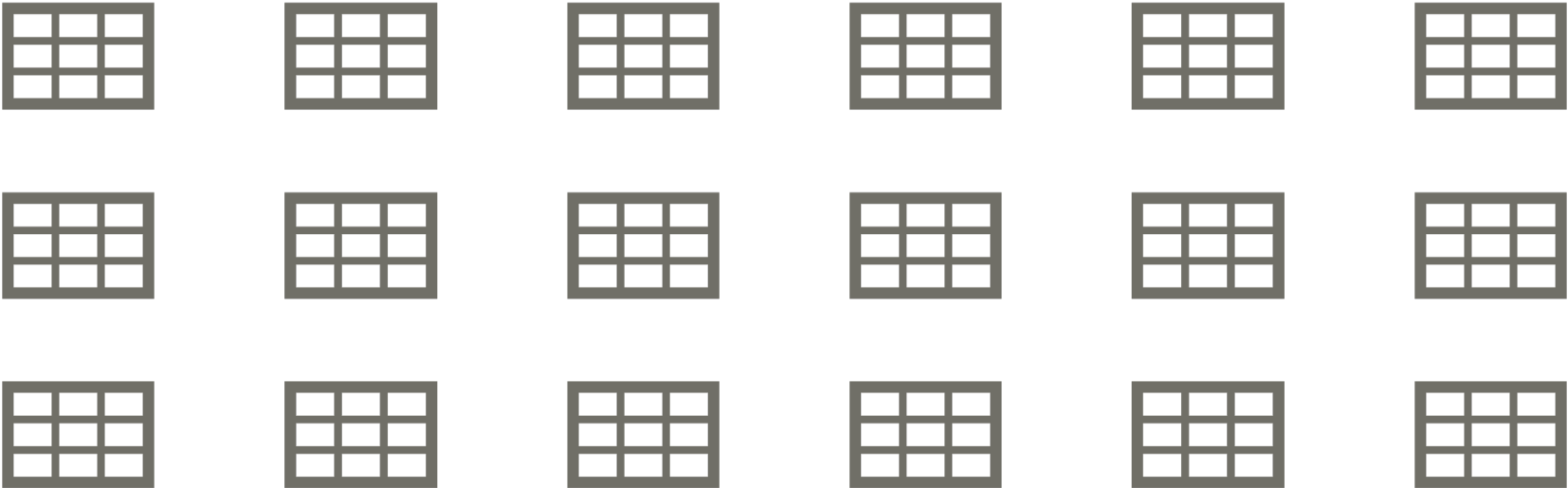
Walk
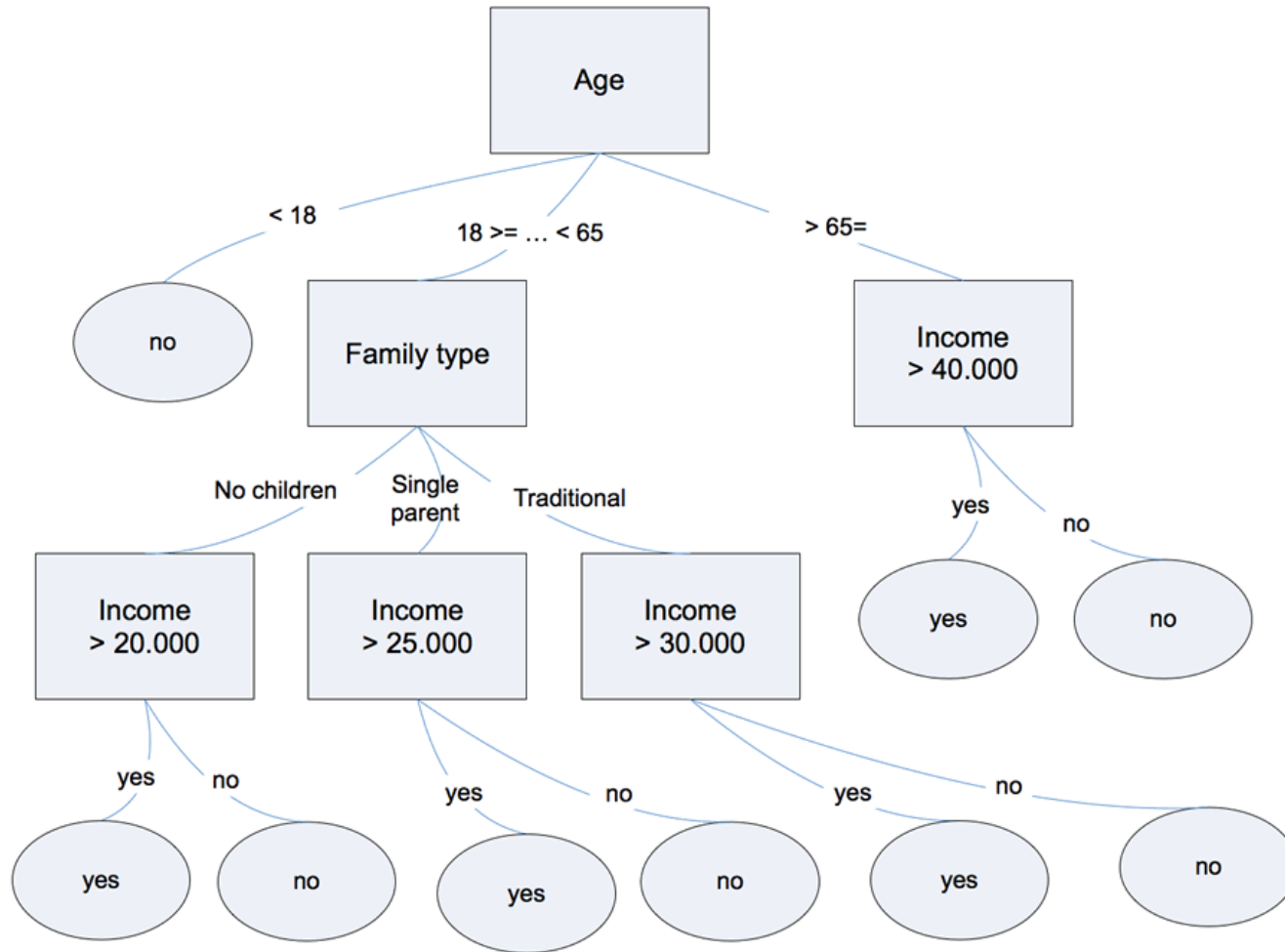
Run
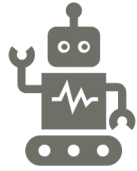
Robotics: Merging software and hardware

Machine Learning & AI

Supervised	Unsupervised	Reinforcement
Classification	Clustering	Q-Learning
Regression	Recommenders	SARSA
Deep Learning	Association Rules	Deep Q Network

Challenges and Pitfalls

Bad/unrepresentative data

Overfitting/underfitting

Big data and technological bottlenecks

Perfectionism/sloppiness

Asking the wrong questions/asking no questions

Interpretability and actionability

Not every decision needs to be made on the strength of analytics

Practical Advice

Listen to your analysts!

Beware the tyranny of past success

Don't expect miracles

Analytics is guiding the decision-making process, not replacing it

Learn to fail in useful ways

Analysis is not (just) about tools, infrastructure, and methods (but it's a big part)

Less is more, and keep things simple (when appropriate)

Use Cases

Determine how likely it is that an individual will require assisted housing

Determine which diplomatic missions are "similar" to one another

Predict the average waiting time at Canadian airports during the day

Determine the likelihood that a project will fail based on a variety of factors

Predict whether a container entering the country is dirty or clean

Predict the failure probability of a nuclear waste repository over long horizons

Suggestions for Internal Communication

1 Create analytics 'goal' roadmap

2 Ensure key KPIs are communicated

3 Assess organizational vs. technical

4 State key assumptions

5 Build strong relationships with stakeholders

6 Construct audience-dependent visualizations

7 Formulate personas or user stories

8 Craft engaging stories

Exploratory vs. Explanatory analysis

Exploratory: understanding the **DATA** (associated with reports)

Explanatory: communicating a **STORY** (associated with dashboards and data viz)
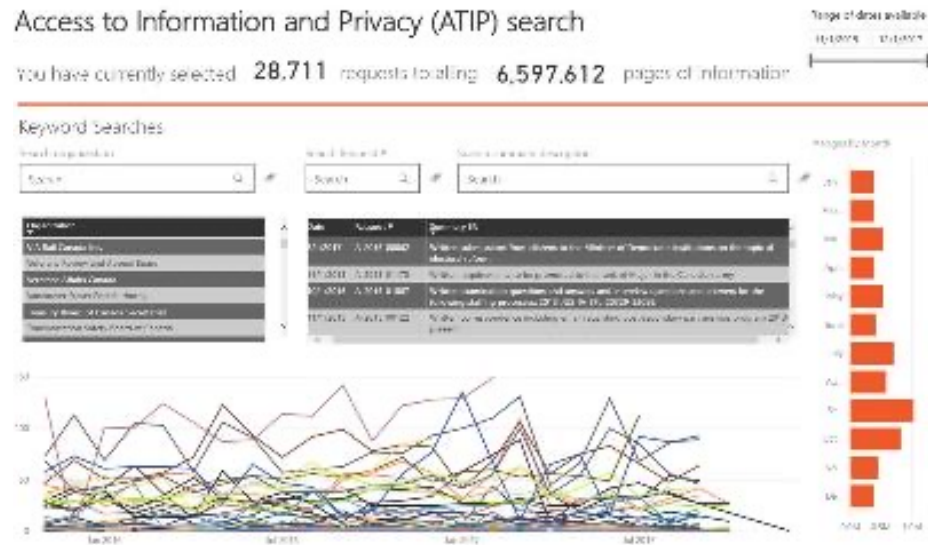
Exploratory

VS.
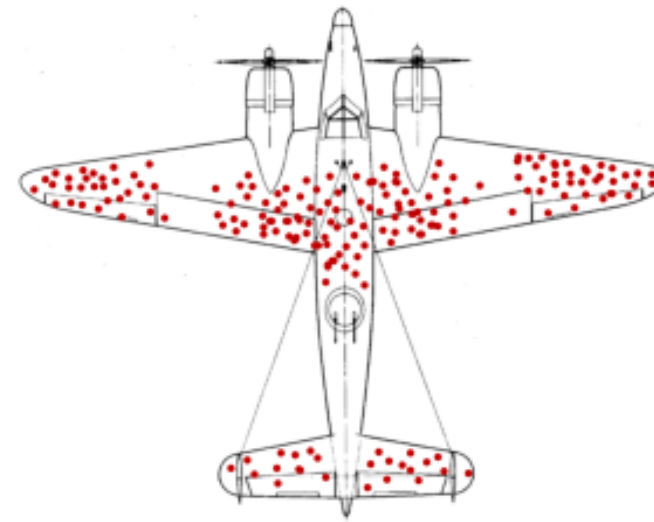
Explanatory

Assess Organizational vs Technical

Organization

← →

Structure
Priorities
Politics
Decision Making

Tech

State Key Assumptions
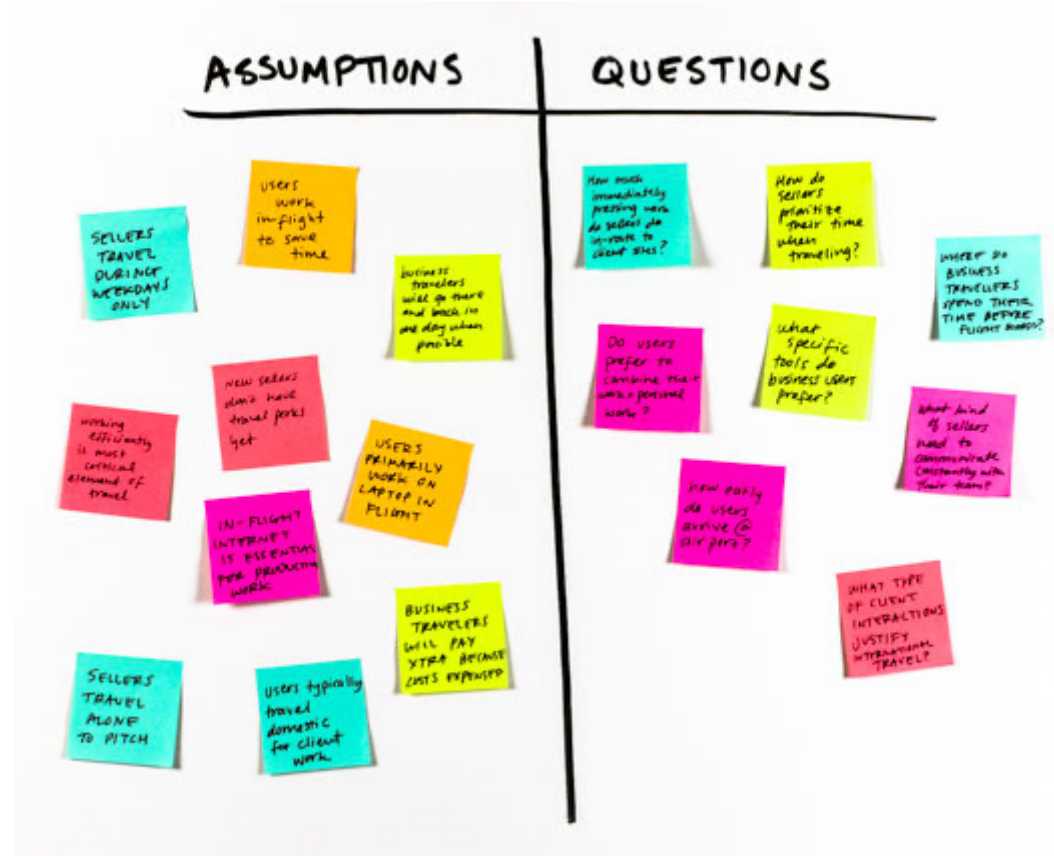
Build Relationships with Stakeholders
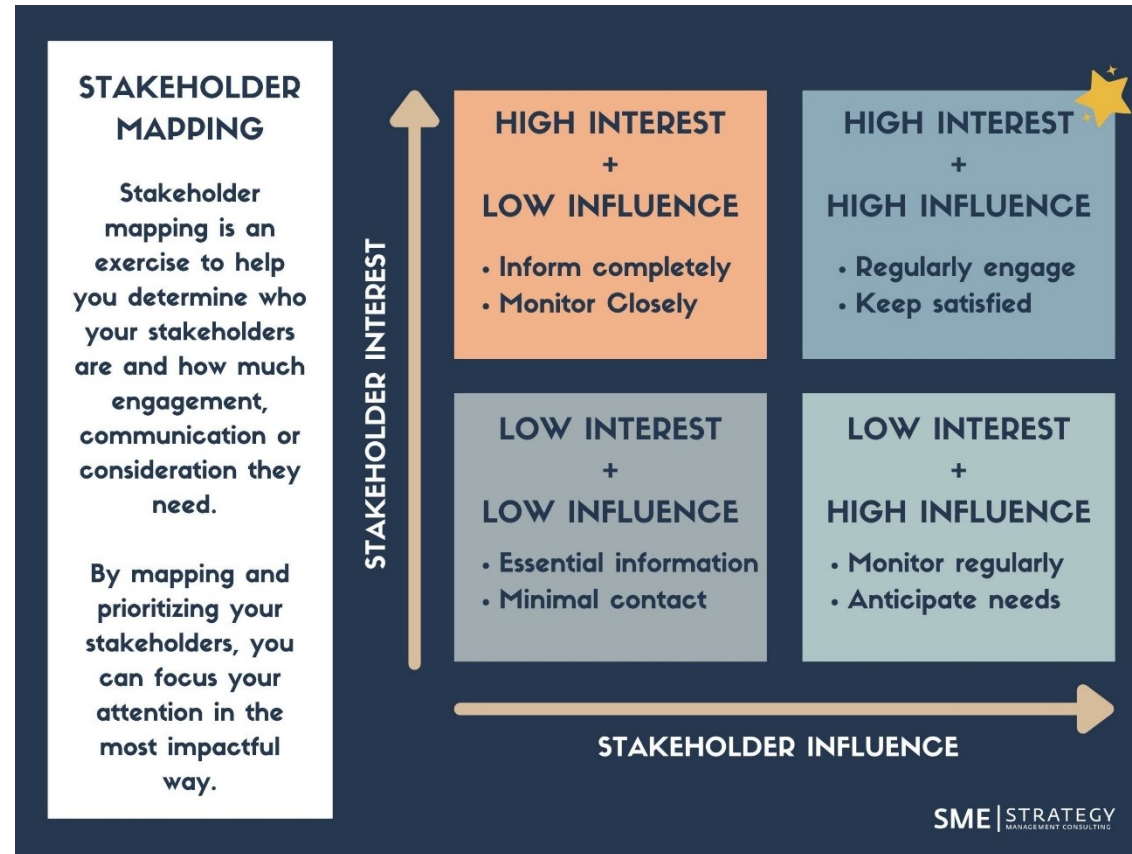
Craft Engaging Stories
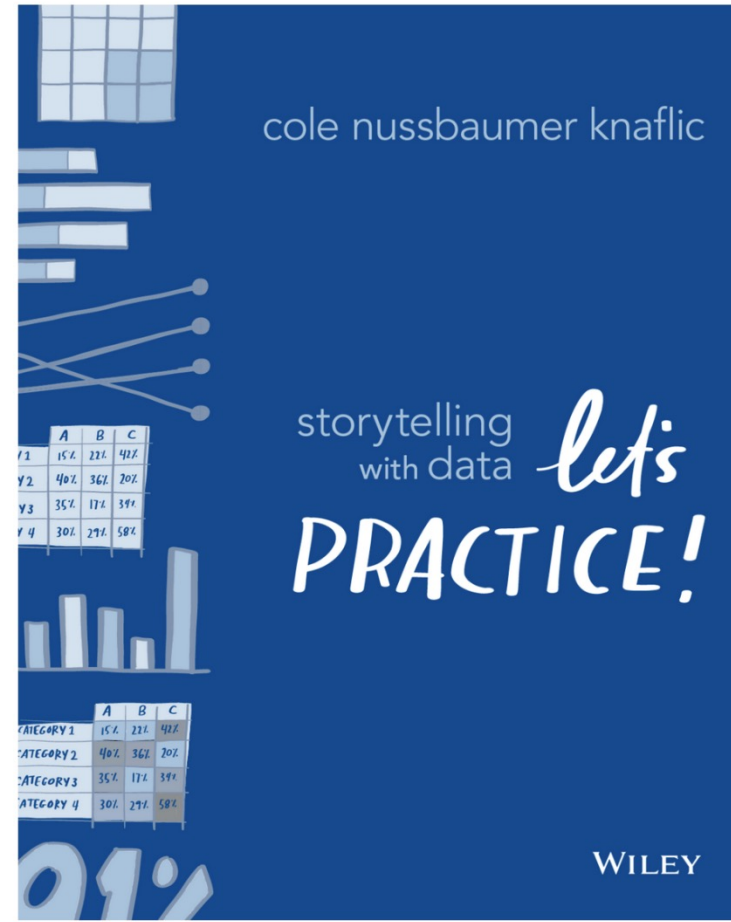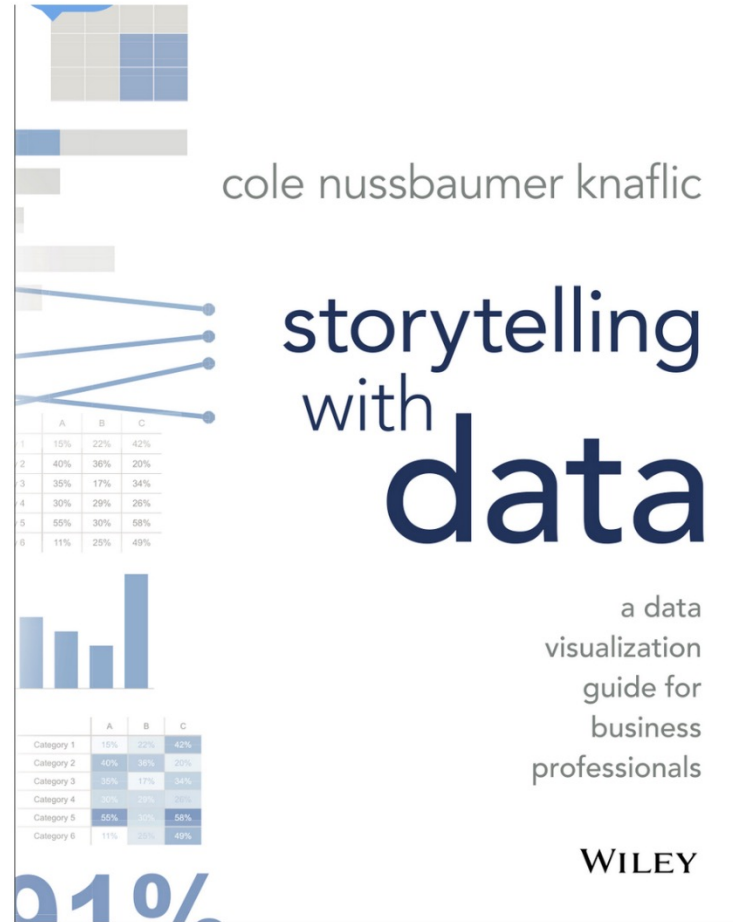
Evolving a Visualization
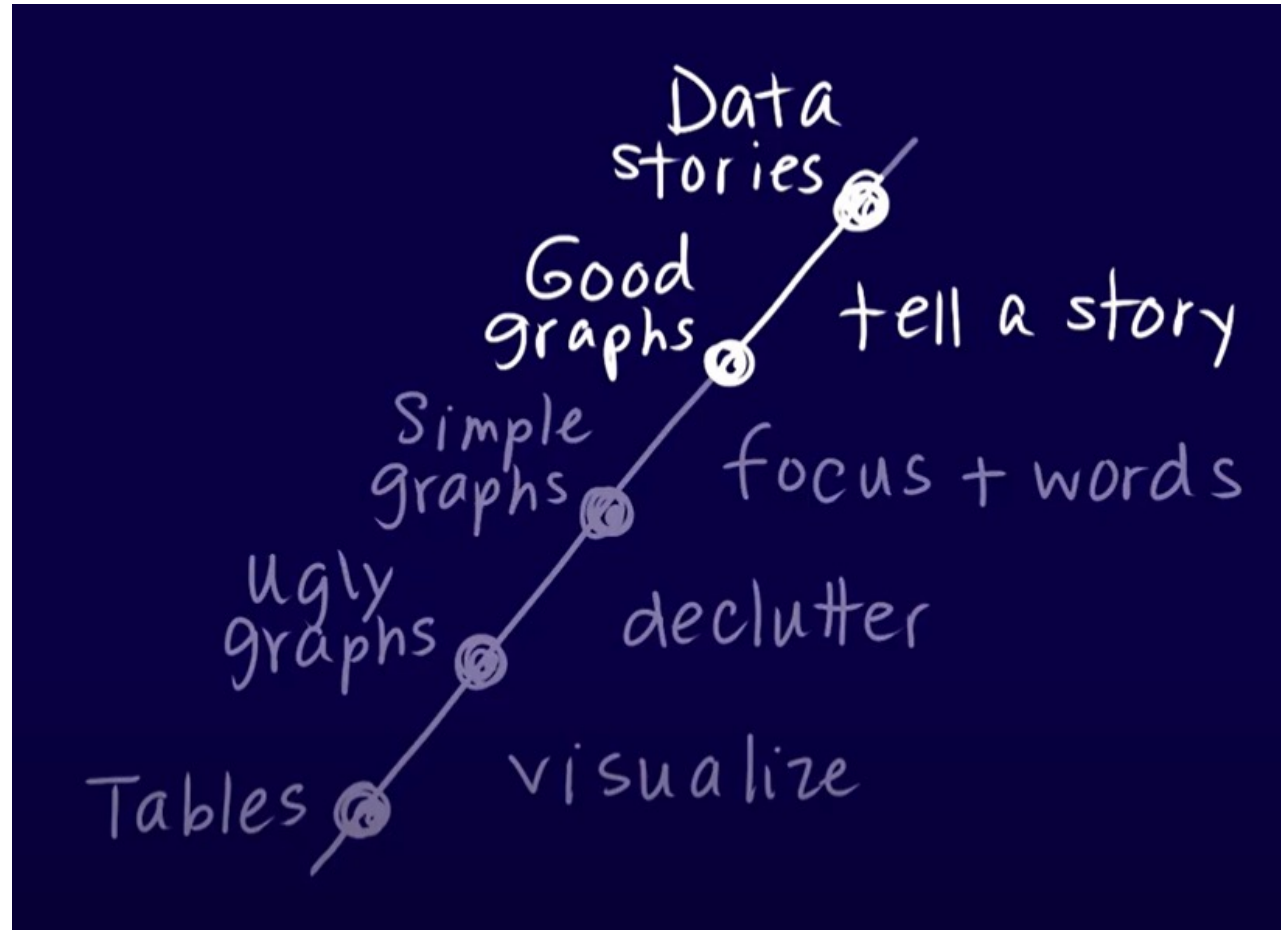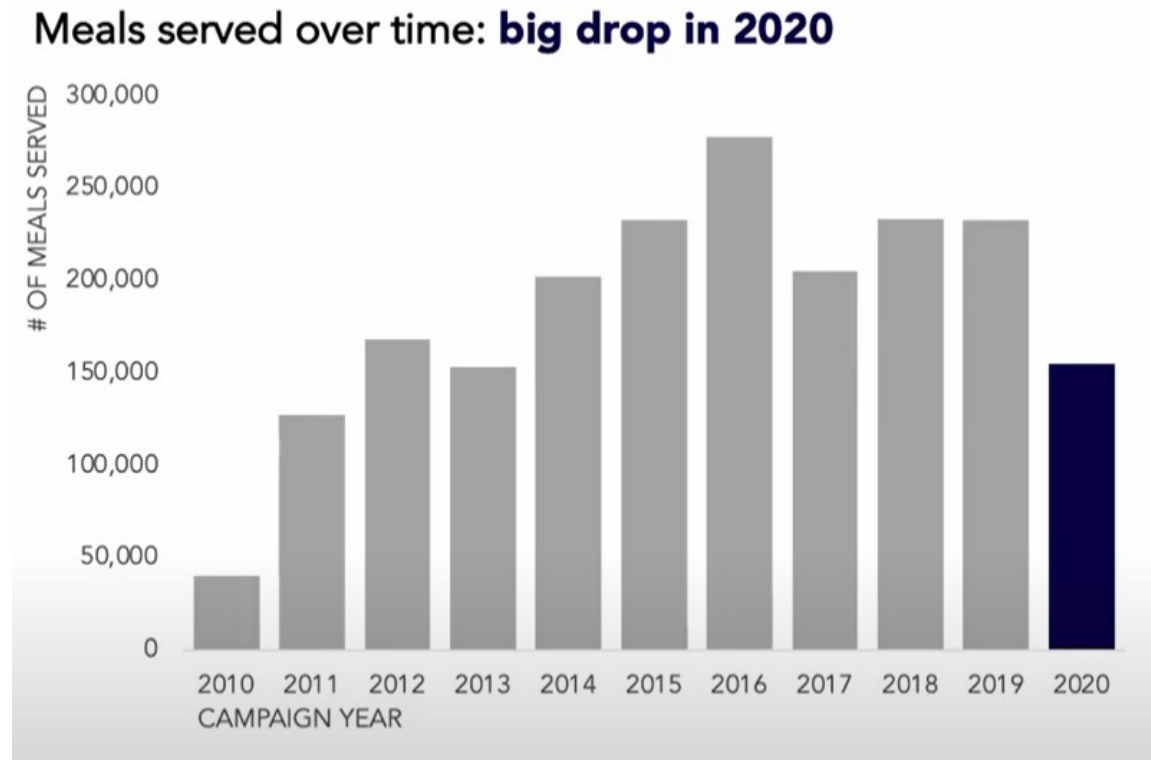
Evolving a Visualization

Meals served over time

Campaign Year	Meals Served
2010	40,139
2011	127,020
2012	168,193
2013	153,115
2014	202,102
2015	232,897
2016	277,912
2017	205,350
2018	233,389
2019	232,797
2020	154,830

Evolving a Visualization
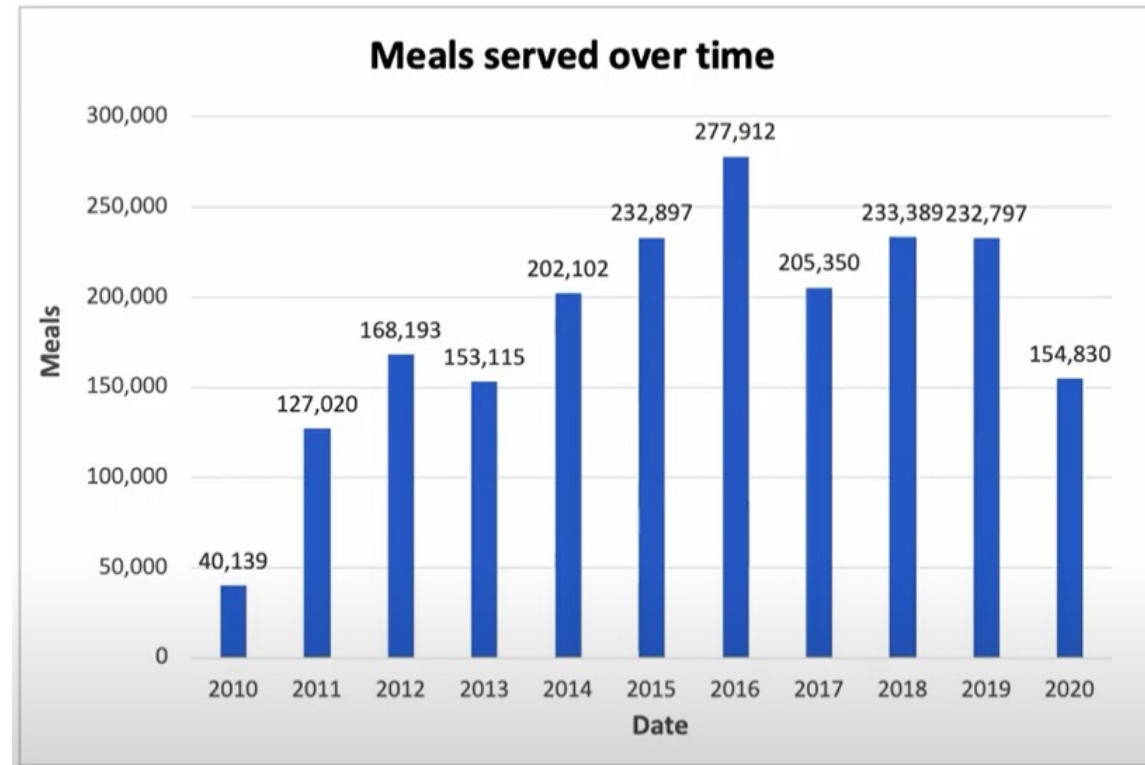
Evolving a Visualization

Meals served over time: **big drop in 2020**

Model Visualizations: Line Graph

Annual giving campaign progress

Source: Storytelling with Data
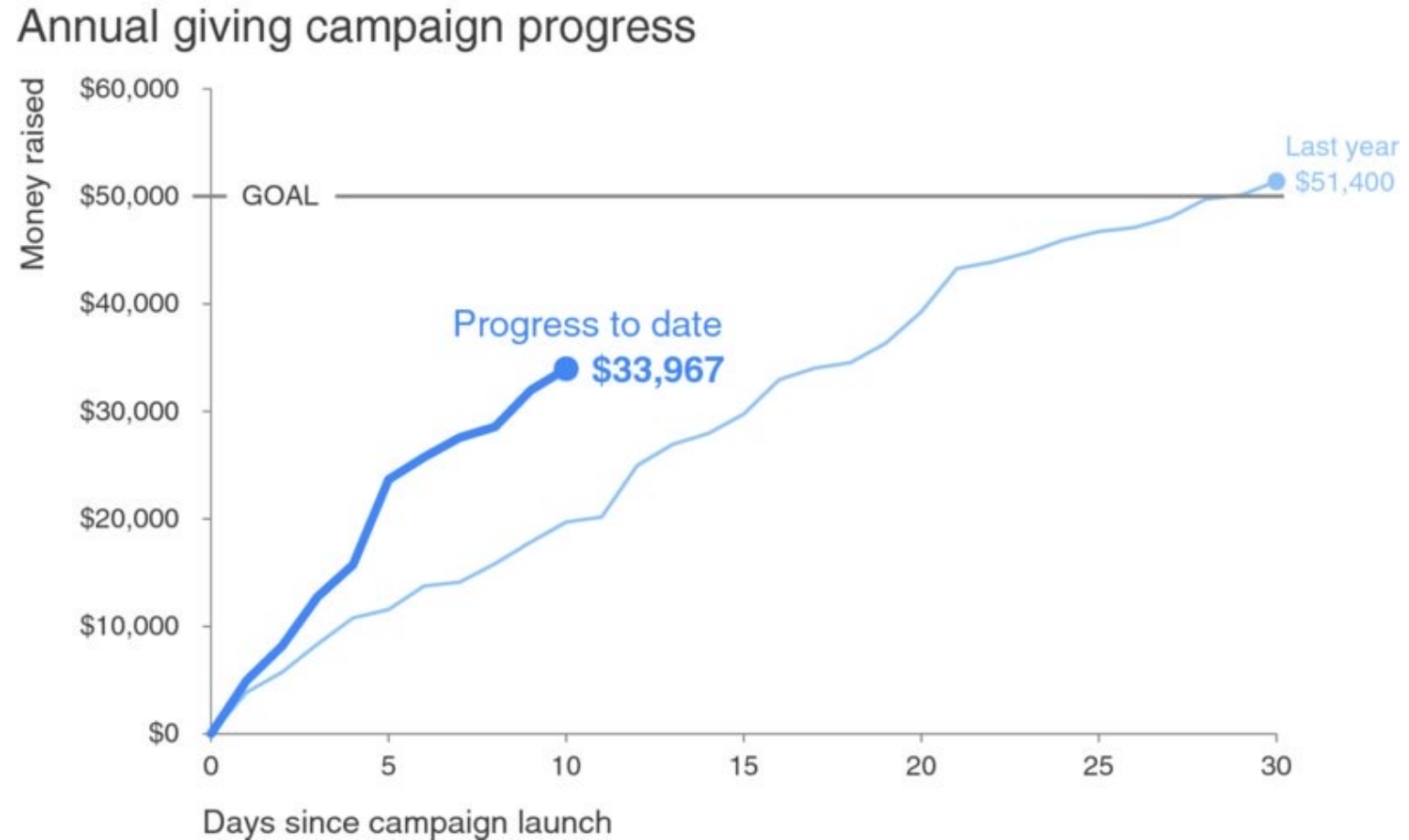
Model Visualizations: Line Graph, Forecast

Sales over time
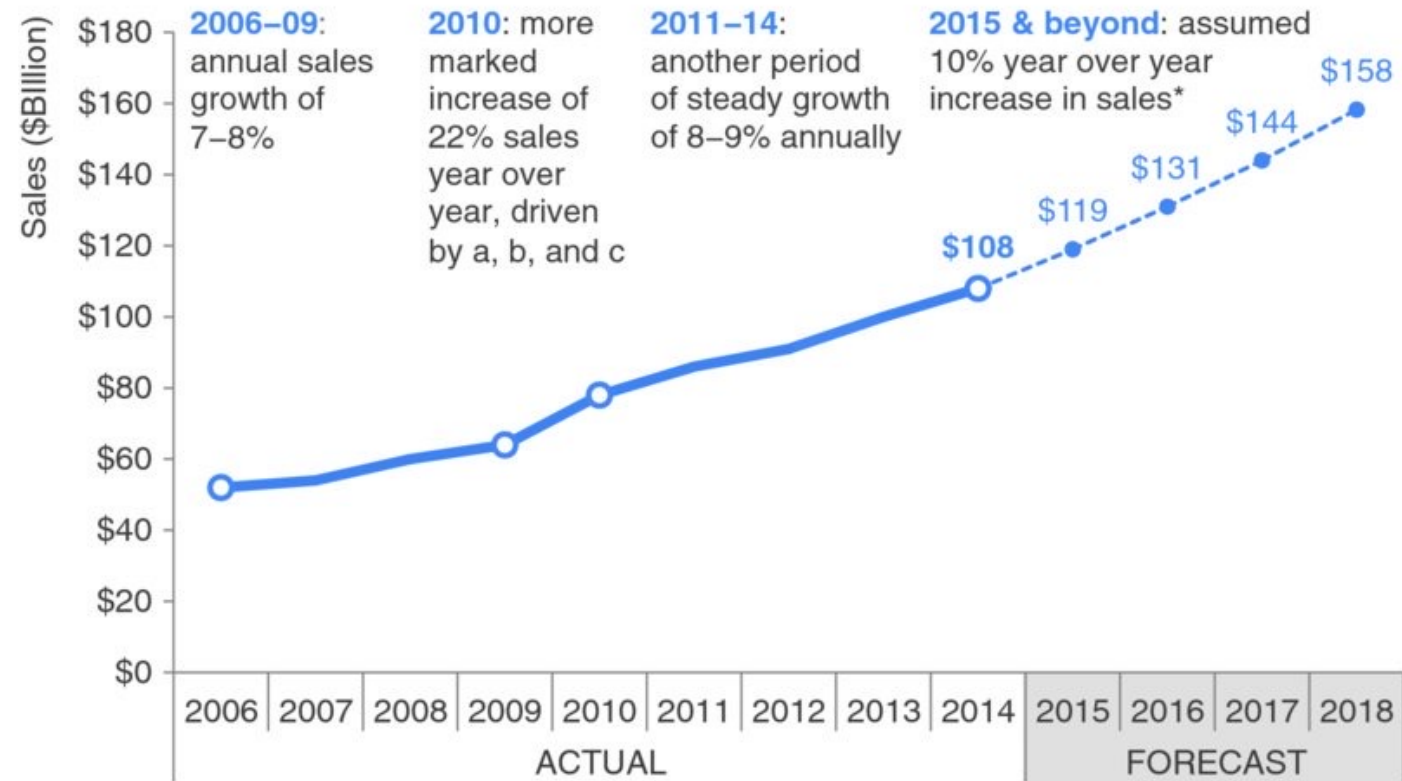
Source: Storytelling with Data

Model Visualizations: Stacked Bars

Goal attainment over time

■ Miss ■ Meet ■ Exceed

As of Q3 2015, **more than 1/3 of projects are missing goals**
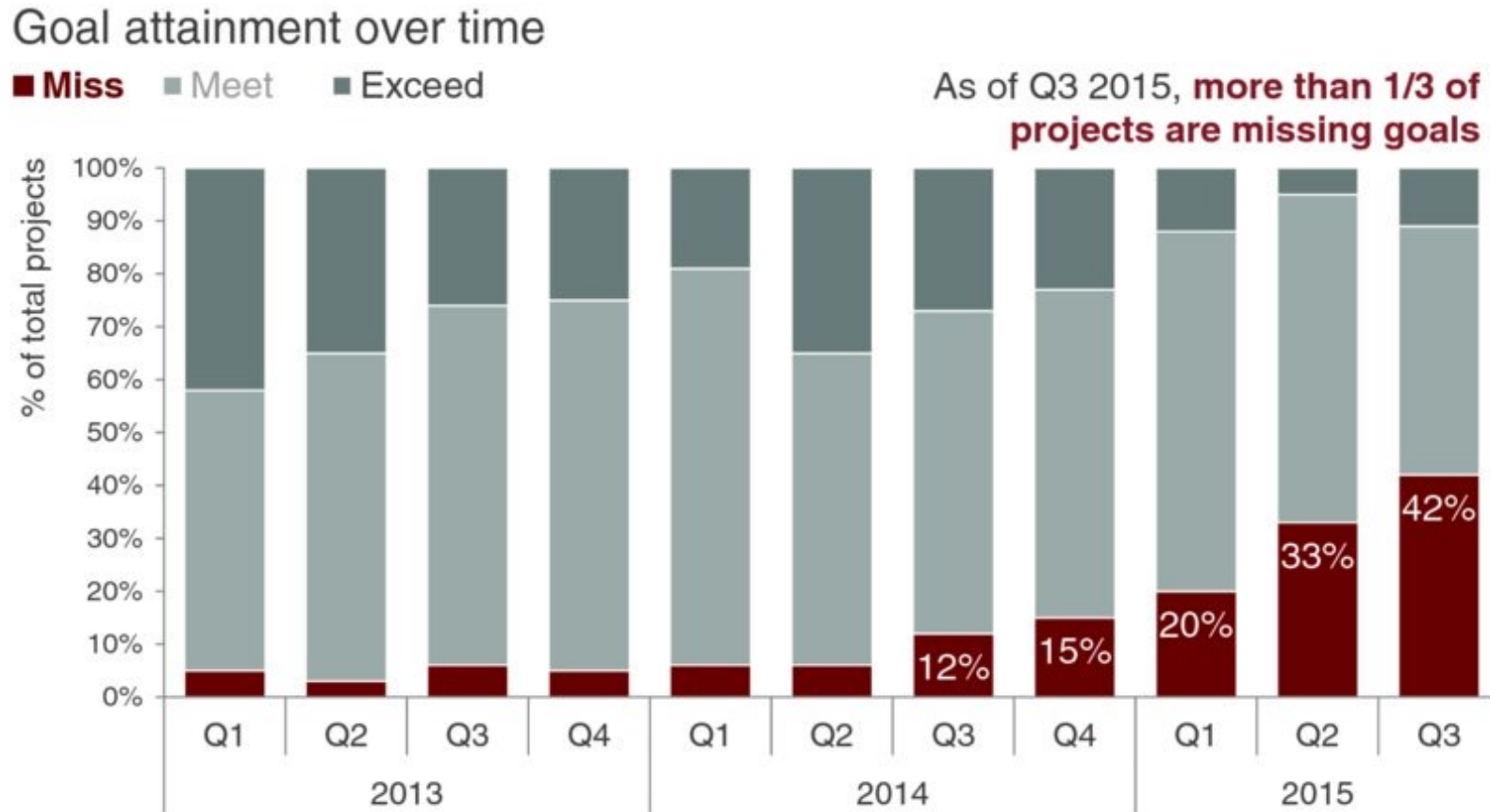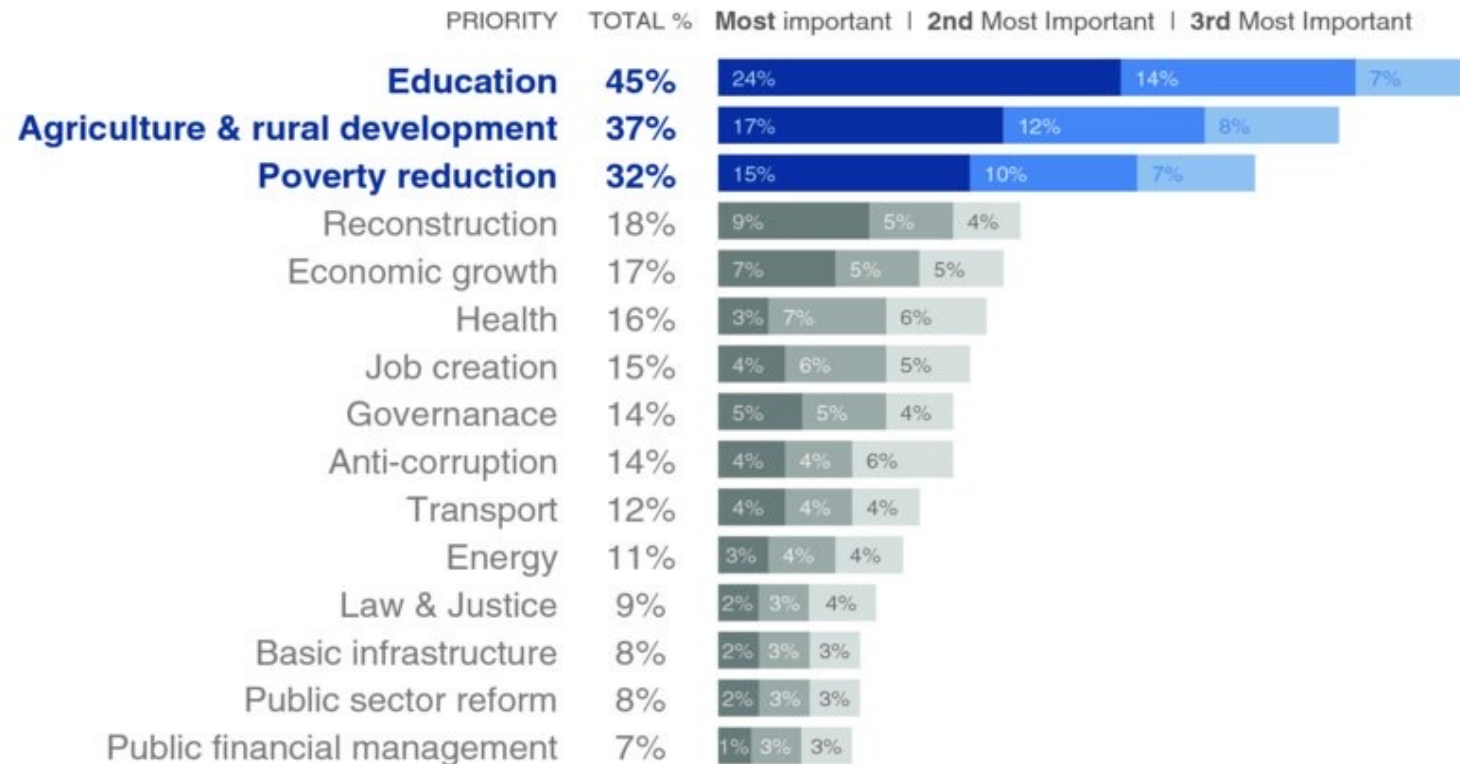
Data source: XYZ Dashboard; the total number of projects has increased over time from 230 in early 2013 to nearly 270 in Q3 2015.

Model Visualizations: Horizontal Stacked Bars

Top 15 development priorities, according to survey

PRIORITY	TOTAL %	Most important	2nd Most Important	3rd Most Important
Education	**45%**	24%	14%	7%
Agriculture & rural development	**37%**	17%	12%	8%
Poverty reduction	**32%**	15%	10%	7%
Reconstruction	18%	9%	5%	4%
Economic growth	17%	7%	5%	5%
Health	16%	3%	7%	6%
Job creation	15%	4%	6%	5%
Governanace	14%	5%	5%	4%
Anti-corruption	14%	4%	4%	6%
Transport	12%	4%	4%	4%
Energy	11%	3%	4%	4%
Law & Justice	9%	2%	3%	4%
Basic infrastructure	8%	2%	3%	3%
Public sector reform	8%	2%	3%	3%
Public financial management	7%	1%	3%	3%

N = 4,392. Based on responses to item, *When considering development priorities, which one development priority is the most important? Which one is the second most important priority? Which one is the third most important priority?* Respondents chose from a list. Top 15 shown.

Source: Storytelling with Data

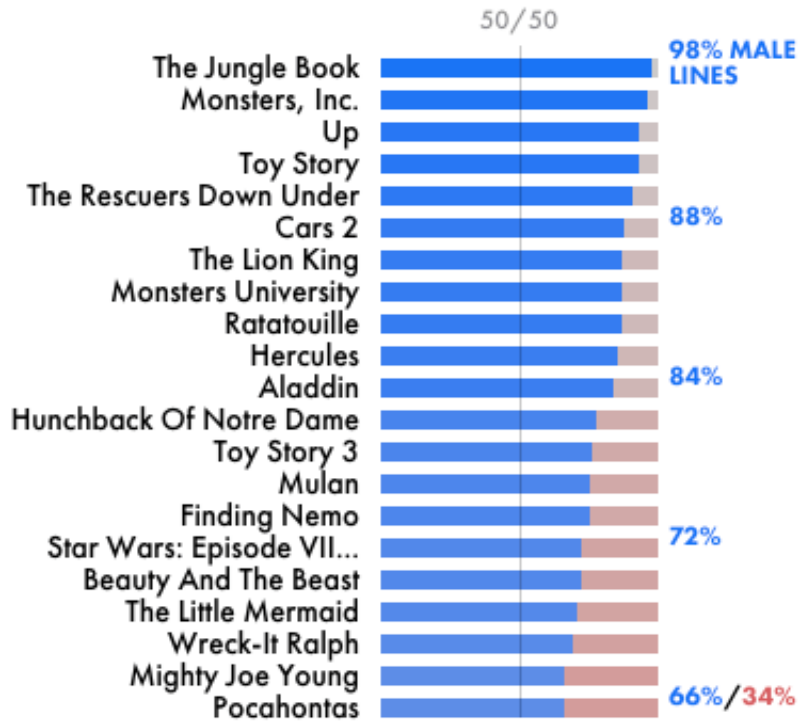# Model Visualizations: Disney

Disney **Screenplay Dialogue, Broken-down by Gender**
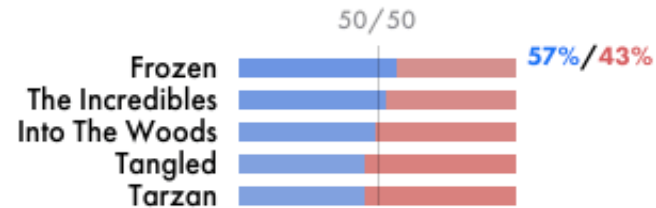
2,005 Screenplays: Dialogue Broken-down by Gender

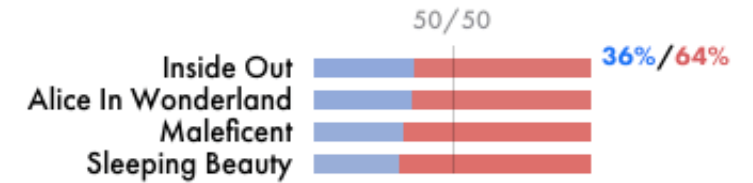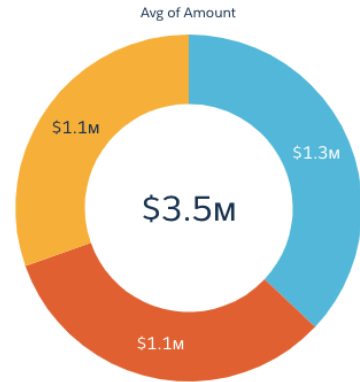Only High-Grossing Films: Ranked in the Top 2,500 by US Box Office*

Men have 60%+ Lines

50/50

The Jungle Book	**98% MALE LINES**
Monsters, Inc.	
Up	
Toy Story	
The Rescuers Down Under	
Cars 2	88%
The Lion King	
Monsters University	
Ratatouille	
Hercules	
Aladdin	84%
Hunchback Of Notre Dame	
Toy Story 3	
Mulan	
Finding Nemo	
Star Wars: Episode VII...	72%
Beauty And The Beast	
The Little Mermaid	
Wreck-It Ralph	
Mighty Joe Young	
Pocahontas	66%/34%

Gender Balance, +/- 10%

50/50

Frozen	57%/43%
The Incredibles	
Into The Woods	
Tangled	
Tarzan	

Women have 60%+ Lines

50/50

Inside Out	36%/64%
Alice In Wonderland	
Maleficent	
Sleeping Beauty	

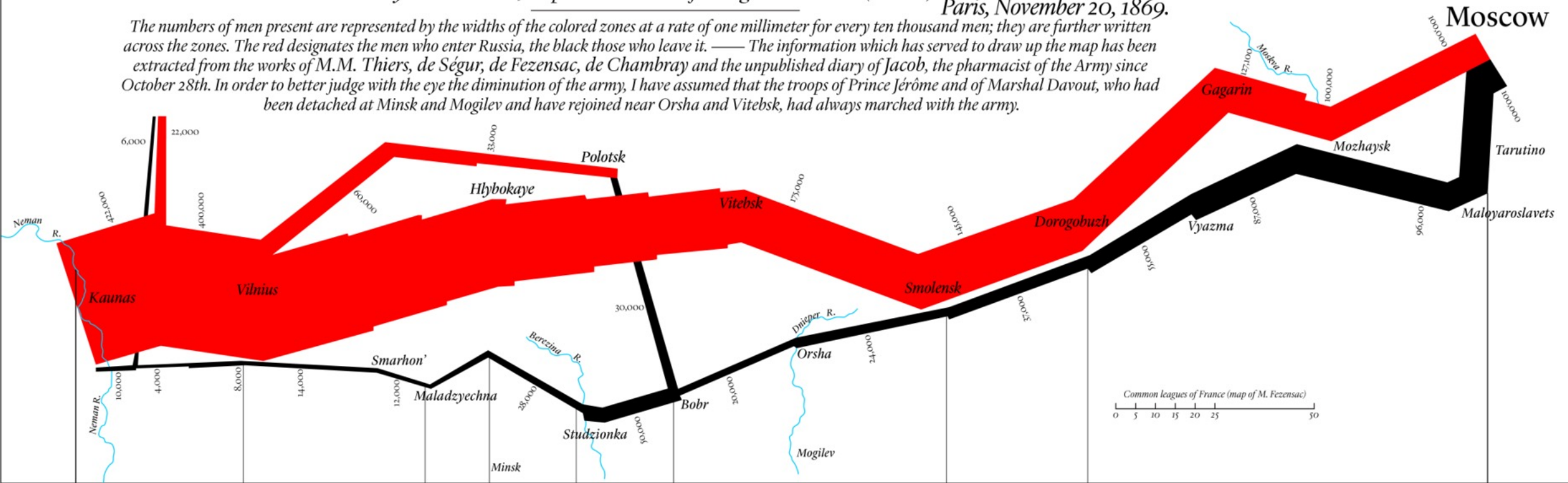# Bad Donut vs Good Donut

Source: PolicyViz
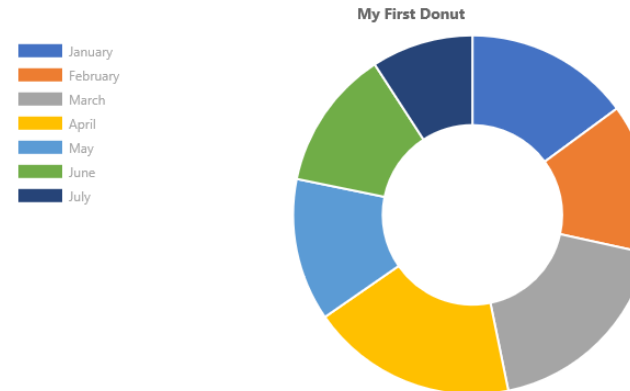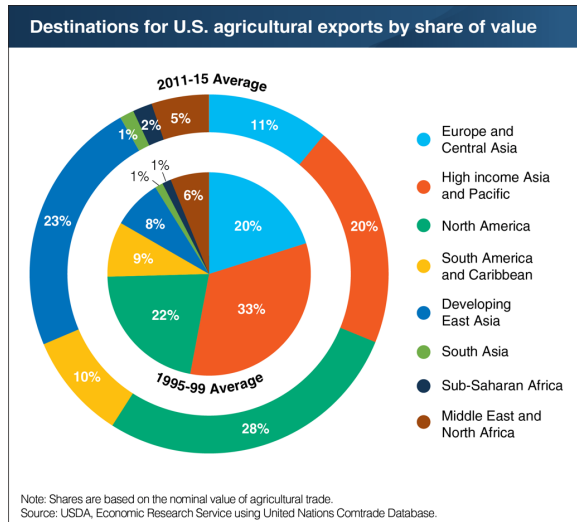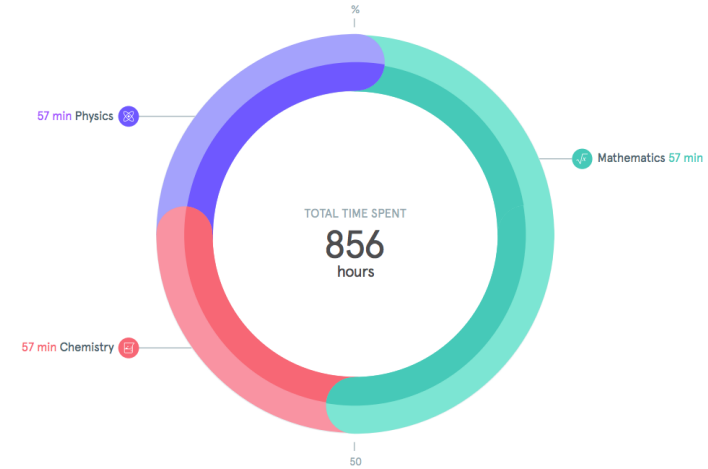
Effective Historical Visualizations

Figurative Map of the successive losses in men of the French Army in the Russian campaign 1812 ~ 1813
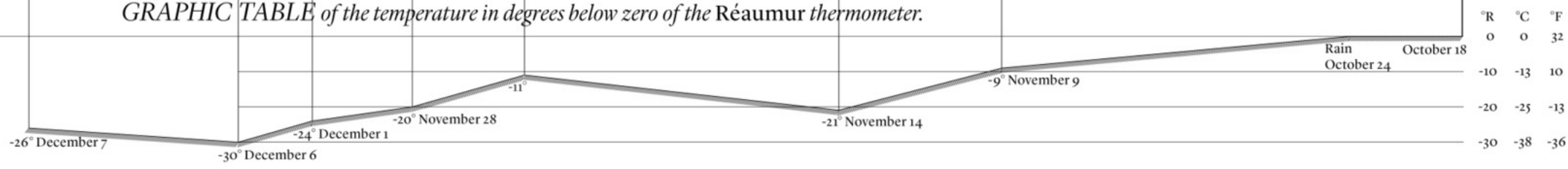Drawn by M. Minard, Inspector General of Bridges and Roads (retired). Paris, November 20, 1869.

The numbers of men present are represented by the widths of the colored zones at a rate of one millimeter for every ten thousand men; they are further written across the zones. The red designates the men who enter Russia, the black those who leave it. —— The information which has served to draw up the map has been extracted from the works of M.M. Thiers, de Ségur, de Fezensac, de Chambray and the unpublished diary of Jacob, the pharmacist of the Army since October 28th. In order to better judge with the eye the diminution of the army, I have assumed that the troops of Prince Jérôme and of Marshal Davout, who had been detached at Minsk and Mogilev and have rejoined near Orsha and Vitebsk, had always marched with the army.
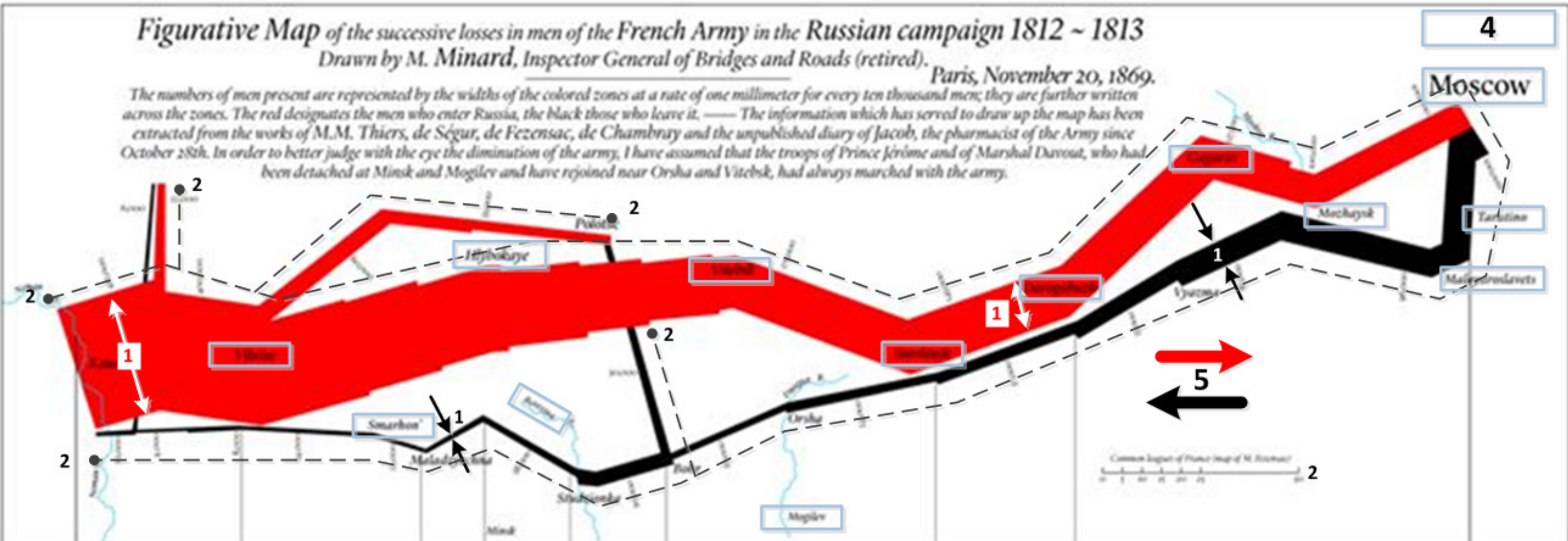
GRAPHIC TABLE of the temperature in degrees below zero of the Réaumur thermometer.
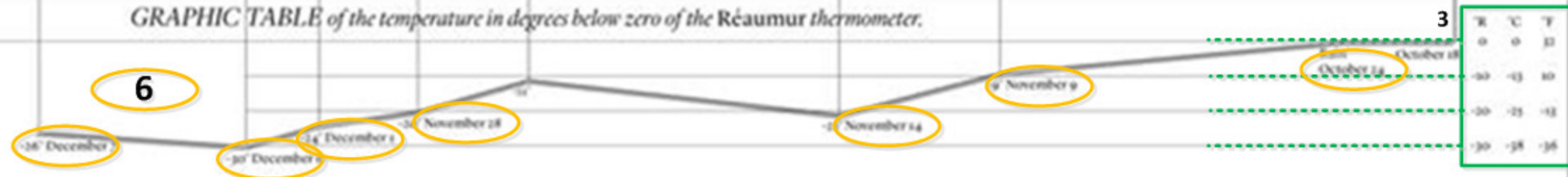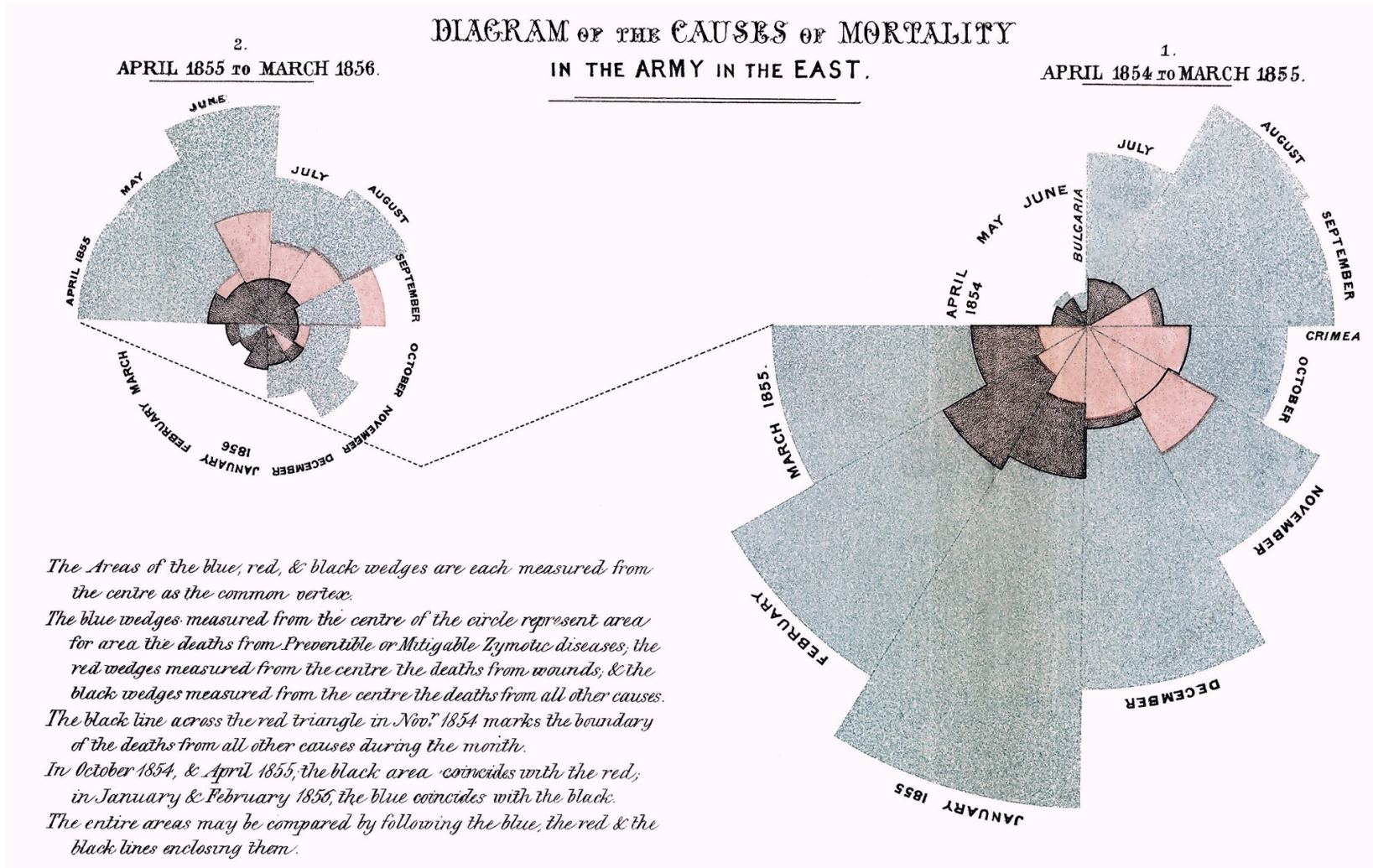
The Cossacks pass the frozen Neman at a gallop.

# Effective Historical Visualizations

Effective Historical Visualizations

During WWII, mathematician **A. Wald** undertook a study to help protect British bombers flying over enemy territory.

Data included: the **number** and **location** of **bullet holes** on returning aircraft, and the goal was to use this information to determine where to add armor to best protect the plane's structure.

A chart was created to show where the maximum number of bullet holes were located on **returning aircraft**. This chart showed greatest damage on the **aircraft extremities**, not on the main wing and tail spars, engines, and core fuselage areas.
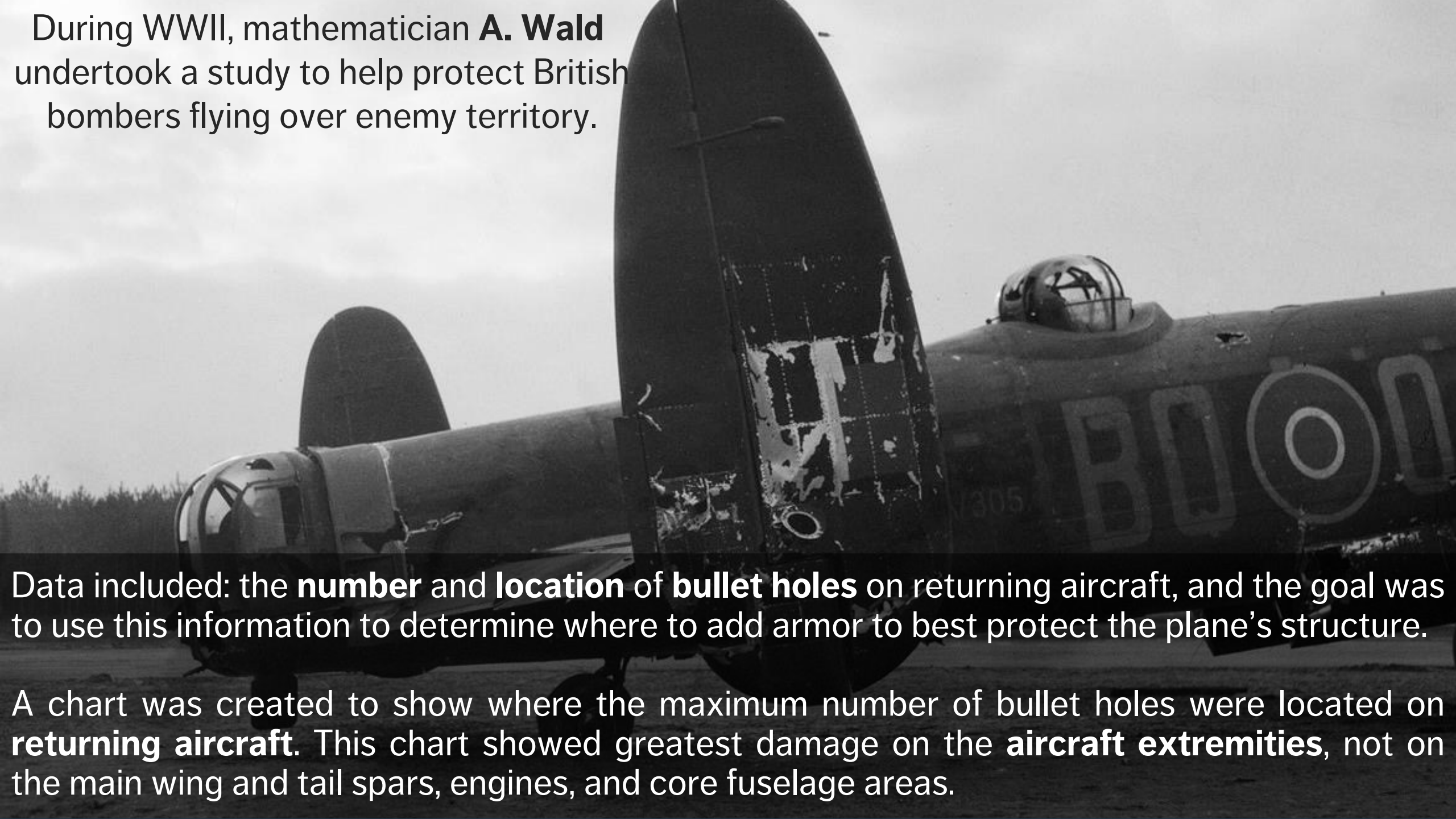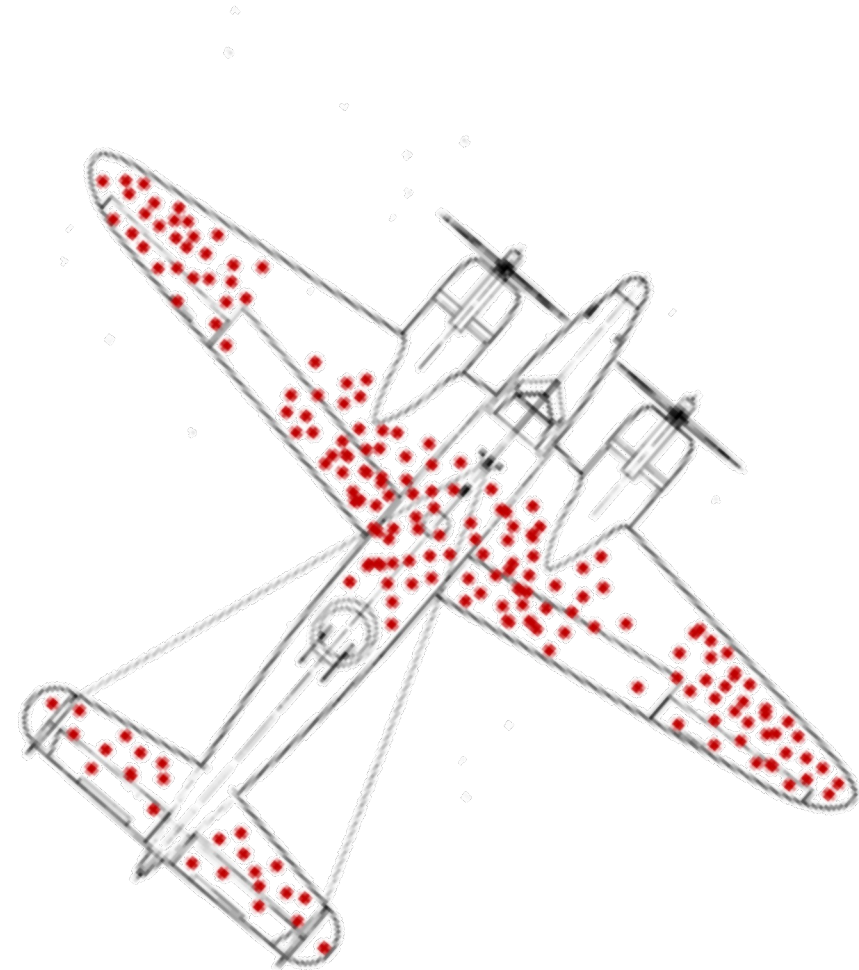
Wald's Story

As such, the Air Ministry wanted to add armor to the **extremities**. Wald suggested they were **dead wrong**.

To avoid **"survivorship bias",** armor should be added to the areas with the **fewest holes**: if no returning planes had holes in their wing spars and engines, then even a few holes in those locations were **deadly**.

Take-Away: the data that is missing may be as important to story than the data that is there. Storytelling is not always an obvious endeavour.

Taking What Can Be Measured As Truth

EVERY RECORDED METEORITE IMPACT ON EARTH FROM 2,500 BCE TO 2012 — Where have they fallen?
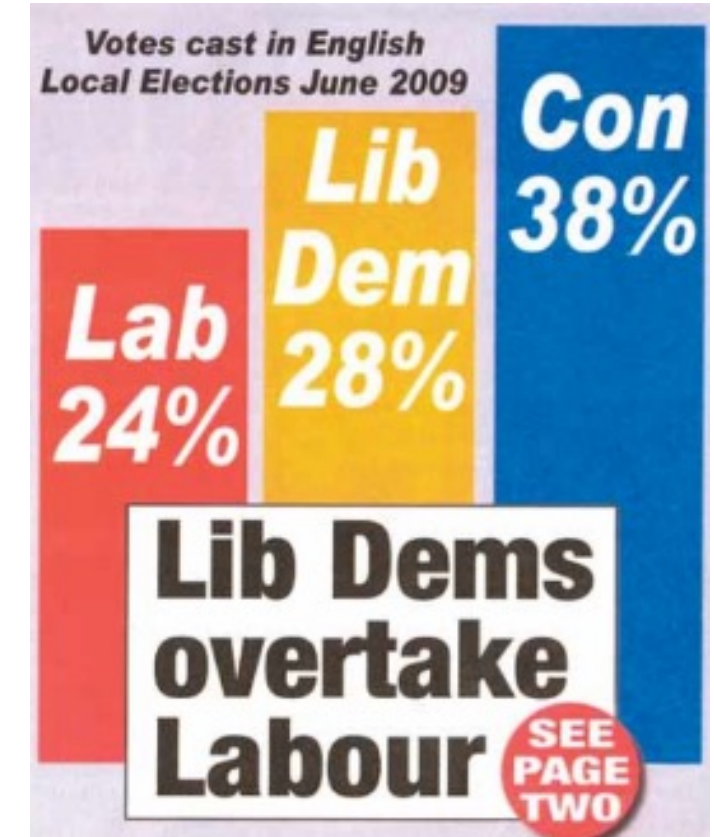
© OpenStreetMap contributors
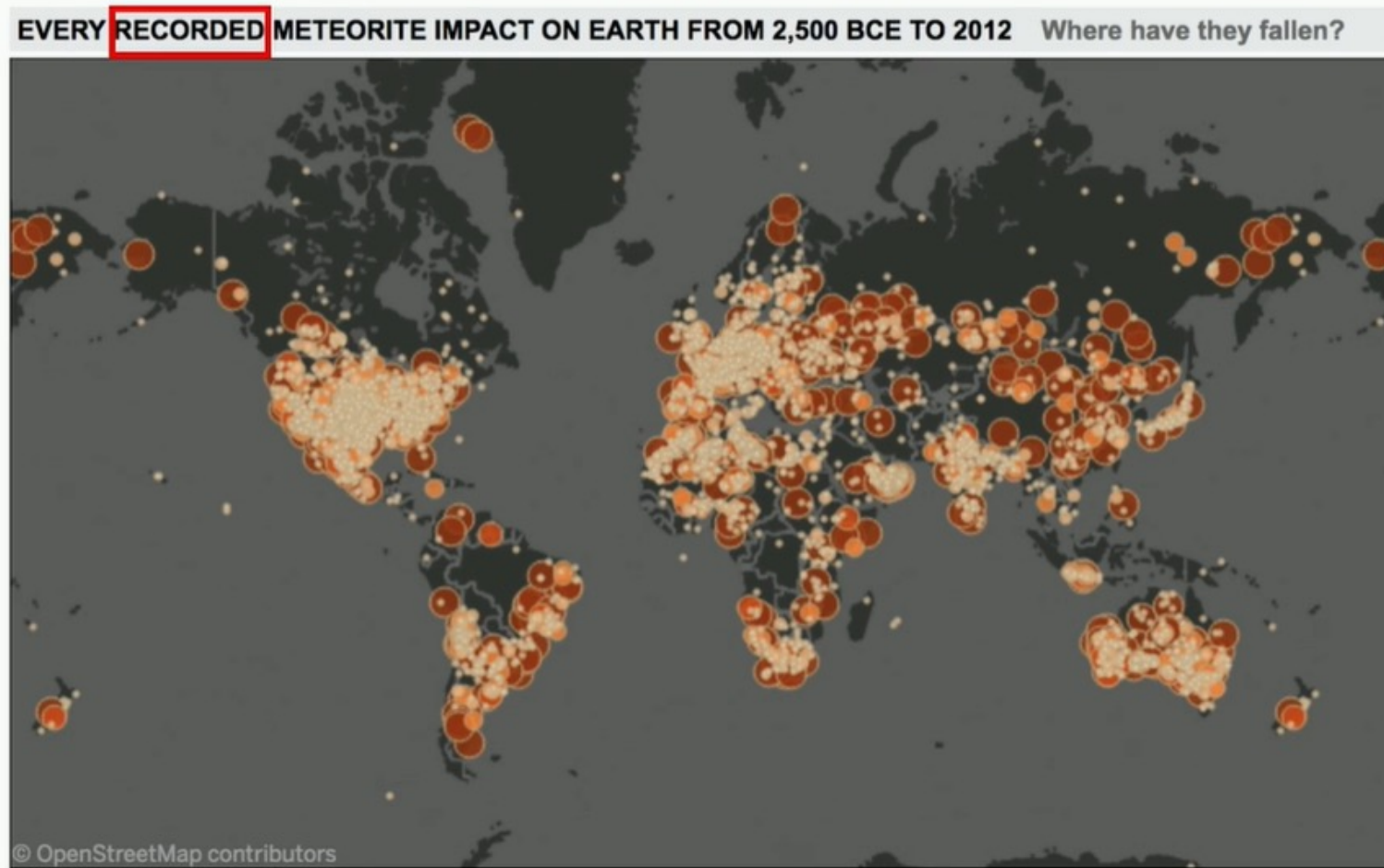
Source: The Head Game

Misleading Charts
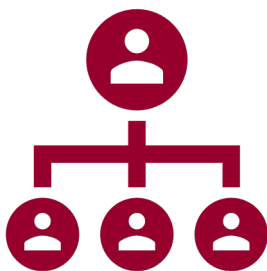
Module 6
Data Engineering and Data Governance

Desktop Data Analysis

Business Intelligence needs are pushing the development of **desktop data analysis** tools and pipelines, such as:

- PowerBI
- Tableau

Democratization of data + increase in data/digital literacy.

This is likely going to push organizations forward as well.

Not **necessarily** a substitute for 'industrial' or 'professional' data pipelines.

BI Gateway to AI/ML

To some extent getting a solid professional/industrial BI pipeline up and running is a major steppingstone in an organization.
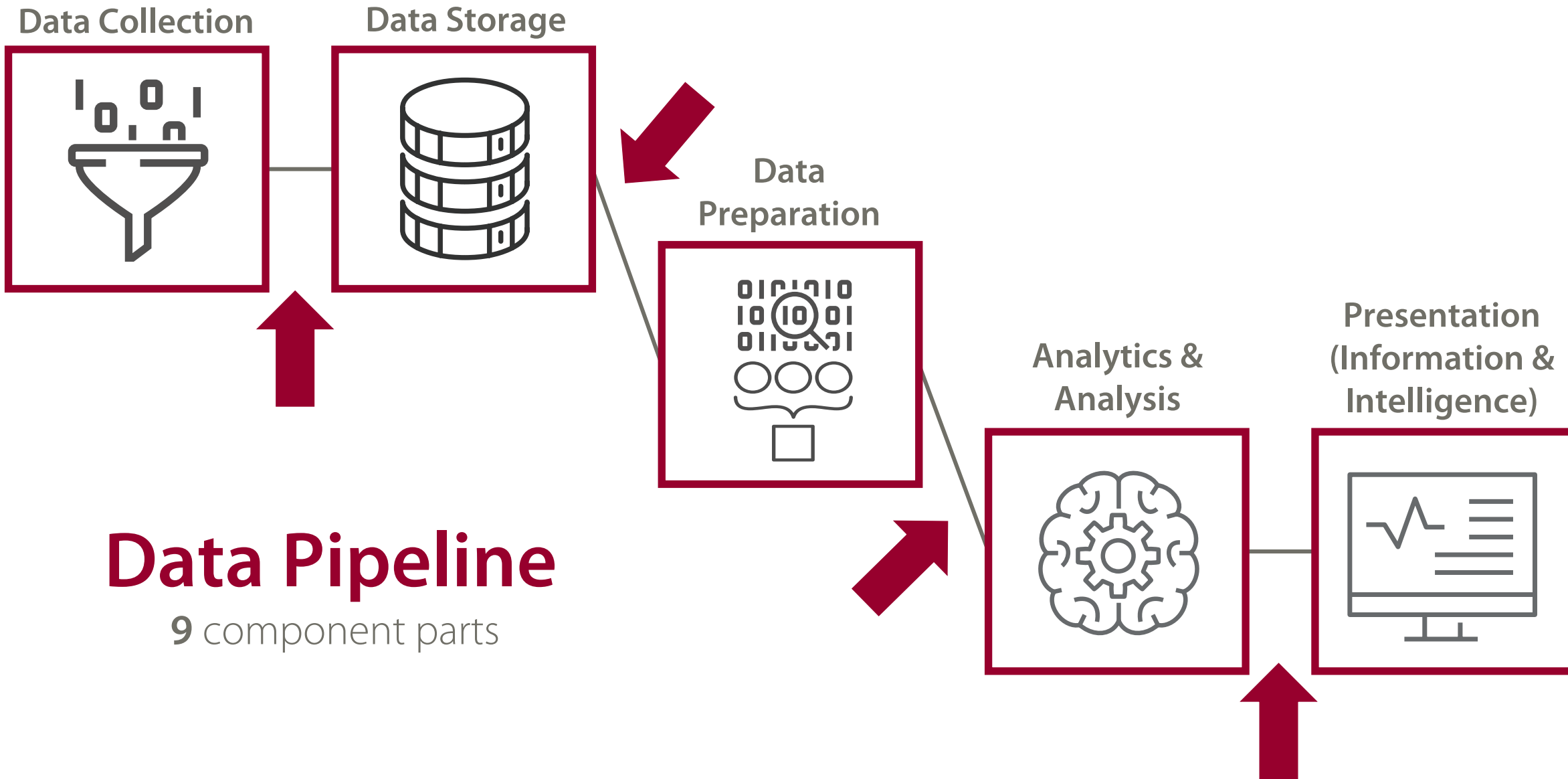
BUT – the data architecture and tools you need for AI/ML/DS analysis may not be the same as those for BI.

You will MAY need to redesign some parts of your BI pipeline to support AI/ML/DS

In particular – your database architecture: Data Lakes vs DataMart vs NoSQL

Things to Think About When Selecting Analysis Tools

A. **Capability: w**hat is their functionality + performance – do they have all the techniques, do they have the processing power

B. **Integration: h**ow do they connect to other parts of your pipeline

C. **User-Experience:** what is the user experience like – what background/level of expertise do you need to operate this tool, how easy is it to use this tool?

D. **Cost:** short and long term

Data Collection

Data Storage

Data Preparation

Analytics & Analysis

Presentation (Information & Intelligence)

Data Pipeline

9 component parts

Data Hierarchy of Needs

THE DATA SCIENCE
HIERARCHY OF NEEDS

LEARN/OPTIMIZE

AGGREGATE/LABEL

EXPLORE/TRANSFORM

MOVE/STORE

COLLECT

AI, DEEP LEARNING

A/B TESTING, EXPERIMENTATION, SIMPLE ML ALGORITHMS

ANALYTICS, METRICS, SEGMENTS, AGGREGATES, FEATURES, TRAINING DATA

CLEANING, ANOMALY DETECTION, PREP

RELIABLE DATA FLOW, INFRASTRUCTURE, PIPELINES, ETL, STRUCTURED AND UNSTRUCTURED DATA STORAGE

INSTRUMENTATION, LOGGING, SENSORS, EXTERNAL DATA, USER GENERATED CONTENT
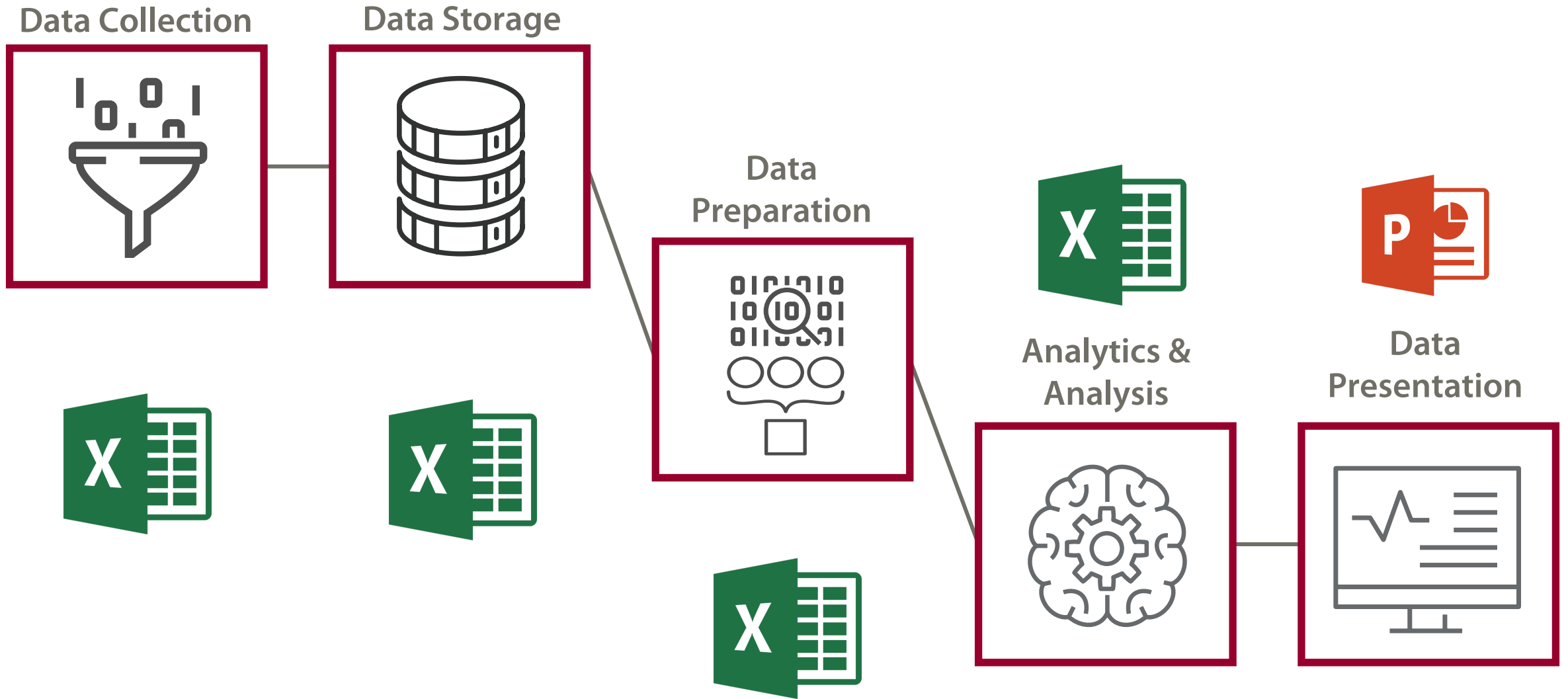
@mrogati

Open-Source Data Driven Stack
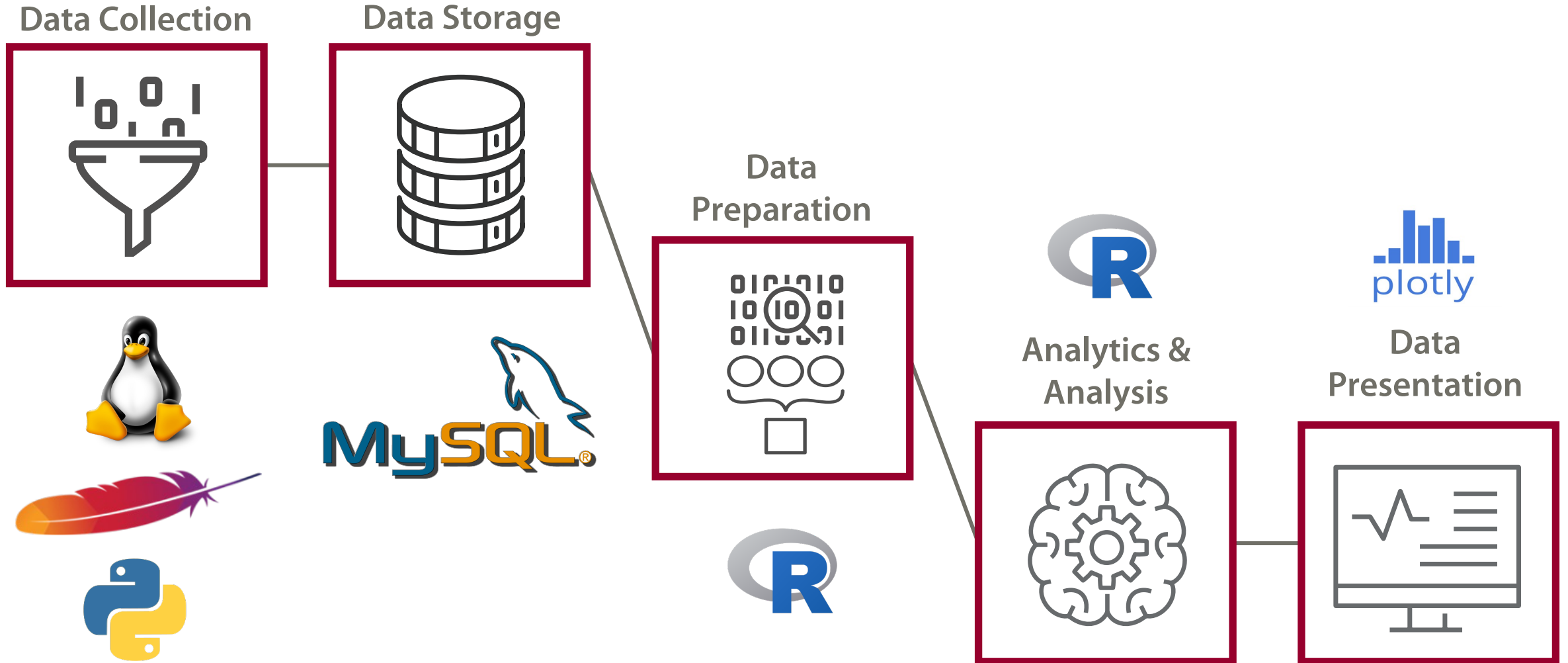
Common GoC Data Stack

Data Pipeline Technologies

|	On-Premises (LAN)	Public Cloud	Private Cloud
Amateur	Shared Directory + Excel + Power Point + 'Desktop' Access	**Server Based End-to-End Automated Pipeline Tech:** On-Premises Azure, On-Premises IBM RedHat	Home Brewed Solutions using Servers Stood Up on Cloud – e.g., AWS, GCP
Semi-Pro	**Desktop DataScience:** Desktop PowerBI SQL-Lite (Desktop) MS Access Stand-Alone In-House DBMS – Read + Write		
Professional	**Server Based End-to-End Automated Pipeline Tech:** On-Premises Azure, On-Premises IBM RedHat	**End-to-end SaaS data pipelines:** e.g., COTS Pachyderm or more bespoke: e.g., SaaSCoder	**End-to-End Cloud Data Pipeline Infrastructure (Serverless/NoServer):** AWS, GCP, Azure

Pipeline Creation Phases

1. Research + Design
2. Implementation
3. Testing
4. Production + Management
5. Research + Design

Agile!

Build vs Buy
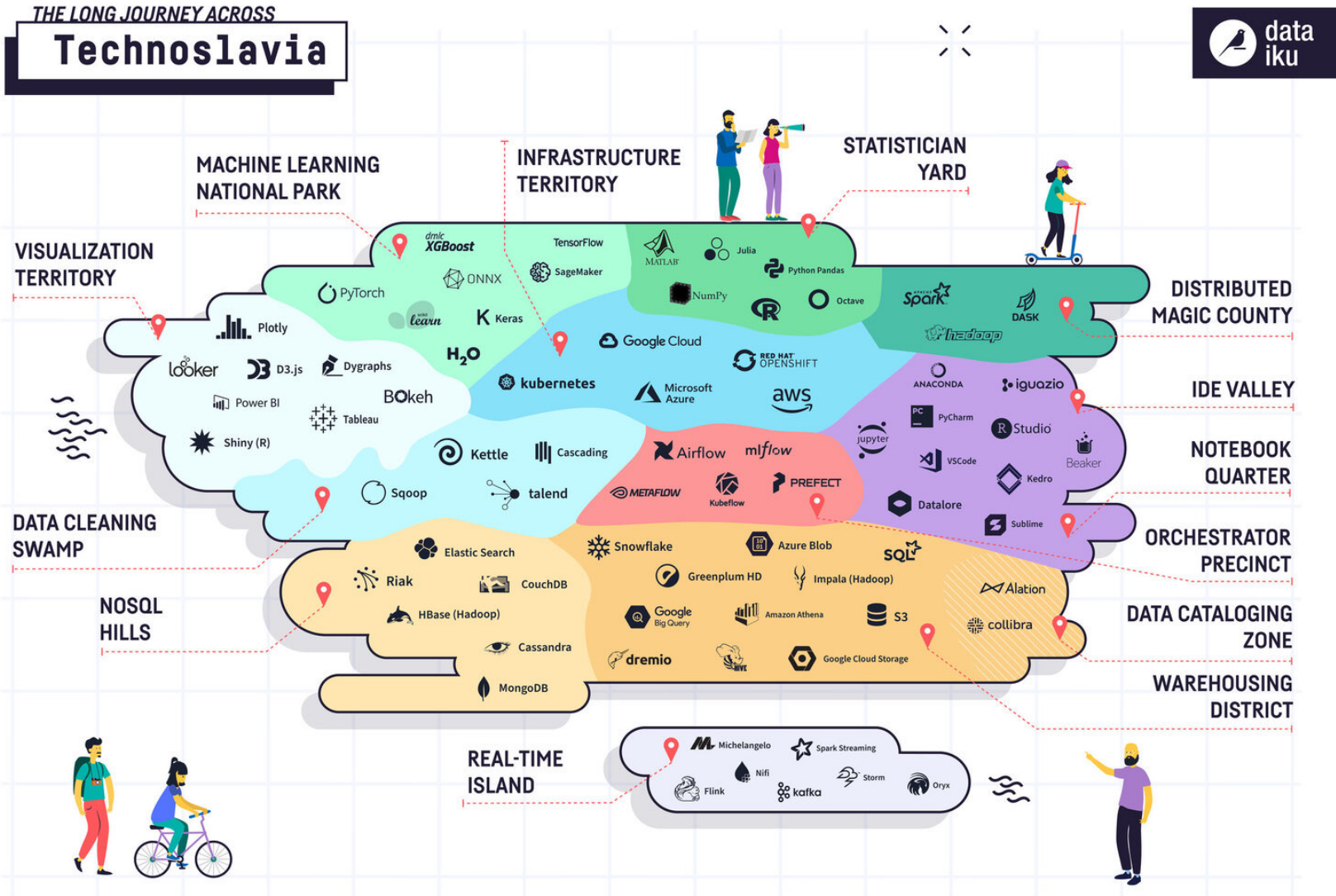
Reasons for Wanting to Build vs Buy

1. Want to gain a competitive advantage.

2. Need to catch up to competitors/other departments.

3. Attempting to encourage more engagement to customers or employees.

4. Potentially creating a new revenue stream.

Some people build then, once built, realize that their solution is poor.

- Boomerang implementers

Who Makes This Decision Typically?

C-Suite, Data Owner, Product Manager, Project Manger

- ideally, a technical advisor?

Consequences of **building:**

- industry standards change
- maintenance nightmares
- dependence on talent
- tacit knowledge creation

Buy Decision Timeline

If you look at things from a costing standpoint, you can determine quickly.

Could need:
- an application person
- infrastructure person
- data architecture person
- UI/UX person

Building Bias: technical individuals might get excited about the idea of a build because of their background, because they're passionate about building.

Build vs Buy

Do your due diligence; easy to be biased towards a provider.

Get the proper buy-in to do the investigation to begin with.

Don't just go by the number of features:
- How many customers have been successful with the vendor?

Finding the right partner:
- Custom software development
- Strong security and infrastructure expertise
- Good integration into existing data governance
- A partner that you can lean on

Source: Making Big Data Work

Consequences of the Wrong Decision

Technical debt of insourcing and outsourcing.

If stuck with the wrong vendor, could be put unknowingly on the hook.
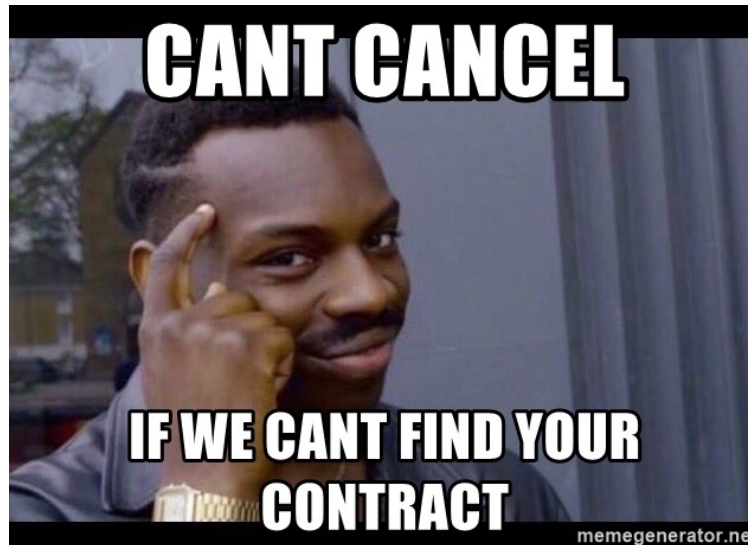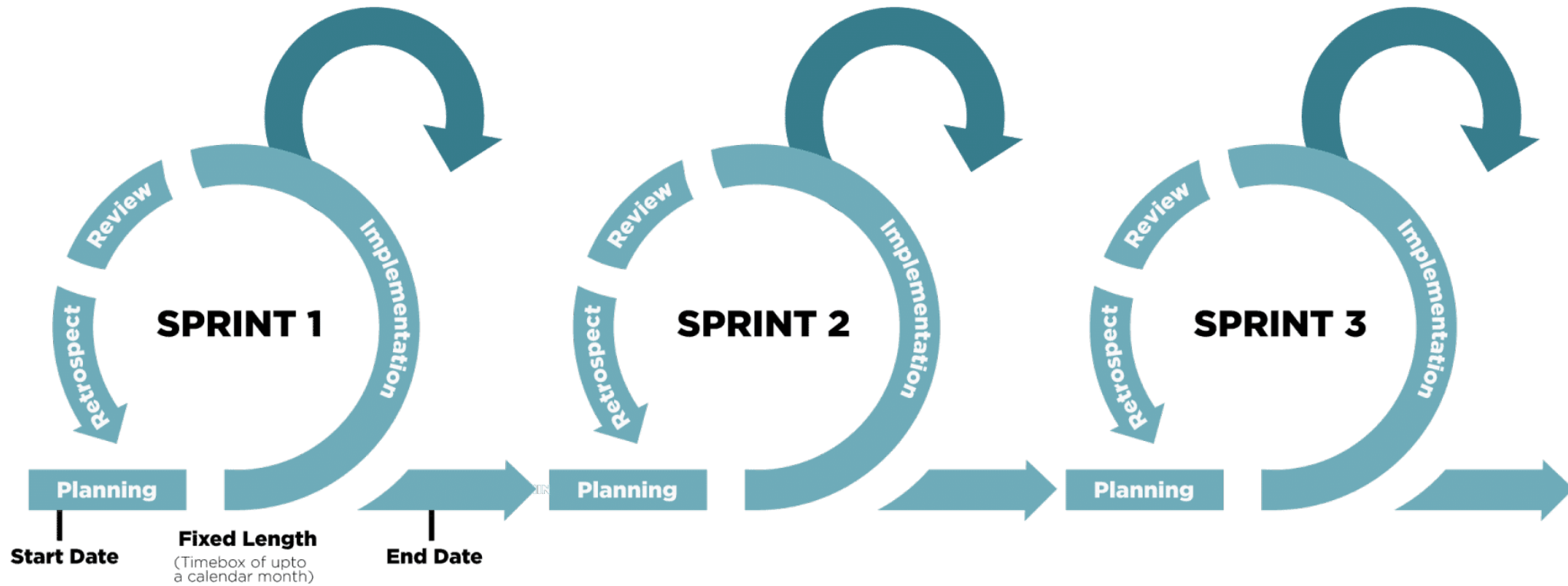
Red Flags from Vendors

Promised features to be delivered after you buy:

- Inconsistent delivery of the roadmap.

- Expensive embedded architectural requirements.

- Inability to use the software without use of professional services.

Source: Making Big Data Work

Agile

The Curse of Categorical Data

GoC data tends to be very heavy on the categories and text data.

Traditional analysis methods:

- not categorical data heavy.
- did not focus on doing complex analyses with (complex) categorical data.

This means we need to work harder to produce good strategies to deal with this type of data (hint: ML likes categories).
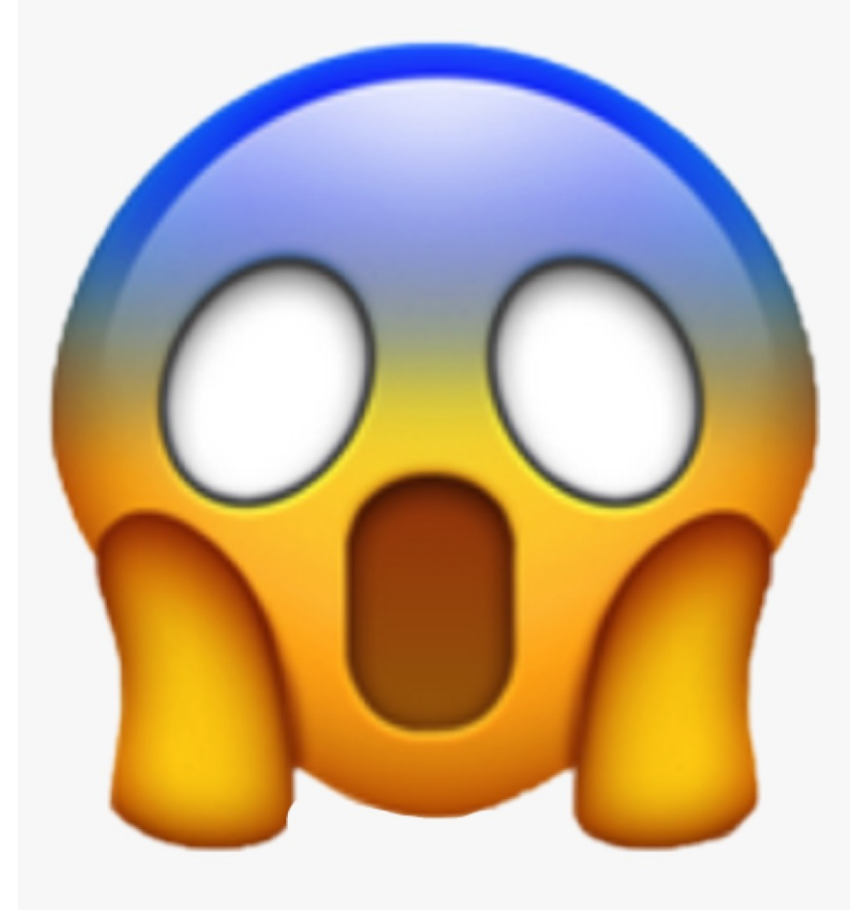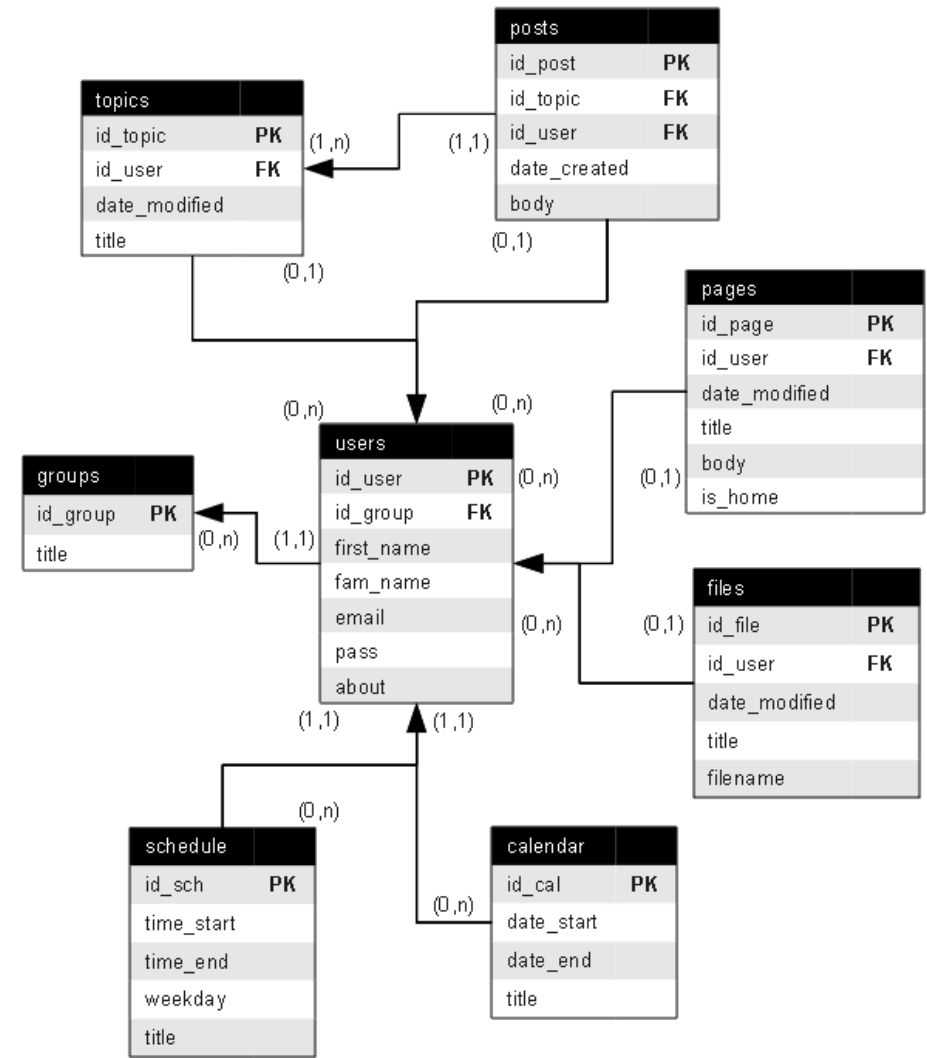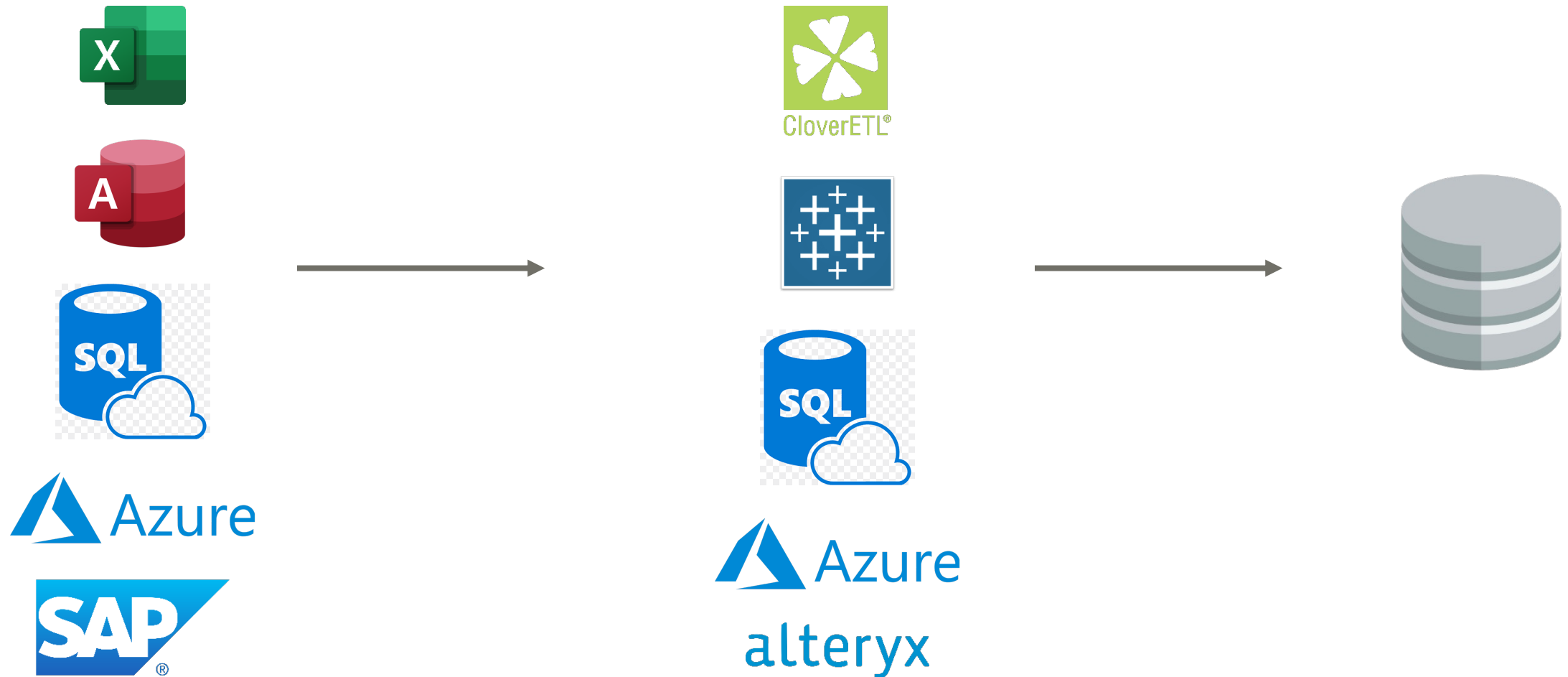
Data Storage

Four different options that are currently popular in terms of fundamental data and knowledge modelling or structuring strategies:

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

More important than you might think at first glance…

How to Clean Data: Tooling Perspective

What is Data Governance?

Data Governance is a concept that enables an organization to ensure that high data quality exists throughout the complete lifecycle of the data.

Focusing on Data Governance allows data throughout the organization to:

- be **available** when needed
- be **usable** when accessed
- be **consistent** when analyzed
- have **integrity** and be of high **quality**
- be **secure** and **trustworthy**

What is Data Governance?

Data governance encompasses:

- **people**, **processes**, **information technology**

It is required to create a consistent and proper handling of an organization's data across the enterprise.

It provides the foundation, strategy, and structure to ensure that data is managed as an asset and transformed into meaningful information.

Exercise: in pairs, list times where you have had issues because of data availability, usability, consistency, integrity, quality, security, or trustworthiness!

Data Governance in the GoC

Central point of reference for GoC (Digital Government website):

- [Strategic plans, policies, standards and guidelines related to government digital services](#)

Report to the Clerk of the Privy Council:

- [A Data Strategy Roadmap for the Federal Public Service](#)

Treasury Board Secretariat (selected):

- [Policy on Service and Digital](#)
- [Government of Canada Strategic Plan for Information Management and Information Technology 2017 to 2021](#)
- [Digital Operations Strategic Plan: 2018-2022](#)
- [Government of Canada Cloud Adoption Strategy: 2018 update](#)

Industry Canada:

- [Canada's Digital Charter in Action: A Plan by Canadians, for Canadians](#)
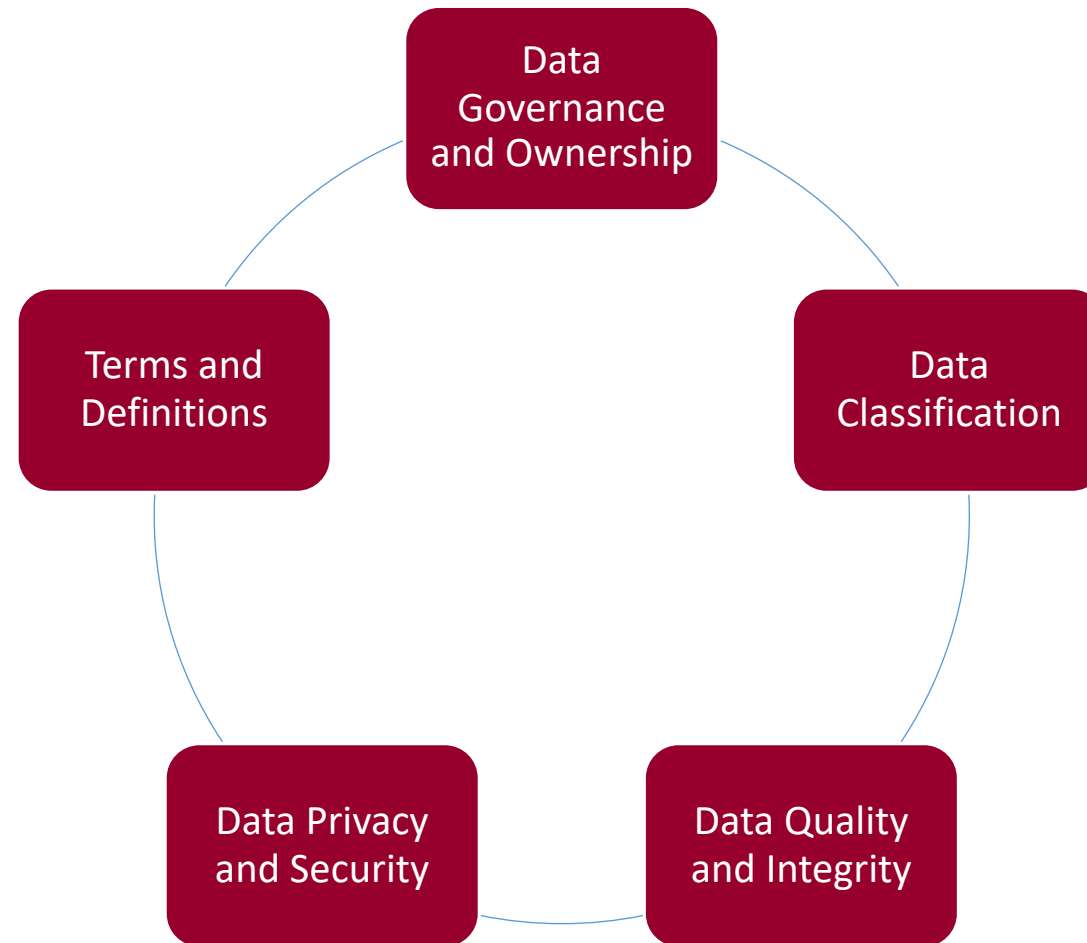
Data Governance

Data Management Association (DAMA)

DMBOK 2 (Data Management Body of Knowledge):

- Well detailed & thought through
- Sections reasonably aligned with GoC approach
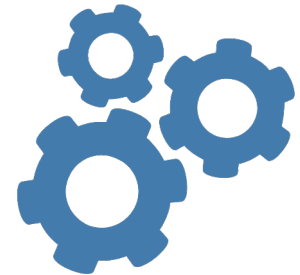- Backed by professional organization
- Not government focused

Data Governance
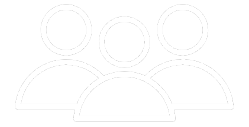
Aligning Data Needs With Strategy

Exercise: in groups, answer the following questions

- does your group create or generate data? if so, what data?
- do you use data from sources outside of your group? if so, which ones?
- how many sources of data (e.g., databases) does your group use, roughly speaking?
- do you publish analysis of data internally to your group? externally? both?

Accountability and Responsibility

	Data Owner	Data Steward	Data Custodian	Data Contributor	Data Consumer
Definition	A person who has governance and compliance responsibility for a set of data assets	A person who has business accountability for a set of data assets	A person who has technical accountability for a set of data assets	A person who creates or collects data that is relevant to the organization	A person who uses data to enable business outcomes
Accountability summary	▪ Compliance ▪ Risk ▪ Oversight ▪ Approval ▪ Champion ▪ Issue resolution	▪ Accuracy ▪ Consistency ▪ Business requirements ▪ Metadata definition and management ▪ Data Quality ▪ Data Curation ▪ Fitness for Purpose ▪ Governance ▪ Inventory ▪ RDM ▪ Role Management	▪ Security ▪ Access Management ▪ Availability ▪ Capacity ▪ Continuity ▪ Safeguarding ▪ Implementation ▪ Technical standards ▪ Configuration ▪ Control ▪ Modeling ▪ Versioning ▪ Change Management	▪ Data Acquisition and entry ▪ Data Quality ▪ Metadata Preparation ▪ Ethical and secure gathering of data ▪ Identification of issues ▪ Identification of new data sources	▪ Ethical use ▪ Report on Data Quality ▪ Report on fitness for purpose ▪ Identification of business and data rules ▪ Identification and reporting of data control ▪ Use in line with governance

Goals of Data Governance

1 — Create Self-Service Data Culture

2 — Establish Internal Rules for Data Use

3 — Implement Compliance Requirements

4 — Improve Internal and External Comms

5 — Increase Value of Data

6 — Reduce Costs

7 — Continually Manage Risks

8 — Ensure Continued Existence
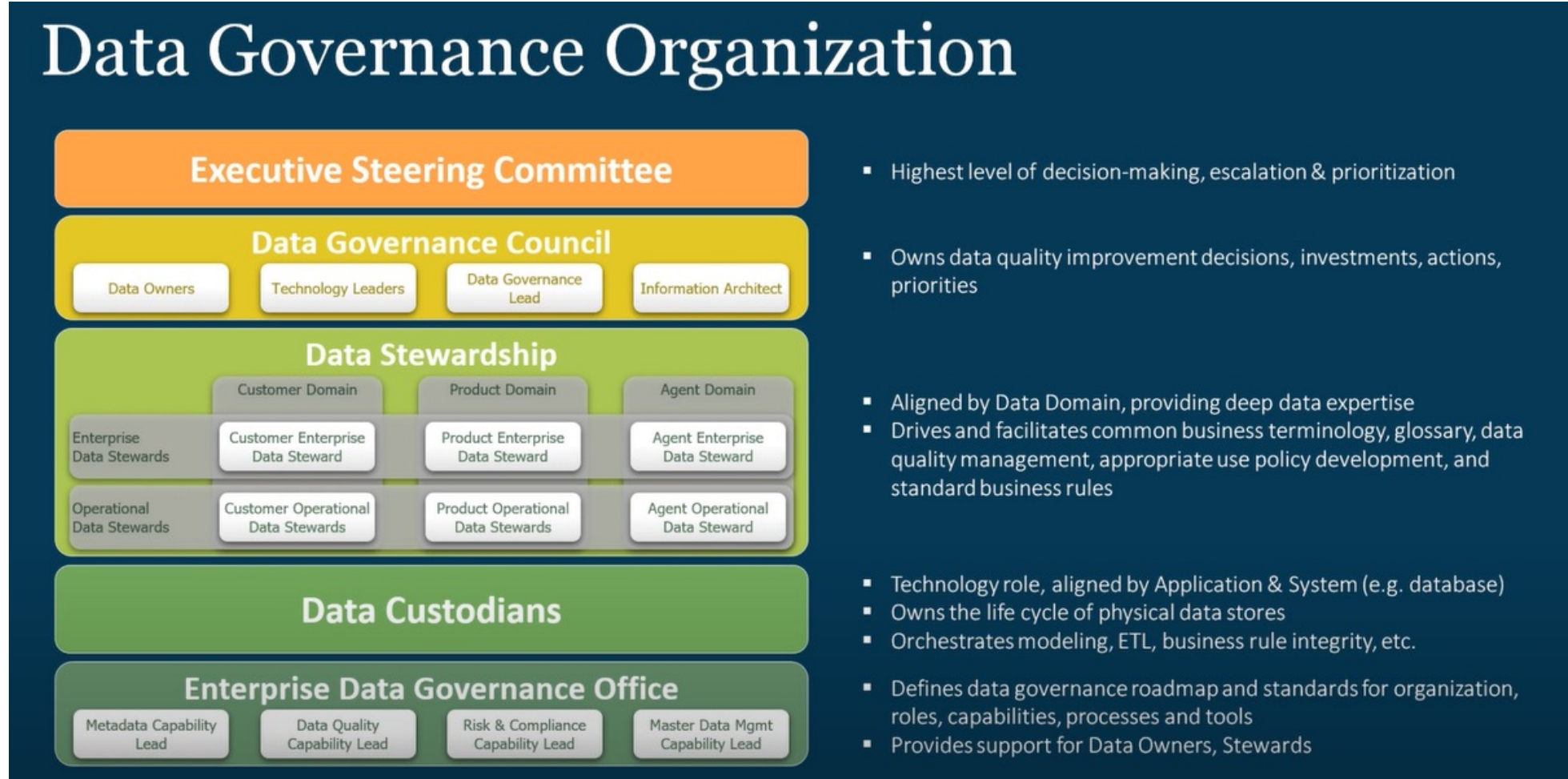
Common Issue with Data Governance

Pioneered, planned, and projected by the most technical resources

Verbiage is all over the place

Perception of 'donut meetings'

Possible Structure

Source: Informatica

Ad Hoc Assessment: Roadblock Recognition

Data Acquisition	Data Maintenance	Data Dissemination
Data Acquisition	Mostly Manual	Constant IT Red Tape
Multiple Data Sources	Data Integration Issues	Data Cleaning Mis-ordered
Manual Progress	Lack of Data Cleansing	Inconsistent Data
No Standards	Data Deduplication	Changes not Communicated
No Defined Process	Little to No Accessibility	Shoulder Tap Culture
No Data Validation	No Data Dictionaries	Reports Send Via Email/Chat

Luck and Information (Reprise)
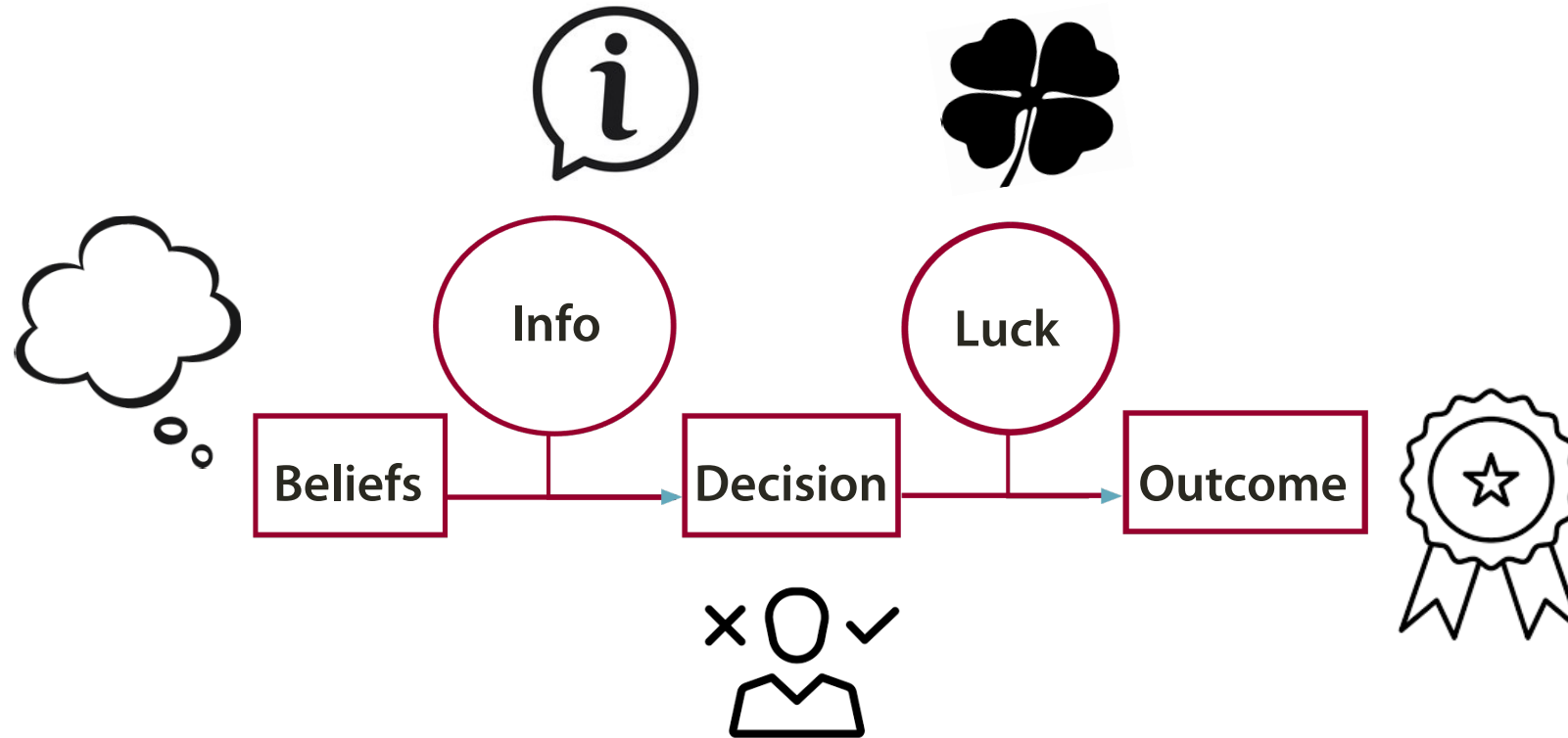
Suggested References

Storytelling with Data, by Cole Nussbaumer Knaflic

Weapons of Math Destruction, by Cathy O'Neil

How to Decide: Simple Tools for Making Better Choices, by Annie Duke

The Signal and the Noise, by Nate Silver

Superforecasting: The Art and Science of Prediction, by Philip E. Tetlock

Business Analytics: Data Analysis & Decision-Making, by Albright and Winston

Data Understanding, Data Analysis, and Data Science, by Patrick Boily

Roundtable

Unseal the decision