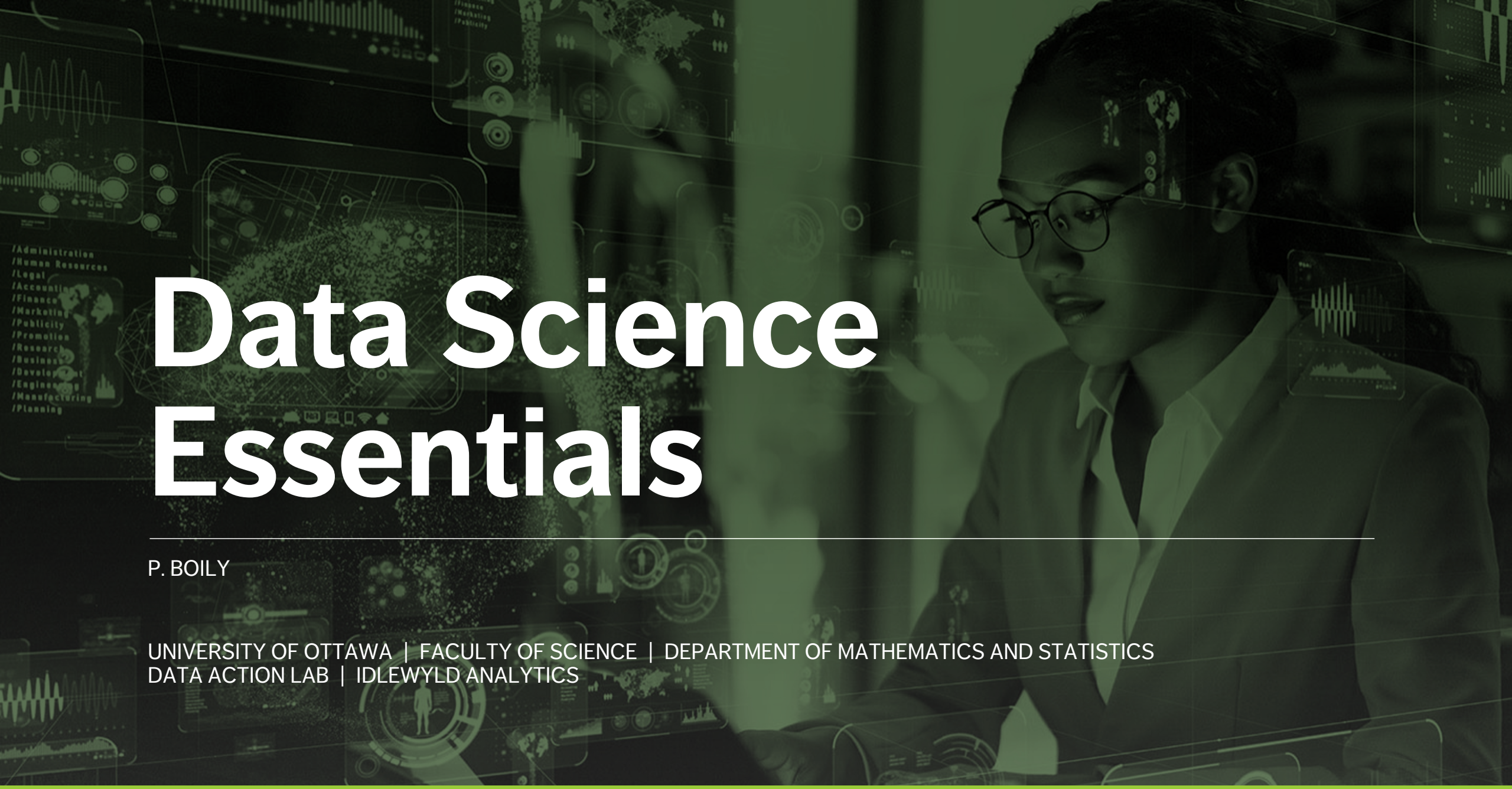
Data Science Essentials

Instructor: Patrick Boily

uOttawa

Institut de développement professionnel
Professional Development Institute



Data Science Essentials

P. BOILY

UNIVERSITY OF OTTAWA | FACULTY OF SCIENCE | DEPARTMENT OF MATHEMATICS AND STATISTICS
DATA ACTION LAB | IDLEWYLD ANALYTICS

Instructor – Patrick Boily

Employment

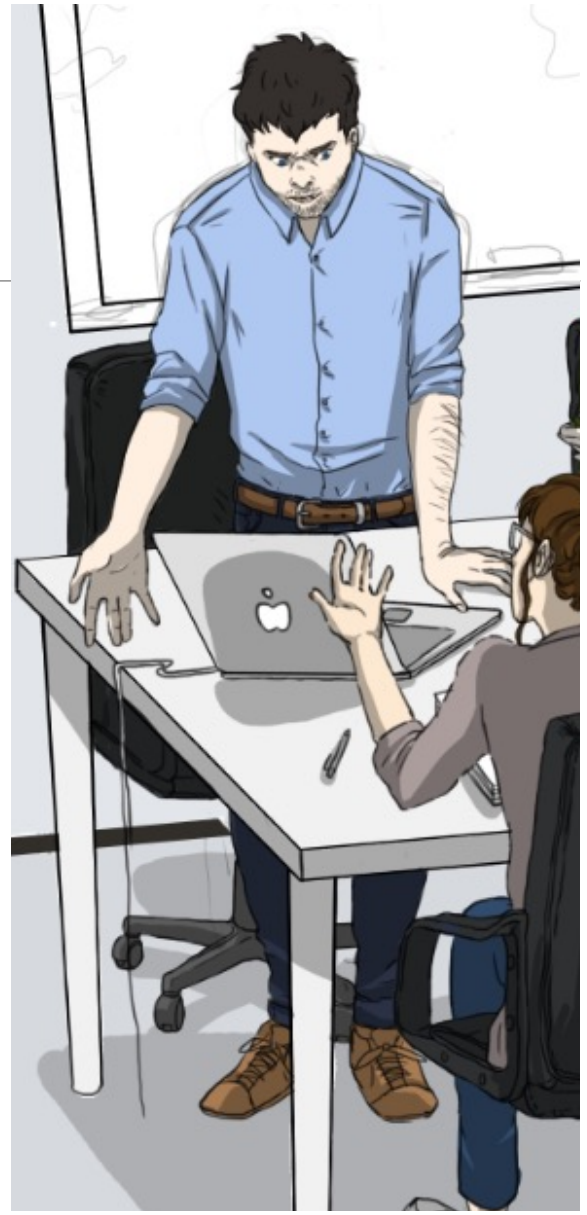
- Professor Math/Stat [‘19 – now, uOttawa]
- President [‘16 – now, Idlewyld Analytics]
- Manager and Senior Consultant [’12 – ‘19, CQADS, Carleton]
- Public Service [’08 – ’12, ASFC | StatCan | TC | TPSGC]
- 60+ uni course; 250+ workshop days

Projects

- GAC; NWMO; CATSA; etc.
- 40+ projects

Specialization

- Data visualization; data cleaning (... unfortunately)
- Application of wide breadth of techniques to all kinds of data
- Mathematical/statistical modeling



Course Material

Course Webpage:

https://data-action-lab.com/101-dse

Course Notes:

https://idlewyldanalytics.com

Contact Info:

pboily@uottawa.ca

Slack Workspace:

https://dspdi.slack.com

Course Description

This course gives participants the opportunity to master foundational knowledge and skills needed for data analysis, along with a discussion of common challenges and pitfalls.

Participants will be introduced to various methods of data preparation, and to some intrinsic limitations of data and data analysis, and to easily avoidable pre-analysis mistakes.

Following the course, the participants have the option of working on a guided project, getting feedback from the instructor.

Additional Information

Exposure to programming frameworks would be beneficial but not necessary. Participants must be comfortable (not necessarily experts) with the concepts introduced in a Probability and Statistics university-level course.

Participants are required to bring a laptop/personal computer on which the current version of R/RStudio (Posit) are installed (for which they may require administrative authorisation to install packages).

Participants doing the guided project must be familiar with R and/or Python.

Learning Outcomes

At the end of this course, participants will be able to:

- select appropriate methods to prepare their data for analysis
- anticipate challenges and limitations inherent to data and desired analysis outcomes
- apply data cleaning strategies to their data
- conduct simple analyses
- build simple data science pipelines to provide actionable insights

Course Outline

Technical and Non-Technical Aspects of Data Work

1. Quantitative Skills
   - Software and Tools
   - Multiple I's Approach
   - Roles and Responsibilities
   - Analysis Cheat Sheet

Data Science Basics

2. Preliminaries
3. Conceptual Frameworks
4. Data Science Ethics
5. Analytics Workflows
6. Getting Insight From Data

Session 1
Session 2
Session 3
Session 4

Course Outline

Data Preparation

7. Data Quality and Data Wrangling
8. Missing Values
9. Anomalous Observations
10. Dimensionality & Data Transformations

Miscellanea

11. Data Engineering
12. Data Management

Session 1

Session 2

Session 3

Session 4

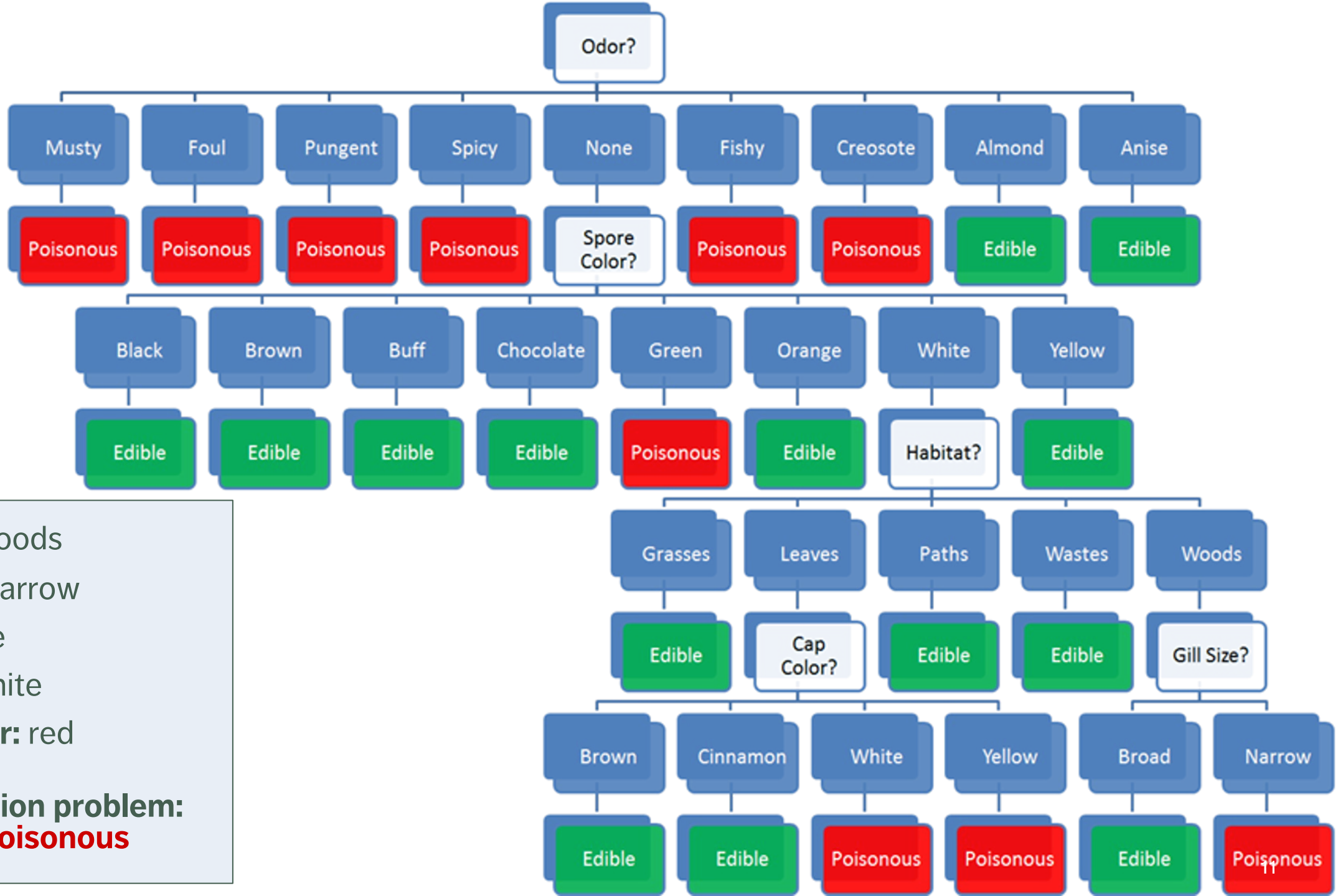
[https://archive.ics.uci.edu/ml/datasets/Mushroom]

Poisonous Mushroom Problem

Amanita muscaria

Habitat: woods
Gill Size: narrow
Odor: none
Spores: white
Cap Colour: red

Classification
Is Amanita muscaria edible, or poisonous?



Habitat: woods

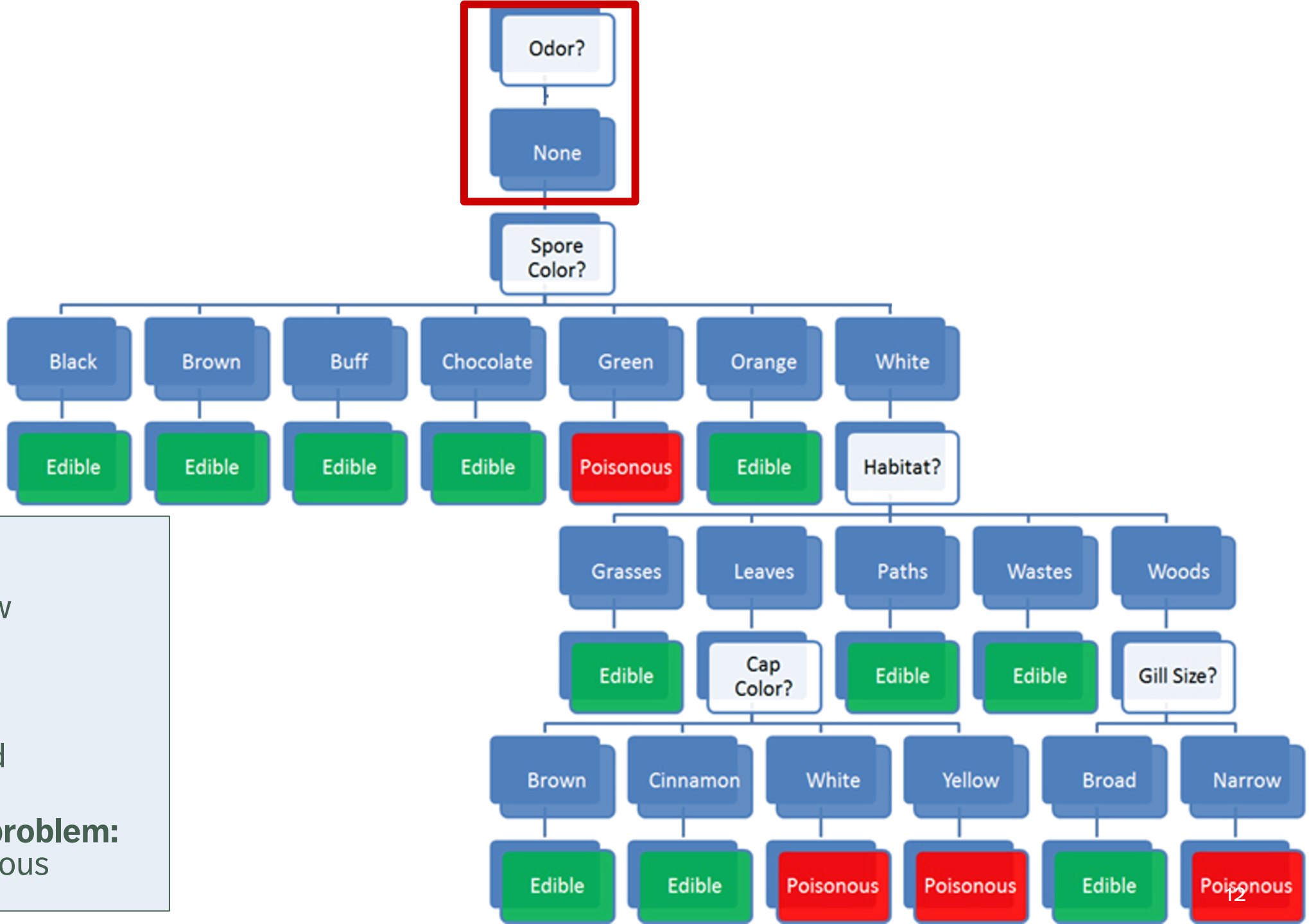
Gill Size: narrow

Odor: none

Spores: white

Cap Colour: red

Classification problem:
edible or **poisonous**



Habitat: woods

Gill Size: narrow

Odor: **none**

Spores: white

Cap Colour: red

Classification problem: edible or poisonous

Habitat: woods

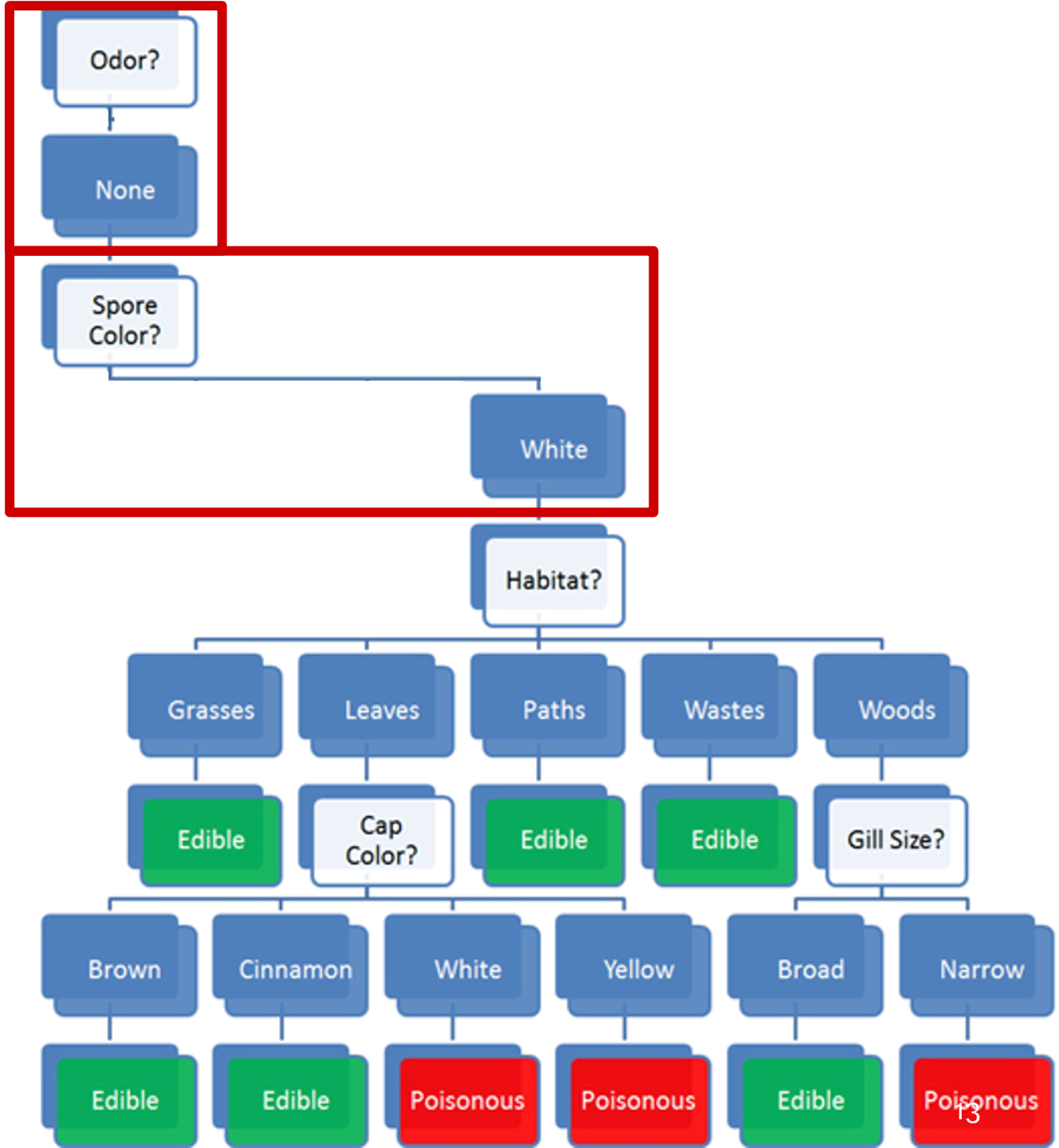
Gill Size: narrow

Odor: none

Spores: **white**

Cap Colour: red

Classification problem:
edible or poisonous



Habitat: woods

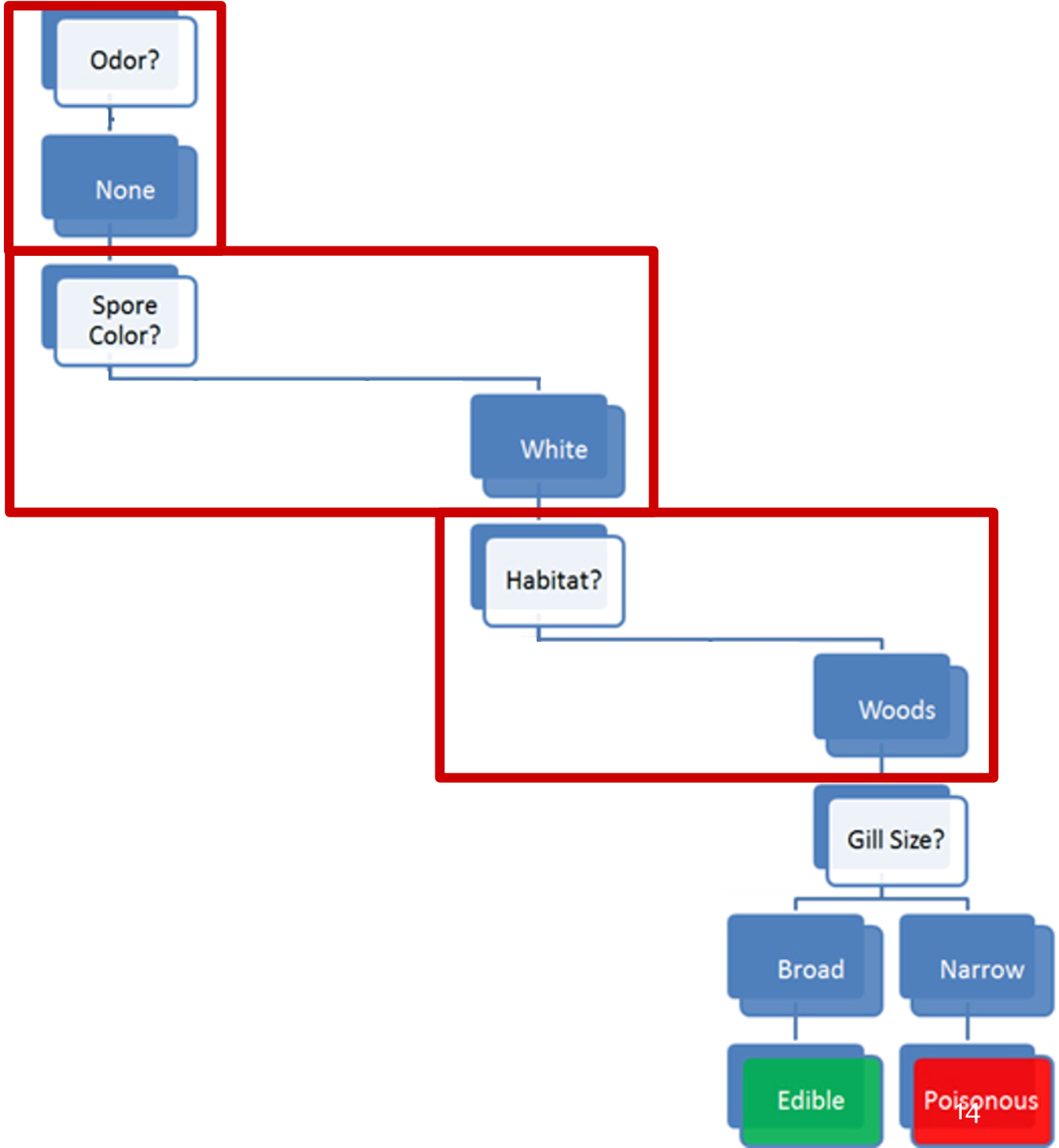
Gill Size: narrow

Odor: none

Spores: white

Cap Colour: red

Classification problem:
edible or poisonous



Habitat: woods

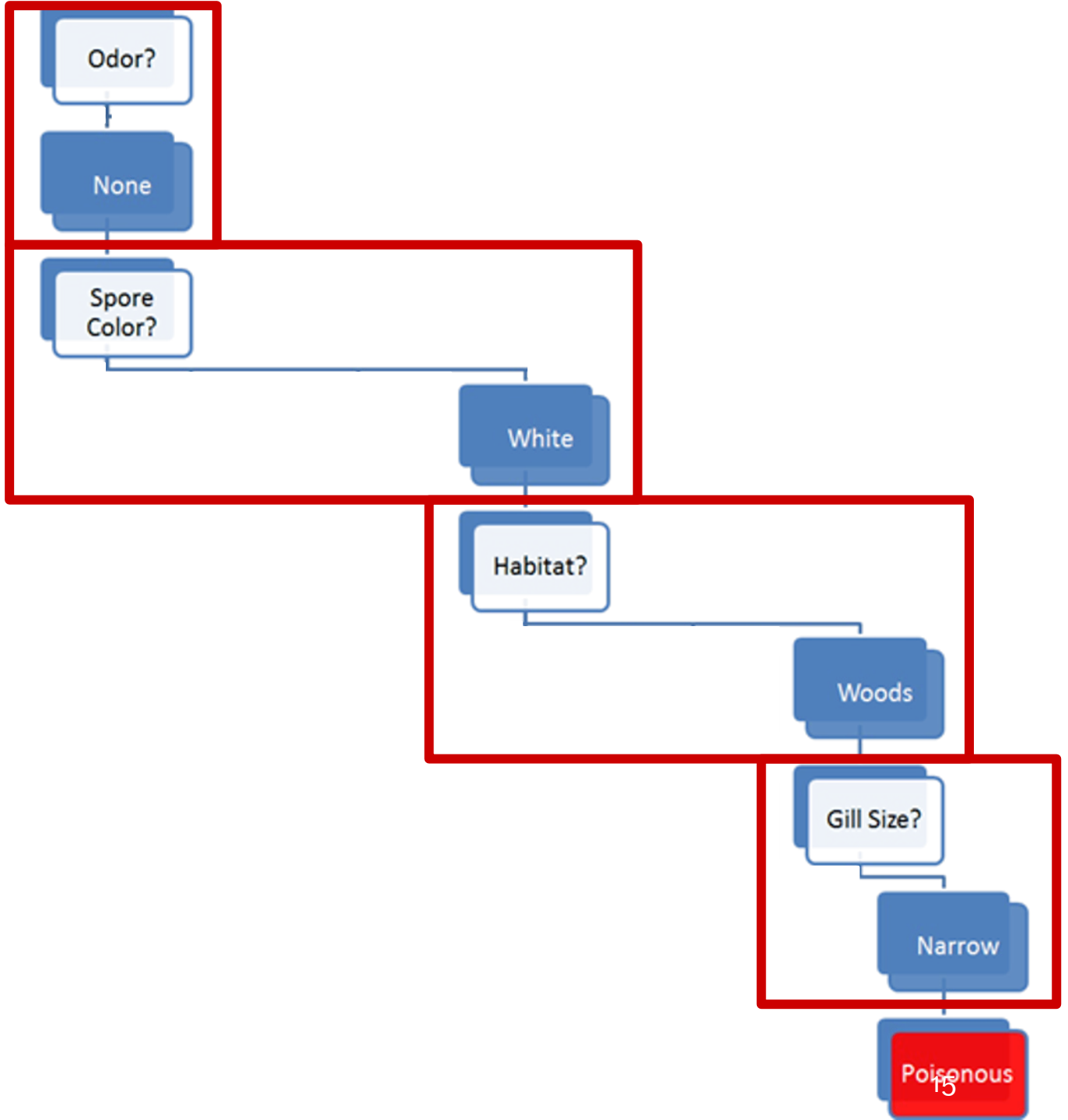
Gill Size: narrow

Odor: **none**

Spores: white

Cap Colour: red

Classification problem:
edible or poisonous



Habitat: woods

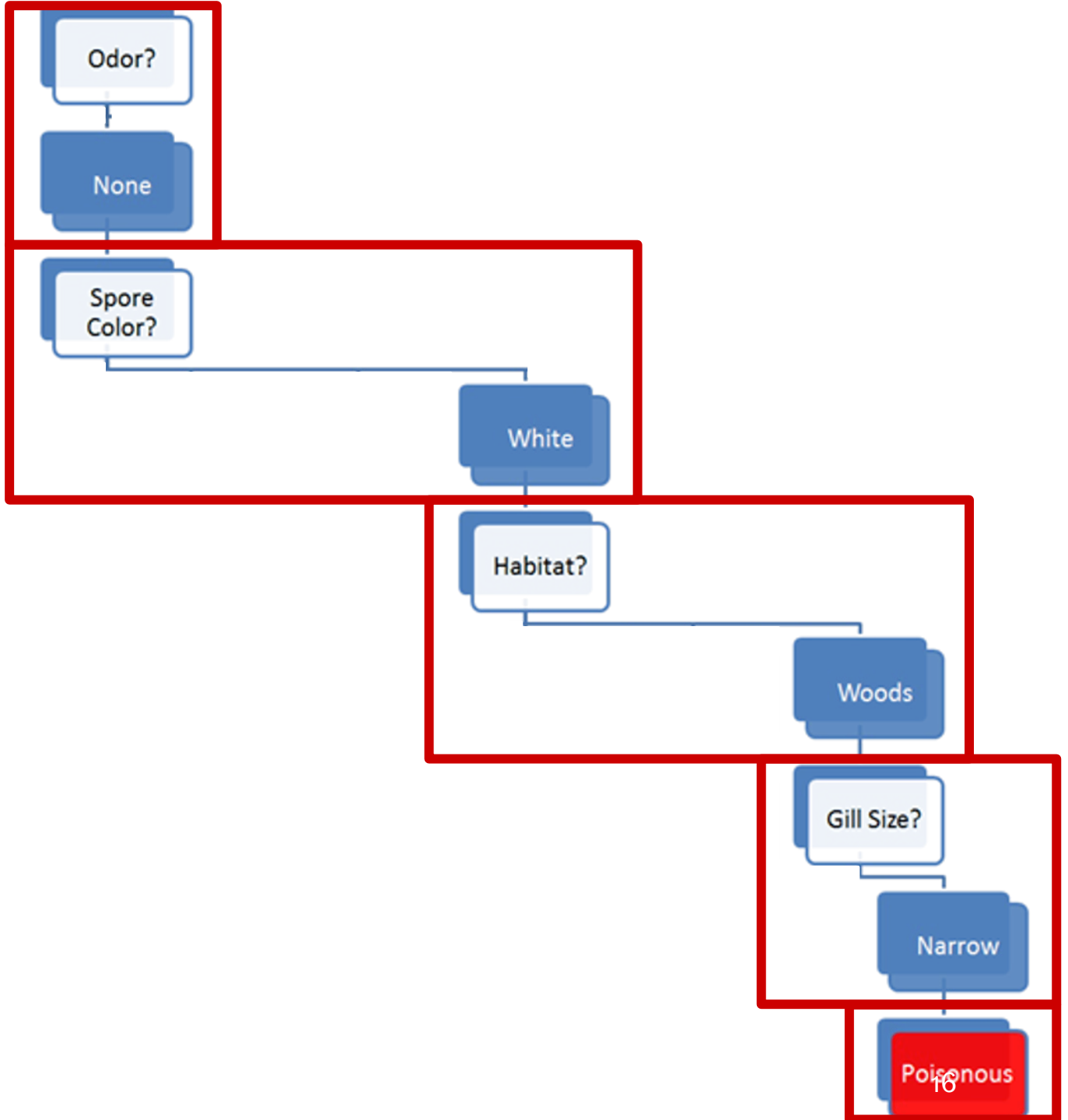
Gill Size: narrow

Odor: none

Spores: white

Cap Colour: red

Classification problem:
edible or **poisonous**



Discussion

Would you have trusted an “**edible**” prediction?

Where is the model coming from?

What would you need to know to trust the model?

What’s the cost of making a classification mistake, in this case?

Sister Courses

DATA VISUALIZATION AND DASHBOARDS

1. Data Visualization Concepts
2. Dashboarding
3. Storytelling with Data
4. Visualizations with ggplot2

INTRODUCTION TO MACHINE LEARNING

1. Statistical Learning
2. Classification
3. Clustering
4. Data Issues and ML Challenges

Session 1

DATA SCIENCE ESSENTIALS

Technical and Non-Technical Aspects of Data Work

DATA SCIENCE ESSENTIALS

1. Technical & Non-Technical Aspects of Data Work

Quantitative Skills

Out-of-academia context:

- apply **quantitative methods** to (business) problems in order to obtain **actionable insight**
- difficult for any given individual to have expertise in **every** field of mathematics, statistics, computer science, data science, data engineering, etc.

With a graduate degree in math/stats, for instance:

- **expertise** in 2-3 areas
- **decent understanding** of related disciplines
- **passing knowledge** in various domains

Flexibility is an ally, perfectionism... only up to a point.

Quantitative Skills

Suggestions:

- **keep up with trends**
- become **conversant in your non-expertise areas**
- know **where to find information**

In many instances (70%?), only the basics (2nd–3rd year mandatory courses at uOttawa, say) are sufficient to meet government/industry needs.

Focus: make sure you really **understand** the basics, stepping stones.

In the rest of the cases, more sophisticated knowledge is required.

Quantitative Skills

- survey sampling and data collection
- data processing and data cleaning
- data visualization
- mathematical modelling
- statistical methods
- regression analysis
- queueing models
- machine learning
- deep learning
- reinforcement learning
- stochastic modelling (MC simulations)
- optimization and operations research
- survival analysis
- Bayesian data analysis
- anomaly detection and outlier analysis
- feature selection/dimensions reduction
- trend extraction and forecasting
- cryptography and coding theory
- design of experiment
- graph and network theory
- text mining/natural language processing
- etc.

Software and Tools

Modern quantitative work typically involves **programming** (or the use of point-and-click software, at the very least).

But programming languages **go in and out of style.**

It is important not just to understand the syntax of a particular language, but also how computer languages and computing infrastructure work in general.

ALSO: avoid getting caught up in programming wars … they're more or less all functionally equivalent!

Software and Tools

Programming (and Related)

- Python, R, C/C++/C#, Perl, Julia, regexps (, Visual Basic?), Java, Ruby, etc.

Database Management

- SQL and variants, ArangoDB, MongoDB, Redis, Amazon DynamoDB (, Access?), Big Query, Redshift, Synapse, etc.

Data Visualization

- ggplot2, seaborn, plot.ly, Power BI, Tableau, D3.js, Google Data Studio, proprietary software, etc.

Simulations, Statistical Analysis, Data Analysis, Machine Learning

- tidyverse, scikit-learn, numpy, pandas, scipy, MATLAB, Simulink, SAS, SPSS, STATA (, Excel?), Visio, TensorFlow, keras, Spark, Scala, etc.

Typesetting and Reporting

- LaTeX, R Markdown, Adobe Illustrator, GIMP (, Word?, PowerPoint?), etc.

Software and Tools

Q: At StatCan, R or SAS?

A: Not easy to answer as StatCan is in a slow transition period. The Agency is better equipped for SAS (with "Big Data" options, such as SAS Grid).

R is [...] not as ideal for large files (e.g., Census data), so it is not an option in such cased because it is still too slow (unless you have very powerful servers). But we would prefer to use the R packages, so it's a dilemma.

TL;DR: R is our future, but SAS is still very much our present. In times of transition, **analysts/employees who know both are better positioned**.

Multiple I's Approach to Data Work

Technical and quantitative proficiency (or expertise) is **necessary** to do good quantitative work *in the real world*, but it is **not sufficient** – optimal real-world solutions may not always be the optimal academic or analytical solutions.

This might be the biggest surprise for those transitioning out of academia.

What works for one person, one job application, one project, one client, etc. may not work for another – **beware the tyranny of past success**!

The focus of quantitative work must include the delivery of **useful analyses/ products** (Multiple "I"s).

Multiple I's Approach to Data Work

- **intuition**
  understanding context
- **initiative**
  establishing an analysis plan
- **innovation**
  new ways to obtain results, if required
- **insurance**
  trying multiple approaches
- **interpretability**
  providing explainable results
- **insights**
  providing actionable results
- **integrity**
  staying true to objectives and results
- **independence**
  self-learning and self-teaching
- **interactions**
  strong analyses through teamwork
- **interest**
  finding and reporting on interesting results
- **intangibles**
  thinking "outside the box";
- **inquisitiveness**
  not only asking the same questions repeatedly

Multiple I's Approach to Data Work

Prospective employees/analysts are not solely gauged on technical know-how, but also on the ability to **contribute positively** to the workplace/project:

- communication
- team work and multi-disciplinary abilities
- social niceties and flexibility
- non-technical interests

Employers rarely chose robots when human beings are available; stakeholders are more likely to accept data recommendations from **well-rounded people**.

You should also evaluate eventual employers/clients on these axes.

Roles and Responsibilities

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

Successful projects require **teams** of highly-skilled individuals who understand the **data**, the **context**, and the **challenges**.

Team *size* could vary from a few to several dozens; typically easier to manage small-ish teams (with 1-4 members, say, with **role overlaps**).

Domain Experts / SMEs

- are authorities in a particular area or topic
- guide team through unexpected complications and knowledge gaps

Roles and Responsibilities

Project Managers / Team Leads

- understand the process enough to recognize whether what is being done makes sense
- provide realistic estimates of the time and effort required to complete tasks
- act as intermediary between team and clients/stakeholders
- responsible for project deliverables.

Data Translators

- have a good grasp on the data and the data dictionary
- help SMEs transmit the underlying context to the data science team

Data Engineers / Database Specialists

- work with clients and stakeholders to acquire useable data sources
- may participate in the analyses, but are not necessarily specialists

Roles and Responsibilities

Data Analysts

- clean and process data
- prepare initial visualizations
- have a decent understanding of quantitative methods (at most 1 area of expertise)
- conduct preliminary analyses

Data Scientists

- work with processed data to build sophisticated models
- focus on actionable insights
- have a sound understanding of algorithms/quantitative methods (2 or 3 areas of expertise)
- can apply them to a variety of data scenarios
- can be counted on to catch up on new material quickly

Roles and Responsibilities

Computer Engineers

- design and build computer systems and pipelines
- are involved in software development and deployment of data science solutions

AI/ML Quality Assurance/Quality Control Specialists

- design testing plans for solutions that implement AI/ML models
- help the team determine whether the models are able to learn

Communication Specialists

- communicate actionable insights to managers, policy analysts, decision-makers, stake holders
- may participate in the analyses, but are not necessarily specialists (often data translators)
- keep abreast of popular accounts of quantitative results and developments

Analysis Cheat Sheet

1. Business solutions are not always academic solutions.
2. Data and models don't always support the stakeholder's hopes, wants, needs.
3. Timely communication is key – externally and internally.
4. Data scientists need to be flexible (within reason), and willing and able to learn something new, quickly.
5. Not every problem calls for data science methods.
6. We should learn from both our good and our bad experiences.

Analysis Cheat Sheet

7. Manage projects and expectations.
8. Maintain a healthy work-life balance.
9. Respect the stakeholders, the project, the methods, and the team.
10. Data science is not about how smart we are; it is about how we can provide actionable insight.
11. When what the client wants can't be done, offer alternatives.
12. "There ain't no such thing as a free lunch."

Suggested Reading

Technical and Non-Technical Aspects of Data Work

Data Understanding, Data Analysis, Data Science
Non-Technical Aspects of Data Work

First Principles

- The “Multiple I’s” Approach
- Roles and Responsibilities
- Consulting/Analysis Cheatsheet

Lessons Learned

Exercises

Technical and Non-Technical Aspects of Data Work

1. Which of the quantitative skills presented in this section do you possess? Which interest you? Which do you plan on learning about?
2. Which of the software skills presented in this section do you possess? Which interest you? Which do you plan on learning about?
3. What data role do you hold in your organization? Which role do you think you are currently best suited for? Which role do you aspire to?
4. Have you encountered the Analysis Cheat Sheet lessons in your work? Have you encountered others?

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

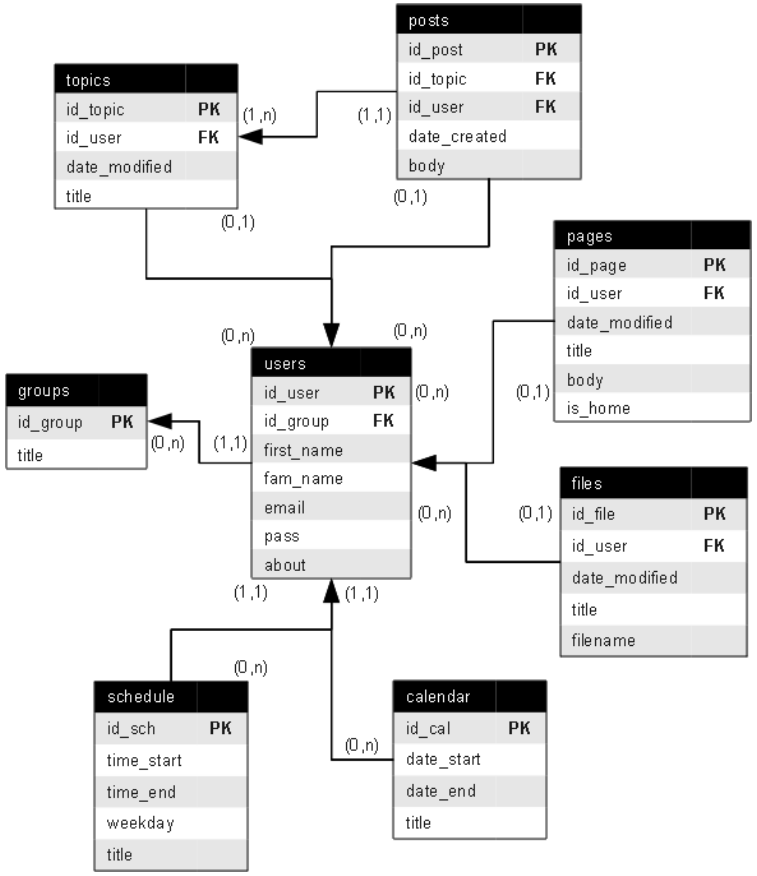
The dataset of physical objects could start with:

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?"

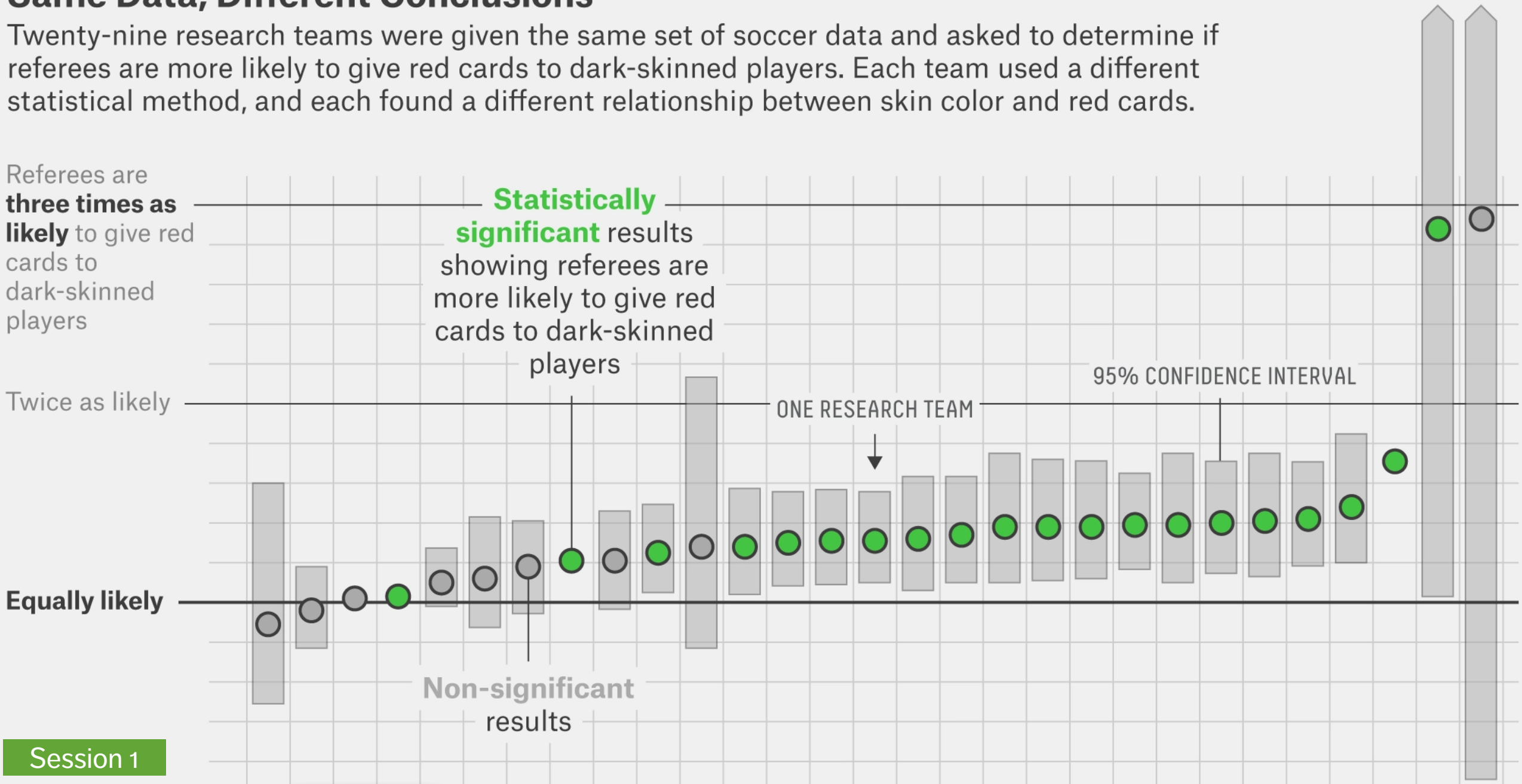
Opinions on the topic are varied – to some, DS/ML/AI provide examples of **brilliant successes**, while to others it is the **dangerous failures** that are at the forefront.

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
Data Science Basics

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

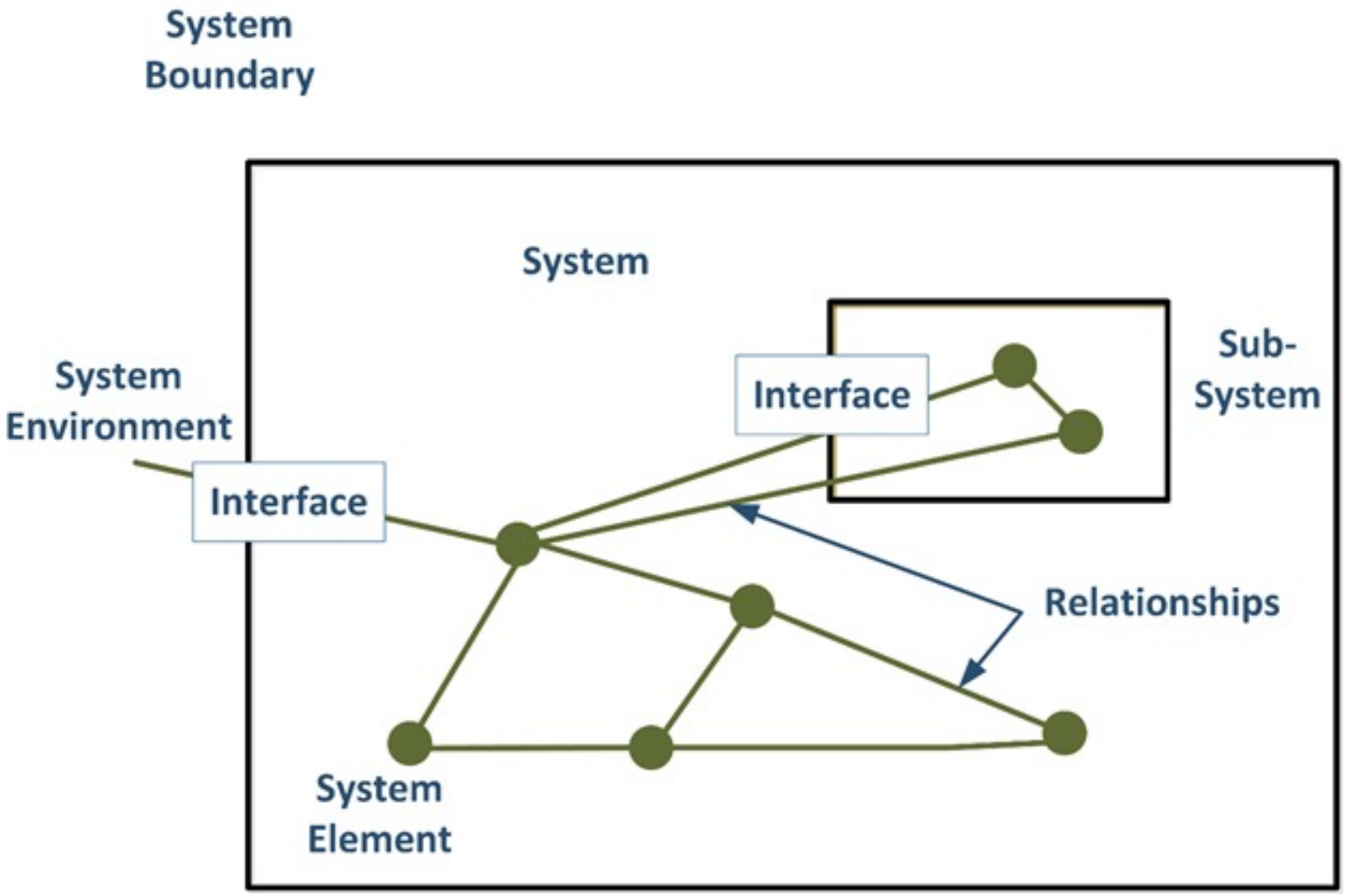
General Systems Theory describes **disparate** natural phenomena using a **common conceptual framework**, all as systems of interacting objects.

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.

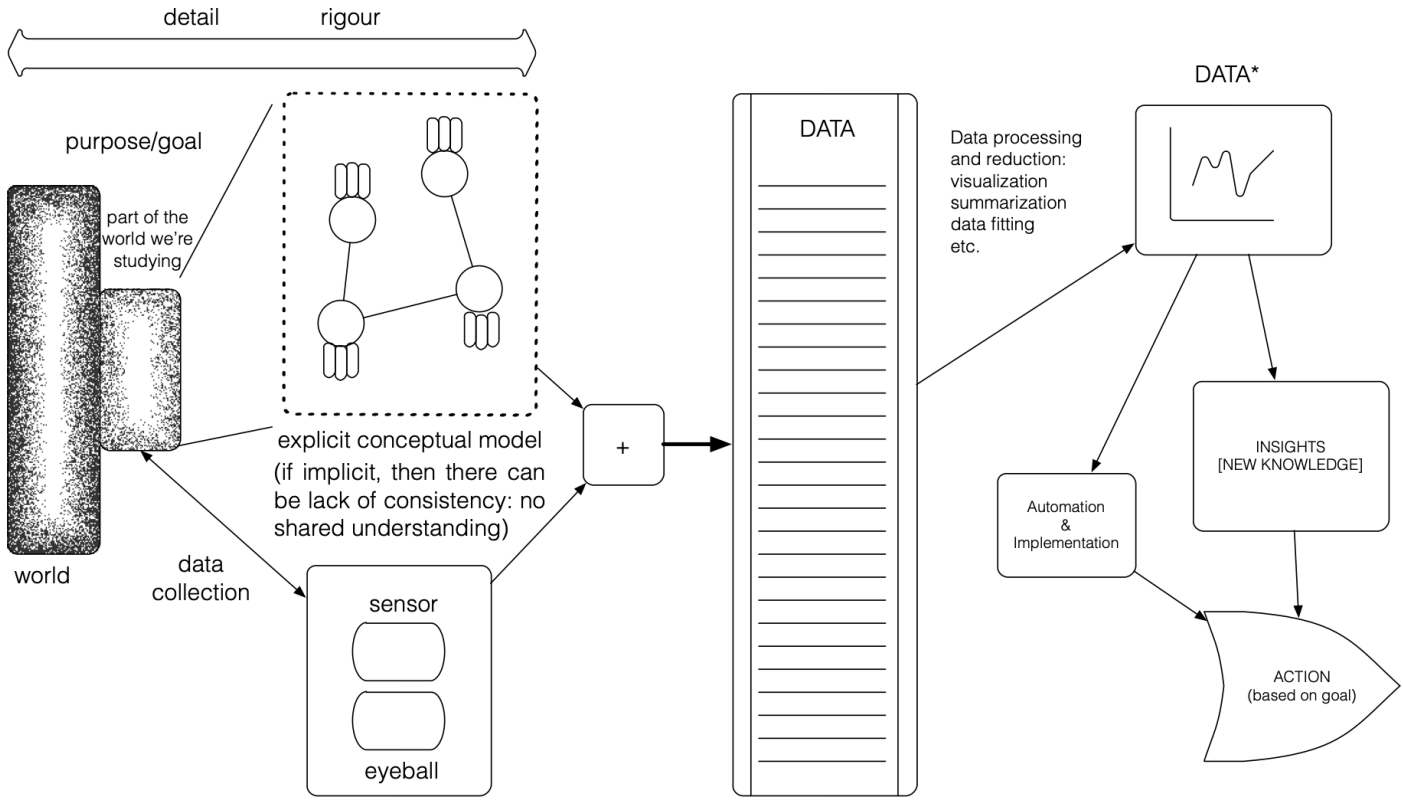
Conceptual Models

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**

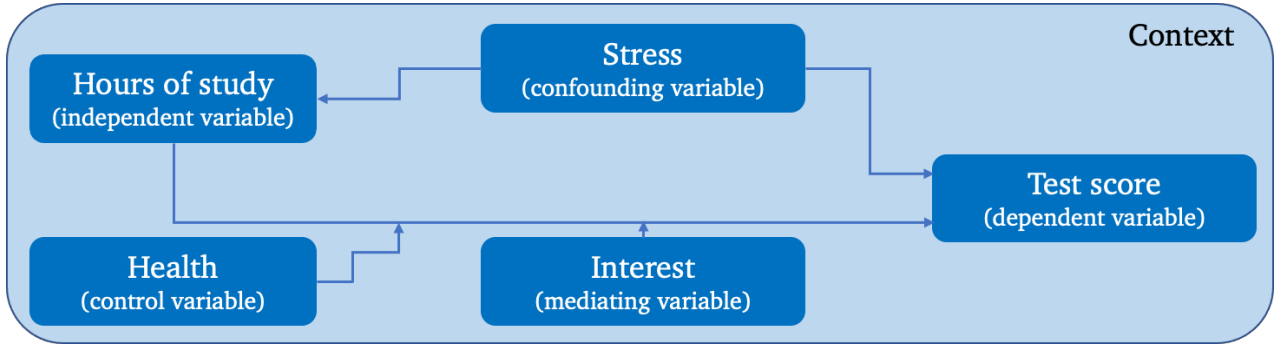


Conceptual Models

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).



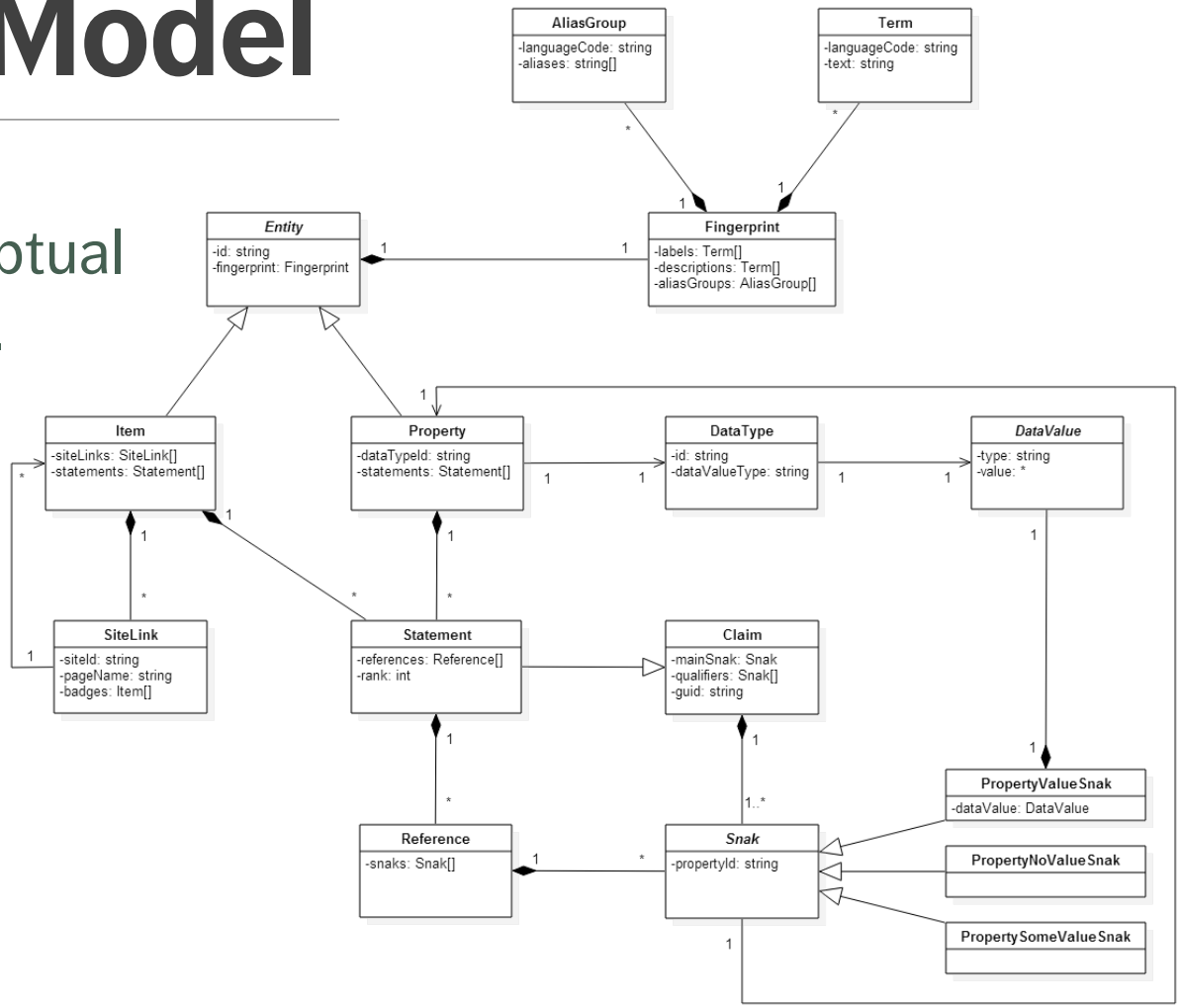
Formal Conceptual Model

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
Data Science Basics

Conceptual Frameworks for Data Work

- 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?

Session 2

DATA SCIENCE ESSENTIALS

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
Data Science Basics

Ethics in the Data Science Context

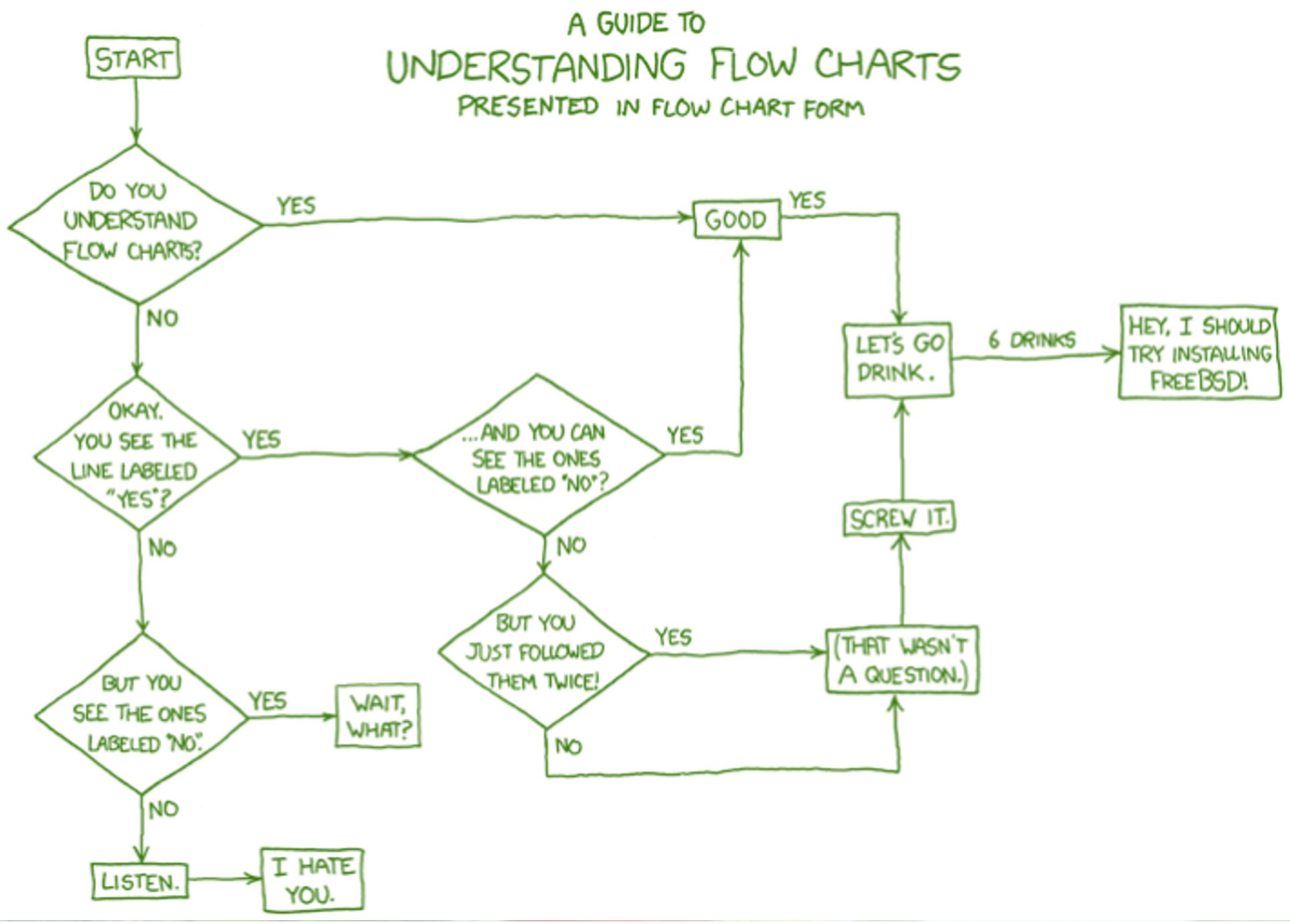
- The Need for Ethics
- What Is/Are Ethics?
- Ethics and Data Science
- Guiding Principles

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

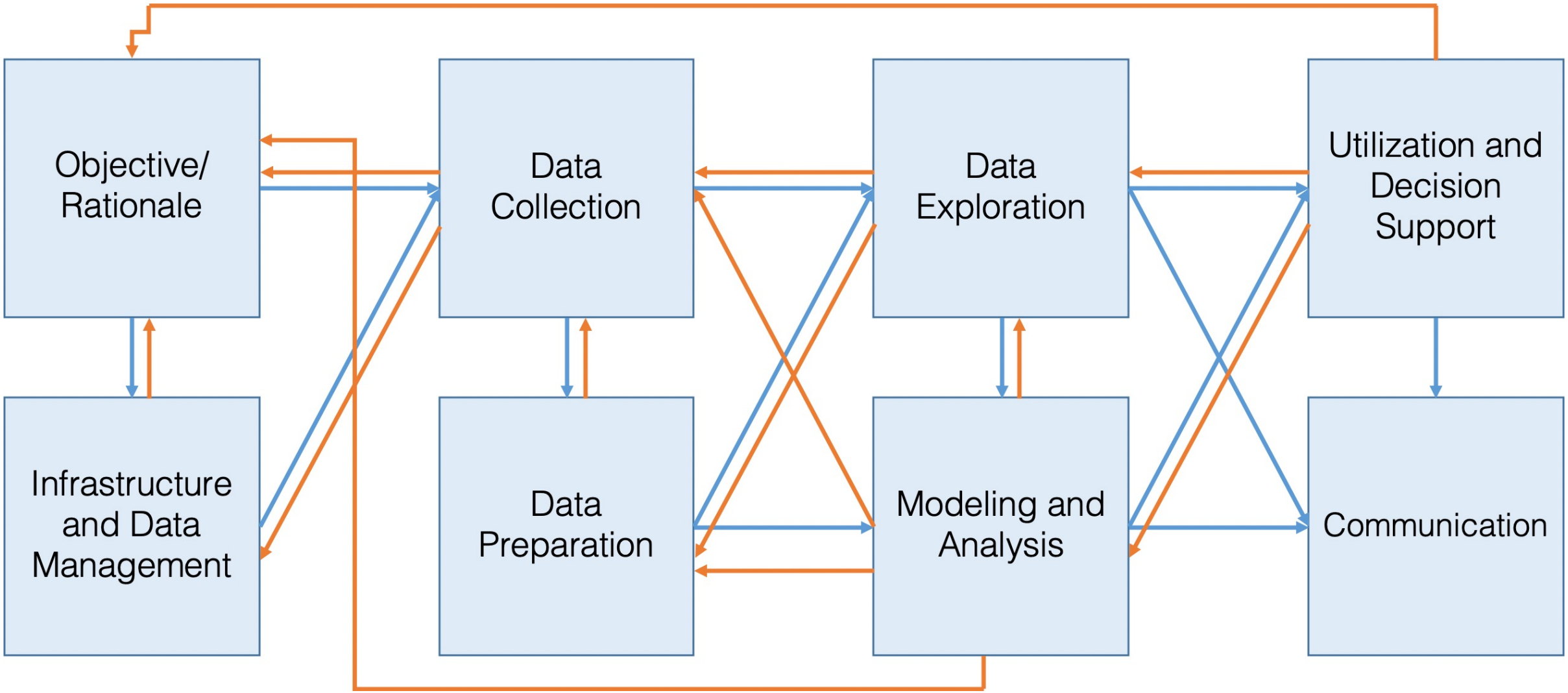
As with the **scientific method**, there is a “step-by-step” guide to data analysis:

- 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.

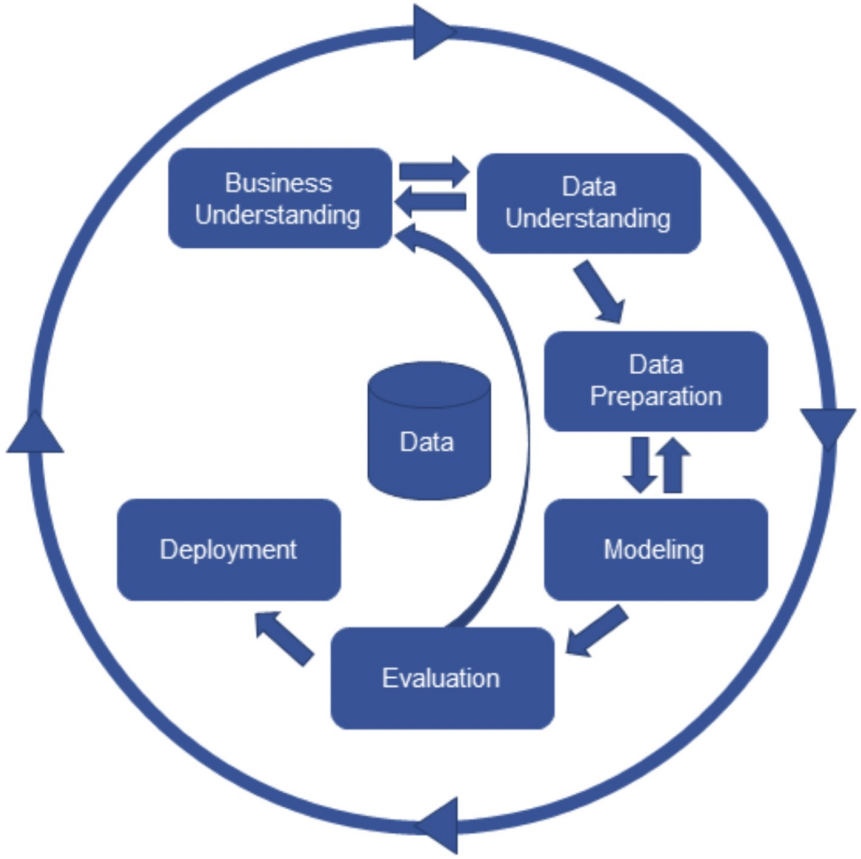
In practice, the process is quite often **messy**, with steps added in and taken out of the sequence, repetitions, re-takes, etc.

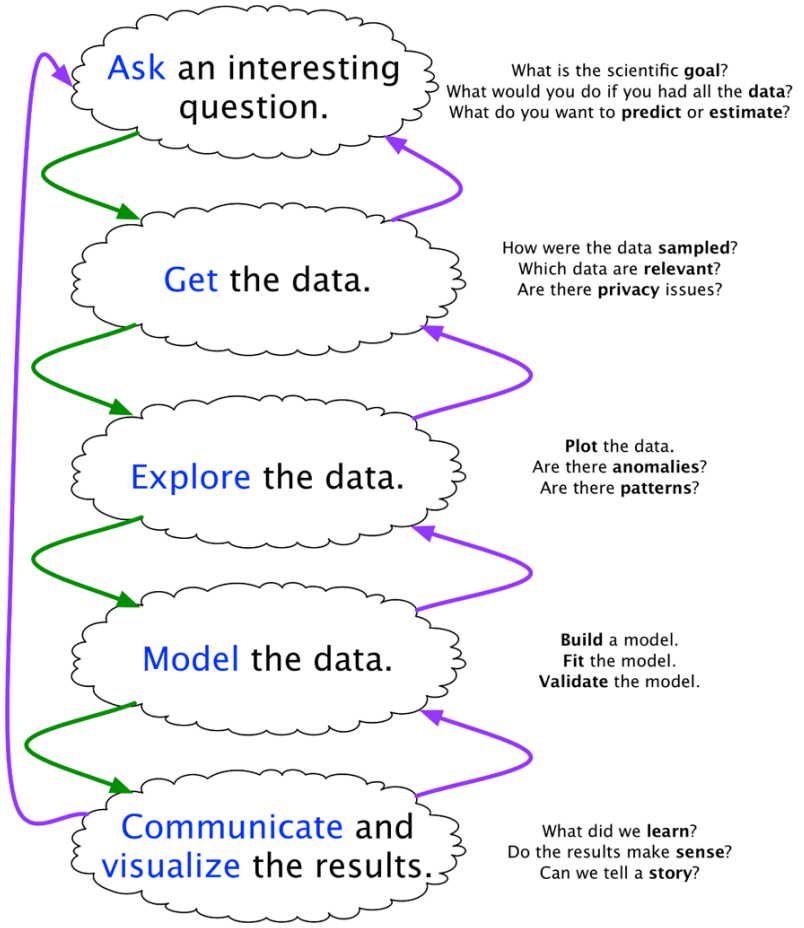
Surprisingly, it tends to work… when **conducted correctly**.


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.

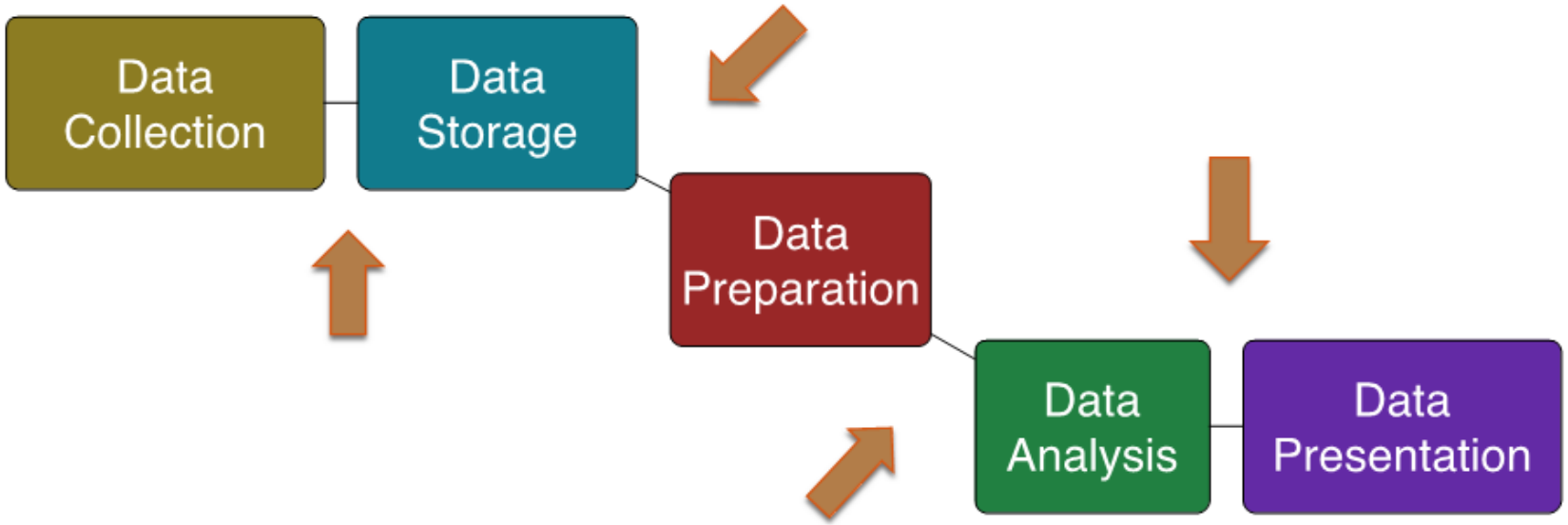
Data pipelines usually consist of 9 components (5 **stages** and 4 **transitions**):

- 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
Data Science Basics

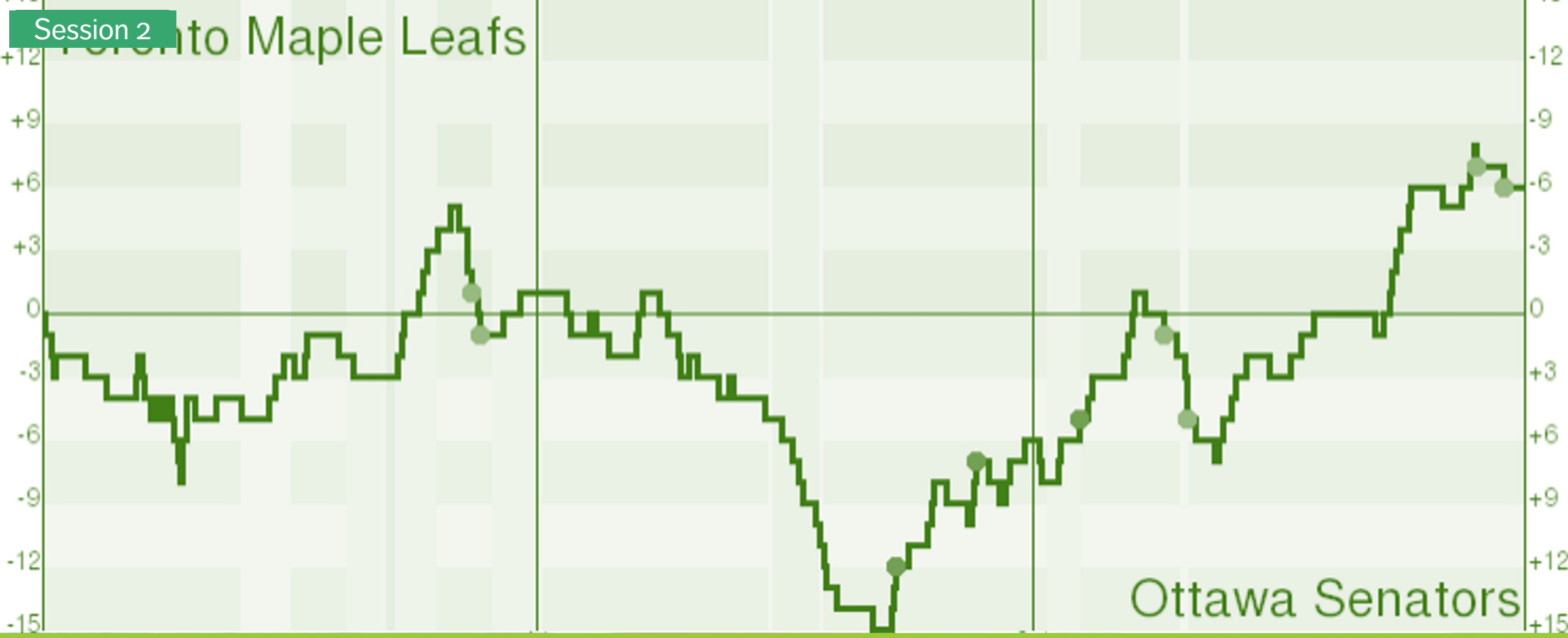
Analytics Workflows

- The "Analytical" Methods
- Data Collection, Storage, Processing, and Modeling
- Model Assessment and Life After Analysis
- Automated Data Pipelines

Exercises

Analytics Workflows

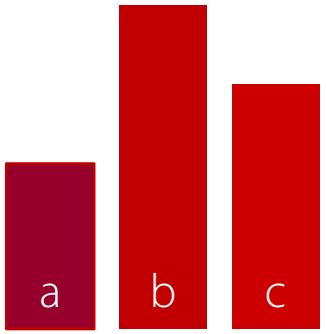
1. Install R / RStudio (Posit), and packages from the list the instructor will provide.
2. Test the installation with examples from the Programming Primer (sections 2 – 4) to make sure that the software performs as expected.



6. Getting Insight From Data

Analytics Modes

Descriptive



Show **what** happened

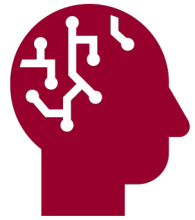
Diagnostic

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

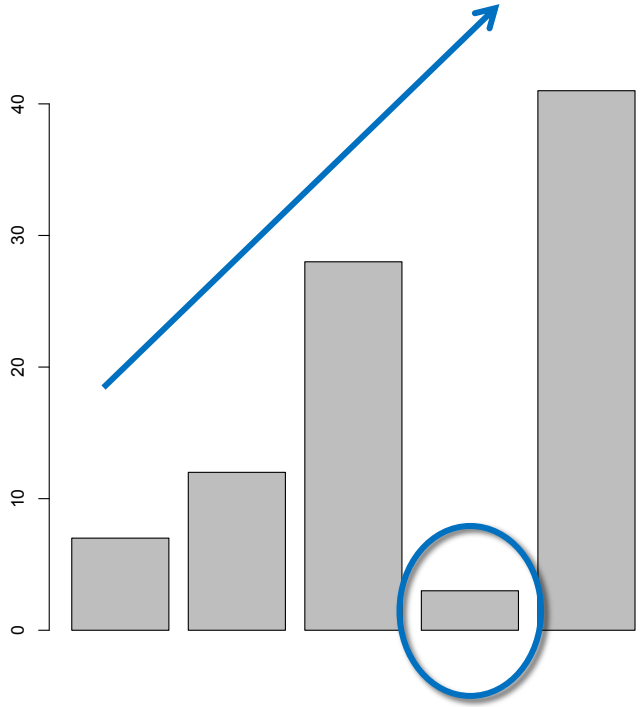
Analysis Through Visualization

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)

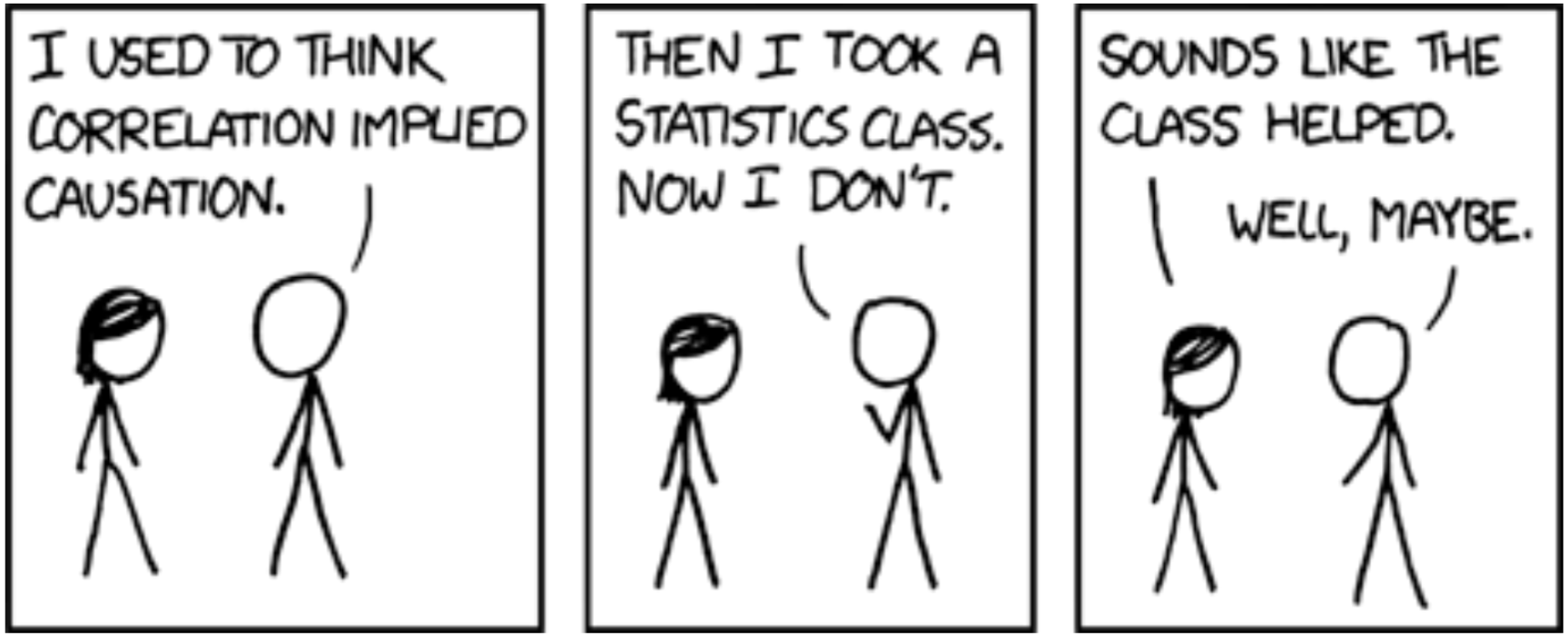
Spread (or dispersion) measures include:

- 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:

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

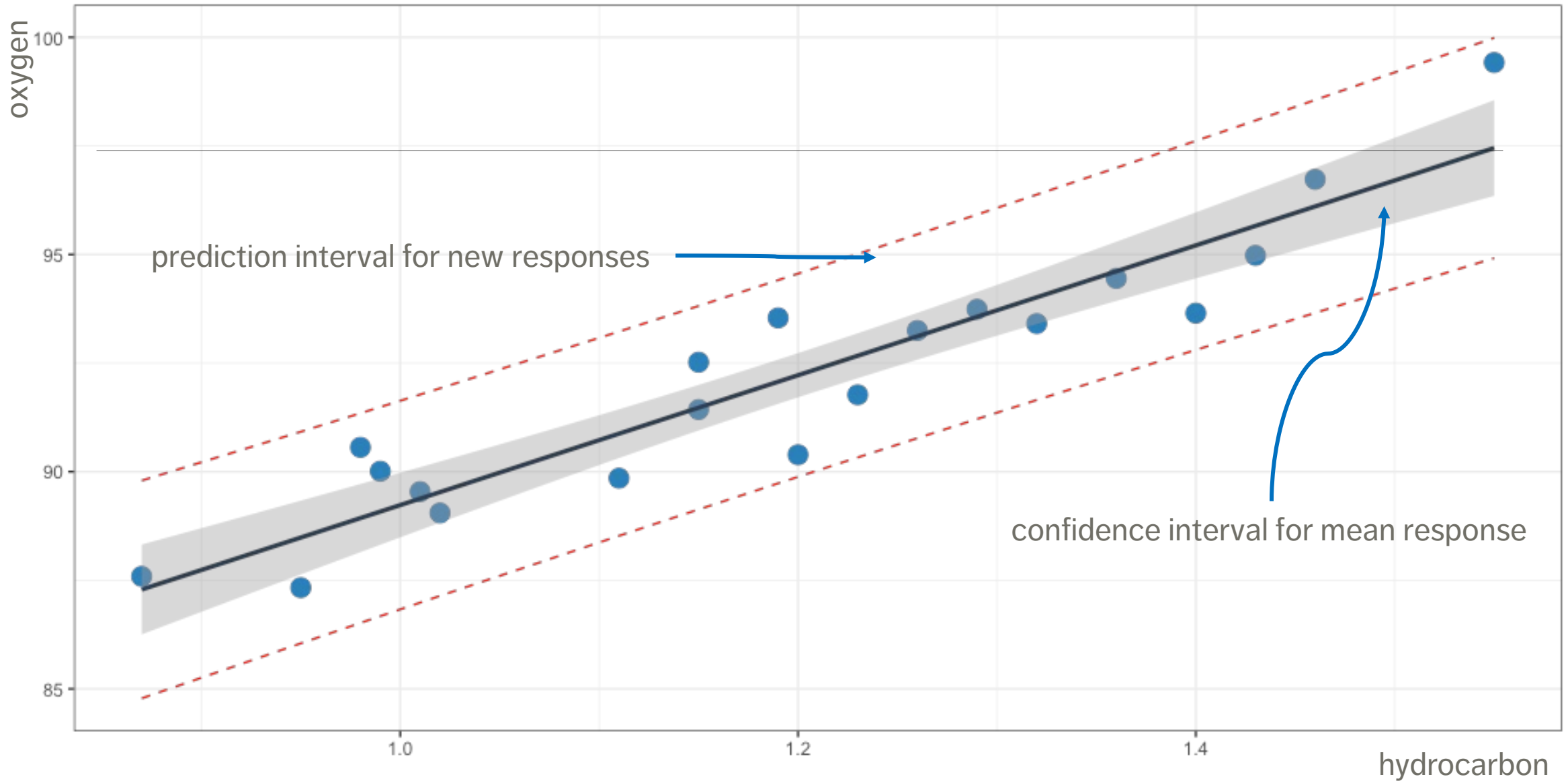
where $\boldsymbol{\beta} \in \mathbb{R}^p$ is to be determined based on the **training set**, and for which

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

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.

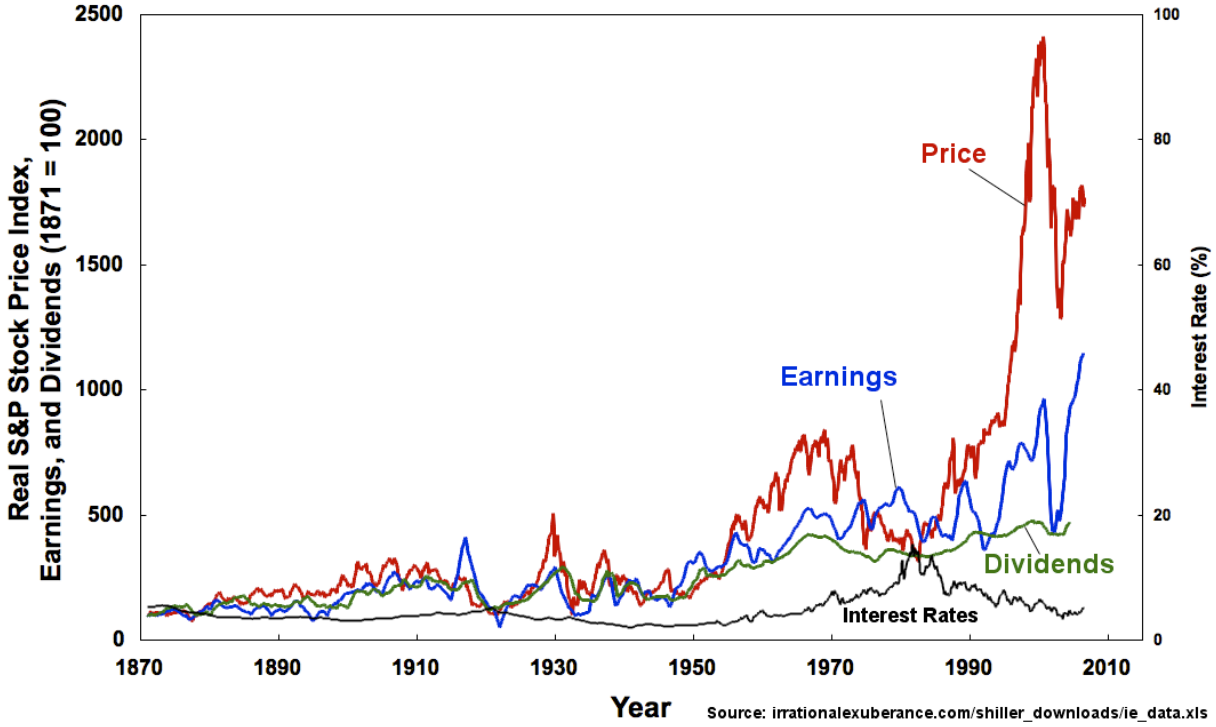
Time Series Analysis

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 ?



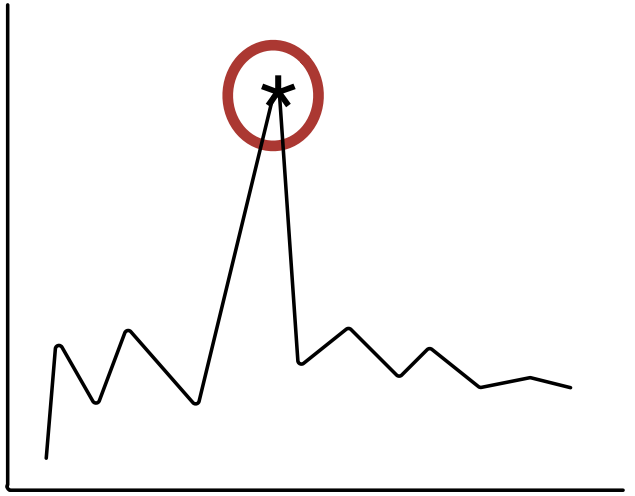
Anomaly Detection

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
Data Science Basics

Getting Insight From Data

- Asking the Right Questions
- Basic Data Analysis Techniques
- Common Statistical Procedures in R
- Quantitative Methods

***Probability and Applications** (advanced)

***Introductory Statistical Analysis** (advanced)

***Survey Sampling** (advanced)

***Regression Analysis** (coming soon)

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.

Session 3

DATA SCIENCE ESSENTIALS

Data Preparation

DATA SCIENCE ESSENTIALS

7. Data Quality and Data Wrangling

The Hot Mess

"Data is messy, you know."
"Even after it's been cleaned?"
"*Especially* after it's been cleaned."

Data **cleaning**, **processing**, **wrangling** are essential aspects of data science projects; analysts may spend **up to 80%** of their time on **data preparation**.

Data Wrangling and Tidy Data

Tidy data has a specific structure:

- each variable is in a single column
- each observation is in a single row
- each type of observational unit is in a single table

Country	2011	2012	2013
FR	7000	6900	7000
DE	5800	6000	6200
US	15000	14000	13000

VS.

Country	Year	n
FR	2011	7000
DE	2011	5800
US	2011	15000
FR	2012	6900
DE	2012	6000
US	2012	14000
FR	2013	7000
DE	2013	6200
US	2013	13000

Data Wrangling Functionality

Data wrangling functions should allow the analyst to:

- extract a subset of variables from the data frame
- extract a subset of observations from the data frame
- sort the data frame along any combination of variables in increasing or decreasing order
- to create new variables from existing variables
- to create (so-called) pivot tables, by observation groups
- database functionality (joins, etc.)
- etc.

Approaches to Data Cleaning

There are two **philosophical** approaches to data cleaning and validation:

- methodical
- narrative

The **methodical** approach consists of running through a **check list** of potential issues and flagging those that apply to the data.

The **narrative** approach consists of **exploring** the dataset and trying to spot unlikely and irregular patterns.

Approaches to Data Cleaning

Methodical (syntax)

- Pros: checklist is **context-independent**; pipelines **easy to implement**; common errors and invalid observations **easily identified**
- Cons: may prove **time-consuming**; cannot identify new types of errors

Narrative (semantics)

- Pros: process may simultaneously yield **data understanding**; false starts are (at most) as costly as switching to mechanical approach
- Cons: may miss important sources of errors and invalid observations for datasets with **high number of features**; domain knowledge may bias the process by neglecting uninteresting areas of the dataset

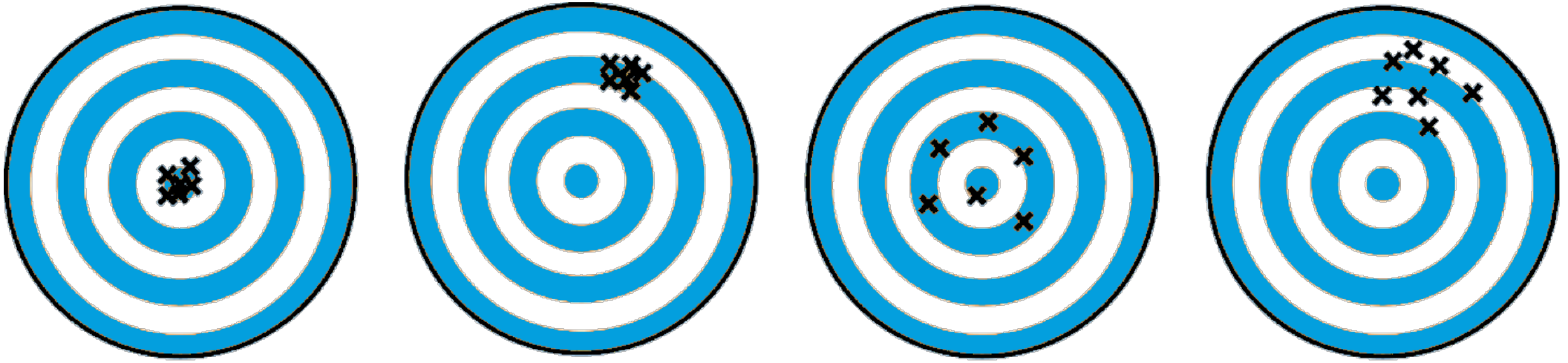
Data Soundness

The ideal dataset will have as few issues as possible with:

- **validity:** data type, range, mandatory response, uniqueness, value, regular expressions
- **completeness:** missing observations
- **accuracy and precision:** related to measurement and data entry errors; target diagrams (accuracy as bias, precision as standard error)
- **consistency:** conflicting observations
- **uniformity:** are units used uniformly throughout?

Checking for data quality issues at an early stage can save headaches later in the analysis.

Data Soundness



accurate and precise

precise but not accurate

accurate but not precise

neither accurate nor very precise

Common Error Sources

When dealing with **legacy**, **inherited** or **combined** datasets (that is, datasets over which there is no collection and initial processing control):

- missing data given a code
- 'NA'/'blank' given a code
- data entry error
- coding error
- measurement error
- duplicate entries
- heaping

Session 3

[Bill Watterson, *Calvin and Hobbes*, 15 Sep. 1990]

Detecting Invalid Entries

Potentially invalid entries can be detected with the help of:

- **Univariate Descriptive Statistics**
  count, range, z-score, mean, median, standard deviation, logic check
- **Multivariate Descriptive Statistics**
  n-way table, logic check
- **Data Visualization**
  scatterplot, scatterplot matrix, histogram, joint histogram, etc.

Detecting Invalid Entries

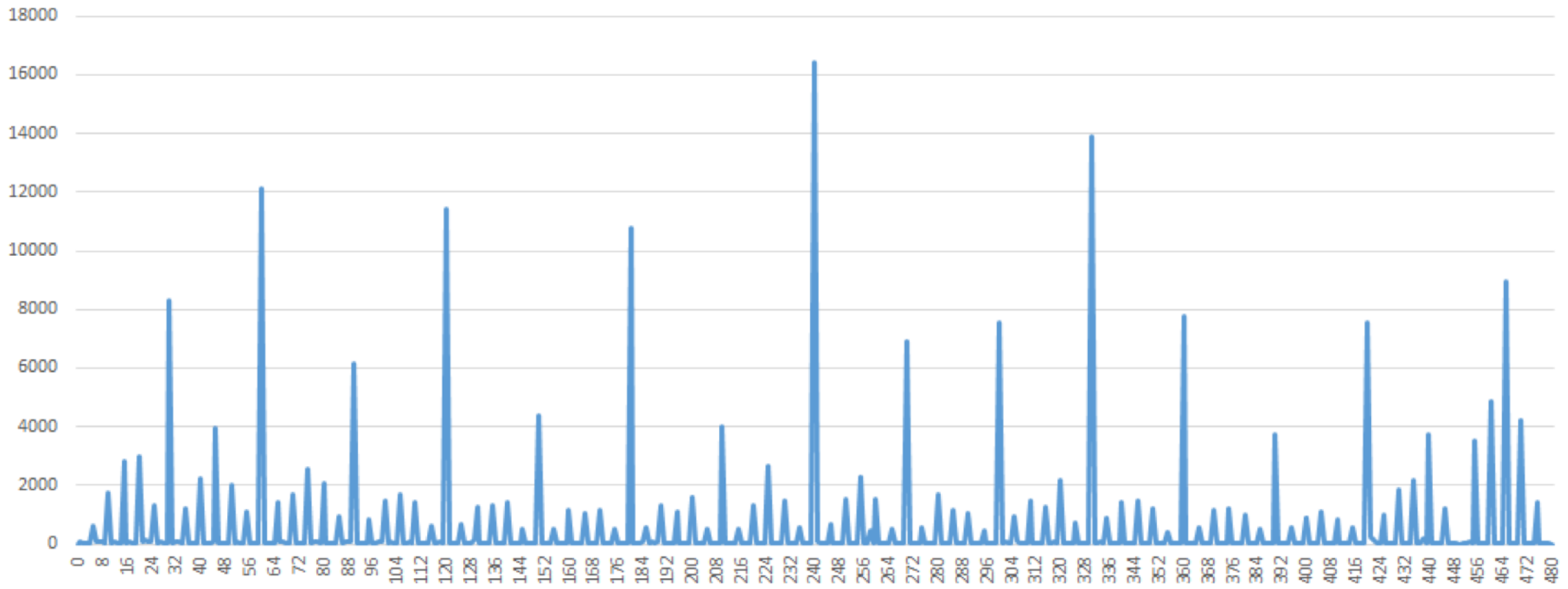
Univariate tests do not always tell the **whole** story.

This step might allow for the identification of potential outliers.

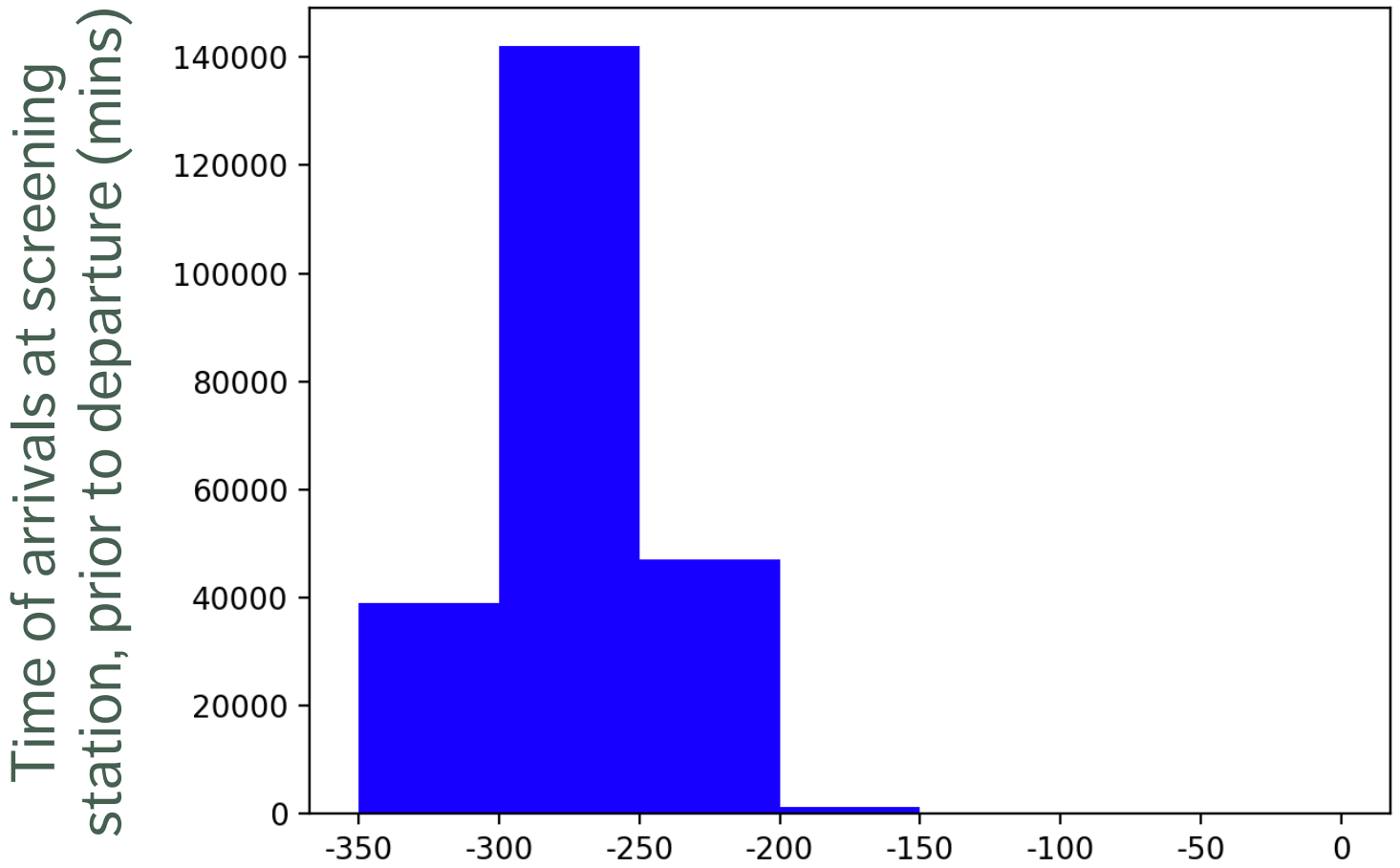
Failure to detect invalid entries ≠ all entries are valid.

Small numbers of invalid entries recoded as “missing.”

Detecting Invalid Entries



Detecting Invalid Entries



Detecting Invalid Entries

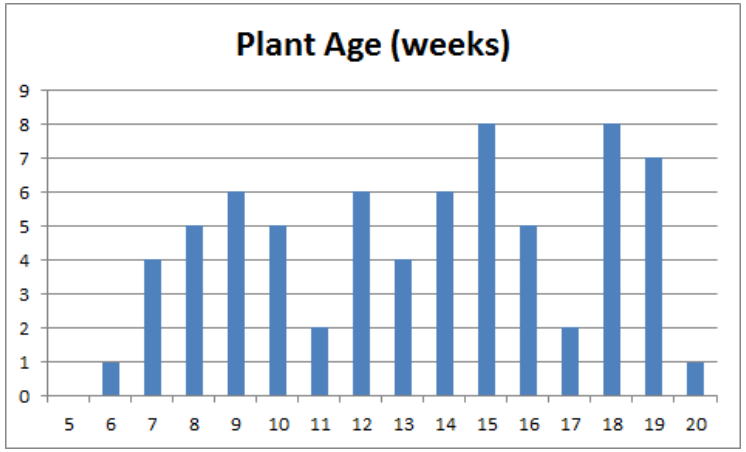
Sex	
Male	19
Female	17
(blank)	2
Total	**38**

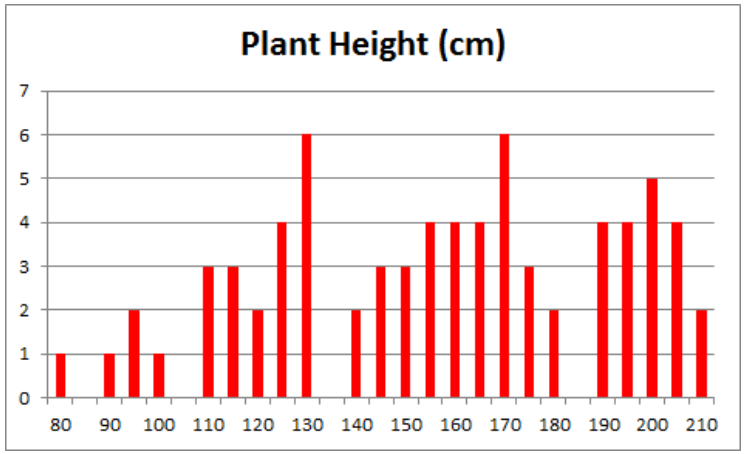
Pregnant	
Yes	7
No	27
99	1
(blank)	3
Total	**38**

vs.

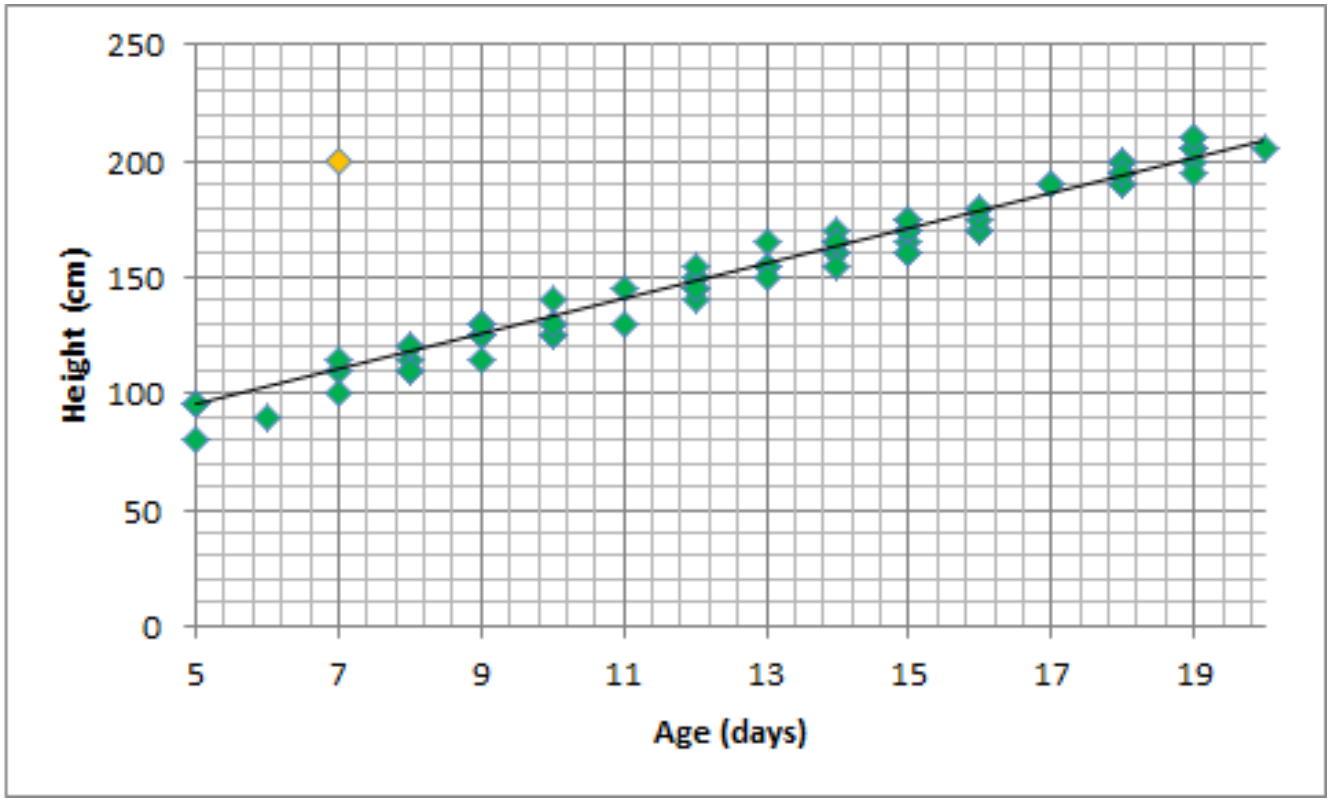
	Pregnant				Total
	Yes	No	99	(blank)	
Male	1	17	1	0	**19**
Female	6	9	0	2	**17**
(blank)	0	1	0	1	**2**
Total	**7**	**27**	**1**	**3**	**38**

Detecting Invalid Entries



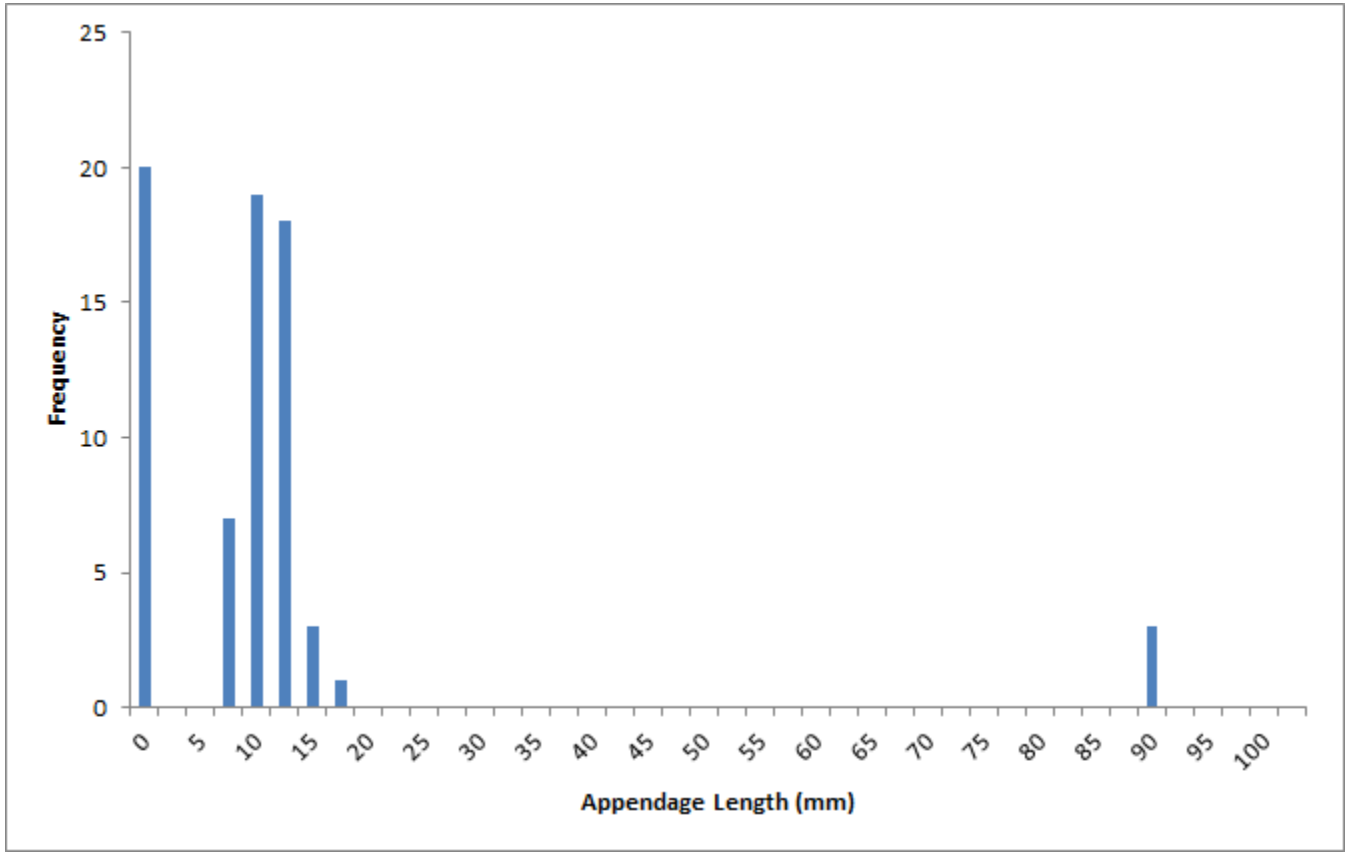


vs.



Detecting Invalid Entries

Appendage length (mm)	
Mean	10.35
Standard Deviation	16.98
Kurtosis	16.78
Skewness	4.07
Minimum	0
First Quartile	0
Median	8.77
Third Quartile	10.58
Maximum	88
Range	88
Interquartile Range	10.58
Mode	0
Count	71



Suggested Reading

Data Quality

Data Understanding, Data Analysis, Data Science
Data Preparation

Introduction

General Principles

- Approaches to Data Cleaning
- Pros and Cons
- Tools and Methods

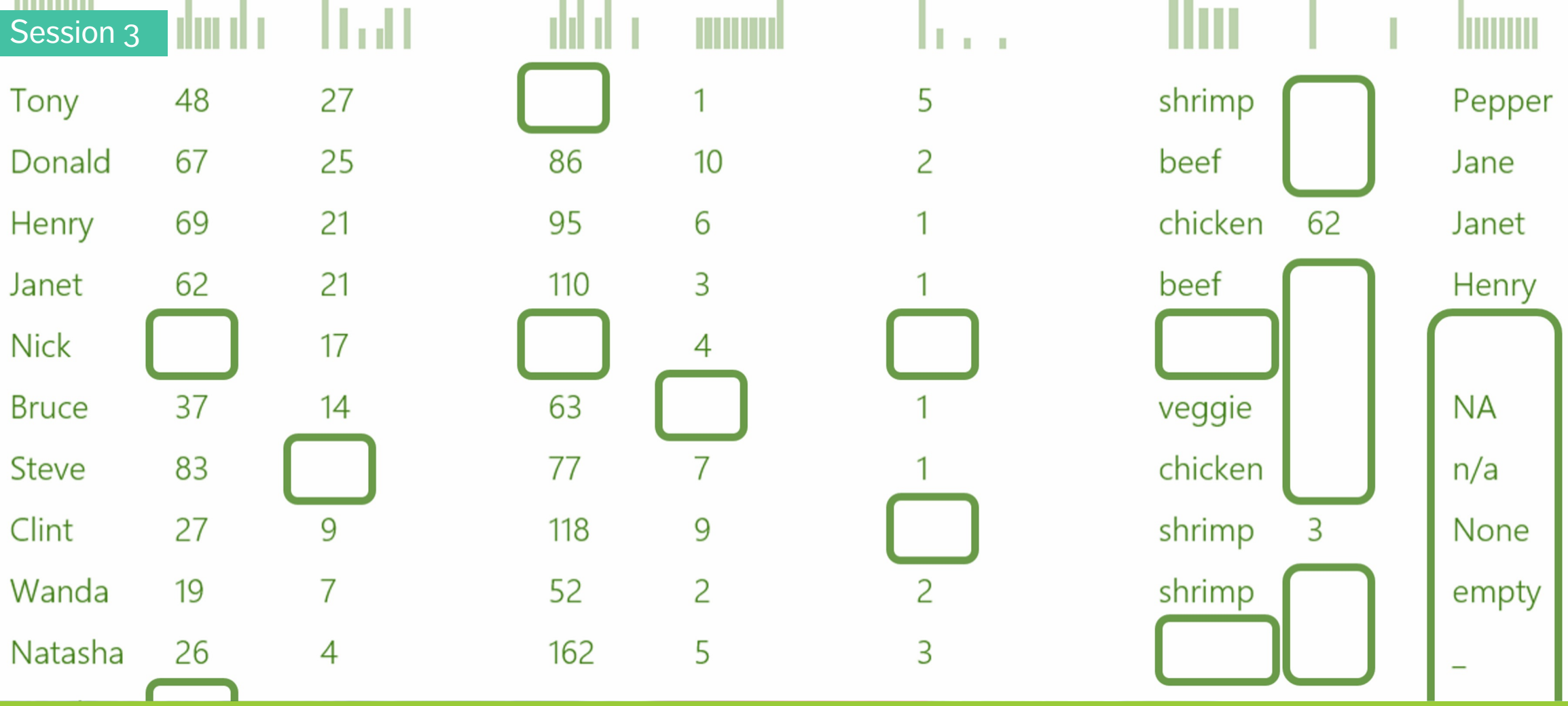
Data Quality

- Common Error Sources
- Detecting Invalid Entries

Exercises

Data Quality

1. Recreate the examples of The Tidyverse.
2. Turn the dataset found in the file cities.txt into a tidy dataset.
3. Does the dataset found in the file cities.txt appear to be of good quality (is it sound? does it have invalid entries?)
4. Create a list of items that could be used in a methodical data cleaning checklist. Use data that you have encountered in the past as inspiration (numerical, categorical, text data).



Tony	48	27		1	5	shrimp	Pepper
Donald	67	25	86	10	2	beef	Jane
Henry	69	21	95	6	1	chicken	62	Janet
Janet	62	21	110	3	1	beef	Henry
Nick		17		4			
Bruce	37	14	63		1	veggie	NA
Steve	83		77	7	1	chicken	n/a
Clint	27	9	118	9		shrimp	3	None
Wanda	19	7	52	2	2	shrimp	empty
Natasha	26	4	162	5	3		_

8. Missing Values

Types of Missing Observations

Blank fields come in 4 flavours:

- **nonresponse**
  an observation was expected but none had been entered
- **data entry issue**
  an observation was recorded but was not entered in the dataset
- **invalid entry**
  an observation was recorded but was considered invalid and has been removed
- **expected blank**
  a field has been left blank, but expectedly so

Types of Missing Observations

Too many missing values (of the first three type) can be indicative of **issues with the data collection process** (more on this later).

Too many missing values (of the fourth type) can be indicative of **poor questionnaire design**.

Finding missing values can help you deal with other data science problems.

The Case for Imputation

Not all analytical methods can easily accommodate missing observations:

- **discard** the missing observation
  - not recommended, unless the data is MCAR in the dataset as a whole
  - acceptable in certain situations (e.g., small number of missing values in a large dataset)
- come up with a **replacement (imputation) value**
  - main drawback: we never know what the true value would have been
  - often the best available option

Missing Values Mechanism

Missing Completely at Random (MCAR)

- item absence is independent of its value or of auxiliary variables
- **example:** an electrical surge randomly deletes an observation in the dataset

Missing at Random (MAR)

- item absence is not completely random; can be accounted by auxiliary variables with complete info
- **example:** if women are less likely to tell you their age than men for societal reasons, but not because of the age values themselves)

Missing Values Mechanism

Not Missing at Random (NMAR)

- reason for nonresponse is related to item value (also called **non-ignorable non-response**)
- **example:** if illicit drug users are less likely to admit to drug use than teetotallers

In general, the missing mechanism **cannot be determined** with any certainty; we may need to make assumptions (domain expertise can help).

Imputation Methods

- list-wise deletion
- mean or most frequent imputation
- regression or correlation imputation
- stochastic regression imputation
- last observation carried forward
- next observation carried backward
- k-nearest neighbours imputation
- multiple imputation
- etc.

Imputation Methods

List-wise deletion: remove units with at least one missing values

- **assumption:** MCAR
- **cons:** can introduce bias (if not MCAR), reduction in sample size, increase in standard error

Mean/most frequent imputation: substitute missing values by average/most frequent value

- **assumption:** MCAR
- **cons:** distortions of distribution (spike at mean) and relationships among variables

Imputation Methods

Regression/correlation imputation: substitute missing values using fitted values based on other variables with complete information

- **assumption:** MAR
- **cons:** artificial reduction in variability, over-estimation of correlation

Stochastic regression imputation: regression/correlation imputation with a random error term added

- **assumption:** MAR
- **cons:** increased risk of type I error (false positives) due to small std error

Imputation Methods

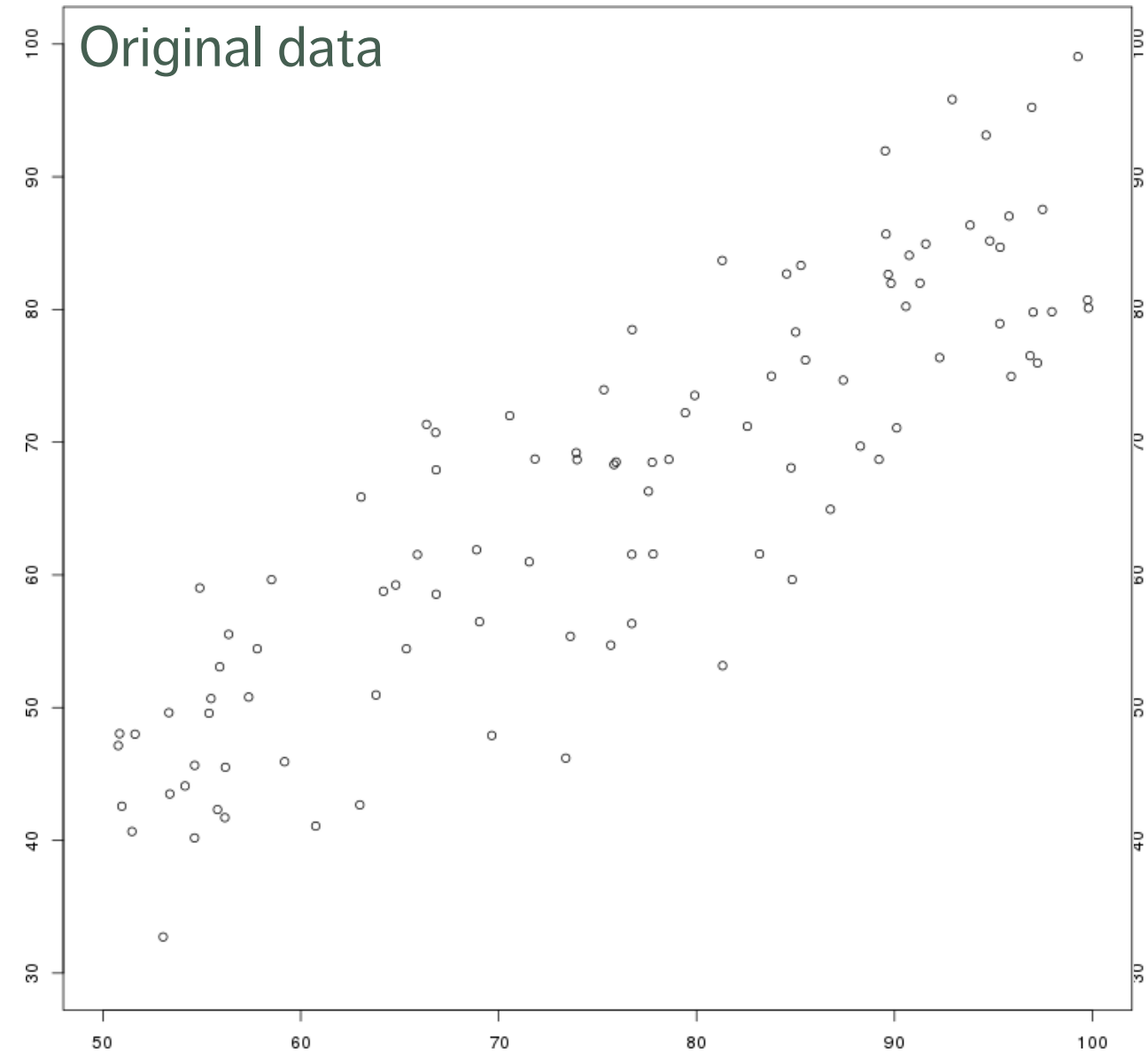
Last observation carried forward: substitute the missing values with latest previous values (in a longitudinal study)

- **assumption:** MCAR, values do not vary greatly over time
- **cons:** may be too "generous", depending on the nature of study

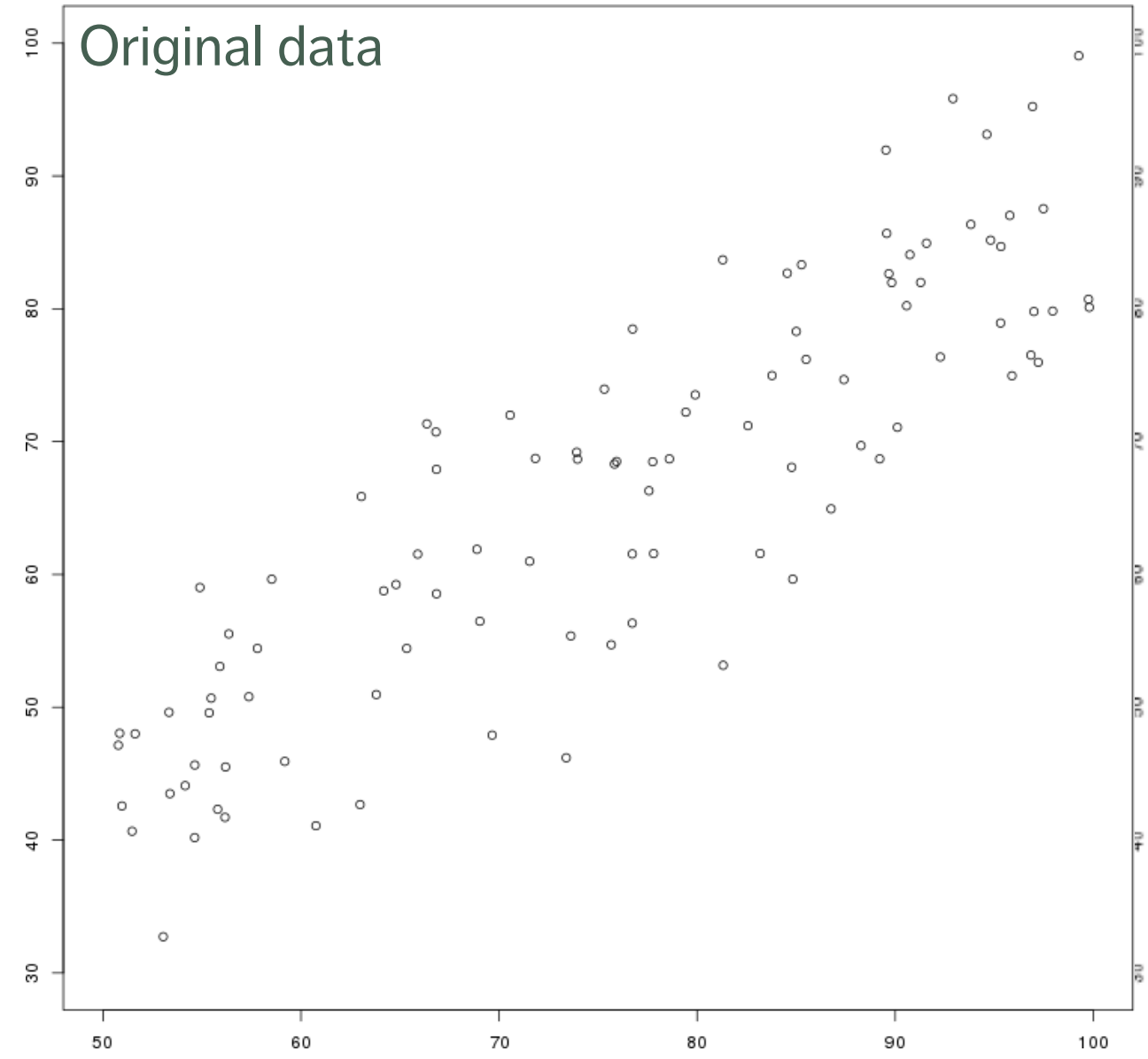
k nearest neighbour imputation (kNN): substitute the missing entry with the average from the group of the k most similar complete cases

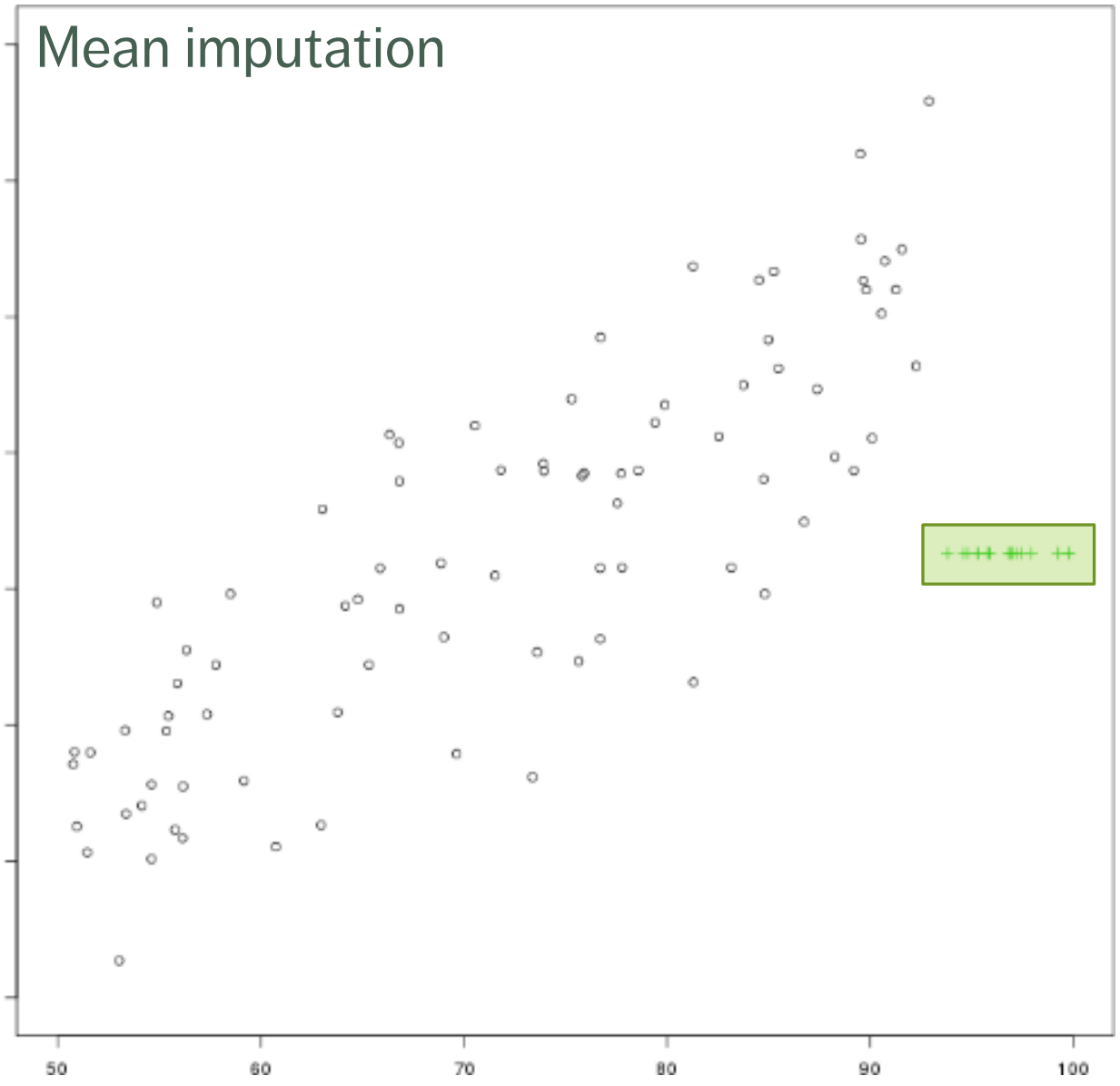
- **assumption:** MAR
- **cons:** difficult to choose appropriate value for k; possible distortion in data structure

Artificial data: the y values of all points for which $x > 92$ have been erased by mistake.

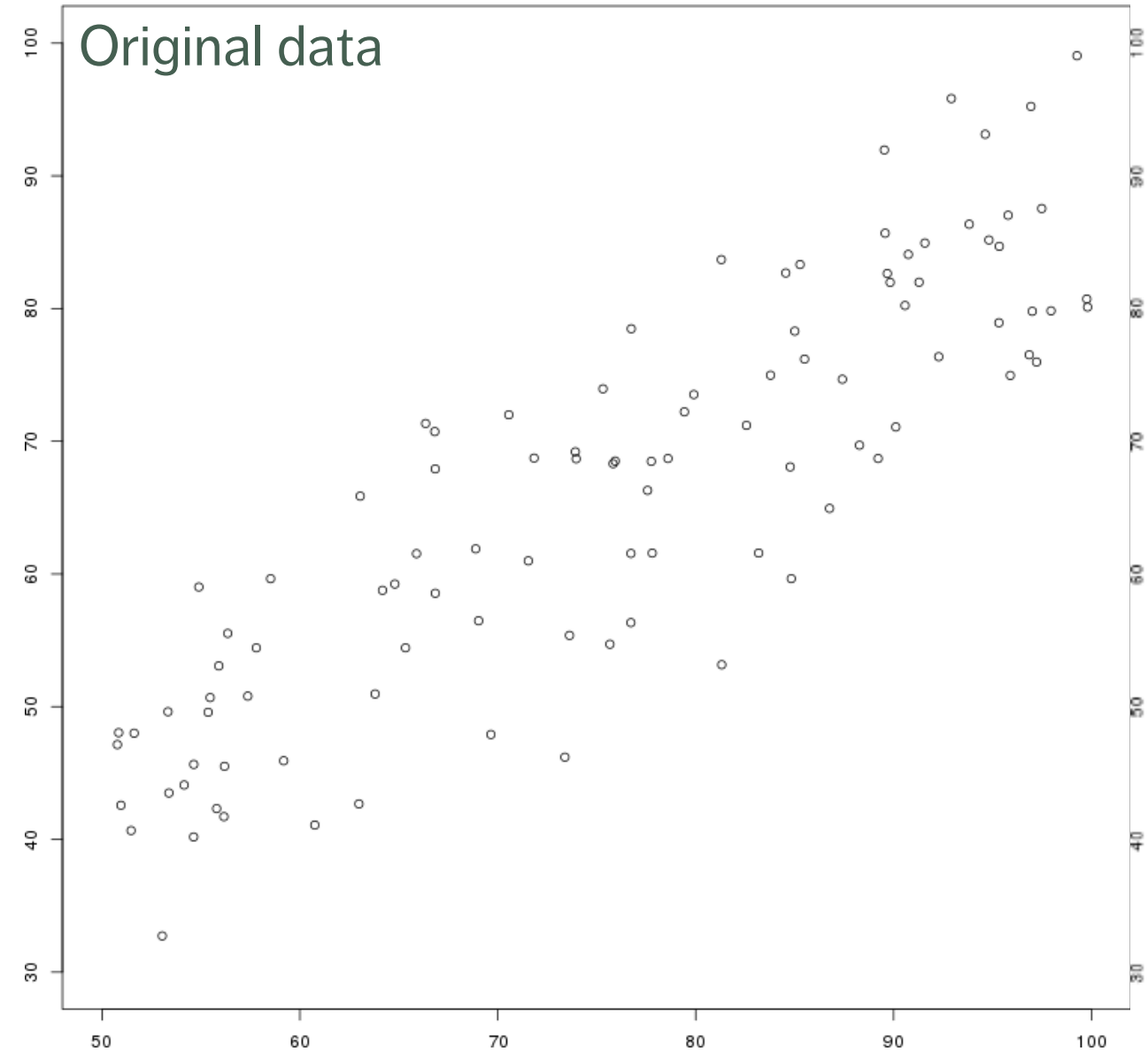


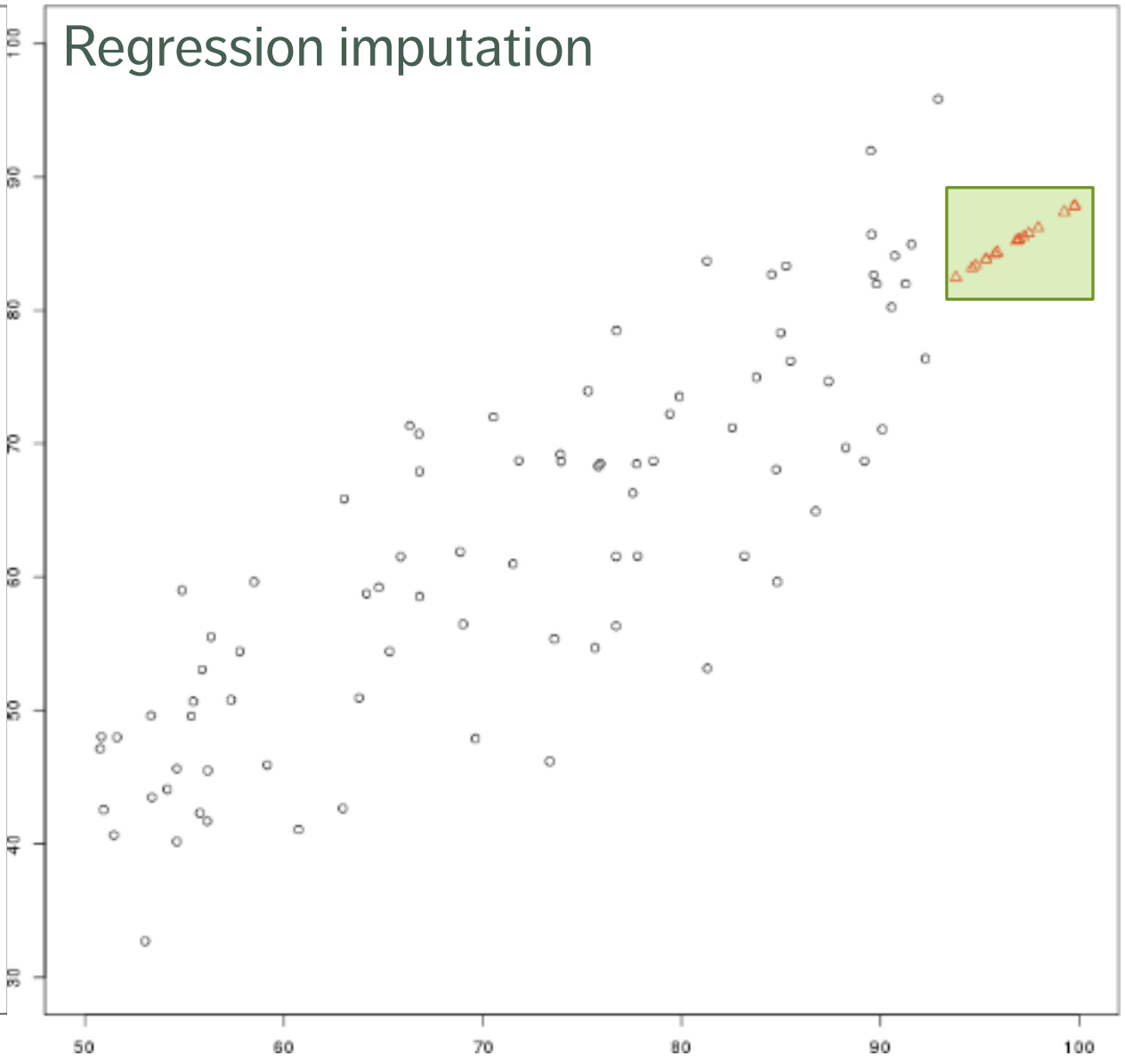
Artificial data: the y values of all points for which $x > 92$ have been erased by mistake.



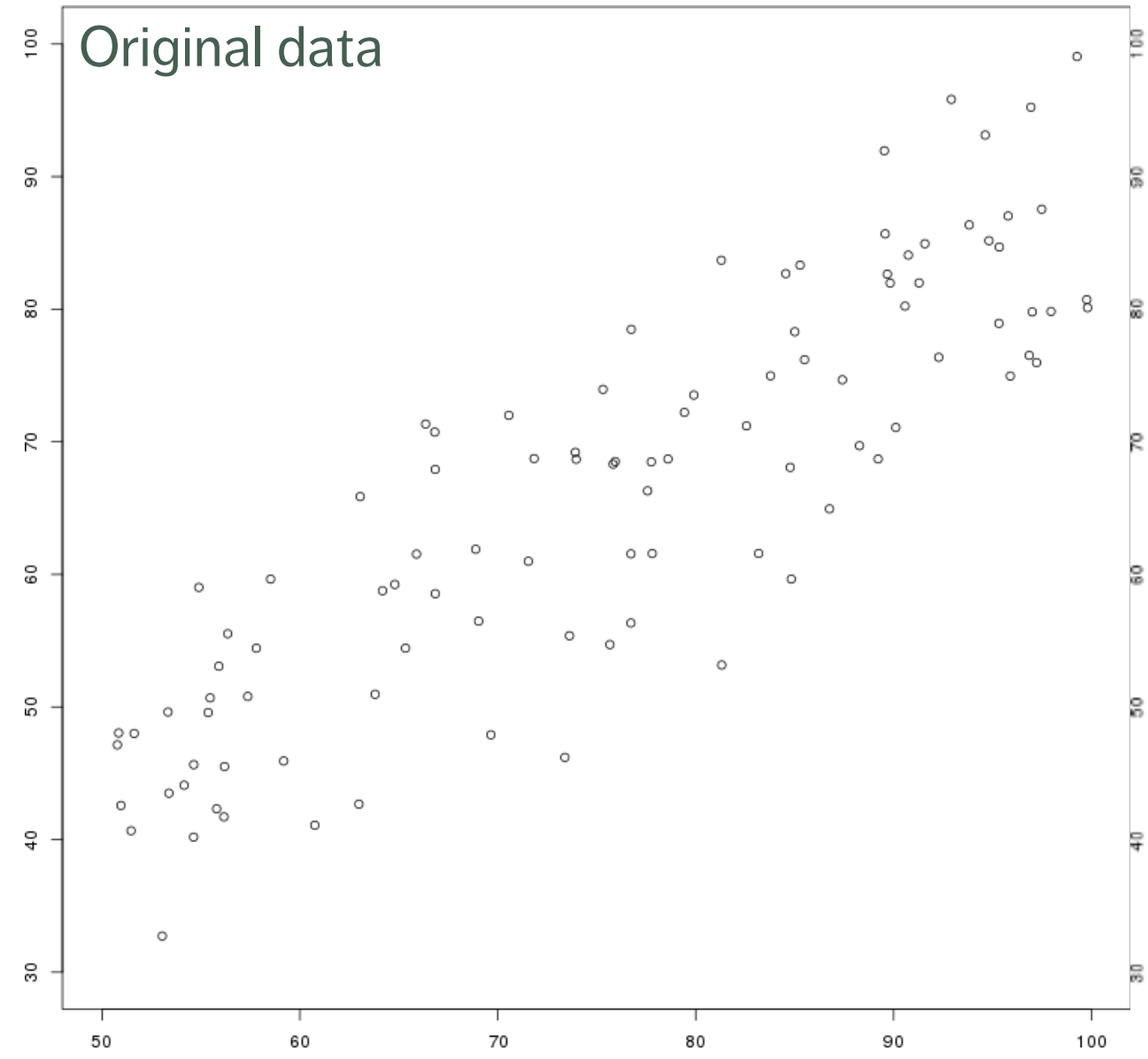


Artificial data: the y values of all points for which $x > 92$ have been erased by mistake.





Artificial data: the y values of all points for which $x > 92$ have been erased by mistake.



Multiple Imputation

Imputations increase the noise in the data.

In **multiple imputation**, the effect of that noise can be measured by consolidating the analysis outcome from multiple imputed datasets

Steps:

1. repeated imputation creates m versions of the dataset
2. each of these datasets is analyzed, yielding m outcomes
3. the m outcomes are pooled into a single result for which the mean, variance, and confidence intervals are known

Multiple Imputation

Advantages

- **flexible**; can be used in a various situations (MCAR, MAR, even NMAR in certain cases)
- accounts for **uncertainty** in imputed values
- fairly easy to implement

Disadvantages

- m may need to be fairly **large** when there are many missing values in numerous features, which slows down the analyses
- if the analysis output is not a single value but some complicated mathematical object, this approach is unlikely to be useful

Take-Aways

Missing values **cannot simply be ignored**.

The missing mechanism **cannot typically be determined** with any certainty.

Imputation methods work best when values are **MCAR** or **MAR**, but imputation methods tend to produce biased estimates.

In single imputation, imputed data is treated as the actual data; multiple imputation can help reduce the noise.

Is stochastic imputation best? In our example, yes – but … **No-Free Lunch theorem**!

Suggested Reading

Missing Values

Data Understanding, Data Analysis, Data Science
Data Preparation

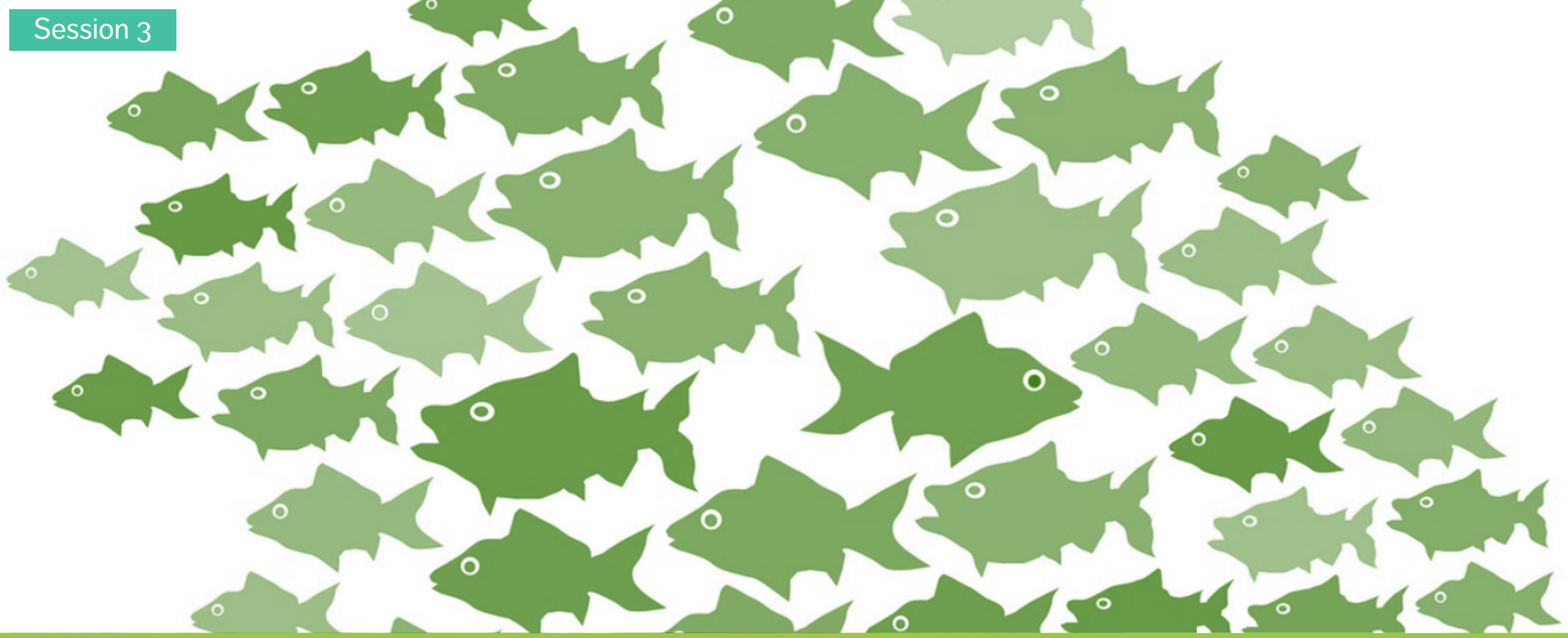
Missing Values

- Missing Value Mechanisms
- Imputation Methods
- Multiple Imputation

Exercises

Missing Values

1. Recreate the examples of Imputation Methods.
2. Recreate the missing value imputation process (data cleaning) used in Example: Algae Bloom.
3. Conduct kNN imputation on the grades dataset with various values of k.
4. Conduct multiple imputation on the grades dataset using stochastic regression in order to estimate the slope and intercept for the line of best fit.



9. Anomalous Observations

Anomalous Observations

In practice, an **anomalous observation** may arise as

- a **"bad" object/measurement:** data artifacts, spelling mistakes, poorly imputed values, etc.
- a **misclassified observation:** according to the existing data patterns, the observation should have been labeled differently;
- an observation whose measurements are found in the **distribution tails** of a large enough number of features;
- an **unknown unknown**: a completely new type of observations whose existence was heretofore unsuspected.

Anomalous Observations

Observations could be anomalous in one context, but not in another:

- A 6-foot tall adult male is in the 86th percentile for Canadian males (tall, but not unusual)
- in Bolivia, the same man would be in the 99.9th percentile (very tall and unusual)

Anomaly detection points towards **interesting questions** for analysts and subject matter experts: in this case, why is there such a large discrepancy in the two populations?

Outliers

Outlying observations are data points which are **atypical** in comparison to

- the unit's remaining features (*within-unit*),
- the field measurements for other units (*between-units*)

Outliers are observations which are **dissimilar to other cases** or which **contradict known dependencies** or rules.

Careful study is needed to determine whether outliers should be retained or removed from the dataset.

Detecting Anomalies

Outliers may be anomalous along any of the unit's variables, or in combination.

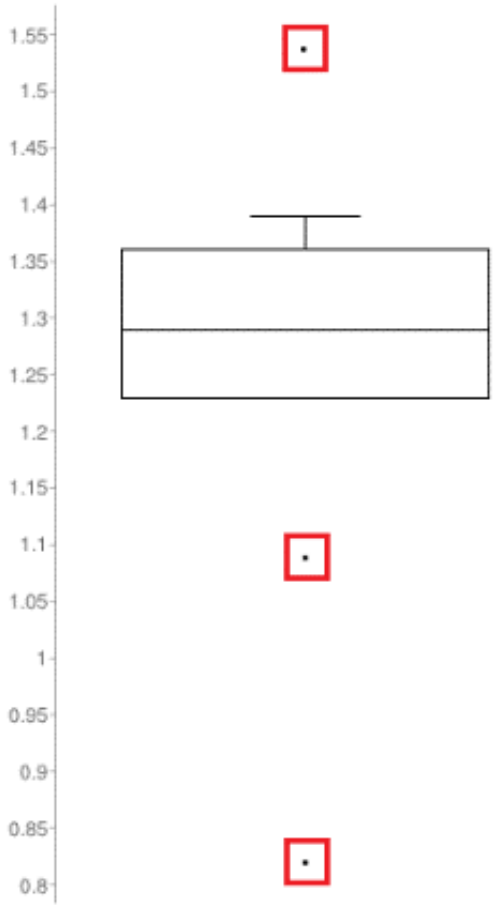
Anomalies are by definition **infrequent**, and typically shrouded in **uncertainty** due to small sample sizes.

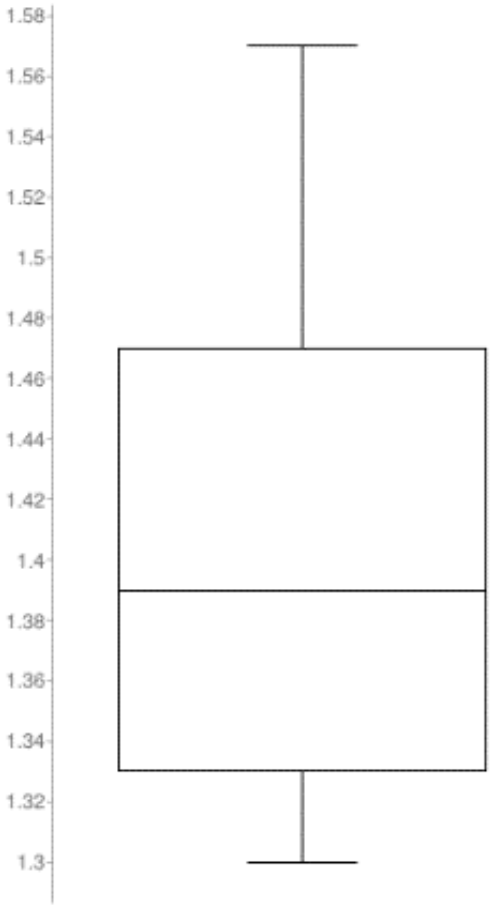
Differentiating anomalies from noise or data entry errors is **hard**.

Boundaries between normal and deviating units may be **fuzzy**.

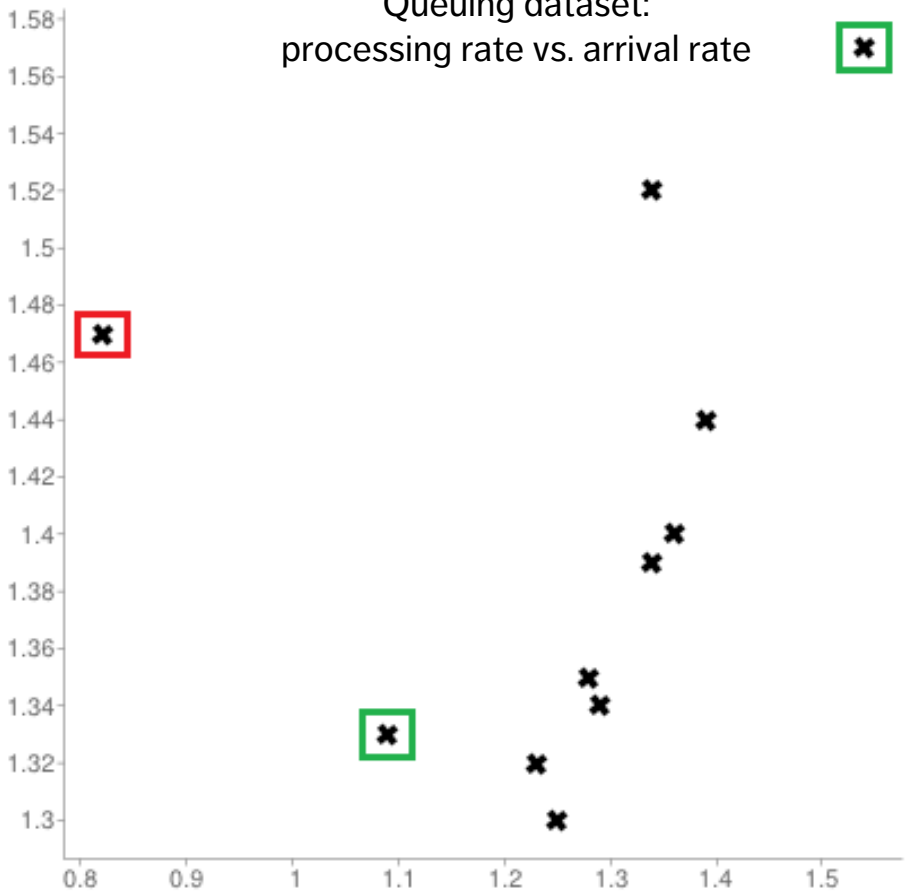
Anomalies associated with malicious activities are typically **disguised**.

Visual Outlier Detection





Queuing dataset:
processing rate vs. arrival rate



Detecting Anomalies

Numerous methods exist to identify anomalous observations; **none of them are foolproof** and judgement must be used.

Graphical methods are easy to implement and interpret:

- **Outlying Observations**

  box-plots, scatterplots, scatterplot matrices, 2D tour, Cooke's distance, normal qq plots

- **Influential Data**

  some level of analysis must be performed (leverage)

Careful: once anomalous observations have been removed from the dataset, previously “regular" units may become anomalous.

Anomaly Detection Algorithms

Supervised methods use a historical record of labeled anomalous observations:

- domain expertise is required to tag the data
- classification or regression task
- rare occurrence problem

		Predicted Class: Normal	Predicted Class: Anomaly
Actual Class	Normal	TN	FP
	Anomaly	FN	TP

Unsupervised methods don't use external information:

- traditional methods and tests
- can also be seen as a clustering or association rules problem

Anomaly Detection Algorithms

The mis-classification cost is often assumed to be symmetrical, which can lead to **technically correct but useless** outputs.

For instance, most (99.999+%) air passengers do not bring weapons with them on flights; a model that predicts that no passenger is smuggling a weapon would be 99.999+% accurate, but it would miss the point completely.

For the **security agency**, the cost of wrongly thinking that a passenger is:

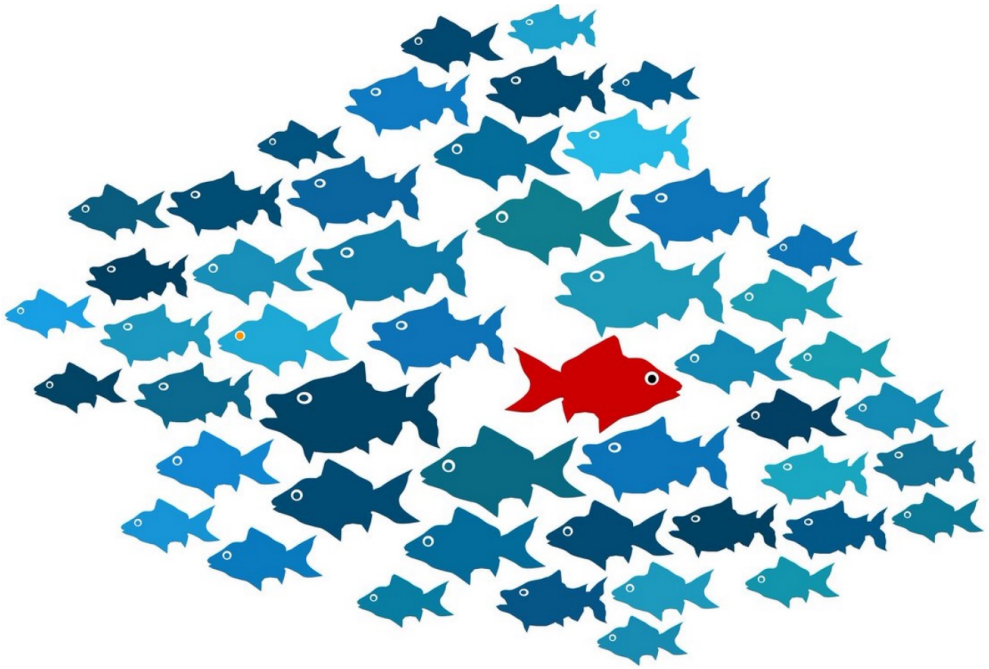
- smuggling a weapon $\Longrightarrow$ cost of a single search
- NOT smuggling a weapon $\Longrightarrow$ catastrophe (potentially)

The wrongly targeted individuals may have a different take on this!

Anomaly Detection Algorithms

If all participants in a workshop except for one can view the video conference lectures, then the one individual/internet connection/computer is **anomalous** – it behaves in a manner which is different from the others.

But this **DOES NOT MEAN** that the different behaviour is necessarily the one we are interested in...



Simple Outlier Tests

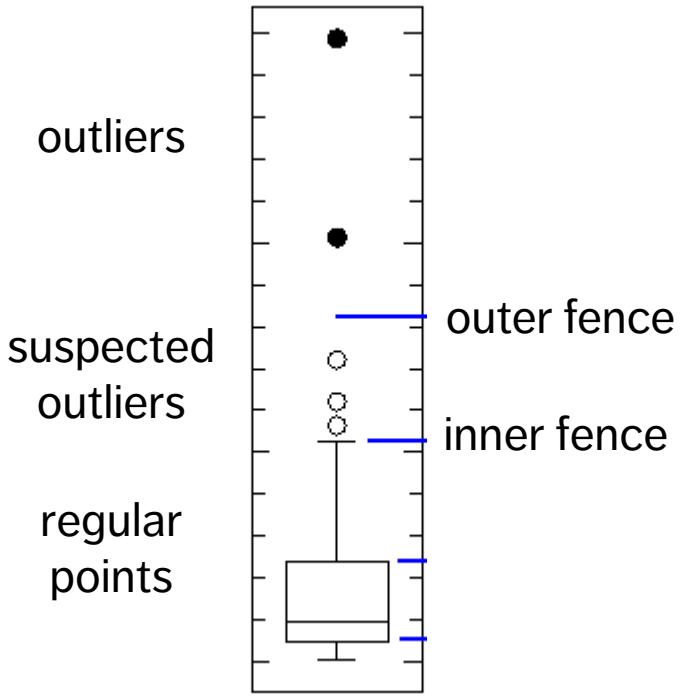
Tukey's Boxplot test: for normally distributed data, regular observations typically lie between the **inner fences**

$$Q_1 - 1.5\times(Q_3 - Q_1) \text{ and } Q_3 + 1.5\times(Q_3 - Q_1).$$

Suspected outliers lie between the **inner fences** and the **outer fences**

$$Q_1 - 3\times(Q_3 - Q_1) \text{ and } Q_3 + 3\times(Q_3 - Q_1).$$

Outliers lie beyond the **outer fences**.



Simple Outlier Tests

The **Dixon Q Test** is used in experimental sciences to find outliers in (extremely) small datasets (dubious validity).

The **Mahalanobis Distance** (linked to the leverage) can be used to find multi-dimensional outliers (when relationships are linear).

Other simple tests:

- **Grubbs** (univariate)
- **Tietjen-Moore** (for a specific # of outliers)
- **generalized extreme studentized deviate** (for unknown # of outliers)
- **chi-square** (outliers affecting goodness-of-fit)

Sophisticated Anomaly Detection

- **DBSCAN**, **OR$_h$**, and **LOF** (unsupervised outlier detection)
- **rank-power** method (supervised outlier detection)
- **distance** or **density-based** methods (with exotic distance measures)
- **autoencoders and reconstruction error** (deep learning method)
- **rare-occurrence** methods (oversampling, undersampling, CREDOS, PN, SHRINK, SMOTE, DRAMOTE, SMOTEBoost, RareBoost, MetaCost, AdaCost, CSB, SSTBoost, etc.)
- **AVF**, **Greedy** algorithms (categorical data)
- **PCA**, **DOBIN**, and other **projection** methods (for high-dimensional data)
- **subspace** methods and **ensemble** methods

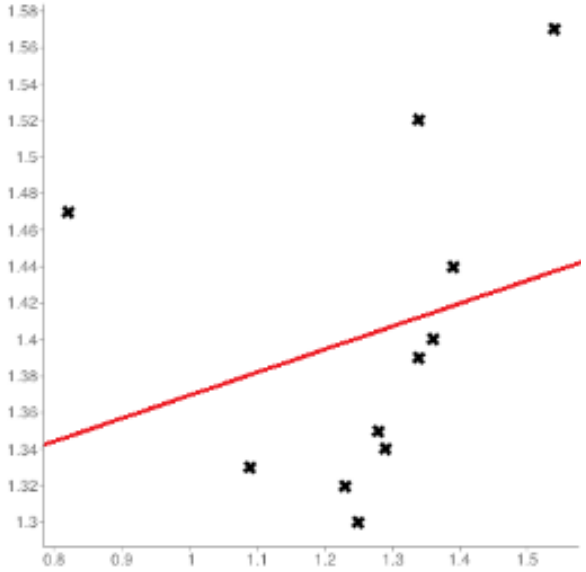
Influential Observations

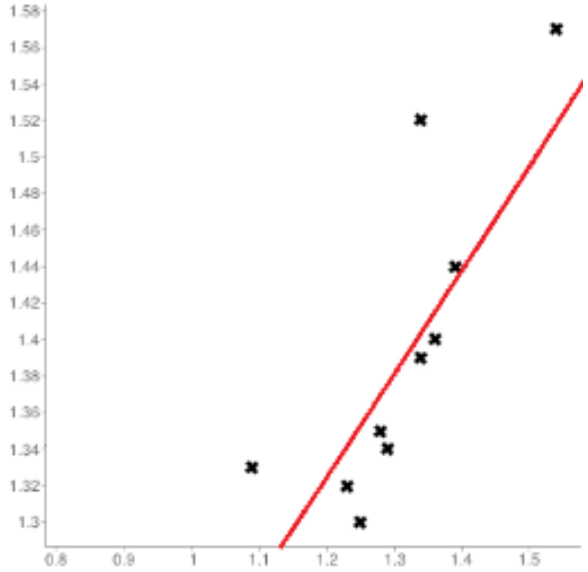
Influential data points are observations whose absence leads to **markedly different** analysis results.

When influential observations are identified, **remedial measures** (such as data transformations) may be required to minimize their undue effects.

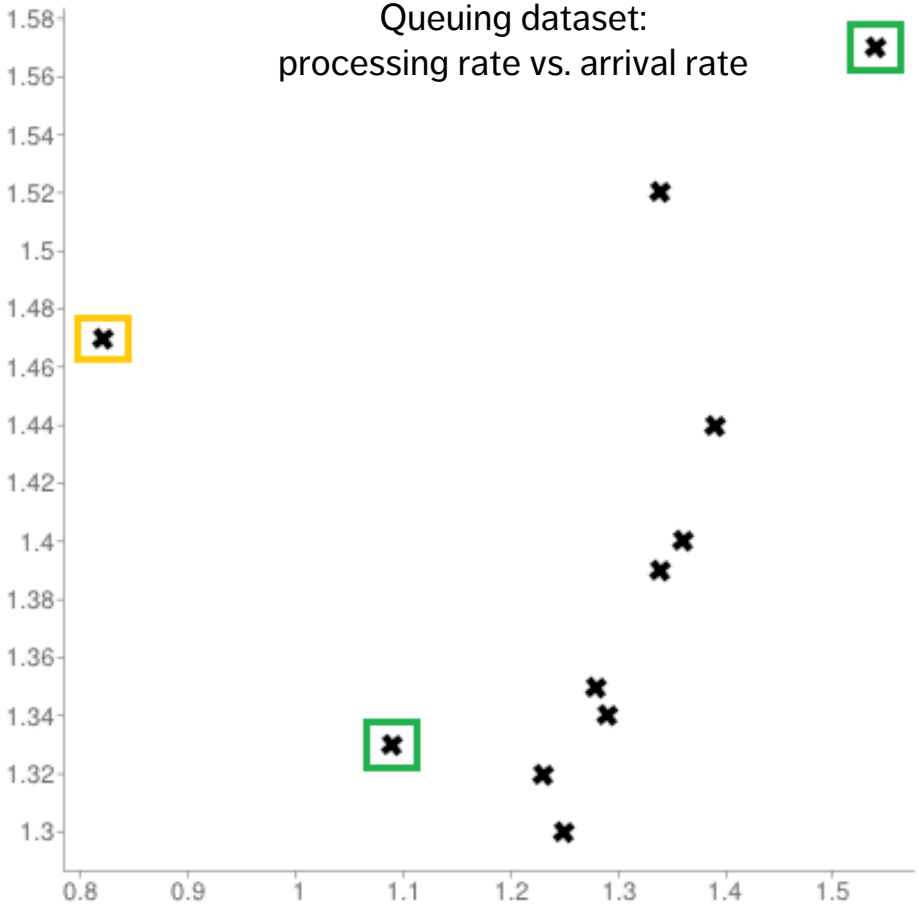
Outliers may be influential data points; influential data points need not be outliers (and *vice-versa*).

Influential Observations





Queuing dataset:
processing rate vs. arrival rate



Anomaly Detection Remarks

Identifying influential points is an **iterative process** as the various analyses have to be run numerous times.

Fully automated identification and removal of anomalous observations is **NOT recommended**.

Use data transformations if the data is **NOT normally distributed**.

Whether an observation is an outlier or not depends on **various factors**; what observations end up being influential data points depends on the **specific analysis to be performed**.

Suggested Reading

Anomalous Observations

Data Understanding, Data Analysis, Data Science
Data Preparation

Anomalous Observations

- Anomaly Detection
- Outlier Tests
- Visual Outlier Detection

***Anomaly Detection and Outlier Analysis** (advanced)

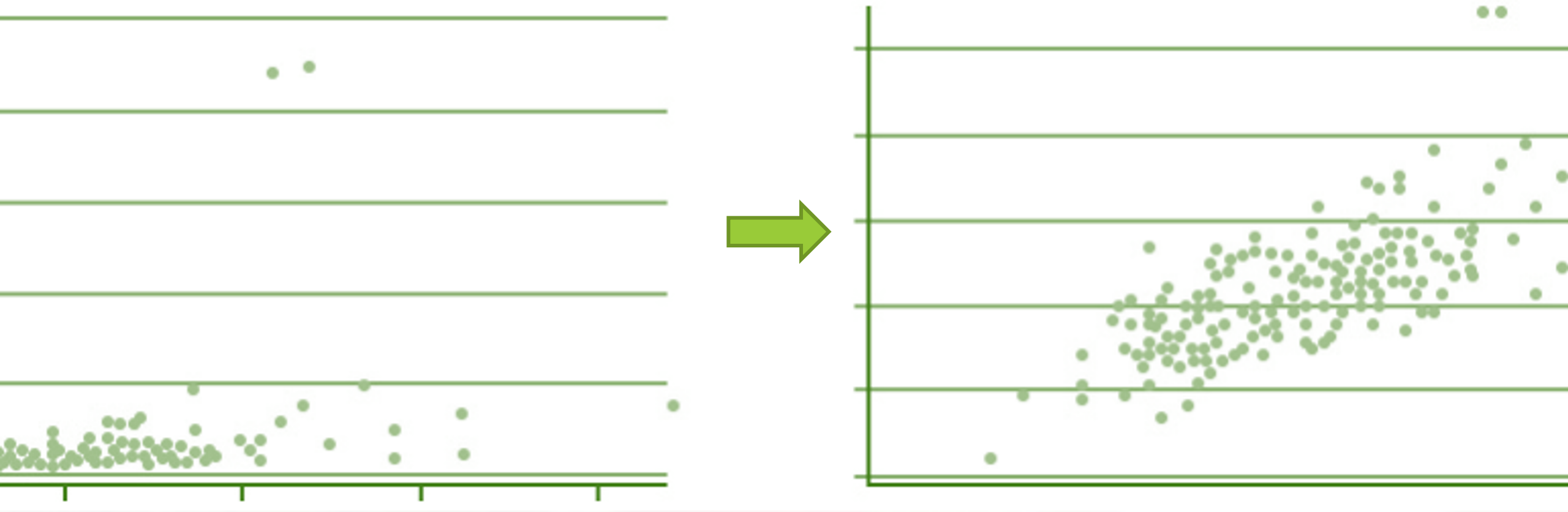
Exercises

Anomalous Observations

1. Recreate the anomaly detection process used in Example: Algae Bloom.
2. Find anomalous observations in the cities.txt and grades datasets (if applicable).
3. Find anomalous observations in a dataset of your choice.

Session 4

DATA SCIENCE ESSENTIALS



10. Dimensionality and Data Transformations

Dimensionality of Data

In data analysis, the **dimension** of the data is the number of attributes that are collected in a dataset, represented by the **number of columns**.

We can think of the number of variables used to describe each object (row) as a vector describing that object: the dimension is simply the **size** of that vector.

(**Note:** "dimension" is used differently in business intelligence contexts)

High Dimensionality and Big Data

Datasets can be "big" in a variety of ways:

- too large for the **hardware** to handle (cannot be stored, accessed, manipulated properly due to # of observations, # of features, the overall size)
- dimensions can go against **modeling assumptions** (# of features ≫ # observations)

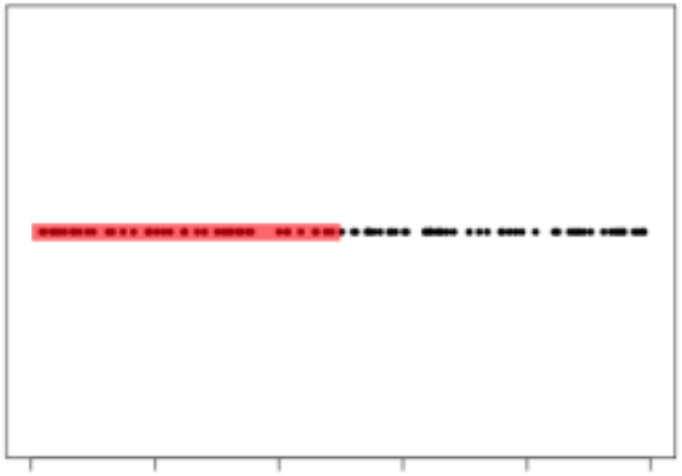
Examples:

- Multiple sensors recording 100+ observations per second in a large geographical area over a long time period = **very big dataset**
- In a corpus' *Term Document Matrix* (cols = terms, rows = documents), the number of terms is usually substantially higher than the number of documents, leading to **sparse data**

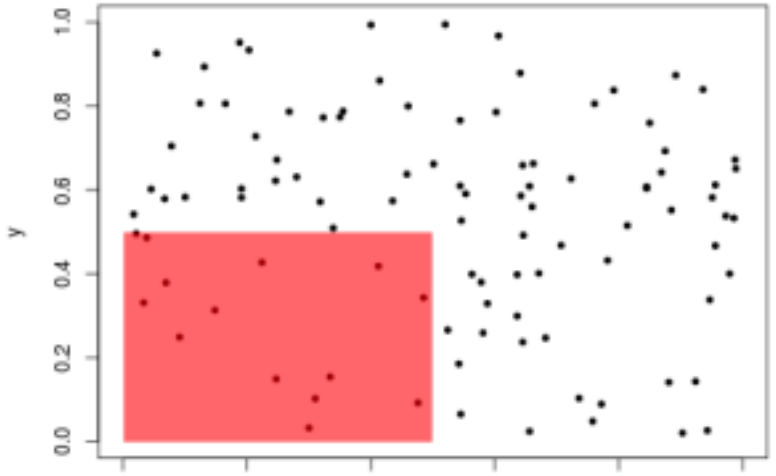
[http://simplystatistics.org]

Curse of Dimensionality

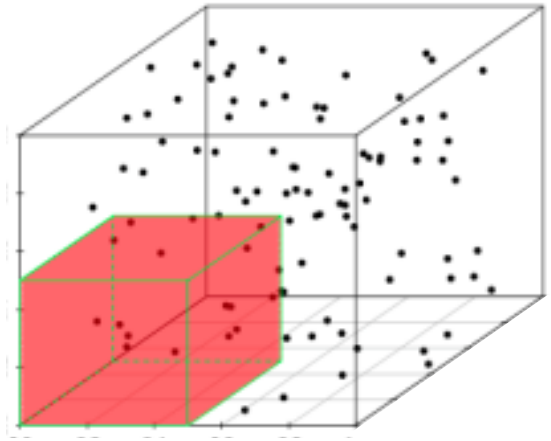
42% of data is captured



14% of data is captured



7% of data is captured



$N = 100$ observations, uniformly distributed on $[0,1]^d, d = 1, 2, 3.$
% of observations captured by $[0,1/2]^d, d = 1,2,3.$

Sampling Observations

Question: does every row of the dataset need to be used?

If rows are selected randomly (with or without replacement), the resulting sample might be **representative** of the entire dataset.

Drawbacks:

- if the signal of interest is rare, sampling might drown it altogether
- if aggregation is happening down the road, sampling will necessarily affect the numbers (passengers vs. flights)
- even simple operations on a large file (finding the # of lines, say) can be taxing on the memory – **prior information on the dataset structure can help**

Feature Selection

Removing **irrelevant/redundant** variables is a common data processing task.

Motivations:

- modeling tools do not handle these well (variance inflation due to multicolinearity, etc.)
- dimension reduction (# variables ≫ # observations)

Approaches:

- filter vs. wrapper
- unsupervised vs. supervised

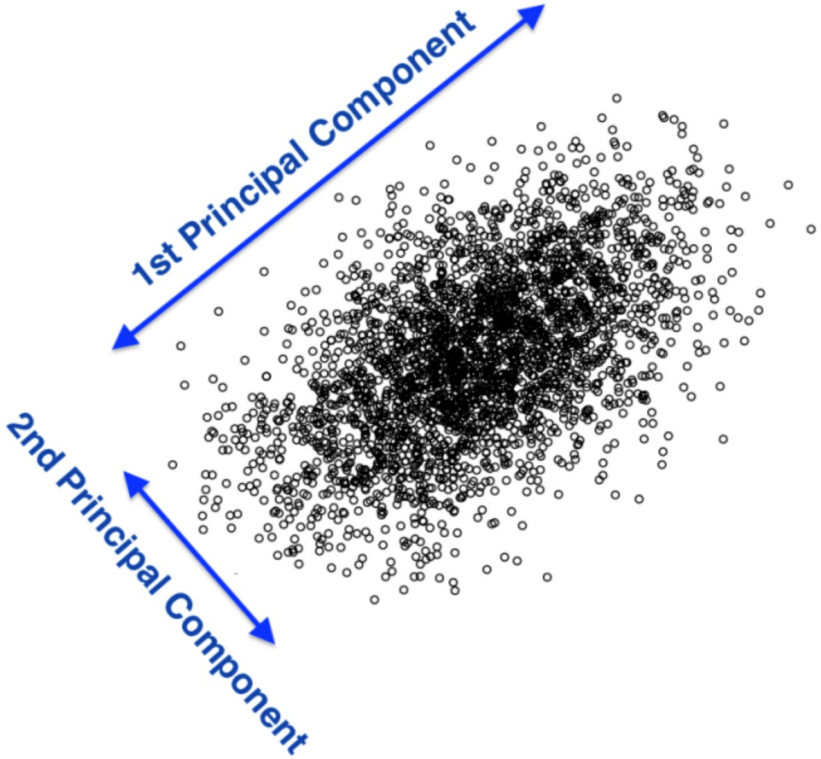
[A. Ng, K. Soo, *Numsense!*]

Dimension Reduction: PCA

Motivational Example: Nutritional Content of Food

What is the best way to differentiate food items? Vitamin content, fat, or protein level? A bit of each?

Principal Component Analysis (PCA) can be used to find the combinations of variables along which the data points are **most spread out** (dimension reduction).



[A. Ng, K. Soo, *Numsense!*]

Dimension Reduction: PCA

Presence of nutrients appears to be **correlated** among food items.

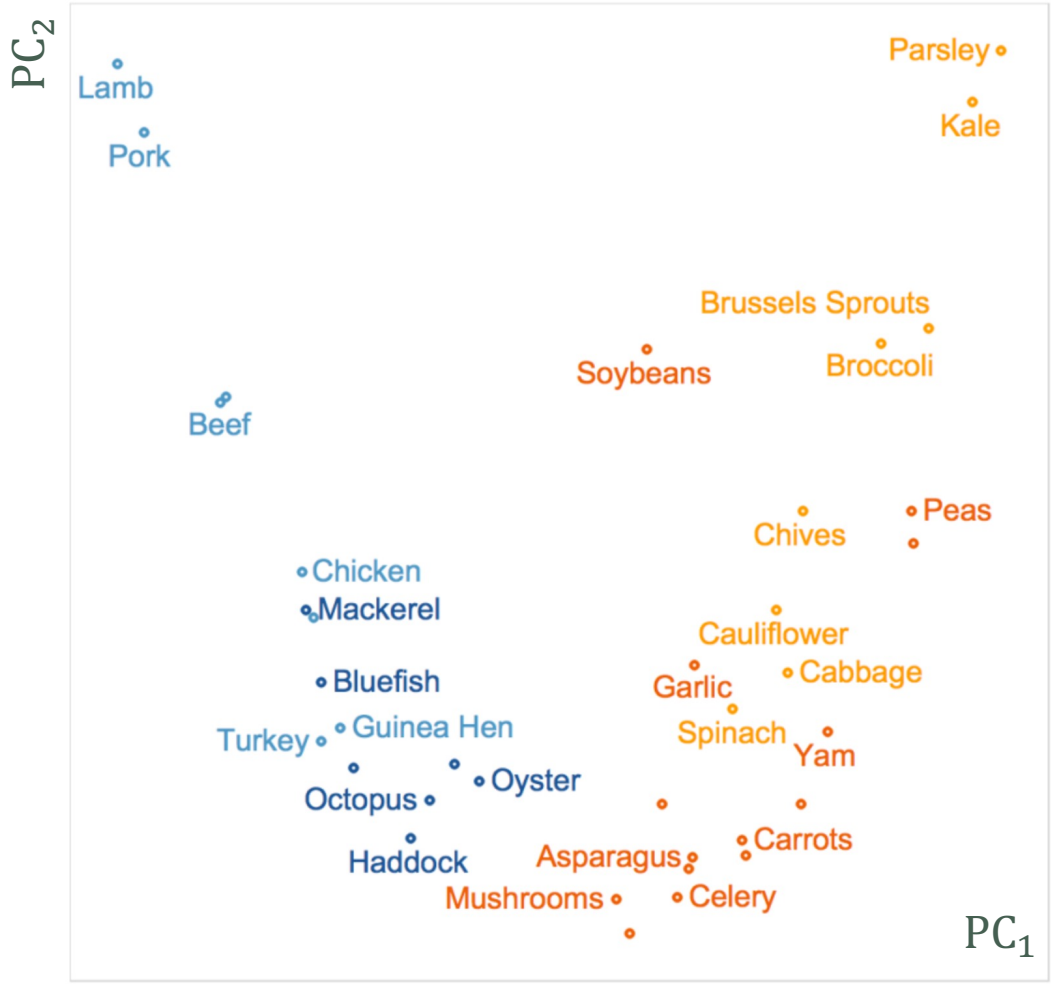
In the (small) sample consisting of Lamb, Pork, Kale, and Parsley, *Fat* and *Protein* levels seem in step, as do *Fiber* and *Vitamin C*.

In a larger dataset, the correlations are $r = 0.56$ and $r = 0.57$.

How much could 2 **derived** variables explain?

$$PC_1 = -0.45 \times \text{Fat} - 0.55 \times \text{Protein} + 0.55 \times \text{Fiber} + 0.44 \times \text{Vitamin C}$$
$$PC_2 = 0.66 \times \text{Fat} + 0.21 \times \text{Protein} + 0.19 \times \text{Fiber} + 0.70 \times \text{Vitamin C}$$

PCA Differentiation



PC_1 differentiates vegetables from meats; PC_2 differentiates 2 **sub-categories** within these:

- **meats** are concentrated on the left (low PC_1 values)
- **vegetables** are concentrated on the right (high PC_1 values)
- **seafood** have lower *Fat* content (low PC_2 values) and are concentrated at the bottom
- **non-leafy veggies** have lower *Vitamin C* content (low PC_2 values) and are also bunched at the bottom

Common Transformations

Models sometimes require that certain data assumptions be met (normality of residuals, linearity, etc.).

If the raw data does not meet the requirements, we can either:

- abandon the model
- attempt to **transform** the data

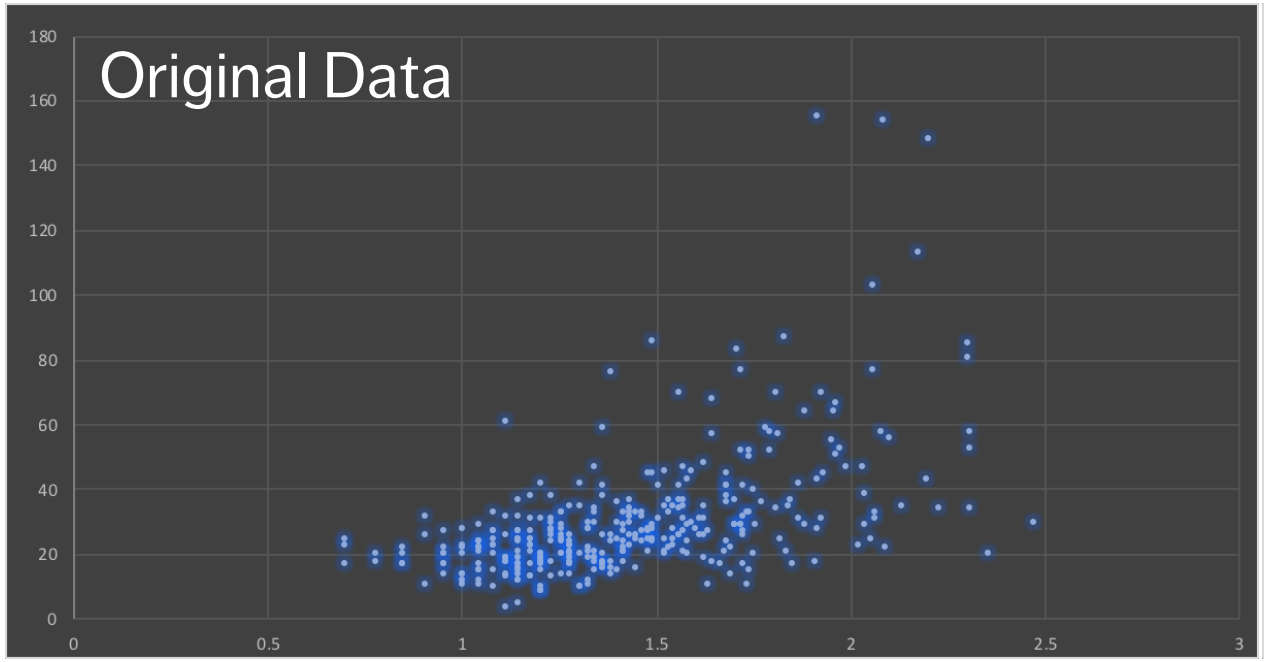
The second approach requires an **inverse transformation** to be able to draw conclusions about the **original data**.

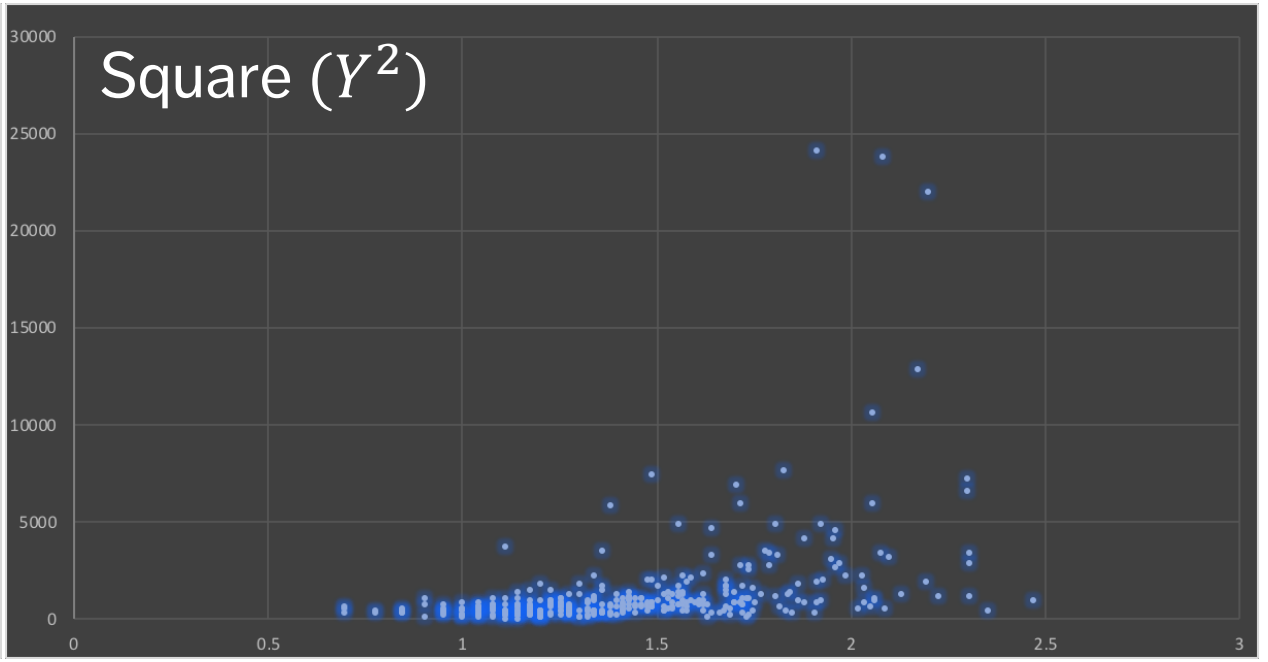
Common Transformations

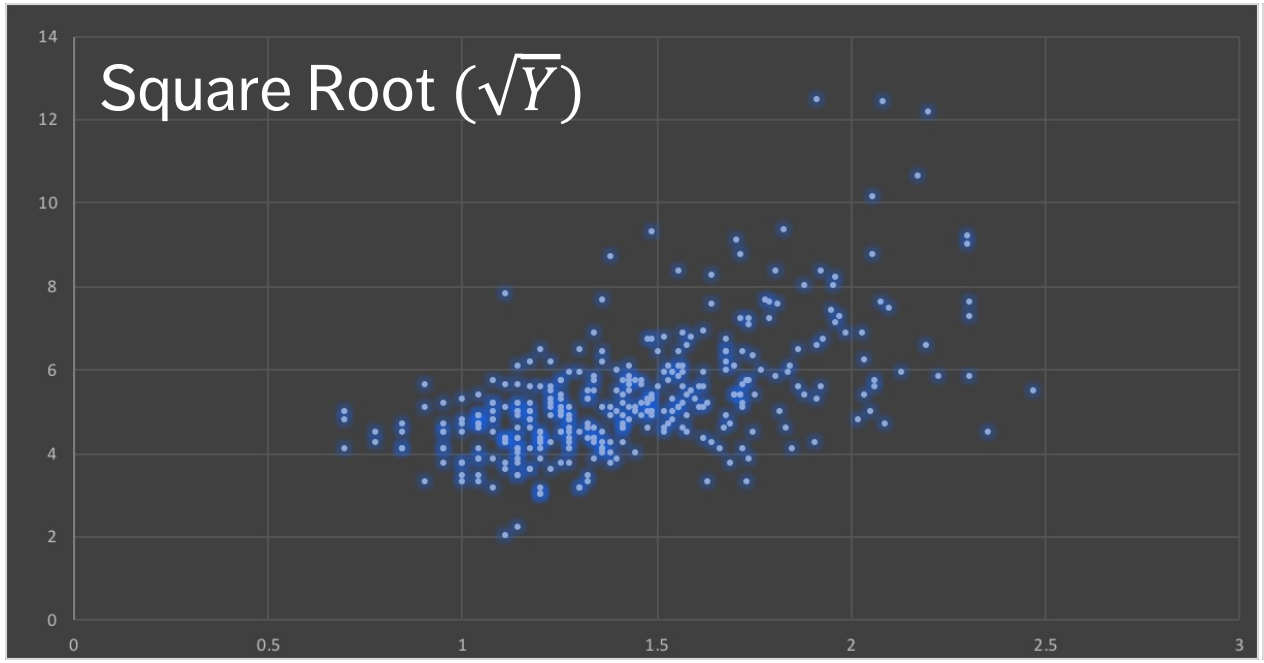
In the data analysis context, transformations are **monotonic:**

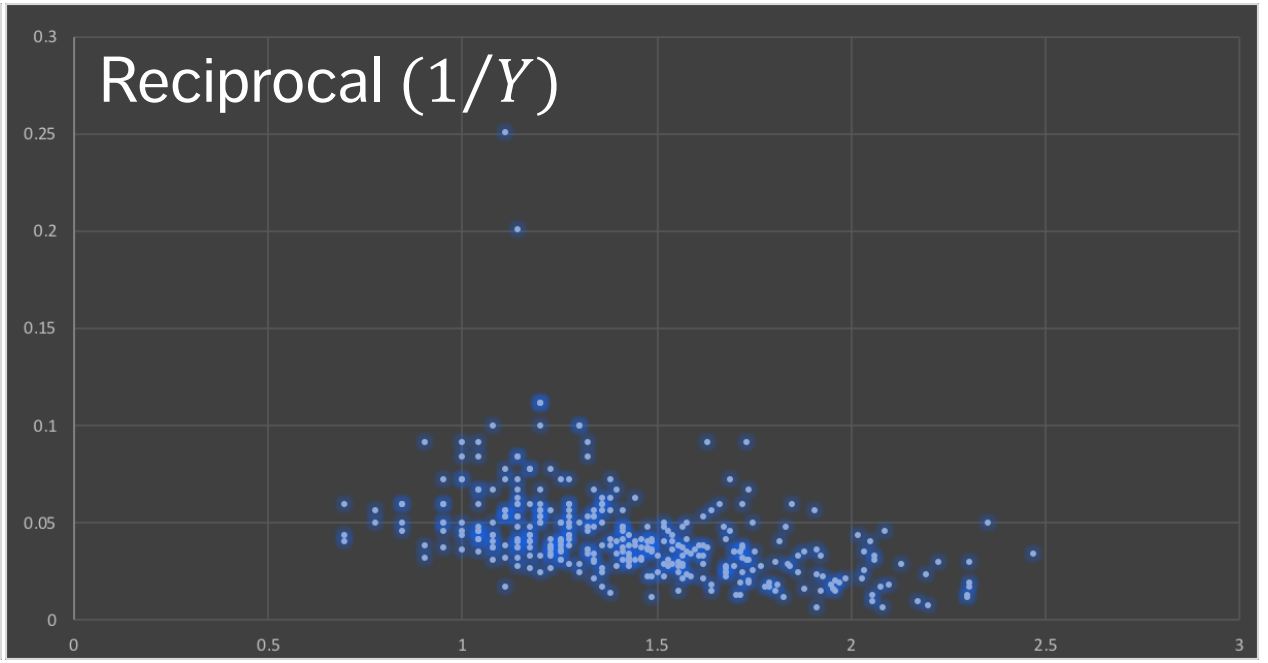
- logarithmic
- square root, inverse, power: W^k
- exponential
- Box-Cox, etc.

Transformations on X may achieve linearity, but usually at some price (correlations are not preserved, for instance). Transformations on Y can help with non-normality and unequal variance of error terms.


Original Data
180
160
140
120
100
80
60
40
20
0
0
0.5
1
1.5
2
2.5
3


Square (Y^2)
30000
25000
20000
15000
10000
5000
0
0
0.5
1
1.5
2
2.5
3


Square Root ($\sqrt{Y}$)
14
12
10
8
6
4
2
0
0
0.5
1
1.5
2
2.5
3


Reciprocal ($1/Y$)
0.3
0.25
0.2
0.15
0.1
0.05
0
0
0.5
1
1.5
2
2.5
3

gm = geometric mean

Box-Cox Transformation

Assume the usual model $Y_j = \sum_i \beta_i X_{j,i} + \varepsilon_j$ with either

- skewed residuals
- not-constant variance
- non-linear trend

The **Box-Cox transformation** $Y_j \mapsto Y_j'(\lambda)$ suggests a choice: select λ which maximizes the corresponding log-likelihood

$$Y_j'(\lambda) = \begin{cases} \mathrm{gm}(\boldsymbol{Y}) \times \ln(Y_j), & \lambda = 0 \\ \lambda^{-1} \mathrm{gm}(\boldsymbol{Y})^{1-\lambda} \times (Y_j{}^{\lambda} - 1), & \lambda \neq 0 \end{cases}$$

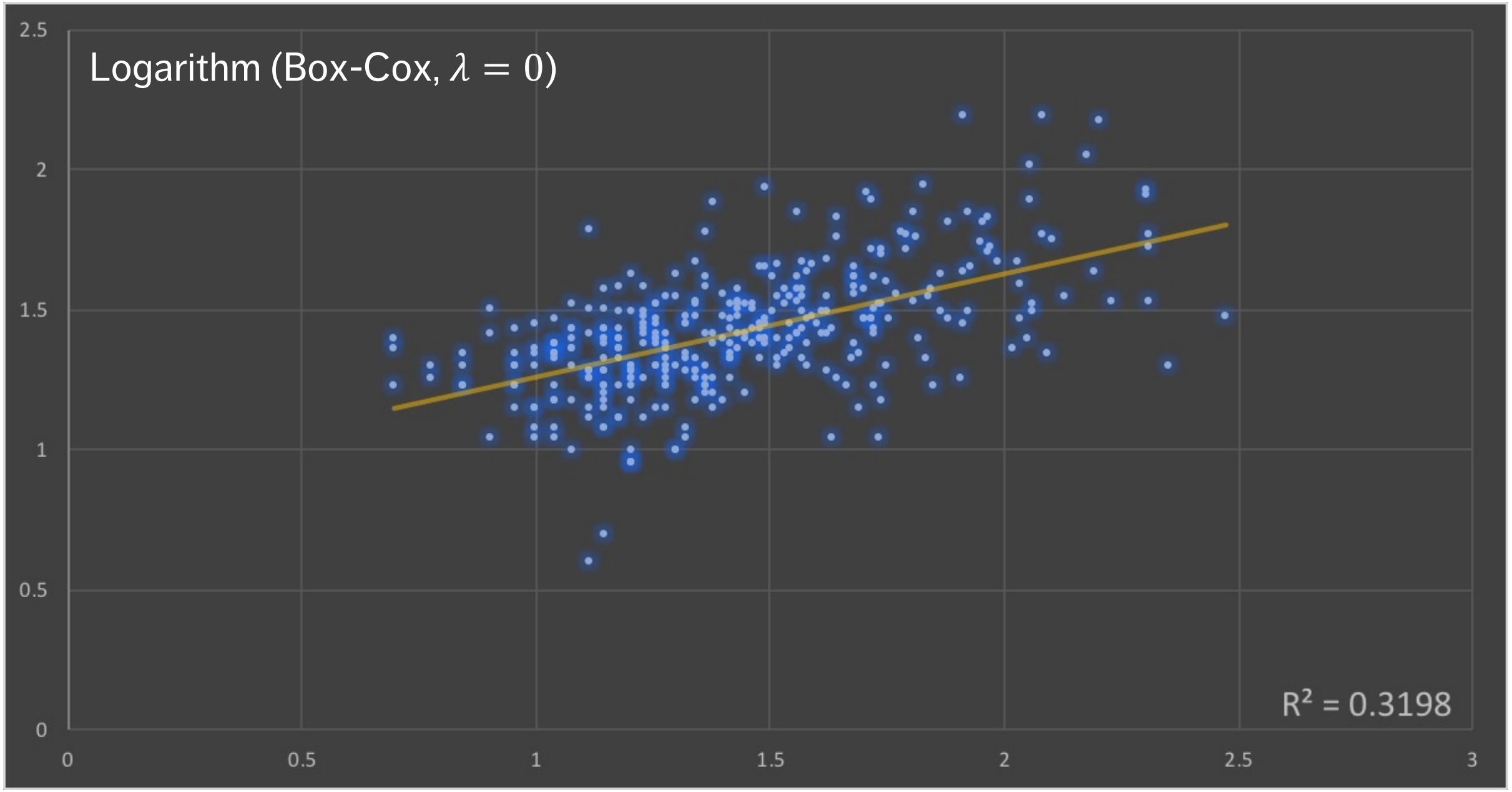
Box-Cox Transformation

The procedure provides a **guide** to select a transformation.

Theoretical/practical **rationales** may exist for a particular choice of λ.

Residual analysis is still required to ensure that the choice was appropriate.

Better to work with (interpret) the transformed data.


Logarithm (Box-Cox, $\lambda = 0$)
2.5
2
1.5
1
0.5
0
0
0.5
1
1.5
2
2.5
3
$R^2 = 0.3198$

Scaling

Numeric variables may have different **scales** (i.e., weights and heights).

The variance of a large-range variable is typically greater than that of a small-range variable, introducing a bias (for instance).

Standardization creates a variable with mean 0 and std. dev. 1:

$$Y_i = \frac{X_i - \bar{X}}{s_X}$$

Normalization creates a new variable in the range $[0,1]$: $Y_i = \frac{X_i - \min X}{\max X - \min X}$

Discretizing

To reduce computational complexity, a numeric variable may need to be replaced by an **ordinal** variable (from *height* value to "*short*", "*average*", "*tall*", for instance).

Domain expertise can be used to determine the bins' limits (although that may introduce unconscious bias to the analyses)

In the absence of such expertise, limits can be set so that either

- the bins each contain the same number of observations
- the bins each have the same width
- the performance of some modeling tool is maximized

Creating Variables

New variables may need to be introduced:

- as **functional relationships** of some subset of available features
- because modeling tool may require **independence of observations**
- because modeling tool may require **independence of features**
- to simplify the analysis by looking at **aggregated summaries** (often used in text analysis)

Time dependencies ⟶ time series analysis (lags?)

Spatial dependencies ⟶ spatial analysis (neighbours?)

Suggested Reading

Dimensionality and
Data Transformations

Data Understanding, Data Analysis, Data Science

Data Preparation

Data Transformations

- Common Transformations
- Box-Cox Transformations
- Scaling
- Discretizing
- Creating Variables

*Feature Selection and Dimension Reduction (advanced)

Exercises

Dimensionality and
Data Transformations

1. Using Example: Algae Bloom as a basis, scale, discretize, and create new variables out of the algae blooms dataset.
2. Scale, discretize, and create new variables out of the grades and cities.txt datasets.
3. Scale, discretize, and create new variables out of a dataset of your choice.

Miscellanea

DATA SCIENCE ESSENTIALS

11. Data Engineering

Background

One of the data science challenge: putting large troves of data into formats that can be **read** by algorithms.

Data engineering is related to processing an ever-increasing supply of data.

After processing, data scientists develop **proofs-of-concept**; AI/ML engineers translate these into **deployable models**.

Data/ML engineering have been around a while (software logs); with the rise of **cloud computing**, some argue that expertise in these fields is becoming more sought after than expertise in data analysis (at least, in some circles).

Data Roles (Reprise)

Data Engineers

- receive data from a source
- structure, distribute, and store data into data lakes and warehouses
- create tools and data models which data scientists can use to query the data

ML Engineers

- apply and deploy data models
- bridge gaps between data engineers and data scientists
- take proof-of-concept ideas to large scale

Data Scientists

- receive data procured/provided by DE
- extract value from the data
- build proof-of-concept predictive models
- measure and improve results
- build analytical models

Data Roles

In smaller organizations, data engineering and data science are typically **blended** into the same role.

Larger companies have **dedicated** data engineers on staff, who build **data pipelines** and manage **data warehouses** (populating them with data and creating table schemas to keep track of the stored data).

In general, DE ≠ DS.

Data Pipelines

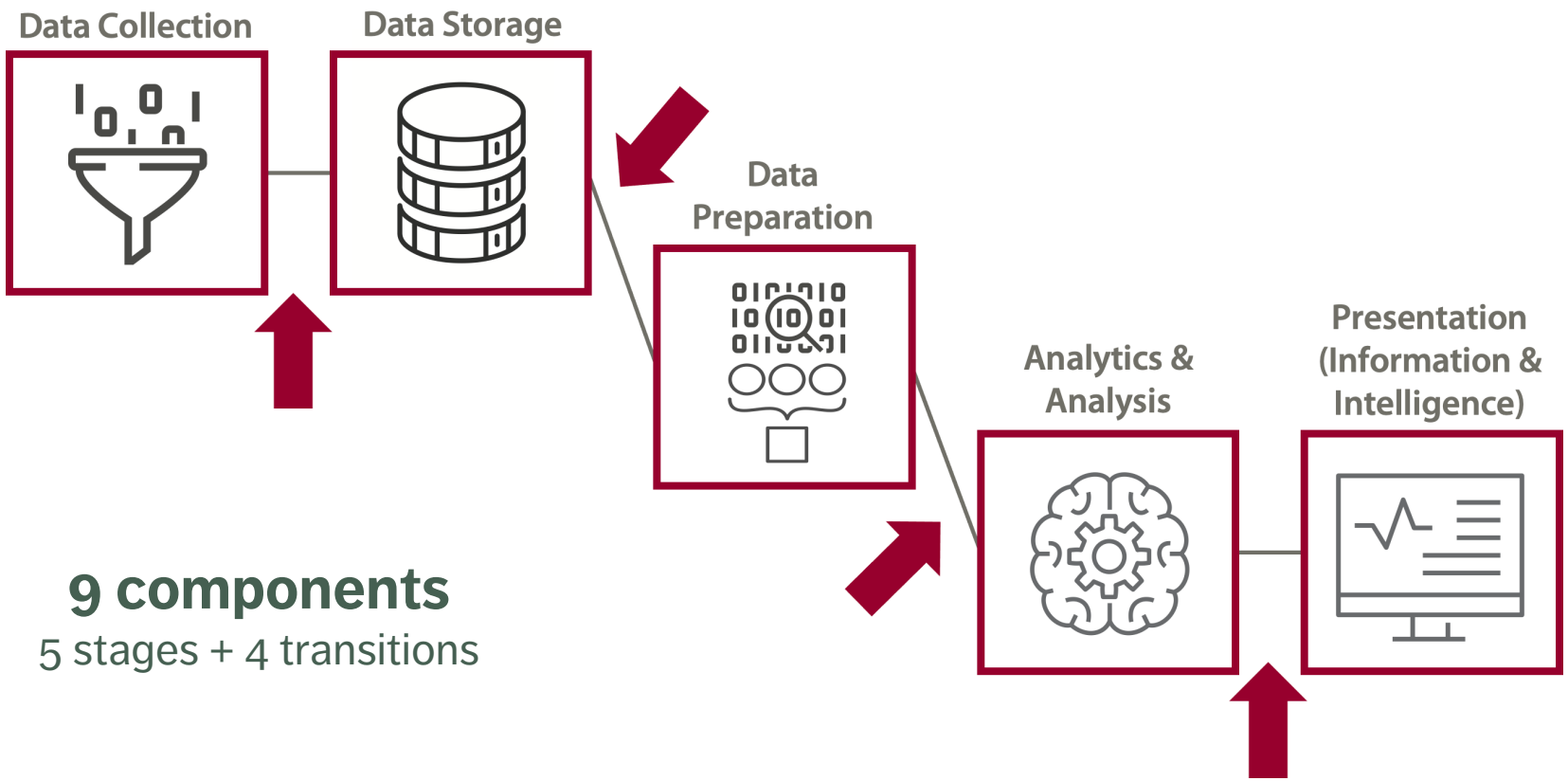
Data engineering

- operations that create **interfaces** and **mechanisms** for the flow and access of information
- setting up **data infrastructure**, preparing it for further analysis by data scientists

Data can arise from many **sources** (and types of sources), and in a variety of formats and size.

Transforming this into a process that data scientists can use and from which they can derive meaning is known as **building a data pipeline**.

Data Pipelines



9 components

5 stages + 4 transitions

Data Pipelines

Main data engineering challenge:

- building a pipeline that **runs in (close to) real-time whenever it is requested**
- so that users get **up-to-date information** from the source with **minimal delays**

Working pipeline proof-of-concept solutions are passed on to ML engineers for **deployment** and **production**. Some of the work surrounding this includes:

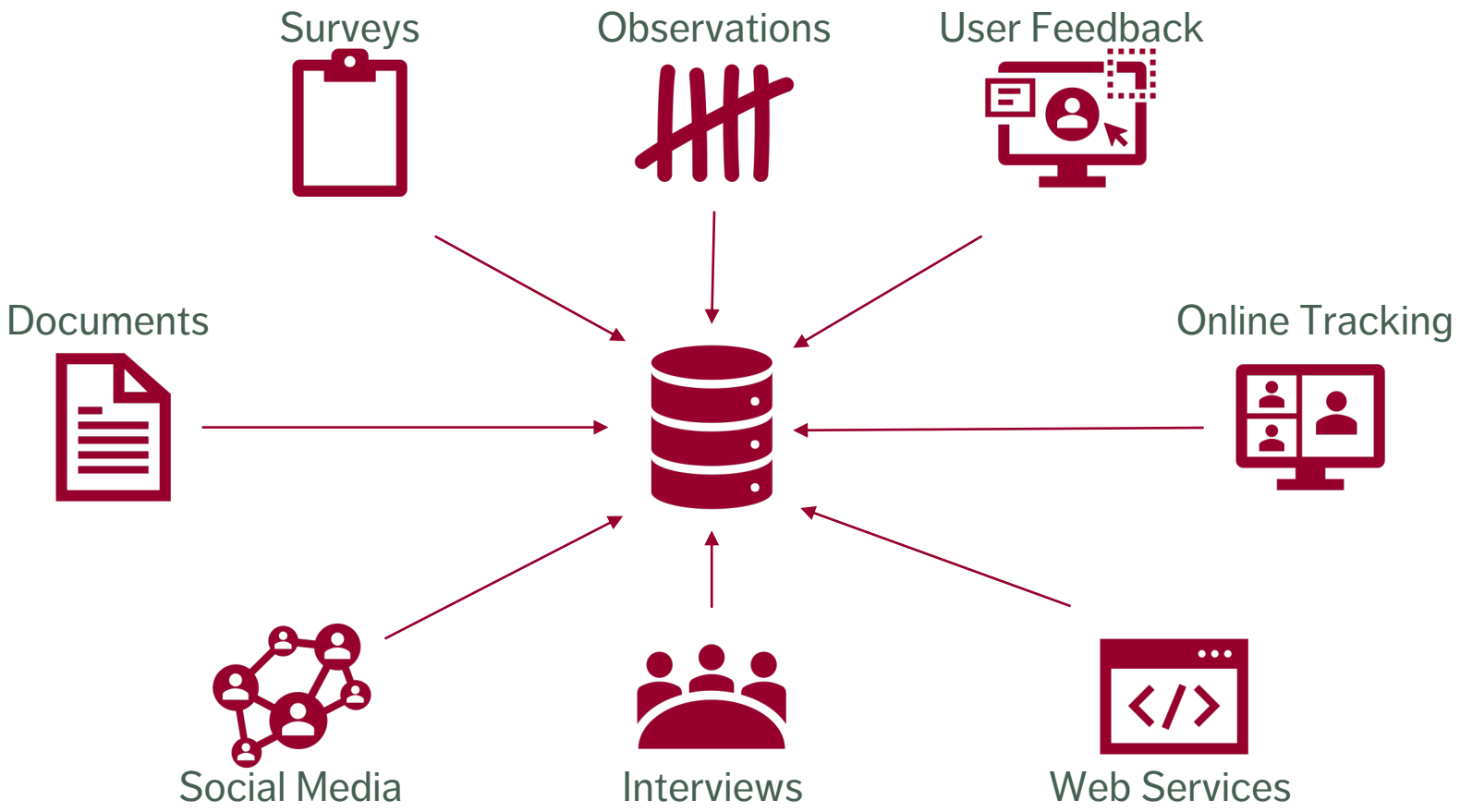
- data quality checks
- optimizing query performance
- creating a continuous integration/continuous delivery ecosystem around model changes
- ingesting data from various sources into the data model
- carrying machine learning and data science techniques to distributed systems.

Data Pipelines

Common **themes** (operations/framework/tasks/sources) for pipeline steps :

- **data collection:** applications, mobile apps, microservices, Internet of Things (IoT) devices, websites, instrumentation, logging, sensors, external data, user generated content, etc.
- **data storage:** Master Data Management (MDM), warehouse, data lake, etc.
- **data integration/preparation:** ETL, stream data integration, etc.
- **data analysis:** machine learning, predictive analytics, A/B testing, experiments, artificial intelligence (AI), deep learning, etc.
- **delivery and presentations:** dashboards, reports, microservices, push notifications, email, SMS, etc.

Data Collection



ETL – Extract

Extract | Transform | Load

ETL – Transform

Changing Structure

Altering Data Types

Aggregating Data

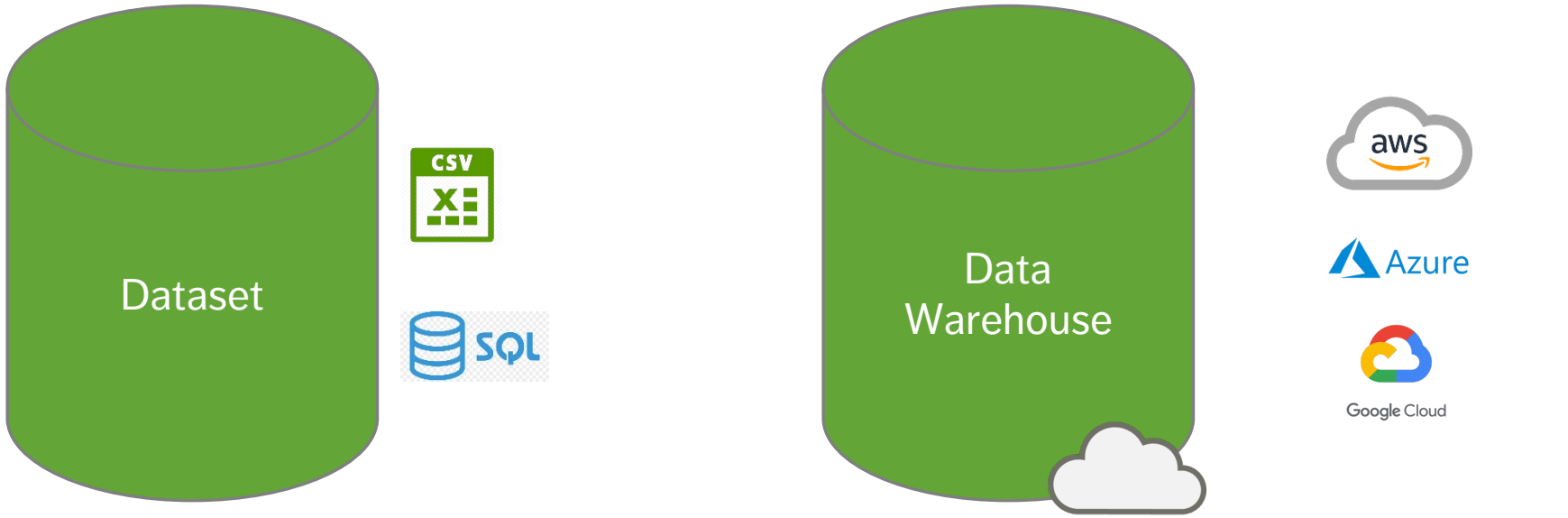
Cleaning

Joining

Grouping

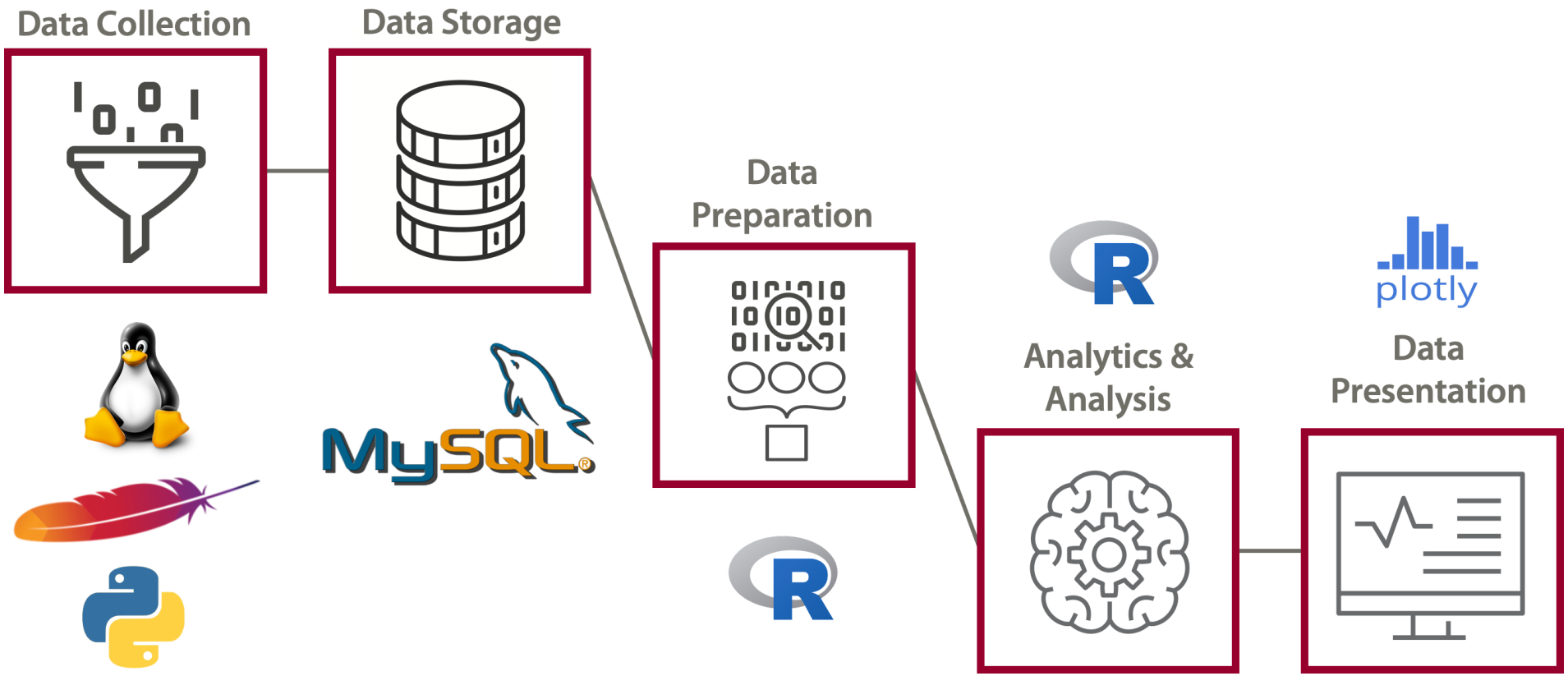
Extract | **Transform** | Load

ETL – Load

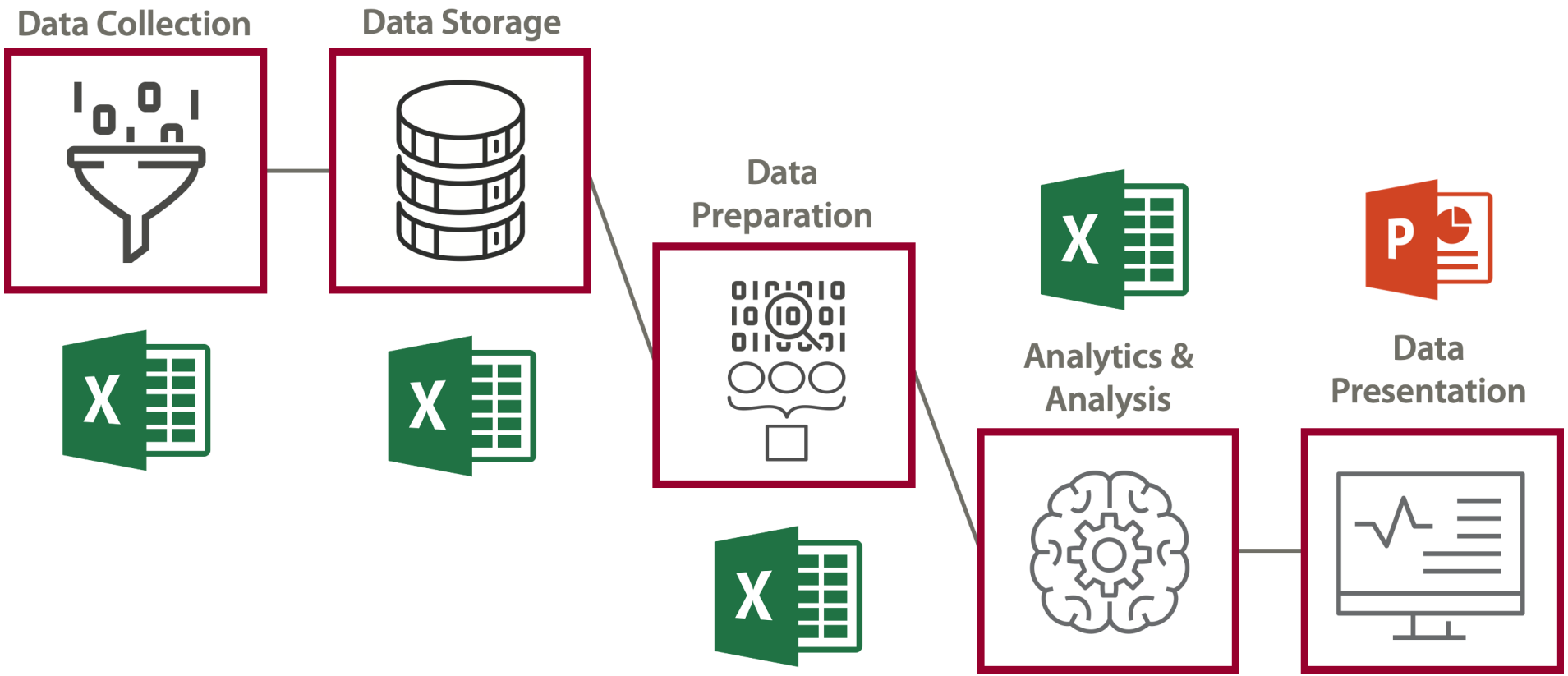


Extract → Transform → **Load**

Data Pipeline: Open Source



Data Pipeline: GoC (?)



Data Pipeline Tools

Pipelines let users split large tasks into a series of smaller sequential steps, which can help **optimize** each step.

If using TensorFlow for the analysis component of a DL pipeline which consists of a single large script, then **everything** from data collection to presentation has to be done with TensorFlow; may not be optimal.

Data pipeline tools select the best framework/language for each pipeline component/task:

- Luigi (Spotify)
- Airflow (AirBnB)
- scikit-learn
- pandas/tidyverse
- etc.

Data Engineering Tools

It is unlikely that one data engineer could achieve mastery over all possible data engineering tools, but teams might get a lot of **coverage:**

- **analytical databases** (Big Query, Redshift, Synapse, etc.)
- **ETL** (Spark, Databricks, DataFlow, DataPrep, etc.)
- **scalable compute engines** (GKE, AKS, EC2, DataProc, etc.)
- **process orchestration** (AirFlow/Cloud Composer, Bat, Azure Data Factory, etc.)
- **platform deployment and scaling** (Terraform, custom tools, etc.)
- **visualization tools** (Power BI, Tableau, Google Data Studio, D3.js, ggplot2, etc.)
- **programming** (tidyverse, numpy, pandas, matplotlib, scikit-learn, scipy, Spark, Scala, Java, SQL, T-SQL, H-SQL, PL/SQL, etc.)

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.

Data Governance

Goals:

- create self-service data culture
- establish internal rules for data use
- implement compliance requirements
- improve internal and external comms
- increase value of data
- reduce costs
- continually manage risks
- ensure continued existence

Suggested Reading

Data Engineering

Data Understanding, Data Analysis, Data Science
Data Engineering and Management

Background and Context

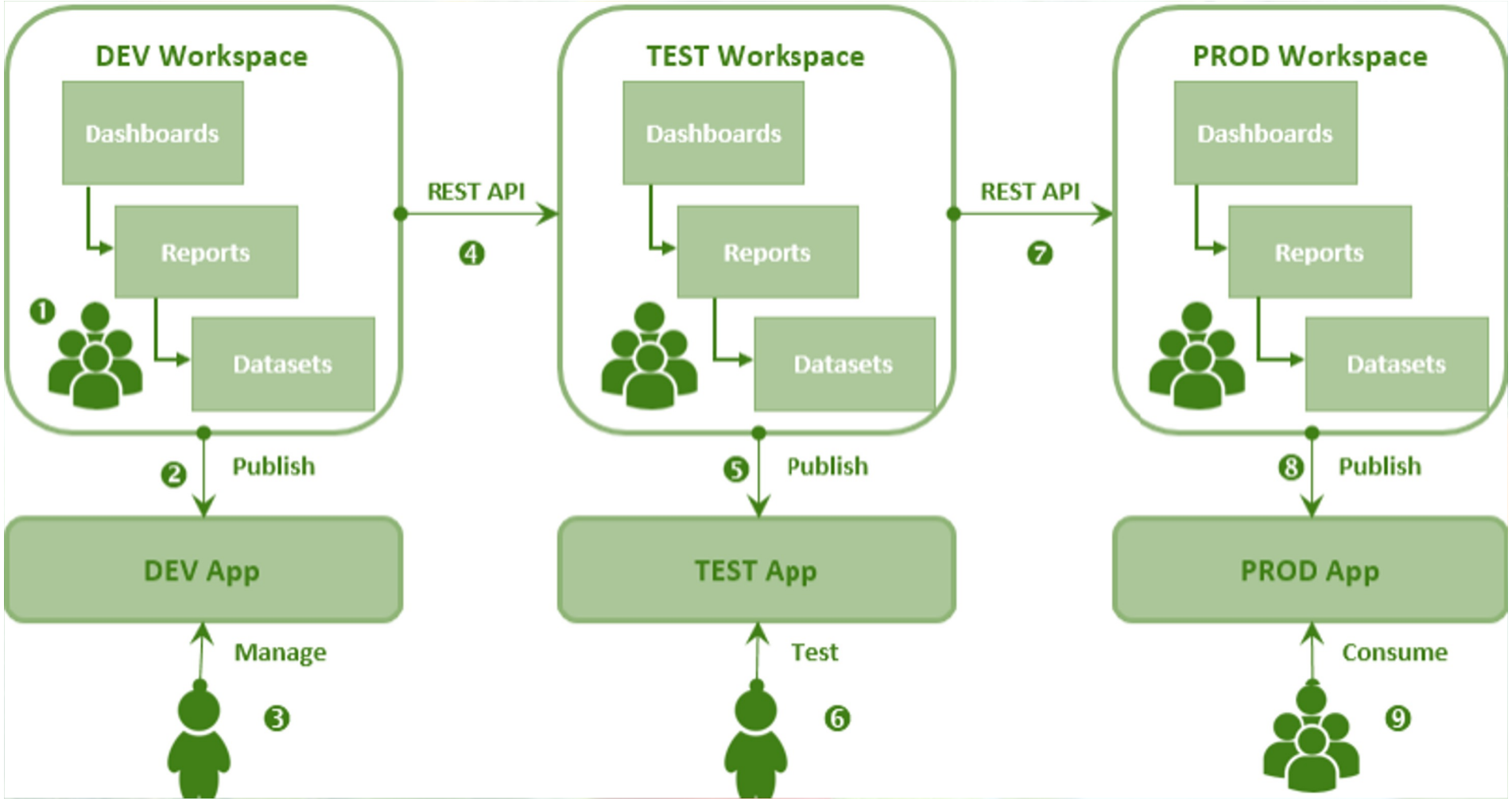
Data Engineering

- Data Pipelines
- Automatic Deployment and Operations
- Scheduled Pipelines and Workflows
- Data Engineering Tools

Exercises

Data Engineering

1. What does your (or your organization's) data science pipeline look like? Could it be improved?
2. Identify instances where you have had issues due to data availability, usability, consistency, integrity, quality, security, or trustworthiness.
3. Complete any of the previous exercises you have not had the chance to finish.



12. Data Management

Fundamental Concepts

Data and **knowledge** must be structured so that it can be:

- stored and accessible
- added to
- usefully and efficiently extracted from that store (extract – transform – load)
- operated over by **humans** and **computers** (programs, bots, A.I.)

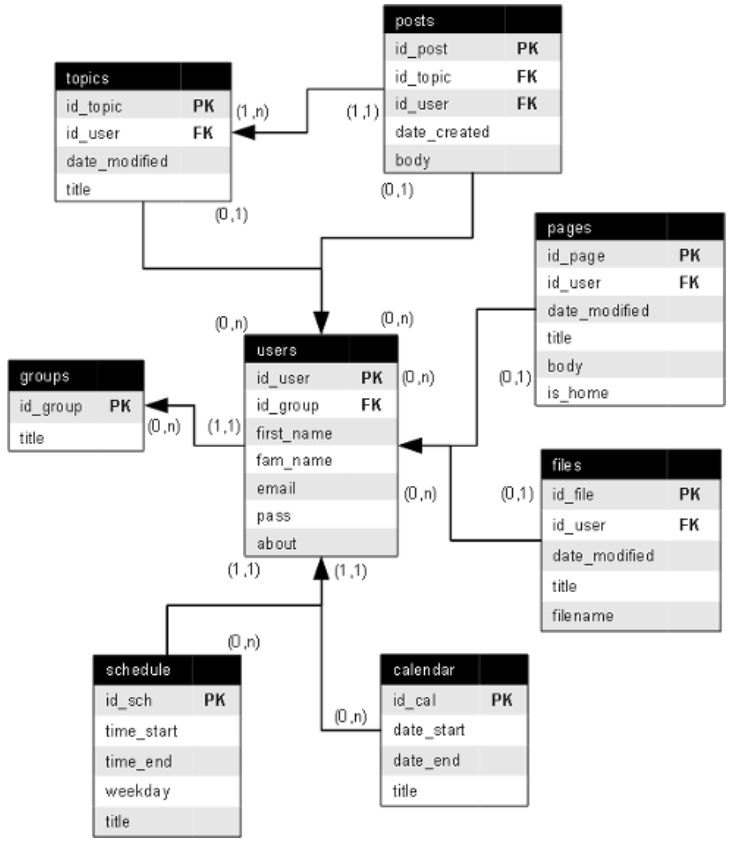
Data Modeling

Data models are **abstract/logical** descriptions of a system, using terms that are implementable as the structure of a type of data management software.

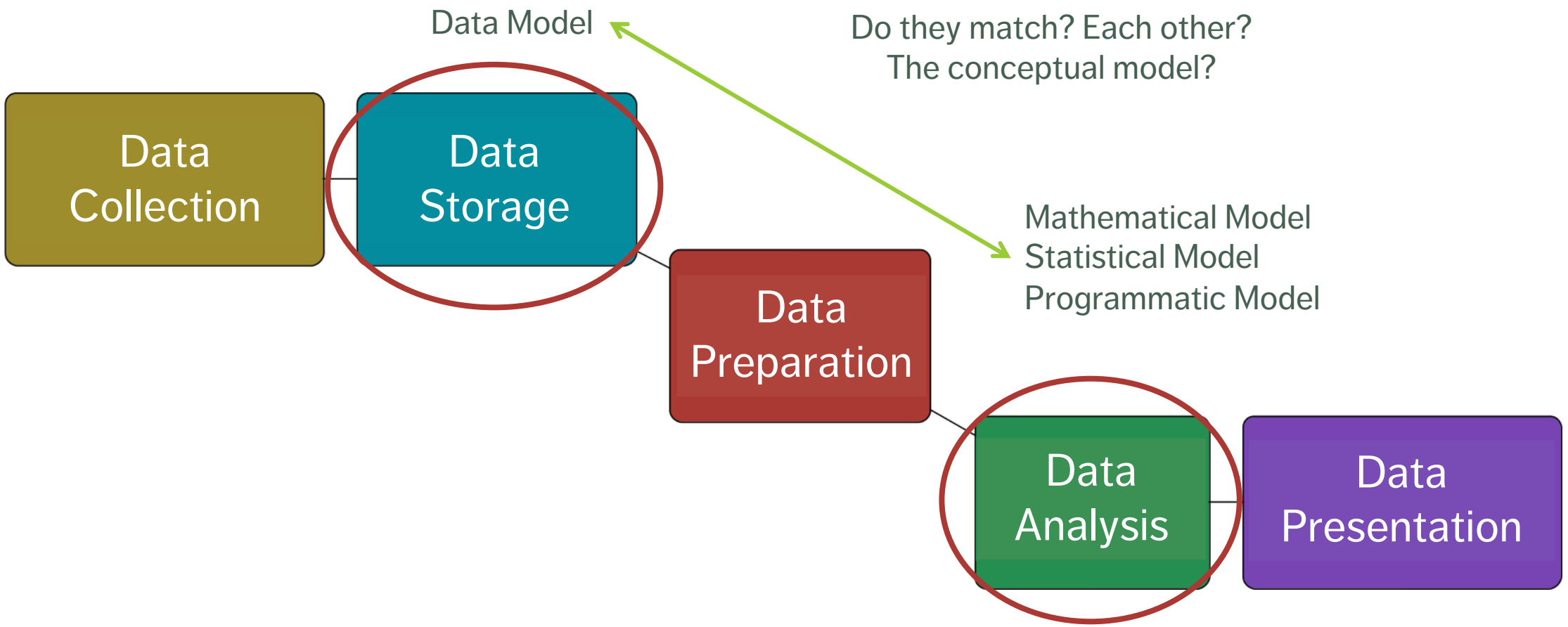
This is half-way between a **conceptual model** and a **database implementation**.

The data itself is about **instances** – the model is about the **object types**.

Another option to consider: **ontologies**.



Automated Data Pipeline



[Representation of Langerhans cells in the *Cell Ontology*, A. M. Masci *et al.*]

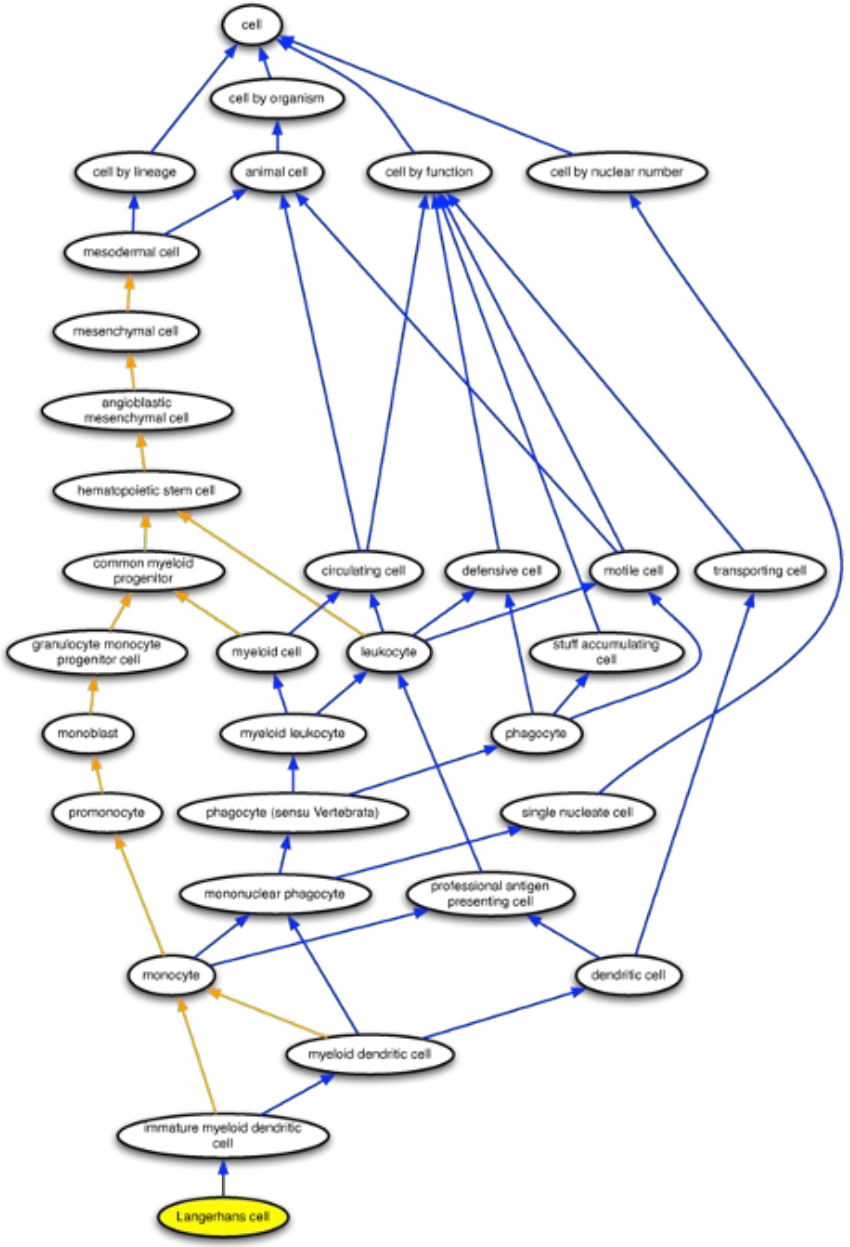
Contextual Metadata

Something gets lost when we move from conceptual models to either a data or a knowledge model.

One way of keeping the context is to provide rich **metadata** – data **about** the data.

Metadata is crucial when it comes to carrying out strategies for working across datasets.

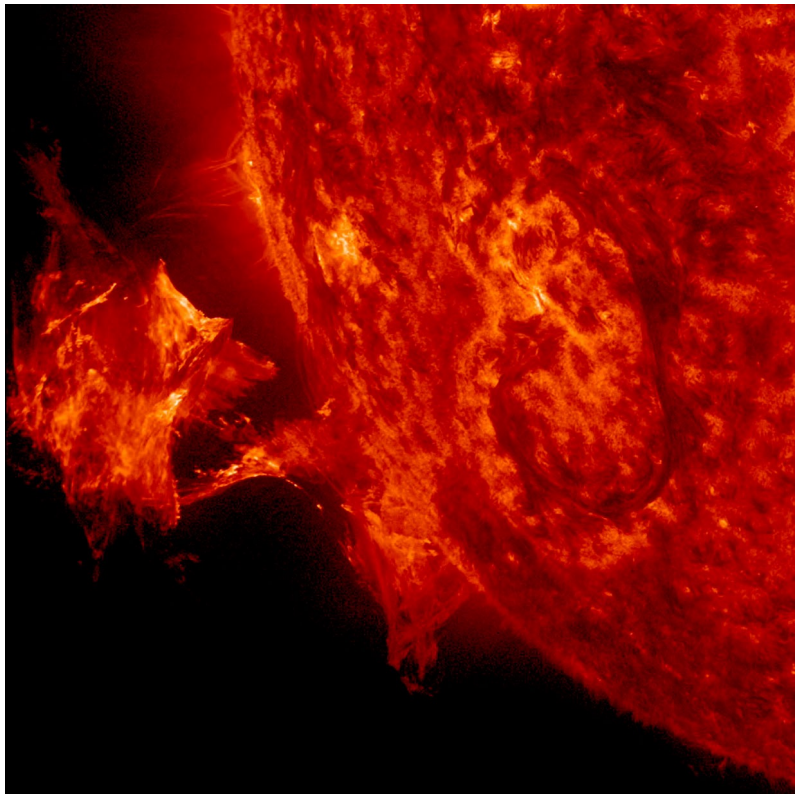
Ontologies can also play a role here.



Structured/Unstructured Data

A major motivator for new developments in database types and other data storing strategies is the increasing availability of **unstructured** data and '**blob**' data:

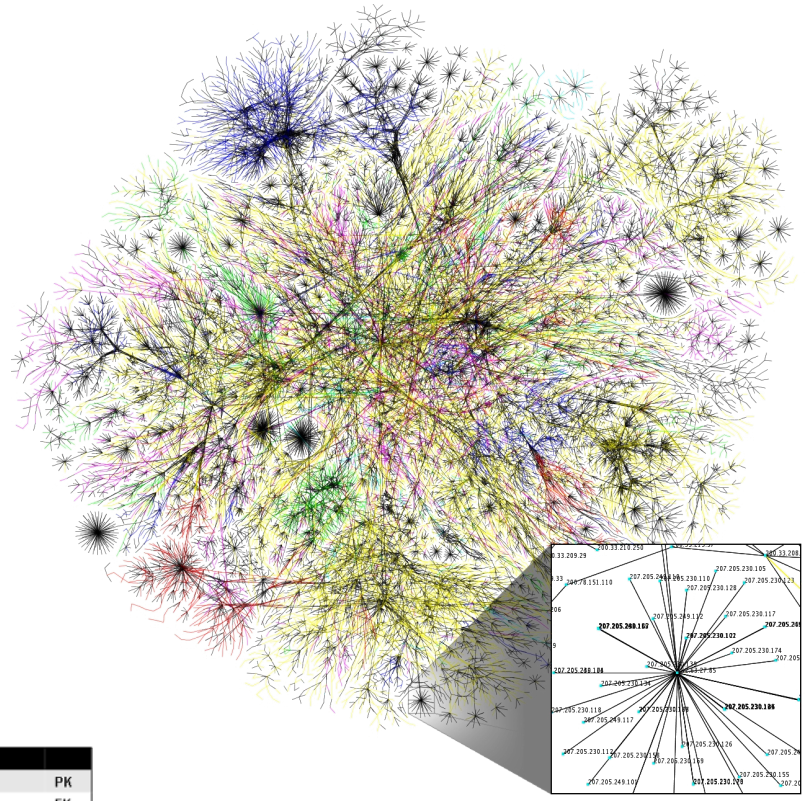
- **structured data:** labeled, organized, discrete structure is constrained and pre-defined
- **unstructured data:** not organized, no specific pre-defined structure data model (text)
- **blob data: B**inary **L**arge **Ob**ject (blob) – images, audio, multi-media

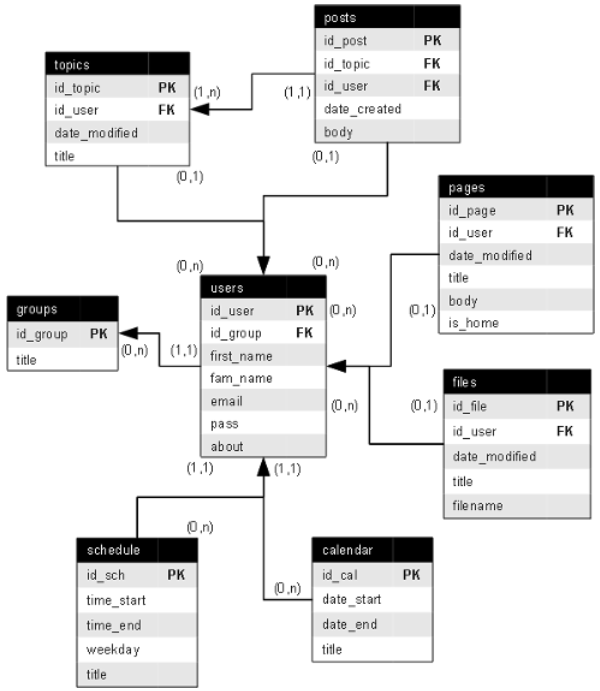


Data Modeling

Different options are currently popular in terms of fundamental **data** and **knowledge** modeling or structuring strategies:

- key-value pairs (e.g., JSON)
- triples (e.g., RDF)
- graph databases
- relational databases
- spreadsheets





Stores and Databases

Relational Database:

- widely supported, well understood, works well for many types of systems and use cases, difficult to change once implemented, doesn't deal with relationships well

Key-Value Stores:

- can take any sort of data, no need to know much about its structure in advance, missing values don’t take up space, can get messy, difficult to find specific data

Graph Databases:

- fast and intuitive for heavily relation-based data, might be the only option in this case as traditional databases may slow to a crawl, probably overkill in other cases, not yet widely supported

Flat Files and Spreadsheets

Pros:

- very efficient if collecting data only once, about one particular type of object
- some types of analysis require all the data in one place
- easy to read into analysis software and do operations over the entire dataset

Cons:

- very hard to manage data integrity if continually collecting data
- not ideal for system data involving multiples types of objects and relationships
- can be very difficult to carry out data querying operations

Tools and Buzzwords

- MongoDB, ArangoDB
- Document store
- JSON, YAML
- API, GraphQL
- Linked Data
- Semantic Web
- Ontology Web Language (OWL)
- Protégé
- SQL, etc.

Data Model Implementation

To implement your data/knowledge model, one needs access to **data storage and management software**.

This can be a challenge for individuals: such software usually runs on **servers**.

Servers are good because they allows multiple users to access a single database **simultaneously**, from different client programs, but it makes it difficult to "play" with the data.

This is where **SQLite** comes into play.

Data Management Software

Data management software provides users with an easy way to interact with their data.

It's essentially a **human** – **data** interface.

Through this interface, users can:

- add data to their data collection
- extract subsets of data from their collection based on certain criteria
- delete or edit data in their collection

Names / Terminology

Previously:

- database
- data warehouse
- data marts
- database management system
- (SQL)

Now:

- data lake
- data pool
- data swamp?
- data graveyard?
- (NoSQL)

Increasingly: distinction between data **store** and data **management software**.

From Data Model to Implementation

Once the (logical) data mode is **completed**

1. **instantiate the model** in chosen software (e.g., create tables in MySQL)
2. **load the data**
3. **query the data:**
   - traditional relational databases use **Structured Query Language** (SQL)
   - others use different query languages (AQL, semantic engines, etc.) or rely on bespoke computer programs (e.g., written in R, Python)

Database Management

Once data has been collected, it must also be **managed**.

Fundamentally, this means that the database must be **maintained**, so that the data is

- **accurate,**
- **precise,**
- **consistent**
- **complete**

Don’t let your data lake turn into a data swamp!

Cloud Service Provider

1. Store **large** amounts of data
2. Run expensive and advanced processes with **click of a button**
3. **Flexible** and **scalable**
4. Enable **low-code** data wrangling

Cloud vs. On-Premise

Cloud

- hands-off
- pay-as-you-go model
- questionable data ownership

On-Premise (On-Prem)

- self-maintained
- all costs absorbed
- fully-controlled security

Suggested Reading

Data Management

Data Understanding, Data Analysis, Data Science

Data Science Basics

Getting Insight From Data

- Structuring and Organizing Data

Data Engineering and Management

Data Management

- Databases
- Data Modeling
- Data Storage

Reporting and Deployment

- Reports and Products
- Cloud and On-Premise Architecture

Exercises

Data Management

1. Does your organization have data? If so, is it hosted on-premise or on the cloud? How is it accessed? Structured?
2. Complete any of the previous exercises you have not had the chance to finish.

Suggested Exercises and Guided Projects

DATA SCIENCE ESSENTIALS

Between Sessions

Session 1 to Session 2

- complete the exercises of session 1
- download the datasets from the website
- read Programming Primer (sections 1 – 4)
- install R / RStudio (Posit)
- install the following R packages: dplyr, xts, knitr, tidyverse, ggplot2, pastecs, Hmisc, e1071, psych, quantmod, ggm, kerndwd, MASS, DMwR, ROCR, car, forcats, corrplot

Session 2 to Session 3

- complete the exercises of session 2

Session 3 to Session 4

- complete the exercises of session 3

After Session 4

- complete the exercises of session 4
- attempt the guided projects

Guided Project I

Select a data project of interest (personally or professionally) and provide a planning draft for it, touching on the topics discussed in this course. The following questions can help:

1. What are some questions associated with the project?
2. What is the conceptual model of the underlying situation?
3. What kind of dataset(s) exist that could help you answer these questions?
4. Are there data or analytical limitations?
5. Do you need to collect new data to handle such questions?
6. How is the data stored/accessed? What are the infrastructure requirements?
7. What do deliverables look like?
8. How would successes be quantified/qualified?
9. What are your timelines and availability?
10. What skillsets are required to work on this project?
11. Would you work on this alone or as part of a team?
12. How costly would it be to initiate and complete this project?
13. What does the data analysis pipeline look like?
14. What software and analytical methods will be used?

Guided Project II

Write a paper discussing some of the ethical issues surrounding the use of artificial intelligence, data science, and/or machine learning (M.L.) algorithms.

Establish a list of the 3 most important ethical principles that the use of such algorithms should abide by. Explain why you have selected each of these principles.

Describe (at least) 2 real-life instances of the use of A.I./D.S./M.L. in the public sector, the private sector, or in academia, when the ethical principles you have chosen were violated. Discuss how the failure to abide by your selected ethical principles have caused (or could cause) harm to individuals, organizations, countries, etc.

Suggest how the projects discussed above could have been modified so that their use of A.I./D.S./M.L. algorithms would abide by your selected ethical principles.

Guided Project III

This project uses the Gapminder Tools (there is also an offline version).

1. Take some time to explore the tool. In the online version, the default starting point is a bubble chart of 2020 life expectancy vs. income, per country (with bubble size associated with total population). In the offline version, select the "Bubbles" option.
2. Can you identify the available variable categories and some of the variables? [You may need to dig around a bit.]
3. Why do you think that Gapminder has selected Life Expectancy and Income as the default plotting variables?
4. Replace Life Expectancy by Babies per woman. Observe and discuss the changes from the default plot.
5. Formulate a few questions that could be answered with the default data.
6. Formulate a few questions that could be answered using some of the other variables.
7. At what point in the data science workflow do you think that visualizations of this nature could be useful?
8. Do these visualizations provide a sound understanding of the system under investigation (the geopolitical Earth)?

Guided Project III (cont.)

9. What do you think the data sources are for the underlying dataset? [You may need to dig around the internet to answer this question].
10. Are all variables and measurements equally trustworthy? How could you figure this out?
11. Is the underlying dataset structured or unstructured?
12. Provide a potential data model for the dataset.
13. What are the types of the 4 default variables (Life Expectancy, Income, Population, World Regions)?
14. Play around with the charts for a bit. Can you find pairs of variables that are positively correlated? Negatively correlated? Uncorrelated?
15. Among those variables that are correlated, do any seem to you to exhibit a dependent-independent relationship? How could you identify such pairs?
16. Can you provide an eyeball estimate of the mean, the median, and the range of various numerical variables?
17. Can you provide an eyeball estimate of the mode of the categorical variables?
18. Can you identify epochal moments (special temporal points) in the data where a shift occurs, say?
19. Is the tool and its underlying dataset useable? What factors does your answer depend on?

Guided Project III (cont.)

20. Do you think that there could be problems with the reported values? For instance, select Sweden and the United States from the checkbox menu on the right and follow their path from 1799 to 2018/2020. From what point onwards are the values sensible? What do you think is happening at the start of the series?
21. Follow Eritrea for the same duration. Look up the country's independence date from Ethiopia. What do you think the measurements prior to that date represent?
22. Follow Austria for the same duration. Look up the historical timeline of the country's boundaries (Austria-Hungary, Anschluss, modern borders, etc.). What does that imply for the measurements?
23. Follow Finland for the same duration. What happens in 1809? Does that tell you anything about the way data is coded in the dataset?
24. De-select all countries and let the simulation run from 1799 to 2018/2020. Can you identify instances where a large subset of observations behaves in unexpected manners? If so, do you think that this is due to data cleaning/data processing issues?
25. Continue exploring the dataset. You may change which variables are displayed or work with some of the other visualization methods. Overall, do you think that the dataset is sound? Would you use it to run analyses? What are some of its strengths and weaknesses?

Guided Project IV

Select a dataset from the list below (or any other set of interest to you):

- GlobalCitiesPBI.csv
- 2016collisionsfinal.csv
- polls_us_election_2016.csv
- HR_2016_Census_simple.xlsx

For your dataset(s):

1. Create a "data dictionary" to explain the different fields and variables. Can you find a source for these datasets online?
2. Develop a list of questions you would like answered about the datasets.
3. Investigate individual variables (through simple charts, univariate statistics, etc.).
4. Repeat the process with bivariate investigations (though simple charts, joint distributions, variable interactions, etc.).
5. Do you trust the dataset, or not? Support your answer. If you do not trust the dataset, flag potential invalid entries, anomalous observations, missing values, or outliers. How should these entries be treated?
6. Does any of your analysis suggest that some of the variables should be transformed? Do any of the questions you developed in step 2 support such transformations? If so, transform the data appropriately.

References

DATA SCIENCE ESSENTIALS

References

C. C. Aggarwal, *Data Classification: Algorithms and Applications*. CRC Press, 2015.

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

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

P. Boily, *Data Understanding, Data Analysis, and Data Science*. Data Action Lab, 2022.

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

Coursera, "Introduction to Data Engineering."

References

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

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

H. Konsek, "Automating Data Pipelines: Types, Use Cases, Best Practices." Soft Kraft.

J. Kunigk, I. Buss, P. Wilkinson, and L. George, *Architecting Modern Data Platforms: A Guide to Enterprise Hadoop at Scale*. O'Reilly Media, 2018.

T. Malaska and J. Seidman, *Foundations for Architecting Data Solutions: Managing Successful Data Projects*. O'Reilly Media, 2018.

References

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

T. Orchard and M. Woodbury, *A missing information principle: theory and applications*. University of California Press, 1972.

R. W. Paul and L. Elder, *Understanding the Foundations of Ethical Reasoning*, 2nd ed. Foundation for Critical Thinking, 2006.

What is Data Engineering? Everything You Need to Know in 2022. phData, 2022.

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

References

T. Raghunathan, J. Lepkowski, J. Van Hoewyk, and P. Solenberger, "A multivariate technique for multiply imputing missing values using a sequence of regression models," *Survey Methodology*, vol. 27, no. 1, pp. 85–95, 2001.

D. B. Rubin, *Multiple imputation for nonresponse in surveys*. Wiley, 1987.

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

simplystatistics.org, "An interactive visualization to teach about the curse of dimensionality."

S. van Buuren, *Flexible imputation of missing data*. CRC Press, 2012.

A. Watt, *Database Design*. BCCampus, 2014.

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

Innovation, Science and Economic Development Canada:

- Canada's Digital Charter in Action: A Plan by Canadians, for Canadians