Introduction to Data Science

Instructor: Patrick Boily
Slides: P. Boily (IACS, DAL, uOttawa), M. Kashef (datascience2go), J. Schellinck (Sysabee, DAL, AI Guides)

uOttawa
Institut de développement professionnel
Professional Development Institute

Introduction to Data Science

Instructor: Patrick Boily

Slides: P. Boily (IACS, DAL, uOttawa), M. Kashef (datascience2go), J. Schellinck (Sysabee, DAL, AI Guides)

Outline

Module 1

Data Insight Fundamentals

Module 2

Data Collection and Data Management

Module 3

Data Visualization and Data Communication

Module 4

Data Processing and Data Cleaning

Module 5

Data Exploration and Data Analysis

Module 6

Data Mining and Machine Learning

Instructor

Bio

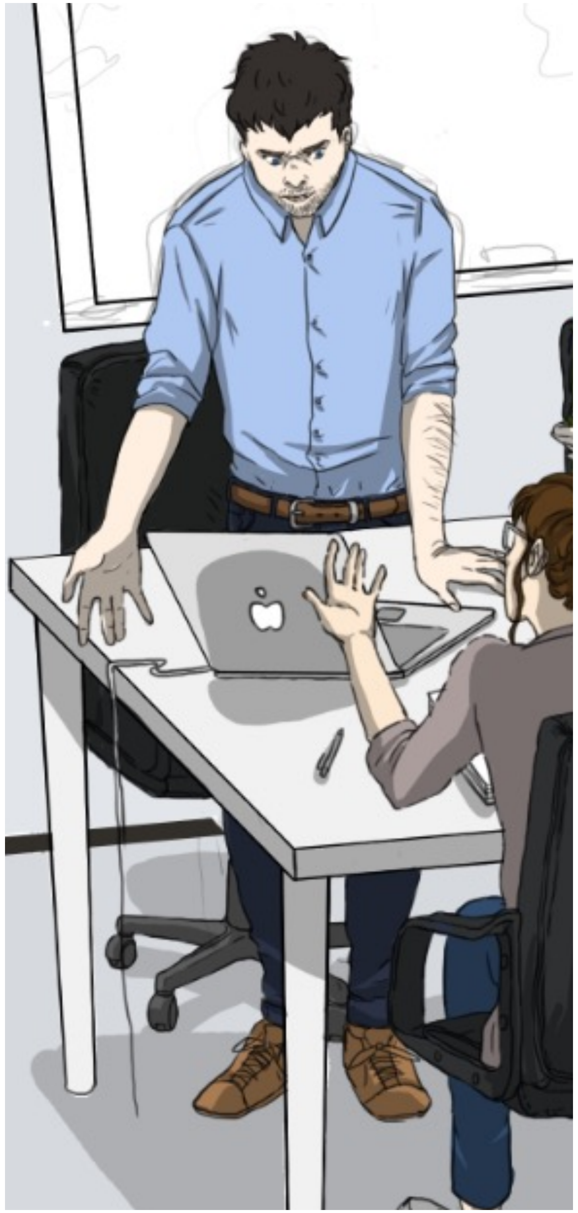
- Prof. Math/Stat ['19 – now, uOttawa]
- Manager and Senior Consultant ['12 – '19, CQADS, Carleton]
- Lecturer ['99 – '19, uOttawa | UQO | Carleton]
- Public Service ['08 – '12, ASFC | StatCan | TC | TPSGC]
- 60+ uni course; 250+ workshop days

Projects

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

Specialization

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



Suggested References

Data Understanding, Data Analysis, and Data Science

P. Boily

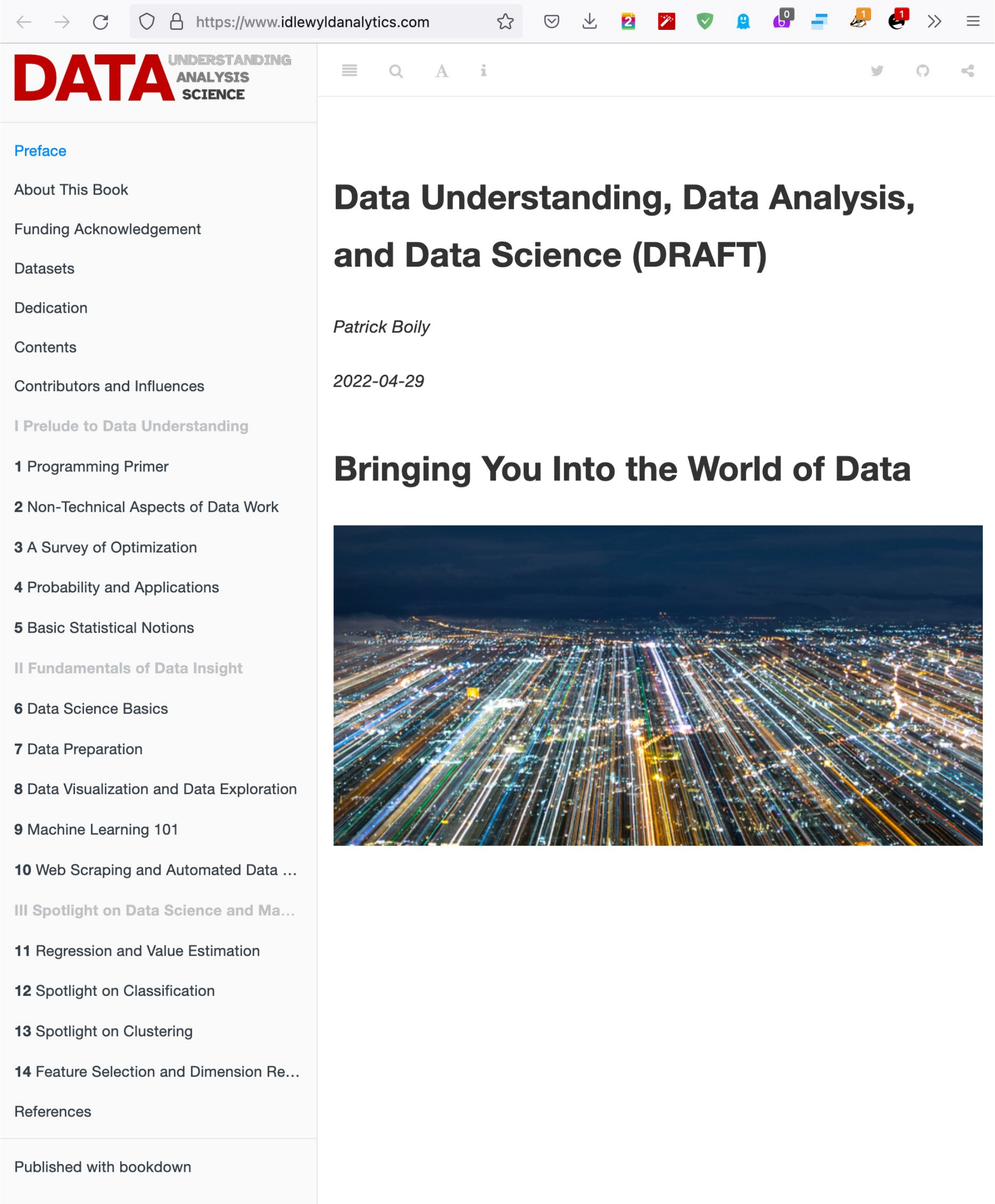
idlewyldanalytics.com

Data Science Basics (suggested exercise: #4)

Data Preparation (suggested exercise: #4)

Data Visualization & Data Exploration (sugg. ex: #7)

Machine Learning 101 (suggested exercise: #18)



https://www.idlewyldanalytics.com
DATA UNDERSTANDING ANALYSIS SCIENCE

Published with bookdown

Data Understanding, Data Analysis, and Data Science (DRAFT)

Patrick Boily

2022-04-29

Bringing You Into the World of Data

Roundtable

Quick Intro

Experience

Why this course?

Module 1

Data Insight Fundamentals

“Reports that say that something hasn't happened are always interesting to me, because as we know, there are **known knowns**; there are things we know that we know. There are **known unknowns**; that is to say, there are things that we now know we don't know. But there are also **unknown unknowns** – there are things we do not know we don't know.”

Donald Rumsfeld, US Department of Defense News Briefing, 2002

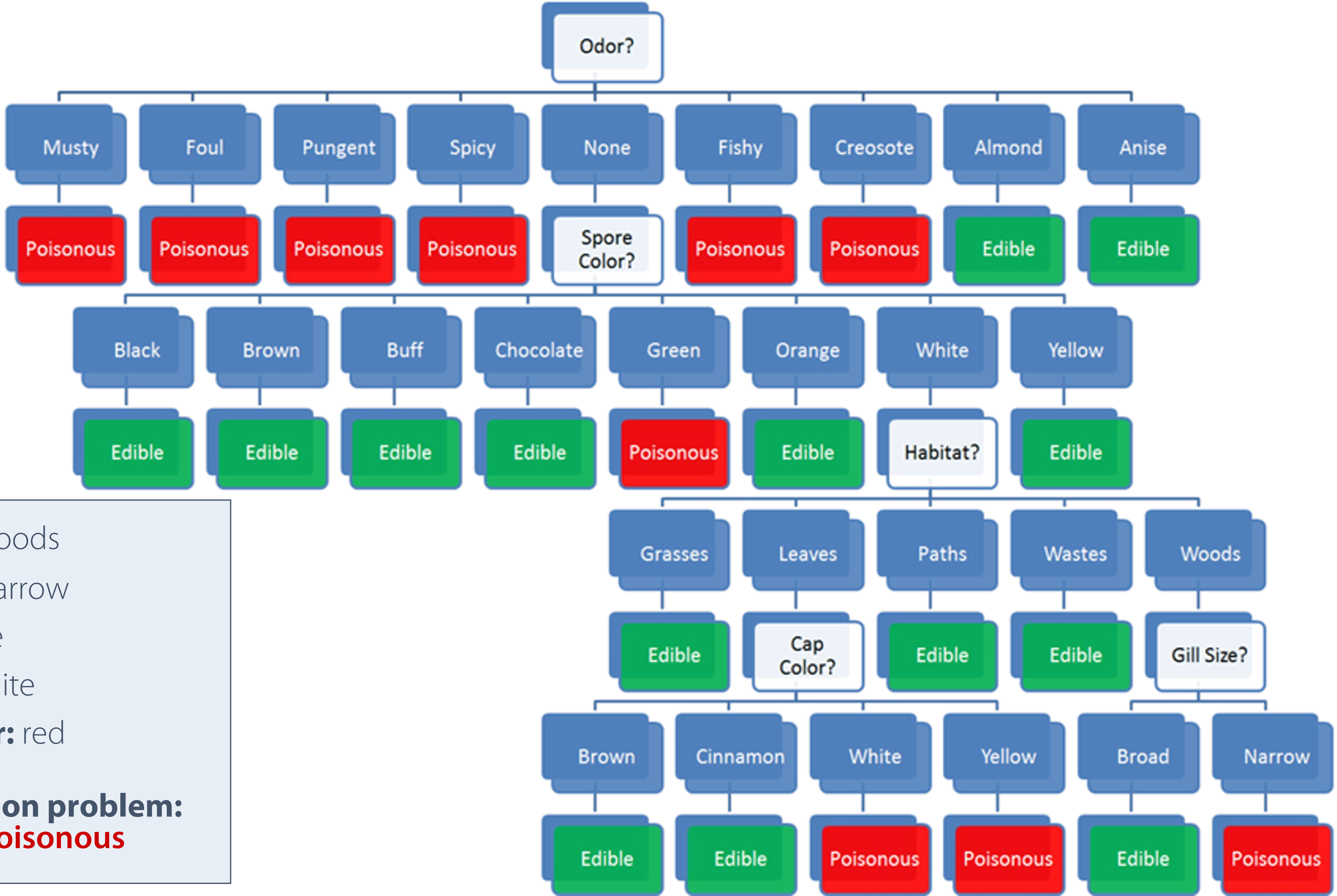
[https://archive.ics.uci.edu/ml/datasets/Mushroom]

Poisonous Mushroom Dataset

Amanita muscaria

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

Classification problem:
Is Amanita muscaria edible, or poisonous?



Habitat: woods

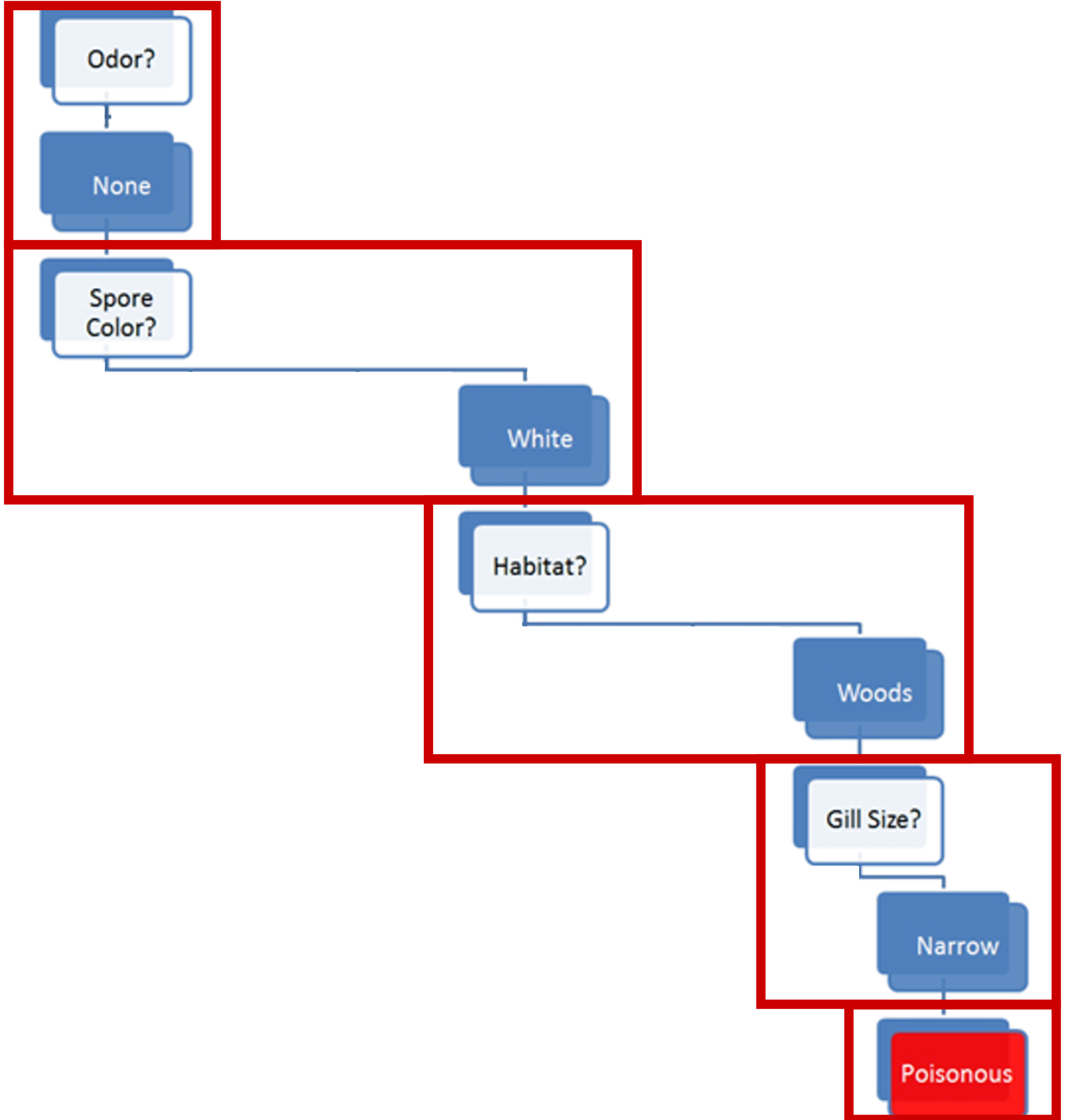
Gill Size: narrow

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**

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?

Asking the Right Questions

Is this an image of a cat or a dog?

Will the customer click this link?

What topics are described in this article?

What's the sentiment of this tweet?

Is this credit card transaction suspicious?

Is this insulin reading unusual?

What will the temperature be next Friday?

What will sales for next quarter be?

Source: kdnuggets

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?

Exercise: Roadmap to Framing Questions

Possible Initial Question 1: Should I buy a house? (vague)

Possible Initial Question 2: Should I buy a single house in Scotland?

Additional Rules

Avoid **glazing over the data** before you settle on the question.

You can be **blinded by love**; you can be **blinded by solutions**.

Do you **fully understand** what you're asking?

Source: kdnuggets

Source: Healthy Families BC

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 **everyone** in my team/department understand the question, regardless of their backgrounds?

3. Does the question need more than one sentence to express?

4. Is the question '**balanced**' – is the scope **so broad** that the question will never truly be answered; **so narrow** that the resulting impact is minimal?

5. Is the question being **skewed to what may be easier to answer** for my team's particular skillset(s)?

Source: The Head Game

Are these good questions?

Question	Specific?	What's the range in answers to this question?
How does **rain** affect goal percentage at a **soccer match**?	No, could be any soccer field	Could completely vary based on location, teams, level of players
Did the **Toronto Maple Leafs** *beat* the **Edmonton Oilers**?		
Did you like watching the **Tokyo Olympics**?		
What **types of recovery drinks** do hockey players drink?		
How many **medals** will Canada achieve at the Paris **2024 Olympics**?		
Should we fund the **Canadian Basketball** team *more* than the **Canadian Hockey** team?		

What is Data Science?

Data science is the collection of processes by which we extract useful and **actionable** insights from data.

T. Kwartler (paraphrased)

Data science is the **working intersection** of statistics, engineering, computer science, domain expertise, and “hacking.” It involves two main thrusts: **analytics** (counting things) and **inventing new techniques** to draw insights from data.

H. Mason (paraphrased)

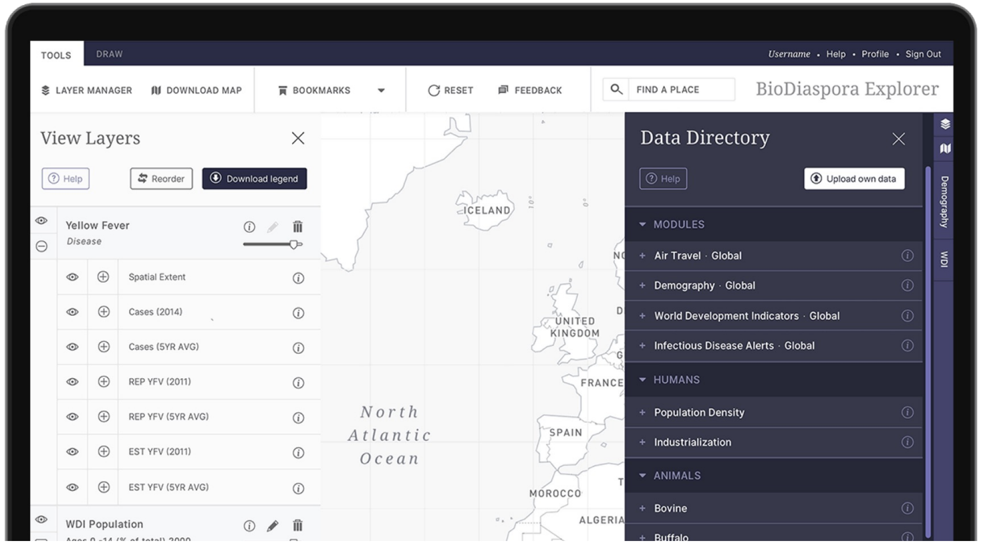
Case Study: BlueDot

Digital health company that tracks the spread of infectious diseases **globally** and assesses their risk of spread and impact worldwide.

Correctly predicted 10 out of 12 of the first cities that were going to see their first COVID case

Using advanced data science techniques, they've built a **global early warning system for infectious diseases:**

- Mapping out 200+ diseases 24/7, processing 100,000+ articles a day in over 65 languages
- Understanding impact of the spread
- Alerting clients to better inform policy decisions

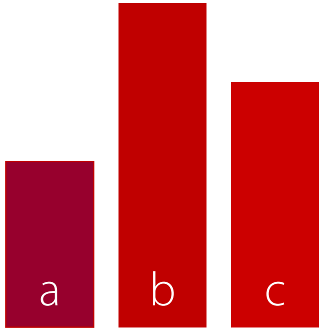


Source: BlueDot

Analytics Modes

Analytics can be broken down into four core **key buckets:**

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

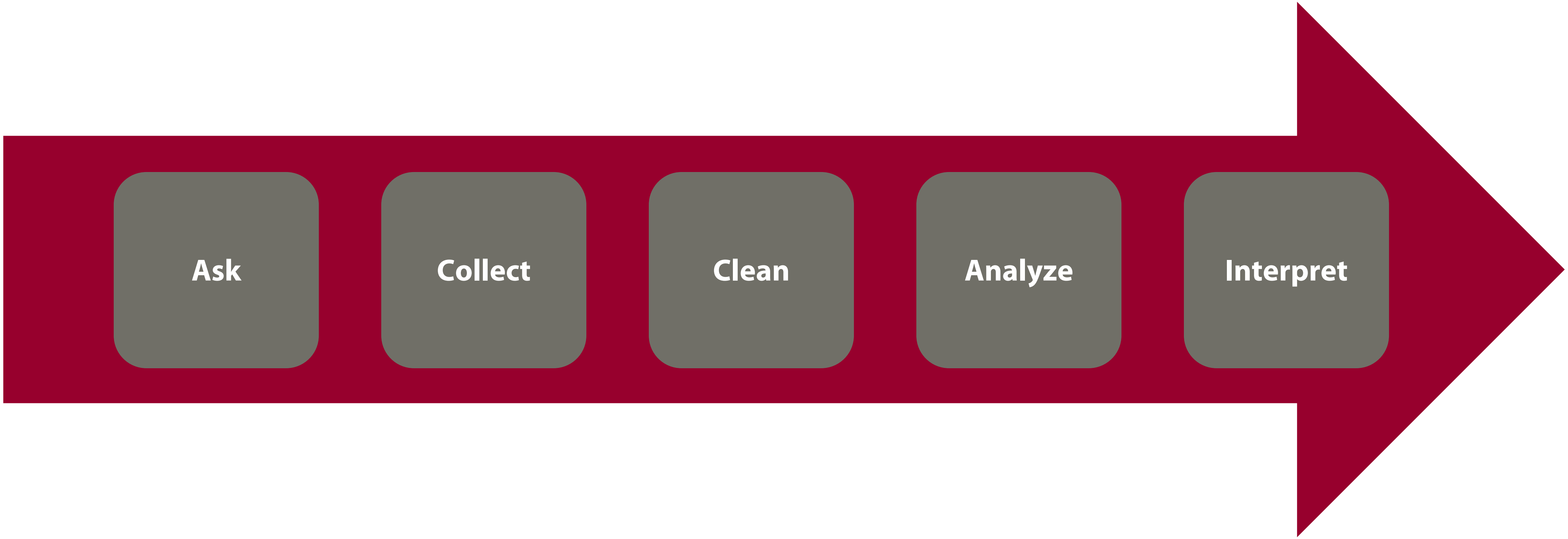
Data Science Ecosystem

Data analysis is a **team sport**, with team members needing a good understanding of both **data** and **context**

- data management
- data preparation
- analysis
- communications

Even **slight improvements** over a current approach can find a useful place in an organization – data science is not solely about **Big Data**, disruption, the "Cloud", etc.!

Data Science Workflow



Ask

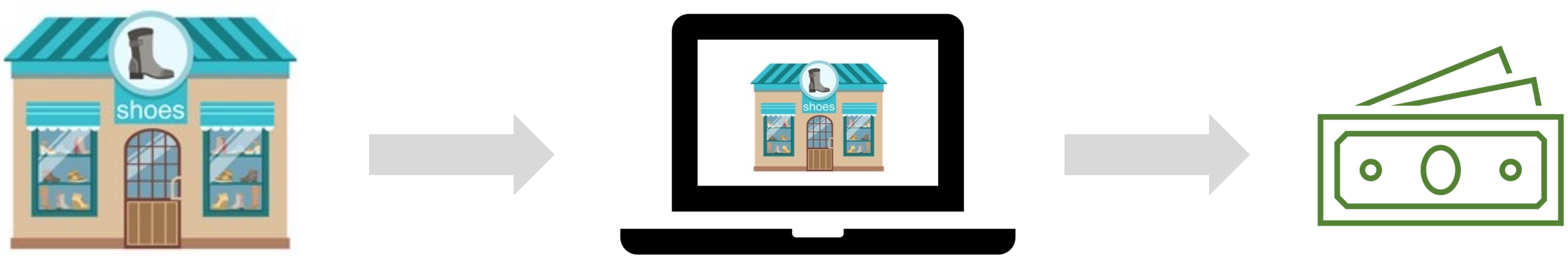
Collect

Clean

Analyze

Interpret

Data Science Process – Example

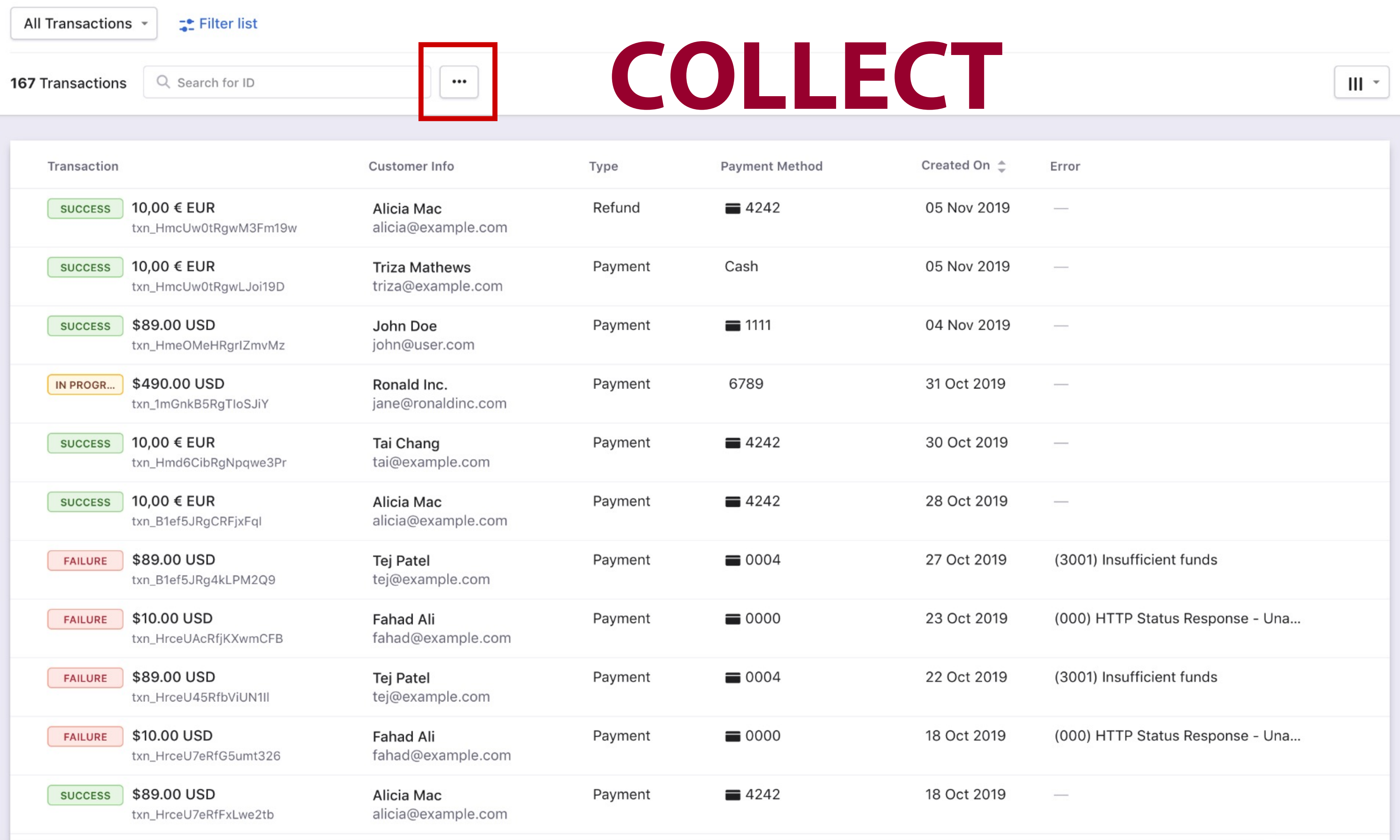


Example: Data Science Process

ASK

What are our customers' habits?

Example: Data Science Process



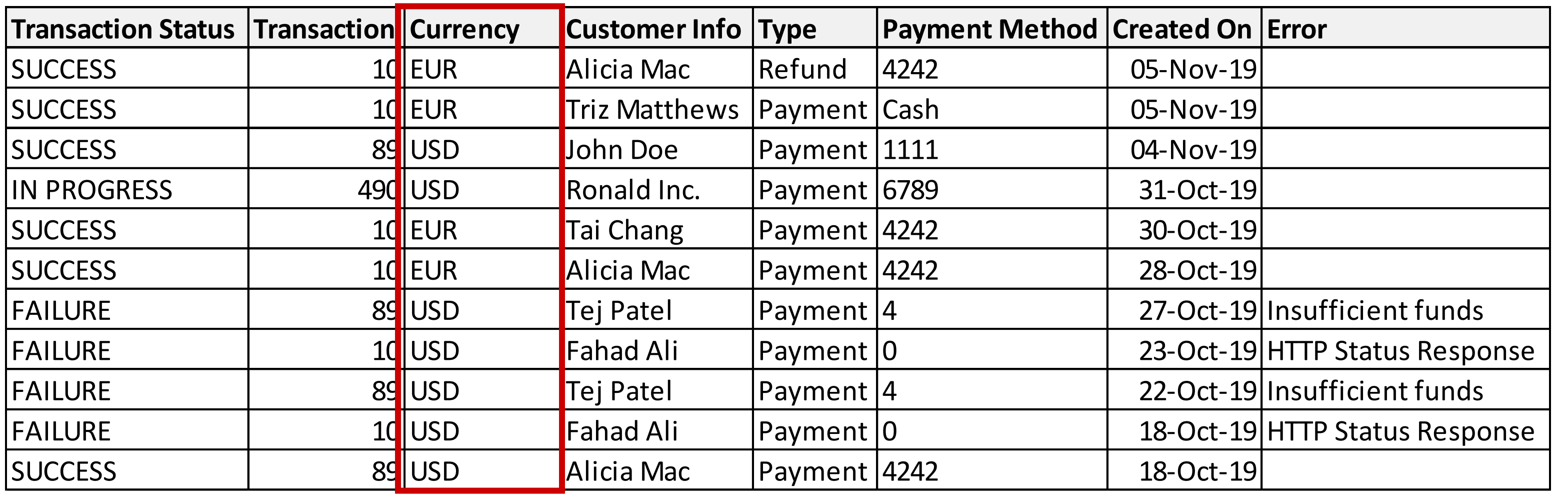
COLLECT

167 Transactions

Transaction	Customer Info	Type	Payment Method	Created On	Error
SUCCESS 10,00 € EUR txn_HmcUw0tRgwM3Fm19w	Alicia Mac alicia@example.com	Refund	▬ 4242	05 Nov 2019	—
SUCCESS 10,00 € EUR txn_HmcUw0tRgwLJoi19D	Triza Mathews triza@example.com	Payment	Cash	05 Nov 2019	—
SUCCESS $89.00 USD txn_HmeOMeHRgrIZmvMz	John Doe john@user.com	Payment	▬ 1111	04 Nov 2019	—
IN PROGR... $490.00 USD txn_1mGnkB5RgTIoSJiY	Ronald Inc. jane@ronaldinc.com	Payment	6789	31 Oct 2019	—
SUCCESS 10,00 € EUR txn_Hmd6CibRgNpqwe3Pr	Tai Chang tai@example.com	Payment	▬ 4242	30 Oct 2019	—
SUCCESS 10,00 € EUR txn_B1ef5JRgCRFjxFql	Alicia Mac alicia@example.com	Payment	▬ 4242	28 Oct 2019	—
FAILURE $89.00 USD txn_B1ef5JRg4kLPM2Q9	Tej Patel tej@example.com	Payment	▬ 0004	27 Oct 2019	(3001) Insufficient funds
FAILURE $10.00 USD txn_HrceUAcRfjKXwmCFB	Fahad Ali fahad@example.com	Payment	▬ 0000	23 Oct 2019	(000) HTTP Status Response - Una...
FAILURE $89.00 USD txn_HrceU45RfbViUN1Il	Tej Patel tej@example.com	Payment	▬ 0004	22 Oct 2019	(3001) Insufficient funds
FAILURE $10.00 USD txn_HrceU7eRfG5umt326	Fahad Ali fahad@example.com	Payment	▬ 0000	18 Oct 2019	(000) HTTP Status Response - Una...
SUCCESS $89.00 USD txn_HrceU7eRfFxLwe2tb	Alicia Mac alicia@example.com	Payment	▬ 4242	18 Oct 2019	—

Example: Data Science Process

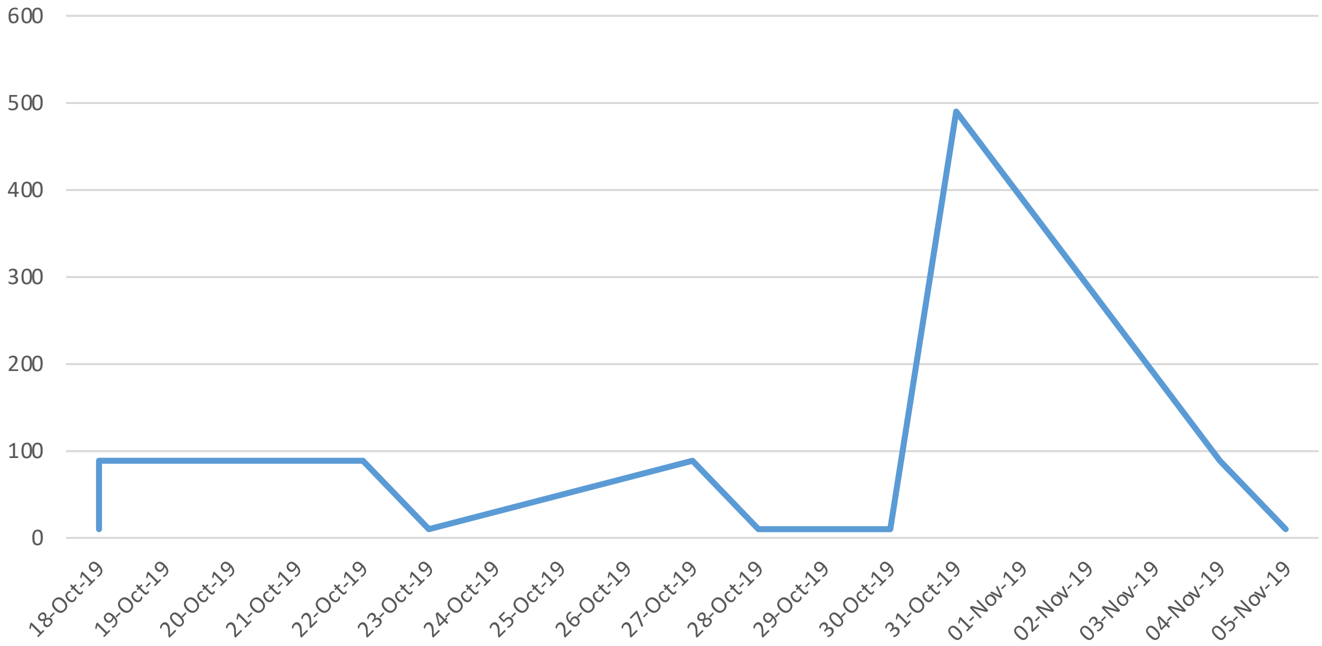
CLEAN



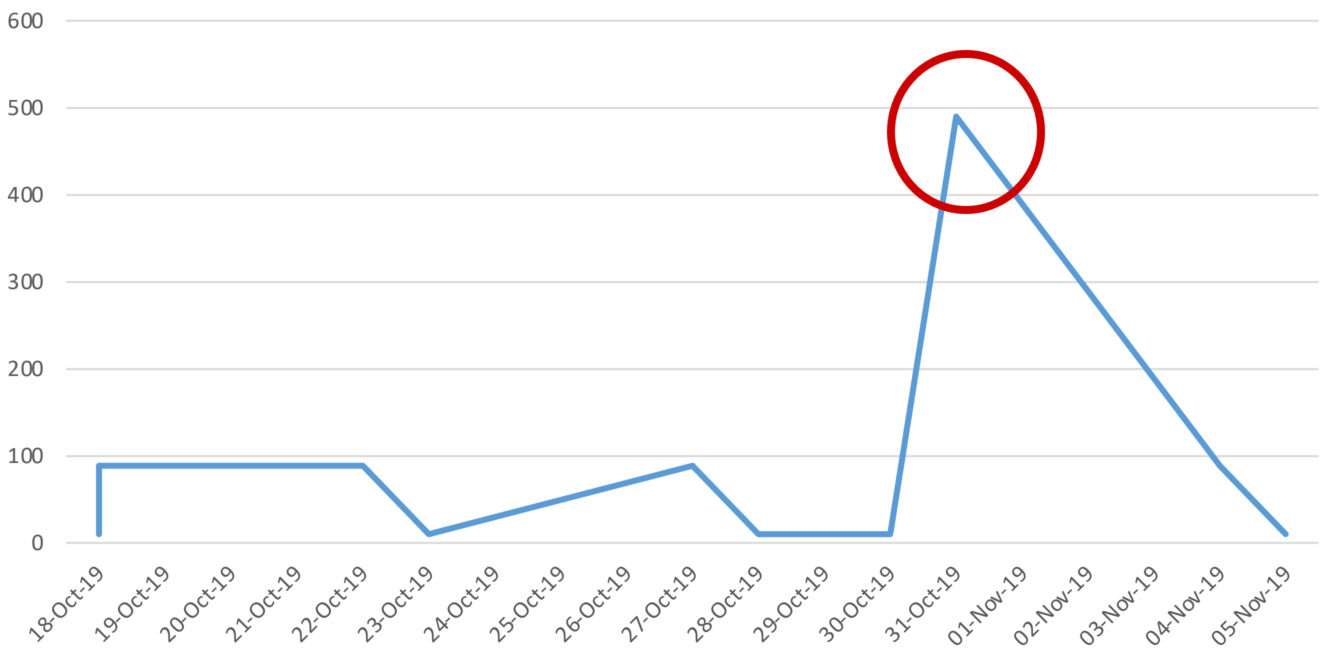
Transaction Status	Transaction	Currency	Customer Info	Type	Payment Method	Created On	Error
SUCCESS	10	EUR	Alicia Mac	Refund	4242	05-Nov-19	
SUCCESS	10	EUR	Triz Matthews	Payment	Cash	05-Nov-19	
SUCCESS	89	USD	John Doe	Payment	1111	04-Nov-19	
IN PROGRESS	490	USD	Ronald Inc.	Payment	6789	31-Oct-19	
SUCCESS	10	EUR	Tai Chang	Payment	4242	30-Oct-19	
SUCCESS	10	EUR	Alicia Mac	Payment	4242	28-Oct-19	
FAILURE	89	USD	Tej Patel	Payment	4	27-Oct-19	Insufficient funds
FAILURE	10	USD	Fahad Ali	Payment	0	23-Oct-19	HTTP Status Response
FAILURE	89	USD	Tej Patel	Payment	4	22-Oct-19	Insufficient funds
FAILURE	10	USD	Fahad Ali	Payment	0	18-Oct-19	HTTP Status Response
SUCCESS	89	USD	Alicia Mac	Payment	4242	18-Oct-19	

Example: Data Science Process

ANALYZE



Example: Data Science Process



INTERPRET

Representations

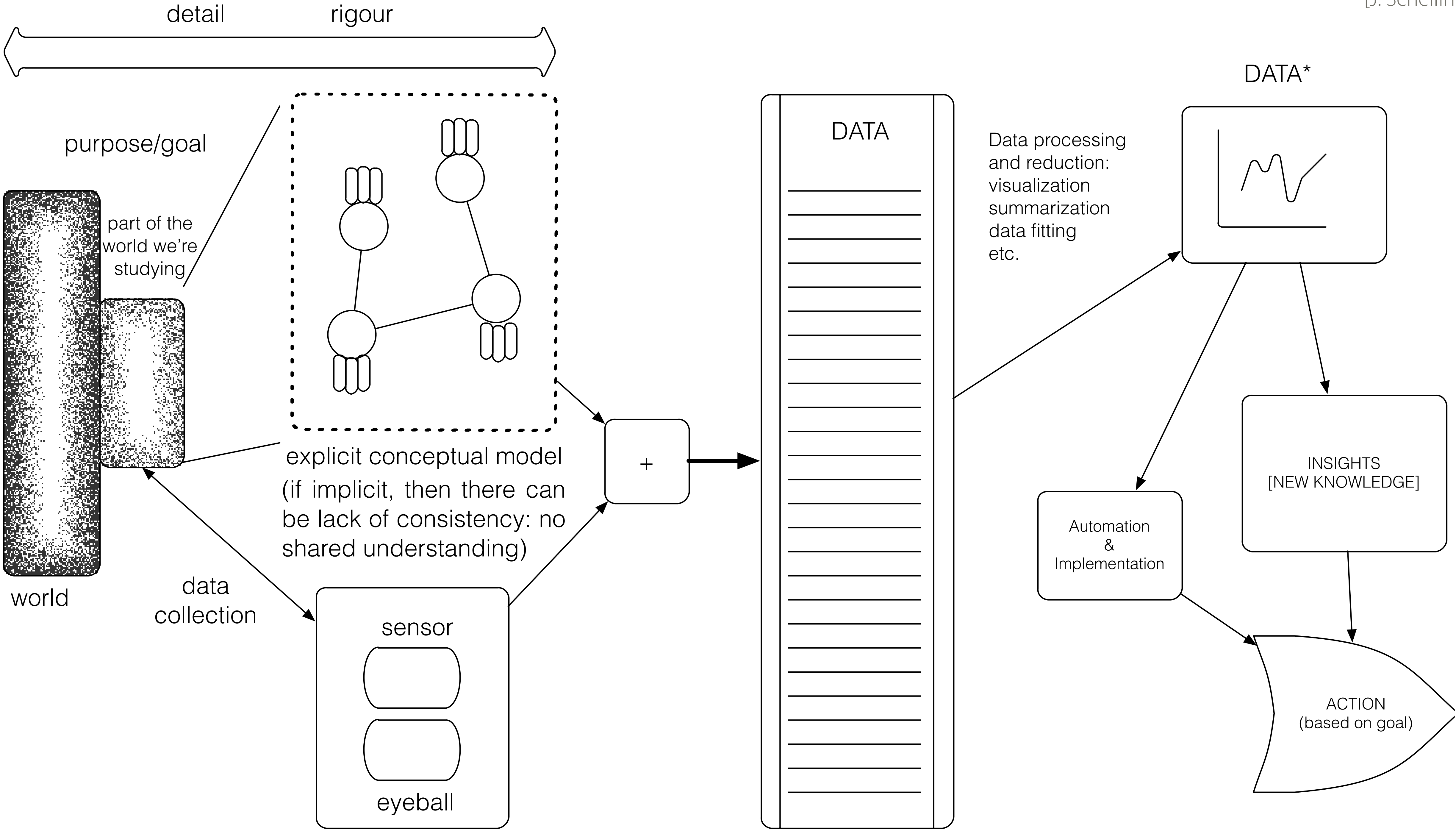
A **representation** is an object that stands in for another object.

A representation may or may not physically resemble the object it represents.

Representations of the world help us to **understand**, **navigate**, and **manipulate** the world.

Ceci n'est pas une pipe.

[J. Schellinck]



detail

rigour

purpose/goal

part of the world we're studying

world

explicit conceptual model (if implicit, then there can be lack of consistency: no shared understanding)

data collection

sensor

eyeball

+

DATA

Data processing and reduction:
visualization
summarization
data fitting
etc.

DATA*

Automation & Implementation

INSIGHTS [NEW KNOWLEDGE]

ACTION (based on goal)

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

Real World

Theory

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

Model

Systemic Thinking Take-Aways

Systems can approximate certain aspects of the Universe.

System models provide the basis under which data is identified and collected, but data itself is **approximate** and **selective**.

Knowledge gaps happen – be ready to re-visit your set-up regularly.

Implicit conceptual modeling can lead to problematic situations.

If the data, the system, and the world are **out of alignment**, data analysis insights might ultimately prove useless.

What are Ethics?

“Ethics” refers to the **study** and **definition** of **right** and **wrong conducts:**

- “not […] social convention, religious beliefs, or laws”. (R.W. Paul, L. Elder)

Influential ethical theories:

- Kant's **golden rule** (do onto others…), **consequentialism** (the ends justify the means), **utilitarianism** (act in order to maximize positive effect), etc.
- **Confucianism**, **Taoism**, **Buddhism** (?), etc.
- **Ubuntu**, **Maori**, **OCAP**, etc.

Discussion: What harm can come from data?

[University of Virginia, Centre for Big Data Ethics, Law, and Policy]

Ethics in the Data Context

Data ethics questions:

- **Who**, if anyone, owns data?
- Are there **limits** to how data can be used?
- Are there **value-biases** built into certain analytics?
- Are there categories that should **not** be used in analyzing personal data?
- Should some data be **publicly available** to **all** researchers?

Analytically, the **general** is preferred to the anecdotal – decisions made based on machine learning and A.I. (security, financial, marketing, etc.) may affect real beings in **unpredictable ways**.

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

Gapminder Exercises

We will conduct the exercises using Gapminder Tools.

The online version is available at https://www.gapminder.org/tools/
[there is also an offline version].

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.

Do the exercises for Module 1.

Module 2

Data Collection and Data Management

What is Data?

4,529 ‘red’ 25.782 ‘Y’

Objects and Attributes

Object: apple

Shape: spherical

Colour: red

Function: food

Location: fridge

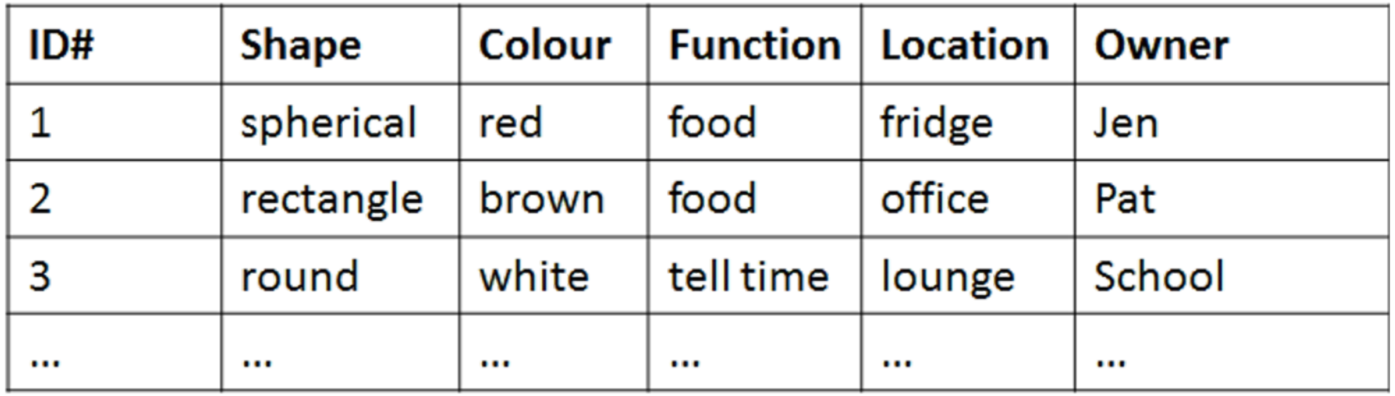
Owner: Jen

A person or an object is **not simply the sum of its attributes**!

From Attributes to Datasets

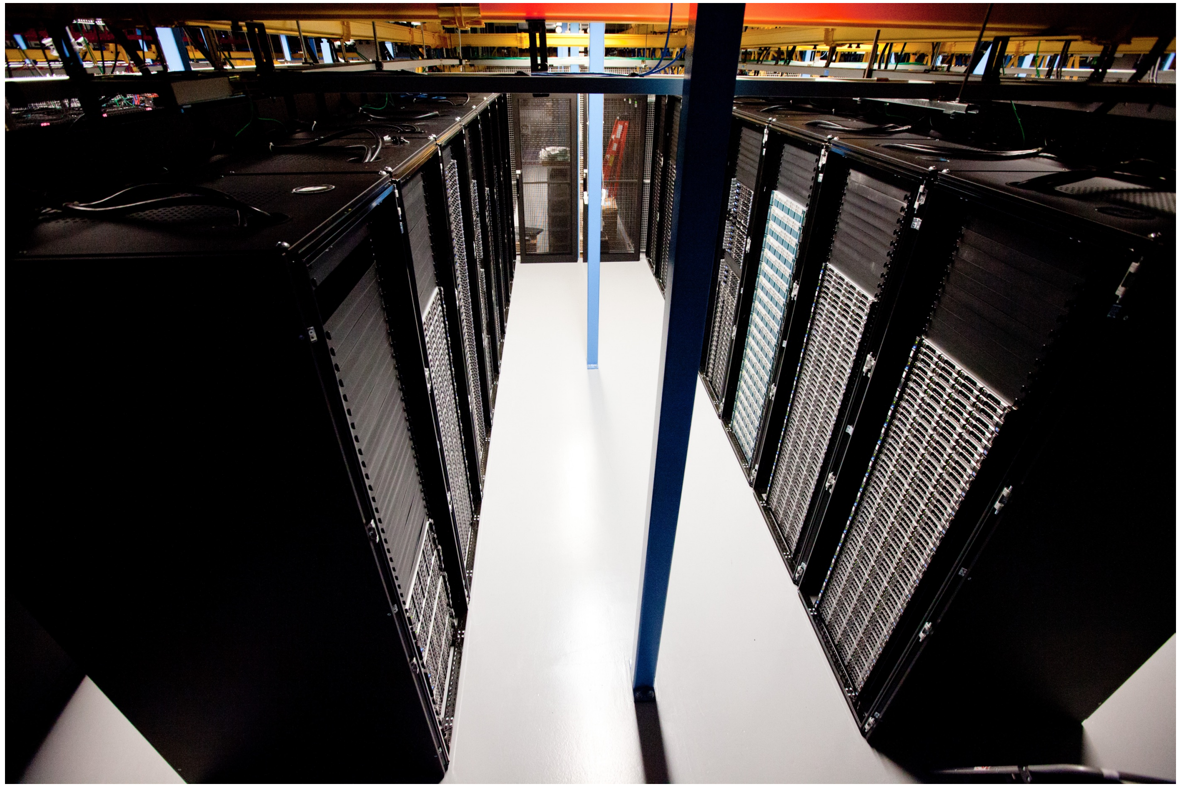
Attributes are **fields** (columns) in a database; objects are **instances** (rows).

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



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
…	…	…	…	…	…

Data is Real



Data is a representation, but data is still **physical**.

It has physical properties.

Physical space and energy are required to process and work with it.

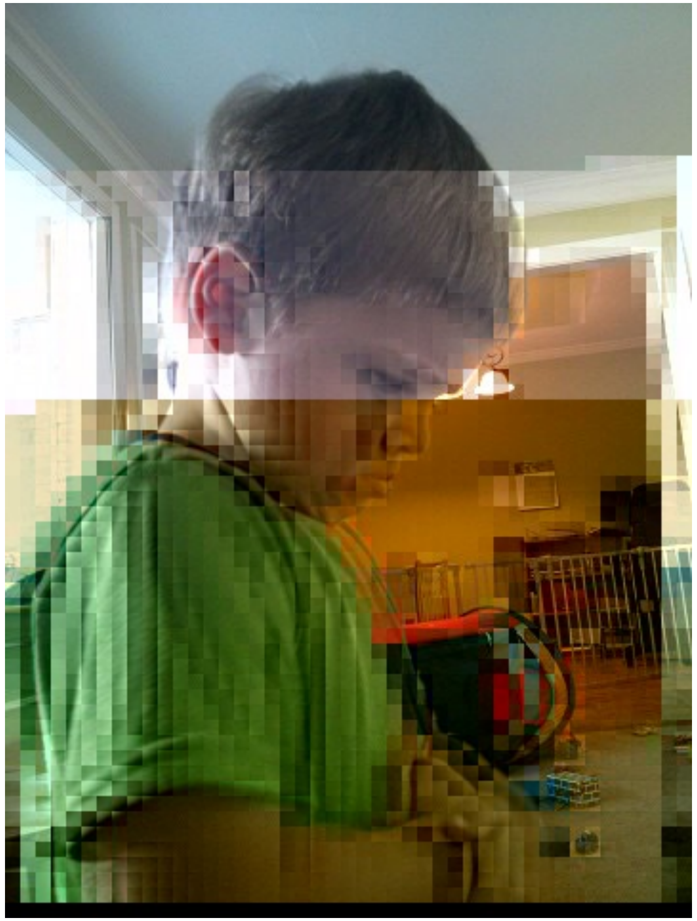
Data Decay

Data ages over time – it has a **shelf life**.

We use the phrase “rotten data” or “decaying data”

- **literally** – the data storage medium might decay
- **metaphorically** – when the data no longer accurately represents the relevant objects and relationships or even when those objects no longer exist in the same way

Data must be kept ‘fresh’ and ‘current’, not ‘stale’ (context and model dependent!)



[G. Smith, The Exaggerated Promise of So-Called Unbiased Data Mining]

"A Dartmouth graduate student used an MRI machine to study the brain activity of a salmon as it was shown photographs and asked questions. The most interesting thing about the study was not that a salmon was studied, but that **the salmon was dead**. Yep, a dead salmon purchased at a local market was put into the MRI machine, and some patterns were discovered. There were inevitably patterns—and they were invariably meaningless."

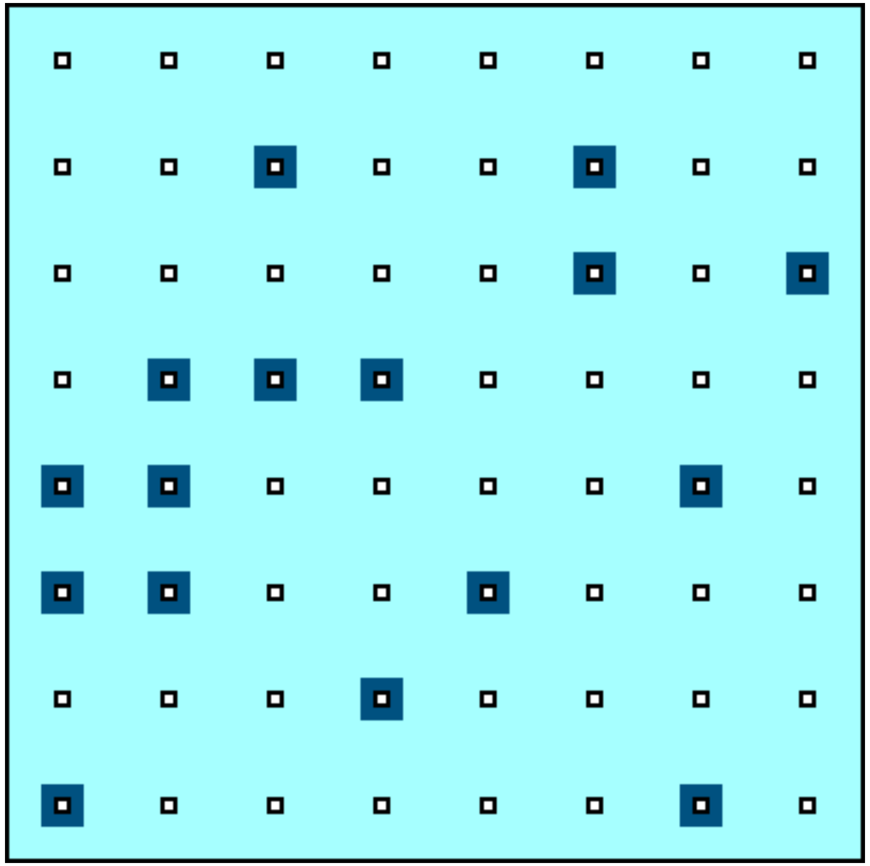
What’s a Sample?

A **sample** is a portion of a ‘population’ from which the data is collected

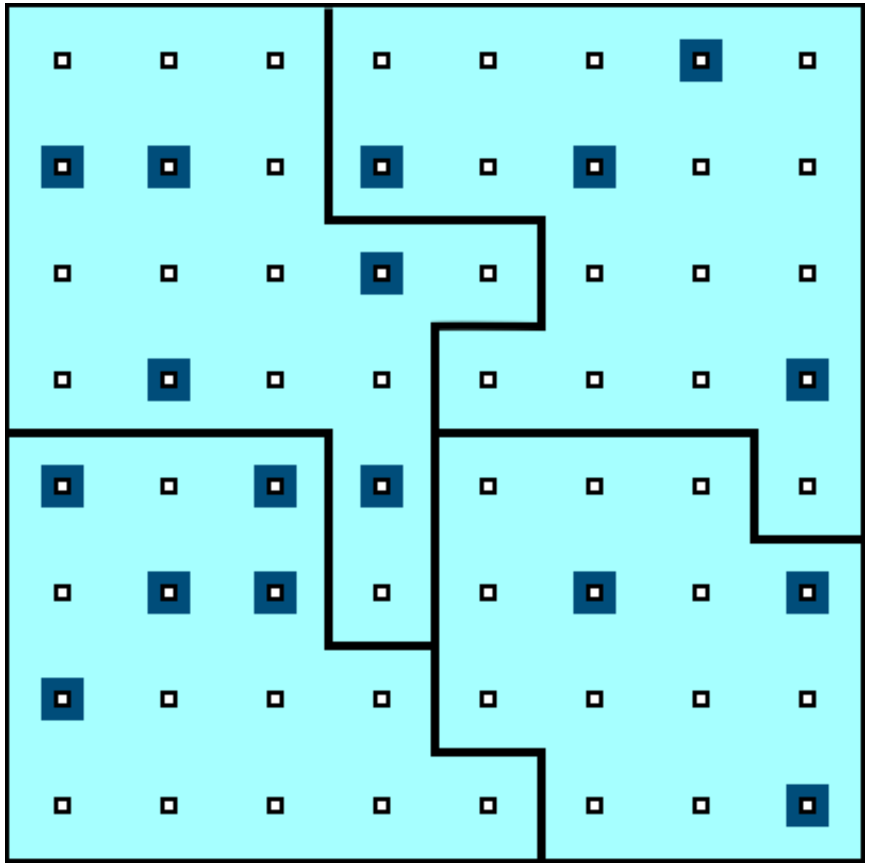
Biased	Unbiased
One or more parts of the population are favoured over others	Everyone has an equal change of being chosen
Does not accurately represent the population	Accurately represents the population
Leads to invalid conclusions	Provides a valid conclusion

Source: Making Big Data Work

Sampling Designs

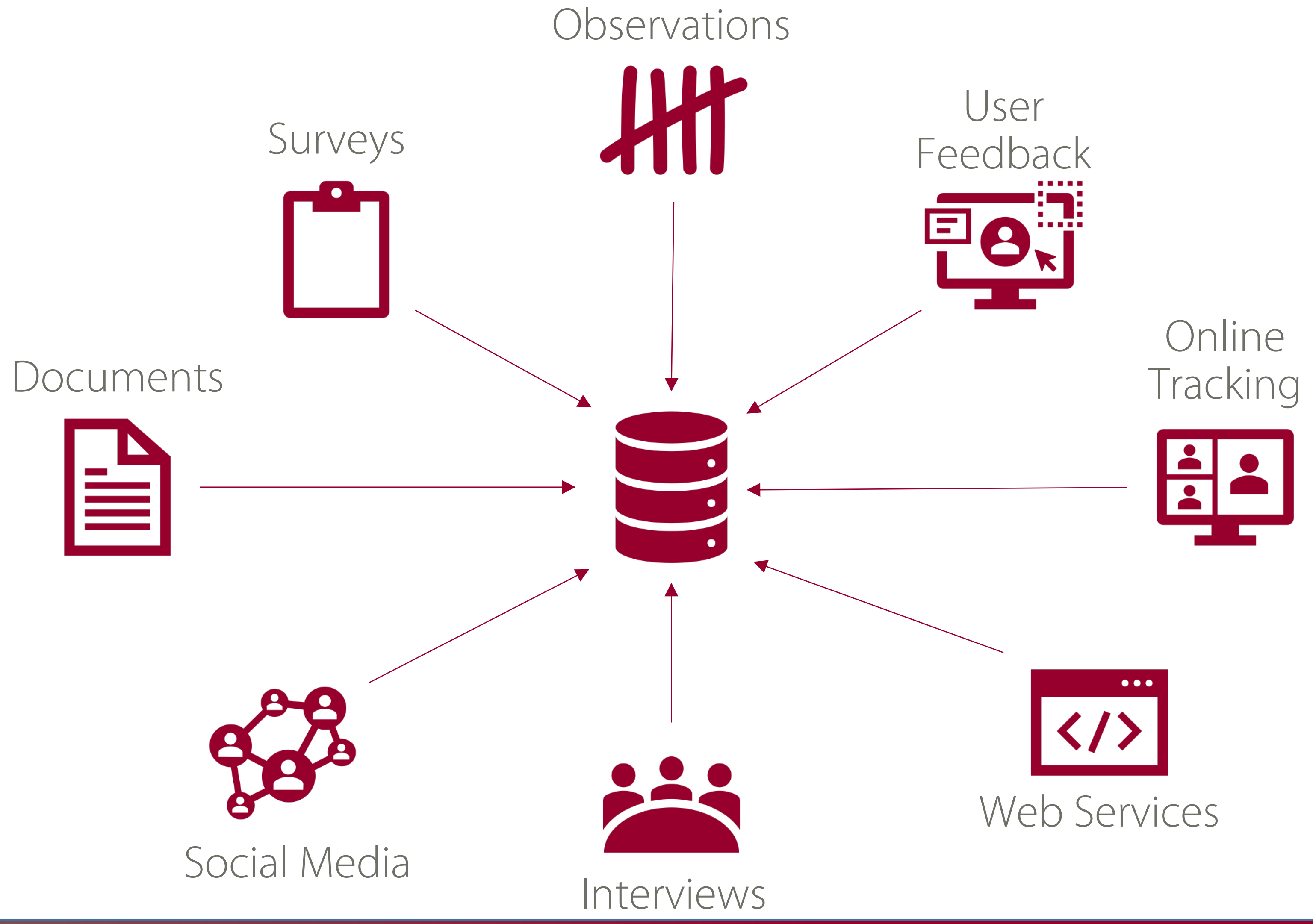


Simple Random Sampling (SRS)



Stratified Random Sampling (STS)

Collect/Create Data



Web Scraping – Example

Let’s say you want to know what people think of a new phone.

Standard approach: market research (e.g. telephone survey, reward system, etc.).

Pitfalls:

- unrepresentative sample: the selected sample might not represent the intended population
- systematic non-response: people who don’t like phone surveys might be less (or more) likely to dislike the new phone
- coverage error: people without a landline can’t be reached, say
- measurement error: are the survey questions providing suitable info for the problem at hand?

Web Scraping – Example

These solutions can be **costly**, **time-consuming**, **ineffective**.

Proxies are indicators that are strongly related to the information of interest, without measuring it directly.

If **popularity** is defined as large groups of people preferring one product over a competitor, then sales statistics on a commercial website may provide a proxy for popularity.

Rankings on Amazon could provide a **more comprehensive** view of the phone market than a traditional survey.

Web Scraping – Example

Representativeness of the **listed products**

- are all phones listed?
- if not, is it because that website doesn't sell them?
- is there some other reason?

Representativeness of the **customers**

- are there specific groups buying/not-buying online products?
- are there specific groups buying from specific sites?
- are there specific groups leaving/not-leaving reviews?

Truthfulness of customers and **reliability** of reviews.

Scraping Dos and Don’ts

1. Stay identifiable

2. Reduce traffic

3. Do not bother server with multiple requests

4. Write modest scrapers (efficient and polite)

Use **application programming interface** (APIs) as much as possible!

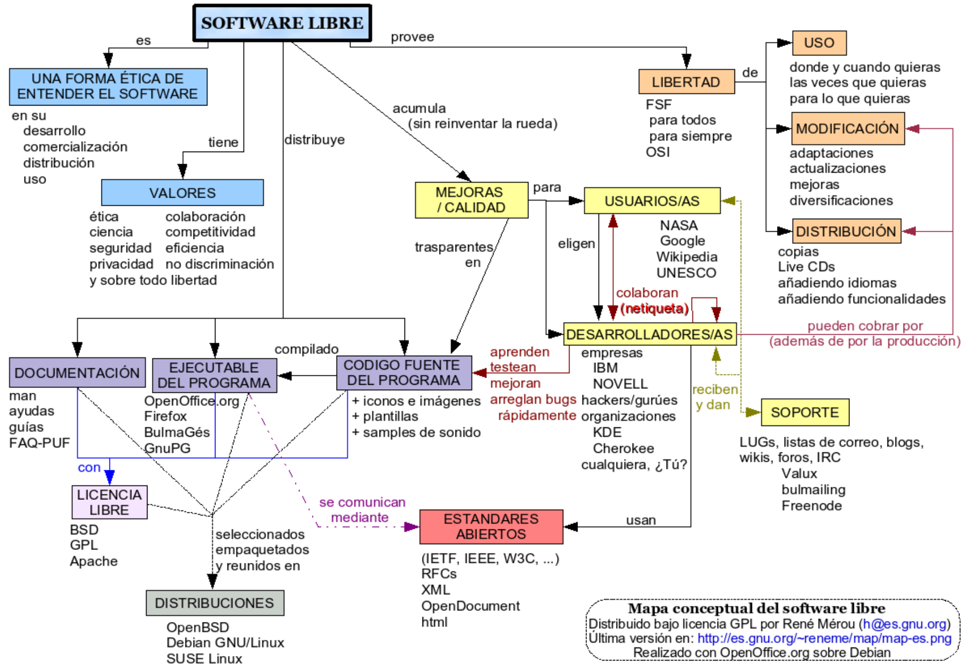
Conceptual Model

A **conceptual model** is, roughly speaking:

- a model that is not implemented, which exists only conceptually
- a diagram or verbal description of a system (e.g. boxes and arrows, mind maps, lists, definitions)

Focus is :

- not on capturing specific behaviors but emphasizing **possible states**
- on object types, not on specific instances; the goal is **abstraction.**



Mapa conceptual del software libre
Distribuido bajo licencia GPL por René Mérou (h@es.gnu.org)
Última versión en: http://es.gnu.org/~reneme/map/map-es.png
Realizado con OpenOffice.org sobre Debian

Fundamental Concepts

It is important to structure **data** and **knowledge** so that it can be:

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

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

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

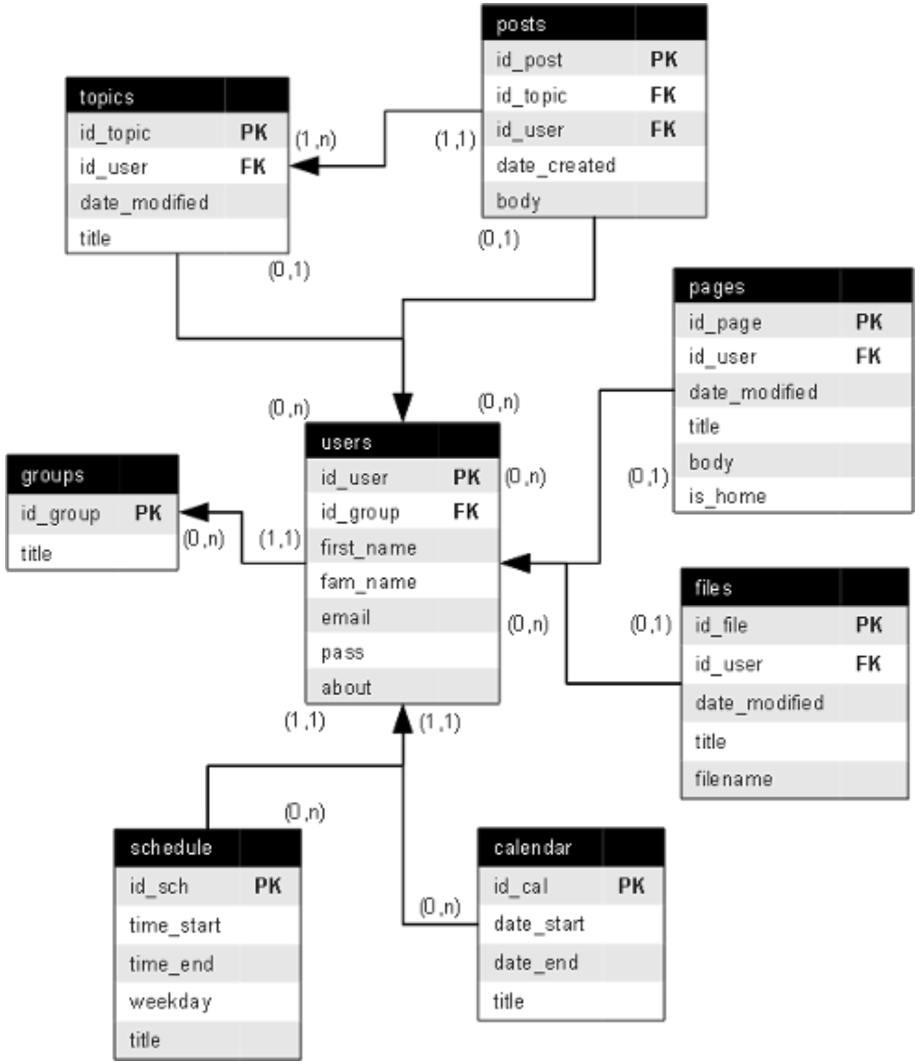
Data Modeling

Data models are **abstract/logical** descriptions of a system, constructed in terms that can then be implemented 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**.



posts	
id_post	PK
id_topic	FK
id_user	FK
date_created	
body	

topics	
id_topic	PK
id_user	FK
date_modified	
title	

pages	
id_page	PK
id_user	FK
date_modified	
title	
body	
is_home	

groups	
id_group	PK
title	

users	
id_user	PK
id_group	FK
first_name	
fam_name	
email	
pass	
about	

files	
id_file	PK
id_user	FK
date_modified	
title	
filename	

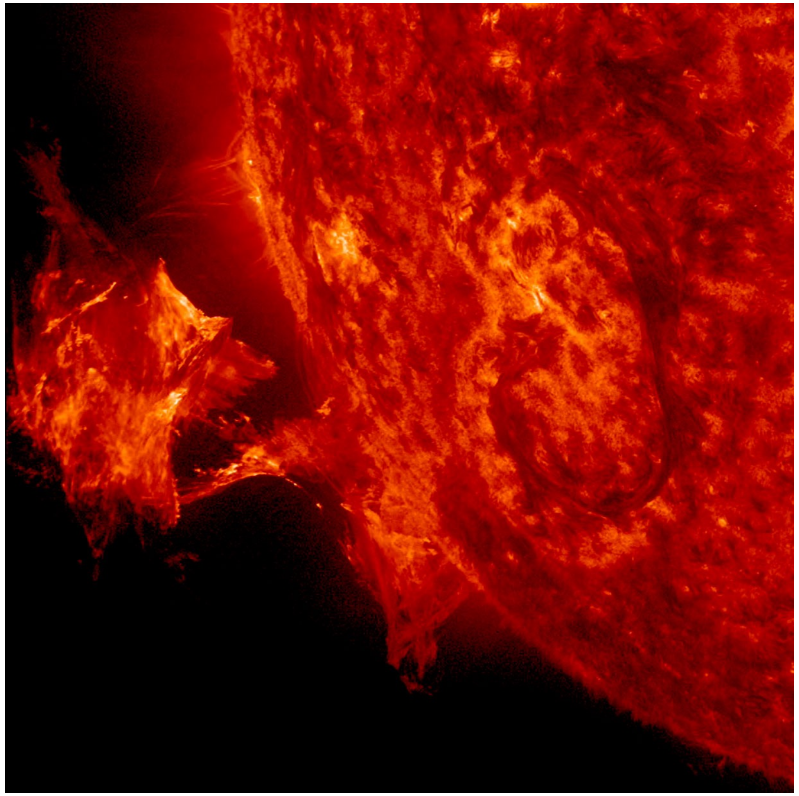
schedule	
id_sch	PK
time_start	
time_end	
weekday	
title	

calendar	
id_cal	PK
date_start	
date_end	
title	

Structured/Unstructured Data

A major motivator for new developments in database types and 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 **O**bject (BLOb) – images, audio, multi-media



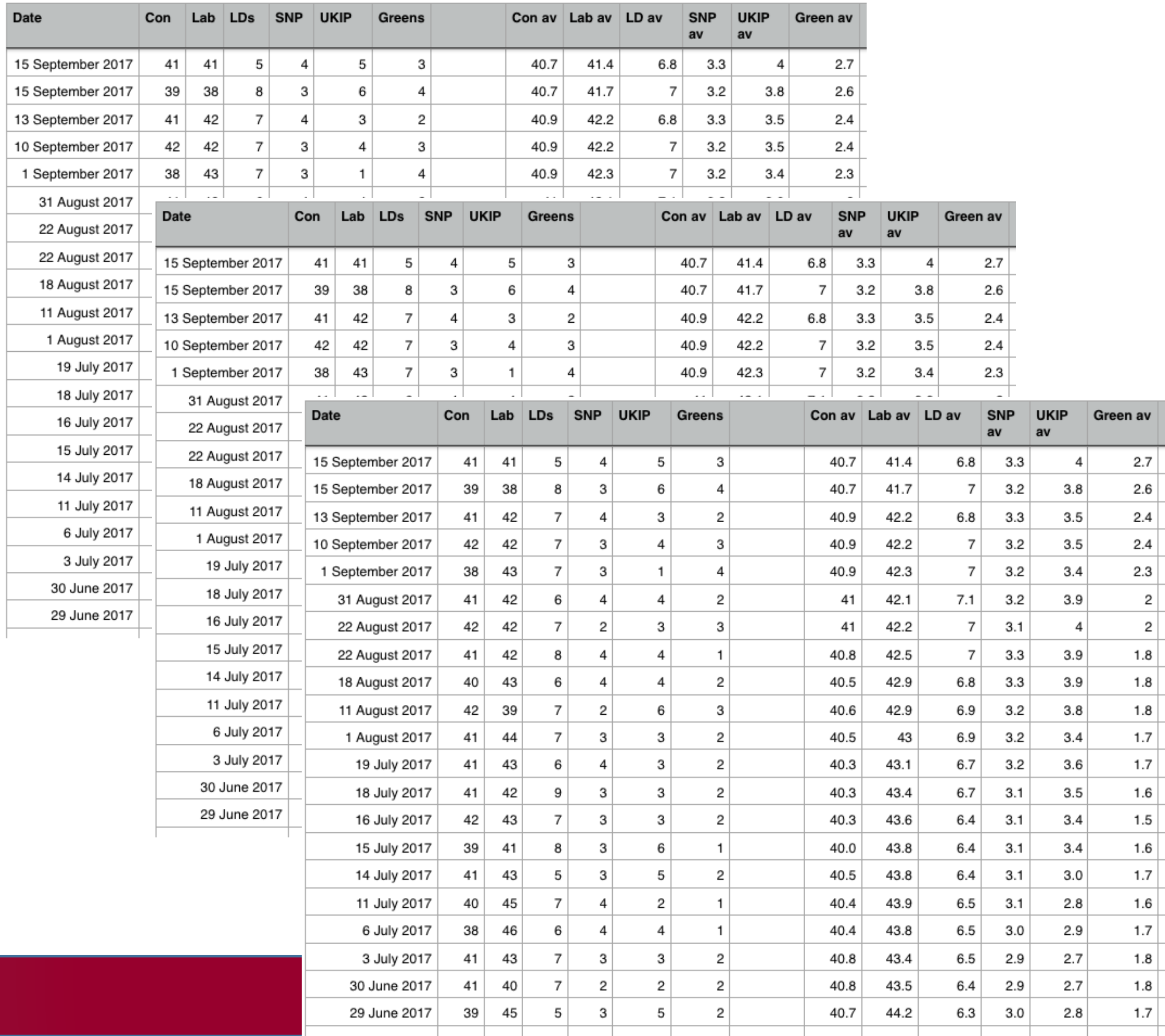
Flat Files and Spreadsheets

What about keeping data in a single giant table (spreadsheet)?

Or multiple spreadsheets?

How bad can it be?

Wayne Eckerson coined the term ‘spreadmart’ to describe a situation with many (ad hoc) spreadsheets as a data strategy.



Date	Con	Lab	LDs	SNP	UKIP	Greens		Con av	Lab av	LD av	SNP av	UKIP av	Green av
15 September 2017	41	41	5	4	5	3		40.7	41.4	6.8	3.3	4	2.7
15 September 2017	39	38	8	3	6	4		40.7	41.7	7	3.2	3.8	2.6
13 September 2017	41	42	7	4	3	2		40.9	42.2	6.8	3.3	3.5	2.4
10 September 2017	42	42	7	3	4	3		40.9	42.2	7	3.2	3.5	2.4
1 September 2017	38	43	7	3	1	4		40.9	42.3	7	3.2	3.4	2.3
31 August 2017													
22 August 2017													
22 August 2017													
18 August 2017													
11 August 2017													
1 August 2017													
19 July 2017													
18 July 2017													
16 July 2017													
15 July 2017													
14 July 2017													
11 July 2017													
6 July 2017													
3 July 2017													
30 June 2017													
29 June 2017													

Date	Con	Lab	LDs	SNP	UKIP	Greens		Con av	Lab av	LD av	SNP av	UKIP av	Green av
15 September 2017	41	41	5	4	5	3		40.7	41.4	6.8	3.3	4	2.7
15 September 2017	39	38	8	3	6	4		40.7	41.7	7	3.2	3.8	2.6
13 September 2017	41	42	7	4	3	2		40.9	42.2	6.8	3.3	3.5	2.4
10 September 2017	42	42	7	3	4	3		40.9	42.2	7	3.2	3.5	2.4
1 September 2017	38	43	7	3	1	4		40.9	42.3	7	3.2	3.4	2.3
31 August 2017													
22 August 2017													
22 August 2017													
18 August 2017													
11 August 2017													
1 August 2017													
19 July 2017													
18 July 2017													
16 July 2017													
15 July 2017													
14 July 2017													
11 July 2017													
6 July 2017													
3 July 2017													
30 June 2017													
29 June 2017													

Date	Con	Lab	LDs	SNP	UKIP	Greens		Con av	Lab av	LD av	SNP av	UKIP av	Green av
15 September 2017	41	41	5	4	5	3		40.7	41.4	6.8	3.3	4	2.7
15 September 2017	39	38	8	3	6	4		40.7	41.7	7	3.2	3.8	2.6
13 September 2017	41	42	7	4	3	2		40.9	42.2	6.8	3.3	3.5	2.4
10 September 2017	42	42	7	3	4	3		40.9	42.2	7	3.2	3.5	2.4
1 September 2017	38	43	7	3	1	4		40.9	42.3	7	3.2	3.4	2.3
31 August 2017	41	42	6	4	4	2		41	42.1	7.1	3.2	3.9	2
22 August 2017	42	42	7	2	3	3		41	42.2	7	3.1	4	2
22 August 2017	41	42	8	4	4	1		40.8	42.5	7	3.3	3.9	1.8
18 August 2017	40	43	6	4	4	2		40.5	42.9	6.8	3.3	3.9	1.8
11 August 2017	42	39	7	2	6	3		40.6	42.9	6.9	3.2	3.8	1.8
1 August 2017	41	44	7	3	3	2		40.5	43	6.9	3.2	3.4	1.7
19 July 2017	41	43	6	4	3	2		40.3	43.1	6.7	3.2	3.6	1.7
18 July 2017	41	42	9	3	3	2		40.3	43.4	6.7	3.1	3.5	1.6
16 July 2017	42	43	7	3	3	2		40.3	43.6	6.4	3.1	3.4	1.5
15 July 2017	39	41	8	3	6	1		40.0	43.8	6.4	3.1	3.4	1.6
14 July 2017	41	43	5	3	5	2		40.5	43.8	6.4	3.1	3.0	1.7
11 July 2017	40	45	7	4	2	1		40.4	43.9	6.5	3.1	2.8	1.6
6 July 2017	38	46	6	4	4	1		40.4	43.8	6.5	3.0	2.9	1.7
3 July 2017	41	43	7	3	3	2		40.8	43.4	6.5	2.9	2.7	1.8
30 June 2017	41	40	7	2	2	2		40.8	43.5	6.4	2.9	2.7	1.8
29 June 2017	39	45	5	3	5	2		40.7	44.2	6.3	3.0	2.8	1.7

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!

Tools and Buzzwords

SQL, SQLite, MySQL, NoSQL

MongoDB, ArangoDB

Document store

JSON, YAML

API, GraphQL

Linked Data

Semantic Web

Ontology Web Language (OWL)

Protégé

etc.

Cloud vs. On-Premise

Cloud	On-Premise (On-Prem)
Hands-off	Self-Maintained
Pay-as-you-go Model	All Costs Absorbed
Questionable Data Ownership	Fully-Controlled Security

Roundtable: About Your Data

?

Does it exist?

Where does it live?

How is it structured and accessed?

Gapminder Exercises

Do the exercises for Module 2.

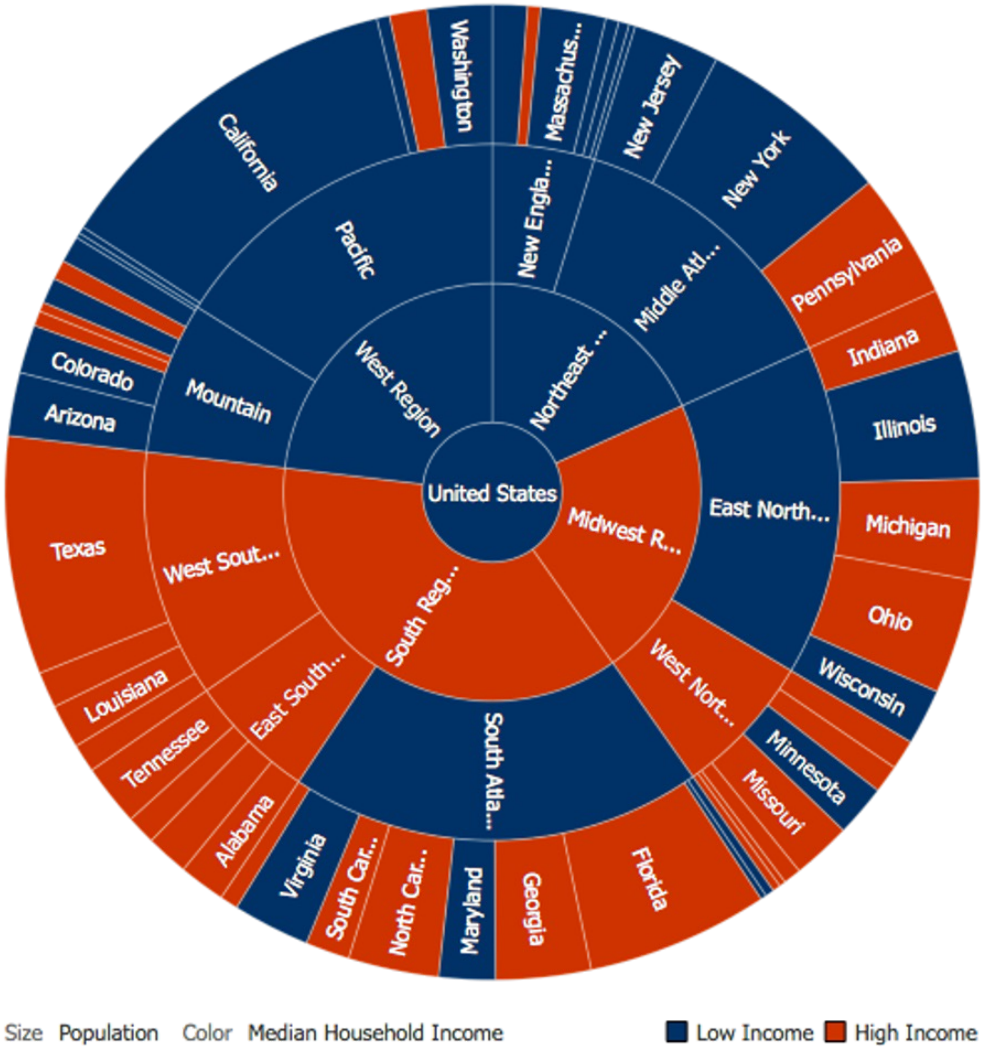
Module 3

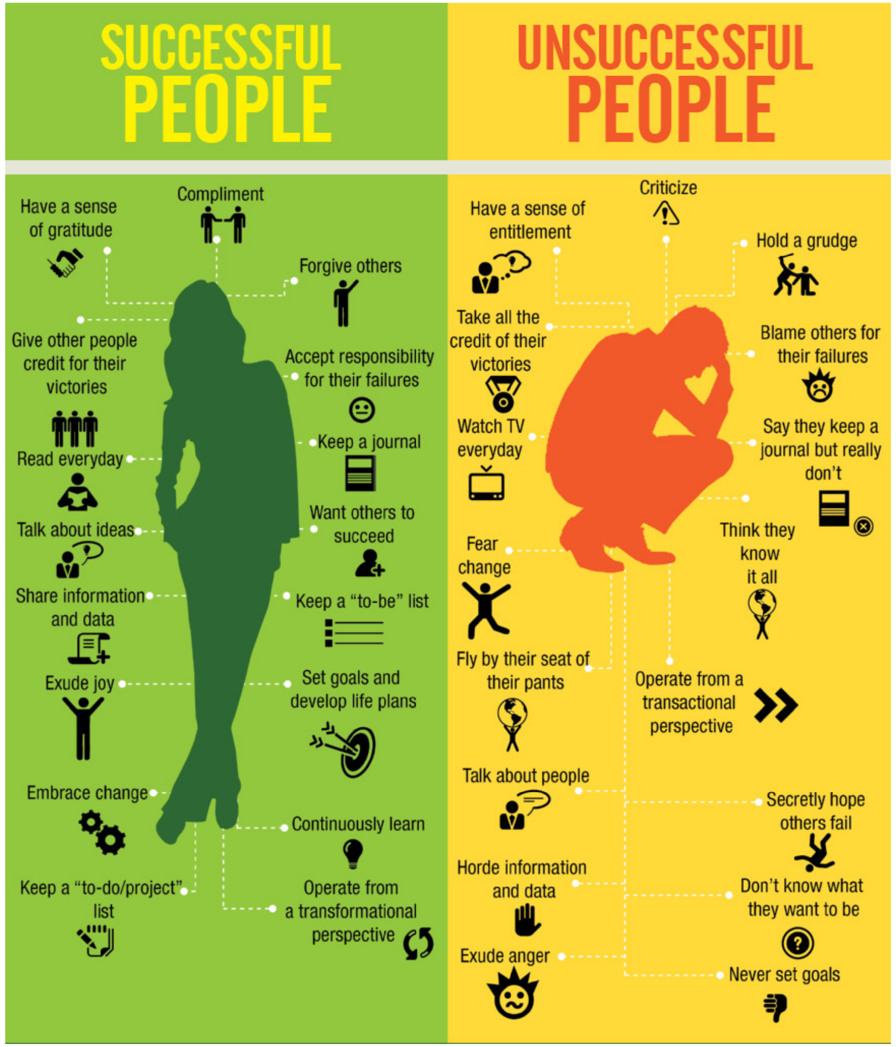
Data Visualization and Data Communication

[Author unknown]

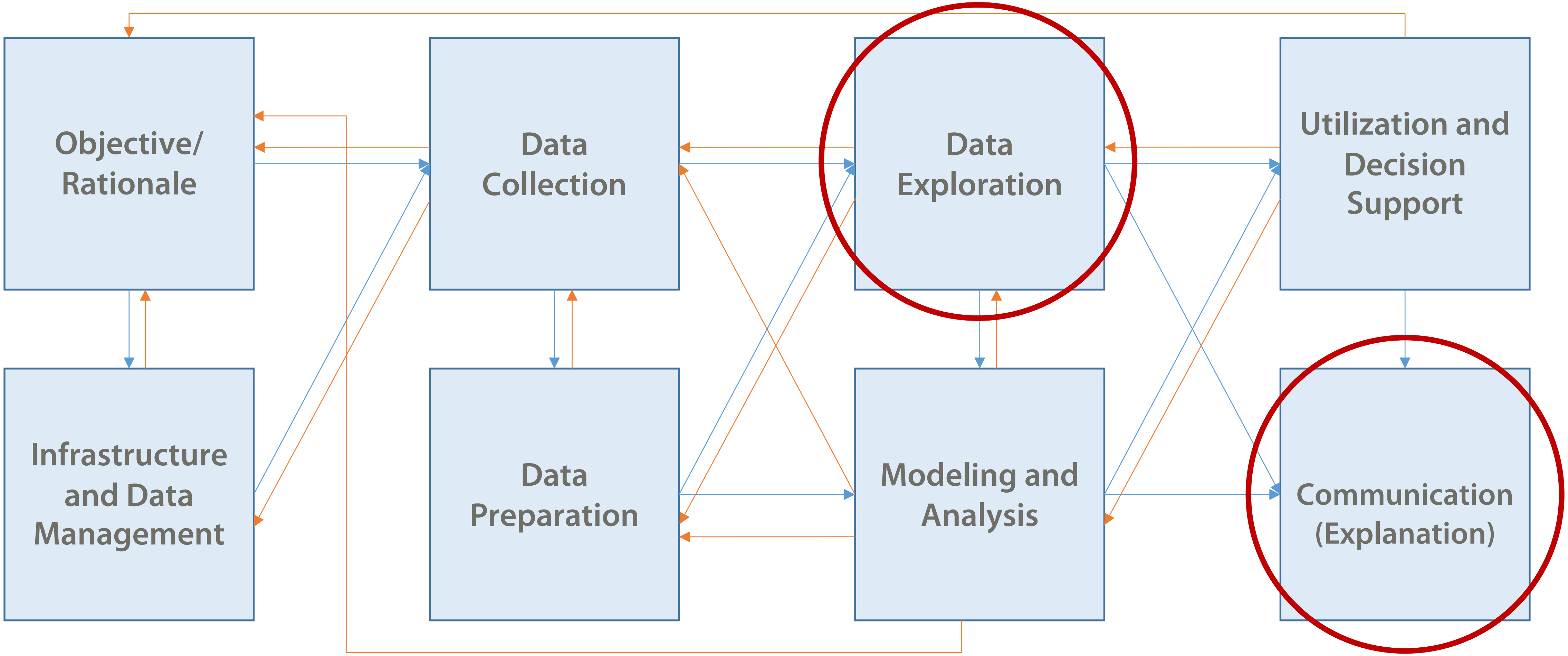
[J.P. Blackard]

Data Visualizations vs Infographics




SUCCESSFUL PEOPLE
Have a sense of gratitude
Compliment
Forgive others
Give other people credit for their victories
Accept responsibility for their failures
Read everyday
Keep a journal
Talk about ideas
Want others to succeed
Share information and data
Keep a "to-be" list
Exude joy
Set goals and develop life plans
Embrace change
Continuously learn
Keep a "to-do/project" list
Operate from a transformational perspective
UNSUCCESSFUL PEOPLE
Criticize
Have a sense of entitlement
Hold a grudge
Take all the credit of their victories
Blame others for their failures
Watch TV everyday
Say they keep a journal but really don't
Fear change
Think they know it all
Fly by their seat of their pants
Operate from a transactional perspective
Talk about people
Secretly hope others fail
Horde information and data
Don't know what they want to be
Exude anger
Never set goals

The (Messy) Analysis Process



Objective/ Rationale

Data Collection

Data Exploration

Utilization and Decision Support

Infrastructure and Data Management

Data Preparation

Modeling and Analysis

Communication (Explanation)

Pre-Analysis Uses

Data visualization can be used to set the stage for analysis:

- **detecting anomalous entries**
  invalid entries, missing values, outliers
- **shaping the data transformations**
  binning, standardization, Box-Cox transformations, PCA-like transformations
- **getting a sense for the data**
  data analysis as an art form, exploratory analysis
- **identifying hidden data structure**
  clustering, associations, patterns informing the next stage of analysis

[E. Tufte, Beautiful Evidence]

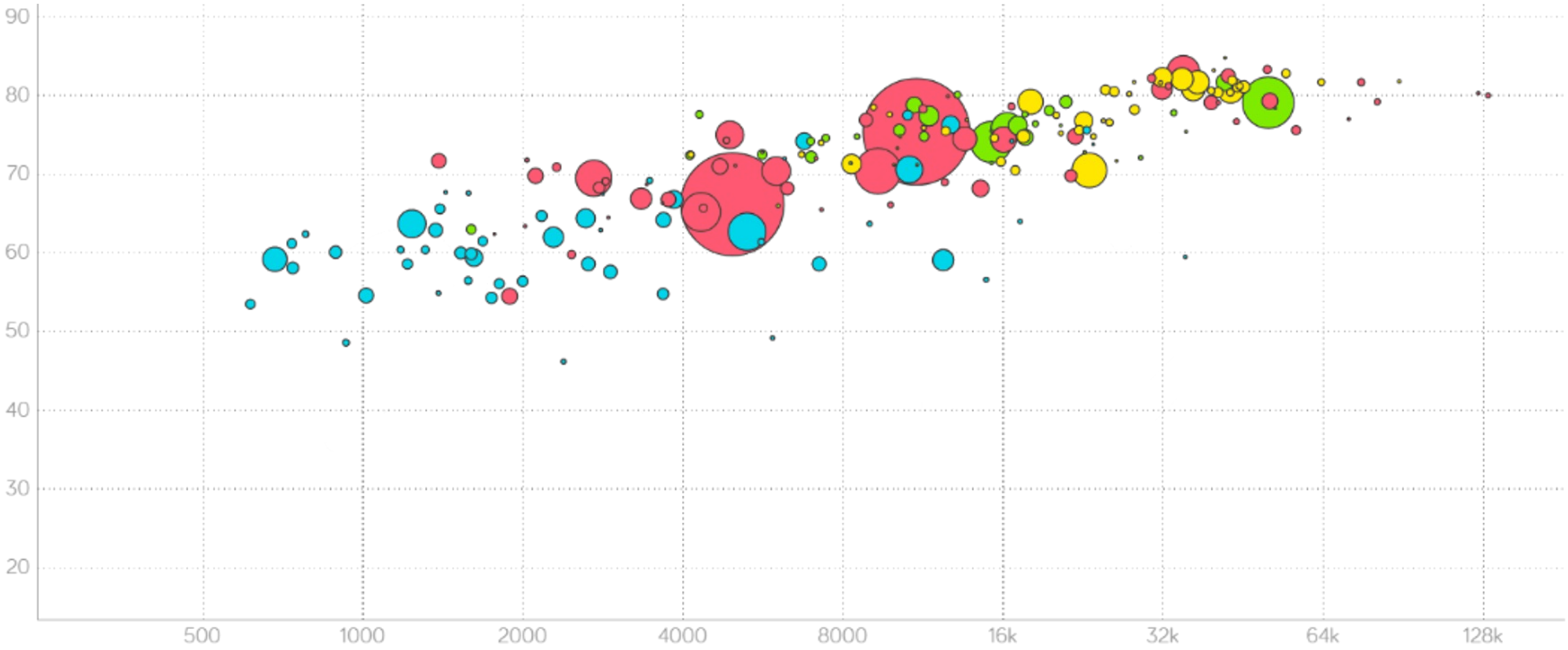
Fundamental Principles of Data Viz

There is a **symmetry** to visual displays of evidence. Consumers should be seeking exactly what producers should be providing, namely:

- meaningful comparisons
- potential causal networks and underlying structure
- multivariate links
- integrated and relevant data
- honest documentation
- primary focus on content

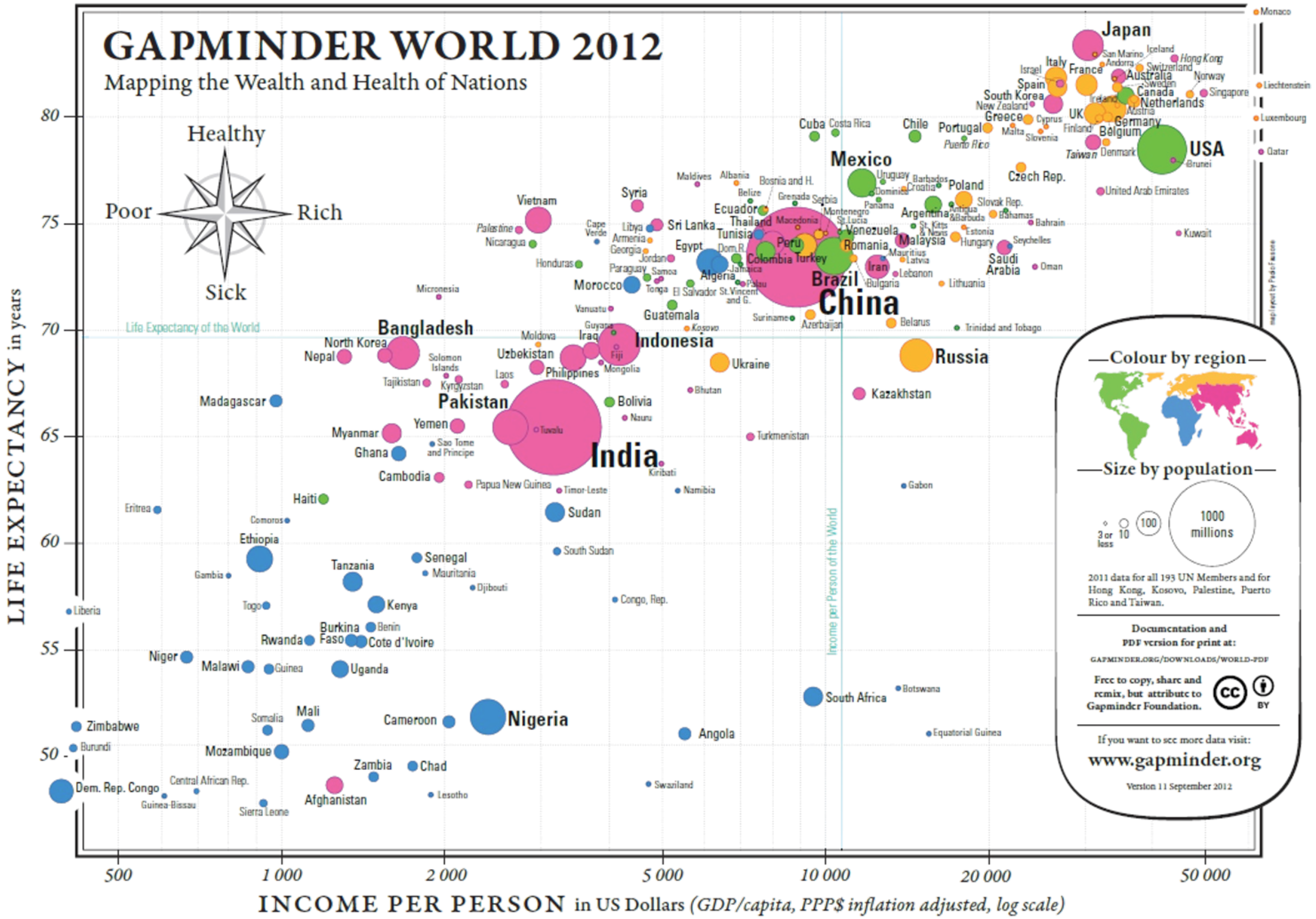
Graphics should be **clear** and **engaging**.

Don’t be afraid to try something new if it helps **convey the message**.

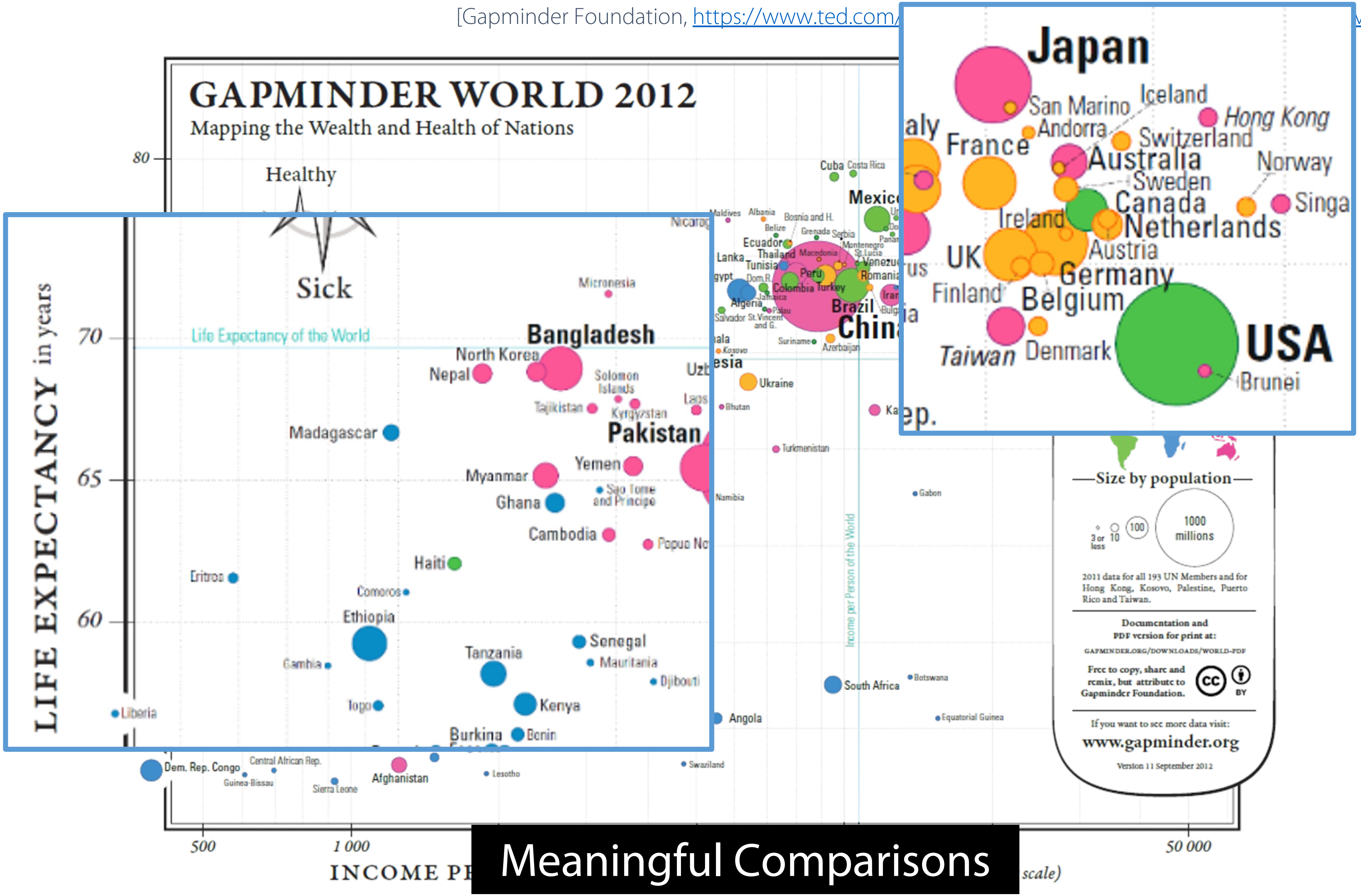


Non-Integrated Data

[Gapminder Foundation, https://www.ted.com/talks/hans_rosling_shows_the_best_stats_you_ve_ever_seen]

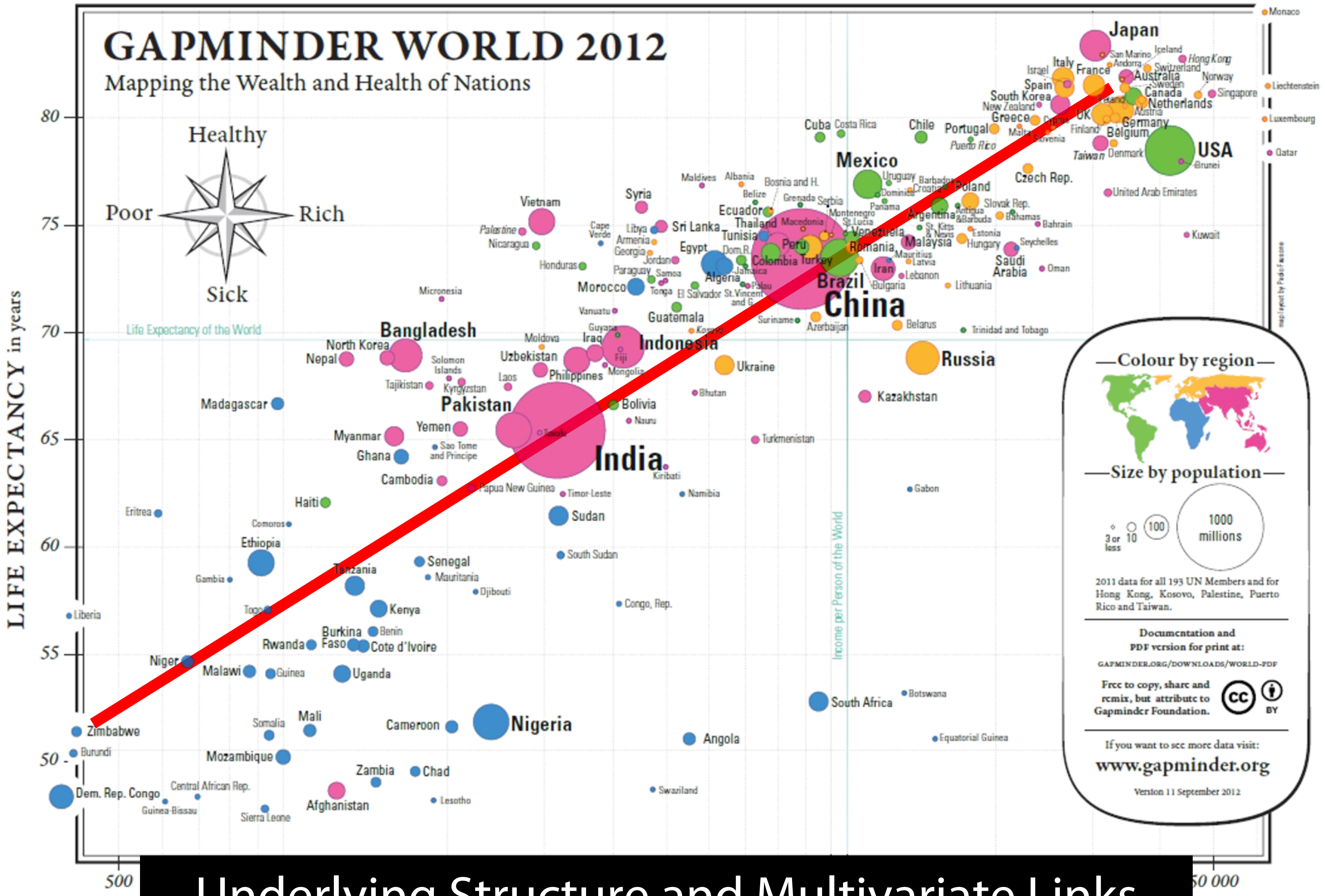


[Gapminder Foundation, https://www.ted.com/ ve_ever_seen]



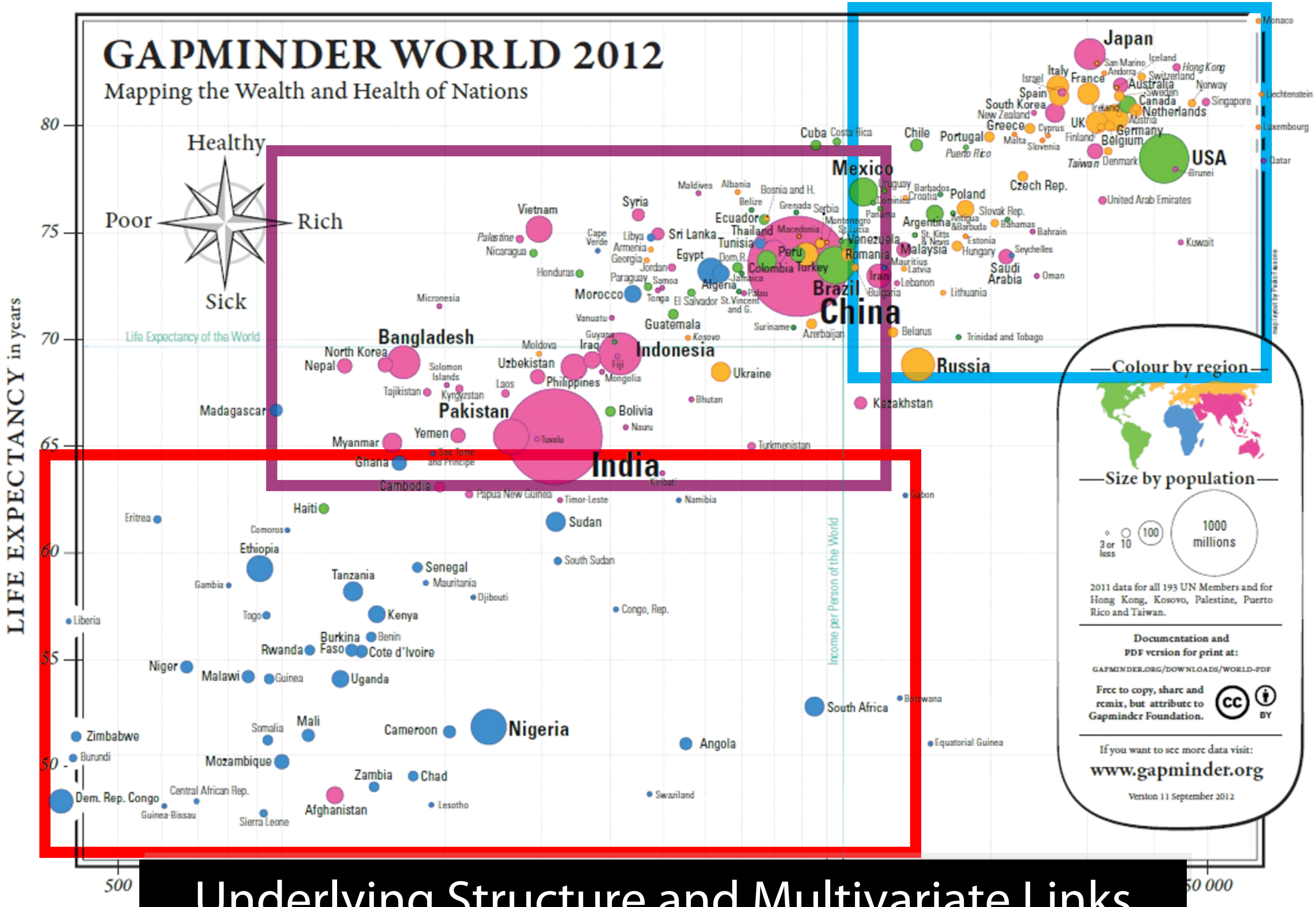
Meaningful Comparisons

[Gapminder Foundation, https://www.ted.com/talks/hans_rosling_shows_the_best_stats_you_ve_ever_seen]



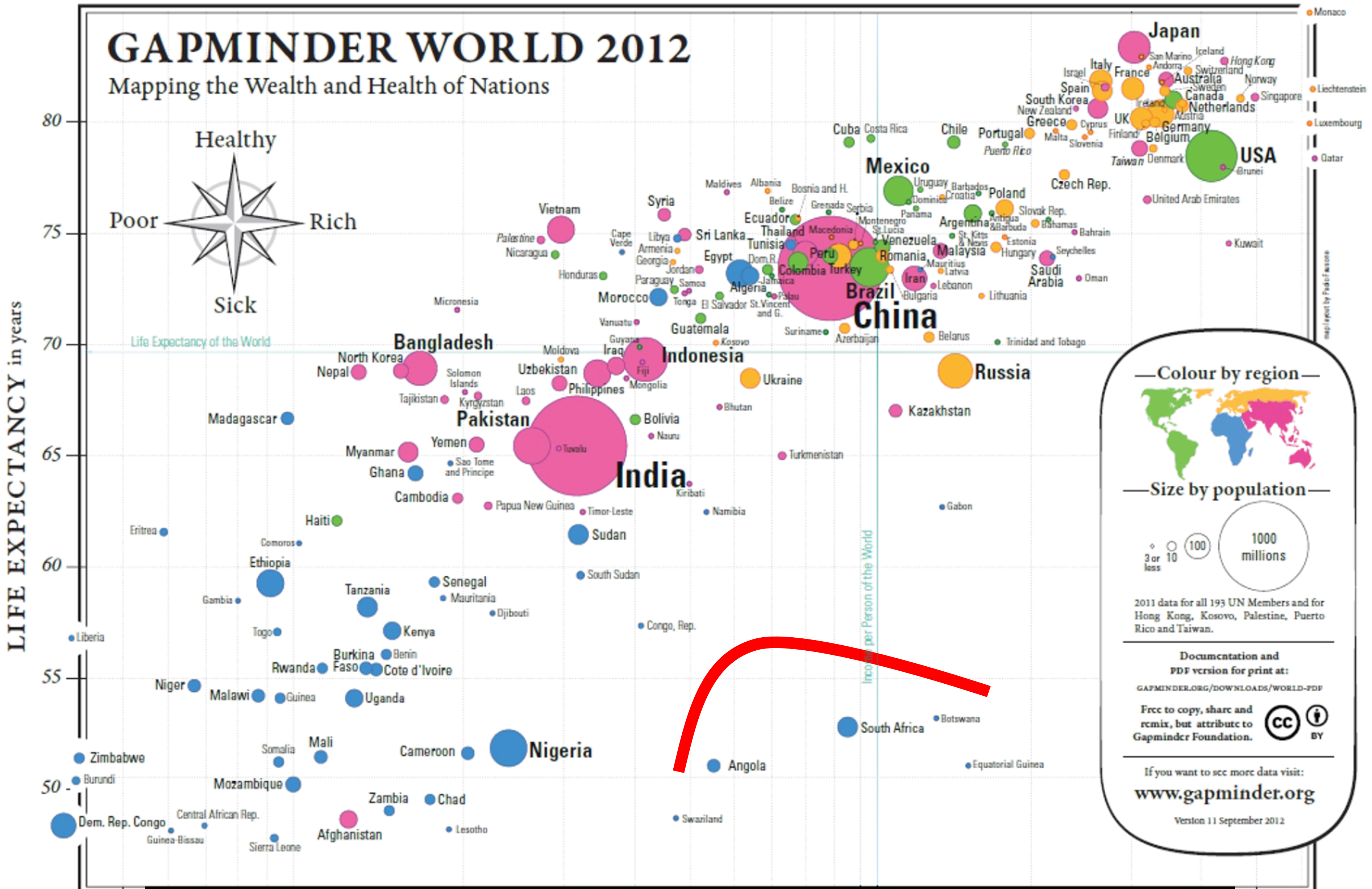
Underlying Structure and Multivariate Links

[Gapminder Foundation, https://www.ted.com/talks/hans_rosling_shows_the_best_stats_you_ve_ever_seen]



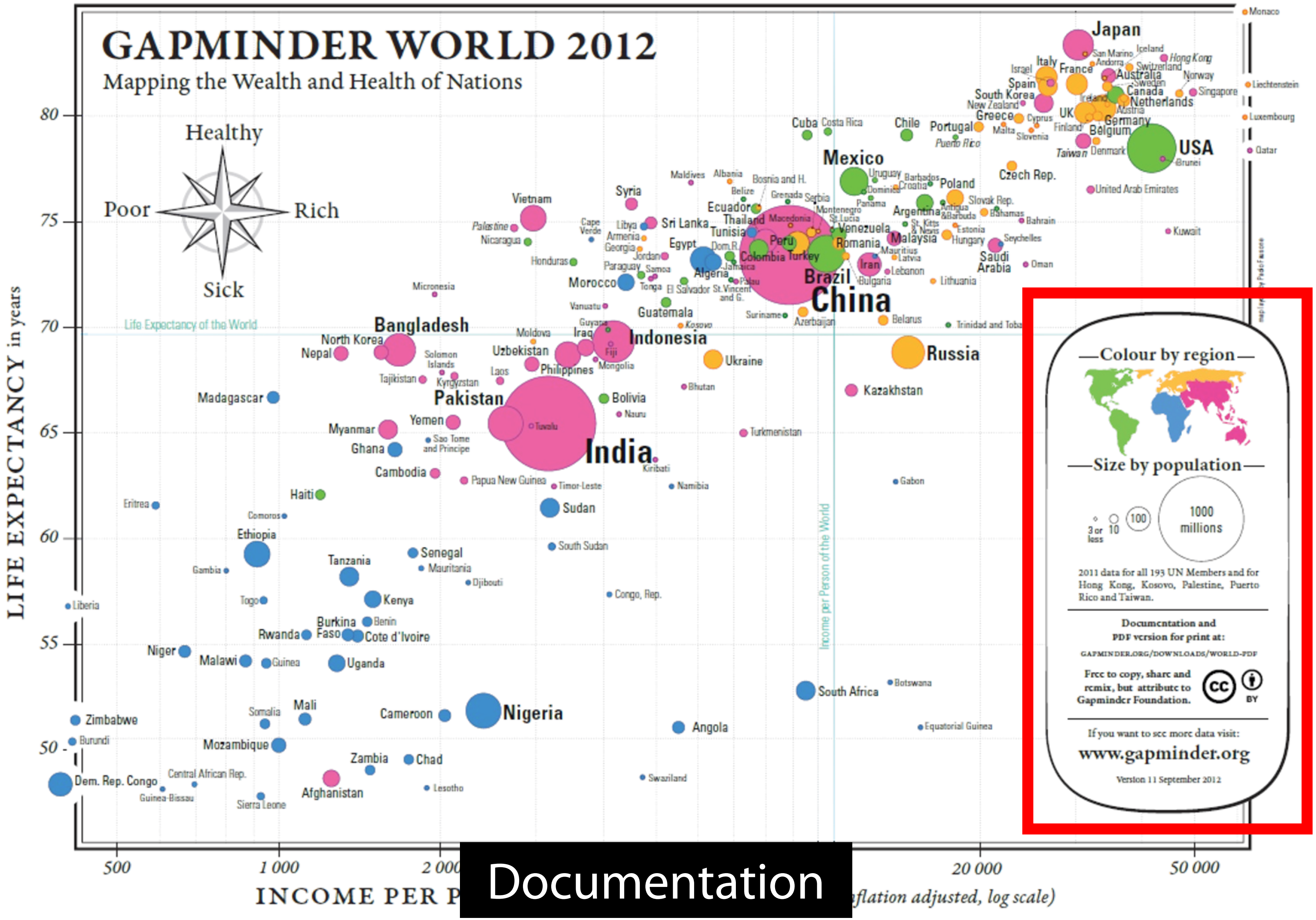
Underlying Structure and Multivariate Links

[Gapminder Foundation, https://www.ted.com/talks/hans_rosling_shows_the_best_stats_you_ve_ever_seen]



Underlying Structure and Multivariate Links

[Gapminder Foundation, https://www.ted.com/talks/hans_rosling_shows_the_best_stats_you_ve_ever_seen]



2011 data for all 193 UN Members and for Hong Kong, Kosovo, Palestine, Puerto Rico and Taiwan.

Documentation and PDF version for print at:
GAPMINDER.ORG/DOWNLOADS/WORLD-PDF

Free to copy, share and remix, but attribute to Gapminder Foundation.

If you want to see more data visit:
www.gapminder.org

Version 11 September 2012

Documentation

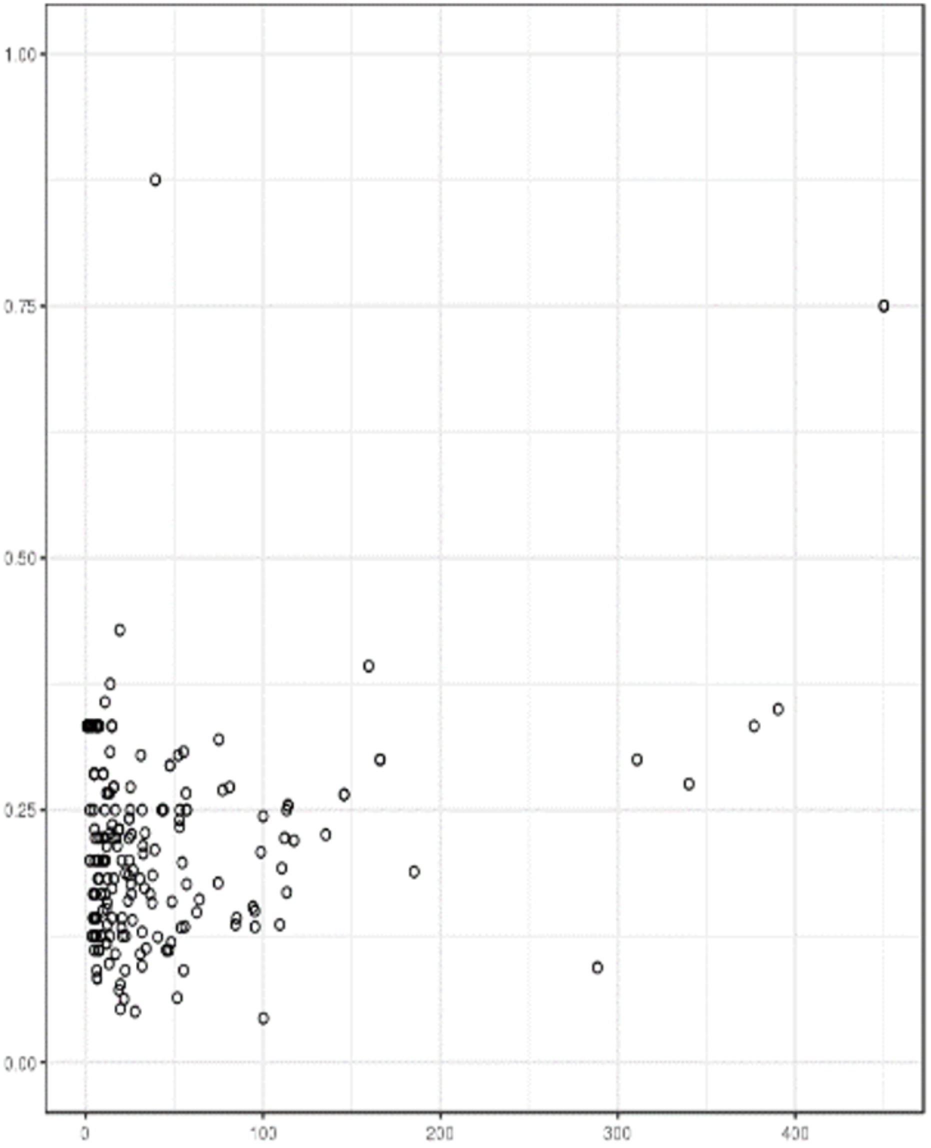
[http://promise.site.uottawa.ca/SERepository/datasets/cm1.arff]

Representing Observations

2 variables can be represented by position.

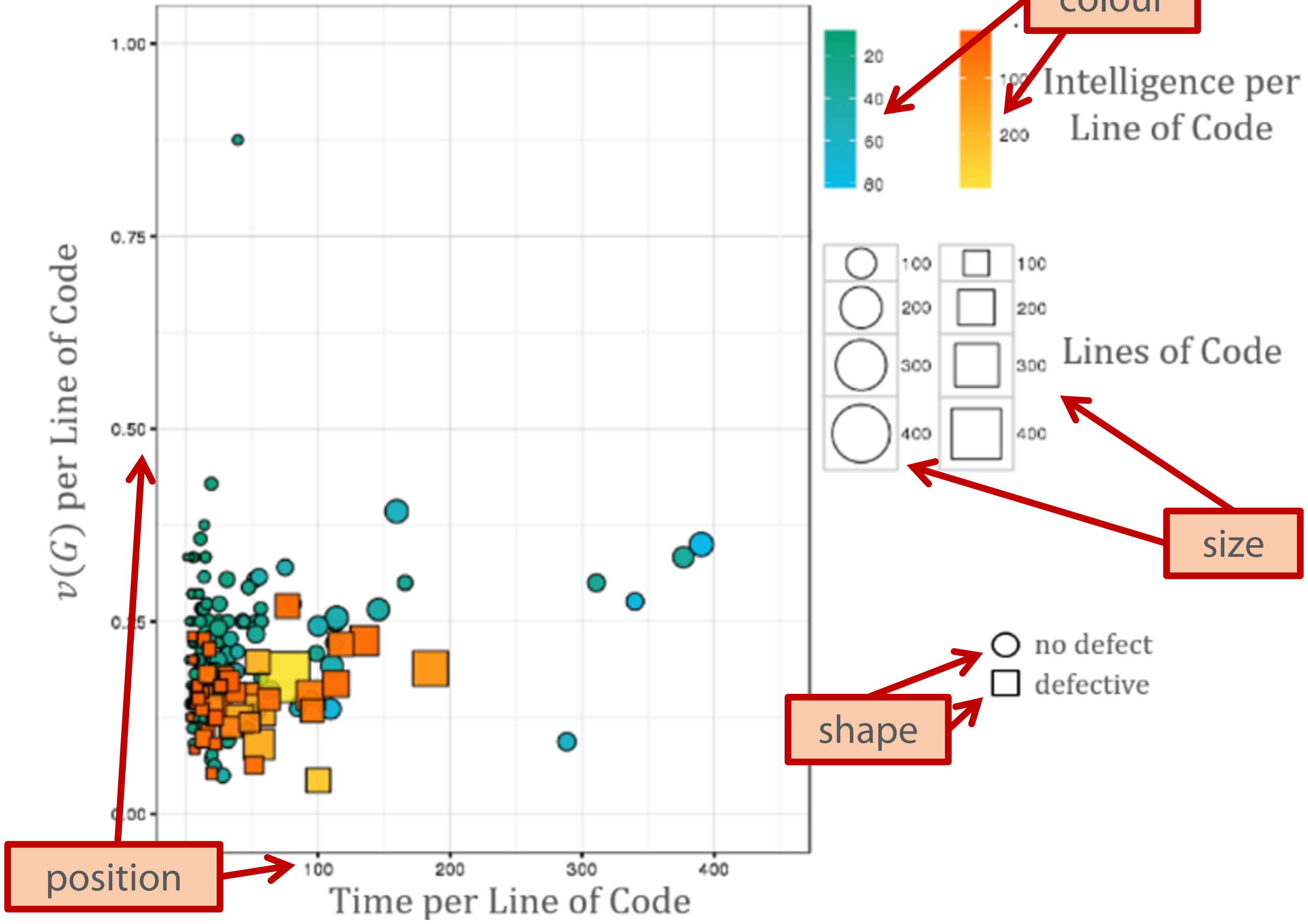
Additional factors can be depicted through:

- size
- color
- value
- texture
- line orientation
- shape
- (motion?)



NASA CM1 Dataset (subset)

NASA CM1 Dataset (subset)



colour

size

shape

position

[http://dataphys.org/list/]

A Word About Accessibility

Charts cannot usually be translated to Braille. Describing the features and emerging structures in a visualization is a possible solution… **if they can be spotted**.

Analysts must produce clear and meaningful visualizations, but they must also describe them and their features in a fashion that allows all to "see" the insights. This requires analysts to have “seen” all the insights, which is not always possible.

Conditions: colourblindness, low vision, motor impairment, cognitive disability, ADHD, etc.

Best Practices: high contrast elements, zoom/magnifications, keyboard navigation, assistive design, short summaries, un/re-do functionality, text-to-voice, etc. [Elavsky]

A Word About Accessibility

Data Perception:

- texture-based representations
- text-to-speech
- sound/music
- odor-based or taste-based representations (?!?)

Sonifications:

- TRAPPIST Sounds : TRAPPIST-1 Planetary System Translated Directly Into Music
- Listening to data from the Large Hadron Collider, L. Asquith

Chart Types

Simple text and tables

Scatterplot

Line chart

Bar charts

Stacked bar charts

100% bar charts

Area charts

Treemaps

Gauge charts

Heatmaps and choropleth maps

Geographical maps

Parallel coordinates

Chernoff faces

Word clouds

Network diagrams

Dendrograms and trees

Sparklines

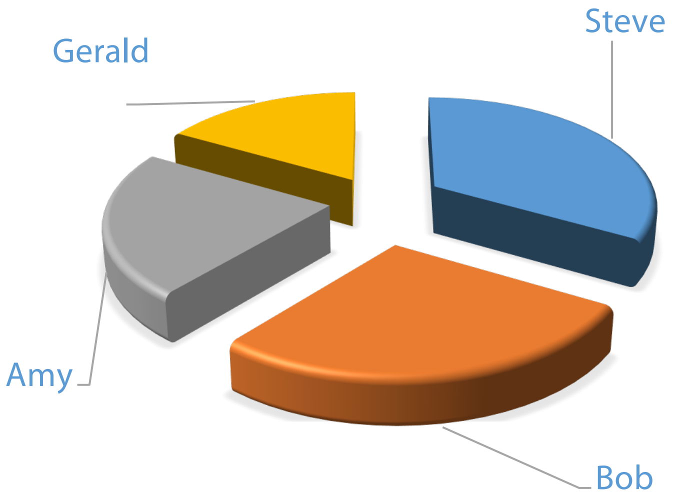
Interactive charts

Small multiples

etc.

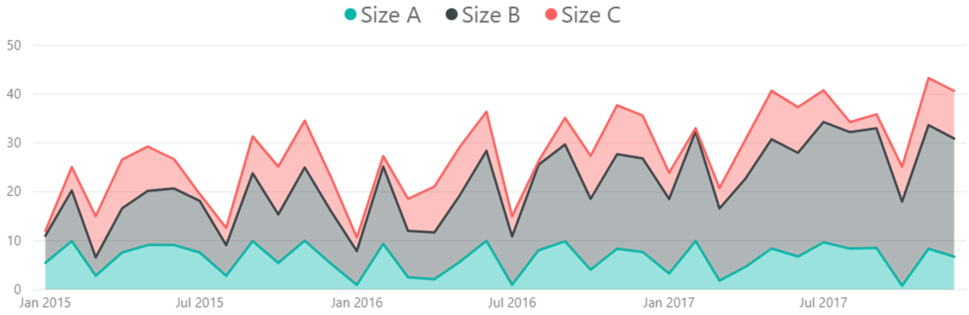
Charts to Avoid

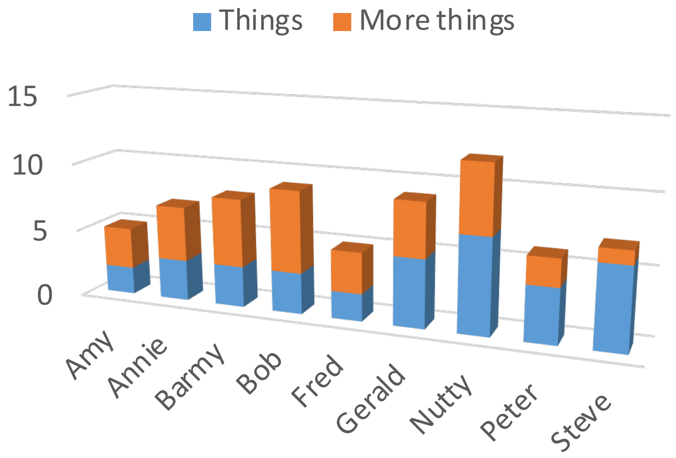
AVOID (?) anything with an arc (except gauge charts): pie, donut, etc. Human brains have a hard time **comparing arcs** – which is larger, Steve or Bob?



AVOID 3D charts: it is difficult to compare them visually (and they add too much clutter).

AVOID stacked area charts: way too confusing.





Decluttering

CLUTTER IS THE ENEMY!

- every element on a page adds **cognitive load**
- identify anything that isn't adding value and **remove**
- think of cognitive load as mental effort required to process information (lower is better)
- Tufte refers to the **data-to-ink** ratio – "the larger the share of a graphic's ink devoted to data, the better"
- in Resonate, Duarte refers to this as "**maximizing the signal-to-noise ratio**" where the signal is the information or the story we want to communicate.

Decluttering

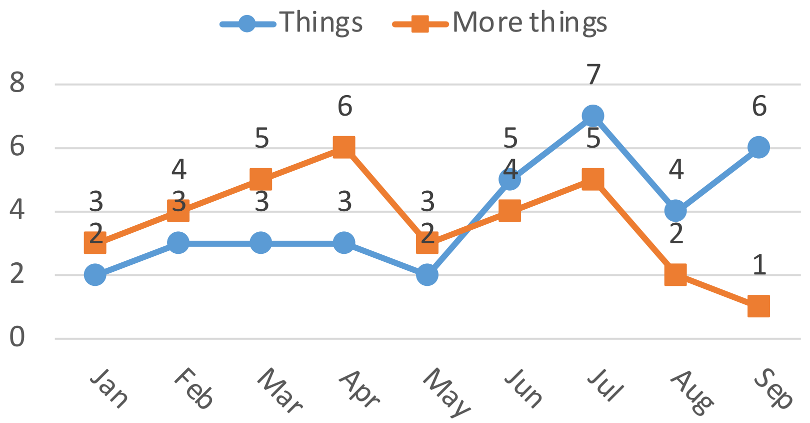
Use **Gestalt Principles** to organize/highlight data in a chart.

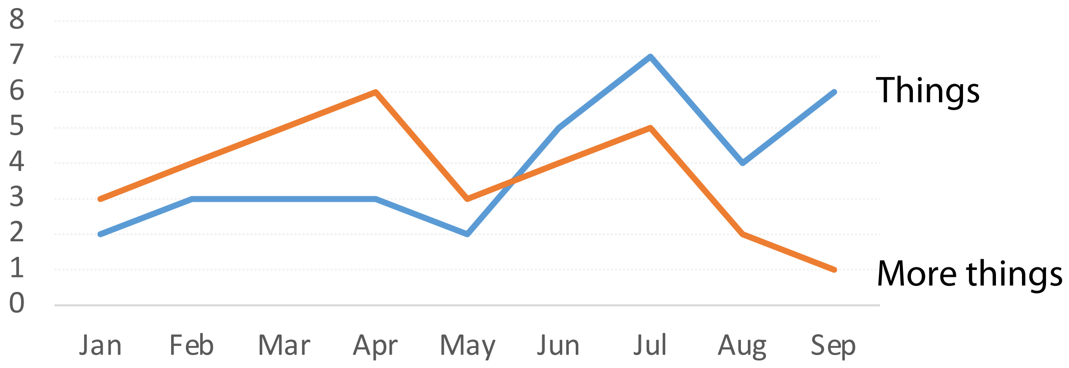
Align all the elements (graphs, text, lines, titles, etc.)

- DON'T rely on eye, use position boxes and values

Charts:

- remove border, gridlines, data markers
- clean up axis labels
- label data directly





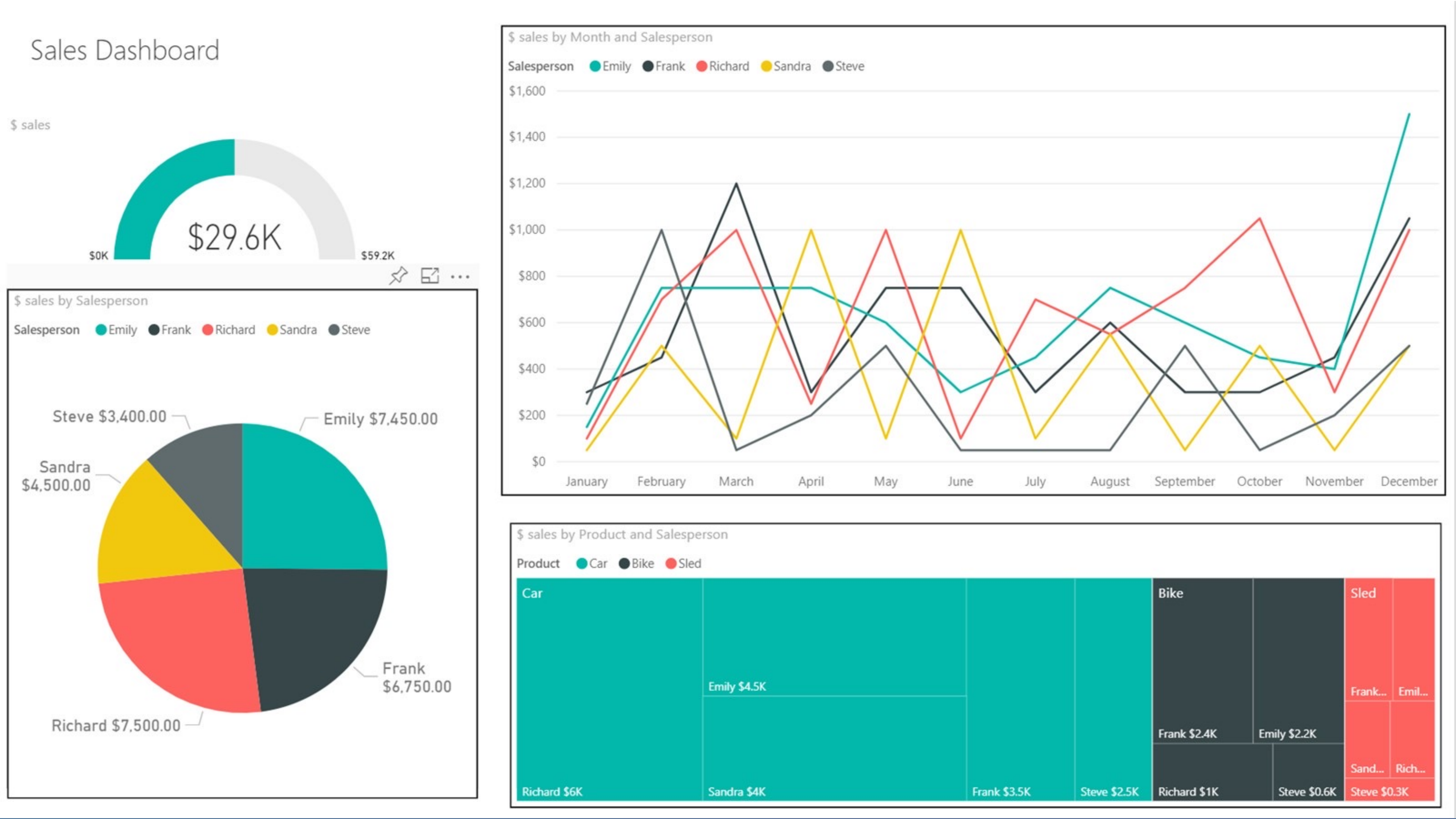
Decluttering

Use **consisten**t font, font size, colour and alignment.

Don’t rotate text to anything other than 0 or 90 degrees.

Use **white space**

- margins should remain free of text and visuals
- don’t stretch visuals to edge of page or too close to other visuals
- think of white space as a border

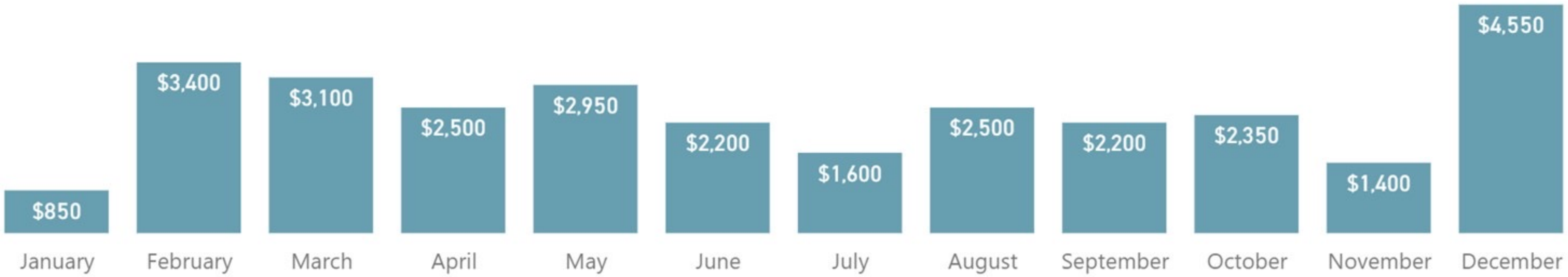

Sales Dashboard
$ sales
$29.6K
$0K
$59.2K
$ sales by Salesperson
Salesperson ● Emily ● Frank ● Richard ● Sandra ● Steve
Steve $3,400.00
Emily $7,450.00
Sandra $4,500.00
Frank $6,750.00
Richard $7,500.00
$ sales by Month and Salesperson
Salesperson ● Emily ● Frank ● Richard ● Sandra ● Steve
$1,600
$1,400
$1,200
$1,000
$800
$600
$400
$200
$0
January
February
March
April
May
June
July
August
September
October
November
December
$ sales by Product and Salesperson
Product ● Car ● Bike ● Sled
Car
Emily $4.5K
Richard $6K
Sandra $4K
Frank $3.5K
Steve $2.5K
Bike
Frank $2.4K
Emily $2.2K
Richard $1K
Steve $0.6K
Sled
Frank...
Emil...
Sand...
Rich...
Steve $0.3K

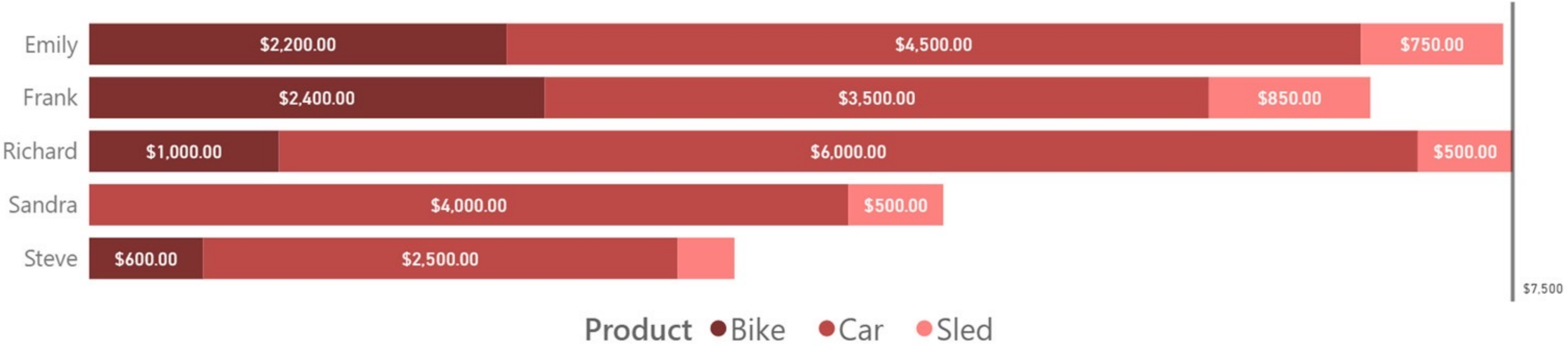
Sales Dashboard

Annual Sales for 2017

Total Sales

$29.6K





Data Viz Best Practices (Overview)

Effective data visualizations **provide insights** and **facilitate understanding**.

The basic principles and **Gestalt principles** can guide your visualization design and consumption.

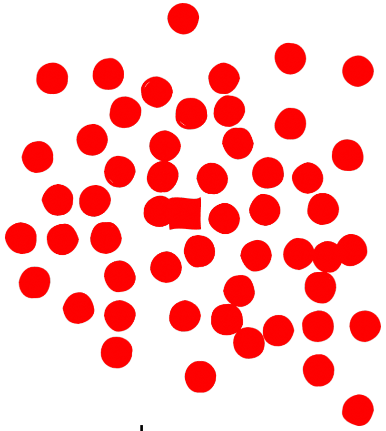
Be creative but keep your data and your representations **honest**.

Be mindful of attempts to **distort trends** and conclusions with flashy visuals.

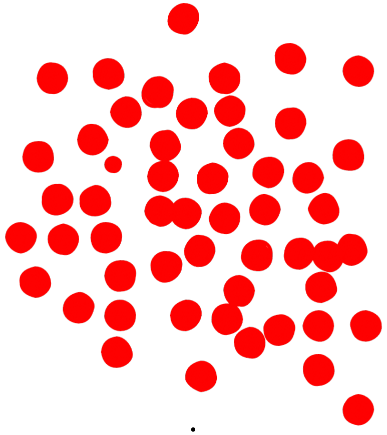
Data and code should be made available along with the displays.

[P. Boily, S. Davies, J. Schellinck, The Practice of Data Visualization]

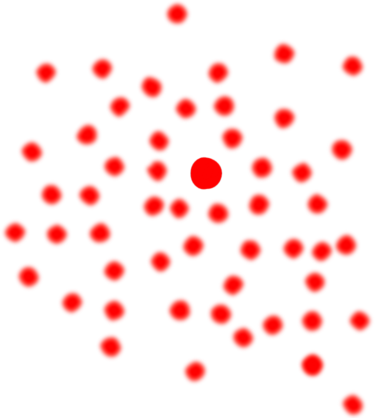
Pre-Attentive Attributes



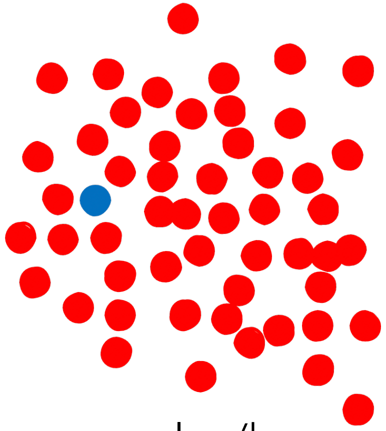
shape



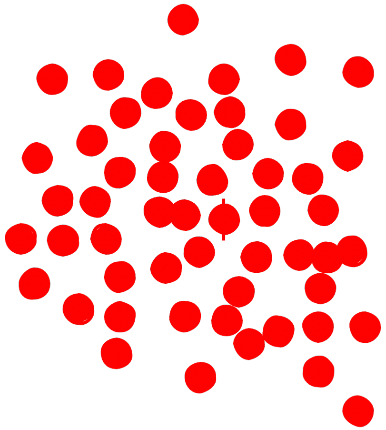
size



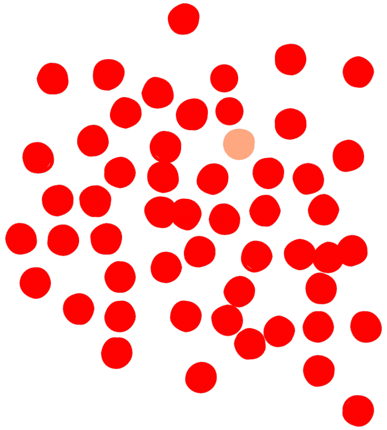
sharpness



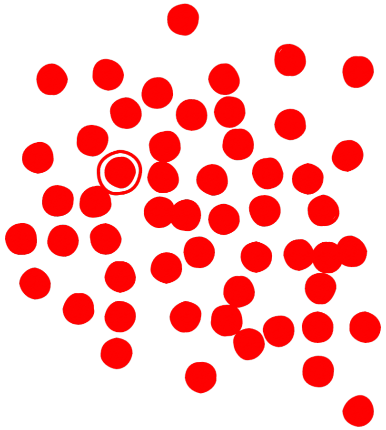
color/hue



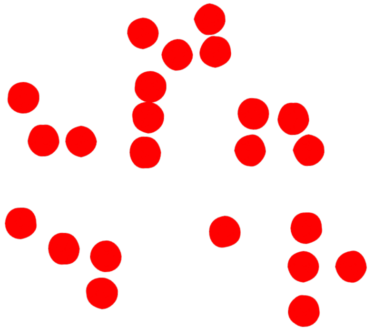
markings



intensity/value



enclosure



numerosity

[P. Boily, S. Davies, J. Schellinck, The Practice of Data Visualization]

Example: Pre-Attentive Attributes

How many 6's are there on the next slide?

[P. Boily, S. Davies, J. Schellinck, The Practice of Data Visualization]

2869408609876

9348586748676

2967303986739

3967496749674

28**6**9408**6**0987**6**

934858**6**748**6**7**6**

29**6**730398**6**739

39**6**749**6**7496 74

[P. Boily, S. Davies, J. Schellinck, The Practice of Data Visualization]

28**6**9408**6**0987**6**

934858**6**748**6**7**6**

29**6**730398**6**739

39**6**749**6**749**6**74

[P. Boily, S. Davies, J. Schellinck, The Practice of Data Visualization]

2869408609876

9348586748676

2967303986739

3967496749674

[P. Boily, S. Davies, J. Schellinck, The Practice of Data Visualization]

2869408609876

934858**67**48**67**6

29**67**303986**7**39

39**67**49**67**49**67**4

2869408609876

9348586748676

2967303986739

3967496749674

Colour Schemes

When it comes to colour, **less is more**: use it sparingly (graphic designers are taught to “get it right, in black and white”).

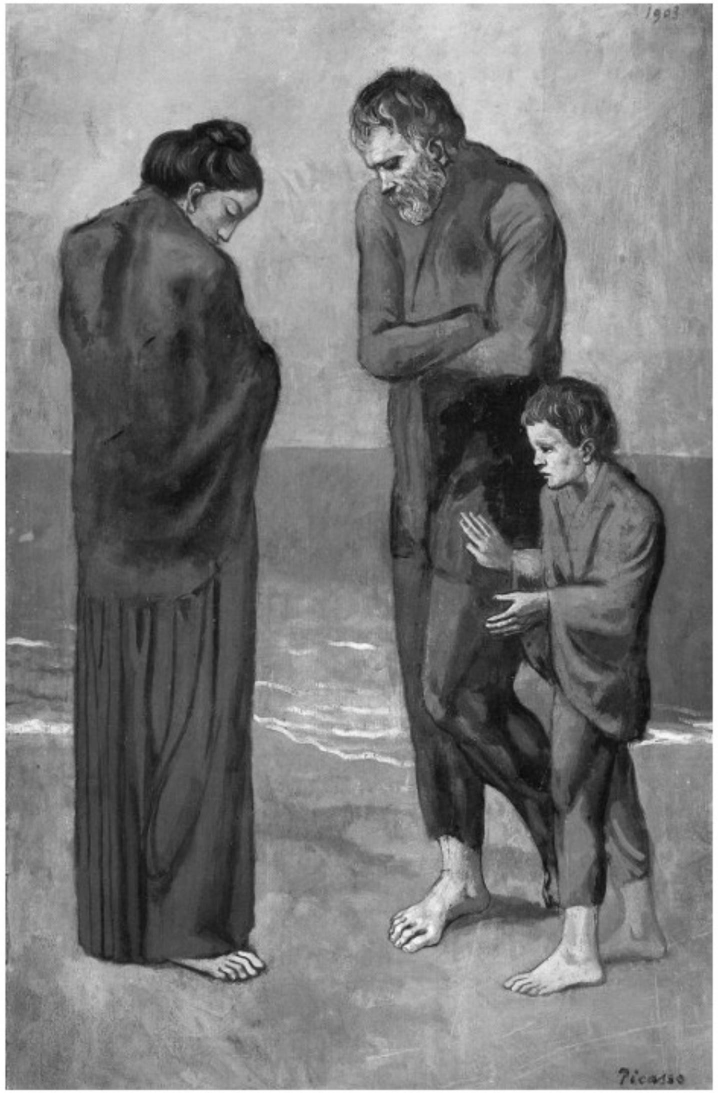
Based on the Gestalt Principles, **monochrome schemes** can be particularly effective.

When appropriate, pick scheme based on corporate identity (this maximizes buy in).

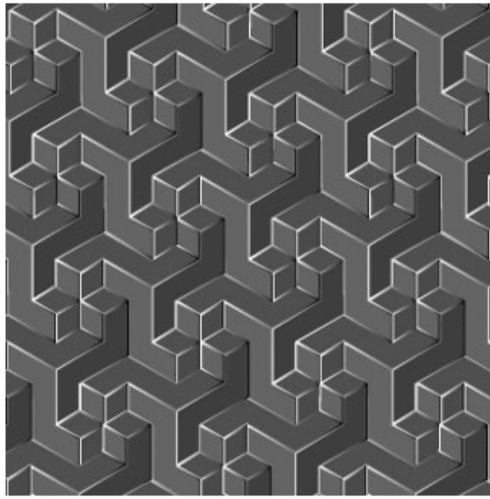
Create a template (and stick to it).

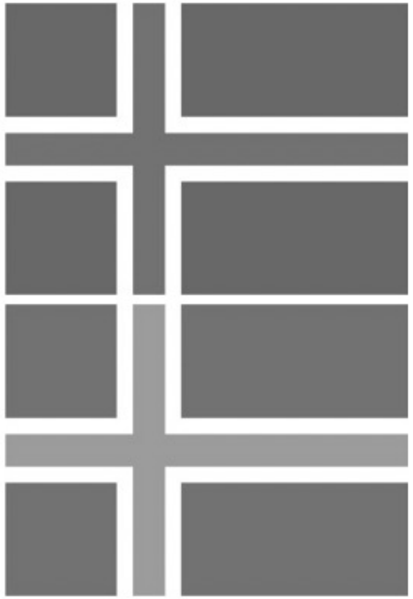
Upload images to see what charts look like in various flavours of colour-blindness:

- https://www.color-blindness.com/coblis-color-blindness-simulator (there are other tools)


1903
Picasso

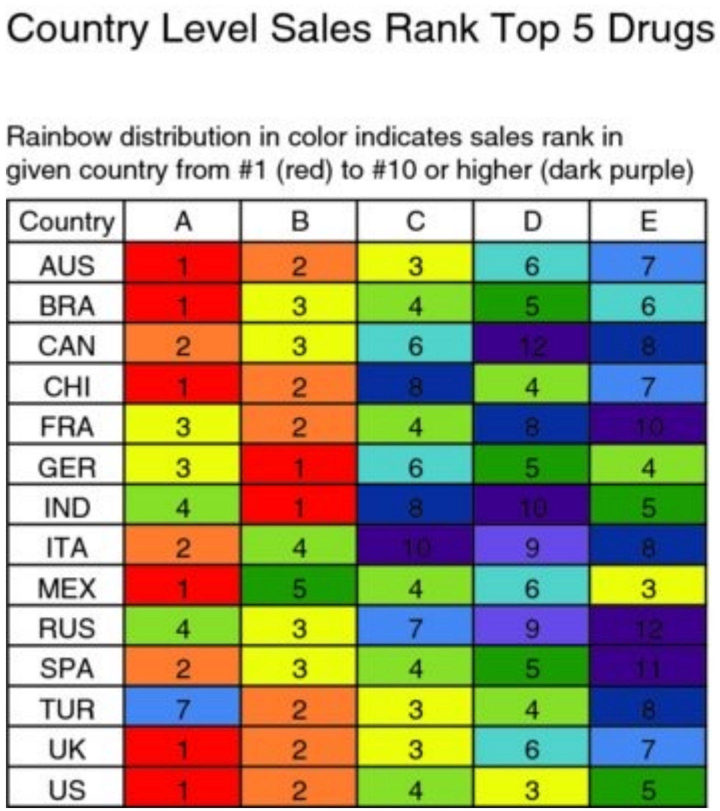
Claude Monet





Kellogg's
FROSTED FLAKES
THEY'RE GR-R-REAT!
Kellogg's
CORN FLAKES
THE ORIGINAL & BEST
HEART HEALTHY
FIBER
Kellogg's
Raisin Bran
9g PROTEIN
REAL FRUIT

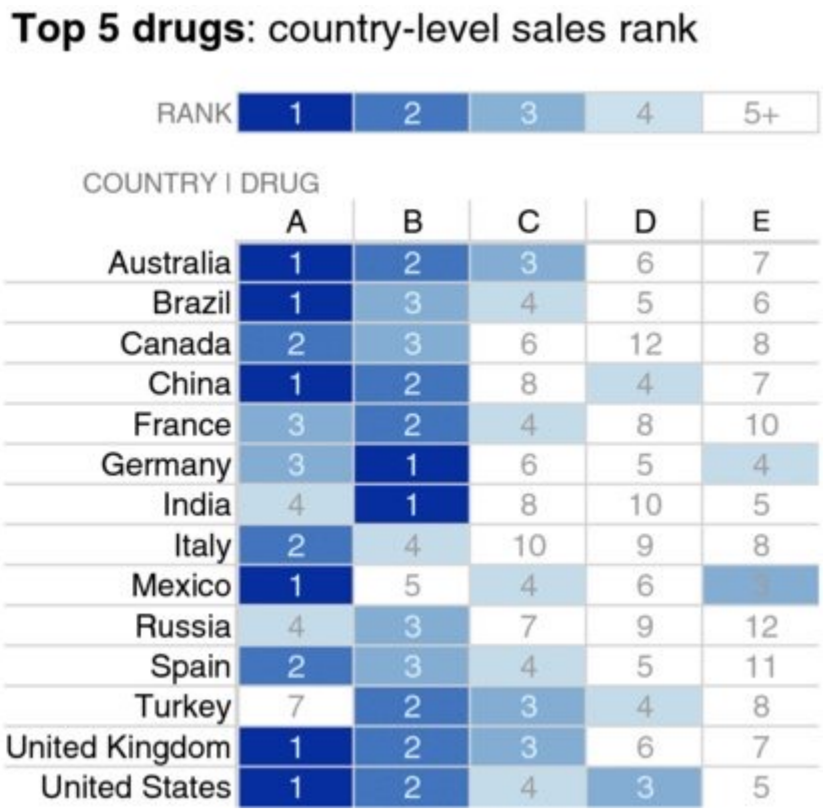
Mastering Colour



Country Level Sales Rank Top 5 Drugs

Rainbow distribution in color indicates sales rank in given country from #1 (red) to #10 or higher (dark purple)

Country	A	B	C	D	E
AUS	1	2	3	6	7
BRA	1	3	4	5	6
CAN	2	3	6	12	8
CHI	1	2	8	4	7
FRA	3	2	4	8	10
GER	3	1	6	5	4
IND	4	1	8	10	5
ITA	2	4	10	9	8
MEX	1	5	4	6	3
RUS	4	3	7	9	12
SPA	2	3	4	5	11
TUR	7	2	3	4	8
UK	1	2	3	6	7
US	1	2	4	3	5



Top 5 drugs: country-level sales rank

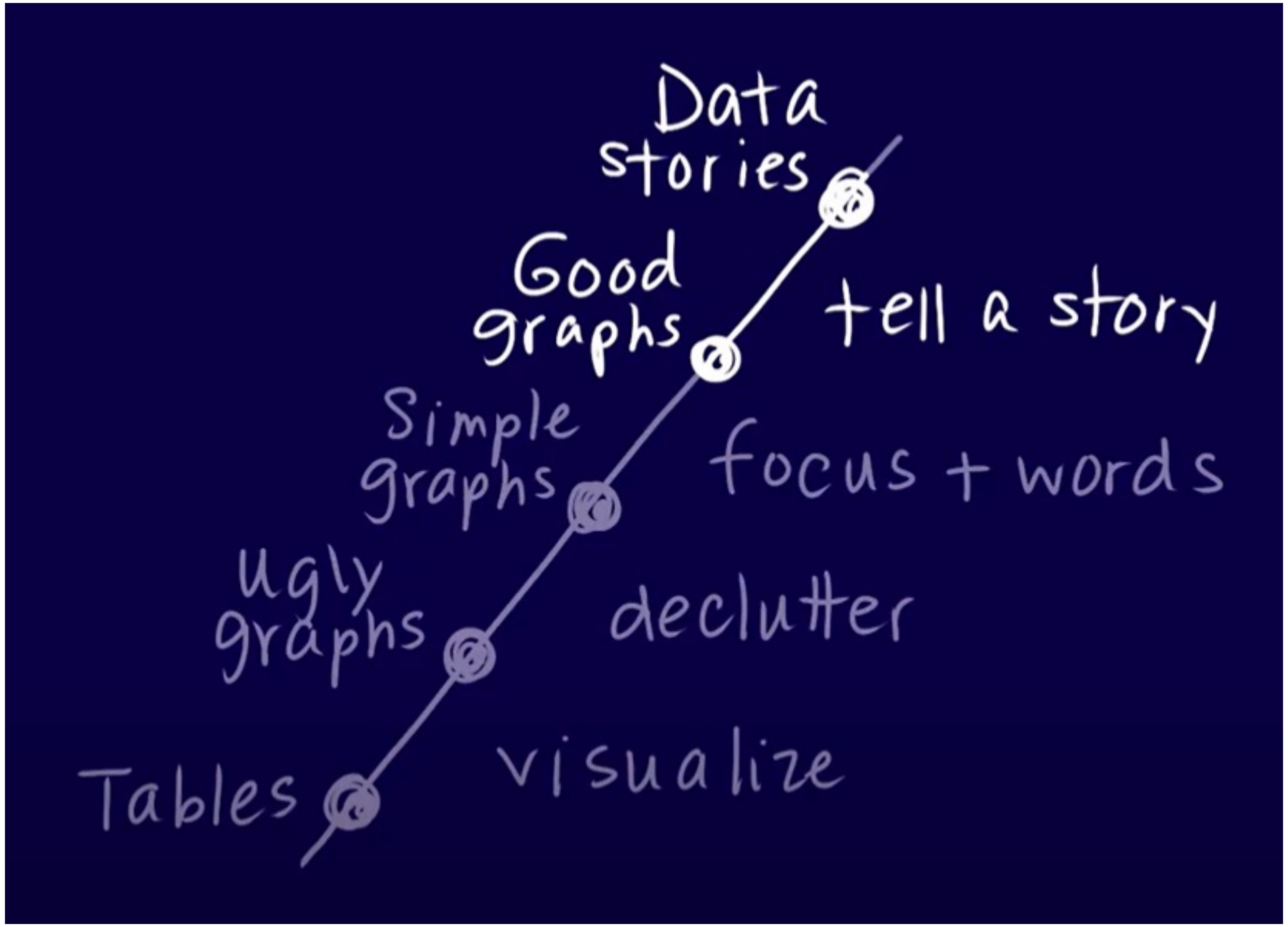
RANK	1	2	3	4	5+

COUNTRY I DRUG

	A	B	C	D	E
Australia	1	2	3	6	7
Brazil	1	3	4	5	6
Canada	2	3	6	12	8
China	1	2	8	4	7
France	3	2	4	8	10
Germany	3	1	6	5	4
India	4	1	8	10	5
Italy	2	4	10	9	8
Mexico	1	5	4	6	3
Russia	4	3	7	9	12
Spain	2	3	4	5	11
Turkey	7	2	3	4	8
United Kingdom	1	2	3	6	7
United States	1	2	4	3	5

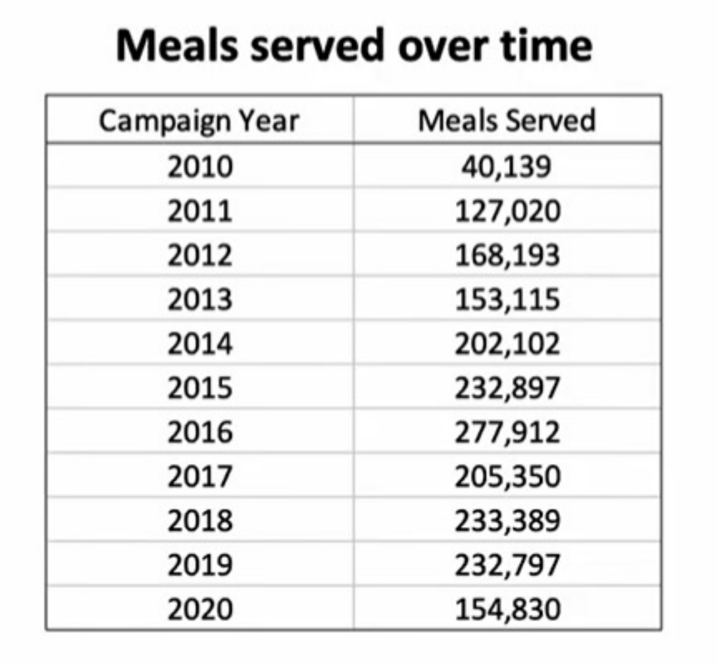
Source: Storytelling with Data

Evolving a Visualization



Source: Storytelling with Data

Evolving a Visualization

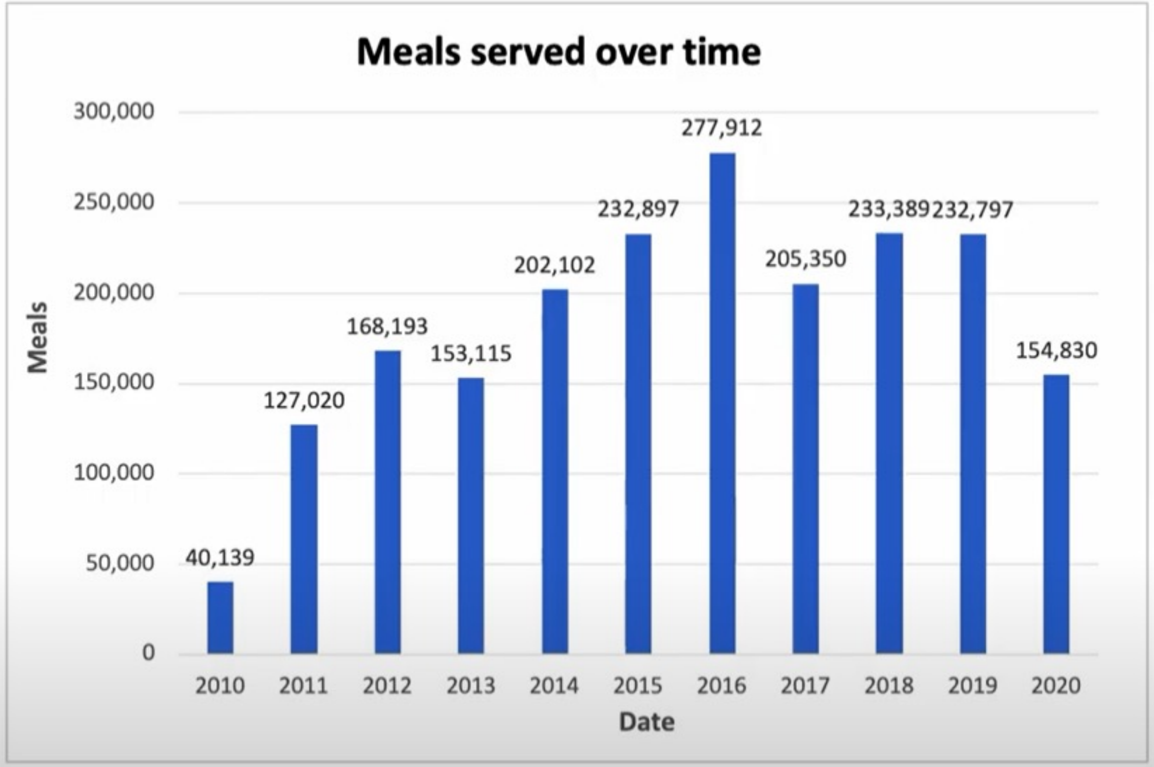


Meals served over time

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

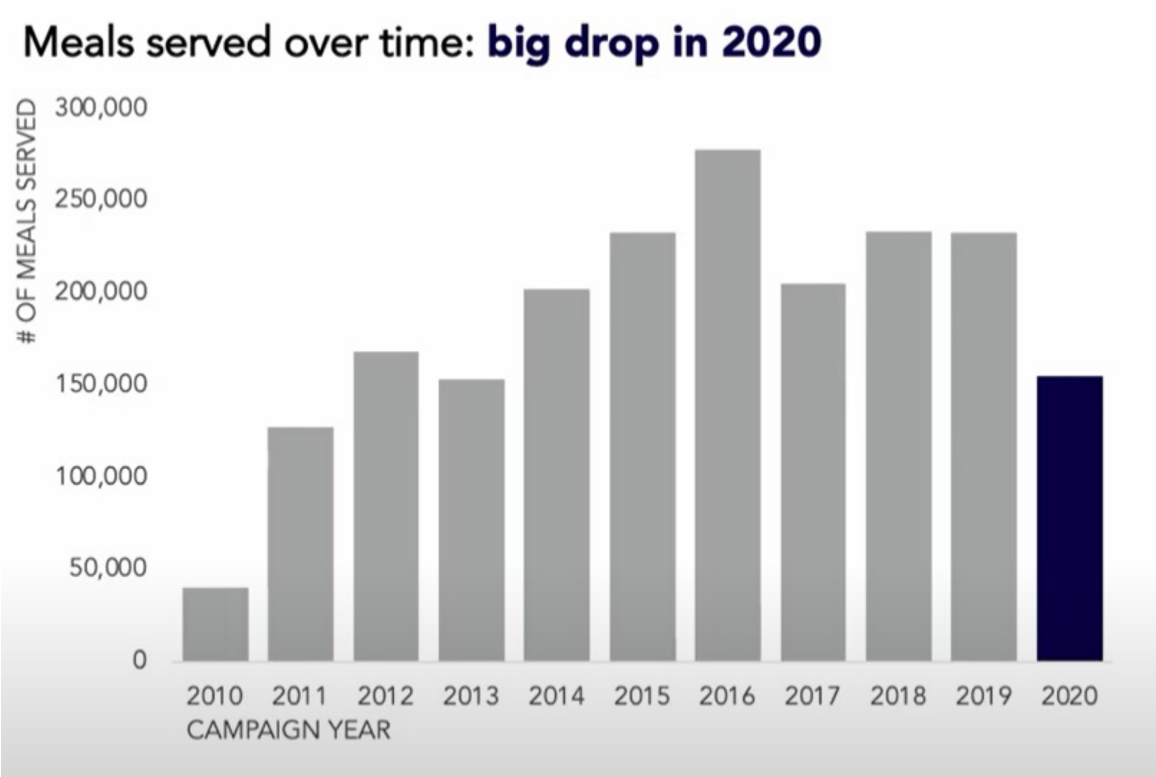
Source: Storytelling with Data

Evolving a Visualization



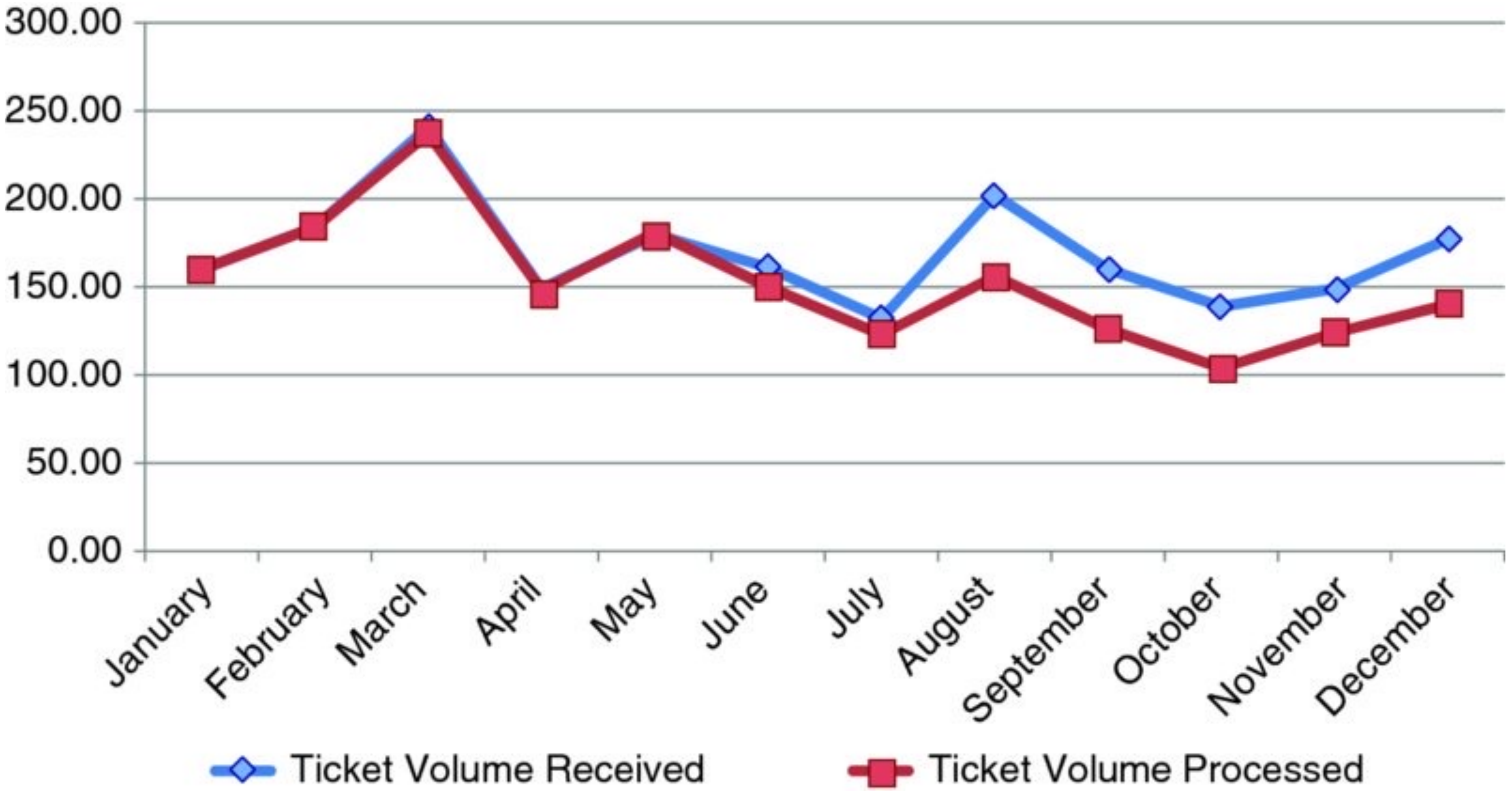
Source: Storytelling with Data

Evolving a Visualization



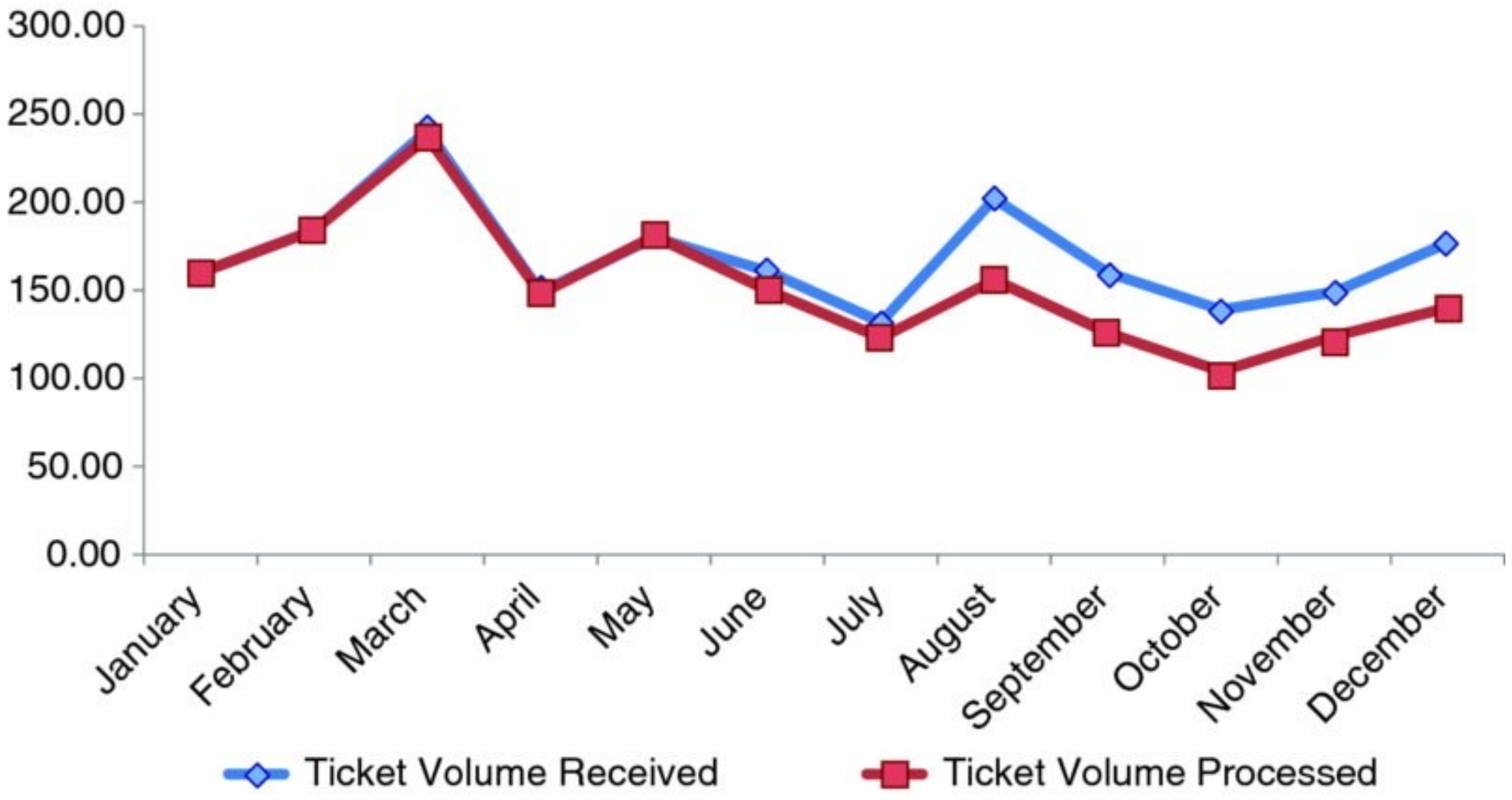
Source: Storytelling with Data

Decluttering – Step-by-Step Example



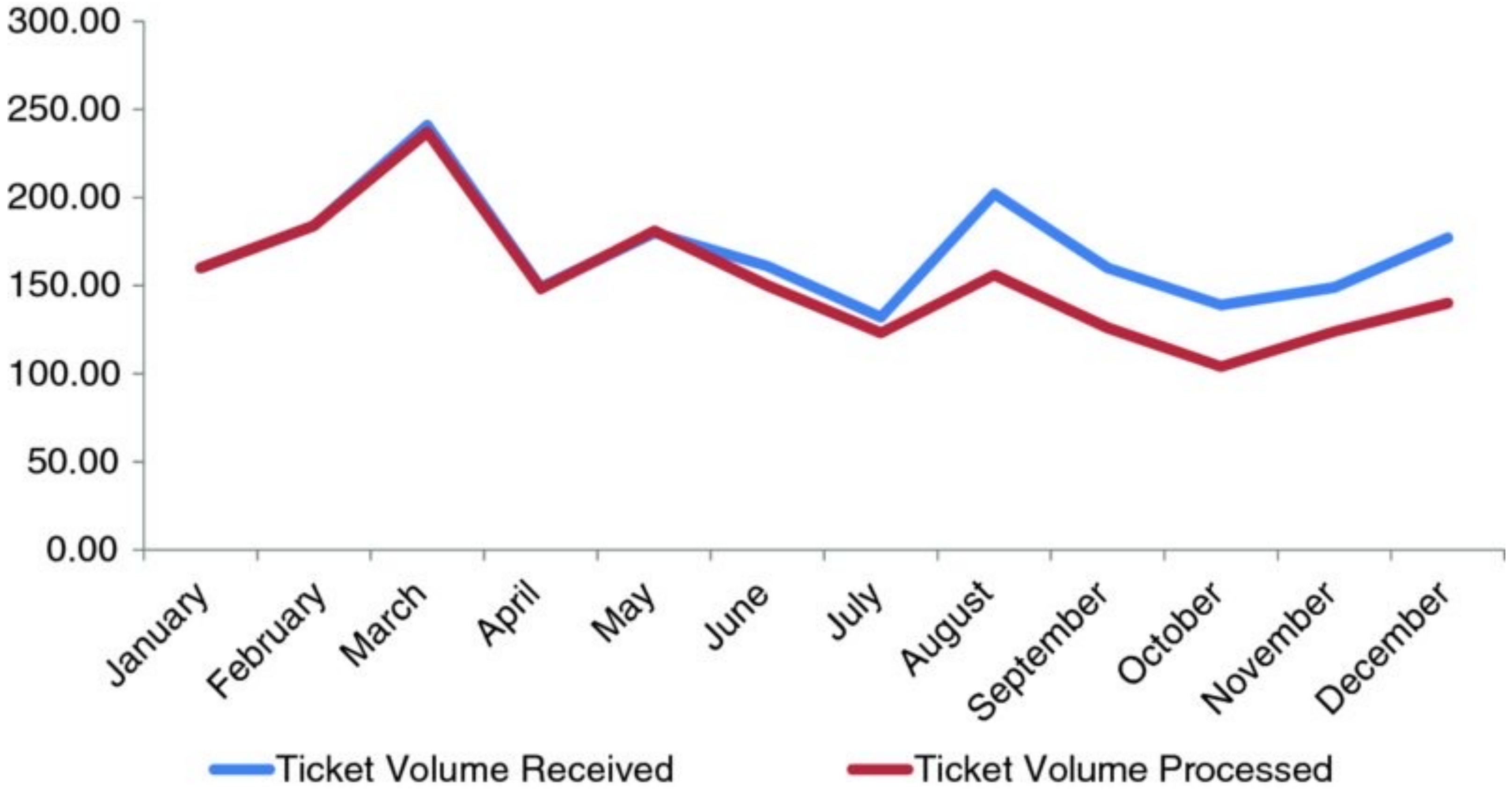
Source: Storytelling with Data

1. Remove Chart Border & Gridlines



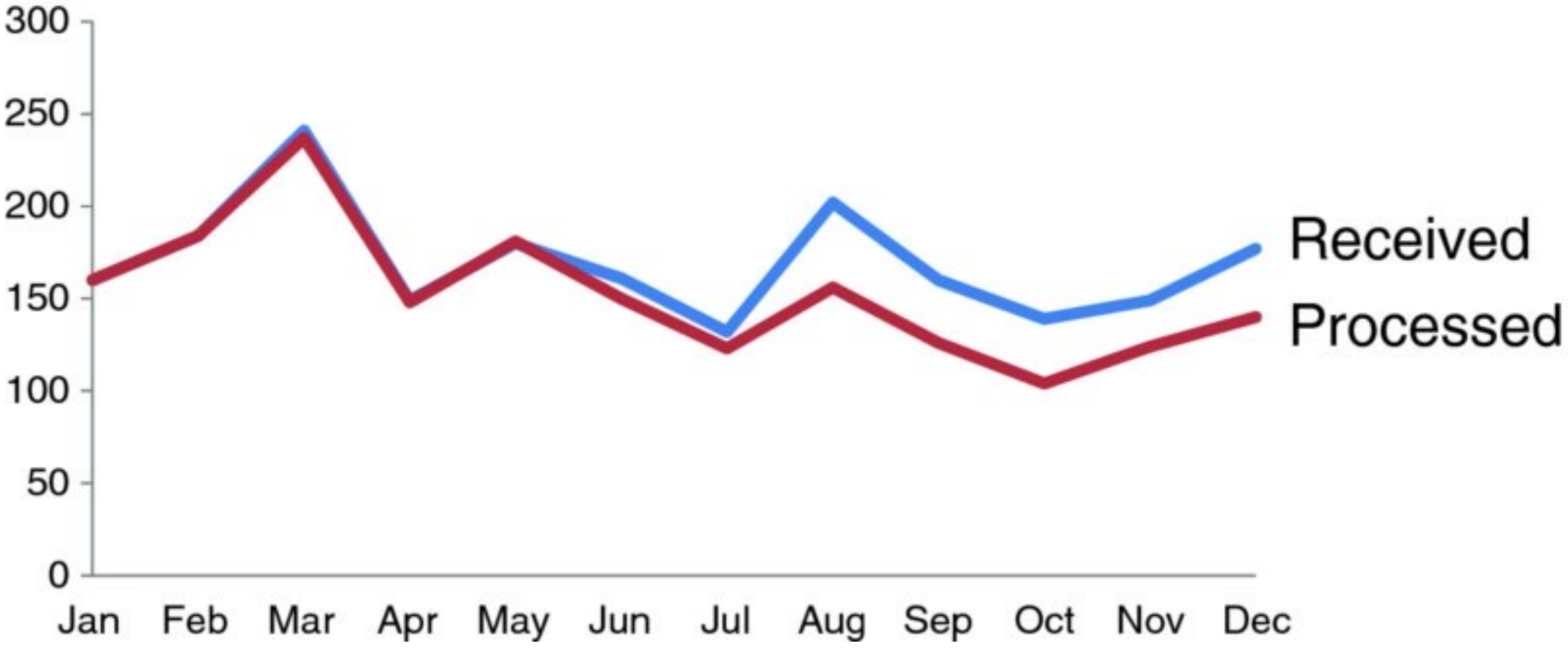
Source: Storytelling with Data

2. Remove Data Markers



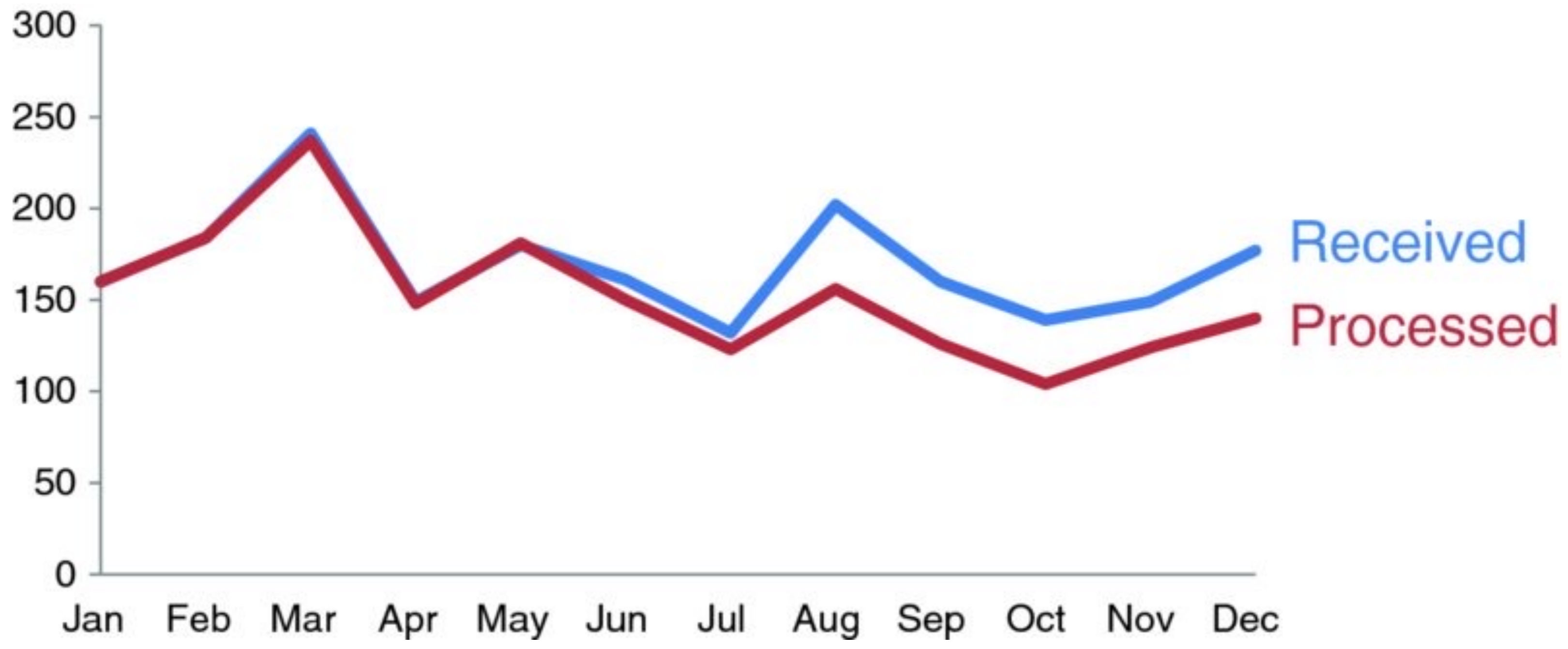
Source: Storytelling with Data

3. Clean Up Axis Labels



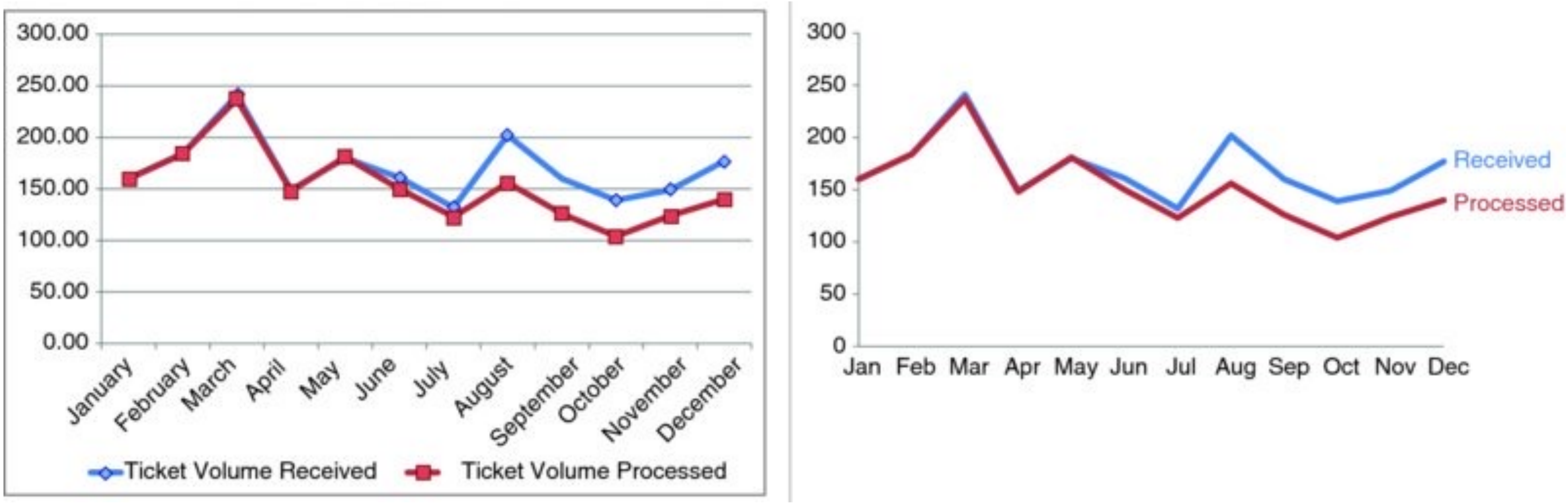
Source: Storytelling with Data

4. Colour Code the Lines



Source: Storytelling with Data

5. Before & After



Source: Storytelling with Data

Dashboards

A **dashboard** is any visual display of data used to monitor conditions and/or facilitate understanding.

In a car's dashboard, a small number of **key indicators** (speed, gasoline level, lights, etc.) need to be understood **immediately**.

A dashboard design that does not take these two characteristics under consideration can have **catastrophic consequences**. The same is true for data dashboards.

Dashboards Best Practices

The most amount of time someone will spend on a dashboard is **10-15 minutes**

- no more than 7-8 pages per dashboard (fewer is better!)

Short-term memory makes it difficult to see **more than 4 visual chunks** at once

- no more than 4-5 charts on a single page (fewer is better)

Pre-attentive features can help **direct the eye**

- each chart should have 1 iconic memory trigger

Long-term memory is more easily triggered by a **combination** of words and visuals

- explain: tell us, in a few words, what we are supposed to be seeing

Exercise

Consider the following dashboards.

Can you figure out, at a glance, who their audience is?

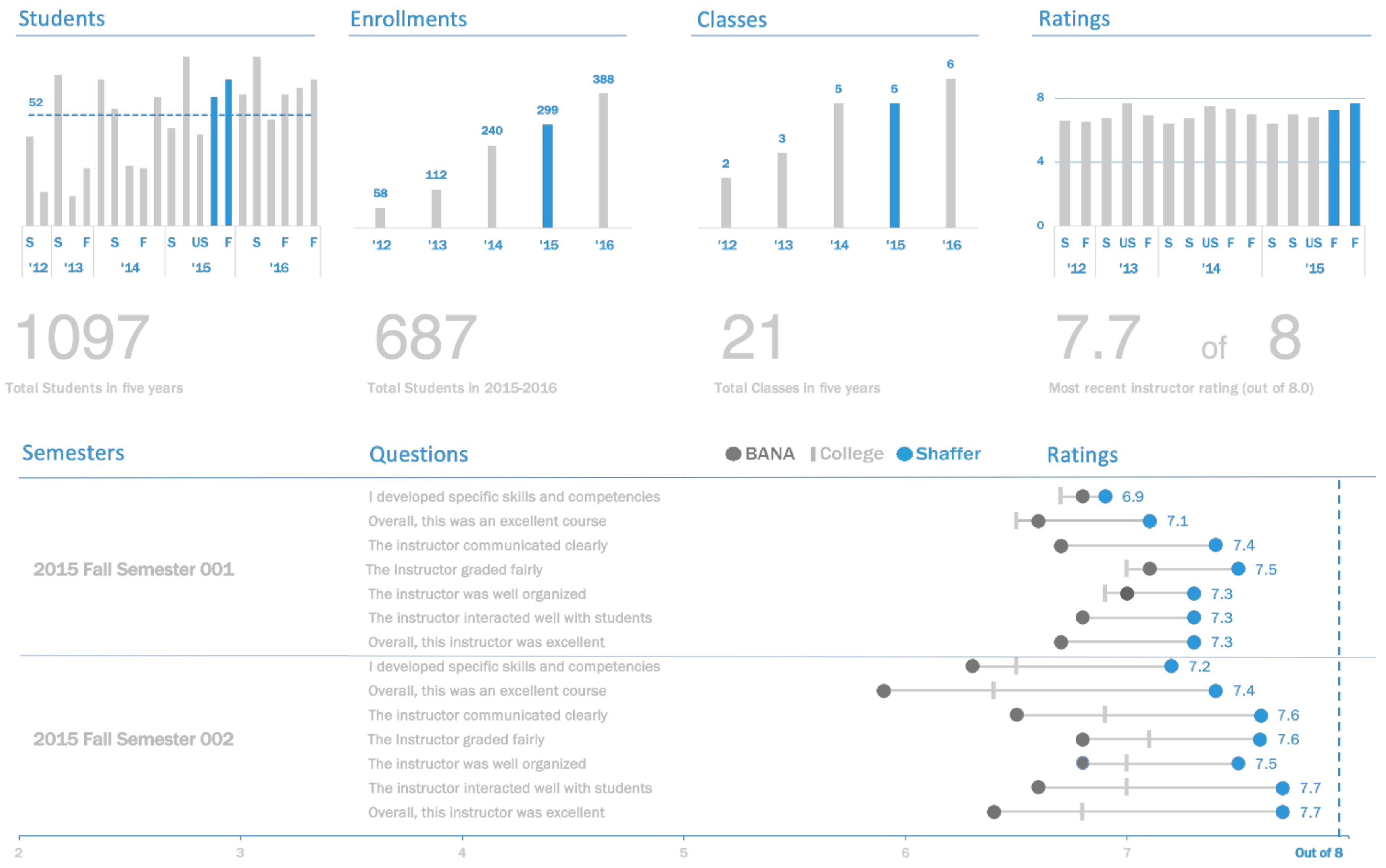
What are their strengths?

What are their limitations?

How would you improve them?

Course Metrics

[https://bigbookofdashboards.com/dashboards.html]



1097
Total Students in five years

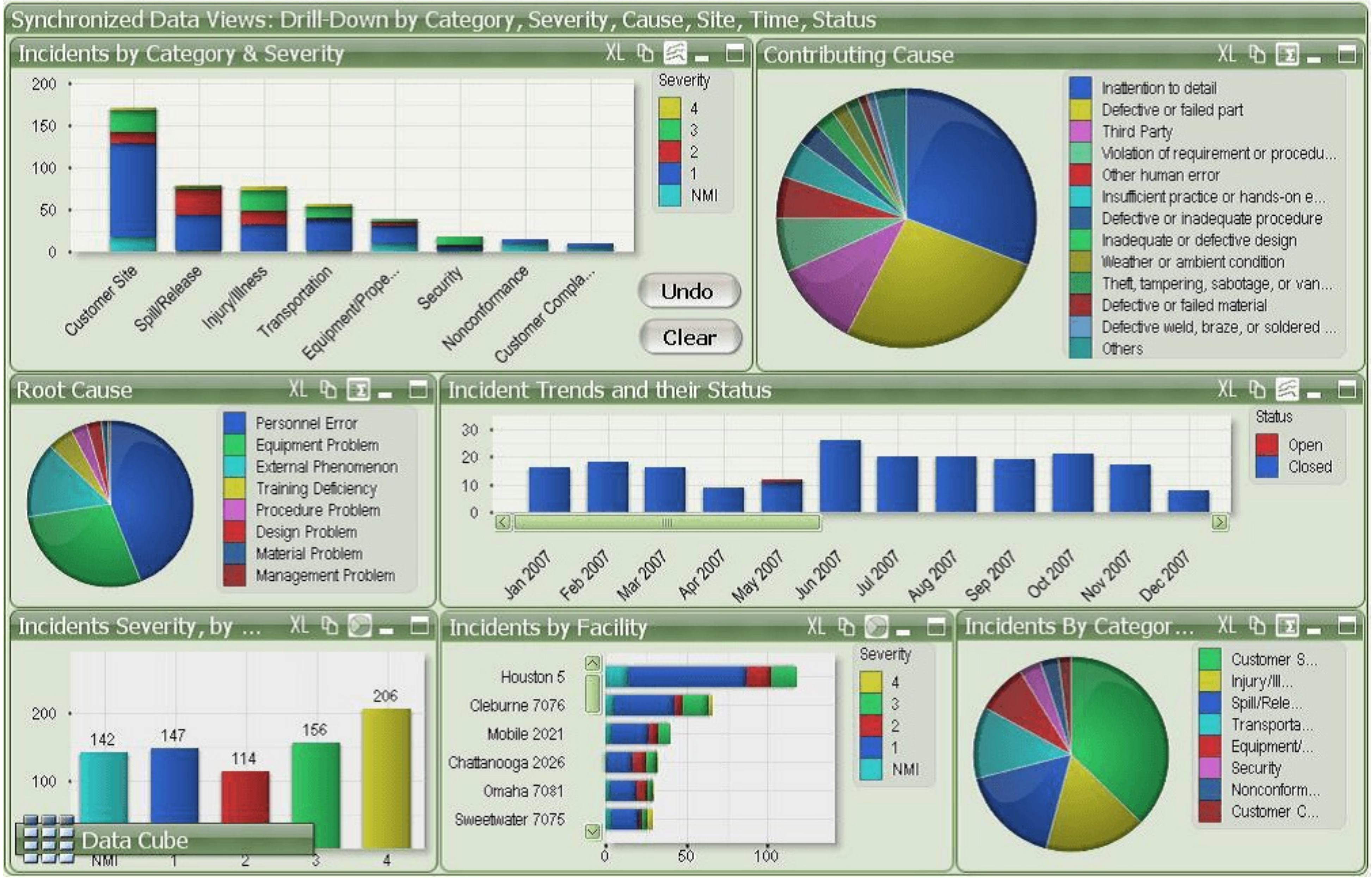
687
Total Students in 2015-2016

21
Total Classes in five years

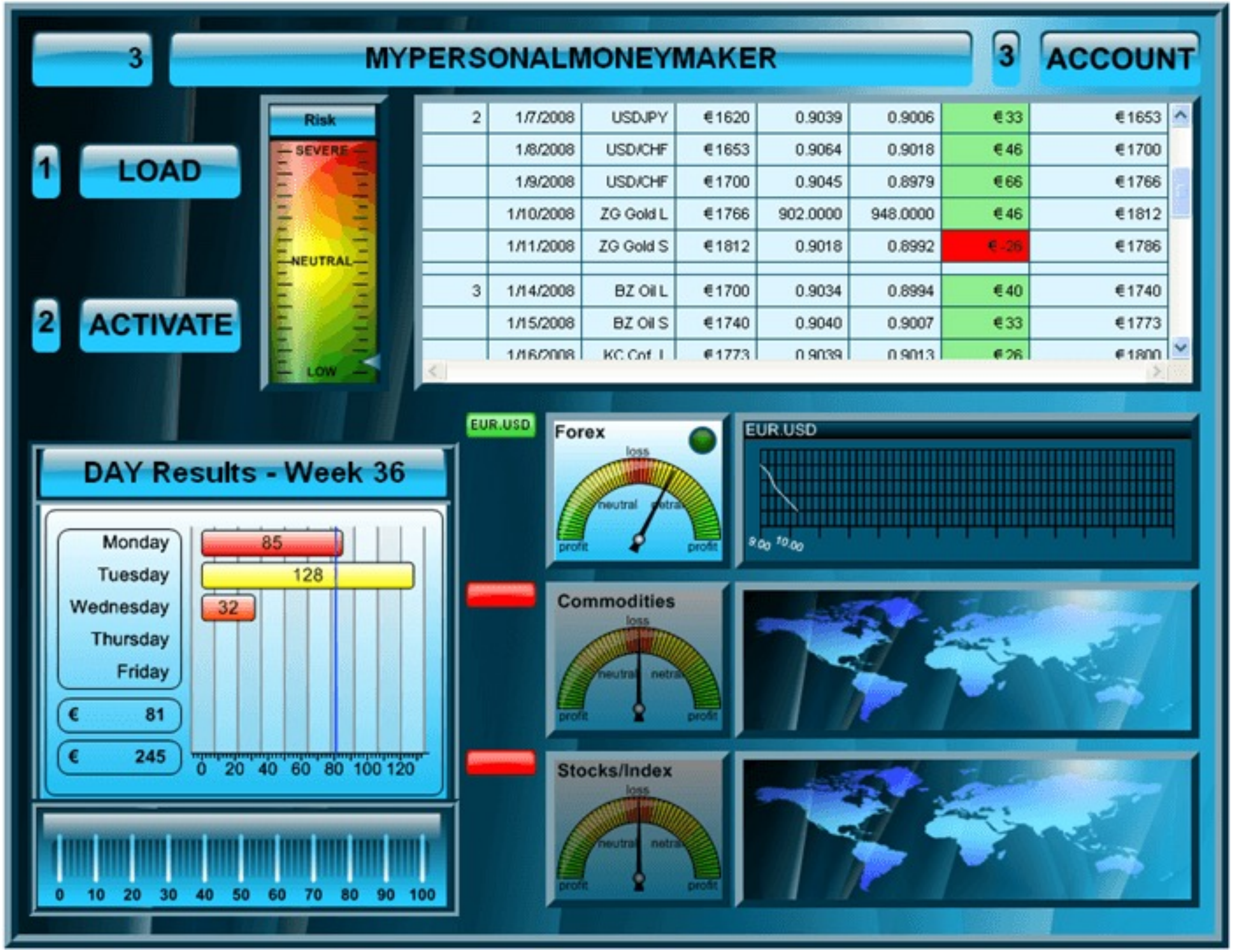
7.7 of 8
Most recent instructor rating (out of 8.0)

Course Metrics Dashboard created by Jeffrey A. Shaffer. Data from University of Cincinnati Course Evaluations. Blue indicates the 2 most recent rating periods.

[https://www.matillion.com/wp-content/uploads/2014/11/qlikview-poor-use-of-dashboard-software.png]



[https://www.geckoboard.com/assets/2-terrible-dashboard-example-min.png]

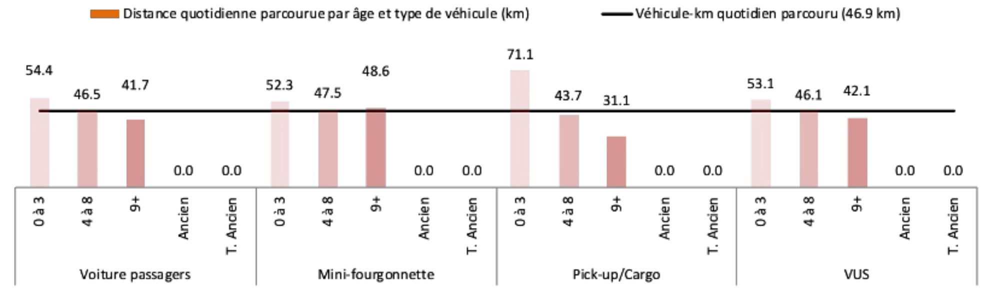


2	1/7/2008	USDJPY	€1620	0.9039	0.9006	€33	€1653
	1/8/2008	USD/CHF	€1653	0.9064	0.9018	€46	€1700
	1/9/2008	USD/CHF	€1700	0.9045	0.8979	€66	€1766
	1/10/2008	ZG Gold L	€1766	902.0000	948.0000	€46	€1812
	1/11/2008	ZG Gold S	€1812	0.9018	0.8992	€ -26	€1786
3	1/14/2008	BZ Oil L	€1700	0.9034	0.8994	€40	€1740
	1/15/2008	BZ Oil S	€1740	0.9040	0.9007	€33	€1773
	1/16/2008	KC Cof L	€1773	0.9039	0.9013	€26	€1800

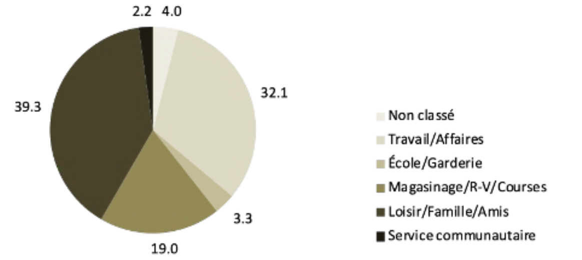
Ontario – 1er trimestre 2012

Caractéristiques des déplacements

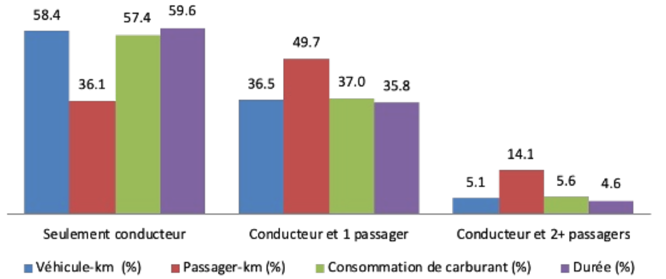
Véhicule-km quotidien parcouru par âge et type de véhicule



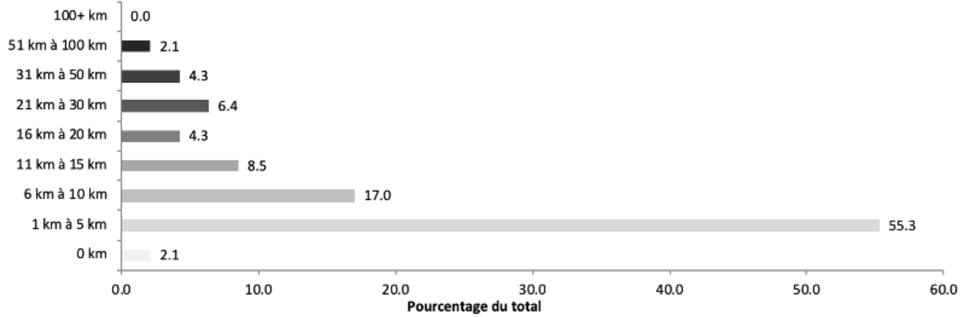
Passager-km quotidien parcouru par but des déplacements (%)



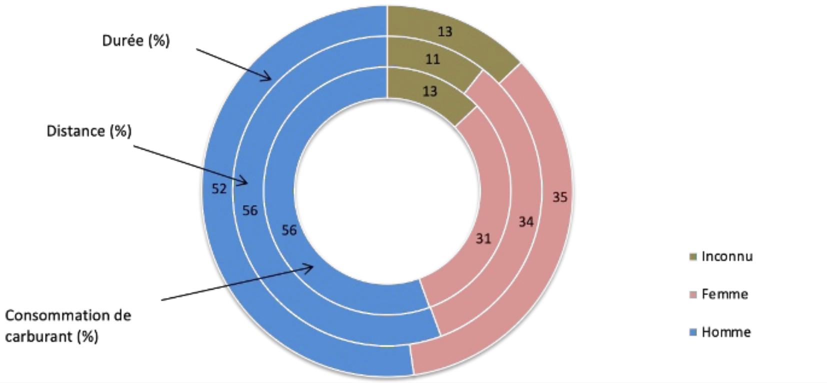
Distance, passager-km parcouru, durée et consommation de carburant par occupation



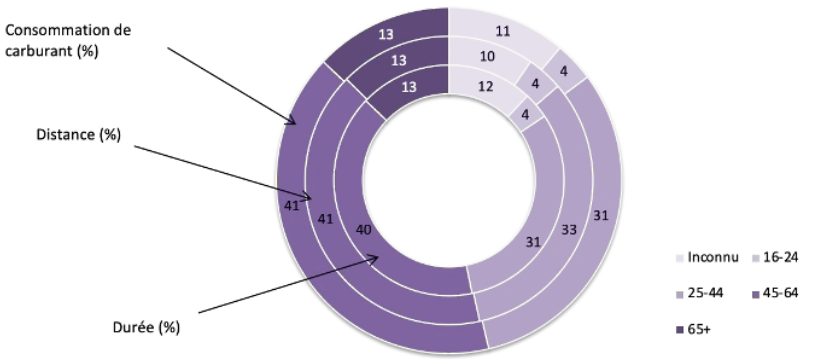
Proportion de déplacements par segments de distance



Durée, distance et consommation de carburant par sexe

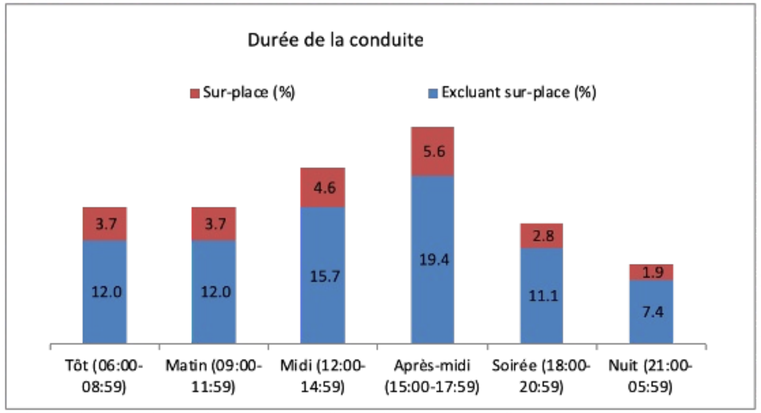


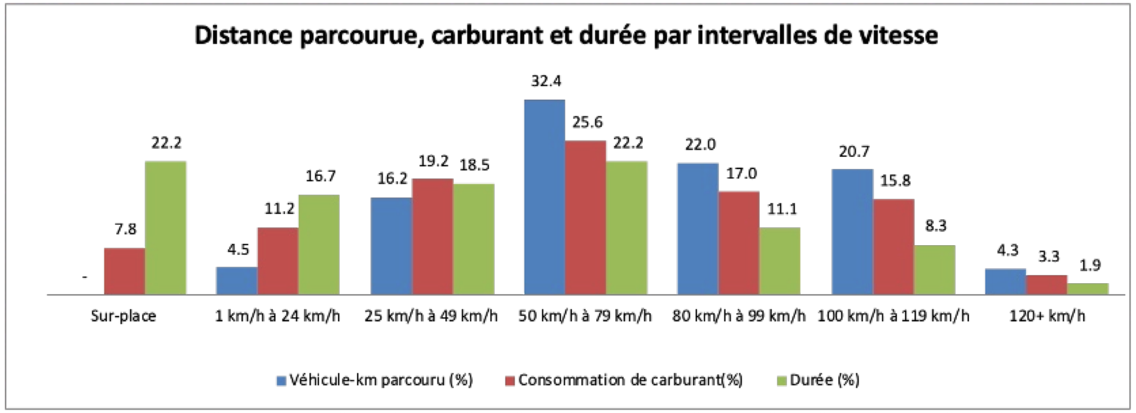
Consommation de carburant, distance et durée par âge des conducteurs

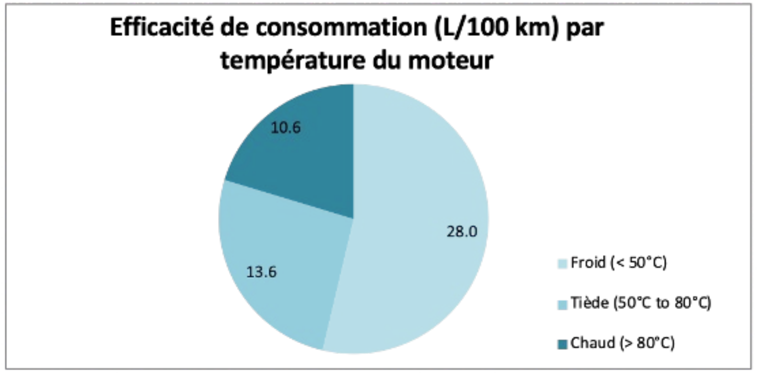


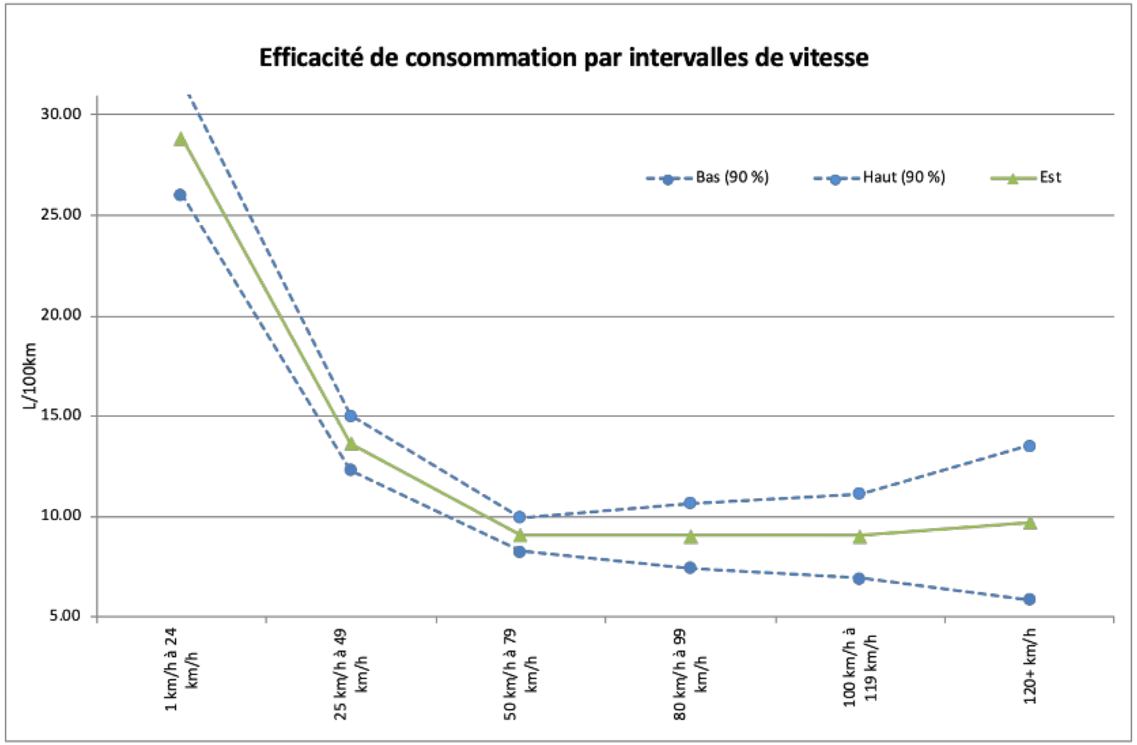
Ontario – 1er trimestre 2012

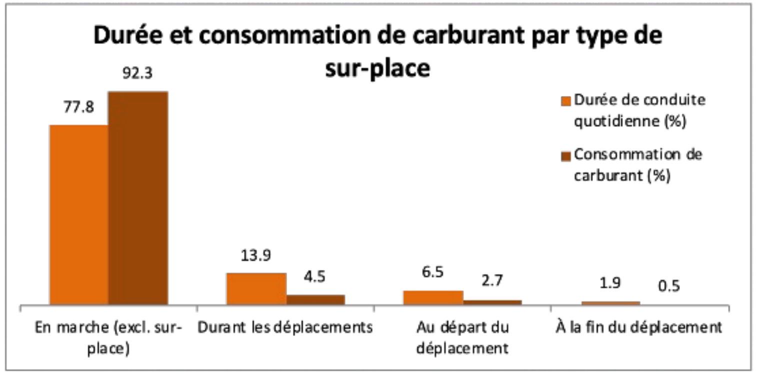
Sous-caractéristiques des déplacements


Durée de la conduite


Distance parcourue, carburant et durée par intervalles de vitesse

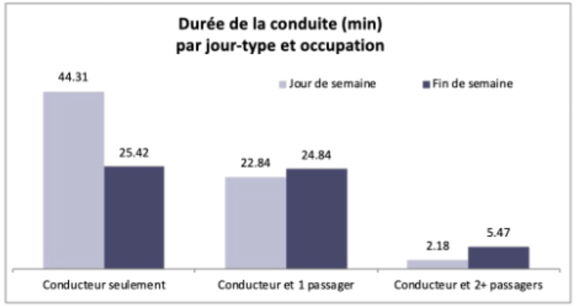

Efficacité de consommation (L/100 km) par température du moteur

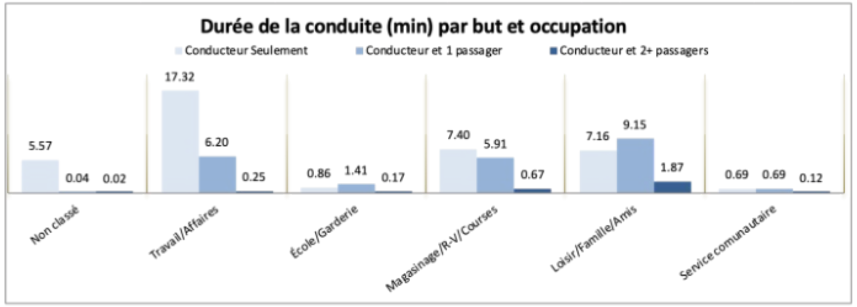

Efficacité de consommation par intervalles de vitesse

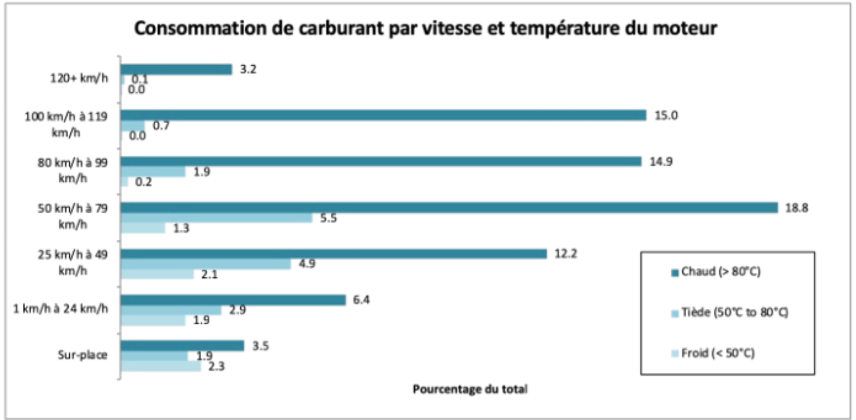

Durée et consommation de carburant par type de sur-place

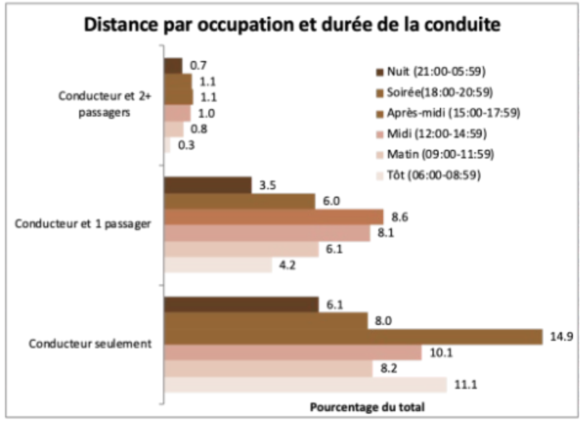
Ontario – 1er trimestre 2012

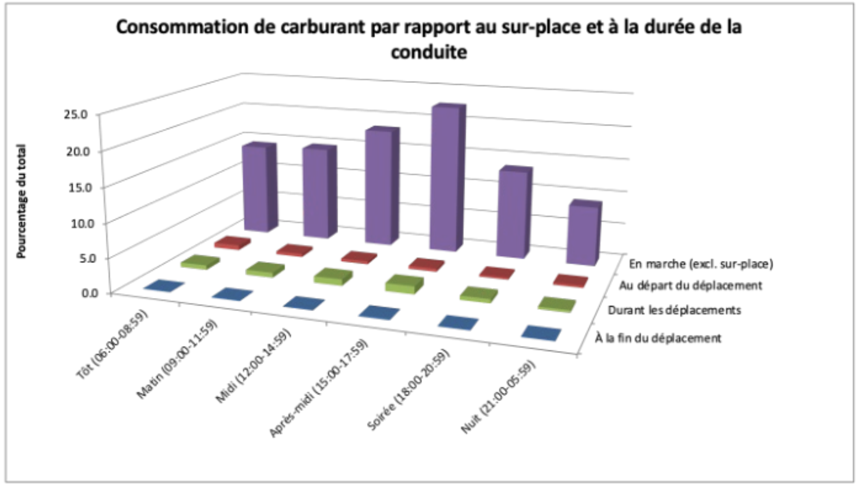
Caractéristiques mixtes sur les déplacements


Durée de la conduite (min) par jour-type et occupation


Durée de la conduite (min) par but et occupation


Consommation de carburant par vitesse et température du moteur


Distance par occupation et durée de la conduite


Consommation de carburant par rapport au sur-place et à la durée de la conduite

Exercise

In teams or individually, identify a scenario for which a dashboard could prove useful.

Determine specific questions that the dashboard could help answer or insights that it could provide.

Identify data sources and data elements that could be fed into your dashboard.

Design a display (with pen and paper) with mock charts.

What are the strengths and limitations of your dashboard? Is it functional? Elegant?

Roundtable: Data Viz Posters

Hits?

Misses?

Thoughts?

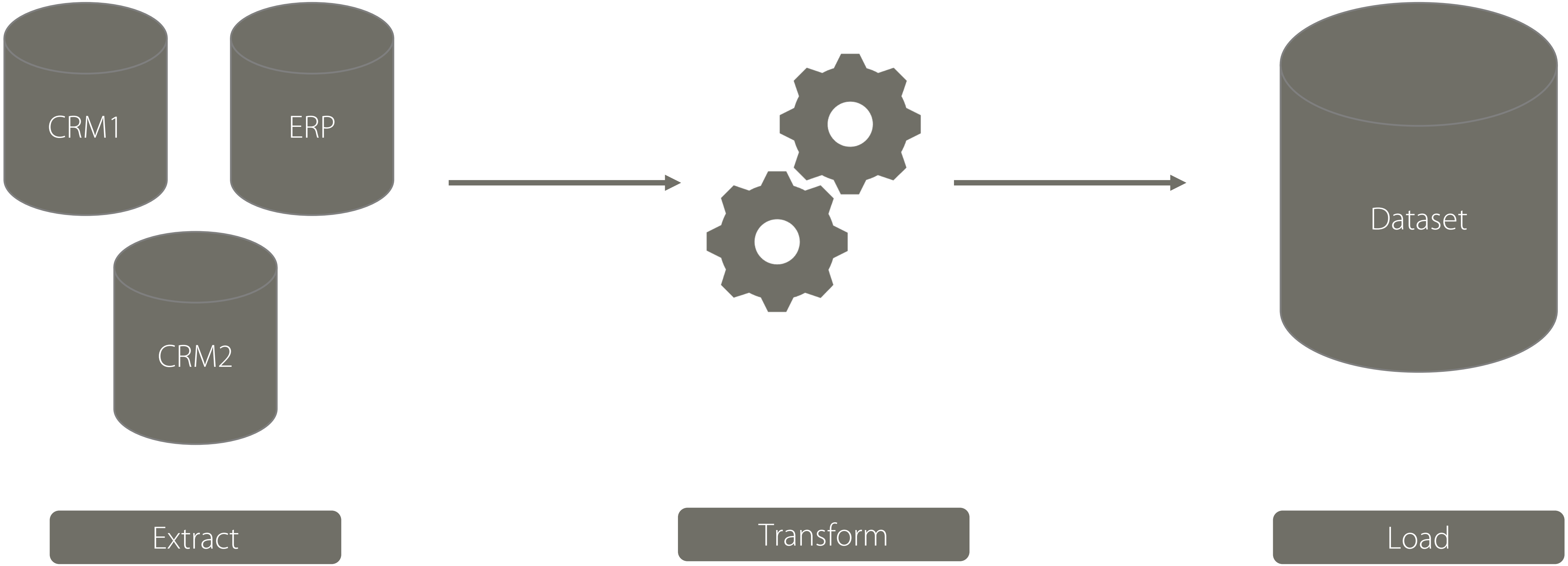
Gapminder Exercises

Do the exercises for Module 3.

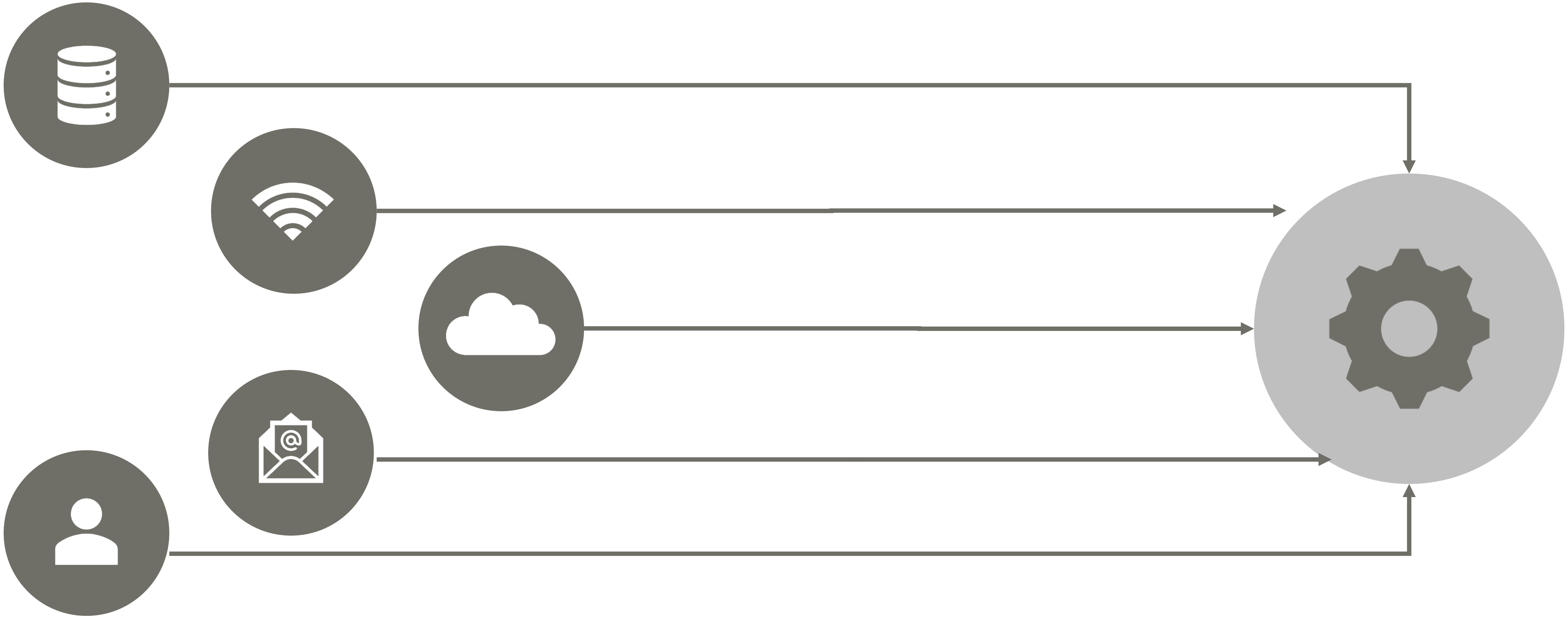
Module 4

Data Processing and Data Cleaning

ETL



Extract



Extract
Transform
Load

Transform

Changing Structure

Altering Data Types

Aggregating Data

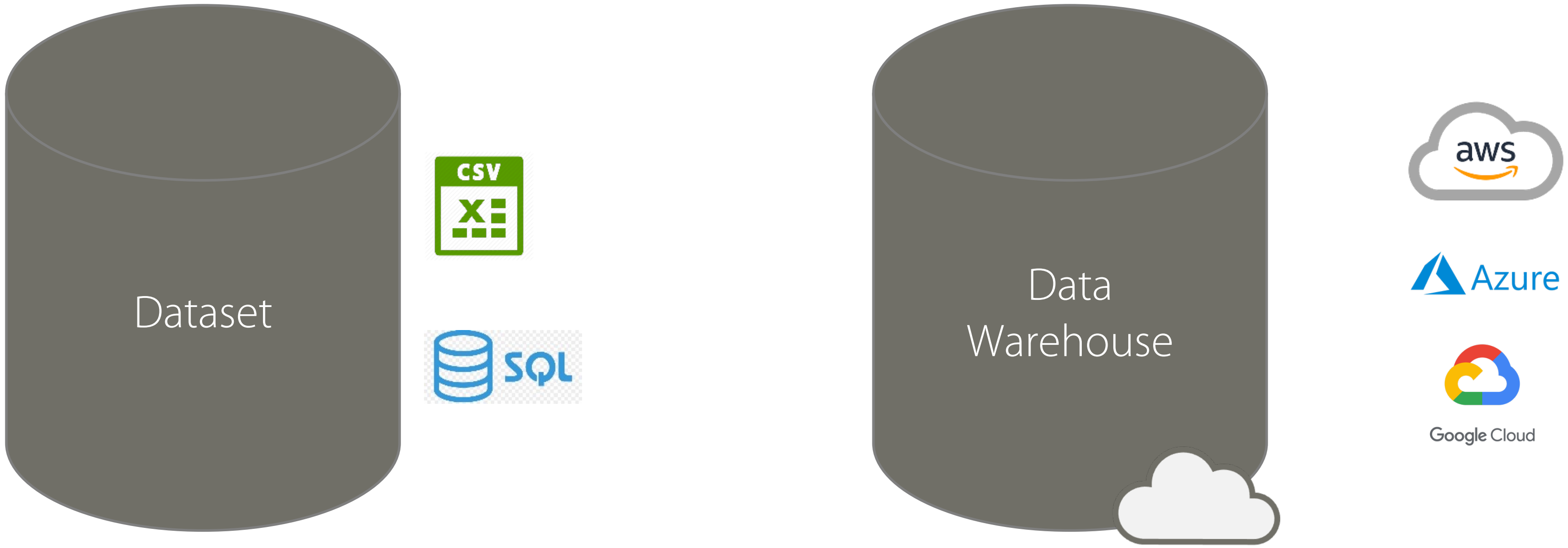
Cleaning

Joining

Grouping

Extract
Transform
Load

Load


Dataset

Data Warehouse

Extract → Transform → Load

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.

Pros and Cons

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

Tools and Methods

Methodical

- list of potential problems (Data Cleaning Bingo)
- code which can be re-used in different contexts

Narrative

- visualization
- data summary
- distribution tables
- small multiples
- data analysis

Data Cleaning Bingo

random missing values	outliers	values outside of expected range - numeric	factors incorrectly/iconsistently coded	date/time values in multiple formats
impossible numeric values	leading or trailing white space	badly formatted date/time values	non-random missing values	logical inconsistencies across fields
characters in numeric field	values outside of expected range - date/time	DCB!	inconsistent or no distinction between null, 0,not available, not applicable,missing	possible factors missing
multiple symbols used for missing values	???	fields incorrectly separated in row	blank fields	logical iconsistencies within field
entire blank rows	character encoding issues	duplicate value in unique field	non-factor values in factor	numeric values in character field

Approaches to Data Cleaning

The narrative approach is akin to working out a crossword puzzle with a pen and putting down potentially wrong answers **occasionally**, to see where that takes you.

The mechanical approach is akin to working it out with a pencil, a dictionary, and never jotting down an answer unless you are certain it is correct.

You'll solve more puzzles (and it will be flashier) the first way, but you'll rarely be wrong the second way.

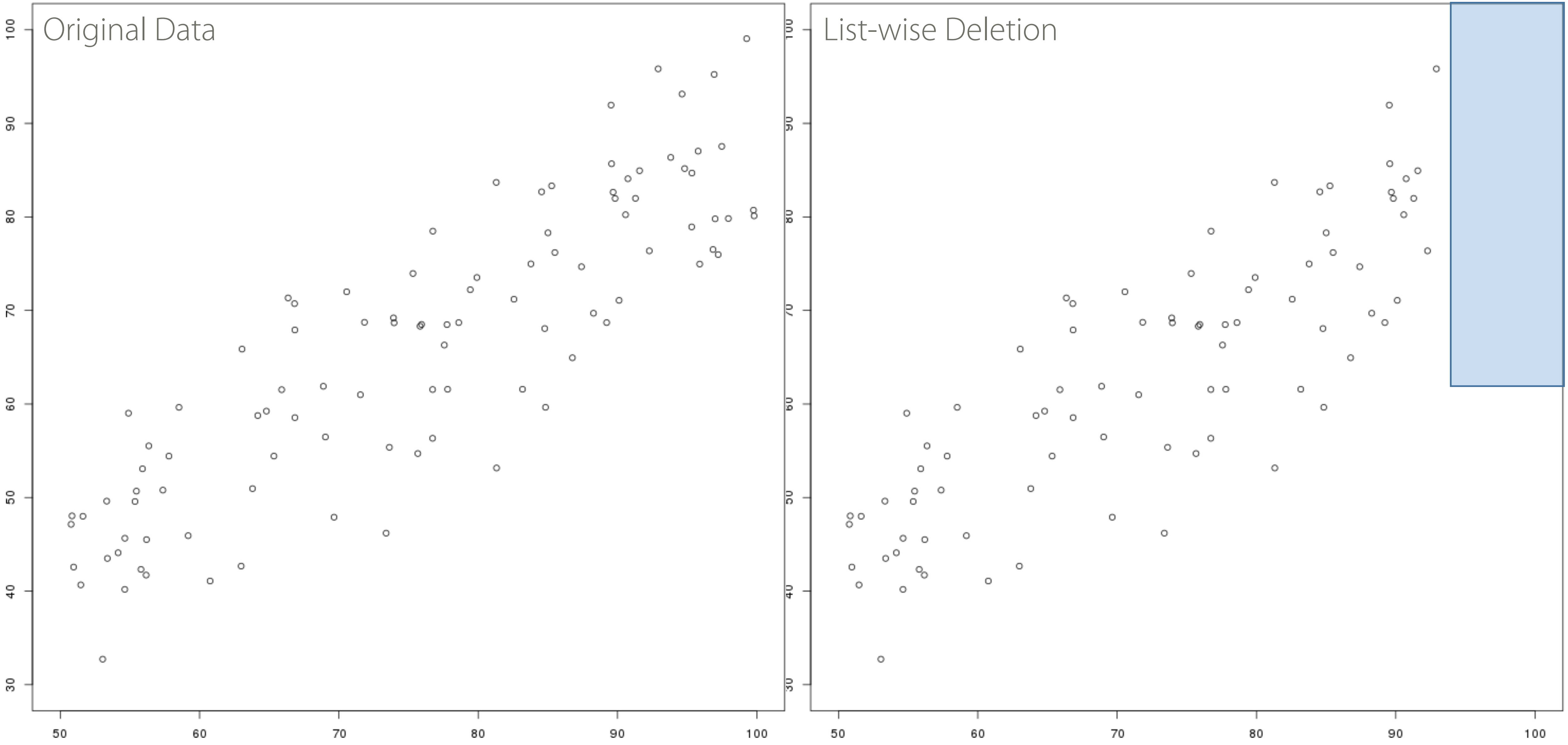
It's the same thing with data: analysts must be comfortable **with both approaches**.

The Case for Imputation

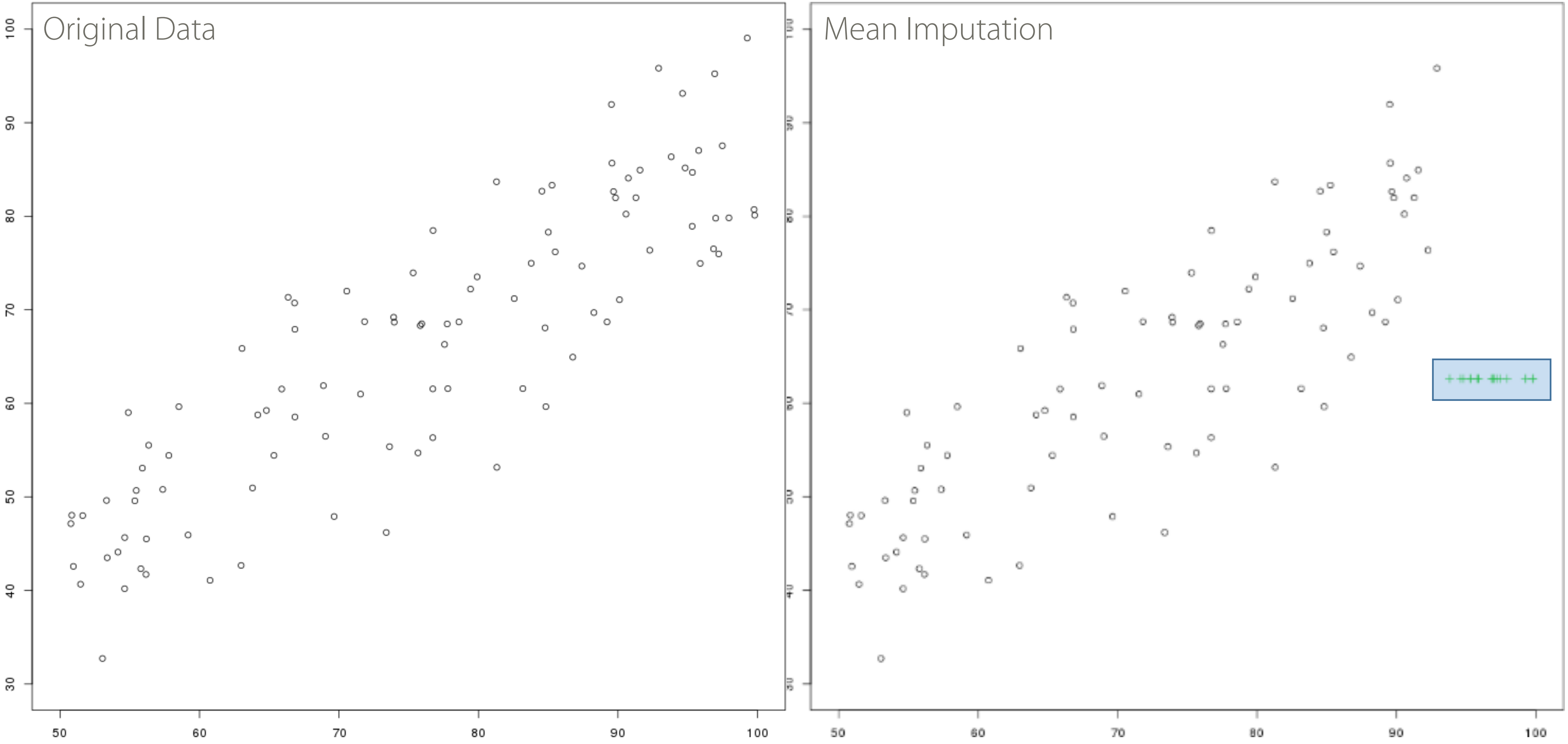
Not all analytical methods can easily accommodate missing observations – 2 options:

- **Discard** the missing observation
  - not recommended, unless the data is missing completely randomly in the dataset
  - acceptable in certain situations (small number of missing values in a large dataset)
- Establish a **replacement (imputation) value**
  - main drawback: we never know what the true value would have been
  - often the best available option

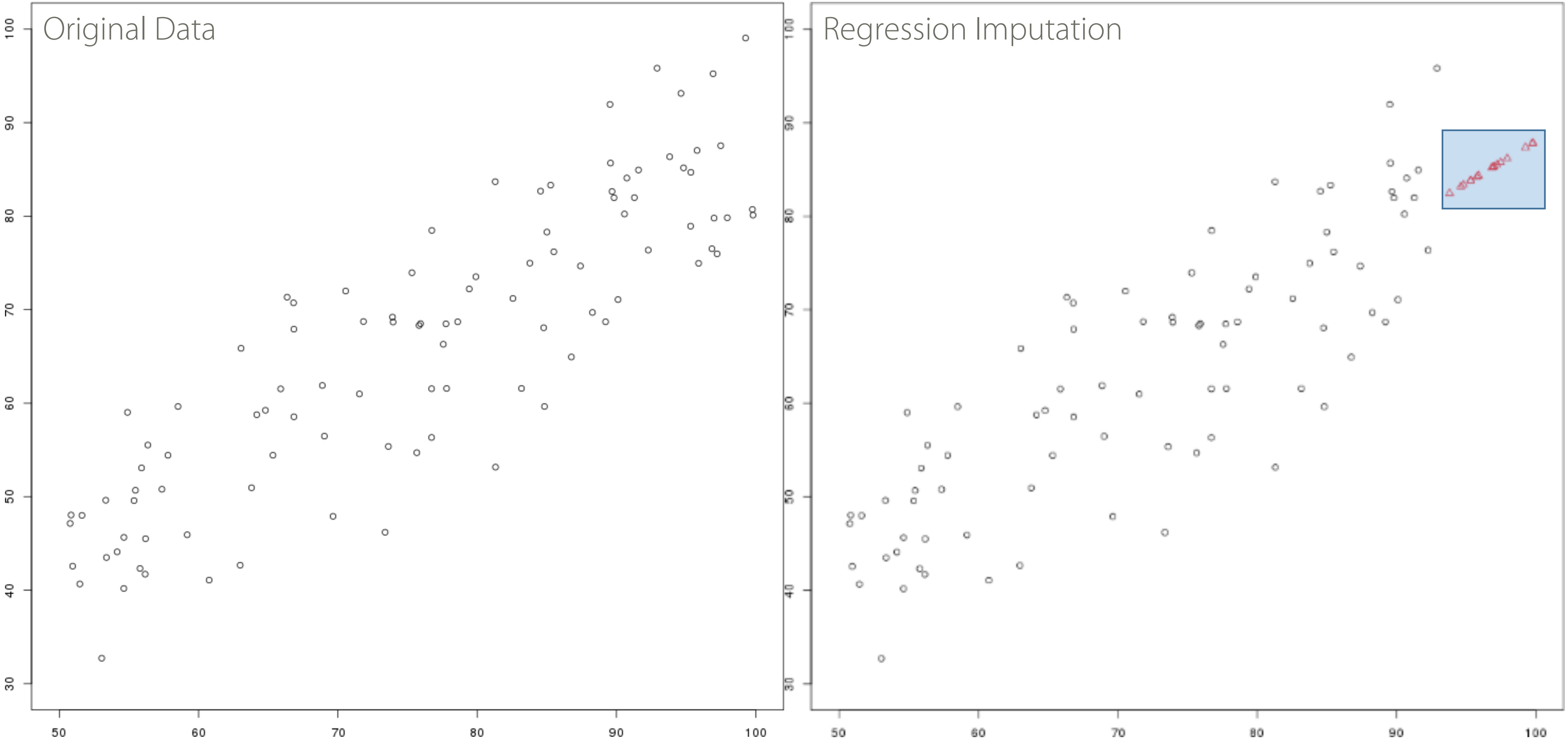
Artificial data: the y values of all points for which $x > 92$ have been erased by mistake.



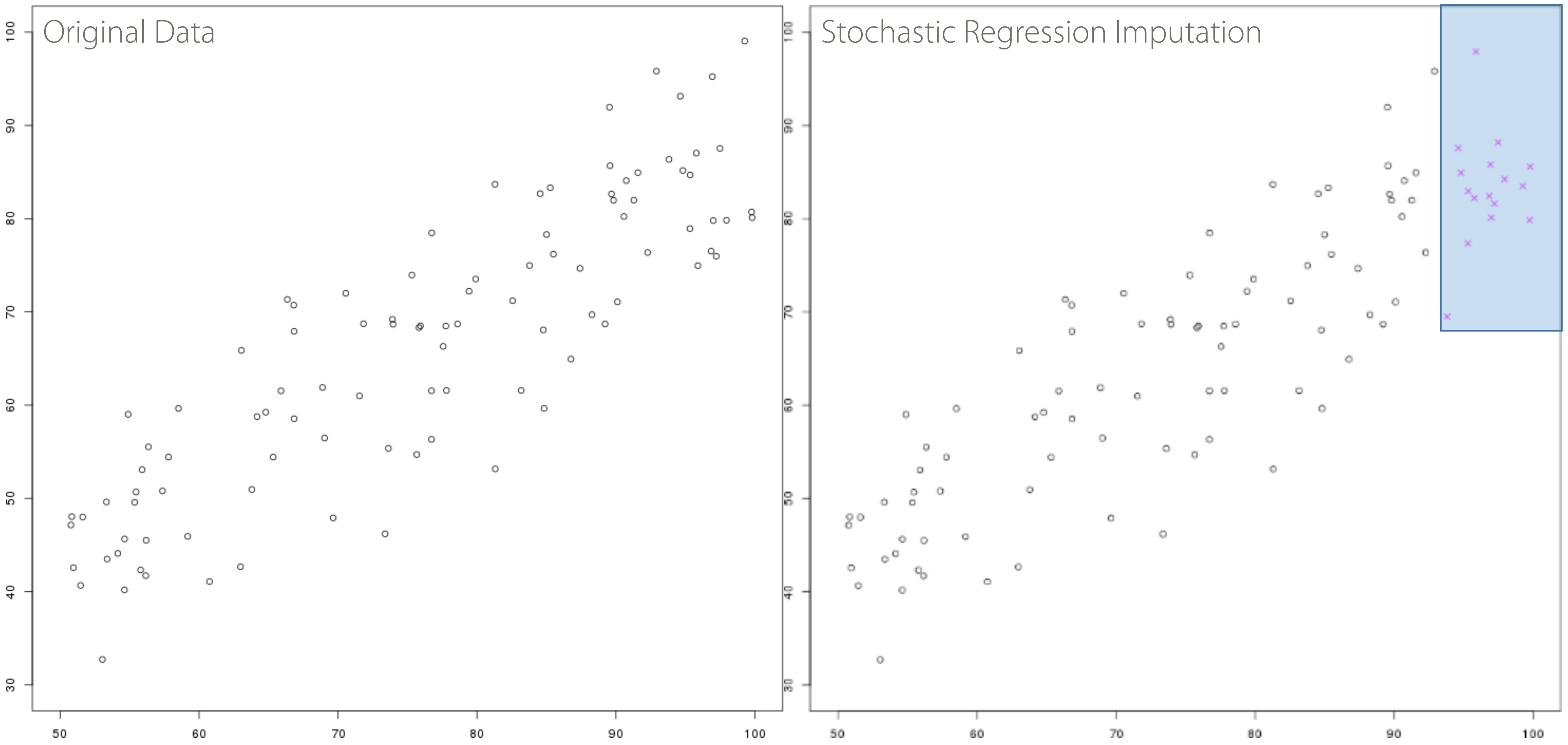
Artificial data: the y values of all points for which $x > 92$ have been erased by mistake.



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.



Imputation Take-Aways

Missing values cannot simply be ignored.

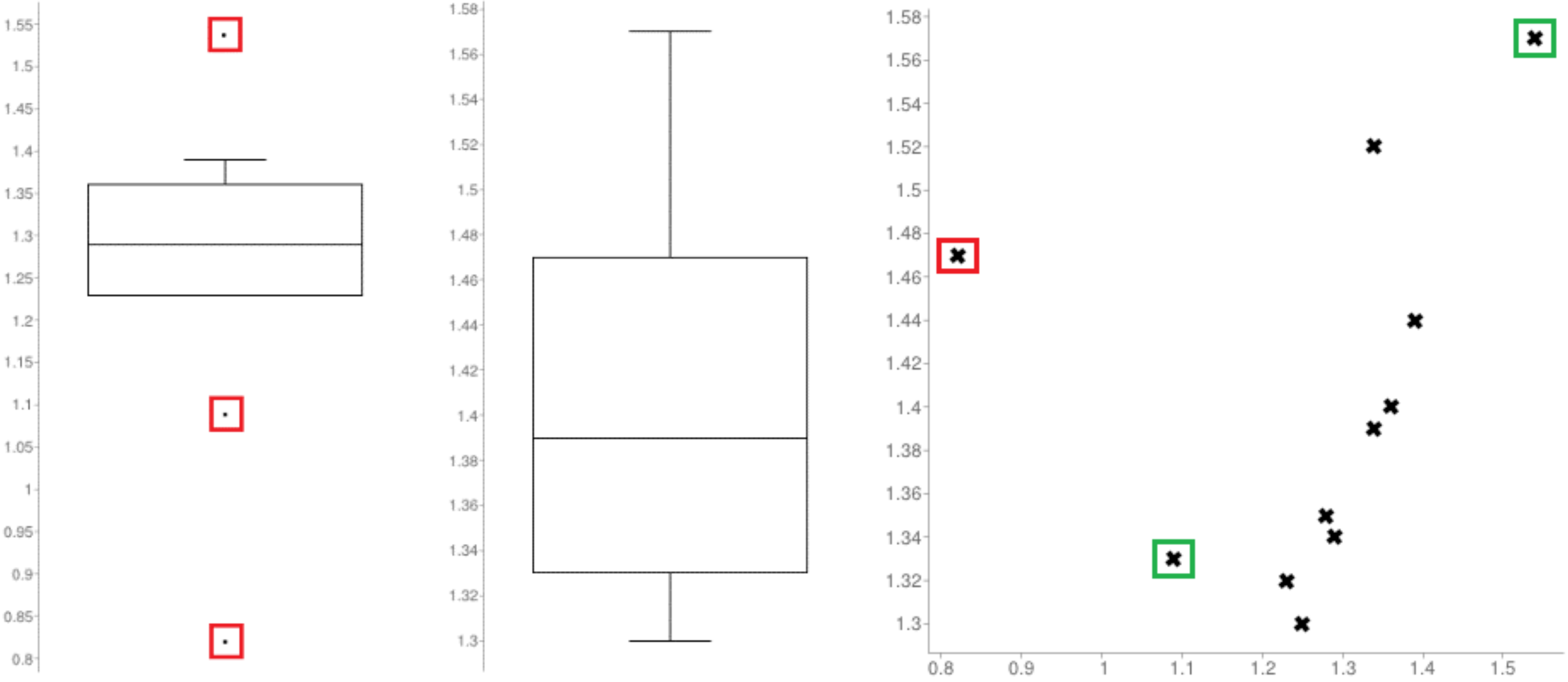
The missing mechanism cannot typically be determined with any certainty.

Imputation methods work best when values are missing completely at random or missing at random, 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.

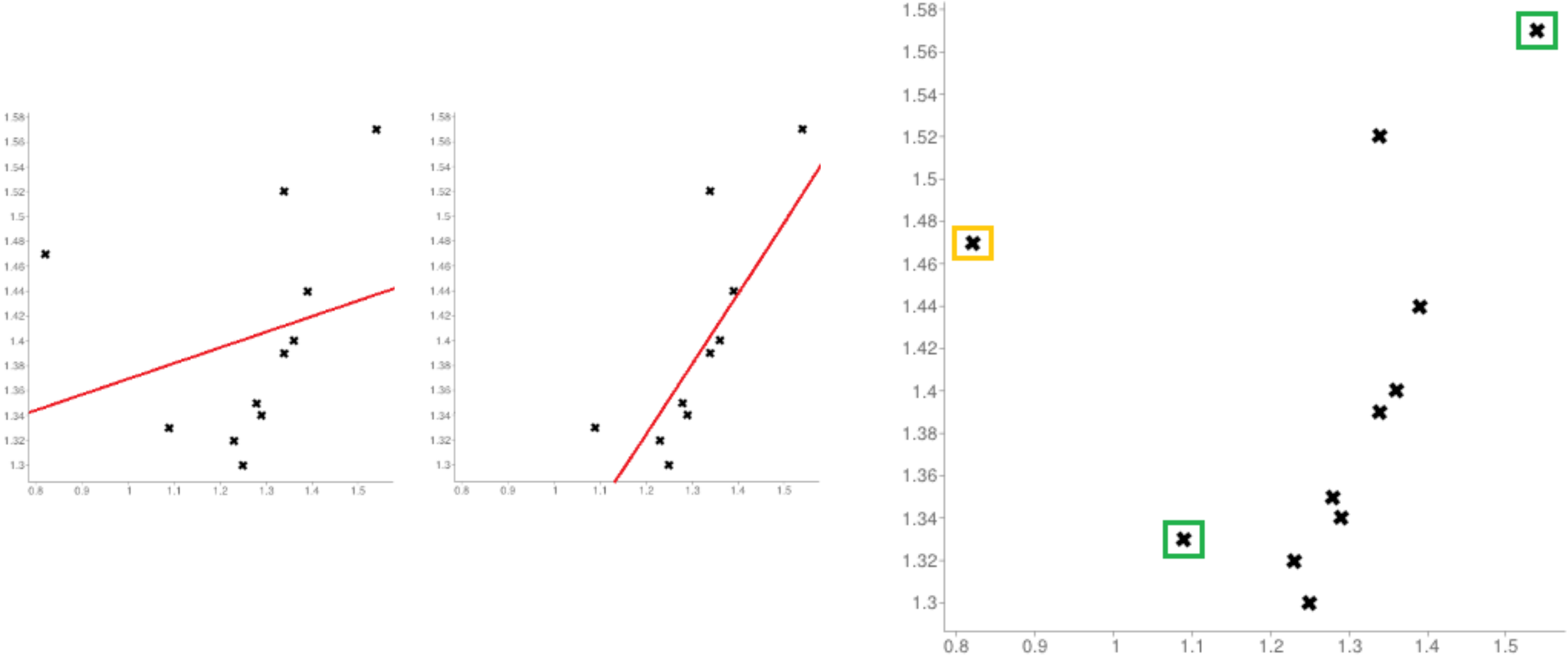
Is stochastic imputation best? In our example, yes – but **No-Free Lunch theorem**!

Outliers



Queuing dataset: processing rate vs. arrival rate

Influential Observations



Queuing dataset: processing rate vs. arrival rate

Feature Selection

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

Motivations:

- modeling tools do not handle these well (variance inflation due to multi-colinearity, etc.)
- dimension reduction (# variables ≫ # observations)
- mitigating the Curse of Dimensionality

Approaches:

- filter vs. wrapper
- unsupervised vs. supervised

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 could 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

Sound Data

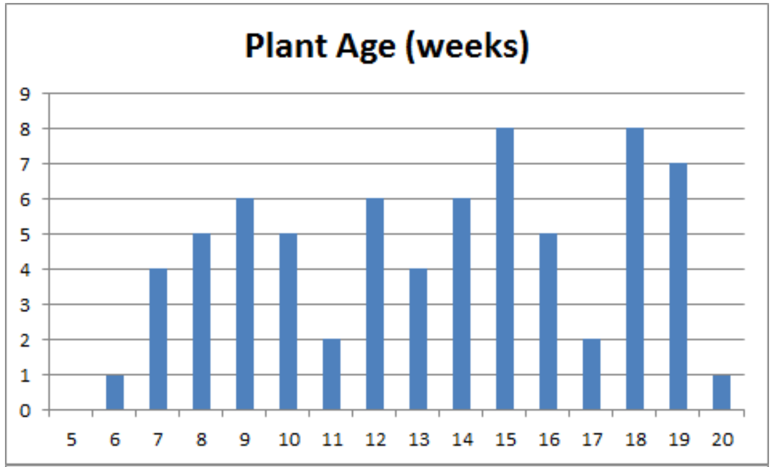
The ideal dataset will have as few issues as possible with:

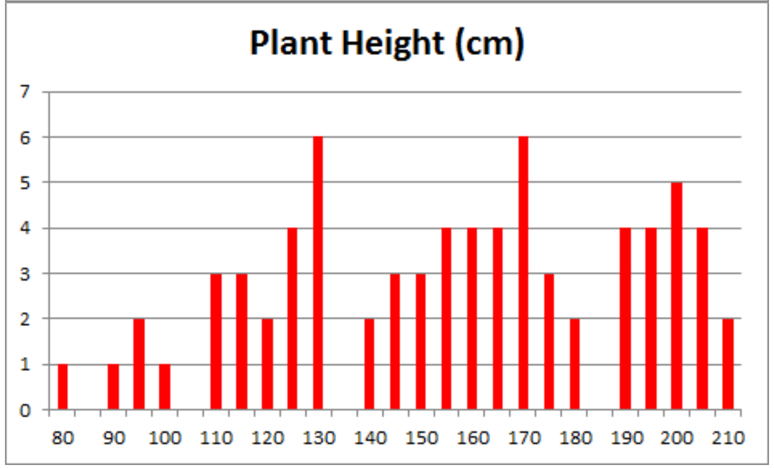
- **validity:** data type, range, mandatory response, uniqueness, value, regular expressions
- **completeness:** missing observations
- **accuracy and precision:** related to measurement and/or data entry errors; target diagrams (accuracy as bias, precision as standard error)
- **consistency:** conflicting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

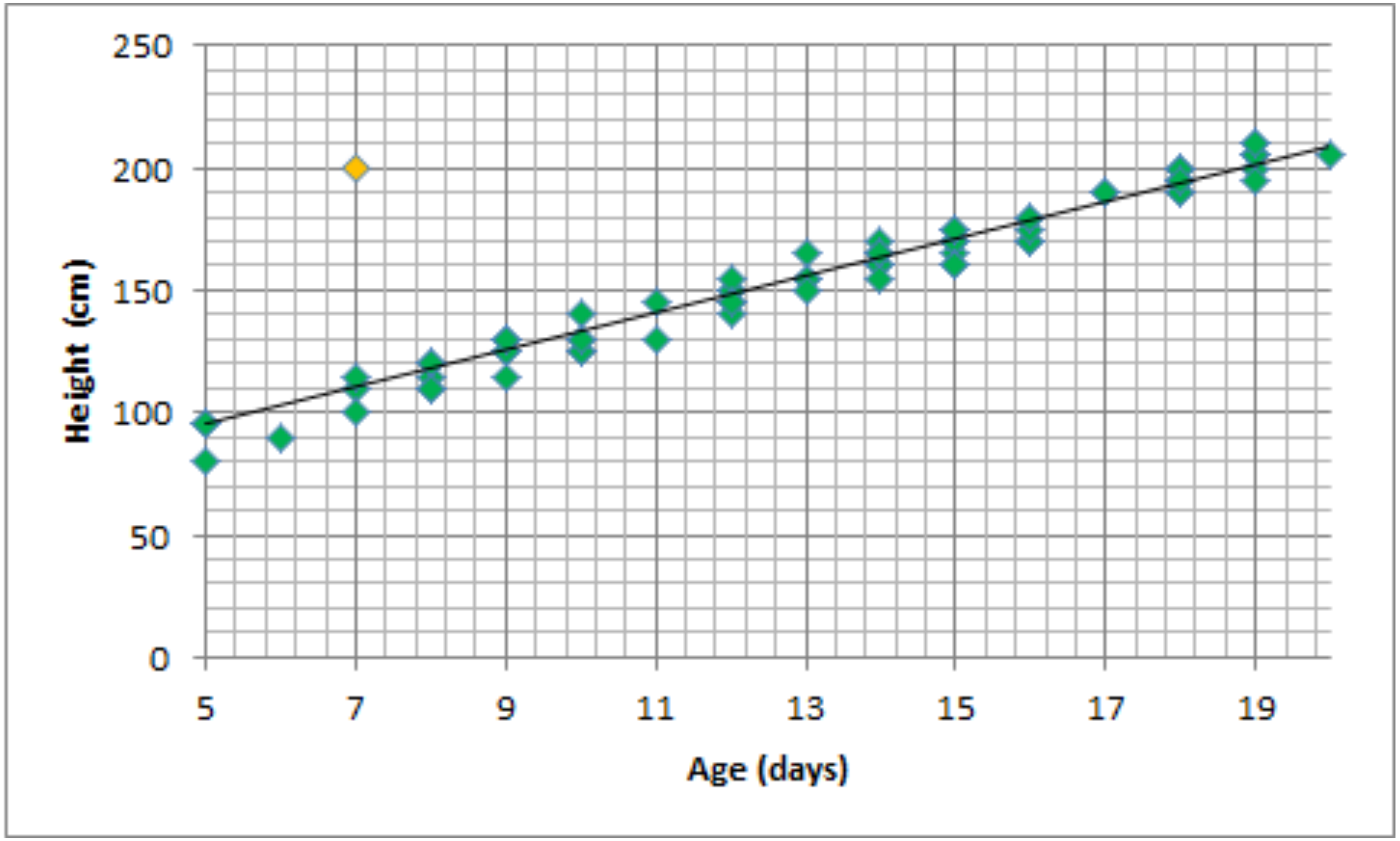
Checking for data quality issues **early** can save headaches at a later analytical stage.

[Personal file]

Detecting Invalid Entries

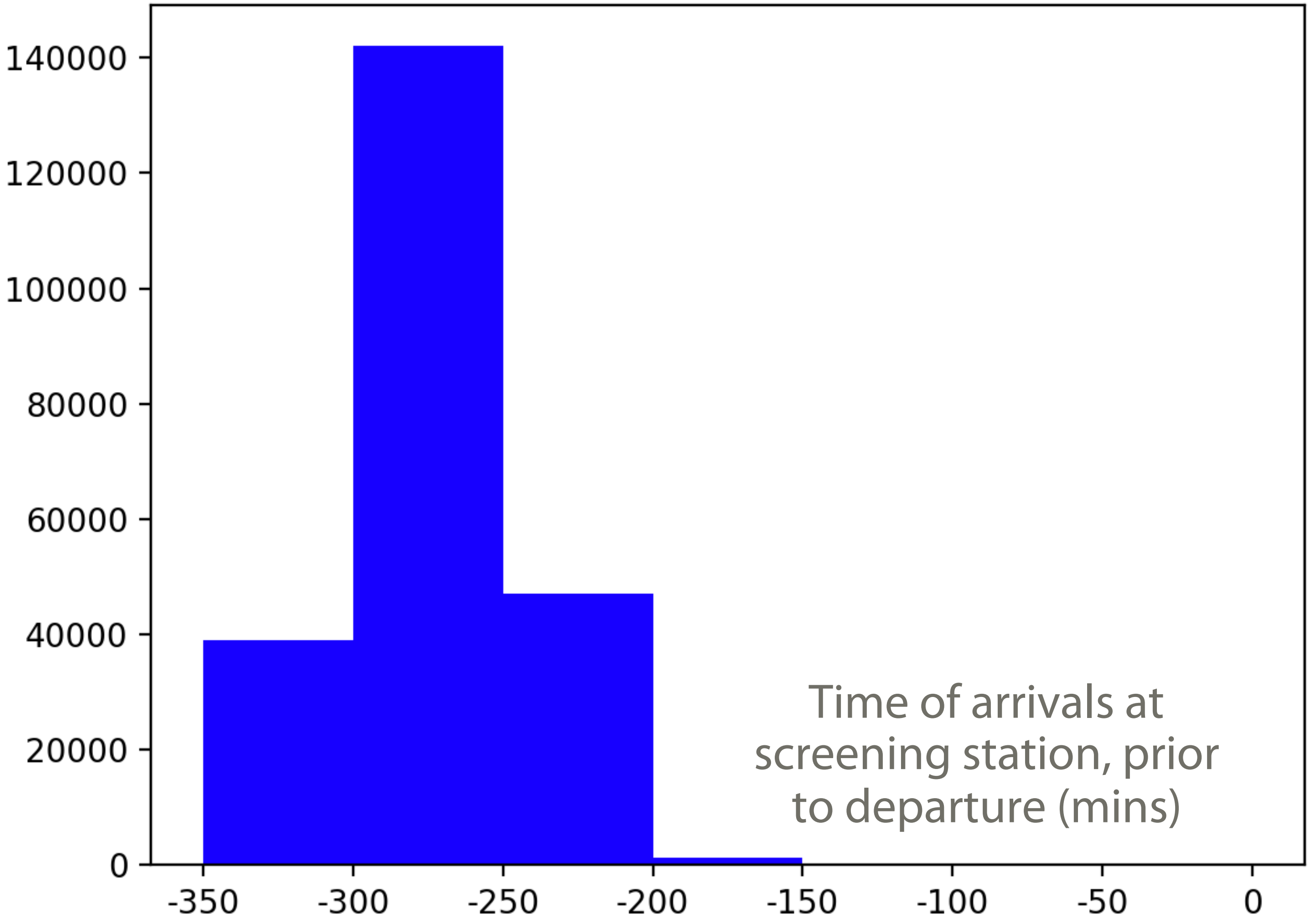






[Personal file]

Detecting Invalid Entries



Time of arrivals at screening station, prior to departure (mins)

Data Quality Take-Aways

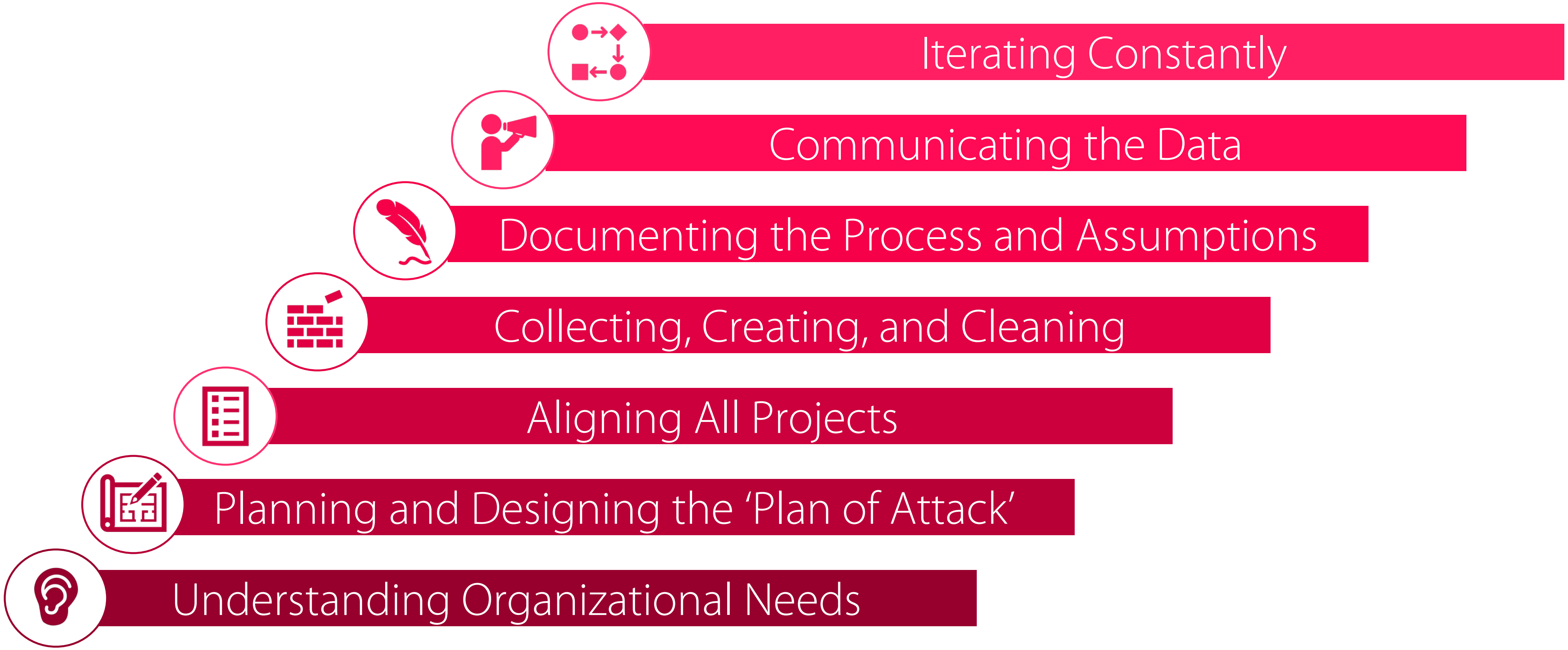
Don't wait until **after** the analysis to find out there was a problem with data quality.

Univariate tests don’t always tell the whole story.

Visualizations can help.

Context is crucial – you may need more context about the data in order to make sense of what you see… but whatever the situation, you need to understand the data quality.

Putting it All Together



Iterating Constantly

Communicating the Data

Documenting the Process and Assumptions

Collecting, Creating, and Cleaning

Aligning All Projects

Planning and Designing the ‘Plan of Attack’

Understanding Organizational Needs

Gapminder Exercises

Do the exercises for Module 4.

Module 5

Data Exploration and Data Analysis

Exploratory Data Analysis (Big Picture)

It is important to understanding what the data looks like before conducting analyses

EDA = Visualize + **Compute** Basic Statistics

Intuition for Data Analysis

Why are some people ‘poor’?

Intuition for Data Analysis

Surface level answer:
Because they don’t have money

Intuition for Data Analysis

- Inequality
- Wars & Conflict
- Food & Water
- Education
- Nature of Work

Intuition for Data Analysis

Assuming vs Analyzing

- Inequality
- Wars & Conflict
- Food & Water
- Education
- Nature of Work

Core Process

Inspect

Clean

Verify

Report

Source: Medium

Inspect – Criteria for 'Letting Go'

1 'Sparse' Irrelevant Data

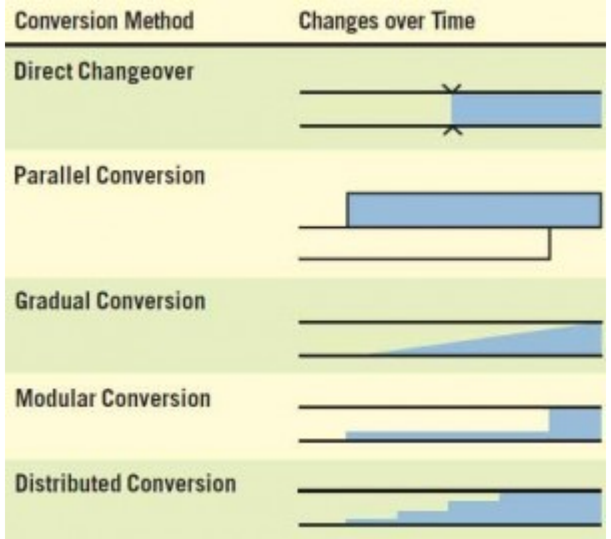
2 System or Standard Shifts

3 Discontinued Data Flows

4 Severe Gaps in Continuous Data

5 Inconsistencies Beyond Repair

6 'Junk' Data from System Migrations



Conversion Method	Changes over Time
Direct Changeover	
Parallel Conversion	
Gradual Conversion	
Modular Conversion	
Distributed Conversion	

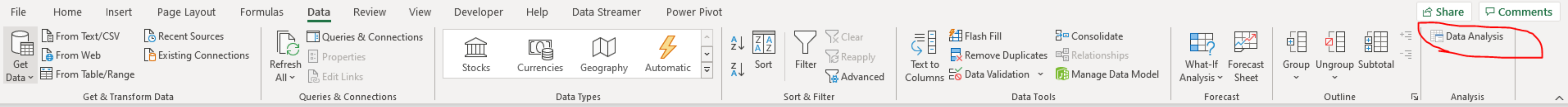
Source: w3 Computing

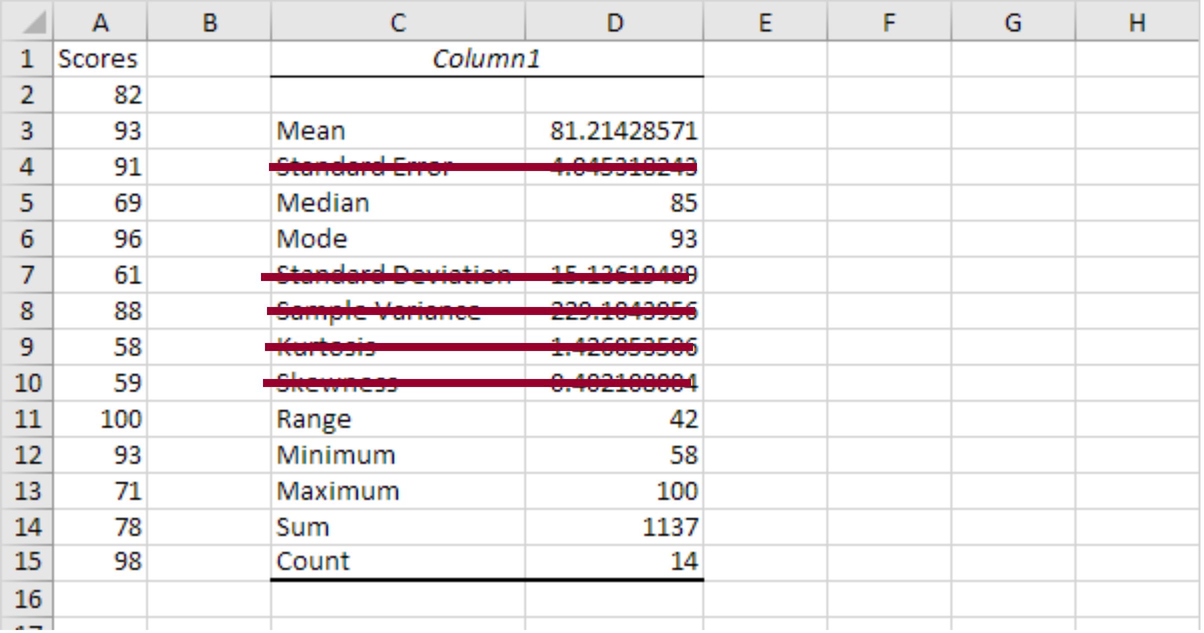
Clean – Tips for Cleaning & Salvaging

1. Alter Data Types
2. Set Range Constraints
3. Create Non-Blank Restrictions
4. Implement Cross-Field Rules
5. Remove Duplicates
6. Normalize Data
7. Fix Typos
8. Impute Missing Values

Source: Medium

Verify – Previewing Numerical Data





	A	B	C	D	E	F	G	H
1	Scores		*Column1*					
2	82							
3	93		Mean	81.21428571				
4	91		~~Standard Error~~	~~4.045318243~~				
5	69		Median	85				
6	96		Mode	93				
7	61		~~Standard Deviation~~	~~15.13619489~~				
8	88		~~Sample Variance~~	~~229.1043956~~				
9	58		~~Kurtosis~~	~~1.426053506~~				
10	59		~~Skewness~~	~~0.402108004~~				
11	100		Range	42				
12	93		Minimum	58				
13	71		Maximum	100				
14	78		Sum	1137				
15	98		Count	14				
16								

Verify – Data Scavenger Hunting

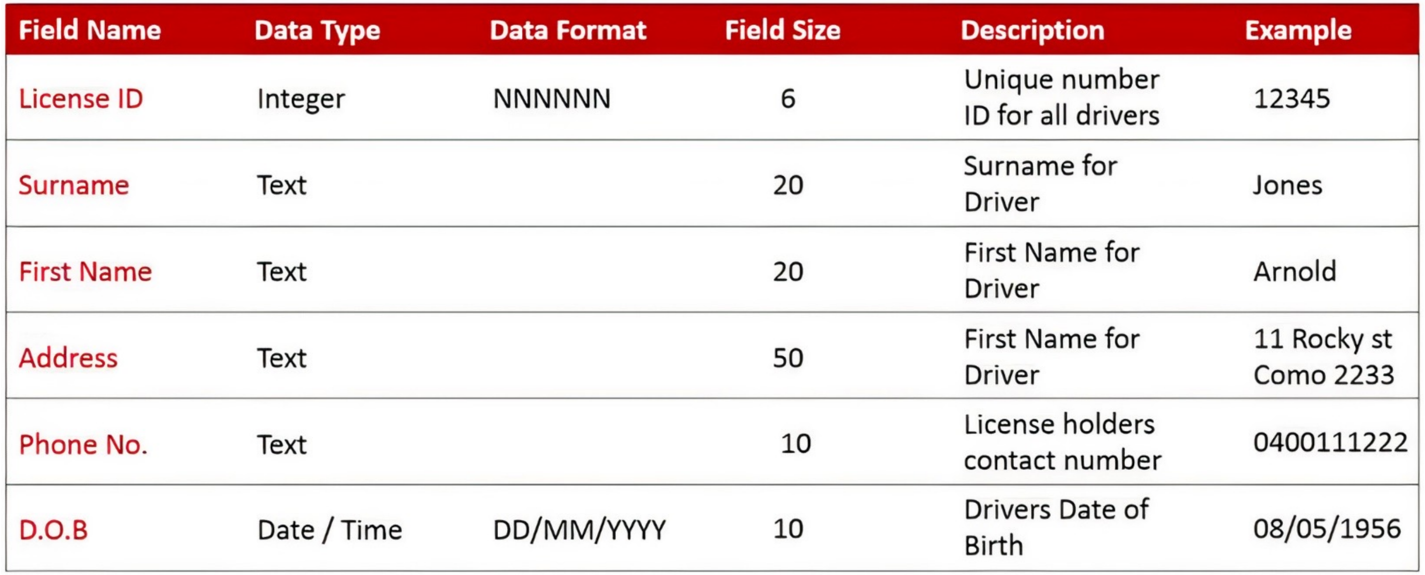
1. Use ‘Look-alike’ Data

2. Leverage ‘Open’ Datasets

3. Create ‘Synthetic’ Data

4. Extrapolate Data if Statistically Sig.

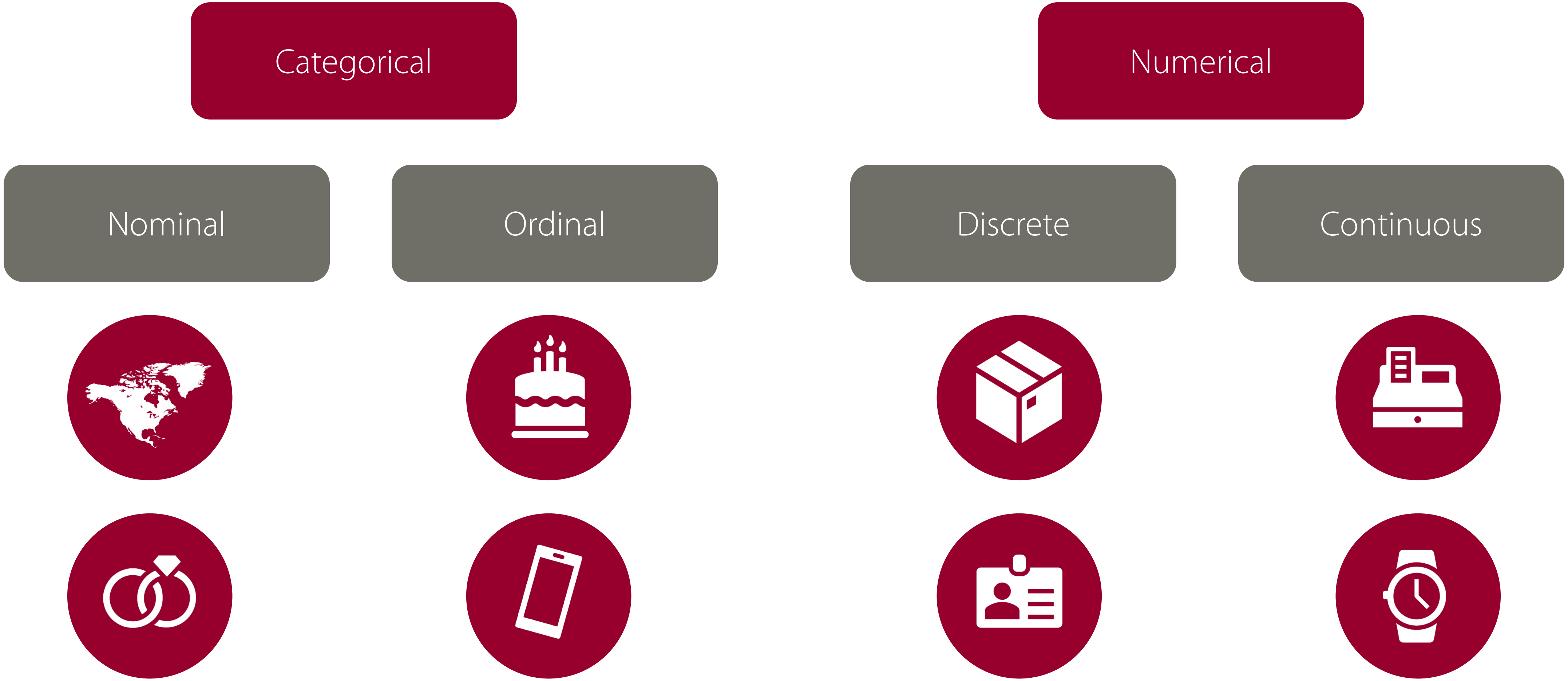
Report – Create a Data Dictionary



Field Name	Data Type	Data Format	Field Size	Description	Example
License ID	Integer	NNNNNN	6	Unique number ID for all drivers	12345
Surname	Text		20	Surname for Driver	Jones
First Name	Text		20	First Name for Driver	Arnold
Address	Text		50	First Name for Driver	11 Rocky st Como 2233
Phone No.	Text		10	License holders contact number	0400111222
D.O.B	Date / Time	DD/MM/YYYY	10	Drivers Date of Birth	08/05/1956

Source: Medium

Types of Data



Categorical

Nominal

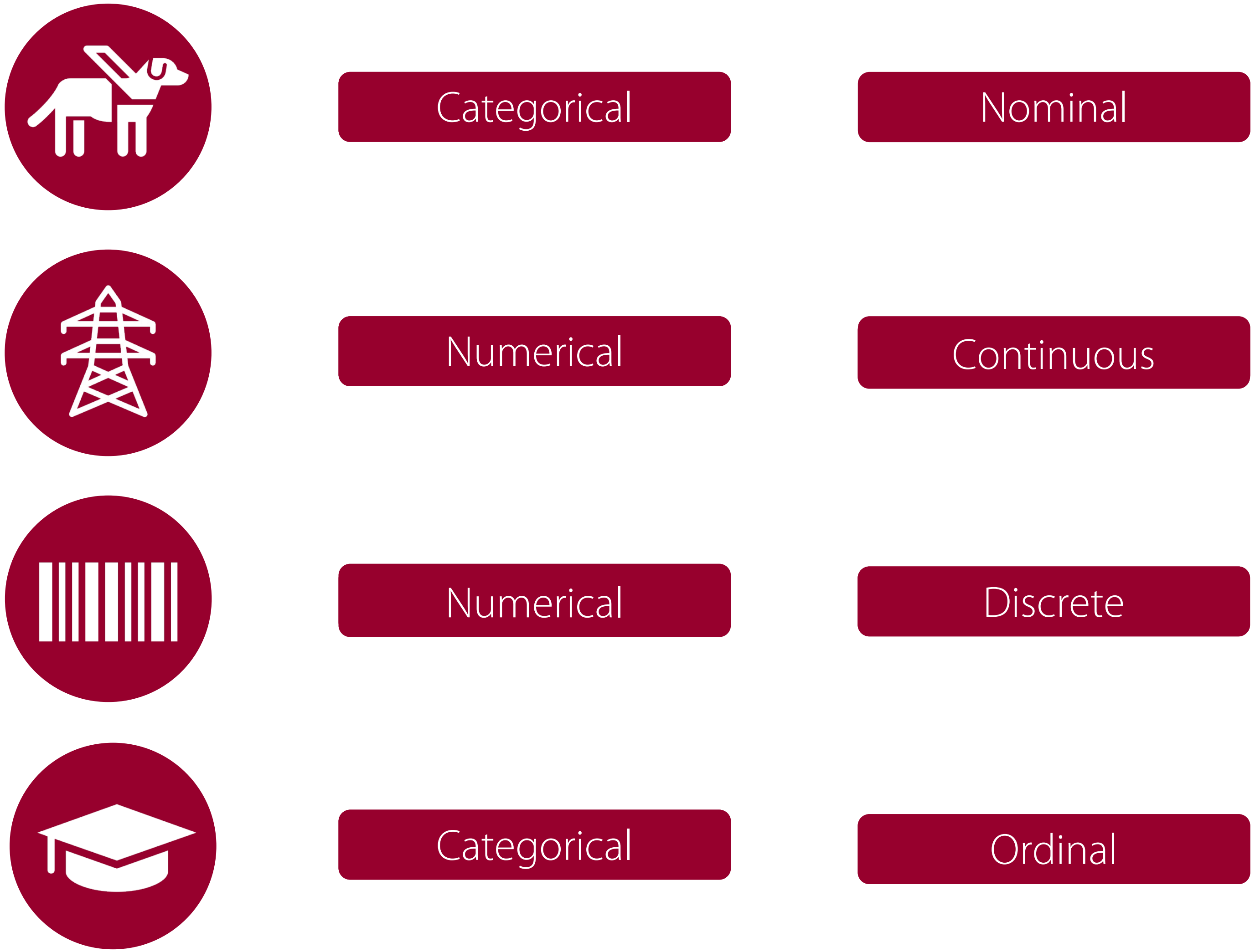
Ordinal

Numerical

Discrete

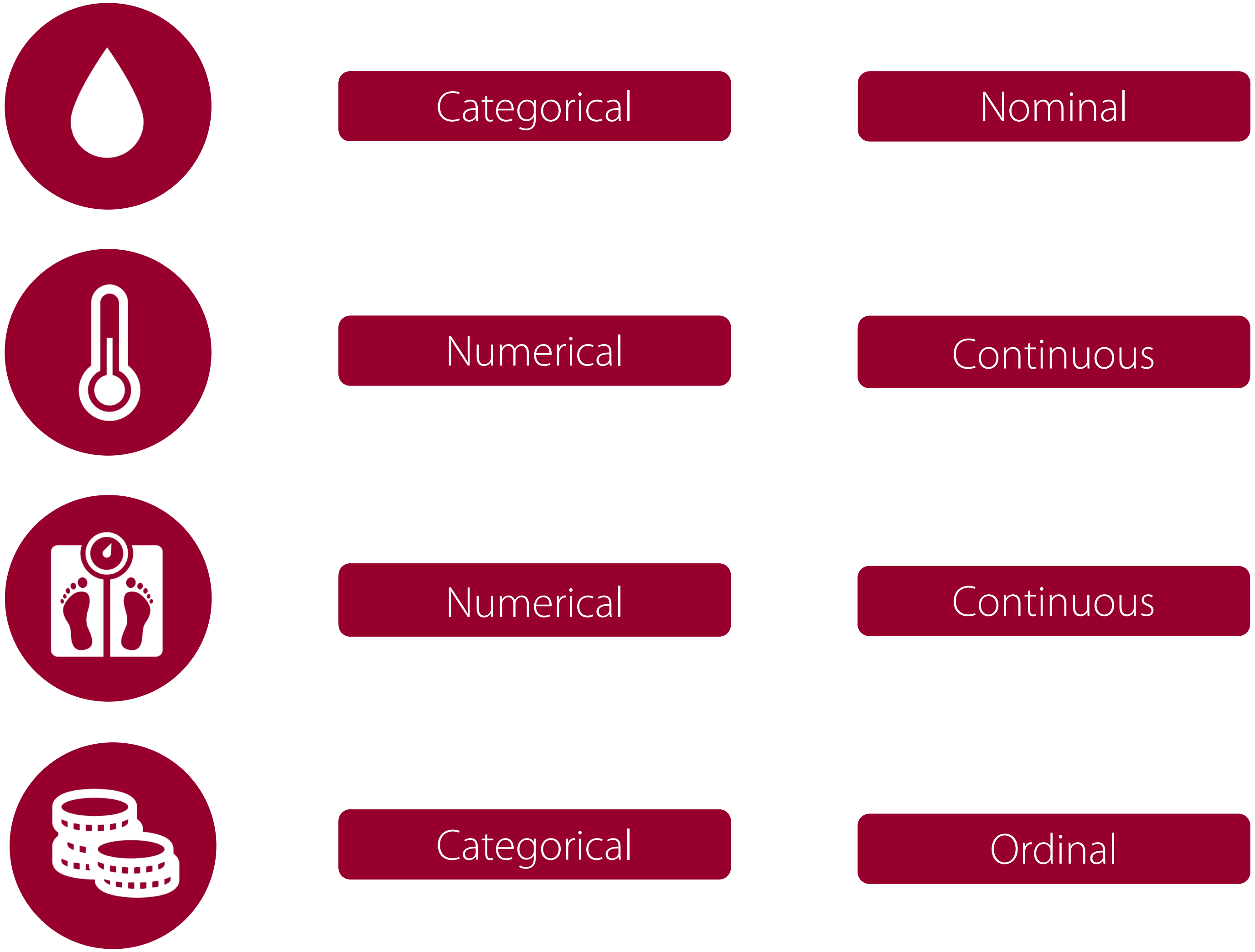
Continuous

Types of Data



	Categorical	Nominal
	Numerical	Continuous
	Numerical	Discrete
	Categorical	Ordinal

Types of Data



Categorical — Nominal

Numerical — Continuous

Numerical — Continuous

Categorical — Ordinal

Special Role of Categorical Data

Categorical data plays a special role:

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

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

Data Summarizing

Min: smallest value

Max: largest value

Median: “middle” value

Mode: most frequent value

Unique Values: list of unique values

etc.

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

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

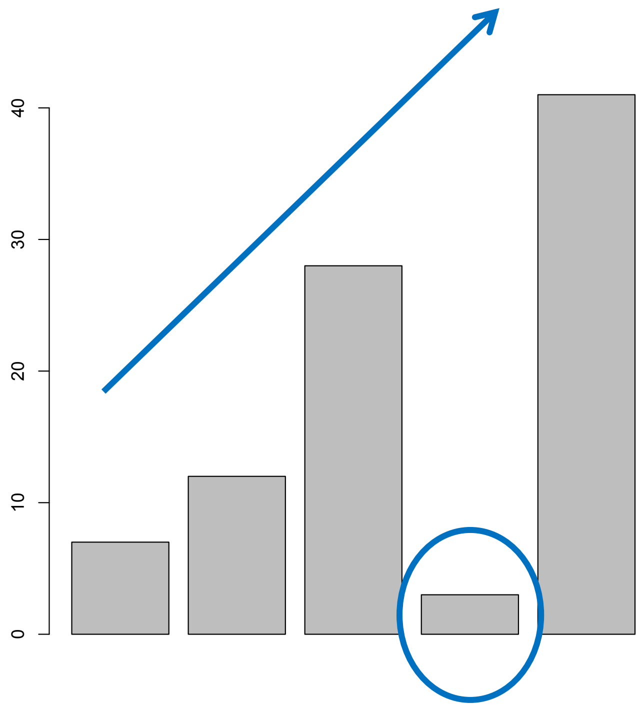
Analysis (broad definition):

- 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.

We will discuss further.



Descriptive Statistics

1 Mean
2 Median
3 Mode

Central Tendency

4 Standard Deviation
5 Range
6 Interquartile Range

Variability

7 Frequency
8 Percentage
9 Proportion

Nominal Metrics

Nominal Data

Frequencies

Count number of events

Proportion

Divide frequency by total number of events

Percentage

Multiply proportion by 100

Descriptive Statistics – Central Tendency

Mean

Sum of values divided by the number of observations.

Median

Number that divides the data set in half.

Mode

The most frequently occurring number.

$7 \; \{ \; 2, 1, 6, 4, 5, 5, 5$

↓

$\boxed{4} \qquad 28$

$2, 1, 6, 4, 5, 5, 5$

↓

$1, 2, 4, \boxed{5}, 5, 5, 6$

$2, 1, 6, 4, 5, 5, 5$

↓

$1, 2, 4, \boxed{5, 5, 5}, 6$

Descriptive Statistics – Variability

Standard Deviation

Amount of variation between the mean and rest of the data points.

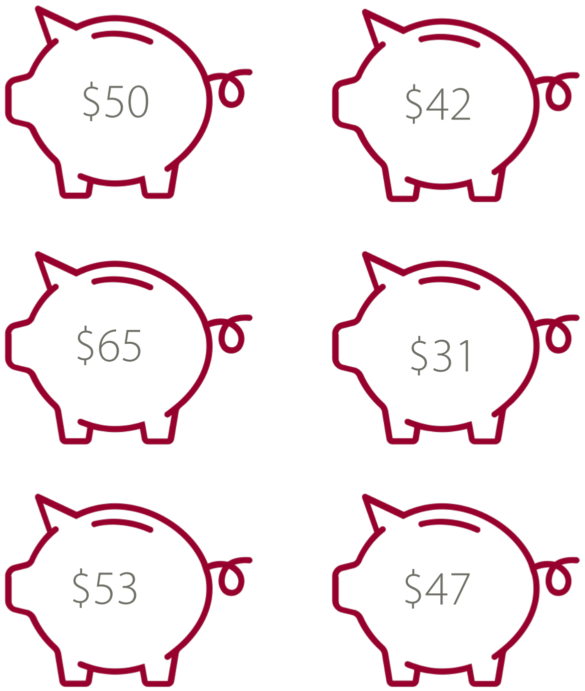
Range

Difference between Min and Max values.

Interquartile Range

Middle fifty of the data – where the majority of the data lies.

Descriptive Statistics – Variability



Quartile	Result	Definition
0	31	Minimum Value
1	43.25	25th Percentile
2	48.5	50th Percentile (median)
3	52.25	75th Percentile
4	65	Maximum Value

Range = 65 – 31 = 34
IQR = 52.25 – 43.25 = 9
Std Dev = 10.36

Visual Summary - Boxplot

The boxplot is a **graphical summary** of a univariate distribution.

Draw a box along the observation axis, with **endpoints** at Q_1 and Q_3, and with a "**belt**" at the median.

Plot a line extending from Q_1 to the smallest obs. less than $1.5\times\mathrm{IQR}$ below Q_1.

Plot a line extending from Q_3 to the smallest obs. more than $1.5\times\mathrm{IQR}$ above Q_3.

Any suspected outlier is plotted separately.

Visual Summary – Histogram

Histograms can also provide an indication of the distribution of a variable.

They should include/contain the following information:

- the **range** of the histogram is $r = Q_4 - Q_0$;
- the **number of bins** should approach $k = \sqrt{n}$, where n is the number of observations;
- the **bin width** should approach r/k, and
- the **frequency of observations** in each bin should be added to the chart.

Example

Consider the daily number of car accidents in Sydney over a 40-day period:

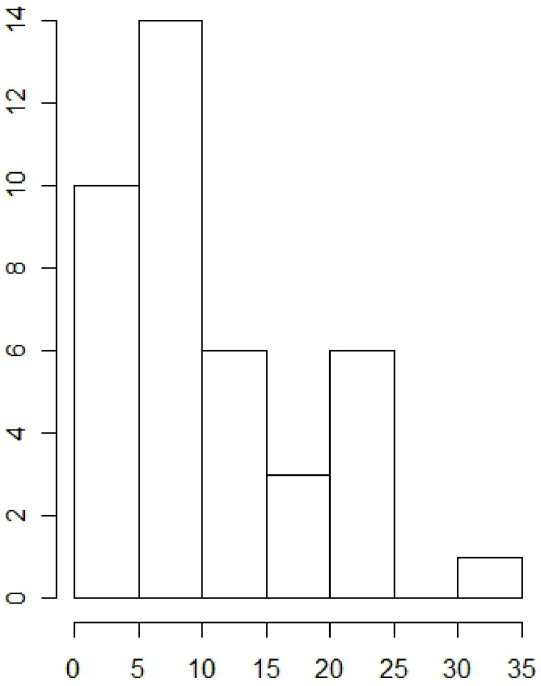
6, 3, 2, 24, 12, 3, 7, 14, 21, 9, 14, 22, 15,
2, 17, 10, 7, 7, 31, 7, 18, 6, 8, 2, 3, 2, 17,
7, 7, 21, 13, 23, 1, 11, 3, 9, 4, 9, 9, 25

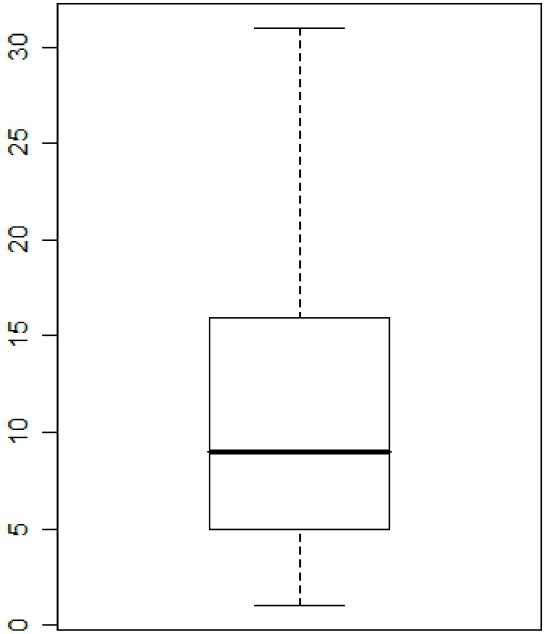
The sorted values are:

1 2 2 2 2 3 3 3 3 4 6 6 7 7 7 7 7
7 8 9 9 9 9 10 11 12 13 14 14 15 17
17 18 21 21 22 23 24 25 31

min	Q_1	med	Q_3	max
1	5.5	9	15.5	31

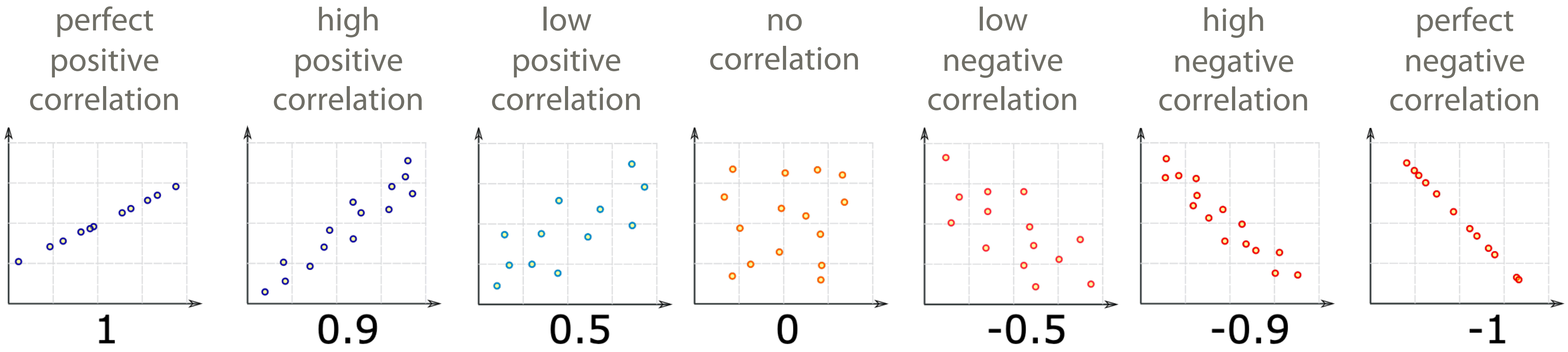
Is it more likely that we have between 5-15 accidents on a given day, or between 25-35?





[Unknown author]

Correlation



perfect positive correlation
1

high positive correlation
0.9

low positive correlation
0.5

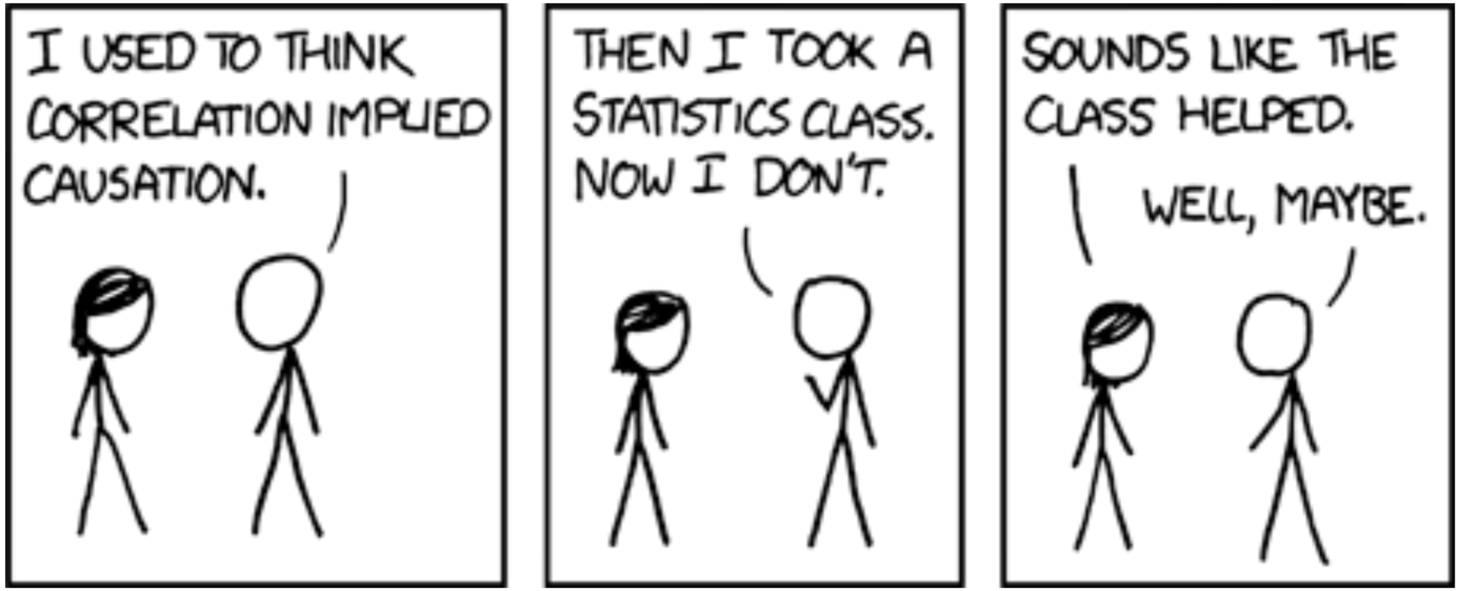
no correlation
0

low negative correlation
-0.5

high negative correlation
-0.9

perfect negative correlation
-1

[https://xkcd.com/552/]


I USED TO THINK CORRELATION IMPLIED CAUSATION.
THEN I TOOK A STATISTICS CLASS. NOW I DON'T.
SOUNDS LIKE THE CLASS HELPED.
WELL, MAYBE.

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

Regression Modeling

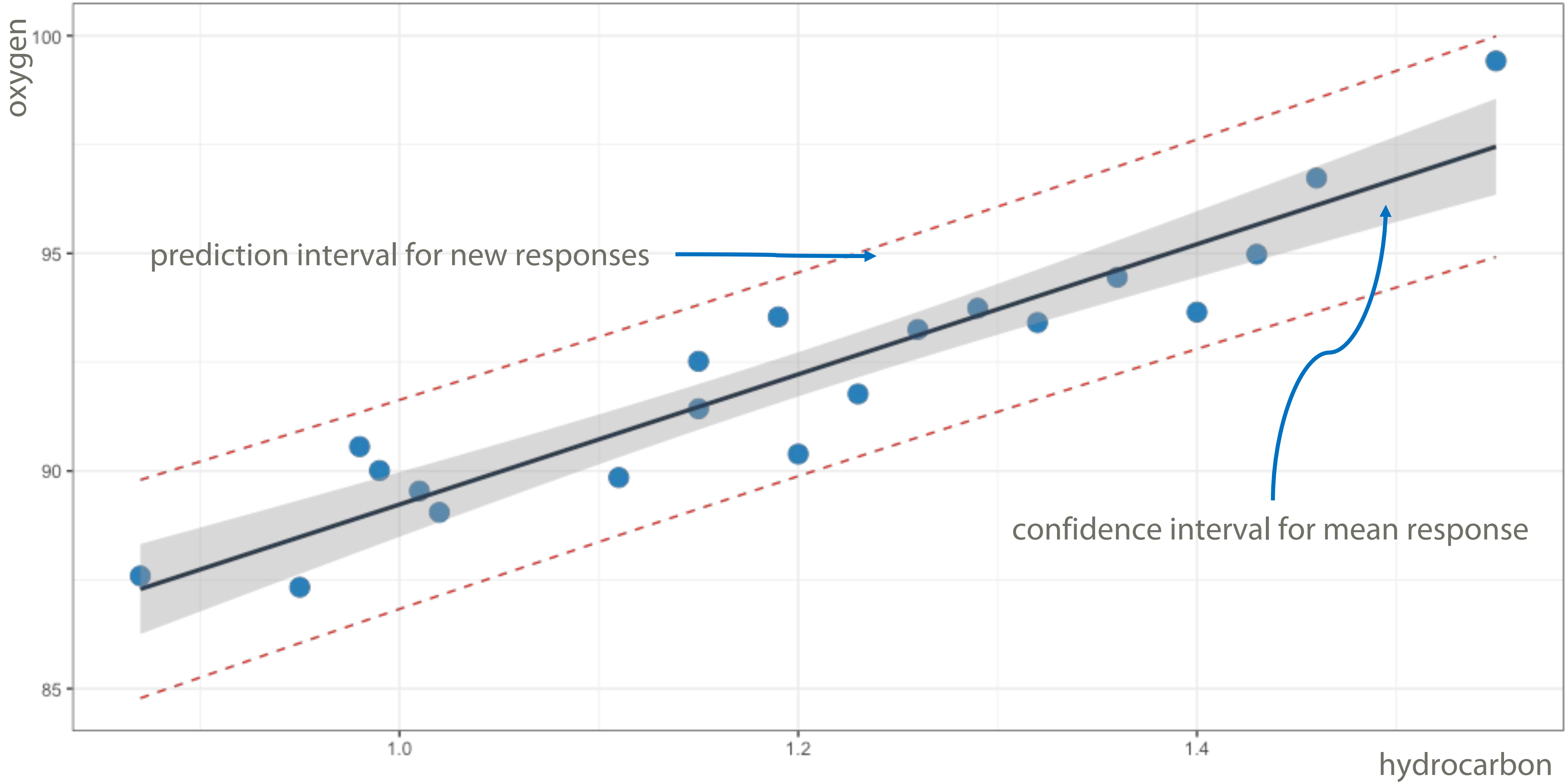
The most common data modeling methods are **regressions**, both linear and logistic.

About 80% of real data applications use a regression as their **final model**, typically after very careful **data preparation**, **encoding**, and **creation of variables**.

There are several reasons for their frequent use:

- generally straightforward to **understand** and to **train**
- mean square error (MSE) objective function has a closed-form linear solution
- system of equations can usually be solved through matrix inversion or linear manipulation

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



prediction interval for new responses

confidence interval for mean response

Other Analytical Approaches

Categorical analysis

Monte-Carlo simulations

Design of experiments

Bayesian data analysis

Times series analysis

Machine learning

Optimization

Queueing models

etc.

Gapminder Exercises

Do the exercises for Module 5.

Module 6

Data Mining and Machine Learning

What is Machine Learning?

Starting around the 1940s, researchers began the earnest study of how to **teach machines to learn**.

The goal of **machine learning** was (is?) to create machines that can **learn**, **adapt**, and **respond** to novel situations

A wide variety of techniques, accompanied by a great deal of theoretical underpinning, was created to achieve this goal.

What is Artificial/Augmented Intelligence?

Artificial Intelligence (A.I.) is non-human intelligence that has been engineered rather than one that has evolved naturally.

A.I. research is research carried out in pursuit of this goal.

Pragmatically speaking, A.I. is “computers carrying out tasks that only humans can usually do”.

Augmented Intelligence is human intelligence that is supported or enhanced by machine intelligence.

The Mining Analogy

What are we mining? data (**earth**)

What are we using to mine? data mining techniques (**digging tools**)

What are we mining for? looking for patterns/knowledge (**raw minerals**)

What do we do with the raw material? describe patterns/relationships (**refine minerals into something useful**)

What is the output, or product? models (**Ge, Ga, Si to build transistors**)

What do we do with the product? apply models to evidence-based decision support (**use transistor in electrical systems**)

Learning in General

Beyond "just taking a quick look," humans learn through:

- answering questions
- testing hypotheses
- creating concepts
- making predictions
- creating categories and classifying objects
- grouping objects

The **central Data Science/Machine Learning problem** is:

can (should) we design algorithms that can learn?

Types of Learning

Supervised Learning (**learning with a teacher**)

- classification, regression, rankings, recommendations
- uses labeled training data (**student gives an answer to each test question based on what they learned from worked-out examples**)
- performance is evaluated using testing data (**teacher provides the correct answers**)

Unsupervised Learning (**grouping similar exercises together as a study aid**)

- clustering, association rules discovery, link profiling, anomaly detection
- uses unlabeled observations (**teacher is not involved**)
- accuracy cannot be evaluated (**students might not end up with the same groupings**)

Types of Learning

Semi-Supervised Learning (**teacher providing worked-out examples and a list of unsolved problems**)

Reinforcement Learning (**embarking on a research project with an advisor?**)

In supervised learning, there’s a target against which to train the model.

In unsupervised learning, we don’t know what the target is, or if there is one.

The distinction is crucial.

Learning Tasks

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

Clustering: do diplomatic missions 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.

Case Study: Association Rules Mining

The Danish National Patient Registry contains 68 million health observations on 6.2 million patients over a 15-year time span (‘96 –‘10).

Objectives:

- finding connections between different diagnoses
- determining how a diagnosis at some point in time might allow for the prediction of another diagnosis at a later point in time

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

Methodology

1. Compute strength of correlation for **pairs of diagnoses** over a 5-year interval on a **representative subset** of the data

2. Test pairs for **directionality** (one repeatedly occurring before the other)

3. Determine reasonable **diagnosis trajectories** (thoroughfares) by combining smaller frequent trajectories with overlapping diagnoses

4. Validate the trajectories by **comparison with non-Danish data**

5. Cluster the thoroughfares to identify **central medical conditions** around which disease progression is organized

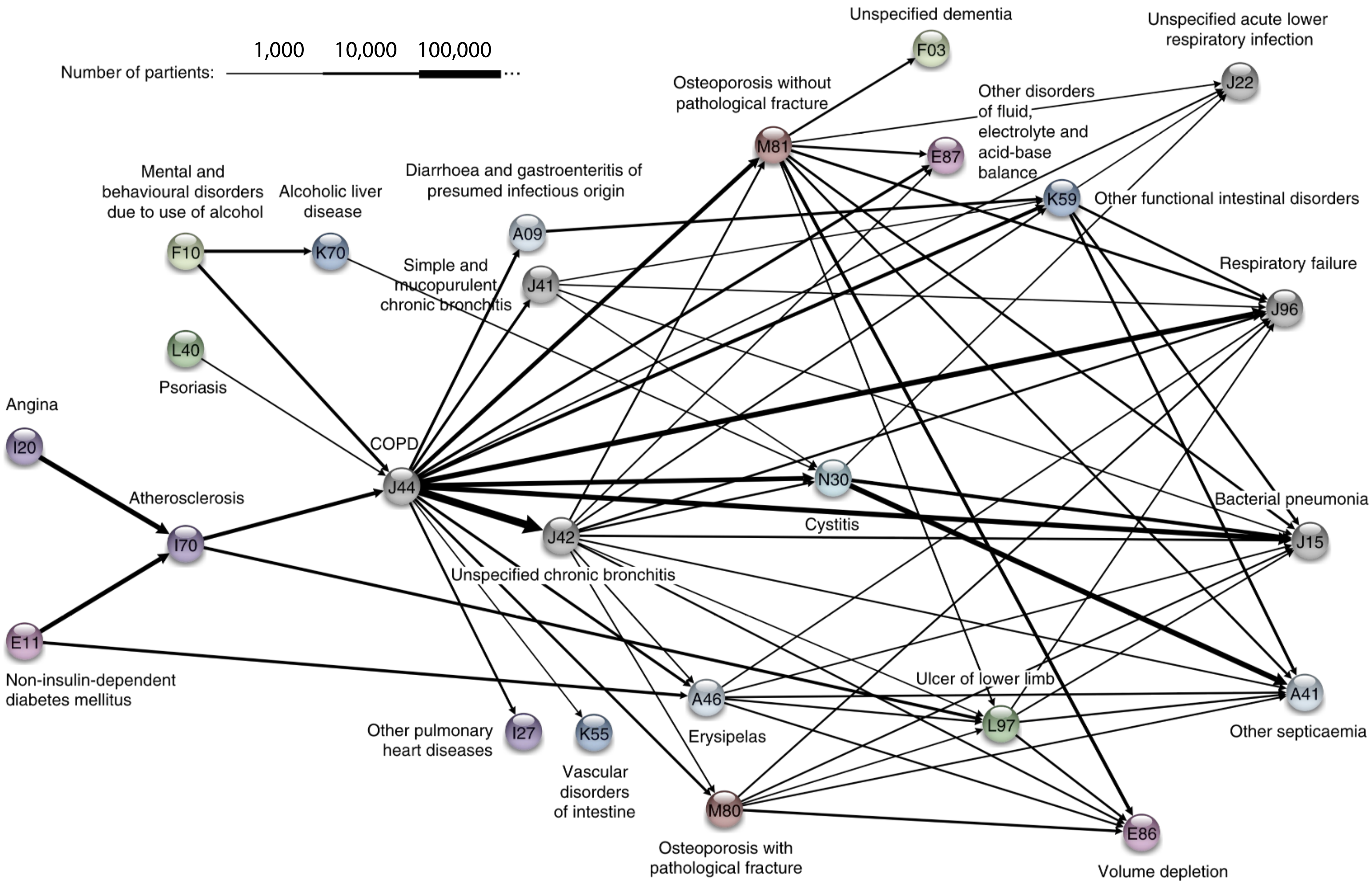
Results

Data was reduced to **1,171 thoroughfares**, with **5 key diagnoses:**

- diabetes
- chronic obstructive pulmonary disease (COPD)
- cancer
- arthritis
- cardiovascular disease.

The data analysis showed, for example:

- diagnoses of **anemia** followed later by the discovery of **colon cancer**
- **gout** was identified as a step toward **cardiovascular disease**
- **COPD** is under-diagnosed and under-treated


Number of partients:
1,000
10,000
100,000
Unspecified dementia
F03
Unspecified acute lower respiratory infection
J22
Osteoporosis without pathological fracture
M81
Other disorders of fluid, electrolyte and acid-base balance
E87
Mental and behavioural disorders due to use of alcohol
F10
Alcoholic liver disease
K70
Diarrhoea and gastroenteritis of presumed infectious origin
A09
K59
Other functional intestinal disorders
Simple and mucopurulent chronic bronchitis
J41
Respiratory failure
J96
L40
Psoriasis
Angina
I20
COPD
J44
N30
Atherosclerosis
I70
Bacterial pneumonia
Cystitis
J42
J15
Unspecified chronic bronchitis
E11
Non-insulin-dependent diabetes mellitus
Ulcer of lower limb
A46
A41
L97
Other septicaemia
Other pulmonary heart diseases
I27
K55
Erysipelas
Vascular disorders of intestine
M80
Osteoporosis with pathological fracture
E86
Volume depletion

Case Study Take-Aways

Data makes it possible to view diseases in a larger context, which could yield tangible **health benefits** beyond one-size-fits-all medicine.

The **sooner** a health risk pattern is identified, the **better** we can prevent and treat critical diseases.

Instead of looking at each disease in isolation, we can talk about a **complex system** with many different **interacting factors**.

The **order** in which different diseases appear can help find patterns and complex correlations outlining the direction for each individual person.

Association Rules Basics

Association Rule Discovery is unsupervised learning that finds connections among attributes (and combinations of attributes).

Example: we might analyze a dataset on the physical activities and purchasing habits of North Americans and discover that

- runners who are also triathletes (the **premise**) tend to drive Subarus, drink microbrews, and use smartphones (the **conclusion**), or
- individuals who have purchased home gym equipment are unlikely to be using it 1 year later (to name some fictitious possibilities)

Market Basket Analysis

Supermarkets record the contents of shopping carts at check-outs to determine items which are **frequently purchased together**.

Examples:

- **bread** and **milk** are often purchased together, but that's not so interesting given how often they are purchased individually
- **hot dog buns** and **wieners** are also often purchased as a pair, but more rarely purchased individually

A supermarket could then have a sale on hot dogs to drive in customers, while raising the price on condiments, to drive in sales.

Applications

Related Concepts

- looking for pairs (triplets, etc) of words that represent a joint concept
- {Ottawa, Senators}, {Michelle, Obama}, {veni, vidi, vici}, etc.

Plagiarism

- looking for sentences that appear in various documents
- looking for documents that share sentences

Bio-markers

- diseases that are frequently associated with a set of bio-markers

Applications

Making predictions and decisions based on these rules.

Alter circumstances or environment to take advantage of these correlations (often mis-used).

Use the connections to modify the likelihood of certain outcomes.

Imputing missing data.

Text autofill and autocorrect.

Causation and Correlation

Association rules can help automate **hypothesis discovery**, but we must remain correlation-savvy (which is less prevalent among analysts than one would hope…).

If attributes A and B are shown to be **correlated**, then the possibilities are:

- A and B are correlated entirely by chance in this dataset
- A is a relabeling of B
- A causes B and/or B causes A
- combinations of other attributes $C_1, \ldots, C_n$ (known or not) cause A & B
- etc?

[E. Siegel, Predictive Analytics: The Power to Predict Who Will Click, Buy, Lie, or Die]

Causation and Correlation

Insight	Organization
Pop-Tarts before a hurricane	Walmart
Higher crime, more Uber rides	Uber
Typing with proper capitalization indicates creditworthiness	A financial services startup company
Users of the Chrome and Firefox browsers make better employees	A human resources professional services firm, over employee data from Xerox and other firms
Men who skip breakfast get more coronary heart disease	Harvard University medical researchers
More engaged employees have fewer accidents	Shell
Smart people like curly fries	Researchers at the University of Cambridge and Microsoft Research
Female-named hurricanes are more deadly	University researchers
Higher status, less polite	Researchers examining Wikipedia behavior

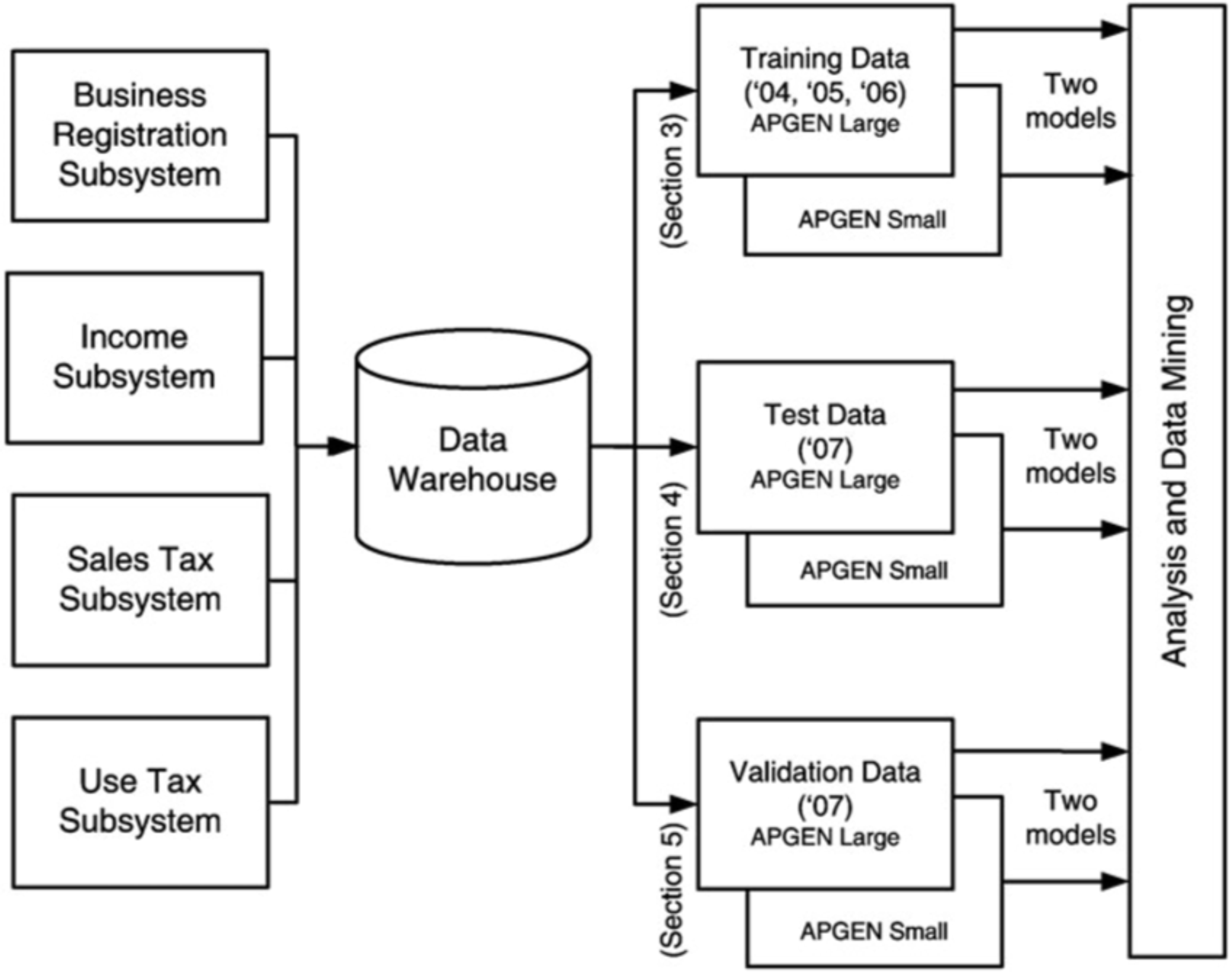
Case Study: Minnesota Tax Audit

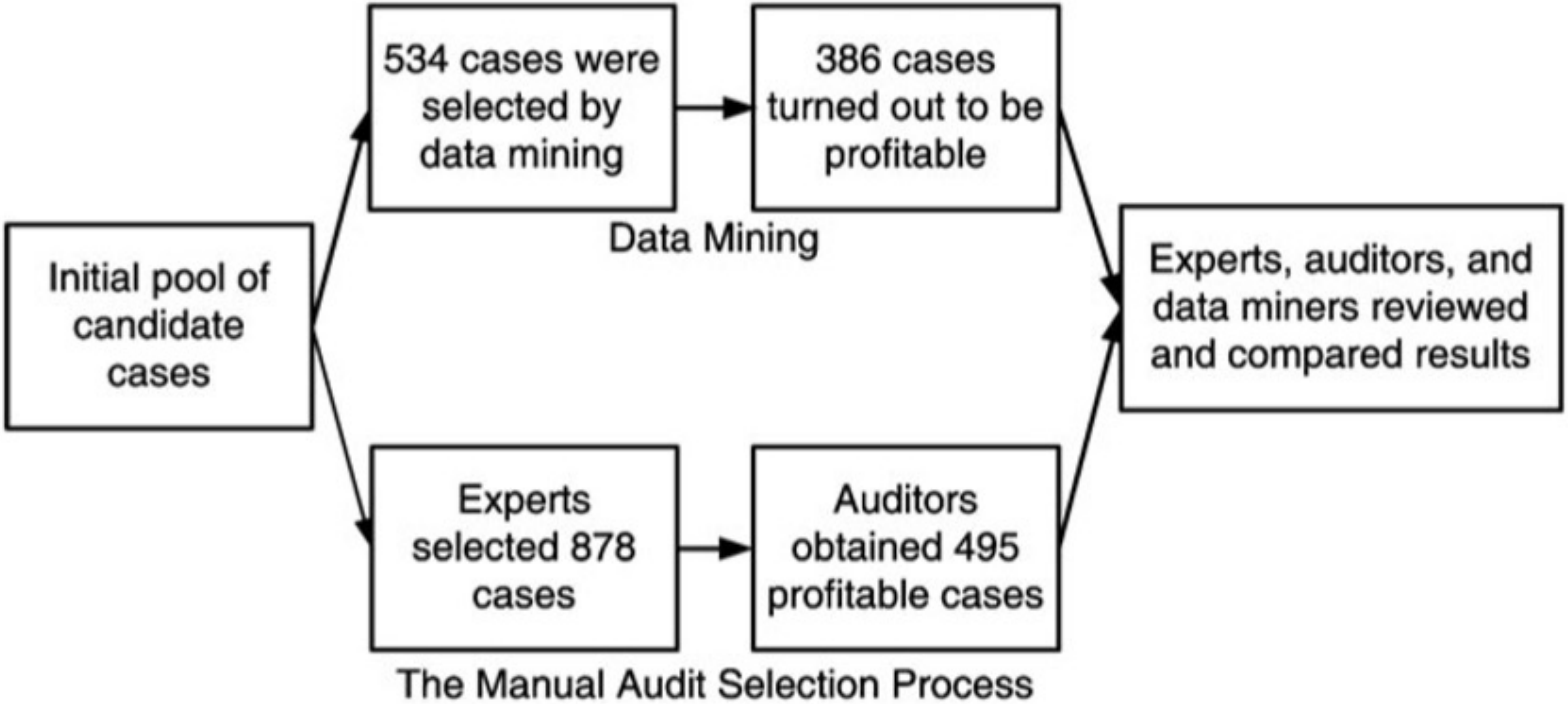
Large gaps between revenue owed (in theory) and revenue collected (in practice) are problematic for governments.

Revenue agencies implement various fraud detection strategies (such as audit reviews) to bridge that gap.

Business audits are costly – are there **algorithms that can predict whether an audit is likely to be successful or a waste of resources**?

Data mining-based tax audit selection: a case study of a pilot project at the Minnesota Department of Revenue
Hsu, W., Pathak, N., Srivatsava, J., Tschida, Bjorklund, E. [2013], Real Word Data Mining Applications, Annals of Information Systems, v.17, Springer.


Business Registration Subsystem
Income Subsystem
Sales Tax Subsystem
Use Tax Subsystem
Data Warehouse
(Section 3)
(Section 4)
(Section 5)
Training Data ('04, '05, '06) APGEN Large
APGEN Small
Test Data ('07) APGEN Large
APGEN Small
Validation Data ('07) APGEN Large
APGEN Small
Two models
Two models
Two models
Analysis and Data Mining



Initial pool of candidate cases

534 cases were selected by data mining

386 cases turned out to be profitable

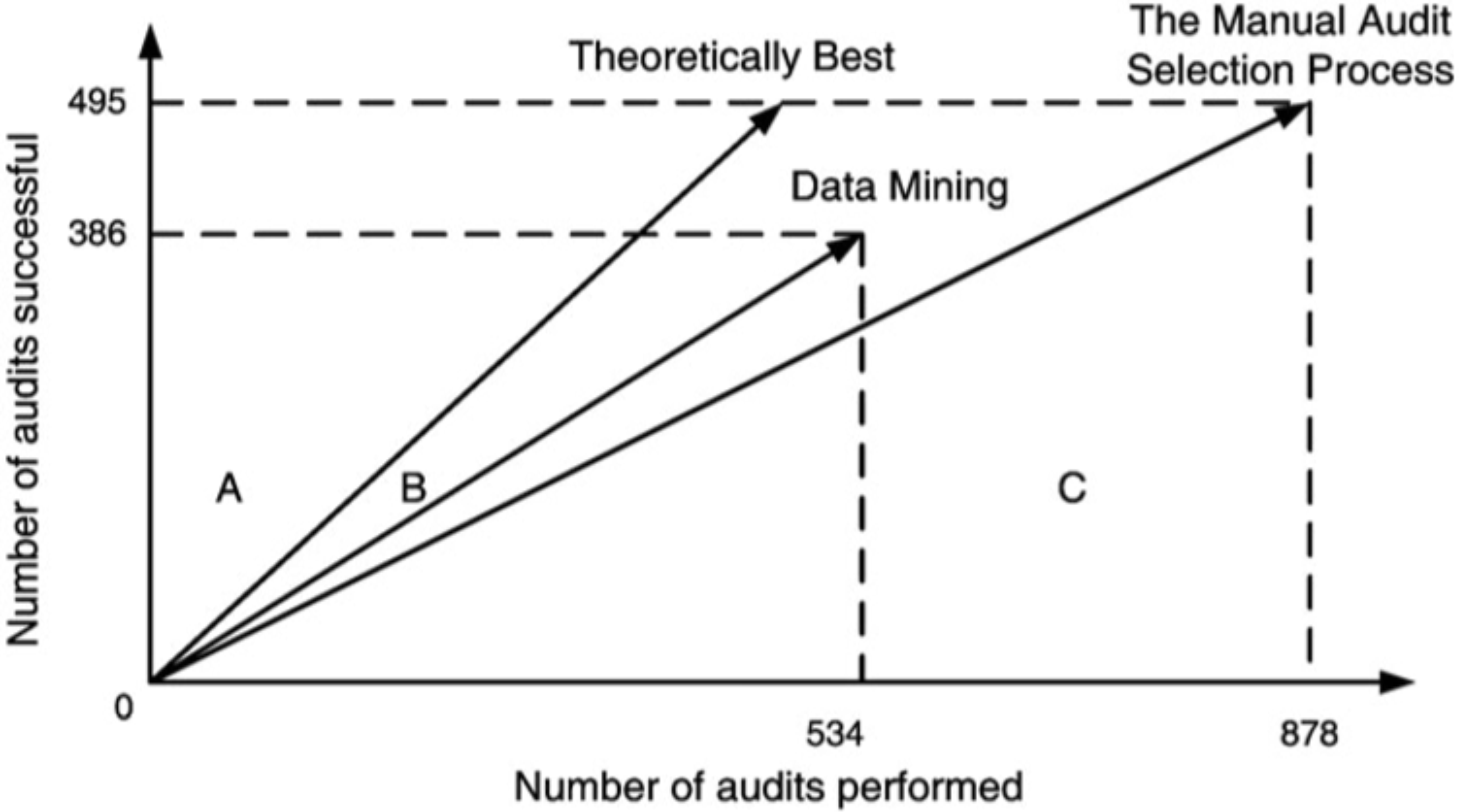
Data Mining

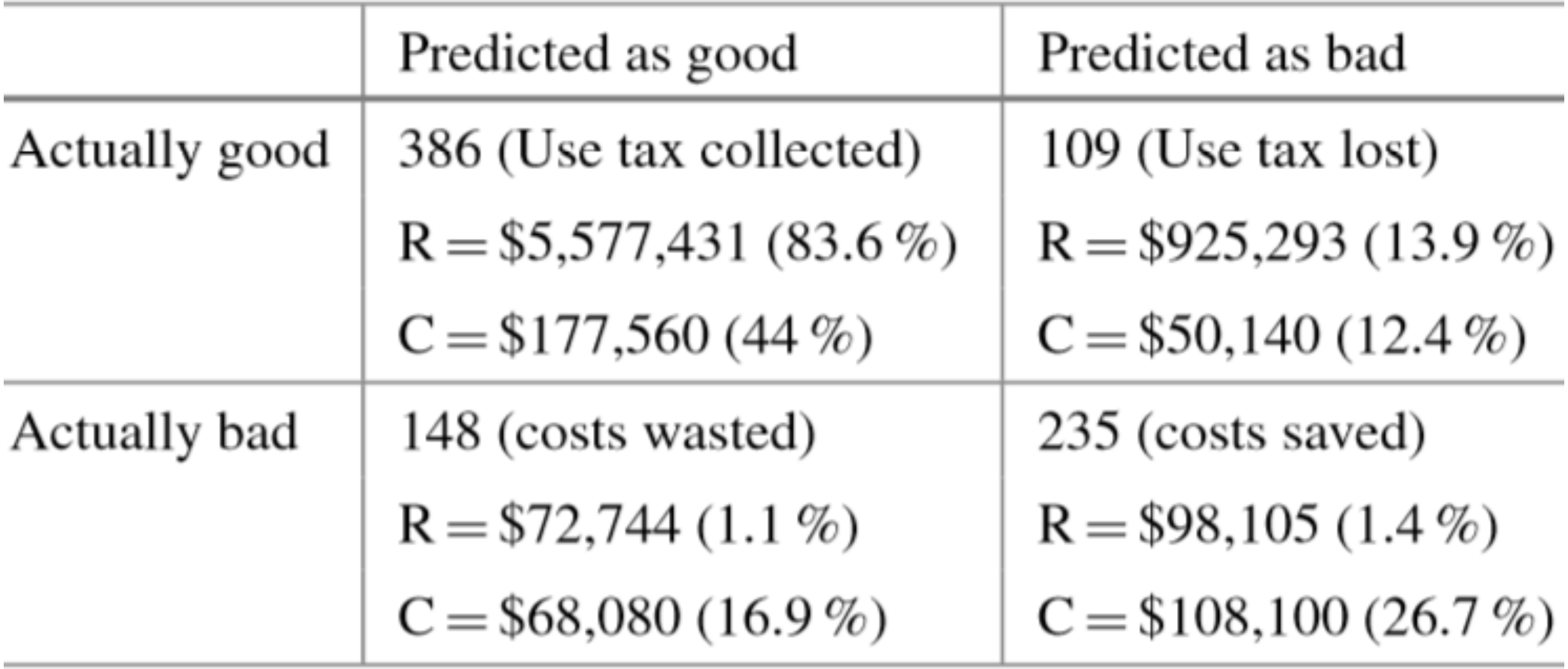
Experts selected 878 cases

Auditors obtained 495 profitable cases

The Manual Audit Selection Process

Experts, auditors, and data miners reviewed and compared results


The Manual Audit Selection Process
Theoretically Best
Data Mining
495
386
0
534
878
A
B
C
Number of audits successful
Number of audits performed



	Predicted as good	Predicted as bad

Actually good	386 (Use tax collected)
	R = $5,577,431 (83.6 %)
	C = $177,560 (44 %)
		109 (Use tax lost)
		R = $925,293 (13.9 %)
		C = $50,140 (12.4 %)

Actually bad	148 (costs wasted)
	R = $72,744 (1.1 %)
	C = $68,080 (16.9 %)
		235 (costs saved)
		R = $98,105 (1.4 %)
		C = $108,100 (26.7 %)

Classification Overview

In **classification**, a sample set of data (the **training set**) is used to determine **rules** and **patterns** that divide the data into pre-determined groups (**classes**).

Classification is a **supervised learning** task.

The training data usually consists of a **randomly selected** subset of the **labeled** (target) data.

Value estimation (regression) is akin to classification, but the target variable is **numerical**.

Classification Overview

In the **testing phase**, the model is used to assign a class to observations for which the **label is hidden**, but ultimately known (the **testing set**).

The **performance** of a classification model is evaluated on the testing set, never on the training set.

Technical challenges include:

- selecting the features to include in the model
- selecting the algorithm
- etc.

Applications

Medicine and Health Science

- predicting which patient is at risk of suffering a second, fatal heart attack within 30 days based on health factors (blood pressure, age, sinus problems, etc.)

Social Policies

- predicting the likelihood of requiring assisting housing in old age based on demographic information/survey answers

Marketing and Business

- predicting which customers are likely to switch to another cell phone company based on demographics and usage

Other Uses

Predicting that an object belongs to a particular class.

Organizing and grouping instances into categories.

Enhancing the detection of relevant objects

- avoidance: “this object is an incoming vehicle”
- pursuit: “this borrower is unlikely to default on her mortgage”
- degree: “this dog is 90% likely to live until it’s 7 years old”

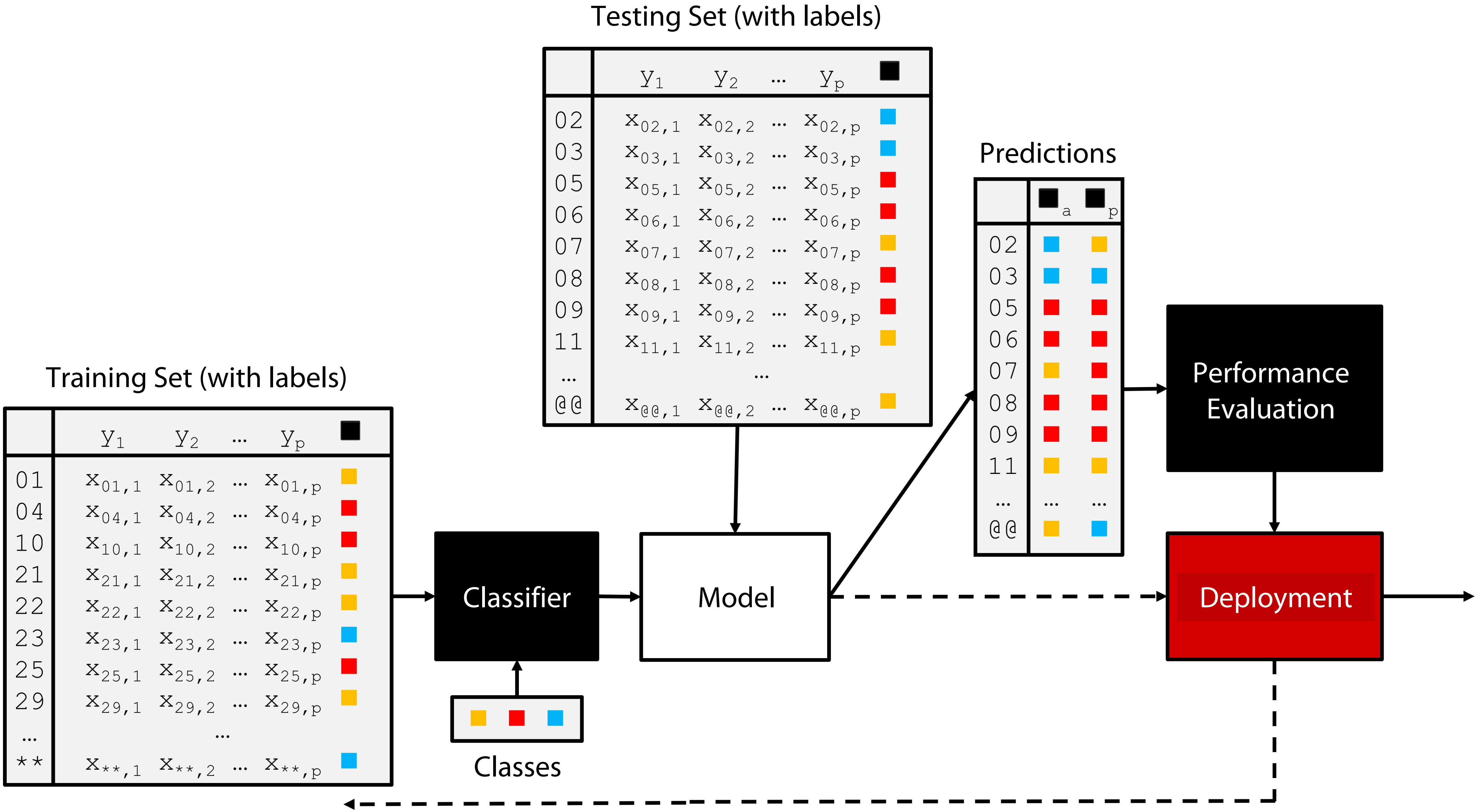
In the absence of testing data, classification may be **descriptive** but not predictive.

Example

Scenario: a motor insurance company has a fraud investigation dept. that studies up to 30% of all claims made, yet money is still getting lost on fraudulent claims.

Questions: can we predict

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



Training Set (with labels)

	y_1	y_2	…	y_p	⬛
01	$x_{01,1}$	$x_{01,2}$	…	$x_{01,p}$	🟨
04	$x_{04,1}$	$x_{04,2}$	…	$x_{04,p}$	🟥
10	$x_{10,1}$	$x_{10,2}$	…	$x_{10,p}$	🟥
21	$x_{21,1}$	$x_{21,2}$	…	$x_{21,p}$	🟨
22	$x_{22,1}$	$x_{22,2}$	…	$x_{22,p}$	🟨
23	$x_{23,1}$	$x_{23,2}$	…	$x_{23,p}$	🟦
25	$x_{25,1}$	$x_{25,2}$	…	$x_{25,p}$	🟥
29	$x_{29,1}$	$x_{29,2}$	…	$x_{29,p}$	🟨
…			…		
**	$x_{**,1}$	$x_{**,2}$	…	$x_{**,p}$	🟦

Classifier

🟨	🟥	🟦

Classes

Model

Testing Set (with labels)

	y_1	y_2	…	y_p	⬛
02	$x_{02,1}$	$x_{02,2}$	…	$x_{02,p}$	🟦
03	$x_{03,1}$	$x_{03,2}$	…	$x_{03,p}$	🟦
05	$x_{05,1}$	$x_{05,2}$	…	$x_{05,p}$	🟥
06	$x_{06,1}$	$x_{06,2}$	…	$x_{06,p}$	🟥
07	$x_{07,1}$	$x_{07,2}$	…	$x_{07,p}$	🟨
08	$x_{08,1}$	$x_{08,2}$	…	$x_{08,p}$	🟥
09	$x_{09,1}$	$x_{09,2}$	…	$x_{09,p}$	🟥
11	$x_{11,1}$	$x_{11,2}$	…	$x_{11,p}$	🟨
…			…		
@@	$x_{@@,1}$	$x_{@@,2}$	…	$x_{@@,p}$	🟨

Predictions

	⬛$_a$	⬛$_p$
02	🟦	🟨
03	🟦	🟦
05	🟥	🟥
06	🟥	🟥
07	🟨	🟥
08	🟥	🟥
09	🟥	🟥
11	🟨	🟨
…	…	…
@@	🟨	🟦

Performance Evaluation

Deployment

Classification Methods

Logistic Regression

- classical model
- affected by variance inflation and variable selection process

Neural Networks

- hard to interpret
- requires all variables to be of the same type
- easier to train since backpropagation (chain rule)

Decision Trees

- may overfit the data if not pruned correctly (manually?)

Classification Methods

Naïve Bayes Classifiers

- quite successful for text mining applications (spam filter)
- assumptions not often met in practice

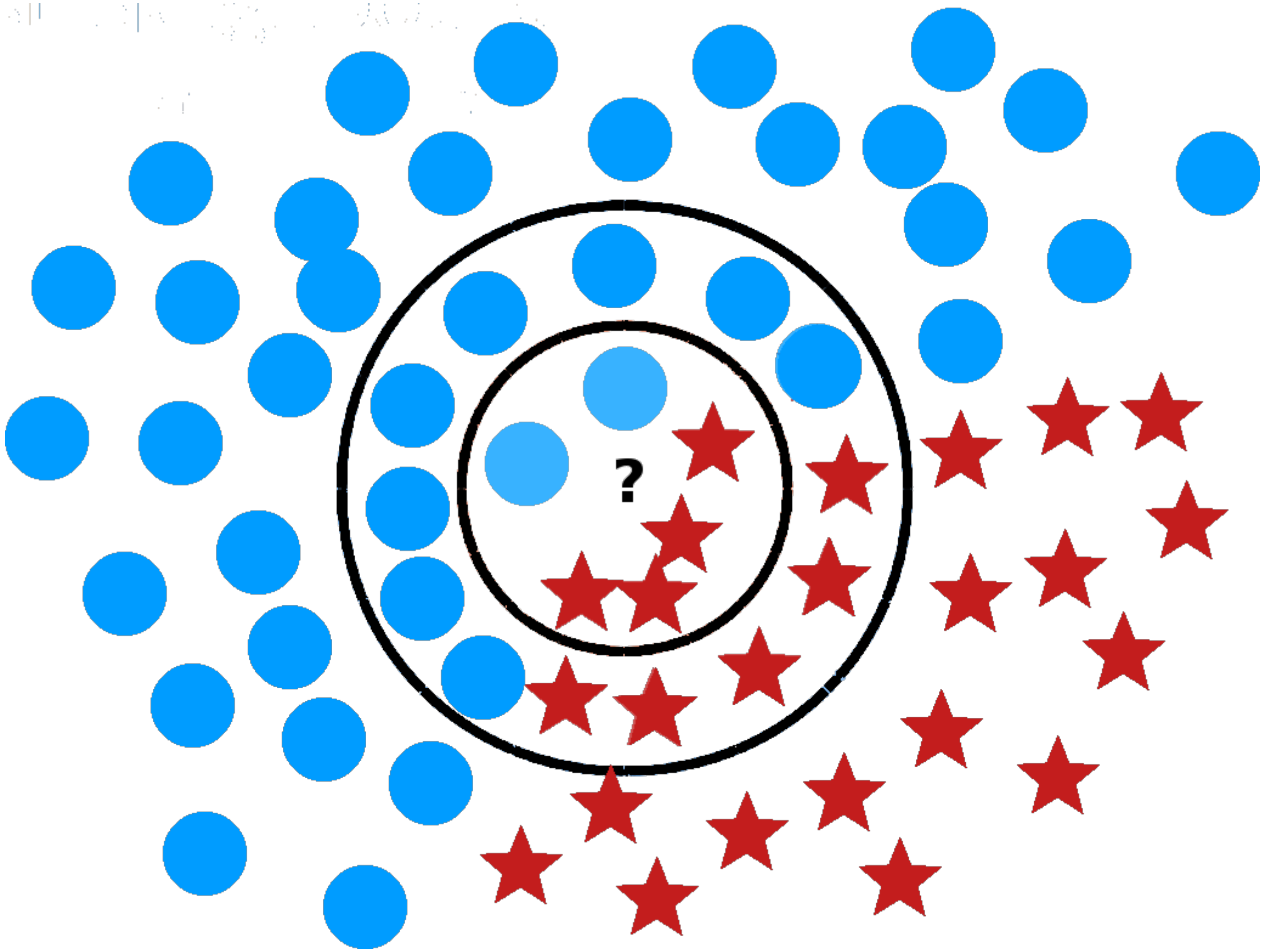
Support Vector Machines

- may be difficult to interpret (non-linear boundaries)
- can help mitigate big data difficulties

Nearest Neighbours Classifiers

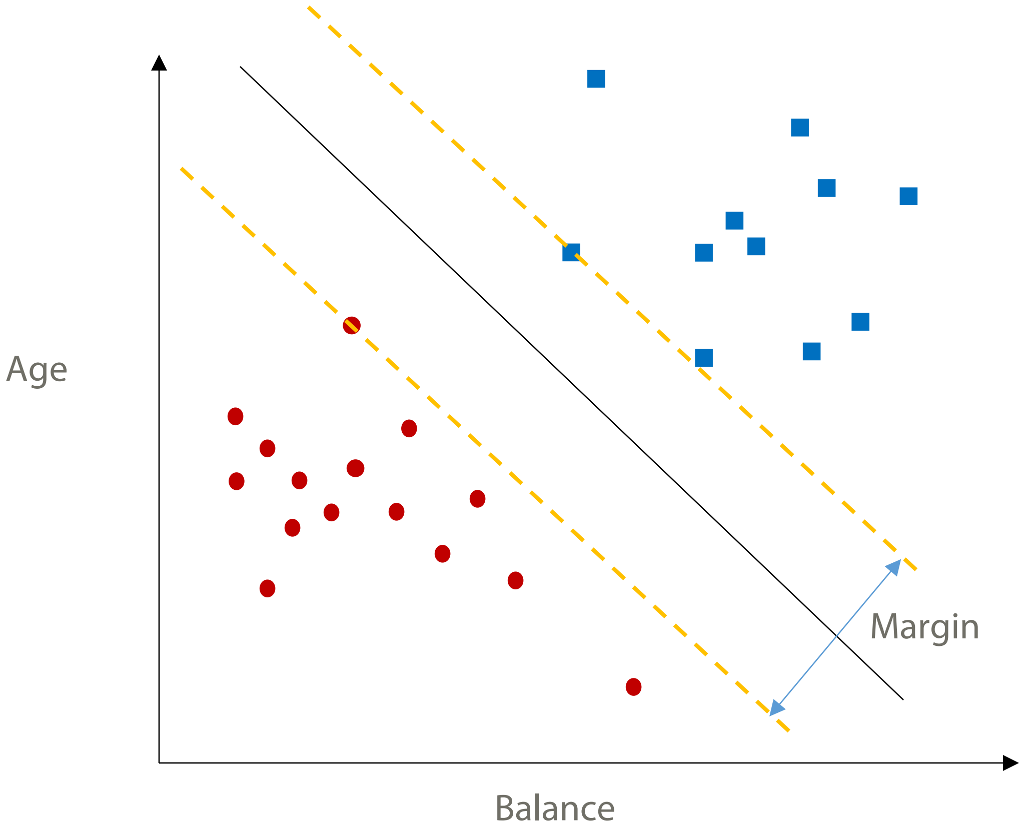
- require very little assumptions about the data
- not very stable (adding points may substantially modify the boundary)

k −Nearest Neighbours



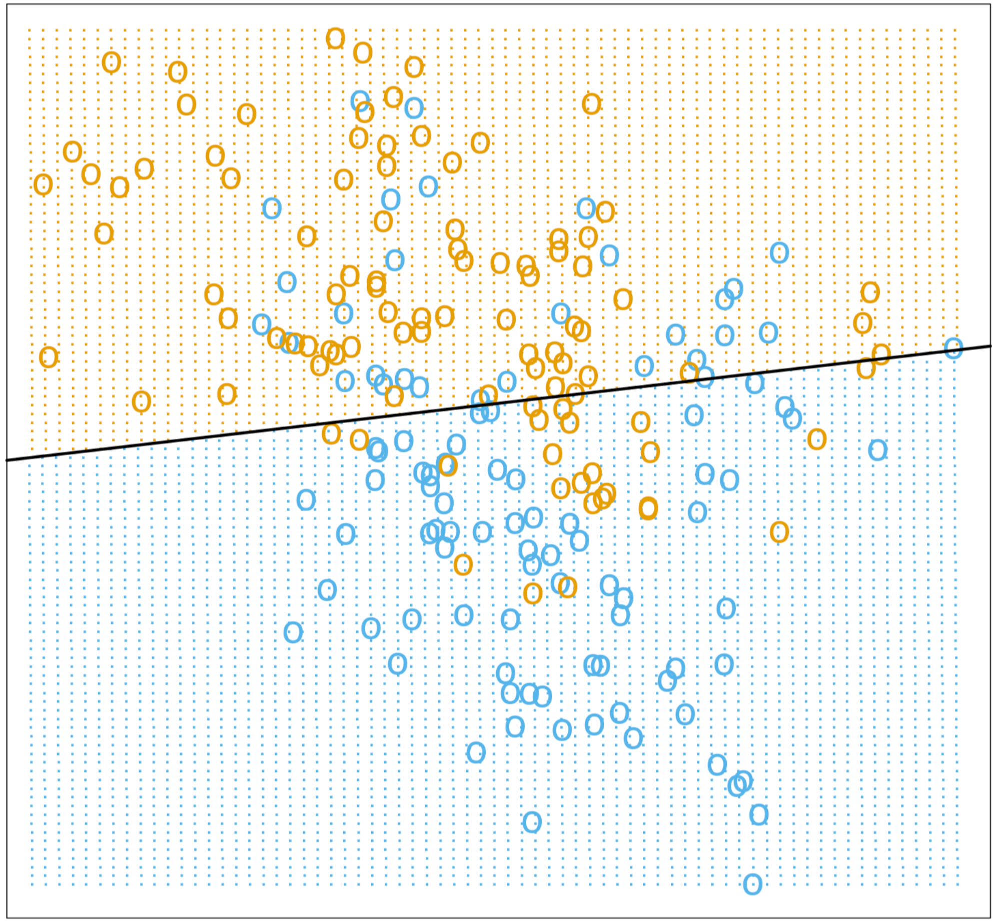
[Adapted from Foster & Provost's Data Science for Business]

Support Vector Machines

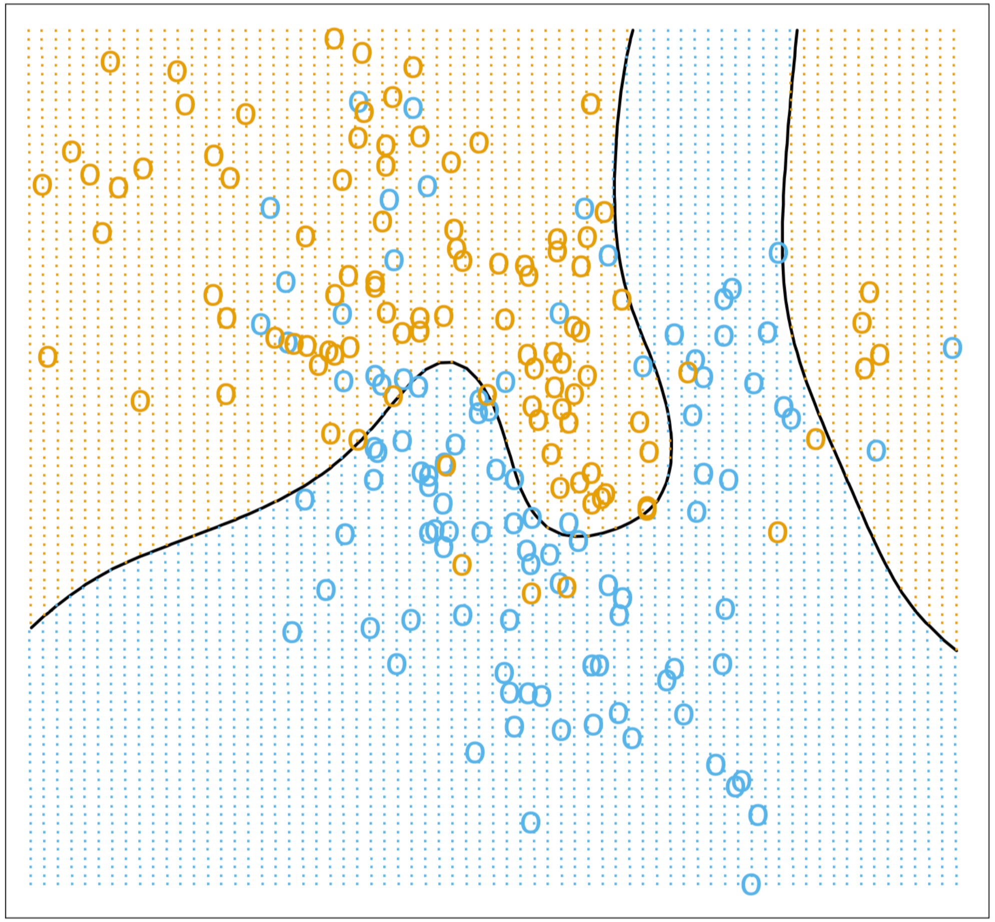


[Images taken from The Elements of Statistical Learning]

Boundary Classifiers



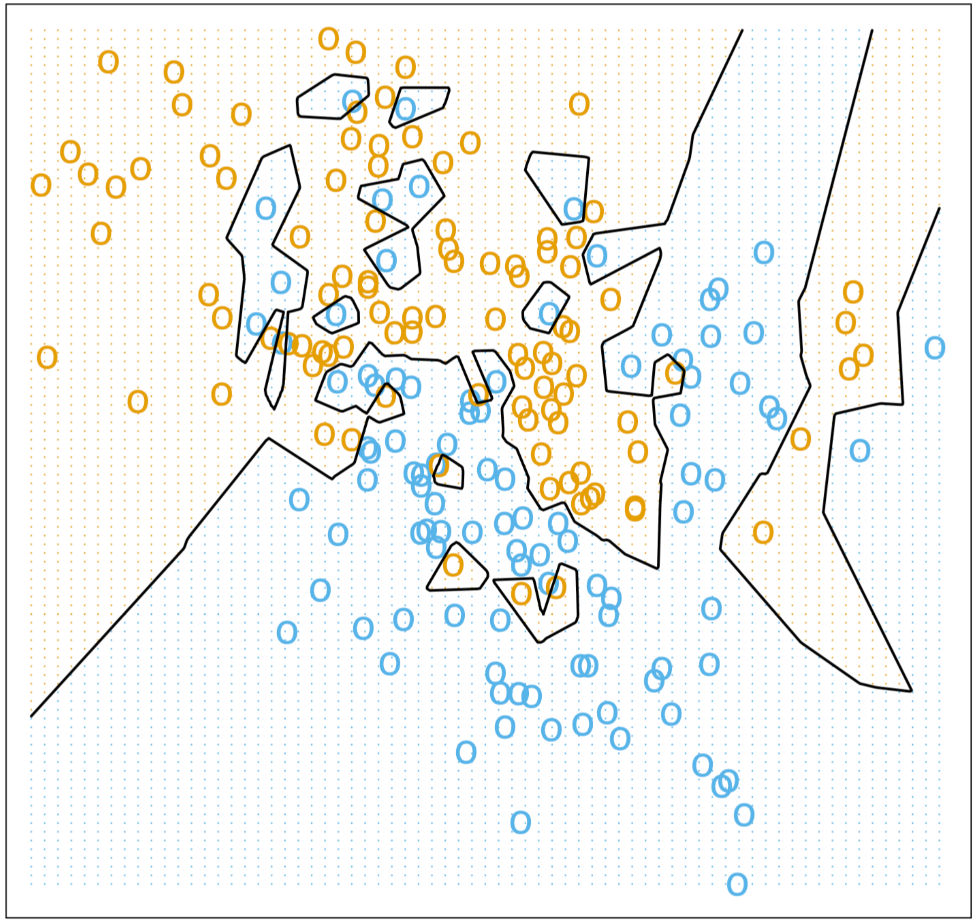
Linear Regression Classifier



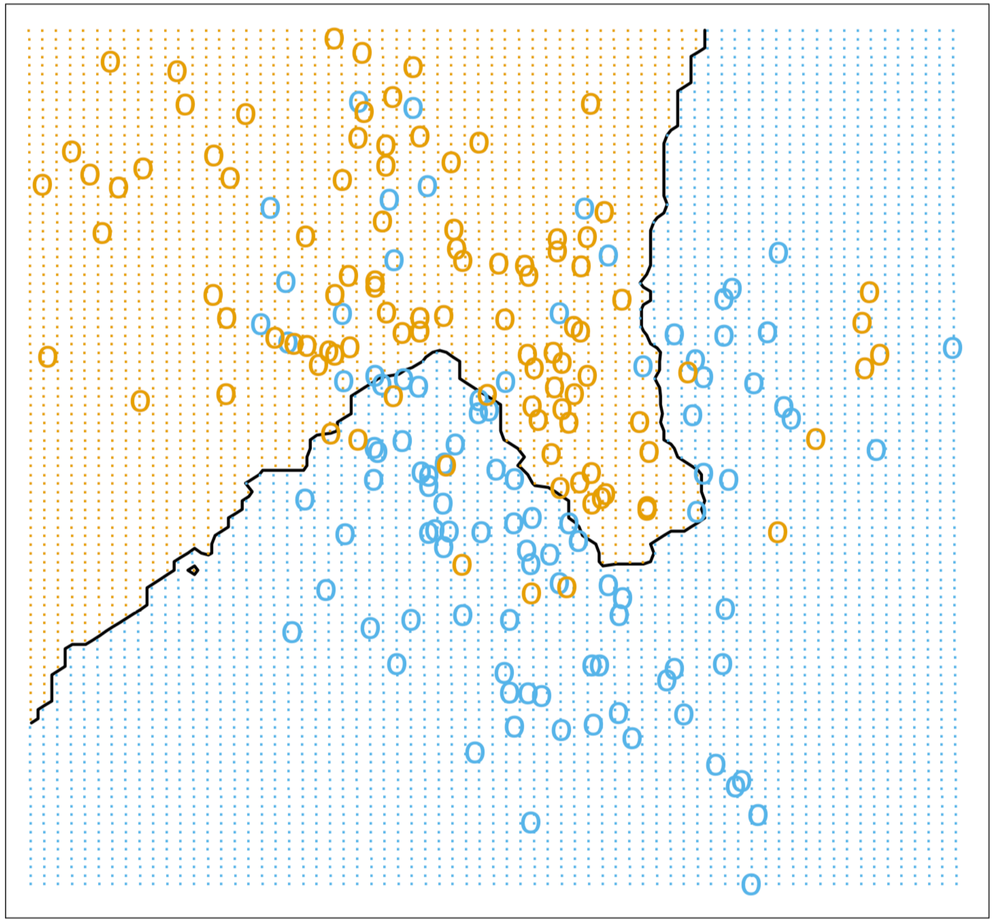
Optimal Bayes Classifier

[Images taken from The Elements of Statistical Learning]

Nearest Neighbours Classifiers



1NN Classifier



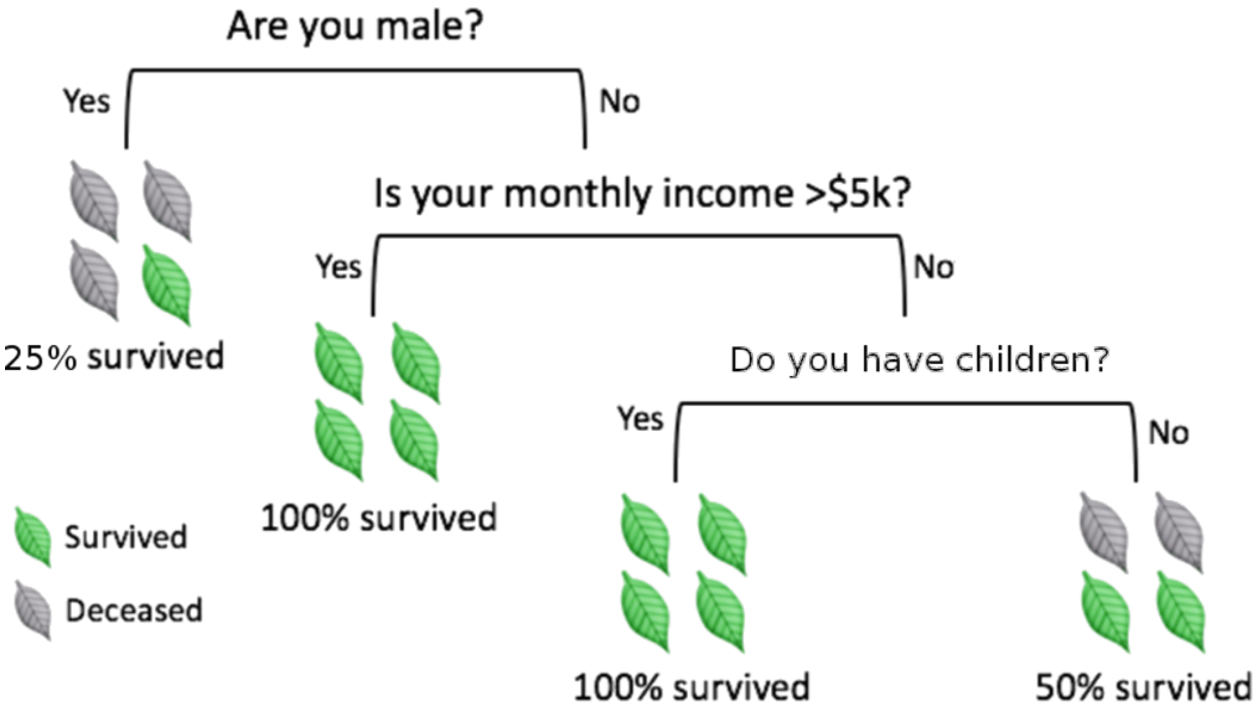
15NN Classifier

[adapted from A. Ng, K. Soo, Numsense!, Surviving a Disaster]

Decision Trees

Decision trees are perhaps the most intuitive of these methods.

Classification is achieved by following a path **up the tree**, from its **root**, through its **branches**, and ending at its **leaves**.



Performance Evaluation

Classifiers are evaluated on a **testing** set.

Ideally, a good classifier would have high rates of both **True Positives** (TP) and **True Negatives** (TN), and low rates of both **False Positives** (FP, Type I error) and **False Negatives** (FN, Type II error).

Evaluation metrics mean very little on their own: context requires comparison with other classifiers, and other evaluation metrics.

Performance Evaluation

sensitivity = $TP/(TP+FN)$

specificity = $TN/(FP+TN)$

precision = $TP/(TP+FP)$

recall = $TP/(TP+FN)$

negative predictive value = $TN/(TN+FN)$

false positive rate = $FP/(FP+TN)$

false discovery rate = $FP/(FP+TP)$

false negative rate = $FP/(FN+TP)$

accuracy = $(TP+TN)/T$

		Predicted		
		Category I	Category II	Total
Actuals	Category I	TP	FN	AP
	Category II	FP	TN	AN
	Total	PP	PN	T

Other metrics:

F_1-score, ROC AUC, informedness, markedness, Matthews' Correlation Coefficient (MCC), etc.

Performance Evaluation

		Predicted			
		A	B		*Total*
Actuals	A	54	10	64	79.0%
	B	6	11	17	21.0%
Total		60	21	81	
		74.1%	25.9%		

Classification Rates	
Sensitivity:	0.84
Specificity:	0.65
Precision:	0.90
Negative Predictive Value:	0.52
False Positive Rate:	0.35
False Discovery Rate:	0.10
False Negative Rate:	0.16

Performance Metrics	
Accuracy:	0.80
F1-Score:	0.87
Informedness (ROC):	0.49
Markedness:	0.42
M.C.C.:	0.46
Pearson's chi2:	0.01
Hist. Stat:	0.10

		Predicted			
		A	B		*Total*
Actuals	A	54	0	54	66.7%
	B	16	11	27	33.3%
Total		70	11	81	
		86.4%	13.6%		

Classification Rates	
Sensitivity:	1.00
Specificity:	0.41
Precision:	0.77
Negative Predictive Value:	1.00
False Positive Rate:	0.59
False Discovery Rate:	0.23
False Negative Rate:	0.00

Performance Metrics	
Accuracy:	0.80
F1-Score:	0.87
Informedness (ROC):	0.41
Markedness:	0.77
M.C.C.:	0.56
Pearson's chi2:	0.33
Hist. Stat:	0.40

Case Study: OK Cupid

Chris McKinlay, a 35 year old UCLA Math PhD Student, was looking for a romantic partner online with little luck

- OK Cupid algorithms use only the questions that both potential matches decide to answer, and the questions he had chosen (more or less at random up to that point) were not popular

Between June 2012 and December 2013, he:

- used statistical sampling to find questions which mattered to the kind of partner he had in mind;
- constructed a new profile that answered only those questions;
- matched only with women in LA who might be right for him.

K. Poulsen, How a Math Genius Hacked OK Cupid to Find True Love, WIRED

Process

This story provides a **great** example of the data mining process, from start to finish:

1. **Collect** data
2. Collect **more** and **slightly better** and **different** data
3. Collect **still more** data
4. Figure out a data mining technique that would be **relevant** to what he wanted to know (clustering)
5. **Validate** the results of the analysis
6. **Investigate** the results, and narrow down which results were interesting
7. Analyze the interesting results **some more**, and use this to solve the original problem
8. Use the data to **improve other areas** of his profile as well
9. Sit back and reap the benefits of data mining?

Methodology and Results

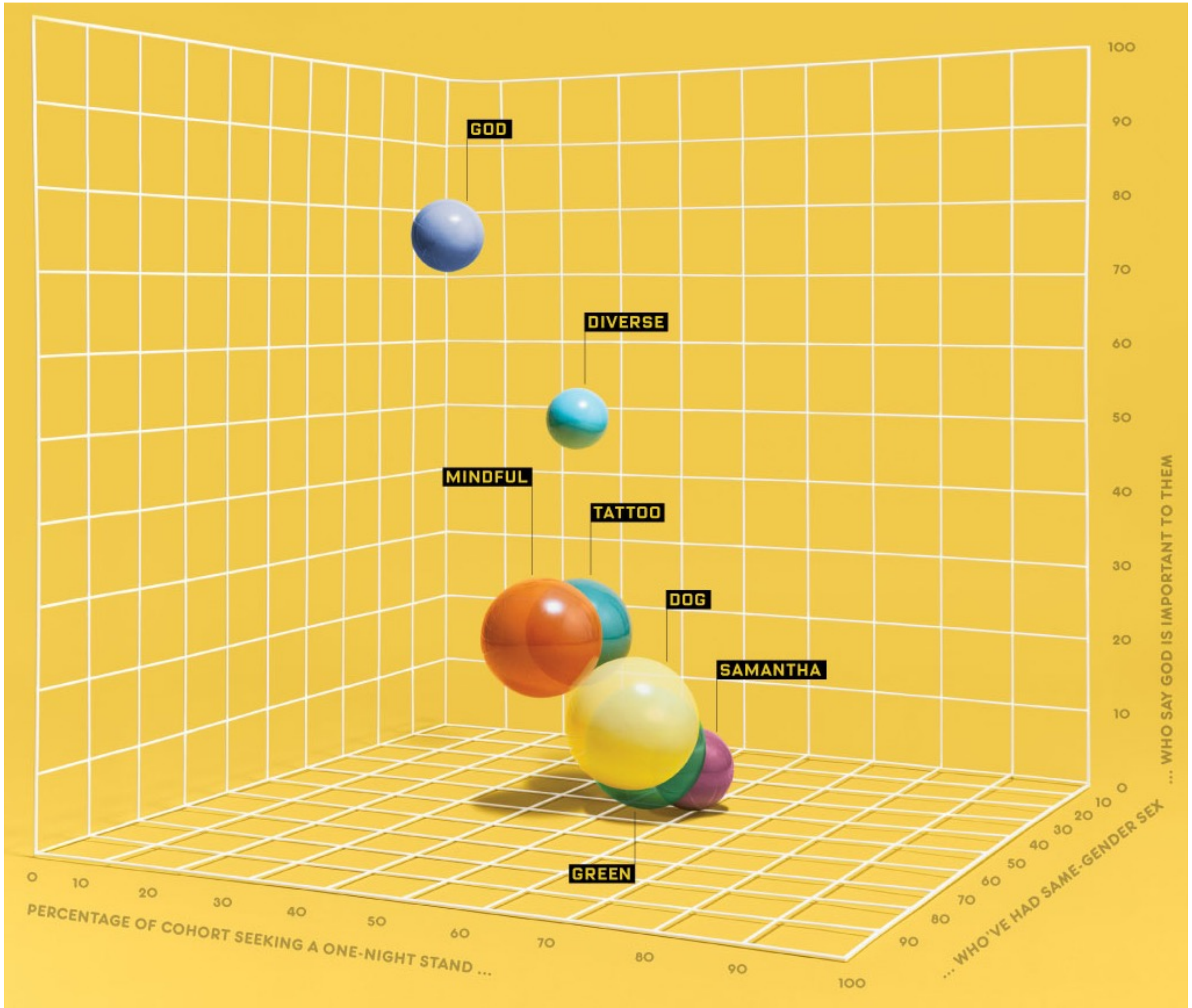
Used k-mode to cluster 20,000 women into seven statistically distinct clusters based on their questions and answers.

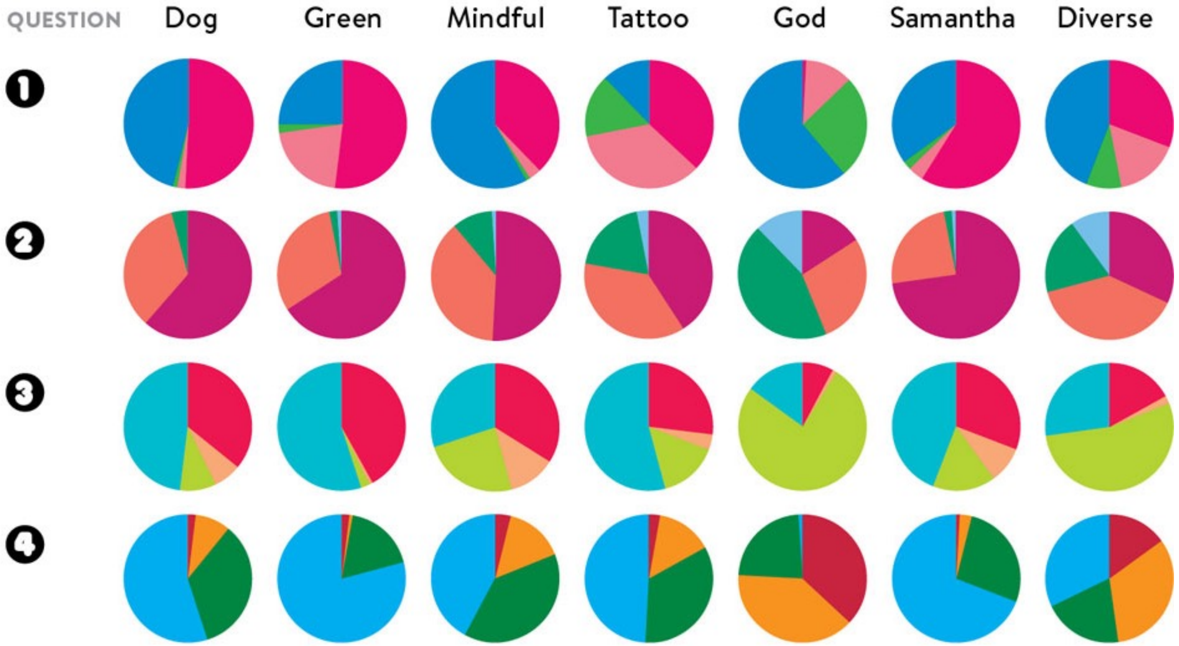
Validated the clustering with another 5,000 profiles from the site.

Analyzed the clusters to find two that interested him

- women in their mid-twenties who looked like indie types, musicians and artists
- slightly older women who held professional creative jobs, like editors and designers.

Used results to derive **which questions he should answer** in his profile, leading to more matches based on his profile, to more first dates, to some second dates, and … to a lone third date.


GOD
DIVERSE
MINDFUL
TATTOO
DOG
SAMANTHA
GREEN
PERCENTAGE OF COHORT SEEKING A ONE-NIGHT STAND ...
0
10
20
30
40
50
60
70
80
90
100
... WHO'VE HAD SAME-GENDER SEX
0
10
20
30
40
50
60
70
80
90
100
... WHO SAY GOD IS IMPORTANT TO THEM
0
10
20
30
40
50
60
70
80
90
100



1. About how long do you want your next relationship to last?

One night
A few months to a year
Several years
The rest of my life

2. Say you've started seeing someone you really like. As far as you're concerned, how long will it take before you have sex?

1-2 dates
3-5 dates
6 or more dates
Only after the wedding

3. Have you ever had a sexual encounter with someone of the same sex?

Yes, and I enjoyed myself
Yes, and I did not enjoy myself
No, and I would never
No, but I'd like to

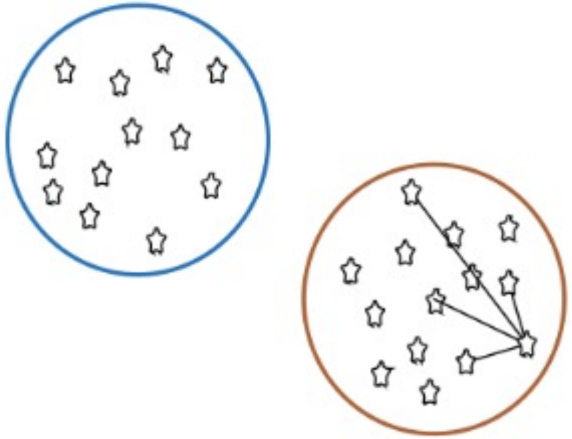
4. How important is religion/God in your life?

Extremely important
Somewhat important
Not very important
Not important at all

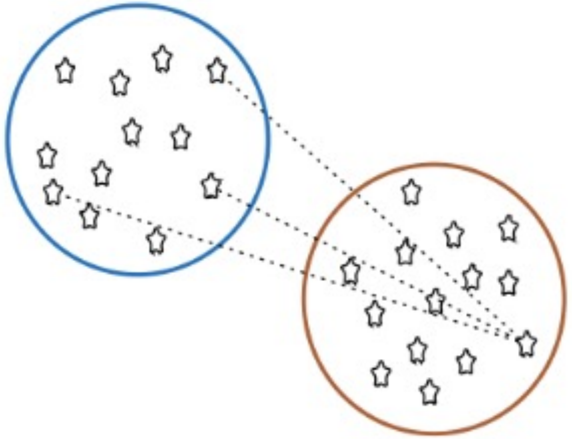
Clustering Overview

In **clustering**, the data is divided into **naturally occurring groups**. Within each group, the data points are **similar**; from group to group, they are **dissimilar**.

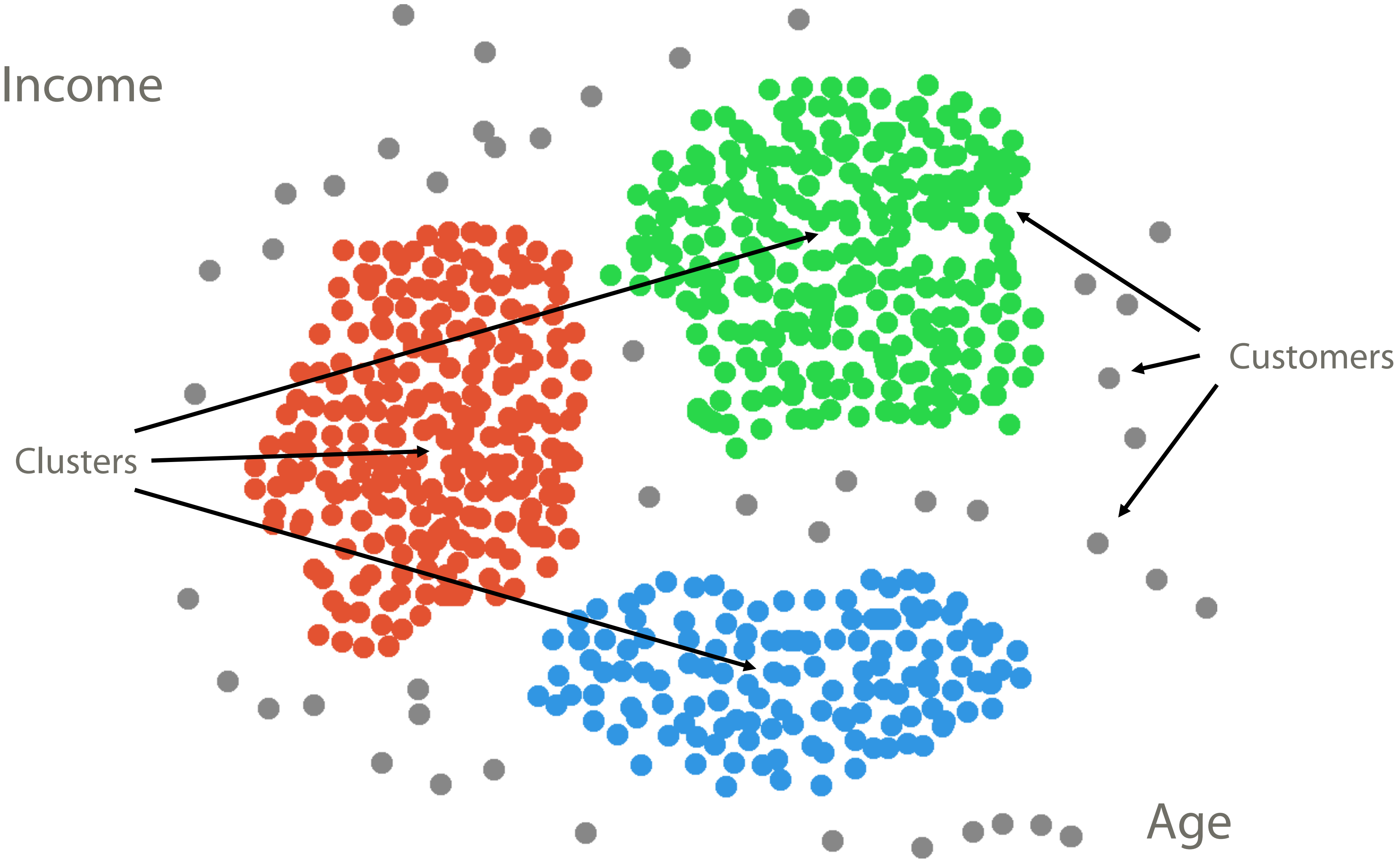
The grouping labels are not determined ahead of time, so clustering is an example of **unsupervised** learning.



average distance to points in own cluster (**low is good**)



average distance to points in neighbouring cluster (**high is good**)


Income
Customers
Clusters
Age

Clustering Overview

Clustering is a relatively **intuitive** concept for human beings as our brains do it unconsciously

- facial recognition
- searching for patterns, etc.

In general, people are very good at **messy** data, but computers and algorithms have a harder time.

Part of the difficulty is that there is **no agreed-upon definition of what constitutes a cluster**

- "I may not be able to define what it is, but I know one when I see one"

Clustering Overview

Clustering algorithms can be **complex** and **non-intuitive**, based on varying notions of similarities between observations

- in spite of that, the temptation to explain clusters a posteriori is strong

They are also (typically) **non-deterministic:**

- the same algorithm, applied twice (or more) to the same dataset, can discover completely different clusters
- the order in which the data is presented can play a role
- so can starting configurations

Applications

Text Documents

- grouping similar documents according to their topics, based on the patterns of common and unusual words

Product Recommendations

- grouping online purchasers based on the products they have viewed, purchased, liked, or disliked
- grouping products based on customer reviews

Marketing and Business

- grouping client profiles based on their demographics and preferences

Other Uses

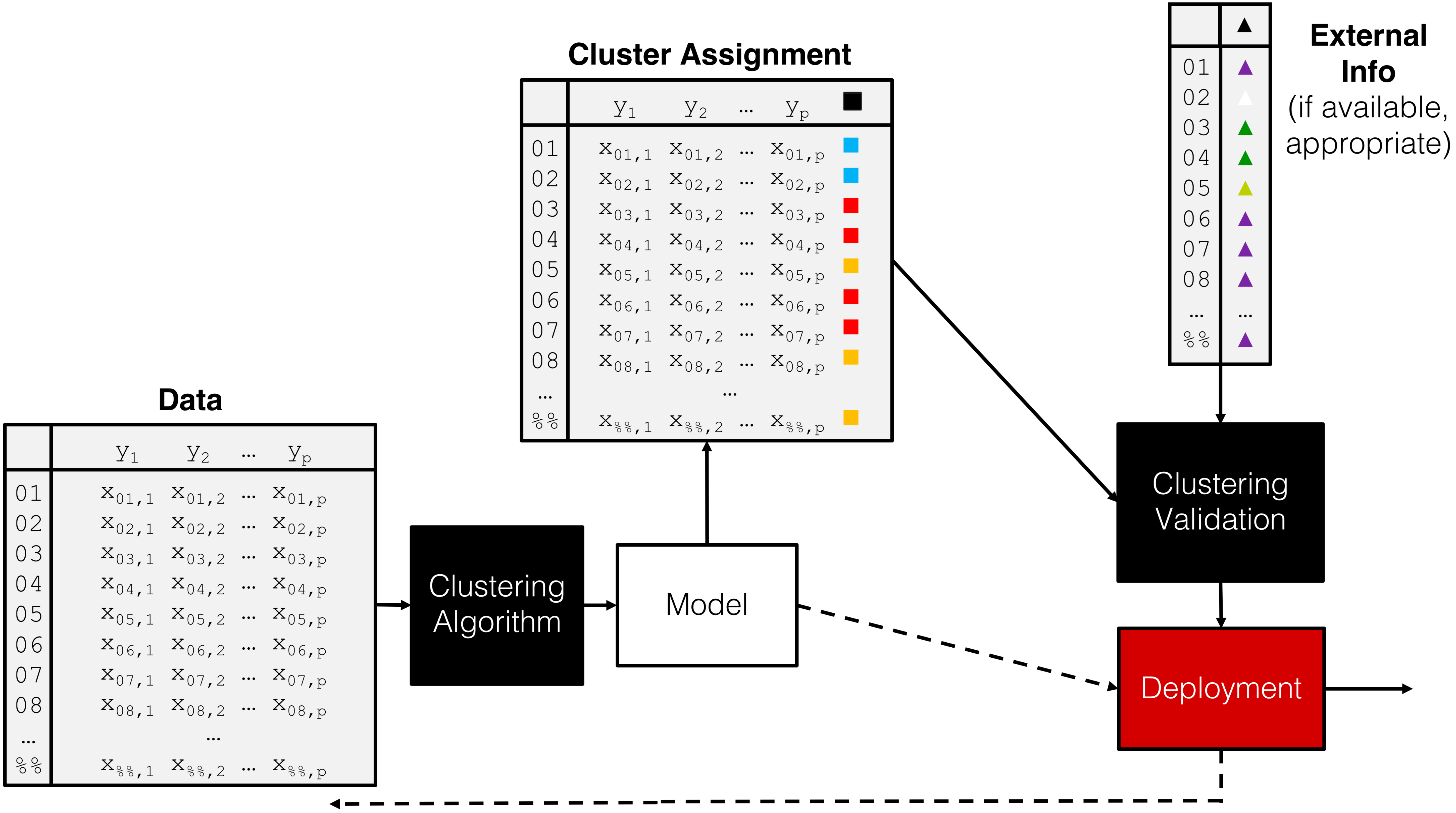
Dividing a larger group (or area, or category) into **smaller** groups, with members of the smaller groups guaranteed to have similarities of some kind

- tasks may then be solved separately for each of the smaller groups
- this may lead to increased accuracy once the separate results are aggregated

Creating (new) taxonomies **on the fly**, as new items are added to a group of items

- this would allow for easier product navigation on a website like Netflix, for instance.

See Spotlight on Clustering, in Data Understanding, Data Analysis, and Data Science for more examples.



Data

	y_1	y_2	…	y_p
01	$x_{01,1}$	$x_{01,2}$	…	$x_{01,p}$
02	$x_{02,1}$	$x_{02,2}$	…	$x_{02,p}$
03	$x_{03,1}$	$x_{03,2}$	…	$x_{03,p}$
04	$x_{04,1}$	$x_{04,2}$	…	$x_{04,p}$
05	$x_{05,1}$	$x_{05,2}$	…	$x_{05,p}$
06	$x_{06,1}$	$x_{06,2}$	…	$x_{06,p}$
07	$x_{07,1}$	$x_{07,2}$	…	$x_{07,p}$
08	$x_{08,1}$	$x_{08,2}$	…	$x_{08,p}$
…		…		
%%	$x_{\%\%,1}$	$x_{\%\%,2}$	…	$x_{\%\%,p}$

Clustering Algorithm

Model

Cluster Assignment

	y_1	y_2	…	y_p	⬛
01	$x_{01,1}$	$x_{01,2}$	…	$x_{01,p}$	🟦
02	$x_{02,1}$	$x_{02,2}$	…	$x_{02,p}$	🟦
03	$x_{03,1}$	$x_{03,2}$	…	$x_{03,p}$	🟥
04	$x_{04,1}$	$x_{04,2}$	…	$x_{04,p}$	🟥
05	$x_{05,1}$	$x_{05,2}$	…	$x_{05,p}$	🟨
06	$x_{06,1}$	$x_{06,2}$	…	$x_{06,p}$	🟥
07	$x_{07,1}$	$x_{07,2}$	…	$x_{07,p}$	🟥
08	$x_{08,1}$	$x_{08,2}$	…	$x_{08,p}$	🟨
…		…			
%%	$x_{\%\%,1}$	$x_{\%\%,2}$	…	$x_{\%\%,p}$	🟨

External Info
(if available,
appropriate)

	▲
01	▲
02	△
03	▲
04	▲
05	▲
06	▲
07	▲
08	▲
…	…
%%	▲

Clustering Validation

Deployment

Clustering Methods

k-Means

- classical (and over-used) model
- assumptions made about the shape of clusters

Hierarchical Clustering

- easy to interpret, deterministic

Latent Dirichlet Allocation

- used for topic modeling

Expectation-Maximization

Clustering Methods

Balanced Iterative Reducing and Clustering using Hierarchies

- aka BIRCH

Density-Based Spatial Clustering of Applications with Noise

- graph-based

Affinity Propagation

- selects the optimal number of clusters automatically

Spectral Clustering

- recognizes non-blob clusters

Clustering Validation

What does it mean for a clustering scheme to be **better** than another?

What does it mean for a clustering scheme to be **valid**?

What does it mean for a single cluster to be **good**?

How many clusters are there in the data, really?

Right vs. wrong is meaningless: seek **optimal vs. sub-optimal**.

Clustering Challenges

Automation

relatively intuitive for humans, but harder for machines

Lack of a clear-cut definition

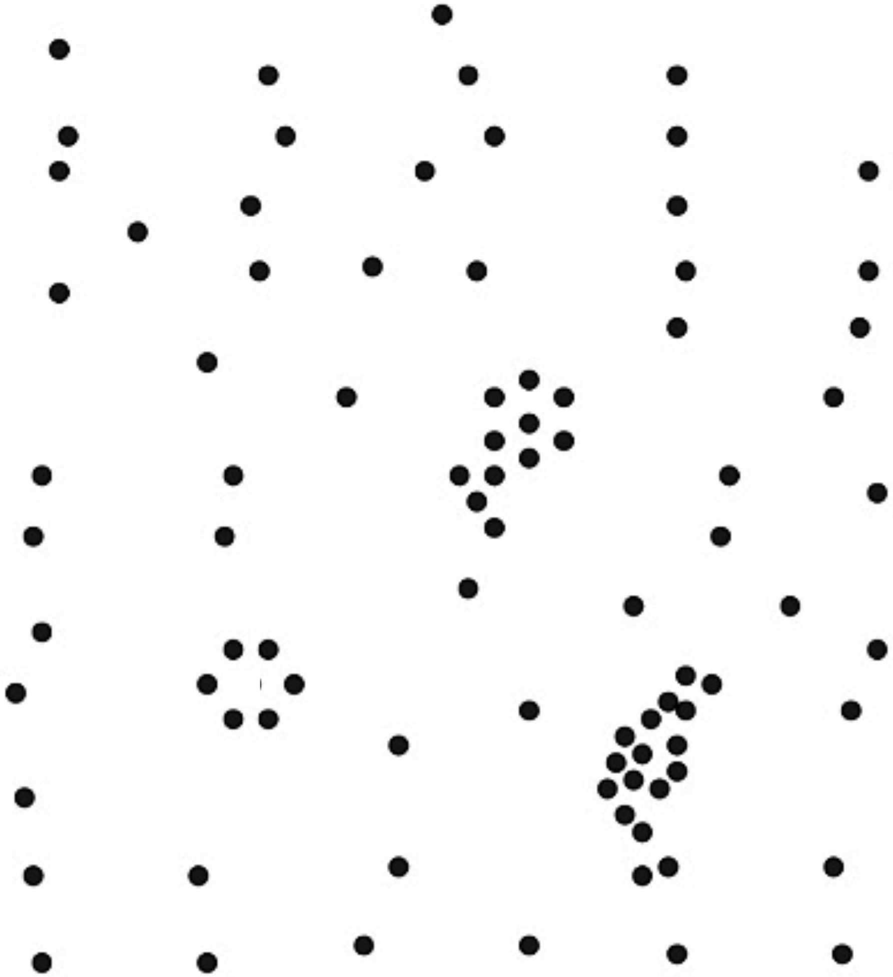
no universal agreement as to what constitutes a cluster

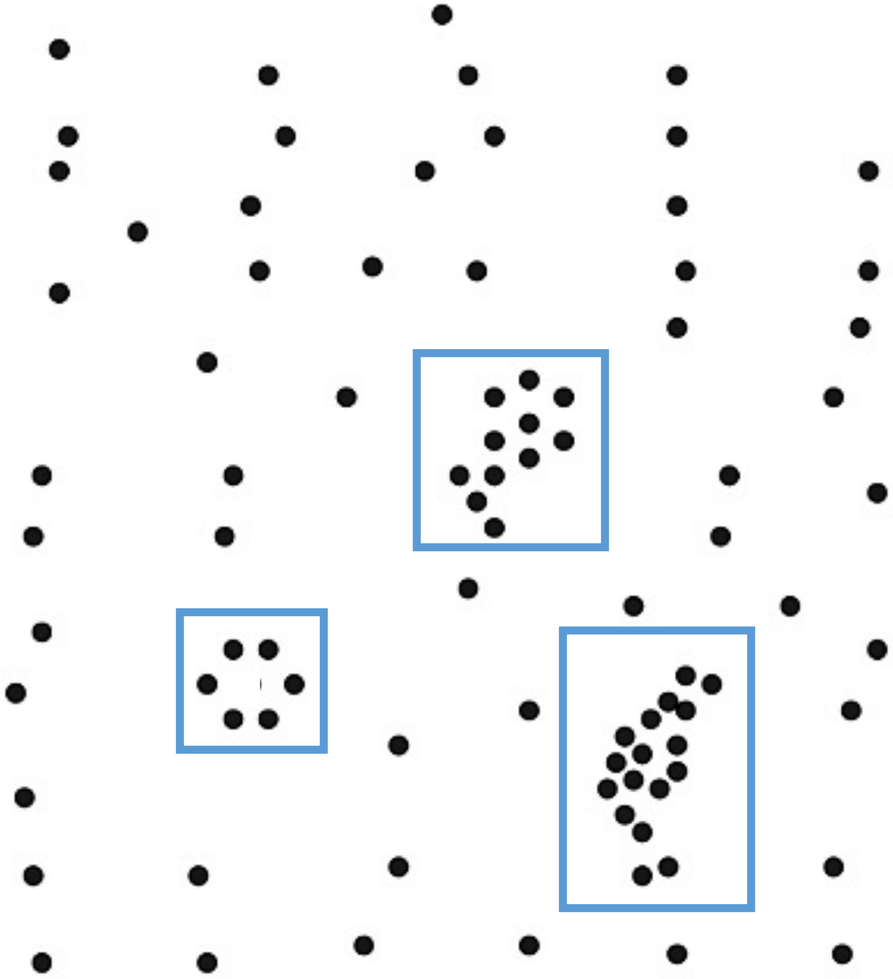
Lack of repeatability

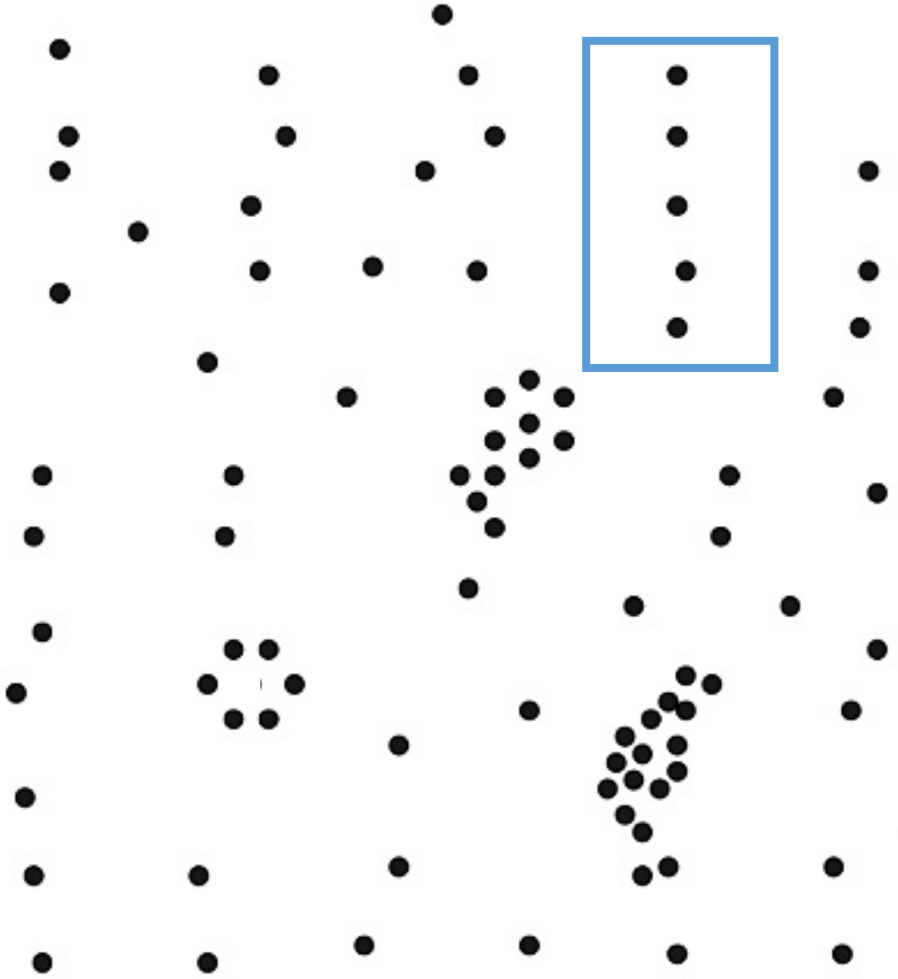
non-deterministic: the same algorithm, applied twice to the same dataset can discover completely different clusters

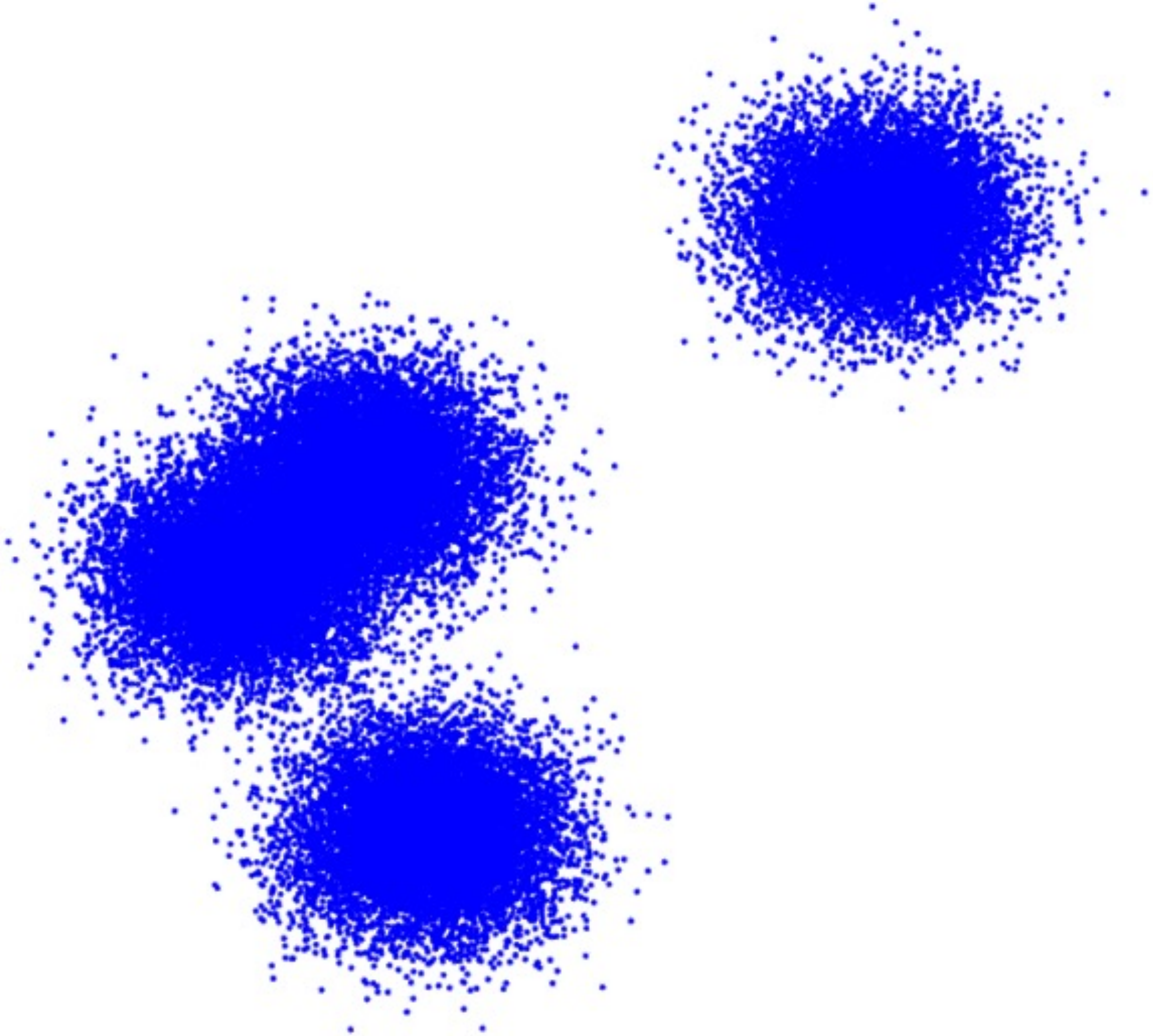
Number of clusters

optimal number of clusters difficult to determine









Clustering Challenges

Cluster description

should clusters be described using representative instances or average values?

Model validation

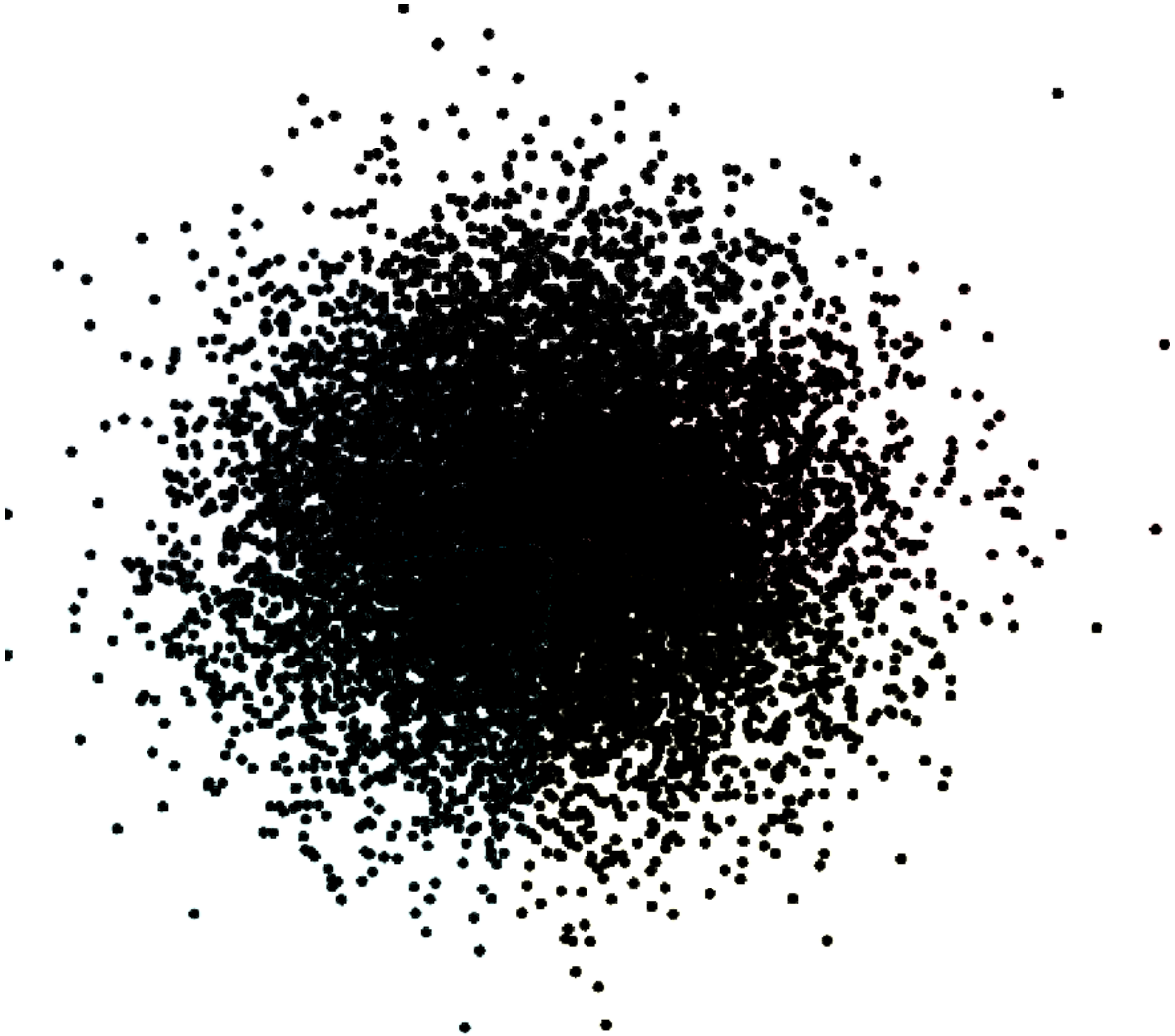
no true clustering information against which to contrast the clustering scheme, so how do we determine if it is appropriate?

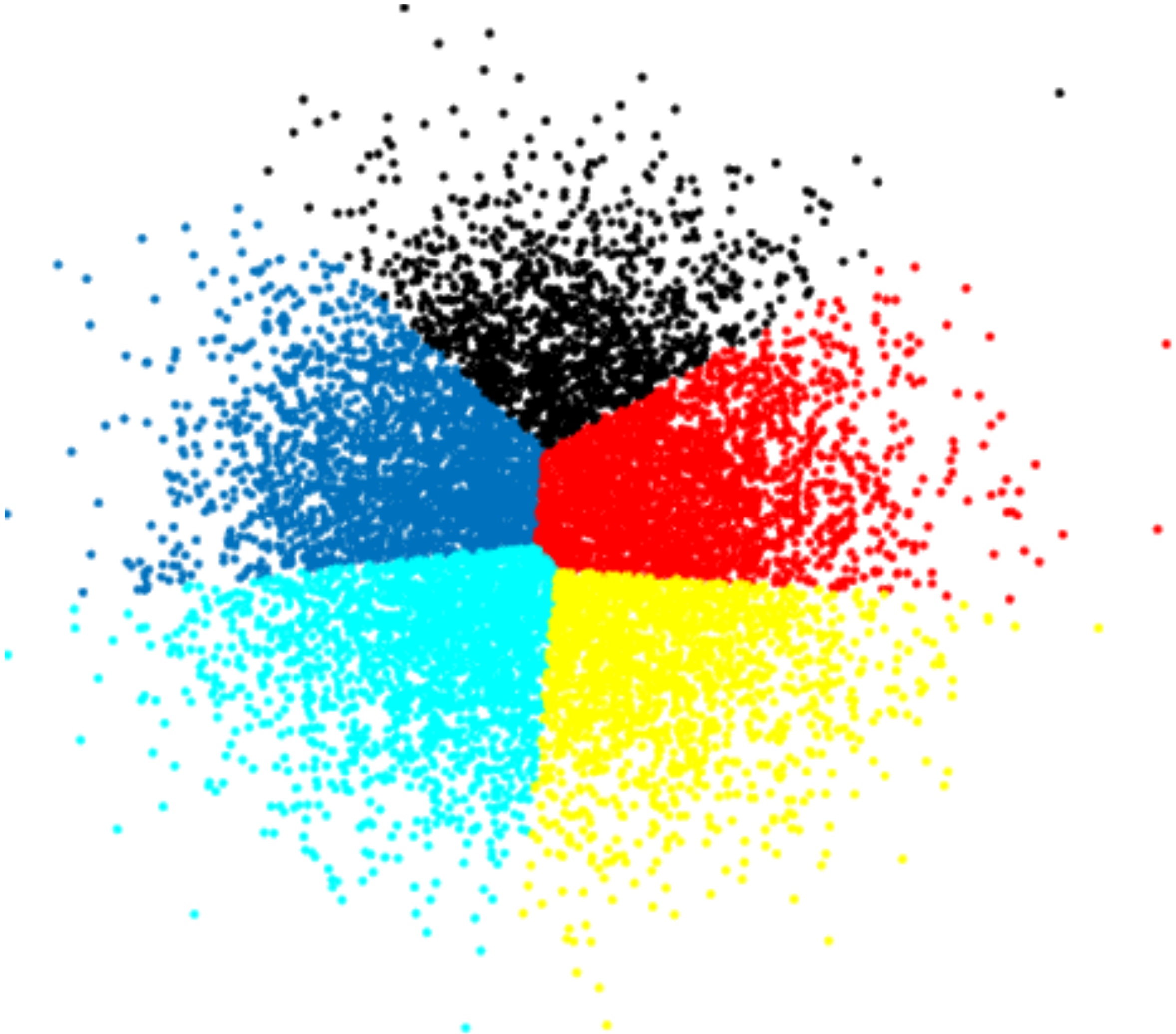
Ghost clustering

most methods will find clusters even if there are none in the data

A posteriori rationalization

once clusters have been found, it is tempting to try to "explain" them …





Gapminder Exercises

Do the exercises for Module 6.

Bad Data

Does the dataset pass the **smell test**? (invalid entries, etc.)

Detecting **lies** and **mistakes** (reporting errors, use of polarizing language)

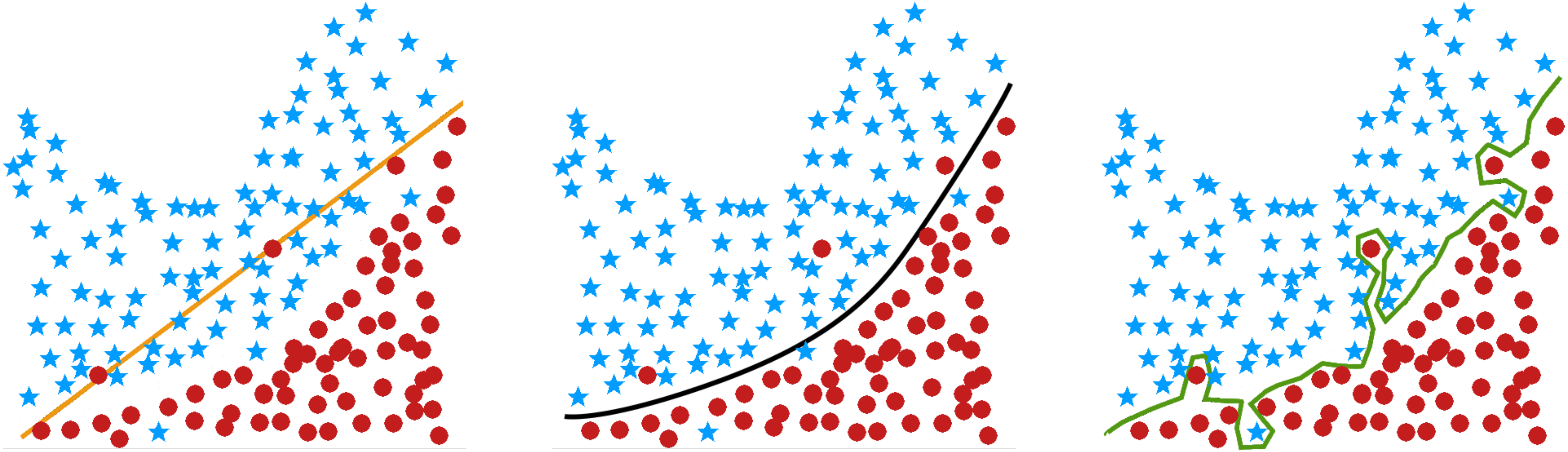
Is **close enough, good enough**?

Sources of **bias** and **errors**

Seeking **perfection** (academic, professional, government, service data)

Data science pitfalls: analysis without understanding, using only one tool (by choice/fiat), analysis for the sake of analysis, unrealistic expectations of data science.

Goldilocks and the Three Models



underfit

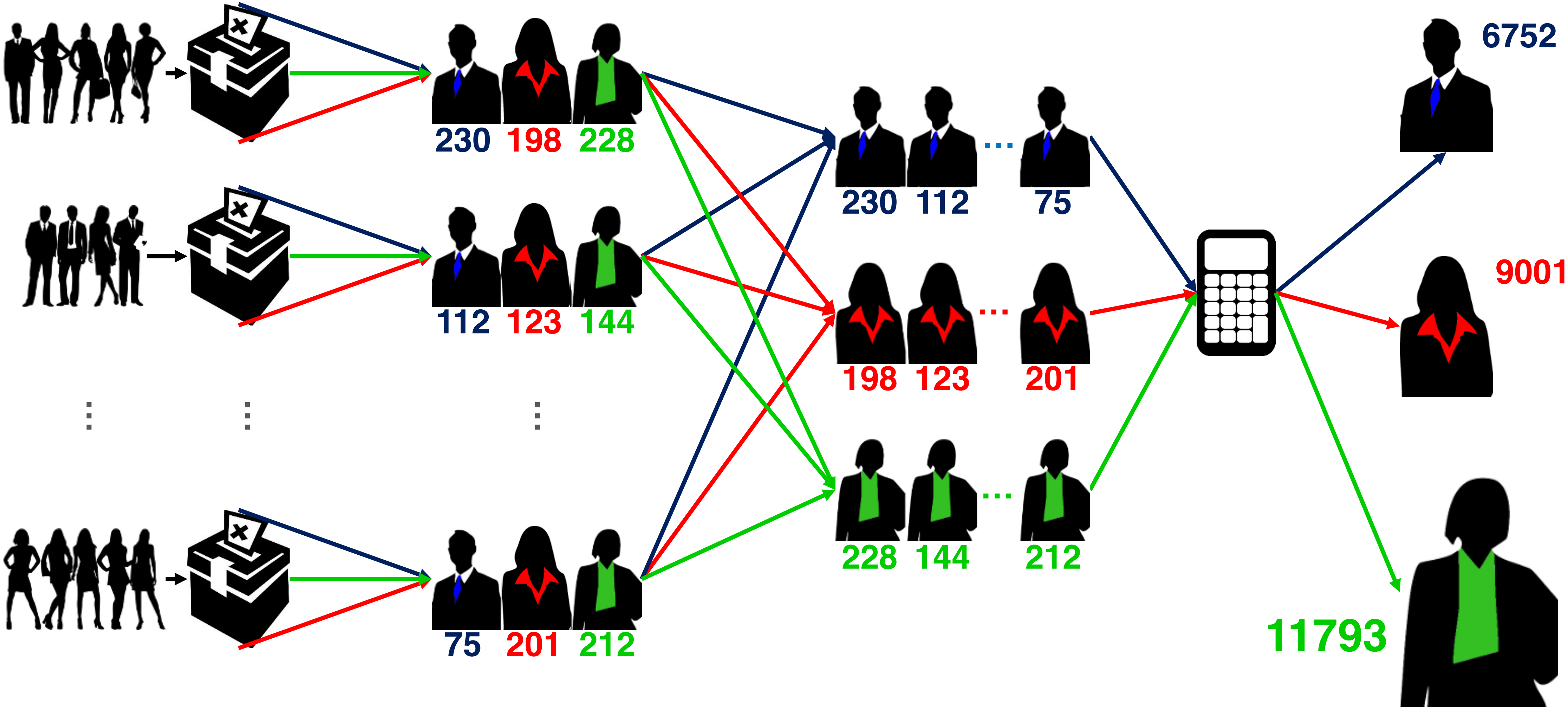
just right

overfit

Big Data vs. Small Data

Parameters	Traditional Data	Big Data
Volume	GB	TB or PB
Generate	per hour, per day	every second or microsecond
Structure	structured	semi-structured or unstructured
Data Source	centralized	fully distributed
Data Integration	easy	difficult
Data Store	RDBMS	HDFS, NoSQL
Data Store	interactive	batch or near real time
Access Update Scenario	repeated read and write	write once repeated read
Data Structure	static schema	dynamic schema
Scaling Potential	non-linear	somewhat close to linear

Analogy: Election



Analogy: Pizzeria

The gains from parallelism depend on whether serial algorithms can be adapted to make use of parallel hardware.

Pizzeria analogy for limitations of parallelization/bottleneck:

- multiple cooks can prepare toppings in parallel
- but baking the crust can't be parallelized
- doubling oven space will increase the number of pizzas that can be made simultaneously but won't substantially speed up any one pizza
- sometimes bottlenecks prevent any gains from parallelism: people line up on both sides of a table to get some soup but there's only one ladle

Biases, Fallacies, and Interpretation

When consulting (or conducting) studies, you should try to determine how the following biases could have come into play:

- **Selection bias** (what data was included, how was it selected?)
- **Omitted-variable bias** (were relevant variables ignored?)
- **Detection bias** (did prior knowledge affect the results?)
- **Funding bias** (who's paying for this?)
- **Publication bias** (what's not being published?)
- **Data-snooping bias** (trying too hard?)
- **Analytical bias** (did the choice of specific method affect the results?)
- **Exclusion bias** (are specific observations/units being excluded?)
- etc. (there are tons)

Biases, Fallacies, and Interpretation

Correlation is not causation (but it is a hint!)

Extreme patterns can mislead

Stay within a study's range

Keep the base rate in mind

Odd results happen (Simpson's Paradox)

Randomness plays a role

Human component to any analytical activity

Small effects can be (statistically) significant

Beware of sacrosanct statistics (p-value, etc.)

[A.K. Maheshwari, Business Intelligence and Data Mining]

Data Science Myths & Mistakes

Mistake #1 – Selecting the wrong problem.

Mistake #2 – Getting buried under tons of data without metadata understanding.

Mistake #3 – Not planning the data analysis process.

Mistake #4 – Insufficient business and domain knowledge.

Mistake #5 – Using incompatible data analysis tools.

Mistake #6 – Using tools that are too specific.

Mistake #7 – Ignoring individual predictions/records in favour of aggregated results.

Mistake #8 – Running out of time.

Mistake #9 – Measuring results differently than the sponsor.

Mistake #10 – Naïvely believing what one’s told about the data.

What We Didn’t Talk About

Tons of other classification and clustering algorithms

Recommender systems

Data streams

Natural language processing (in depth)

Feature selection and dimension reduction (curse of dimensionality)

Data engineering

… and much, much more!

The Future of DS/ML/AI

Self-driving vehicles

Machine translation and language understanding

Detection and prevention of climate and ecosystem disturbances

Automated data science (?!)

Detection and prevention of astronomical catastrophic events

Explainable A.I.

What else?

Future Trends

New questions

New tools

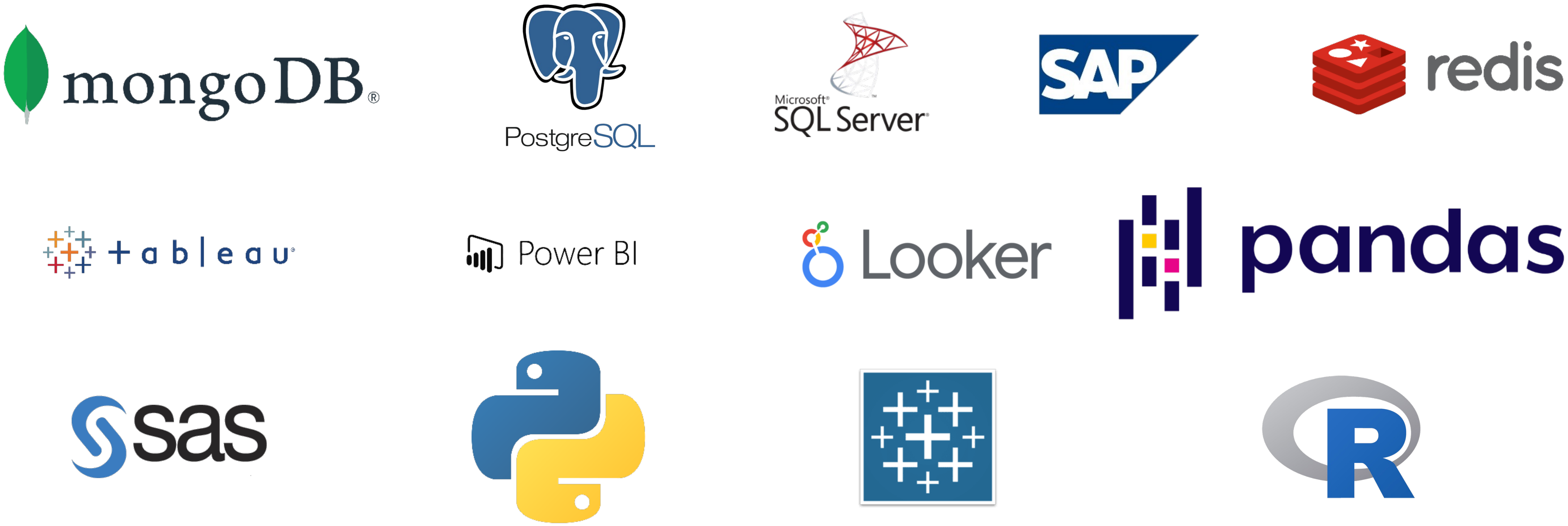
New data sources

Data science as job component

Augmented/swarm intelligence

What else?

Data Science Buffet



[Joe Blitzstein and Hanspeter Pfister, http://cs109.org]

In Conclusion

Data science is a team activity, with subject matter experts.

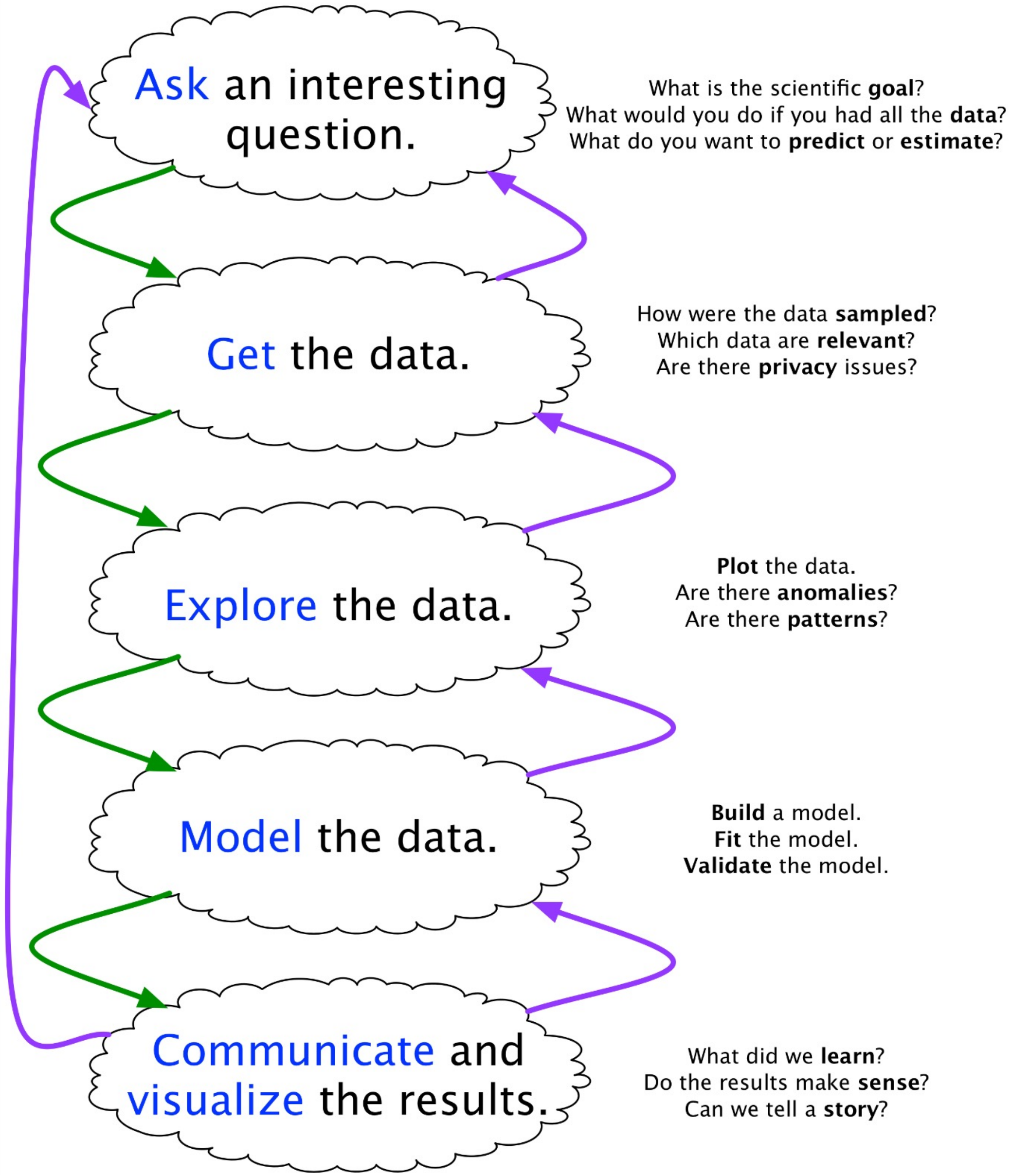
Ethical considerations are paramount and need not conflict with profitability.

Let the data speak (but be careful).

Look for actionable insights.

Supervised vs. unsupervised vs…

Much time must be spent on data preparation.



Ask an interesting question.

What is the scientific **goal**?
What would you do if you had all the **data**?
What do you want to **predict** or **estimate**?

Get the data.

How were the data **sampled**?
Which data are **relevant**?
Are there **privacy** issues?

Explore the data.

Plot the data.
Are there **anomalies**?
Are there **patterns**?

Model the data.

Build a model.
Fit the model.
Validate the model.

Communicate and visualize the results.

What did we **learn**?
Do the results make **sense**?
Can we tell a **story**?