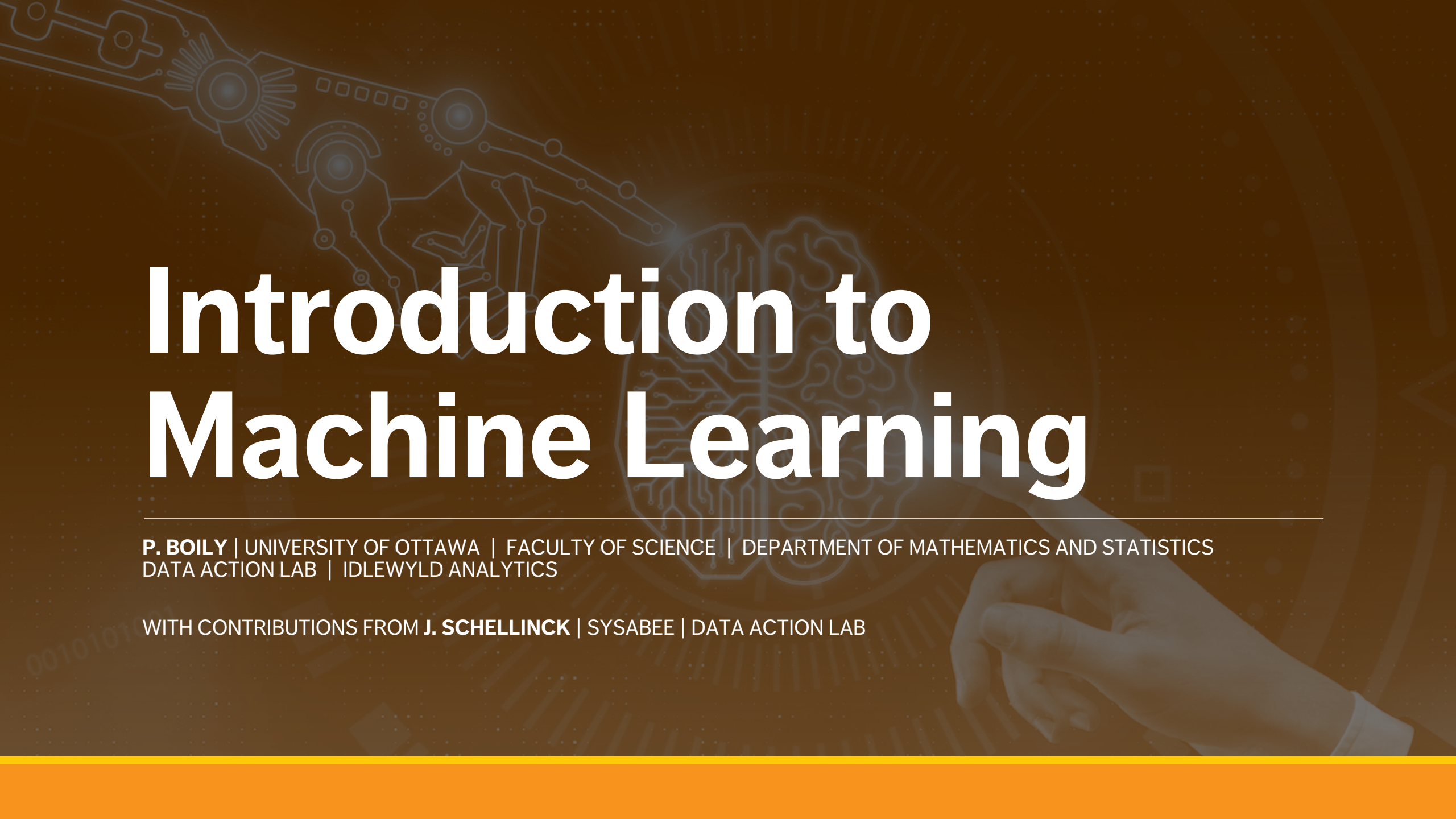
Introduction to Machine Learning

Instructor: Patrick Boily

uOttawa

Institut de développement professionnel
Professional Development Institute



Introduction to Machine Learning

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

WITH CONTRIBUTIONS FROM **J. SCHELLINCK** | SYSABEE | DATA ACTION LAB

Instructor – Patrick Boily

Employment

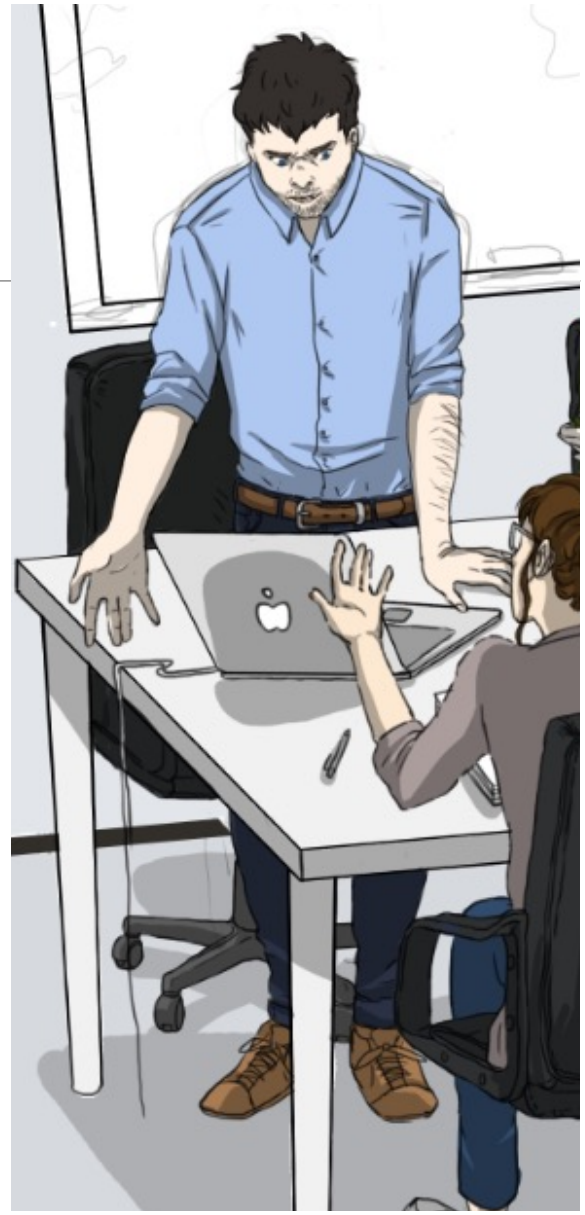
- Professor Math/Stat [‘19 – now, uOttawa]
- President [‘16 – now, Idlewyld Analytics]
- Manager and Senior Consultant [’12 – ‘19, CQADS, Carleton]
- Public Service [’08 – ’12, ASFC | StatCan | TC | TPSGC]
- 60+ uni course; 250+ workshop days

Projects

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

Specialization

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



Course Material

Course Webpage:

https://data-action-lab.com/103-iml

Course Notes:

https://idlewyldanalytics.com

Contact Info:

pboily@uottawa.ca

Slack Workspace:

https://dspdi.slack.com

Course Description

This course leads the participants to analyze and discuss the general tasks and problems of statistical learning (machine learning), as well as their pitfalls.

In this course, participants will be introduced to simple association rules mining, classification, and clustering algorithms.

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

Additional Information

Participants are expected to be familiar with the concepts introduced in the courses *Data Science Essentials* (data preparation, data cleaning), and *Data Visualization and Dashboards* (data exploration), and their pre-requisites.

Familiarity with optimization methods would be beneficial but is not required.

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

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

Learning Outcomes

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

- differentiate between situations which require a supervised learning approach and those which require an unsupervised approach (or some combination of both)
- identify strategies used to overcome common real-world statistical learning issues and challenges
- recognize the variety of machine learning algorithms available to them
- implement simple machine learning algorithms to provide actionable insights
- build a simple data analysis pipeline incorporating machine learning components

Course Outline

Statistical Learning

1. Types of Learning;
   Machine Learning Tasks

Association Rules Discovery

2. Association Rules Overview;
   Case Study: Danish Medical Data
3. Association Rules Concepts

Classification

4. Classification Overview;
   Case Study: Minnesota Tax Audits
5. Decision Trees and Other Algorithms
6. Performance Evaluation

Session 1
Session 2
Session 3
Session 4

Course Outline

Clustering

7. Clustering Overview; *Case Study*: Livehoods
8. k-Means and Other Algorithms
9. Validation and Notes

Issues and Challenges

10. Bad Data and Big Data
11. Underfitting and Overfitting/Transferability
12. Miscellanea

Session 1

Session 2

Session 3

Session 4

Sister Courses

DATA SCIENCE ESSENTIALS

1. Non-Technical Aspects
2. Data Science Basics
3. Data Preparation
4. Data Engineering

DATA VISUALIZATION AND DASHBOARDS

1. Data Viz Concepts
2. Dashboarding
3. Storytelling with Data
4. Data Viz with ggplot2

POWER BI FOR BEGINNERS

1. The Tool
2. Exploration
3. Monitoring
4. Storytelling

Session 1

INTRODUCTION TO MACHINE LEARNING

Statistical Learning

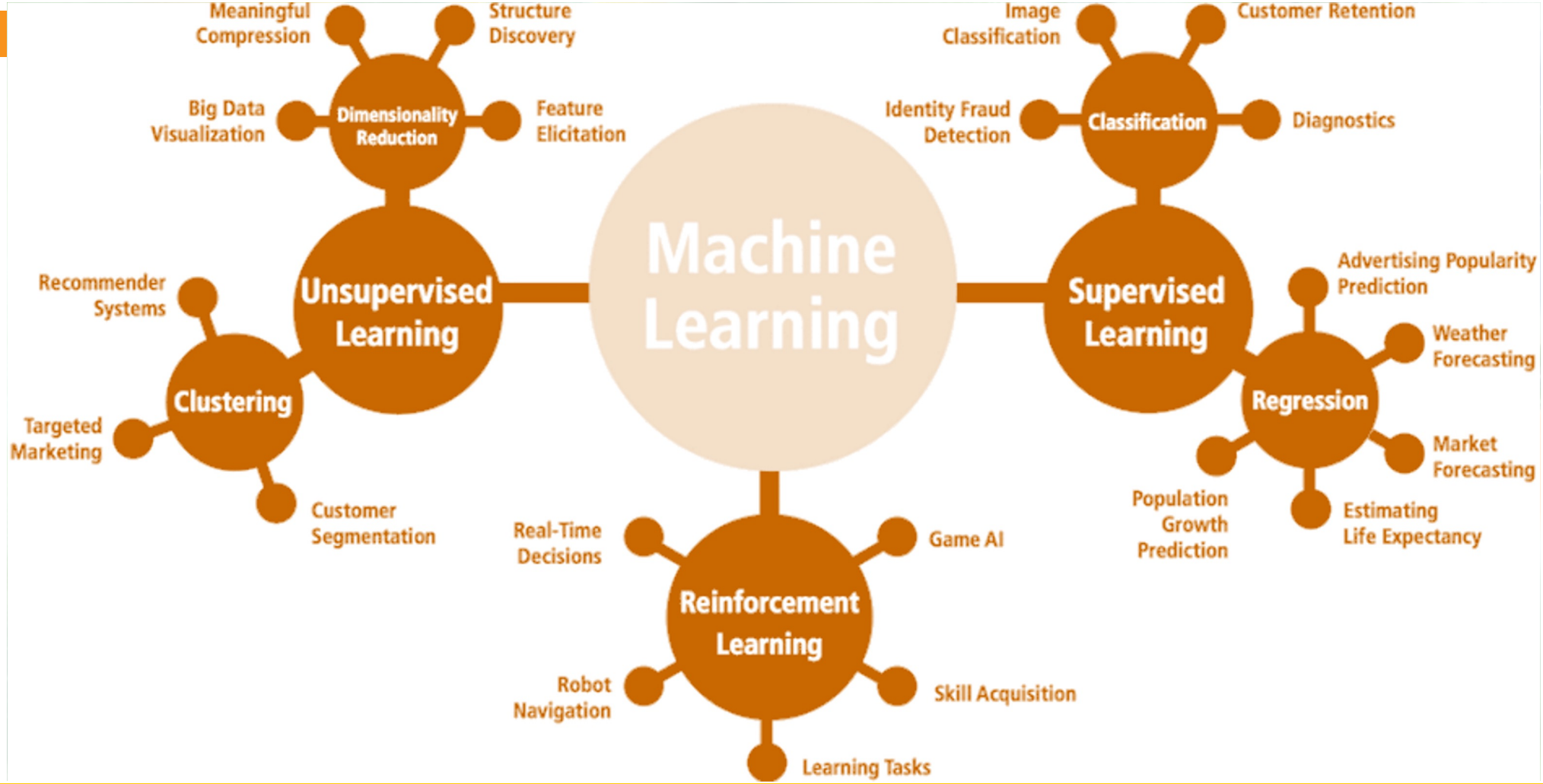
INTRODUCTION TO MACHINE LEARNING

We learn from failure, not from success!

[B. Stoker, Dracula]

Data is not information, information is not knowledge, knowledge is not understanding, understanding is not wisdom.

[C. Stoll (attributed), *Nothing to Hide: Privacy in the 21st Century*, 2006]



1. Types of Learning and Machine Learning Tasks

A Modern Challenge

One challenge of working in the **data science** (DS), **machine learning** (ML), and **artificial intelligence** (AI) fields : most quantitative work can be described as DS/ML/AI (this is often stretched to a ridiculous extent).

DS/ML/AI consists of quantitative processes (what H. Mason has called “the **working intersection** of statistics, engineering, computer science, domain expertise, and”hacking”) that help users **learn actionable insights** about their situation without completely abdicating their decision-making responsibility.

A Hierarchical Framework

Robinson suggests an "**inclusive hierarchical structure**":

1. in a first stage, DS provides "**insights**" *via* visualization and (manual) inferential analysis;
2. in a second stage, ML yields "**predictions**" (or "**advice**"), while reducing (not eliminating) the operator's analytical, inferential and decisional workload;
3. in the final stage, AI **removes the need for oversight**, allowing for automatic "actions" to be taken by a mostly unattended system (general AI vs. augmented intelligence).

In practice, stakeholders should probably not seek to abdicate **all** of their agency in the decision-making process.

Learning

Humans learn (at all stages) by first **taking in their environment**, and then by:

- answering questions about it;
- testing hypotheses;
- creating concepts;
- making predictions;
- creating categories, and
- classifying and grouping its various objects and attributes.

Statistical/Machine Learning

The main goal of DS/ML/AI is to try to **teach** machines to extract insight from data, properly and efficiently, and free of biases and pre-conceived notions – in other words, **can (should?) we design algorithms that can learn?**

The simplest DS/ML/AI method is **exploring representative data** to:

- provide a summary through basic statistics – mean, mode, histograms, etc.;
- make its multi-dimensional structure evident through data data visualization; and
- look for consistency, considering what is in there and what is missing.

Supervised Learning

Supervised learning (SL) is akin to "**learning with a teacher**": students give an answer to each exam question based on what they learned from worked-out examples provided by the teacher/textbook; the teacher provides the correct answers and marks the exam questions using a key.

Typical tasks include:

- **classification**
- regression
- rankings
- recommendations

Supervised Learning

In SL, algorithms use **labeled training data** to build (or train) a **predictive model**; each algorithm's performance is evaluated using **test data** for which the label is known but not used in the prediction.

There are fixed **targets** against which to train the model (such as age categories, or plant species) – these categories/classes (and their number) are known **prior to the analysis**.

Unsupervised Learning

Unsupervised learning (UL) is akin to "**self-learning by grouping similar exercises together as a study guide**": the teacher is not involved in the discovery process and students might end up with different groupings.

Typical tasks include:

- **clustering**
- **association rules discovery**
- link profiling
- anomaly detection

Unsupervised Learning

Unsupervised algorithms use **unlabeled data** to find **natural patterns** in the data; the drawback is that accuracy **cannot be evaluated** with the same degree of satisfaction.

In UL, we don't know what the target is, or even if there is one – we are simply looking for **natural groups/associations** in the data, such as:

- junior students who like literature, have longish hair, and know how to cook **vs.**
- students who are on a sports team and have siblings **vs.**
- financial professionals with a penchant for superhero movies, craft beer and Hello Kitty backpack **vs.** ...

Other Learning Frameworks

Some DS/ML/AI techniques fit into both camps; others can be either SL or UL, but there are other **conceptual approaches** (usually for AI tasks):

- **semi-supervised learning** in which some data points have labels but most do not, which often occurs when acquiring data is costly ("the teacher provides worked-out examples and a list of unsolved problems to try out; the students try to find similar groups of unsolved problems and compare them with the solved problems to find close matches")
- **reinforcement learning**, where an agent attempts to collect as much (short-term) reward as possible while minimizing (long-term) regret ("embarking on a Ph.D. with an advisor… with all the highs and the lows and **maybe** a diploma at the end of the process?")

Statistical Learning/Machine Learning

The term “statistical learning” is not used frequently in practice (except by mathematicians and statisticians); the tendency is to speak instead of **machine learning**.

If a distinction must be made, it could be argued that:

- statistical learning arises from statistical-like models, and the emphasis is usually placed on **interpretability**, **precision**, and **uncertainty**, whereas
- machine learning arise from artificial intelligence studies, with emphasis on **large scale applications** and **prediction accuracy**.

The dividing line between the terms is blurry – the vocabulary used by practitioners mostly betrays their educational backgrounds.

DS/ML Questions

Outside of academia, DS/ML/AI methods are only really interesting when they help users ask and answer useful questions. Compare, for instance:

- **Analytics** – "How many clicks did this link get?"
- **Data Science** – "Based on the previous history of clicks on links of this publisher's site, can I predict how many people from Manitoba will read this specific page in the next three hours?" or "Is there a relationship between the history of clicks on links and the number of people from Manitoba who will read this specific page?"
- **Quantitative Methods** – "We have no similar pages whose history could be consulted to make a prediction, but we have reasons to believe that the number of hits will be strongly correlated with the temperature in Winnipeg. Using the weather forecast over the next week, can we predict how many people will access the specific page during that period?"

DS/ML Questions

DS/ML models are **predictive/descriptive** (not **explanatory/prescriptive**): they show connections, and exploit correlations to make predictions, but they don't reveal **why** such connections exist (Bayesian networks).

Quantitative methods, on the other hand, usually assume a certain level of **causal understanding** based on various **first principles**. That distinction is not always understood properly by analysts and stakeholders.

ML Tasks

Common ML tasks (with representative questions) include:

- **classification** – which undergraduates are likely to succeed at the graduate level?
- **probability estimation** – how likely is it that a given candidate will win an election?
- **value estimation** – how much is a given client going to spend at a restaurant?
- **similarity matching** – which prospective clients are most similar to established best clients?
- **clustering** – do signals from a sensor form natural groups?
- **association rules discovery** – what books are commonly purchased together online?
- **profiling** and **behaviour description** – what is the typical cell phone usage of a certain customer's segment?
- **link prediction** – J. and K. have 20 friends in common: perhaps they'd be great friends?

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

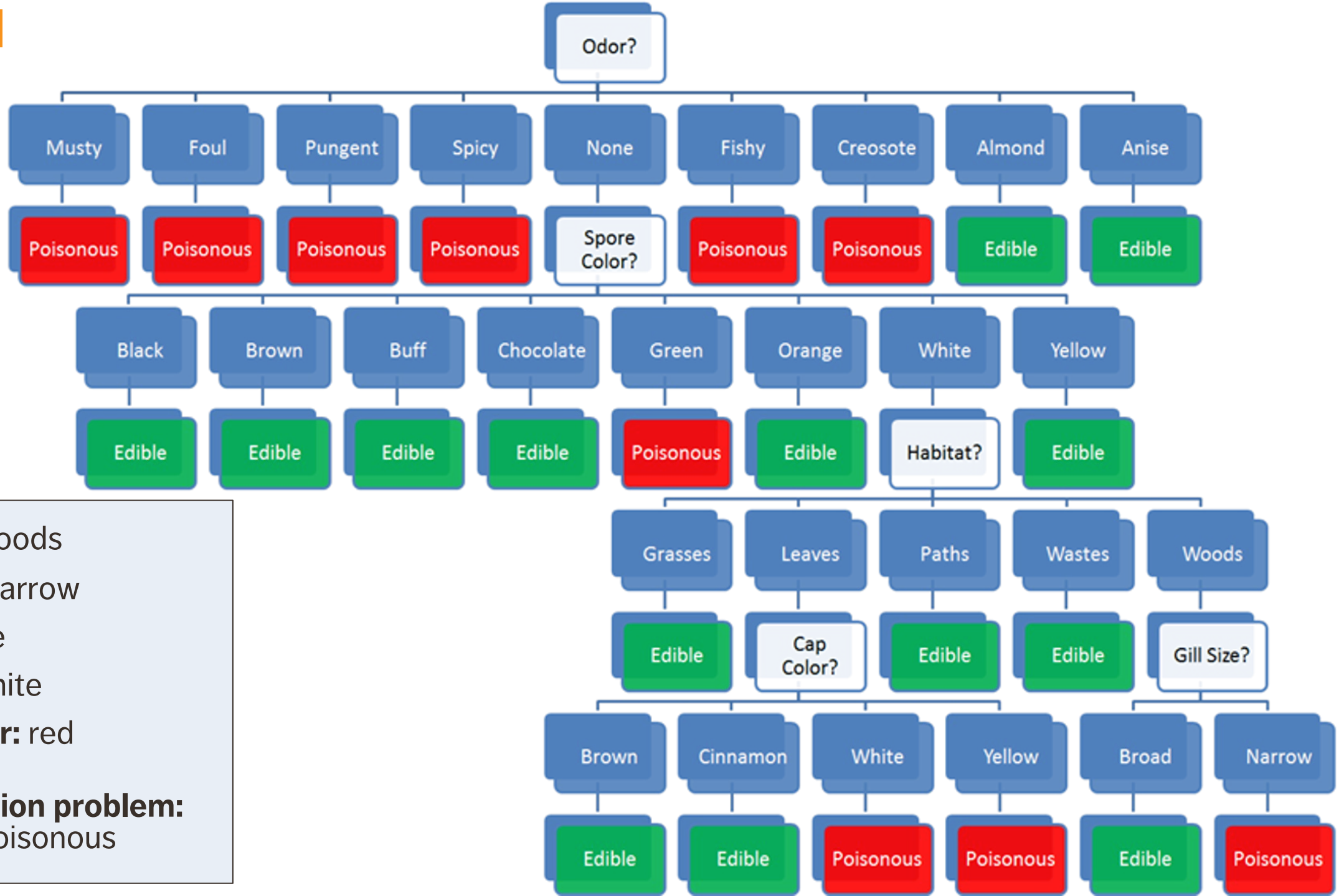
Mushroom Classification Problem

Amanita muscaria

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

Classification problem:
Is Amanita muscaria edible or poisonous?

Session 1



Habitat: woods

Gill Size: narrow

Odor: none

Spores: white

Cap Colour: red

Classification problem:
edible or poisonous

Session 1

Habitat: woods

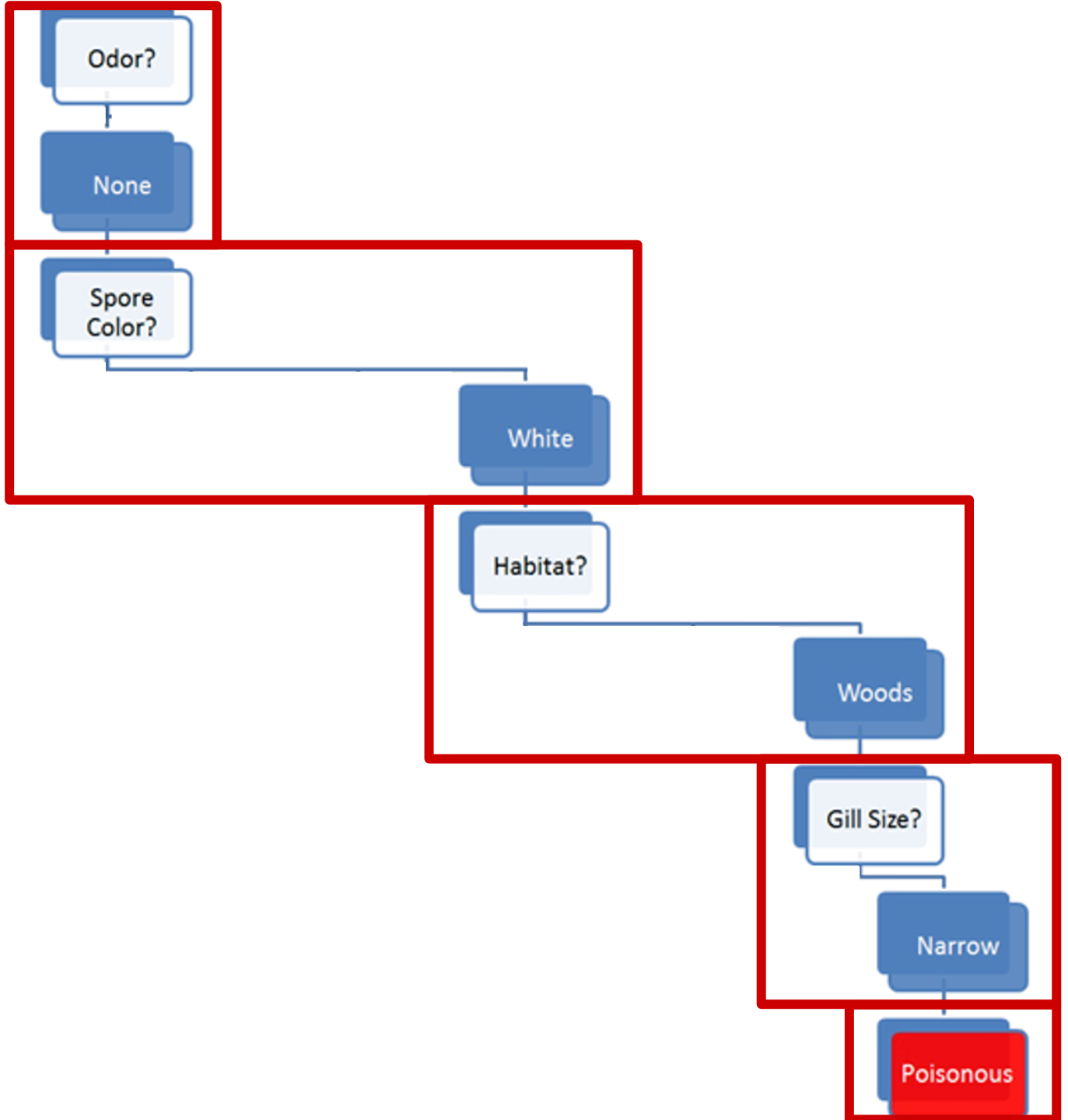
Gill Size: narrow

Odor: none

Spores: white

Cap Colour: red

Classification problem:
edible or **poisonous**



Mushroom Classification Problem

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?

Suggested Reading

Types of Learning and Machine Learning Tasks

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

D. Woods, "Bitly's Hilary Mason on "what is a data scientist?"," *Forbes*, Mar. 2012.

Data Understanding, Data Analysis, Data Science
Regression and Value Estimation

*Statistical Learning (advanced)

- Supervised Learning Framework
- Systematic Compontent and Regression Function

Exercises

Types of Learning and Machine Learning Tasks

1. Of what types of machine learning tasks are the following problems representative?
   - Identifying risk factors associated to breast/prostate cancer.
   - Predicting whether a patient will have a second, fatal heart attack within 30 days of the first on the basis of demographics, diet, clinical measurements, etc.
   - Establishing the relationship between salary and demographic information in population survey data.
   - Predicting the yearly inflation rate using various indicators.

2. What are some examples of supervised, unsupervised, semi-supervised, reinforcement machine learning tasks in the business world? In a public policy/government setting? In a scientific setting?

Exercises

Types of Learning and Machine Learning Tasks

3. Assuming that DS/ML/AI techniques are used in the following cases, identify whether the required task falls under SL or UL.
   a. Estimating the repair time required for an aircraft based on a trouble ticket.
   b. Deciding whether to issue a loan to an applicant based on demographic and financial data (with reference to a database of similar data on prior customers).
   c. In an online bookstore, making recommendations to customers concerning additional items to buy based on the buying pattern in prior transactions.
   d. Identifying a network data packet as dangerous (virus, hacker attack) based on comparison to other packets with a known threat status.
   e. Identifying segments of similar customers.

Exercises

Types of Learning and Machine Learning Tasks

3. Assuming that DS/ML/AI techniques are used in the following cases, identify whether the required task falls under SL or UL.
   f. Predicting whether a company will go bankrupt based on comparing its financial data to those of similar bankrupt and non-bankrupt firms.
   g. Automated sorting of mail by zip code scanning.
   h. It is more difficult and expensive to win new customers than it is to retain existing customers. Scoring each customer on their likelihood to quit can help an organization design effective interventions, such as discounts or free services, to retain profitable customers in a cost-effective manner.
   i. Some medical practitioners conduct unnecessary tests and/or over-bill their government or insurance companies. Using audit data, it may be possible to identify such providers and take appropriate action.

Exercises

Types of Learning and Machine Learning Tasks

3. Assuming that DS/ML/AI techniques are used in the following cases, identify whether the required task falls under SL or UL.
   j. A market basket analysis can help develop predictive models to determine which products often sell together. This knowledge of affinities between products can help retailers create promotional bundles to push non-selling items along a set of products that sell well.
   k. Diagnosing the cause of a medical condition is the crucial first step in medical engagement. In addition to the current condition, other factors can be considered, including the patient's health history, medication history, family's history, and other environmental factors. A predictive model can absorb all of the information available to date (for this patient and others) and make probabilistic diagnoses, in the form of a decision tree, taking away most of the guess work involved.
   l. Schools can develop models to identify students who are at risk of not returning to school. Such students can be flagged to be on the receiving end of potential corrective measures.

Exercises

Types of Learning and Machine Learning Tasks

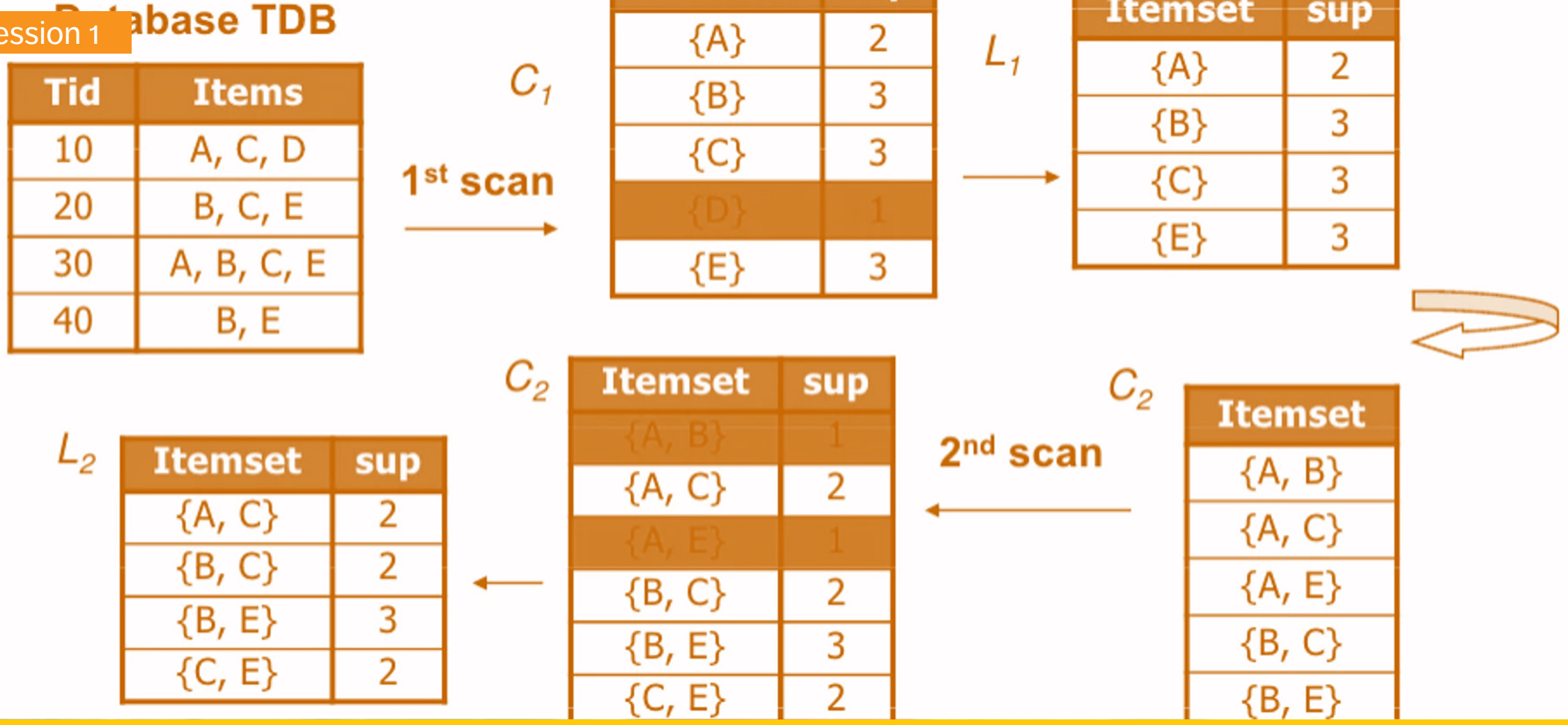
3. Assuming that DS/ML/AI techniques are used in the following cases, identify whether the required task falls under SL or UL.
    13. In addition to customer data, telecom companies also store call detail records (CDR), which precisely describe the calling behaviour of each customer. The unique data can be used to profile customers, who may be marketed to based on the similarity of their CDR to other customers'.
    14. Statistically, all equipment is likely to break down at some point in time. Predicting which machine is likely to shut down is a complex process. Decision models to forecast machinery failure could be constructed using past data, which can lead to savings provided by preventative maintenance.
    15. Identifying which tweets contain disinformation and which tweets are legitimate.

Association Rules Mining

INTRODUCTION TO MACHINE LEARNING

Correlation isn’t causation... but it’s a big hint!

[E. Tufte]



2. Association Rules Overview

Overview

Association rules discovery (ARD) is a type of unsupervised learning that finds **connections** among the attributes and levels of a dataset's observations.

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 smart phones (the **conclusion**)
- individuals who have purchased home gym equipment are unlikely to be using it 1 year later

Overview

The presence of a **correlation** between the premise and the conclusion does not imply the existence of a **causal relationship** between them.

It is difficult to prove causation *via* data analysis; in practice, decision-makers pragmatically (often erroneously) focus on "**there's no smoke without fire**."

Example: being a triathlete does not cause one to drive a Subaru, but Subaru Canada thought that the connection was strong enough to offer to reimburse the registration fee at an IRONMAN 70.3 competition (at least in 2018)!

Market Basket Analysis

ARD is aka as **market basket analysis**.

Example: purchase of bread and milk, but that is unlikely to be of interest given the frequency of market baskets containing milk (**or** bread).

If the presence of milk is **independent** of the presence of bread (and *vice-versa*), and if 70% of baskets contain milk and 90% contain bread, say, we would expect **at least** 90% × 70% = 63% of all baskets to contain **both**.

If we observe both in 72% of baskets, say (a 1.15-fold increase), we conclude that there is a **weak correlation** between the milk and bread purchases.

Market Basket Analysis

Sausages and buns are not purchased as frequently as milk and bread, but they might still be purchased as a pair more often than one would expect.

If the presence of sausage is **independent** of the presence of buns (and *vice-versa*), and if 10% of baskets contain sausages and 5% contain buns, say, we would expect **at least** 10% × 5% = 0.5% of all baskets to contain **both**.

If we observe both in 4% of baskets, say (an 8-fold increase), we conclude that there is a **strong correlation** between the sausage and buns purchases.

Market Basket Analysis

How can we **act** on this insight? Supermarkets could advertise a sale on sausages while **simultaneously** (and quietly) raising the price of buns. This could have the effect of bringing in a higher number of customers into the store, hoping to increase the **sale volumes** for both items while keeping the **combined price of the two items constant**.

Little Story: a supermarket found an association rule linking the purchase of beer and diapers and consequently moved its beer display closer to its diapers display, having confused correlation and causation.

What do you think might actually be happening here?

Applications

Typical uses include:

- finding **related concepts** in text documents – looking for pairs (triplets, etc) of words that represent a joint concept: {San Jose, Sharks}, {Michelle, Obama}, etc.;
- detecting **plagiarism** – looking for specific sentences that appear in multiple documents, or for documents that share specific sentences;
- identifying **biomarkers** – finding diseases frequently associated with a set of biomarkers;

Applications

Typical uses include:

- making predictions and decisions based on association rules (there are pitfalls)
- altering circumstances to take advantage of correlations (suspected causal effect)
- using connections to modify the likelihood of certain outcomes (see above)
- imputing missing data
- text autofill and autocorrect
- etc.

Case Study

Danish Medical Data

Jensen *et al.*
Temporal disease trajectories condensed from population-wide registry data covering 6.2 million patients
Nature Communications, vol. 5, 2014

Objective

Using data from the *Danish National Patient Registry*, the authors sought connections between different **diagnoses:** how does a diagnosis at some point in time allow for the prediction of another diagnosis at a later time?

Case Study

Danish Medical Data

Jensen *et al.*
Temporal disease trajectories condensed from population-wide registry data covering 6.2 million patients
Nature Communications, vol. 5, 2014

Methodology

1. compute the **strength of correlation** for pairs of diagnoses over a 5 year interval (on a representative subset of the data)
2. test diagnoses pairs for **directionality** (one diagnosis repeatedly occurring before the other)
3. determine reasonable **diagnosis trajectories** (thoroughfares) by combining smaller (but frequent) trajectories with overlapping diagnoses
4. **validate** the trajectories by comparison with non-Danish data
5. **cluster** the thoroughfares to identify a small number of **central medical conditions** (key diagnoses) around which disease progression is organized

Case Study

Danish Medical Data

Jensen *et al.*
Temporal disease trajectories condensed from population-wide registry data covering 6.2 million patients
Nature Communications, vol. 5, 2014

Data

The *Danish National Patient Registry* is an electronic health registry containing administrative information and diagnoses, covering the whole population of Denmark, including private and public hospital visits of all types:

- inpatient (overnight stay)
- outpatient (no overnight stay)
- emergency visits.

The data set covers 15 years of such visits, from January ’96 to November ’10, and consists of 68 million records for 6.2 million patients.

Case Study

Danish Medical Data

Jensen *et al.*
Temporal disease trajectories condensed from population-wide registry data covering 6.2 million patients
Nature Communications, vol. 5, 2014

Challenges and Pitfalls

- Access to the **patient registry** was protected and could only be granted after approval by *the National Board of Health*.
- There are gender-specific differences in diagnostic trends, but many diagnoses were made predominantly in different sites, suggesting the stratifying by **site** as well as by **gender**.
- In the process of forming small diagnoses chains, they had to compute the correlations using **large groups** for each pair of diagnoses (1 million diagnosis pairs = 80+ million samples) to compensate for **multiple testing** (1000s years' worth of CPU run time) – pre-filtering steps were used to avoid this pitfall.

Case Study

Danish Medical Data

Jensen *et al.*
Temporal disease trajectories condensed from population-wide registry data covering 6.2 million patients
Nature Communications, vol. 5, 2014

Project Summary and Results

The dataset was reduced to **1,171 significant trajectories**.

These thoroughfares were clustered into patterns centred on 5 key diagnoses for disease progression:

- **diabetes**
- **chronic obstructive pulmonary disease** (COPD)
- **cancer**
- **arthritis**
- **cerebrovascular disease**

Case Study

Danish Medical Data

Jensen *et al.*
Temporal disease trajectories condensed from population-wide registry data covering 6.2 million patients
Nature Communications, vol. 5, 2014

Project Summary and Results

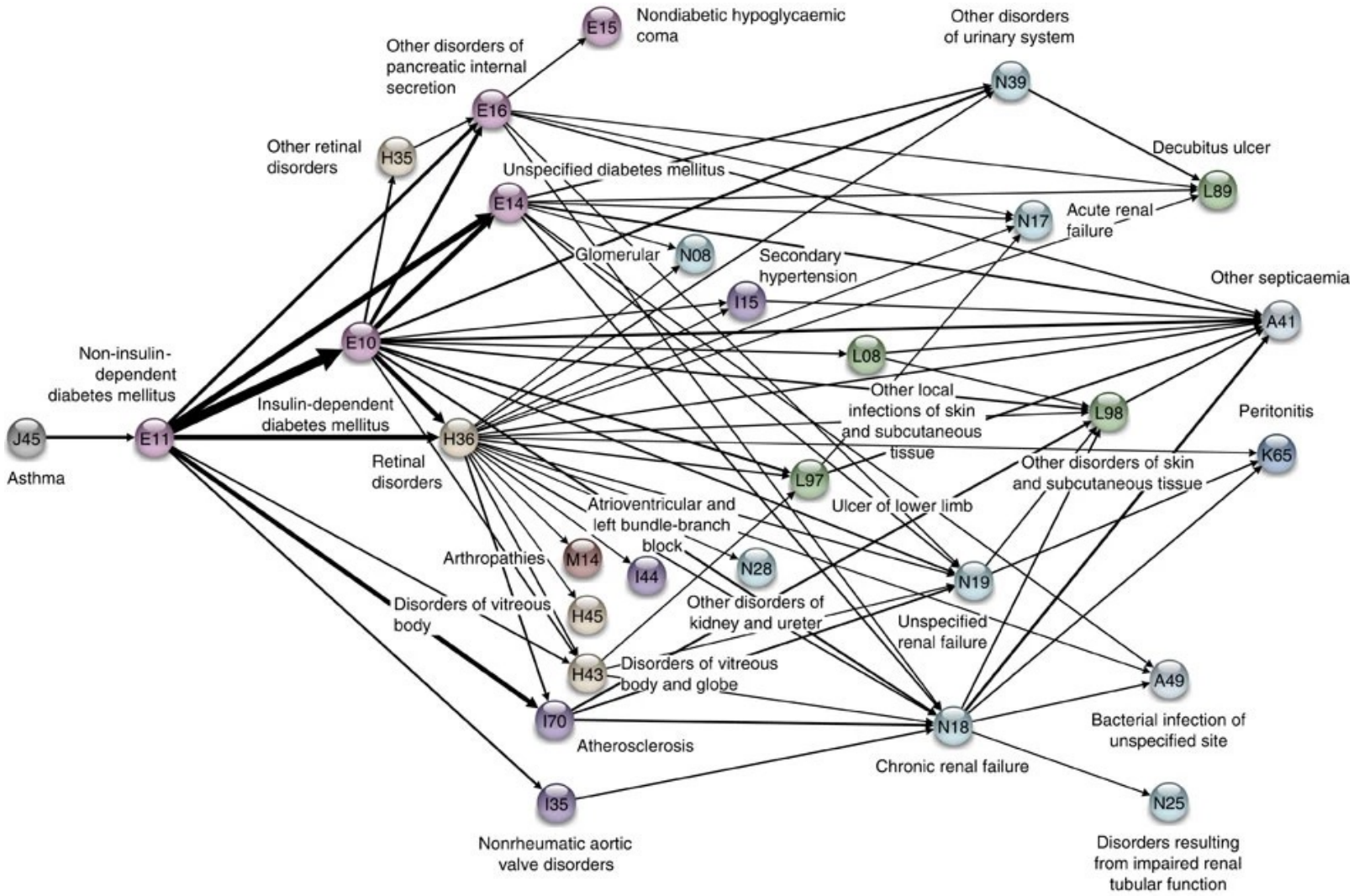
Early diagnoses for these central factors can help reduce the risk of adverse outcome linked to future diagnoses of other conditions.

Among the specific results, the following “surprising” insights were found:

- a diagnosis of anemia is typically followed months later by the **discovery of colon cancer**
- a diagnosis of gout was identified as **a step on the path** toward eventual diagnosis of cardiovascular diseases
- COPD is **under-diagnosed** and **under-treated**

Case Study

Danish Medical Data



Jensen *et al.*
Temporal disease trajectories condensed from population-wide registry data covering 6.2 million patients
Nature Communications, vol. 5, 2014

Suggested Reading

Association Rules Overview

Data Understanding, Data Analysis, Data Science
Machine Learning 101

Association Rules Mining

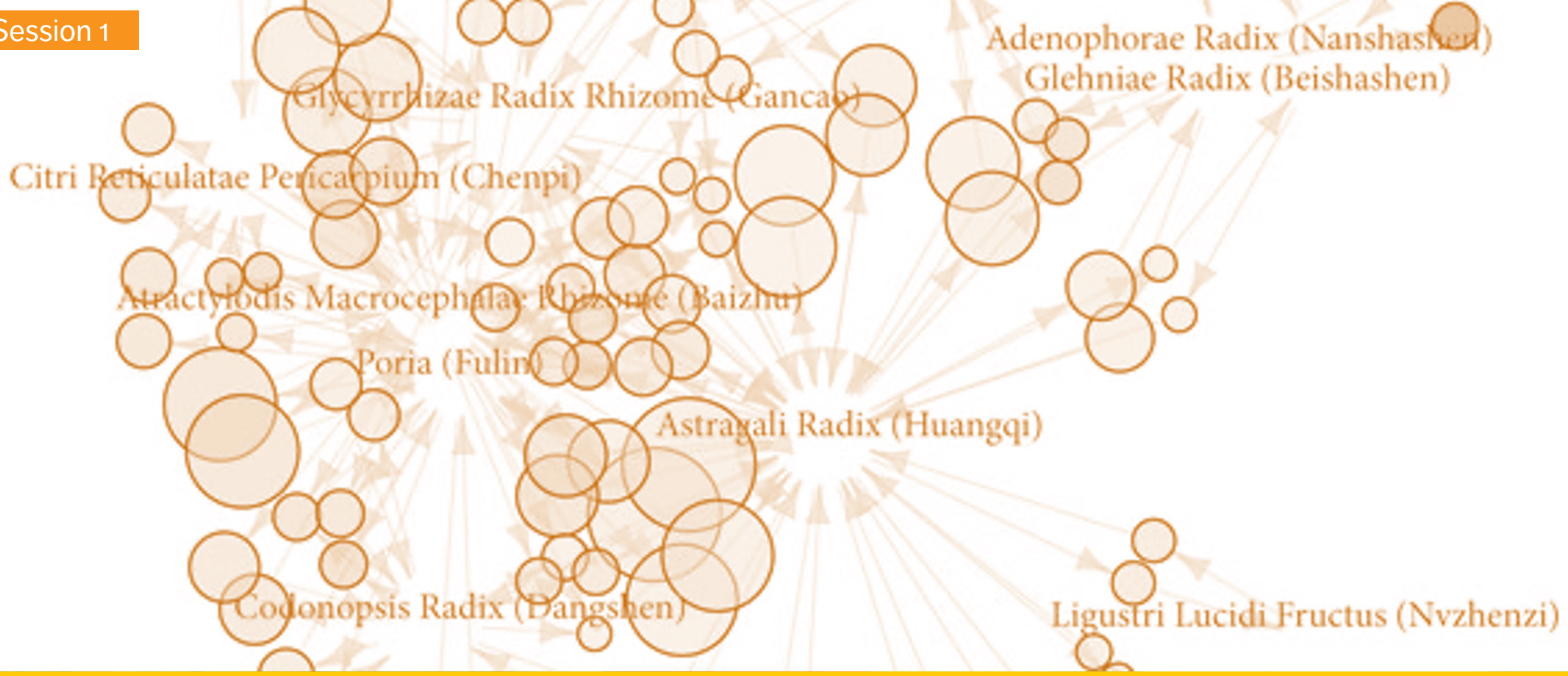
- Overview
- Case Study: Danish Medical Data

Exercises

Association Rules Overview

1. Of what types of machine learning tasks are the following problems representative?
   - Identifying risk factors associated to breast/prostate cancer.
   - Predicting whether a patient will have a second, fatal heart attack within 30 days of the first on the basis of demographics, diet, clinical measurements, etc.
   - Establishing the relationship between salary and demographic information in population survey data.
   - Predicting the yearly inflation rate using various indicators.

2. What are some examples of supervised, unsupervised, semi-supervised, reinforcement machine learning tasks in the business world? In a public policy/government setting? In a scientific setting?



3. Association Rules Concepts

Correlation and Causation

Association rules can automate **hypothesis discovery**, but one must remain correlation-savvy (less prevalent than one might hope...).

If attributes A and B are correlated in a dataset, there are various possibilities:

- A and B are correlated entirely by chance in this particular dataset
- A is a re-labeling of B (or *vice-versa*)
- A causes B (or *vice-versa*)
- some other attributes $C_1, \ldots, C_n$ (which may not be available in the data) cause A and B
- etc.

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

Correlation and Causation

Insight	Organization
Pop-Tarts sales shoot up 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

Definitions

premise conclusion

A **rule** $X \rightarrow Y$ is a statement of the form "if X then Y" built from any logical combinations of a dataset attributes.

A rule **does not need to be true for all observations** in the dataset – there could be instances where the premise is satisfied but the conclusion is not.

Some of the "best" rules are those which are only accurate 10% of the time, as opposed to rules which are only accurate 5% of the time, say.

It depends on the context.

Definitions

To determine a rule's strength, we compute various **rule metrics**, such as the:

- **support** (the frequency at which a rule occurs in a dataset) – low coverage values indicate rules that rarely occur
- **confidence** (the reliability of the rule: how often does the conclusion occur in the data given that the premises have occurred) –high confidence rules are "truer"
- **interest** (the difference between its confidence and the relative frequency of its conclusion) – rules with high absolute interest are more "interesting"
- **lift** (the increase in the frequency of the conclusion which can be explained by the premises) – with a high lift (> 1), the conclusion occurs more frequently than expected
- also **conviction**, **all-confidence**, **leverage**, **collective strength**, etc.

[**Note:** $\text{Freq}(A) \in \{0,1,\ldots,N\} =$ # of observations for which A holds]

Definitions

If N is the number of observations in a dataset, then:

$$\text{Support}(X \to Y) = \frac{\text{Freq}(X \cap Y)}{N} \in [0,1]$$

Proportion of instances where the premise and the conclusion occur together

$$\text{Confidence}(X \to Y) = \frac{\text{Freq}(X \cap Y)}{\text{Freq}(X)} \in [0,1]$$

Proportion of instances where the conclusion occurs when the premise occurs

$$\text{Interest}(X \to Y) = \text{Confidence}(X \to Y) - \frac{\text{Freq}(Y)}{N} \in [-1,1]$$

$$\text{Lift}(X \to Y) = \frac{N^2 \cdot \text{Support}(X \to Y)}{\text{Freq}(X) \cdot \text{Freq}(Y)} \in (0, N^2)$$

... ?!?

$$\text{Conviction}(X \to Y) = \frac{1 - \text{Freq(Y)}/N}{1 - \text{Confidence}(X \to Y)} \geq 0$$

Interpretation of the Lift: 70% of those born before 1976 own a copy, whereas 56% of those born after 1976 own a copy.

$$1.2 \approx \frac{0.70}{0.56}$$

Example

RM: if an individual is born before 1976 (X), then they own a copy of the Beatles' *Sergeant Peppers' Lonely Hearts Club Band*, in some format (Y).

Assume that :

- $N = 15{,}356$
- $\text{Freq}(X) = 3888$
- $\text{Freq}(\text{Y}) = 9092$
- $\text{Freq}(X \cap Y) = 2720$

$$\text{Support}(\text{RM}) = \frac{2720}{15,536} \approx 18\%$$

$$\text{Confidence}(\text{RM}) = \frac{2720}{3888} \approx 70\%$$

$$\text{Interest}(\text{RM}) = \frac{2720}{3888} - \frac{9092}{15,356} \approx 0.11$$

$$\text{Lift}(\text{RM}) = \frac{15,356^2 \cdot 0.18}{3888 \cdot 9092} \approx 1.2$$

$$\text{Conviction}(\text{RM}) = \frac{1 - 9092/15,356}{1 - 2720/3888} \approx 1.36$$

Interpreting Association Rules

All this seems to point to the rule RM being not entirely devoid of meaning, but to what extent, exactly? **This is a difficult question to answer**.[170]

It is difficult to provide thresholds, but evaluation of a lone rule is **meaningless**.

It is recommended to conduct a **preliminary exploration** of the space of association rules (using domain expertise) in order to determine reasonable threshold ranges for the specific situation; candidate rules would then be discarded or retained depending on these metric thresholds.

This requires the ability to “easily” generate potential candidate rules.

Generating Association Rules

The real challenge of association rules discovery is to **generate** candidate rules without wasting time generating rules which are likely to be discarded.

An **itemset** for a dataset is a list of attributes with values. A set of **rules** can be created from the itemset by adding "**IF** ... **THEN**" blocks to the instances.

From $\{\text{membership} = \text{True}, \text{age} = \text{Youth}, \text{purchasing} = \text{Typical}\}$, we can get:

- **IF** (purchasing = Typical AND membership = True) **THEN** age = Youth
- **IF** age = Youth **THEN** membership = True
- etc.
- n **items** $\Rightarrow$ $2^n - 1$ **rules** (combinatorial explosion)

Brute Force Algorithm

1. Generate item sets (of size 1, 2, 3, 4, etc.).
2. Create rules from each item set.
3. Calculate the support, confidence, interest, lift, conviction, etc., for each rule.
4. Retain only the rules with “high enough” coverage, accuracy, interest, lift, conviction, or other appropriate metrics.
5. These rules are considered to be **true** for the dataset – they are **new knowledge derived from the data**.

A Priori Algorithm

The combinatorial explosion is a problem – it disqualifies the **brute force** approach for any dataset with a realistic number of attributes.

How can we generate a small number of **promising** candidate rules?

The ***a priori*** algorithm is an early attempt to overcome that difficulty.

Initially, it was developed to work for **transaction data** (i.e. goods as columns, customer purchases as rows); every reasonable dataset can be transformed into a transaction dataset using dummy variables.

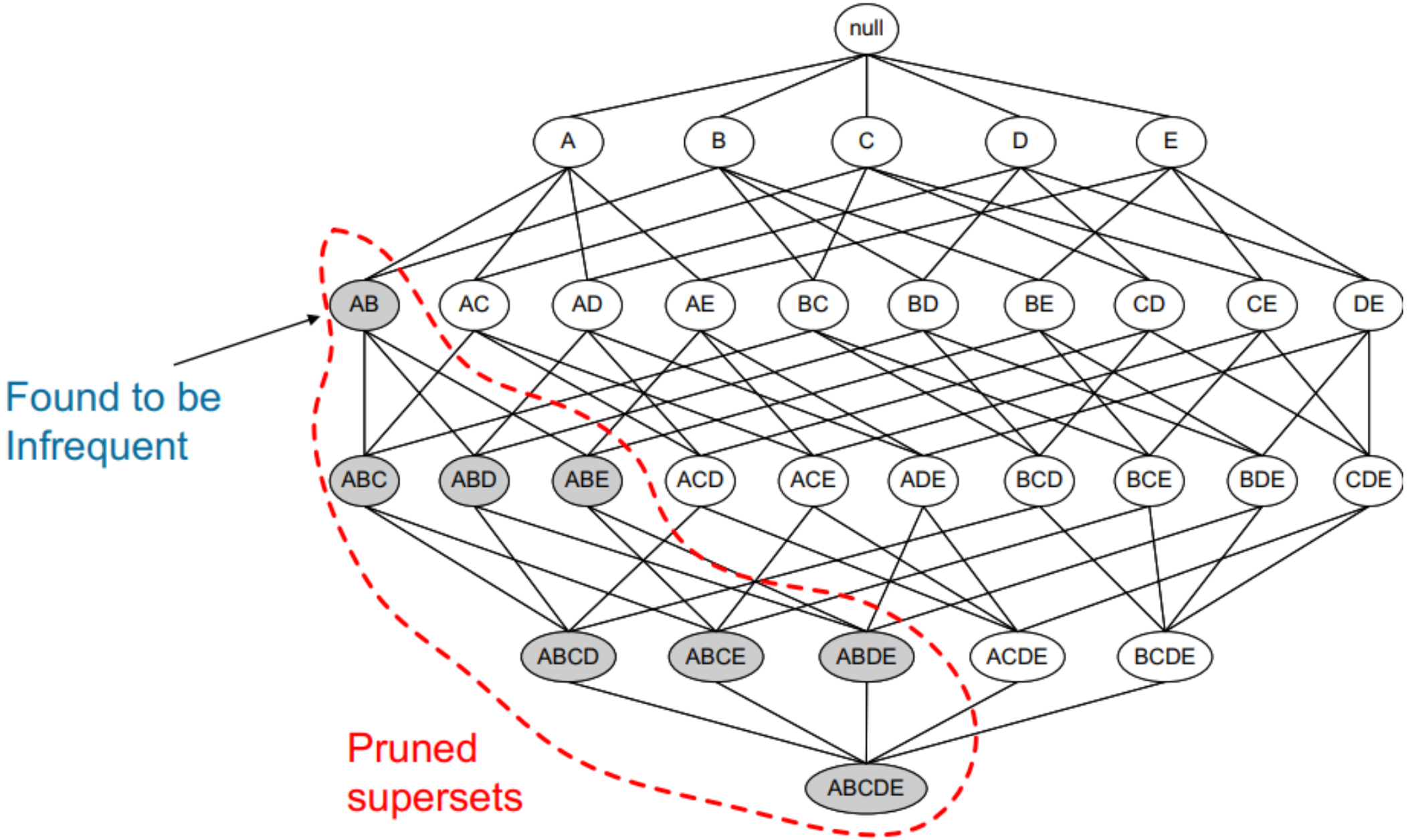
A Priori Algorithm

The a priori algorithm attempts to find **frequent itemsets** from which to build candidate rules, instead of building rules from **all** possible itemsets.

It starts by identifying frequent **individual items** in the database and extends those that are retained into larger and larger **item supersets**, who are themselves retained only if they occur **frequently enough** in the data.

The main idea is that “all non-empty subsets of a frequent itemset must also be frequent”, or equivalently, that all supersets of an infrequent itemset must also be infrequent.

[T. Chou, Apriori: Association Rule Mining In-Depth Explanation and Python Implementation]



A Priori Algorithm

The algorithm terminates when no further itemsets extensions are retained, which always occurs given the finite number of levels in categorical datasets:

- **strengths:** easy to implement and to parallelize
- **limitations:** slow, requires frequent scans, not ideal for infrequent and rare itemsets

More efficient algorithms have since displaced it in practice:

- **max-miner** tries to identify frequent itemsets without enumerating them – it performs jumps in itemset space instead of using a bottom-up approach
- **eclat** is faster and uses depth-first search, but requires extensive memory storage

Validation

How **reliable** are association rules?

What is the likelihood that they occur entirely **by chance**?

How **relevant** are they?

Can they be generalized **outside** the dataset, or to **new** data streaming in?

Statistically sound association discovery can help reduce the risk of finding spurious associations to a user-specified significance level.

Validation

We end this section with a few comments:

- frequent rules correspond to instances that occur repeatedly in the dataset, algorithms that generate itemsets often try to **maximize coverage**; when **rare events** are more meaningful we need algorithms that can generate rare itemsets – **this is not a trivial problem**;
- continuous data has to be binned into **categorical** data to generate rules; as there are many ways to accomplish that task, the same dataset can give rise to completely different rules – this could create some **credibility issues** with clients and stakeholders;
- other algorithms: AIS, SETM, aprioriTid, aprioriHybrid, PCY, Multistage, Multihash, etc.

Suggested Reading

Association Rules Concepts

Data Understanding, Data Analysis, Data Science
Machine Learning 101

Association Rules Mining

- Generating Rules
- The A Priori Algorithm
- Validation
- Toy Example: Titanic Dataset

R Examples

- Association Rules Mining: Titanic Dataset

Exercises

Association Rules Concepts

1. Evaluate the following candidate rules in the music dataset:
   - if an individual owns a classical music album (W), they also own a hip-hop album (Z), given that $\text{Freq}(W) = 2010$, $\text{Freq}(Z) = 6855$, $\text{Freq}(W \cap Z) = 132$.
   - if an individual owns both a Beatles and a classical music album, then they were born before 1976, given that $\text{Freq}(Y \cap W) = 1852$, $\text{Freq}(Y \cap W \cap X) = 1778$.

2. Out of the 3 rules that have been established ($X \rightarrow Y$, $W \rightarrow Z$, $Y \,\&\, W \rightarrow X$), which do you think is more useful? Which is more surprising?

Exercises

Association Rules Concepts

3. A store that sells accessories for cellular phones runs a promotion on faceplates. Customers who purchase multiple faceplates from a choice of 6 different colours get a discount. Managers, who would like to know what colours will be purchased together, collected purchases in Transactions.csv.

   Consider the following rules:

   - {**r**ed, **w**hite} ⇒ {**g**reen}
   - {**g**reen} ⇒ {**w**hite}
   - {**r**ed, **g**reen} ⇒ {**w**hite}
   - {**g**reen} ⇒ {**r**ed}
   - {**o**range} ⇒ {**r**ed}
   - {**w**hite, **b**lack} ⇒ {**y**ellow}
   - {**b**lack} ⇒ {**g**reen}

Exercises

Association Rules Concepts

3. (cont.) For each rule, compute the **support**, **confidence**, **interest**, **lift**, and **conviction**. Amongst the rules for which the support is positive (> 0), which one has the highest lift? Confidence? Interest? Conviction? Build an additional 5-10 candidate rules, and evaluate them. Which of the 12-17 candidate rules do you think would be most useful for the store managers? How would one determine reasonable threshold values for the support, coverage, interest, lift, and conviction of rules derived from a given dataset?

Exercises

Association Rules Concepts

4. Go over the titanic association rules example found in DUDADS (see suggested reading). Repeat the process with the UniversalBank.csv dataset (you may need to clean and visualize the dataset first, as well as categorize the numerical variables; can you come up with a reasonable guess as to what each of the variables represent?). Find “true knowledge” about the dataset in the form of reliable and meaningful association rules (use metrics as appropriate).

Session 2

INTRODUCTION TO MACHINE LEARNING

Classification

INTRODUCTION TO MACHINE LEARNING

The diversity of problems that can be addressed by classification algorithms is significant, and covers many domains. It is difficult to comprehensively discuss all the methods in a single book.

[C.C. Aggarwal]

4. Classification Overview

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, or classes (supervised learning; predictive analytics).

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

Value estimation (regression) is similar to classification when the target variable is **numerical**.

Overview

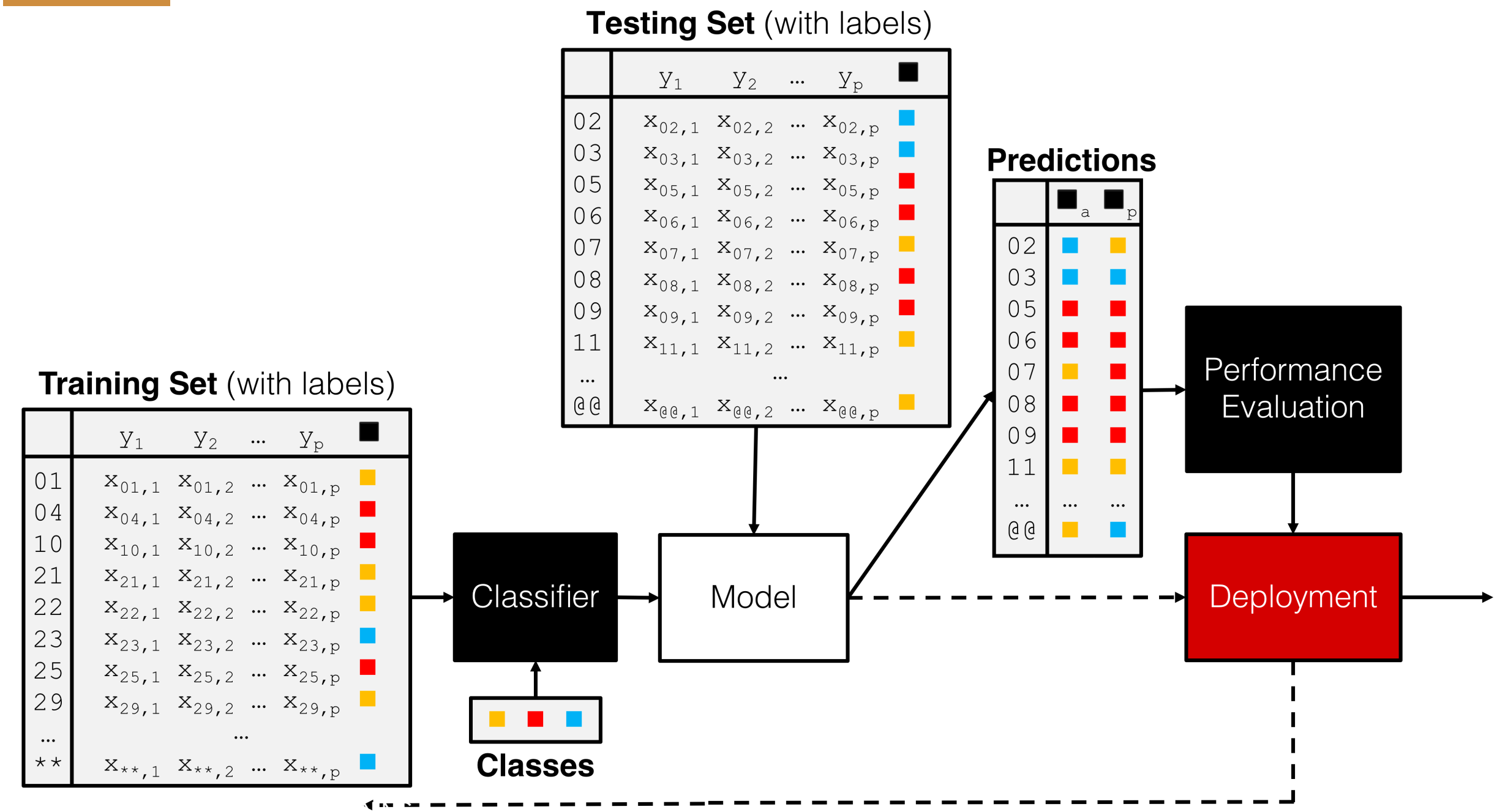
In the **testing** phase, the model is used to assign a class to observations for which the label is hidden, but ultimately known (the **testing** set).

The performance of a classification model is evaluated on the testing set, **never** on the training set. In the **absence** of testing data, classification may be **descriptive** but not predictive.

Technical issues include:

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

Session 2


Testing Set (with labels)
Predictions
Training Set (with labels)
Performance Evaluation
Classifier
Model
Deployment
Classes

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

Applications

Other uses include:

- 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"
- Predicting the inflation rate for the coming two years based on a number of economic indicators.

Examples

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?

Examples

Scenario:

Customers who make a large number of calls to a mobile phone company's customer service number have been identified as churn risks. The company is interested in reducing said churn.

Questions: can we predict

- the overall lifetime value of a customer?
- which customers are more likely to churn in the near future?
- what retention offer a particular customer will best respond to?

Case Study

Minnesota Tax Audits

Hsu *et al.*
Data Mining Based Tax Audit Selection: A Case Study of a Pilot Project at the Minnesota Department of Revenue
Real World Data Mining Applications, 2015

Objective

The U.S. Internal Revenue Service (IRS) estimated that there were large gaps between **revenue owed** and **revenue collected** for 2001 and for 2006.

Using DoR data, the authors sought to increase **efficiency** in the audit selection process and to **reduce the gap** between revenue owed and revenue collected.

Case Study

Minnesota Tax Audits

Hsu *et al.*
Data Mining Based Tax Audit Selection: A Case Study of a Pilot Project at the Minnesota Department of Revenue
Real World Data Mining Applications, 2015

Methodology

1. **data selection and separation:** experts selected several hundred cases to audit and divided them into training, testing and validating sets
2. **classification modeling** using MultiBoosting, Naïve Bayes, C4.5 decision trees, multilayer perceptrons, support vector machines, etc.
3. **evaluation of all models** on the testing set – models performed poorly until the size of the business being audited was recognized to have an effect, leading to two separate tasks (large/small businesses).
4. **model selection/validation** compared the estimated accuracy between different classification model predictions and the actual field audits (MultiBoosting with Naïve Bayes was selected as the final model; suggesting improvements to increase audit efficiency).

Case Study

Minnesota Tax Audits

Hsu *et al.*
Data Mining Based Tax Audit Selection: A Case Study of a Pilot Project at the Minnesota Department of Revenue
Real World Data Mining Applications, 2015

Data

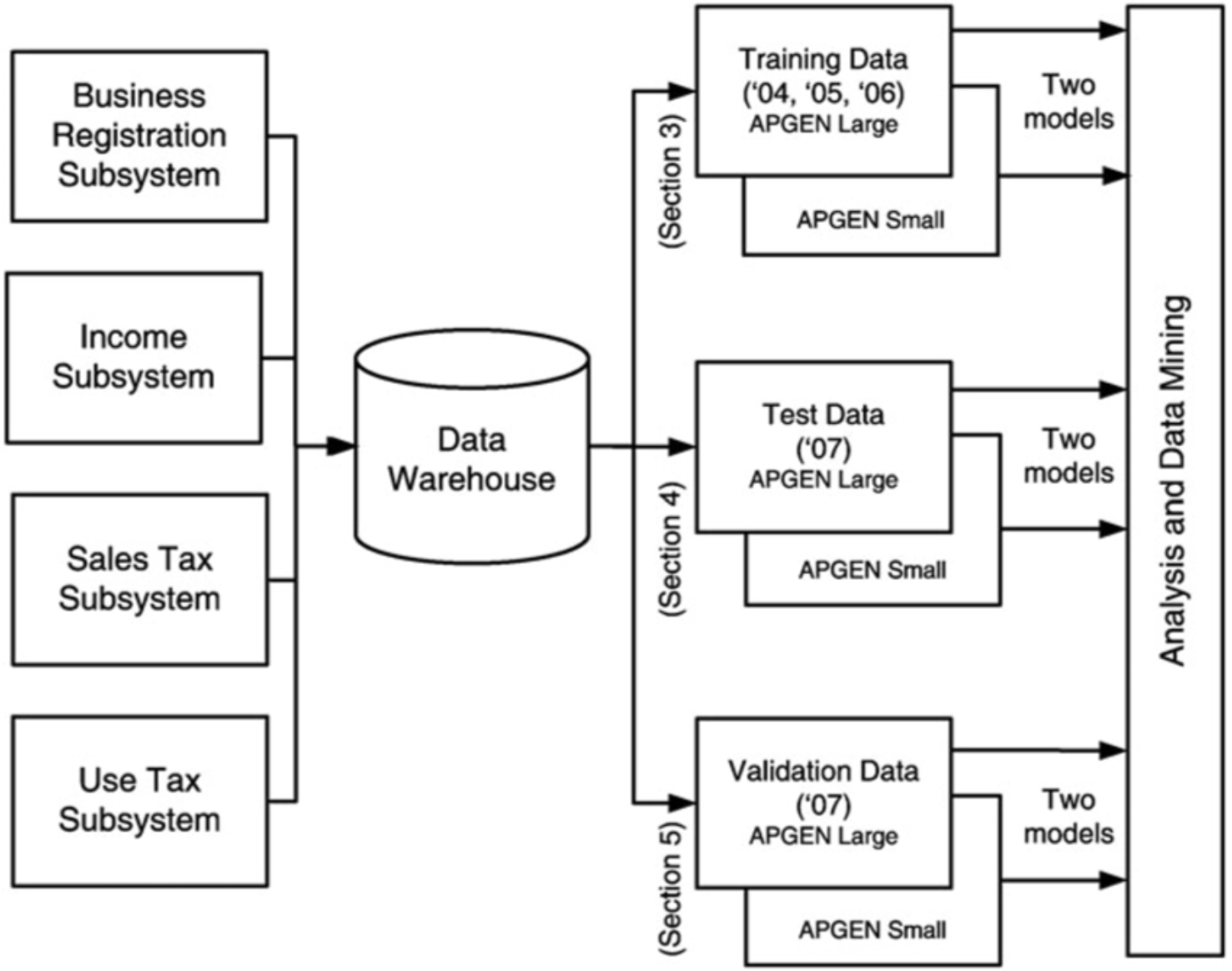
Selected tax audit cases from 2004 to 2007, collected by the audit experts, which were split into training, testing and validation sets:

- the **training data** set consisted of *Audit Plan General* (APGEN) *Use Tax* audits and their results for the years 2004-2006
- the **testing data** consisted of APGEN Use Tax audits conducted in 2007 and was used to test or evaluate models (for Large and Smaller businesses) built on the training dataset
- while **validation** was assessed by actually conducting field audits on predictions made by models built on 2007 Use Tax return data processed in 2008.

Case Study

Minnesota Tax Audits

Hsu *et al.*
Data Mining Based Tax Audit Selection: A Case Study of a Pilot Project at the Minnesota Department of Revenue
Real World Data Mining Applications, 2015



Case Study

Minnesota Tax Audits

Hsu *et al.*
Data Mining Based Tax Audit Selection: A Case Study of a Pilot Project at the Minnesota Department of Revenue
Real World Data Mining Applications, 2015

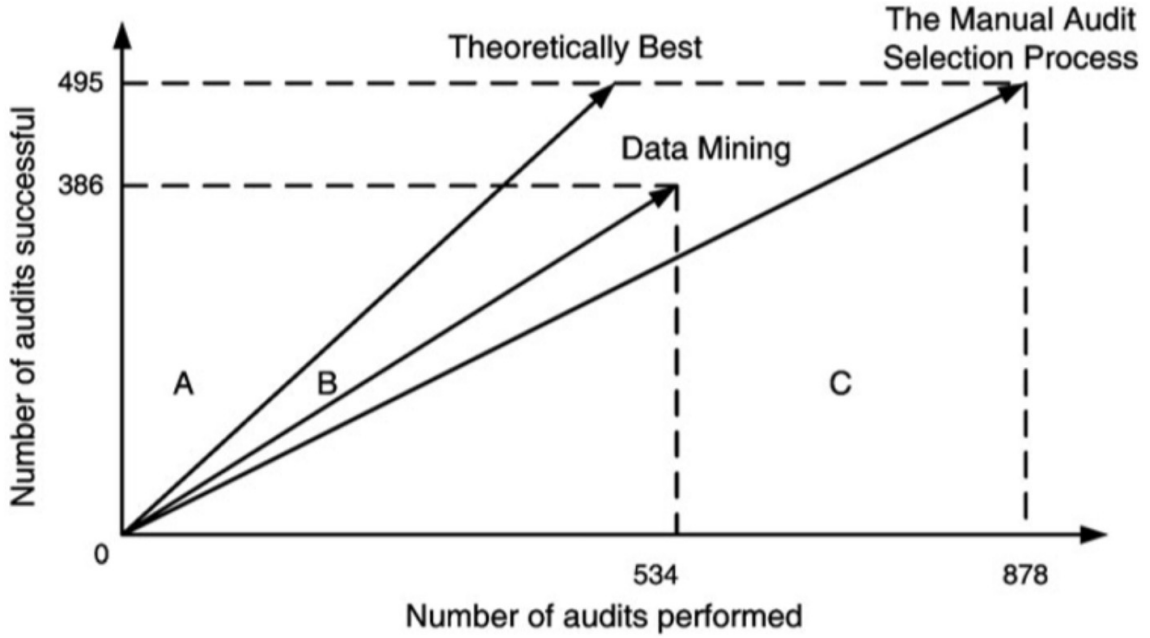
Strengths and Limitations of the Algorithms

- Naïve Bayes classification assumes independence of the features, which rarely occurs in real-world situations. This approach is also known to potentially introduce bias to classification schemes. In spite of this, classification models built using it have a successfully track record.
- MultiBoosting is an **ensemble technique** that uses committee (i.e. groups of classification models) and "group wisdom" to make predictions; unlike other ensemble techniques, it is different from other ensemble techniques in the sense that it forms a committee of sub-committees, which has a tendency to reduce both bias and variance of predictions.

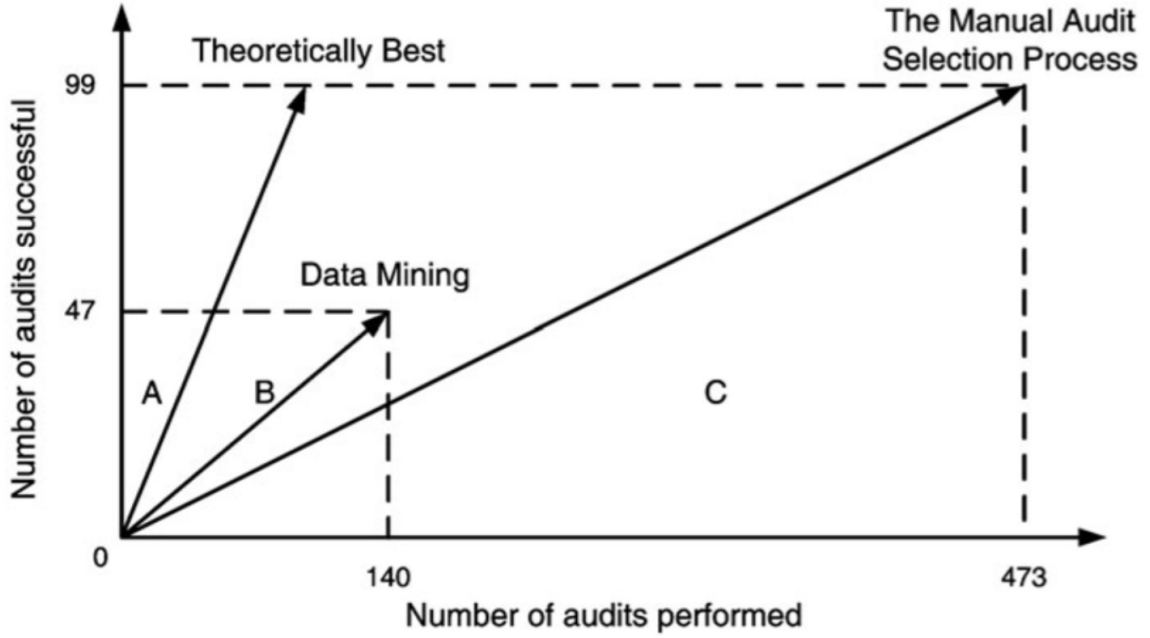
Case Study

Minnesota Tax Audits

APGEN Large



APGEN Small



Hsu *et al.*
Data Mining Based Tax Audit Selection: A Case Study of a Pilot Project at the Minnesota Department of Revenue
Real World Data Mining Applications, 2015

Case Study

Minnesota Tax Audits

Hsu *et al.*
Data Mining Based Tax Audit Selection: A Case Study of a Pilot Project at the Minnesota Department of Revenue
Real World Data Mining Applications, 2015

Take-Aways

- Many models were churned out before the team made a final selection.
- Past performance of a model family in a previous project can guide the selection, but remember the *No Free Lunch (NFL) Theorem*: nothing works best all the time!
- The feature selection process could very well require a number of visits to domain experts before the feature set yields promising results.
- Data analysis teams should seek out individuals with a good understand of both data and context.
- Domain-specific knowledge has to be integrated in the model in order to beat random classifiers, on average.
- Even slight improvements over the current approach can find a useful place in an organization – data science is not solely about Big Data and disruption!

General Classification Comments

Classification is linked to **probability estimation**

- approaches based on regression models could prove fruitful

Rare occurrences (often more interesting or important):

- historical data at Fukushima's nuclear reactor prior to the meltdown could not have been used to learn about meltdowns, for instance
- predicting no meltdown will yield correct predictions roughly 99.99% of the time, but will miss the point of the exercise

No Free-Lunch Theorem: no classifier works best for all data.

With big datasets, algorithms must also consider efficiency.

Suggested Reading

Classification Overview

Data Understanding, Data Analysis, Data Science
Machine Learning 101

Classification and Value Estimation

- Overview
- Case Study: Minnesota Tax Audit

Spotlight on Classification

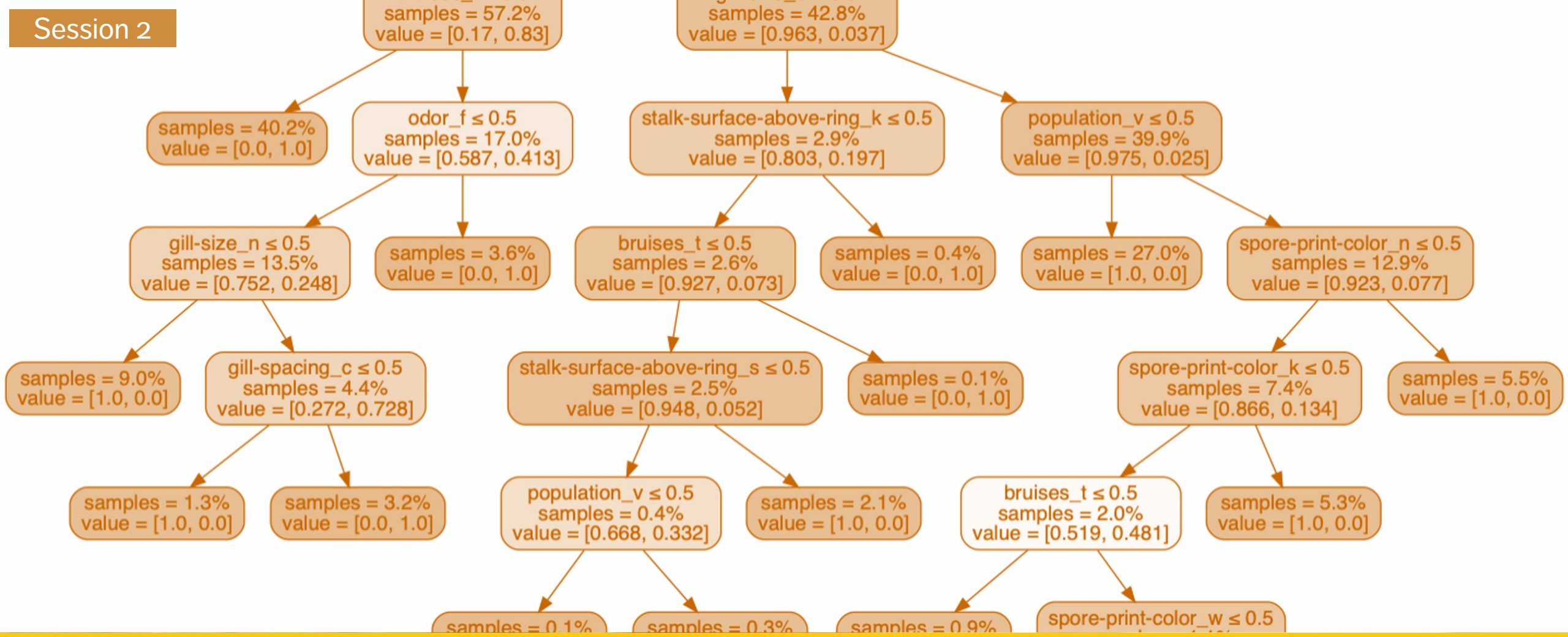
*Overview (advanced)

- Formalism

Exercises

Classification Overview

1. How would you use standard statistical modeling techniques to answer the questions presented in the two scenarios in the slides?
2. Identify scenarios and questions that could use classification and/or value estimation in your every day work activities.



5. Decision Trees and Other Algorithms

Classification Algorithms

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)

Bayesian Methods

Decision Trees

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

Classification Algorithms

Naïve Bayes Classifiers

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

Support Vector Machines

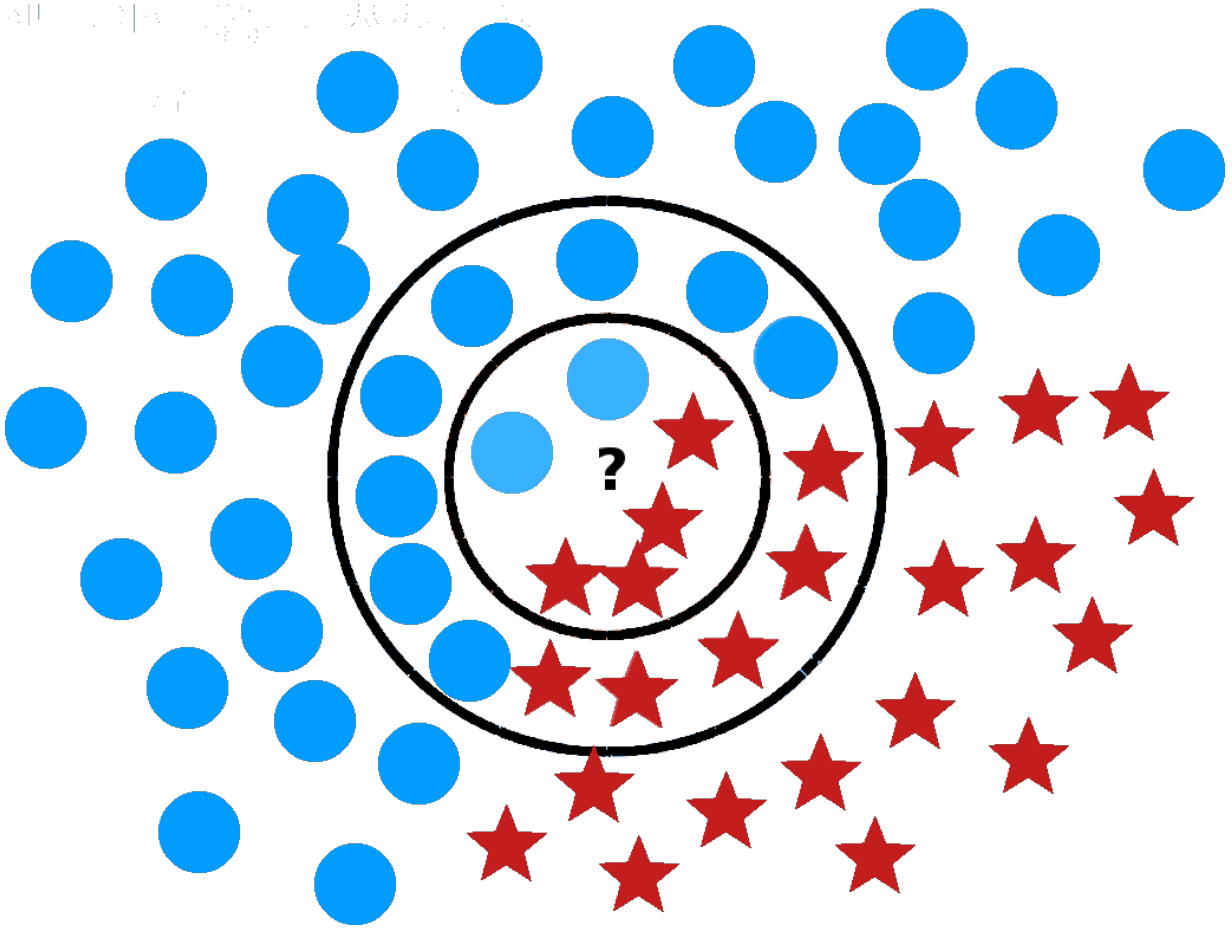
- may be difficult to interpret (non-linear boundaries)
- can help mitigate big data difficulties

Boosting Methods

Nearest Neighbours Classifiers

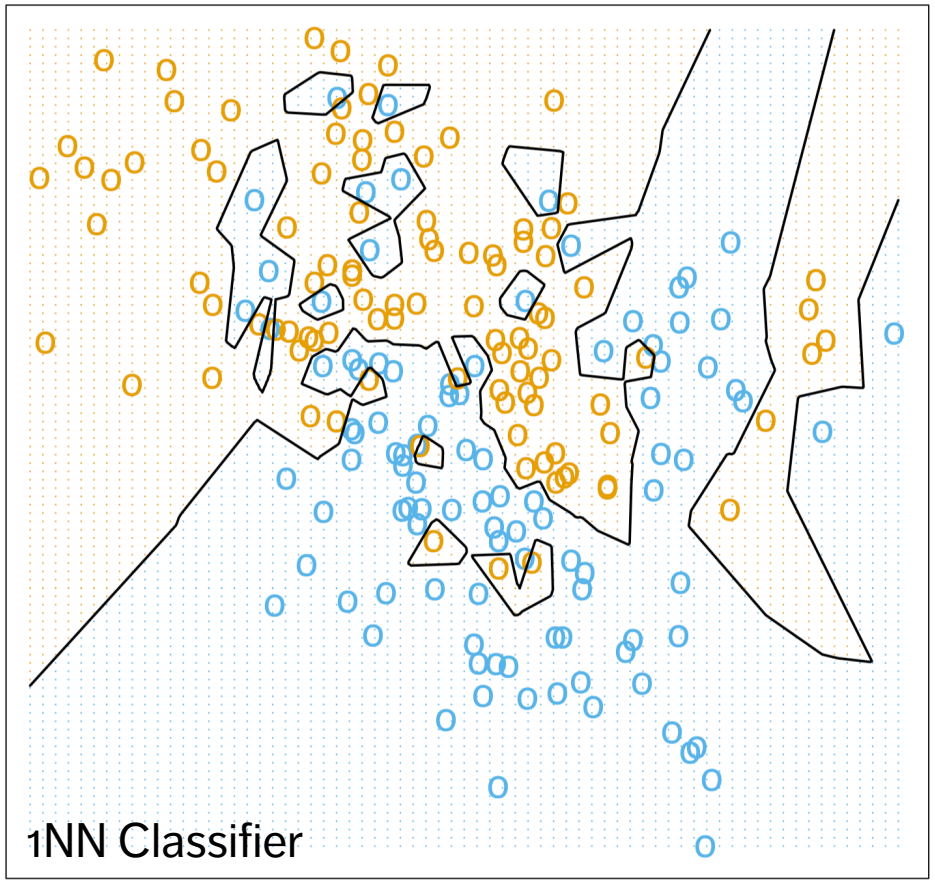
- require very little assumptions about the data
- not very stable (adding points may substantially modify the boundary)

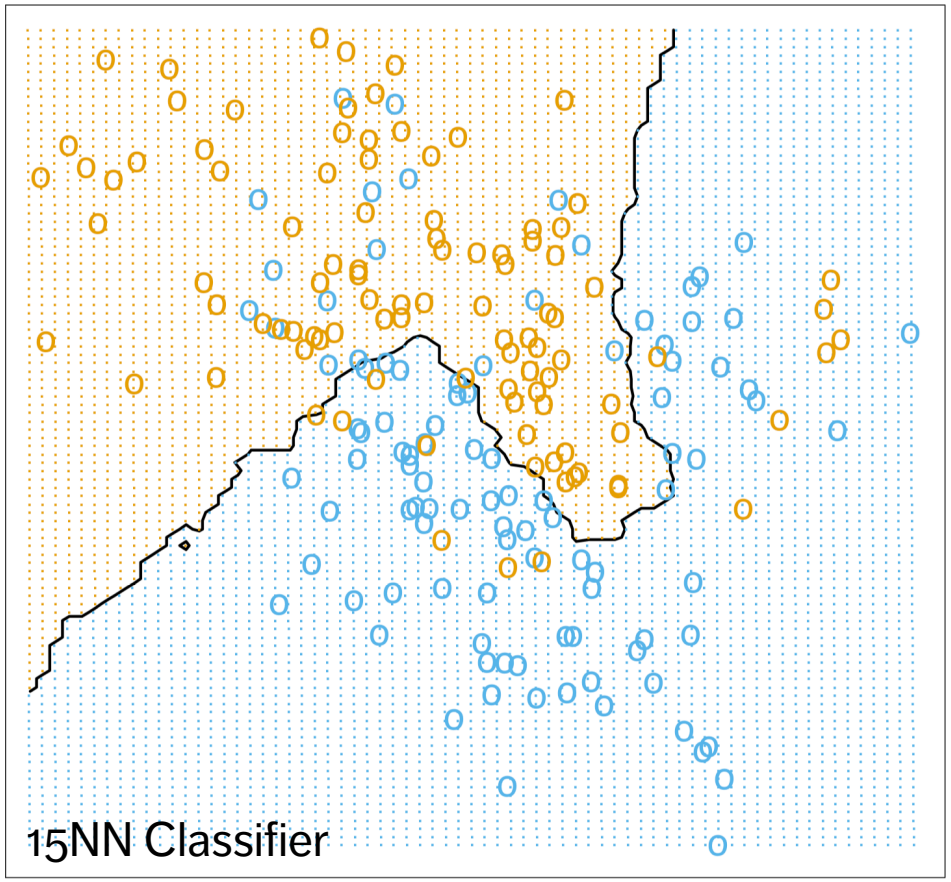
k –Nearest Neighbours Classifier



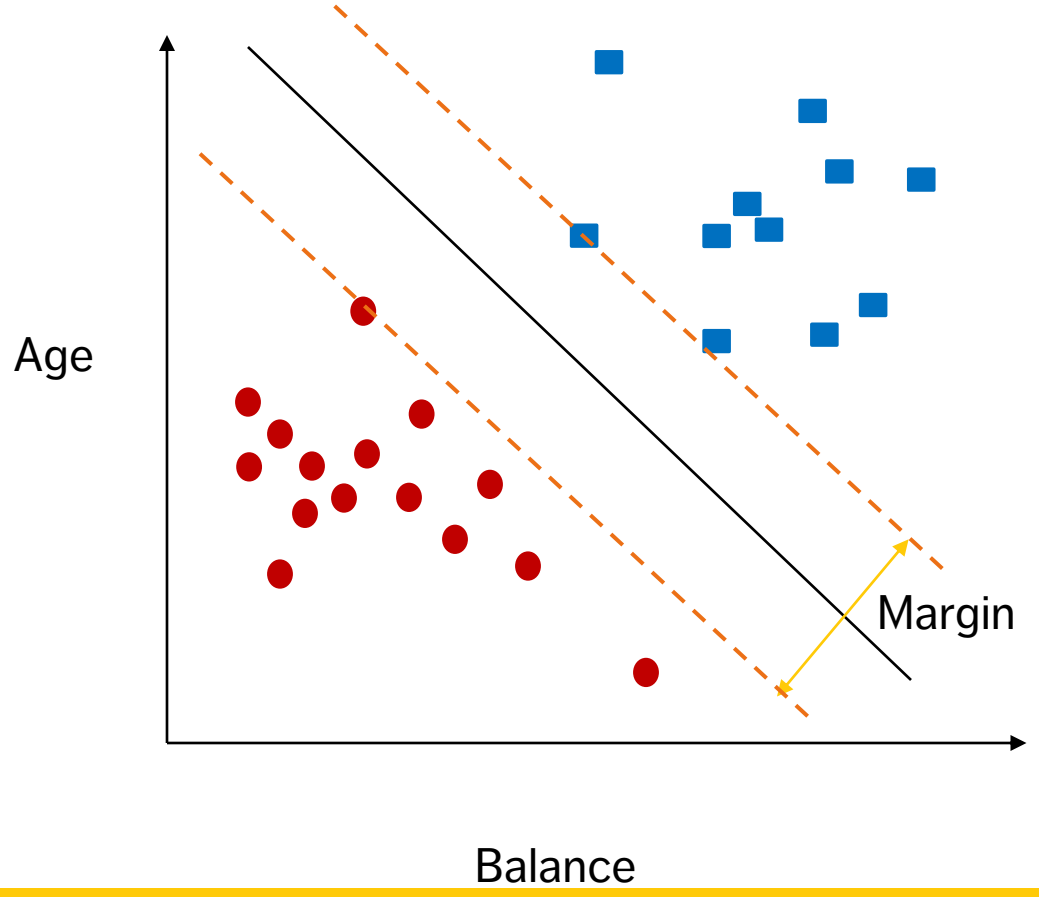
[Tibshirani, Hastie & Friedman, *The Elements of Statistical Learning*]

k –Nearest Neighbours Classifier



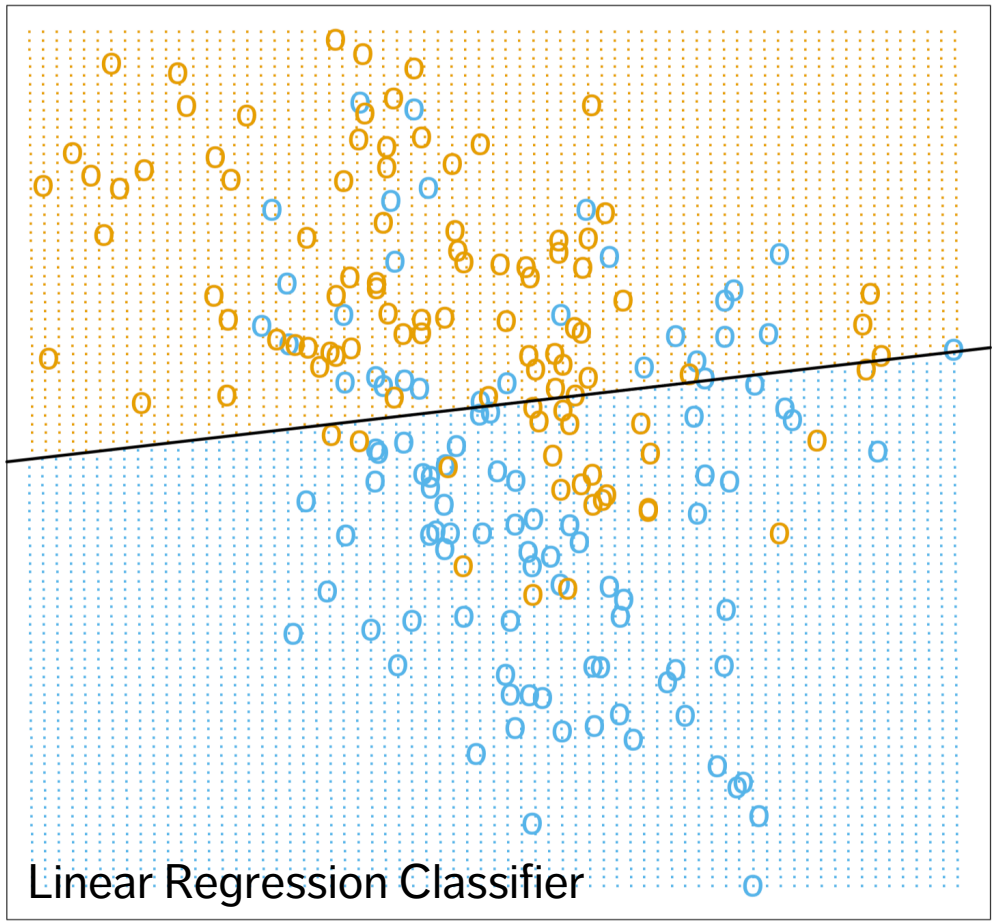


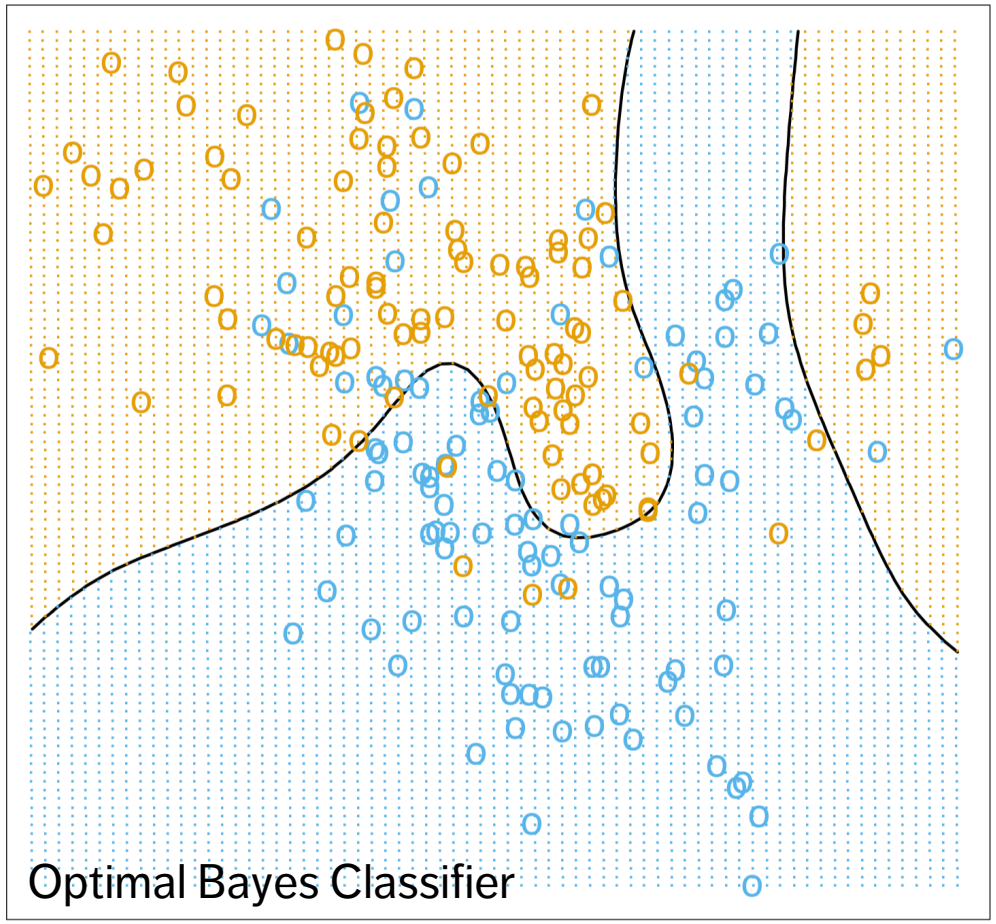
Support Vector Machines



[Tibshirani, Hastie & Friedman, *The Elements of Statistical Learning*]

Other Classifiers

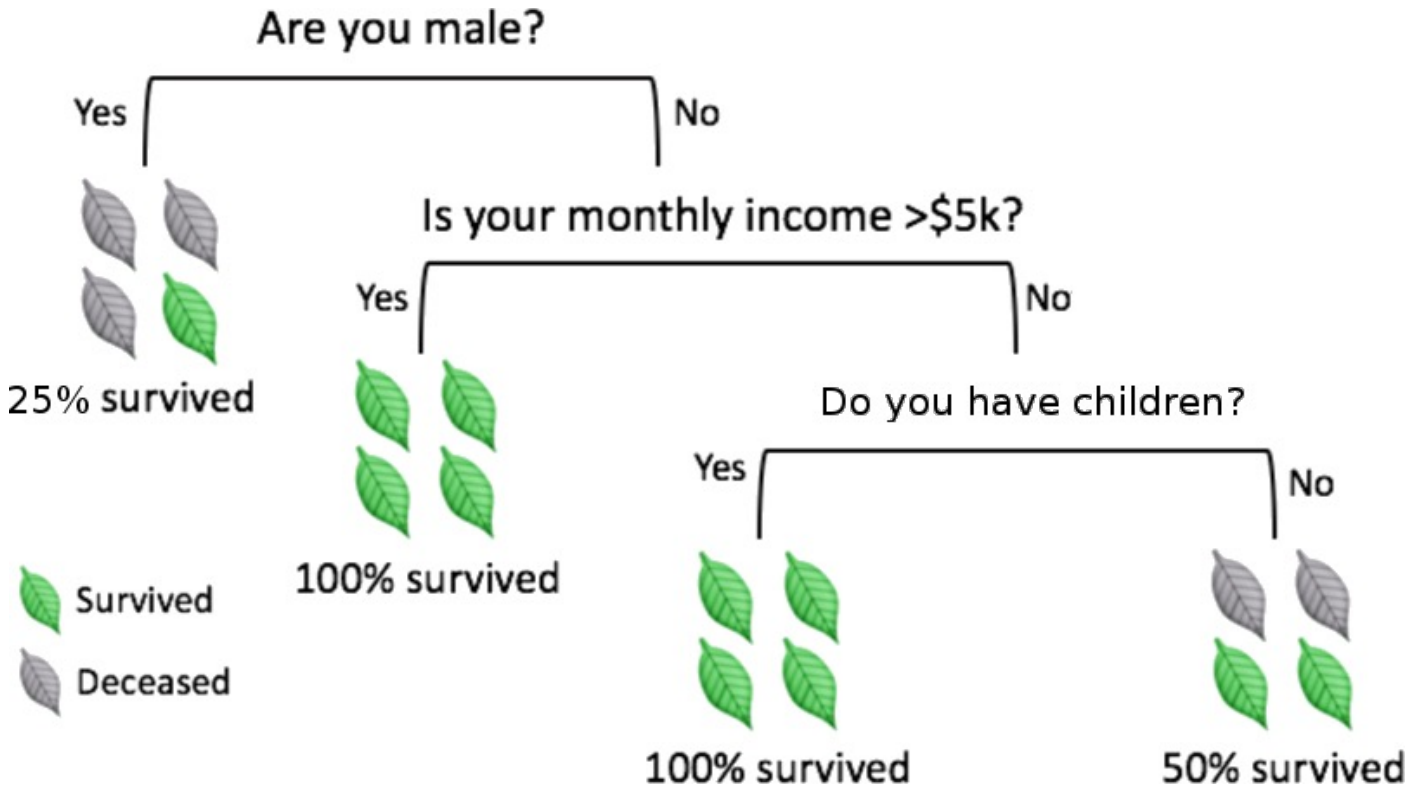




[adapted from A. Ng, K. Soo, *Numsense!*, Surviving a Disaster]

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**.



Decision Trees

To make a **prediction** for a new instance, follow the path down the tree, and read the prediction directly once a leaf is reached.

Creating the tree and traversing it might be **time-consuming** if there are too many variables.

Prediction accuracy can be a concern in trees whose growth is **unchecked**. In practice, the criterion of **purity** at the leaf-level is linked to bad prediction rates for new instances.

- other criteria are often used to prune trees, which may lead to **impure** leaves (i.e. with non-trivial entropy).

Decision Tree Algorithm (ID3)

Task: grow a decision tree using a training set (a subset of the data for which the correct classification of the target is known).

Overview:

1. Split the training data (**parent**) set into (**children**) subsets, using the different levels of a particular attribute
2. Compute the **information gain** for each subset
3. Select the **most advantageous** split
4. Repeat for each node until some **leaf** criterion is met (each item in the leaf has the same classification is one possibility)

Information Gain

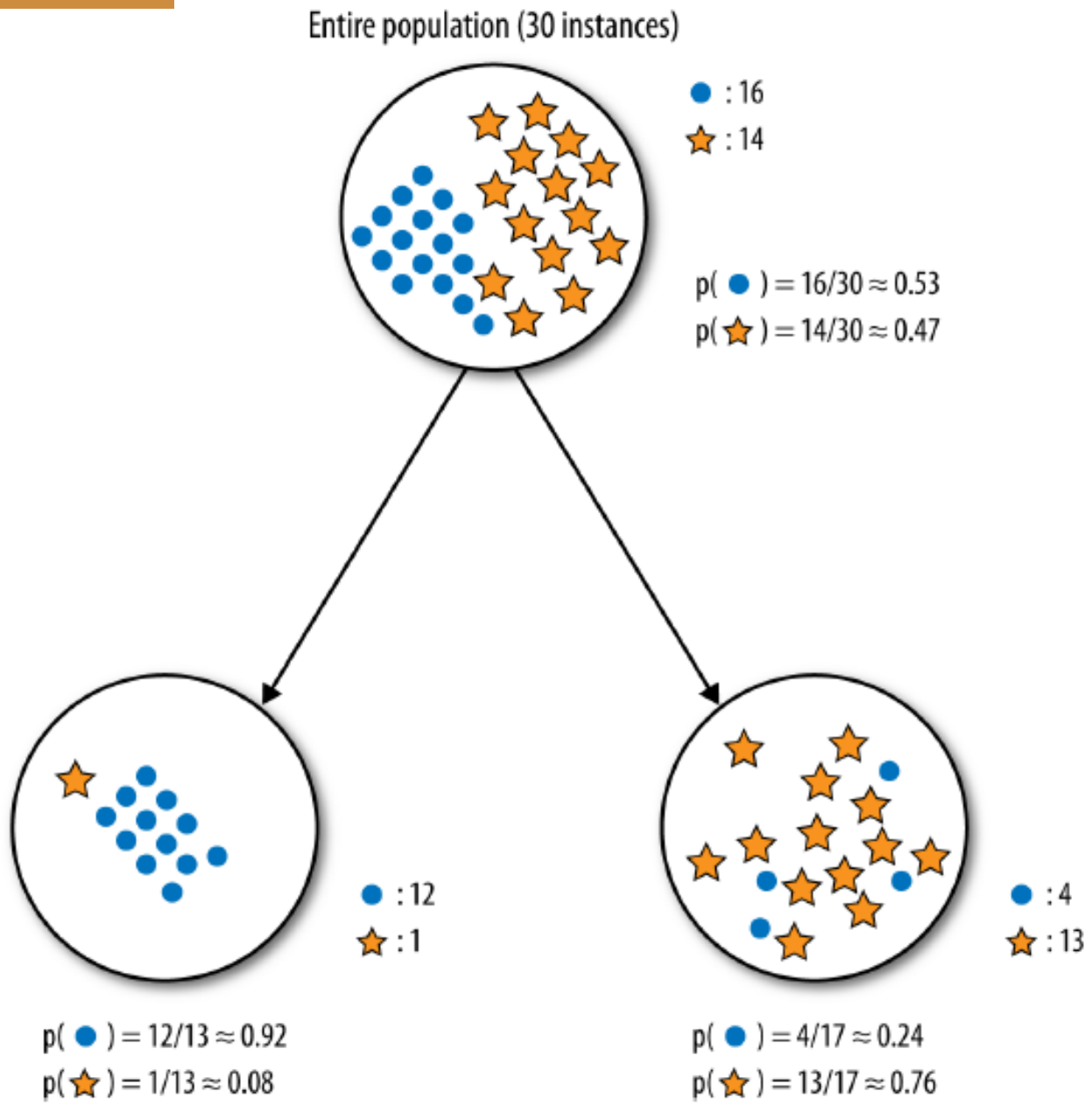
Entropy is a measure of disorder in a set S. Let p_i be the % of observations in S belonging to category i, for $i = 1, \dots, n$. The entropy of S is given by

$$E(S) = -p_1 \log p_1 - p_2 \log p_2 - \cdots - p_n \log n.$$

If the **parent set** S consisting of m records is split into k **children sets** $C_1, \dots, C_k$ containing $q_1, \dots, q_k$ records (resp.), then the **information gain** from the split is given by

$$\mathrm{IG}(S; C) = E(S) - \frac{1}{m}[q_1 E(C_1) + \cdots + q_k E(C_k)].$$

[Provost & Fawcett, *Data Science for Business*]



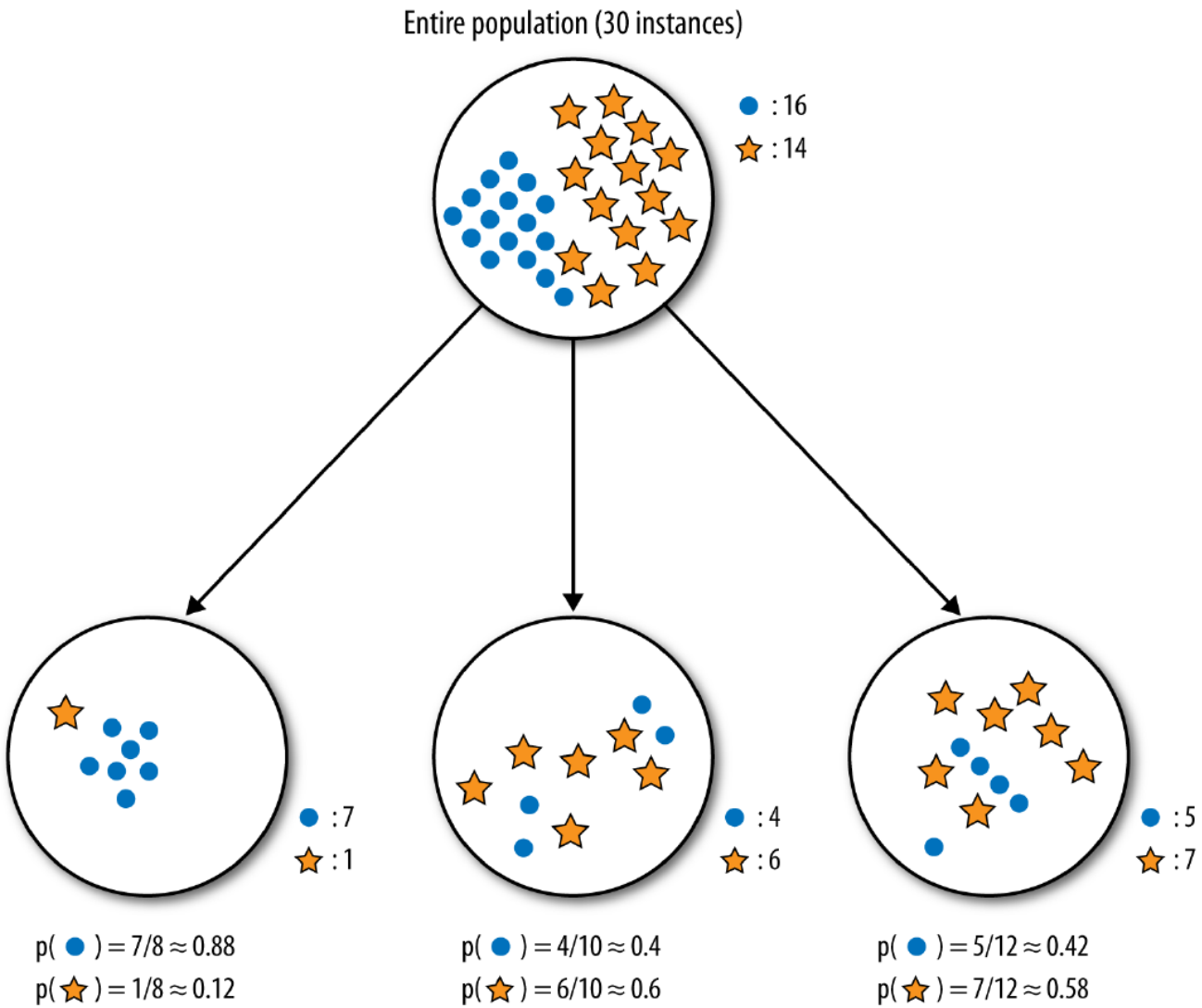
$$E(S) = -p_\circ \log p_\circ - p_* \log p_*$$
$$= -\frac{16}{30}\log\frac{16}{30} - \frac{14}{30}\log\frac{14}{30} \approx 0.99$$

$$E(L) = -p_\circ \log p_\circ - p_* \log p_*$$
$$= -\frac{12}{13}\log\frac{12}{13} - \frac{1}{13}\log\frac{1}{13} \approx 0.39$$

$$E(R) = -p_\circ \log p_\circ - p_* \log p_*$$
$$= -\frac{4}{17}\log\frac{4}{17} - \frac{13}{17}\log\frac{13}{17} \approx 0.79$$

$$\text{IG} = E(S) - \frac{1}{30}[q_L E(L) + q_R E(R)]$$
$$\approx 0.99 - \frac{1}{30}[13(0.39) + 17(0.79)]$$
$$\approx \mathbf{0.37}$$

[Provost & Fawcett, *Data Science for Business*]



$$E(S) = -p_\circ \log p_\circ - p_* \log p_*$$
$$= -\frac{16}{30}\log\frac{16}{30} - \frac{14}{30}\log\frac{14}{30} \approx 0.99$$

$$E(L) = -p_\circ \log p_\circ - p_* \log p_*$$
$$= -\frac{7}{8}\log\frac{7}{8} - \frac{1}{8}\log\frac{1}{8} \approx 0.54$$

$$E(C) = -p_\circ \log p_\circ - p_* \log p_*$$
$$= -\frac{4}{10}\log\frac{4}{10} - \frac{6}{10}\log\frac{6}{10} \approx 0.97$$

$$E(R) = -p_\circ \log p_\circ - p_* \log p_*$$
$$= -\frac{5}{12}\log\frac{5}{12} - \frac{7}{12}\log\frac{7}{12} \approx 0.98$$

$$\mathrm{IG} = E(S) - \tfrac{1}{30}[q_L E(L) + q_C E(C) + q_R E(R)]$$
$$\approx 0.99 - \tfrac{1}{30}[8(0.54) + 10(0.97) + 12(0.98)]$$
$$\approx \mathbf{0.13}$$

Decision Trees Strengths

White box model

- predictions can always be explained by following the appropriate paths

Can be used with **incomplete** datasets

Built-in feature selection

- less relevant features don't tend to be used as splitting features

Makes **no assumption** about

- independence, constant variance, underlying distributions, co-linearity

Decision Trees Limitations

Not as accurate as other algorithms (usually)

Not robust: small changes in the training dataset can lead to a completely different tree, with a completely different predictions

Particularly vulnerable to **overfitting** in the absence of **pruning**

- pruning procedures are typically convoluted

Optimal decision tree learning is **NP-complete**

Biased towards categorical features with **high** number of levels

Decision Trees Notes

Splitting metrics:

- information gain, Gini impurity, variance reduction, etc.

Common variants:

- Iterative Dichotomiser 3, C4.0, C4.5, CHAID, MARS, conditional inference trees, CART

Decision trees can also be combined together using boosting algorithms (**AdaBoost**) or **Random Forests**, providing a type of voting procedure (Ensemble Learning).

Suggested Reading

Decision Trees and Other Algorithms

Data Understanding, Data Analysis, Data Science
Machine Learning 101

Classification and Value Estimation

- Classification Algorithms
- Decision Trees
- Toy Example: Kyphosis Dataset

R Examples

- Classification: Kyphosis Dataset

Spotlight on Classification

*Simple Classification Methods (advanced)
*Rare Occurrences (advanced)
*Other Supervised Approaches (advanced)
*Ensemble Learning (advanced)

Exercises

Decision Trees and Other Algorithms

1. Go over the kyphosis classification example found in DUDADS (see suggested reading). Repeat the process with the titanic dataset (you may wish to visualize the dataset first) in order to build a decision tree that will help you determine if a passenger survived the sinking or not.

Exercises

Decision Trees and Other Algorithms

2. UniversalBank is looking at converting its **liability** customers (i.e., customers who only have deposits at the bank) into **asset** customers (i.e., customers who have a loan with the bank). In a previous campaign, *UniversalBank* was able to convert 9.6% of 5000 of its liability customers into asset customers. The marketing department would like to understand what combination of factors make a customer more likely to accept a personal loan, in order to better design the next conversion campaign.

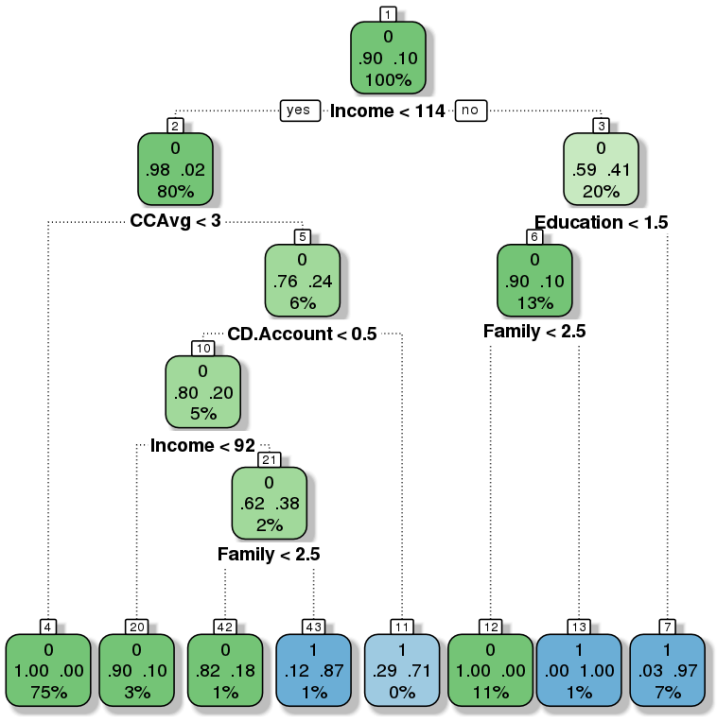
   The dataset contains data on 5000 customers, including the following measurements: age, years of professional experience, yearly income (in $K), family size, value of mortgage with the bank, whether the client has a certificate of deposit with the bank, a credit card, etc.

Exercises

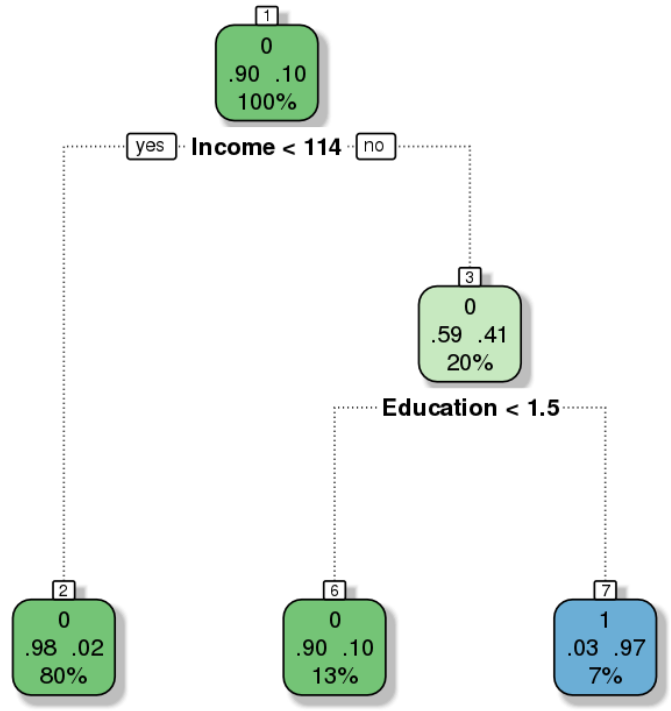
Decision Trees and Other Algorithms

2. (cont.) We build 2 decision trees on a training subset of 3000 records to predict whether a customer is likely to accept a personal loan (1) or not (0).

Tree *A*



Tree *B*



Exercises

Decision Trees and Other Algorithms

a. How many variables are used in the construction of tree A? Of tree B?

b. Is the following decision rule valid or not for tree A:
   `IF (Income` ≥ 114`) AND (Education` ≥ 1.5`)`
   `THEN (Personal Loan` $= 1$`)`?

c. Is the following decision rule valid or not for tree B:
   `IF (Income` < 92`) AND (CCAvg` ≥ 3`)`
   `AND (CD.Account` < 0.5`)`
   `THEN (Personal Loan` $= 0$`)`?

d. What prediction would tree A make for a customer with:
   - yearly income of 94,000$USD (`Income = 94`),
   - 2 kids (`Family = 4`),
   - no certificate of deposit with the bank (`CD.Account = 0`),
   - a credit card interest rate of 3.2% (`CCAvg = 3.2`), and
   - a graduate degree in Engineering (`Education = 3`).

e. What about tree B?

Actual \ Predicted	A	B	C	D	Total
A	50	10	30	20	110
B	15	20	30	15	80
C	20	10	30	40	100
D	15	15	30	50	110
Total	100	55	120	125	800

6. Performance Evaluation

Model Selection

As a consequence of the **No-Free-Lunch Theorem**, no single classifier can be the best performer for every problem.

Model selection must take into account:

- the **nature** of the available data
- the **relative frequencies of the classification sub-groups**
- the **stated classification goals**
- how easily the model lends itself to **interpretation** and **statistical analysis**
- how much **data preparation** is required

Model Selection

Model selection must take into account (continued):

- whether it can accommodate various data types and missing observations
- whether it performs well with large datasets, and
- whether it is **robust** against small data departures from theoretical assumptions.

Past success is not a guarantee of future success – it is the analyst's responsibility to try a **variety of models**.

But how can the "**best**" model be selected?

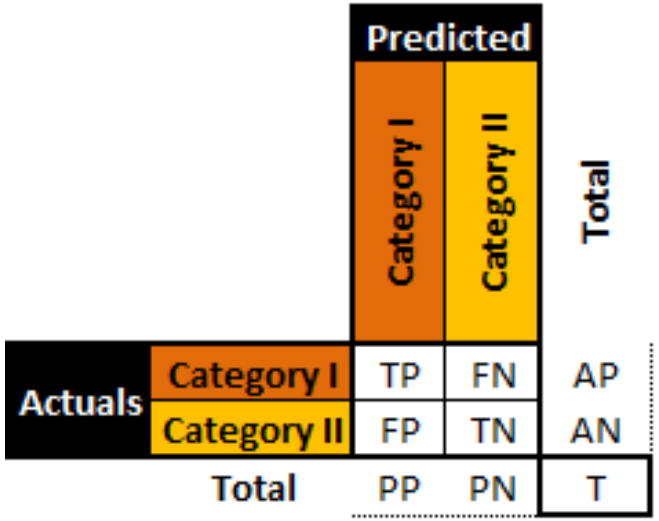
Classification Errors

When attempting to determine what kind of music a new customer would prefer, there is no real **cost** in making a mistake; if, on the other hand, the classifier attempts to determine the presence or absence of cancerous cells in lung tissue, mistakes are **more consequential**.

Several metrics can be used to assess a classifier's performance, depending on the context.

Binary classifiers are simpler and have been studied far longer than multi-level classifiers; consequently, a larger body of evaluation metrics is available for these classifiers.

Binary Classifiers



TP, TN, FP, FN: **True Positives**, **True Negatives**, **False Positives**, and **False Negatives**, respectively.

Perfect classifiers would have $FP, FN = 0$, but that rarely ever happens in practice (and not ideal, in a way).

Metrics:

- 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 = $(TP + TN)/T$

Other metrics:

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

Session 2

		Predicted A	Predicted B	Total	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	Predicted B	Total	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

Both classifiers have an accuracy of 80%; the second classifier makes some wrong predictions for A, but never for B; the first classifier makes mistakes for both classes. The second classifier mistakenly predicts occurrence A as B on 16 occasions, but the first one only does so 6 times. Which one is best depends on the **cost of misclassification**.

Multi-Level Classifiers

It is preferable to select metrics that generalize more readily to **multi-level classifiers.**

Accuracy: proportion of correct predictions amid all the observations

- value ranges from 0% to 100%
- the higher the accuracy, the better the match
- a predictive model with high accuracy may be useless thanks to the **Accuracy Paradox**

Matthews Correlation Coefficient (MCC): useful even when the classes are of very different sizes

- correlation coefficient between actual and predicted classifications
- range varies from -1 to 1
- if $MCC = 1$, predicted and actual responses are identical
- if $MCC = 0$, the classifier performs no better than a random prediction ("flip of a coin").

Multi-Level Classifiers

MCC: 69.7%
Accuracy: 78.3%
Pearson: 0.13161
Hist: 30.0%

			Predicted: Maltreatment: Unfounded	Predicted: Maltreatment: Suspected	Predicted: Maltreatment: Substantiated	Predicted: Risk: No	Predicted: Risk: Yes	Predicted: Risk: Unknown	Total	
Actuals	Maltreatment	Unfounded	4,577	-	-	198	6	-	*4,781*	*29.2%*
		Suspected	-	965	-	29	2	-	*995*	*6.1%*
		Substantiated	-	-	6,187	116	35	2	*6,339*	*38.7%*
	Risk	No	894	-	763	949	19	9	*2,632*	*16.1%*
		Yes	123	-	520	122	111	5	*880*	*5.4%*
		Unknown	212	-	303	184	21	24	*745*	*4.6%*
	Total		*5,805*	*965*	*7,772*	*1,597*	*194*	*40*	***16,372***	
			35.5%	*5.9%*	*47.5%*	*9.8%*	*1.2%*	*0.2%*		

Regression Performance Evaluation

For numerical targets y with predictions $\hat{y}$, metrics include:

- **mean squared** and **mean absolute errors**

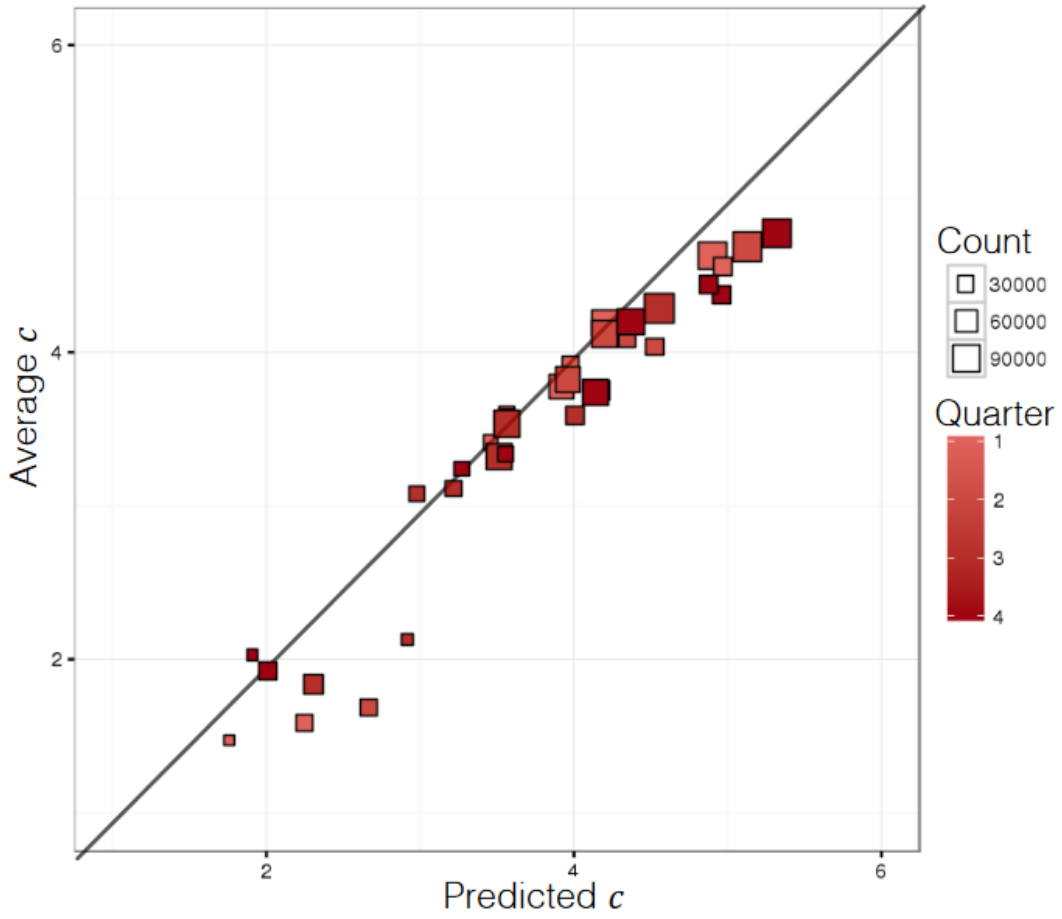
$$\text{MSE} = \text{mean}\{(\hat{y}_i - y_i)^2\},\ \text{MAE} = \text{mean}\{|\hat{y}_i - y_i|\}$$

- **normalized mean squared** and **normalized mean absolute errors**

$$\text{NMSE} = \frac{\text{mean}\{(\hat{y}_i - y_i)^2\}}{\text{mean}\{(\bar{y} - y_i)^2\}}, \text{NMAE} = \frac{\text{mean}\{|\hat{y}_i - y_i|\}}{\text{mean}\{|\bar{y} - y_i|\}}$$

- **mean average percentage error** $\text{MAPE} = \text{mean}\left\{\frac{|\hat{y}_i - y_i|}{y_i}\right\}$
- **correlation** $\rho_{\hat{y},y}$

Regression Performance Evaluation



Suggested Reading

Performance Evaluation

Data Understanding, Data Analysis, Data Science
Machine Learning 101

Classification and Value Estimation

- Performance Evaluation

Regression and Value Estimation

*Statistical Learning (advanced)

- Model Evaluation

Exercises

Performance Evaluation

We continue the UniversalBank example. The confusion matrices for the predictions of trees A and B on the remaining 2000 testing observations are shown below.

1. Using the appropriate matrices, compute the performance evaluation metrics for each of the trees (on the testing set).

2. If customers who would not accept a personal loan get irritated when offered a personal loan, what tree should *the* marketing group use to maintain good customer relations?

Tree A

		Predicted		Total	
		A	B		
Actuals	A	1792	19	1811	90.55%
	B	18	171	189	9.45%
Total		1810	190	2000	
		90.50%	9.50%		

Tree B

		Predicted		Total	
		A	B		
Actuals	A	1801	10	1811	90.55%
	B	64	125	189	9.45%
Total		1865	135	2000	
		93.25%	6.75%		

Session 3

INTRODUCTION TO MACHINE LEARNING

Clustering

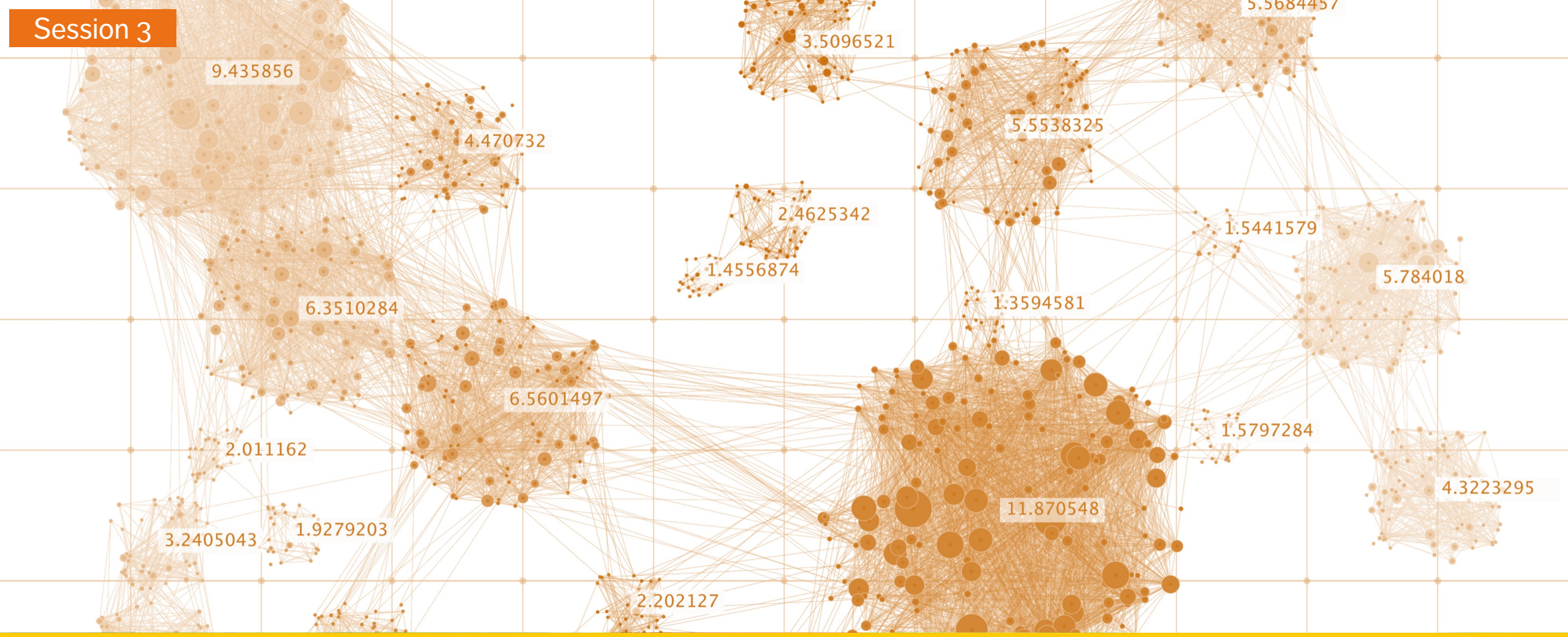
INTRODUCTION TO MACHINE LEARNING

Clustering is in the eye of the beholder, and as such, researchers have proposed many induction principles and models whose corresponding optimisation problem can only be approximately solved by an even larger number of algorithms.

[V. Estivill-Castro, *Why So Many Clustering Algorithms?*]

Woes clusters. Rare are solitary woes; they love a train, they tread each other's heel.

[E. Young]



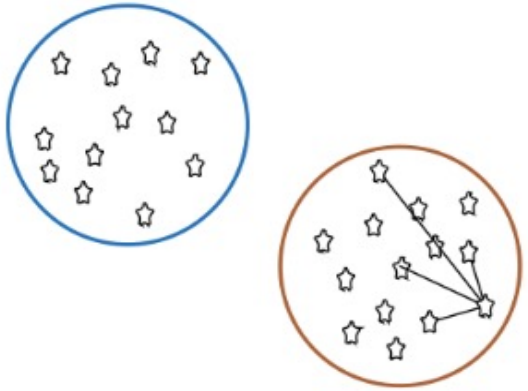
7. Clustering Overview

Overview

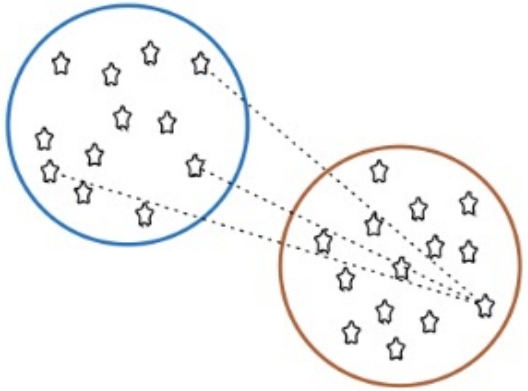
In **clustering**, the data is divided into **naturally occurring groups**. Within each group, the data points are **similar**; from group to group, they are **dissimilar**.

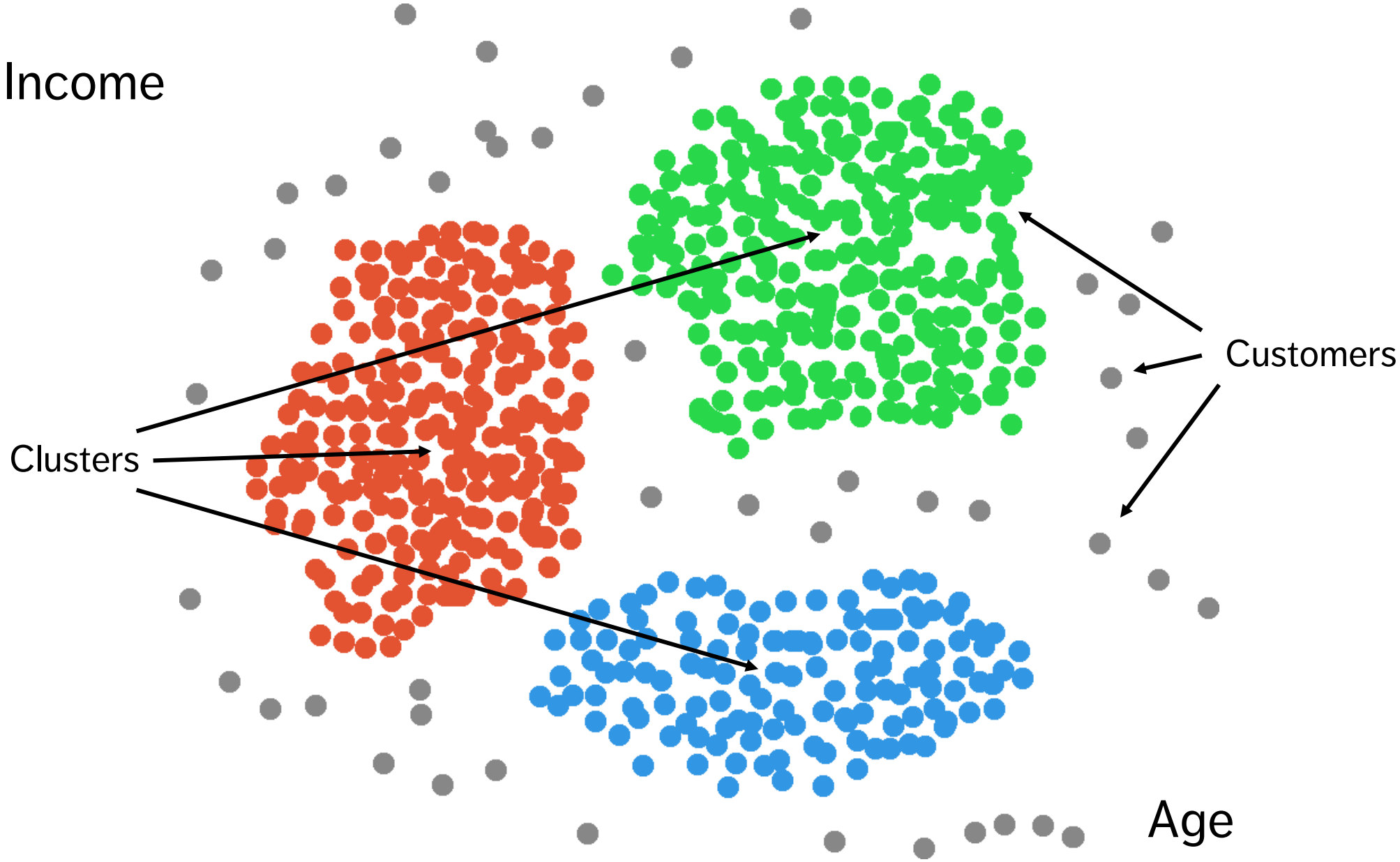
The grouping labels are not determined ahead of time, so clustering is an example of **unsupervised** learning.

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



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




Income
Customers
Clusters
Age

Overview

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

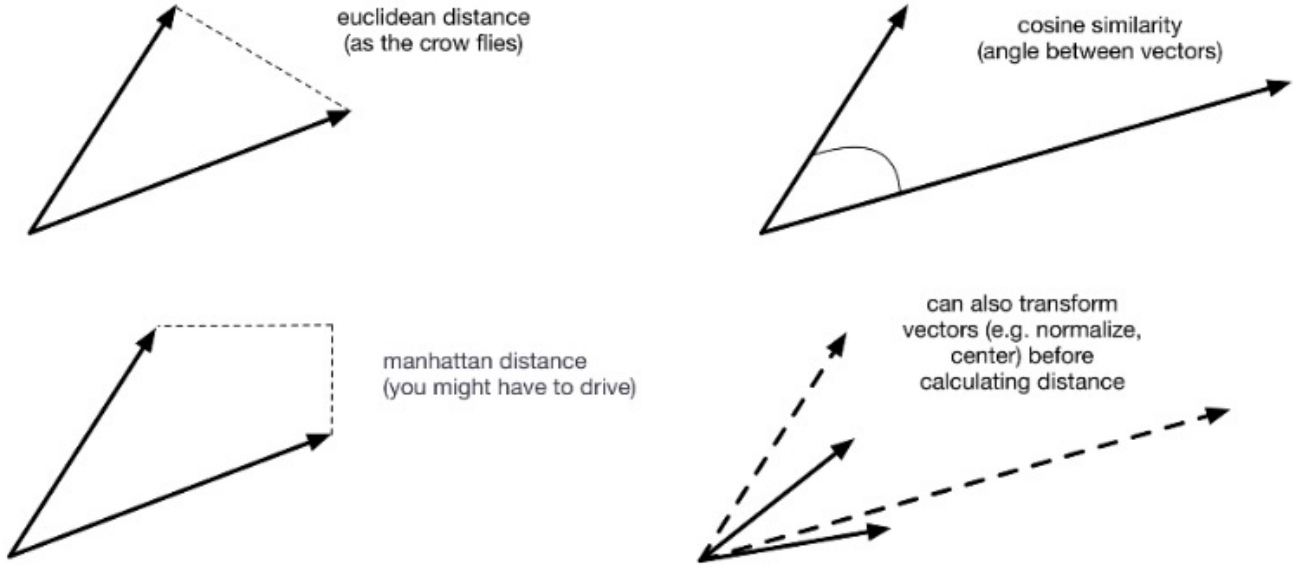
- in spite of that, the temptation to explain clusters *a posteriori* is **strong**

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

Clustering Requirement

A measure of **similarity** w (or a distance d) between observations:



IMPORTANT: data must be scaled before it is fed into clustering algorithms.

Typically, $w \to 1$ as $d \to 0$, and $w \to 0$ as $d \to \infty$.

Distance Measures (Metrics)

Categorical Variables*

- Hamming distance
- Russel/Rao index
- Jaccard
- Dice's coefficient
- etc.

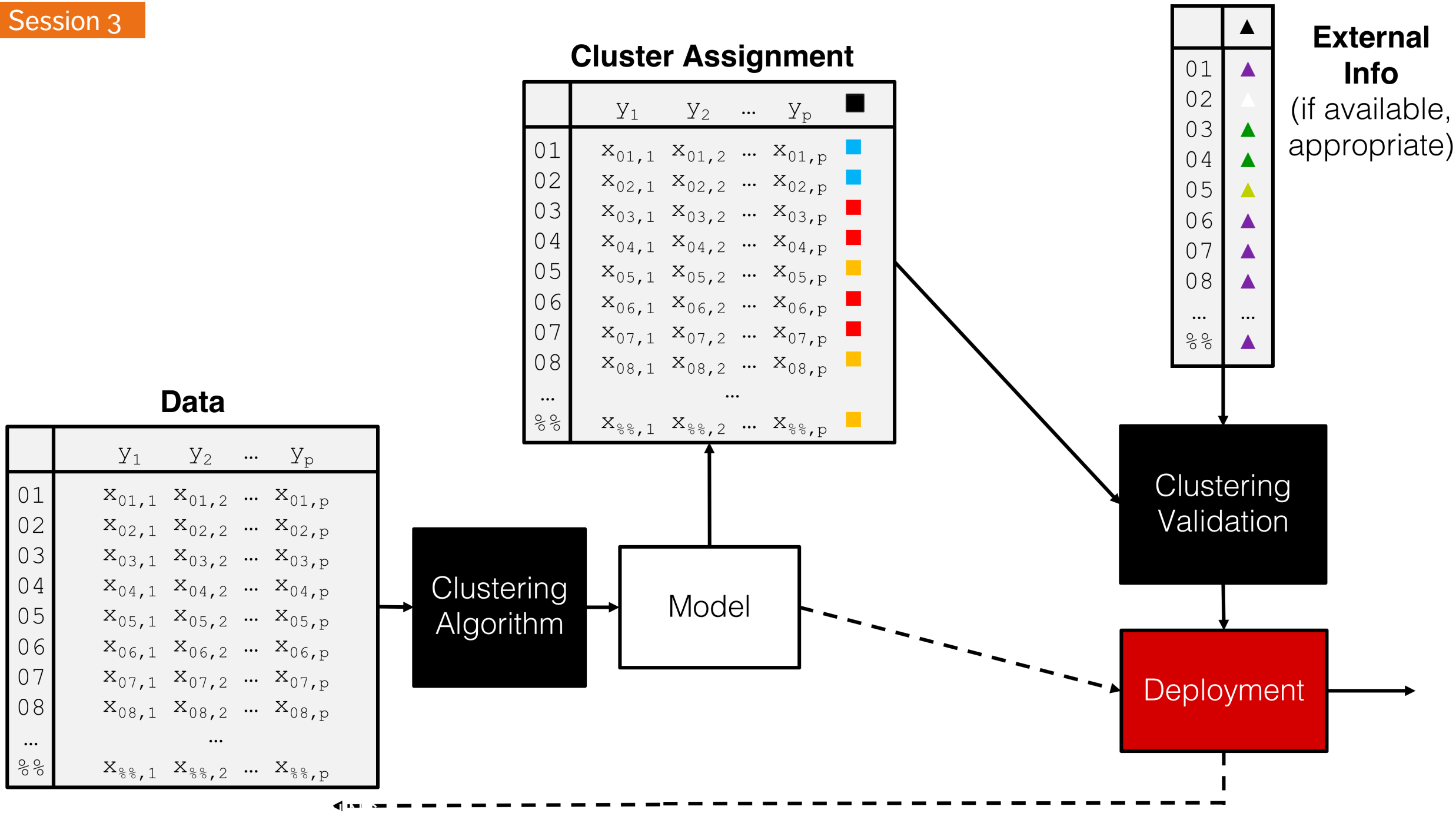
No steadfast rule to determine which distance to use; competing schemes are often produced with diff. metrics.

Numerical Variables

- Euclidean
- Manhattan
- correlation
- cosine
- etc.

We may need to create hybrid metrics for dataset with both categorical and numerical variables.

Session 3


Cluster Assignment
External Info (if available, appropriate)
Data
Clustering Algorithm
Model
Clustering Validation
Deployment

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

Applications

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 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

Case Study

Livehoods

Cranshaw *et al.*
The Livehoods Project: Utilizing Social Media to Understand the Dynamics of a City
ICWSM, 2012

Objective

When we think of similarity at the urban level, we typically think in terms of neighbourhoods. Is there some other way to identify similar parts of a city?

The researchers aims to draw the boundaries of **livehoods**, areas of similar character within a city, by using clustering models. Unlike **static** administrative neighborhoods, the livehoods are defined based on the **habits** of their inhabitants.

Case Study

Livehoods

Cranshaw *et al.*
The Livehoods Project: Utilizing Social Media to Understand the Dynamics of a City
ICWSM, 2012

Methodology

The authors use **spectral clustering** to discover **distinct geographic areas** of the city based on collective **movement patterns**.

Livehood clusters are built as follows:

1. a **geographic distance** is computed based on pairs of check-in venues' coordinates;
2. a **social similarity** is computed between each pair of **venues** using cosine measurements;
3. spectral clustering produces **candidate livehoods**;
4. interviews are conducted with residents in order to **explore**, **label**, and **validate** the clusters discovered by the algorithm.

Case Study

Livehoods

Cranshaw *et al.*
The Livehoods Project: Utilizing Social Media to Understand the Dynamics of a City
ICWSM, 2012

Data

The data comes from two sources, combining approximately 11 million check-ins from the dataset of Chen et al. (a recommendation site for venues based on users' experiences) and a new dataset of 7 million Twitter check-ins downloaded between June and December of 2011.

For each check-in, the data consists of the **user ID**, the **time**, the **latitude and longitude**, the **name of the venue**, and its **category**.

In this case study, data from the city of Pittsburgh, Pennsylvania, is examined *via* 42,787 check-ins of 3840 users at 5349 venues.

Case Study

Livehoods

Cranshaw *et al.*
The Livehoods Project: Utilizing Social Media to Understand the Dynamics of a City
ICWSM, 2012

Strengths and Limitations of the Approach

- The technique used in this study is **agnostic** towards the particular source of the data: it is not dependent on meta-knowledge about the data.
- The algorithm may be prone to "majority" bias, possibly misrepresenting/hiding minority behaviours.
- The dataset is built from a **limited** sample of check-ins shared on Twitter and are therefore biased towards the types of visits/locations that people typically want to share **publicly**.
- Tuning the clusters is non-trivial: experimenter bias may combine with "confirmation bias" of the interviewees in the validation stage – if the researchers are residents of Pittsburgh, will they see clusters when there were none?

Case Study

Livehoods

Cranshaw *et al.*
The Livehoods Project: Utilizing Social Media to Understand the Dynamics of a City
ICWSM, 2012

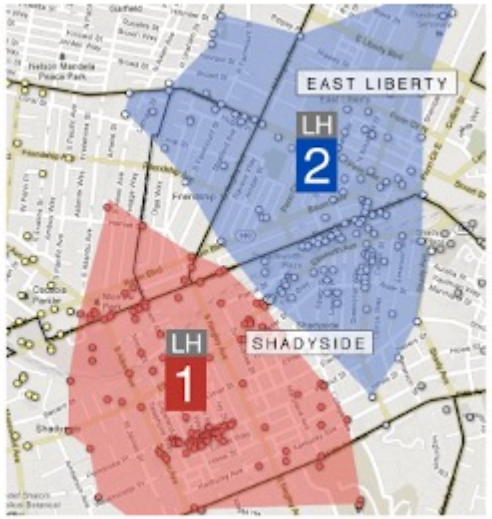
Results, Evaluation, and Validation

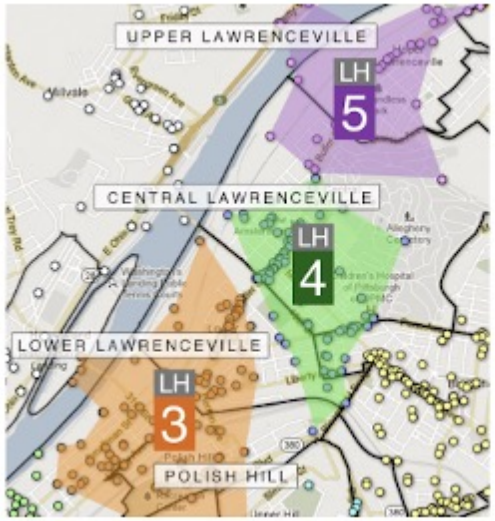
Over 3 areas of the city, 9 livehoods have been identified and validated by 27 Pittsburgh residents

- **Municipal Neighborhoods Borders:** livehoods are dynamic, and evolve as people's behaviours change, unlike fixed neighbourhoods set by the city government.
- **Demographics:** the interviews displayed strong evidence that the demographics of the residents and visitors of an area play a strong role in explaining the livehood divisions.
- **Development and Resources:** economic development can affect the character of an area. Similarly, the resources provided by a region has a strong influence on the people that visit it, and hence its resulting character.

Case Study

Livehoods

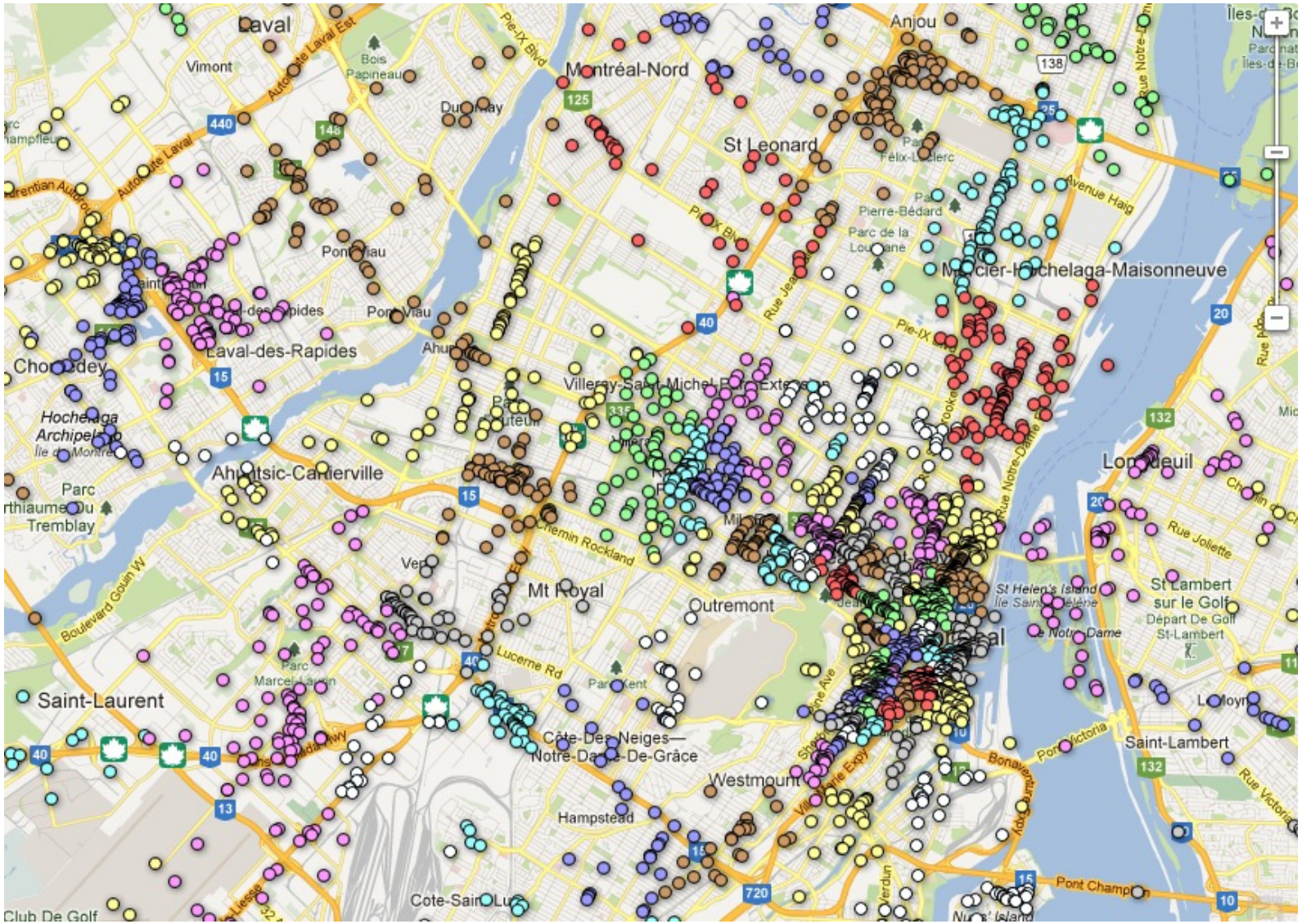




Cranshaw *et al.*
The Livehoods Project: Utilizing Social Media to Understand the Dynamics of a City
ICWSM, 2012

Case Study

Livehoods



Cranshaw *et al.*
[The Livehoods Project: Utilizing Social Media to Understand the Dynamics of a City](#)
ICWSM, 2012

General Remarks

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”

Suggested Reading

Clustering Overview

Data Understanding, Data Analysis, Data Science
Machine Learning 101

Clustering

- Overview
- Case Study: Livehoods

Spotlight on Clustering

*Overview (advanced)

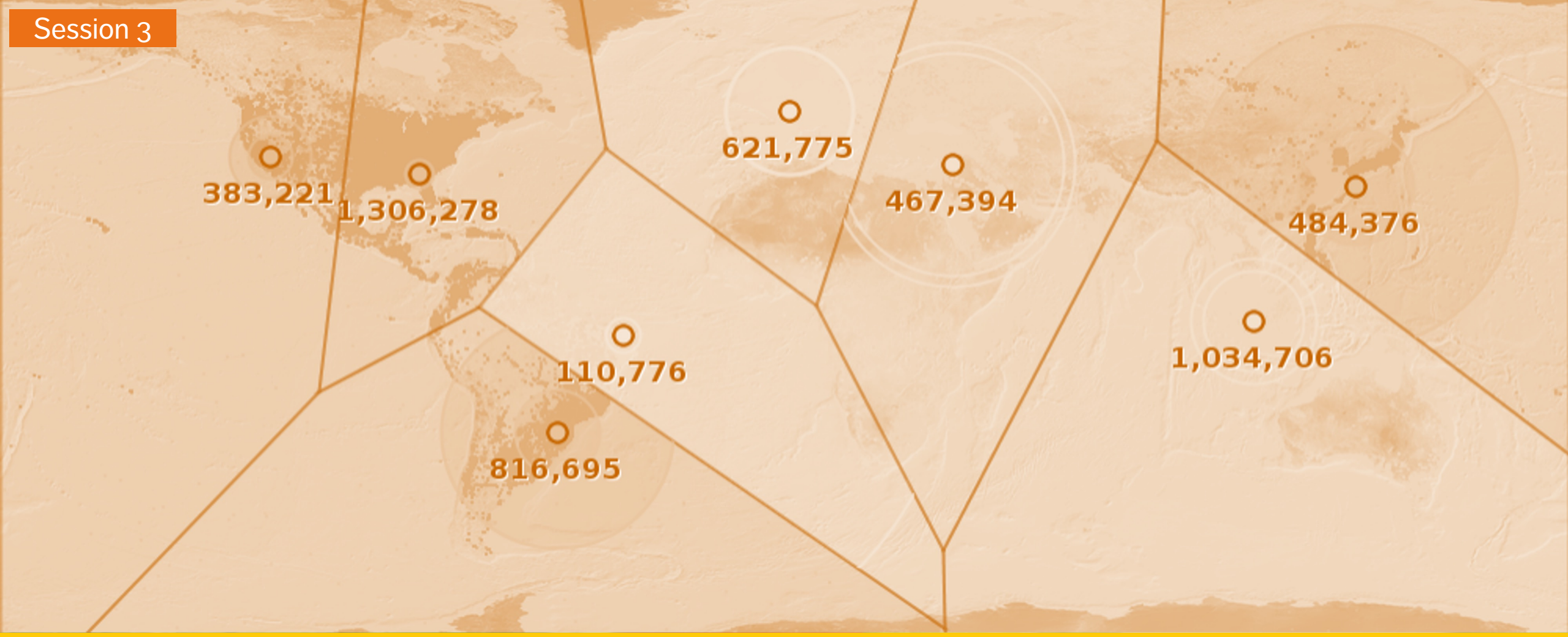
- Unsupervised Learning
- Clustering Framework
- A Philosophical Approach to Clustering

Exercises

Clustering Overview

1. What does the (potential) non-replicability of clustering imply for validation? For client and/or stakeholder buy-in?

2. Identify scenarios and questions that could use classification and/or value estimation in your every day work activities.



8. k-Means and Other Algorithms

Clustering Algorithms

k-Means

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

Hierarchical Clustering

- easy to interpret, deterministic

Cluster Ensembles

Latent Dirichlet Allocation

- used for topic modeling

Expectation Maximization

Clustering Algorithms

Balanced Iterative Reducing and Clustering using Hierarchies

Density-Based Spatial Clustering of Applications with Noise

- graph-based

Affinity Propagation

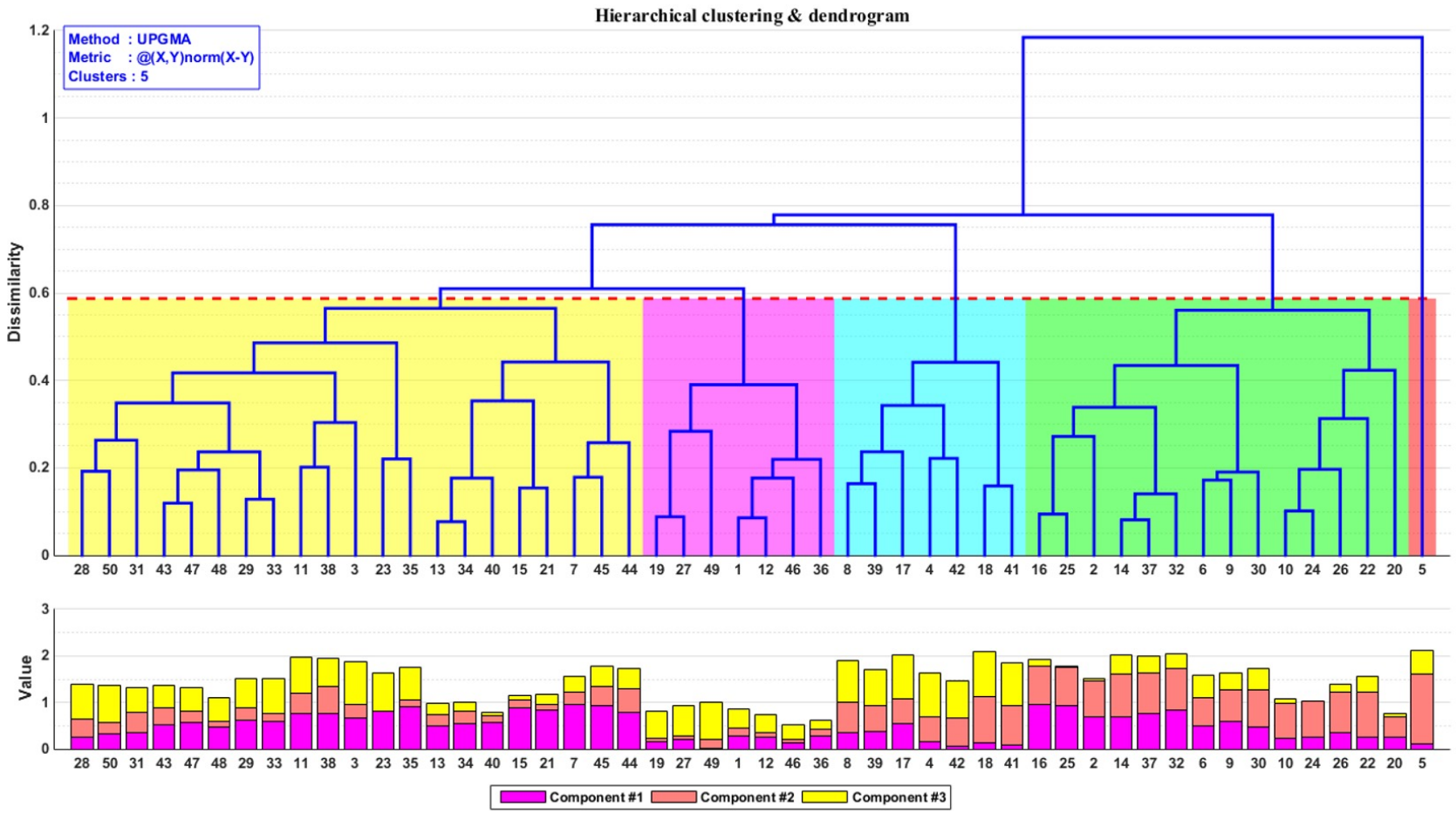
- selects the optimal number of clusters automatically

Spectral Clustering

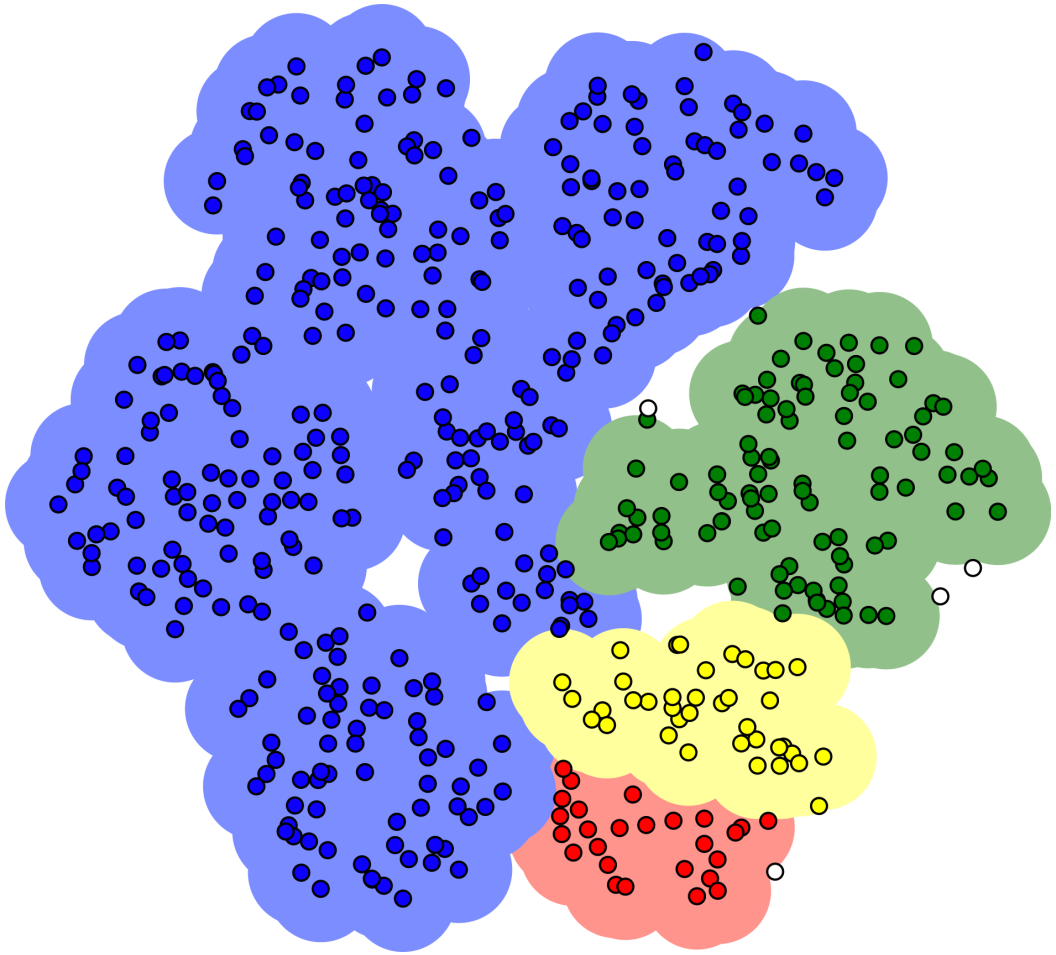
- recognizes non-blob clusters

Fuzzy Clustering

Hierarchical Clustering

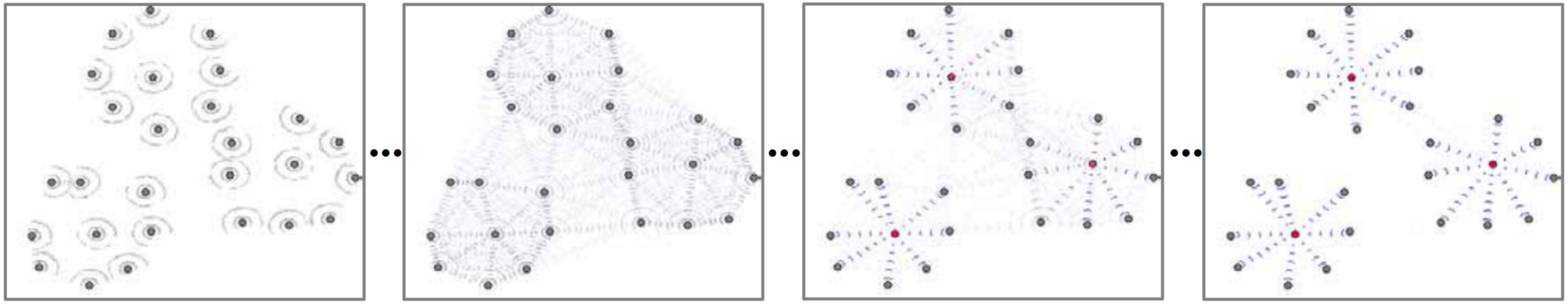


DBSCAN



[D. Dueck]

Affinity Propagation



initialization

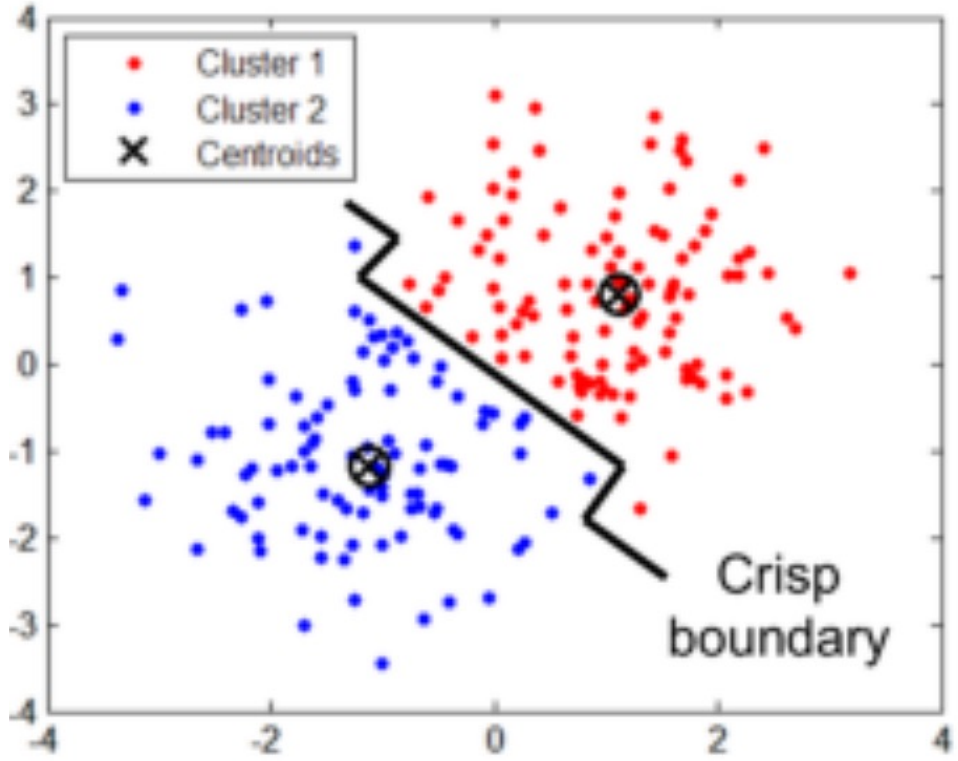
iteration X
(uncertainty)

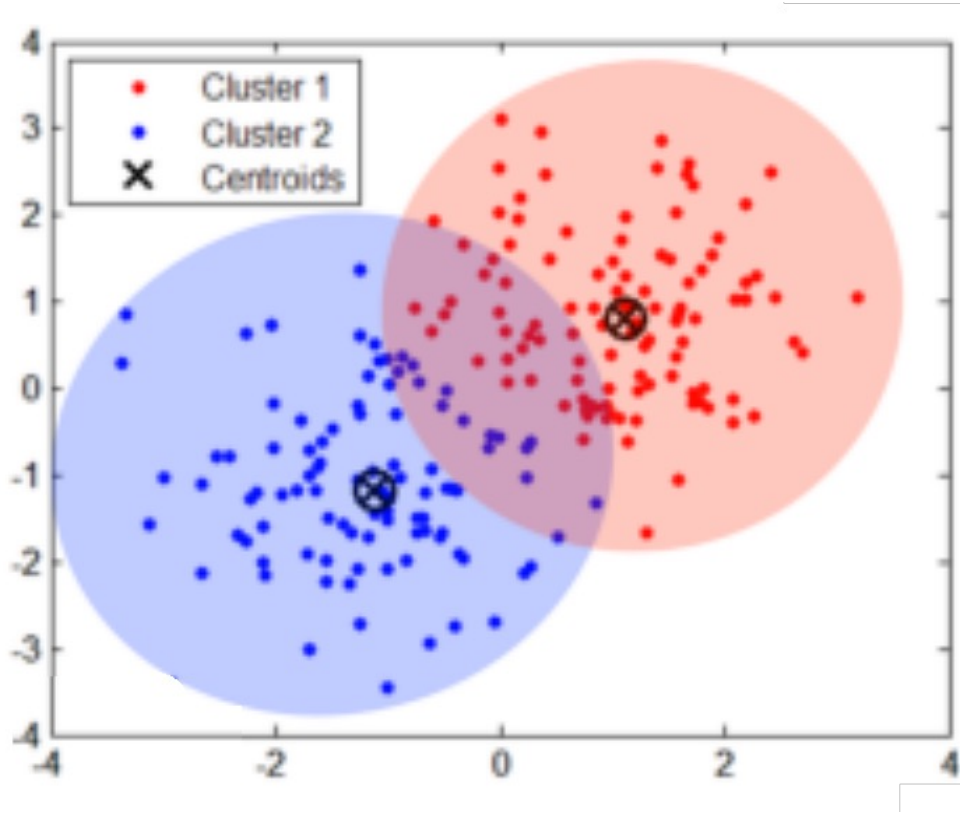
iteration Y
(emergence)

final iteration
(convergence)

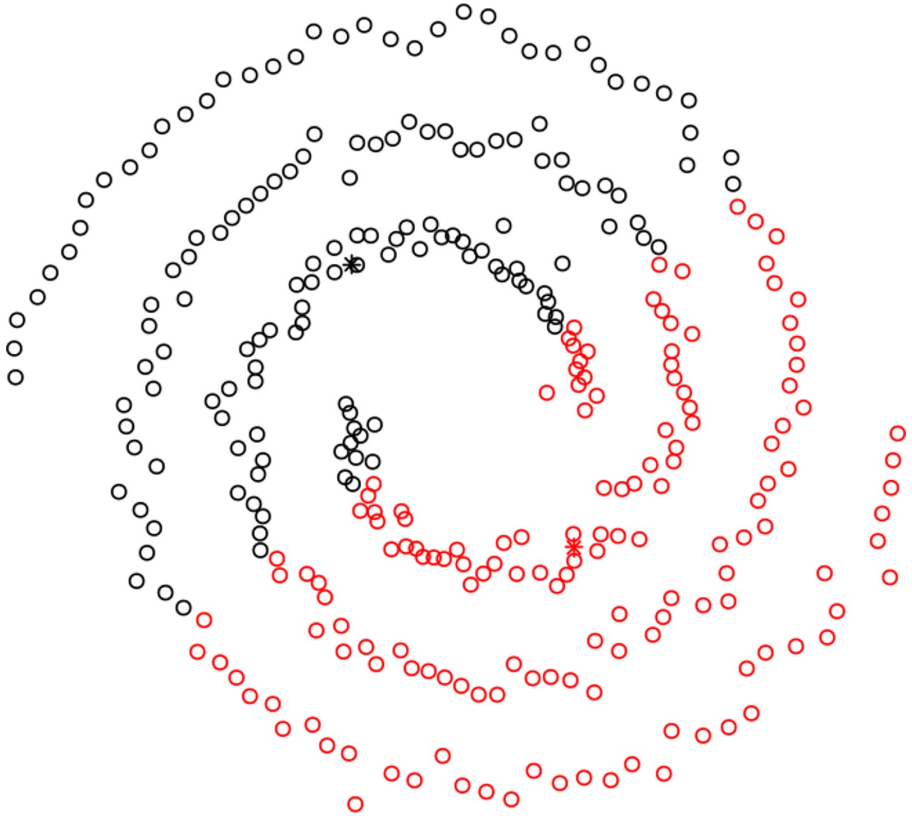
[A. Bagnaro, F. Baltar, and G. Brownstein]

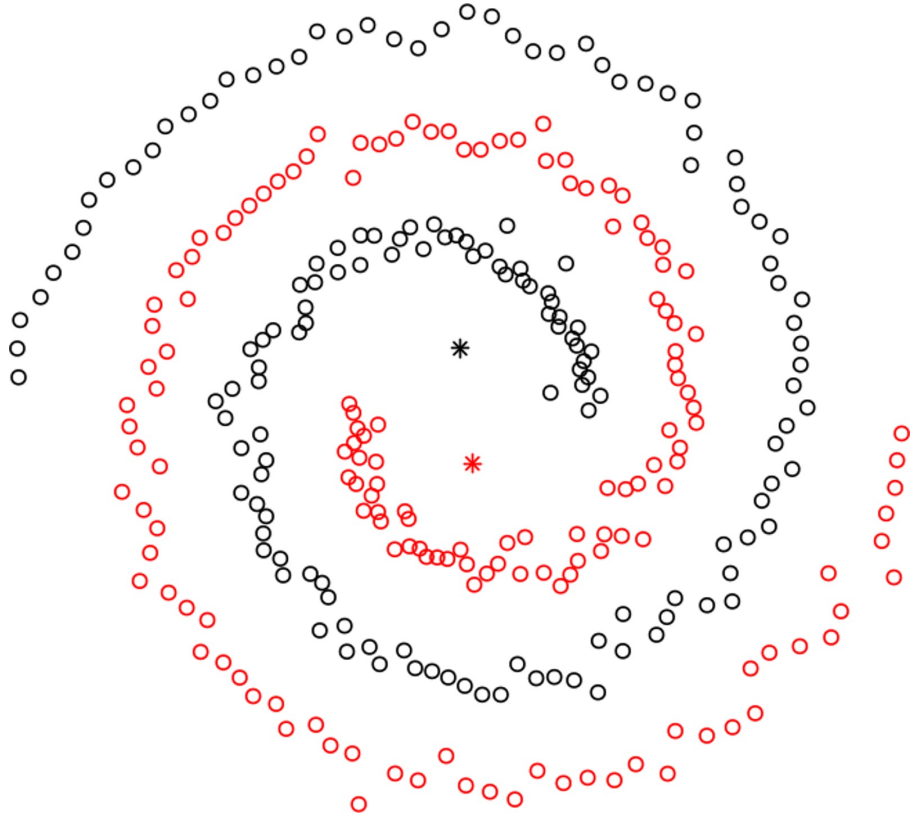
k-Means and Fuzzy c-Means





k-Means and Spectral Clustering

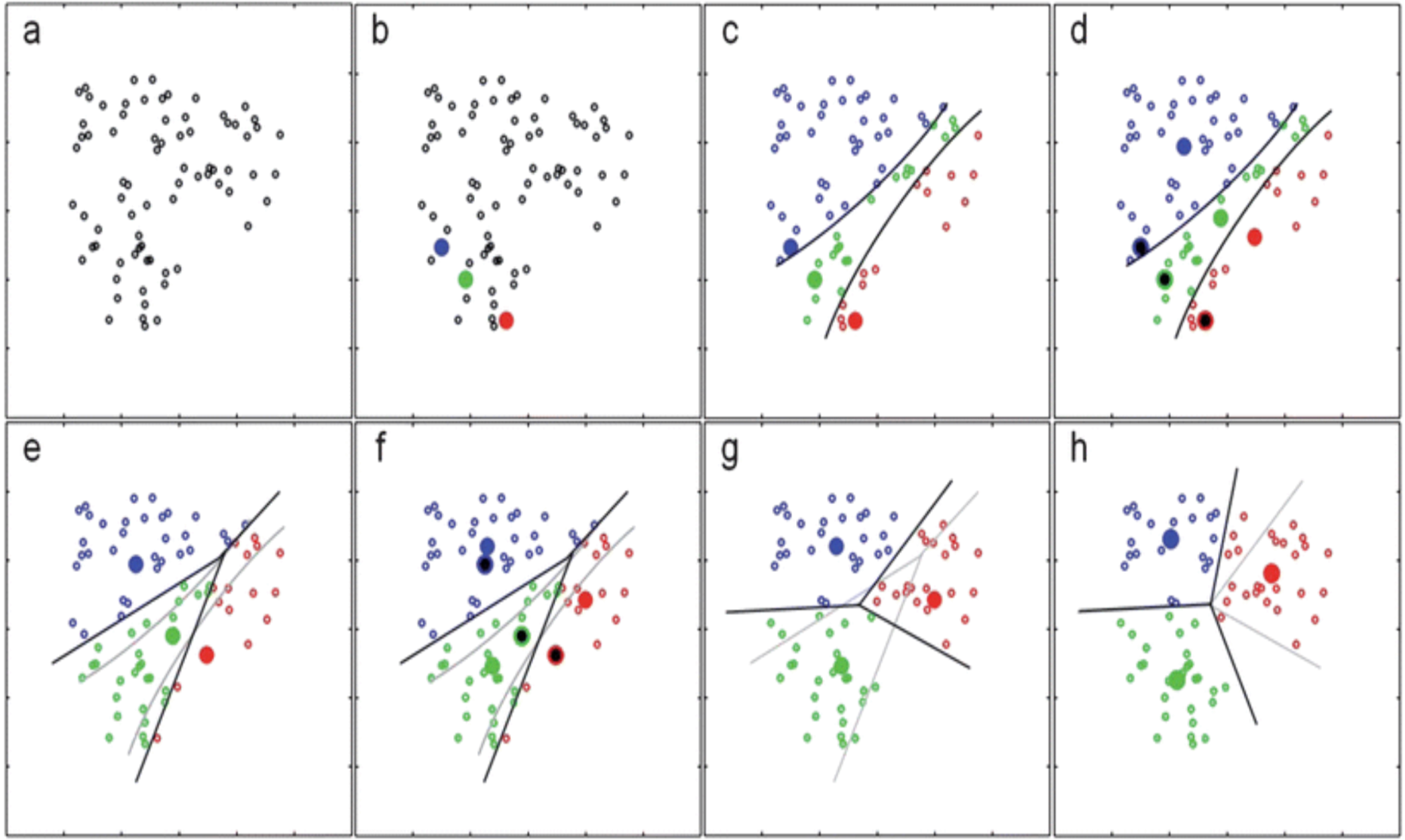




k-Means Algorithm

1. Select the desired **number of clusters**, say k
2. Randomly choose k instances as initial **cluster centres**
3. Calculate the **distance** from each observation to each centre
4. Place each instance in the cluster whose centre it is **nearest** to
5. Compute the **centroid** for each cluster
6. Repeat steps 3 – 5 with the new centroids
7. Repeat step 6 until the clusters are **stable**

[A.H. Du]



k-Means Strengths

Easy to implement (without having to actually compute pairwise distances).

- extremely common as a consequence
- elegant and simple

In many contexts, k-means is a **natural** way to look at grouping observations.

Helps provide a **basic understanding of the data structure** in a first pass.

k-Means Limitations

Data points can only be assigned to **one** cluster

- this can lead to overfitting
- robust solution: consider the probability of belonging to each cluster

Underlying clusters are assumed to be **blob-shaped**

- k-means will fail to produce useful clusters if that assumption is not met in practice

Clusters are assumed to be separate (discrete)

- k-means does not allow for **overlapping** or **hierarchical** groupings

k-Means Limitations

There are many ways to pick the **optimal number** of clusters k.

One problem is that the algorithm is stochastic: different initial configurations may yield **different outcomes**, which may yield a different optimal number.

It may also depend on the **size** of data, the choice of **distance**, the choice of **cluster quality metric**, etc.

Suggested Reading

k-Means and Other Algorithms

Data Understanding, Data Analysis, Data Science
Machine Learning 101

Clustering

- Clustering Algorithms
- k-Means
- Toy Example: Iris Dataset

R Examples

- Clustering: Iris Dataset

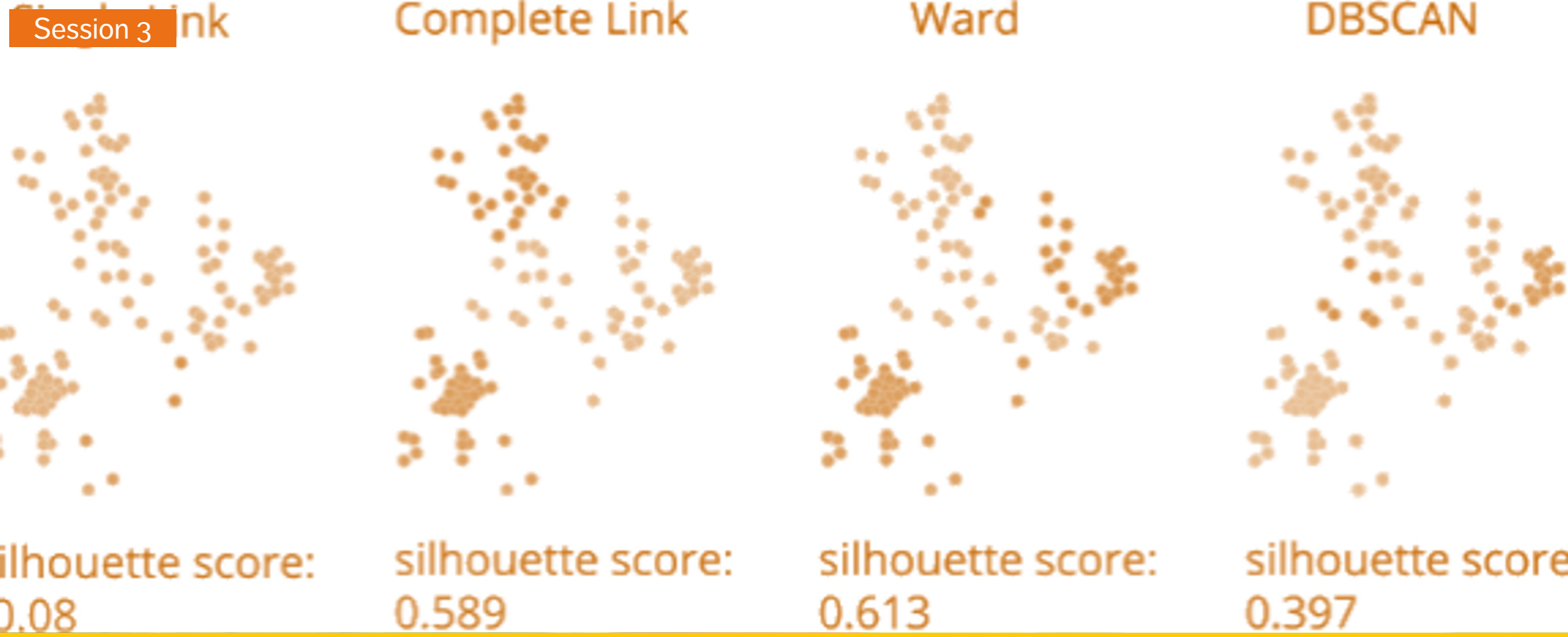
Spotlight on Clustering

*Simple Clustering Methods (advanced)
*Advanced Clustering Approaches (advanced)

Exercises

k-Means and Other Algorithms

1. Go over the iris clustering example found in DUDADS (see suggested reading). Repeat the process with the UniversalBank dataset (you may wish to visualize the dataset first) in order to build a clustering scheme. Determine the optimal number of clusters using the Davies-Bouldin index.



9. Validation and Notes

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 Validation

Optimal clustering scheme:

- maximal separation between clusters
- maximal similarity within groups
- agrees with human eye test
- useful at achieving its goals

Validation types

- **external** (uses additional information)
- **internal** (uses only the clustering results)
- **relative** (compares across clustering attempts)

Clustering Validation

Clustering involves two main activities:

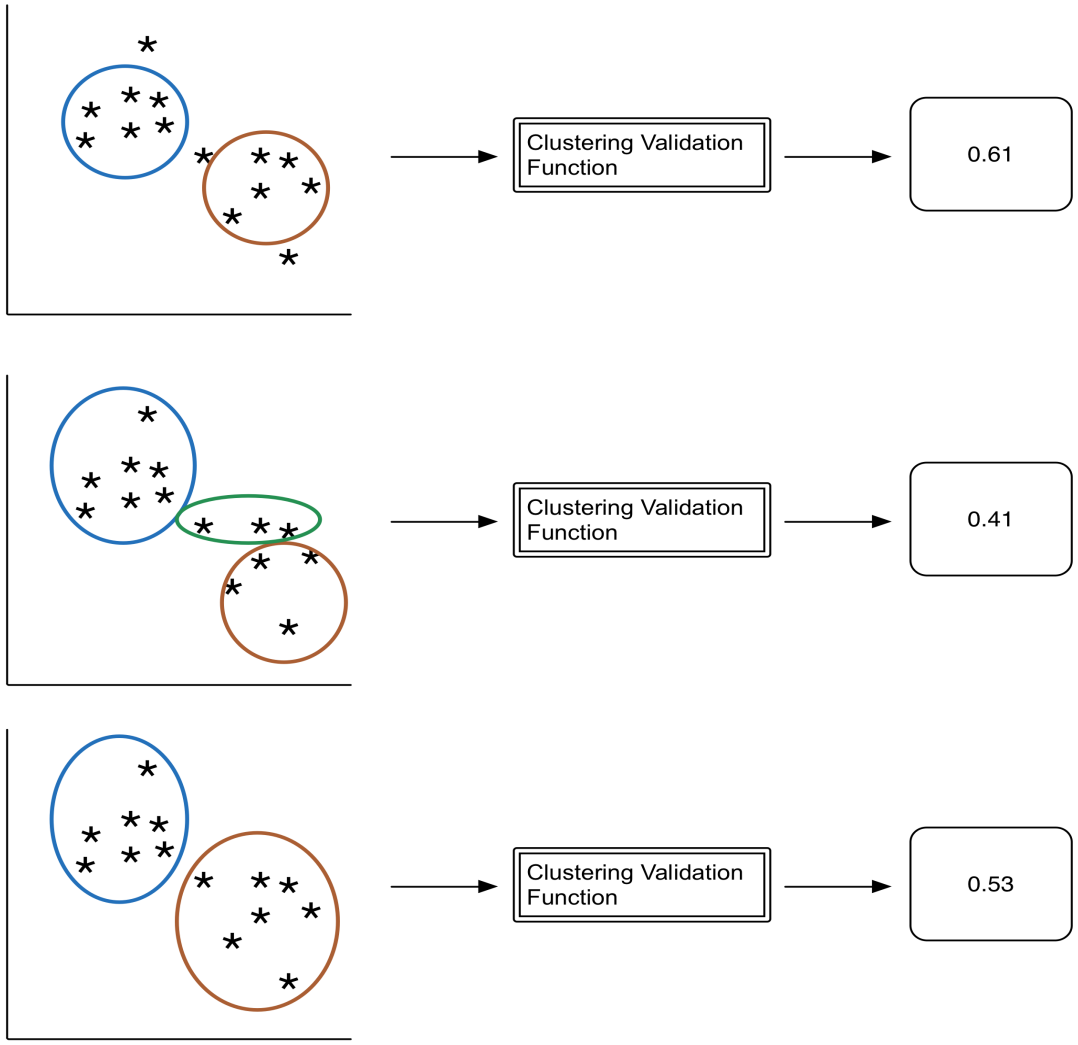
- creating clusters
- **assessing cluster quality**

Clustering functions

- input: instances (vectors)
- output: cluster assignment to each instance

Assessing cluster quality

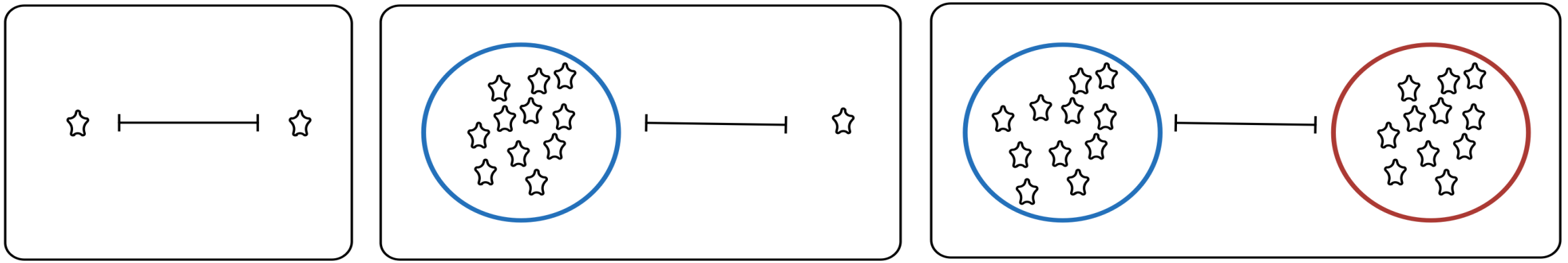
- input: instances + cluster assignments (+ similarity matrix, usually)
- output: a numeric value



Function Components

There are many clustering and cluster validation functions, but they are all built out of basic measures relating to instance or cluster properties:

- **instance properties**
- **cluster properties**
- **instance – instance relationship properties**
- **cluster – instance relationship properties**
- **cluster – cluster relationship properties**



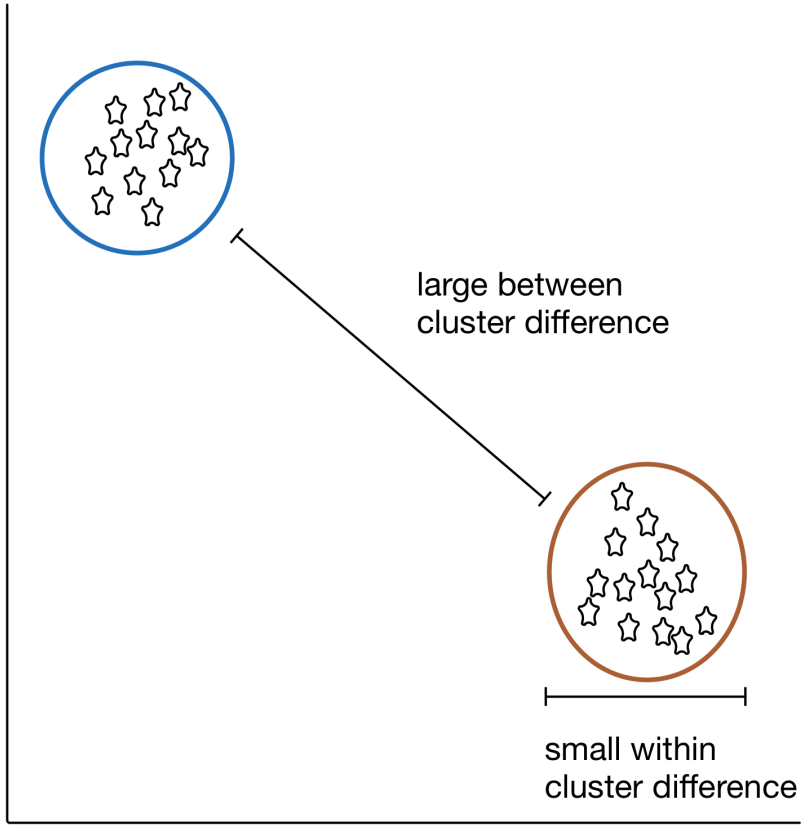
Internal Validation Goals

Within clusters, everything is very similar. Between clusters, there is a lot of difference.

The problem: there are many ways for clusters to deviate from this ideal.

How do we weigh the good aspects (e.g., high **within-cluster similarity**) relative to the bad (e.g., low **between-cluster separation**).

Thus, the large # of **cluster quality metrics** (CQM).



Internal Validation CQM

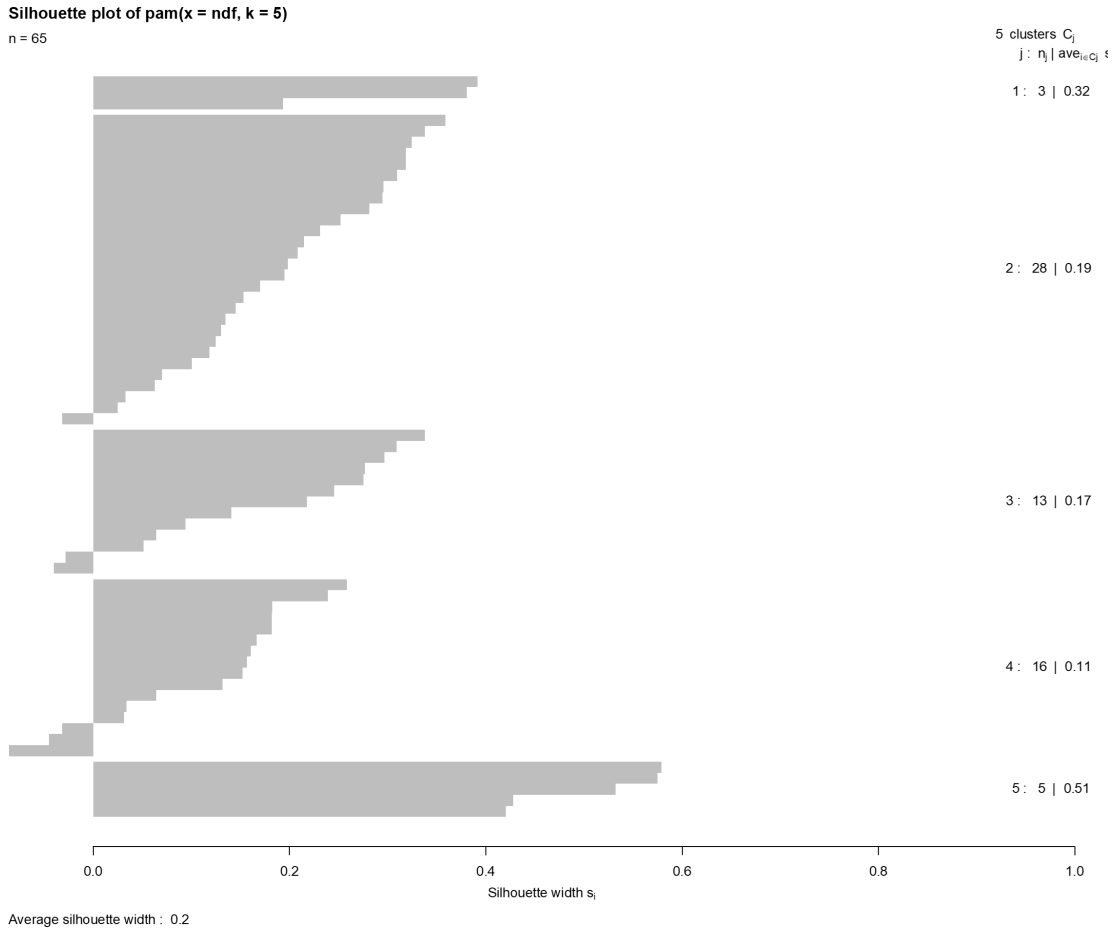
Davies-Bouldin index

Dunn's index

Silhouette metric

Within Sum of Squares

etc. (there are tons!)

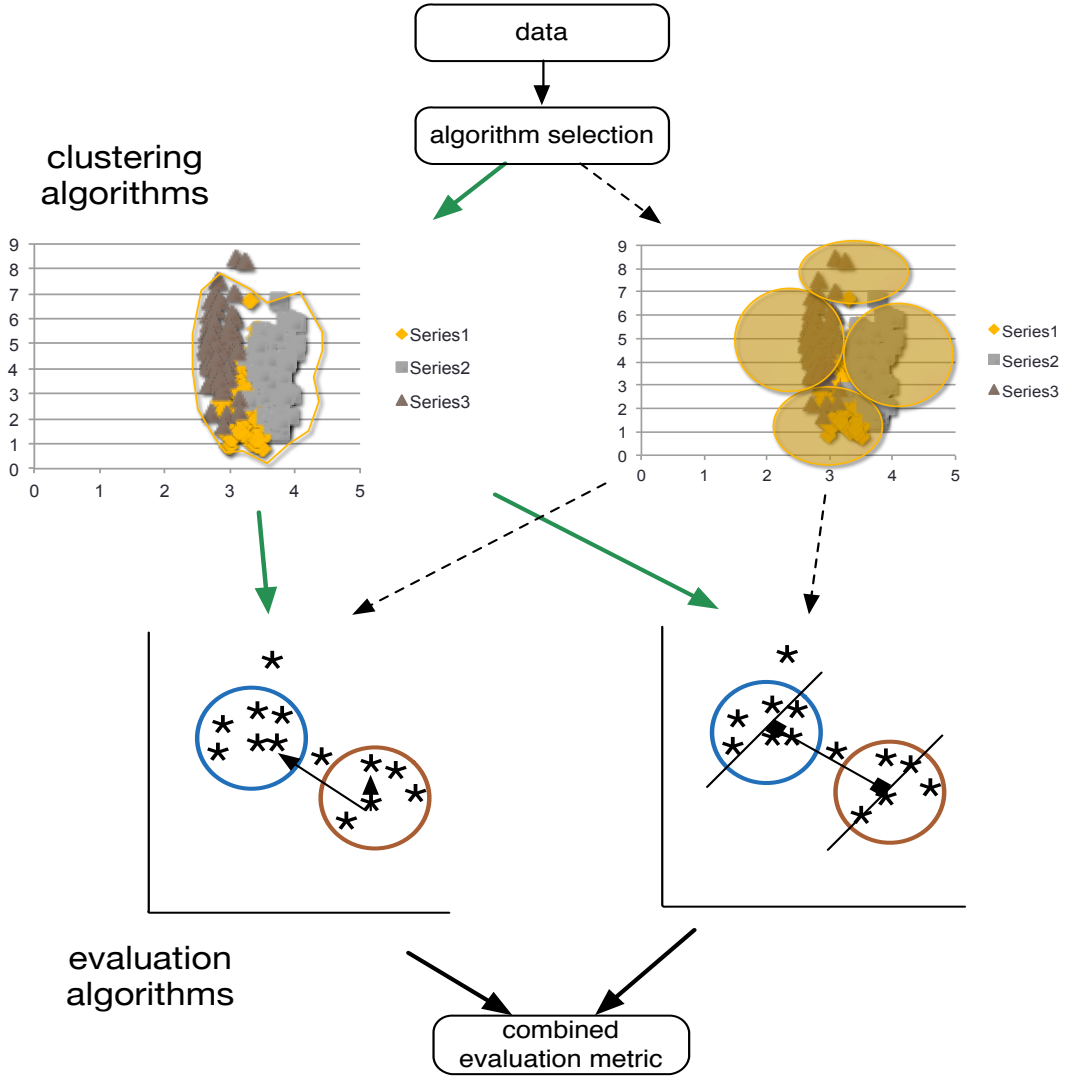


Relative Validation

Getting a single validation measure for a single clustering is not that useful – could the results be better? Is this the best we can hope for?

We could **compare results** across runs or parameter settings.

The main difficulty is to determine how to compare results of **individual runs**.



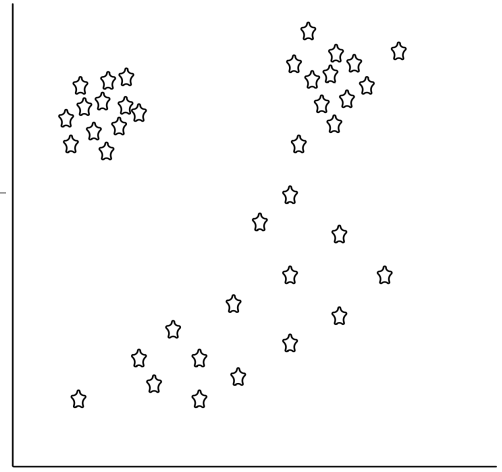
Ensemble Methods

Some options:

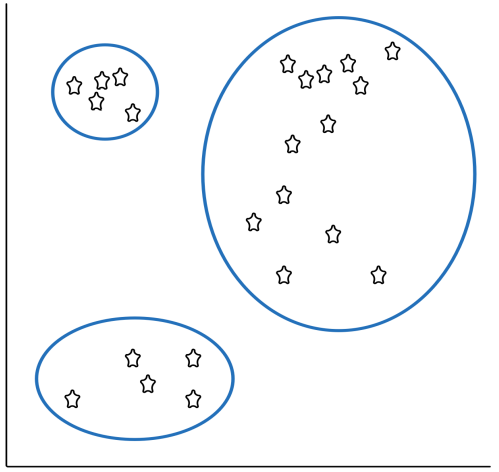
- multiple samples from the same source
- different subsets of columns are used
- different algorithms are used

The **similarity** of the clustering results is measured.

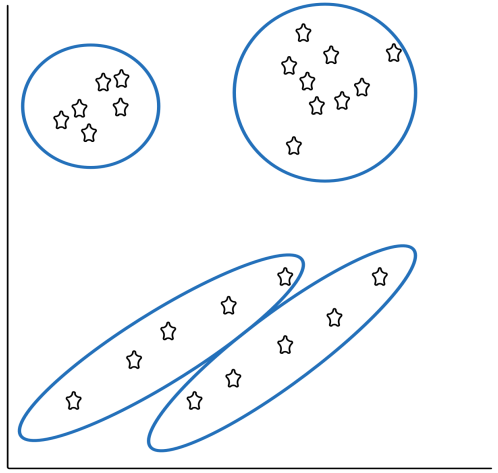
If the results are **not stable** across the clustering outcomes, more investigation is required.



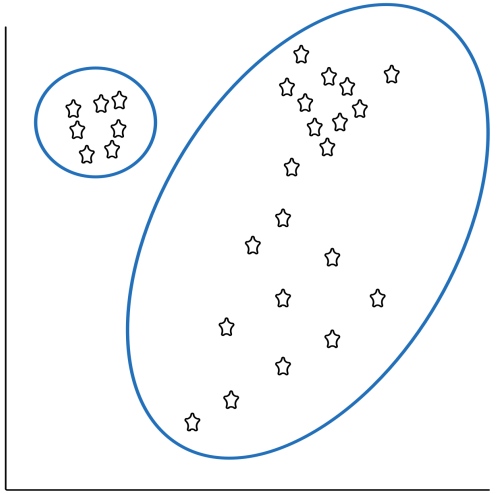
Full dataset



Sample 1 clustering



Sample 2 clustering



Sample 3 clustering

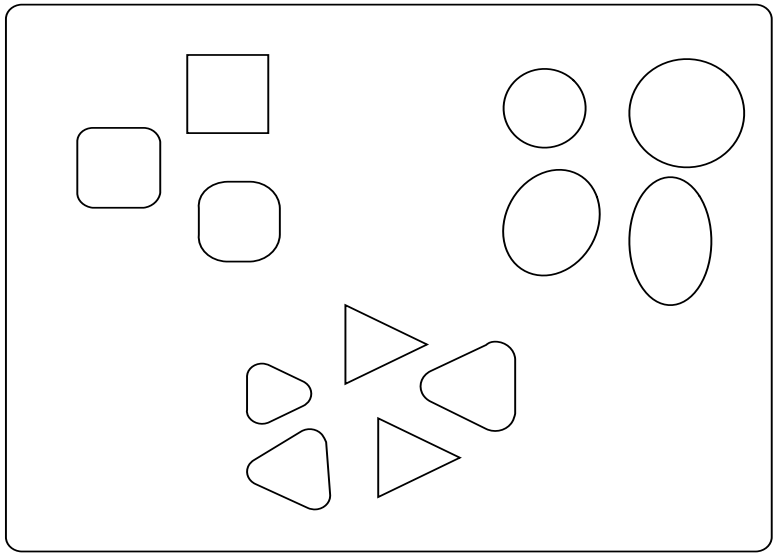
External Validation

Brings in outside info. to **evaluate** the clusters.

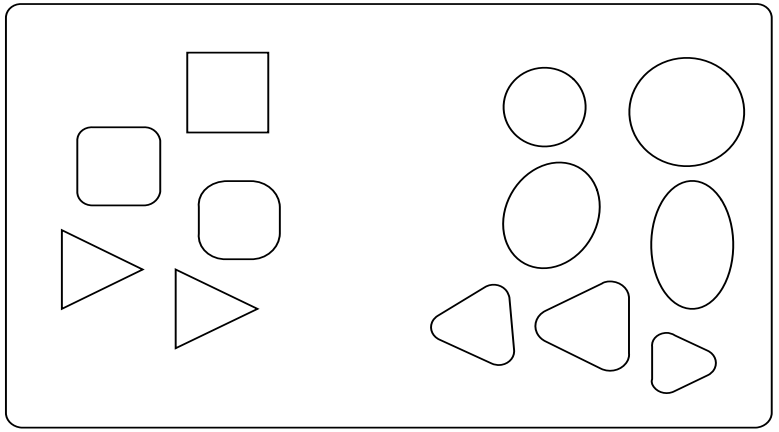
Outside information is typically the 'correct' class.

How is this different from classification then?

Often used to build confidence in the overall approach, based on preliminary or sample results.



Natural Groupings



Clustering Results

Purity

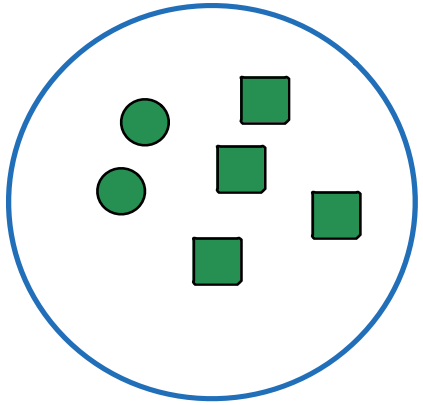
For this external validation metric, each cluster is assigned to the class which is **most frequent** in the cluster.

We calculate the **purity** as follows: number of correctly assigned points / number of points in the cluster.

Some other options: **precision**, **recall**.

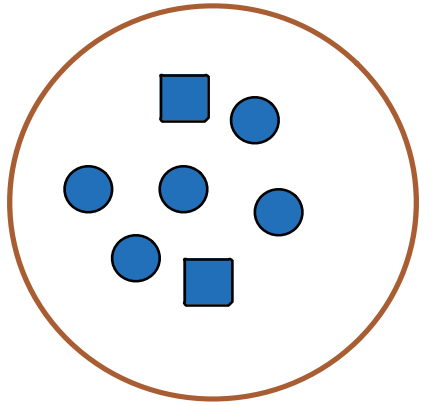
Assuming we are interested in shape…

SQUARE CLUSTER



purity = 66%

CIRCLE CLUSTER



purity = 71%

Clustering Challenges

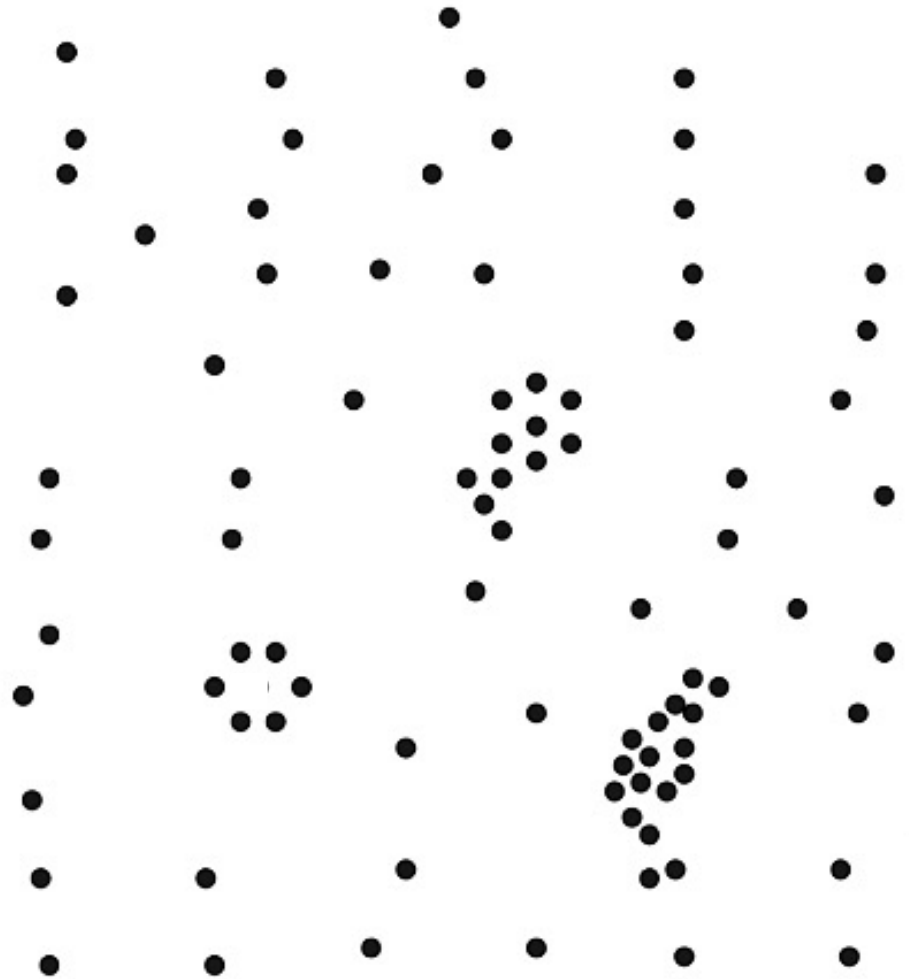
Automation
relatively intuitive for humans, but harder for machines

Lack of a clear-cut definition
no universal agreement as to what constitutes a cluster

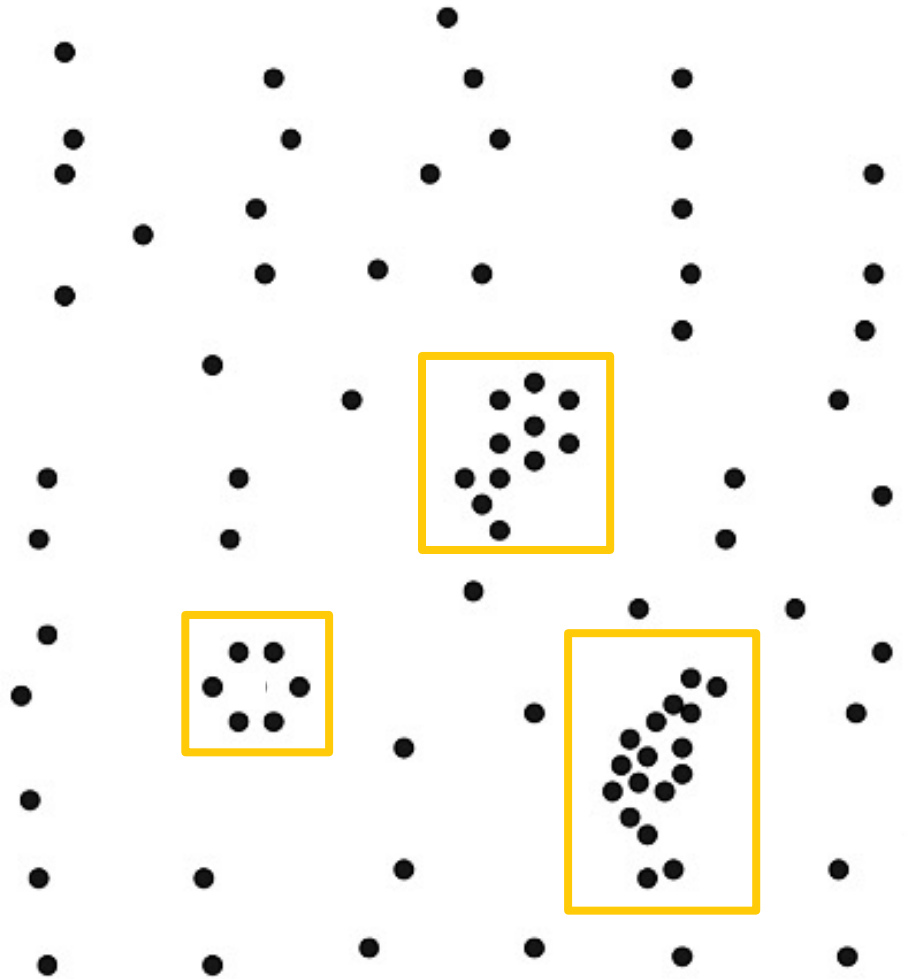
Lack of repeatability
non-deterministic: the same algorithm, applied twice to the same dataset can discover completely different clusters

Number of clusters
optimal number of clusters difficult to determine

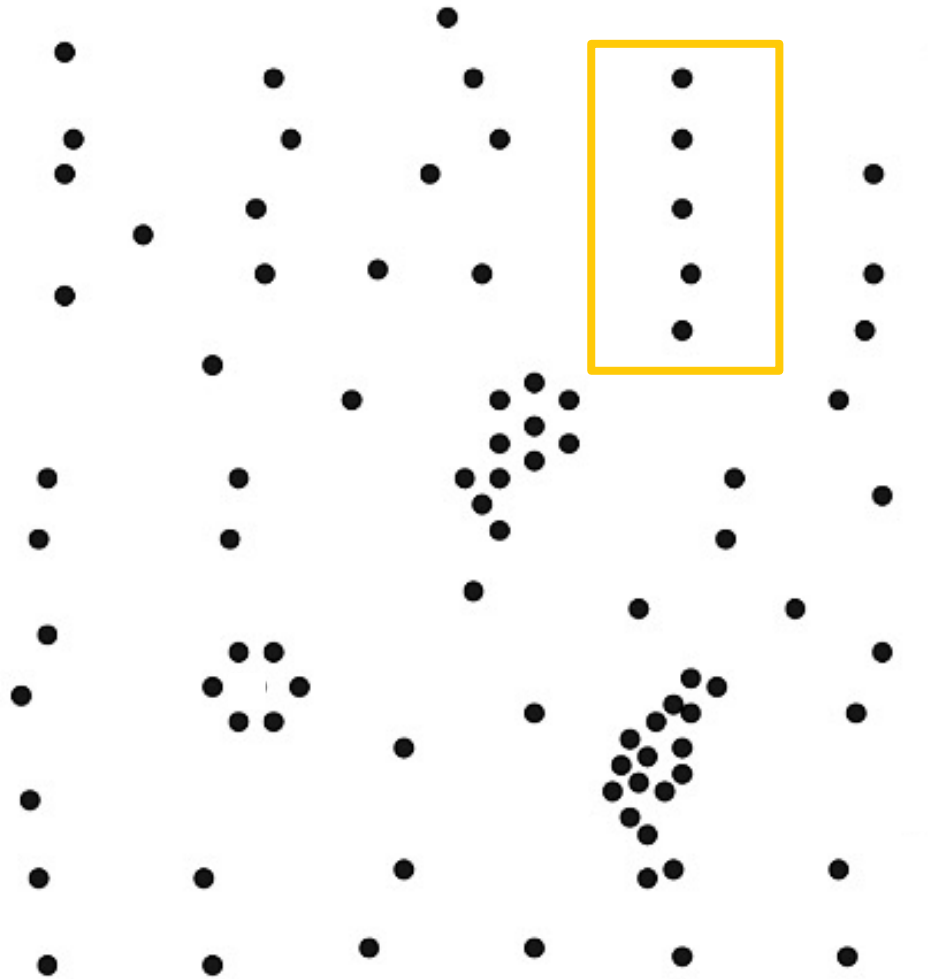
Session 3



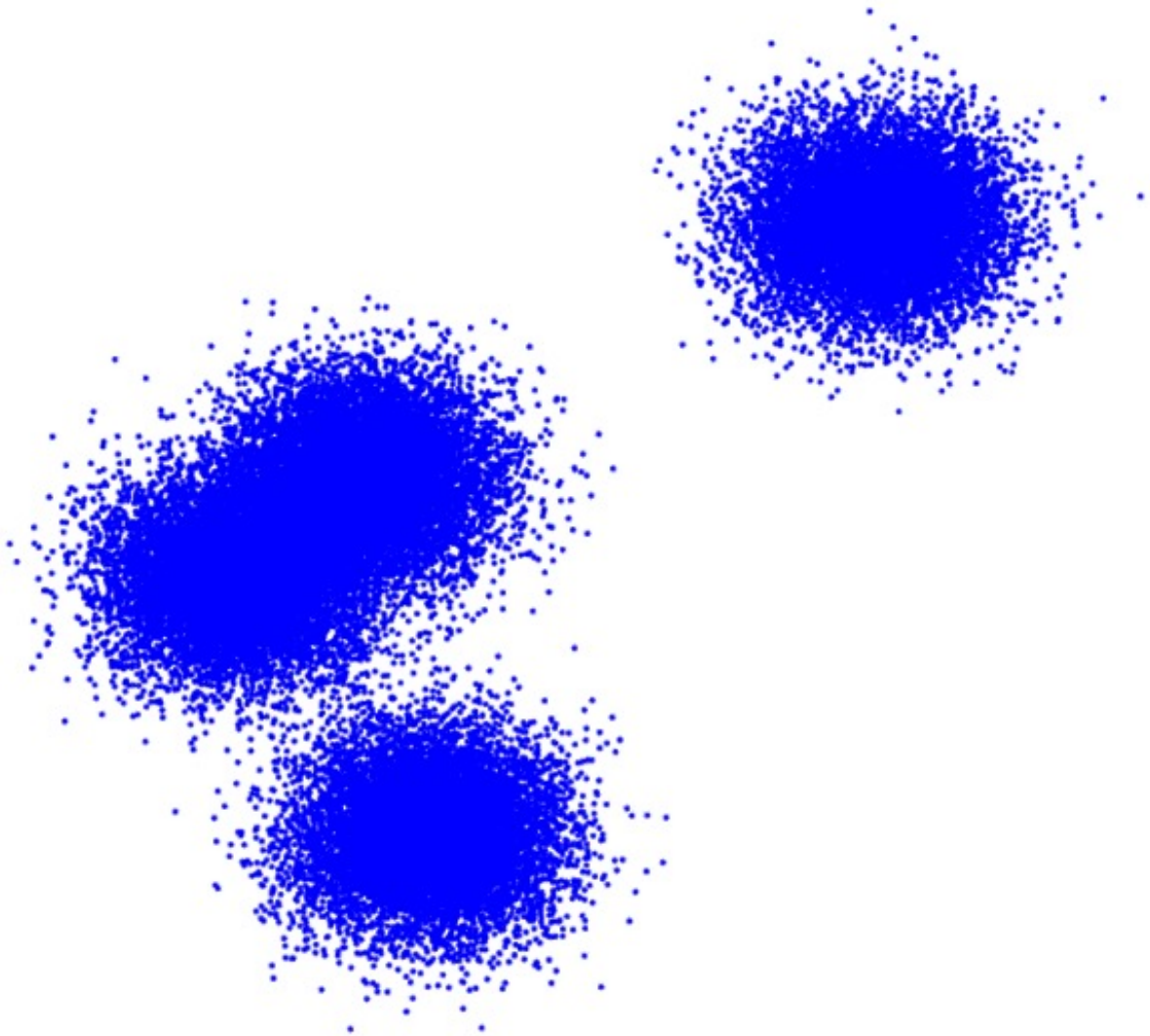
Session 3



Session 3



Session 3



Clustering Challenges

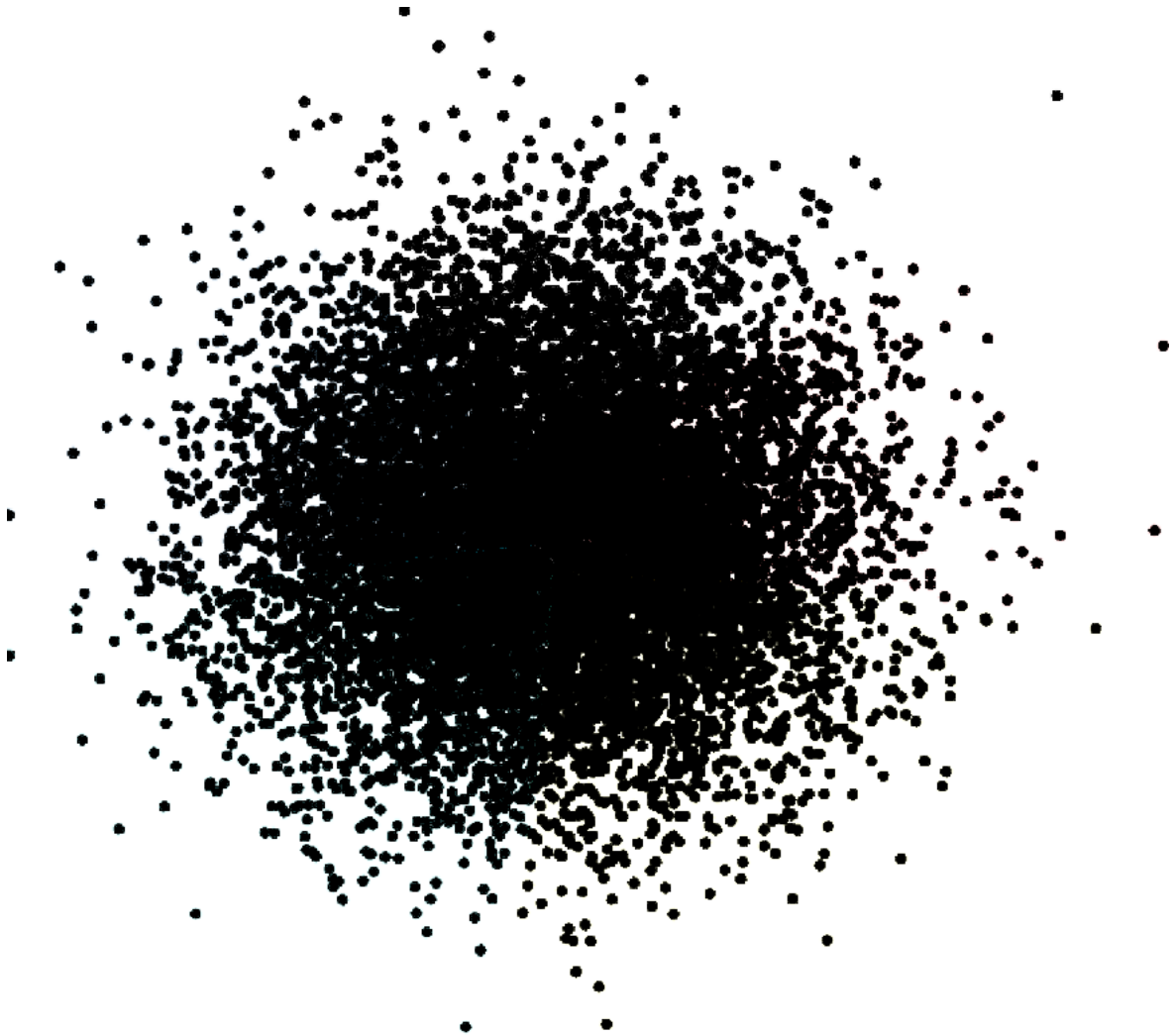
Cluster description
should clusters be described using representative instances or average values?

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

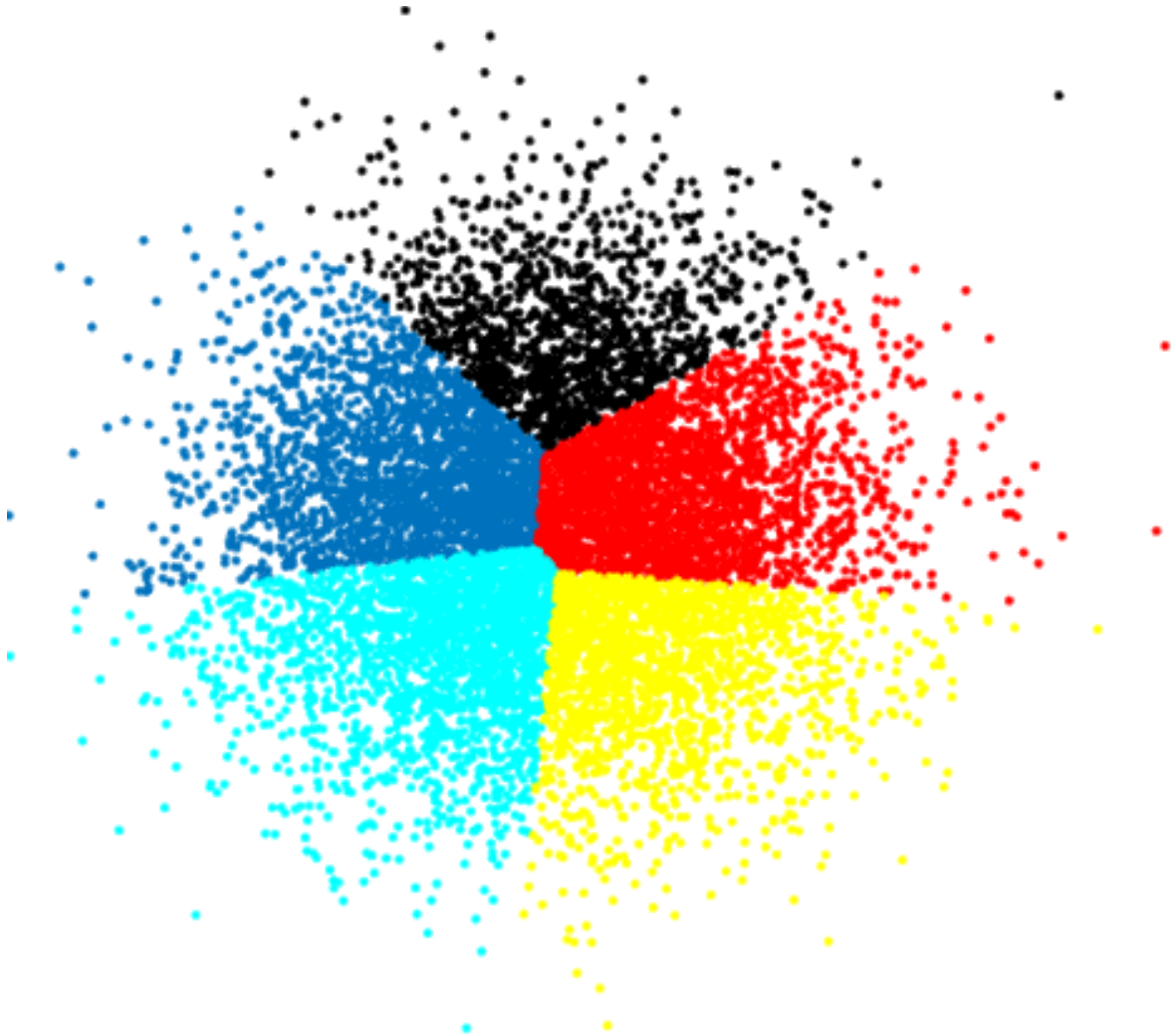
Ghost clustering
most methods will find clusters even if there are none in the data

***A posteriori* rationalization**
once clusters have been found, it is tempting to try to "explain" them ...

Session 3



Session 3



Suggested Reading

Validation and Notes

Data Understanding, Data Analysis, Data Science
Machine Learning 101

Clustering

- Clustering Validation

Spotlight on Clustering

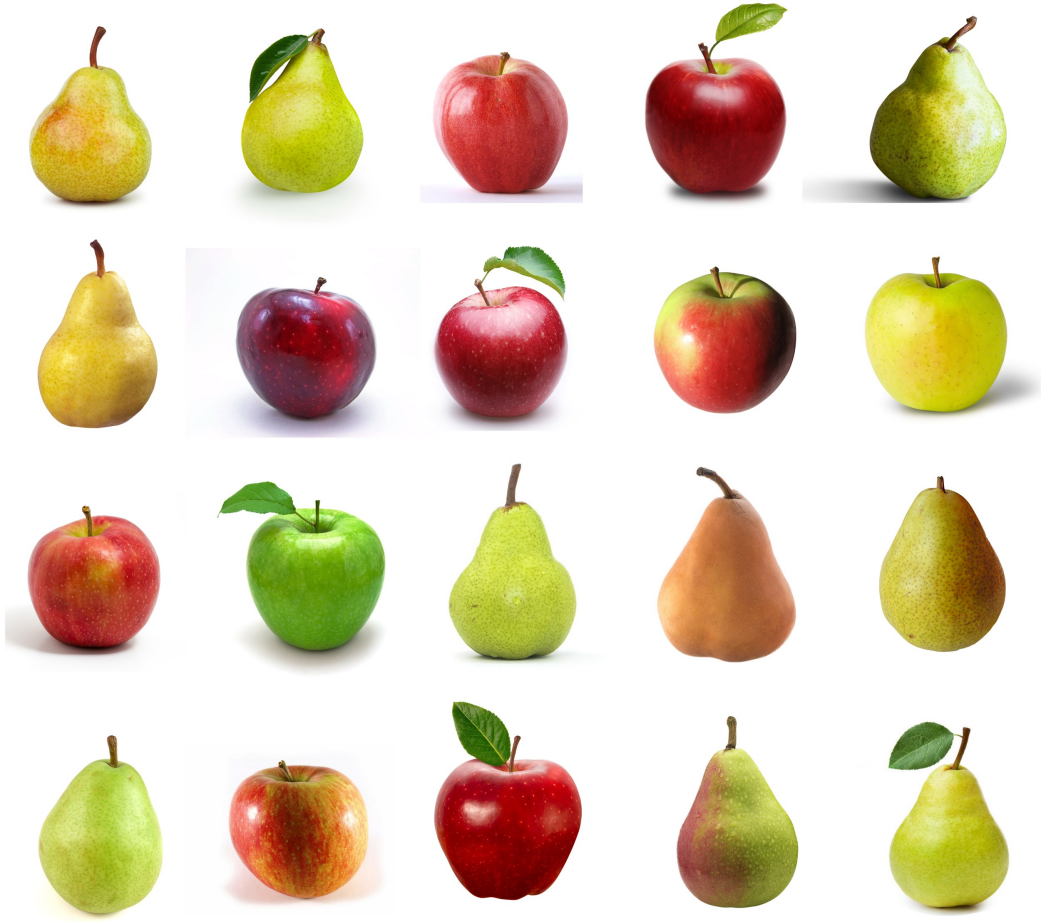
*Clustering Evaluation (advanced)

- Clustering Assessment
- Model Selection

Exercises

Validation and Notes

Consider the fruit image dataset below.



Provide a few clustering schemes for the data, and discuss how you would validate them.

Session 4

INTRODUCTION TO MACHINE LEARNING

Issues and Challenges

INTRODUCTION TO MACHINE LEARNING

We all *say* we like data, but we don't. We like getting insight out of data. That's not quite the same as liking data itself. In fact, I dare say that I don't care for data, and it sounds like I'm not alone. [Q.E. McCallum, *Bad Data Handbook*]

Data, big or small, is only as useful as the questions you ask of it. [M. Jones, P. Silberzahn]

Nothing is always absolutely so. [Sturgeon's First Law]

95% of everything is crud. [Sturgeon's Maxim]

It can be tempting to use data as a crutch in decision-making: “The data says so!” But **sometimes the data lets us down** and that exciting correlation you found is just a by-product of a messy, biased sample. [...] Smart skeptics can help step back, reflect, and ask if **what the data is saying actually fits** with what you know and expect about the world.

[Nicholas Diakopoulos, Harvard Business Review]

Session 4

Dear News Media,

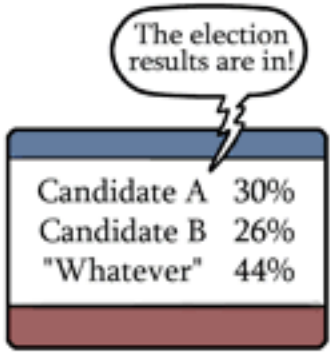
When reporting poll results, please keep in mind the following suggestions:

1. If two poll numbers differ by less than the margin of error, it's not a news story.

2. Scientific facts are not determined by public opinion polls.

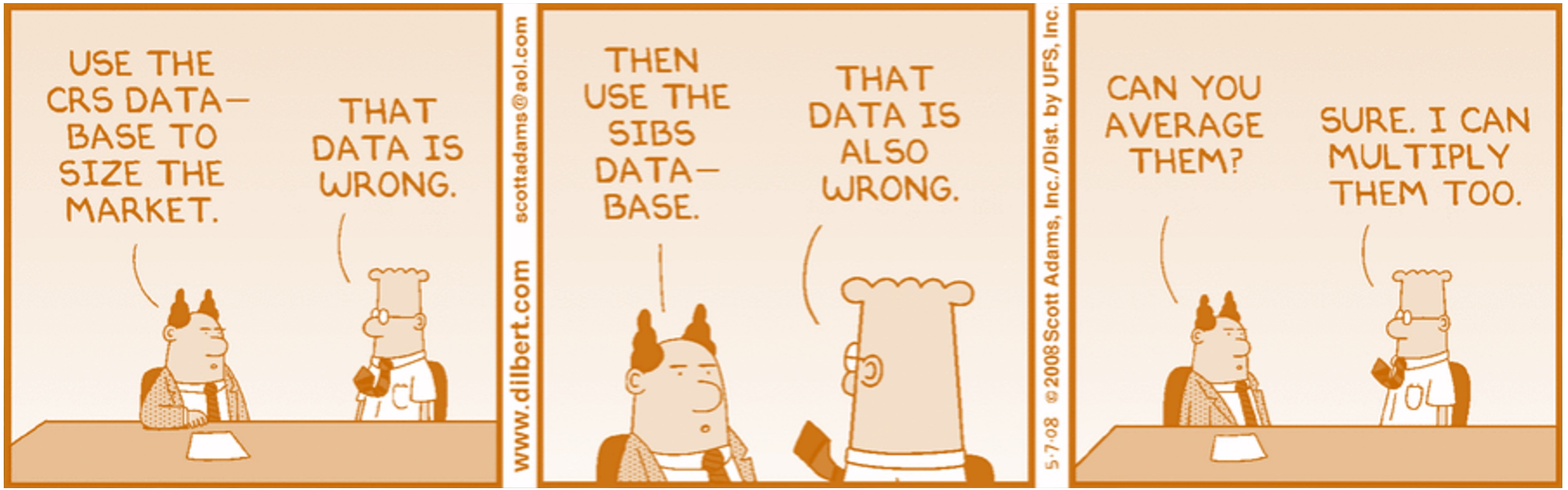
3. A poll taken of your viewers/internet users is not a scientific poll.

4. What if all polls included the option "Don't care"?



Signed,

-Someone who took a basic statistics course.



10. Bad Data and Big Data

Bad Data

Does the dataset pass the **smell test**?

- invalid entries, anomalous observations, etc.

Data formatted for human consumption, not machine readability

Difficulties with **text processing**

- encoding
- application-specific characters

Bad Data

Collecting data **online**

- legality of obtaining data
- storing offline versions

Detecting **lies** and **mistakes**

- reporting errors (lies or mistakes)
- use of polarizing language

Data and reality

- bad data
- bad reality?

Bad Data

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

- imputation bias
- top/bottom coding (replacing extreme values with average values)
- proxy reporting (head of household for household)

Seeking **perfection**

- academic data
- professional data
- government data
- service data

Bad Data

Data science **pitfalls**

- analysis without understanding
- using only one tool (by choice or by fiat)
- analysis for the sake of analysis
- unrealistic expectations of data science
- it's on a need-to-know basis and you don't need to know

Databases vs. files vs. cloud computing

- the cloud will solve all of our problems!

Bad Data

When is **close enough, good enough**?

- completeness
- coherence
- correctness
- accountability

Big Data – A Word of Warning

Big Data is no crystal ball

- "Past performance does not guarantee future results"

Big Data can't dictate personal or organizational values

- The right value answer may be the wrong data science answer
- Data-based conclusions do not live in a vacuum: context matters
- Blind obedience to data-driven results is just as dangerous as rejection based on gut-reaction

Big Data can't solve every problem

- "When all you have is a hammer, everything looks like a nail"

Big Data vs. Small Data

What is the main difference?

- the datasets are **LARGE**
- issues: collection, capture, access, storage, analysis, visualization

Where does the data come from?

- technology advances are lifting the limits on data processing speeds
- information-sensing, mobile devices, cameras and wireless networks

What are the challenges?

- most techniques were built for very small dataset
- direct approach will leave the best analyst waiting years for results

6. **variability**
7. **visualization**

The 5V(7V?) Paradigm

1. **volume:** large amounts of data
2. **velocity:** speed at which data is created, accessed, processed
3. **variety:** different types of available data, can't all be saved in relational databases (tables, pictures,...)
4. **veracity:** quality and accuracy of big data is harder to control
5. **value:** turn the data into something useful

The Big Data Problem

Many computations happen **instantly**, others take a **significant** amount of time.

Crunching very large datasets is a perfect example. Analysis in *R* or *Python* with steadily increasing datasets leads to computer lags. Eventually, the time required becomes **impractically long**.

Optimizing code and using a faster CPU can only provide so much relief.

That is the **Big Data problem**.

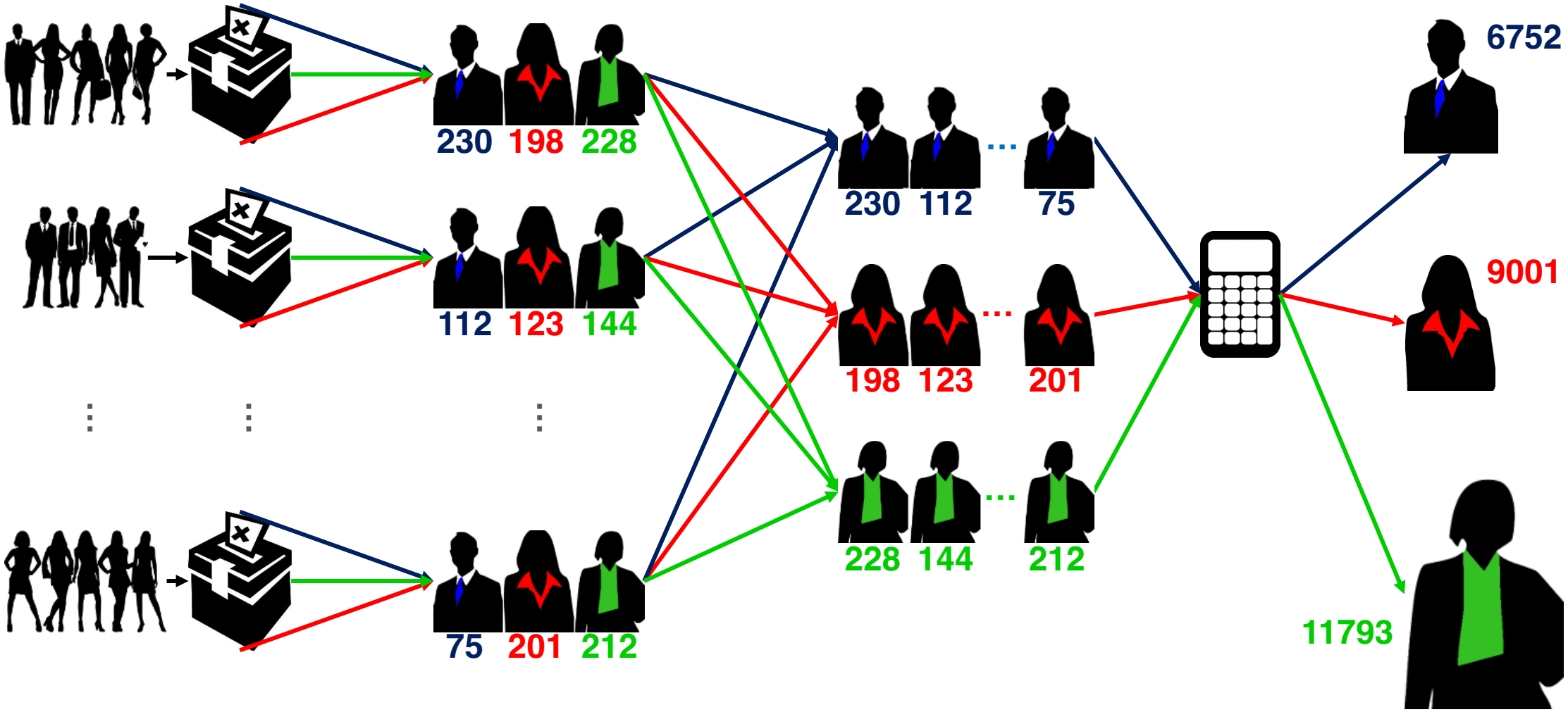
Distributed Computing

Splitting the computations among multiple CPU cores/CPUs can divide the computation time by a factor of 4, or 32, or 1000, or ... This allows algorithms to run on big data to keep analytics, smart services, and recommendations updated **daily**, **hourly**, in **real time**.

Election analogy to parallelization:

- counting votes at different polling stations in a riding
- each station simultaneously counts its own votes and reports their total
- the totals of all polling stations are aggregated at Elections HQ
- one person counting all the ballots would eventually get the same result, but it would take *too long* to get the result.

Analogy: Elections



Analogy: Pizzeria

Parallelism gains 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

Good News

Most practical computational tasks can be and are parallelized.

Modern data scientists use frameworks where distributed computing are already implemented (Apache Spark implements *MapReduce*, for instance).

Take some time to think about this potential issue **before** the start of the data collection/data analysis process – it will save headaches in the long run.

Suggested Reading

Bad Data and Big Data

J. Leskovec, A. Rajamaran, and J. D. Ullman, *Mining of Massive Datasets*. Cambridge Press, 2014.

Data Understanding, Data Analysis, Data Science
Machine Learning 101

Issues and Challenges

- Bad Data

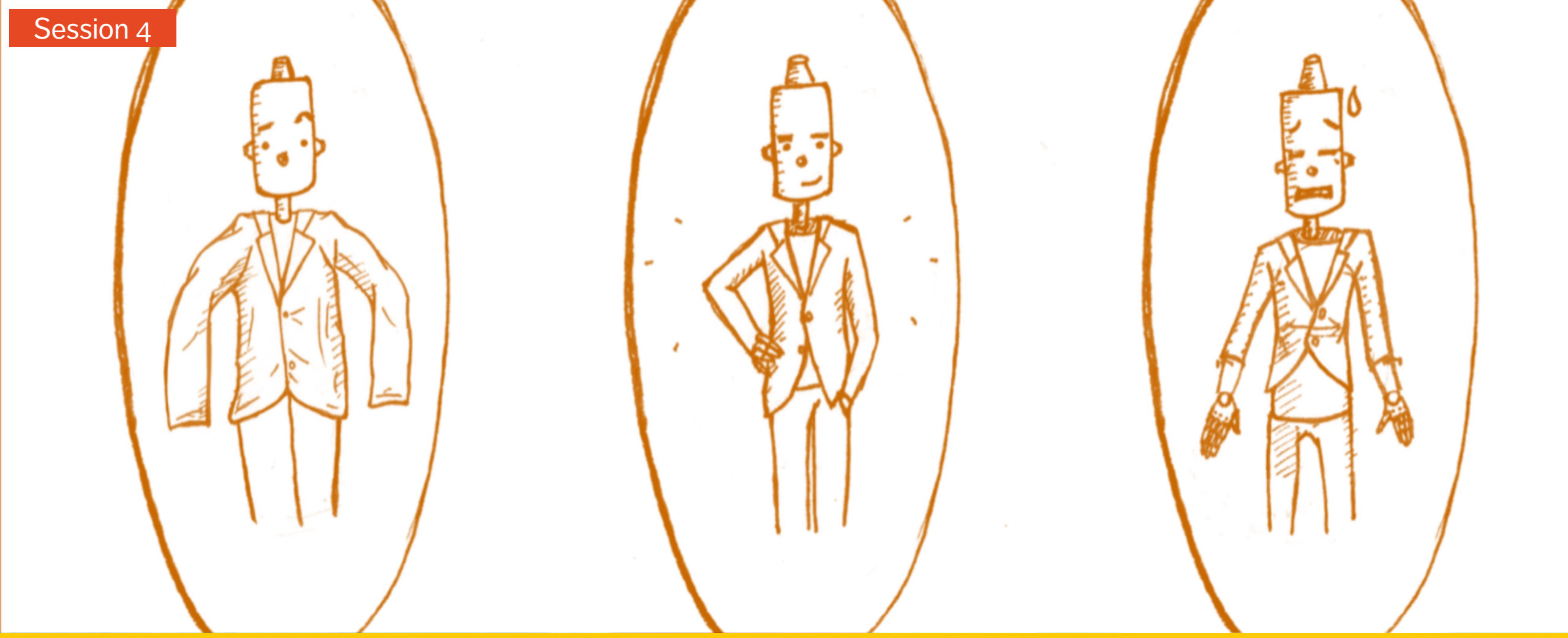
Exercises

Bad Data and Big Data

1. As the saying goes, “garbage in, garbage out”. What are the analytical, business, and public policy consequences of making decisions based on bad data?

2. Whether a dataset is considered small or “big” depends not only on the dataset, but also on the available tools.

   Generate increasingly larger random datasets (3 variables + 1 class) to cluster with `kmeans()` and classify with `rpart()`. Keep track of the runtime. How does the runtime vary with the number of observations? At what size do you predict that the algorithms will be too slow and cumbersome for your needs?



11. Underfitting and Overfitting/Transferability

Fundamentals

Rules or models generated by any technique on a **training set** have to be generalizable to **new data** (or **validation/ testing sets**) to be useful.

Problems arise when knowledge that is gained from **supervised learning** does not generalize properly to the data.

Unsupervised learning can also be affected.

Ironically, this may occur if the rules or models fit the training set **too well** – the results are **too specific to the training set**.

Example of Rules

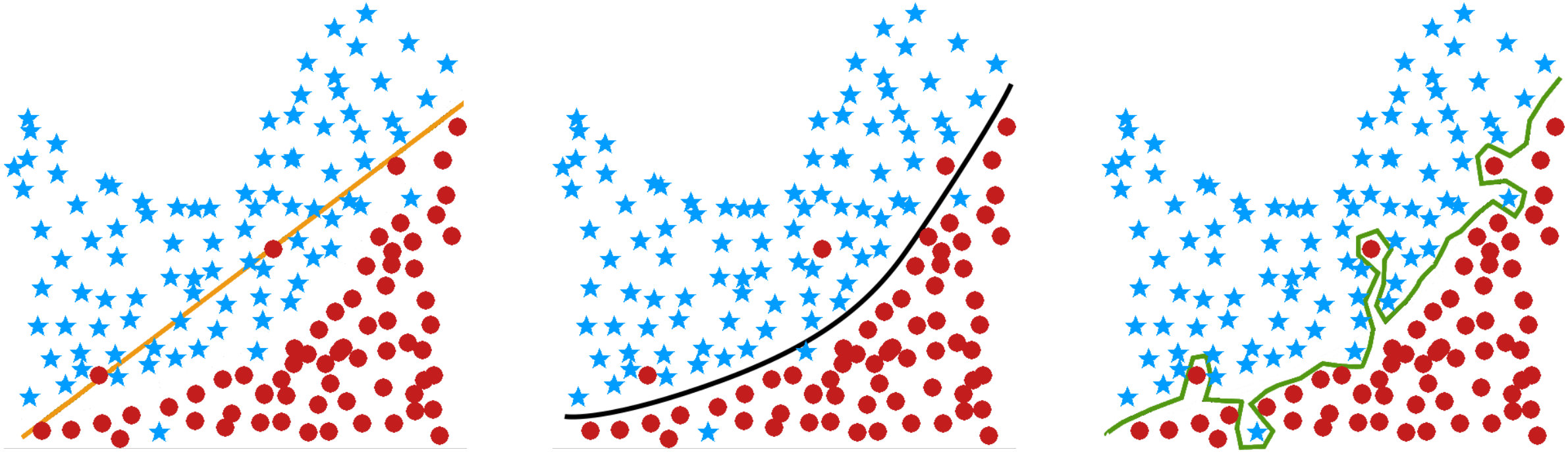
Rule I: based on a survey of 400 Germans, we infer that 43.75% of the world's population has black hair, 37.5% have brown hair, 9% have blond hair, 0.25% have red hair, and 9.5% grey hair.

Rule II: humans' hair colour is either black, brown, blond, red, or grey.

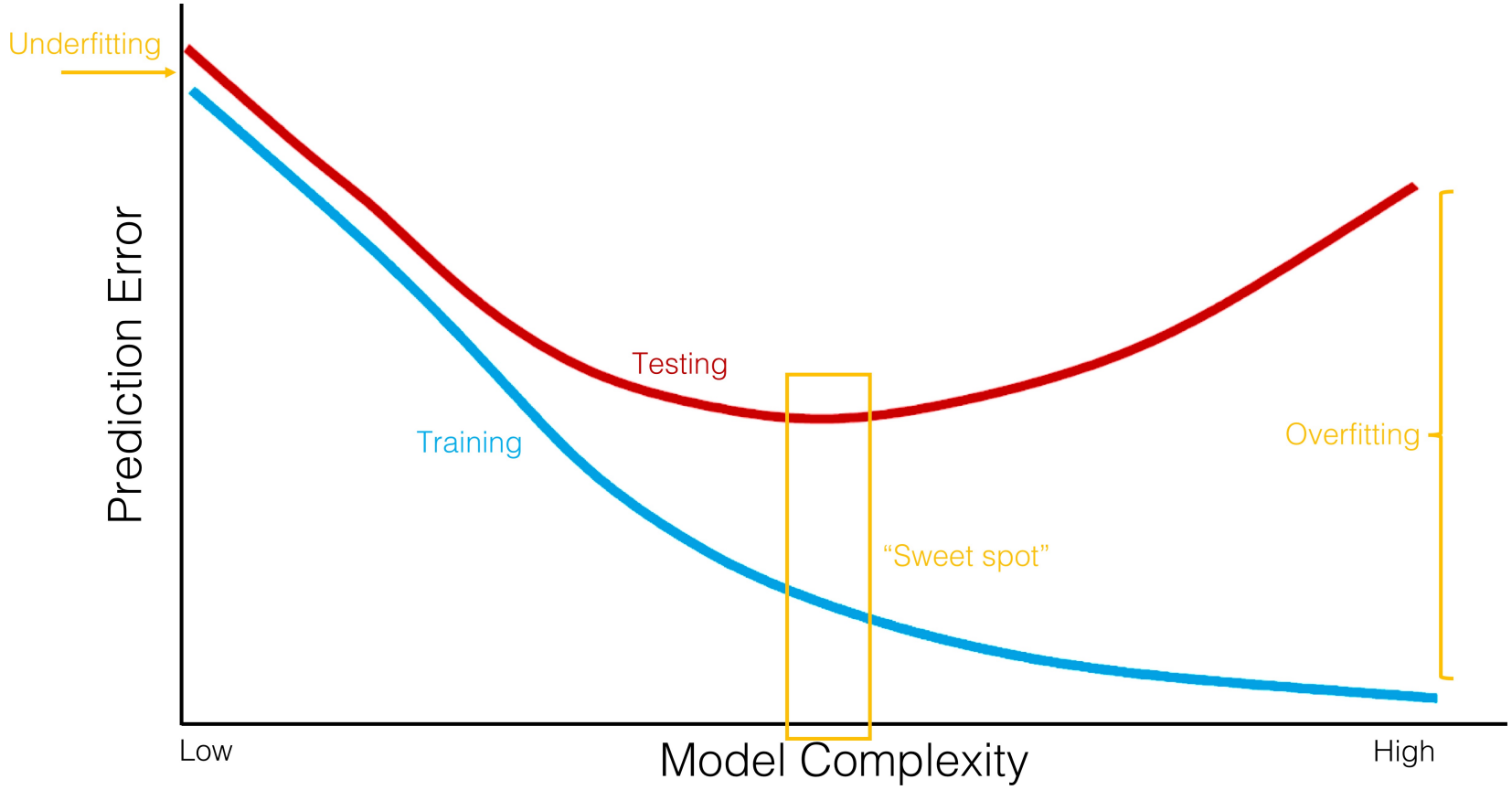
Rule III: approx. 40% of humans have black hair, 40% have brown hair, 5% blond, 2% red and 13% grey.

Which of the 3 rules is most useful? The most vague? Which is overly specific?

Goldilocks and the Three Models



Bias-Variance Trade-Off



ALWAYS evaluate models on unseen (testing) data!

Bias-Variance Trade-Off

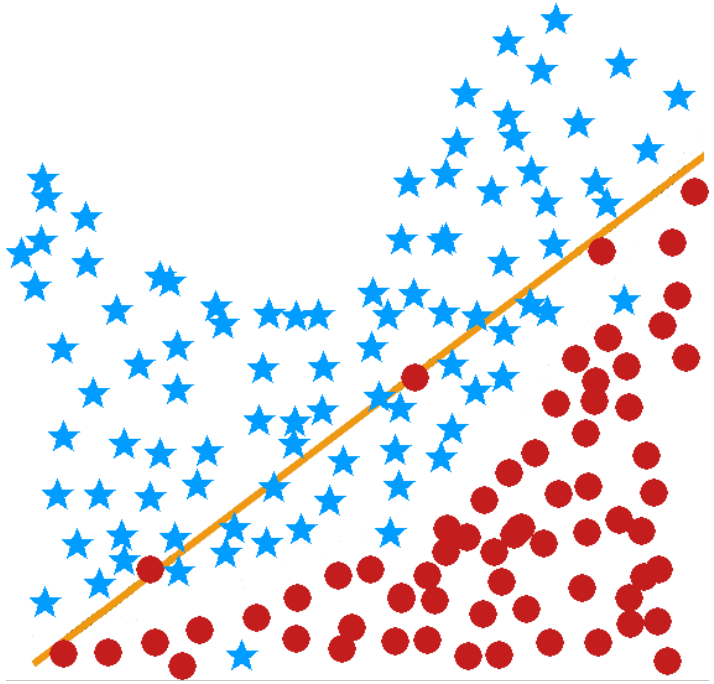
We **build** a model on **historical data** and **evaluate** its performance on **new data**.

Let $\text{Error}_{\text{Test}}$ be the model's performance on test data:

$$\text{Error}_{\text{Test}} = \text{Bias}^2_{\text{Model}} + \text{Variance}_{\text{Model}}$$

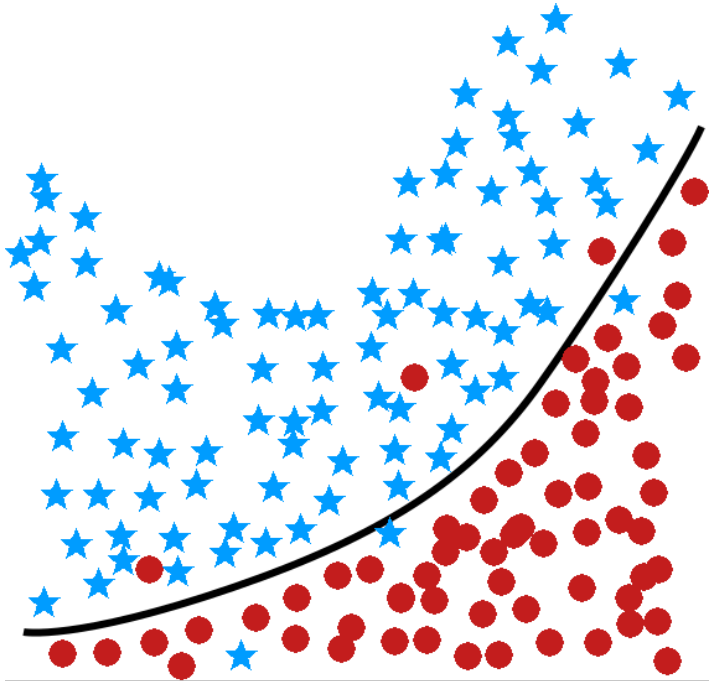
The **bias** measures the model's prediction **accuracy**; the **variance**, its **sensitivity** to (small) changes in the data.

Goldilocks and the Three Models

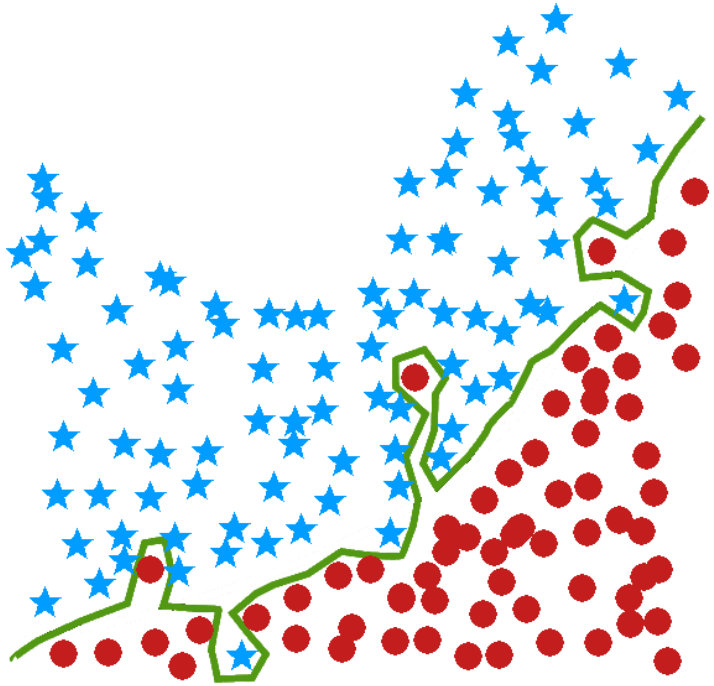


underfit
bias = high
variance = low
error = **high**

predictions are not very accurate



just right
bias = medium
variance = medium
error = **medium**



overfit
bias = low
variance = high
error = **high**

model is too specific to the data

Possible Solutions

Underfitting can be overcome by considering models that are more complex.

Overfitting can be overcome in several ways:

- **using multiple training sets**
  overlap is allowed (or not: see cross-validation)
- **using larger training sets**
  70% - 30% split is suggested
- **optimizing the data instead of the model**
  models are only as good as the data

Recommended Procedures

Small datasets (less than a few hundred observations)

- use 100-200 repetitions of a **bootstrap** procedure

Average-sized datasets (less than a few thousand observations)

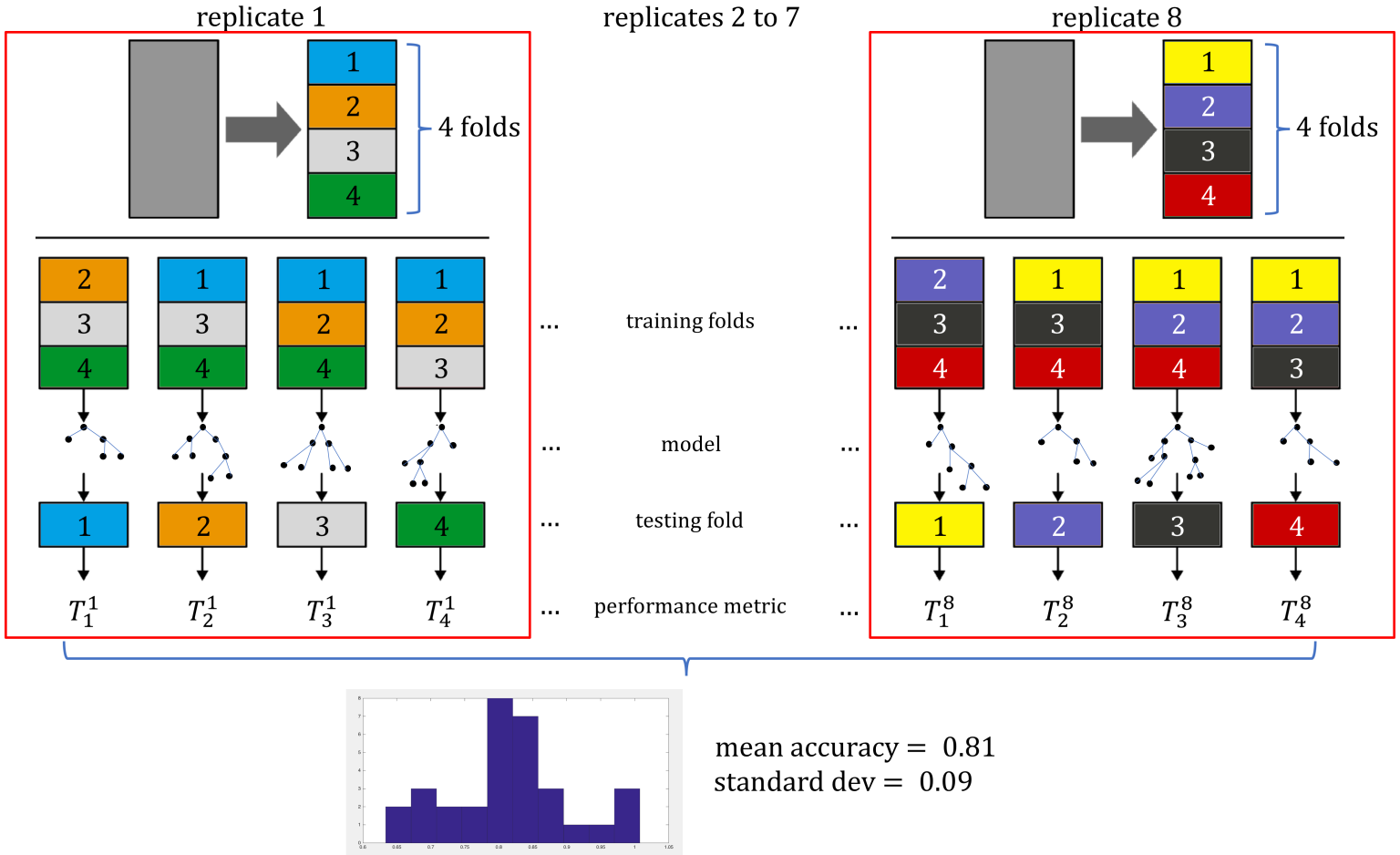
- use a few repetitions of 10-fold **cross-validation** on the training set (see next slide)

Large datasets

- use a few repetitions of **holdout** (70%-30%) split

The **decision boundaries** depend on computing power and the number of tasks in the workflows.

Cross-Validation



mean accuracy = 0.81
standard dev = 0.09

Appropriateness and Transferability

Data science and machine learning models will continue to be used heavily in the coming years.

We have discussed pros and cons of some of the applications on ethical and other non-technical grounds, but there are also **technical challenges**.

DS/ML methods are **not appropriate**:

- If you must absolutely use an existing (**legacy**) datasets instead of an **ideal** dataset ("it's the best data we have!")

Appropriateness and Transferability

DS/ML methods are **not appropriate** (cont.):

- if the dataset has attributes that usefully predict a value of interest, but which are **not available** when a prediction is required

  Example: the total time spent on a website may be predictive of a visitor's future purchases, but the prediction must be made before the total time spent on the website is known.

- if you attempt to predict **class membership** using an **unsupervised** learning algorithm

  Example: clustering loan default data might lead to a cluster that contains multiple defaulters. If new instances get added to this cluster, should they automatically be viewed as loan defaulters? (no)

Non-Transferable Assumptions

Every model makes certain assumptions about what is and is not **relevant** to its workings, but there is a tendency to only gather data which is **assumed** to be relevant to a particular situation.

If data is used in other contexts, or to make predictions depending on attributes without data, validating the results may prove impossible.

- **Example:** can we use a model that predicts mortgage defaulters to also predict car loan defaulters? A car is not a house: they play different roles in an individual's life; the values are of different magnitudes; and so on...
- That being said, is there truly no link between mortgage defaults and car loan defaults?

Suggested Reading

Underfitting and Overfitting/ Transferability

Data Understanding, Data Analysis, Data Science
Machine Learning 101

Issues and Challenges

- Overfitting/Underfitting
- Appropriateness and Transferability

Regression and Value Estimation

*Statistical Learning (advanced)

- Model Evaluation
- Bias-Variance Trade-Off

*Resampling Methods (advanced)

- Cross-Validation

*Model Selection (advanced)

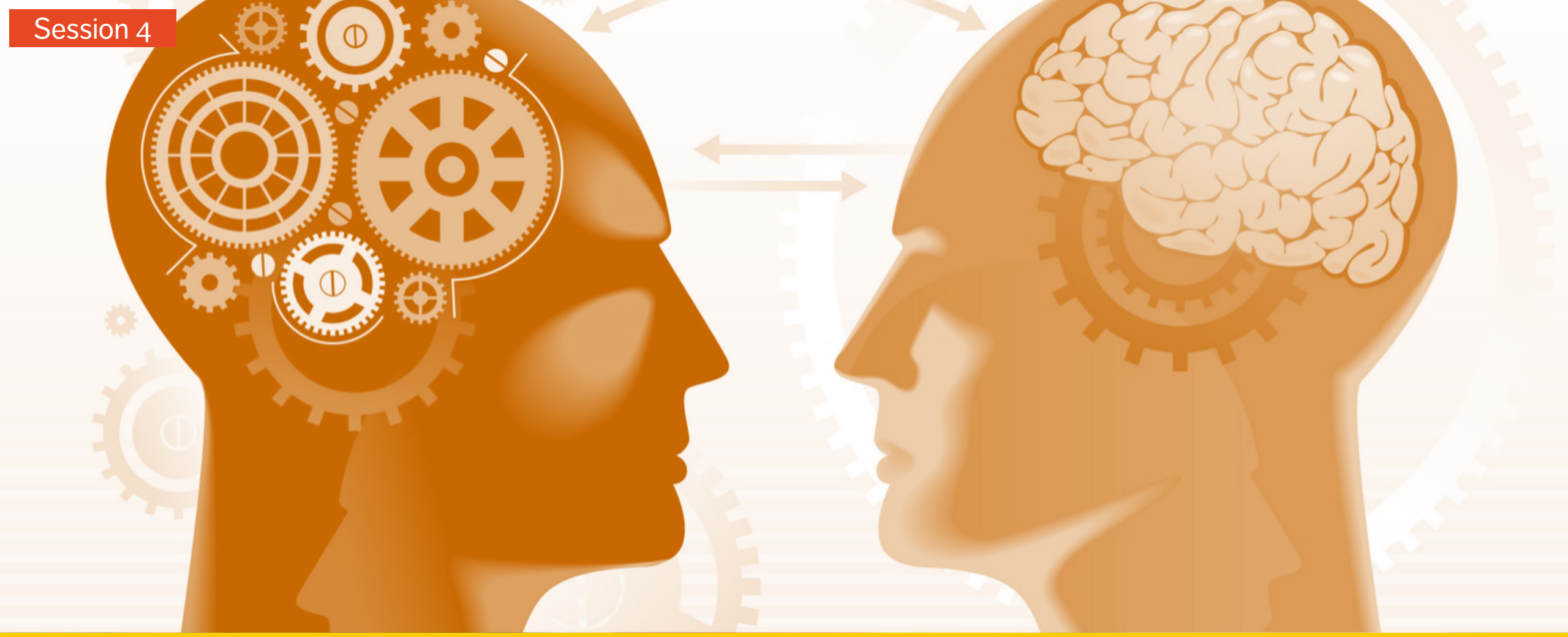
- Selecting the Optimal Model (Validation and Cross-Validation Reprise)

Exercises

Underfitting and Overfitting/ Transferability

1. This exercise illustrates overfitting/underfitting.
   a. Randomly generate $n = 150$ values in $[0,10]$ for the predictor x.
   b. Randomly generate $n = 150$ response values according to $y = 10 + x - 2x^2 + 17x^3 + \varepsilon$, where ε is a random error term of your choice.
   c. Fit a linear model, a quadratic model, a cubic model, and a polynomial model of degree 10 to the data.
   d. Add 3 observations to the data as in steps a. and b. Repeat step c. Do the models change much?
   e. Which model(s) would you trust to make predictions on new data?

2. Modify the Gapminder example from Cross-Validation to select a model in the previous question.



12. Miscellanea

Biases, Fallacies, and Interpretation

When consulting (or conducting) studies, beware:

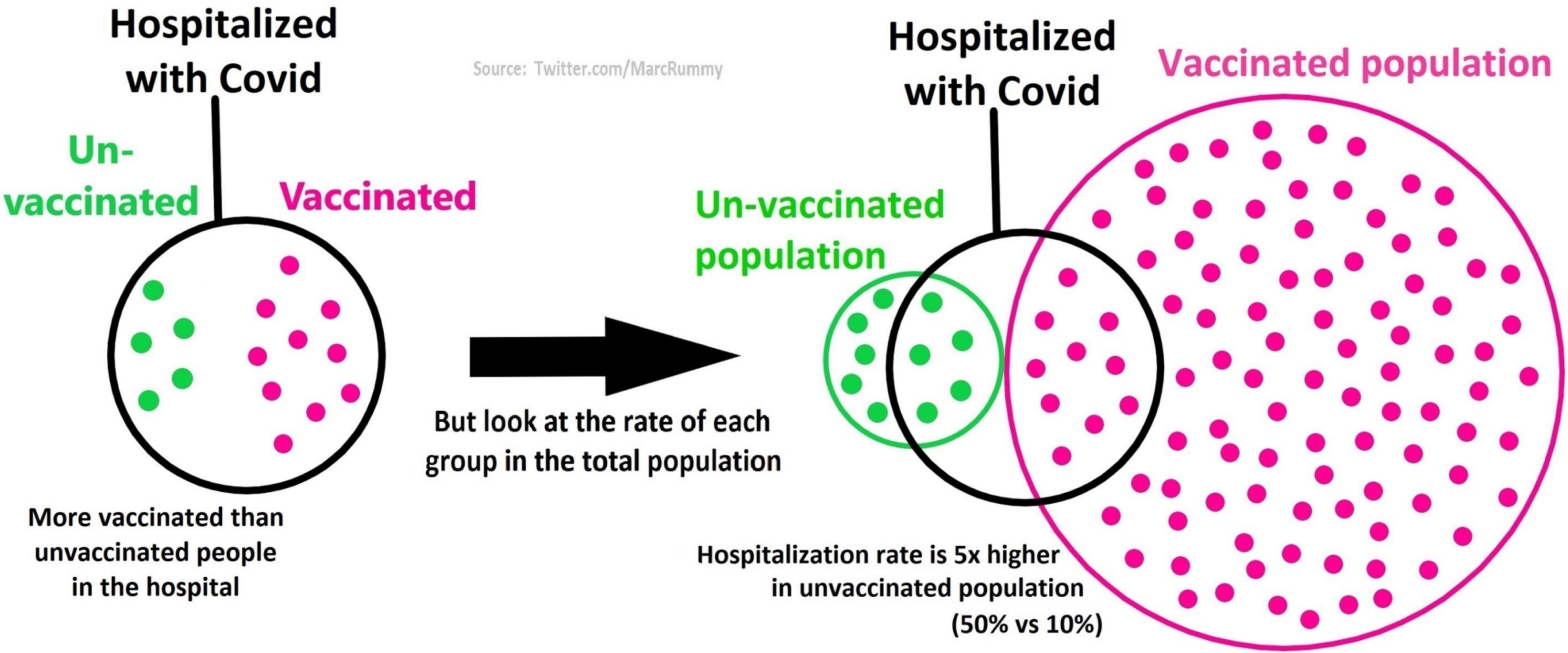
- **selection bias** (what data was included, how was it selected?)
- **omitted-variable bias** (were relevant variables ignored?)
- **detection bias** (did prior knowledge affect the results?)
- **funding bias** (who's paying for this?)
- **publication bias** (what's not being published?)
- **data-snooping bias** (trying too hard?)
- **analytical bias** (did the choice of specific method affect the results?)
- **exclusion bias** (are specific observations/units being excluded?)

But: does the presence of bias necessarily invalidate the results?

Biases, Fallacies, and Interpretation

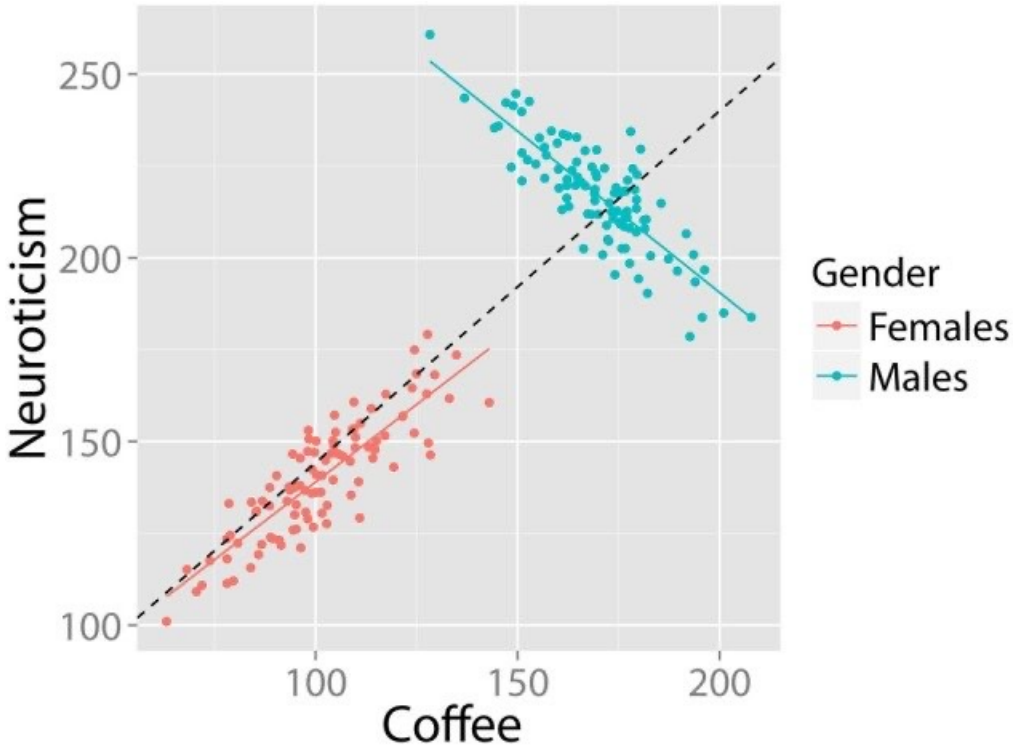
Remember:

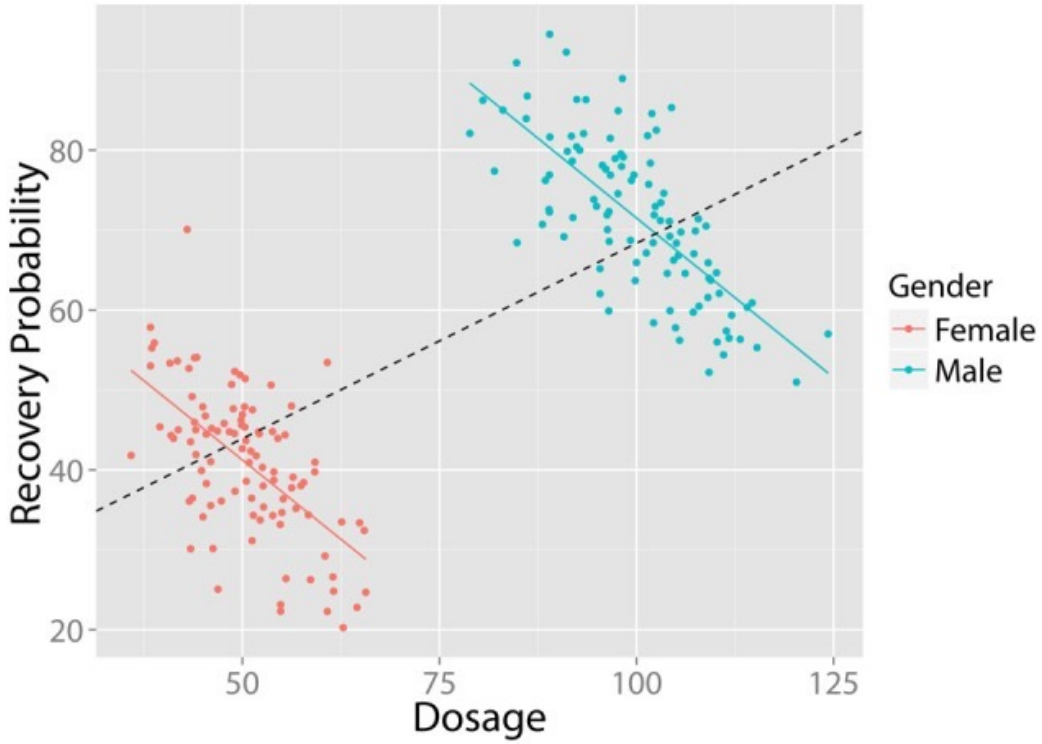
- correlation is not causation (but it is a hint!)
- extreme patterns can mislead
- stay within a study's range
- keep the base rate in mind
- counter-intuitive results are not always wrong (Simpson's Paradox, Benford's Law, etc.)
- randomness plays a role
- there is a human component to any analytical activity
- small effects can still be (statistically) significant
- beware of sacrosanct statistics (p-value, etc.)



Note: The ratios presented are made to illustrate the concept of the base rate fallacy when the vaccination rate is high

[Kievit et al., *Simpson's Paradox in Psychological Science: A Practical Guide*]





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

DS/ML Myths and Mistakes

Myths:

- DS/ML is about algorithms
- DS/ML is about predictive accuracy
- DS/ML requires data warehouses and fancy infrastructure
- DS/ML requires a large quantity of data
- DS/ML requires technical experts (?)

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

DS/ML Myths and Mistakes

Mistakes:

- selecting the wrong problem
- getting buried under tons of data without metadata understanding
- not planning the data analysis process
- having insufficient business and domain knowledge
- using incompatible data analysis tools
- using tools that are too specific
- ignoring individual predictions/records in favour of aggregated results
- running out of time
- measuring results differently than the sponsor/stakeholders
- naïvely believing what one's told about the data

The Future of DS/ML/AI

What we didn't talk about:

- tons of classification and clustering algorithms
- recommender systems
- data streams
- bayesian data analysis
- natural language processing and text mining
- feature selection and dimension reduction (curse of dimensionality)
- data engineering
- ... and much, much more!

The Future of DS/ML/AI

Future tasks:

- 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.

The Future of DS/ML/AI

Future trends:

- new questions
- new tools
- new data sources
- data science as job component
- augmented/swarm intelligence

In Conclusion

DS/ML is a team activity.

Ethical considerations are crucial.

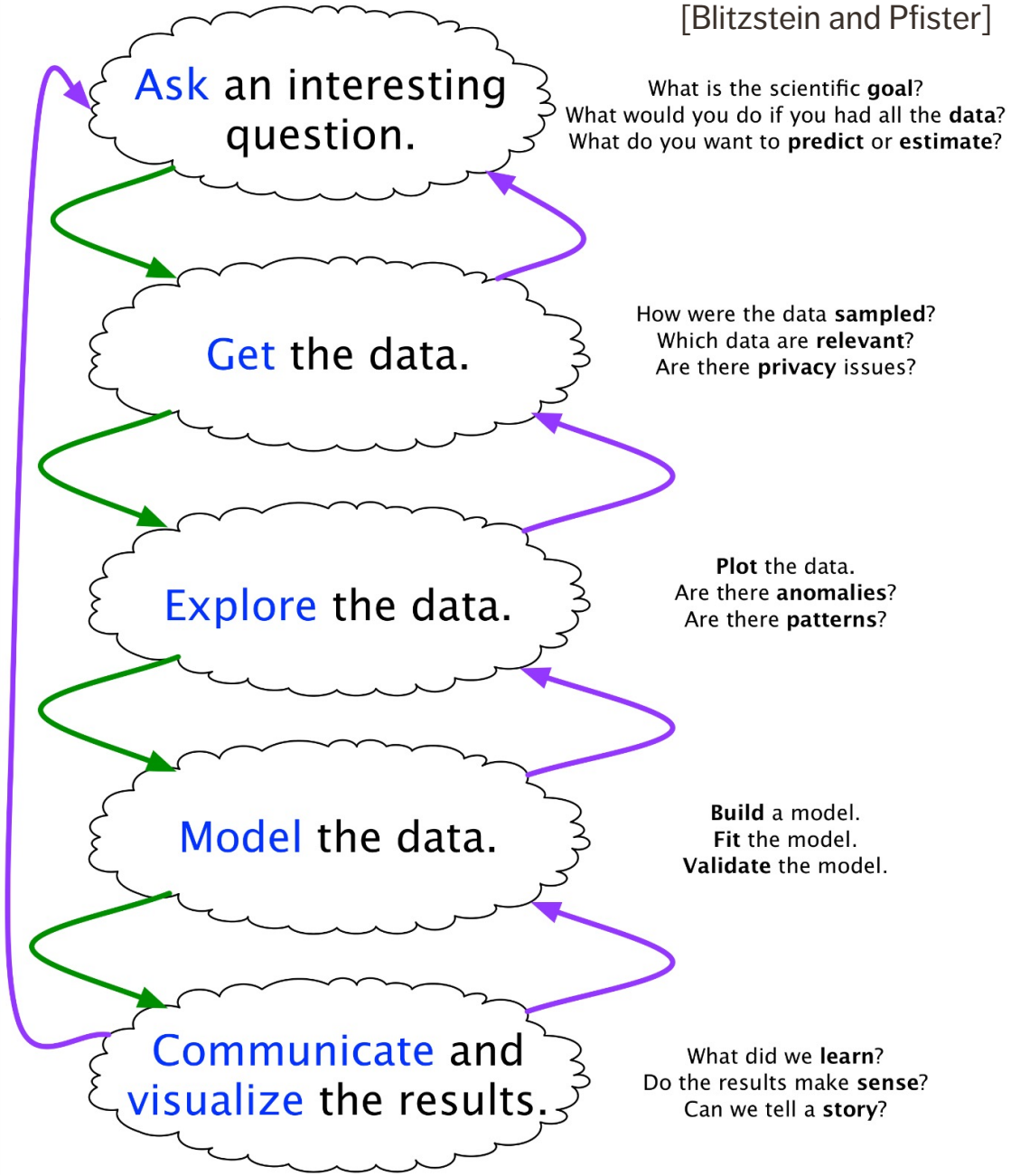
Let the data speak.

Look for actionable insights.

Supervised vs. unsupervised.

Be ready to clean, prepare, & visualize data.

[Blitzstein and Pfister]



Exercises

Miscellanea

1. What is your preferred approach: “tried, tested, and true” or “disruptive data science”? What would it take for you to consider the other side of the coin?

2. True or False?
   a. The predictive performance of a supervised model is evaluated on the training set.
   b. Cross-validation can be used to reduce the risk of overfitting a predictive model.
   c. It is always better to use as many variables as possible in a model.
   d. If observations with missing values are deleted, this may lead to bias and errors.
   e. We can use a clustering algorithm to predict class membership.

Exercises

Miscellanea

2. True or False? (cont.)
   f. If all methods don't yield the same result, it is a proof that the question cannot be answered.
   g. Business and domain knowledge is only necessary when working with old data.
   h. Sponsors and clients need to know all analytical details.
   i. It's impossible to plan the data analysis process before we know what the data looks like.
   j. The available data is not always appropriate/representative of the situation we are modeling.

3. In what ways can you see DS/ML becoming a crucial part of your work? Is this development welcomed? How do you want to be involved?

Suggested Exercises and Guided Projects

INTRODUCTION TO MACHINE LEARNING

Between Sessions

Session 1 to Session 2

- complete the exercises of session 1
- download the datasets from the website
- read Programming Primer
- install R / RStudio (Posit)
- install the following R packages: dplyr, tidyverse, ggplot2, arules, arulesViz, rpart, rpart.plot, rattle, party, flexclust, e1071, psych

Session 2 to Session 3

- complete the exercises of session 2

Session 3 to Session 4

- complete the exercises of session 3

After Session 4

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

Guided Project I

This project uses the Gapminder Tools.

1. In the default configuration, we can identify some potential association rules. Using visual and ballpark estimates, evaluate the performance of the following rules:
   - Income > 8000 → Life Expectancy > 70
   - Income < 8000 AND Life Expectancy < 70 → World Region = Africa
2. Play around with various charts and identify/evaluate 5+ additional AR.
3. Identify groups of similar countries, in 2018 [validate your clusters using various charts]. Were they also similar in 1930? 1970? 2000?
4. In the default configuration, follow the trajectories of Finland, Sweden, Iceland, Norway, and Denmark between 1900 and 2018. Do the countries appear to follow similar trajectories? Are there outliers or anomalous trajectories?
5. Repeat step 4 for Brazil, Paraguay, Uruguay, Venezuela, Colombia, Peru, and Ecuador.
6. Based on your results in steps 4 and 5, would you expect the trajectory for Argentina to be more like those of the Nordic countries or those of the South American countries? Or perhaps neither? Is your answer the same over all time horizons?

Guided Project II

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

- GlobalCitiesPBI.csv
- 2016collisionsfinal.csv
- HR_2016_Census_simple.xlsx
- custdata.tsv

For your dataset(s):

1. Perform the appropriate data understanding, data preparation, data cleaning, and data exploration steps to allow you to determine if it is trustworthy and what it could be used for (see Guided Project IV [*Data Science Essentials*] and Guided Project III [*Data Visualization and Dashboards*].
2. Conduct an association rule mining analysis of the datasets, determining 10-20 strong association rules. Visualize them, validate them, and interpret their results.

Guided Project III

Consider the Algae Bloom Dataset (see this example). We try to build a model to predict the presence/absence of algae based on various chemical properties of river water. The data science motivation for such a model is simple: chemical monitoring is cheap and easy to automate, whereas biological analysis of samples is expensive and slow. Another reason is that analyzing the samples for harmful content does not provide a better understanding of algae drivers: it just tells us which samples contain algae.

1. Load the data and summarize/visualize it: you will be tasked with predicting the presence/absence of algae a1 and a2.
2. Clean the data and impute missing values, as needed.
3. Remove 20% of the observations and save them to a validation set.
4. Create a training/testing pair on the remaining 80% of the observations and train 2 decision trees to predict the presence/absence of algae a1 and a2, respectively. Evaluate the performance of each model. Which models performs best on your training/testing pair?

Guided Project III (cont.)

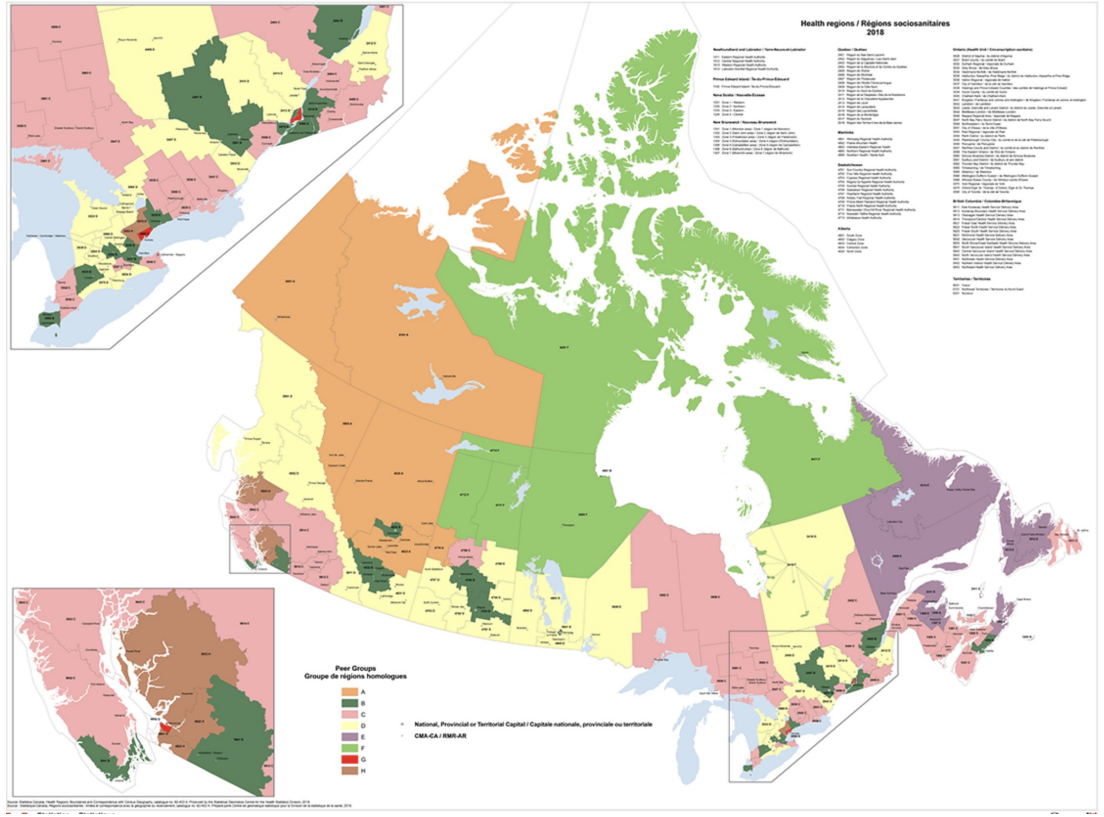
5. Repeat step 4 on at least 20 distinct training/testing pairs. Evaluate the performance of each model, and save them.
6. For each algae, pick the best of the models (how would you determine this) and use it to make predictions for the readings in the validation set. Evaluate these predictions.
7. Instead of picking the best of the 20+ models, find some way to combine the results of the 20 models and to make predictions for the readings in the validation set. Evaluate these predictions.
8. Which of the resulting models of steps 6 or 7 provide the best performance? Which are easier to interpret?
9. Use the same validation set as in step 3. In step 4, use the remaining 80% of the data to build a decision tree (do not split into a training/testing pair first). Use these models to make predictions for the readings in the validation set. Evaluate these predictions. Is there evidence of overfitting?
10. Use the same validation set in step 3. In steps 4 to 7, use decision stumps (decision trees with only 1 branching point) instead of full growth trees. Is there evidence of underfitting?
11. Conduct the analysis steps from 1 to 10 using other classification algorithms. Discuss the results.

Guided Project IV

The population of Canada is divided physically into provincial and territorial areas, most of which are further subdivided into health regions.

The Census information (from 2016) is available for those health regions. The equivalent 2018 dataset has been clustered to produce peer groups: the result is shown here (and on the right).

The data is in HR_2016_Census_simple.xlsx



Guided Project IV (cont.)

1. Load the data and summarize/visualize it (extract the rows with a 4-digit geocode).
2. Clean and scale the data.
3. Run k −means (with Euclidean distance) on the scaled data, using ALL the features, for reasonable value sof k. Use the Davies-Bouldin index and the Within-SS index to determine the optimal number of clusters. Is that clustering scheme plausible?
4. Reduce the dimension of the health region dataset by running a principal component analysis (PCA) and keep the principal components that explain up to 80% of the variability in the data. Repeat step 3. Are the results significantly different than they were?
5. Run k −means on the original health regions data (previous question) and on the reduced data, for the same range of k −values, but replicate the process 30+ times per value of k. What are the optimal k values in the aggregate runs?
6. Save the cluster assignments for each run with the optimal values of k. Two observations A and B have similarity $w(A, B) \in [0,1]$ if A and B lie in the same cluster in $w(A, B)\%$ of the runs. What are some observations with high similarity measurements? With low similarity measurements?

References

INTRODUCTION TO MACHINE LEARNING

References

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

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

D. Woods, "Bitly's Hilary Mason on "what is a data scientist?"," *Forbes*, Mar. 2012.

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

E. Garcia, C. Romero, S. Ventura, and T. Calders, "Drawbacks and solutions of applying association rule mining in learning management systems," 2007.

Wikipedia, "Association rule learning." 2020.

References

G. Piatetsky-Shapiro, “Discovery, analysis, and presentation of strong rules,” 1991.

C. C. Aggarwal and P. S. Yu, “A new framework for itemset generation,” in *Proceedings of the seventeenth ACM SIGACT-SIGMOD-SIGART symposium on principles of database systems*, 1998, pp. 18–24. doi: 10.1145/275487.275490.

P.-N. Tan, V. Kumar, and J. Srivastava, “Selecting the right objective measure for association analysis,” *Inf. Syst.*, vol. 29, no. 4, pp. 293–313, Jun. 2004, doi: 10.1016/S0306-4379(03)00072-3.

M. Hahsler and K. Hornik, “New probabilistic interest measures for association rules,” *CoRR*, vol. abs/0803.0966, 2008.

J. Leskovec, A. Rajamaran, and J. D. Ullman, *Mining of Massive Datasets*. Cambridge Press, 2014.

References

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

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

G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning: With Applications in R*. Springer, 2014.

E. Frank, I.H. Witten, *Data Mining: Practical Machine Learning Tools and Techniques* (2nd ed.), Elsevier, 2005.

References

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

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

Wikipedia, "Cluster analysis algorithms."

R. Yedida, "Evaluating clusters." Beginning with ML, 2019.

References

Q. E. McCallum, *Bad Data Handbook*. O'Reilly, 2013.

A. K. Maheshwari, *Business Intelligence and Data Mining*. Business Expert Press, 2015.

H. Kargupta, J. Han, P.S. Yu, R. Motwani, V. Kumar (eds), *Next Generation of Data Mining*, CRC/ Chapman & Hall, 2019.

N. Silver, *The Signal and the Noise: Why So Many Predictions Fail – But Some Don't*, Penguin Press, 2012.

M. Lewis, *Moneyball: The Art of Winning an Unfair Game*, Norton, 2003.

References

N. Diakopoulos, How Google Flu Trends Is Getting to the Bottom of Messy Data, HBR, July 2013.

R. Sarniak, 9 types of research bias and how to avoid them, Quirk's Media, Aug 2015.

Wikipedia entry for Bias.

Cochrane Handbook for Systematic Reviews of Interventions, Cochrane Methods.

Wikipedia, "Statistical bias".