Affaires mondiales Canada | Global Affairs Canada

Canada

CANADIAN FOREIGN SERVICE INSTITUTE | L'INSTITUT CANADIEN DU SERVICE EXTÉRIEUR

Introduction to Data Analysis

STATISTICAL LEARNING

Patrick Boily
Data Action Lab | uOttawa | Idlewyld Analytics

[with files from Jen Schellinck | Sysabee]

“Data science does not replace statistical modeling and data analysis; it augments them.”

(P. Boily)

LEARNING CONTEXT

STATISTICAL LEARNING

"We learn from failure,
not from success!"
(Bram Stoker, *Dracula*)

WHAT IS DATA SCIENCE? (REPRISE)

Data Science (DS) is the collection of processes by which we extract useful and **actionable insights** from data.

(paraphrased from T. Kwartler)

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

(paraphrased from H. Mason)

Data science is about **asking the right questions** and **accepting imaginative solutions**; in the battle between "tried, tested, and true" and "disruptive data science", with whom do you side?

(P. Boily)

THE MINING ANALOGY

What are we mining? data (*earth*)

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

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

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

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

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

LEARNING IN GENERAL

Beyond “just taking a quick look,” humans learn through:

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

The central Data Science/Machine Learning problem is:

can (should) we design algorithms that can learn?

TYPES OF LEARNING

Supervised Learning (learning with a teacher)

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

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

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

TYPES OF LEARNING

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

Reinforcement Learning (embarking on a Ph.D. with an advisor?)

In **supervised learning**, there's a target against which to train the model. In **unsupervised learning**, we don't know what the target is, or if there is one.

The distinction is **crucial**.

EXAMPLES

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) is a **supervised** learning task.

Making recommendations to customers concerning additional items to buy based on the buying pattern in their prior transactions (and prior transactions of other customers) is an **unsupervised** learning task.

ASSOCIATION RULES

STATISTICAL LEARNING

MR. SNIFF: What are you looking for?
MR. SNOOP: A five-dollar bill.
MR. SNIFF: Are you sure you lost it on this street?
MR. SNOOP: Oh no! I lost it in the next block, but I'm lookin' up here because the light is better.

(*Boys' Life Magazine*, 1932)

MOTIVATING EXAMPLE

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

Objectives:

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

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

METHODOLOGY

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

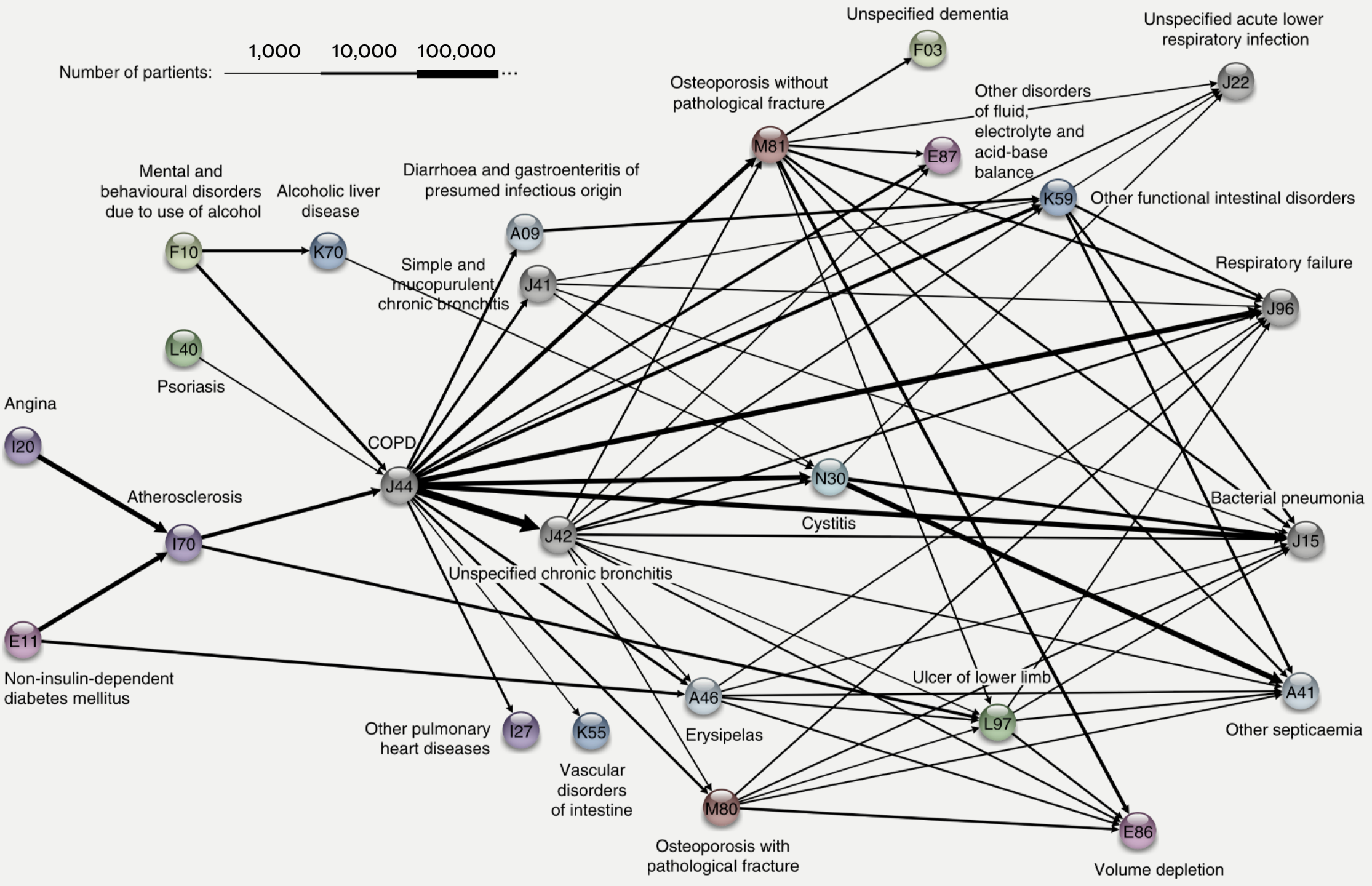
RESULTS

Data was reduced to 1,171 thoroughfares on the course of

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

The data analysis showed, for example:

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


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

TAKE-AWAYS

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

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

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

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

ASSOCIATION RULES BASICS

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

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

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

ORIGINAL APPLICATION

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

Examples:

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

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

APPLICATIONS

Related Concepts

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

Plagiarism

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

Bio-markers

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

APPLICATIONS

Making predictions and decisions based on these rules.

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

Use the connections to modify the likelihood of certain outcomes.

Imputing missing data.

Text autofill and autocorrect.

CAUSATION AND CORRELATION

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

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

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

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

CAUSATION AND CORRELATION

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

DEFINITIONS

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 **need not be true for all observations** in the dataset (i.e. rules are not necessarily 100% accurate).

Sometimes the "best" rules are those which are only accurate 10% of the time (as opposed to rules for which the accuracy is only 5% of the time).

As always, **it depends on the context**.

Technical challenge: coming up with a small set of reasonable rules.

DEFINITIONS

To determine a rule's strength, we compute rule metrics:

- **Support** (coverage) measures the frequency at which a rule occurs in a dataset. A low coverage value indicates that the rule rarely occurs (whether it is true or not).
- **Confidence** (accuracy) measures the reliability of the rule: how often does the conclusion occur in the data given that the premises have occurred. Rules with high confidence are "truer".
- **Interest** measures the difference between a rule's confidence and the relative frequency of its conclusion. Rules with high absolute interest are... well, more interesting.
- **Lift** measures the increase in the frequency of the conclusion due to the premises. In a rule with a high lift (> 1), the conclusion occurs more frequently than it would if it was independent of the premises.

FORMULAS

If N is the number of observations in the dataset:

- $\text{Support}(X \to Y) = \frac{\text{Freq}(X \cap Y)}{N} \in [0,1]$ ⟸ Proportion of instances where the premise and the conclusion occur together
- $\text{Confidence}(X \to Y) = P(Y|X) = \frac{\text{Freq}(X \cap Y)}{\text{Freq}(X)} \in [0,1]$ ⟸ Proportion of instances where the conclusion occurs when the premise occurs
- $\text{Interest}(X \to Y) = \text{Confidence}(X \to Y) - \frac{\text{Freq}(Y)}{N} \in [-1,1]$
- $\text{Lift}(X \to Y) = \frac{N^2 \cdot \text{Support}\,(X \to Y)}{\text{Freq}(X) \cdot \text{Freq}(Y)} \in (0, N^2]$

⟸ … ?!?

EXAMPLE

Music dataset containing data for $N = 15{,}356$ music lovers.

Candidate Rule (RM): "If an individual is born before 1976 (X), then they own a copy of at least one Beatles album, in some format (Y)".

Let's assume that

- Freq$(X) = 3888$ individuals were born before 1976
- Freq$(Y) = 9092$ individuals have a copy of at least one Beatles album
- Freq$(X \cap Y) = 2720$ individuals were born before 1976 and have a copy of at least one Beatles album

EXAMPLE

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

The 4 metrics are:

- $\text{Support}(RM) = \frac{2720}{15{,}356} \approx 18\%$ (RM occurs in 18% of the observations)
- $\text{Confidence}(RM) = \frac{2720}{3888} \approx 70\%$ (RM is true in 70% when born prior to 1976)
- $\text{Interest}(RM) = \frac{2720}{3888} - \frac{9092}{15356} \approx 0.11$ (RM is not very interesting)
- $\text{Lift}(RM) = \frac{15{,}356^2 \cdot 0.18}{3888 \cdot 9092} \approx 1.2$ (weak correlation between being born prior to 1976 and owning a copy of a Beatles' album)

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

NOTES

For **Big(ger) Data**, it can be costly to generate rules in that fashion (especially when the number of attributes increases). How do we generate **promising** candidate rules, in general?

How **reliable** are association rules? What is the likelihood that they occur by **chance**? How **relevant** are they? Can they be generalized to **new** data?

Since frequent rules correspond to instances that occur often in the data, algorithms that generate item sets often try to **maximize coverage**.

When **rare events** are more meaningful, we need algorithms that can generate rare item sets. **This is not a trivial problem**.

NOTES

Continuous vs. **Categorical**: continuous data has to be binned into categorical data in order for association rules to be meaningful. There's more than one way to do that.

Algorithms:

- brute force, AIS, SETM, apriori, aprioriTid, aprioriHybrid, eclat, PCY, multistage, multihash, etc.

EXAMPLE – TITANIC

The Titanic dataset was compiled by Robert Dawson in 1995; it consists of 4 categorical attributes for each of the 2201 people aboard the Titanic when it sank in 1912.

Attributes are:

- **class** (first class, second class, third class, crewmember)
- **age** (adult, child)
- **sex** (male, female)
- **survival** (yes, no)

EXAMPLE – TITANIC

The natural question of interest for this dataset is how survival relates to the other attributes.

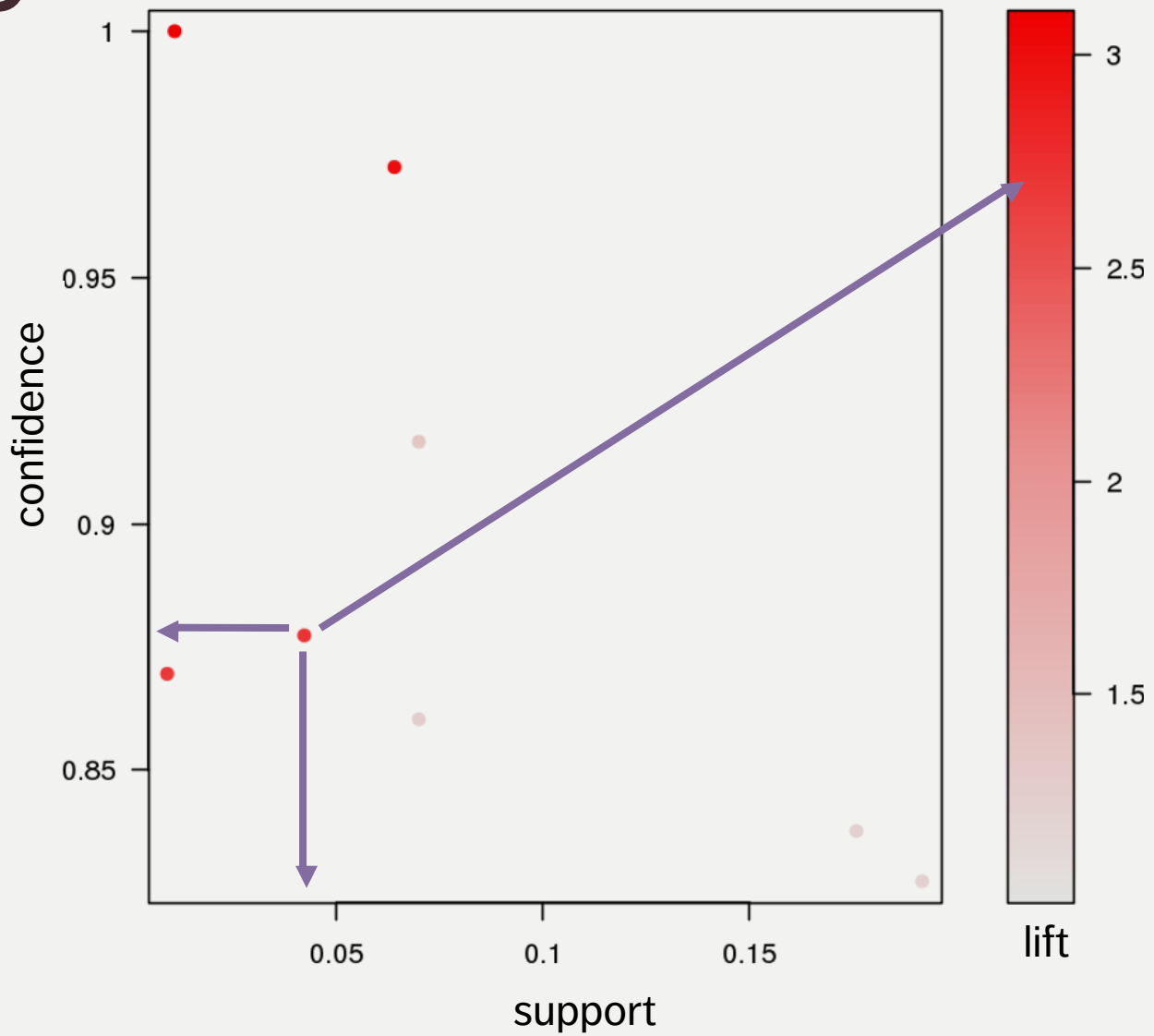
We use the `arules` implementation of *apriori* in `R` to generate and prune candidate rules, eventually leading to **8 rules**.

Is this a supervised or an unsupervised task?

Would any rule be transferrable to new data?

EXAMPLE – TITANIC

Rule	Supp	Conf	Lift
IF class = 2nd AND age = Child THEN survived = Yes	0.01	1	3.10
IF class = 1st AND sex = Female THEN survived = Yes	0.06	0.97	3.01
IF class = 2nd AND sex = Female THEN survived = Yes	0.04	0.88	2.72
IF class = Crew AND sex = Female THEN survived = Yes	0.00	0.87	2.70
IF class = 2nd AND sex = Male AND age = Adult THEN survived = No	0.07	0.92	1.35
IF class = 2nd AND sex = Male THEN survived = No	0.07	0.86	1.27
IF class = 3rd AND sex = Male AND age = Adult THEN survived = No	0.18	0.84	1.24
IF class = 3rd AND sex = Male THEN survived = No	0.19	0.83	1.22



CLASSIFICATION

STATISTICAL LEARNING

"Data science does not replace statistical modeling and data analysis; it augments them."

(P. Boily)

CLASSIFICATION OVERVIEW

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

CLASSIFICATION OVERVIEW

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

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

Technical issues include:

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

APPLICATIONS

Medicine and Health Science

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

Social Policies

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

Marketing and Business

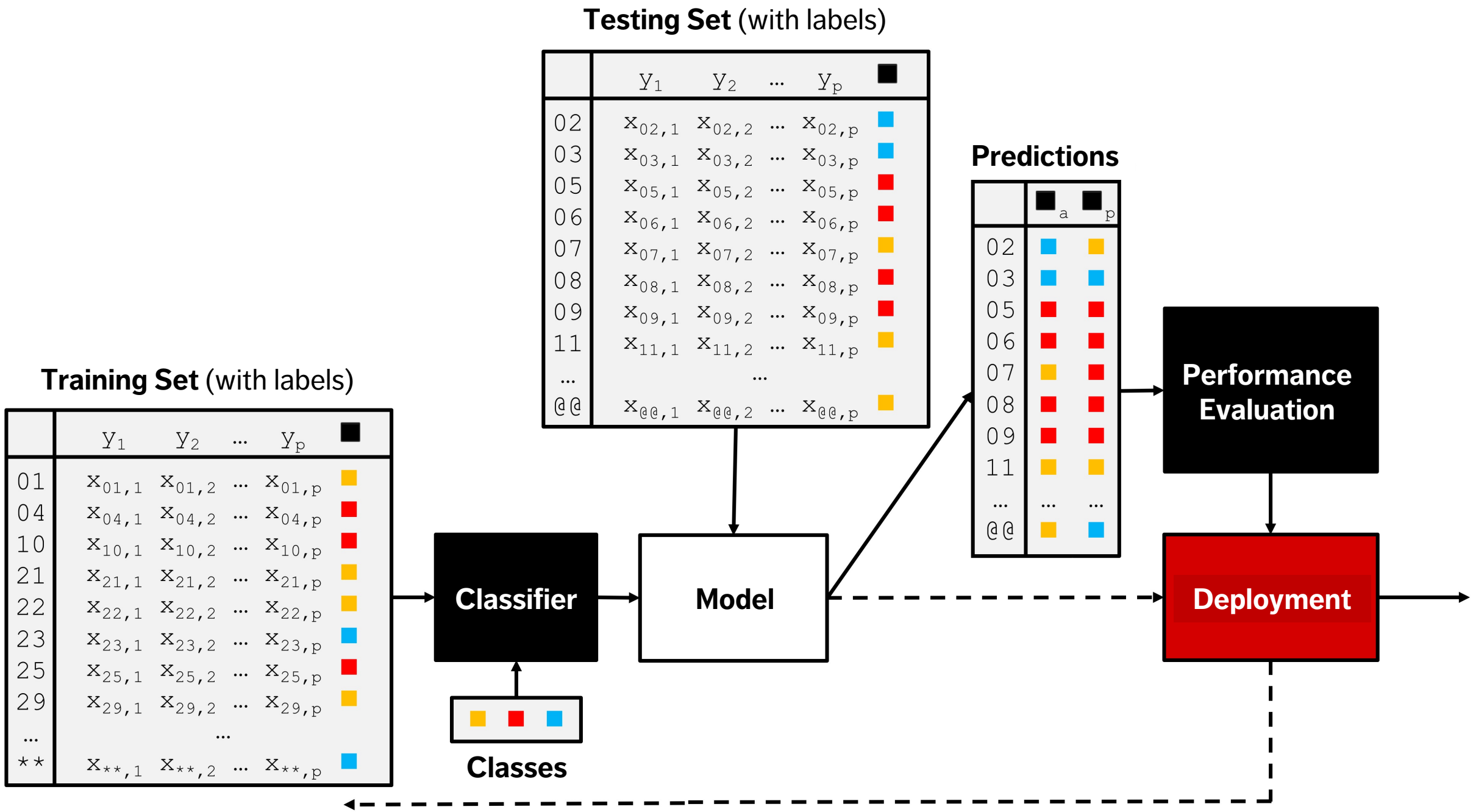
- predicting which customers are likely to switch to another cell phone company based on demographics and usage

EXAMPLE

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

Questions: can we predict

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


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

CLASSIFICATION METHODS

Logistic Regression

Neural Networks

Decision Trees

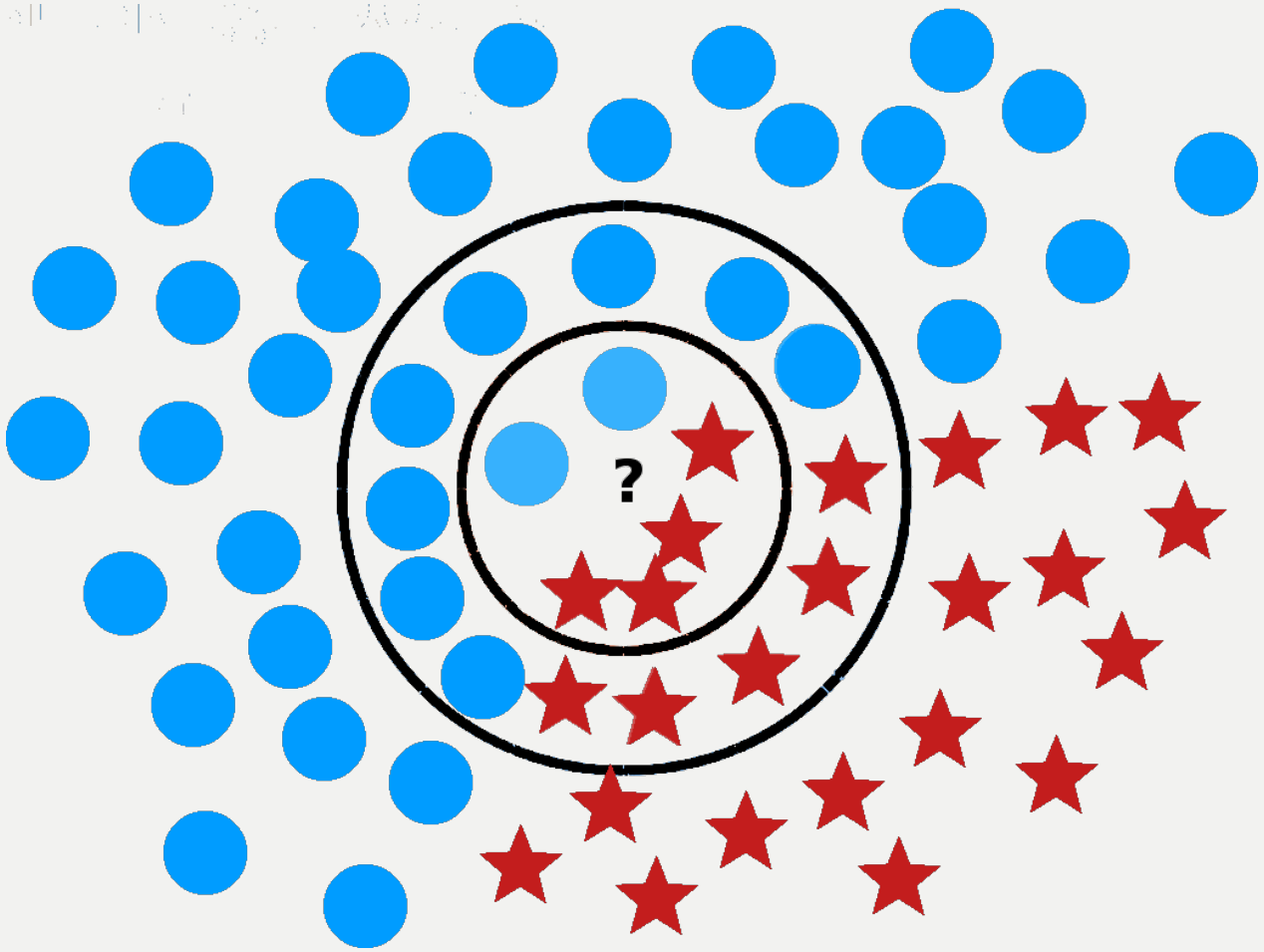
Naïve Bayes Classifiers

Support Vector Machines

Nearest Neighbours Classifiers

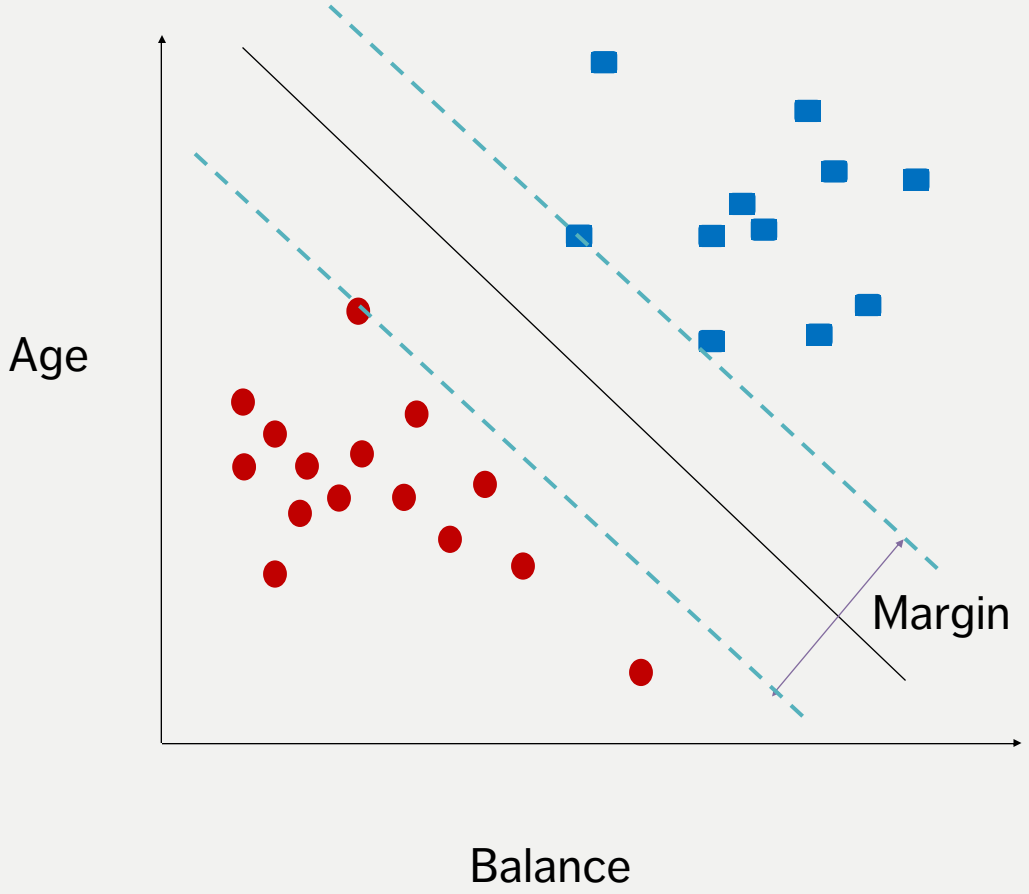
etc.

k –NEAREST NEIGHBOURS



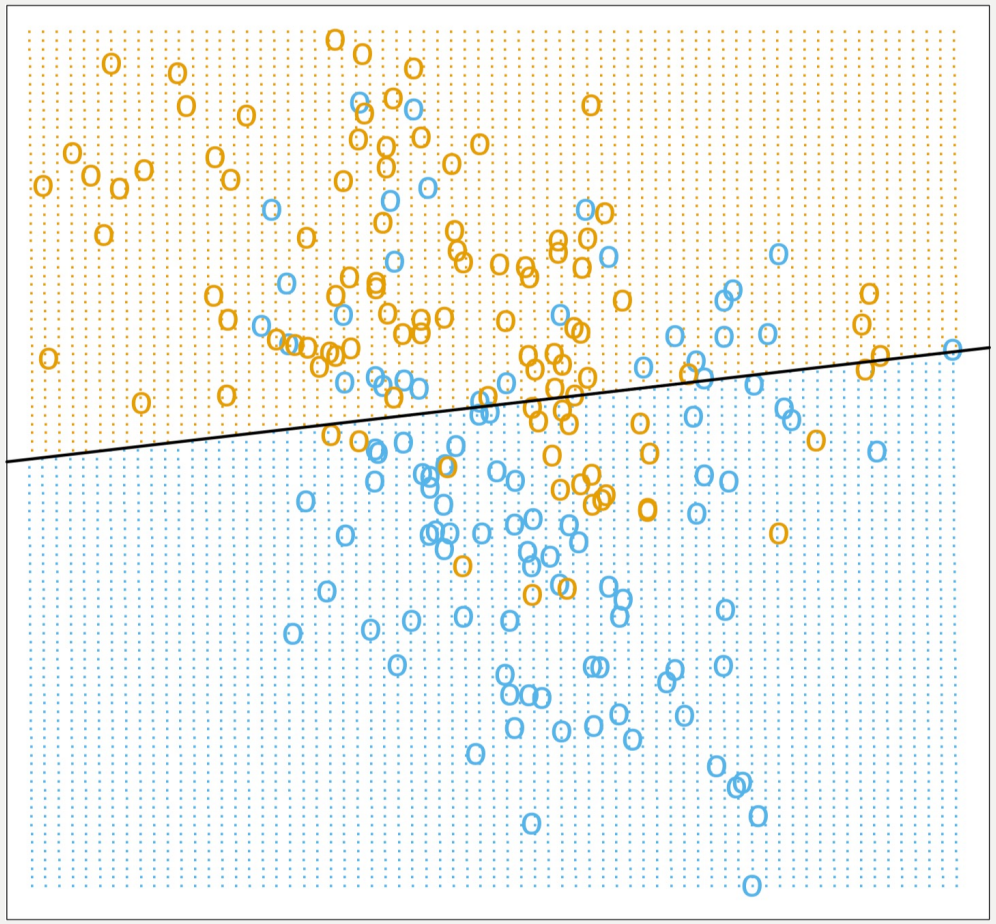
[Adapted from Foster & Provost's *Data Science for Business*]

SUPPORT VECTOR MACHINES

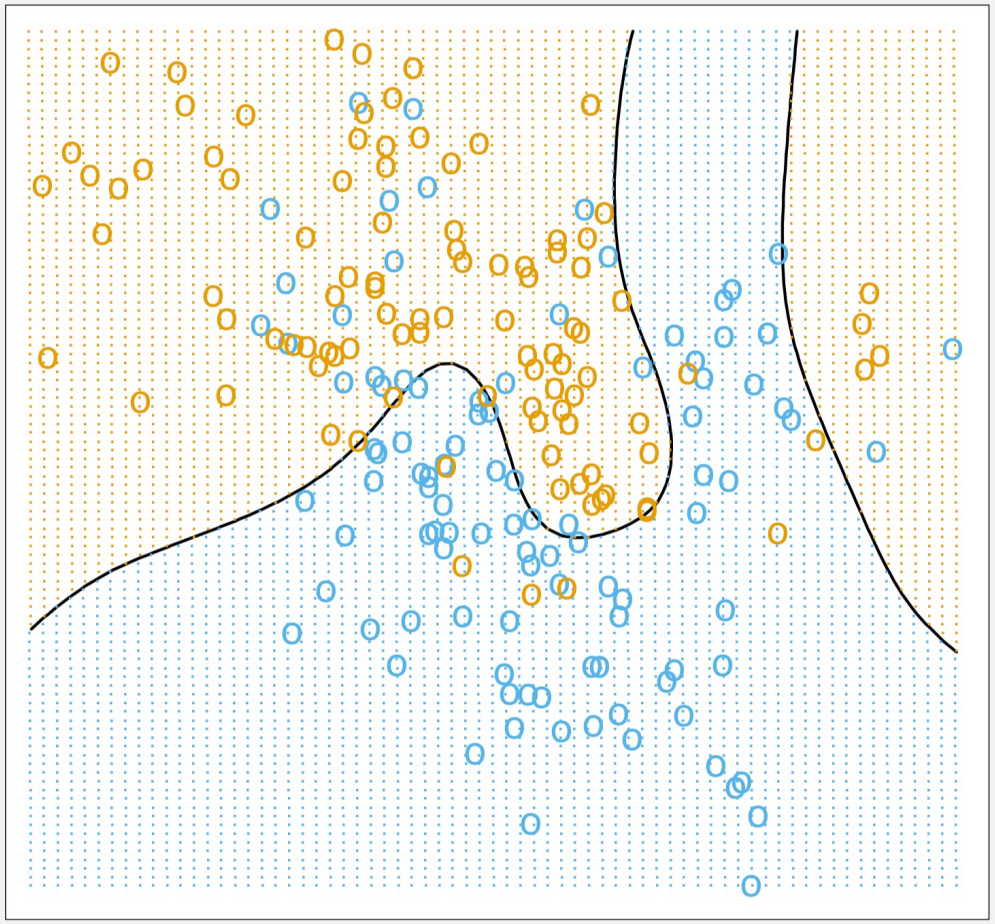


[Images taken from *The Elements of Statistical Learning*]

BOUNDARY CLASSIFIERS



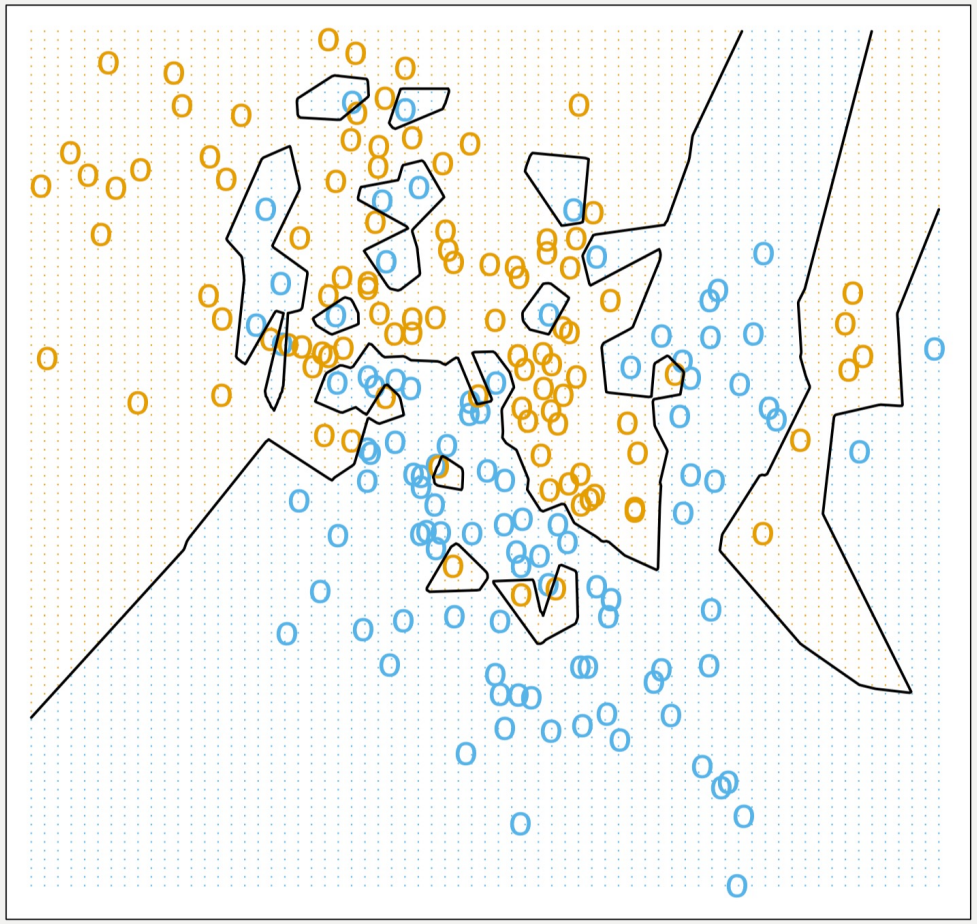
Linear Regression Classifier



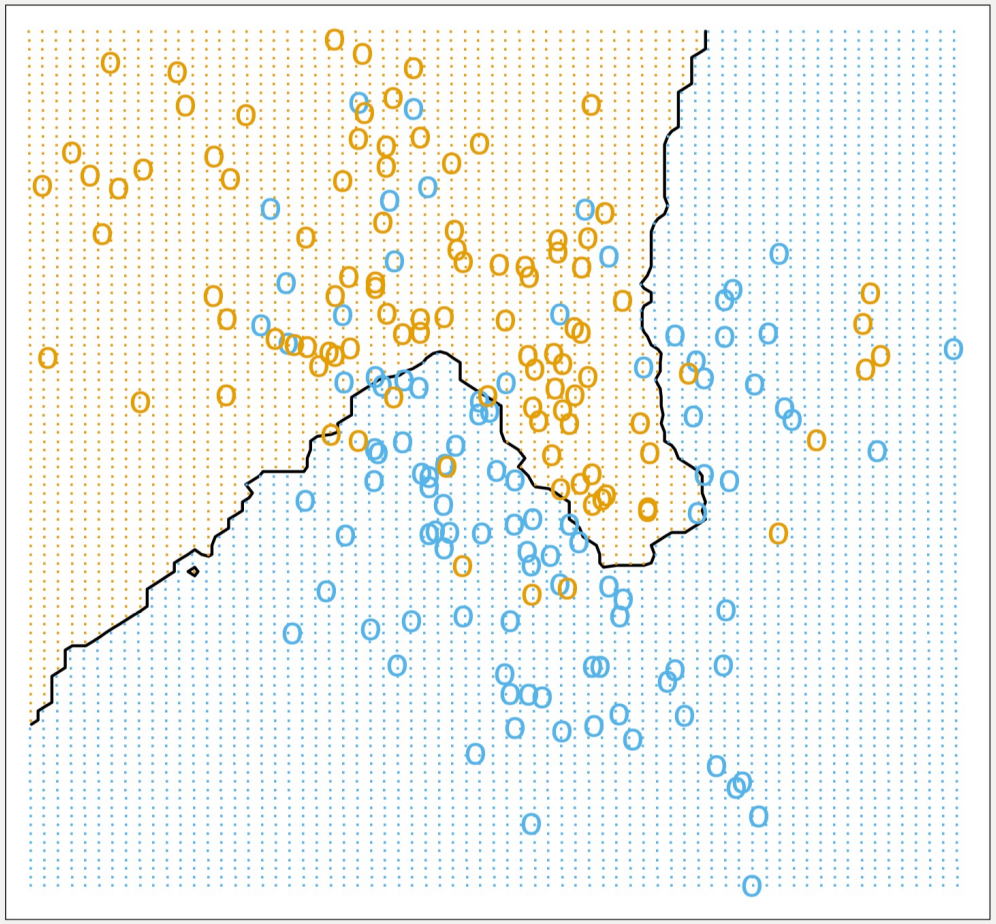
Optimal Bayes Classifier

[Images taken from *The Elements of Statistical Learning*]

NEAREST NEIGHBOURS CLASSIFIERS



1NN Classifier



15NN Classifier

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

DECISION TREES

Decision trees are perhaps the most **intuitive** of these methods.

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

DECISION TREES

To make a **prediction** for a new instance, follow the path down the tree, reading 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 TREES
STRENGTHS AND LIMITATIONS

Strengths:

- **white box** model
- can be used with **incomplete** datasets
- **built-in** feature selection
- makes **no assumption** about data

Limitations:

- **not as accurate** as other algorithms (usually)
- **not robust**
- particularly vulnerable to **overfitting**
- optimal decision tree learning is **NP-complete**
- biased towards categorical features with high number of levels

DECISION TREE NOTES

Splitting Metrics

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

Common Algorithms

- iterative Dichotomiser 3, C4.0, C4.5, CHAID, MARS, CI trees, CART
- decision trees can also be combined together using boosting algorithms (**AdaBoost**) or **Random Forests**, providing a type of voting procedure (**Ensemble Learning**)

OTHER POINTS TO PONDER

Classification is linked to **probability estimation**

- approaches based on regression models could prove fruitful

Rare occurrences (often more interesting/important) continue to plague classification attempts

- historical data at Fukushima's nuclear reactor prior to the meltdown could not have been used to learn about meltdowns

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

With big datasets, algorithms must also consider efficiency.

PERFORMANCE EVALUATION

Classifiers are evaluated on the testing set.

Ideally, a good classifier would have

- high rates of both **True Positives** (TP) and **True Negatives** (TN), and
- low rates of both **False Positives** (FP) and **False Negatives** (FN).

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

PERFORMANCE EVALUATION

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

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

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

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

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

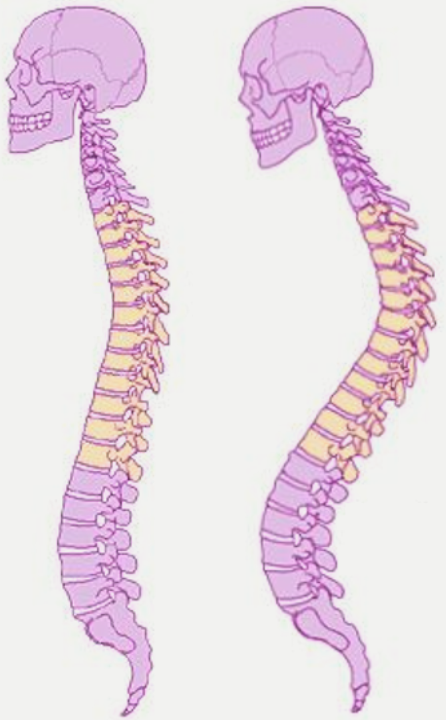
Performance Metrics	
Accuracy:	0.80
F1-Score:	0.87
Informedness (ROC):	0.41
Markedness:	0.77
M.C.C.:	0.56
Pearson's chi2:	0.33
Hist. Stat:	0.40

EXAMPLE – KYPHOSIS

Kyphosis is a medical condition related to the excessive convex curvature of the spine. Corrective spinal surgery is at times performed on children.

The dataset has 81 observations and 4 attributes:

- **kyphosis** (absent or present after operation)
- **age** (at time of operation, in months)
- **number** (of vertebrae involved)
- **start** (topmost vertebra operated on)



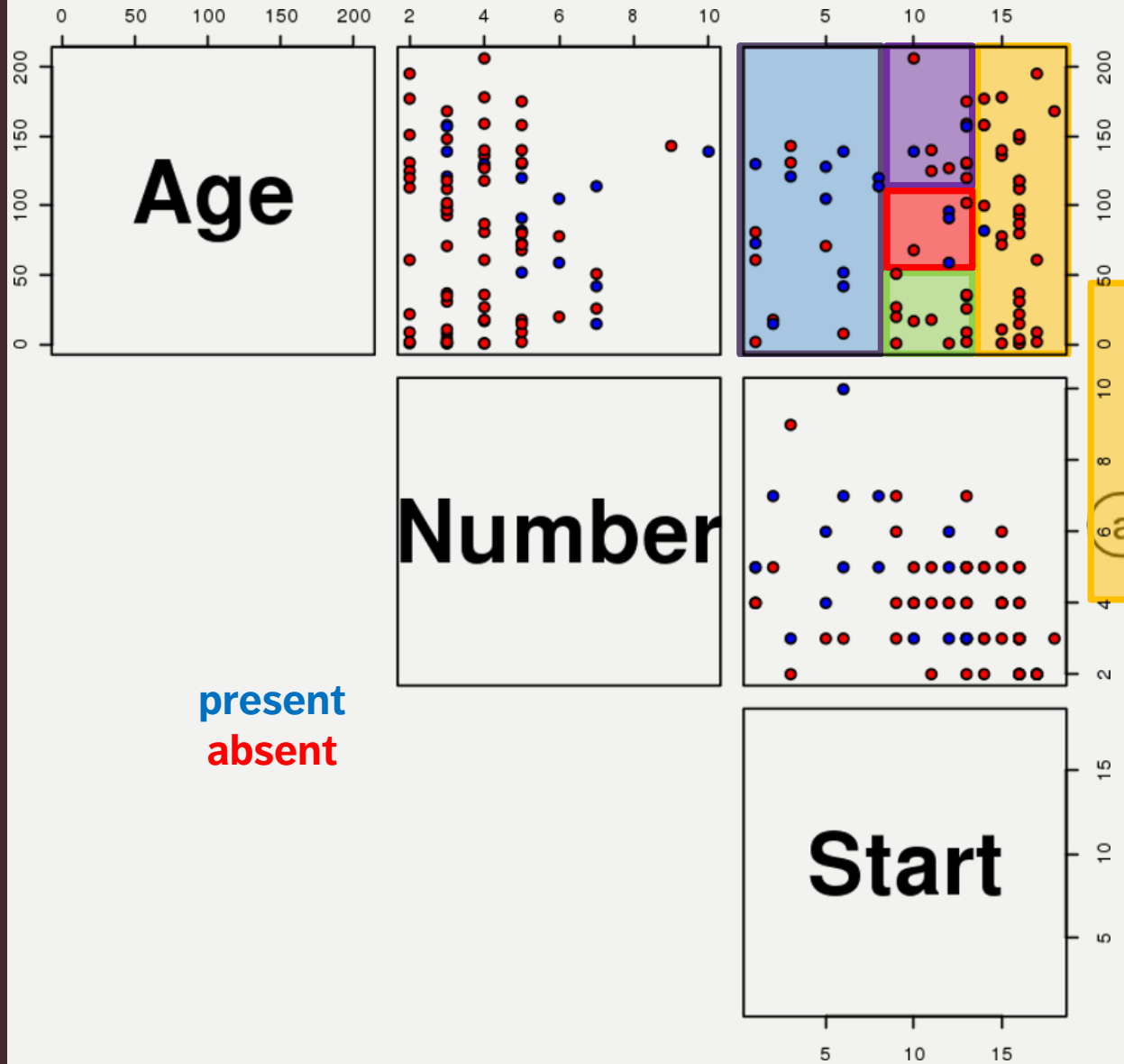
EXAMPLE – KYPHOSIS

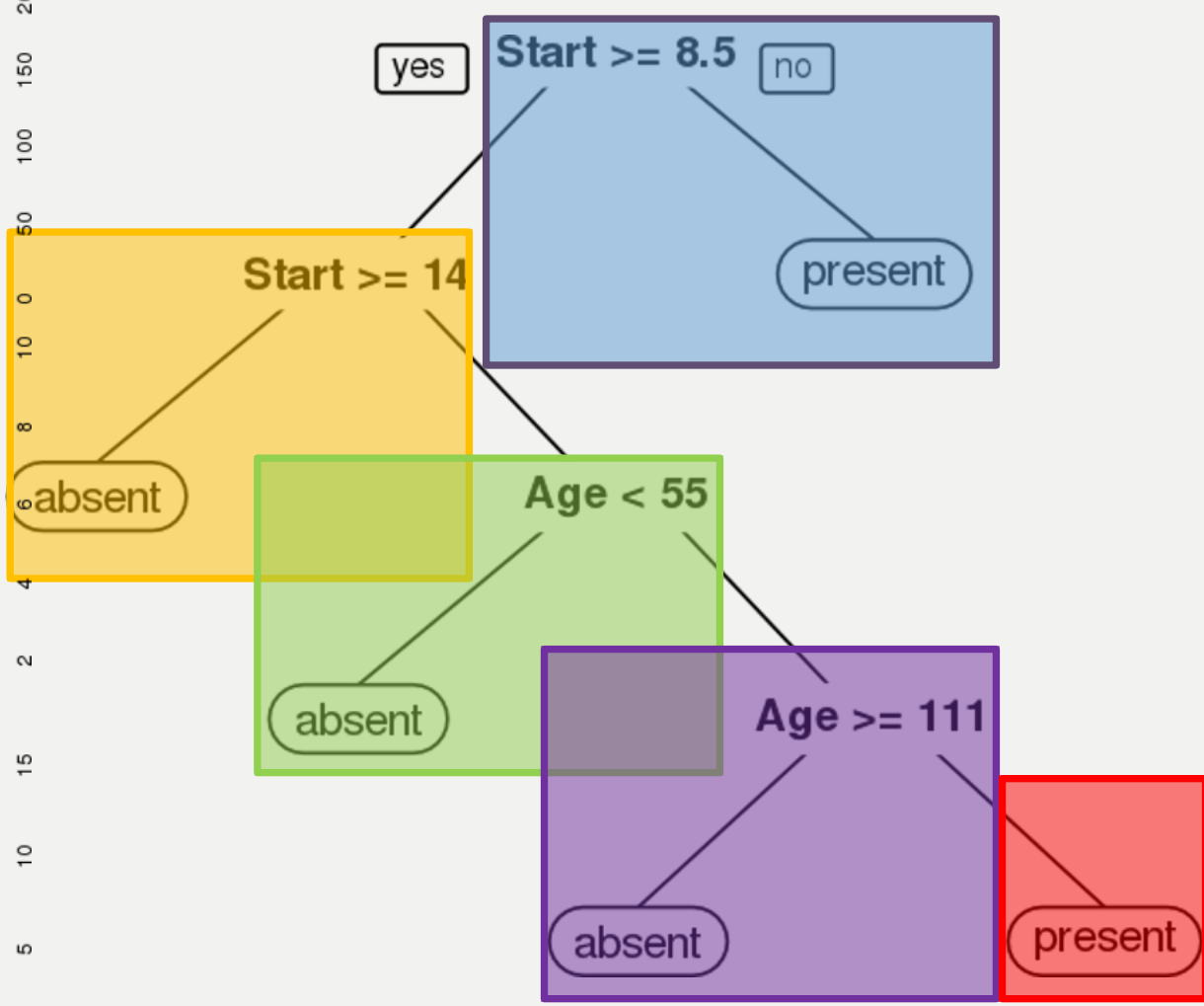
The question of interest for this natural dataset is how the three explanatory attributes might impact the operation's success.

We use the `rpart` implementation of CART to generate candidate decision trees.

Strictly speaking, this is not a predictive supervised task as we treat the entire dataset as a training set (no hold-out testing observations).

EXAMPLE – KYPHOSIS





EXAMPLE – KYPHOSIS

We train a model on 50 randomly selected observations and evaluate the performance on the remaining 31 observations.

		Predicted			Total
		A	B		
Actuals	A	23	3	26	83.9%
	B	3	2	5	16.1%
Total		26	5	31	
		83.9%	16.1%		

Classification Rates	
Sensitivity:	0.88
Specificity:	0.40
Precision:	0.88
Negative Predictive Value:	0.40
False Positive Rate:	0.60
False Discovery Rate:	0.12
False Negative Rate:	0.12

Performance Metrics	
Accuracy:	0.81
F1-Score:	0.88
Informedness (ROC):	0.28
Markedness:	0.28
M.C.C.:	0.28
Pearson's chi2:	0.00
Hist. Stat:	0.00

Is this a good model?

CLUSTERING

STATISTICAL LEARNING

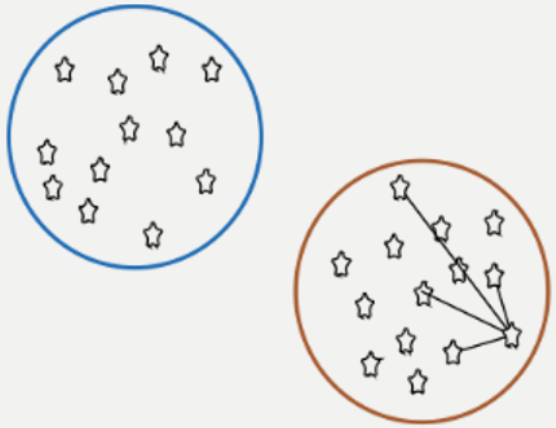
"Data is not information,
information is not knowledge,
knowledge is not understanding,
understanding is not wisdom."
(C. Stoll)

CLUSTERING OVERVIEW

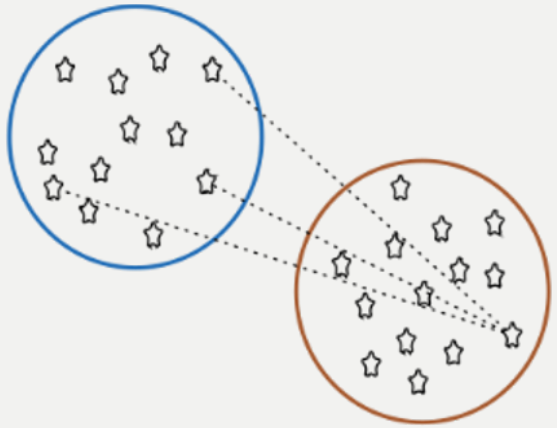
In **clustering**, the data is divided into **naturally occurring groups**. Within each group, the data points are **similar**; from group to group, they are **dissimilar**.

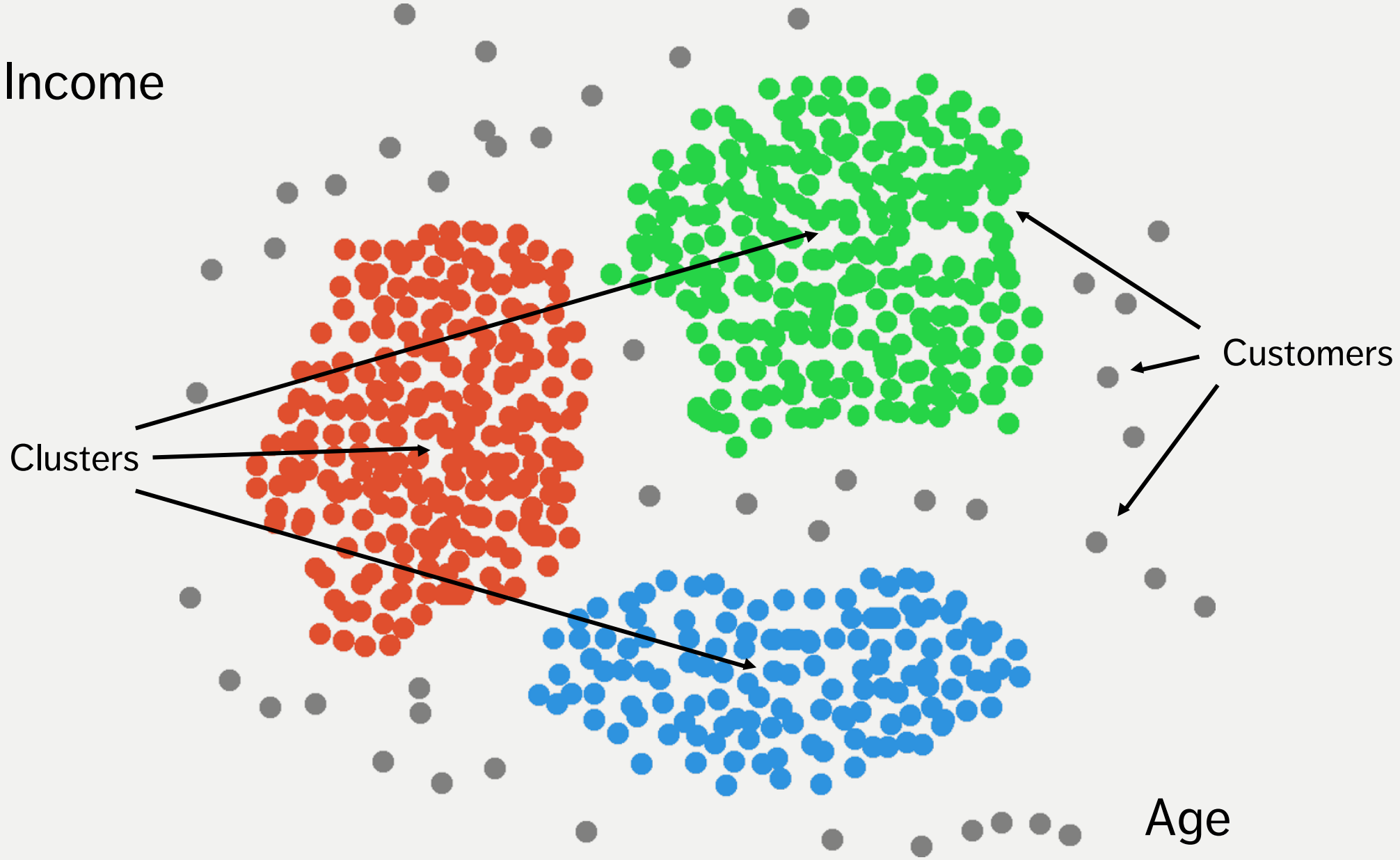
The grouping labels are not determined ahead of time, so clustering is an example of **unsupervised** learning.

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



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




Income
Customers
Clusters
Age

CLUSTERING OVERVIEW

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

- facial recognition
- searching for patterns, etc.

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

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

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

CLUSTERING OVERVIEW

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

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

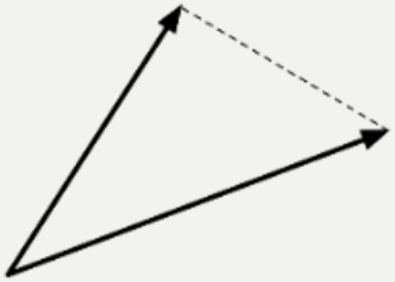
They are also (typically) **non-deterministic**:

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

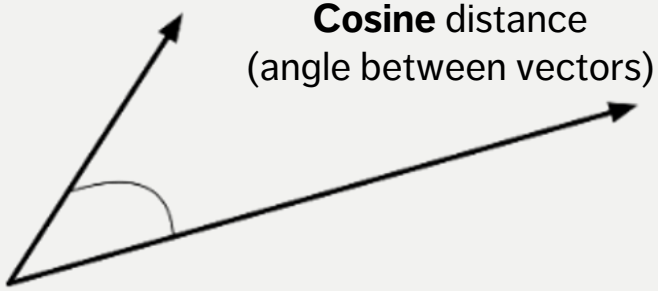
IMPORTANT: data must be scaled before it is fed into clustering algorithms.

CLUSTERING REQUIREMENT

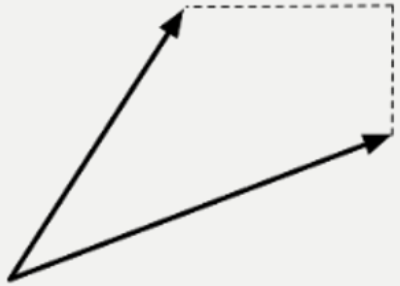
A measure of **similarity** w (or a distance d) between observations.



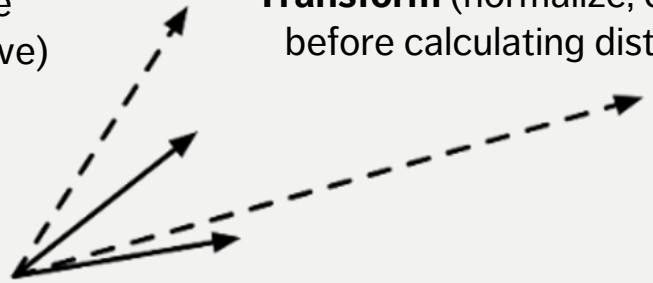
Euclidean distance
(as the crow flies)



Cosine distance
(angle between vectors)



Manhattan distance
(you might have to drive)



Transform (normalize, center)
before calculating distance

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

APPLICATIONS

Text Documents

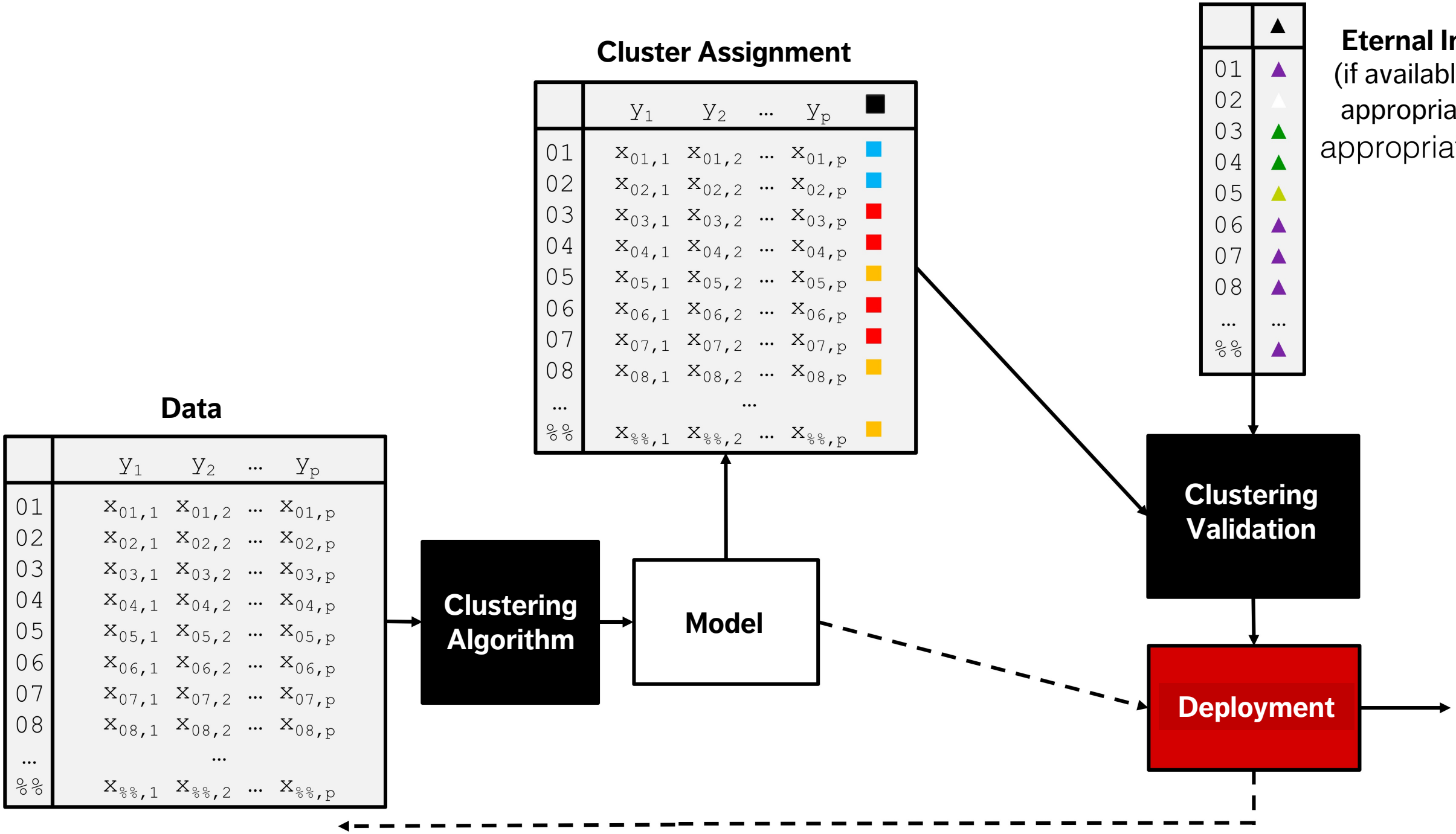
- grouping similar documents according to their topics, based on the patterns of common and unusual words

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


Cluster Assignment
Eternal Info
(if available & appropriate)
appropriate)
Data
Clustering Validation
Clustering Algorithm
Model
Deployment

CLUSTERING METHODS

k-Means

Hierarchical Clustering

Latent Dirichlet Allocation

Expectation-Maximization

Balanced Iterative Reducing and Clustering using Hierarchies

Density-Based Spatial Clustering of Applications with Noise

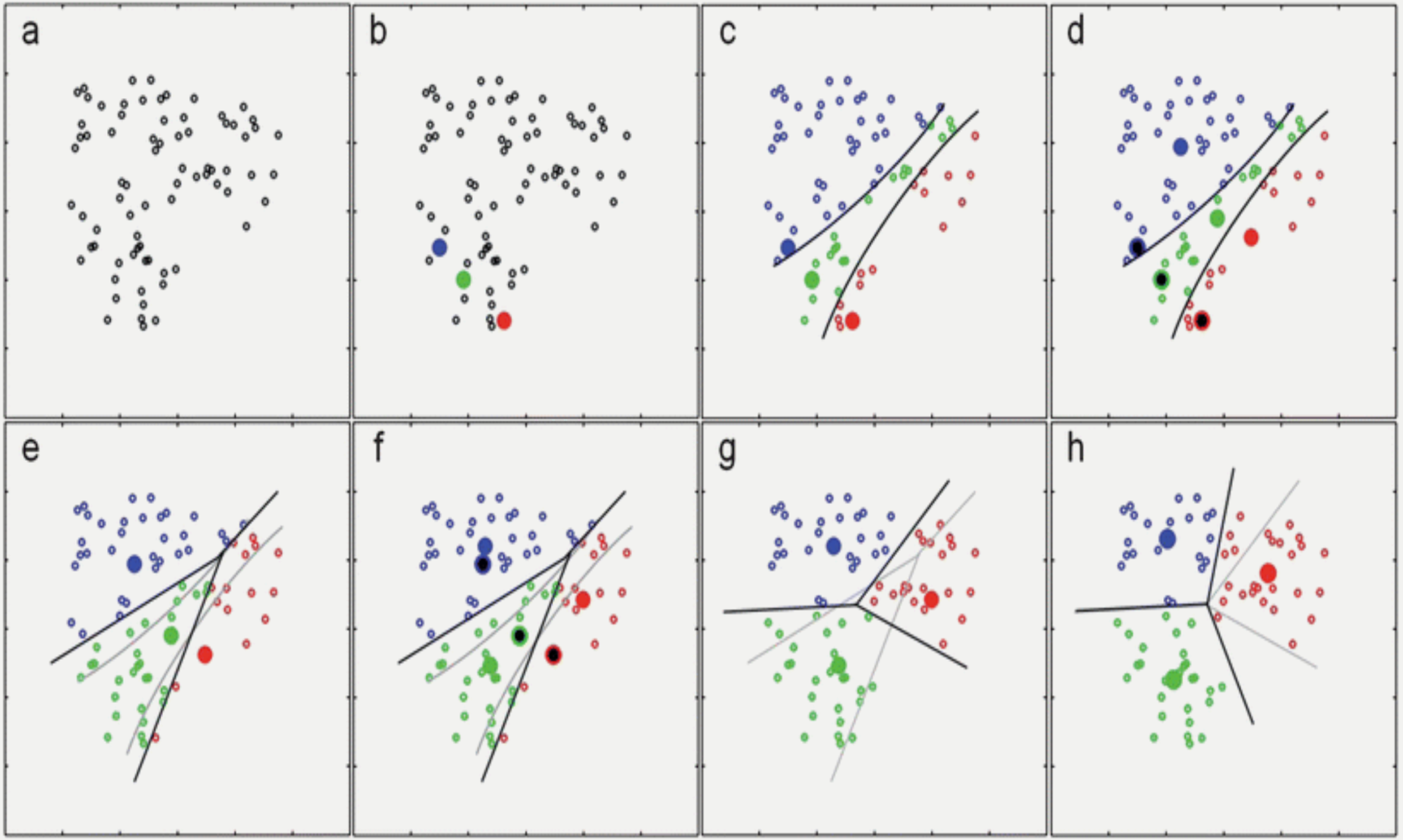
Affinity Propagation

Spectral Clustering, etc.

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**

[https://notebook.community/alanhdu/AccessibleML/4. K-means clustering and unsupervised learning]



k-MEANS STRENGTHS AND LIMITATIONS

Strengths:

- easy to implement
- often a **natural** way to group observations
- helps provide a **basic understanding of the data structure** in a first pass

Limitations:

- points can only be assigned to **one** cluster
- underlying clusters are assumed to be **blob-shaped**
- clusters are assumed to be discrete (separate entities)

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)

EXAMPLE – FRUIT IMAGES

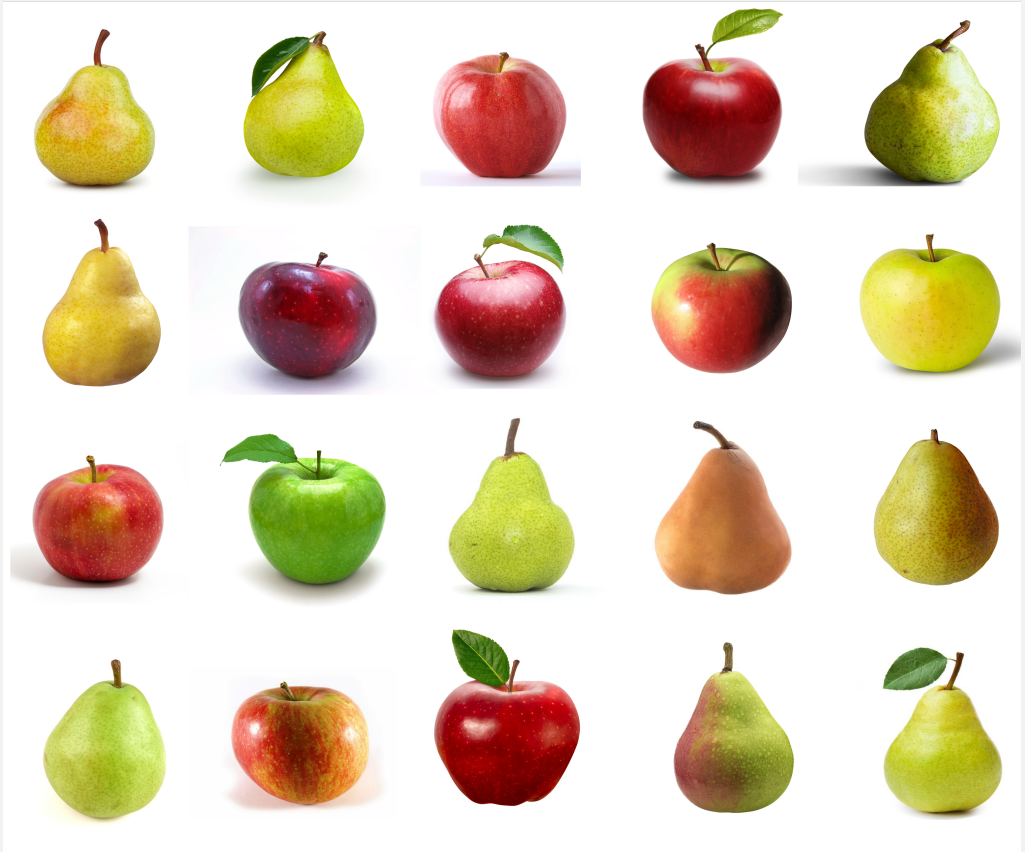
20 images of fruit.

Are there right or wrong groupings of this dataset?

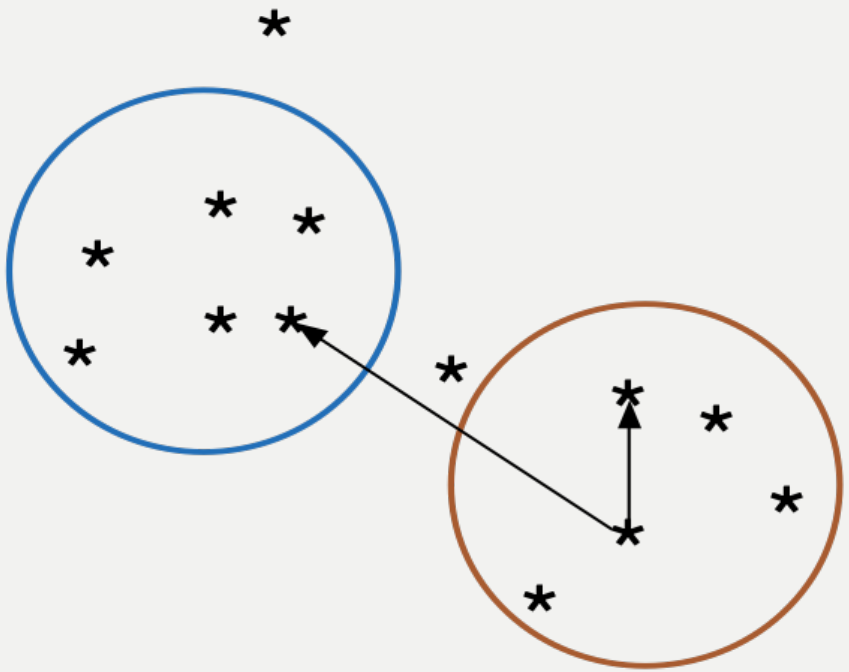
Are there multiple possible ‘natural’ clusterings?

Could different clusterings be used differently?

Will some clusterings be of (objectively) higher **quality** than others?



VALIDITY VS. QUALITY



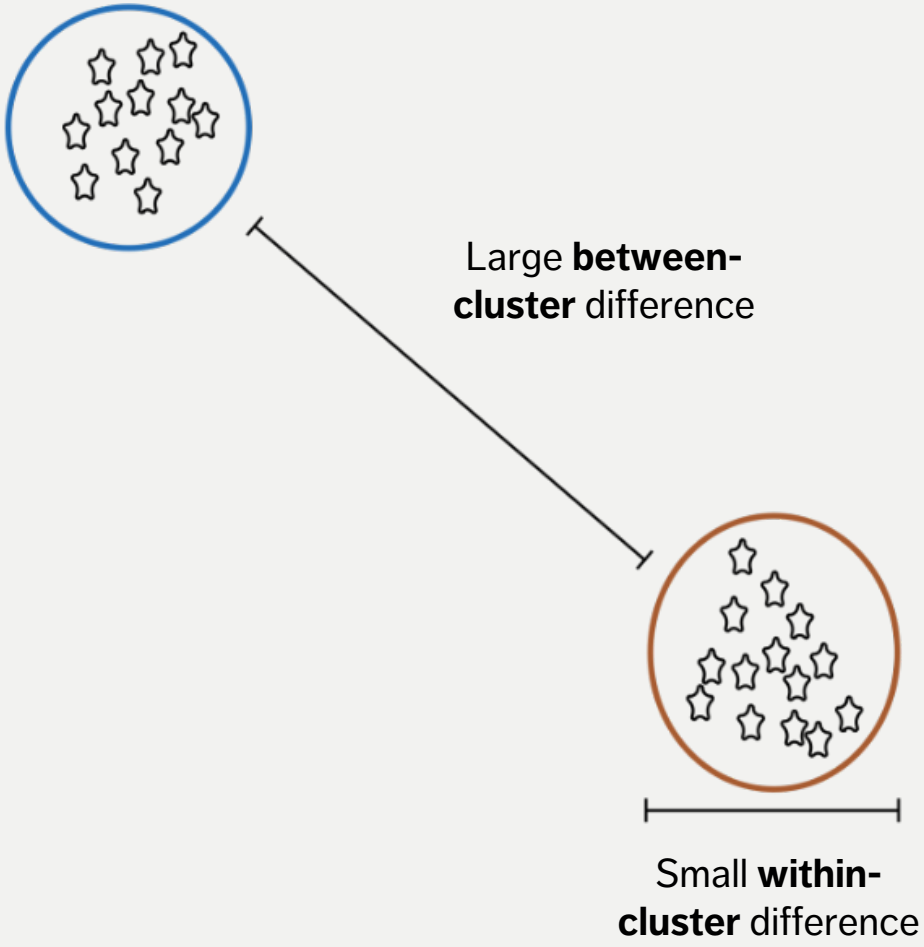
Context is very relevant to the quality of a given clustering, but what if we have no context?

Is there a way to **objectively measure** cluster quality without any specific context?

The term ‘validity’ suggests there is a **correct** clustering, against which we compare performance.

Lewis, Ackerman, de Sa (2012) speak of **Clustering Quality Measures** (CQM) instead.

BROAD GOALS OF CQMs



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).

Result: a huge number of CQMs (30+).

CLUSTERING CHALLENGES

Automation

Lack of a clear-cut definition

Lack of repeatability

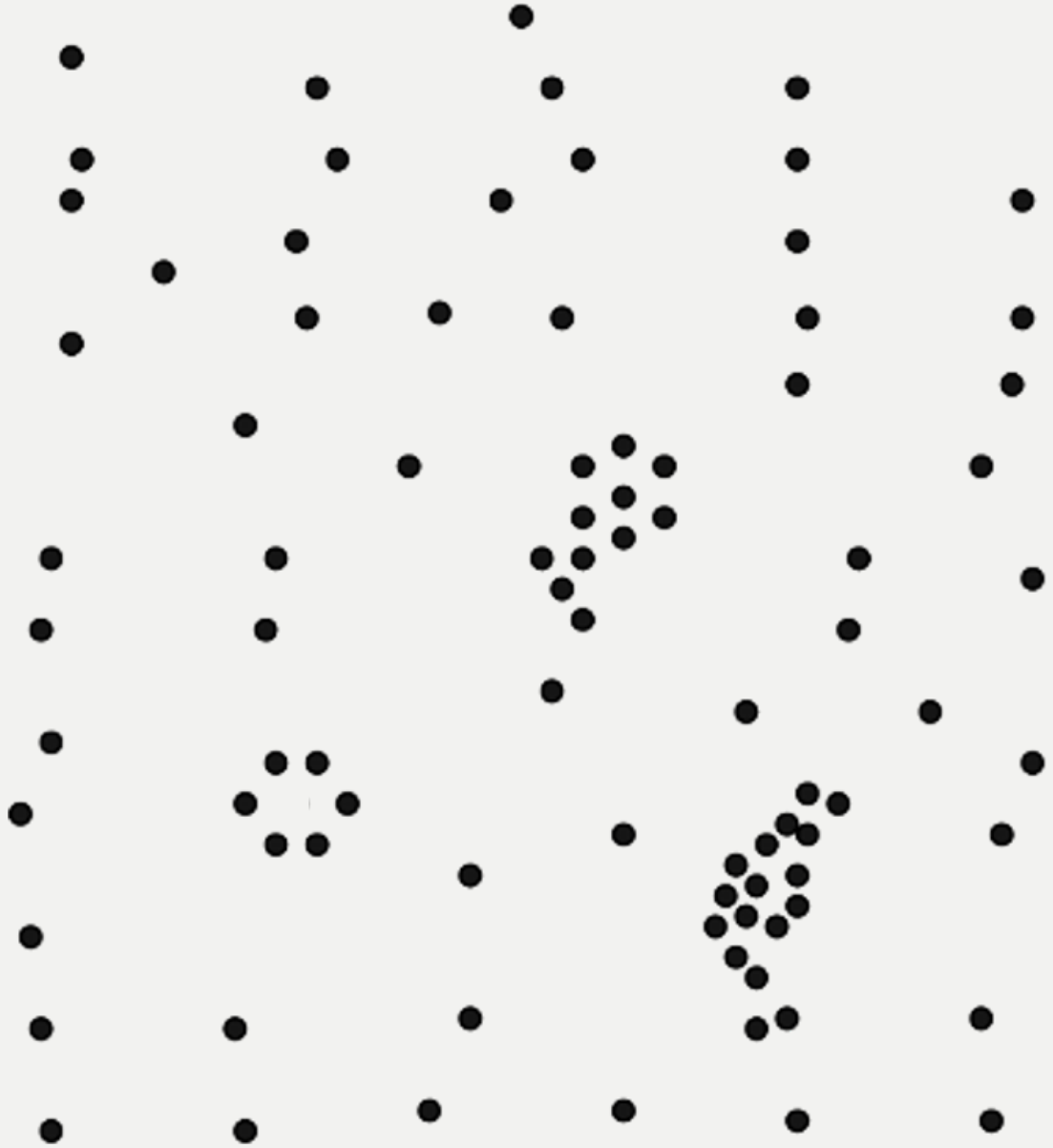
Number of clusters

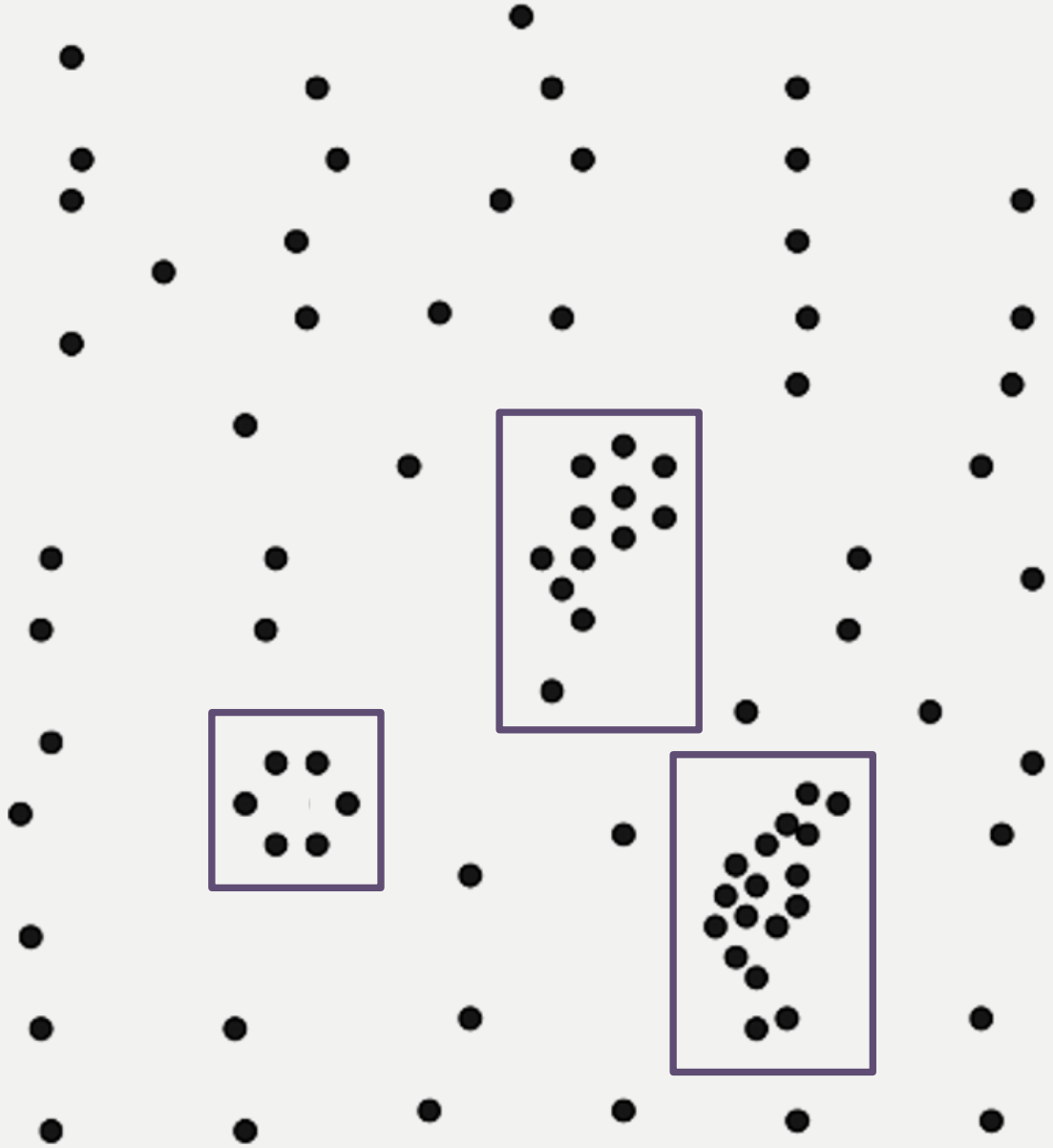
Cluster description

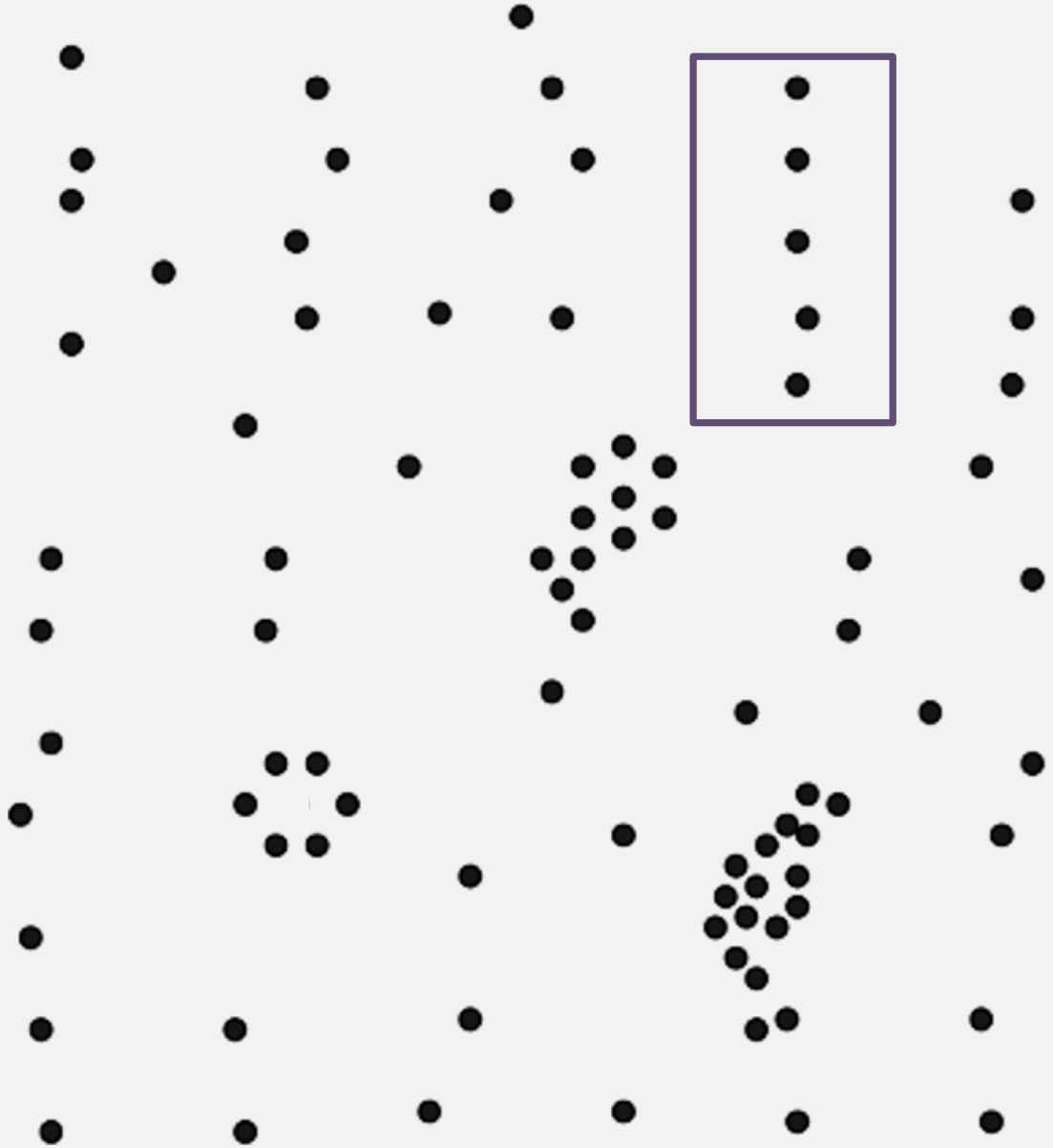
Model validation

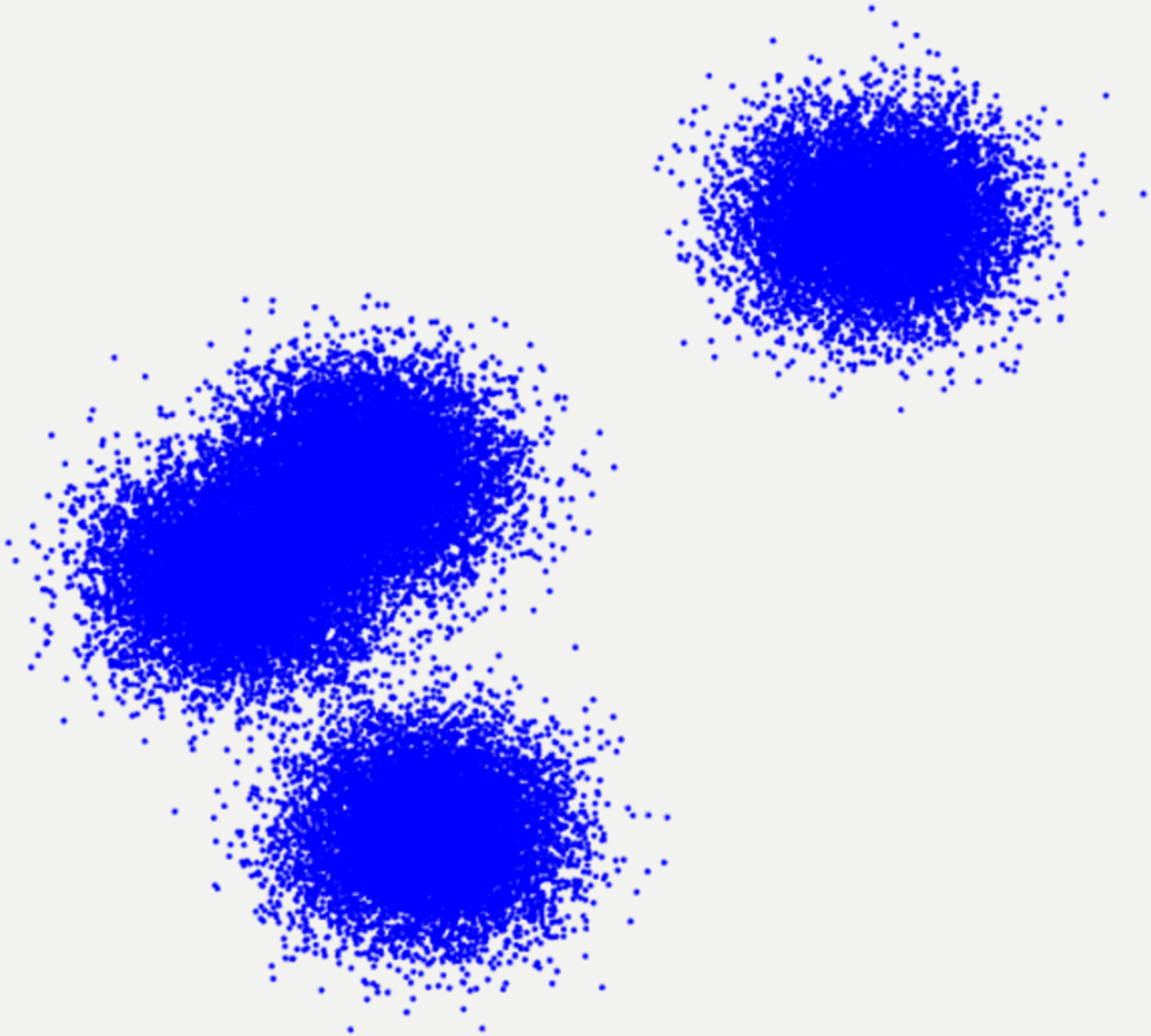
Ghost clustering

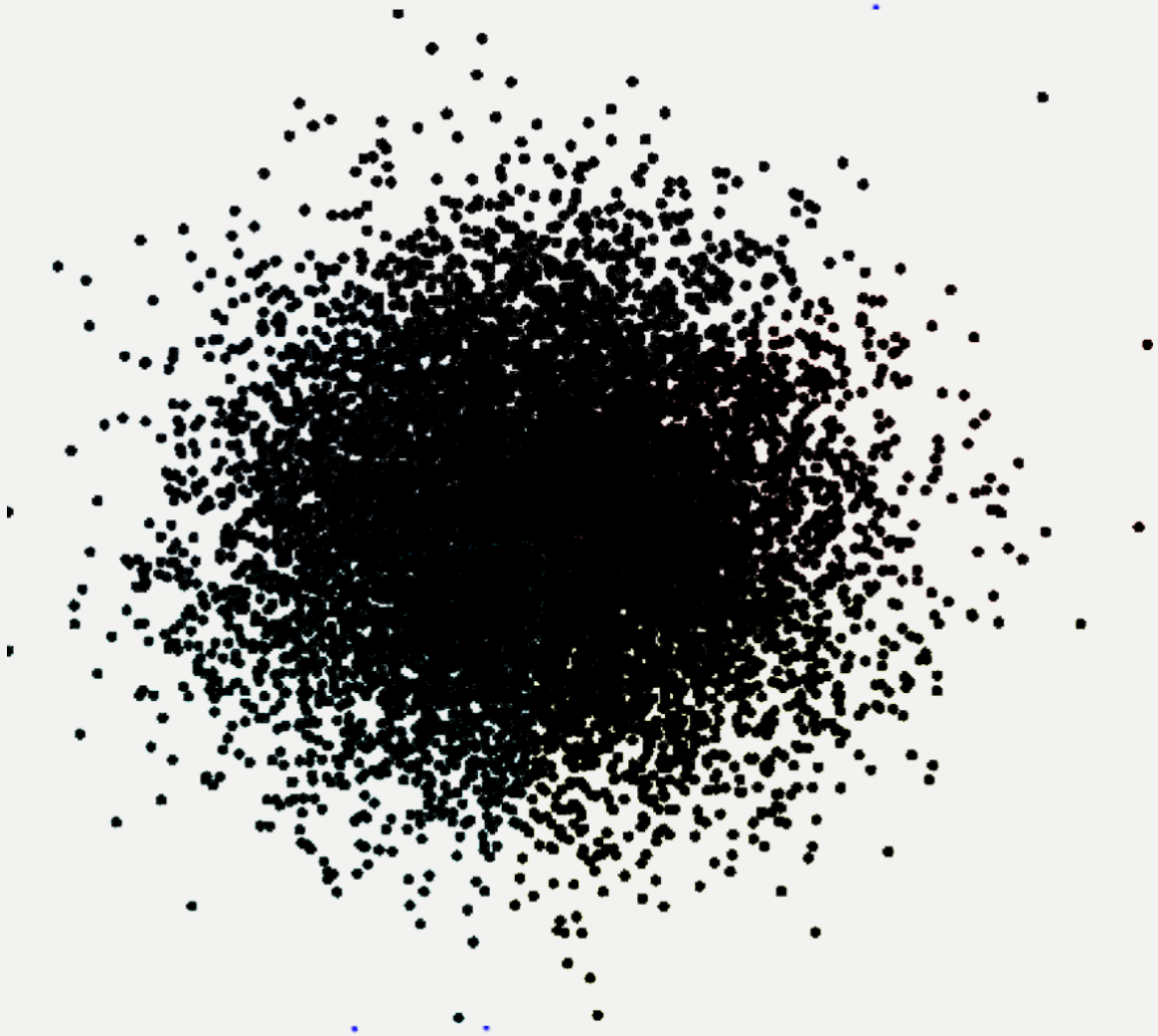
A posteriori rationalization

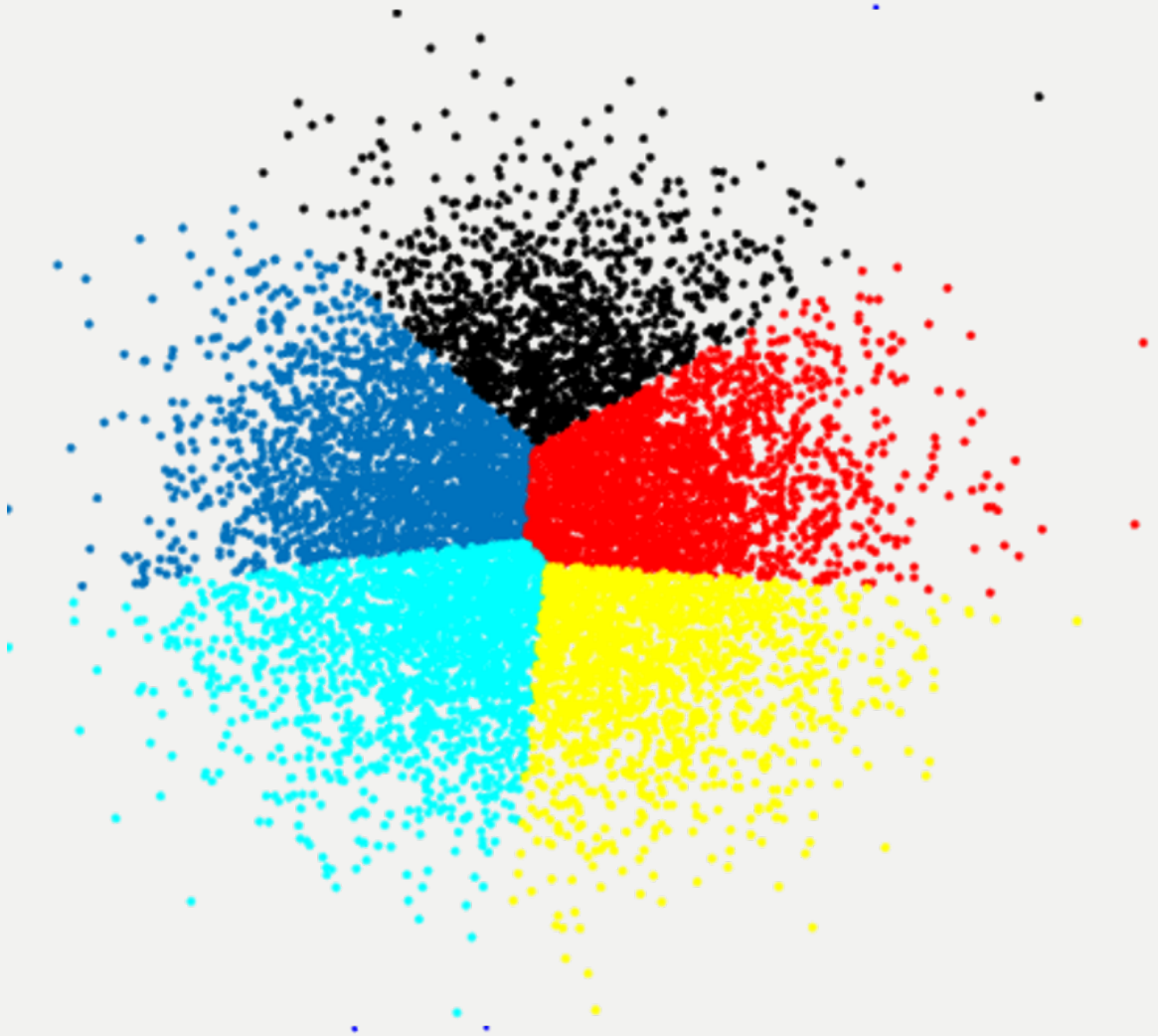












ISSUES & CHALLENGES

STATISTICAL 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 quite care for data, and it sounds like I'm not alone."

(Q.E. McCallum, *Bad Data Handbook*)

BAD DATA

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

Difficulties with **text processing** (encoding, application-specific characters)

Collecting data **online** (legality of obtaining data, storing offline versions)

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

Data and reality: bad data vs. bad reality?

Is **close enough, good enough**? (completeness, coherence, correctness, accountability)

BAD DATA

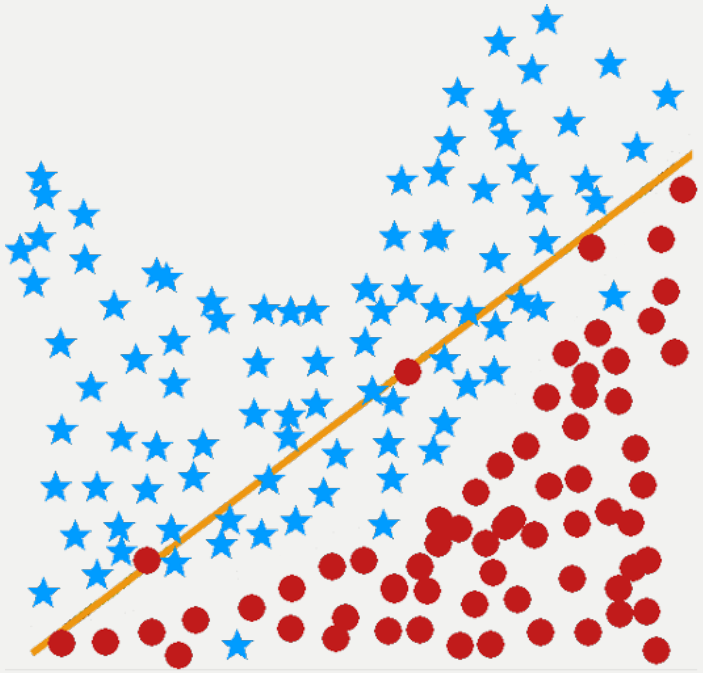
Sources of **bias** and **errors:** imputation bias, replacing extreme values with average values, head of household reports for household, etc.

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

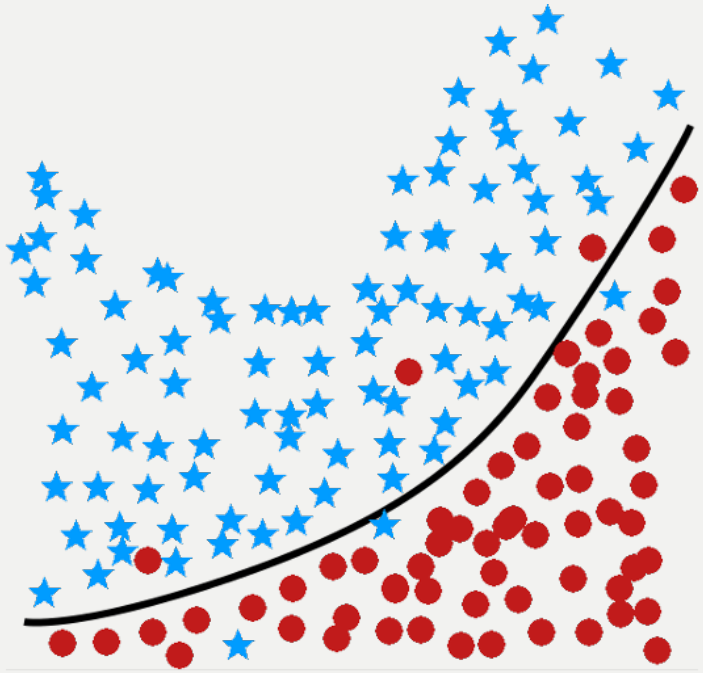
Data science **pitfalls:** analysis without understanding, using only one tool (by choice/fiat), analysis for the sake of analysis, unrealistic expectations of data science, it's on a need-to-know basis and you don't need to know.

Databases vs. files vs. cloud computing (the cloud will solve our problems!)

OVERFITTING



underfit



just right

overfit

OVERFITTING

The hope is for rules or models generated by any technique on a **training set** to be generalizable to **new data** (or **validation/ testing sets**).

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**.

Evaluate on unseen data and be wary of models built on all available data.

OVERFITTING EXAMPLE

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.

BIG DATA VS. SMALL DATA

What is the main difference?

- 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

THE 5-V PARADIGM

Volume: large amounts of data

Velocity: speed at which data is created, accessed, processed

Variety: different types of available data, can't all be saved in relational databases (tables, pictures,...)

Veracity: quality and accuracy of big data is harder to control

Value: turn the data into something useful

Variability
Visualization

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 larger 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**.

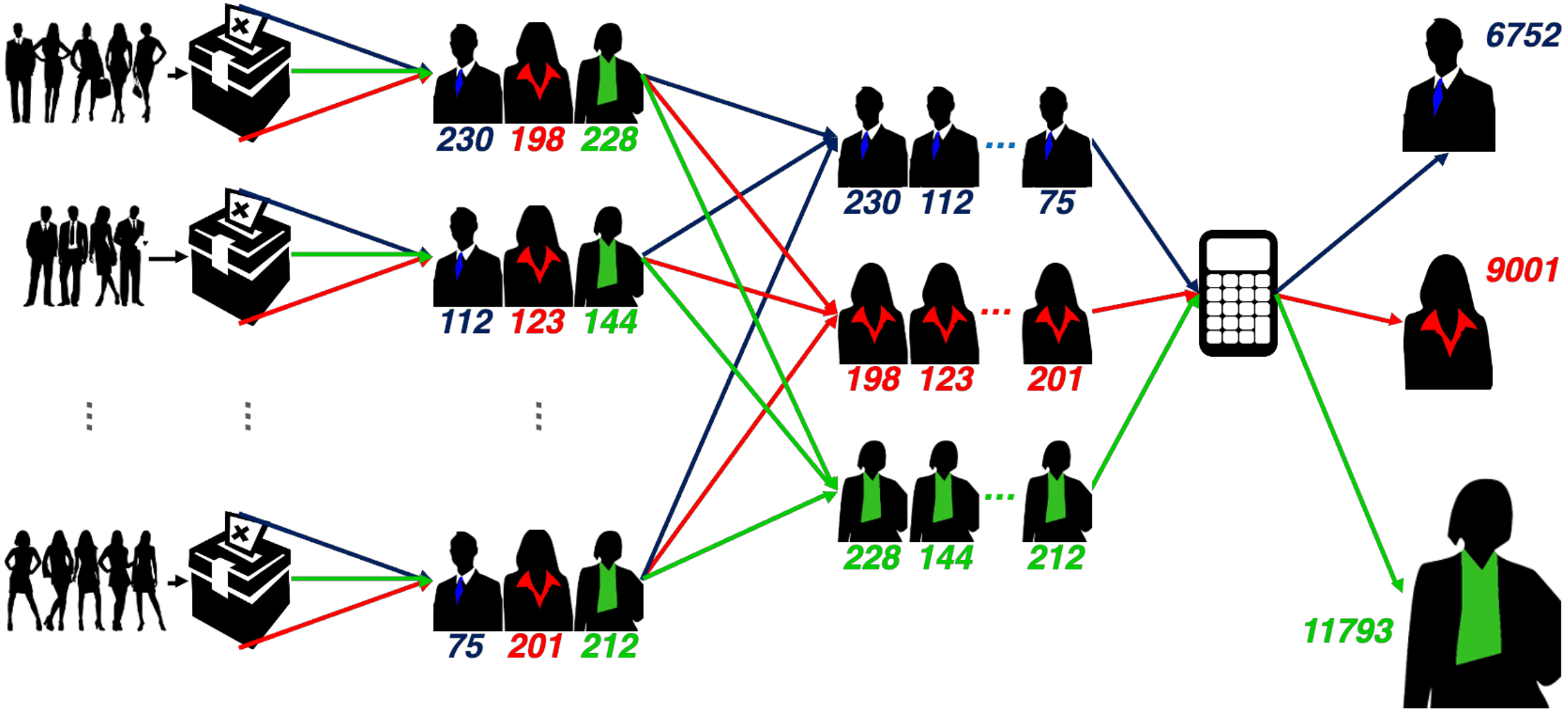
DISTRIBUTED COMPUTING

Splitting the computations among multiple CPU cores/CPUs can divide the computation time by a factor of 4, or 32, or 1000.

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.


230 198 228
112 123 144
75 201 212
230 112 ... 75
198 123 ... 201
228 144 ... 212
6752
9001
11793

SERIAL PROCESSING

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

Pizzeria analogy for limitations of parallelization/bottleneck:

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

Most practical computational tasks can be and are parallelized.

[https://hbr.org/2013/07/how-google-flu-trends-is-getting-to-the-bottom]

"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."

(N. Diakopoulos, *Harvard Business Review*)

APPROPRIATENESS AND TRANSFERABILITY

Data science models are used heavily.

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

Data Science methods are **not** appropriate if:

- if one absolutely must use an existing (**legacy**) datasets instead of an ideal dataset (“it’s the best data we have!”)

APPROPRIATENESS AND TRANSFERABILITY

Data Science methods are **not** appropriate if (continued):

- 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 one will attempt to predict class membership using an unsupervised learning algorithm

  Example: clustering loan default data might lead to a cluster contains many defaulters. If new instances get added to this cluster, should they be viewed as loan defaulters?

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 is impossible.

- **Example:** can we use a model that predicts mortgage defaulters to also predict car loan defaulters?

BIASES, FALLACIES & INTERPRETATION

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

Extreme patterns can mislead.

Stay within a study's range.

Keep the base rate in mind.

Odd results sometimes happen (Simpson's Paradox).

BIASES, FALLACIES & INTERPRETATION

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.).

Does the presence of bias necessarily invalidate the results?

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

DATA SCIENCE MISTAKES

Mistake #1 – Selecting the wrong problem.

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

Mistake #3 – Not planning the data analysis process.

Mistake #4 – Insufficient business and domain knowledge.

Mistake #5 – Using incompatible data analysis tools.

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

DATA SCIENCE MISTAKES

Mistake #6 – Using tools that are too specific.

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

Mistake #8 – Running out of time.

Mistake #9 – Measuring results differently than the sponsor.

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

WHAT WE DIDN'T TALK ABOUT

Tons of classification and clustering algorithms

Deep learning and neural networks

Recommender systems

Data streams

Bayesian data analysis

Text mining and natural language processing

Data engineering

... and so much more!

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.

FUTURE TRENDS

New questions

New tools

New data sources

Data science as job component

Augmented/swarm intelligence