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]

LEARNING CONTEXT

STATISTICAL LEARNING

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

TYPES OF LEARNING

The central Data Science/Machine Learning problem is:

can (should) we design algorithms that can learn?

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)
- a **target** exists against which to train the model

TYPES OF LEARNING

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)
- the concept of a target is **not applicable**

Others:

- **semi-supervised learning** (teacher providing worked-out examples **and** a list of unsolved problems)
- **reinforcement learning** (embarking on a Ph.D. with an advisor?)

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)

ASSOCIATION RULES BASICS

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

Examples:

- bread and milk are often purchased together… is that interesting?
- hot dogs and mustard are also often purchased as a pair, but more rarely purchased individually… is that interesting?

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

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

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 $\underbrace{X}_{\text{premise}}$ then $\underbrace{Y}_{\text{conclusion}}$" 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

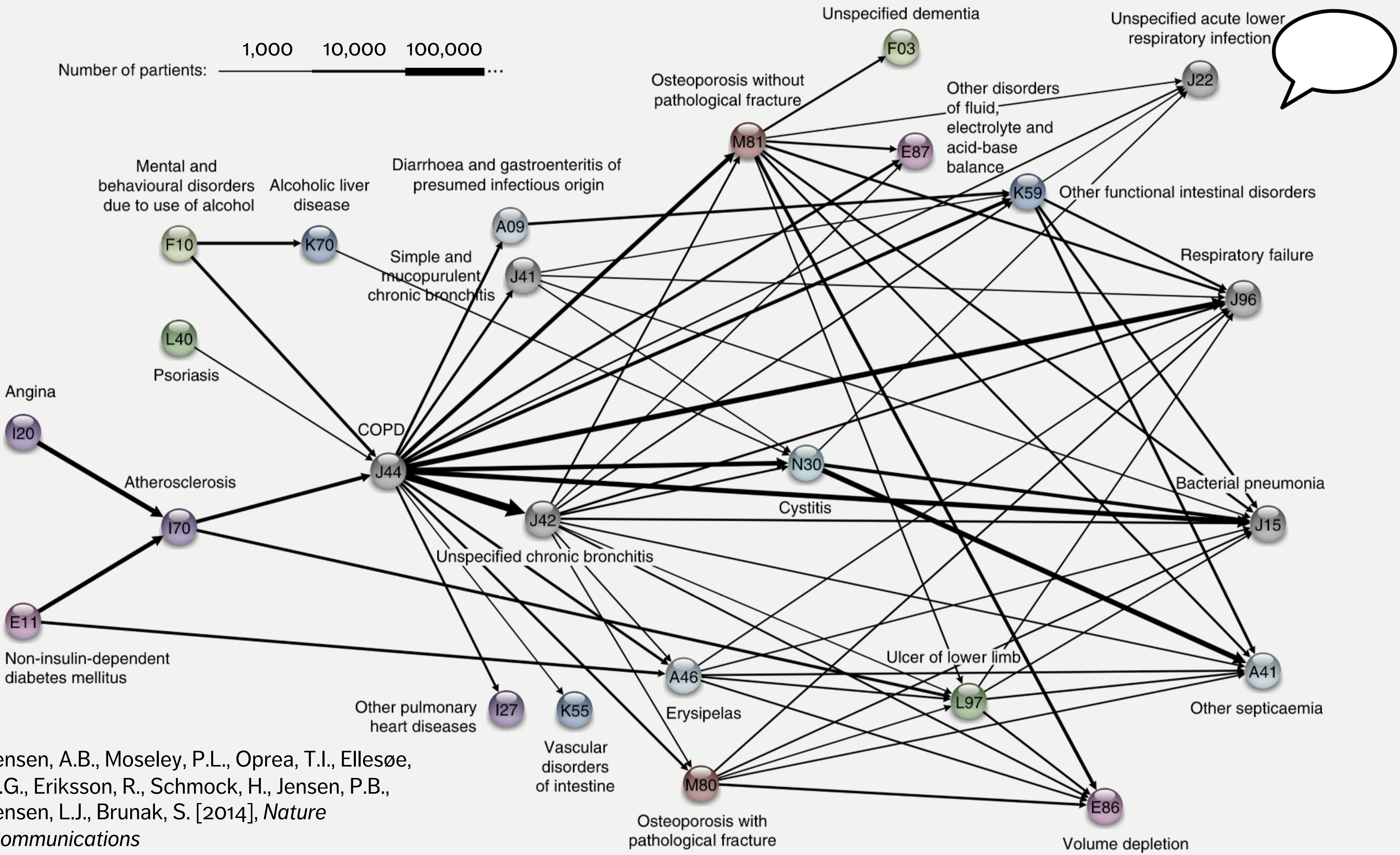
EXAMPLE

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

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.



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*

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

CLASSIFICATION METHODS

Logistic Regression

Neural Networks

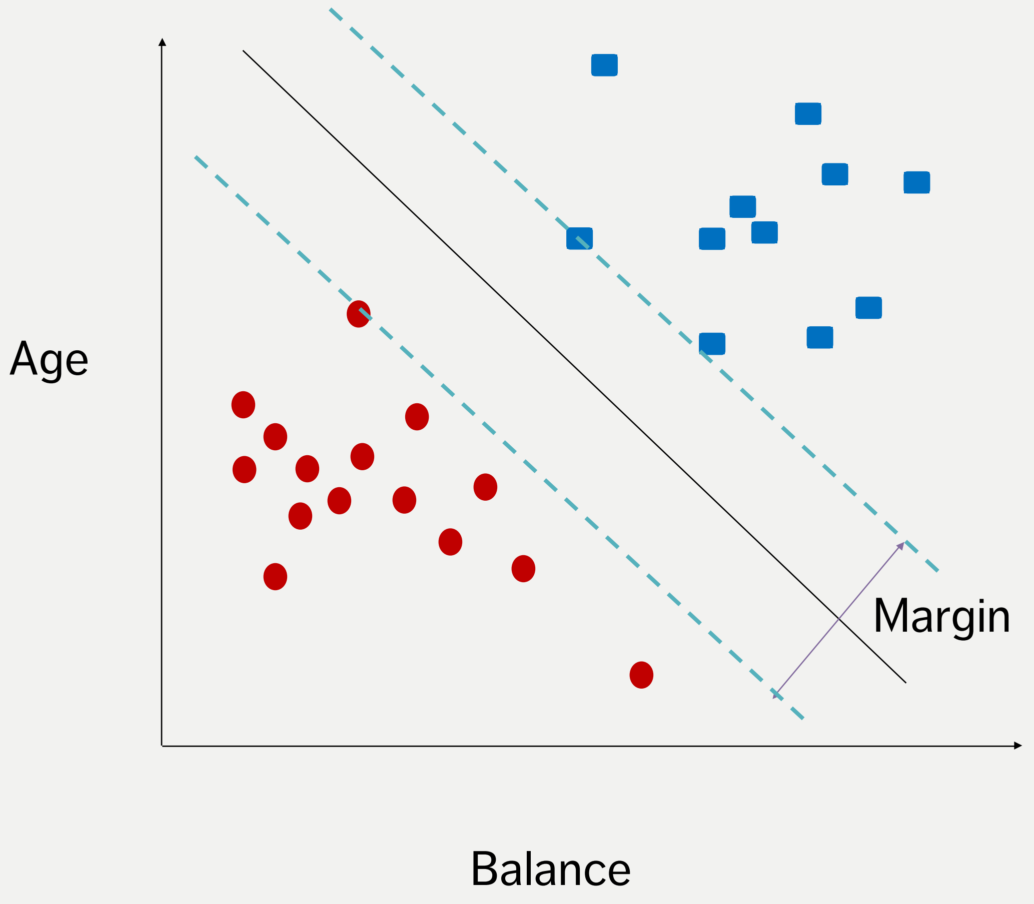
Decision Trees

Naïve Bayes Classifiers

Support Vector Machines

Nearest Neighbours Classifiers

etc.



[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**.

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

		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

CLUSTERING

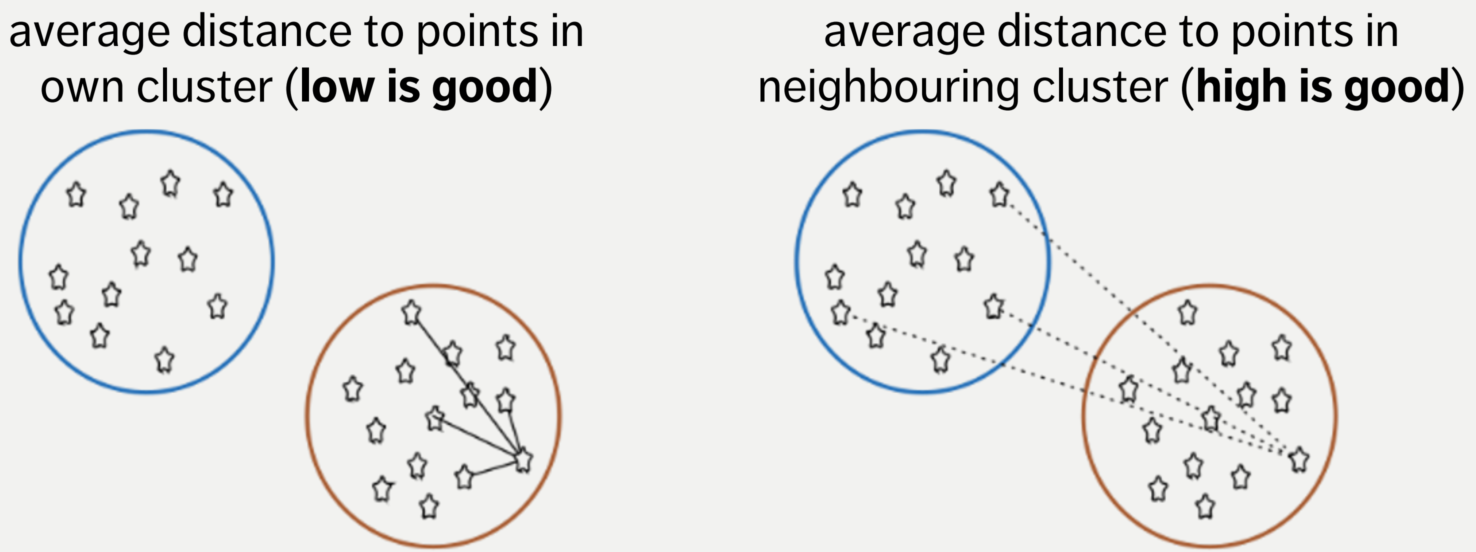
STATISTICAL LEARNING

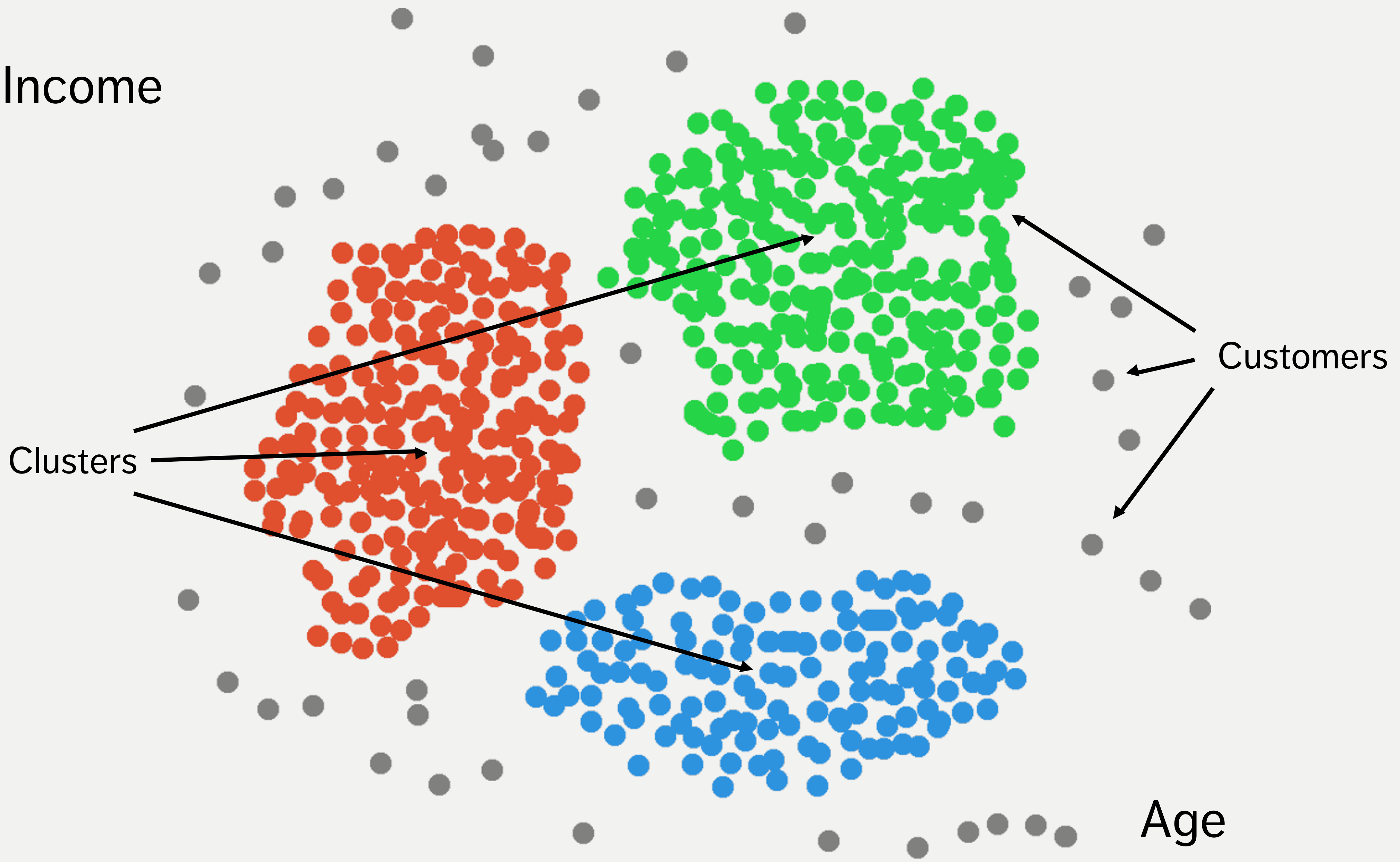
“Data is not information,
information is not knowledge,
knowledge is not understanding,
understanding is not wisdom.”
(C. Stoll)

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.




Income
Clusters
Customers
Age

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

CLUSTERING METHODS

k-Means

Hierarchical Clustering

Latent Dirichlet Allocation

Expectation-Maximization

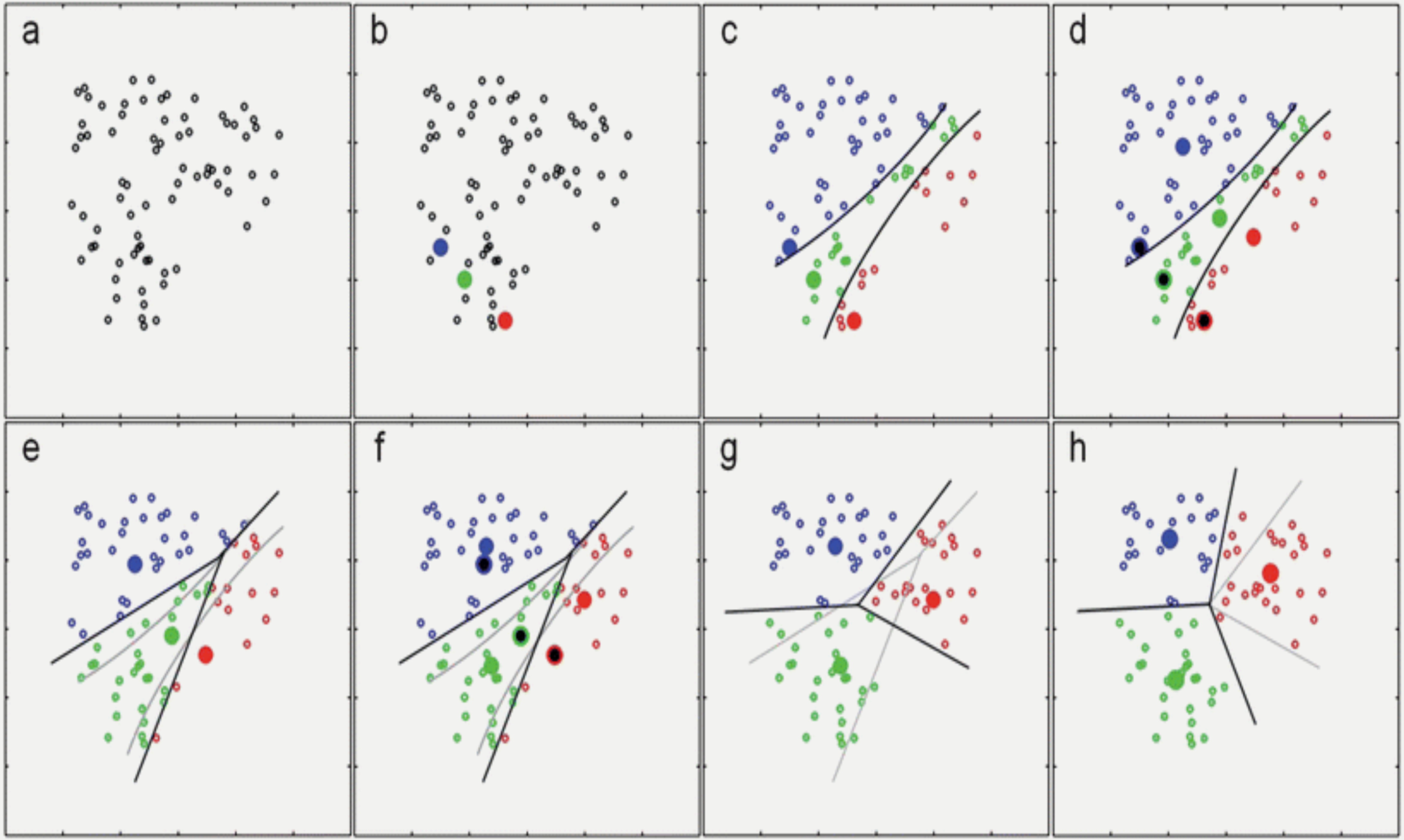
Balanced Iterative Reducing and Clustering using Hierarchies

Density-Based Spatial Clustering of Applications with Noise

Affinity Propagation

Spectral Clustering, etc.

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



CLUSTERING CHALLENGES

Automation

Lack of a clear-cut definition

Lack of repeatability

Number of clusters

Cluster description

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.)

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

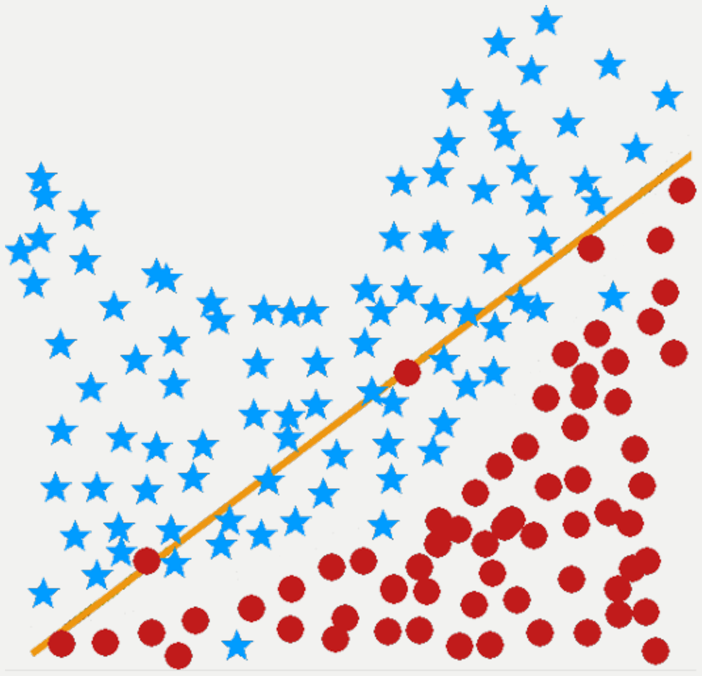
Is **close enough, good enough**?

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

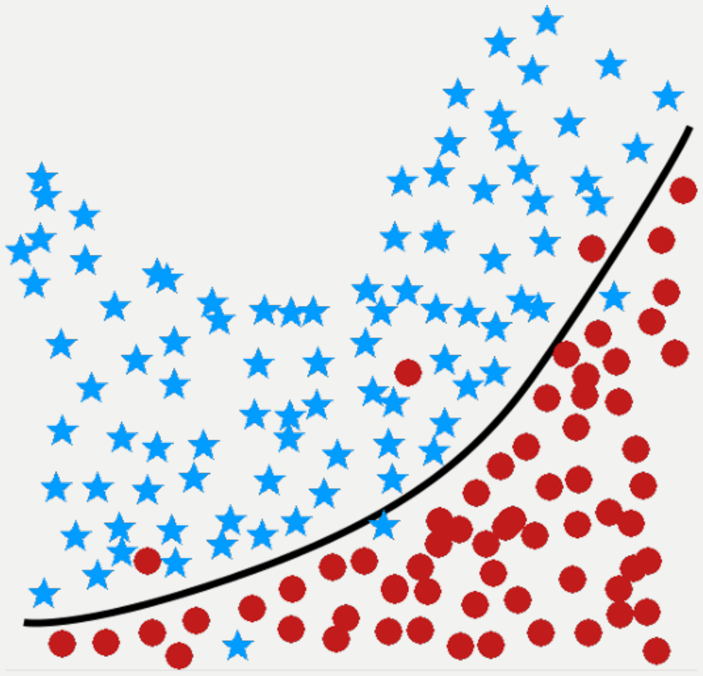
Seeking **perfection** (academic, professional, government, service data)

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

OVERFITTING



underfit



just right

overfit

BIG DATA VS. SMALL DATA

What is the main difference?

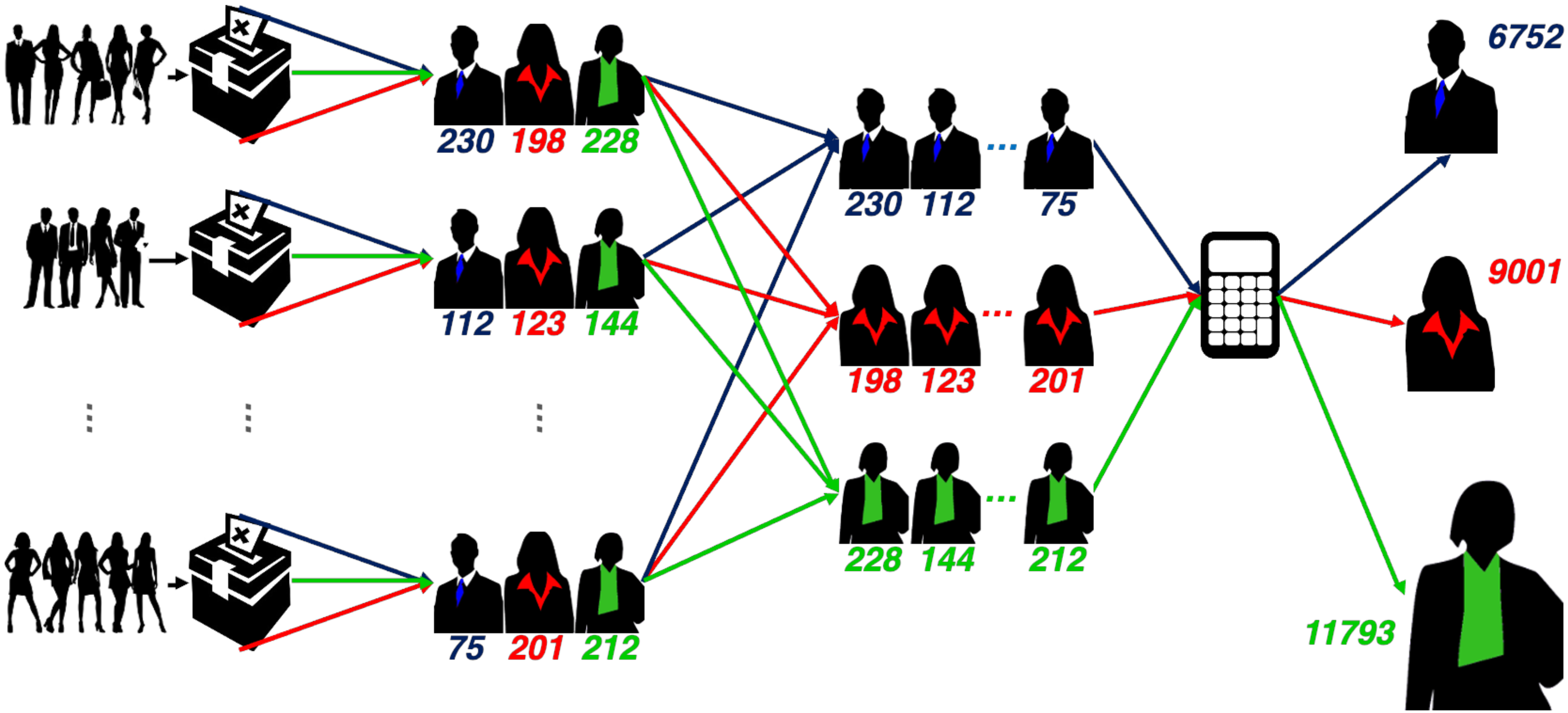
- datasets are **LARGE**
- issues: collection, capture, access, storage, analysis, visualization

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



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

APPROPRIATENESS & TRANSFERABILITY

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!")
- the dataset has attributes that usefully predict a value of interest, but which are not available when a prediction is required
- if one will attempt to predict class membership using an unsupervised learning algorithm

If data/model 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

Extreme patterns can mislead

Stay within a study's range

Keep the base rate in mind

Odd stuff happens (Simpson's Paradox)

Randomness plays a role

Human component to any analytical activity

Small effects can be (statistically) significant

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

Does bias necessarily invalidate the results?