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]

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 BASICS

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

Examples:

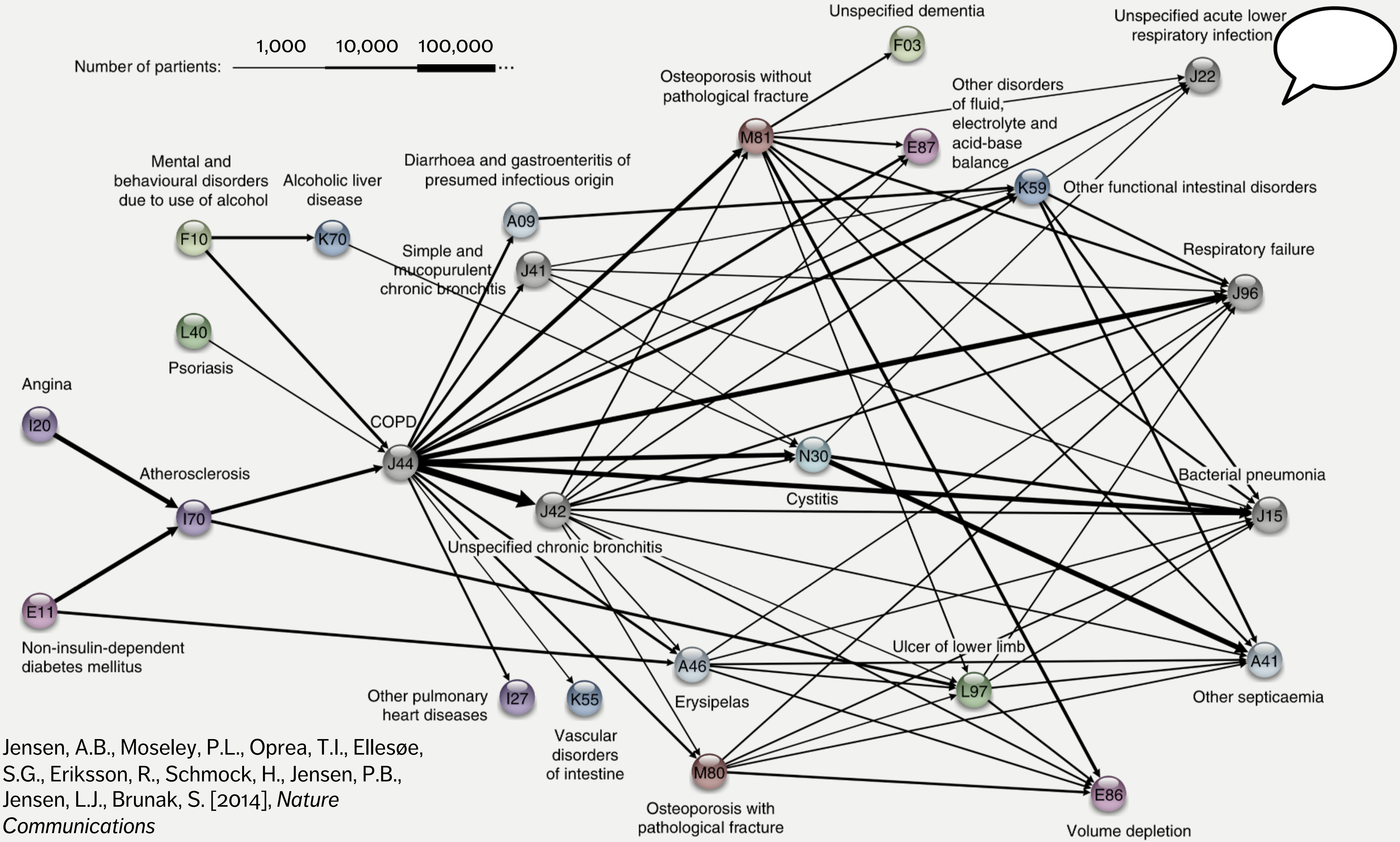
- bread and milk are often purchased together… is that interesting?
- hot dogs and mustard are also often purchased as a pair, but more rarely purchased individually… is that interesting?

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

[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



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 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.

CLASSIFICATION METHODS

Logistic Regression

Neural Networks

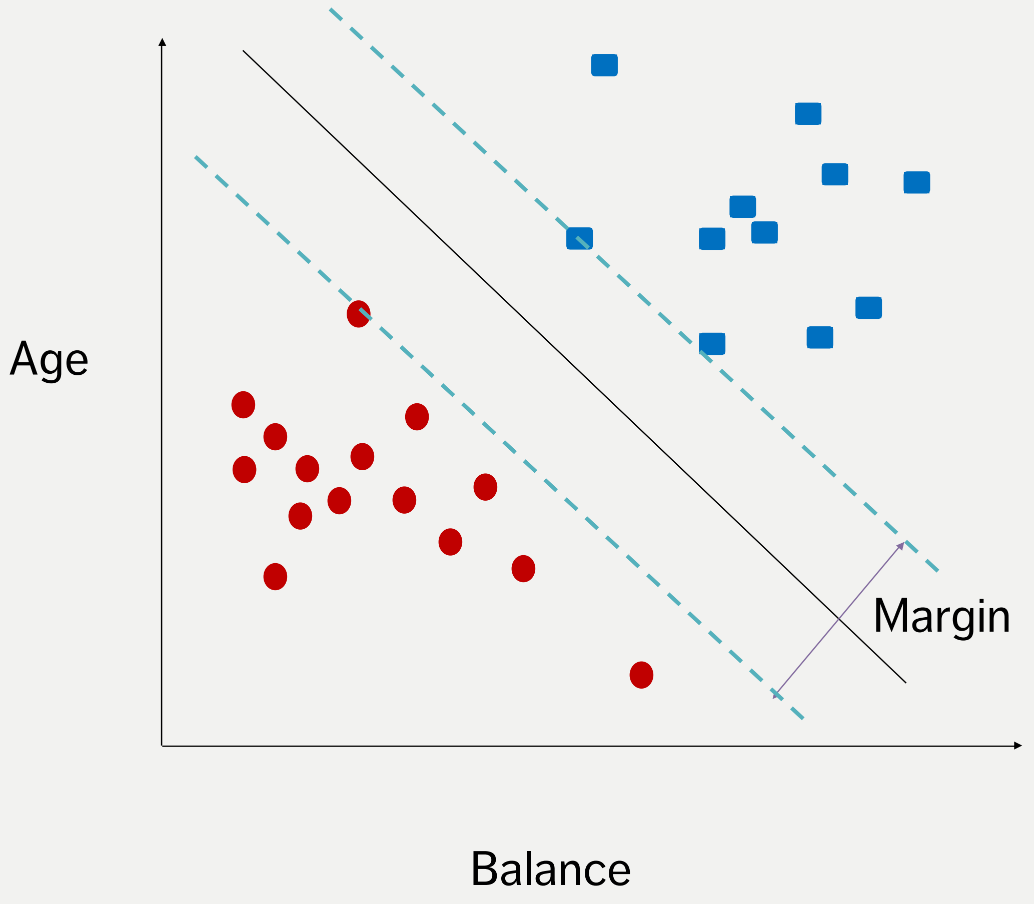
Decision Trees

Naïve Bayes Classifiers

Support Vector Machines

Nearest Neighbours Classifiers

etc.

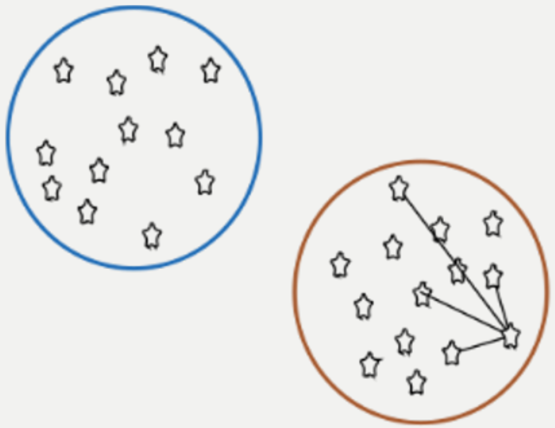


CLUSTERING OVERVIEW

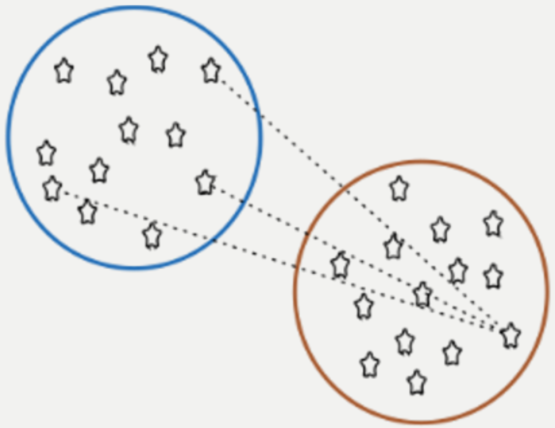
In **clustering**, the data is divided into **naturally occurring groups**. Within each group, the data points are **similar**; from group to group, they are **dissimilar**.

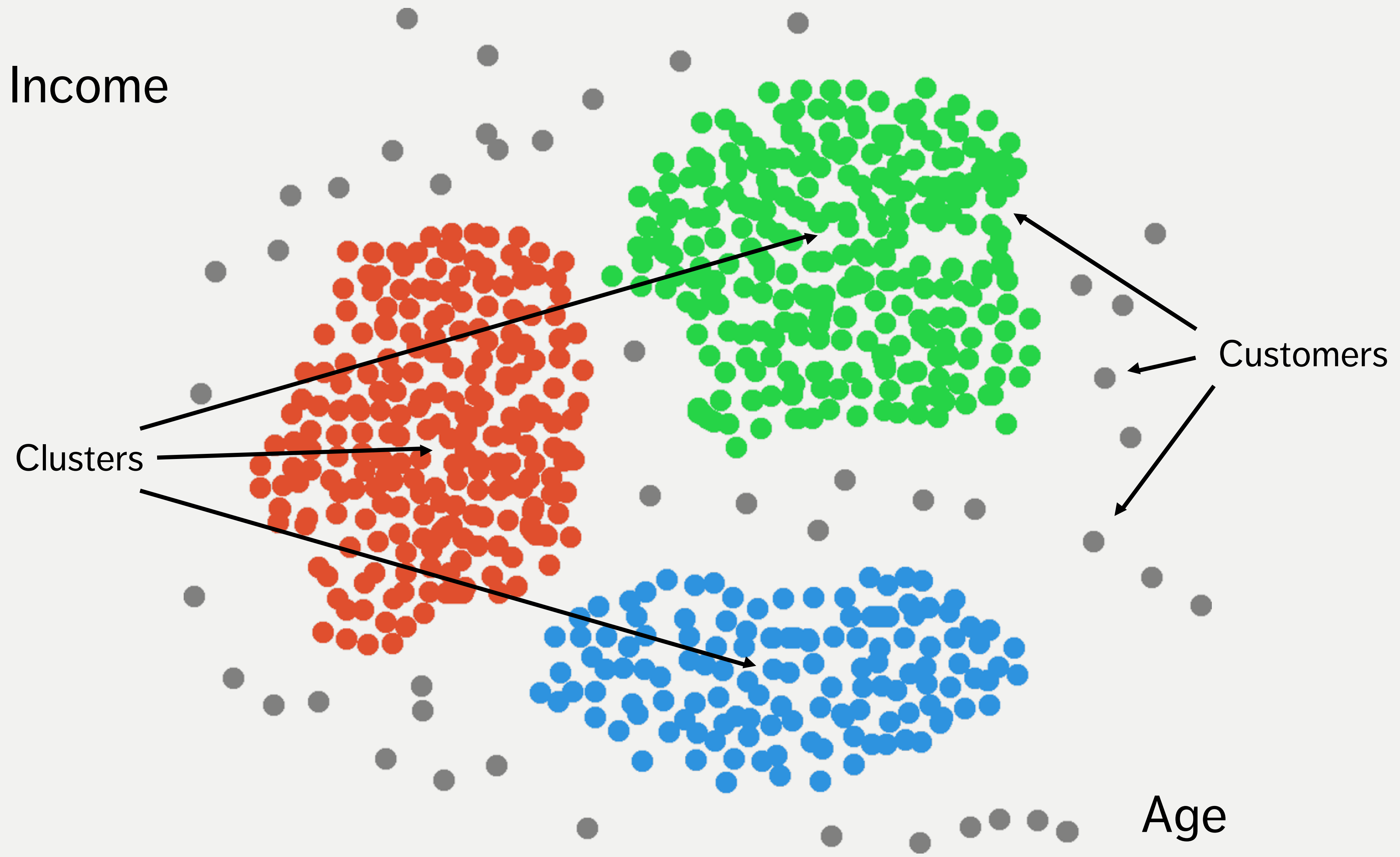
The grouping labels are not determined ahead of time, so clustering is an example of **unsupervised** learning.

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 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.

BAD DATA

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

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

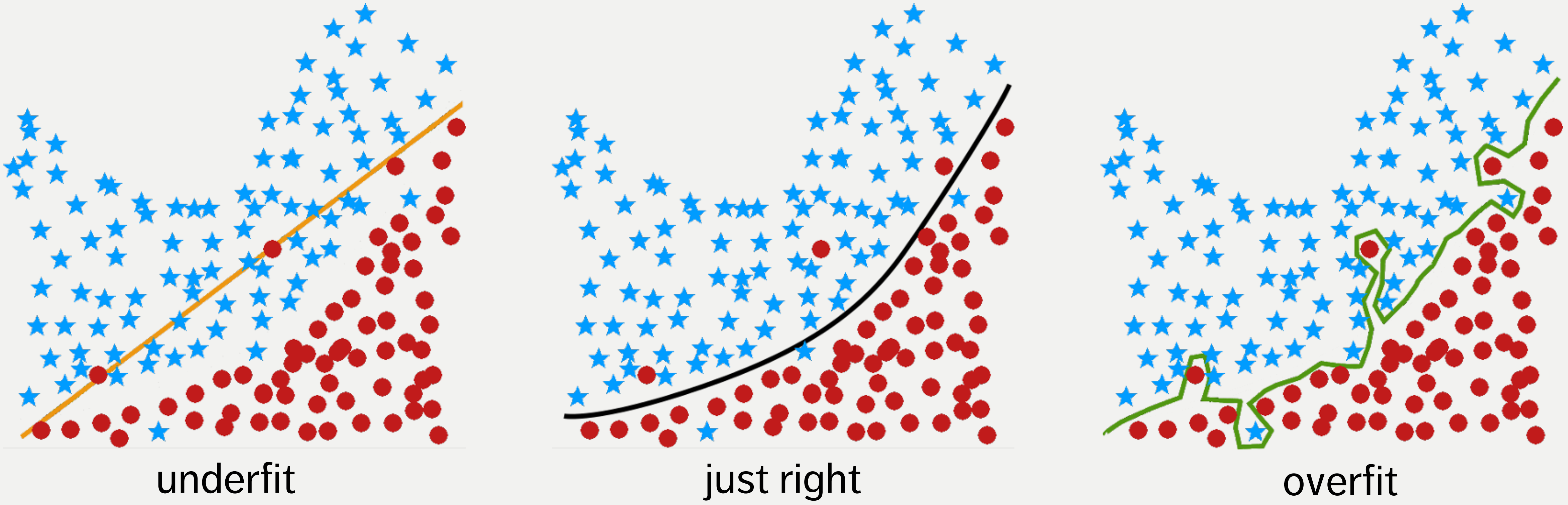
Is **close enough, good enough**?

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



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

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?