

Data Science
Data Science for Mobility



Motivations

Why Data Science Now?

“Necessity is the mother of invention”

Explosive Growth of Data

Pressing need for the automated analysis of massive data

Emerged in the late 1980s

Major developments in the mid 1990s.

Several names over the years. Based on Statistics, Machine Learning, Database technology

- 1960s: data collection, database creation, & network DBMS
- 1970s: relational data model, relational DBMS implementation
- 1980s: RDBMS, advanced data models (extended-relational, OO, deductive, etc.); application-oriented DBMS (spatial, scientific, engineering, etc.)
- 1990s: data mining, data warehousing, multimedia databases, and Web databases
- 2000s: stream data management and mining, web technology (XML, data integration), global information systems
- 2010s: social networks, NoSQL, unstructured data, etc.

- **Customer attrition**
 - Given customer information for the past months
 - Predict who is likely to attrite next month, or estimate customer value
- **Credit assessment**
 - Given a loan application
 - Predict whether the bank should approve the loan
- **Customer segmentation**
 - Given several information about the customers
 - Identify interesting groups among them
- **Community detection**
 - Given a social network of users
 - Identify community based on their connections (friendship relation, discussions, etc.)

Some terms you might have
heard other than data science ...

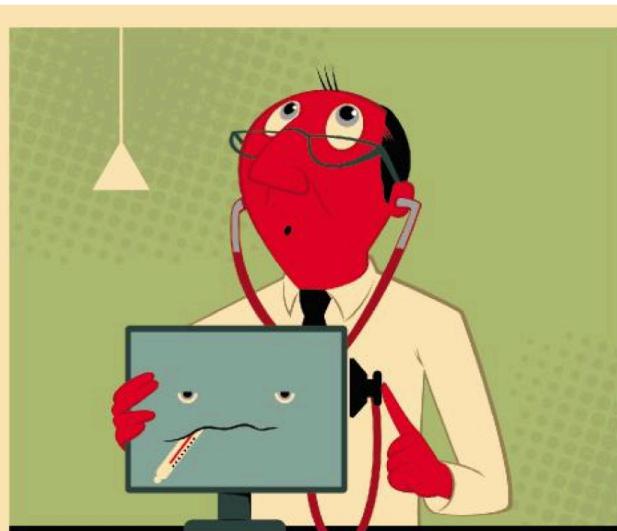
Big Data

BIG DATA

The Parable of Google Flu: Traps in Big Data Analysis

David Lazer,^{1,2*} Ryan Kennedy,^{1,3,4} Gary King,³ Alessandro Vespignani^{3,5,6}

In February 2013, Google Flu Trends (GFT) made headlines but not for a reason that Google executives or the creators of the flu tracking system would have hoped. *Nature* reported that GFT was predicting more than double the proportion of doctor visits for influenza-like illness (ILI) than the Centers for Disease Control and Prevention (CDC), which bases its estimates on surveillance reports from laboratories across the United States (1, 2). This happened despite the fact that GFT was built to predict CDC reports. Given that GFT is often held up as an exemplary use of big data (3–4), what lessons can we draw



Large errors in flu prediction were largely avoidable, which offers lessons for the use of big data.

the algorithm in 2009, and this model has run ever since, with a few changes announced in October 2013 (10, 15).

Although not widely reported until 2013, the new GFT has been persistently overestimating flu prevalence for a much longer time. GFT also missed by a very large margin in the 2011–2012 flu season and has missed high for 100 out of 108 weeks starting with August 2011 (see the graph). These errors are not randomly distributed. For example, last week's errors predict this week's errors (temporal autocorrelation), and the direction and magnitude of error varies with the

<http://simplystatistics.org/2014/05/07/why-big-data-is-in-trouble-they-forgot-about-applied-statistics/>

http://bits.blogs.nytimes.com/2014/03/28/google-flu-trends-the-limits-of-big-data/?_r=0

<http://gking.harvard.edu/files/gking/files/0314policyforumff.pdf>

Data Mining

Data Mining

The non-trivial process of identifying
(1) valid, (2) novel, (3) potentially useful,
and (4) understandable patterns in data.

An Example Using Contact Lens Data

11

Age	Spectacle prescription	Astigmatism	Tear production rate	Recommended lenses
Young	Myope	No	Reduced	None
Young	Myope	No	Normal	Soft
Young	Myope	Yes	Reduced	None
Young	Myope	Yes	Normal	Hard
Young	Hypermetrope	No	Reduced	None
Young	Hypermetrope	No	Normal	Soft
Young	Hypermetrope	Yes	Reduced	None
Young	Hypermetrope	Yes	Normal	hard
Pre-presbyopic	Myope	No	Reduced	None
Pre-presbyopic	Myope	No	Normal	Soft
Pre-presbyopic	Myope	Yes	Reduced	None
Pre-presbyopic	Myope	Yes	Normal	Hard
Pre-presbyopic	Hypermetrope	No	Reduced	None
Pre-presbyopic	Hypermetrope	No	Normal	Soft
Pre-presbyopic	Hypermetrope	Yes	Reduced	None
Pre-presbyopic	Hypermetrope	Yes	Normal	None
Presbyopic	Myope	No	Reduced	None
Presbyopic	Myope	No	Normal	None
Presbyopic	Myope	Yes	Reduced	None
Presbyopic	Myope	Yes	Normal	Hard
Presbyopic	Hypermetrope	No	Reduced	None
Presbyopic	Hypermetrope	No	Normal	Soft
Presbyopic	Hypermetrope	Yes	Reduced	None
Presbyopic	Hypermetrope	Yes	Normal	None

An example of possible pattern

if astigmatism = yes
and tear production rate = normal
and spectacle prescription = myope
then recommendation = hard

- Is it valid?
 - The pattern has to be valid with respect to a certainty level (rule true for the 86%)
- Is it novel?
 - Is the relation between astigmatism and hard contact lenses already well-known?
- Is it useful? Is it actionable?
 - The pattern should provide information useful to the bank for assessing credit risk
- Is it understandable?

Machine Learning

“A computer program is said to learn from experience E with respect to some class of task T and a performance measure P, if its performance at tasks in T, as measured by P, improves because of experience E.”

- Suppose we have the experience E encoded as a dataset,

$$D = x_1, x_2, x_3, \dots, x_N$$

- **Supervised Learning**

- Given the desired outputs $t_1, t_2, t_3, \dots, t_N$ learns to produce the correct output given a new set of input

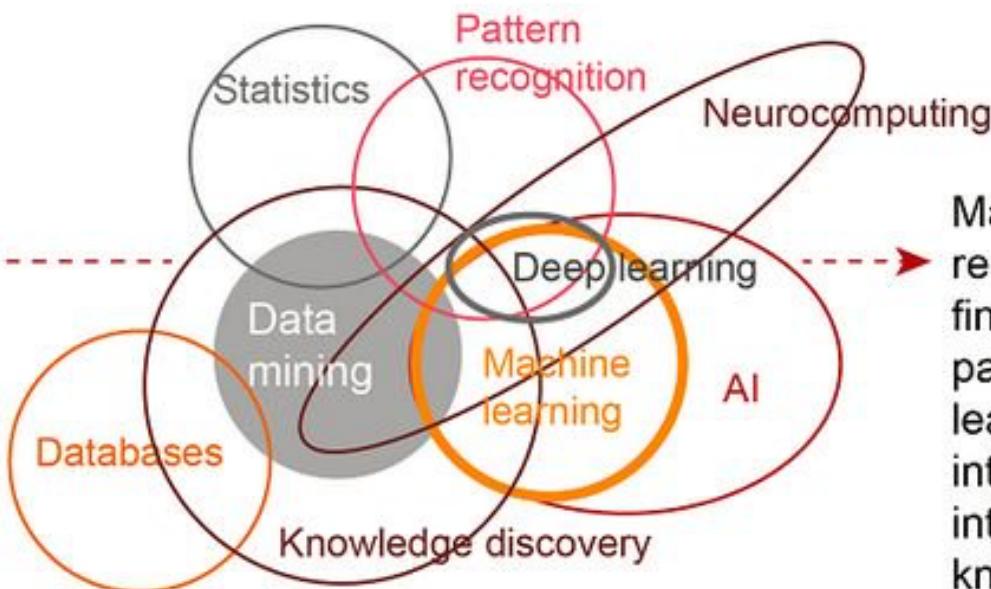
- **Unsupervised learning**

- Exploits regularities in D to build a representation to be used for reasoning or prediction

- **Reinforcement learning**

- Producing actions $a_1, a_2, a_3, \dots, a_N$ which affect the environment, and receiving rewards $r_1, r_2, r_3, \dots, r_N$ learn to act in order to maximize rewards in the long term

How does machine learning relate to artificial intelligence?



Machine learning is a category of research and algorithms focused on finding patterns in data and using those patterns to make predictions. Machine learning falls within the artificial intelligence (AI) umbrella, which in turn intersects with the broader field of knowledge discovery and data mining.

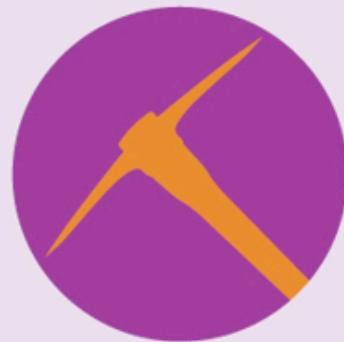
Source: SAS, 2014 and PwC, 2016

Data Science?

What is data science?

Data science can be broken down into four essential parts.

Mining data



Collecting and formatting
the information

Statistics



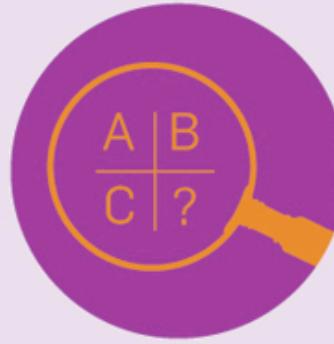
Information analysis

Interpret



Representation or visualization in
the form of presentations,
infographics, graphs or charts

Leverage



Implications of the data,
application of the data, interaction
using the data and predictions
formed from studying it

What is the goal?

- Build computer programs that navigate through databases automatically, seeking regularities or patterns
- There will be problems
 - Most patterns are uninteresting
 - Most patterns are spurious, inexact, or contingent on accidental coincidences in the particular dataset used
 - Real data is imperfect: Some parts will be garbled, and some will be missing
- Algorithms need to be robust enough to cope with imperfect data and to extract regularities that are inexact but useful

Descriptive vs. Predictive analytics

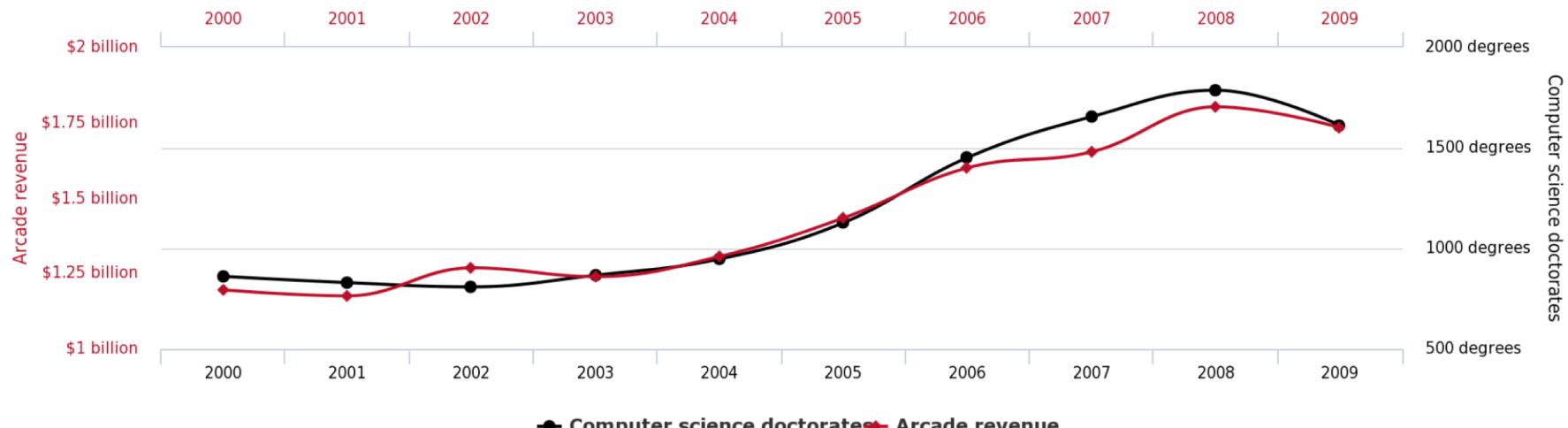
Are the models built for get insight?

Are the models built to predict?

“Prescriptive” analytics
applies descriptive and predictive mining
to recommend a course of action

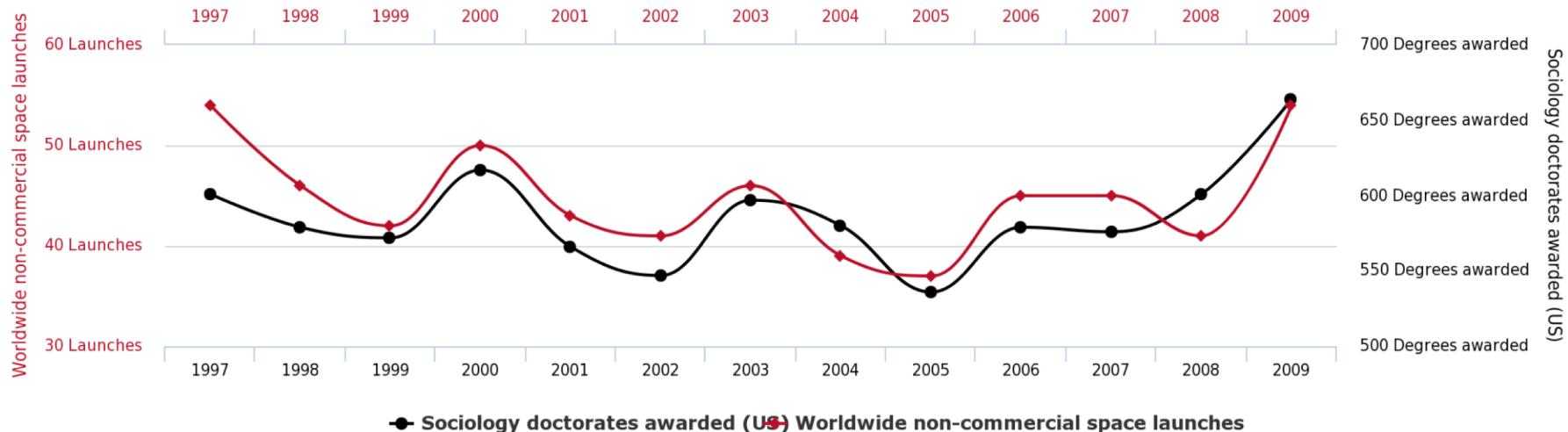
Pitfalls

Total revenue generated by arcades
correlates with
Computer science doctorates awarded in the US



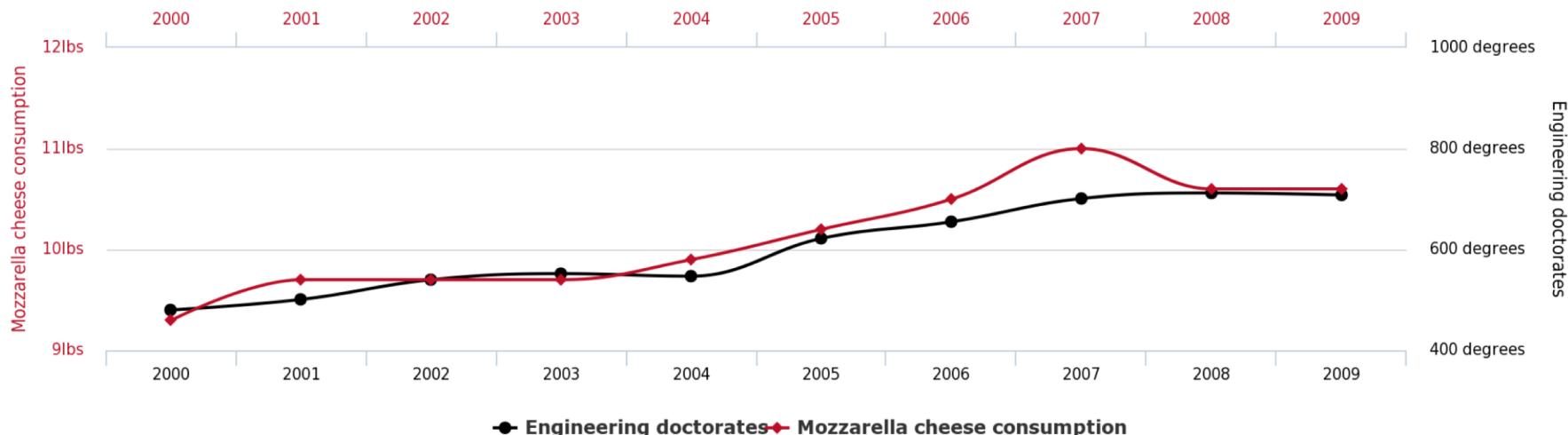
<http://www.tylervigen.com>

Worldwide non-commercial space launches correlates with Sociology doctorates awarded (US)



<http://www.tylervigen.com>

Per capita consumption of mozzarella cheese
correlates with
Civil engineering doctorates awarded



<http://www.tylervigen.com>

The Process

Interesting Question

- What is the scientific goal
- What do you want to predict/estimate?

Get the Data

- How the data was sampled?
- Which data is relevant? Are there any privacy issue?

Explore the Data

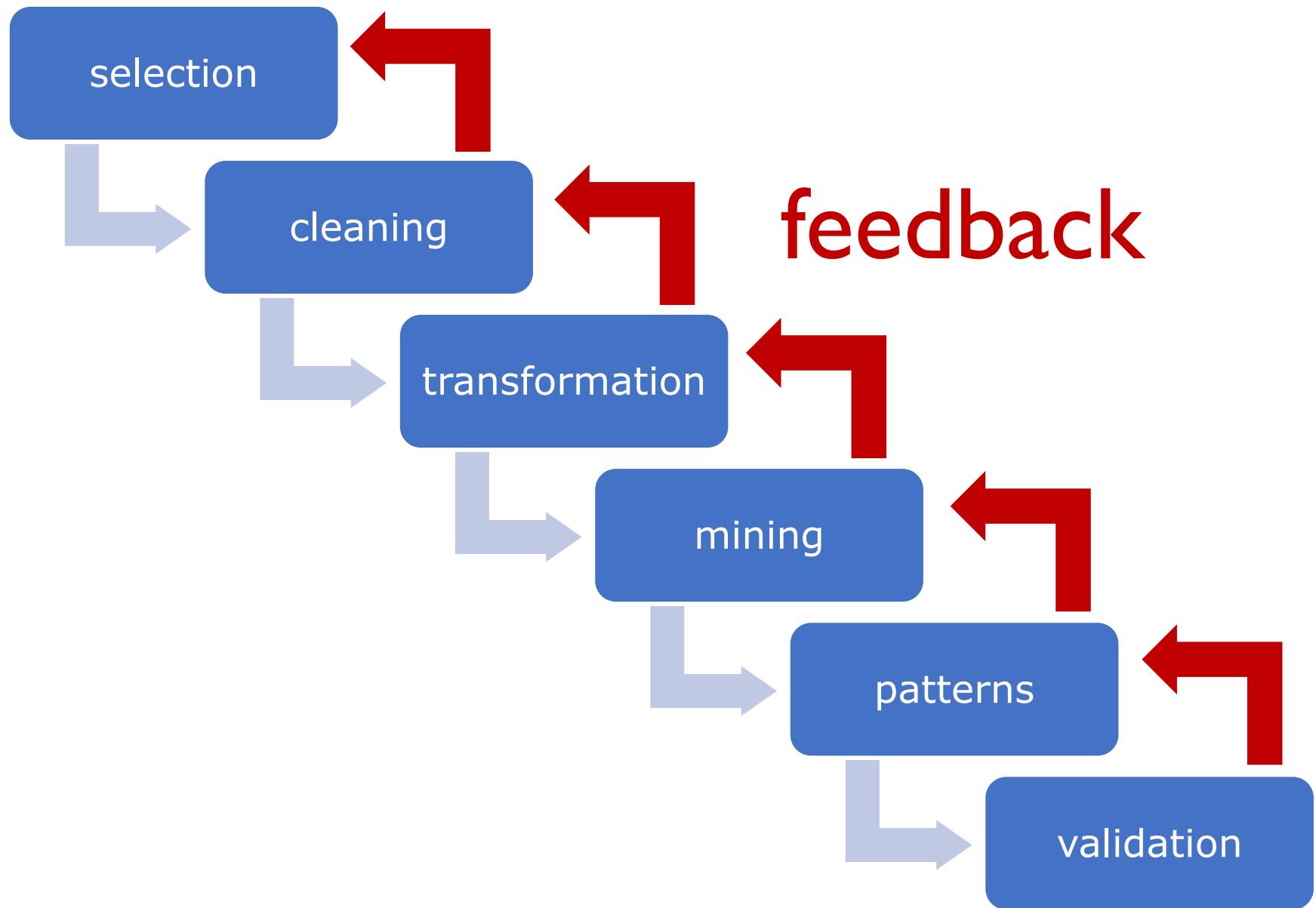
- Plot the data, compute statistics, search for anomalies
- Which data is relevant? Are there any privacy issue?

Build the Model

- Build, fit, and validate the model

Communicate the Results

- What did we learn? Is there a story?

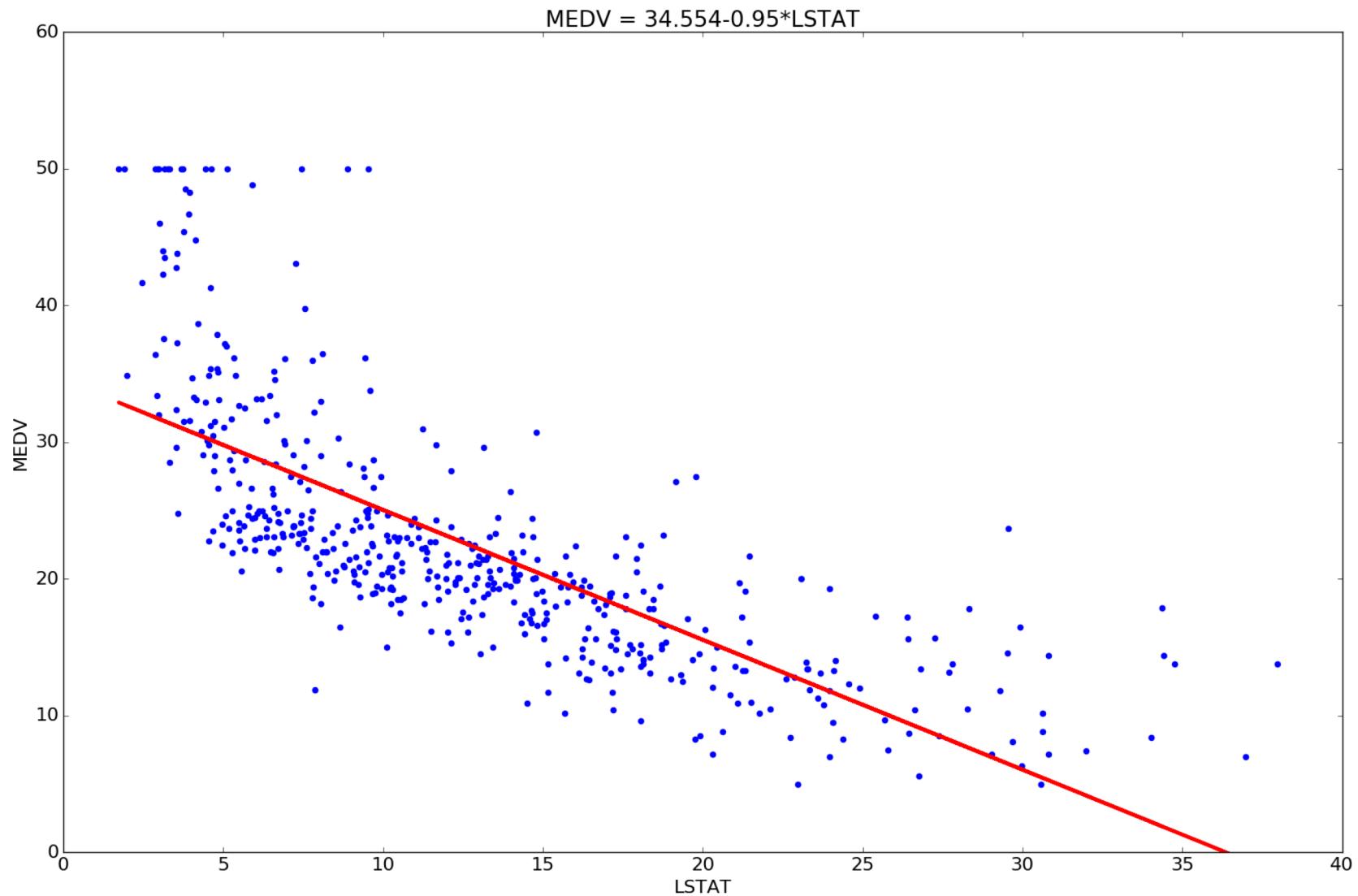


- **Selection**
 - What are data we actually need to answer the posed question?
- **Cleaning**
 - Are there any errors or inconsistencies in the data we need to eliminate?
- **Transformation**
 - Some variables might be eliminated because equivalent to others
 - Some variables might be elaborated to create new variables
(e.g. birthday to age, daily measures into weekly/monthly measures, log?)
- **Mining**
 - Select the mining approach: classification, regression, association, etc.
 - Choose and apply the mining algorithm(s)
- **Validation**
 - Are the patterns we discovered sound? According to what criteria?
 - Are the criteria sound? Can we explain the result?

What type of tasks?

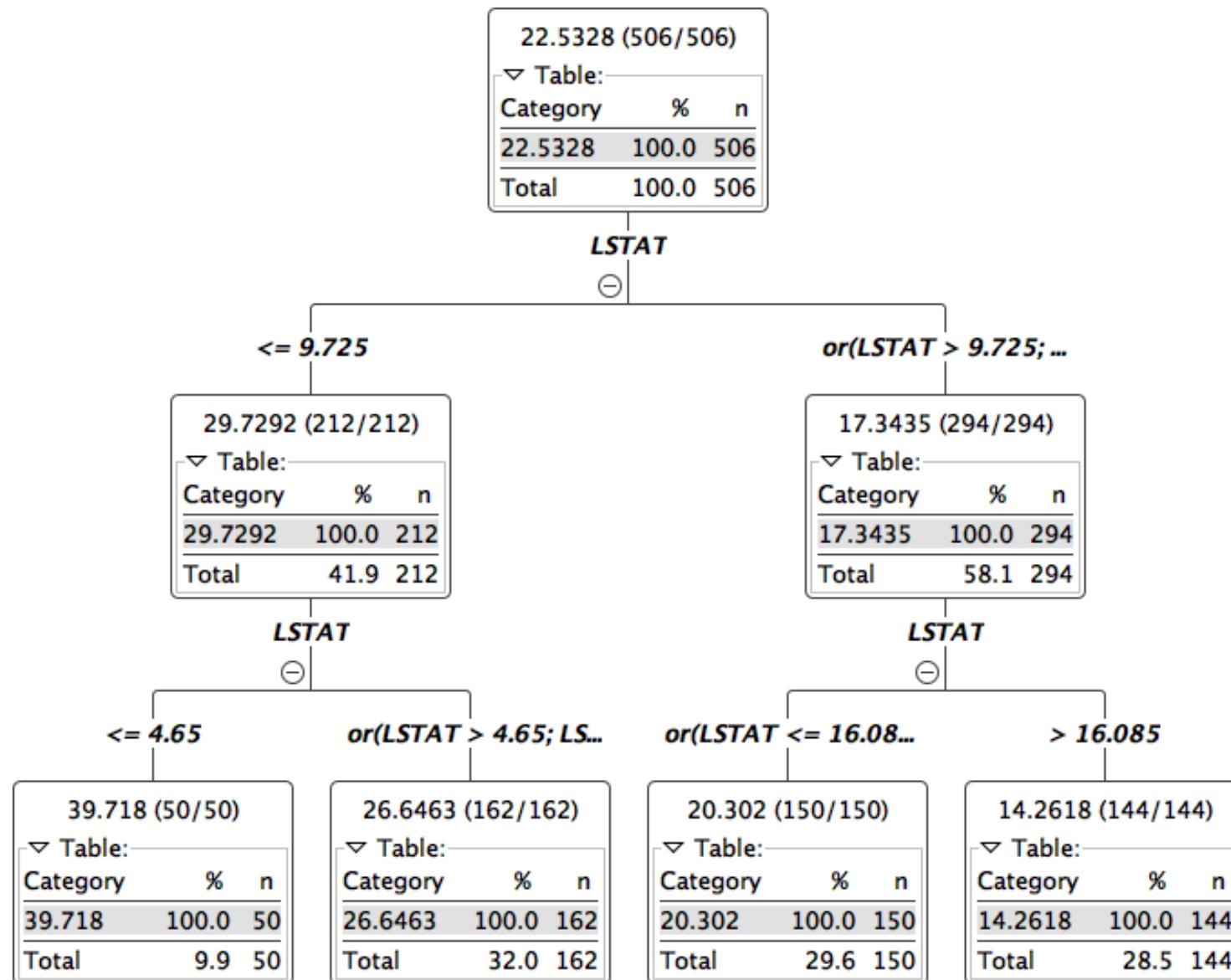
Prediction & Regression

- Information collected by the U.S Census Service concerning housing in the area of Boston (MA)
<https://archive.ics.uci.edu/ml/machine-learning-databases/housing/>
- 506 cases 14 variables: CRIM - per capita crime rate by town; ZN - proportion of residential land zoned for lots over 25,000 sq.ft.; INDUS - proportion of non-retail business acres per town; CHAS - Charles River dummy variable (1 if tract bounds river; 0 otherwise); NOX - nitric oxides concentration (parts per 10 million); RM - average number of rooms per dwelling; AGE - proportion of owner-occupied units built prior to 1940; DIS - weighted distances to five Boston employment centres; RAD - index of accessibility to radial highways; TAX - full-value property-tax rate per \$10,000; PTRATIO - pupil-teacher ratio by town; B - $1000(Bk - 0.63)^2$ where Bk is the proportion of blacks by town; LSTAT - % lower status of the population; MEDV - Median value of owner-occupied homes in \$1000's
- MEDV is the target variable



Input variable: LSTAT - % lower status of the population

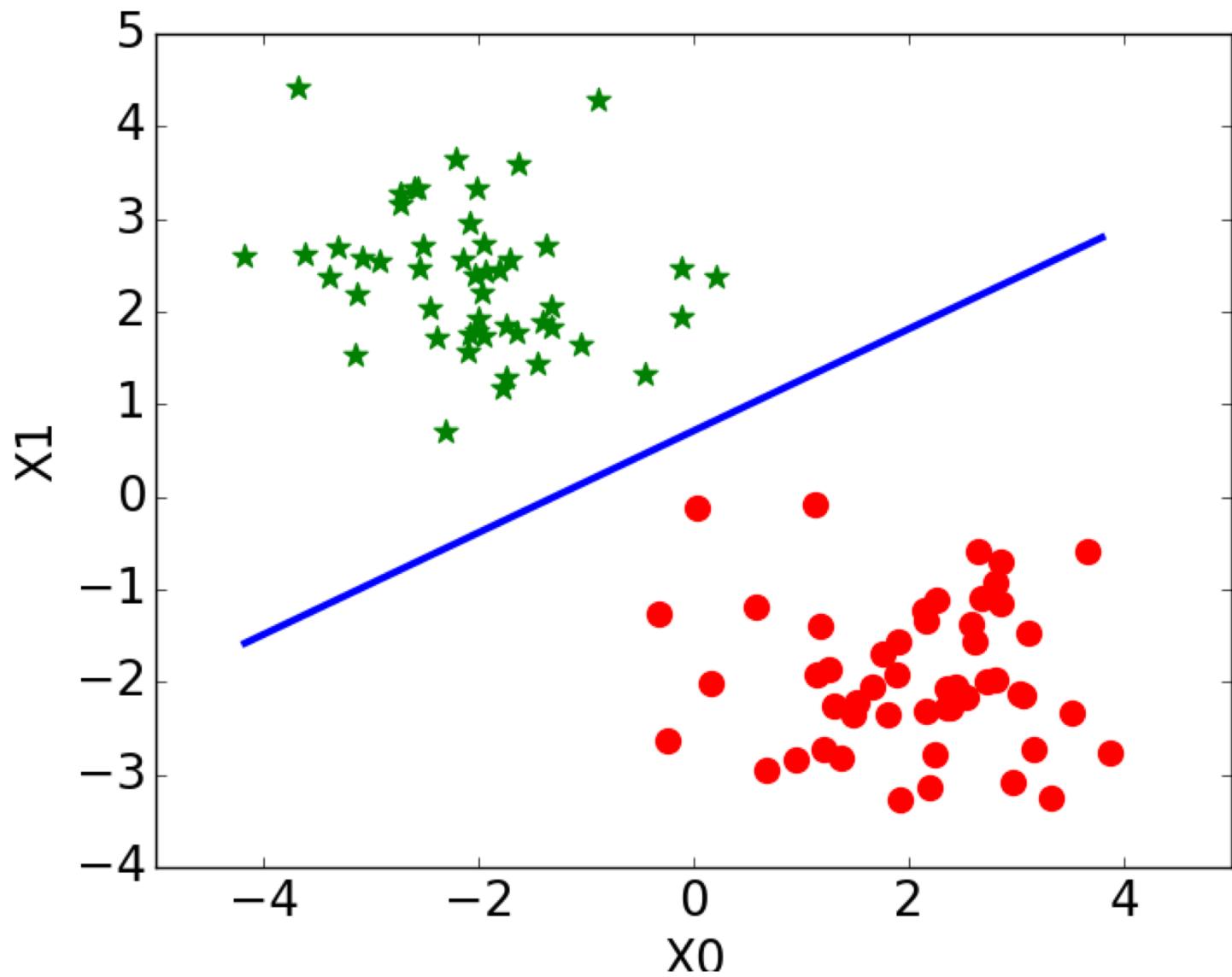
Output variable: MEDV - Median value of owner-occupied homes in \$1000's



Input variable: LSTAT - % lower status of the population

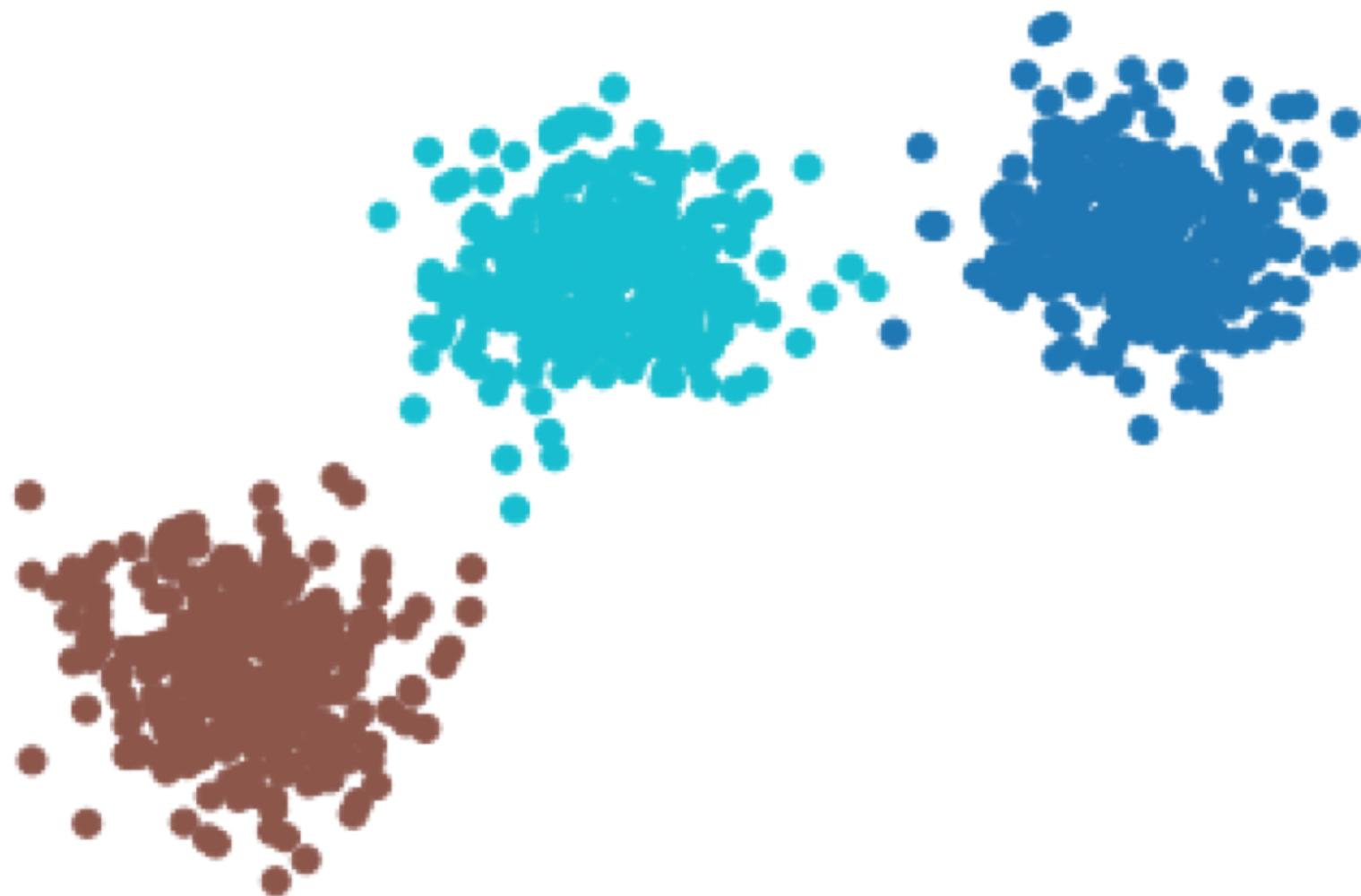
Output variable: MEDV - Median value of owner-occupied homes in \$1000's

Classification



Clustering





Associations

Bread
Peanuts
Milk
Fruit
Jam

Bread
Jam
Soda
Chips
Milk
Fruit

Steak
Jam
Soda
Chips
Bread

Jam
Soda
Peanuts
Milk
Fruit

Is there something interesting to be noted?

Jam
Soda
Chips
Milk
Bread

Fruit
Soda
Chips
Milk

Fruit
Soda
Peanuts
Milk

Fruit
Peanuts
Cheese
Yogurt

- Outlier analysis
 - An outlier is a data object that does not comply with the general behavior of the data
 - It can be considered as noise or exception but is quite useful in fraud detection, rare events analysis
- Trend and evolution analysis
 - Trend and deviation: regression analysis
 - Sequential pattern mining, periodicity analysis
 - Similarity-based analysis
- Text Mining, Topic Modeling, Graph Mining, Data Streams
- Sentiment Analysis, Opinion Mining, etc.
- Other pattern-directed or statistical analyses

Relevant Issues

- A data science/data analytics/data mining process may generate thousands of patterns, but typically not all of them are interesting.
- **Interestingness measures**
 - A pattern is interesting if it is easily understood by humans, valid on new or test data with some degree of certainty, potentially useful, novel, or validates some hypothesis that a user seeks to confirm
- **Objective vs. subjective interestingness measures**
 - Objective measures are based on statistics and structures of patterns
 - Subjective measures are based on user's belief in the data, e.g., unexpectedness, novelty, etc.