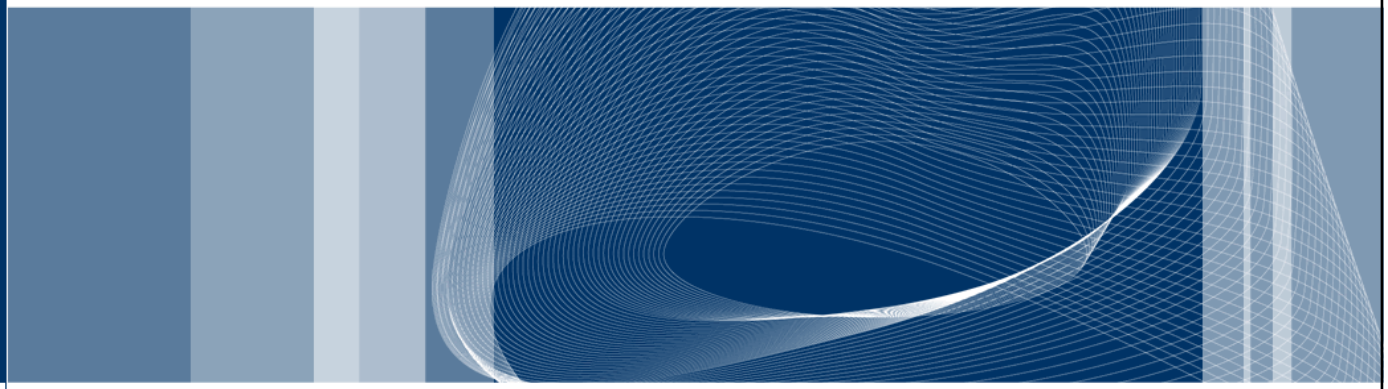
POLITECNICO DI MILANO



Data Mining

Data Mining and Text Mining

Prof. Pierluca Lanzi

Motivations

Why?

"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

Evolution of 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.
- 2020s: …

Examples

- 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
  - Who is discussing what?
- Sentiment Analysis
  - …

Big Data?

Massive datasets

Gigantic quantities of data are being collected in every domain imaginable: science, agriculture, healthcare, transport, sport …

Exponential increases in computing power has led to ever decreasing costs of instrumentalization and storage

Big Data

The quantity of data of quantity is so large that
(1) it facilitates predictions not previously possible, and
(2) it necessitates special processes to deal with
Quantity could be total storage, arrival rate, …

FINAL FINAL

POLICYFORUM

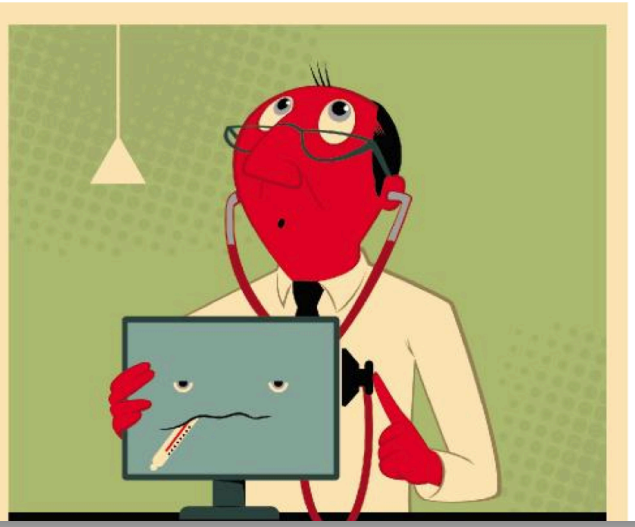
BIG DATA

The Parable of Google Flu: Traps in Big Data Analysis

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

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



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

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,

$$\mathrm{D} = \mathrm{x}_1, \mathrm{x}_2, \mathrm{x}_3, \ldots, \mathrm{x}_\mathrm{N}$$

- Supervised Learning
  - Given the desired outputs $t_1, t_2, t_3, \ldots, t_N$ 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 $a_1, a_2, a_3, \ldots, a_N$ which affect the environment, and receiving rewards $r_1, r_2, r_3, \ldots, r_N$ learn to act in order to maximize rewards in the long term

Data Science?

http://wikibon.org/blog/role-of-the-data-scientist/

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

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 a "Good" Pattern?

- 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?

What is the General Idea?

- Build computer programs that navigate through databases automatically, seeking regularities or patterns

- However …
  - 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

Are the models built for gaining insight?
(about what already happened)

Or are they built for accurate prediction?
(about what might happen)

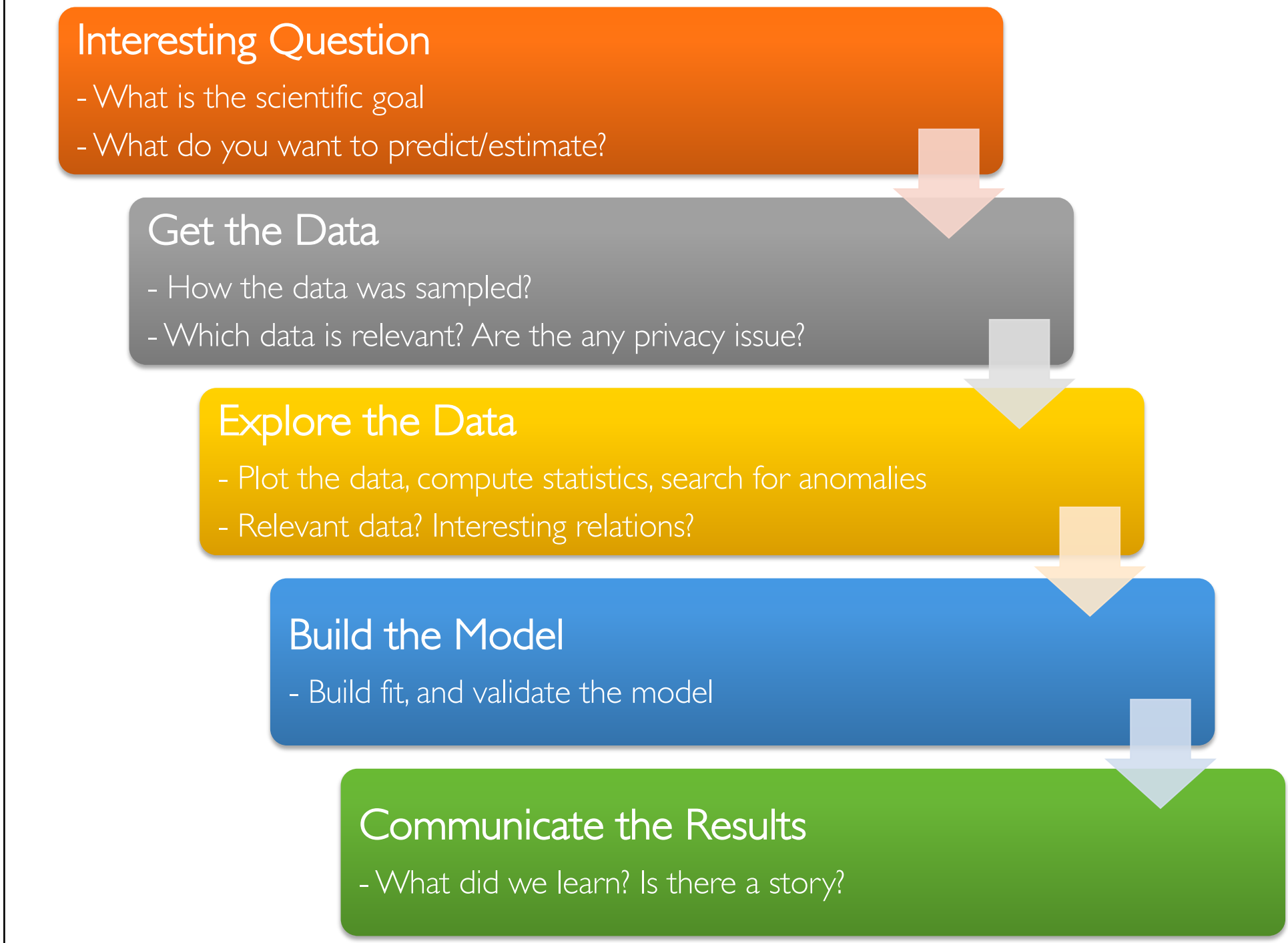
I want to analyze who are the customers that shop at my store more than twice a week

I need a model to help me investing in the stock market!

...

"Prescriptive"
use descriptive and predictive mining
to recommend a course of action

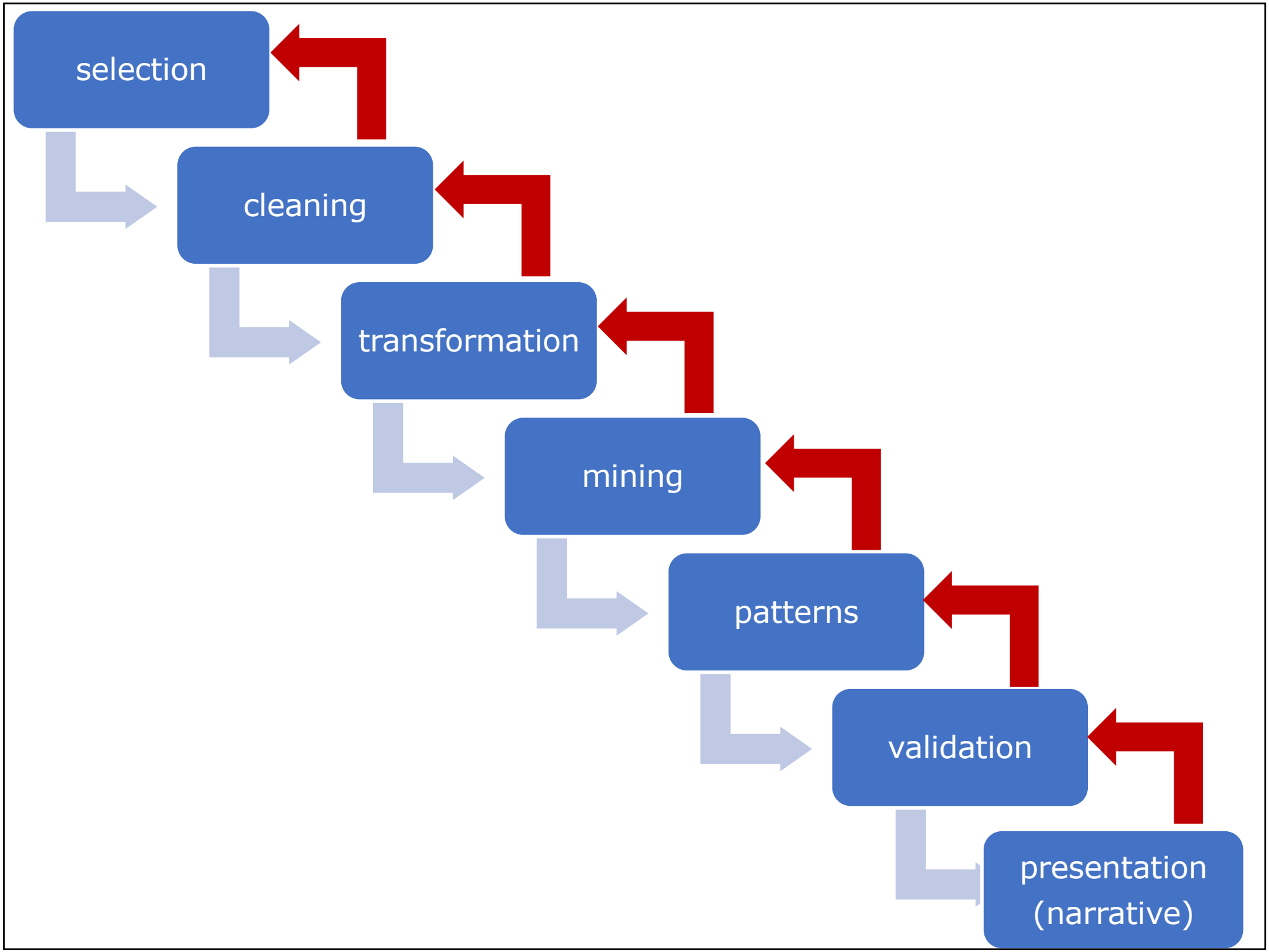
The Process


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 the any privacy issue?
Explore the Data
- Plot the data, compute statistics, search for anomalies
- Relevant data? Interesting relations?
Build the Model
- Build fit, and validate the model
Communicate the Results
- What did we learn? Is there a story?

What did we learn?

Is there a story?

Can we build a narrative?


selection
cleaning
transformation
mining
patterns
validation
presentation
(narrative)

What Are the Main Steps?

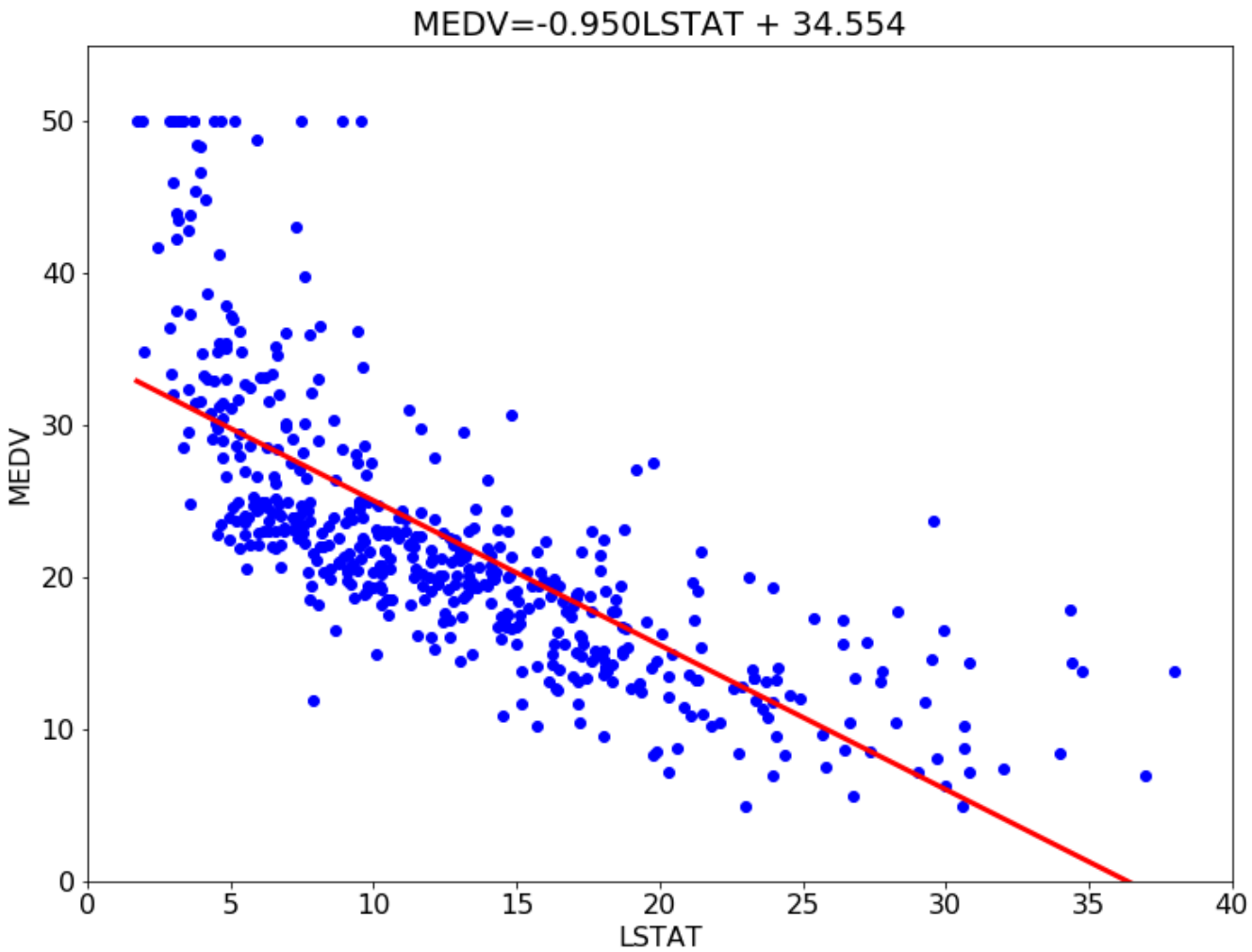
- Selection
  - What are data we actually need to answer the posed question?
- Cleaning
  - Are there any errors or inconsistencies in the data we need to eliminate?
- Transformation
  - Some variables might be eliminated because equivalent to others
  - Some variables might be elaborated to create new 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?
- Presentation & Narrative
  - What did we learn? Is there a story to tell? A take-home message?

Data Mining Tasks

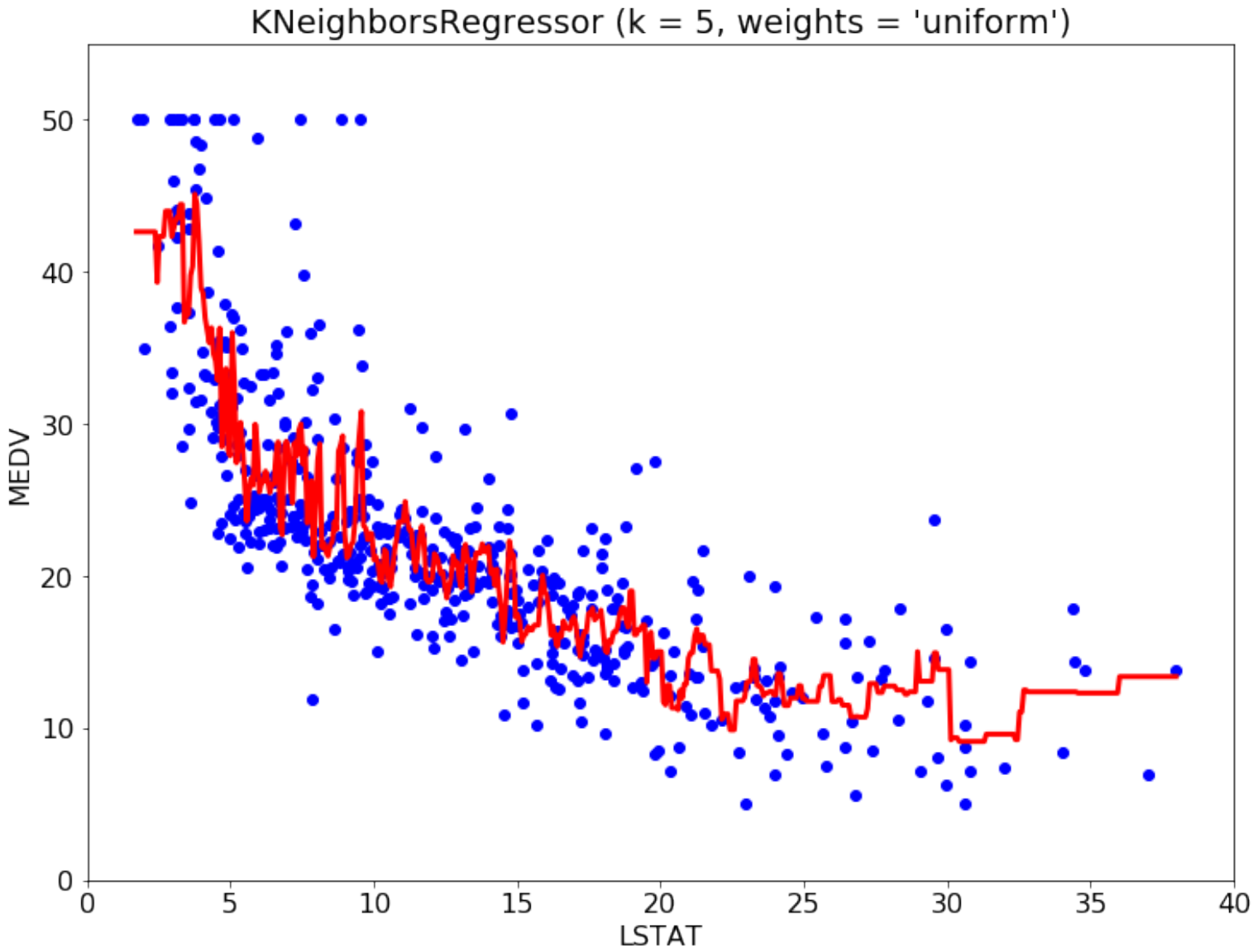
Prediction & Regression

Boston Housing Dataset

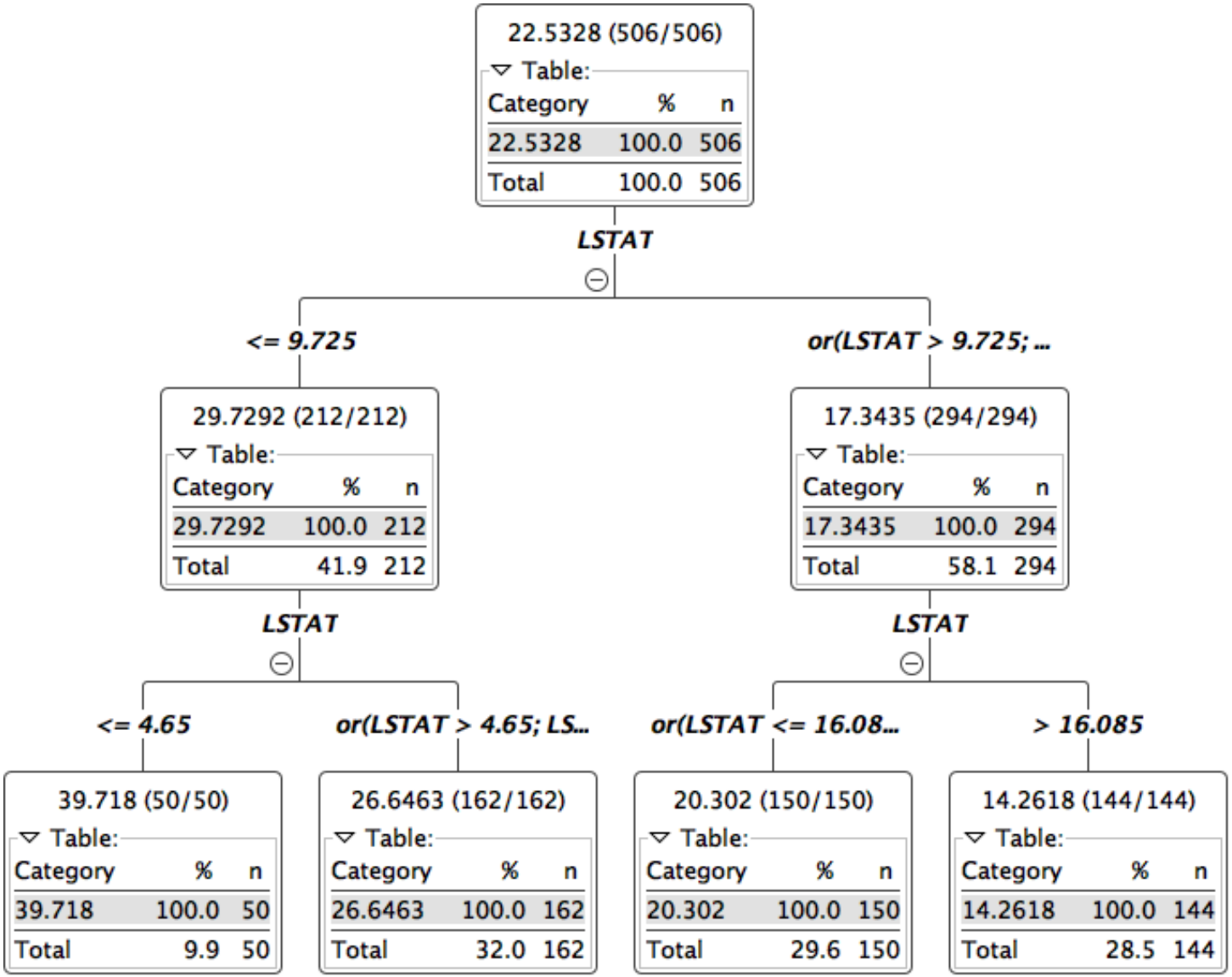
- Information concerning housing in the area of Boston (MA)
  - collected by U.S Census Service:
  - https://archive.ics.uci.edu/ml/machine-learning-databases/housing/
- 506 cases and 14 variables:
  - CRIM - per capita crime rate by town;
  - CHAS - Charles River dummy variable (1 if tract bounds river; 0 otherwise);
  - TAX - full-value property-tax rate per $10,000;
  - PTRATIO - pupil-teacher ratio by town;
  - LSTAT - % lower status of the population;
  - …
  - MEDV - Median value of owner-occupied homes in $1000's
- MEDV is the target variable



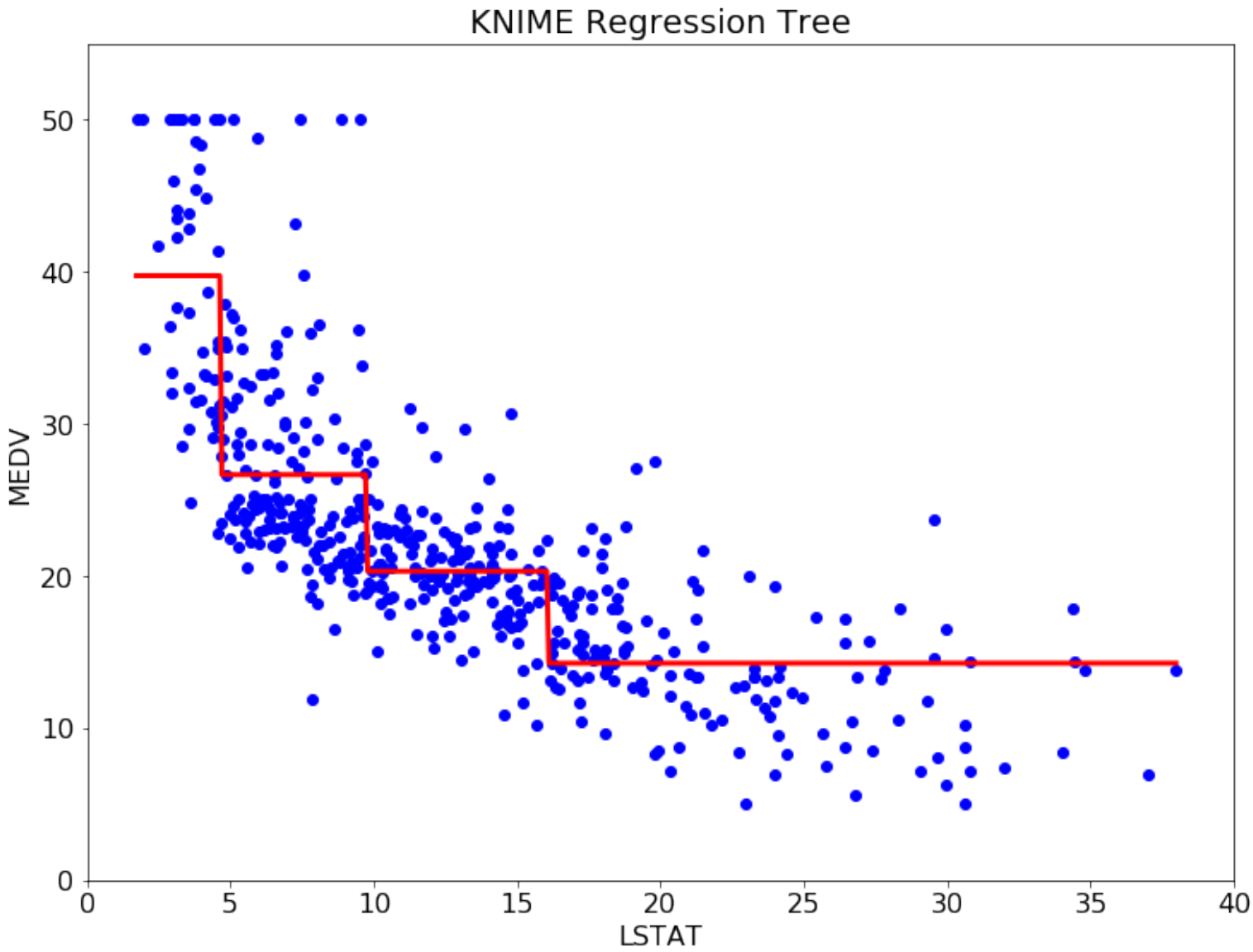
Input variable: LSTAT - % lower status of the population
Output variable: MEDV - Median value of owner-occupied homes in $1000's



K-Nearest Neighbor Regressor with a k of 5

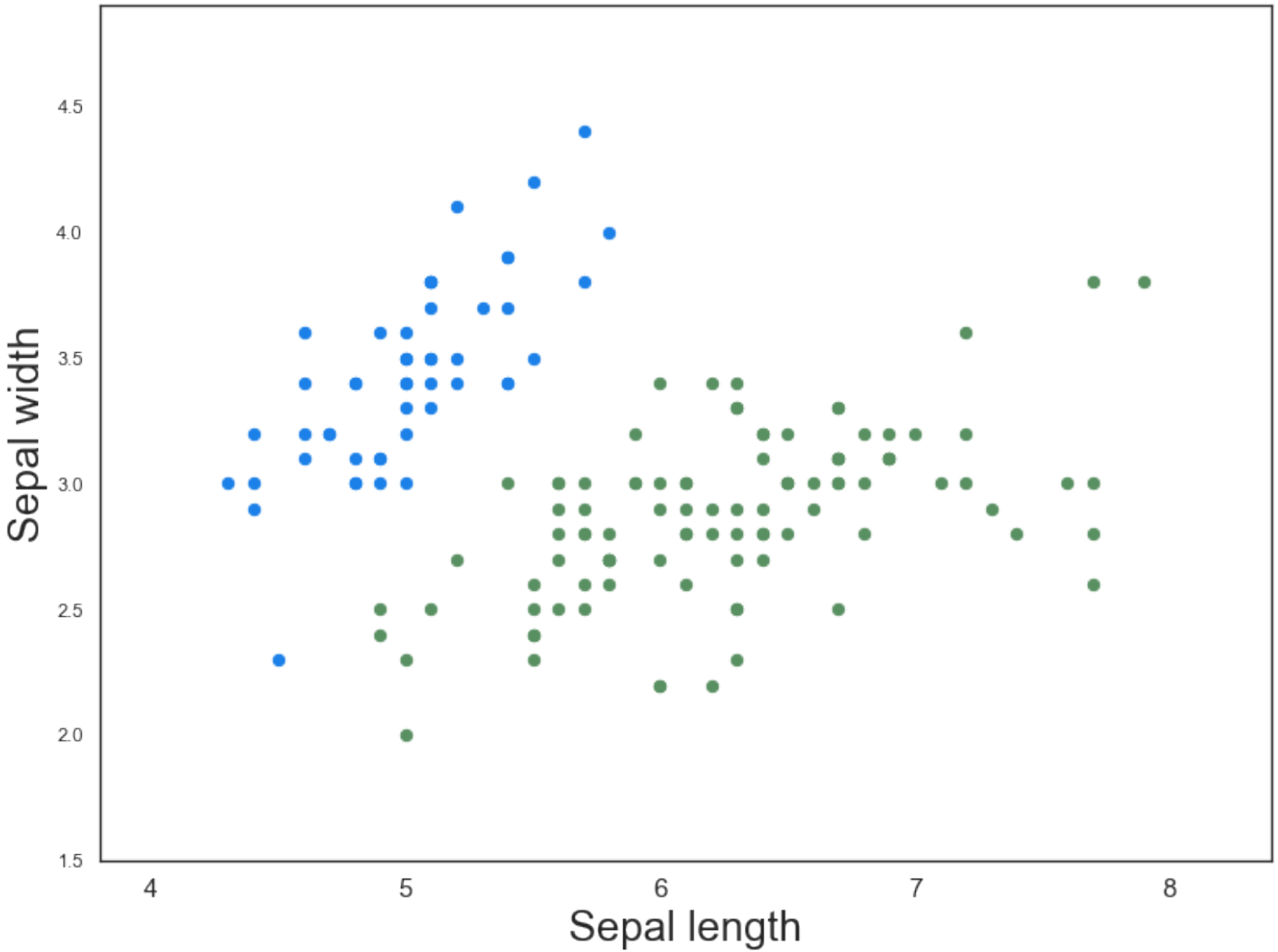


Input variable: LSTAT - % lower status of the population
Output variable: MEDV - Median value of owner-occupied homes in $1000's

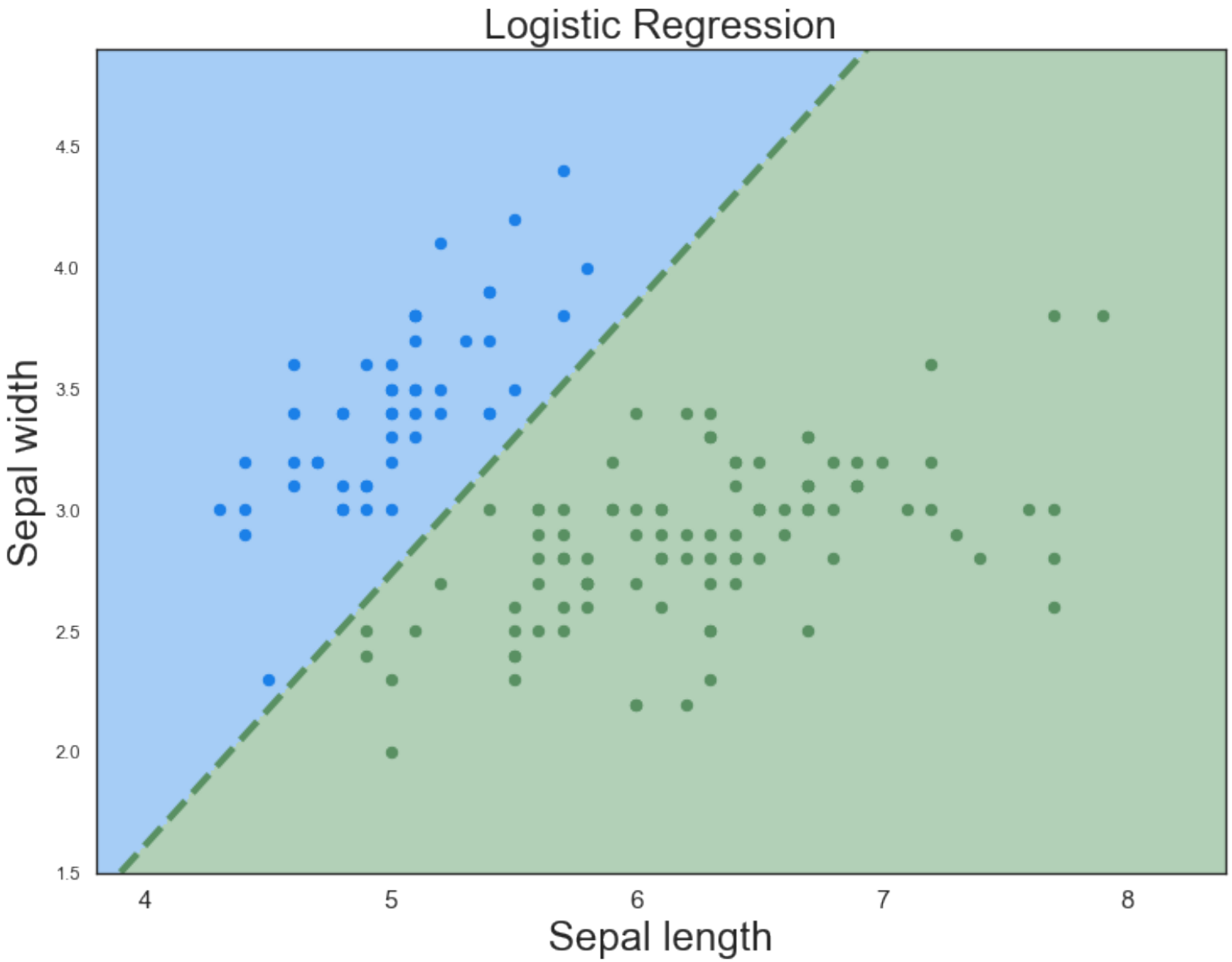


Regression tree computed using KNIME

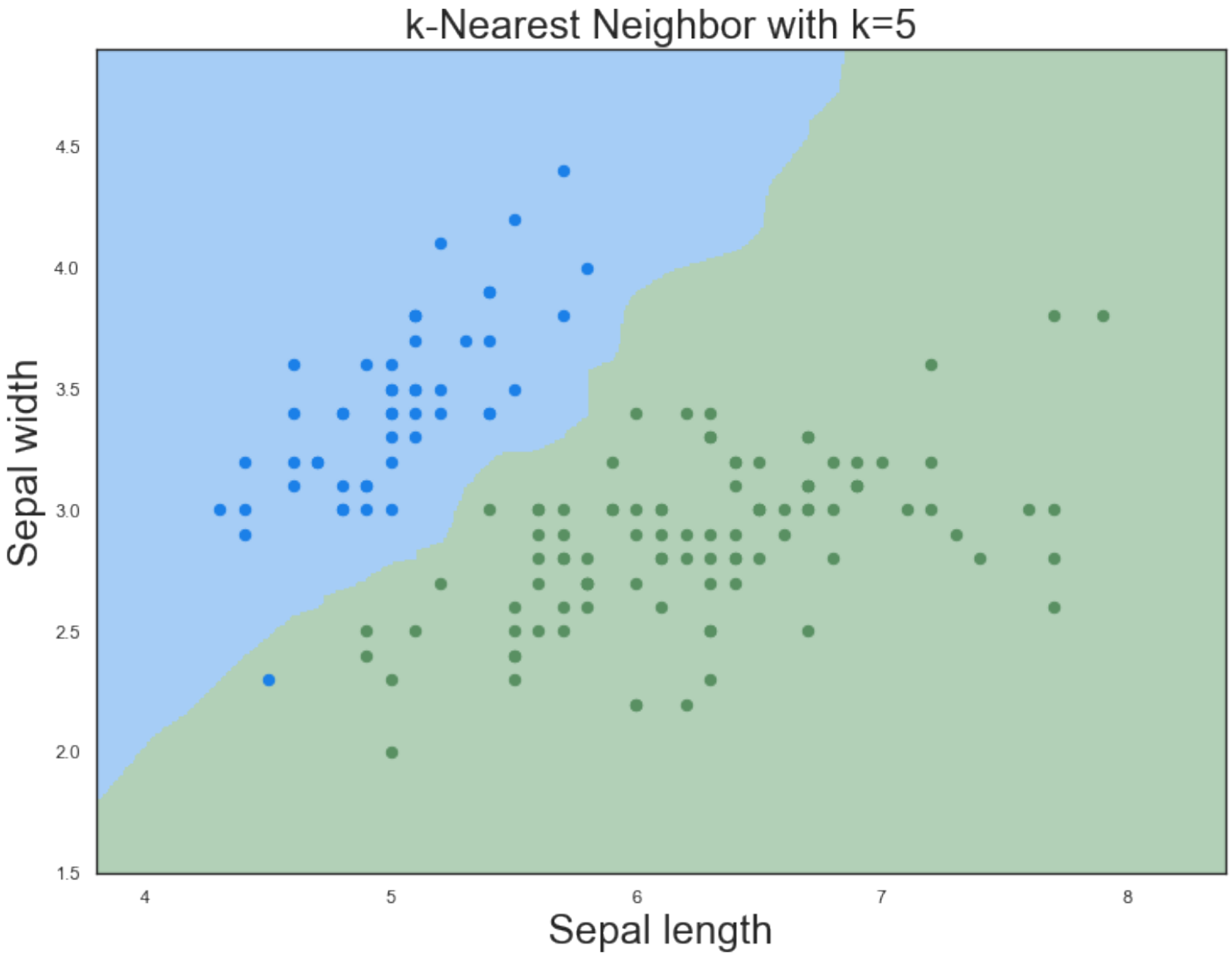
Classification



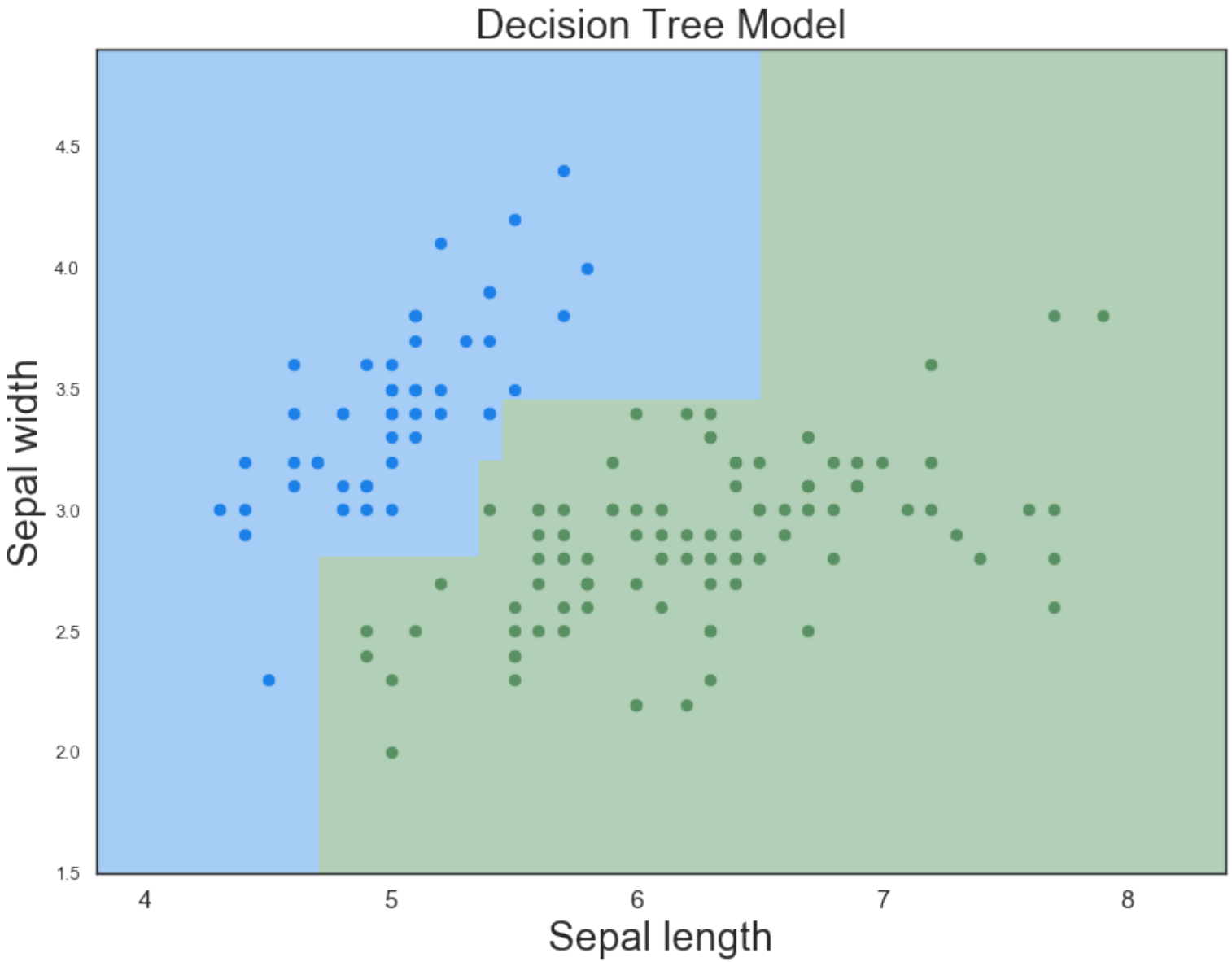
Iris dataset plotted using only two variables and two classes.



Logistic Regression Model

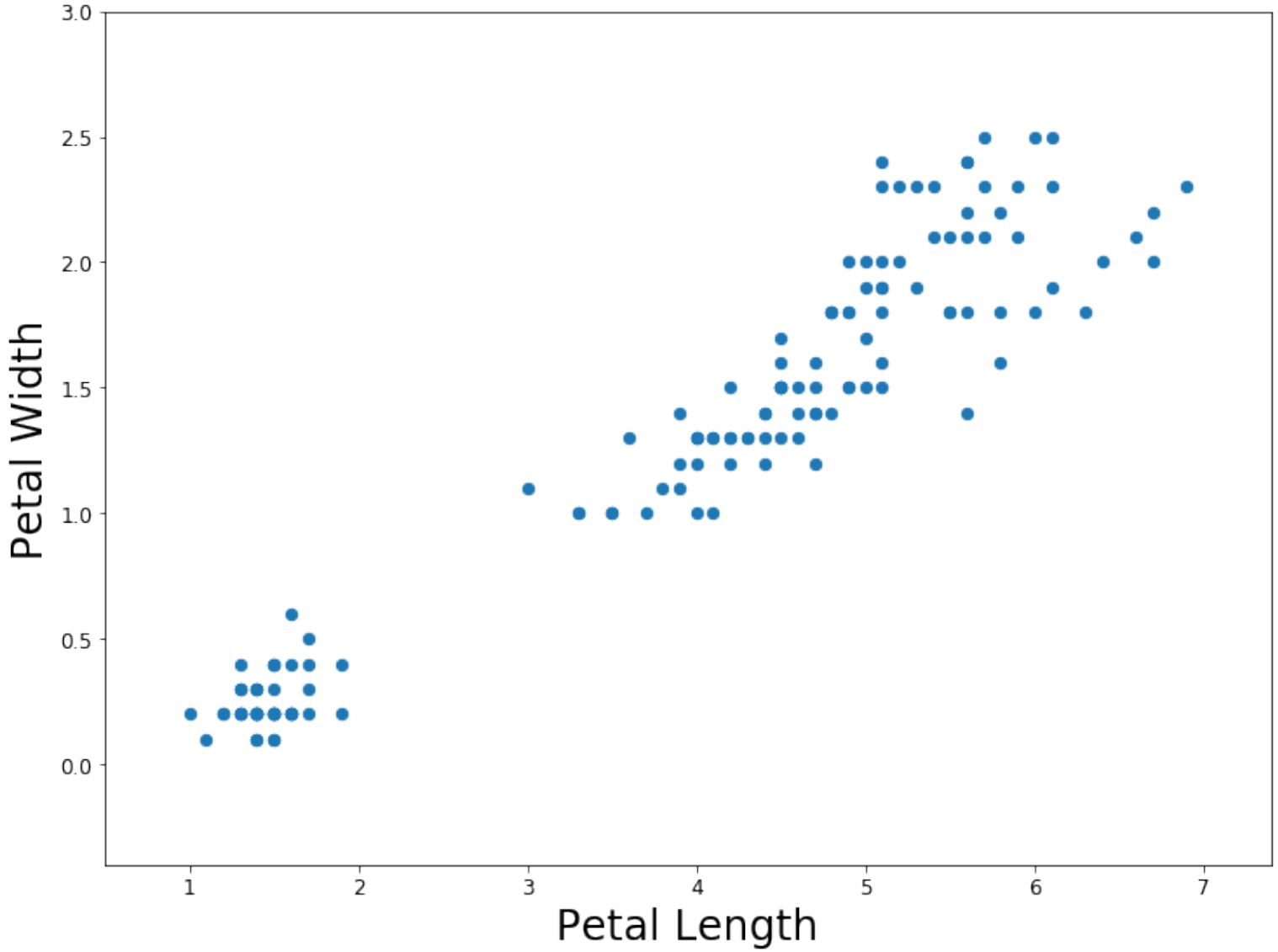


K-Nearest Neighbor model with a k of 5

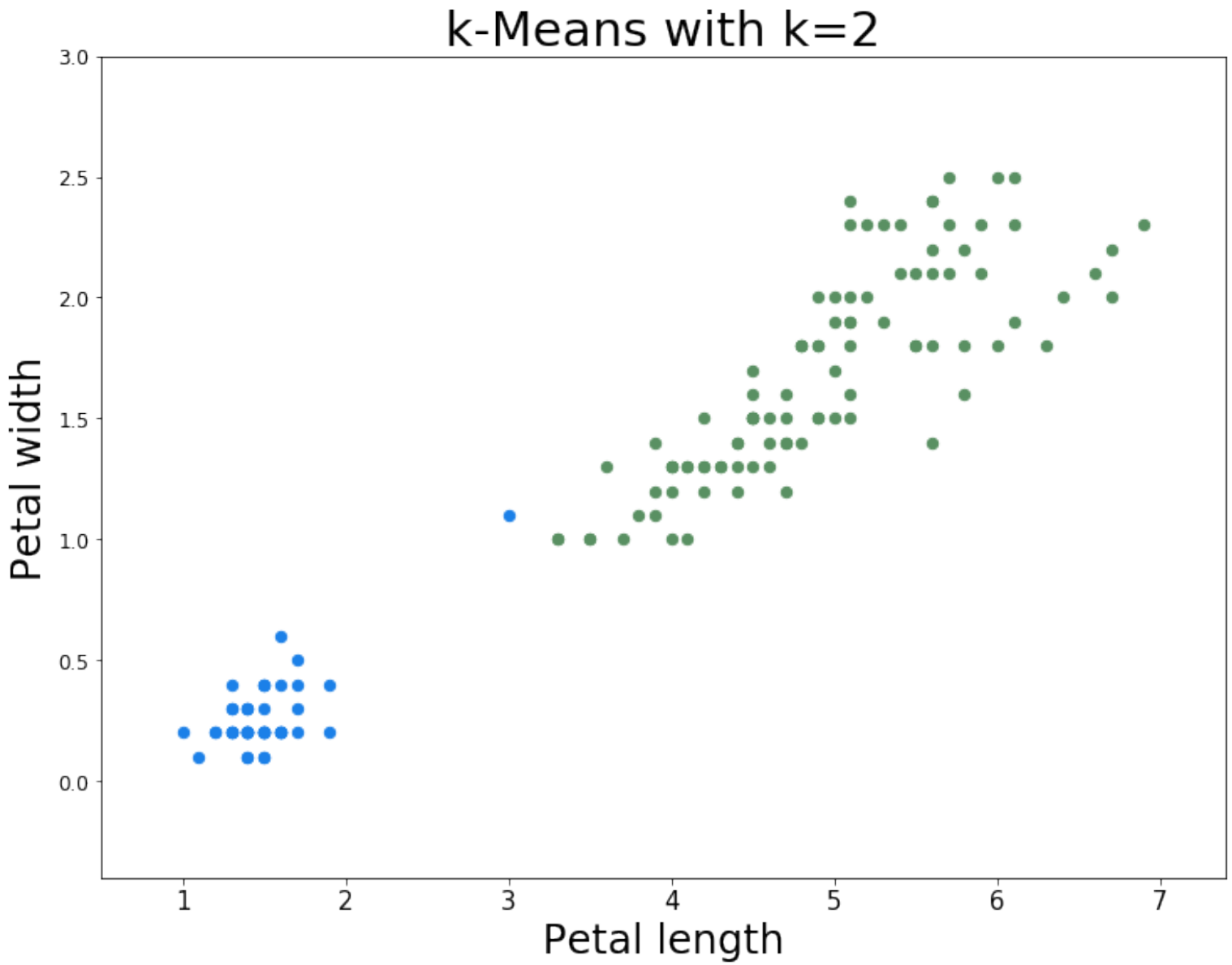


Decision Tree Model

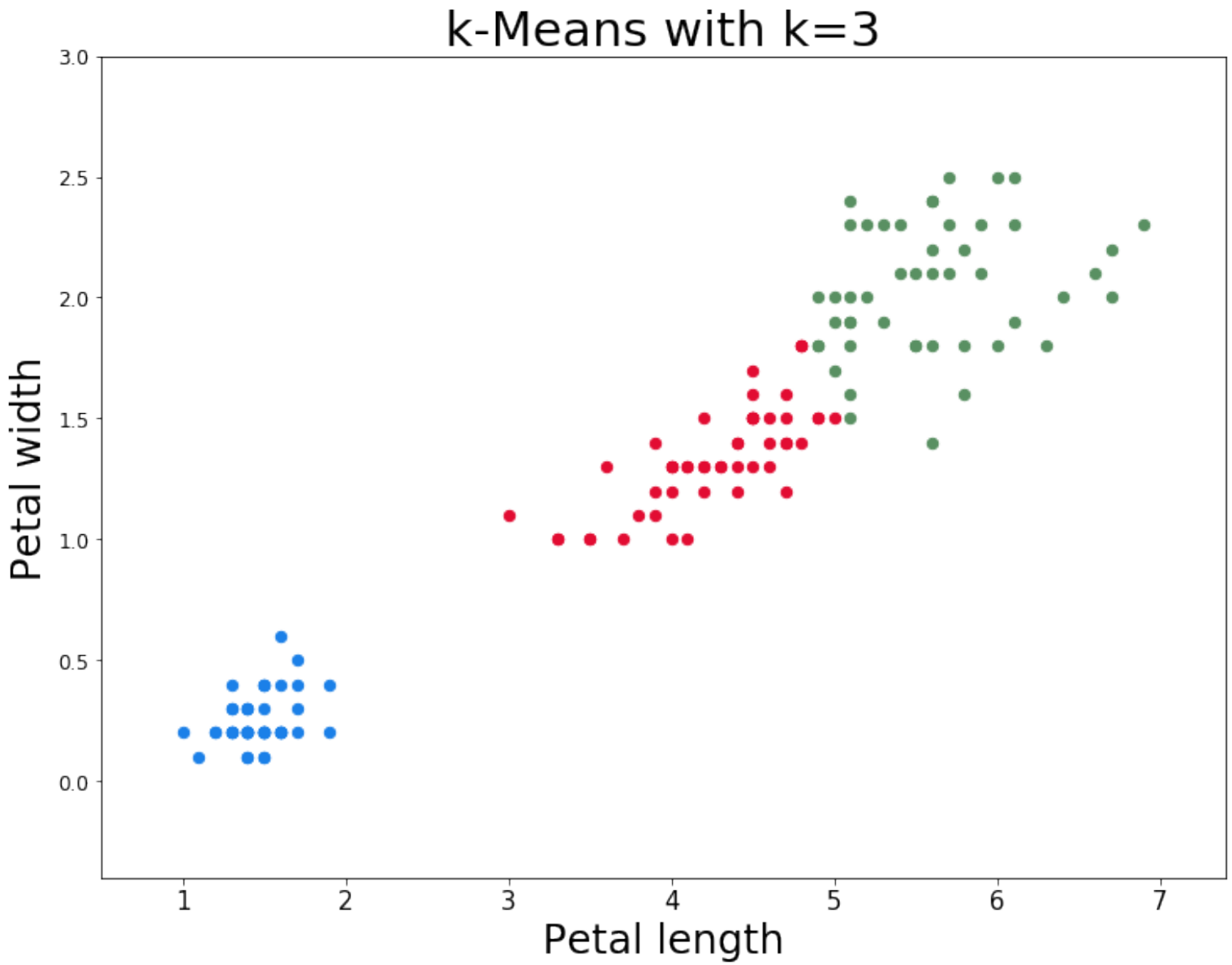
Clustering



Iris dataset plotted using only two variables without taking into account the known class.



k-mean with k=2



k-means with k=3

Associations

Data Mining Tasks: Associations

- {citrus fruit, semi-finished bread, margarine, ready soups}
- {tropical fruit, yogurt, coffee}
- {whole milk}
- {pip fruit, yogurt, cream cheese, meat spreads}
- {other vegetables, whole milk, condensed milk, long life bakery product}
- {whole milk, butter, yogurt, rice, abrasive cleaner}
- {other vegetables, UHT-milk, rolls/buns, bottled beer, liquor (appetizer)}
- {whole milk, cereals}
- ….

Are there interesting products associations?

Frequent Itemsets (Minimum Support 0.1%)


```
     items                          support     count
[1]  {hair spray}                   0.001118454 11
[2]  {flower soil/fertilizer}       0.001931876 19
[3]  {rubbing alcohol}              0.001016777 10
[4]  {frozen fruits}                0.001220132 12
[5]  {prosecco}                     0.002033554 20
[6]  {honey}                        0.001525165 15
[7]  {cream}                        0.001321810 13
[8]  {decalcifier}                  0.001525165 15
[9]  {organic products}             0.001626843 16
[10] {soap}                         0.002643620 26
...
```

Association Rules (Support, Confidence, Lift)


```
{rice, sugar} => {whole milk}
Support 0.001220132
Confidence 1
Lift 3.913649

{canned fish, hygiene articles} => {whole milk}
Support 0.001118454
Confidence 1
Lift 3.913649

{root vegetables, butter, rice} => {whole milk}
Support 0.001016777
Confidence 1
Lift 3.913649
...
```

- 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

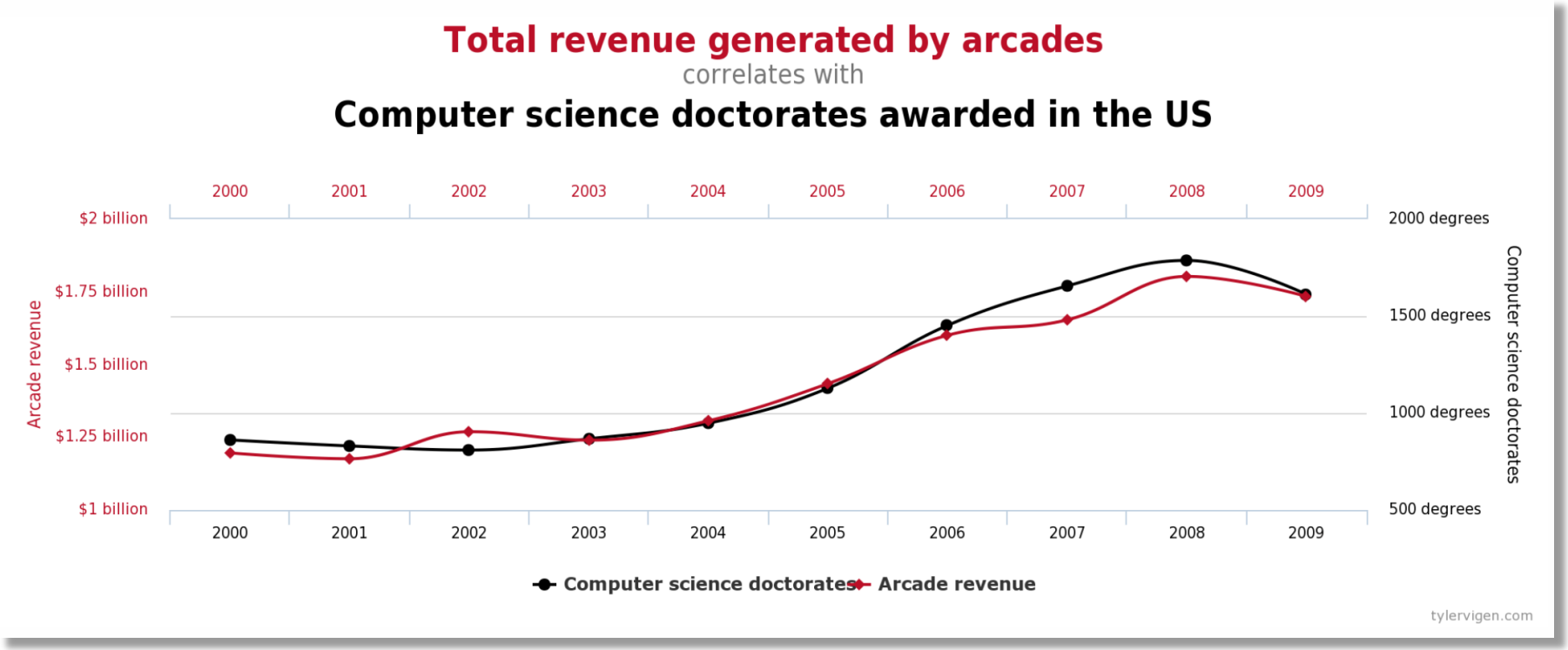
Are all the "Discovered" Patterns Interesting?

- Data Mining 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.

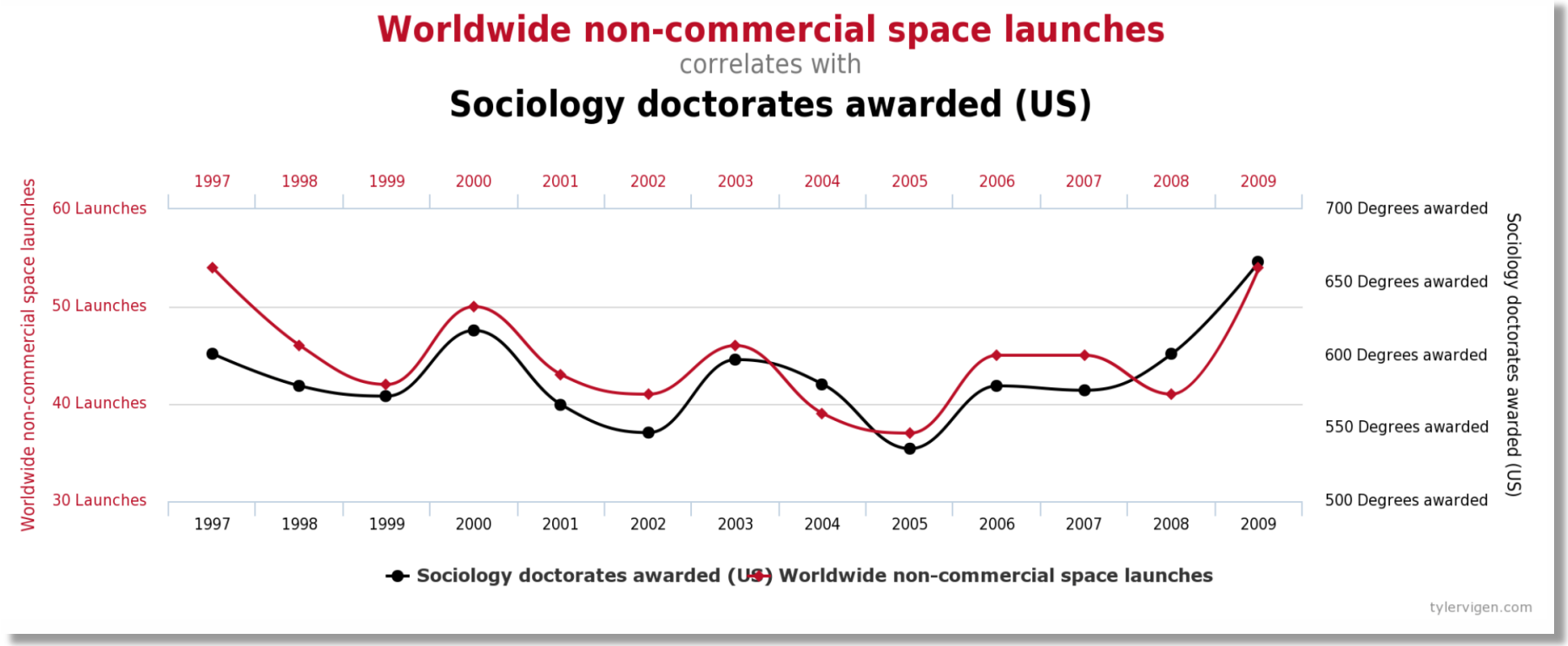
Can We Find All and Only Interesting Patterns?

- Completeness
  - Find all the interesting patterns
  - Can a data mining system find all the interesting patterns?
  - Association vs. classification vs. clustering
- Optimization
  - Search for only interesting patterns:
  - Can a data mining system find only the interesting patterns?
  - Two approaches: (1) first general all the patterns and then filter out the uninteresting ones; (2) generate only the interesting patterns—mining query optimization

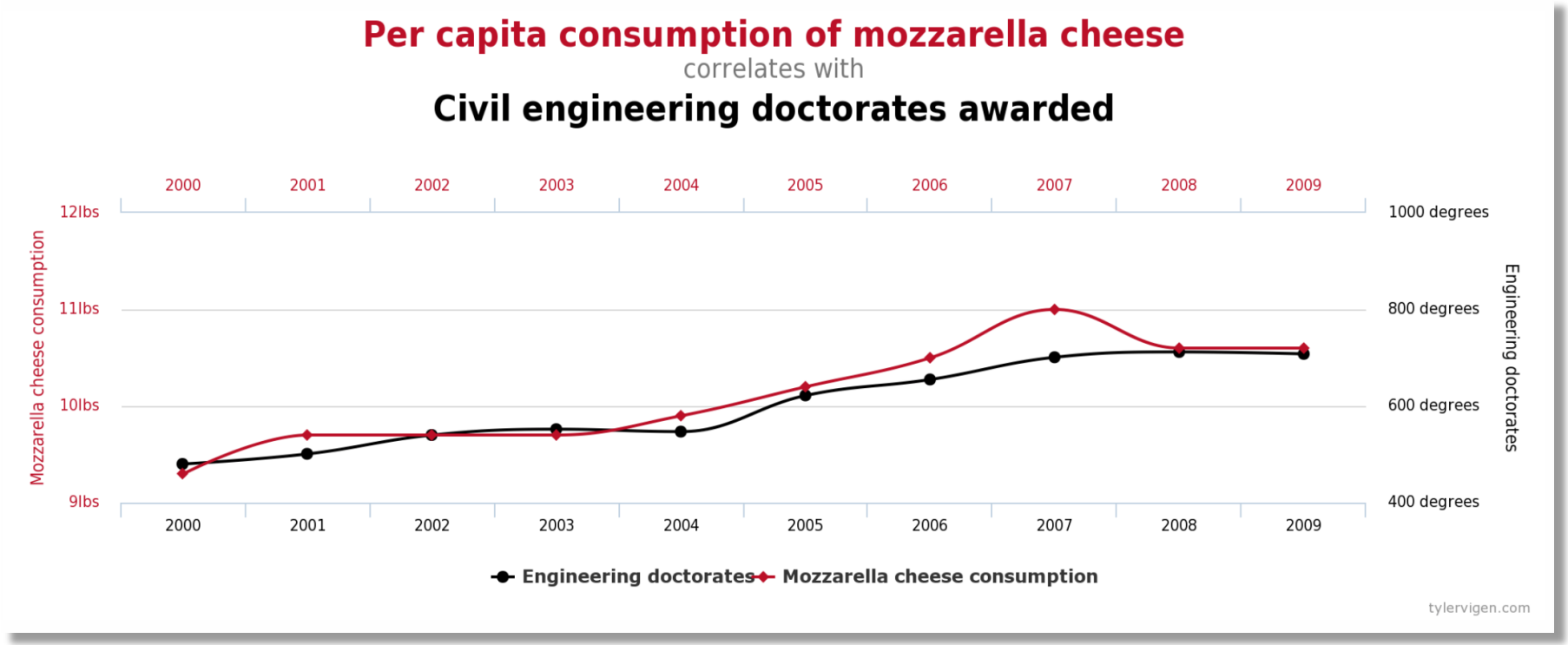
Pitfalls



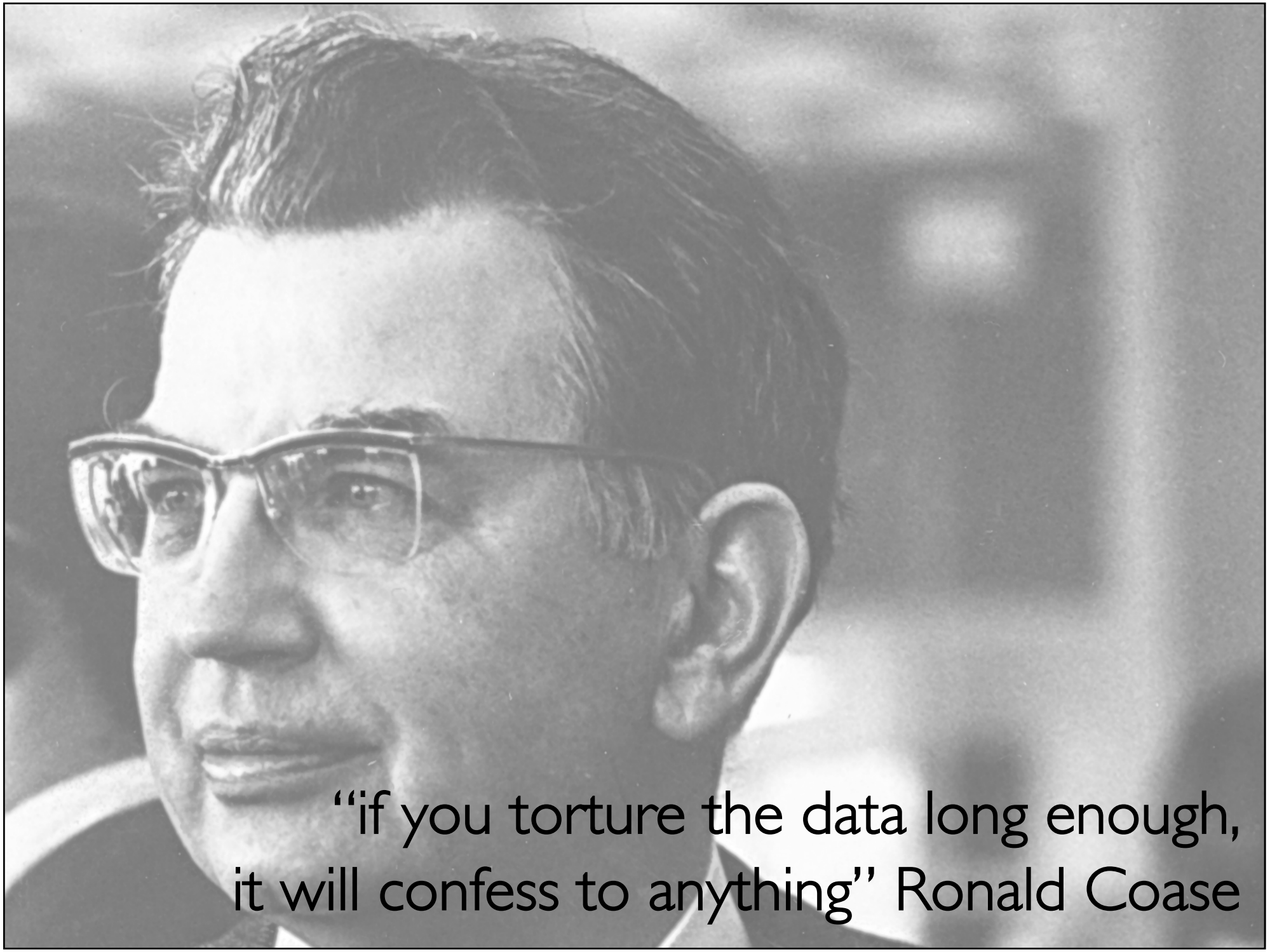
http://www.tylervigen.com



http://www.tylervigen.com



http://www.tylervigen.com


"if you torture the data long enough,
it will confess to anything" Ronald Coase