Introduction to machine learning

Introduction

R. Brooks

Who?

Responsible:	Richard Brooks rib@via.dk
Instructors:	Knud Erik Rasmussen kera@via.dk

The Course

- Machine Learning
  - 5 ECTS
  - Liberal wrt waiving pre-requisites
    - But it is up to you to determine if you have the appropriate background
  - Do I have the appropriate background?
    - Linear algebra: vector/matrix manipulations, properties
    - Probability: common distributions; Bayes' Theorem
    - Statistics: mean/median/mode; maximum likelihood; regression; testing; p-value
    - Work ethics: not like other courses
  - Textbook
    - Introduction to Machine Learning with Python, as well as datasets
      - Book is readily available online
      - Own datasets are welcome

Requirements

- You will probably need to read!
- Assignment:
  - One group assignment within one of the three main topics
    - Classification
    - Clustering/dimensionality reduction
    - Neural networks
  - Groups of 4 – make groups soon or we will make them for you – after September 29, we decide the groups
  - Assignment will form basis for part of the exam
- Exam:
  - The course is evaluated based on two oral examinations.
    - The first examination is a group exam in which the students make a 10 minute presentation about their group assignment. This is followed by approx. 20 minutes of discussion between the students and the examiners. This discussion will evolve around two of the main topics of the course. The group examination takes a total of 30 minutes.
    - After the group exam, each student is then called for an individual 15 minute oral exam. This exam is a discussion about the two main topics that were not covered in the group examination. The student in not allowed to make a presentation at the individual oral exam. The 15 minutes include grading and feedback.
    - The student is given one grade based on both the group exam and the individual exam.

The Overview

36	Introduction + Nearest neighbor + Naïve Bayes + Methodology	RIB
37	Feature Engineering + Data exploration/preparation + Evaluation	RIB*
38	Regularization: Lasso and ridge regression	RIB*
39	PCA + SVD + Preprocessing	RIB*
40	Clustering	RIB*
41	Random Forest & Decision Trees	KERA*
43	Support Vector Machine + Logistic Regression	KERA*
44	SVM + Neural Networks	KERA*
45	Neural Networks	KERA*
46	Neural Networks	KERA/RIB
47	Group Assignment	
48	Group Assignment	

*Online

What is Machine Learning?

- How can we solve a specific problem?
  - As computer scientists you write a program that encodes a set of rules that are useful to solve the problem
  - In many cases it is very difficult to specify those rules, e.g., given a picture determine whether there is a cat in the image

- Lear a problem, inste
  - E
  - F
  - L raining examples ir

Why use learning?

- It is very hard to write programs that solve problems like recognizing a handwritten digit
  - What distinguishes a 2 from a 7?
  - How does our brain do it?
- Instead of writing a program by hand, we collect examples that specify the correct output for a given input
- A machine learning algorithm then takes these examples and produces a program that does the job
  - The program produced by the learning algorithm may look very different from a typical hand-written program.
  - If we do it right, the program works for new cases as well as the ones we trained it on

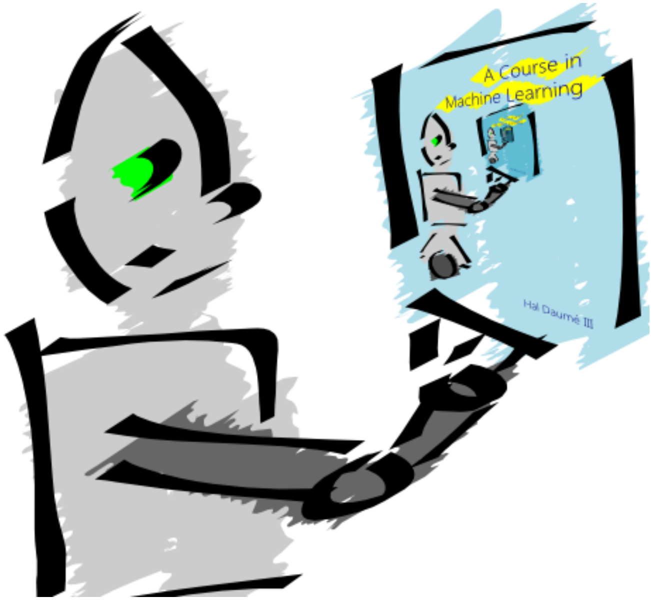
When to use learning

- There is no need to “learn” to calculate payroll or other simple calculations
- Learning is used when:
  - No human experts
    - industrial/manufacturing control
    - mass spectrometer analysis, drug design, astronomic discovery, navigating on Mars
  - Black-box human expertise
    - face/handwriting/speech recognition
    - driving a car, flying a plane
  - Rapidly changing phenomena
    - credit scoring, financial modeling
    - diagnosis, fraud detection
    - routing on a computer
  - Need for customization/personalization
    - personalized news reader
    - movie/book recommendation
    - User biometrics

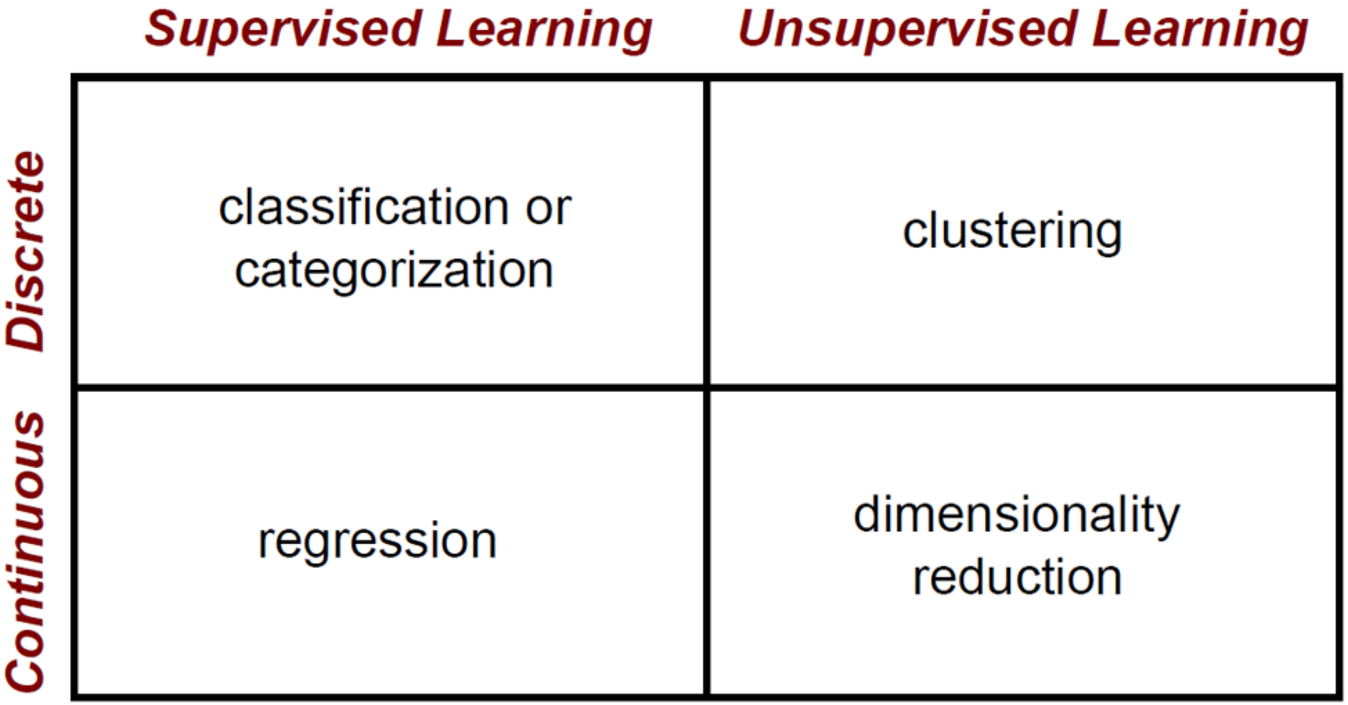
Machine Learning is…

Machine learning is about predicting the future based on the past

-- Hal Daume III

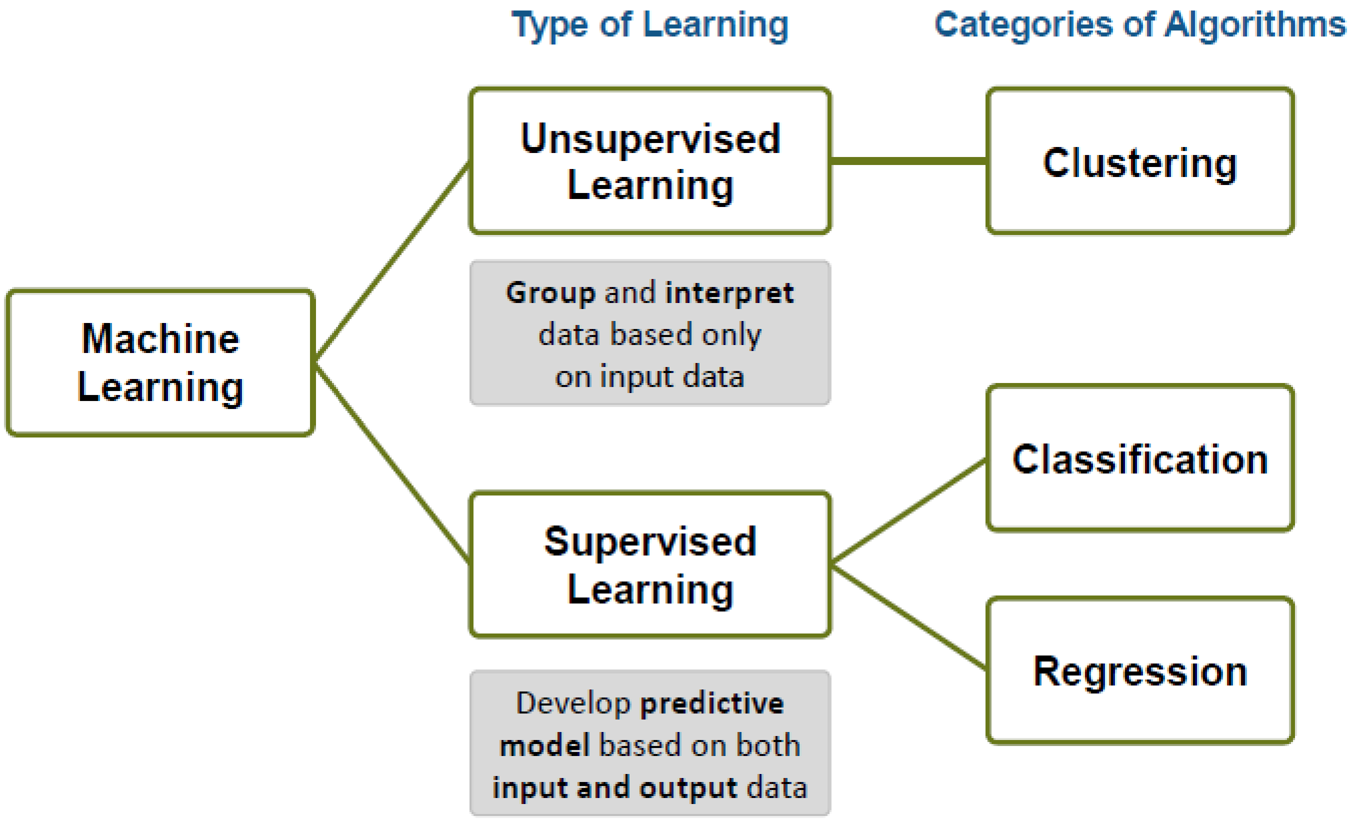


Machine Learning Problems



	Supervised Learning	***Unsupervised Learning***
Discrete	classification or categorization	clustering
Continuous	regression	dimensionality reduction

Machine Learning – An Overview



Type of Learning

Categories of Algorithms

Machine Learning

Unsupervised Learning

Group and **interpret** data based only on input data

Supervised Learning

Develop **predictive model** based on both **input and output** data

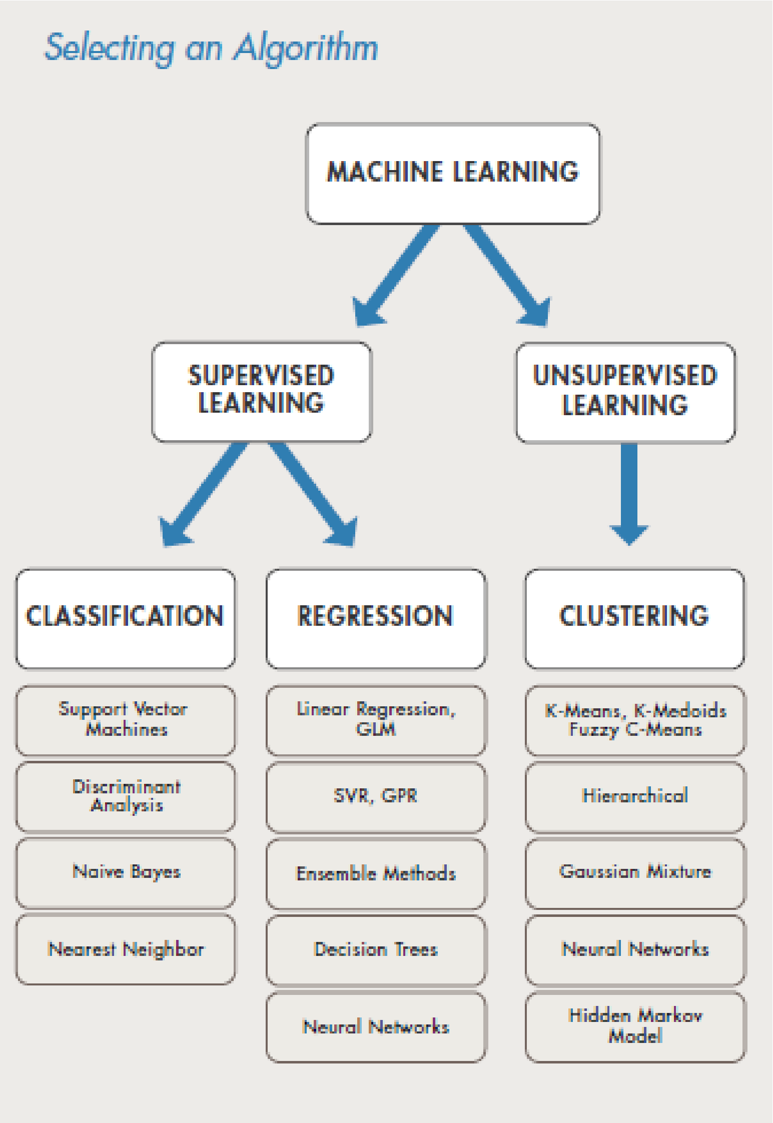
Clustering

Classification

Regression

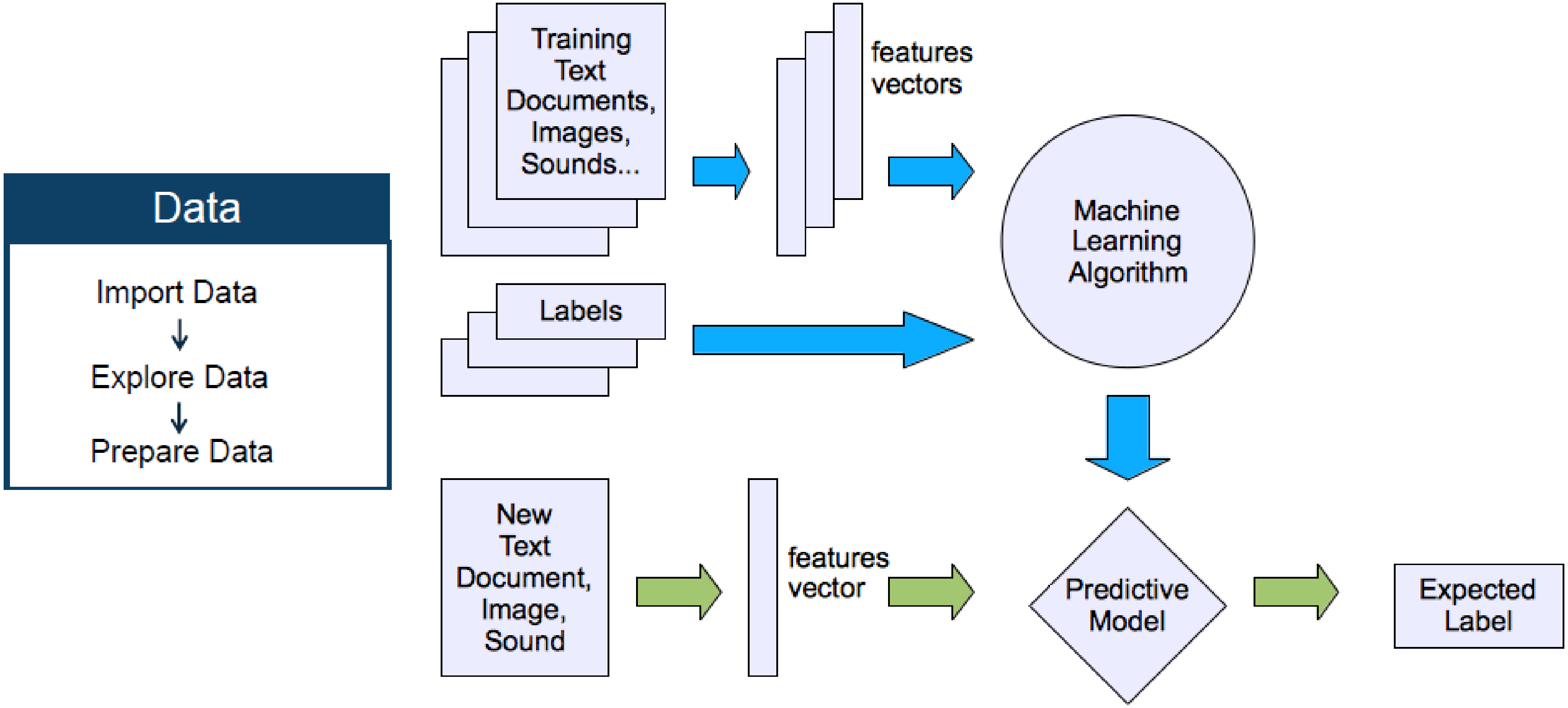
Machine Learning Algorithms

- There is no best method or one size fits all
- Finding the right algorithm is partly just trial and error
- Algorithm selection also depends on the size and type of data you're working with, the insights you want to get from the data, and how those insights will be used



Machine Learning Structure

- Supervised learning



Data

Import Data
↓
Explore Data
↓
Prepare Data

Training Text Documents, Images, Sounds...

features vectors

Labels

Machine Learning Algorithm

New Text Document, Image, Sound

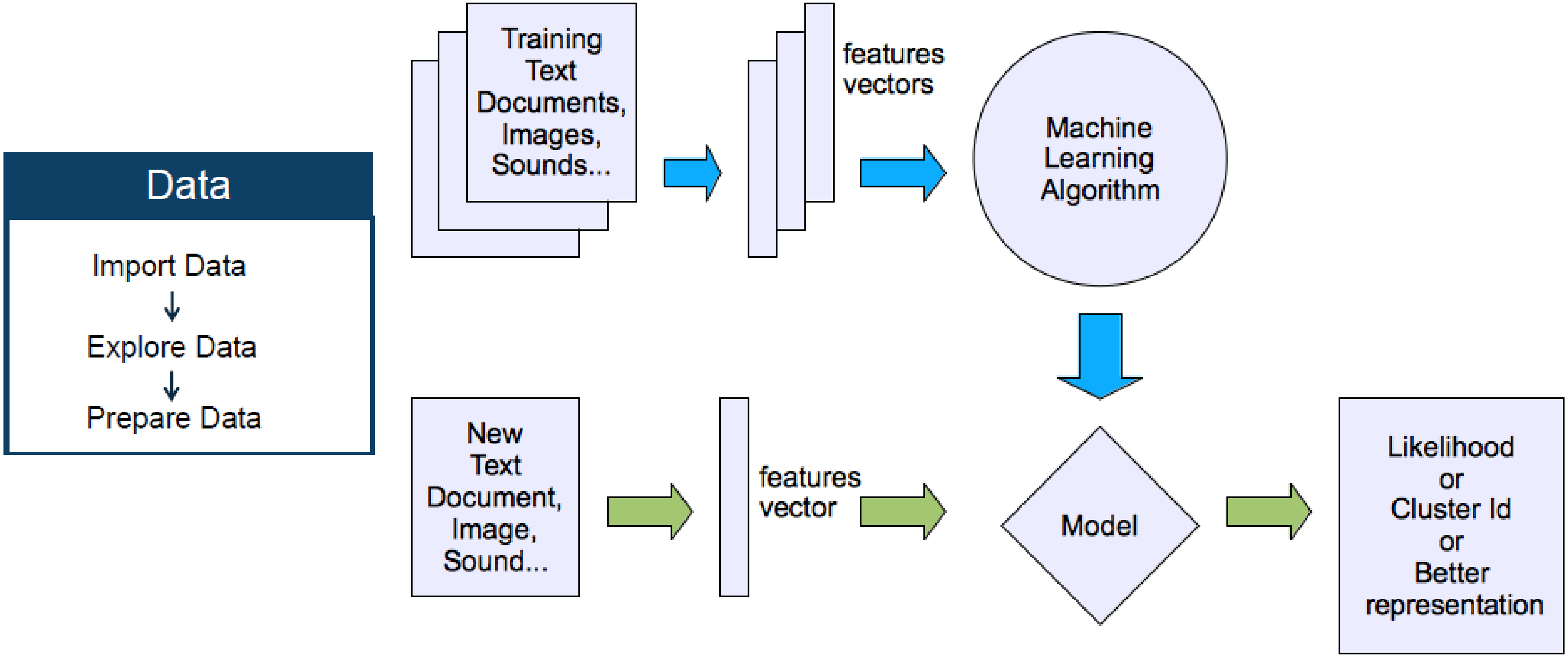
features vector

Predictive Model

Expected Label

Machine Learning Structure

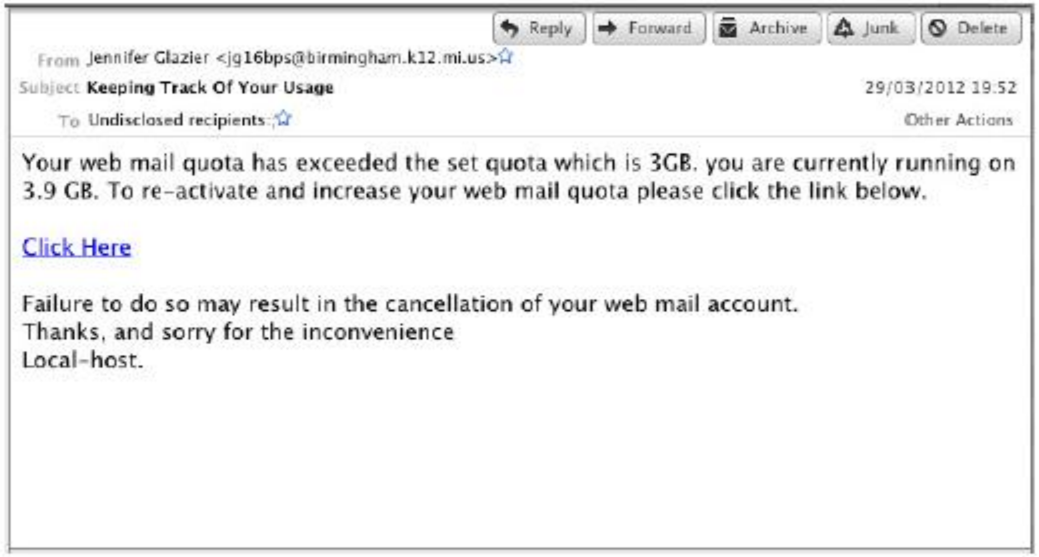
- Unsupervised learning


Data
Import Data
Explore Data
Prepare Data
Training Text Documents, Images, Sounds...
features vectors
Machine Learning Algorithm
New Text Document, Image, Sound...
features vector
Model
Likelihood or Cluster Id or Better representation

Examples: Classification

- **Spam Filtering**

Spam filtering in Mozilla: user trains the mail reader to recognize spam by manually labeling incoming mails as spam/no spam.



From Jennifer Glazier <jg16bps@birmingham.k12.mi.us>

Subject **Keeping Track Of Your Usage** 29/03/2012 19:52

To Undisclosed recipients:;

Your web mail quota has exceeded the set quota which is 3GB. you are currently running on 3.9 GB. To re-activate and increase your web mail quota please click the link below.

Click Here

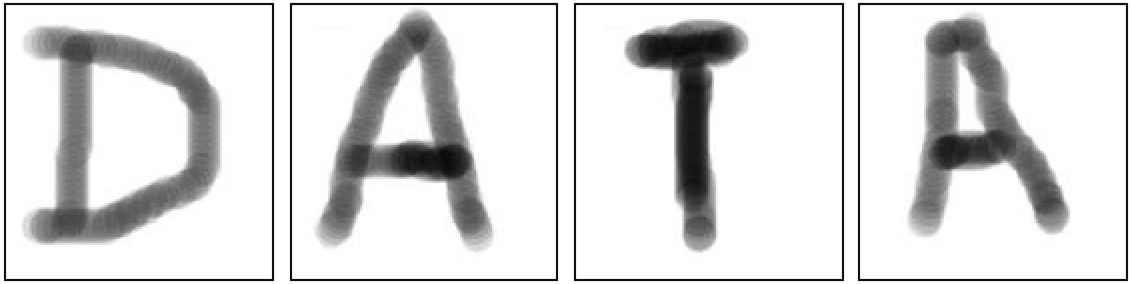
Failure to do so may result in the cancellation of your web mail account.
Thanks, and sorry for the inconvenience
Local-host.

Data: collection of user-classified emails (full text).
Task: build a classifier that automatically categorizes an incoming email as spam/no spam

Examples: Classification

- **Character Recognition**

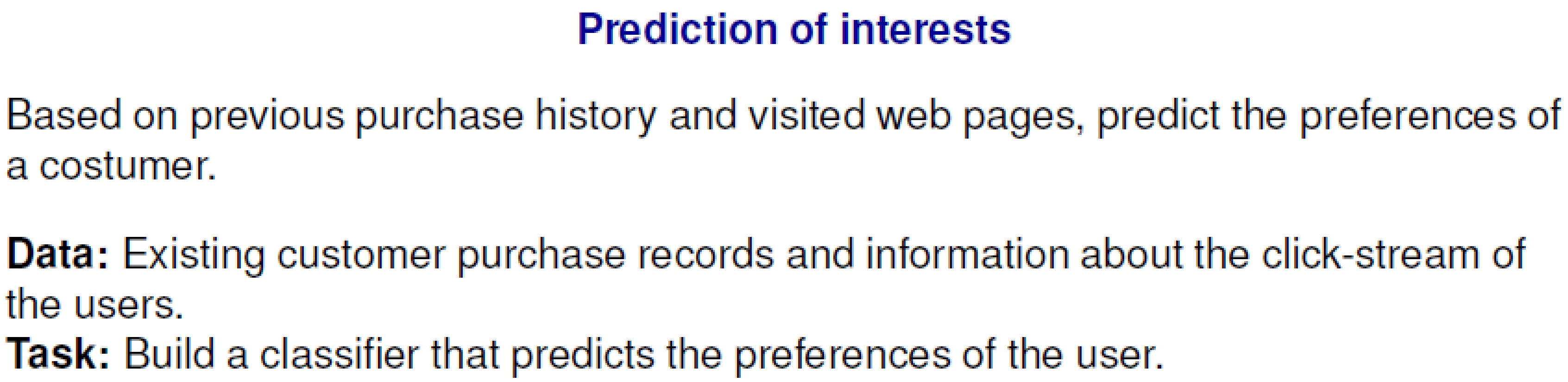
Example for a *Pattern Recognition* problem (pattern recognition is an older discipline than data mining, but now can also be seen as a sub-area of data mining):



Data: collection of handwritten characters, correctly labeled.
Task: build a classifier that identifies new handwritten characters.

Classification/regression

- Prediction of interests



Prediction of interests

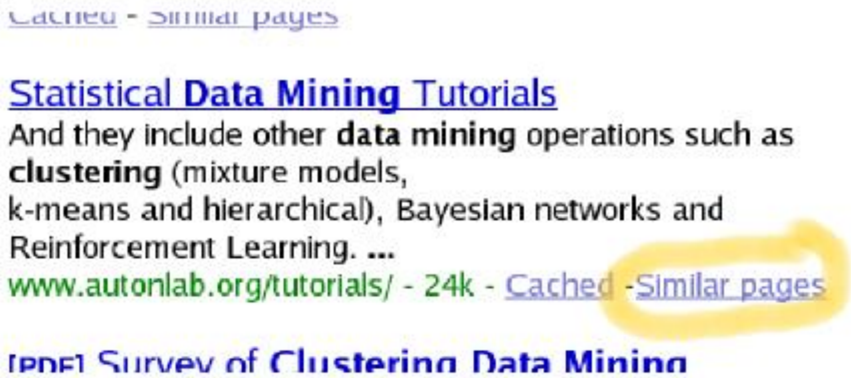
Based on previous purchase history and visited web pages, predict the preferences of a costumer.

Data: Existing customer purchase records and information about the click-stream of the users.
Task: Build a classifier that predicts the preferences of the user.

Examples: Clustering/web-structure mining

- Text Categorisation

Web mining: automatically detect similarity between web pages (e.g. to support search engines or automatic construction of internet directories).



Cached - Similar pages

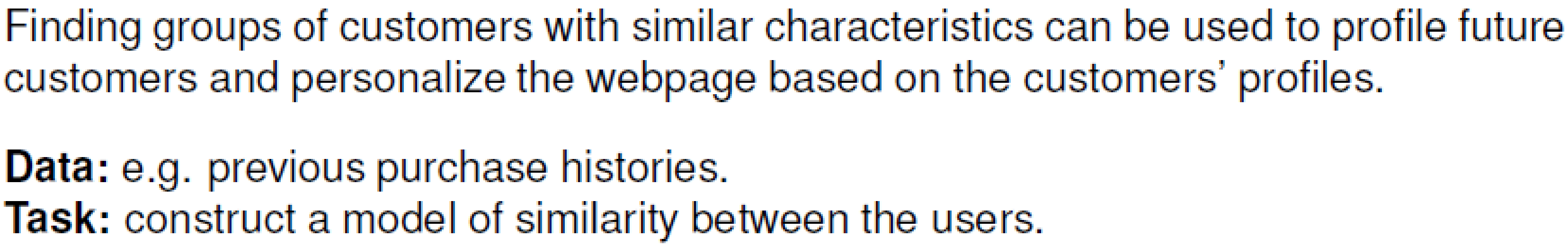
Statistical **Data Mining** Tutorials
And they include other **data mining** operations such as **clustering** (mixture models,
k-means and hierarchical), Bayesian networks and Reinforcement Learning. **...**
www.autonlab.org/tutorials/ - 24k - Cached - Similar pages

[PDF] Survey of **Clustering Data Mining**

Data: the WWW.
Task: Construct a (similarity) model for pages on the WWW.

Examples: Clustering

- Customer Analysis

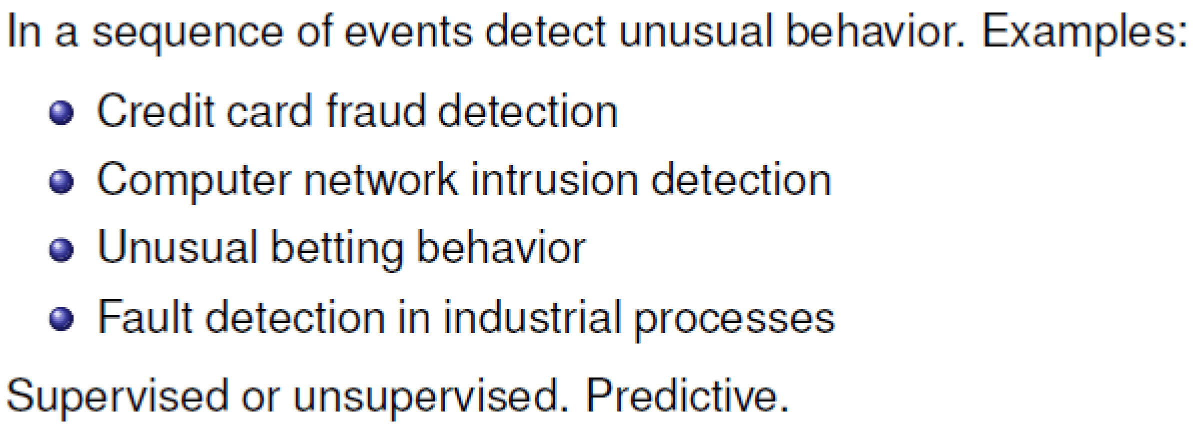


Finding groups of customers with similar characteristics can be used to profile future customers and personalize the webpage based on the customers' profiles.

Data: e.g. previous purchase histories.
Task: construct a model of similarity between the users.

Examples: Clustering/classification

- Intrusion/Outlier Detection



In a sequence of events detect unusual behavior. Examples:

- Credit card fraud detection
- Computer network intrusion detection
- Unusual betting behavior
- Fault detection in industrial processes

Supervised or unsupervised. Predictive.

Machine Learning vs Data Mining

- Data-mining: Typically using very simple machine learning techniques on very large databases because computers are too slow to do anything more interesting with ten billion examples
- Previously used in a negative sense – misguided statistical procedure of looking for all kinds of relationships in the data until finally find one
- Now lines are blurred: many ML problems involve tons of data

Machine Learning vs Statistics

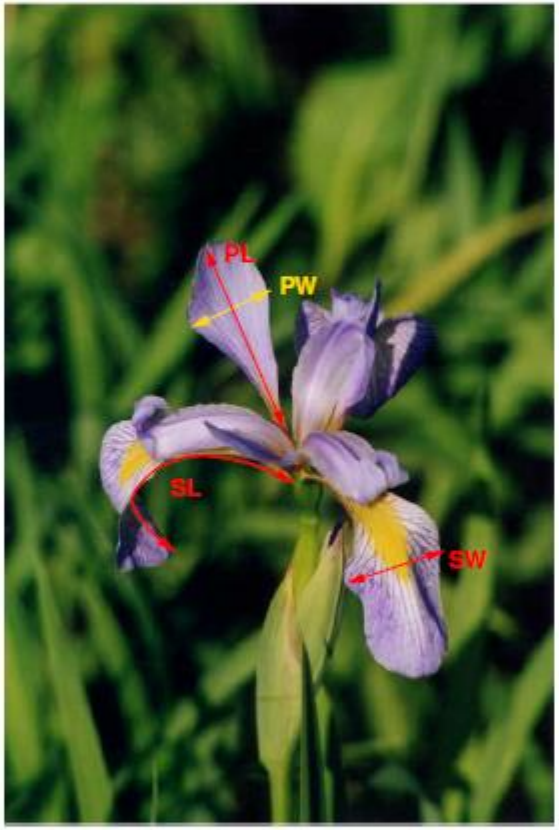
- ML uses statistical theory to build models
- A lot of ML is rediscovery of things statisticians already knew; often disguised by differences in terminology
- But the emphasis is very different
- Can view ML as applying computational techniques to statistical problems.

Initial Case Study

- What grade will I get in this course?
- Data: entry survey and marks from this and previous years
- Process the data
  - Split into training set; and test set
  - Determine representation of input;
  - Determine the representation of the output;
- Choose form of model: linear regression
- Decide how to evaluate the system's performance: objective function
- Set model parameters to optimize performance
- Evaluate on test set: generalization

Example of workflow

Example data: Iris data set



Measurement of petal width/length and sepal width/length for 150 flowers of 3 different species of Iris.

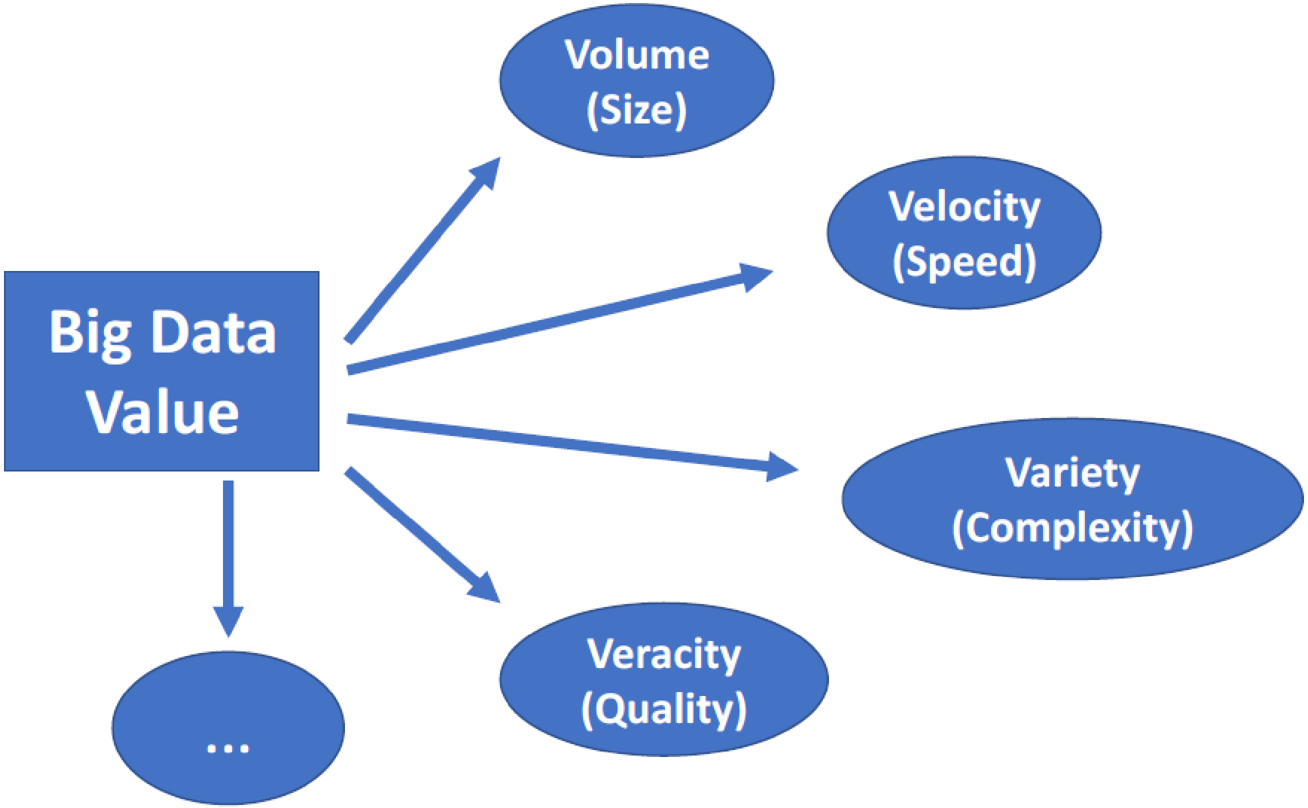
first reported in:
Fisher,R.A. "The use of multiple measurements in taxonomic problems" Annual Eugenics, 7 (1936).

	Attributes			Class variable
SL	SW	PL	PW	Species
5.1	3.5	1.4	0.2	Setosa
4.9	3.0	1.4	0.2	Setosa
6.3	2.9	6.0	2.1	Virginica
6.3	2.5	4.9	1.5	Versicolor
...	...	...	...	...

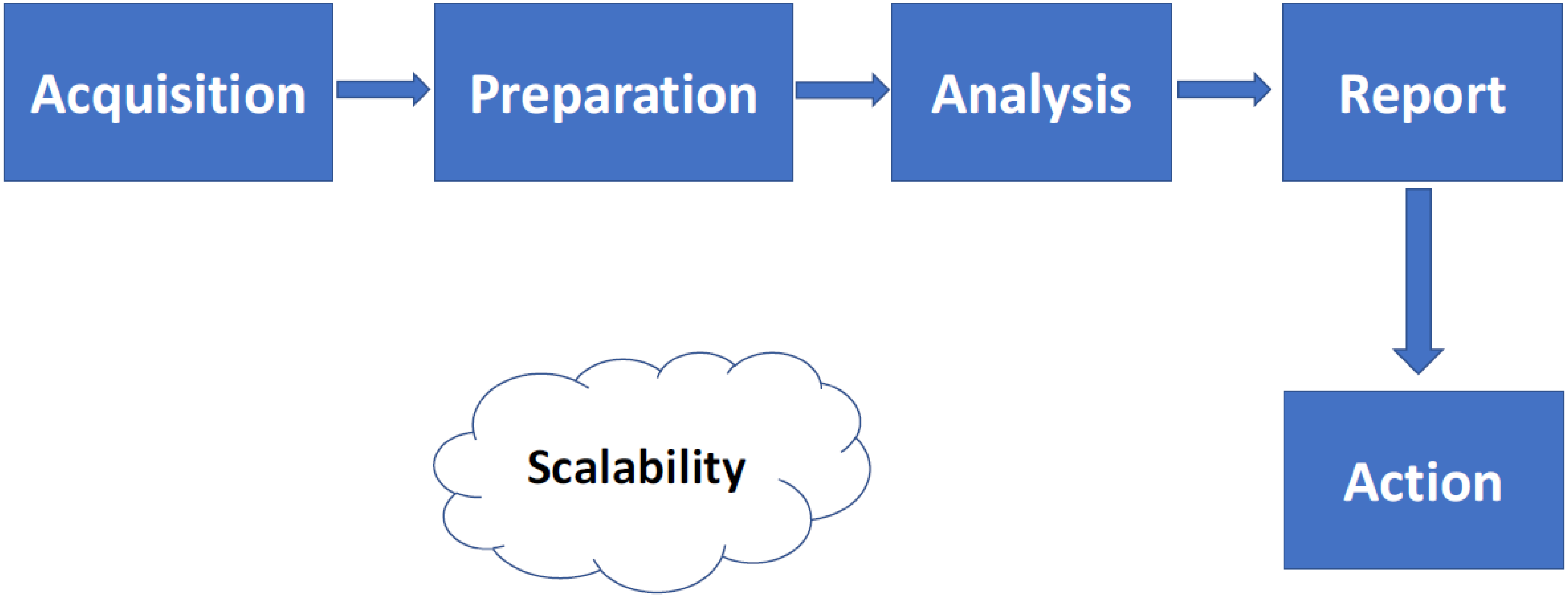
Task: Predict species of new iris flowers based on their measurements (a supervised learning problem)

Data Engineering Flow

Big Data Characteristics

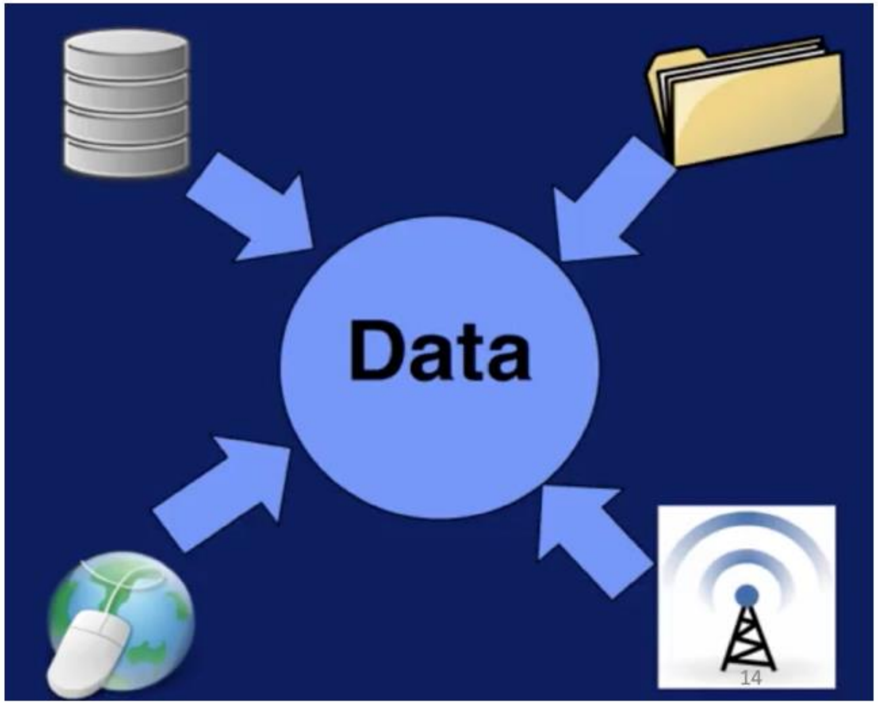


(Big) Data Science: Engineering Flow



Data acquisition

- Identify data sources
- Collect data from different sources
- Synthetize and integrate them



Data Preparation

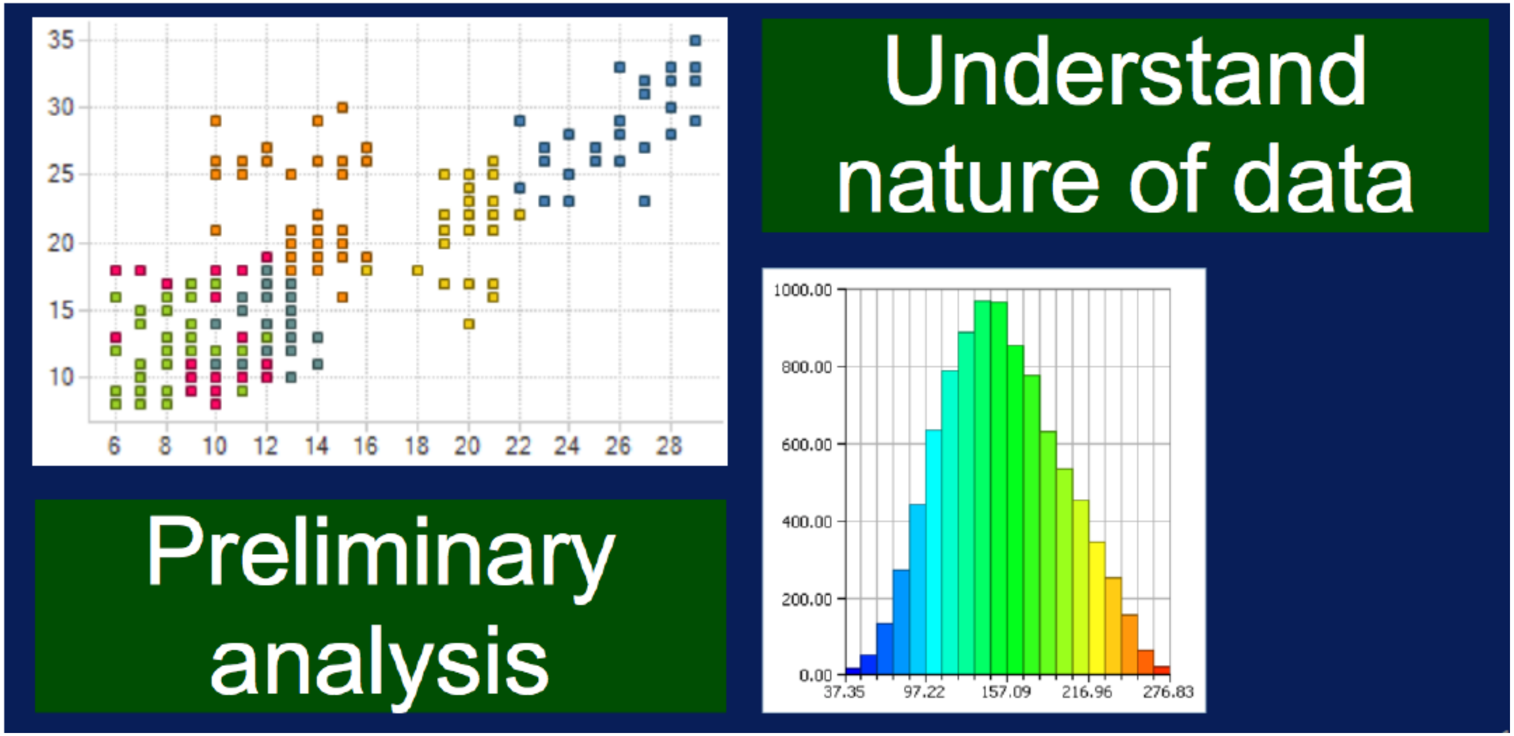
Data Exploration

- Understand the statistics of your data
- Finding correlations, trends, outliers etc.

Data Preprocessing

- Prepare data for analysis
- Clean, extract, transform

Preparation: Data Exploration



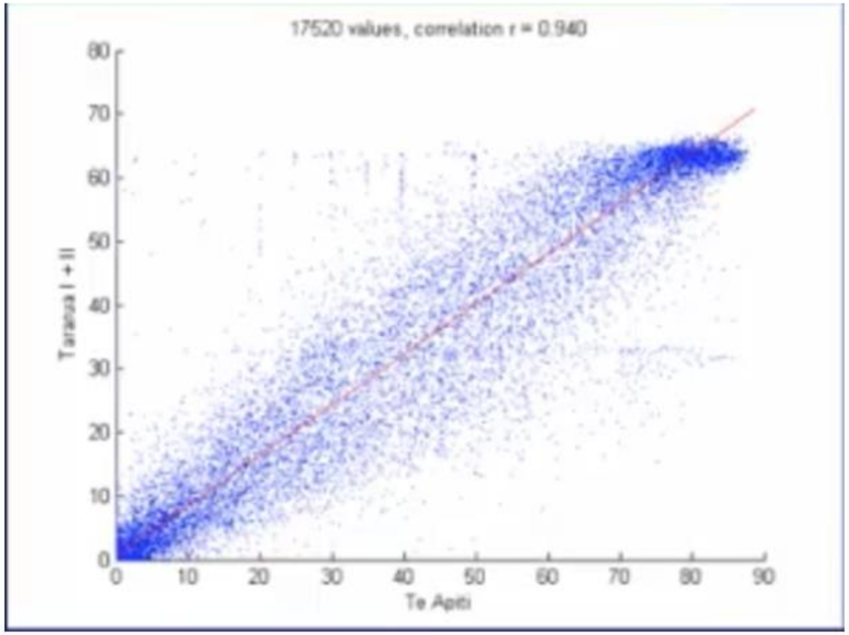
Understand nature of data

Preliminary analysis

Preparation: Data Exploration

Exploring correlations:

- Understand the relationships between variables

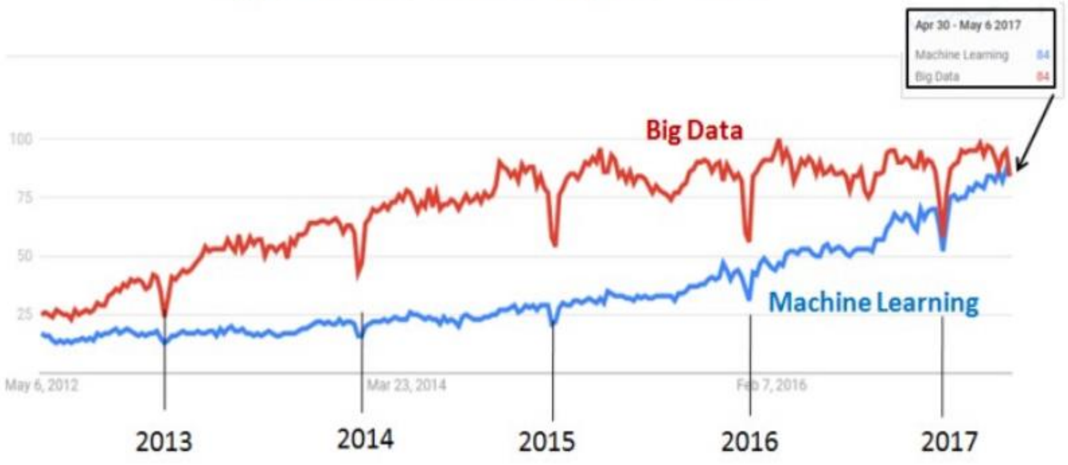


Preparation: Data Exploration

Extracting trends:

- Moving direction of certain events/phenomenon

Google Trends, May 2012 - April 2017
Big Data vs **Machine Learning** search terms

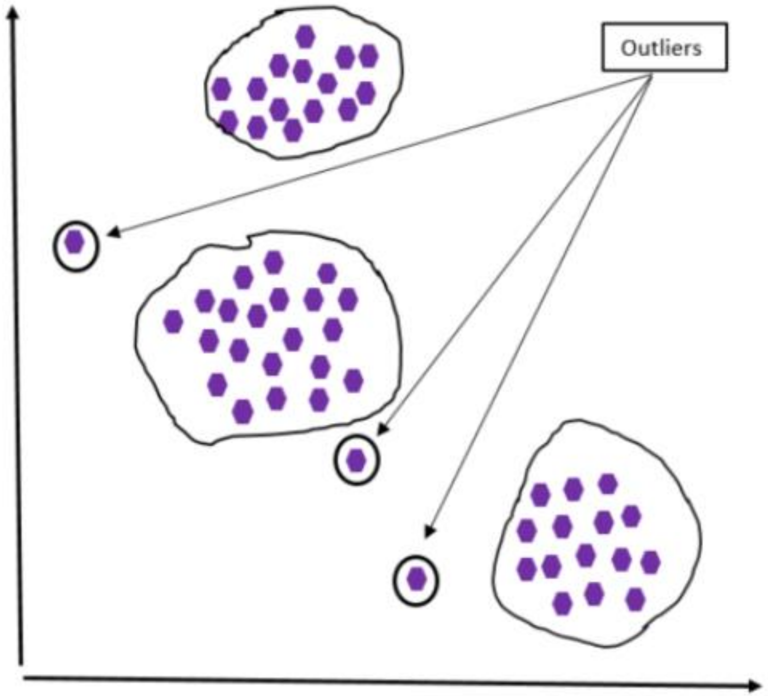


Source: http://www.experiencedata.nl/experience-centermachine-learning-overtaking-big-data/

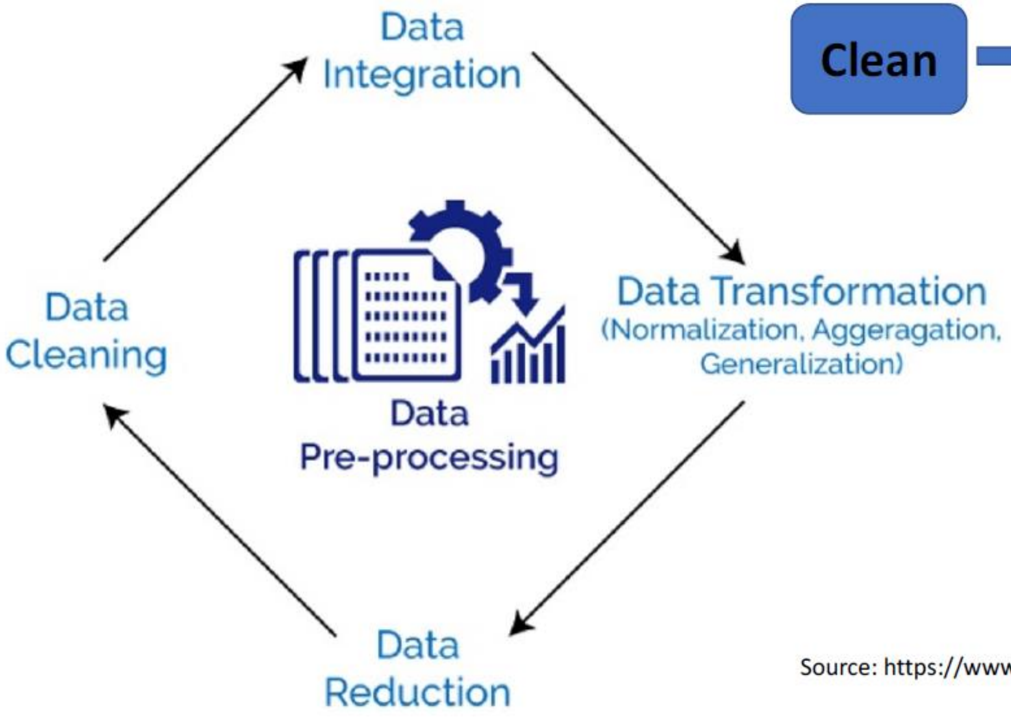
Preparation: Data Exploration

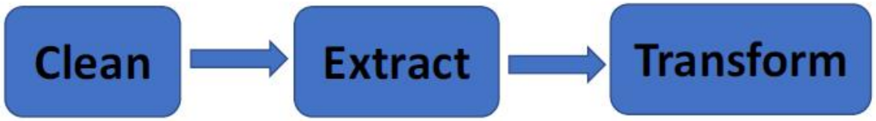
Outlier detection:

- Identify interesting/abnormal data points



Preparation: Data Pre-processing





Source: https://www.electronicsmedia.info/2017/12/20/what-is-data-preprocessing/

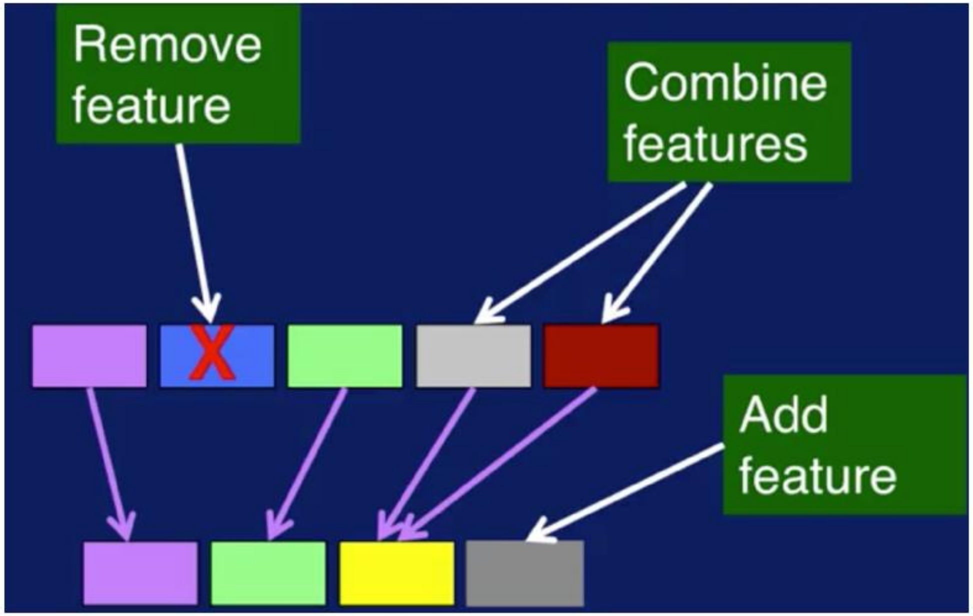
Pre-processing: Data Cleaning

Deal with data quality issues:

- Missing values
- Duplicate values
- Invalid/inconsistent data
- Noise
- Outliers

Pre-processing: Variables Extraction

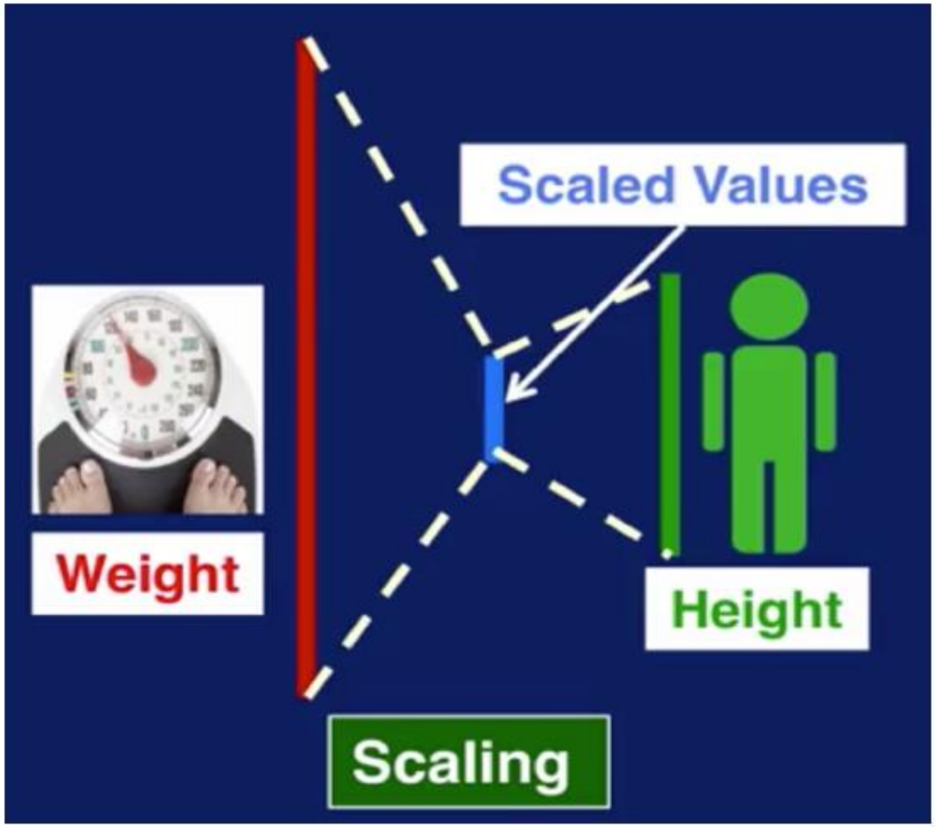
- Remove redundant/irrelevant features
- Combine two or more features
- Add new features



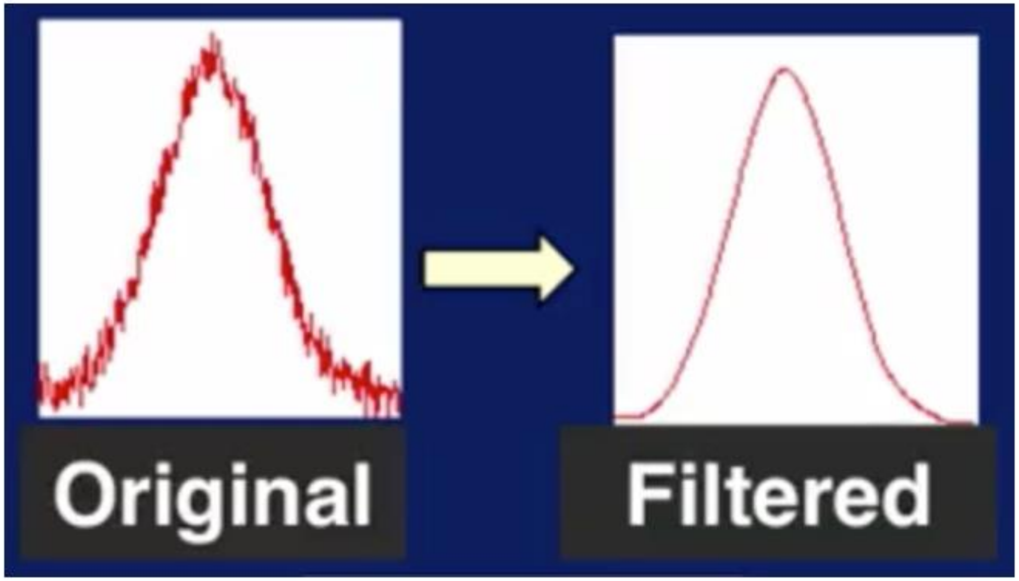
Pre-processing: Data Transformation

Map data into different formats

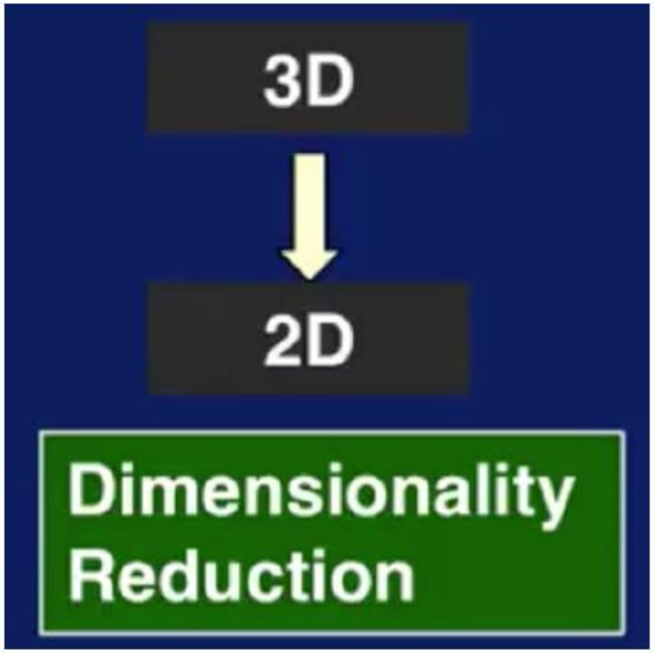
- Scaling
- Aggregation
- Normalization



Pre-processing: Data Transformation

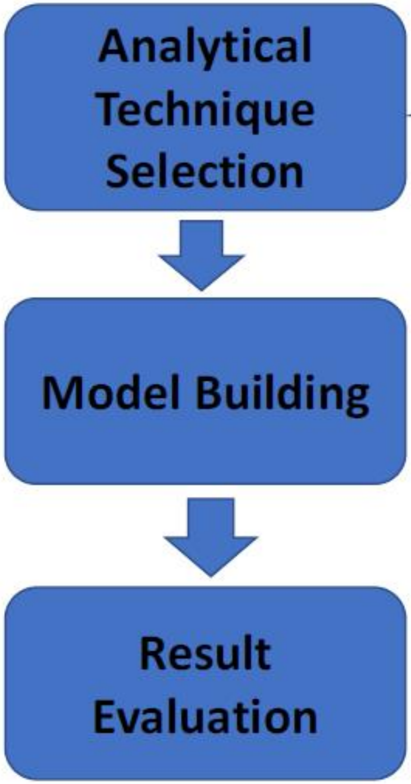


Data filtering



Data reduction

Data Analysis

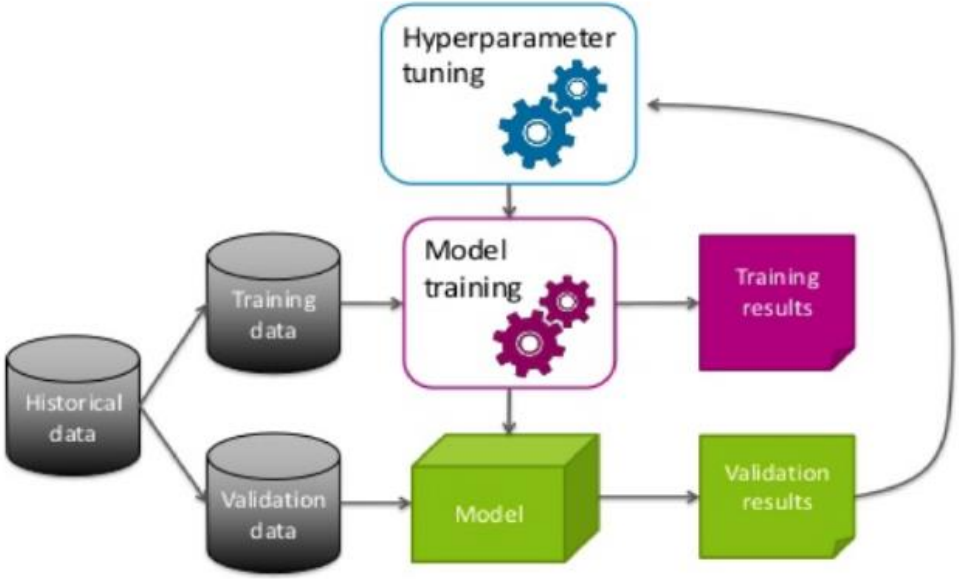


Analytical Technique Selection

Model Building

Result Evaluation

Classification, regression, clustering, association, deep learning



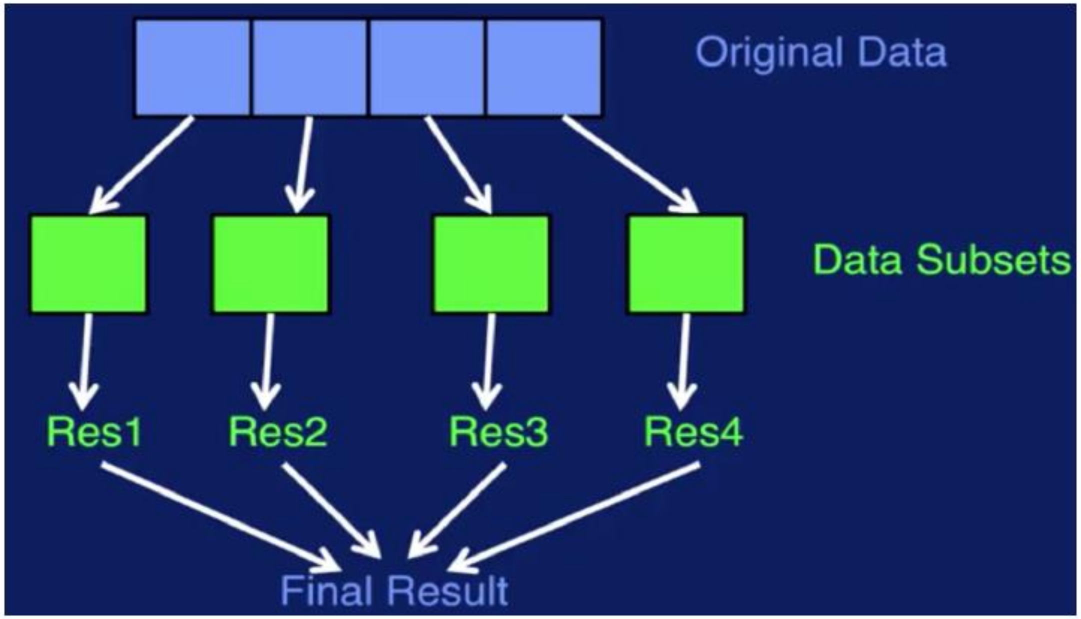
Scaling Machine Learning Algorithms

Scale up

- Use more hardware: big and stronger machine (CPUs, memory, etc.)
- Use specialized hardware: GPUs

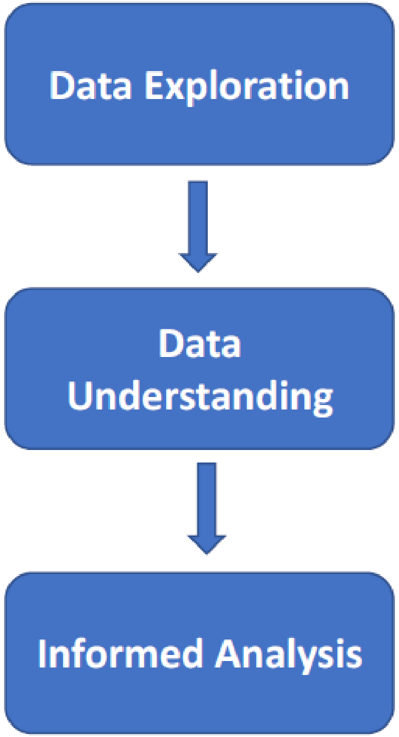
Scale out:

- Distributed approach
- Build cluster of machines
- Distribute data among the machines



Data Exploration

- Summary statistics
  - Extract summary information from the data set
  - Location: mean, median
  - Spread: standard deviation
  - Shape: skewness
- Visualization
  - View data graphically
  - Histogram
  - Scatter plot
  - Line plot


Data Exploration
Data Understanding
Informed Analysis