LONDON METROPOLITAN UNIVERSITY

Machine Learning CS7052 Lecture 2

Dr. Elaheh Homayounvala

week 2

Outline of today's lecture

- Review last week
- An introduction to Machine Learning (ML)
  - States of the Art Applications of ML
  - Types of Learning
- Understanding Data and Data Analysis Process, Nelli's book Ch. 1
- A First Application, Iris, Muller & Guido's book, Ch. 1, pp. 13-23

Review last week

- About the module, Weblearn page, Assessment
- What we'll cover in this module
- What is Machine Learning?
- When do we use ML?
- Some applications of ML

What is Machine Learning?

"Learning is any process by which a system improves performance from experience."

- Herbert Simon

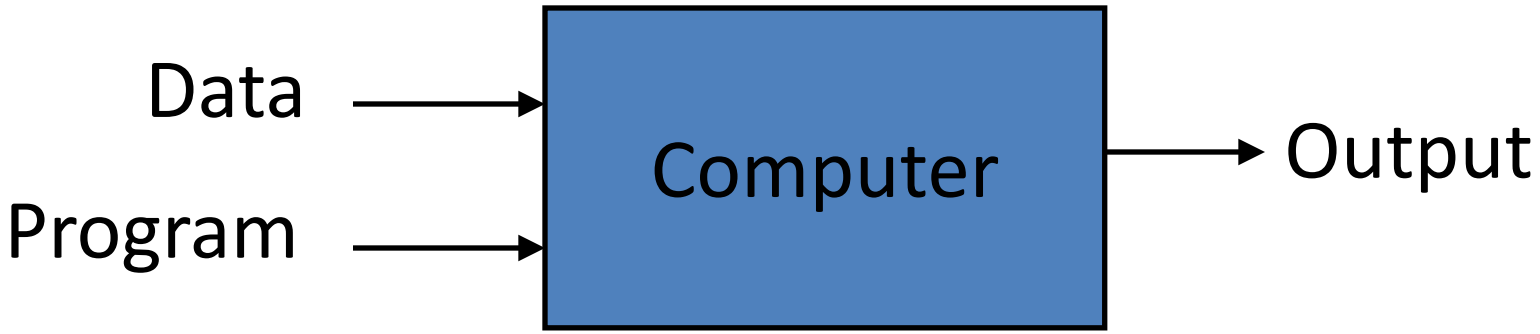
Definition by Tom Mitchell (1998):

Machine Learning is the study of algorithms that

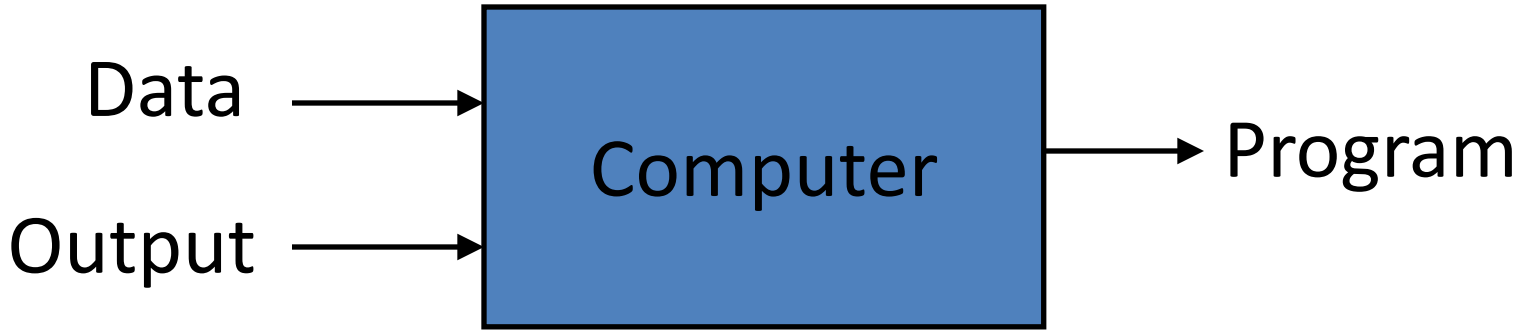
- improve their performance P
- at some task T
- with experience E.

A well-defined learning task is given by <P, T, E>.

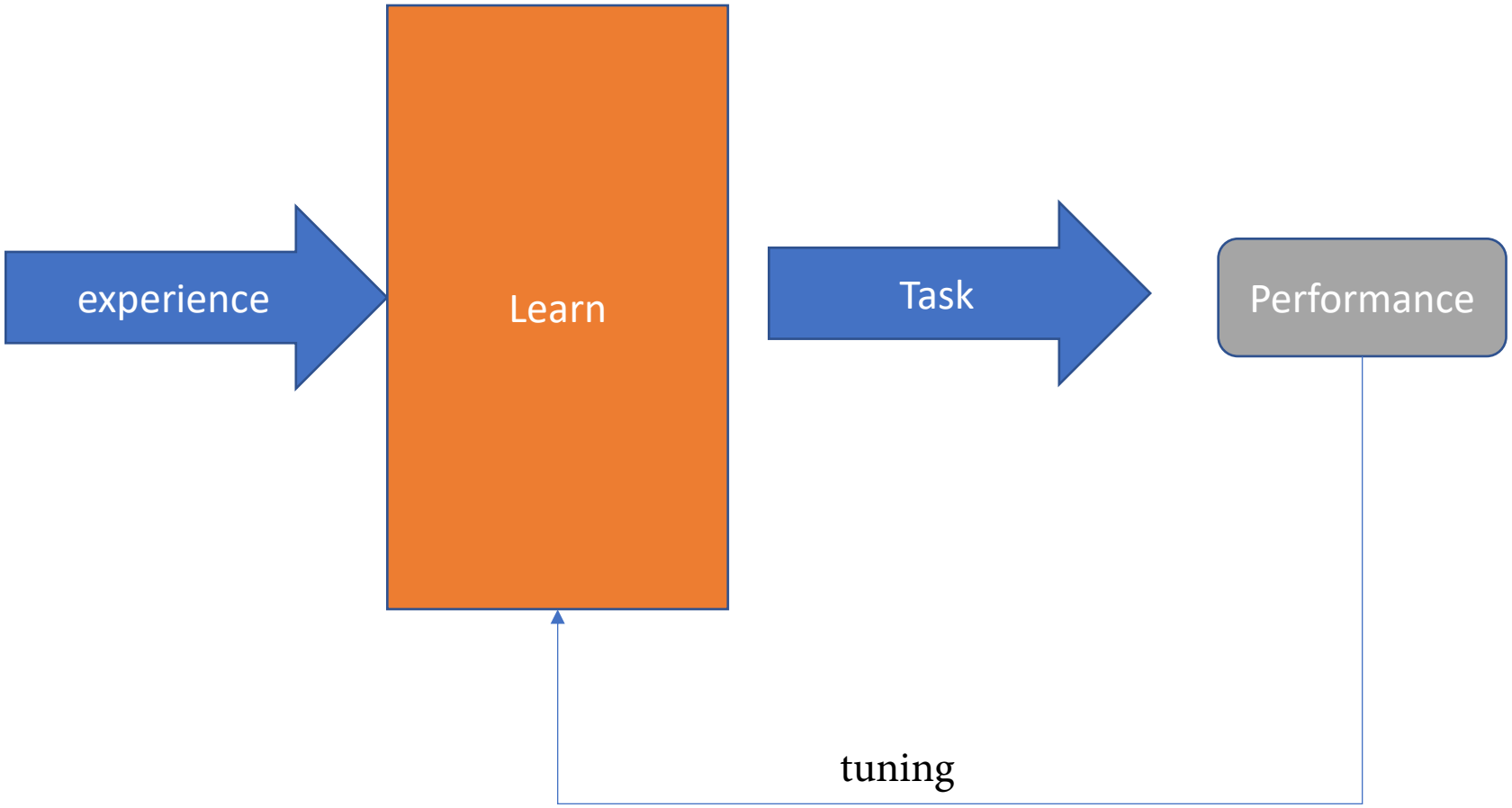
Traditional Programming



Machine Learning



What is ML?



When do we use Machine Learning?

- A pattern exists
- We do not know it mathematically
- We have data on it

State of the Art Applications of Machine Learning

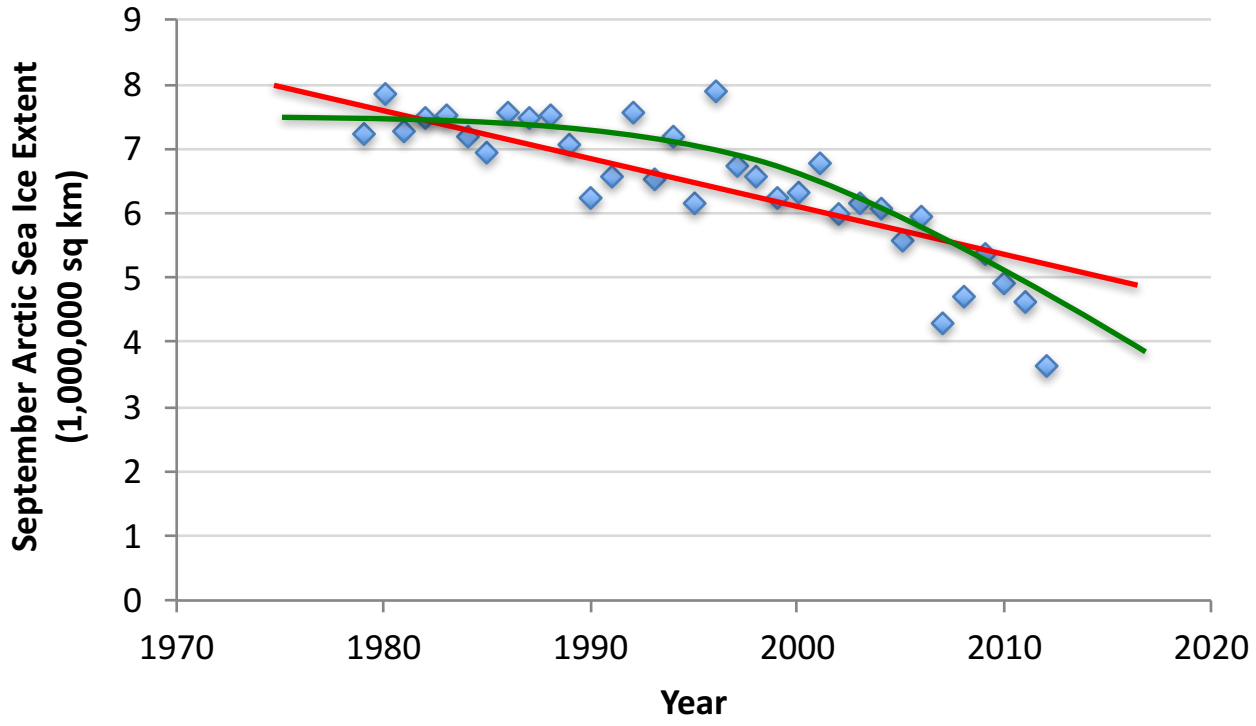
Types of Learning

Types of Learning

- **Supervised (inductive) learning**
  - Given: training data + desired outputs (labels)
- **Unsupervised learning**
  - Given: training data (without desired outputs)
- **Semi-supervised learning**
  - Given: training data + a few desired outputs
- **Reinforcement learning**
  - Rewards from sequence of actions

Supervised Learning: Regression

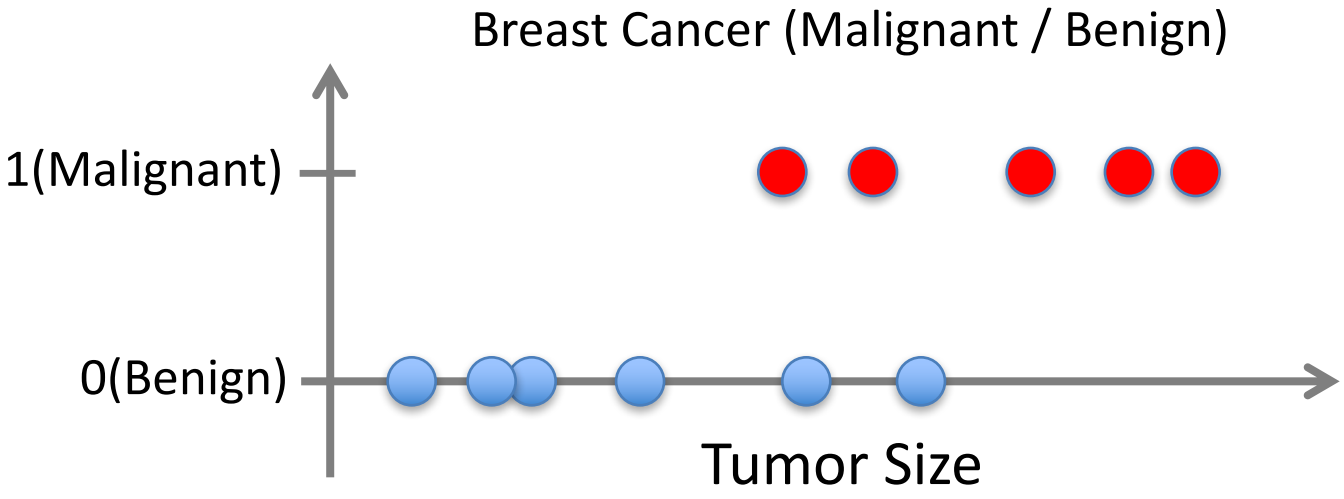
- Given (x_1, y_1), (x_2, y_2), …, (x_n, y_n)
- Learn a function $f(x)$ to predict y given x
  - y is real-valued == regression



Data from G. Witt. Journal of Statistics Education, Volume 21, Number 1 (2013)

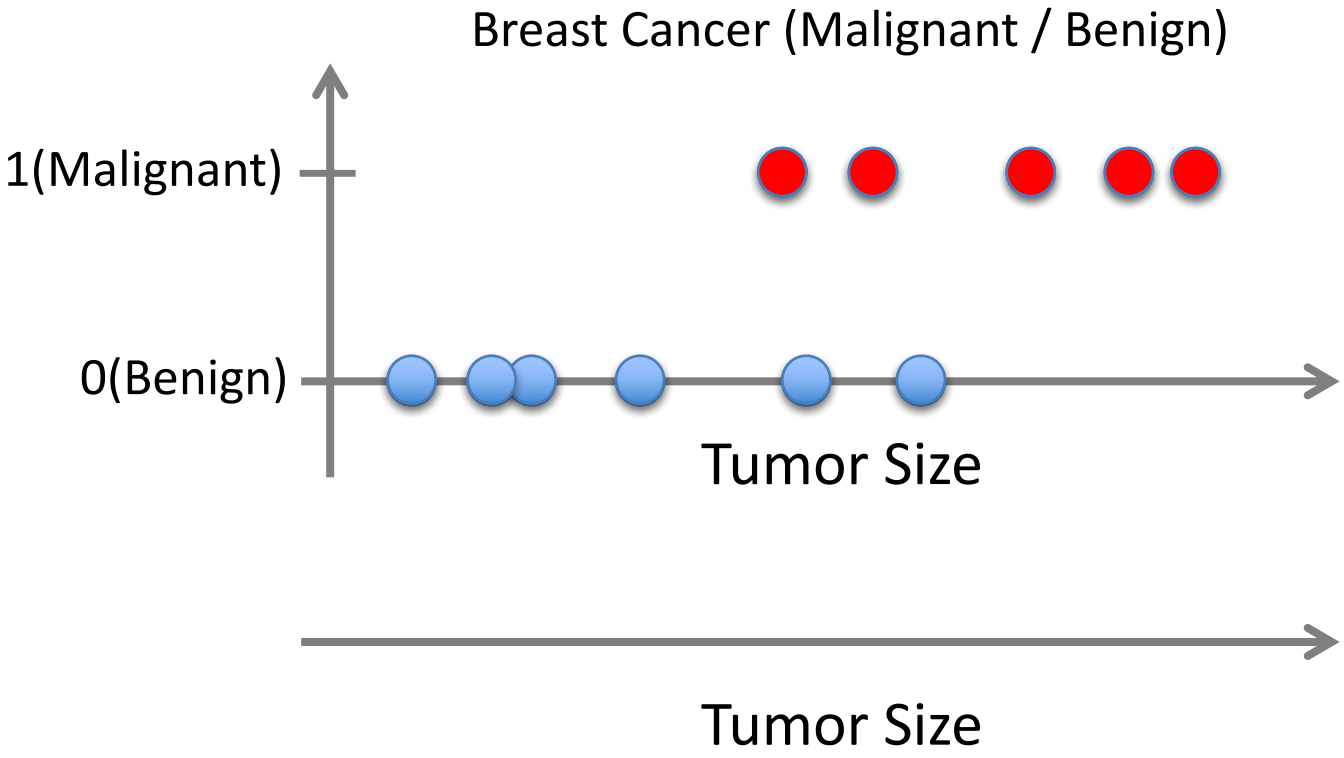
Supervised Learning: Classification

- Given (x_1, y_1), (x_2, y_2), ..., (x_n, y_n)
- Learn a function $f(x)$ to predict y given x
  - y is categorical == classification



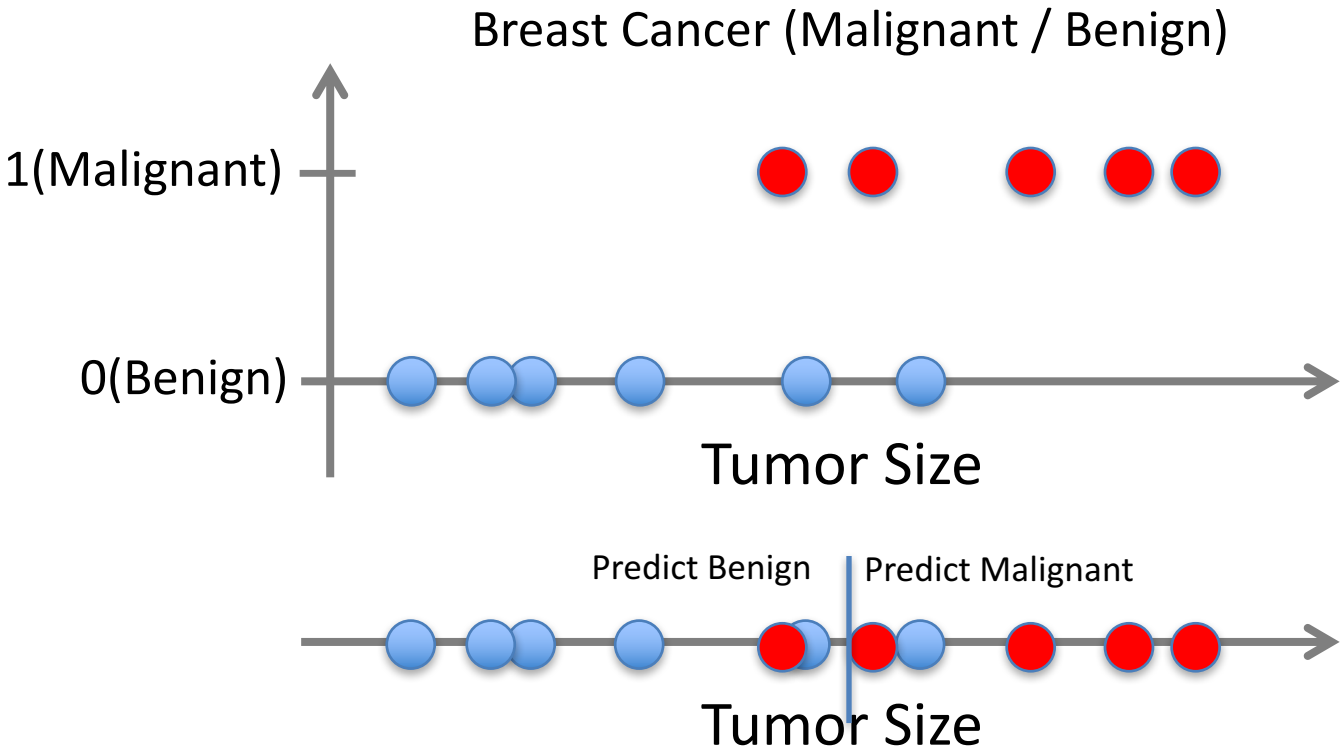
Supervised Learning: Classification

- Given (x_1, y_1), (x_2, y_2), ..., (x_n, y_n)
- Learn a function $f(x)$ to predict y given x
  - y is categorical == classification



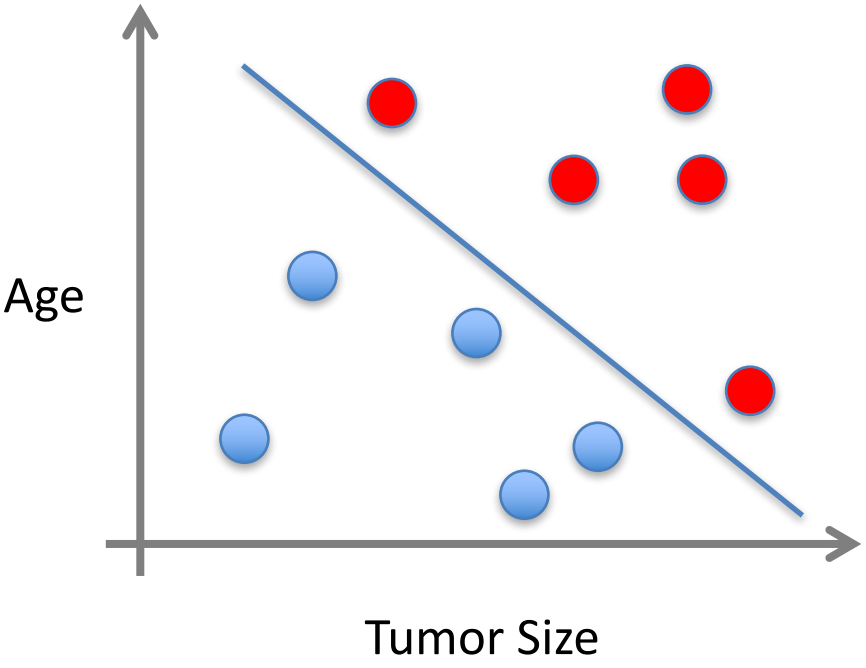
Supervised Learning: Classification

- Given (x_1, y_1), (x_2, y_2), ..., (x_n, y_n)
- Learn a function $f(x)$ to predict y given x
  - y is categorical == classification



Supervised Learning

- x can be multi-dimensional
  - Each dimension corresponds to an attribute

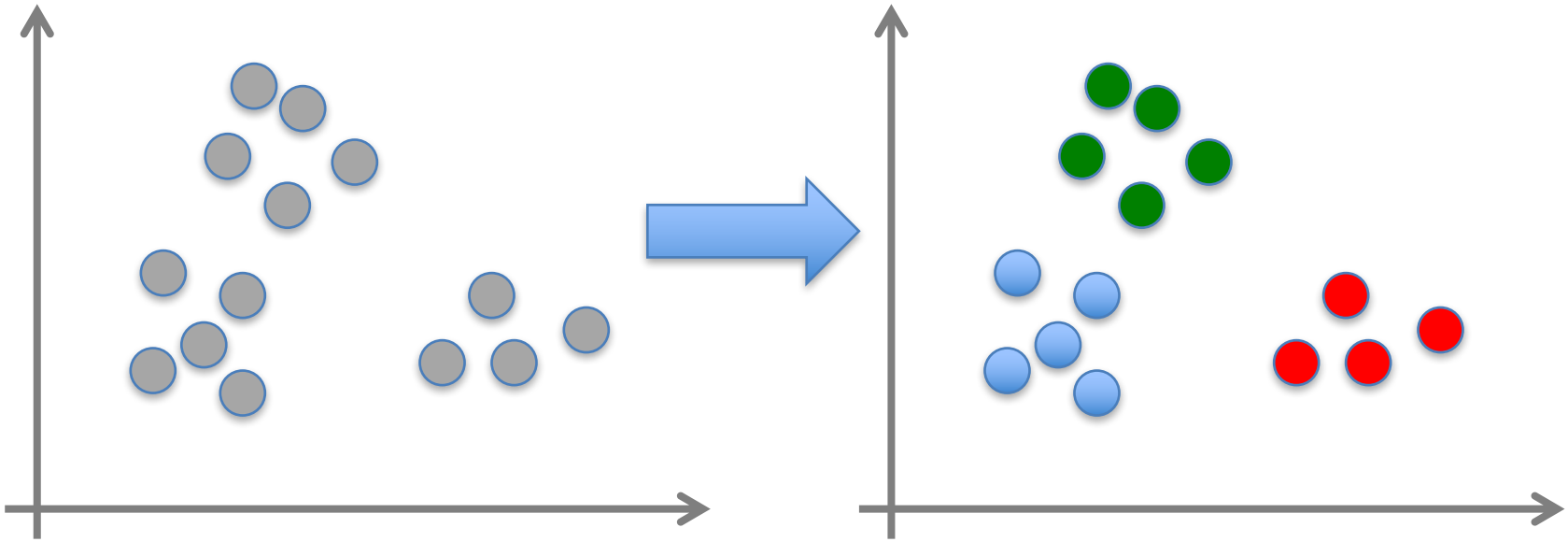


- Clump Thickness
- Uniformity of Cell Size
- Uniformity of Cell Shape

…

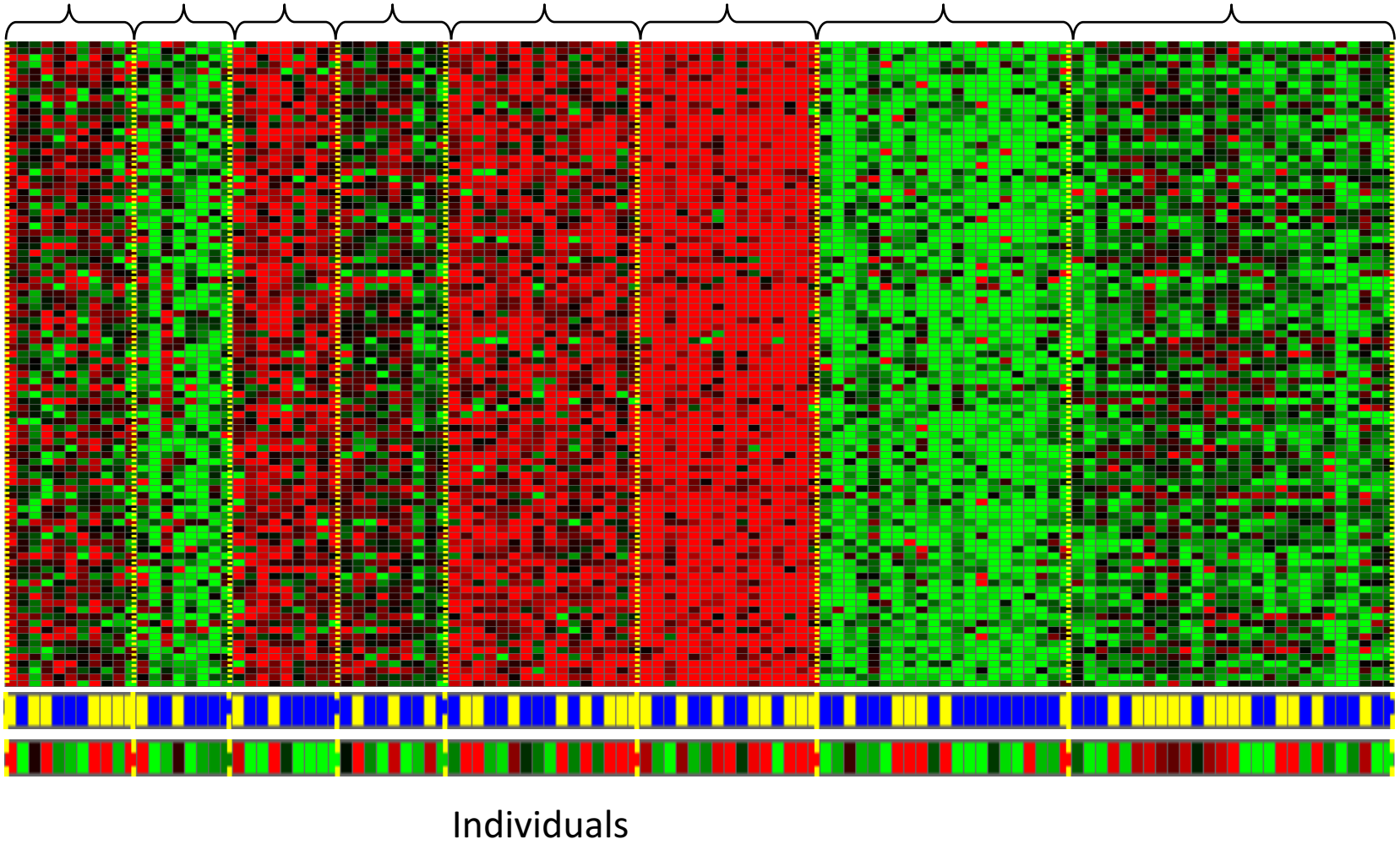
Unsupervised Learning

- Given x_1, x_2, …, x_n (without labels)
- Output hidden structure behind the x's
  - E.g., clustering



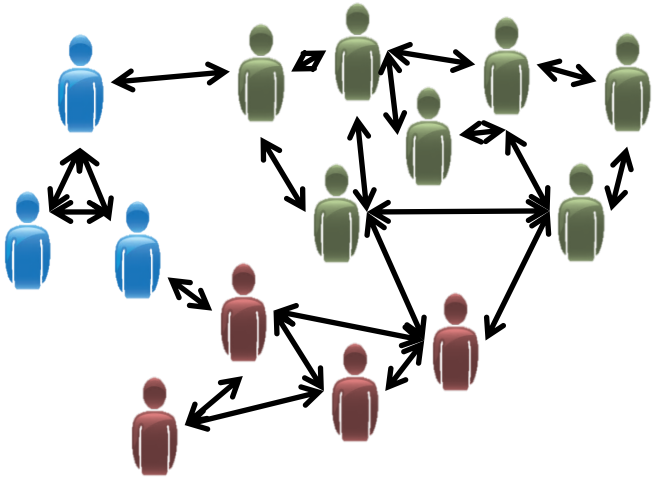
Unsupervised Learning

Genomics application: group individuals by genetic similarity



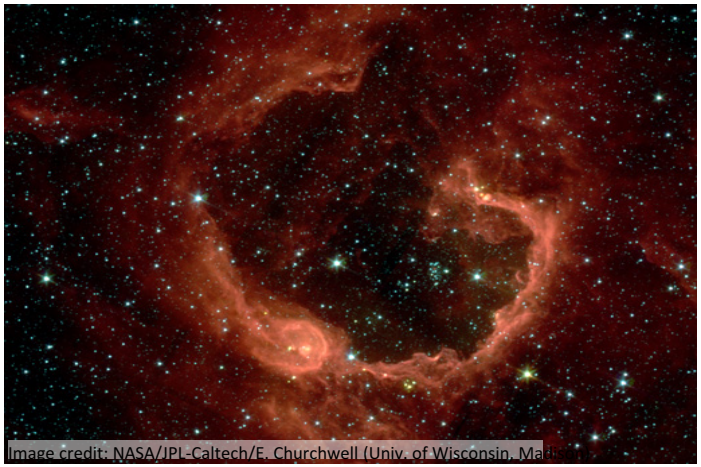
Unsupervised Learning

Organize computing clusters



Social network analysis

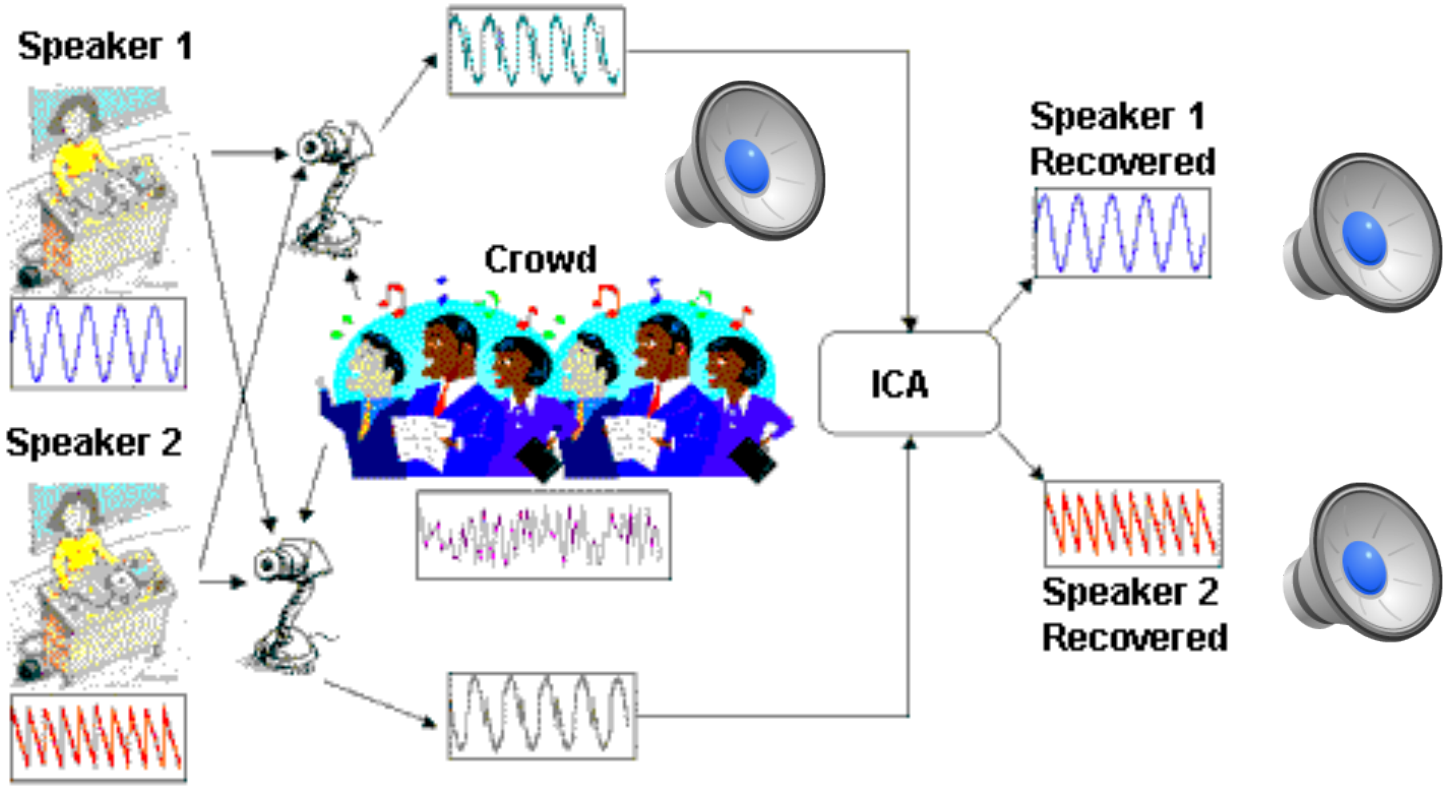
Market segmentation



Astronomical data analysis

Unsupervised Learning

- Independent component analysis – separate a combined signal into its original sources

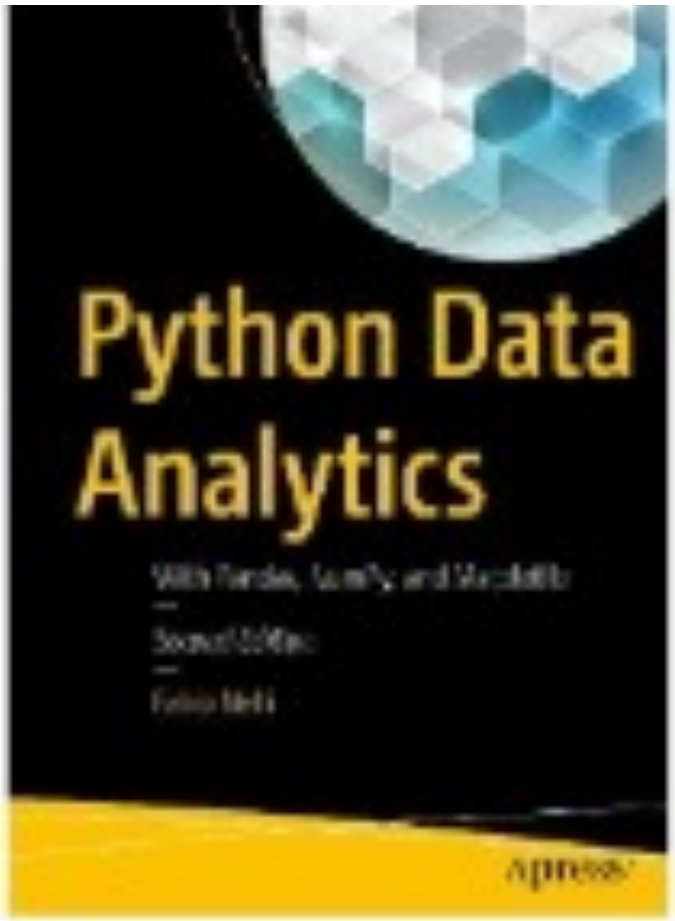


Reinforcement Learning

- Given a sequence of states and actions with (delayed) rewards, output a policy
  - Policy is a mapping from states → actions that tells you what to do in a given state
- Examples:
  - Credit assignment problem
  - Game playing
  - Robot in a maze
  - Balance a pole on your hand

Understanding Data and Data Analysis Process

Nelli's book Chapter 1



Understanding the nature of the Data

- When the data become information
- When the information becomes knowledge

Types of Data

- Categorical (nominal and ordinal)
- Numerical (discrete and continuous)

Data Analysis Process

is a process

- consisting of several steps
- in which the **raw data** are transformed and processed
- in order to produce **data visualisations** and
- make **predictions**
- thanks to **a mathematical model** (in this module, a ML model)
- based on the **collected data**.

- Source: Nelli's book page 6

Data Analysis Process

- Source: Nelli's book page 8

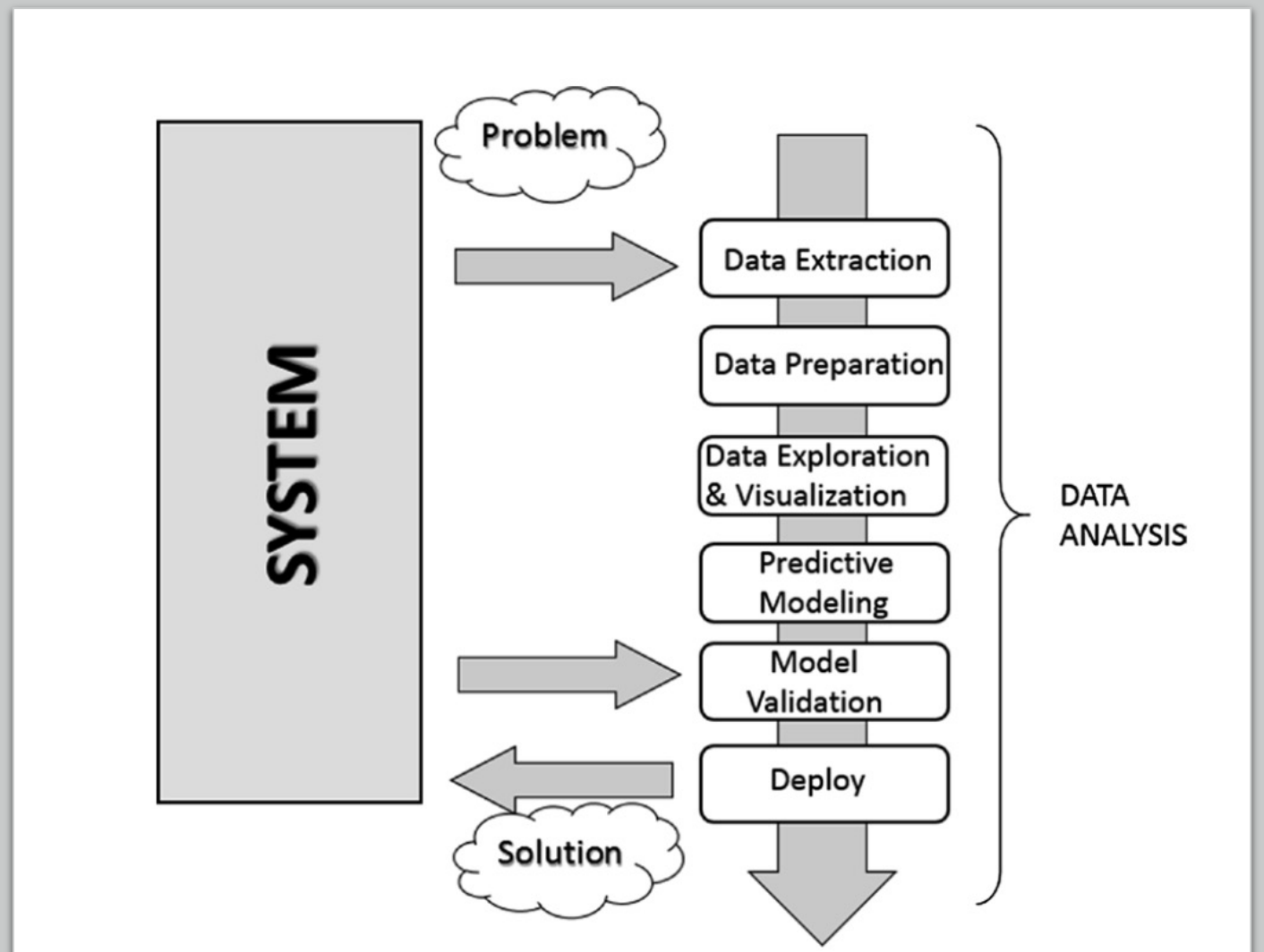


Figure 1-1. *The data analysis process*

Problem Definition

- A new problem to be solved
- Focus on the system you want to study
  - A mechanism
  - An application
  - A process
- Prepare documentation
- Project planning
  - Resources
  - Issues
  - Team

Data Extraction

- Obtain the data
- If sample data to be collected, does it reflect as much as possible the real world?
- Quality and quantity of data
- Data sources (experimental data or otherwise)

Data Preparation

- Obtaining
- Cleaning
- Normalising
- Transforming into an optimized dataset

Issues: invalid, ambiguous, or missing values, replicated fields, and out-of-range data

Data Exploration/Visualisation

Exploring data involves:

- Searching data in a graphical or statistical presentation
- in order to find patterns, connections and relationships
- Data visulaisation is the best tool to highlight possible patterns
- Data visulisations may consists of
  - summerisation,
  - grouping data,
  - exploring the relationship between the various attributes,
  - identifying patterns and trends and more

Predictive Modeling

- Classification models
  - If the result obtained by the model type is categorical
- Regression models
  - If the result obtained by the model type is numeric
- Clustering models
  - If the result obtained by the model type is descriptive

Model Validation

- Validate the model built on the basis of the starting data

- Training data for building the model
- Validation set for validating the model

- Model validation numerically evaluates the effectiveness of the model
- Cross-validation

Deployment

Deploy and document

- Analysis results
- Decision deployment
- Risk analysis
- Measuring the business impact

Quantitative and Qualitative Data Analysis

Source: Nelli's book, page 15

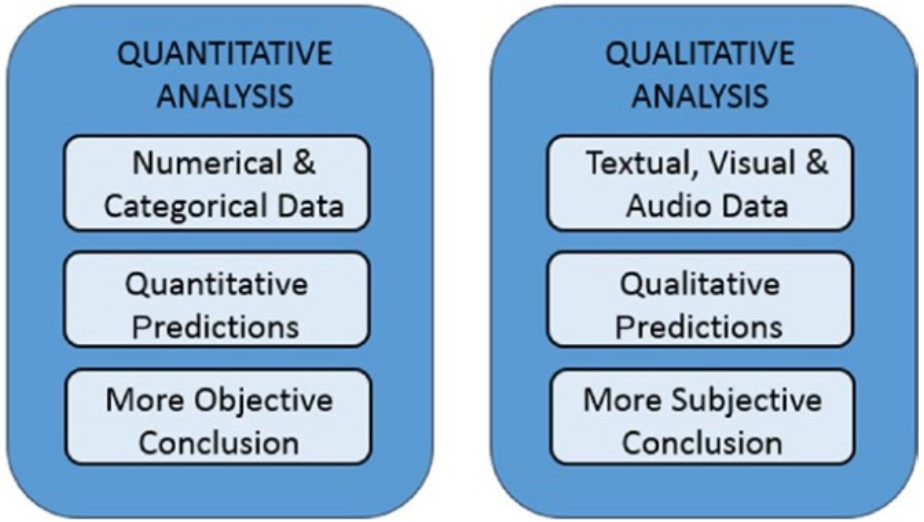


Figure 1-2. *Quantitative and qualitative analyses*

Open Data

- DataHub
- World Health Organisation
- Data.gov
- European Union Open Data Portal
- Amazon Web Service Public datasets
- Facebook Graph
- Healthdata.gov
- Google Trends, Google Finance, Google Books Ngrams
- Machine Learning Repository

LOD cloud diagram

- Nelli's book page 16

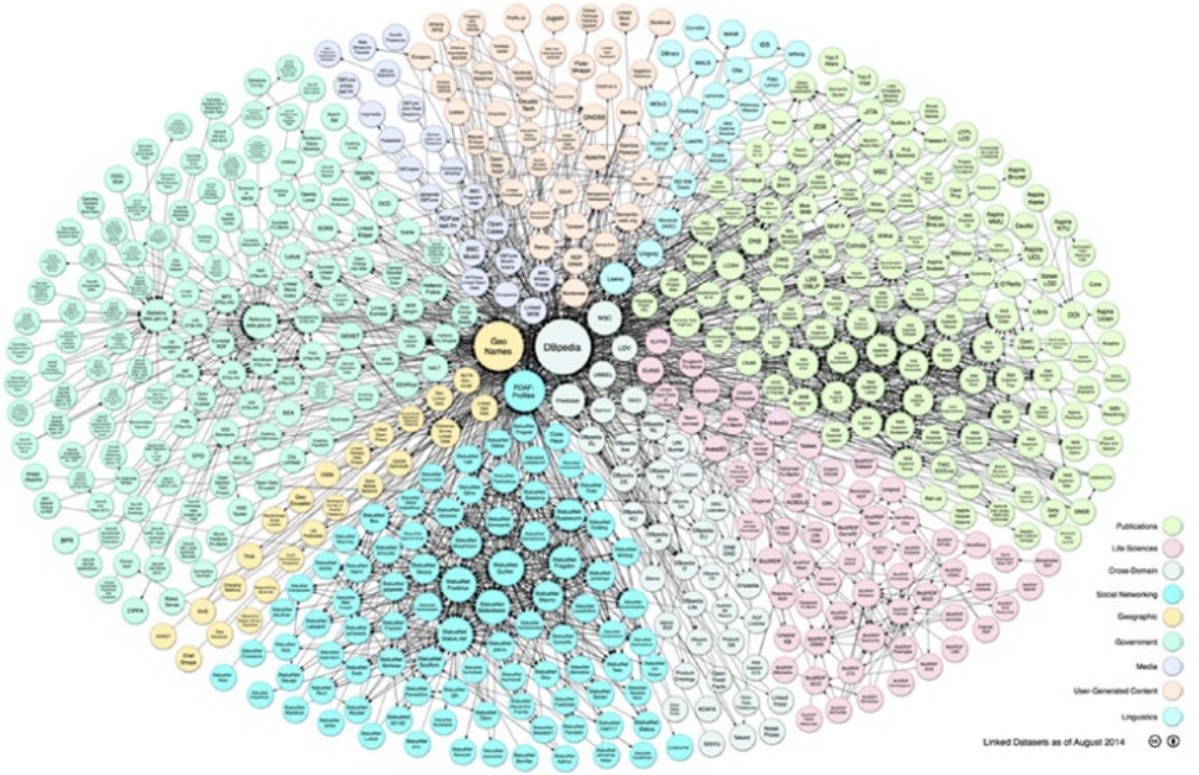
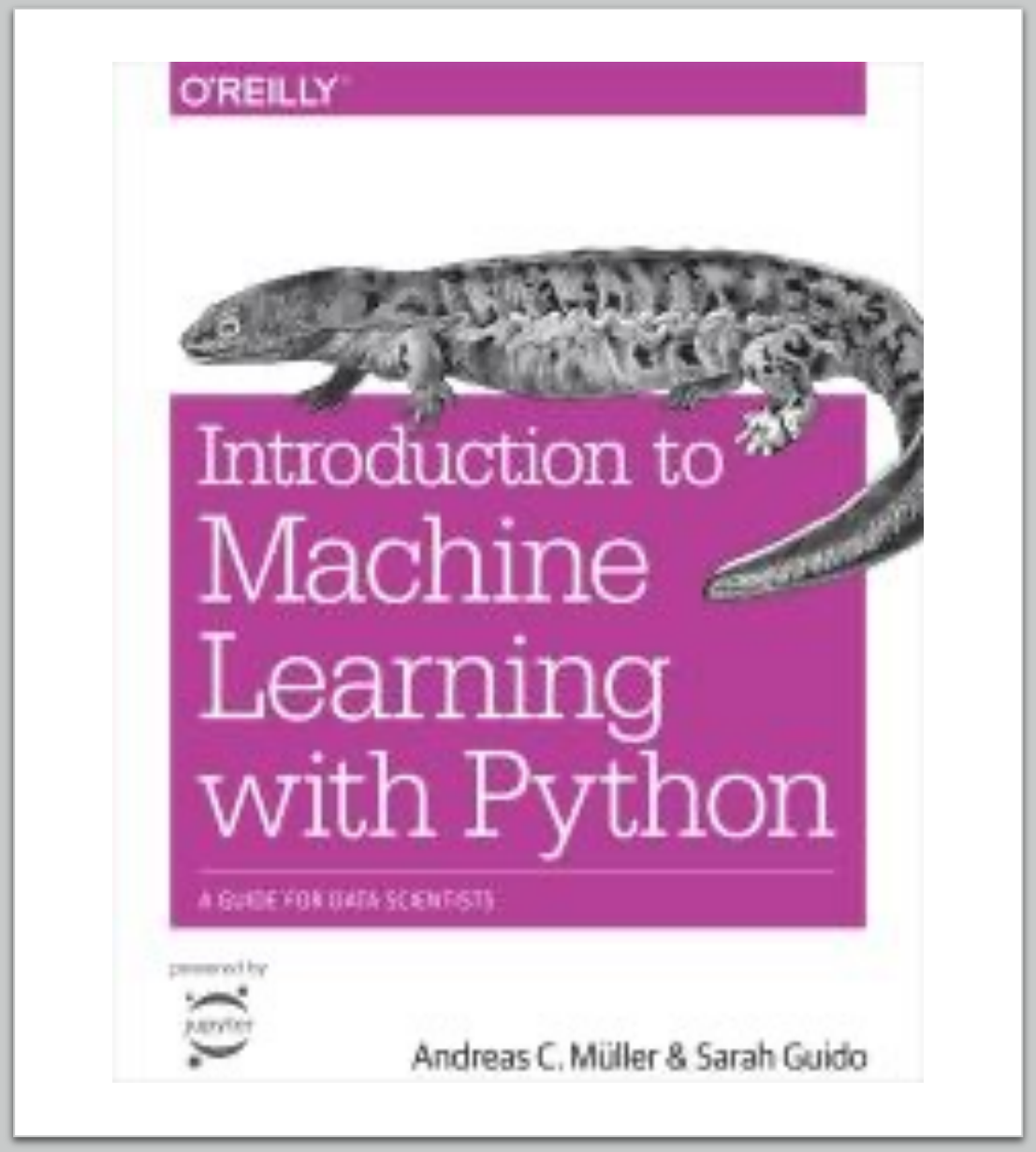


Figure 1-3. *Linking open data cloud diagram 2014, by Max Schmachtenberg, Christian Bizer, Anja Jentzsch, and Richard Cyganiak.* `http://lod-cloud.net/` *[CC-BY-SA license]*

A First Application, Classifying Iris Species

- Muller & Guido's book, Chapter 1, pp. 13-23



A First Application, Iris

- A hobby botanist is interested in distinguishing the species of some iris flower

- Source: Muller & Guido's book, page

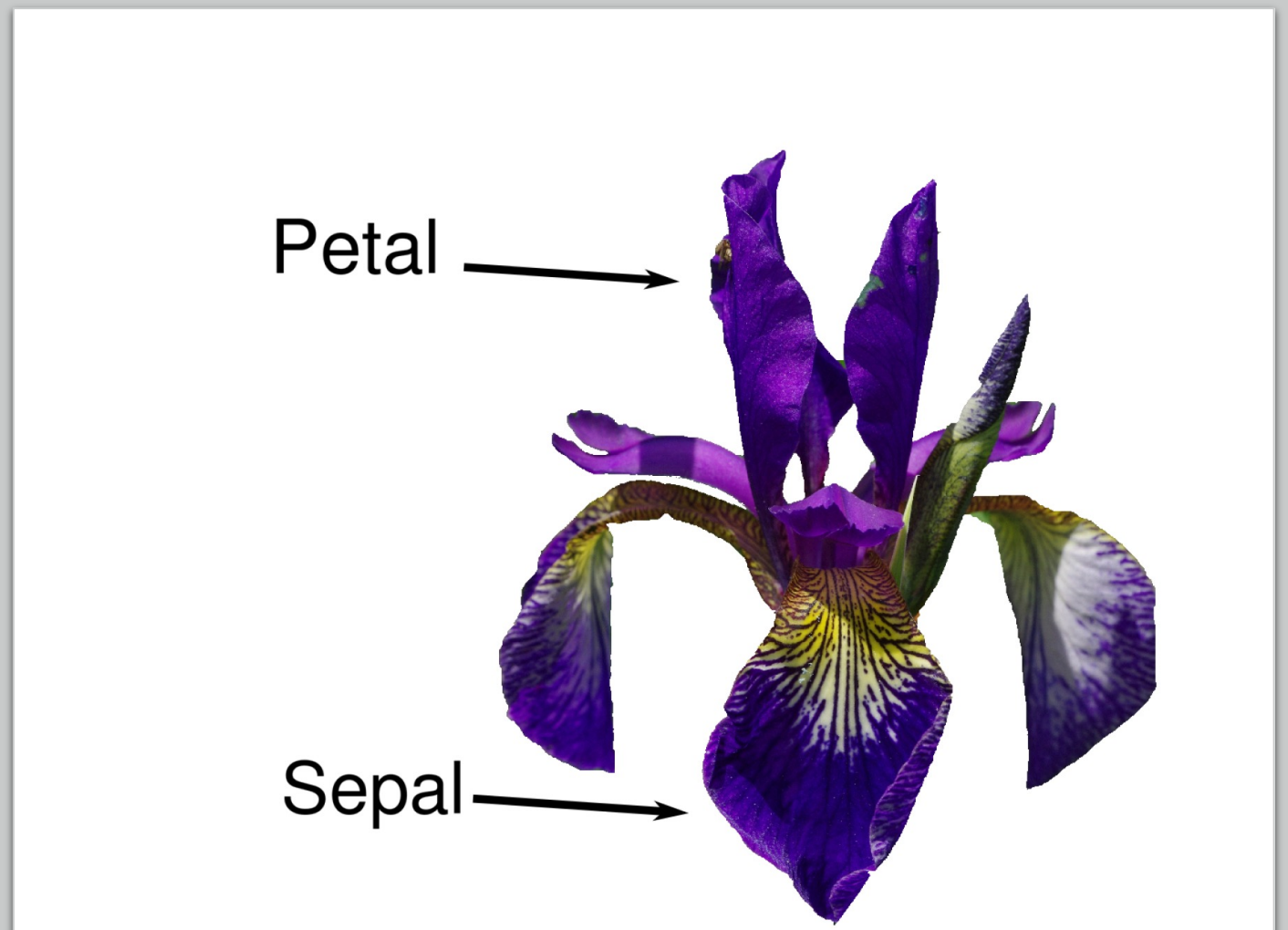


Figure 1-2. Parts of the iris flower

Measurements collected

- The length and the width of the petals in cm
- The length and the width of the sepals in cm

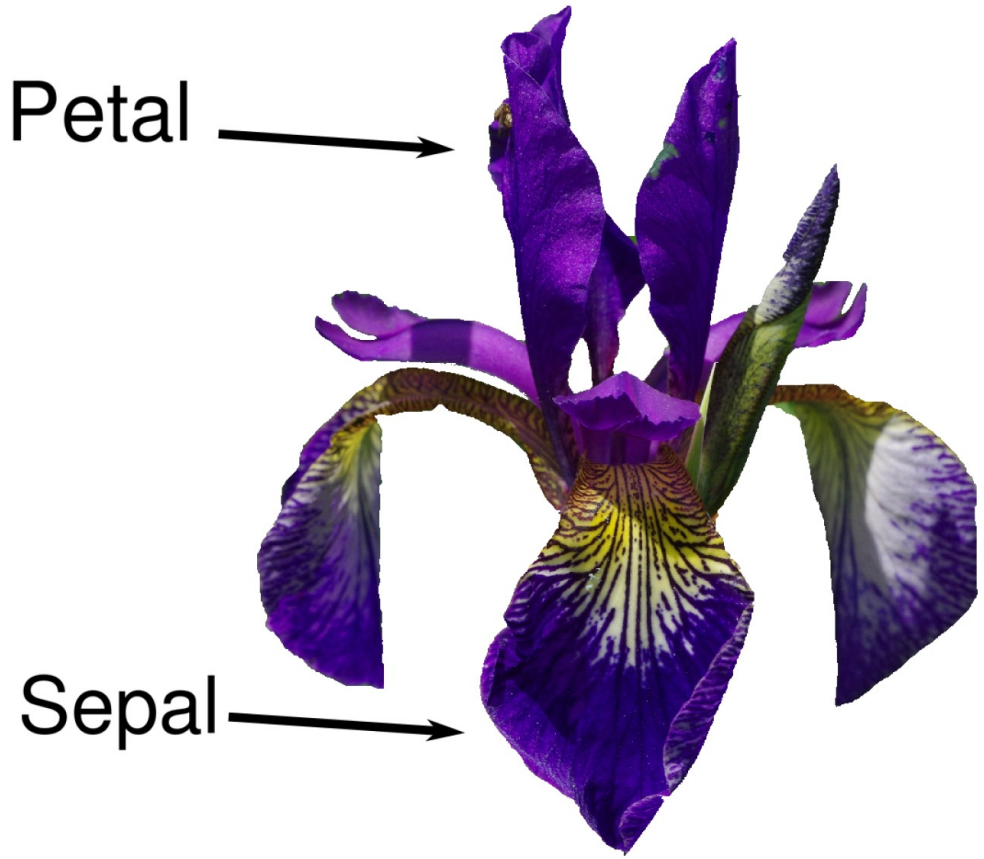


Figure 1-2. Parts of the iris flower

Three species

Our goal

Build a machine learning model

- that can learn from the measurements of these irises whose species is known
- so that we can predict the species for a new iris

Which machine learning model?

- Supervised or unsupervised?
- Classification, clustering, regression?

Today’s workshop

- Meet the data
- Training and testing data
- Look at your data (visualization)
- Building your first model
- Making predictions
- Evaluating the model

- As you can see above, you are going to practice some of the steps in data analysis process