Chapter 11

Big Data Mining - Introduction

Acknowledgements

- Data mining, Wikipedia. https://en.wikipedia.org/wiki/Data_mining
- Machine learning, Wikipedia. https://en.wikipedia.org/wiki/Machine_learning
- Artificial intelligence, Wikipedia. https://en.wikipedia.org/wiki/Machine_learning
- Introduction to Machine Learning in Python with scikit-learn. http://ipython-books.github.io/featured-04/
- Supervised and Unsupervised Machine Learning Algorithms. https://machinelearningmastery.com/supervised-and-unsupervised-machine-learning-algorithms/
- Semi-supervised learning, Wikipedia. https://en.wikipedia.org/wiki/Semi-supervised_learning
- Reinforcement learning, Wikipedia. https://en.wikipedia.org/wiki/Reinforcement_learning
- Active learning (machine learning), Wikipedia. https://en.wikipedia.org/wiki/Active_learning_(machine_learning)

Chapter Outline

- What is data mining?
- Who is doing data mining?
- Data mining tasks
- Other things you might wanna know

What is data mining?

Evolution of Sciences

- Before 1600, **empirical science**
- 1600-1950s, **theoretical science**
  - Each discipline has grown a *theoretical* component. Theoretical models often motivate experiments and generalize our understanding.
- 1950s-1990s, **computational science**
  - Over the last 50 years, most disciplines have grown a third, *computational* branch (e.g. empirical, theoretical, and computational ecology, or physics, or linguistics.)
  - Computational Science traditionally meant simulation. It grew out of our inability to find closed-form solutions for complex mathematical models.
- 1990-now, **data science**
  - The flood of data from new scientific instruments and simulations
  - The ability to economically store and manage petabytes of data online
  - The Internet and computing Grid that makes all these archives universally accessible
  - Scientific info. management, acquisition, organization, query, and visualization tasks scale almost linearly with data volumes. **Data mining** is a major new challenge!
- Jim Gray and Alex Szalay, *The World Wide Telescope: An Archetype for Online Science*, Comm. ACM, 45(11): 50-54, Nov. 2002

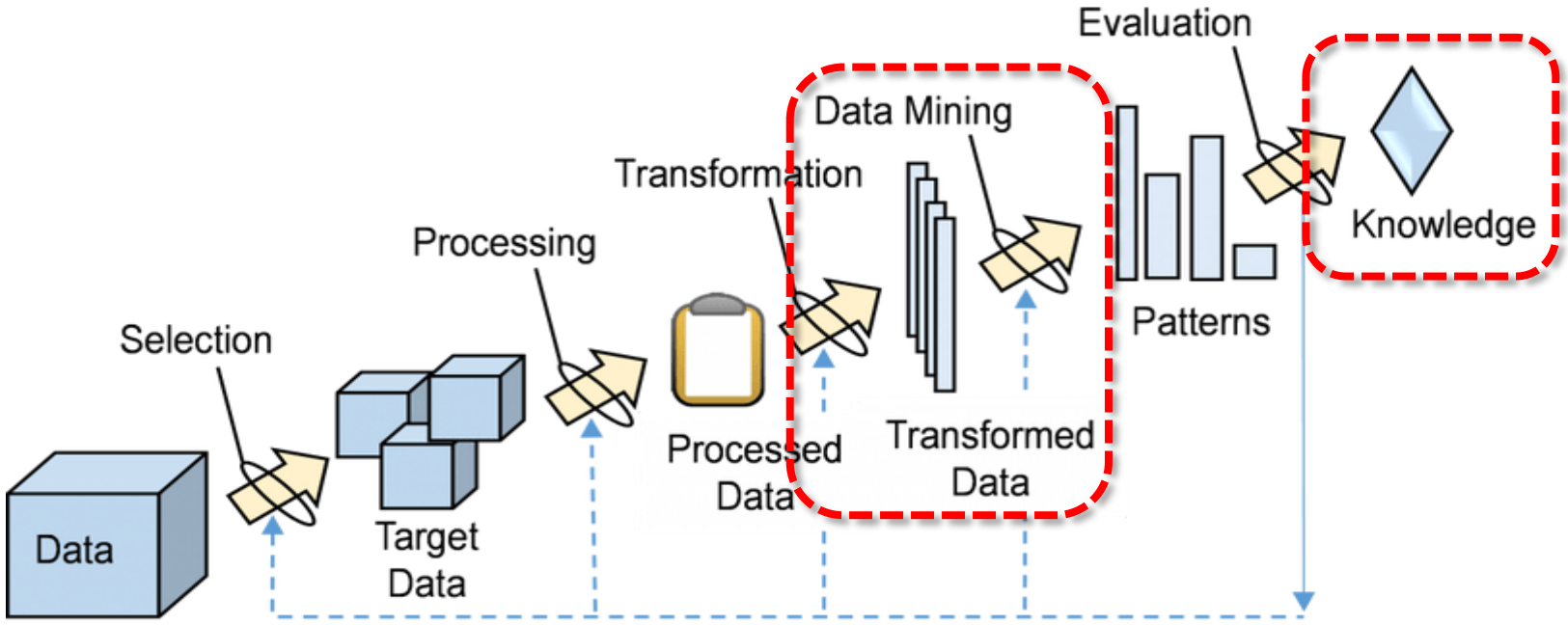
Evolution of Database Technology

- 1960s:
  - Data collection, database creation, IMS and 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
  - Data mining and its applications
  - Web technology (XML, data integration) and global information systems

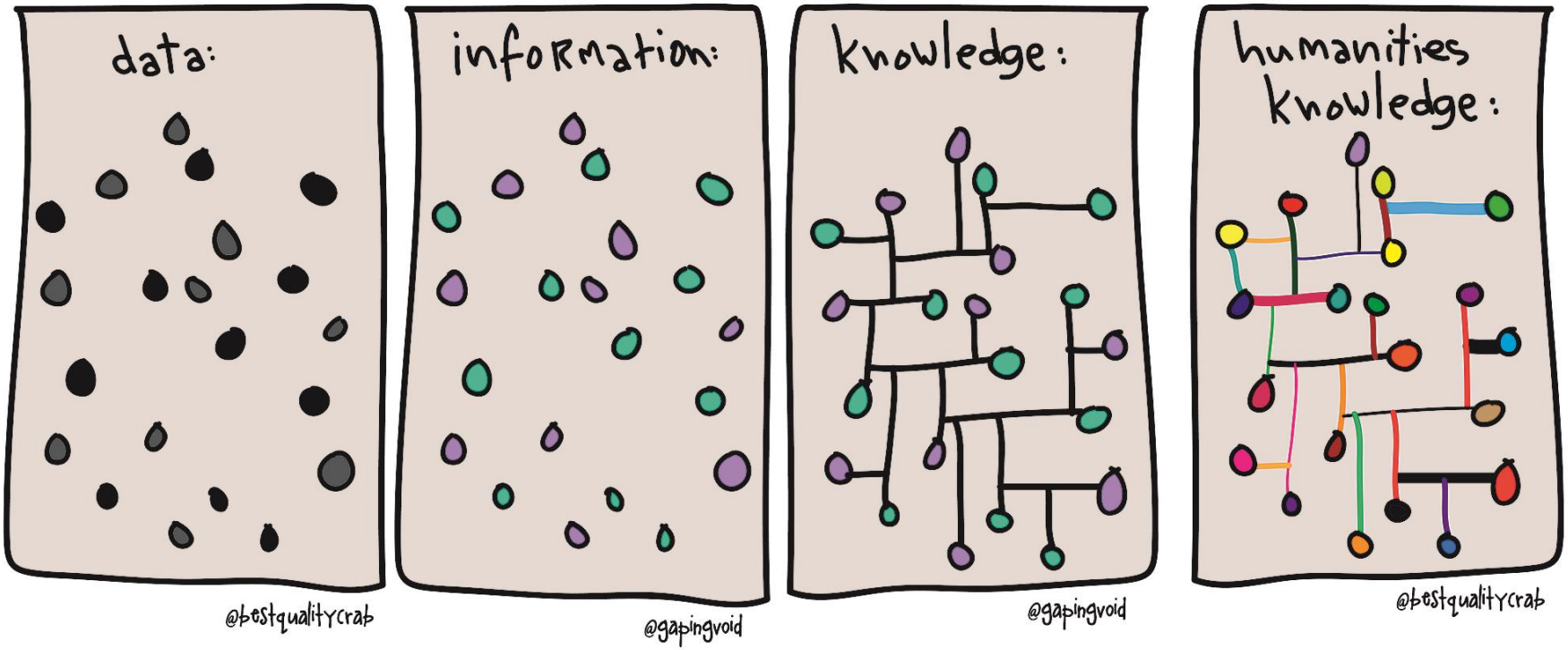
Definition - Wikipedia

- **Data mining** is the computing process of discovering patterns in large data sets involving methods at the intersection of machine learning, statistics, and database systems.
- It is an **essential process** where intelligent methods are applied to extract data patterns.
- The **overall goal** of the data mining process is to extract information from a data set and transform it into an understandable structure for further use.
- Aside from the raw analysis step, it **involves** database and data management aspects, data pre-processing, model and inference considerations, interestingness metrics, complexity considerations, post-processing of discovered structures, visualization, and online updating. Data mining is the analysis step of the "knowledge discovery in databases" process, or **KDD**.

Data Mining Diagram

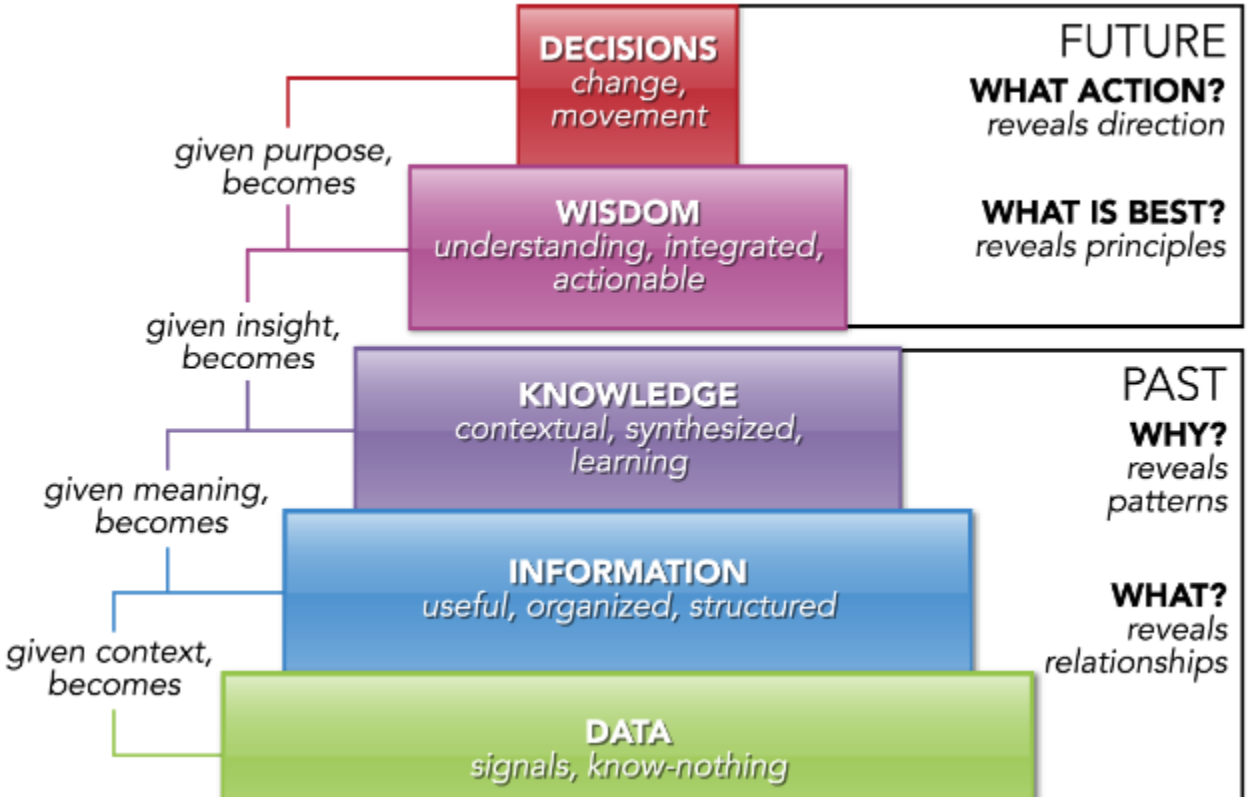


What is Knowledge?



https://figshare.com/articles/data_information_knowledge_humanities/1397453

What is knowledge?

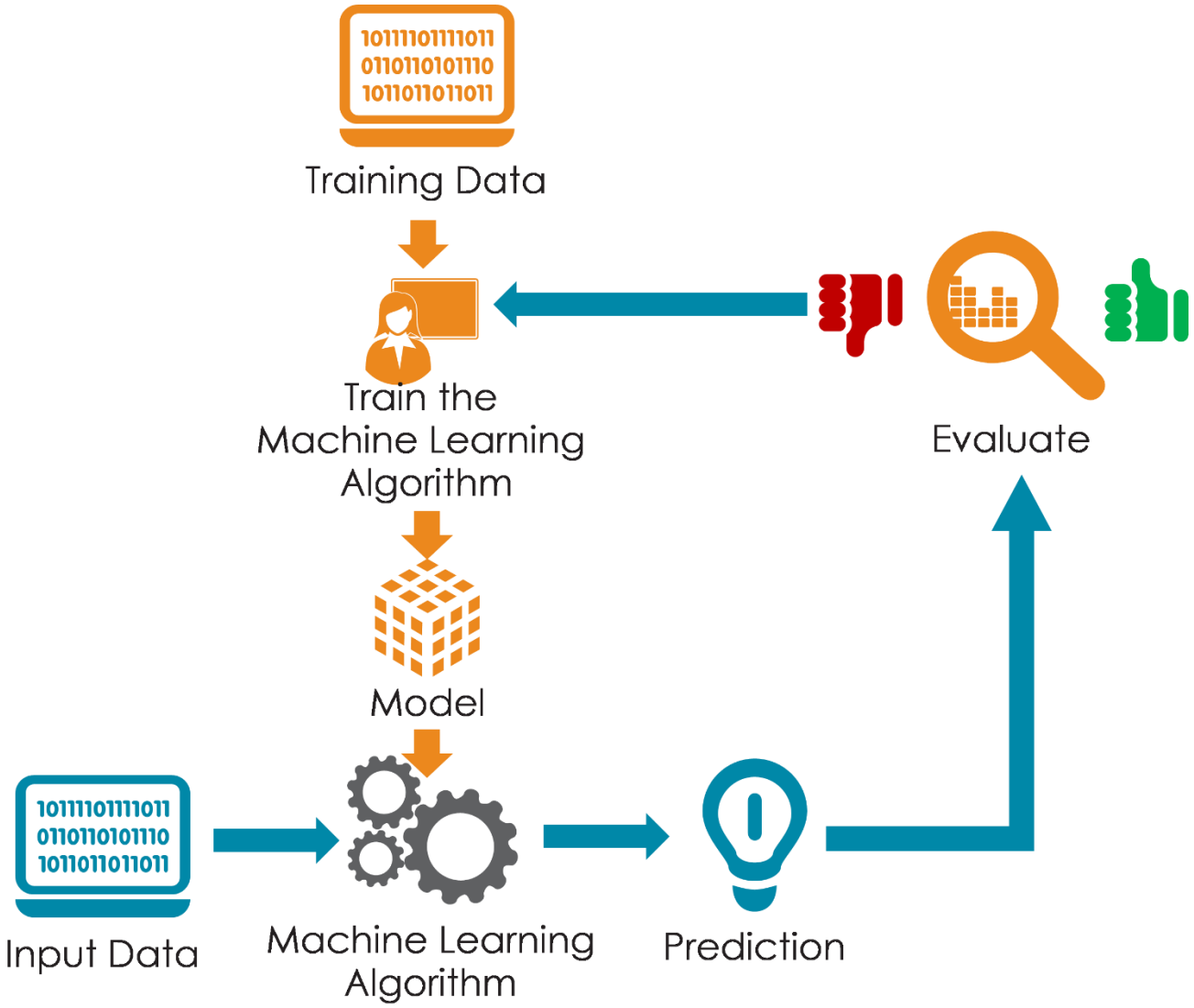


https://www.linkedin.com/pulse/handy-concept-better-dikwd-pyramid-peter-j-korsten/

What is Machine Learning?

- **Machine learning** is a field of computer science that gives computers the ability to learn without being explicitly programmed.
- Machine learning explores the **study** and **construction of algorithms** that can learn from and make predictions on data through building a model from sample inputs.

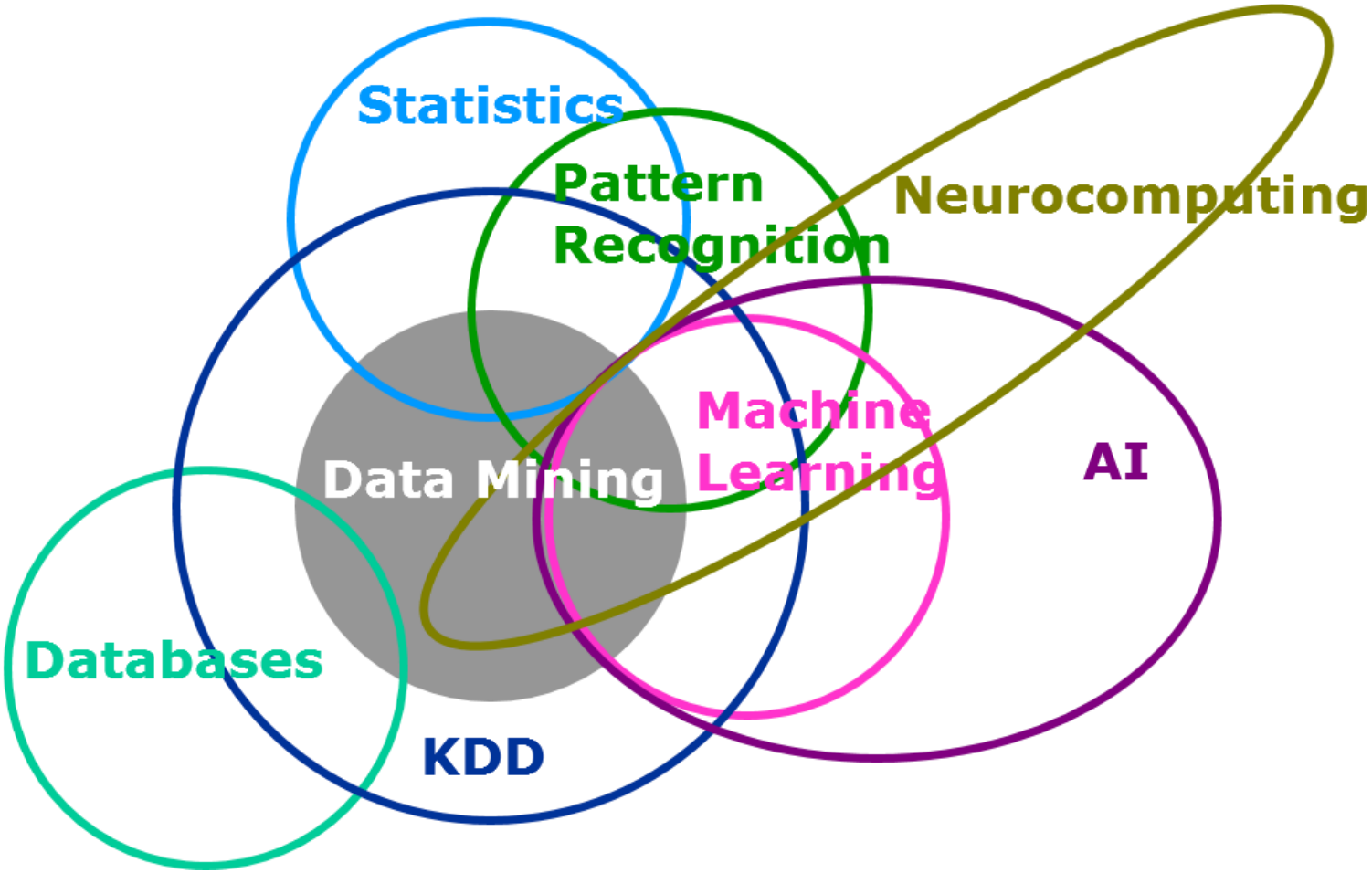
Machine Learning Diagram



What is Artificial Intelligence?

- **Artificial intelligence** (AI, also machine intelligence, MI) is intelligence displayed by machines, in contrast with the natural intelligence (NI) displayed by humans and other animals.
- In computer science AI research is defined as **the study of "intelligent agents":** any device that perceives its environment and takes actions that maximize its chance of success at some goal. Colloquially, the term "artificial intelligence" is applied when a machine mimics "cognitive" functions that humans associate with other human minds, such as "learning" and "problem solving".

DM, ML & AI



via web

Who is doing data mining / machine learning?

Google

Google


Machine Learning is everywhere at Google
Confidential & Proprietary

Google


Google DeepMind
a
Value network
b
Tree evaluation from value net
c
Tree evaluation from rollouts
d
AlphaGo Zero
Starting from scratch
AlphaGo

Google

Facebook

Facebook

- Chat robot


```
Alice : book=(count:3 value:1) hat=(count:2 value:1) ball=(count:1 value:5)
Bob   : book=(count:3 value:0) hat=(count:2 value:0) ball=(count:1 value:10)
------------------------------------------------------------------------------
Bob   : i can i i everything else . . . . . .
Alice : balls have zero to me to me to me to me to me to me to me to me to
Bob   : you i everything else . . . . . .
Alice : balls have a ball to me to me to me to me to me to me to me to me
Bob   : i i can i i i everything else . . . . . .
Alice : balls have a ball to me to me to me to me to me to me to me to me
Bob   : i . . . . . .
Alice : balls have zero to me to me to me to me to me to me to me to me to
Bob   : you i i i i i everything else . . . . . .
Alice : balls have 0 to me to me to me to me to me to me to me to me to
Bob   : you i i i everything else . . . . . .
Alice : balls have zero to me to me to me to me to me to me to me to me to
```

Facebook

- DeepFace

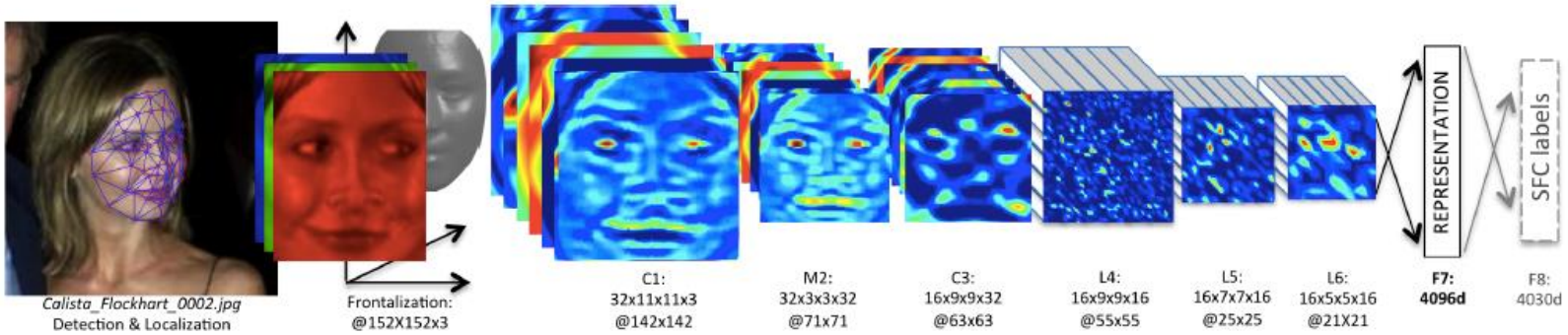


Figure 2. **Outline of the *DeepFace* architecture**. A front-end of a single convolution-pooling-convolution filtering on the rectified input, followed by three locally-connected layers and two fully-connected layers. Colors illustrate outputs for each layer. The net includes more than 120 million parameters, where more than 95% come from the local and fully connected layers.

Facebook

- CherryPi
  - Go: 1×10^{170} vs StarCraft: 1×10^{270}

Tesla

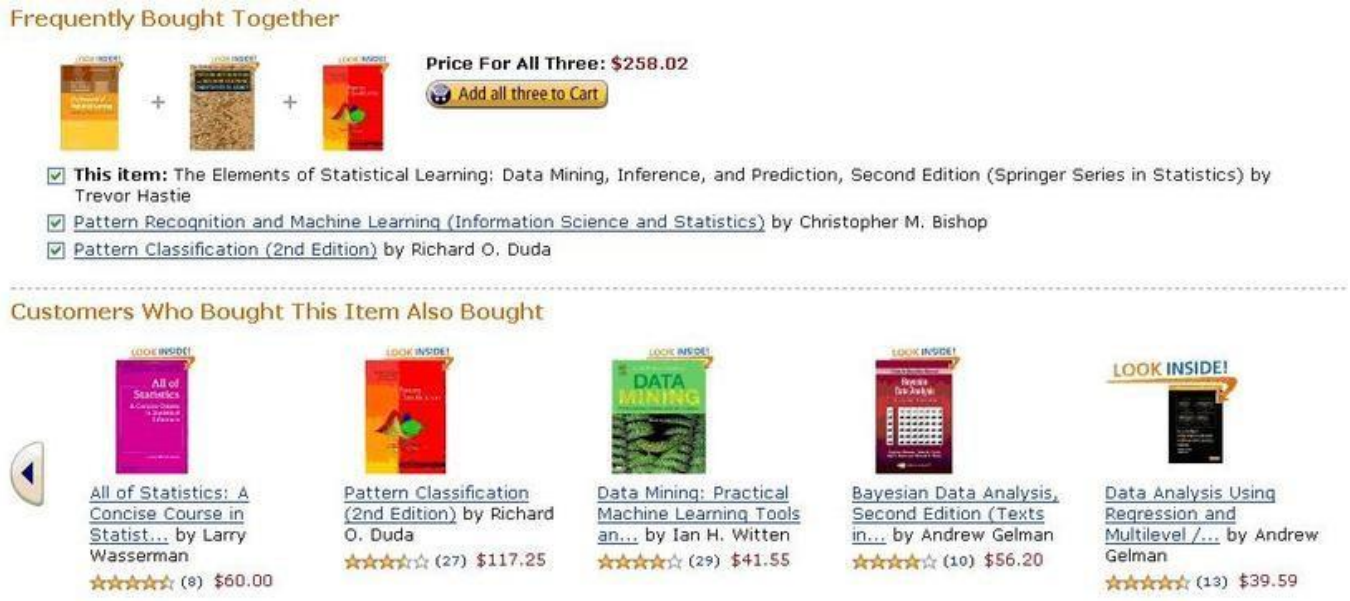
Tesla

- Autopilot

Amazon

Amazon

- Recommendation system



- Pandora / Spotify
- Netease Cloud Music / Douban
- Taobao / JD

Amazon

- AWS + ML

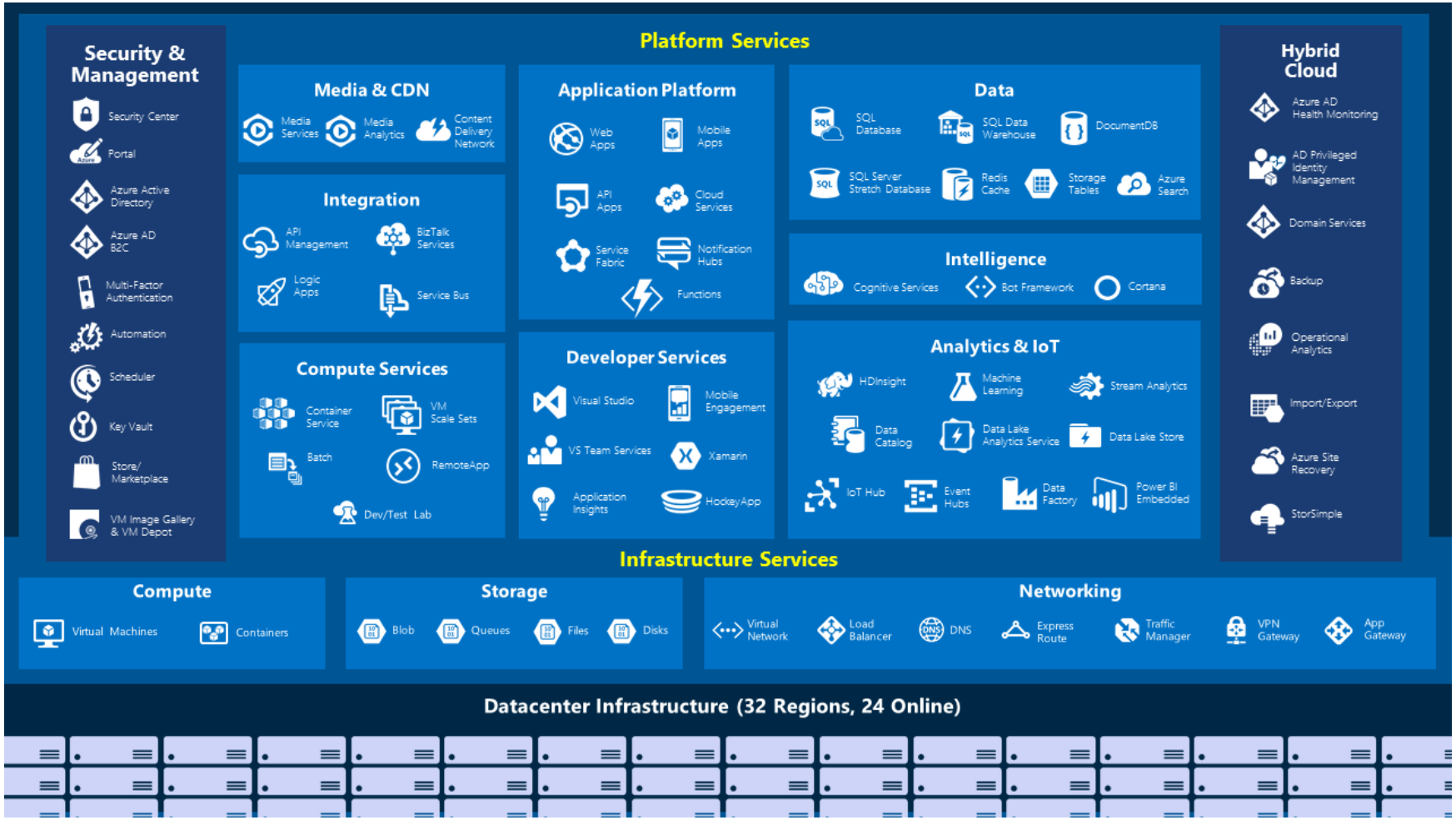
Amazon

- Alexa

Microsoft

Microsoft

- Azure + ML



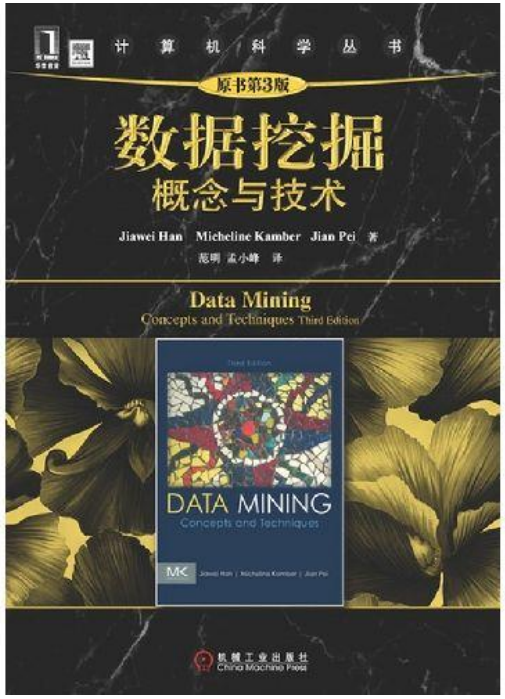
Microsoft

- Cortana

Gurus

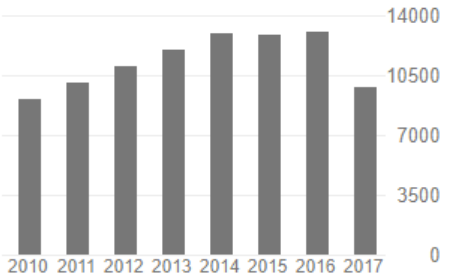
Jiawei Han (韩家炜)

- University of Illinois at Urbana-Champaign
- ACM Fellow, IEEE Fellow
- ICDM, VLDB, **KDD**

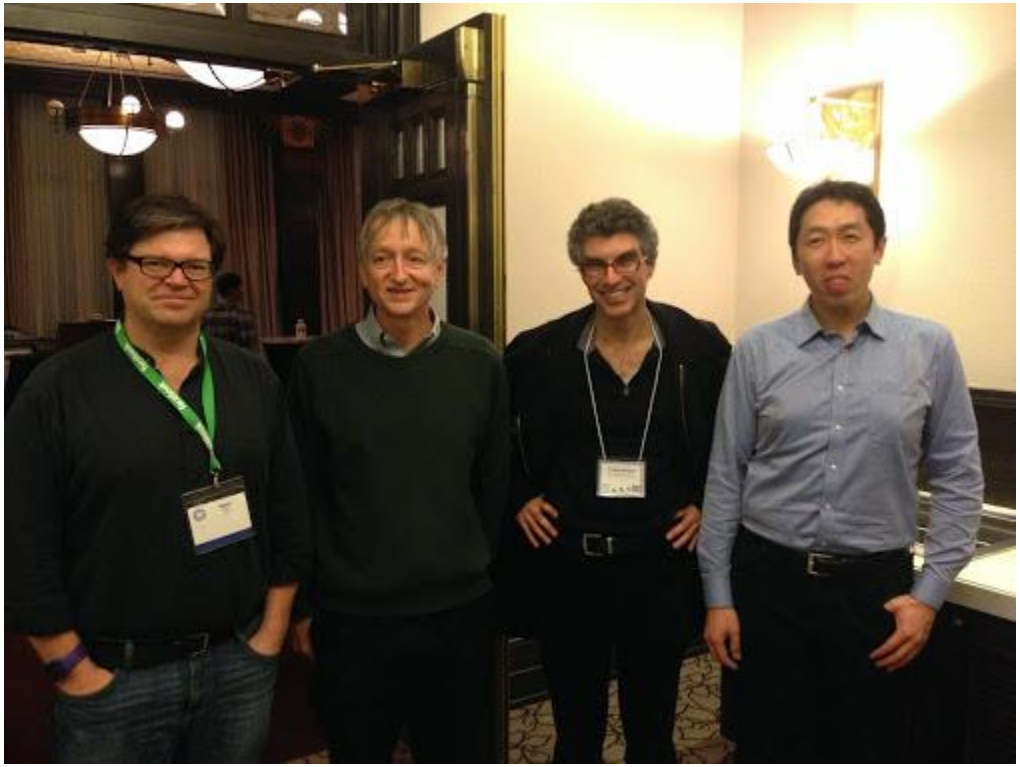


Cited by VIEW ALL

	All	Since 2012
Citations	142305	72095
h-index	154	112
i10-index	660	564



Pop Stars in ML



Geoffrey Hinton

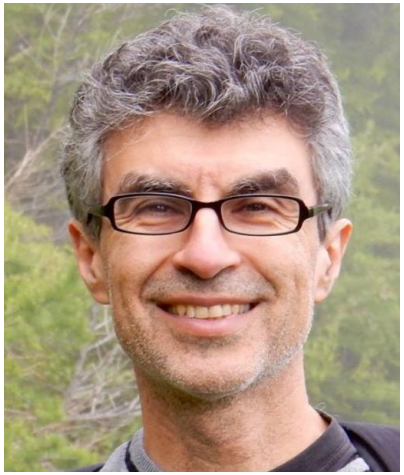
- University of Toronto
- Backpropagation (BP)
- Father of neural networks
- Father of deep learning
- Google (DNNResearch)
  - 2 stuff: Alex Krizhevsky, Ilya Sutskever
  - 1 toolkit: cat face recognition
- Neural Networks Renaissance
- His aunt Joan Hinton

Yann LeCun

- New York University
- Postdoctoral in Hinton's lab
- First director of Facebook AI Research
- Convolutional Neural Network (CNN)
- Generative Adversarial Network (GAN)
- Neural Networks Renaissance

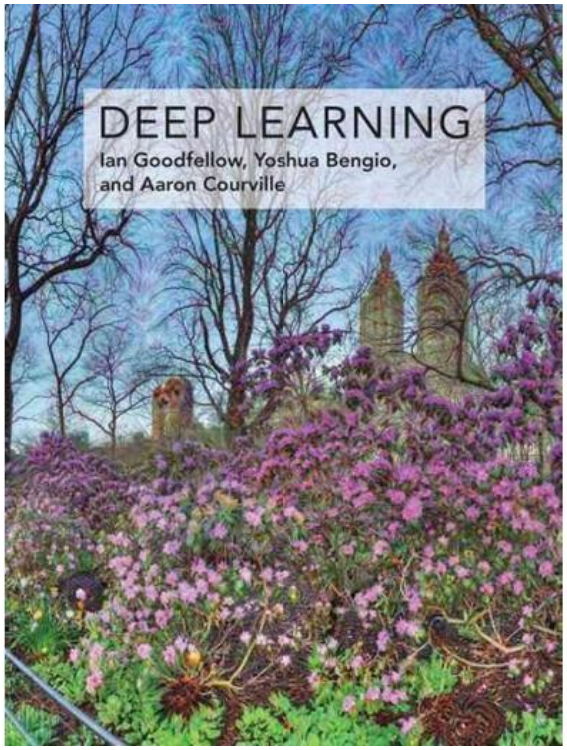
Yoshua Bengio

- Université de Montréal
- Postdoctoral in Michael Jordan's lab
- Colleague of Yann LeCun in AT&T Bell Labs
- Neural Networks Renaissance
- Focus on academia



Ian Goodfellow

- Ph.D. from Université de Montréal
- Under the supvervision of Yoshua Bengio
- Google Brain
- Generative Adversarial Network (GAN)
- OpenAI



Andrew Ng (吴恩达)

- Born in UK
- Parents were Hong Kongers
- Stanford University
  - Carnegie Mellon University (CMU)
  - Massachusetts Institute of Technology (MIT)
  - University of California, Berkeley (UC Berkeley)
- Founder of Google Brain Deep Learning Project
- Co-founder and chairman of Coursera
- Chief Scientist of Baidu
- Wife: Carol Reiley
  - cofounder and President of drive.ai

Data Mining Tasks

Data mining tasks

Supervised Learning
Unsupervised Learning
Semi-supervised Learning
Other Learnings

Supervised Learning

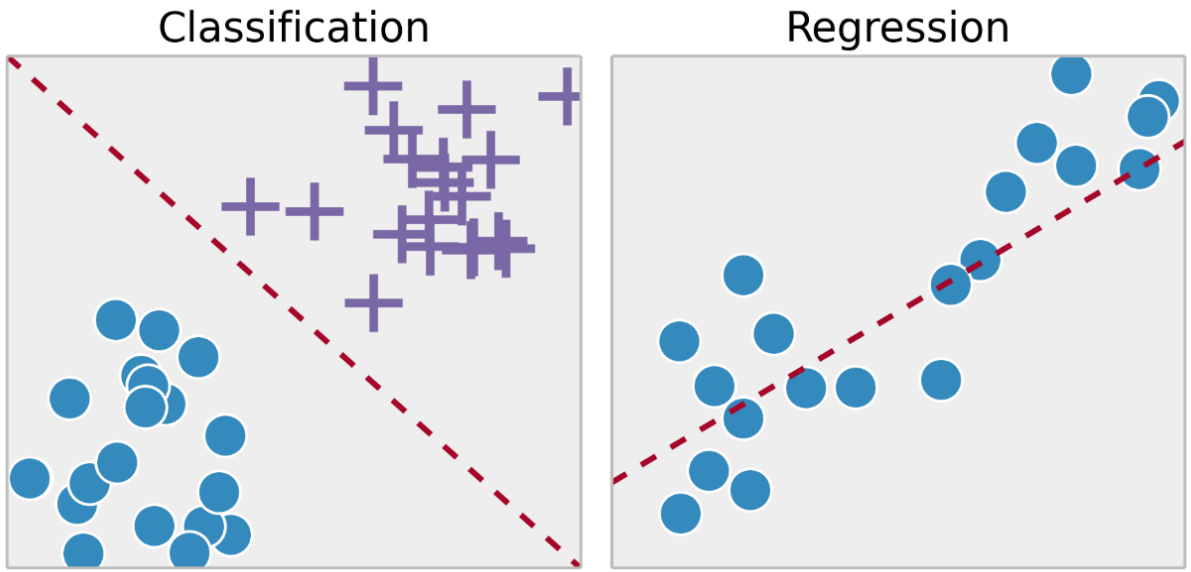
Supervised learning is where you have input variables (**X**) and an output variable (**Y**) and you use an algorithm to learn the mapping function from the input to the output:

$$\mathbf{Y} = f(\mathbf{X})$$

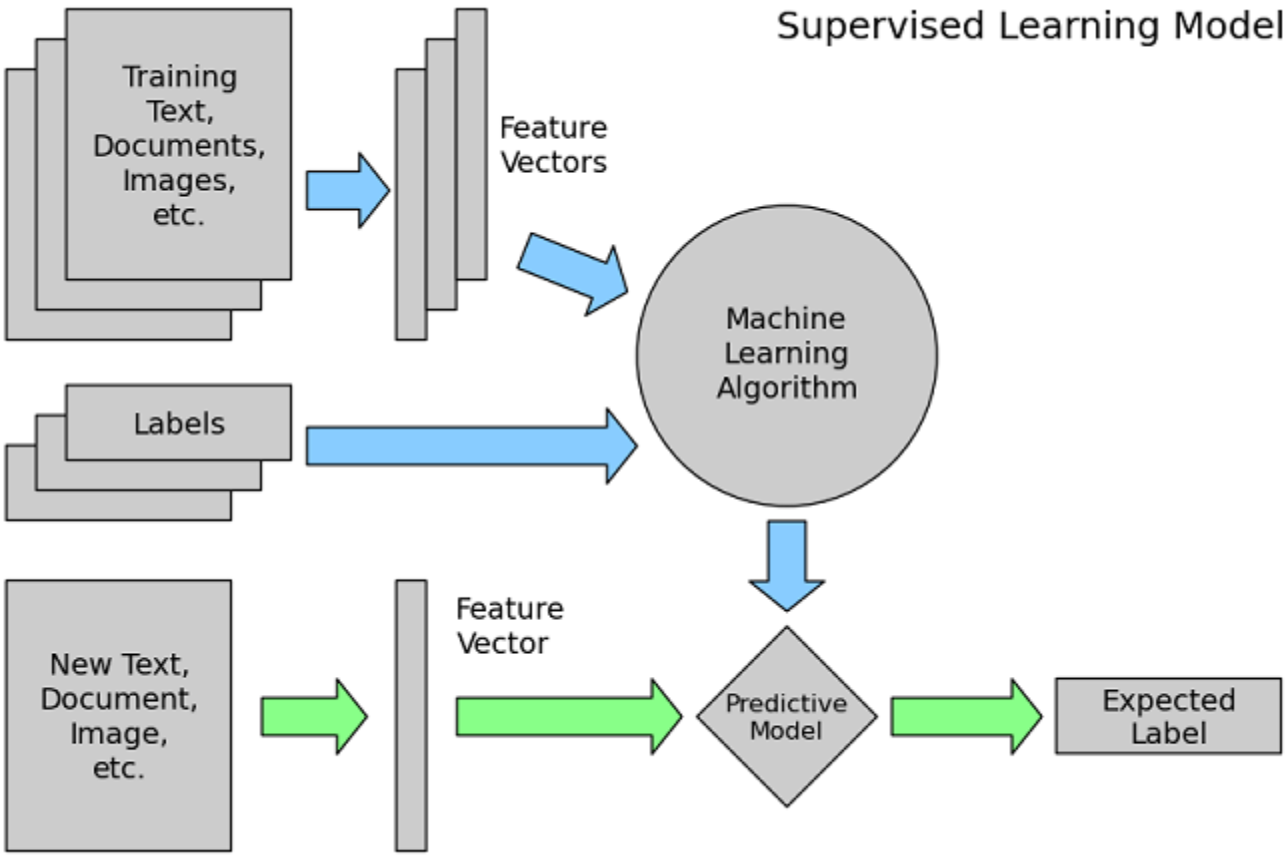
The goal is to approximate the mapping function so well that when you have new input data (**X**) that you can predict the output variables (**Y**) for that data.

Supervised Learning

- **Classification** is when our labels **Y** can only take a finite set of values (categories).
- **Regression** is when our labels **Y** can take any real (continuous) value.



Supervised Learning

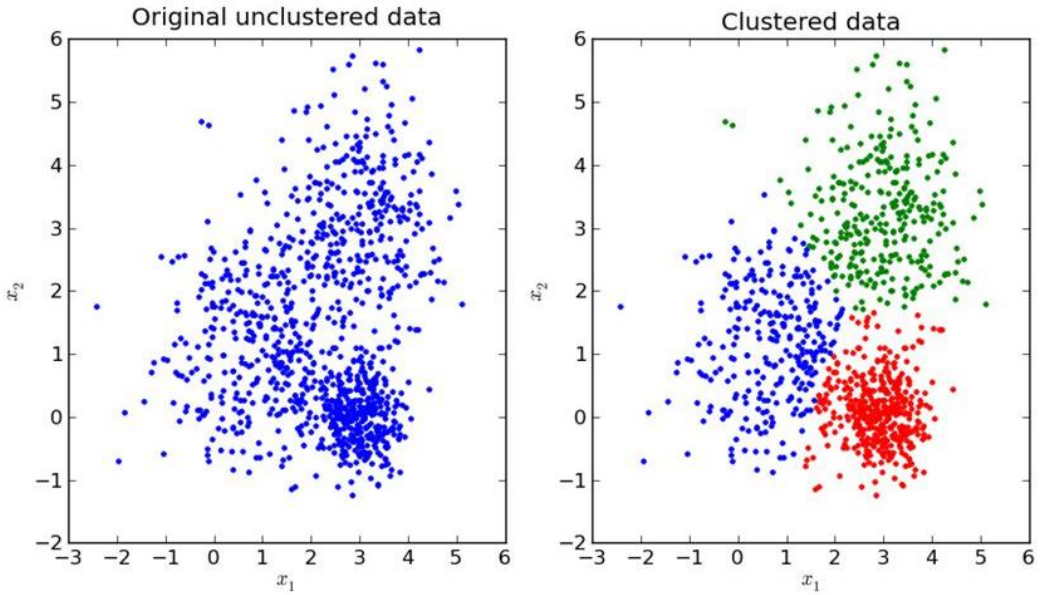


Data mining tasks

Supervised Learning
Unsupervised Learning
Semi-supervised Learning
Other Learnings

Unsupervised Learning

Unsupervised learning is where you only have input data (**X**) and no corresponding output variables. The goal for unsupervised learning is to model the underlying structure or distribution in the data in order to learn more about the data.



Unsupervised Learning

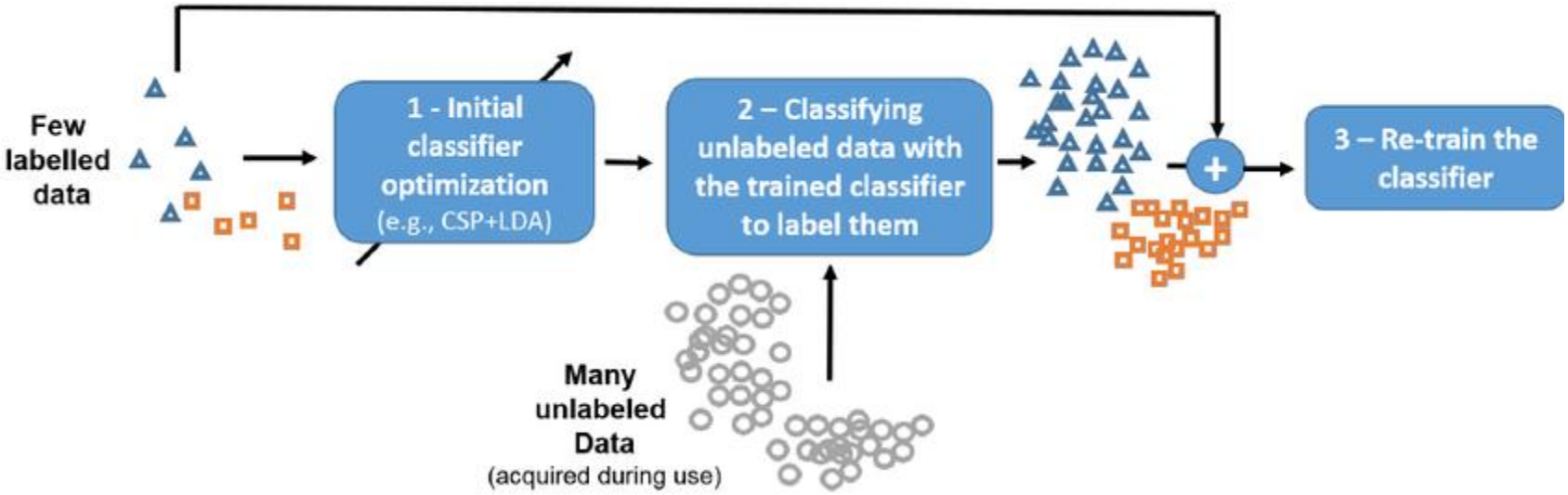
- **Clustering**: Grouping similar points together within clusters.
- **Outlier detection**: detecting and subsequently excluding outliers from a given set of data. An outlier may be defined as a piece of data or observation that deviates drastically from the given norm or average of the data set.
- **Density estimation**: Estimating a probability density that can explain the distribution of the data points.
- **Dimension reduction**: Getting a simple representation of high-dimensional data points by projecting them onto a lower-dimensional space. This technique is notably used for data visualization.
- **Manifold learning** (or nonlinear dimension reduction): Finding a low-dimensional manifold containing the data points.

Data mining tasks

Supervised Learning
Unsupervised Learning
Semi-supervised Learning
Other Learnings

Semi-supervised Learning

- Semi-supervised learning is a class of supervised learning tasks and techniques that also make use of unlabeled data for training – typically a small amount of labeled data with a large amount of unlabeled data.

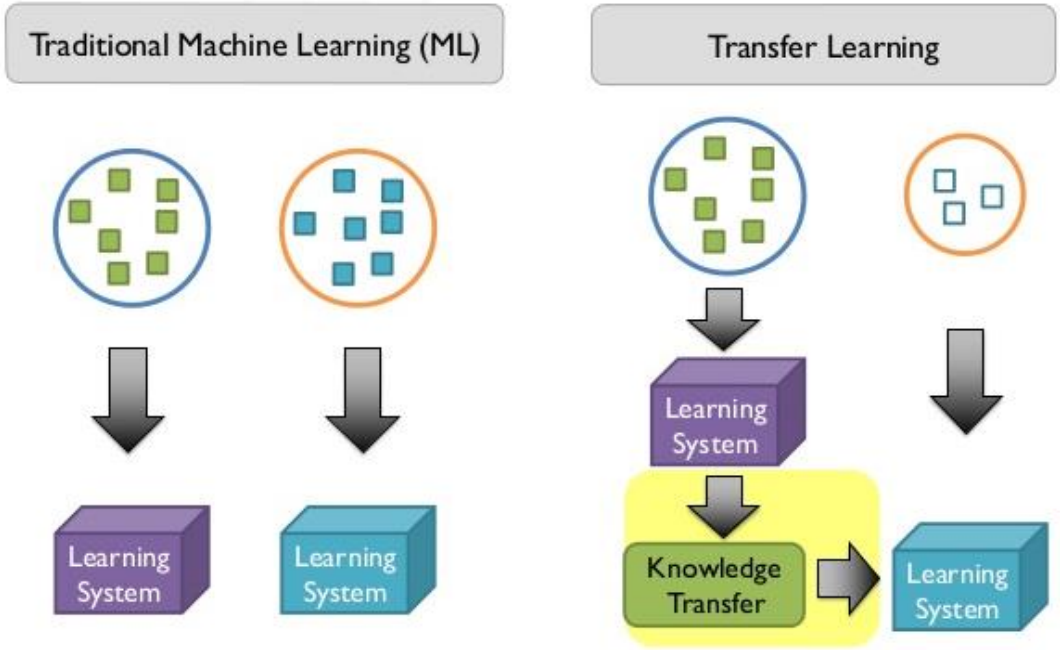


Data mining tasks

Supervised Learning
Unsupervised Learning
Semi-supervised Learning
Other Learnings

Transfer Learning

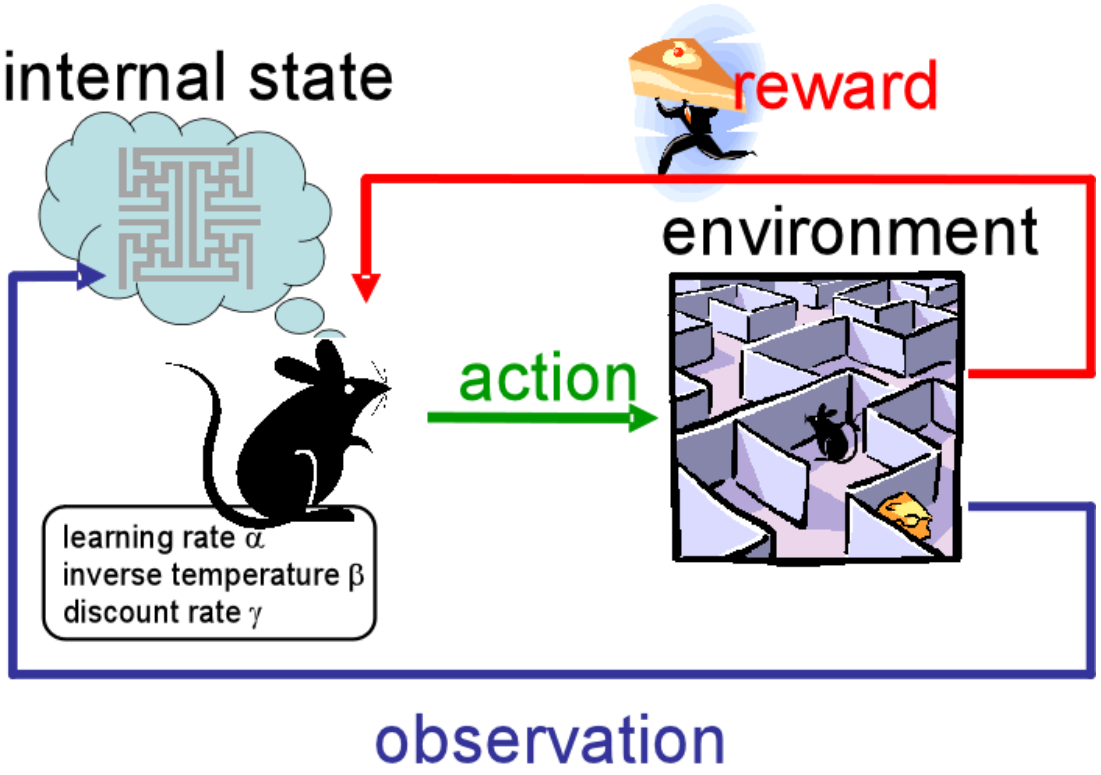
- Transfer learning or inductive transfer is a research problem in machine learning that focuses on storing knowledge gained while solving one problem and applying it to a different but related problem.



Reinforcement Learning

- Reinforcement Learning allows machines and software agents to automatically determine the ideal behavior within a specific context, in order to maximize its performance.
- Simple reward feedback is required for the agent to learn its behavior; this is known as the reinforcement signal.
- This behavior can be learnt once and for all, or keep on adapting as time goes by. If the problem is modelled with care, some Reinforcement Learning algorithms can converge to the global optimum; this is the ideal behavior that maximizes the reward.

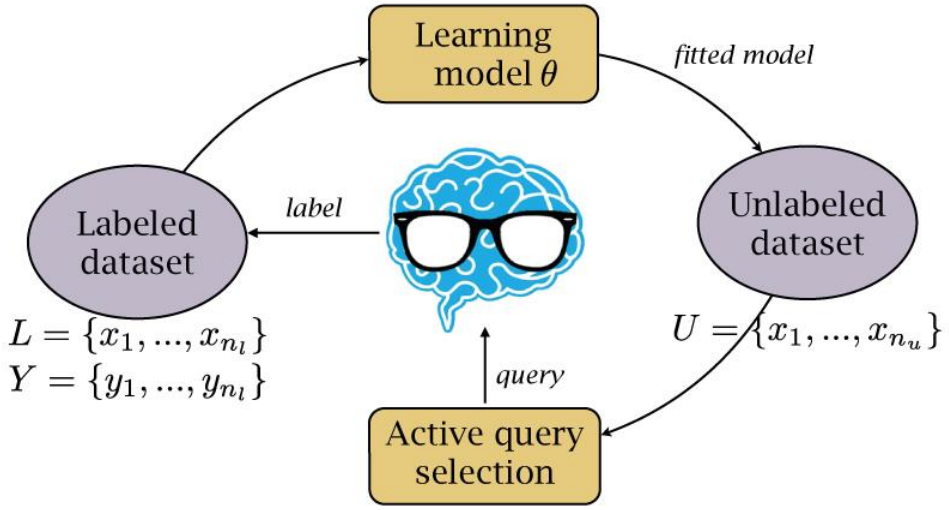
Reinforcement Learning



- Trial-and-error learning
- Latent learning

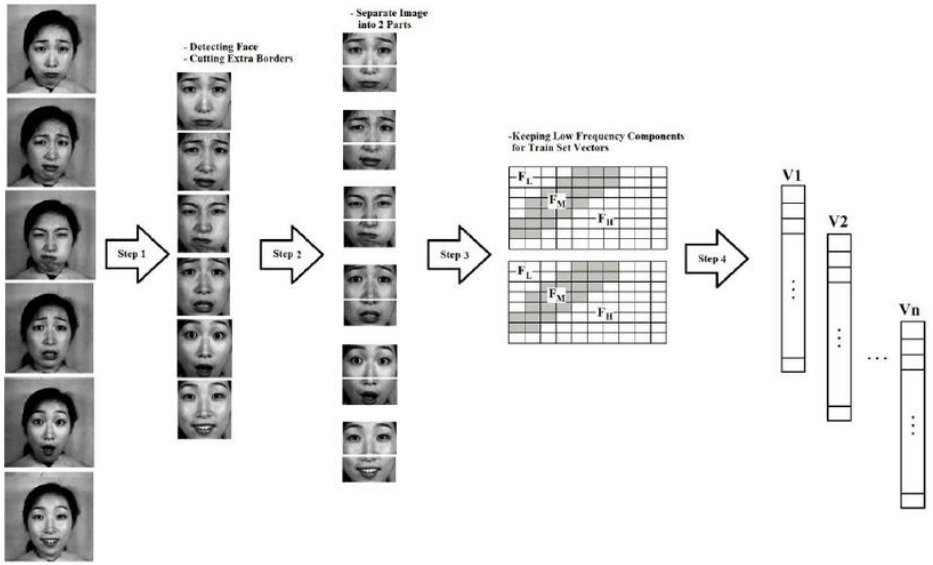
Active Learning

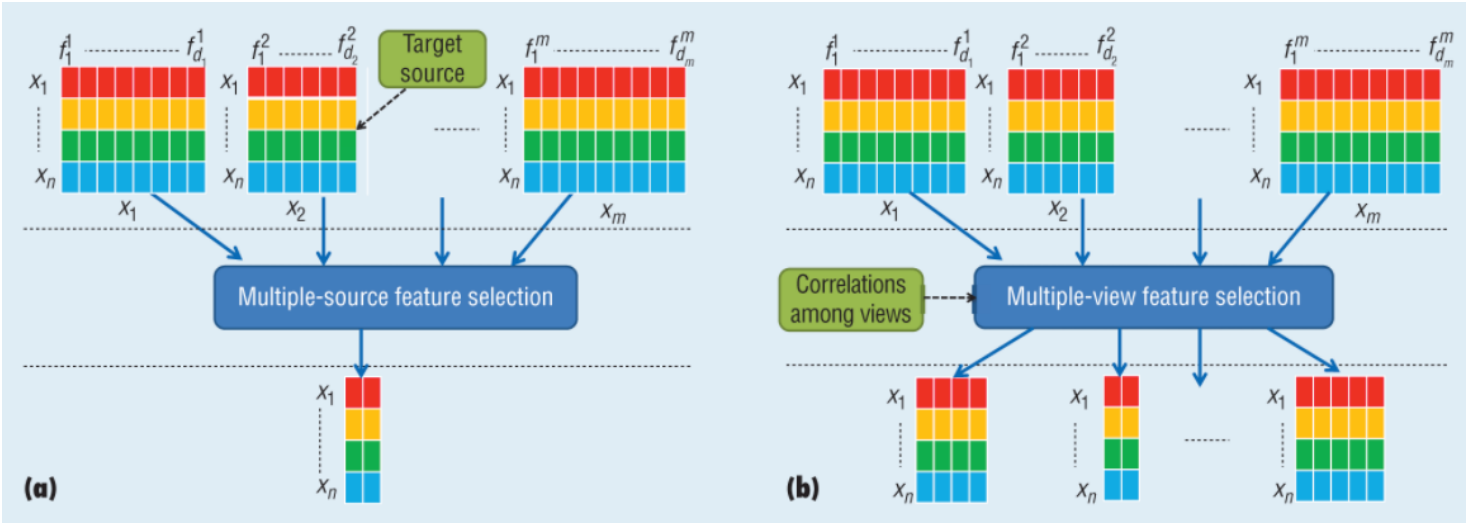
- Active learning is a special case of semi-supervised machine learning in which a learning algorithm is able to interactively query the user (or some other information source) to obtain the desired outputs at new data points. In statistics literature it is sometimes also called optimal experimental design.



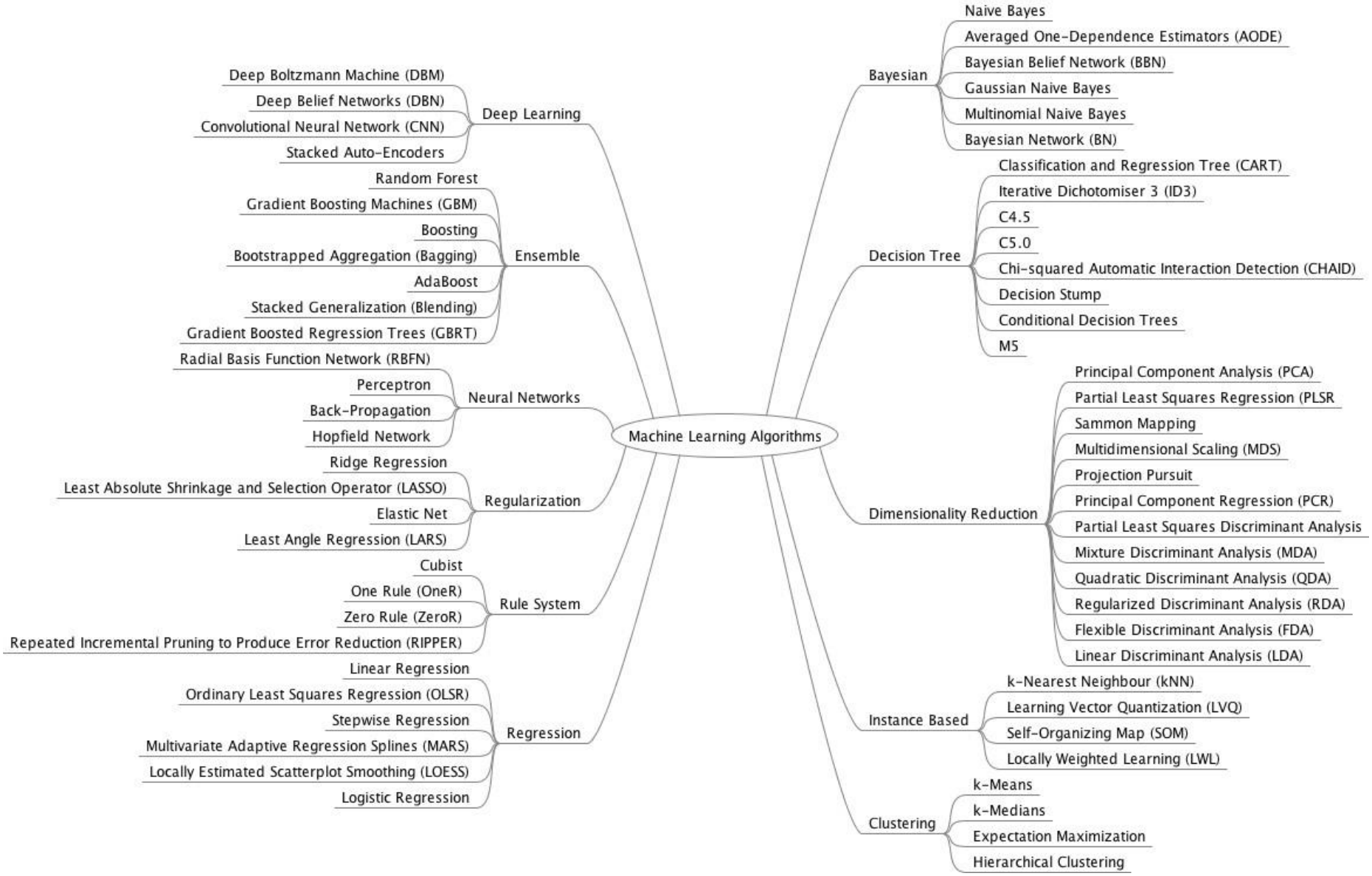
Other things you might wanna know…

Feature Extraction & Selection

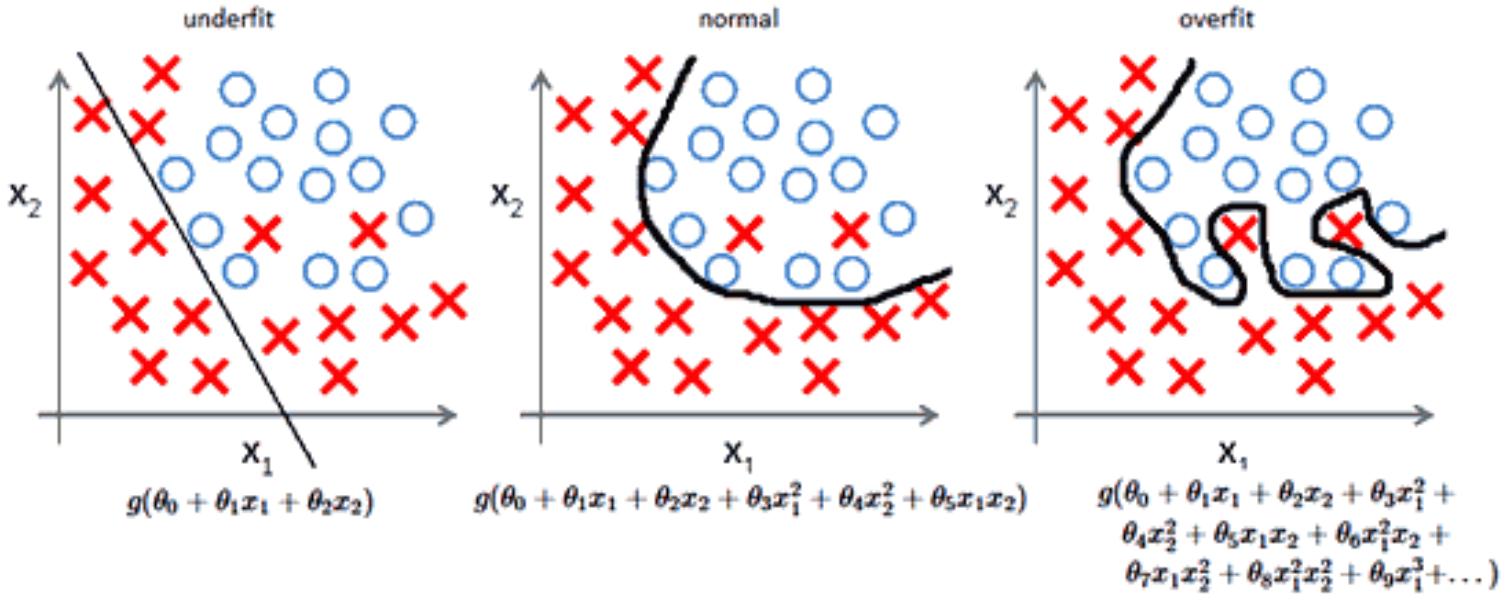




Model Selection

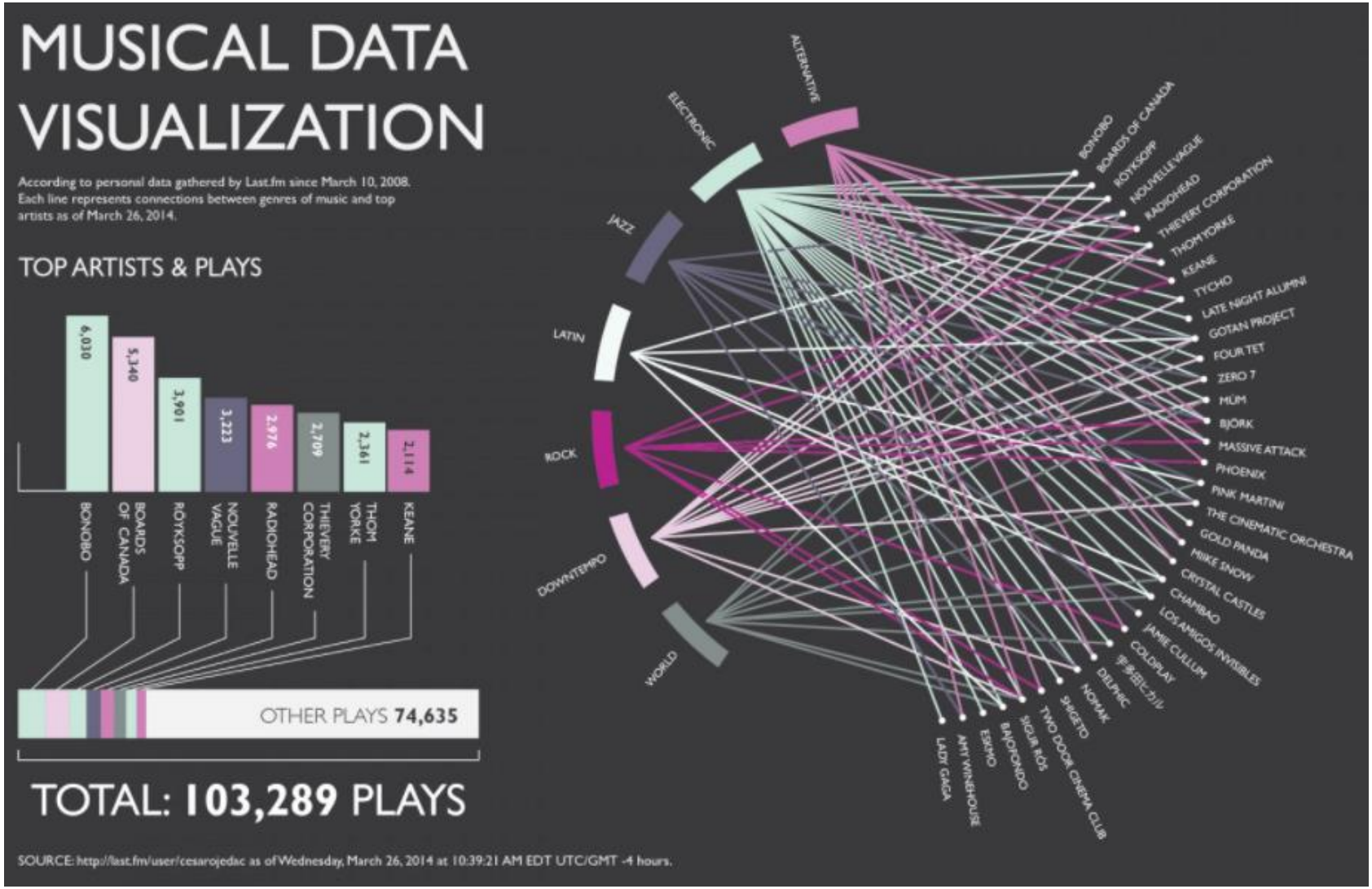


Overfitting & Underfitting

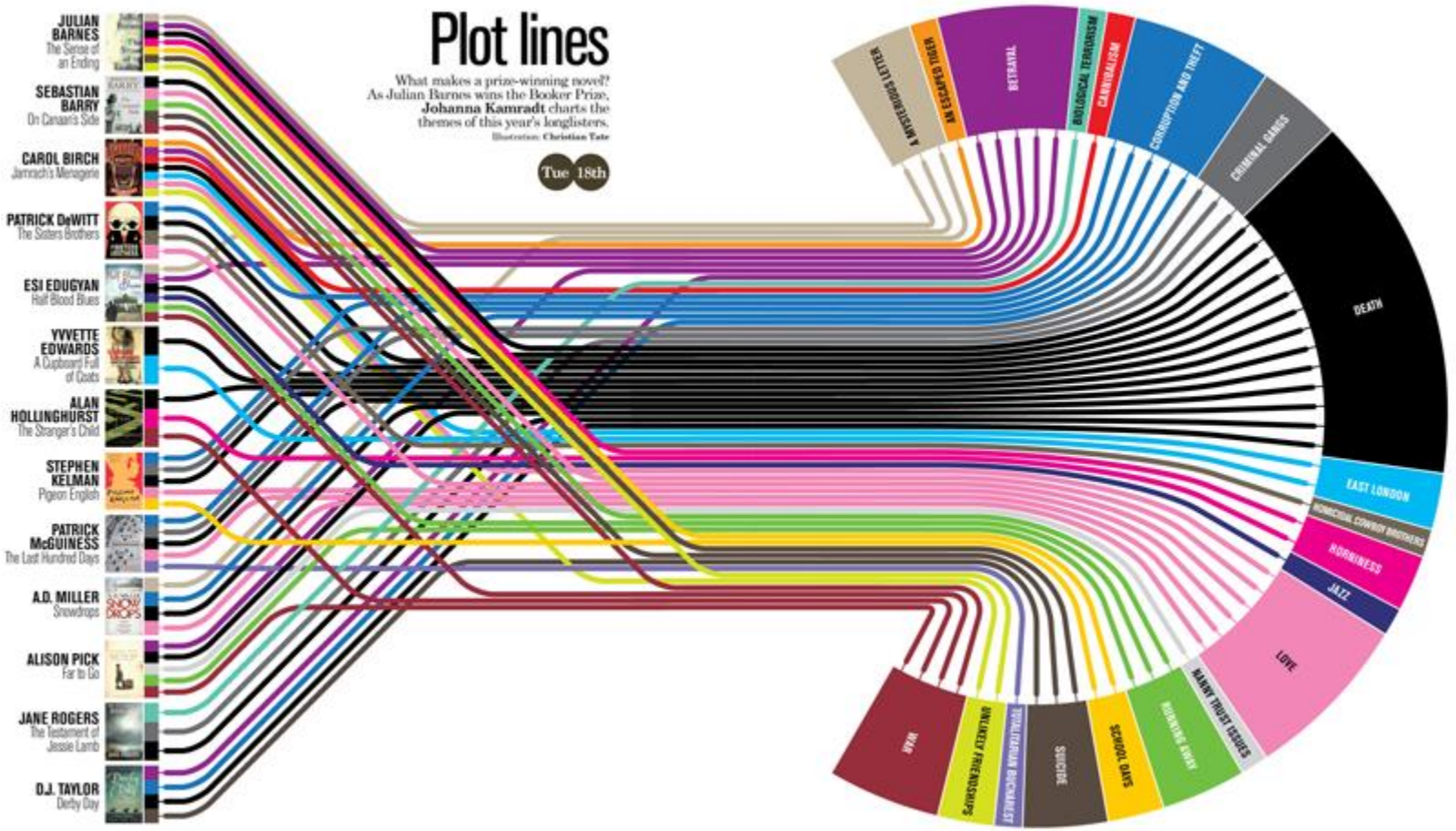


Data Visualization

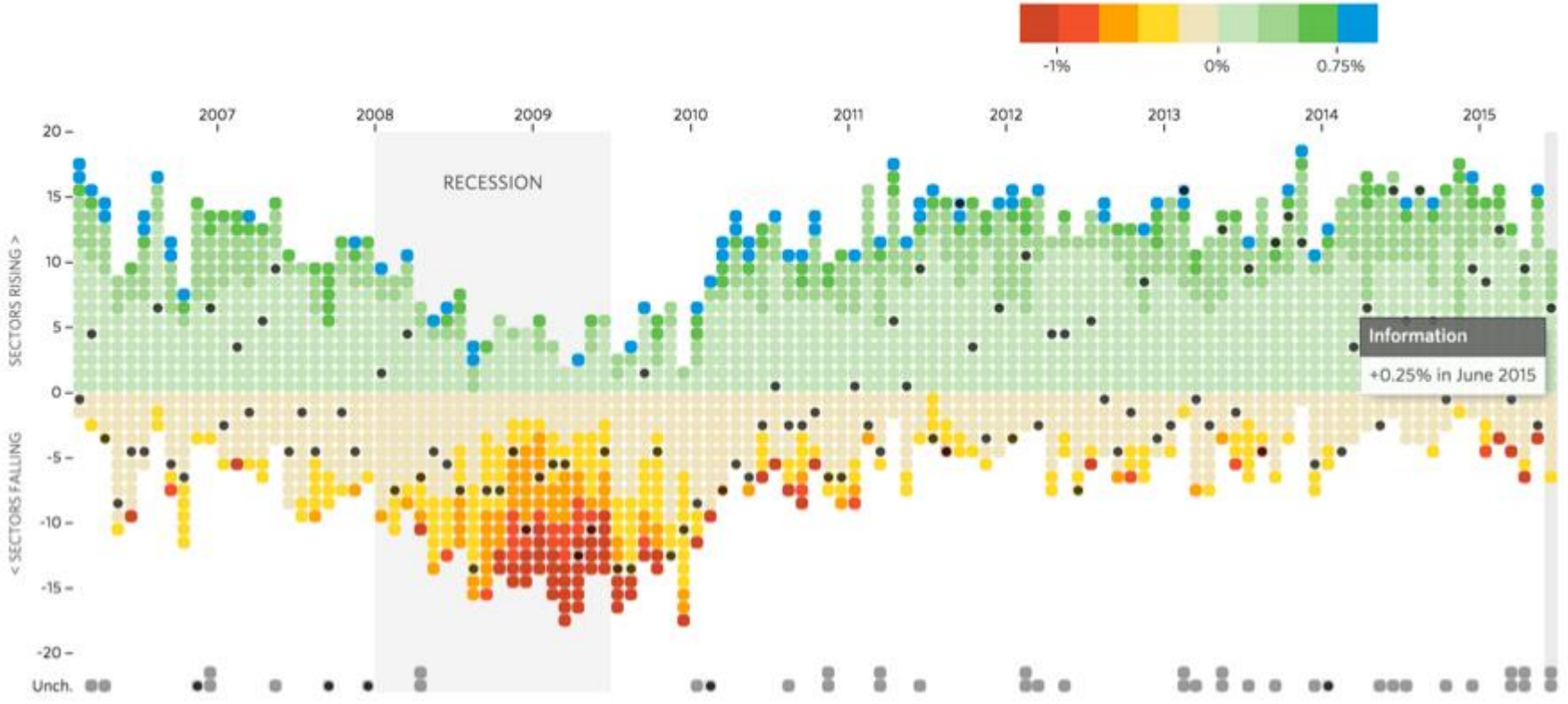
Data Visualization



Data Visualization



Data Visualization



End of Chapter 11