Big Data Analytics & Applications

Bin Li

School of Computer Science

Fudan University

About the Course

- **Introduce real-world data analytics applications**
- **Introduce typical machine learning solutions**

- **Syllabus of the course**
  - Part 1: Introduction to Machine Learning
  - Part 2: Text Data Analytics
  - Part 3: Network Data Analytics
  - Part 4: Image Data Analytics
- **Assessment methods**
  - Interactions and presentations (30%)
  - Course project with final report (70%)

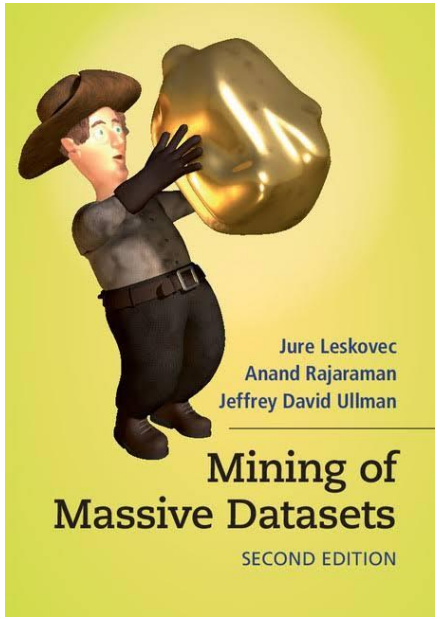
About the Students

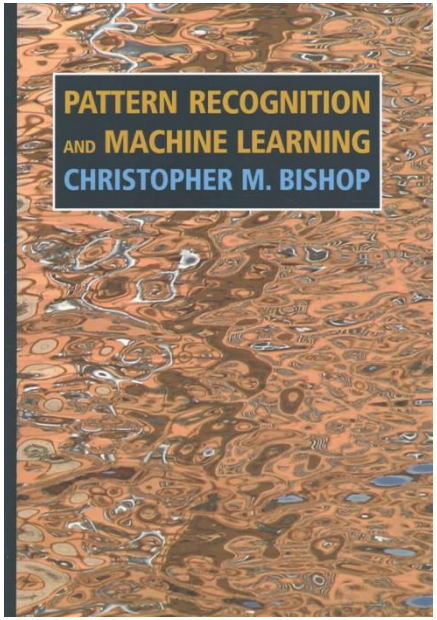
- **Prerequisites of this course**
  - Calculus
  - Linear Algebra
  - Probability Theory and Statistics
  - Programming
  - English Speaking & Writing

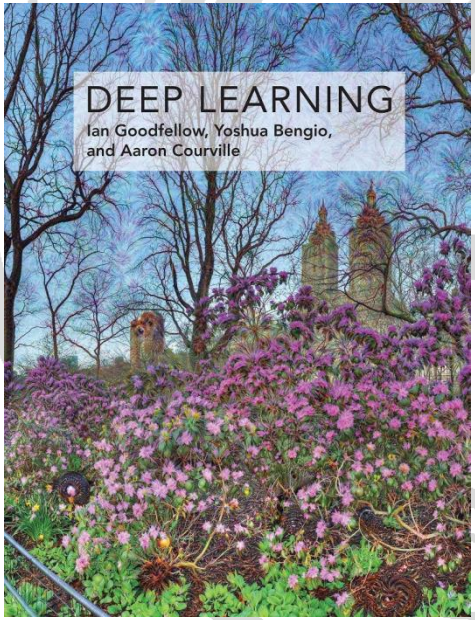
- **Course project requirements**
  - Select one data analytics problem introduced in the course (Collect real-world new dataset by yourself – extra point)
  - Use toolbox (e.g., scikit, TensorFlow) to solve the problem (Develop new method by yourself – extra point)
  - Write project report in English

Reference Books

- **No textbook for this course**
- Reference books and some papers mentioned in the course







Data Explosion

- Can you tell some Chinese counterparts of these Apps?

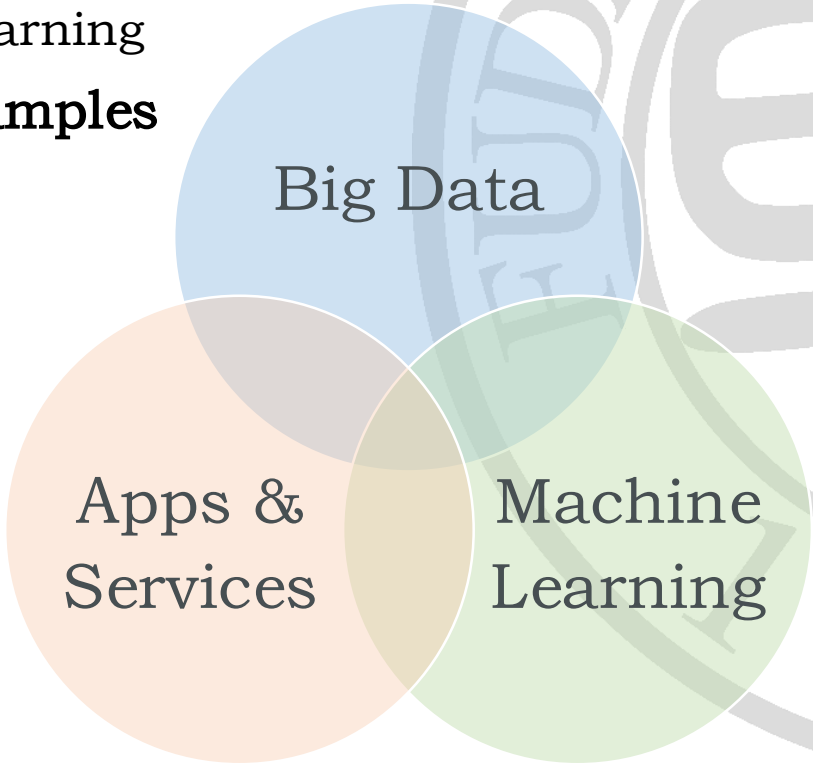
Big Data (why called "Big")

- Big data[1] can be described by the "3V" characteristics
  - Volume: The quantity of generated and stored data
  - Variety: The type and nature of the data
  - Velocity: Big data is often available in real-time

[1] https://en.wikipedia.org/wiki/Big_data

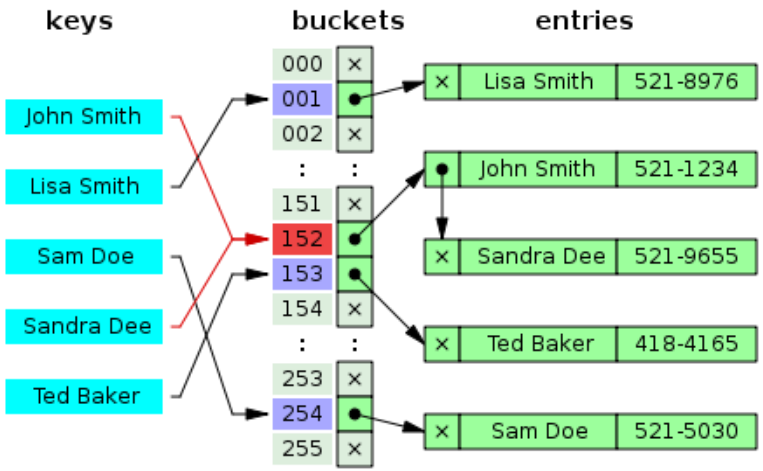
Big Data Analytics (BDA)

- **Three elements for big data analytics**
  - Source: Big Data
  - Demand: Applications & Services
  - Tools: Machine Learning
- **Try to find some examples**



BDA Example: LSH

- Locality-sensitive hashing (LSH) reduces the dimensionality of high-dimensional data such that similar items map to the same buckets with high probability.
  - Similarity search
  - Duplicate detection
  - Dimension reduction
  - Preprocessing for machine learning tasks

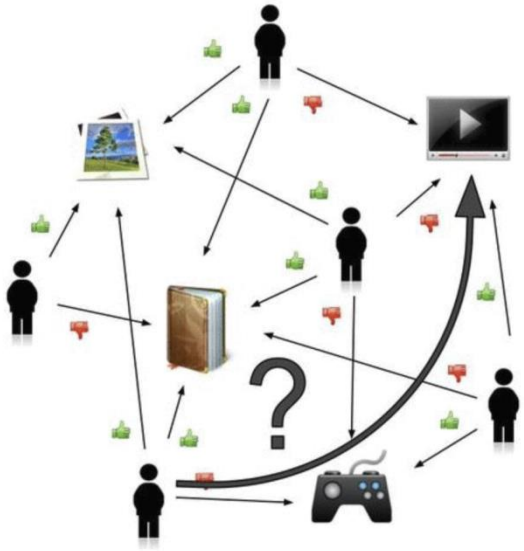


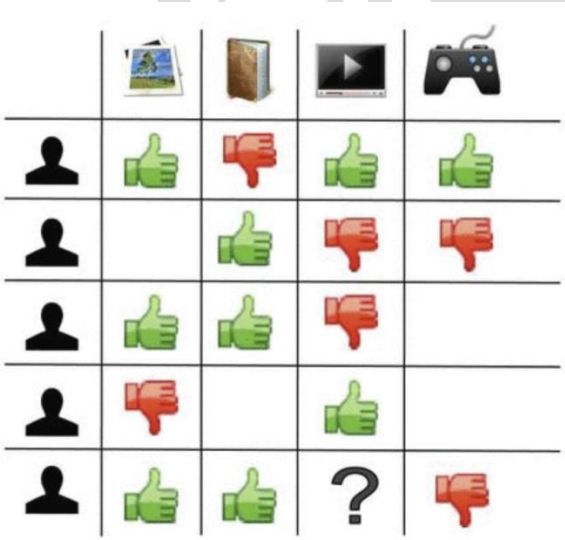
BDA Example: Topic Modeling

- Discover latent topics from massive documents: Each news article has a distribution over K latent topics while each of the K latent topics has a distribution over the words
  - Document (e.g., news) automatic categorization
  - Any bag-of-words objects grouping tasks

BDA Example: Collaborative Filtering

- Collaborative filtering makes predictions (filtering) about the interests of a user by collecting preferences or taste information from many users (collaborating)
  - Recommender systems
  - User behavior prediction
  - Marketing



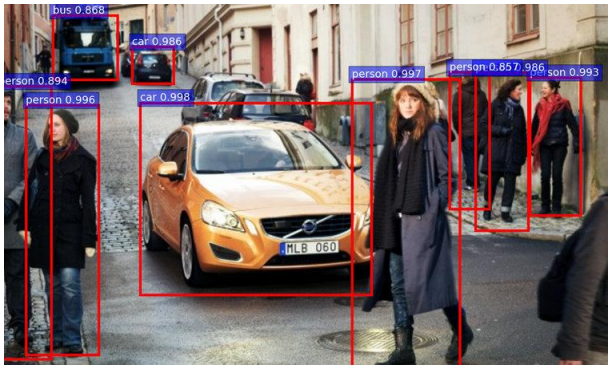


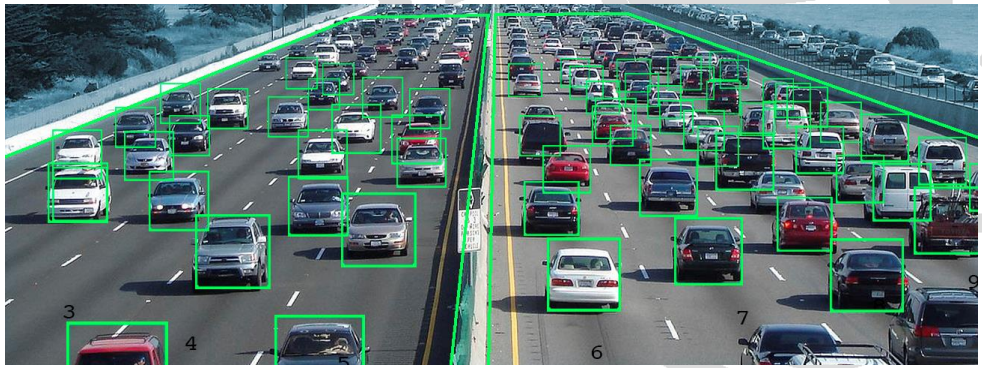
BDA Example: Social Network Analysis

- Social network analysis (SNA) is the process of investigating social structures where nodes denote users and links denote relationships.
  - Community detection
  - Link prediction
  - Social capital detection
  - etc.

BDA Example: Object Detection

- Detect certain objects from massive images or video streams using bounding boxes
  - Security surveillance
  - Intelligent transport
  - Image annotation and retrieval
  - Internet censorship
  - etc.





BDA Example: OCR

- Optical character recognition (OCR) is a technology that extracts and recognizes texts from images.
  - Document scanning
  - Text recognition in the wild
  - Commercial analysis
  - Internet censorship
  - etc.


```
494.91,36.36,494.91,81.45,596.0,81.45,595.27,32.73,三星
614.91,34.91,614.91,77.82,783.64,77.82,783.64,34.91,N7100
524.73,93.82,524.73,146.18,784.36,146.18,784.36,93.82,钢化玻璃膜
164.0,143.27,164.0,157.09,251.27,157.09,251.27,143.27,SAMSUNG
316.73,174.55,316.73,189.09,353.82,189.09,353.82,174.55,17:39
117.45,236.36,117.45,284.36,298.55,284.36,298.55,236.36,17:39
142.91,292.36,142.91,309.82,268.73,309.82,268.73,292.36,9月12日星期三
262.91,354.91,262.91,367.27,345.82,367.27,345.82,354.91,小到中雨转阵雨
321.82,338.91,321.82,350.55,345.82,350.55,345.82,338.91,北京
70.57,378.86,70.57,384.57,87.14,384.57,87.14,378.86,###
88.86,376.57,88.86,385.71,118.57,385.71,118.57,376.57,新浪天气
206.0,371.43,206.0,381.71,324.29,381.71,324.29,371.43,已更新2012/09/1116:59
76.86,545.71,76.86,556.0,106.0,556.0,104.86,546.29,沃·3G
150.57,545.71,150.57,557.71,183.71,557.71,183.71,545.71,沃商店
226.57,546.86,226.57,557.71,264.29,557.71,264.29,546.86,116114
294.57,545.14,294.57,557.71,350.0,557.71,350.0,545.14,手机营业厅
67.71,682.86,67.71,693.14,100.29,693.14,100.29,682.86,联系人
135.71,682.29,134.57,694.29,156.29,694.29,156.29,681.14,手机
195.14,682.86,195.14,693.71,218.57,693.71,218.57,682.86,信息
251.14,682.86,251.14,692.57,282.57,692.57,282.57,682.86,互联网
307.14,681.71,307.14,693.14,349.43,693.14,349.43,681.71,应用程序
232.29,513.14,232.29,522.86,261.43,522.86,261.43,513.14,116114
150.0,514.29,150.0,522.86,172.86,522.86,172.86,514.29,WO
153.43,523.43,153.43,532.0,180.86,532.0,180.86,523.43,###
76.29,522.29,76.29,529.14,105.43,529.14,105.43,522.29,###
76.29,504.0,76.29,514.86,107.14,514.86,107.14,504.0,WO
69.43,341.14,69.43,366.86,125.43,366.86,125.43,341.14,29° C
141.43,340.0,141.43,366.86,195.71,366.86,195.71,340.0,17° C
547.71,701.14,547.71,746.86,784.29,746.86,784.29,701.14,送贴膜工具
568.29,554.86,588.29,608.0,792.29,534.29,775.14,481.71,防爆防刮
551.14,449.14,577.43,510.86,795.71,434.29,785.43,372.57,智能贴合
542.0,344.0,564.29,396.57,788.86,323.43,771.71,265.71,防指纹油
659.14,543.43,659.14,544.0,659.71,544.0,659.71,543.43,###
```


➢ Prediction：复优者联盟

➢ Prediction：专力高档

➢ Prediction：OEF

➢ Prediction：香酥芯麻味

➢ Prediction：令加工糕点

➢ Prediction：七

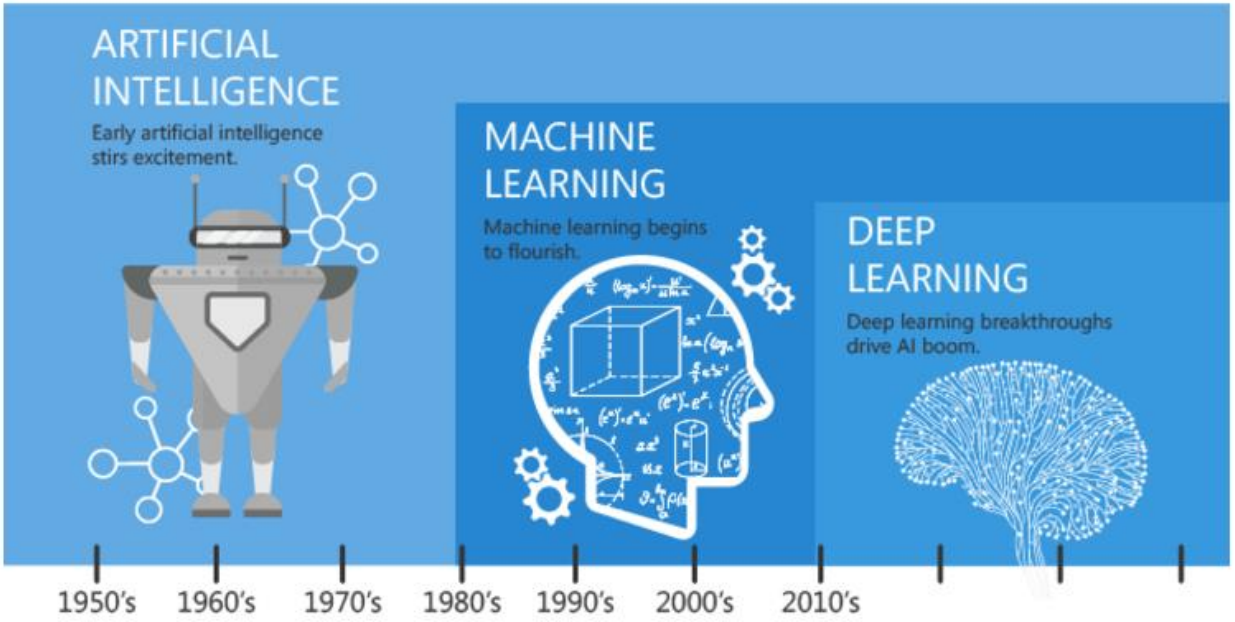
Three Elements: Examples

- **Three elements for big data analytics**
  - Source: Big Data
  - Demand: Applications & Services
  - Tools: Machine Learning

Demand	Source	Tools
Similarity Search	Webpages	Locality-Sensitive Hashing (LSH)
Topic Modeling	Documents	Latent Dirichlet Allocation (LDA)
Object Detection	Surveillance Video	Covolutional Neural Networks
OCR	Commercials	Covolutional Neural Networks
Recommendation	User Ratings	Matrix Factorization
Community Detection	Social networks	Infinite Relational Models

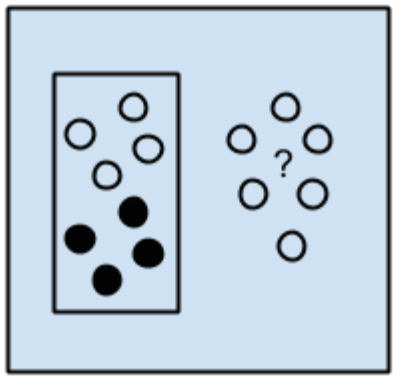
Machine Learning

- Machine learning often uses statistical techniques to give computers the ability to "learn" with data.
- Within the field of data analytics, machine learning is a method used to devise models and algorithms that lend themselves to prediction.



Machine Learning Problems

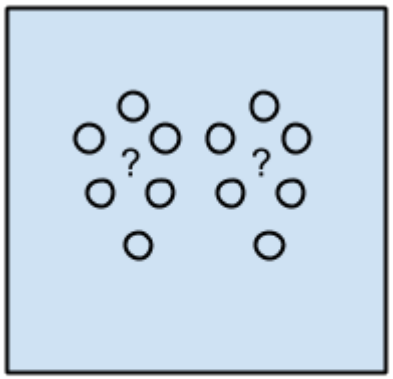
- Problem settings in machine learning



Supervised Learning Algorithms

Classification
Regression

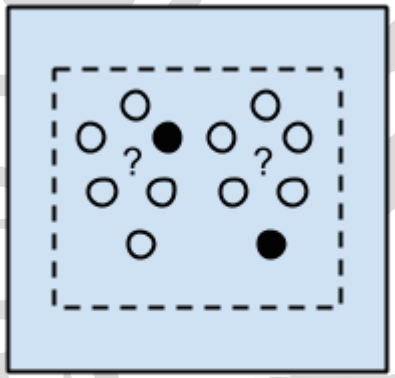
Examples: Logistic Regression, Neural Network, etc.



Unsupervised Learning Algorithms

Clustering
Dimensionality Reduction

Examples: K-Means, Principal Component Analysis, etc.



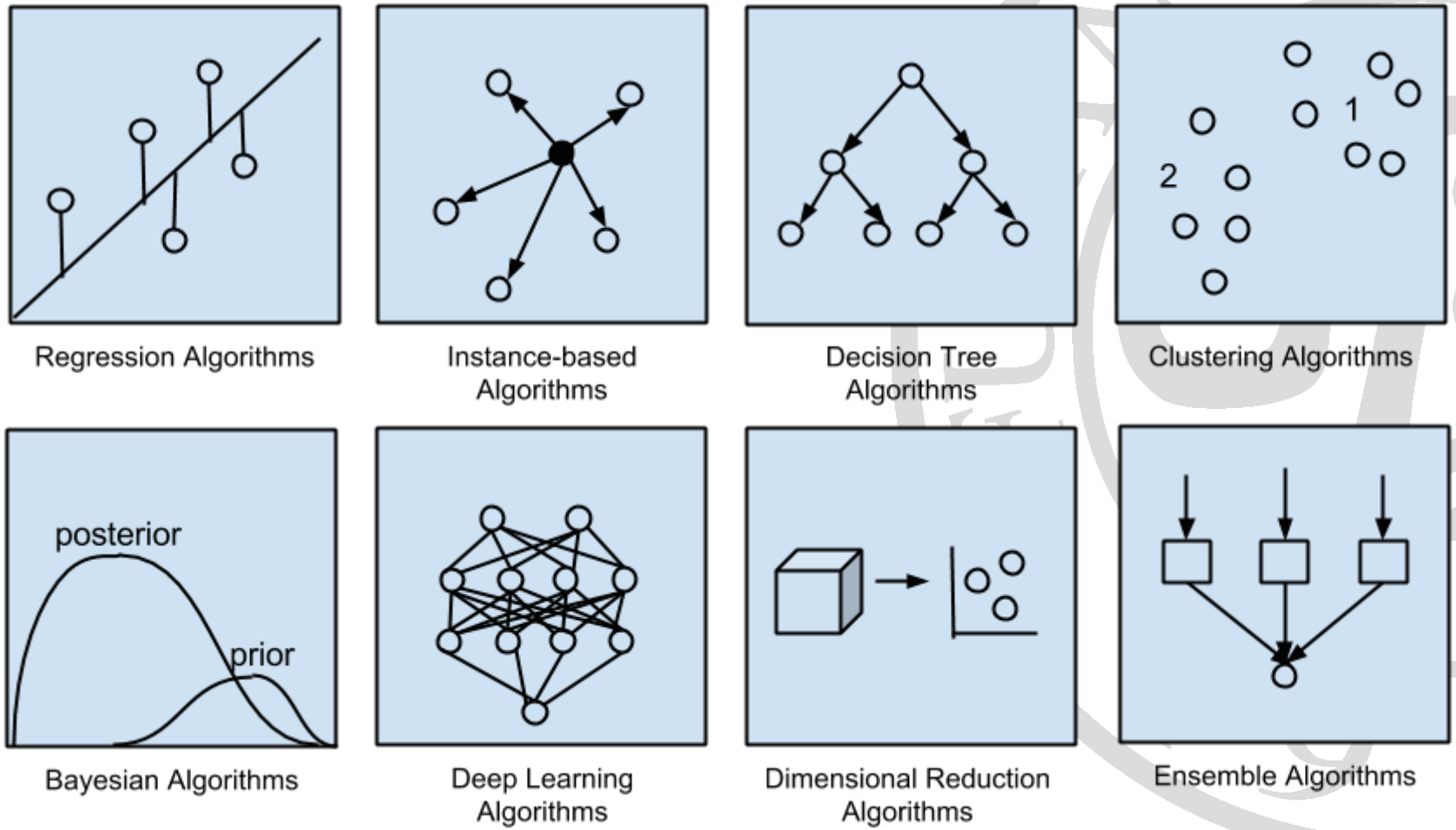
Semi-supervised Learning Algorithms

Classification
Regression

Examples: Transductive SVM, etc.

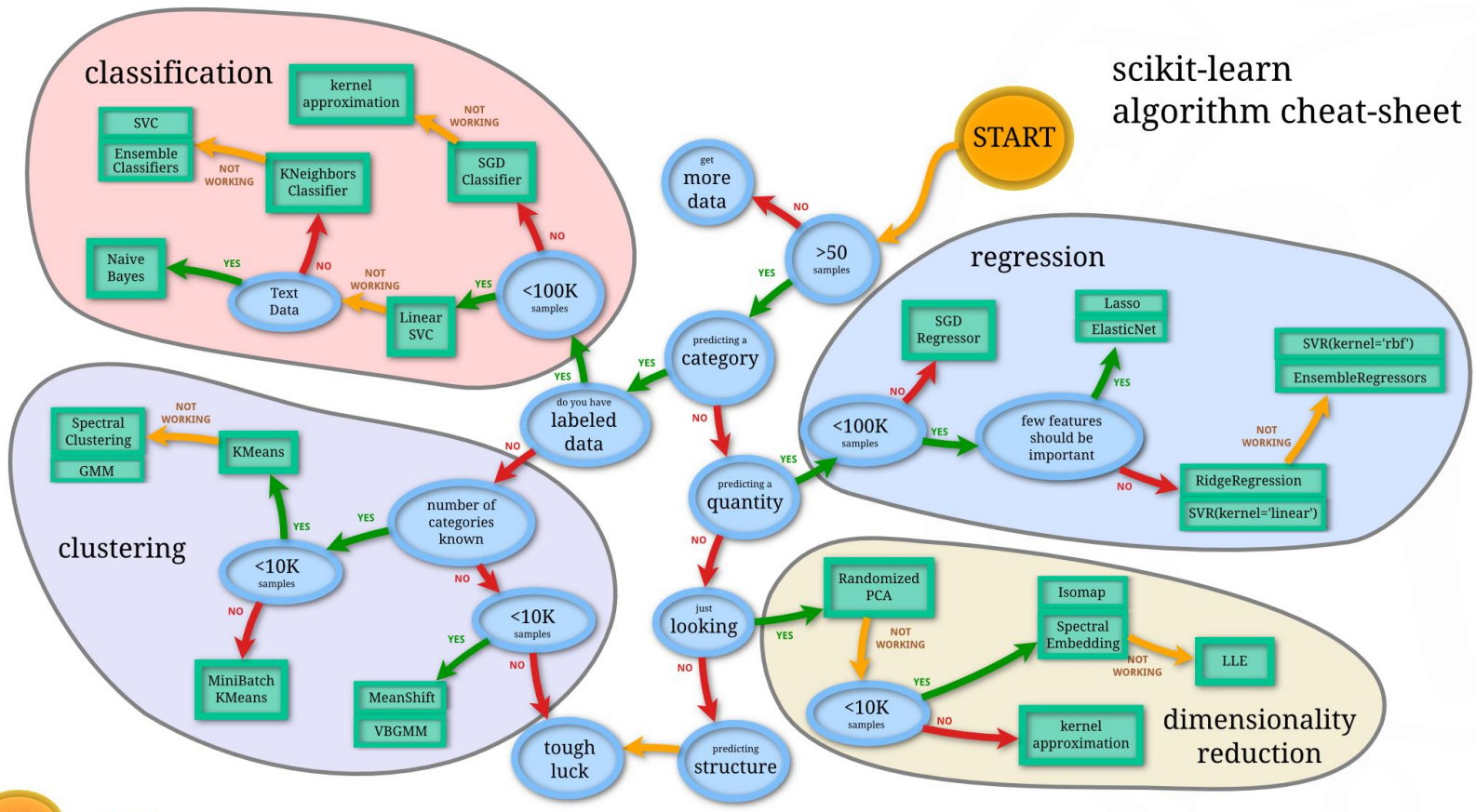
[1] https://machinelearningmastery.com/a-tour-of-machine-learning-algorithms/

Typical ML Algorithms



[1] https://machinelearningmastery.com/a-tour-of-machine-learning-algorithms/

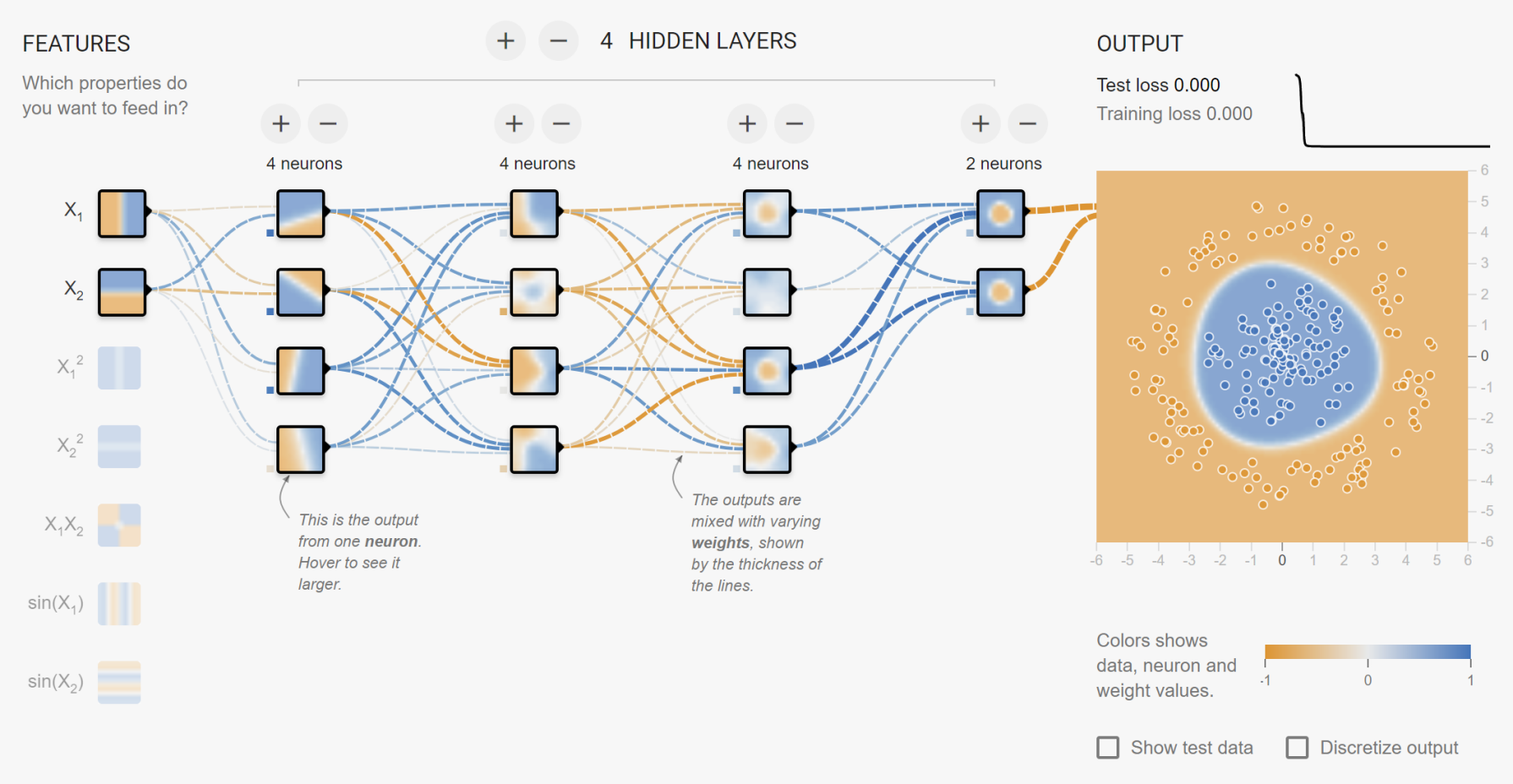
A Map of ML Tool Selection



Toolboxes for DBA

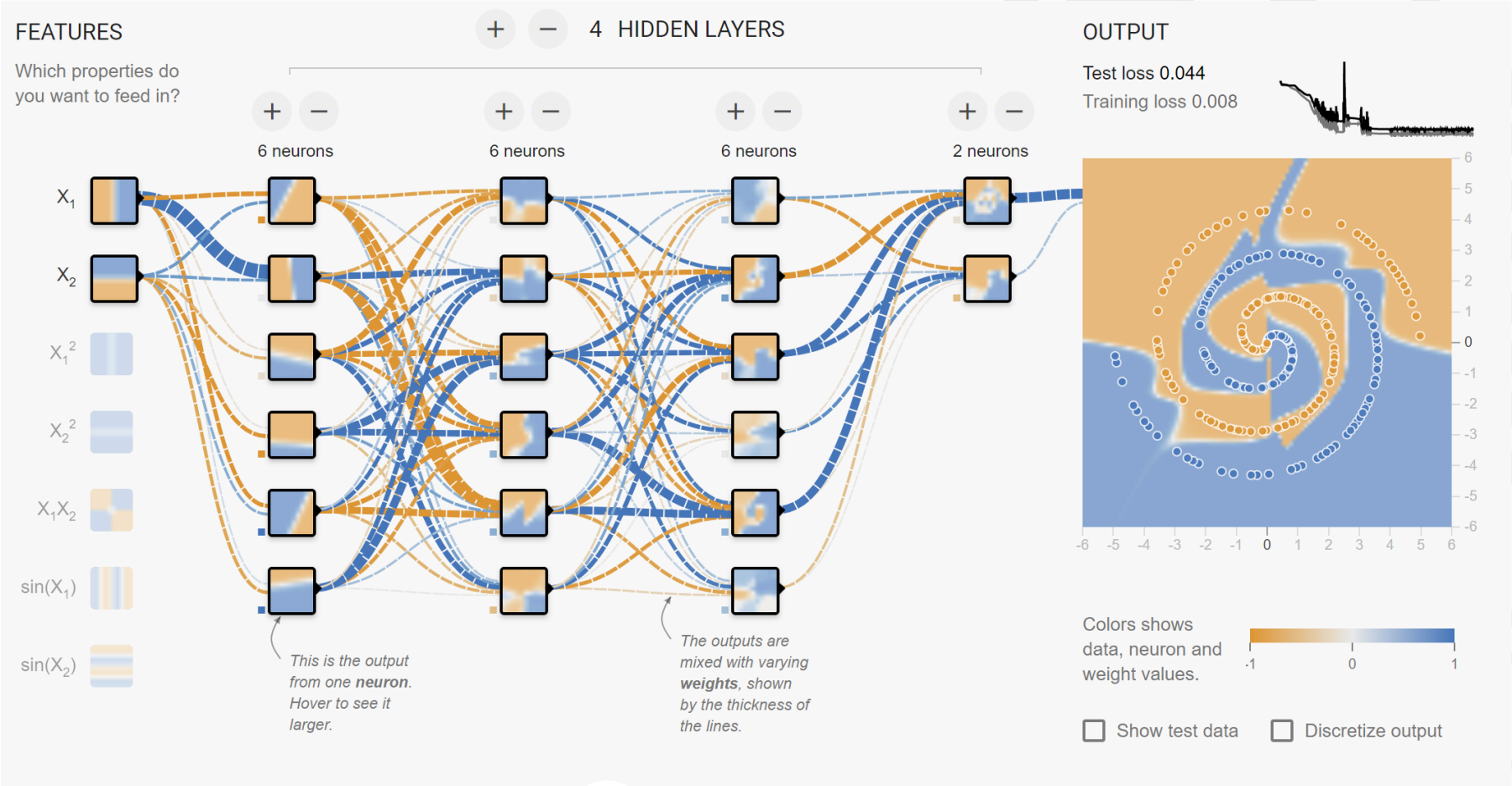
- For big data analytics with traditional machine learning tasks: http://scikit-learn.org
- For big data analytics with deep learning tasks: https://www.tensorflow.org or https://pytorch.org/

Tensorflow Playground



https://playground.tensorflow.org

Tensorflow Playground



https://playground.tensorflow.org

Thanks

Email: libin@fudan.edu.cn