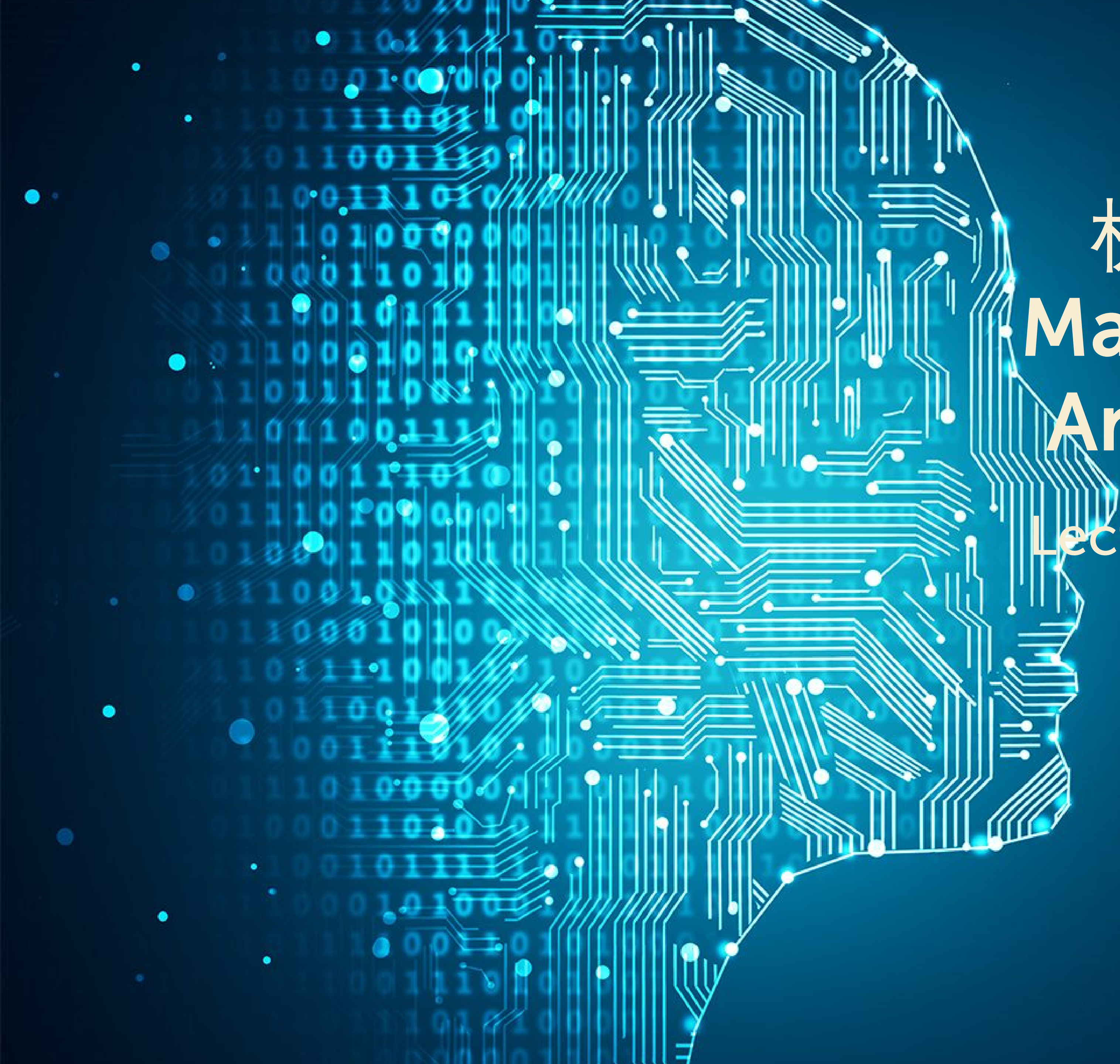


机器学习与人工智能
Machine Learning and Artificial Intelligence

Lecture 1 Introduction/Overview

Yingjie Zhang (张颖婕)
Peking University
yingjiezhang@gsm.pku.edu.cn

2021 Fall

Me...

Emails me: start with “**[MLAI]**” in the subject title

- Email: yingjiezhang@gsm.pku.edu.cn
- Office hour: Wed 10-11am; or by appointment
- Website: sites.google.com/view/yingjiezhang/home

- My Research:
  - Topics: Mobile and Sensor Technologies, Big Data and Smart City, User-generated Content, Sharing Economy, and Social Media.
  - Methodologies: Econometrics, Machine Learning, Text Mining, and Field Experiment.

Teaching Assistant

Guangxin Yang (杨广鑫)

- 2nd PhD Student in Marketing
- Email： ygx@stu.pku.edu.cn.

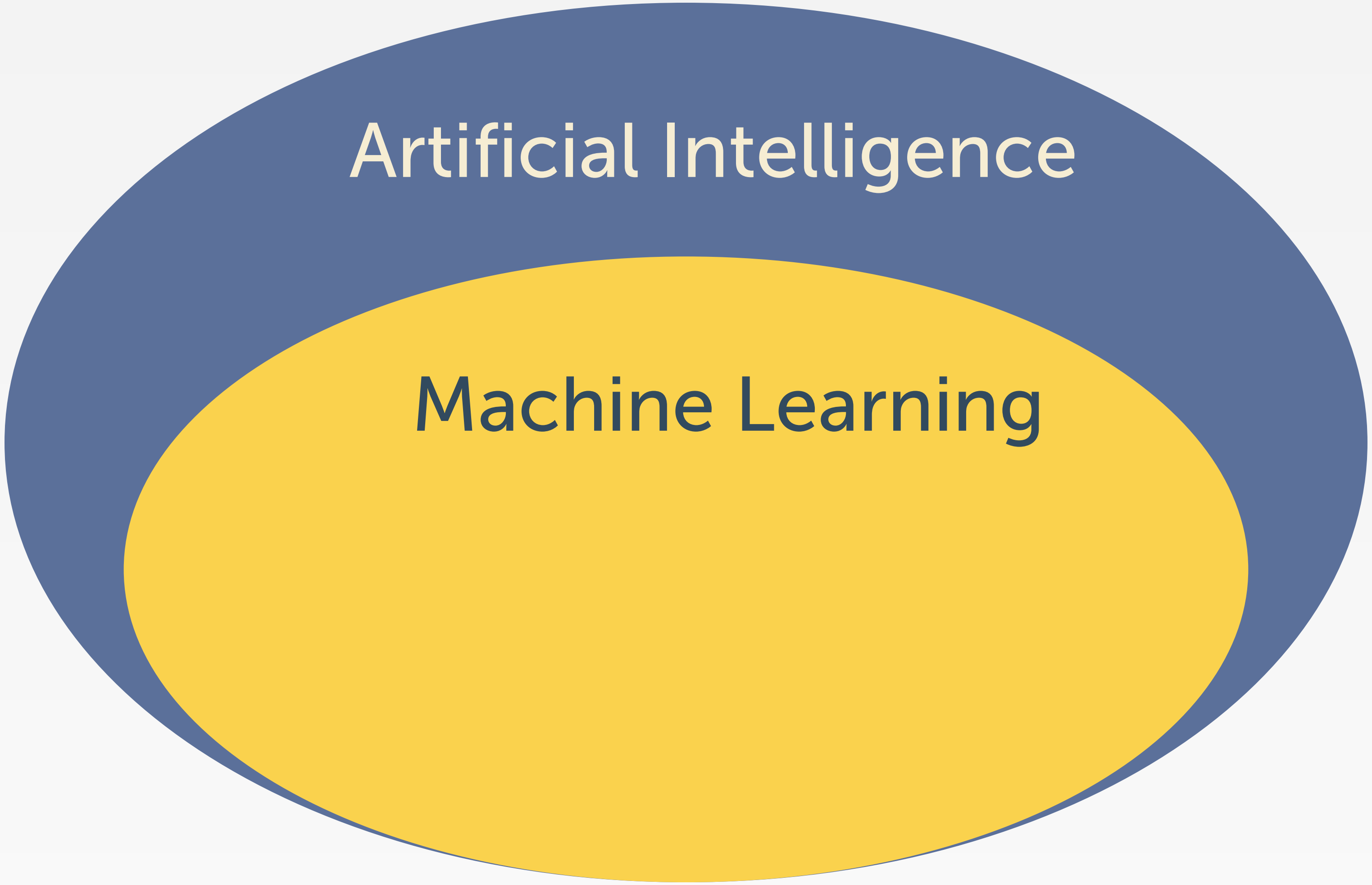
- TA session: Saturday 10:00-11:00; or by appointment

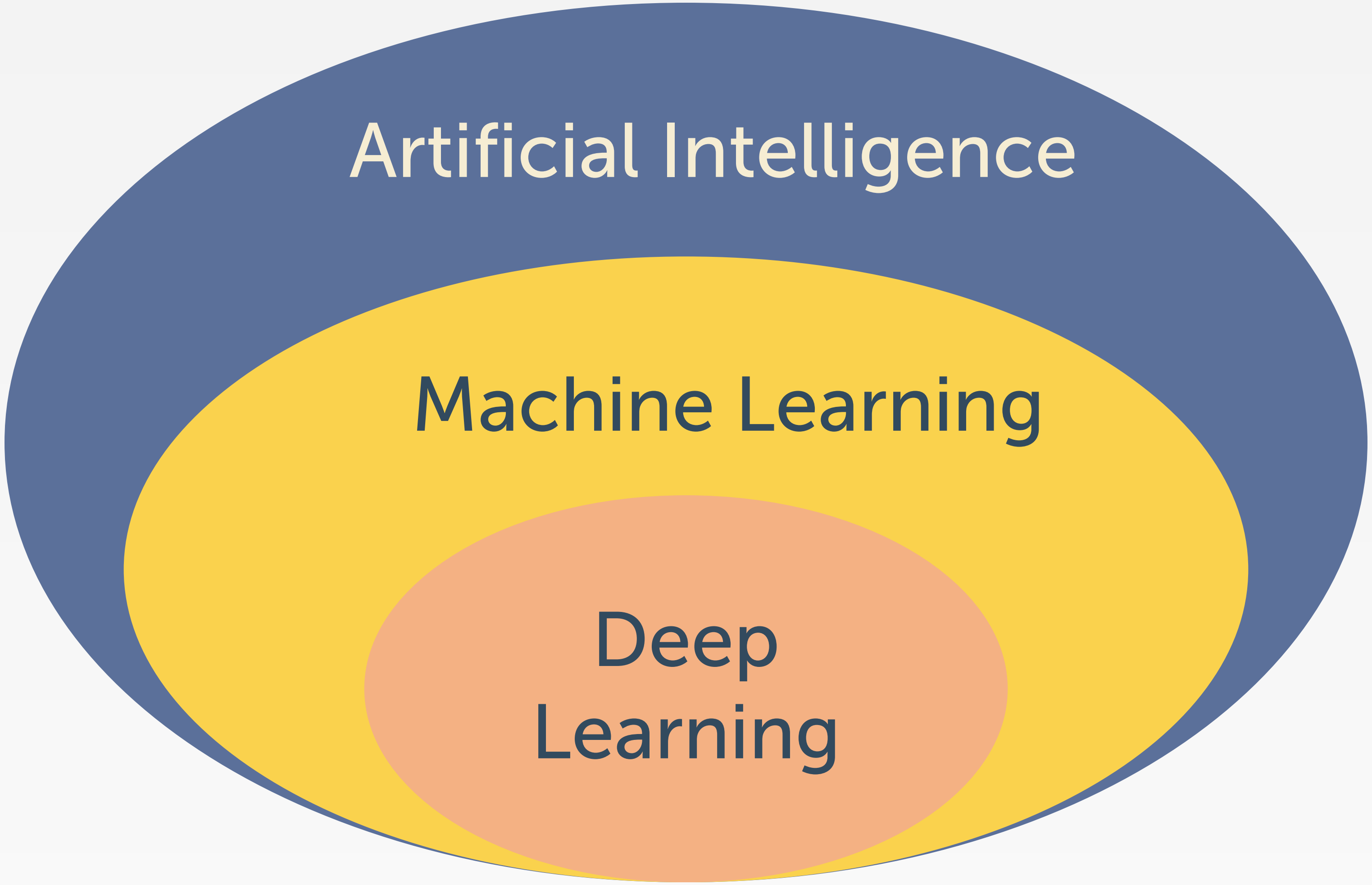
- A novice at ML with you guys. Debug Python code together!

What is MACHINE LEARNING

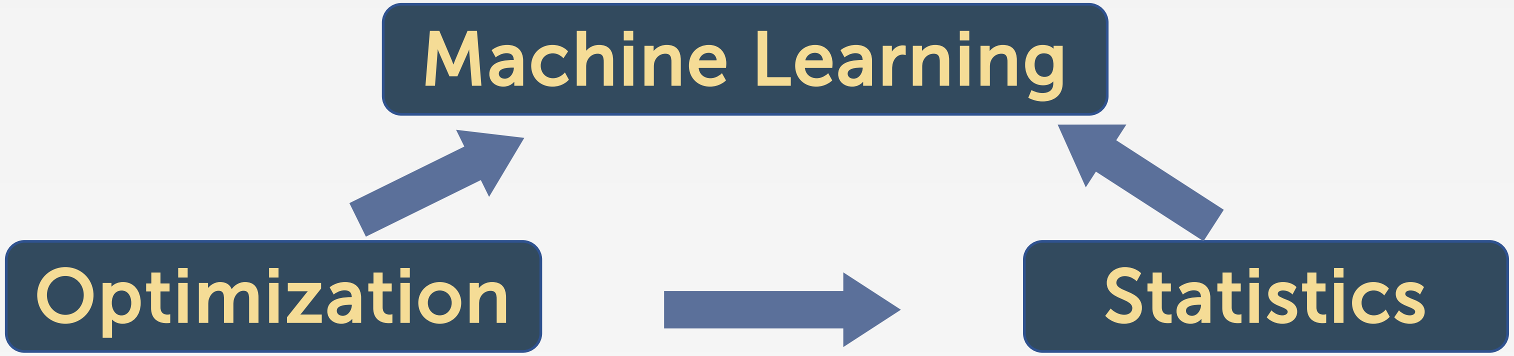
Source: Some slides can be referred to CMU 10601

Artificial Intelligence

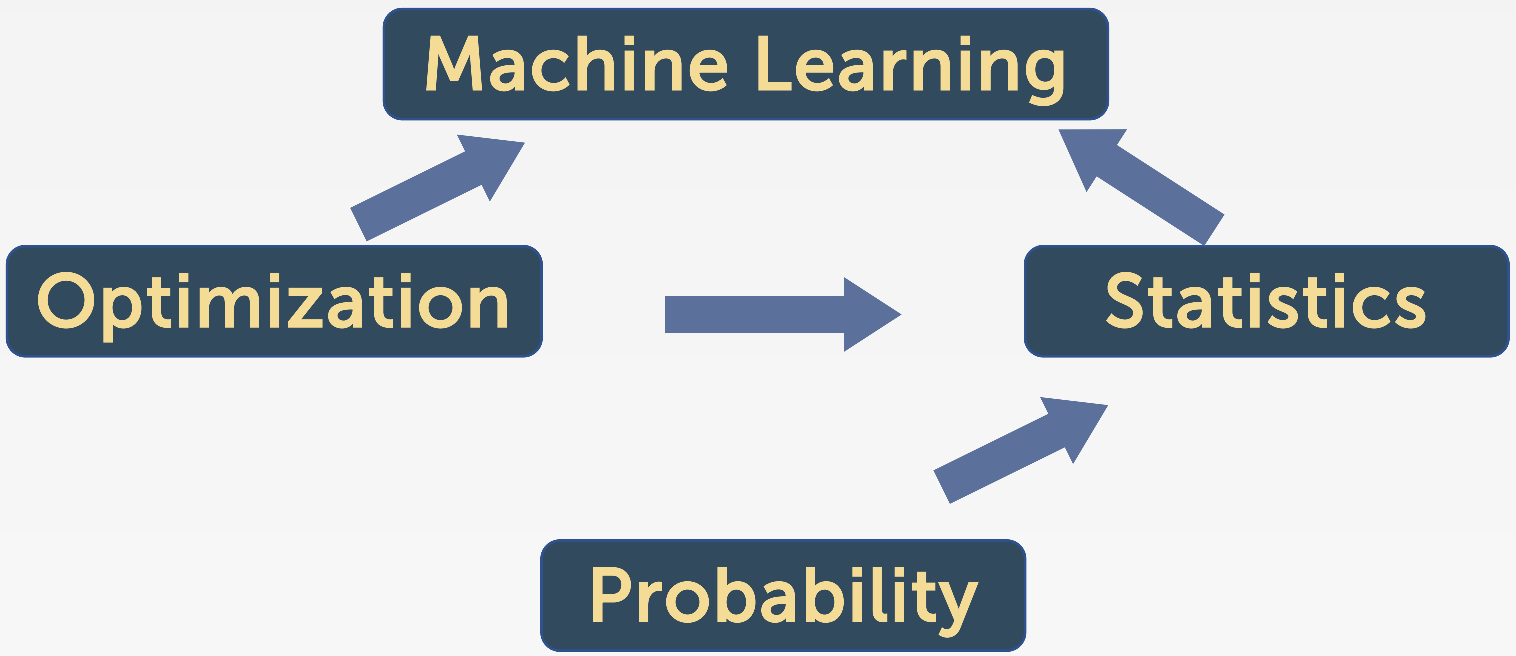

Artificial Intelligence
Machine Learning

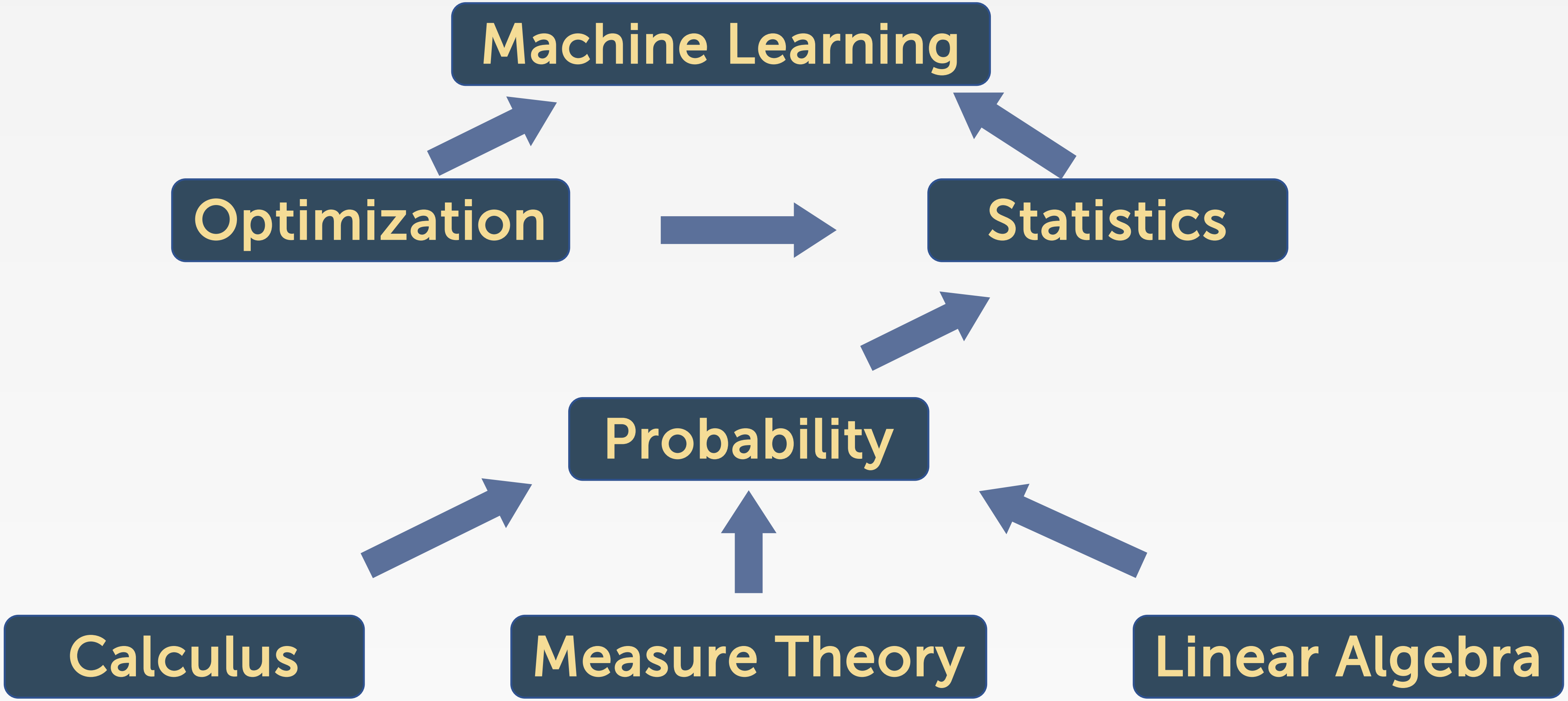

Artificial Intelligence
Machine Learning
Deep
Learning

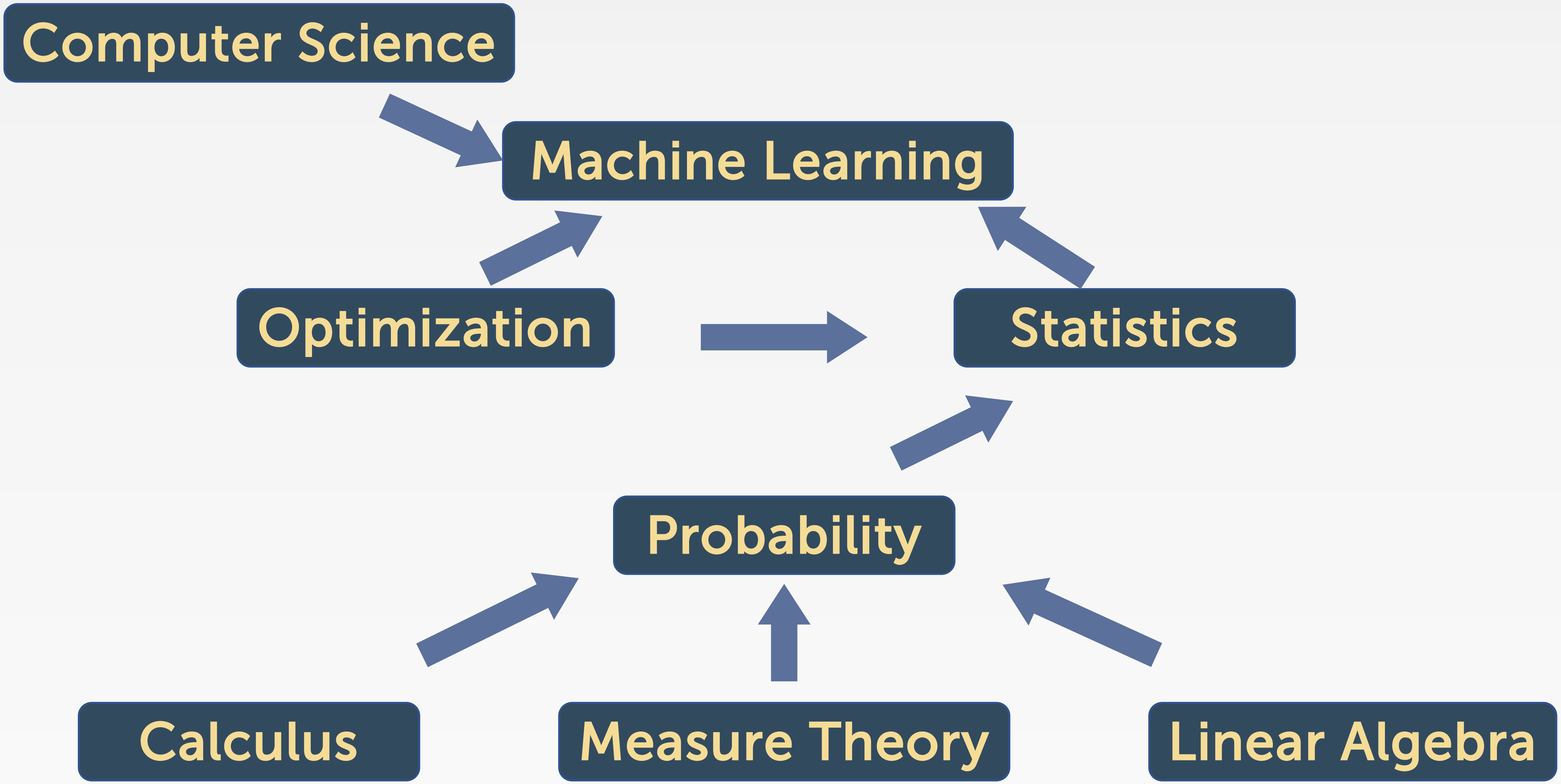
Machine Learning

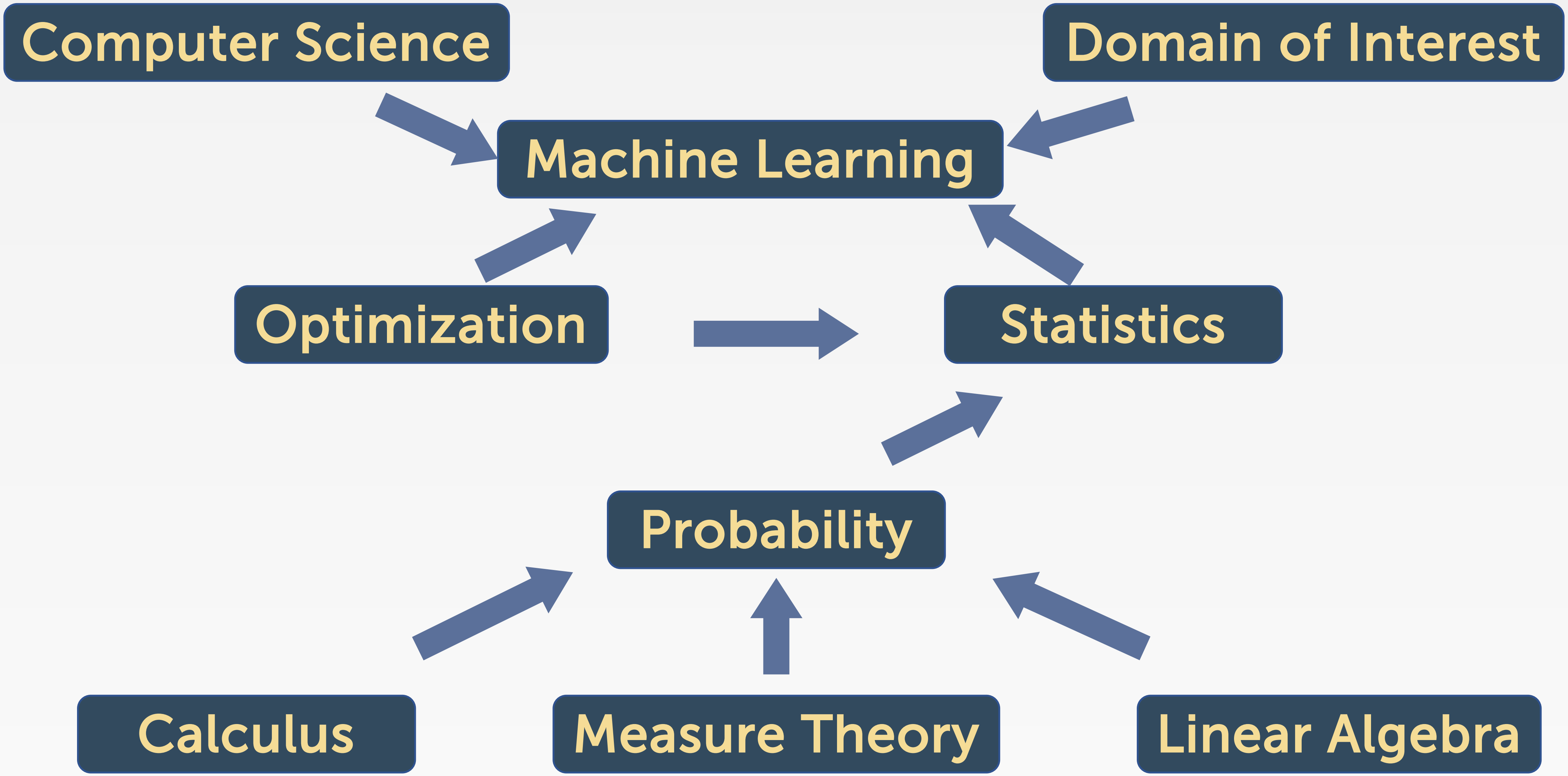


Guanghua School of Management


Machine Learning
Optimization
Statistics
Probability

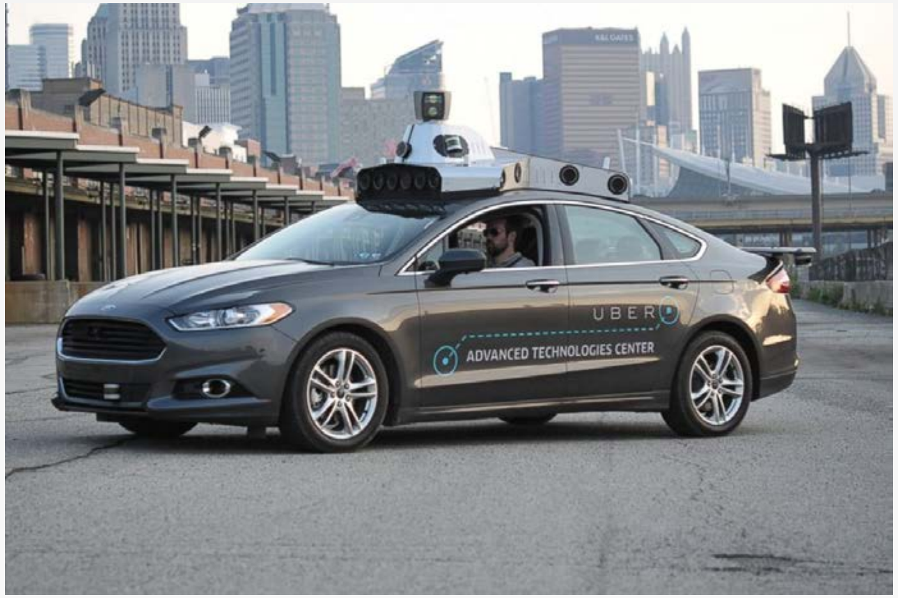

Machine Learning
Optimization
Statistics
Probability
Calculus
Measure Theory
Linear Algebra


Computer Science
Machine Learning
Optimization
Statistics
Probability
Calculus
Measure Theory
Linear Algebra


Computer Science
Domain of Interest
Machine Learning
Optimization
Statistics
Probability
Calculus
Measure Theory
Linear Algebra

ML is everywhere

Learning to drive an autonomous vehicle (robotics)



Tesla Self-Driving cars

Uber Self-Driving services

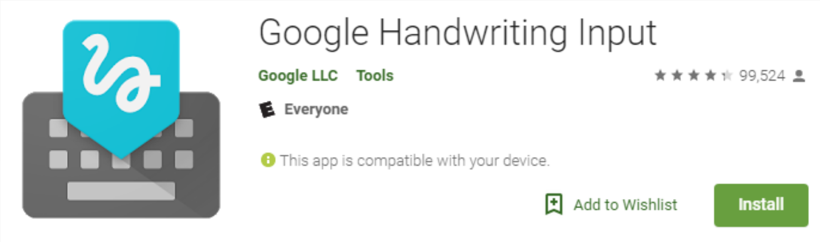
光华管理学院
Guanghua School of Management

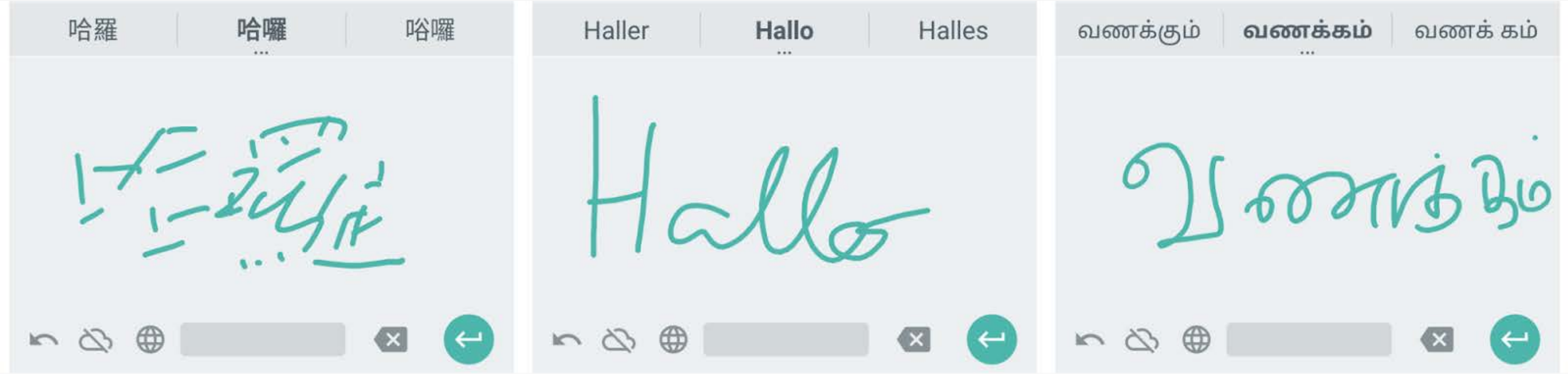
ML is everywhere

Learning to beat the masters at go games (Games/Reasoning)

ML is everywhere

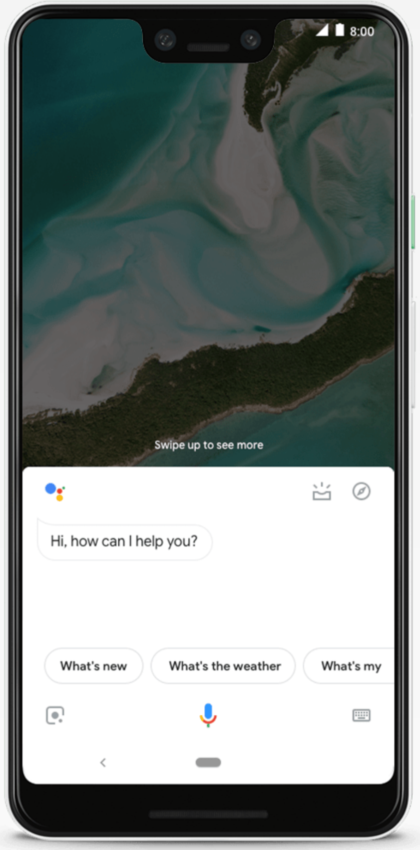
Learning to recognize handwriting (computer vision)





ML is everywhere

Learning to recognize spoken words (speech recognition)

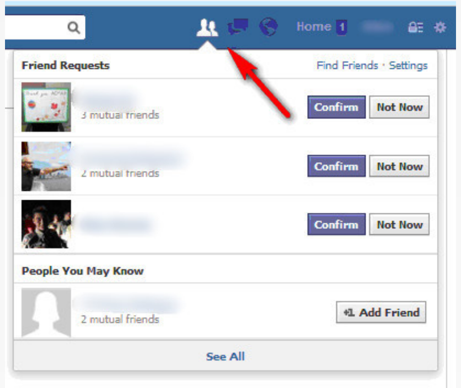


ML is everywhere

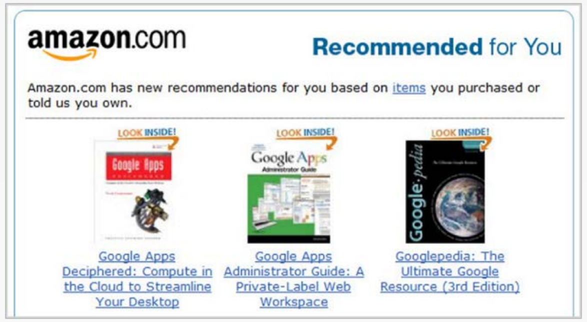
Learning to…

identify spam emails

suggest people you may know



recommend movies



…

Societal/Business Impacts of ML

- Search results are optimized for ad revenues

Societal/Business Impacts of ML

- Search results are optimized for ad revenues

魏则西事件五周年

青年魏则西去世五周年

时间：2016年4月12日

2016年4月12日，21岁的魏则西因滑膜肉瘤去世，在其生前求医过程中，通过百度搜索到武警北京总队第二医院，被该医院宣传的“生物免疫疗法”、“斯坦福技术”所骗，花费不赀却未收获任何效果，贻误合理治疗时机。魏则西去世后，莆田系医院虚假宣传、百度搜索竞价排名、部队医院对外承包混乱等问题引发社会强烈关注。

Societal/Business Impacts of ML

- Search results are optimized for ad revenues
- An autonomous vehicle is permitted to drive unassisted on the road

Societal/Business Impacts of ML

- Search results are optimized for ad revenues
- An autonomous vehicle is permitted to drive unassisted on the road

Societal/Business Impacts of ML

- Search results are optimized for ad revenues
- An autonomous vehicle is permitted to drive unassisted on the road

31岁企业家驾驶蔚来ES8车祸身亡，行驶数据浮出水面

2021年08月15日 19:22:36

来源：北京青年报 1595人参与 420评论

日前，一则蔚来ES8 “自动驾驶” 发生车祸死亡的事件引起网友关注。8月15日，疑似涉事车辆行驶数据曝光。

31岁创始人驾驶ES8车祸身亡

8月14日，认证名为 “美一好” 的个人公众号发布讣告称，2021年8月12日下午2时，上善若水投资管理公司创始人、意统天下餐饮管理公司创始人、美一好品牌管理公司创始人林文钦（昵称 “萌剑客” ），驾驶蔚来ES8汽车启用自动驾驶功能（NOP领航状态）后，在沈海高速涵江段发生交通事故，不幸逝世，终年31岁。

美一好 >

讣告 | 我们的“萌剑客”走了

2021年8月12日下午2时，上善若水投资管理公司创始人、意统天下餐饮管理公司创始人、美一好品牌管理公司创始人林文钦先生（昵称“萌剑客”），驾驶蔚来ES8汽车启用自动驾驶功能（NOP领航状态）后，在沈海高速涵江段发生交通事故，不幸逝世，终年31岁。

Societal/Business Impacts of ML

- Search results are optimized for ad revenues
- An autonomous vehicle is permitted to drive unassisted on the road
- A doctor is prompted by an intelligent system with a plausible diagnosis for her patient

Societal/Business Impacts of ML

- Search results are optimized for ad revenues
- An autonomous vehicle is permitted to drive unassisted on the road
- A doctor is prompted by an intelligent system with a plausible diagnosis for her patient

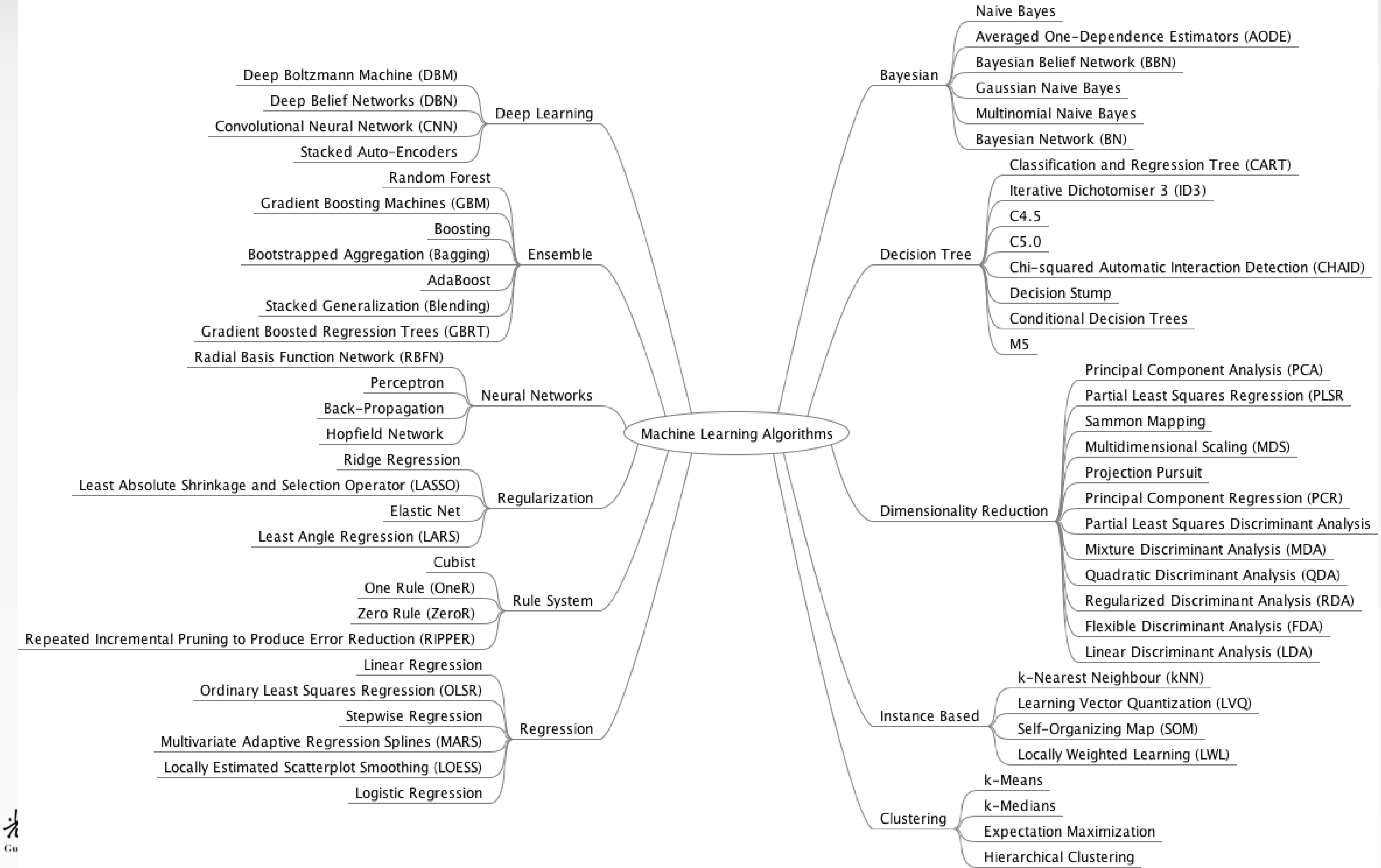
Societal/Business Impacts of ML

- Search results are optimized for ad revenues
- An autonomous vehicle is permitted to drive unassisted on the road
- A doctor is prompted by an intelligent system with a plausible diagnosis for her patient

英国权威医学期刊 diss 医疗AI：在乳腺癌检测上取代放射科医生是痴人说梦

本文作者：我在思考中　2021-09-06 11:04

导语：近日，《英国医学杂志》刊登了一篇研究工作。该团队工作对近年 AI 技术用于乳腺癌筛查的工作进行了检索，希望检验 AI 技术用于乳房 X 光摄像识别的准确度。


Machine Learning Algorithms
Deep Learning
Deep Boltzmann Machine (DBM)
Deep Belief Networks (DBN)
Convolutional Neural Network (CNN)
Stacked Auto-Encoders
Ensemble
Random Forest
Gradient Boosting Machines (GBM)
Boosting
Bootstrapped Aggregation (Bagging)
AdaBoost
Stacked Generalization (Blending)
Gradient Boosted Regression Trees (GBRT)
Neural Networks
Radial Basis Function Network (RBFN)
Perceptron
Back-Propagation
Hopfield Network
Regularization
Ridge Regression
Least Absolute Shrinkage and Selection Operator (LASSO)
Elastic Net
Least Angle Regression (LARS)
Rule System
Cubist
One Rule (OneR)
Zero Rule (ZeroR)
Repeated Incremental Pruning to Produce Error Reduction (RIPPER)
Regression
Linear Regression
Ordinary Least Squares Regression (OLSR)
Stepwise Regression
Multivariate Adaptive Regression Splines (MARS)
Locally Estimated Scatterplot Smoothing (LOESS)
Logistic Regression
Bayesian
Naive Bayes
Averaged One-Dependence Estimators (AODE)
Bayesian Belief Network (BBN)
Gaussian Naive Bayes
Multinomial Naive Bayes
Bayesian Network (BN)
Decision Tree
Classification and Regression Tree (CART)
Iterative Dichotomiser 3 (ID3)
C4.5
C5.0
Chi-squared Automatic Interaction Detection (CHAID)
Decision Stump
Conditional Decision Trees
M5
Dimensionality Reduction
Principal Component Analysis (PCA)
Partial Least Squares Regression (PLSR
Sammon Mapping
Multidimensional Scaling (MDS)
Projection Pursuit
Principal Component Regression (PCR)
Partial Least Squares Discriminant Analysis
Mixture Discriminant Analysis (MDA)
Quadratic Discriminant Analysis (QDA)
Regularized Discriminant Analysis (RDA)
Flexible Discriminant Analysis (FDA)
Linear Discriminant Analysis (LDA)
Instance Based
k-Nearest Neighbour (kNN)
Learning Vector Quantization (LVQ)
Self-Organizing Map (SOM)
Locally Weighted Learning (LWL)
Clustering
k-Means
k-Medians
Expectation Maximization
Hierarchical Clustering

DEFINING a ML Problem

ML brings together different areas

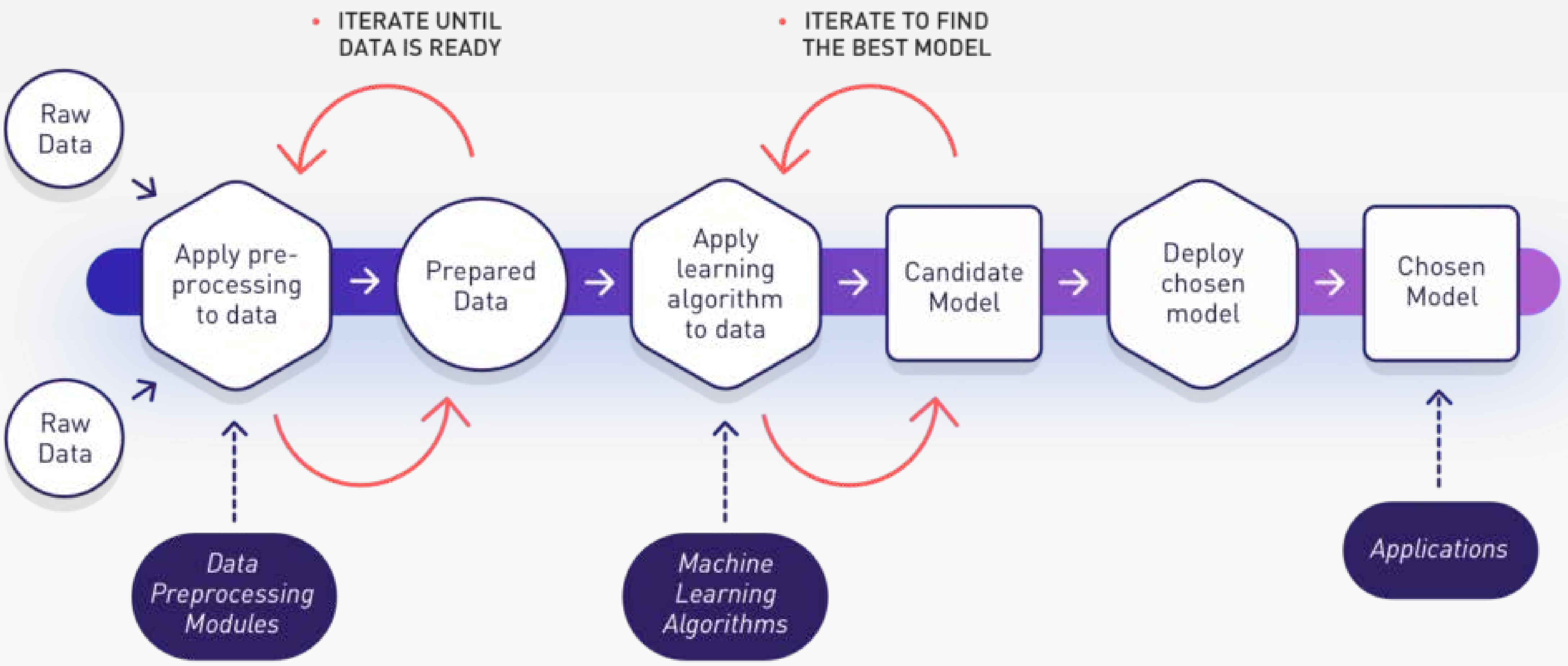
ML brings together different areas

- Statistical methods
  - Infer conclusions from data
  - Estimate reliability of predictions
- Computer science
  - Large-scale computing architectures
  - Algorithms for capturing, manipulating, indexing, combining, retrieving and performing predictions on data
  - Software pipelines that manage the complexity of multiple subtasks

ML brings together different areas

- Statistical methods
  - Infer conclusions from data
  - Estimate reliability of predictions
- Computer science
  - Large-scale computing architectures
  - Algorithms for capturing, manipulating, indexing, combining, retrieving and performing predictions on data
  - Software pipelines that manage the complexity of multiple subtasks
- Economics, biology, psychology
  - How can an individual or system efficiently improve their performance in a given environment?
  - What is learning and how can it be optimized?

Machine Learning Workflow



Define a Learning Problem

- *Definition*: A computer program learns if its performance at tasks in T, as measured by P, improves with experience E.

Define a Learning Problem

- *Definition*: A computer program learns if its performance at tasks in T, as measured by P, improves with experience E.

- Three components
  - Task, T
  - Performance measure, P
  - Training, E

Define a Learning Problem

- Three components
  - Task T
  - Performance measure P
  - Training E

Example I : Handwriting recognition

- T:
- P:
- E:

Define a Learning Problem

- Three components
  - Task T
  - Performance measure P
  - Training E

Example I : Handwriting recognition

- T: recognizing and classifying handwritten words with images
- P: percent of words correctly classified
- E: a database of handwritten words with given classifications

Define a Learning Problem

- Three components
  - Task T
  - Performance measure P
  - Training E

Example II : Self-driving

- T:
- P:
- E:

Define a Learning Problem

- Three components
  - Task T
  - Performance measure P
  - Training E

Example II : Self-driving

- T: driving on public four-lane highways using vision sensors
- P: average distance traveled before an error
- E: a sequence of images and steering commands recorded while observing a human driver

Define a Learning Problem

- Three components
  - Task T
  - Performance measure P
  - Training E

Exercise: Siri response to voice commands

- T: ?
- P: ?
- E: ?

Solution #1 Expert Systems

- Over 20 years ago, we had *rule-based systems*:
  1. Put a bunch of linguists in a room
  2. Have them think about the structure of their native language and write down the rules they devise

Solution #1 Expert Systems

- Over 20 years ago, we had *rule-based systems*:
  1. Put a bunch of linguists in a room
  2. Have them think about the structure of their native language and write down the rules they devise

Give me directions to Starbucks

```
If: "give me directions to X"
Then: directions(here, nearest(X))
```

Solution #1 Expert Systems

- Over 20 years ago, we had *rule-based systems*:
  1. Put a bunch of linguists in a room
  2. Have them think about the structure of their native language and write down the rules they devise

Give me directions to Starbucks

```
If: “give me directions to X”
Then: directions(here, nearest(X))
```

How do I get to Starbucks?

```
If: “how do I get to X””
Then: directions(here, nearest(X))
```

Solution #1 Expert Systems

- Over 20 years ago, we had *rule-based systems*:
  1. Put a bunch of linguists in a room
  2. Have them think about the structure of their native language and write down the rules they devise

Give me directions to Starbucks

```
If: “give me directions to X”
Then: directions(here, nearest(X))
```

How do I get to Starbucks?

```
If: “how do I get to X””
Then: directions(here, nearest(X))
```

Where is the nearest Starbucks?

```
If: “where is the nearest X”
Then: directions(here, nearest(X))
```

Solution #2 Annotate Data and Learn

- Experts:
  - Very good at answering questions about specific cases
  - Not very good at telling HOW they do it
- 1990s: So why not just have them tell you what they do on SPECIFIC CASES and then let MACHINE LEARNING tell you how to come to the same decisions that they did

Solution #2 Annotate Data and Learn

- Collect raw sentences $\{x^{(1)}, \dots, x^{(n)}\}$
- Experts annotate their meaning $\{y^{(1)}, \dots, y^{(n)}\}$

Solution #2 Annotate Data and Learn

- Collect raw sentences $\{x^{(1)}, \ldots, x^{(n)}\}$
- Experts annotate their meaning $\{y^{(1)}, \ldots, y^{(n)}\}$

$x^{(1)}$: Give me directions to Starbucks

$y^{(1)}$: `directions(here, nearest(Starbucks))`

$x^{(2)}$: Send a text to John that I'll be late

$y^{(2)}$: `txtmsg(John, I'll be late)`

$x^{(3)}$: Show me the closest Starbucks

$y^{(3)}$: `map(nearest(Starbucks))`

$x^{(4)}$: Set an alarm for seven in the morning

$y^{(4)}$: `setalarm(7:00AM)`

Define a Learning Problem

- Three components
  - Task T
  - Performance measure P
  - Training E

Exercise: Siri response to voice commands

- T:
- P:
- E:

Define a Learning Problem

- Three components
  - Task T
  - Performance measure P
  - Training E

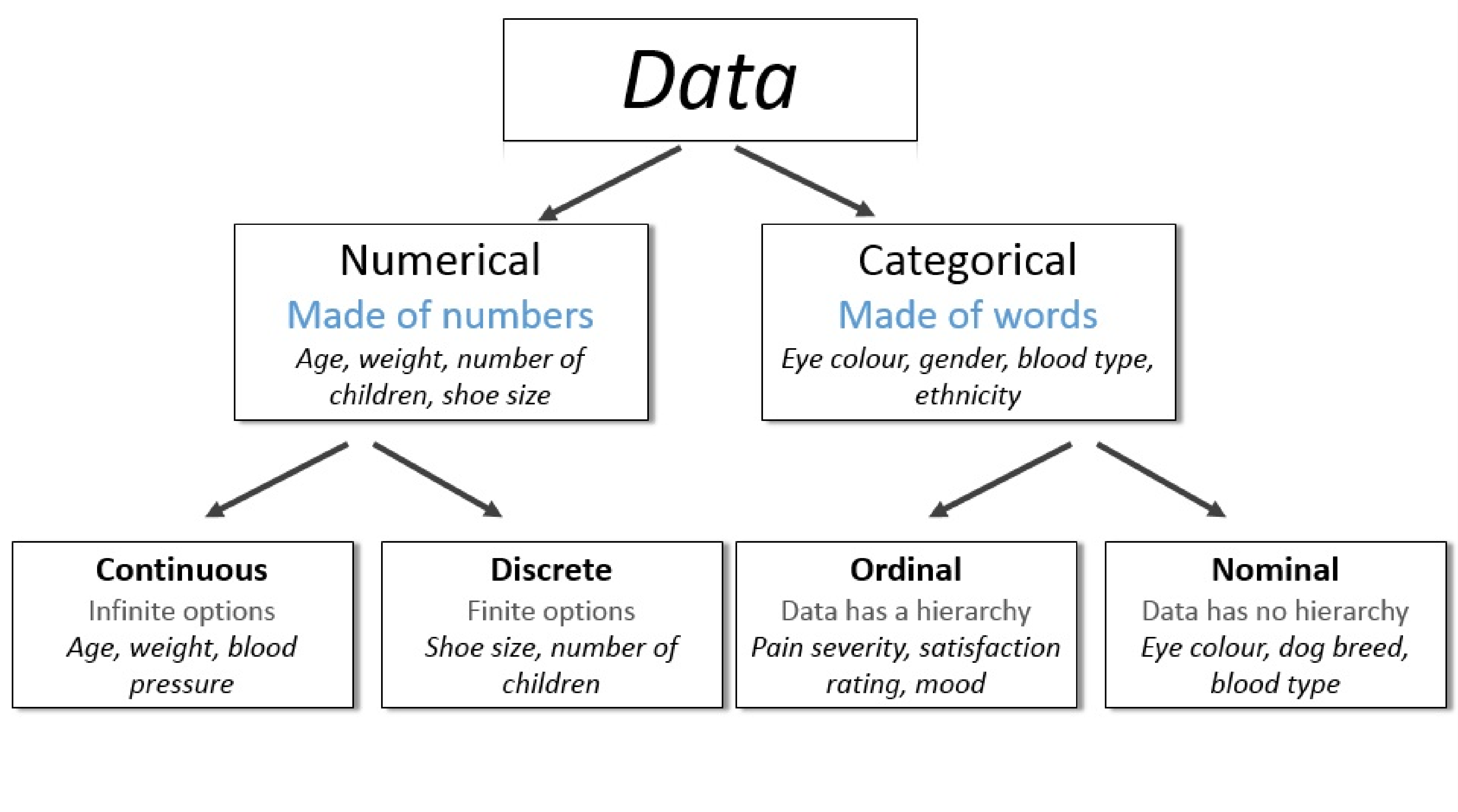
Exercise: Siri response to voice commands

- T: predicting action from speech
- P: percent of correct actions taken in user pilot study
- E: examples of (speech, action) pairs

Problem Formulation

- Formulate a problem in more than one ways:
- Loan applications:
  - Credit score (regression)
  - Default probability (density estimation)
  - Loan decision (classification)

Data Types

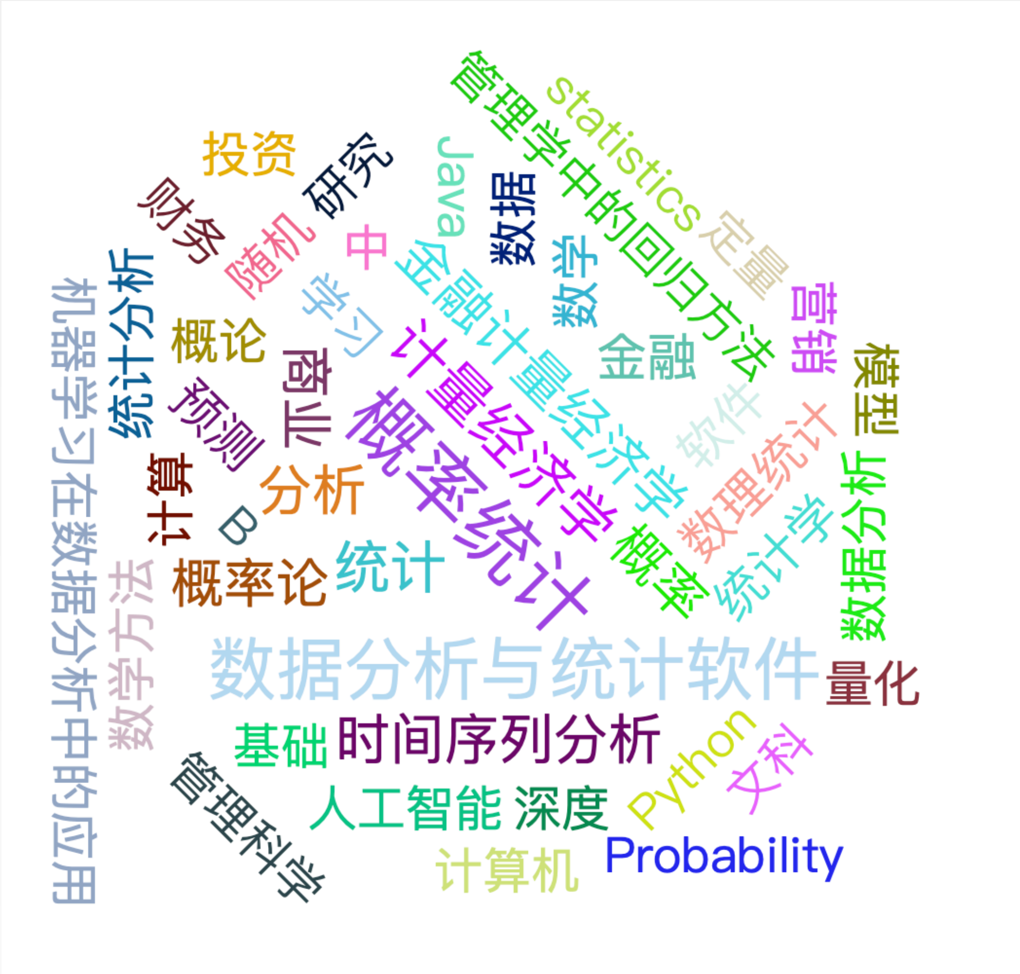

Data
Numerical
Made of numbers
Age, weight, number of children, shoe size
Categorical
Made of words
Eye colour, gender, blood type, ethnicity
Continuous
Infinite options
Age, weight, blood pressure
Discrete
Finite options
Shoe size, number of children
Ordinal
Data has a hierarchy
Pain severity, satisfaction rating, mood
Nominal
Data has no hierarchy
Eye colour, dog breed, blood type

SYLLABUS

学业背景

专业

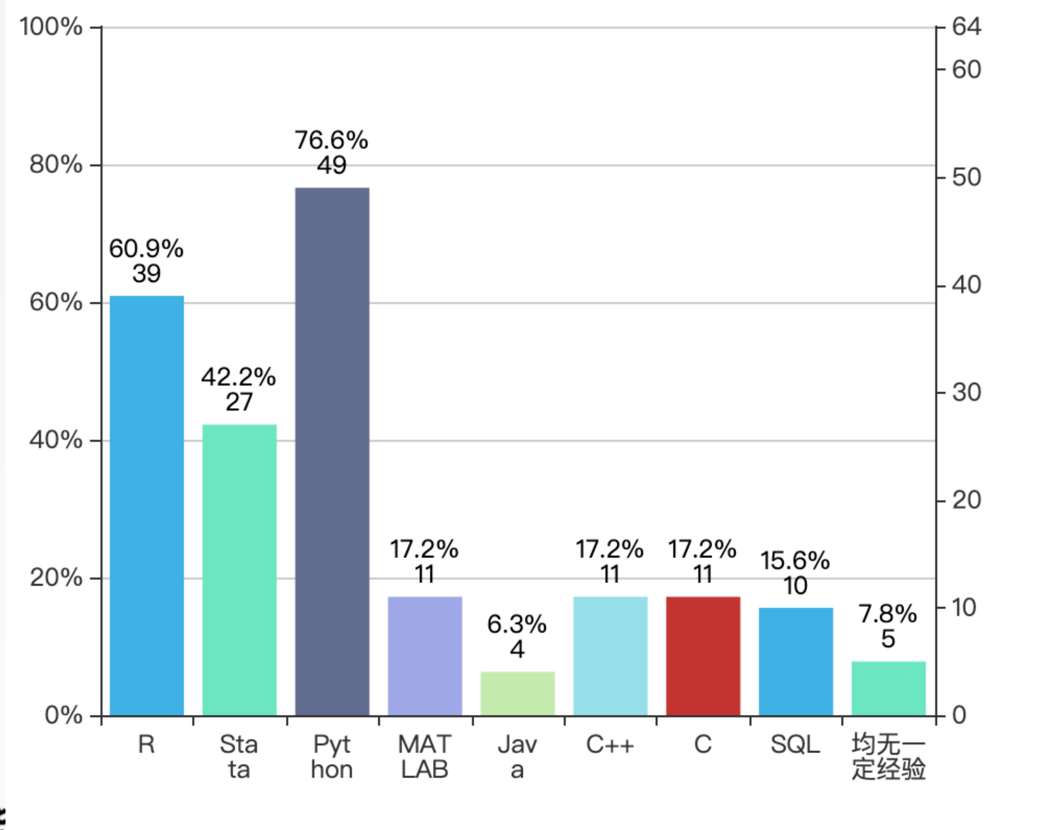
Minor

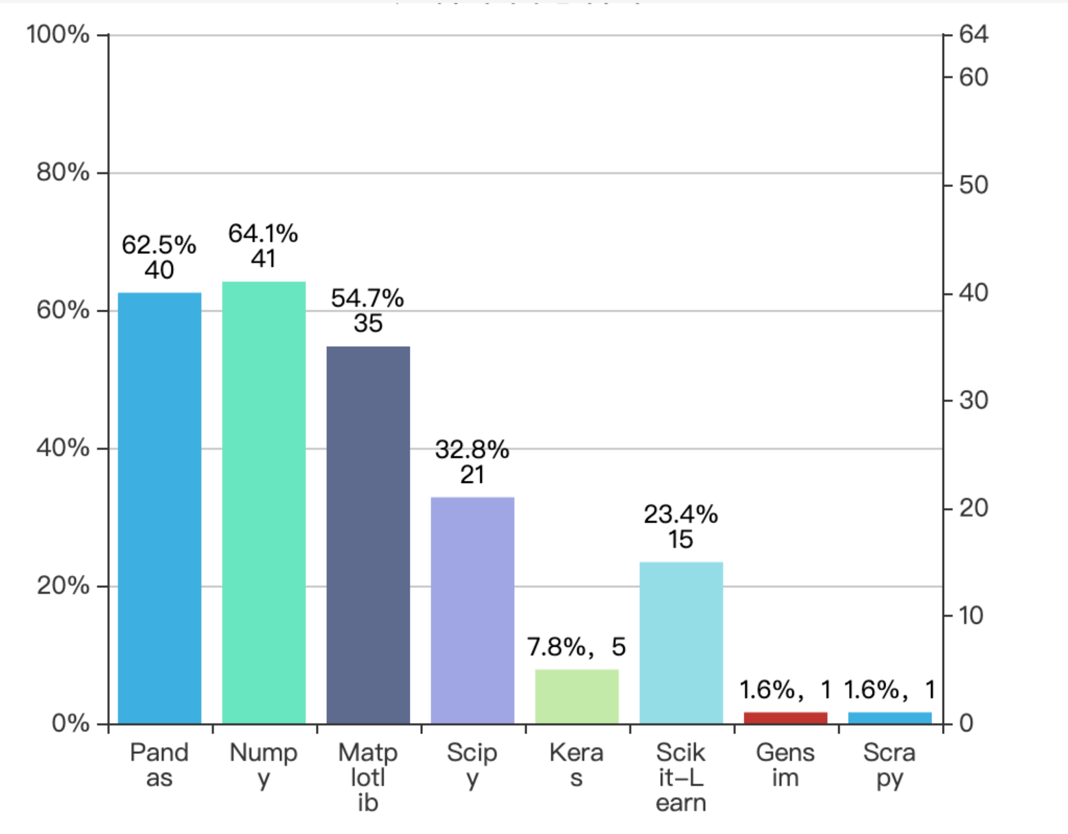


先修课程

编程背景


28%
基本无基础
45%
有一点基础
27%
有比较好的基础（熟练运用至少一款编程语言）


100%
80%
60%
40%
20%
0%
64
60
50
40
30
20
10
0
60.9%
39
42.2%
27
76.6%
49
17.2%
11
6.3%
4
17.2%
11
17.2%
11
15.6%
10
7.8%
5
R
Stata
Python
MATLAB
Java
C++
C
SQL
均无一定经验


100%
80%
60%
40%
20%
0%
64
60
50
40
30
20
10
0
62.5%
40
64.1%
41
54.7%
35
32.8%
21
7.8%，5
23.4%
15
1.6%，1
1.6%，1
Pandas
Numpy
Matplotlib
Scipy
Keras
Scikit-Learn
Gensim
Scrapy

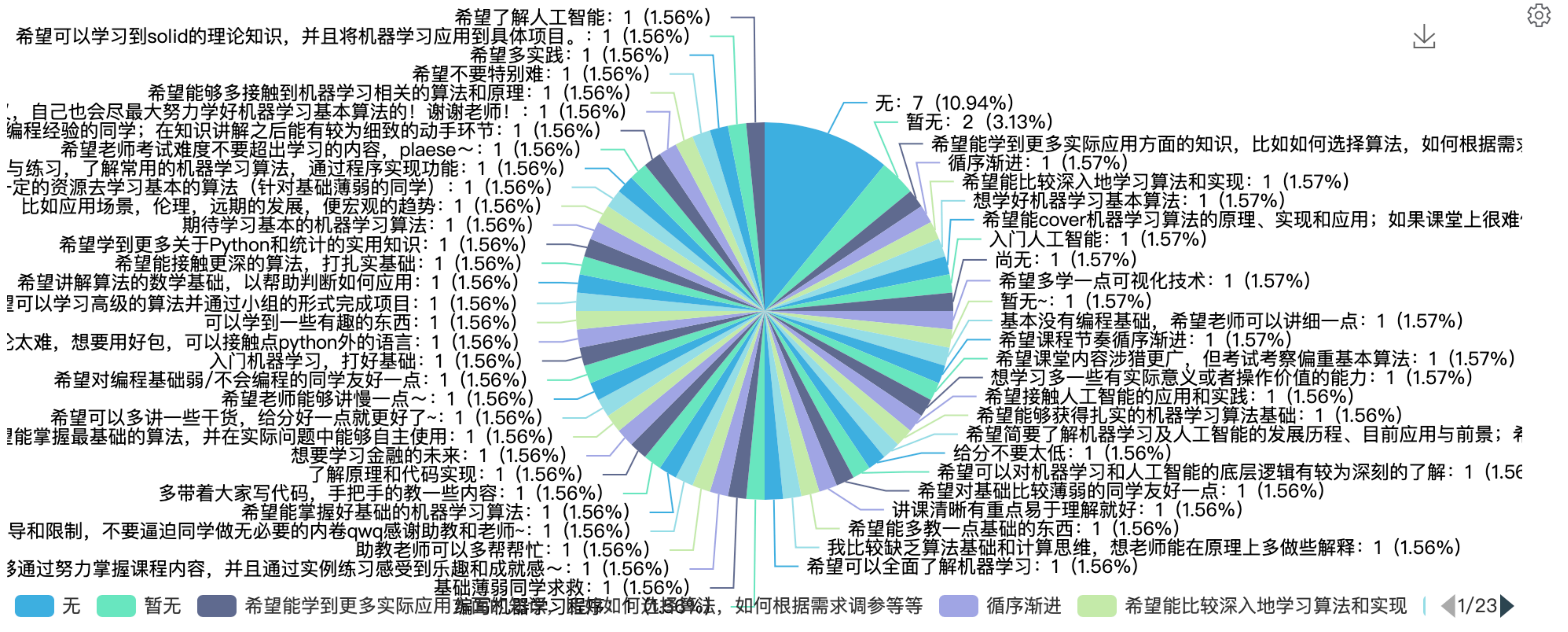
课程期待

42%

心有余而力不足，想学好机器学习基本算法

58%

期待，可以学习更加fancy的算法



Logistics

Logistics

- Homework:
  - 3 individual assignments
  - Late assignment is *NOT* accepted
    - Supervised Learning
    - Unsupervised Learning + Model Evaluation
    - Advanced Topics (e.g., Deep Learning, Reinforcement Learning)
  - Re-grade: You can appeal your grade with a one-page explanation. A re-grade may cause your grade to either go up or go down.

Logistics

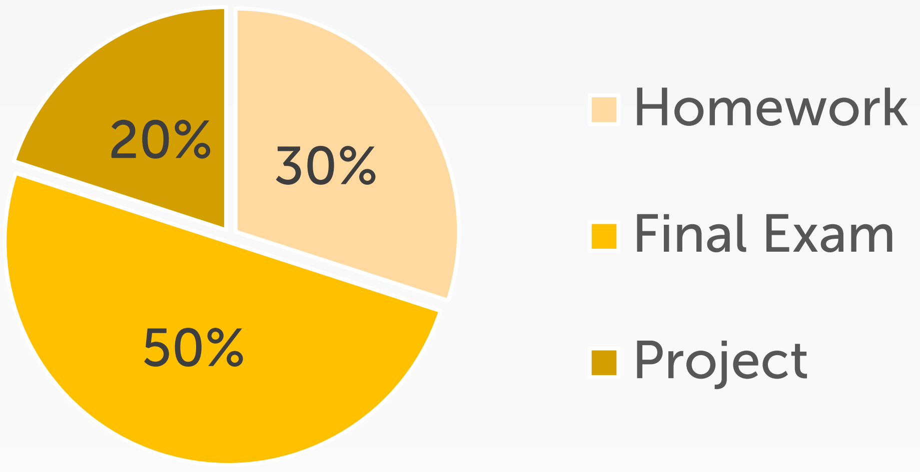
- Homework:
  - 3 individual assignments
  - Late assignment is *NOT* accepted
    - Supervised Learning
    - Unsupervised Learning + Model Evaluation
    - Advanced Topics (e.g., Deep Learning, Reinforcement Learning)
  - Re-grade: You can appeal your grade with a one-page explanation. A re-grade may cause your grade to either go up or go down.
- Group Project
  - At most 5 students per group (No Free-riding)
  - Proposal
  - In-class presentation
  - Final reports

Logistics

- Homework:
  - 3 individual assignments
  - Late assignment is *NOT* accepted
    - Supervised Learning
    - Unsupervised Learning + Model Evaluation
    - Advanced Topics (e.g., Deep Learning, Reinforcement Learning)
  - Re-grade: You can appeal your grade with a one-page explanation. A re-grade may cause your grade to either go up or go down.
- Group Project
  - At most 5 students per group (No Free-riding)
  - Proposal
  - In-class presentation
  - Final reports
- Final Exam
  - Week 12 (in-class)

Logistics

- Academic Honor Code: No plagiarism!
  - form study groups (with arbitrary number of people); discuss and work on homework problems in groups
  - write down the solutions independently
  - write down the names of people with whom you've discussed the homework
- Class participation



Guanghua School of Management

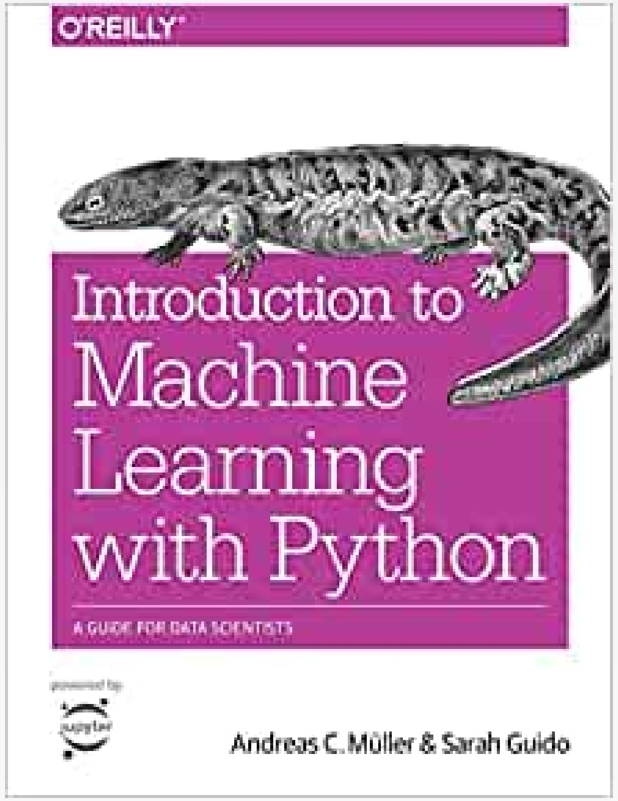
Topics

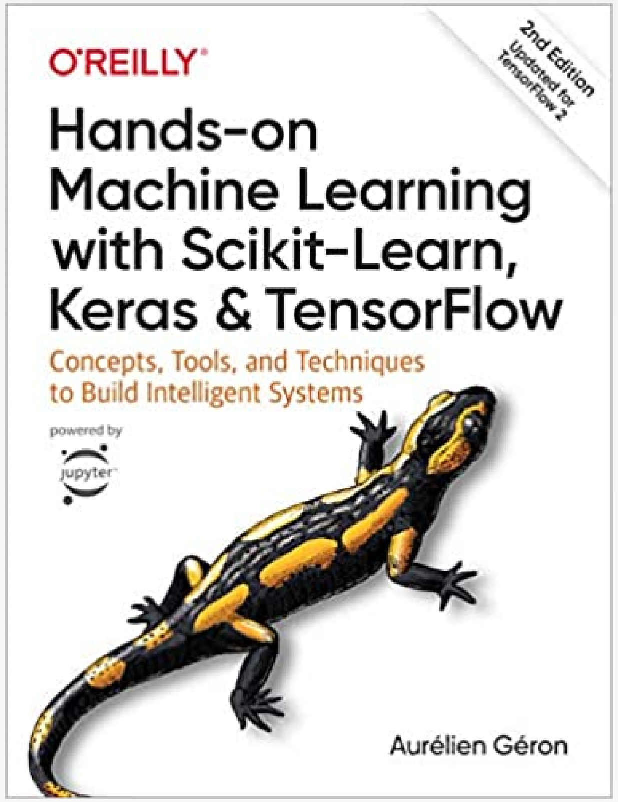
- Regression
- K-Nearest Neighbors
- Decision Trees
- Naïve Bayes
- SVM
- Ensemble Models

- Cross Validation
- Overfitting / Underfitting
- Model Evaluation and Selection

- Unsupervised Learning
  - Clustering
  - PCA
- Reinforcement Learning
- Deep learning
  - Neural network
  - Backpropagation
  - CNNs, LSTM, etc.

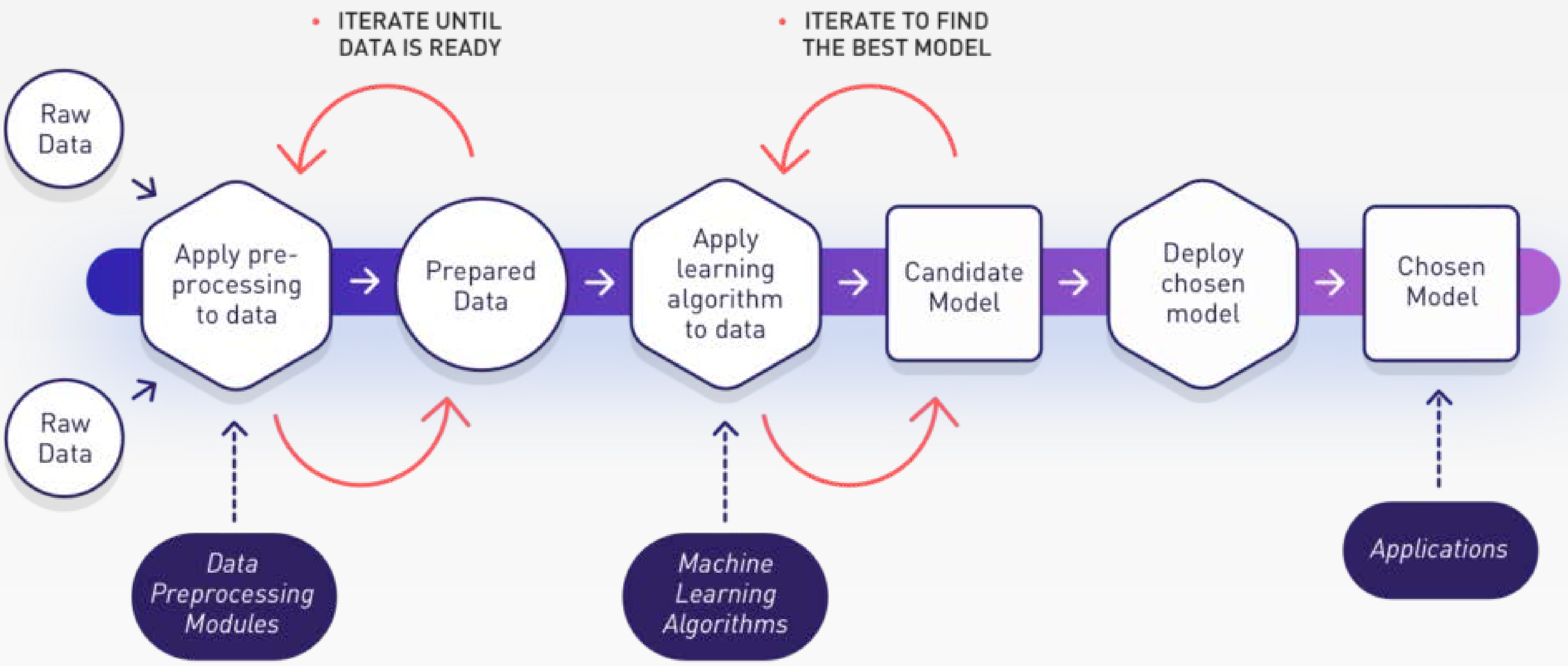
Recommended Books





Overview of ML Models

Machine Learning Workflow



Applied ML

- Understand basic ML concepts and workflow (require basic statistics/probability background)
- Apply properly “black-box” ML components and features
- From theory to real-world practice

Types of ML Systems

Types of ML Systems

Criteria Whether or not they are trained with human supervision

Types of ML Systems

Criteria Whether or not they are trained with human supervision

Supervised Learning

Fraud detection
Prediction of stock markets

Types of ML Systems

Criteria Whether or not they are trained with human supervision

Supervised Learning	Unsupervised Learning
Fraud detection Prediction of stock markets	Customer segmentation Recommendation

Types of ML Systems

Criteria Whether or not they are trained with human supervision

Supervised Learning

Fraud detection
Prediction of stock markets

Unsupervised Learning

Customer segmentation
Recommendation

Semi-supervised Learning

Photo-hosting service
Speech analysis
Web-content classification

Types of ML Systems

Criteria Whether or not they are trained with human supervision

Supervised Learning Fraud detection Prediction of stock markets	**Unsupervised Learning** Customer segmentation Recommendation
Semi-supervised Learning Photo-hosting service Speech analysis Web-content classification	**Reinforcement Learning** Robotics Go games Self-driving cars

Types of ML Systems

Criteria Whether or not they are trained with human supervision

Supervised Learning

Fraud detection
Prediction of stock markets

Unsupervised Learning

Customer segmentation
Recommendation

Semi-supervised Learning

Photo-hosting service
Speech analysis
Web-content classification

Reinforcement Learning

Robotics
Go games
Self-driving cars

Supervised Learning

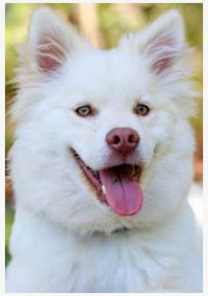
Supervised Learning

- Regressions
- KNN
- Decision Trees
- SVM
- Naïve Bayes
- …

Supervised Learning

Data

Dog



Dog

Not dog

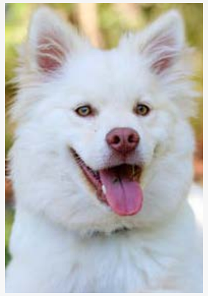
Algorithm f(x)

Label?

Supervised Learning

Data

Dog



Dog

Not dog

Algorithm f(x)

Label?

Training data

Validation data

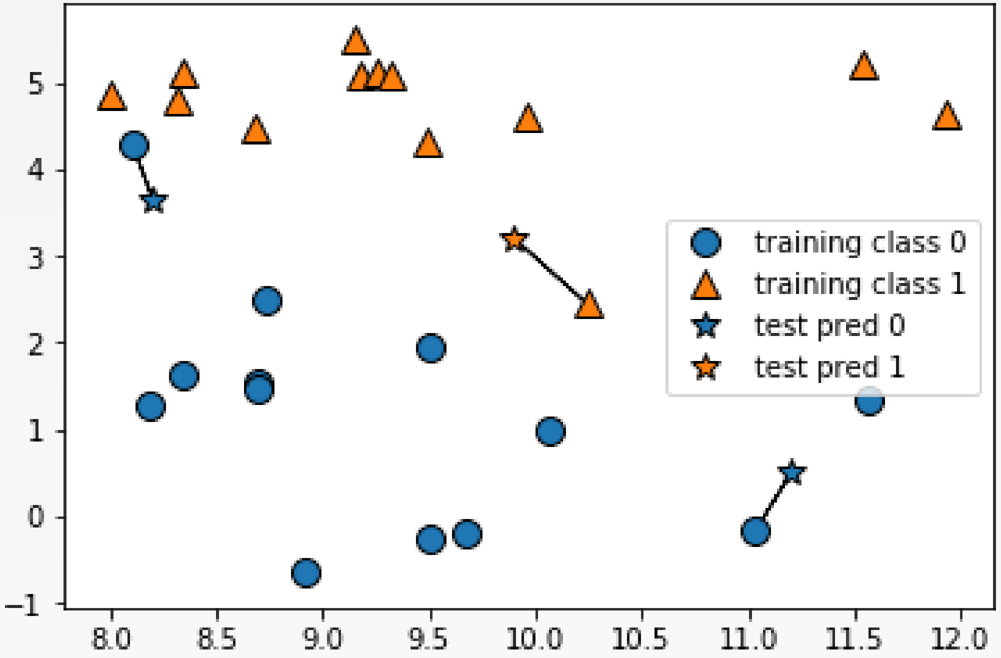
Test data

Data Division

- **Training dataset**: the sample of data used to fit the model
- **Validation Dataset**: the sample of data used to provide an unbiased evaluation of a model fit on the training dataset while tuning model hyperparameters.
- **Test Dataset**: a set of examples used only to assess the performance (i.e. generalization) of a fully specified classifier

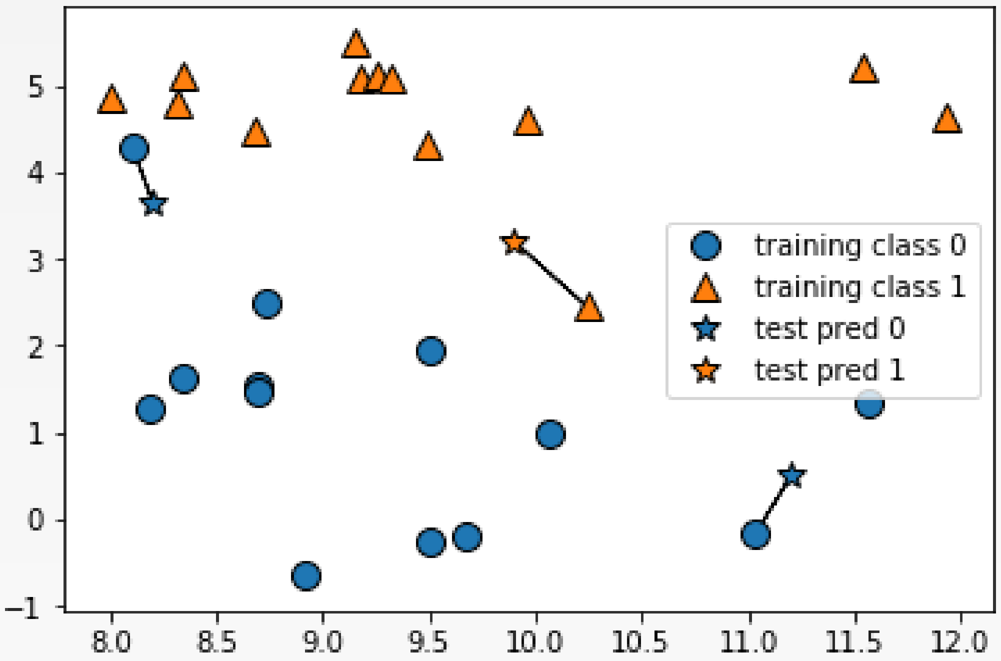
K-Nearest Neighbors

An Illustrative Example

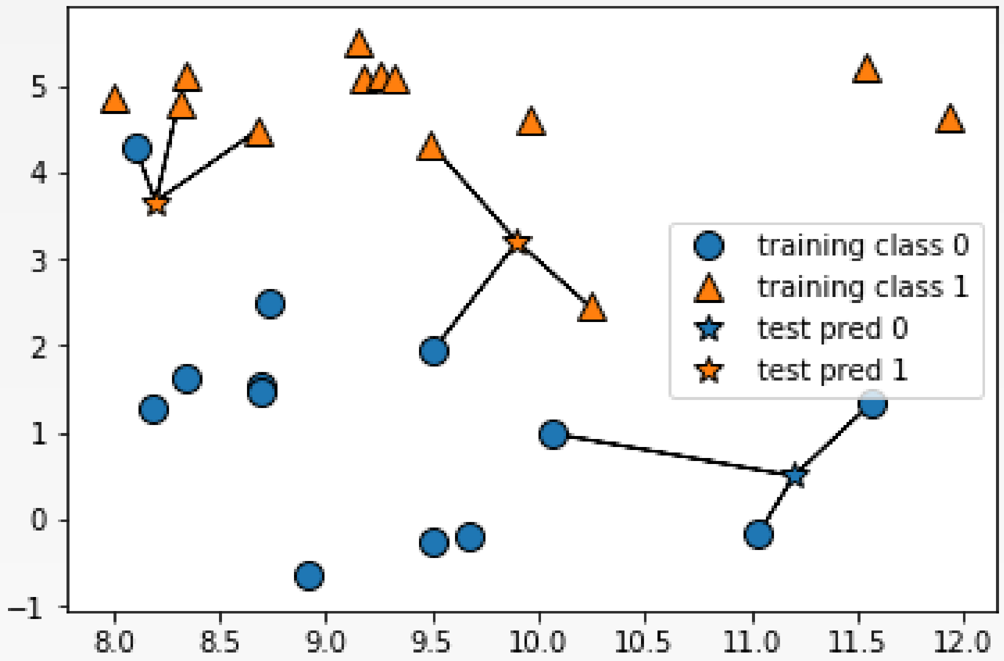


K = 1

An Illustrative Example



K = 1



K = 3

KNN: Classification Algorithm

KNN: Classification Algorithm

- Training: Store all the examples (X_{train}, Y_{train})

KNN: Classification Algorithm

- Training: Store all the examples (X_{train}, Y_{train})
- Prediction: X_{new}
  - Let $X_1, \ldots, X_k$ be the k most similar examples to X_{new}
  - Use certain method (e.g., majority vote) to determine Y_{new} based on $(Y_1, \ldots, Y_k)$

KNN: Classification Algorithm

- Training: Store all the examples (X_{train}, Y_{train})
- Prediction: X_{new}
  - Let $X_1, \dots, X_k$ be the k most similar examples to X_{new}
  - Use certain method (e.g., majority vote) to determine Y_{new} based on $(Y_1, \dots, Y_k)$

Keys

1. A distance metric

Euclidean distance $d(x_j, x_k) = \sqrt{\sum_i (x_{j,i} - x_{k,i})^2}$

Manhattan distance $d(x_j, x_k) = \sum_i |x_{j,i} - x_{k,i}|$

KNN: Classification Algorithm

- Training: Store all the examples (X_{train}, Y_{train})
- Prediction: X_{new}
  - Let $X_1, \dots, X_k$ be the k most similar examples to X_{new}
  - Use certain method (e.g., majority vote) to determine Y_{new} based on $(Y_1, \dots, Y_k)$

Keys

1. A distance metric

Euclidean distance $d(x_j, x_k) = \sqrt{\sum_i (x_{j,i} - x_{k,i})^2}$

Manhattan distance $d(x_j, x_k) = \sum_i |x_{j,i} - x_{k,i}|$

2. Value of "K"

Cross validation: larger k? smaller k?

KNN: Classification Algorithm

- Training: Store all the examples (X_{train}, Y_{train})
- Prediction: X_{new}
  - Let $X_1, \dots, X_k$ be the k most similar examples to X_{new}
  - Use certain method (e.g., majority vote) to determine Y_{new} based on $(Y_1, \dots, Y_k)$

Keys

1. A distance metric

Euclidean distance $d(x_j, x_k) = \sqrt{\sum_i (x_{j,i} - x_{k,i})^2}$

Manhattan distance $d(x_j, x_k) = \sum_i |x_{j,i} - x_{k,i}|$

2. Value of "K"

Cross validation: larger k? smaller k?

3. Aggregation of the classes of neighbor points

Majority vote

KNN: Pros and Cons

KNN: Pros and Cons

- Advantages:
  - Very simple and intuitive
  - The cost of the learning process is zero
  - No assumption about the characteristics/distributions
  - Works on both classification and regression tasks

KNN: Pros and Cons

- Advantages:
  - Very simple and intuitive
  - The cost of the learning process is zero
  - No assumption about the characteristics/distributions
  - Works on both classification and regression tasks

- Drawbacks:
  - Computationally expensive when the dataset is very large
    - Need to calculate the compare distance from new example to all other examples
  - Sensitive to outliers

Python...

Python Quick Checks

- I can read python codes…
- I can write python functions…
- Errors and debugging…

Coding Tips

- Comments?
- Printed messages?
- Functions?

iPython

- Command: jupyter notebook
- Install: Anaconda

- Programming in the browser
- Codes, instructions, and outputs are displayed “in-line”
- Useful for writing codes that tells a story
- Used by scientists and researchers
- …

Python Packages

- Scikit-learn: Python Machine Learning Library

```
from sklearn.tree import DecisionTreeClassifier
```

- Numpy: Scientific Computing Library
  - Typically, data input to scikit-learn will be in the form of a Numpy array

```
import numpy as np
```

- Pandas: Data Manipulation

```
import pandas as pd
```

- Matplotlib: Plotting Library

```
import matplotlib.pyplot as plt
```

- Others: mglearn; graphviz; seaborn

Python Practice

Questions?

For Next Week…

- Python Review
- Regressions
- Bring your laptop with Python (and packages) installed