

# CS420 Machine Learning

Weinan Zhang

Shanghai Jiao Tong University

<http://wnzhang.net>

Spring Semester, 2018

<http://wnzhang.net/teaching/cs420/index.html>

# Self Introduction – Weinan Zhang

- Position
  - Assistant Professor at John Hopcroft Center, CS Dept. of SJTU 2016-now
  - Apex Data and Knowledge Management Lab
  - Research on machine learning and data mining topics
- Education
  - Ph.D. on Computer Science from University College London (UCL), United Kingdom, 2012-2016
  - B.Eng. on Computer Science from ACM Class 07 of Shanghai Jiao Tong University, China, 2007-2011

# Course Administration

- No official text book for this course, some recommended books are
  - 李航《统计学习方法》清华大学出版社，2012.
  - 周志华《机器学习》清华大学出版社，2016.
  - Tom Mitchell. "Machine Learning". McGraw-Hill, 1997
  - Jerome H. Friedman, Robert Tibshirani, and Trevor Hastie. "The Elements of Statistical Learning". Springer 2004.
  - Chris Bishop. "Pattern Recognition and Machine Learning". Springer 2006.
  - Richard S. Sutton and Andrew G. Barto. "Reinforcement Learning: An Introduction". MIT, 2012.

# Course Administration

- A hands-on machine learning course
  - No assignment, no paper exam
  - Select two out of three course works (80%)
    - Kaggle-in-Class competitions on Classification (40%)
    - Kaggle-in-Class competitions on Recommendation (40%)
    - MAgent battle game competition (40%)
  - Poster session (10%)
  - Attending (10%)
    - Could be evaluated by quiz

# Teaching Assistants



- Jiacheng Yang (杨嘉成)
- kipsora [A.T.] gmail.com
- ACM15 student, research intern in Apexlab
- Research on AutoML, reinforcement learning, deep learning
- Papers: NIPS & AAAI (published as demos), ICML (submitted)



- Lianmin Zheng (郑怜悯)
- mercy\_zheng [A.T.] apex.sjtu.edu.cn
- ACM15 student, research intern in Apexlab
- Research on machine learning systems, multi-agent reinforcement learning
- Papers: NIPS & AAAI (published as demos)

# TA Administration

- Join the mail list
  - Please send your
    - Chinese name
    - Student number
    - Email address

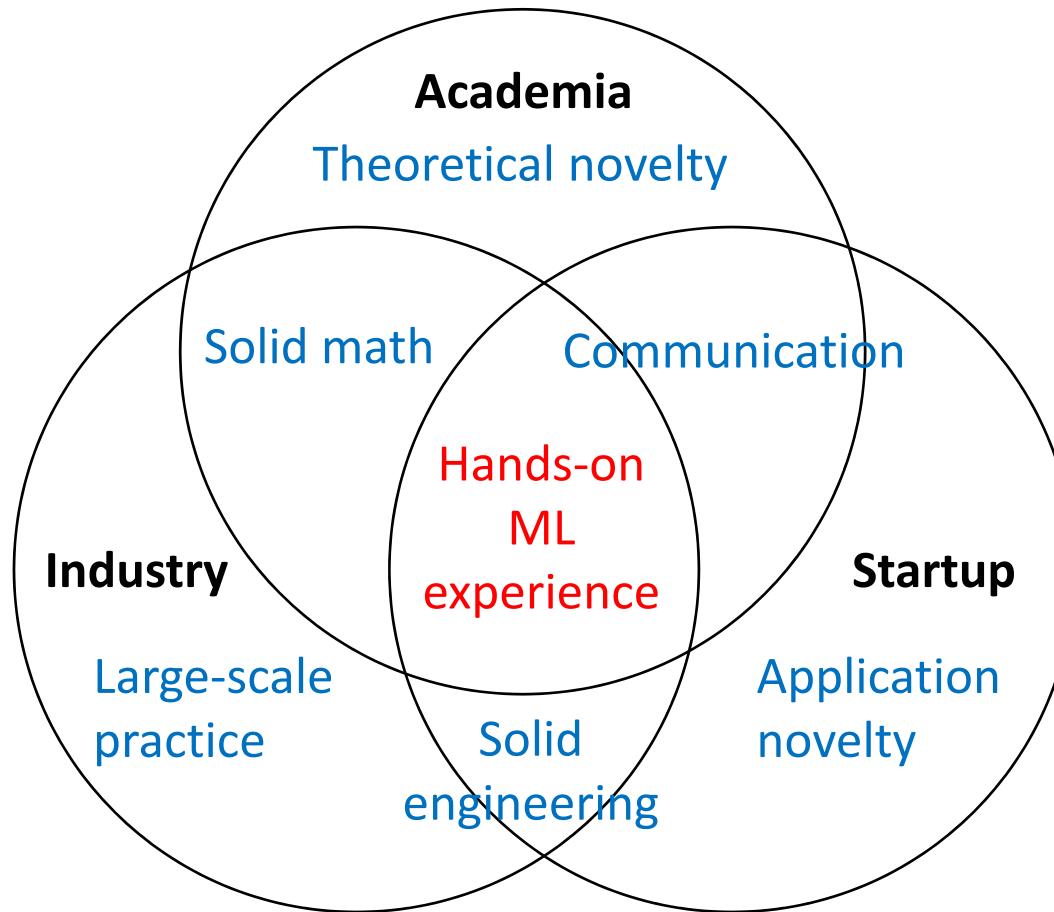
to [mercy\\_zheng \[A.T\] sjtu.edu.cn](mailto:mercy_zheng[A.T]sjtu.edu.cn)  
with email title “Check in CS420 2018”

- Office hour
  - Every Wednesday 7-8pm, 307 Yifu Building
  - TAs will be there for QA

# Goals of This Course

- Know about the big picture of machine learning
- Get familiar with popular ML methodologies
  - Data representations
  - Models
  - Learning algorithms
  - Experimental methodologies
- Get some first-hand ML developing experiences
- Present your own ML solutions to real-world problems

# Why we focus on hands-on ML



- So play with the data and get your hands dirty!

# Course Landscape

- |                                |                           |
|--------------------------------|---------------------------|
| 1. ML Introduction             | 9. Unsupervised Learning  |
| 2. Linear Models               | 10. Model Selection       |
| 3. SVMs and Kernels [cw1]      | 11. RL Introduction [cw3] |
| 4. Neural Networks             | 12. Model-free RL         |
| 5. Tree Models                 | 13. Multi-agent RL        |
| 6. Ensemble Models             | 14. Transfer Learning     |
| 7. Ranking and Filtering [cw2] | 15. Advanced ML           |
| 8. Graphic Models              | 16. Poster Session        |

2018 CS420 Machine Learning, Lecture 1

# Introduction to Machine Learning

Weinan Zhang

Shanghai Jiao Tong University

<http://wnzhang.net>

# Artificial Intelligence

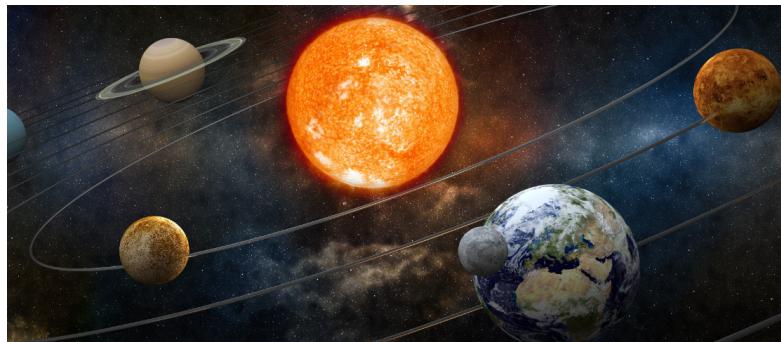
- Artificial intelligence (AI) is intelligence exhibited by machines.
- The subject AI is about the methodology of designing machines to accomplish intelligence-based tasks.
- Intelligence is the computational part of the ability to achieve goals in the world.

# Methodologies of AI

- Rule-based
  - Implemented by direct programming
  - Inspired by human heuristics
- Data-based
  - Expert systems
    - Experts or statisticians create rules of predicting or decision making based on the data
  - Machine learning
    - Direct making prediction or decisions based on the data
    - Data Science

# What is Data Science

- Physics
    - Goal: discover the underlying principle of the world



- **Solution:** build the model of the world from observations

$$F = G \frac{m_1 m_2}{r^2}$$

- Data Science
    - Goal: discover the underlying principle of the data



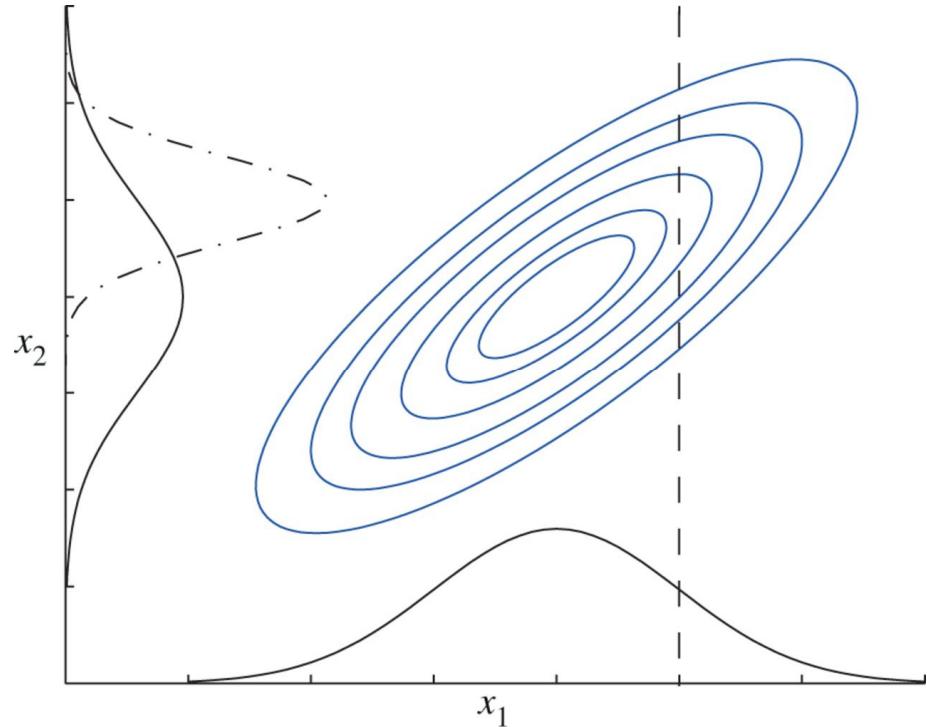
- **Solution:** build the model of the data from observations

$$p(x) = \frac{e^{f(x)}}{\sum_{x'} e^{f(x')}}$$

# Data Science

- Mathematically
  - Find joint data distribution  $p(x)$
  - Then the conditional distribution  $p(x_2|x_1)$
- Gaussian distribution
  - Multivariate

$$p(x) = \frac{e^{-(x-\mu)^\top \Sigma^{-1}(x-\mu)}}{\sqrt{|2\pi\Sigma|}}$$



- Univariate

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

# A Simple Example in User Behavior Modelling

Interest	Gender	Age	BBC Sports	PubMed	Bloomberg Business	Spotify
Finance	Male	29	Yes	No	Yes	No
Sports	Male	21	Yes	No	No	Yes
Medicine	Female	32	No	Yes	No	No
Music	Female	25	No	No	No	Yes
Medicine	Male	40	Yes	Yes	Yes	No

- Joint data distribution

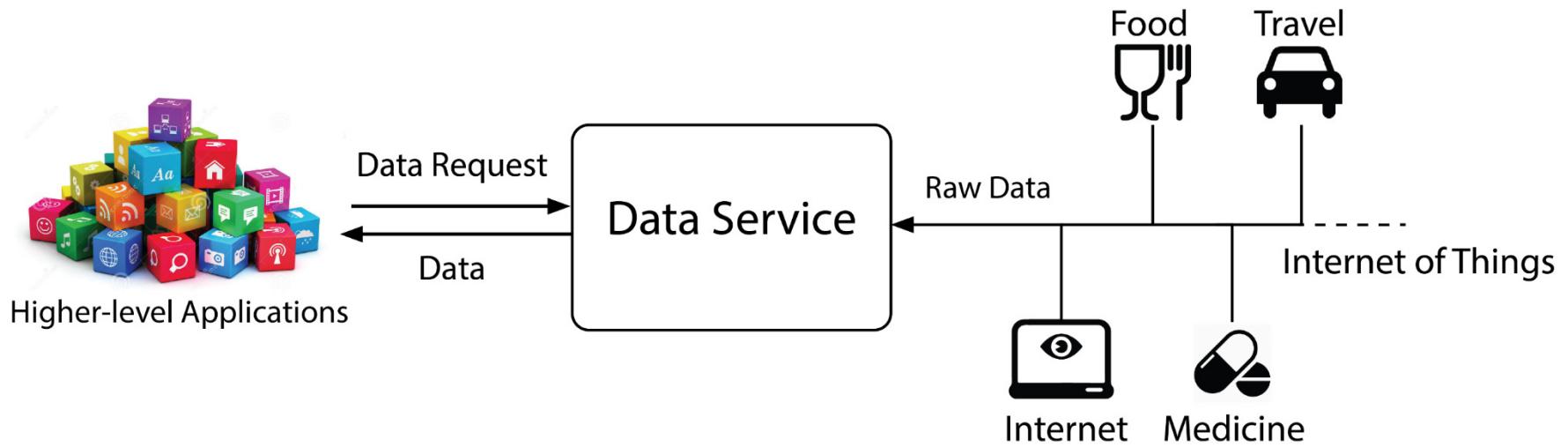
$p(\text{Interest}=\text{Finance}, \text{Gender}=\text{Male}, \text{Age}=29, \text{Browsing}=\text{BBC Sports}, \text{Bloomberg Business})$

- Conditional data distribution

$p(\text{Interest}=\text{Finance} \mid \text{Browsing}=\text{BBC Sports}, \text{Bloomberg Business})$

$p(\text{Gender}=\text{Male} \mid \text{Browsing}=\text{BBC Sports}, \text{Bloomberg Business})$

# Data Technology



Data itself is not valuable, data service is!

# What is Machine Learning

- Learning

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

--- Herbert Simon  
Carnegie Mellon University

Turing Award (1975)  
artificial intelligence, the psychology of human cognition

Nobel Prize in Economics (1978)  
decision-making process within economic organizations



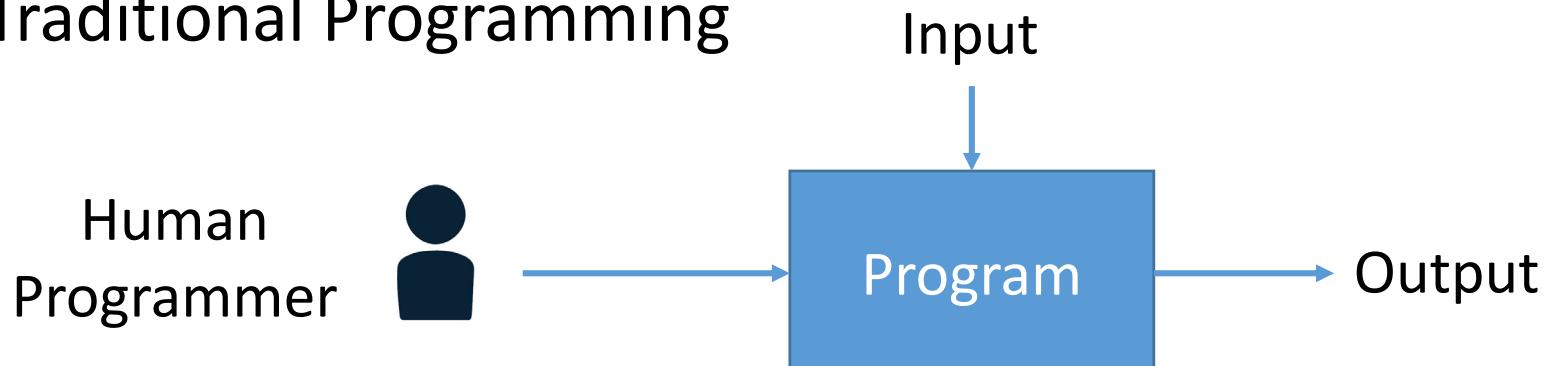
# What is Machine Learning

A more mathematical definition by Tom Mitchell

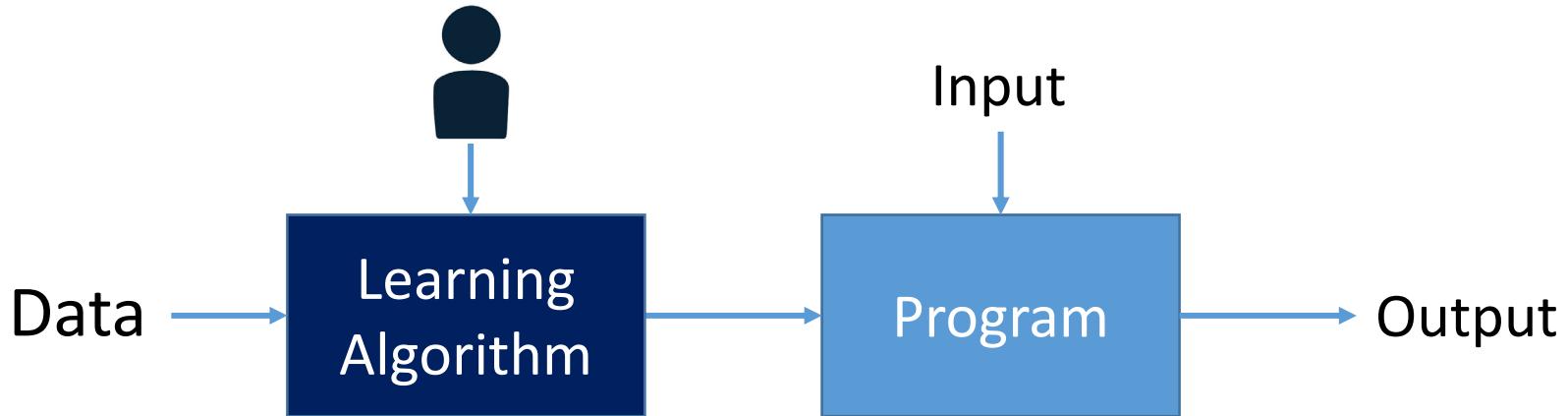
- Machine learning is the study of algorithms that
  - improve their performance  $P$
  - at some task  $T$
  - based on experience  $E$
  - with non-explicit programming
- A well-defined learning task is given by  $\langle P, T, E \rangle$

# Programming vs. Machine Learning

- Traditional Programming



- Machine Learning



# When does ML Make Advantages

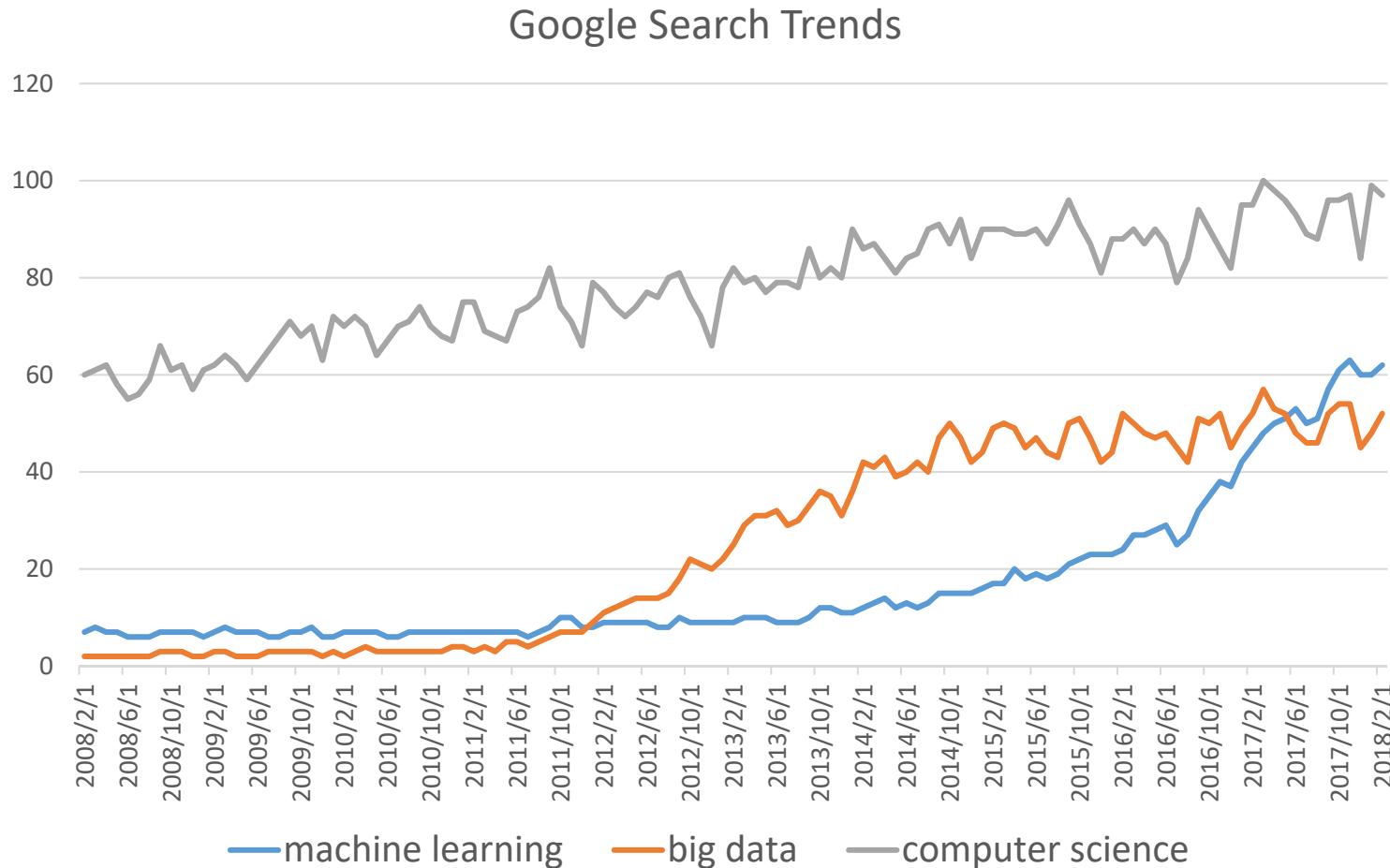
ML is used when

- Models are based on a huge amount of data
  - Examples: Google web search, Facebook news feed
- Output must be customized
  - Examples: News / item / ads recommendation
- Humans cannot explain the expertise
  - Examples: Speech / face recognition, game of Go
- Human expertise does not exist
  - Examples: Navigating on Mars

# Two Kinds of Machine Learning

- Prediction
  - Predict the desired output given the data (supervised learning)
  - Generate data instances (unsupervised learning)
- Decision Making
  - Take actions in a dynamic environment (reinforcement learning)
    - to transit to new states
    - to receive immediate reward
    - to maximize the accumulative reward over time

# Trends



<https://www.google.com/trends>

# Some ML Use Cases

# ML Use Case 1: Web Search

A screenshot of a Google search results page. The search query "shanghai jiao tong university" is entered in the search bar. Below the search bar, a dropdown menu shows several suggestions: "shanghai jiao tong university ranking", "shanghai jiao tong university international students", "shanghai jiao tong university school of medicine", and "shanghai jiao tong university admission". The last suggestion, "shanghai jiao tong university admission", is highlighted with a red rectangular box. Below the suggestions, the text "About 1,000,000 results (0.34 seconds)" is visible.

- Query suggestion

Scholarly articles for **shanghai jiao tong university**

[Shanghai Jiao Tong University](#) - Wang - Cited by 21

[Shanghai Jiao-Tong University](#) - Xue - Cited by 14

[Nanosheet-constructed porous TiO<sub>2</sub>-B for advanced ...](#) - Liu - Cited by 206

---

[Shanghai Jiao Tong University](#)

[en.sjtu.edu.cn/](#) ▾

Site Search. Home; About SJTU; Admission; Academics; Research; Join Us ... Antai College of SJTU

Rose to No.7 in 2016 Financial Times EMBA Ranking ...

[Programs in English](#) · Schools · Fall 2016 SJTU Graduate ... · Scholarships

[上海交通大学](#)

[www.sjtu.edu.cn/](#) ▾ [Translate this page](#)

全面介绍上海交通大学新闻的网站。

- Page ranking

[Shanghai Jiao Tong University - Wikipedia](#)

[https://en.wikipedia.org/wiki/Shanghai\\_Jiao\\_Tong\\_University](https://en.wikipedia.org/wiki/Shanghai_Jiao_Tong_University) ▾

Shanghai Jiao Tong University is a public research university located in Shanghai, China.

Established in 1896 by an imperial edict issued by the Guangxu ...

[Name](#) · [History](#) · [Academics, enrollment, and staff](#) · [Organization](#)

# ML Use Case 2: News Recommendation

美国大选 + 关注

## “特朗普时代”的中美新局

周浩：美国在经济上强势可能带来外交上相对弱势，美国可能给予中国在亚太更大的主动权，以为经济发展蓄势。



更新于2016年11月16日 07:07 德国商业银行首席中国经济师 周浩 为英国《金融时报》中文网撰稿

特朗普强势当选美国总统，给全市场留下了一个费解的难题：到底这位特立独行的美国白人会给世界带来怎样的变化，而未来世界格局中，中美两大经济体又将会以怎样的方式来进行互动。

到目前为止，我们只能通过特朗普在竞选过程中的讲话，部分了解未来美国政策的走向。比如说，特朗普反对TPP，认为目前的全球化策略并没有能够解决美国企业的困境，并表示要对中国商品征收45%的关税，同时要在美墨边境建造“长城”来防止非法移民。特朗普也反对美国目前的世界警察角色，认为这给美国普通家庭带来了负担和悲痛，这意味着美国在全球战略布局中将更多采取收缩策略。此外，特朗普认为美国的能源政策和医疗保险制度是个灾难，认为政府插手太多，造成了巨大的浪费。

- Predict whether a user will like a news given its reading context

### 您可能感兴趣的文章除了：



焦点与希望——选后华盛顿侧记

这是特朗普的1966年



特朗普能被政治精英驯服吗？



从特朗普胜选看美国政治

# ML Use Case 3: Online Advertising

Google search results for "iphone 6s case".

Web Shopping News Images Videos More ▾ Search tools

About 16,900,000 results (0.33 seconds)

**iPhone 6s Cases - case-mate.com**  
Ad [www.case-mate.com/iPhone-6s-Cases](http://www.case-mate.com/iPhone-6s-Cases) ▾  
4.6 ★★★★☆ rating for case-mate.com  
Shop The iPhone 6s Case Collection. Free Standard Shipping!  
Refined Protection · Slim & Tough · Genuinely Crafted · Premium Designs

**iPhone 6s**  
Ad [www.apple.com/](http://www.apple.com/) ▾  
The only thing that's changed is everything. Learn more.  
A9 chip · Two sizes · Now in rose gold  
Pre-order 9.12 · iPhone Upgrade Program · 3D Touch · Cameras

Shop for iphone 6s case on Google Sponsored

 Case-mate - Karat Case Fo... \$49.99 Best Buy ★★★★★ (163)	 Moshi - Iglae Armour Case... \$39.99 Best Buy ★★★★★ (161)	 Logitech - Protection... \$21.99 Best Buy ★★★★★ (90)	 Moshi - Overture Wall... \$49.99 Best Buy ★★★★★ (18)
 Case-mate - Brilliance Cas... \$44.99 Best Buy ★★★★★ (294)	 Case-mate - Wallet Folio C... \$54.99 Best Buy ★★★★★ (173)	 Marc by Marc Jacobs Metall... \$38.00 shopbop ★★★★★ (1)	 Case-mate - Karat Hard Sh... \$49.99 Best Buy ★★★★★ (34)

In the news

[Speck's iPhone 6s CandyShell + MightyShell cases bring best-of-breed protection to Apple's latest iPhones](#)

9 to 5 Mac - 1 day ago

With the iPhone 6s and iPhone 6s Plus debuting next week, it's important to start thinking ...

[Moshi's iPhone 6s and 6s Plus cases offer premium protection](#)  
iMore - 23 hours ago

[Top 5 Best Leather iPhone 6s Cases](#)  
Heavy.com - 12 hours ago

[More news for iphone 6s case](#)

[iPhone 6s Cases & Covers from OtterBox](#)

- Whether the user likes the ads
- How advertisers set bid price

# ML Use Case 3: Online Advertising

## 大陆



### 河南省公安厅彻查“封丘36人入警 35人身份不合规”

中封丘县公安局的36名受训人员，35人是公安局内部的文职或临时人员，与“民警必须具备公务员身份”的国家规定不符，引发该局内部

- 上海至成都沿江高铁提上日程 串联长江沿线22城市
- 2016号歼-20原型机曝光 已滑行测试(图)
- 日媒：中国或派万吨海警船巡钓鱼岛 打消耗战
- 外媒：中国开始研制隐身武装直升机 预计2020年交付
- 习近平关于中美关系的十个判断
- 住建部黑臭水沟整治工作指南：9成百姓满意才能达标
- 陕西：职校“校长”让女学生陪酒 学校被撤除
- 揭秘“团团伙伙”的武钢漩涡和落马高管

## 国际



### 巴塞罗那200万人游行 呼吁加泰罗尼亚独立(图)

- 李伟光：收税是不公平的恶？
- 许章润：超级大国没有纯粹内政
- 刘昀献：国外政党联系群众的路径研究

## 时局观



民革中央副主席：中共从未否定国民党抗战作用

- 施芝鸿：文革基础上搞改革致一个时期市场官场乱象
- 朱维群回应争议：尊重民族差异而不强化
- 伊协副会长：穆斯林不应因宗教功修忽视社会责任

## 领袖圈



奥巴马54岁啦，当7年总统人苍老了头发也白了



海绵城市 未来之城

水危机：青岛告急

探访中国绿化博览会

帝都吸引华人首富

凤凰房产 诚邀加盟

谈华山论剑与中国精神

黑龙江创新驱动三步棋

《印记》之江城夜未眠

办公环境搜查令

圈层生活尽在凤凰会

## 精彩视频

凤凰联播台



菲媒曝菲律宾军演针对中国 直指南海生命线

播放数：2602282

- Whether the user likes the ads
- How advertisers set bid price

# ML Use Case 4: Information Extraction

## Kinect - Fastest Selling Electronic Product in History

Posted on: 3/10/2011 1:09:45 PM by David Lewis

Microsoft's Kinect sensor system has been officially recognised as the fastest selling electrical device in history.



Manufactured to give wireless interactivity with the company's Xbox game platform, the device has sold eight million units in its first two months, outstripping the sales of Apple's iPhone and iPad when they were launched.

The news comes as a welcome relief for Microsoft who have been trailing Apple in the technology stakes over the last few years with the Apple brand being seen as more cool and sexy than Microsoft.

The figures, which have been verified by the Guinness Book of World Records, represent

sales of the camera add-on which uses infrared technology to track the movement of the participant and translate their movements to action in the game.

For some time Microsoft's Xbox was at a disadvantage to Nintendo's Wii system because of the lack of a motion detector but the Kinect addresses the issue well. Microsoft were keen on using a different technological base for their system to avoid being accused of copyright infringement and so the solution was built around infrared technology.

Microsoft says that sales of the Kinect reflect the popularity of the games platform in comparison with the Wii and hope that the availability of Kinect will also boost sales of the Xbox itself.

It notes that sales of games for the Xbox have also rocketed since the device became available with total sales now exceeding ten million.

In January Microsoft reported profits of \$6.63bn (£4.1bn) for the last three months of 2010, down from \$6.66bn a year earlier despite the excellent sales performance of Kinect.

Posted: 3/10/2011 1:09:45 PM by David Lewis | with 0 comments

Webpage



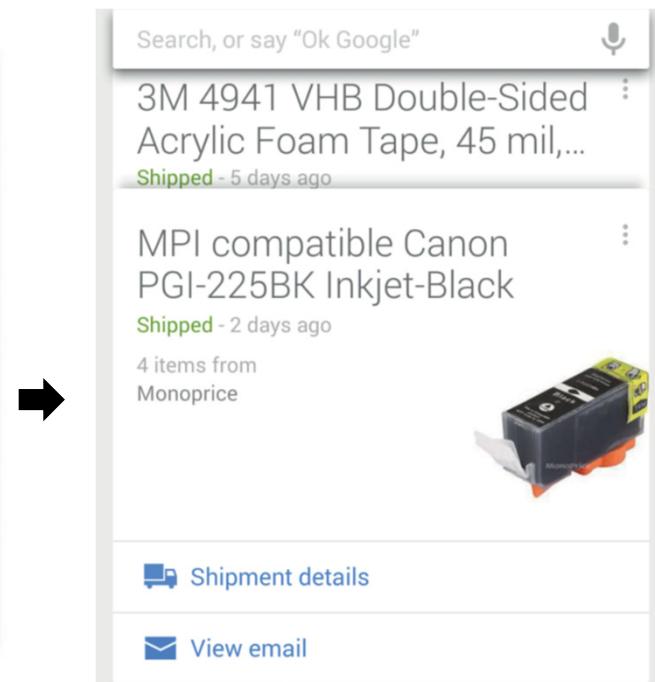
Kinect  
Electronic Product  
Microsoft's Xbox  
Games  
Xbox Game Platform

- 
- 
- 

Keywords

# ML Use Case 4: Information Extraction

- Structural information extraction and illustration



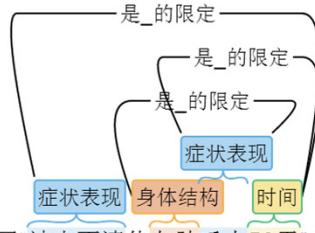
Gmail

Google Now

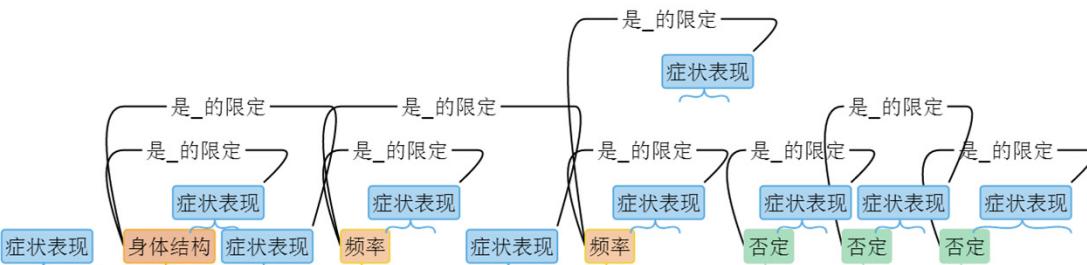
# ML Use Case 4: Information Extraction

- Clinical medicine structural information extraction

出院记录



入院情况：因“神志不清伴左肢乏力50天”入院。



出院情况：患者对言语无反应，全身消瘦衰竭状，偶有睁眼及眼球活动，偶有咳嗽咳痰，无发热，无呕吐，无肢体抽搐，



出院诊断：1.脑梗死2.高血压病3.肺部感染4.心律失常5.心功能IV级6.重度营养不良



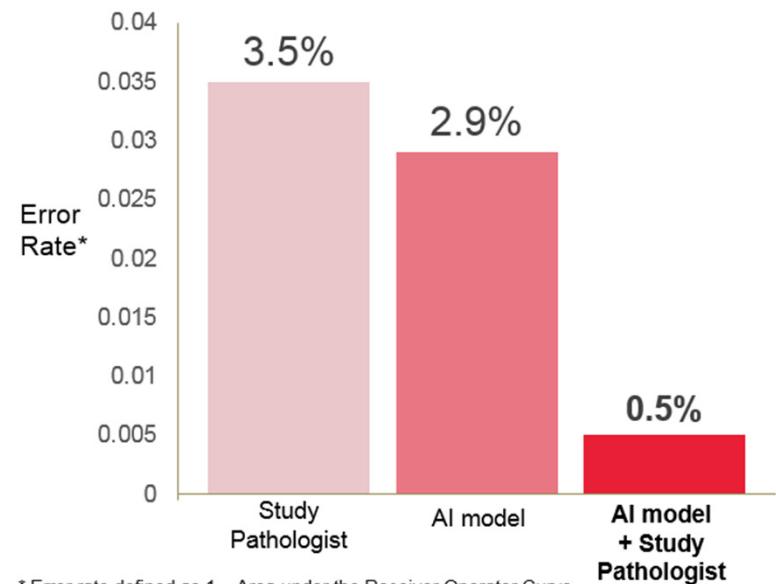
诊治经过：完善相关检查,予吸氧,抗血小板聚集,保护脑细胞,营养神经,保护胃黏膜,改善脑循环及补液对症支持治疗。

# ML Use Case 5: Medical Image Analysis

- Breast Cancer Diagnoses



(AI + Pathologist) > Pathologist



\* Error rate defined as  $1 - \text{Area under the Receiver Operator Curve}$

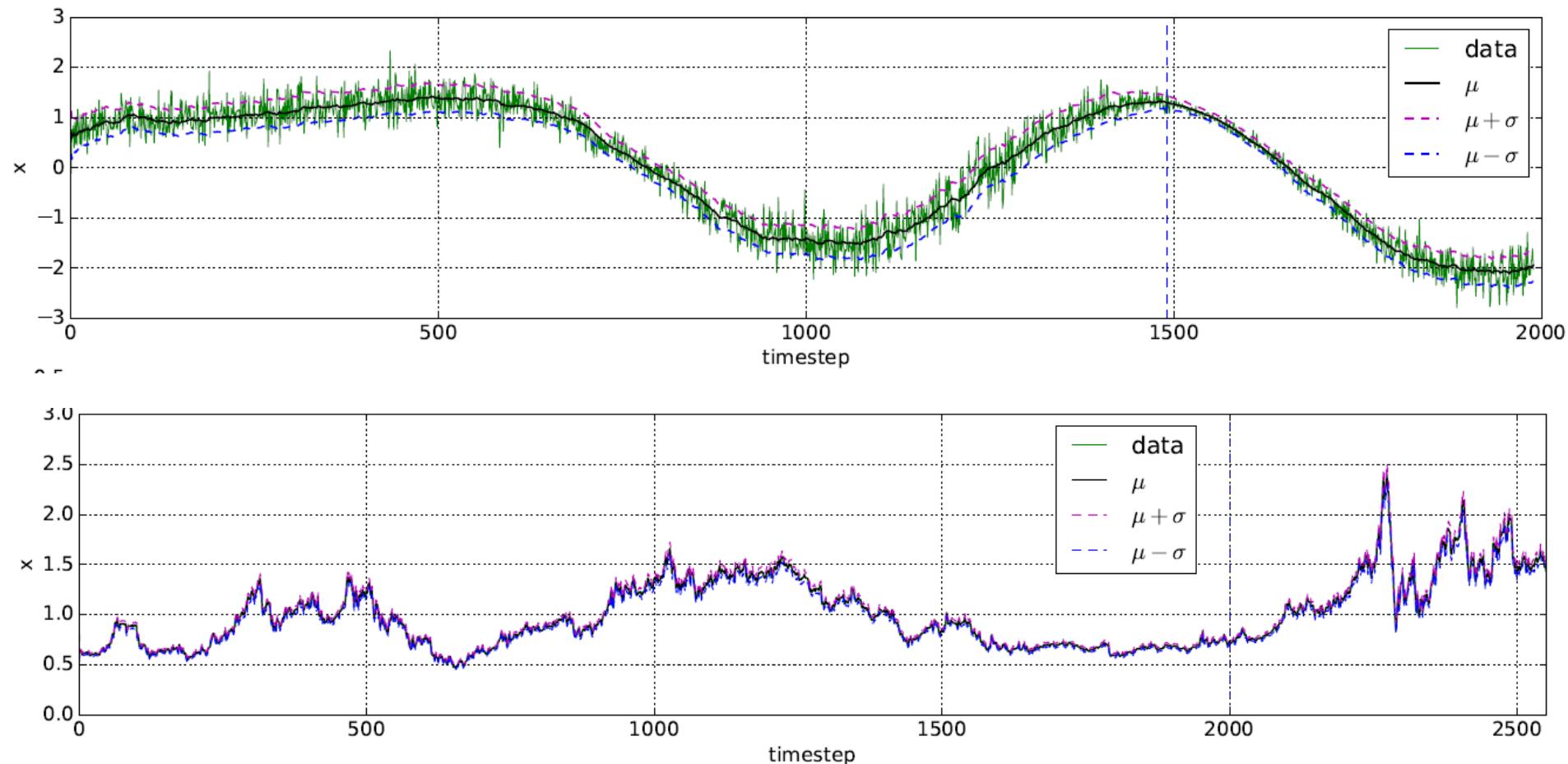
\*\* A study pathologist, blinded to the ground truth diagnoses, independently scored all evaluation slides.

© 2016 PathAI

Wang, Dayong, et al. "Deep learning for identifying metastatic breast cancer." arXiv preprint arXiv:1606.05718 (2016).  
<https://blogs.nvidia.com/blog/2016/09/19/deep-learning-breast-cancer-diagnosis/>

# ML Use Case 6: Financial Data Prediction

- Predict the trend and volatility of financial data



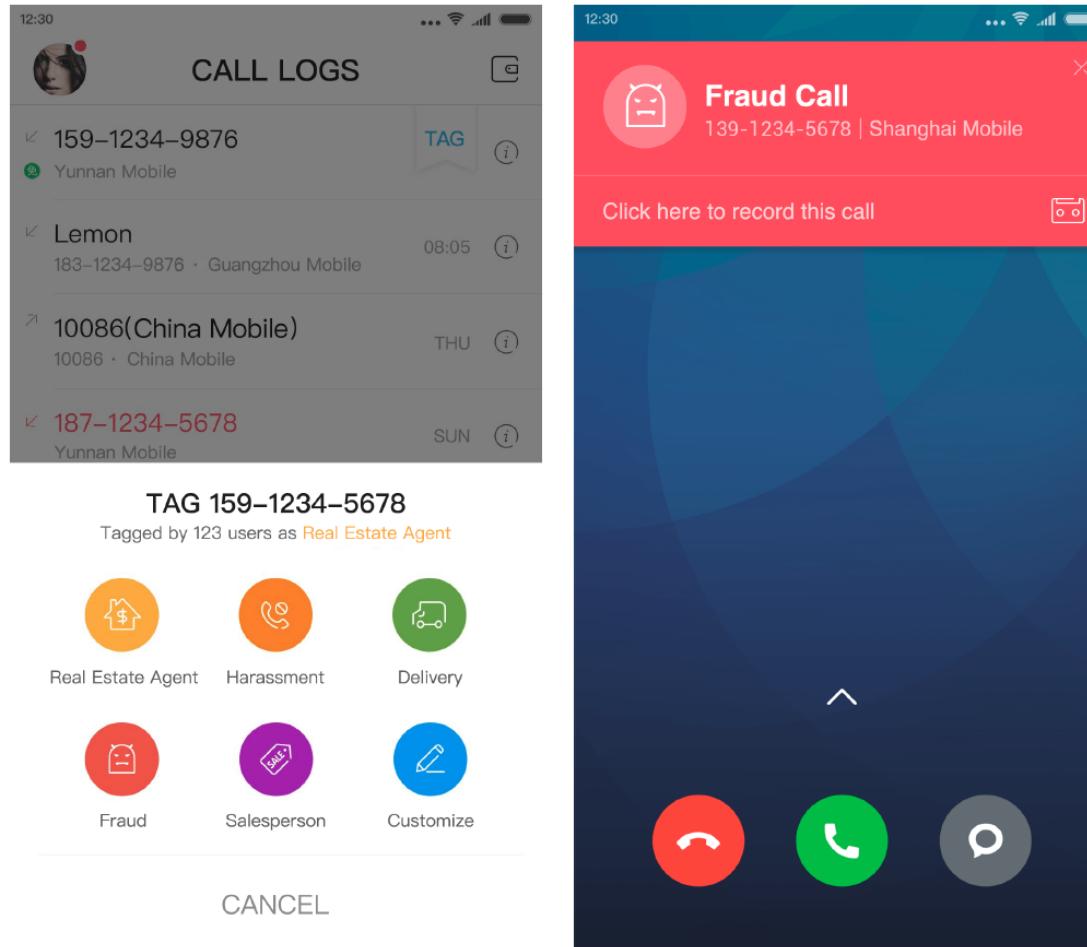
# ML Use Case 7: Social Networks

- Friends/Tweets/Job Candidates suggestion



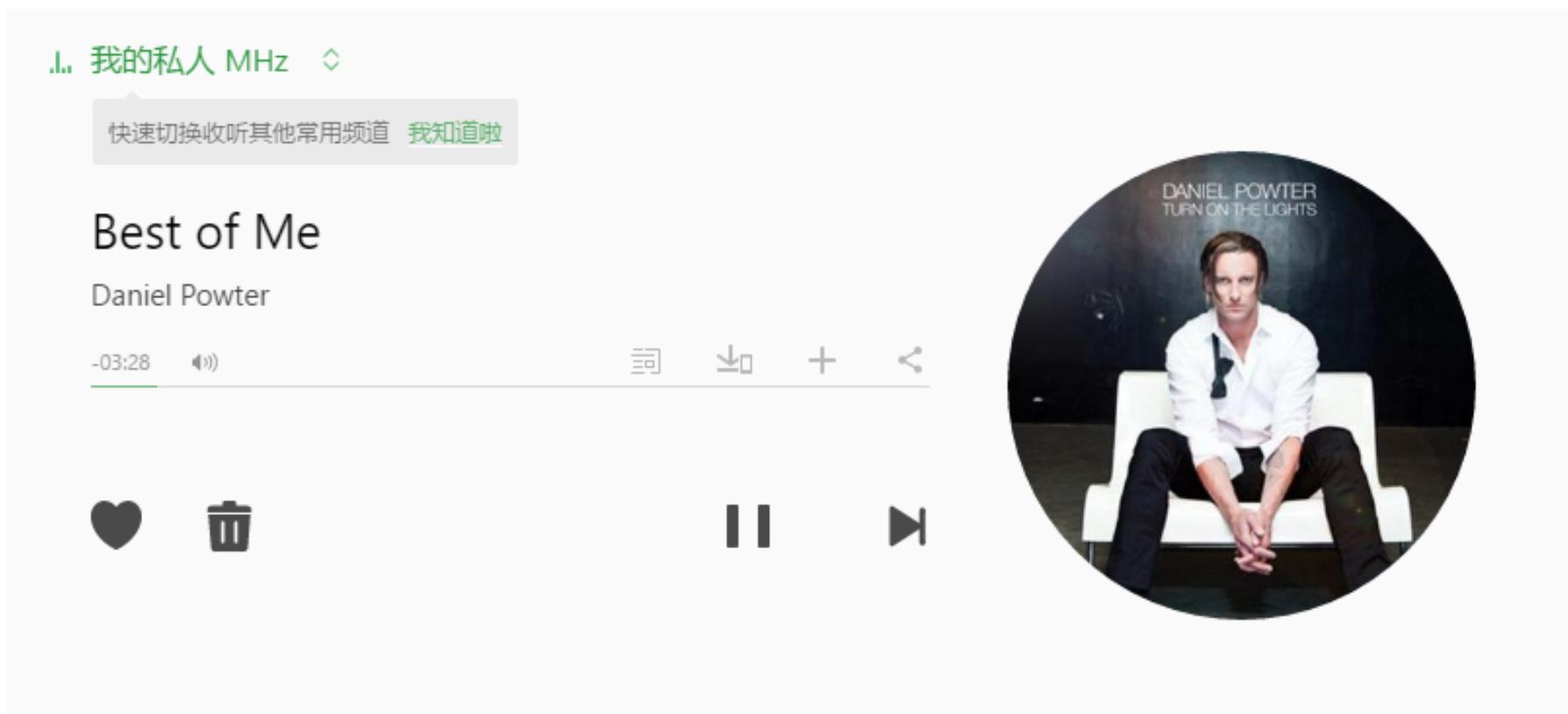
# ML Use Case 8: Anomaly Detection

- Detect malicious calls



# ML Use Case 9: Interactive Recommendation

- Douban.fm music recommend and feedback
  - The machine needs to make decisions, not just prediction



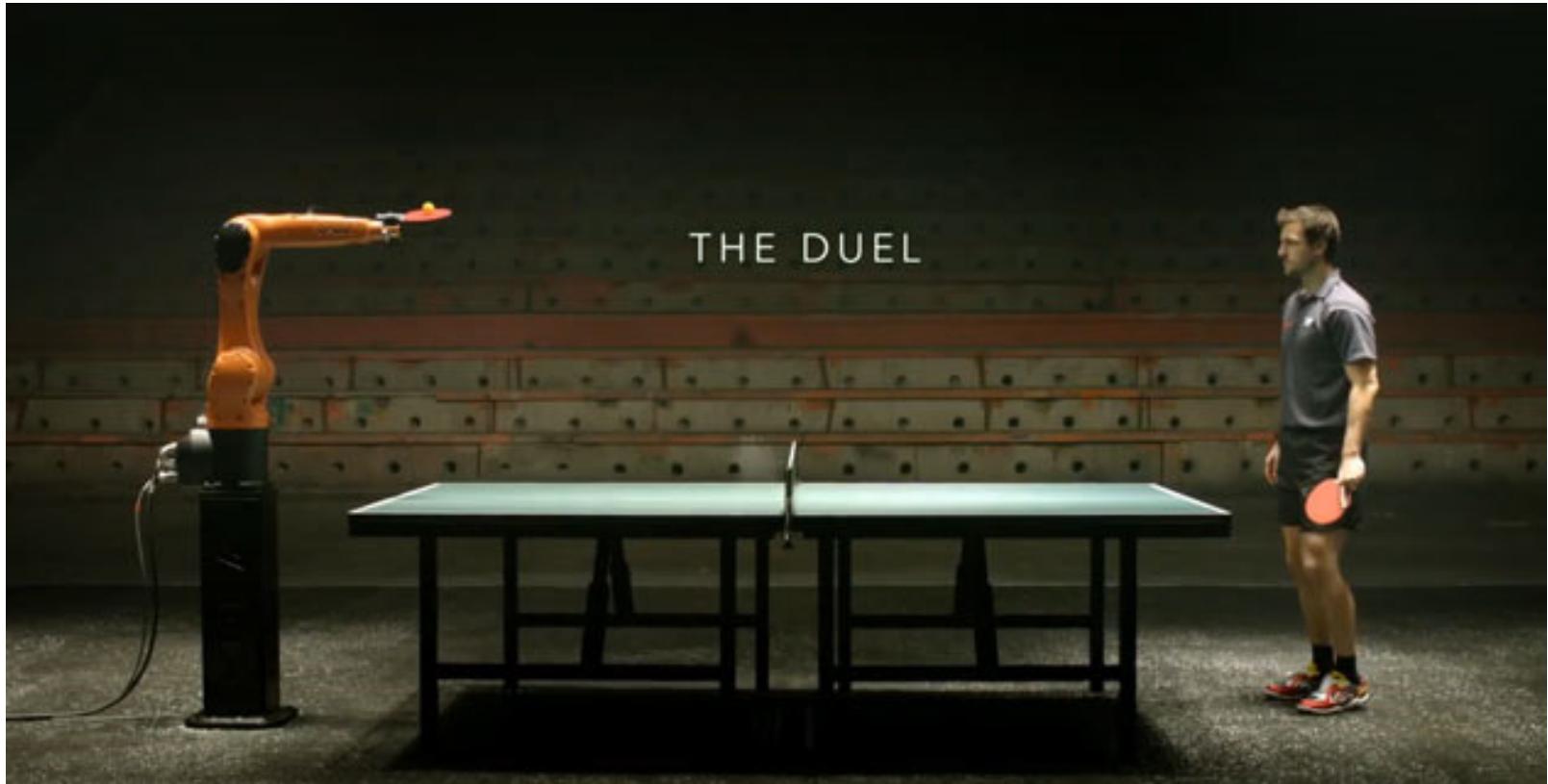
# ML Use Case 10: Robotics Control

- Stanford Autonomous Helicopter
  - <http://heli.stanford.edu/>



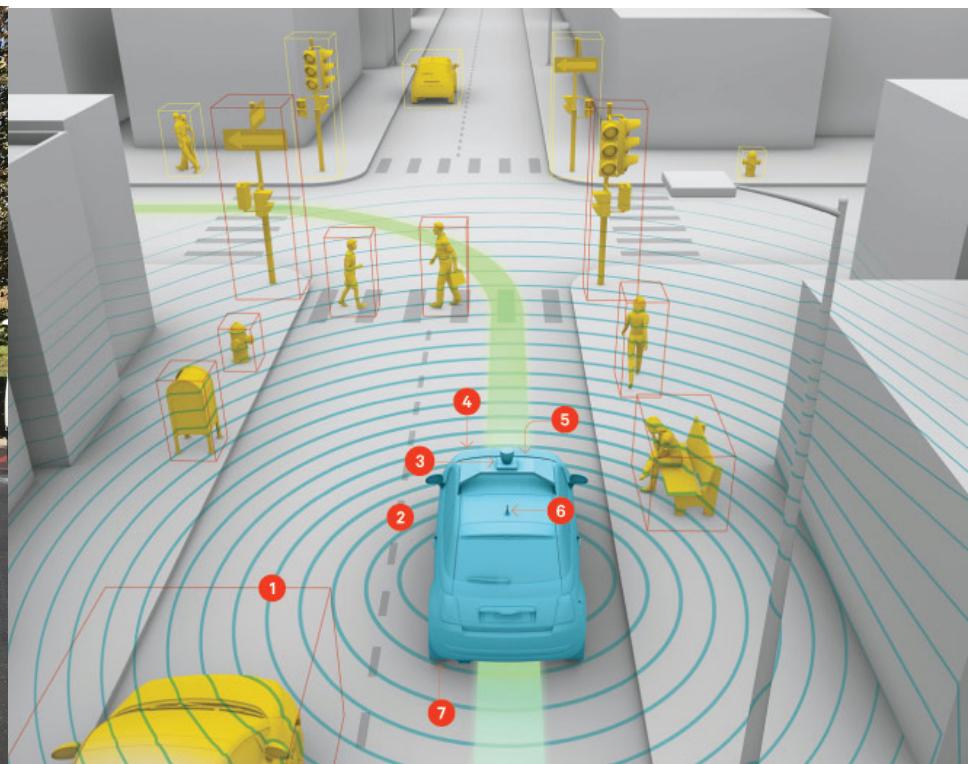
# ML Use Case 10: Robotics Control

- Ping pong robot
  - <https://www.youtube.com/watch?v=tIIJME8-au8>



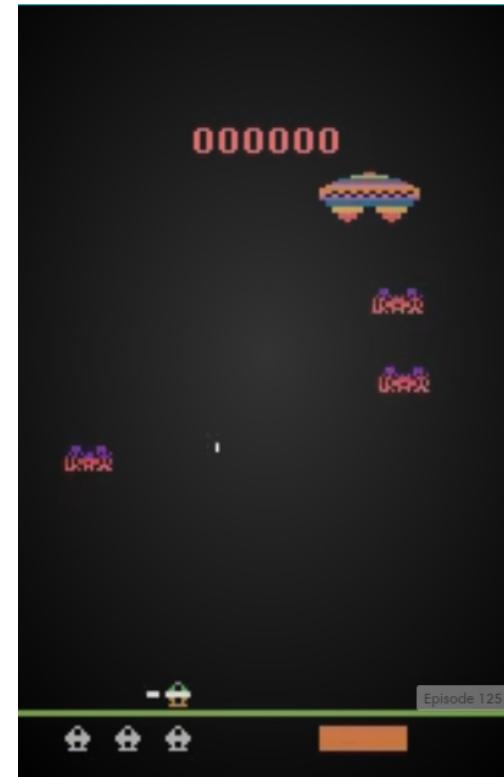
# ML Use Case 11: Self-Driving Cars

- Google Self-Driving Cars
  - <https://www.google.com/selfdrivingcar/>



# ML Use Case 12: Game Playing

- Take actions given screen pixels
  - <https://gym.openai.com/envs#atari>



# ML Use Case 13: AlphaGo



IBM Deep Blue (1996)

- 4-2 Garry Kasparov on Chess
- A large number of crafted rules
- Huge space search



Google AlphaGo (2016)

- 4-1 Lee Sedol on Go
- Deep machine learning on big data

Silver, David, et al. "Mastering the game of Go with deep neural networks and tree search." *Nature* 529.7587 (2016): 484-489.

# ML Use Case 14: Text Generation

- Making decision of selecting the next word/char
- Chinese poem example. Can you distinguish?

南陌春风早，东邻去日斜。

山夜有雪寒，桂里逢客时。

紫陌追随日，青门相见时。

此时人且饮，酒愁一节梦。

胡风不开花，四气多作雪。

四面客归路，桂花开青竹。

Human

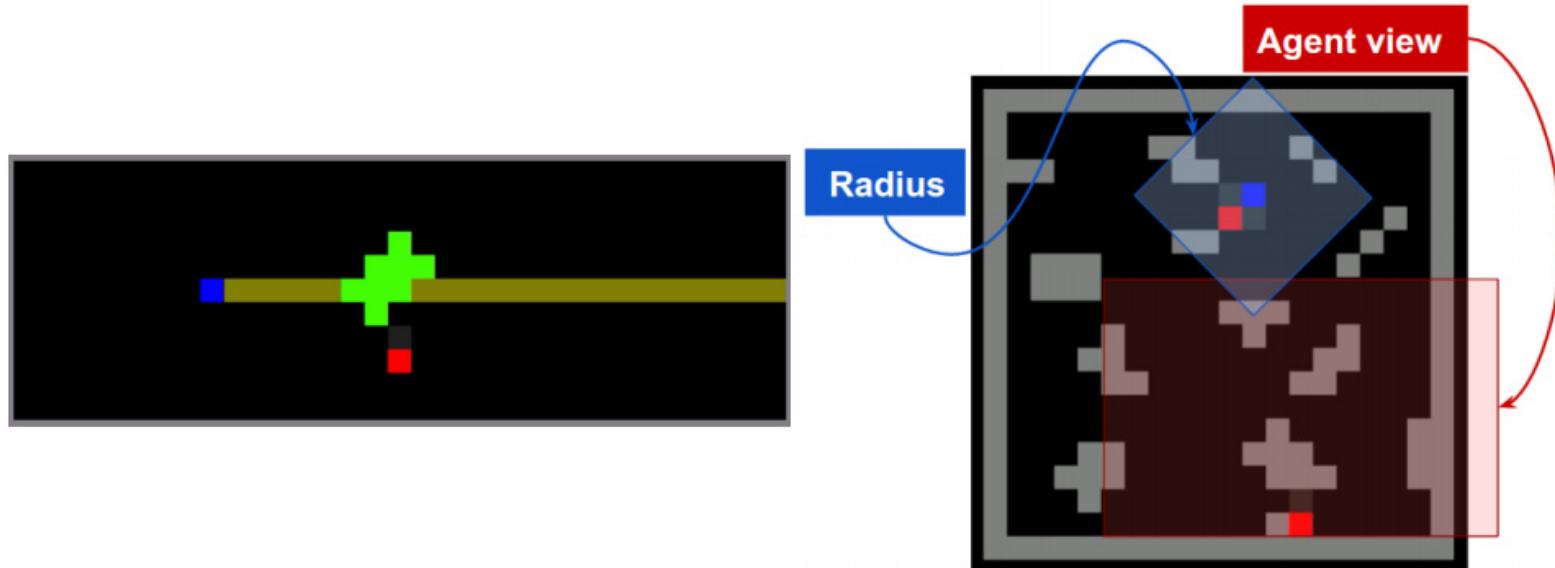
Machine

Jiaxian Guo, Sidi Lu, Weinan Zhang et al. Long Text Generation via Adversarial Training with Leaked Information. AAAI 2018.

Lantao Yu, Weinan Zhang, et al. Seqgan: sequence generative adversarial nets with policy gradient. AAAI 2017.

# ML Use Case 15: Multi-Agent Game Playing

- Multi-agent game playing
  - Learning to cooperate and compete



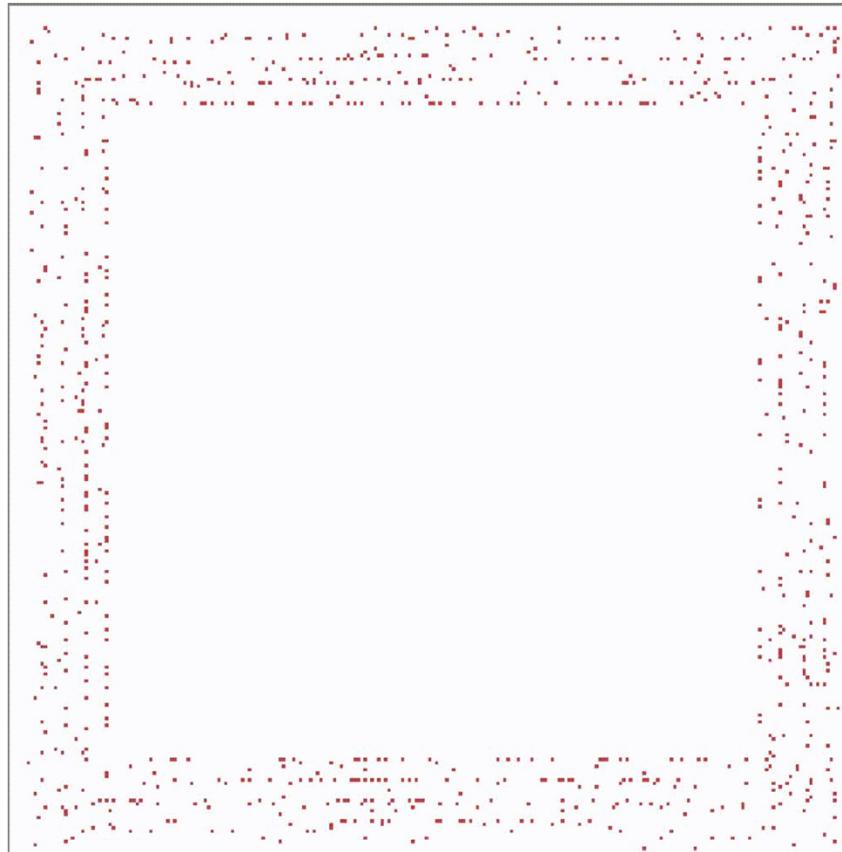
# ML Use Case 15: Multi-Agent Game Playing

- Multi-agent game playing
  - Learning to cooperate and compete



# ML Use Case 16: Many-Agent Interactions

- MAgent game: aligning



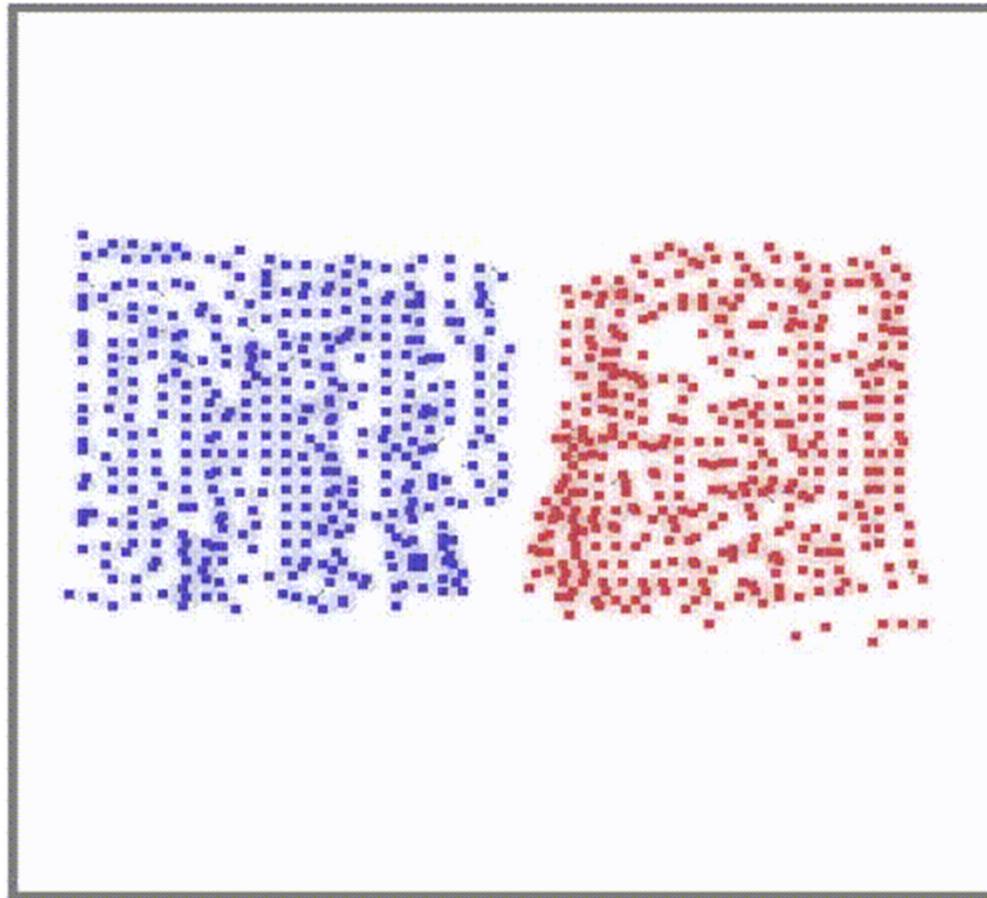
# ML Use Case 16: Many-Agent Interactions

- MAgent game: city simulation



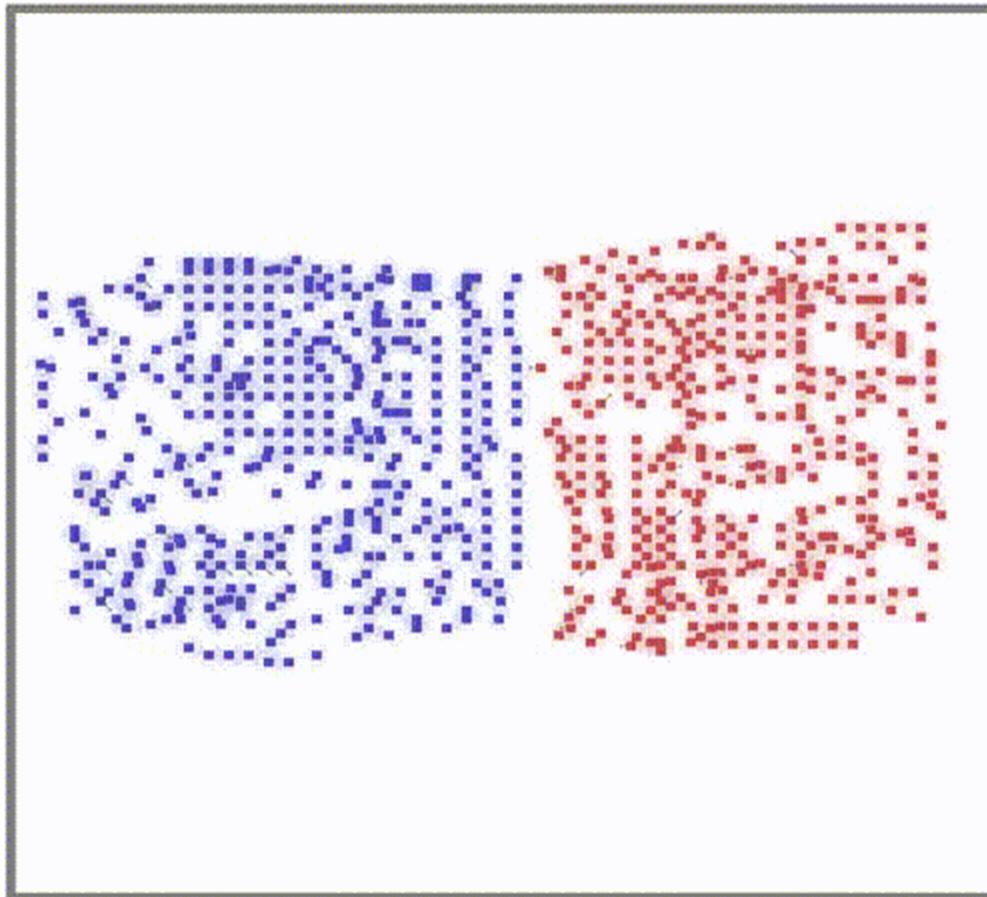
# ML Use Case 16: Many-Agent Interactions

- MAgent game: battle



# ML Use Case 16: Many-Agent Interactions

- MAgent game: battle



# History of Machine Learning

- 1950s
  - Samuel's checker player
  - Machine learning term created
- 1960s:
  - Neural networks: Perceptron
  - Pattern recognition
  - Minsky and Papert prove limitations of Perceptron
- 1970s:
  - Symbolic concept induction
  - Winston's arch learner
  - Expert systems and the knowledge acquisition bottleneck
  - Quinlan's ID3
  - Mathematical discovery with AM



Arthur Samuel coined the term “machine learning” in 1959

# History of Machine Learning

- 1980s:
  - Advanced decision tree and rule learning
  - Explanation-based Learning (EBL)
  - Learning and planning and problem solving
  - Utility problem
  - Analogy
  - Cognitive architectures
  - Resurgence of neural networks (connectionism, backpropagation)
  - Valiant's PAC Learning Theory
  - Focus on experimental methodology
- 1990s
  - Data mining
  - Adaptive software agents and web applications
  - Text learning
  - Reinforcement learning (RL)
  - Inductive Logic Programming (ILP)
  - Ensembles: Bagging, Boosting, and Stacking
  - Bayes Net learning
  - Support vector machines
  - Kernel methods

# History of Machine Learning

- 2000s
  - Graphical models
  - Variational inference
  - Statistical relational learning
  - Transfer learning
  - Sequence labeling
  - Collective classification and structured outputs
  - Computer systems applications
    - Compilers
    - Debugging
    - Graphics
    - Security (intrusion, virus, and worm detection)
  - Email management
  - Personalized assistants that learn
  - Learning in robotics and vision

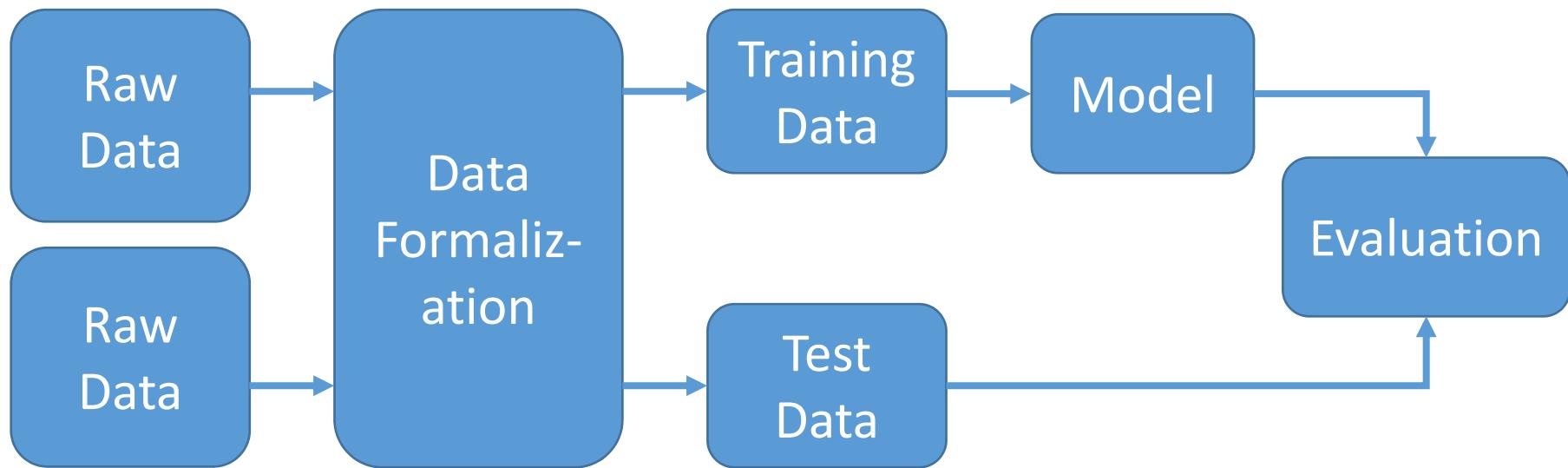
# History of Machine Learning

- 2010s
  - Deep learning
  - Learning from big data
  - Learning with GPUs or HPC
  - Multi-task & lifelong learning
  - Deep reinforcement learning
  - Massive applications to vision, speech, text, networks, behavior etc.
  - ...

# Machine Learning Categories

- Supervised Learning
  - To perform the desired output given the data and labels
- Unsupervised Learning
  - To analyze and make use of the underlying data patterns/structures
- Reinforcement Learning
  - To learn a policy of taking actions in a dynamic environment and acquire rewards

# Machine Learning Process



- Basic assumption: there exist the same patterns across training and test data

# Supervised Learning

- Given the training dataset of **(data, label)** pairs,

$$D = \{(x_i, y_i)\}_{i=1,2,\dots,N}$$

let the machine learn a function from **data** to **label**

$$y_i \simeq f_\theta(x_i)$$

- Function set  $\{f_\theta(\cdot)\}$  is called **hypothesis space**
- Learning is referred to as updating the **parameter**  $\theta$
- How to learn?
  - Update the parameter to make the prediction close to the corresponding label
    - What is the learning objective?
    - How to update the parameters?

# Learning Objective

- Make the prediction closed to the corresponding label

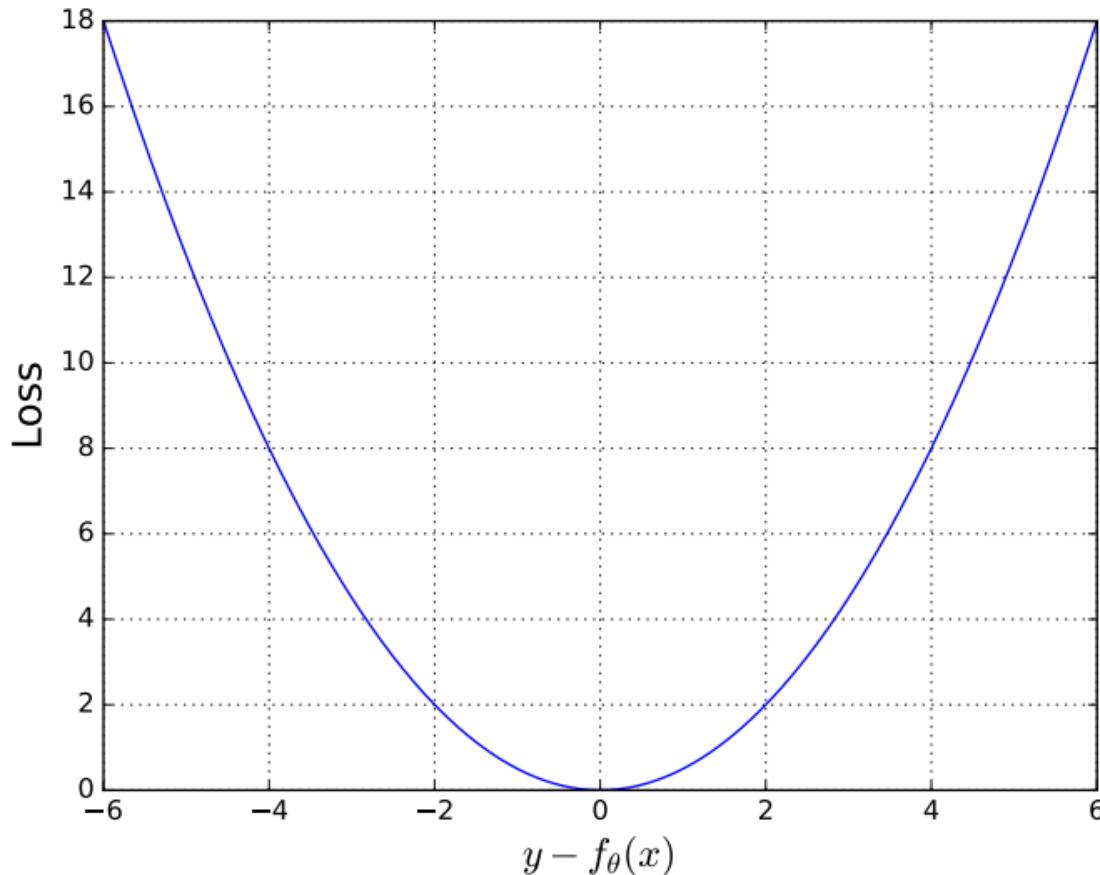
$$\min_{\theta} \frac{1}{N} \sum_{i=1}^N \mathcal{L}(y_i, f_{\theta}(x_i))$$

- Loss function  $\mathcal{L}(y_i, f_{\theta}(x_i))$  measures the error between the label and prediction
- The definition of loss function depends on the data and task
- Most popular loss function: squared loss

$$\mathcal{L}(y_i, f_{\theta}(x_i)) = \frac{1}{2}(y_i - f_{\theta}(x_i))^2$$

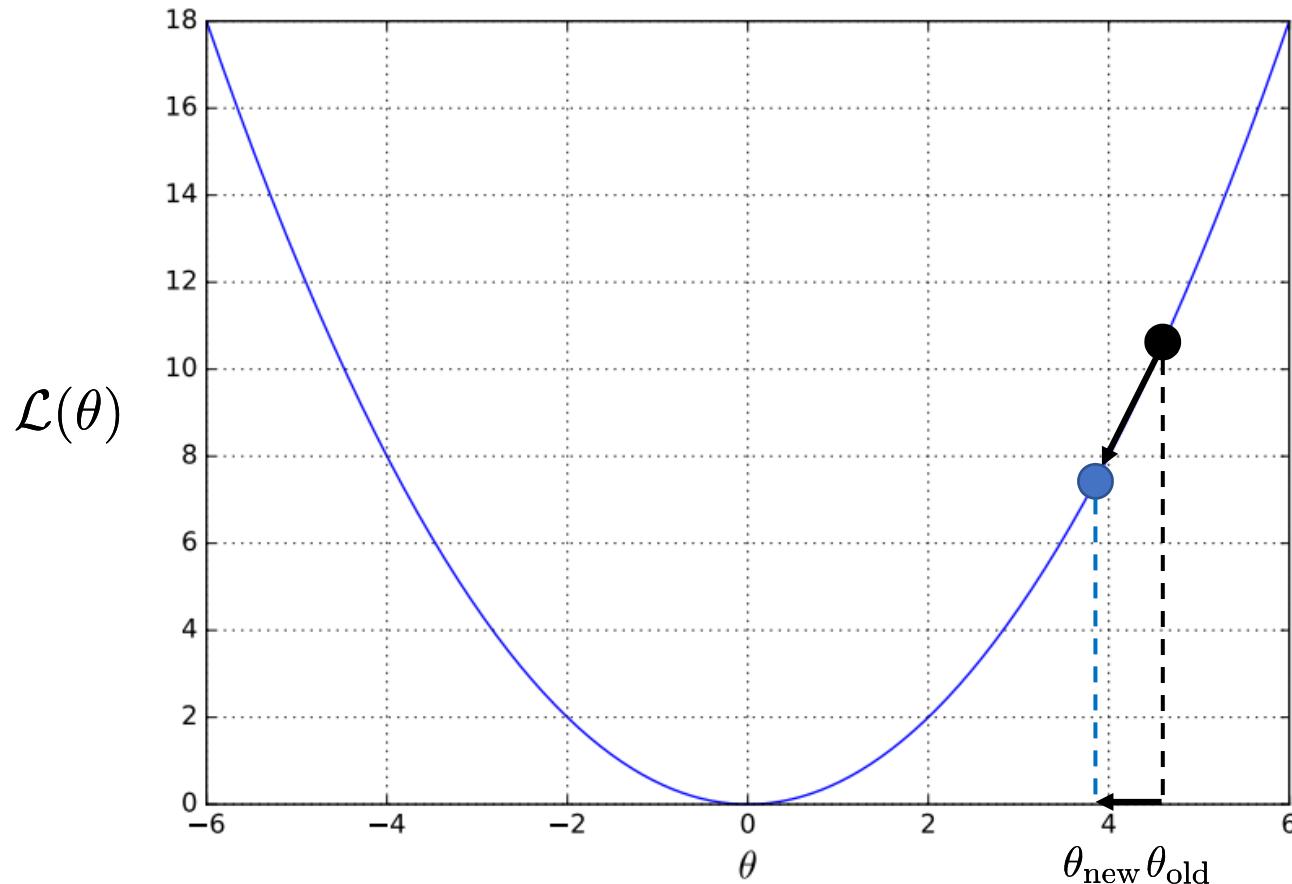
# Squared Loss

$$\mathcal{L}(y_i, f_\theta(x_i)) = \frac{1}{2}(y_i - f_\theta(x_i))^2$$



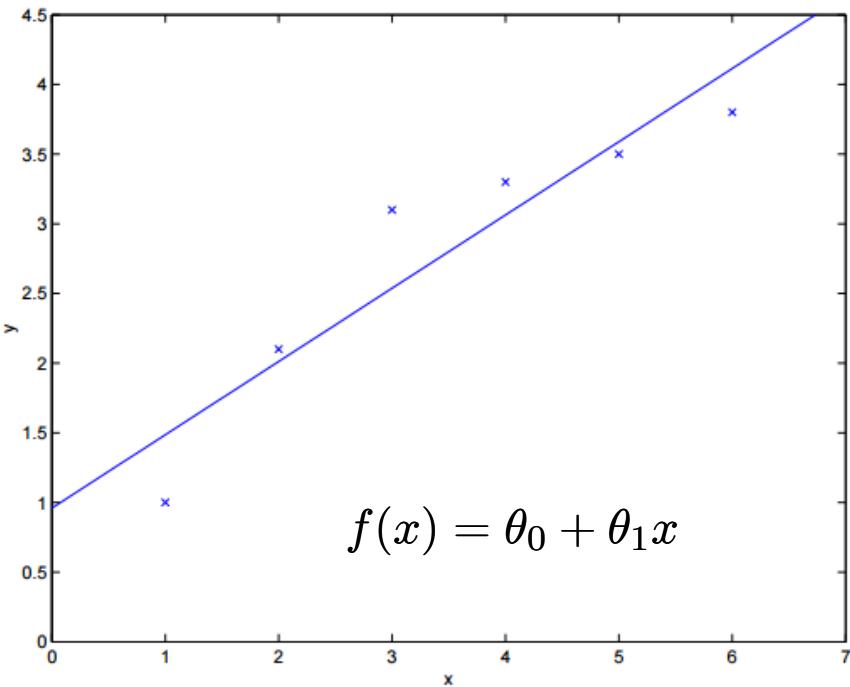
- Penalty much more on larger distances
- Accept small distance (error)
  - Observation noise etc.
  - Generalization

# Gradient Learning Methods

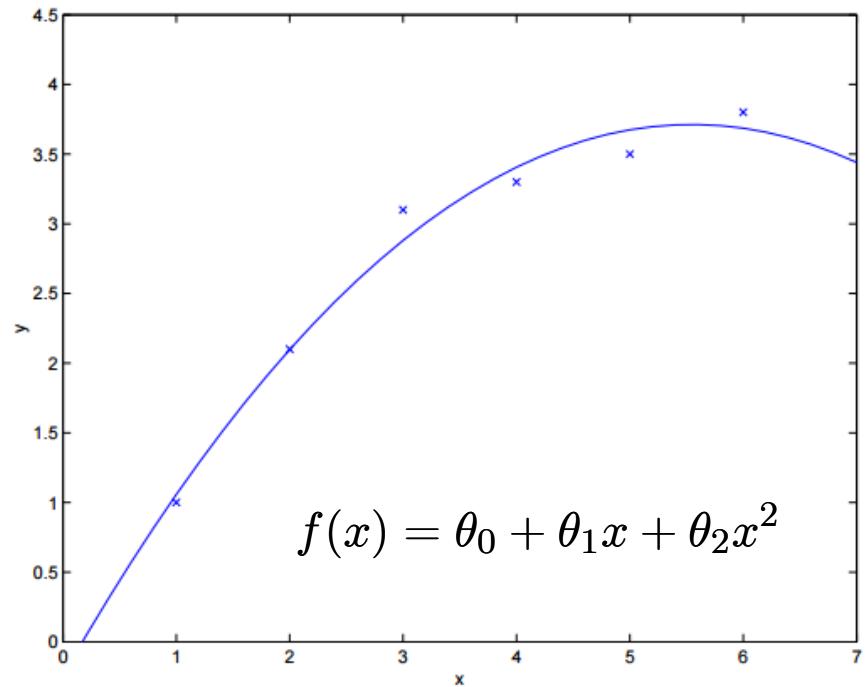


$$\theta_{\text{new}} \leftarrow \theta_{\text{old}} - \eta \frac{\partial \mathcal{L}(\theta)}{\partial \theta}$$

# A Simple Example



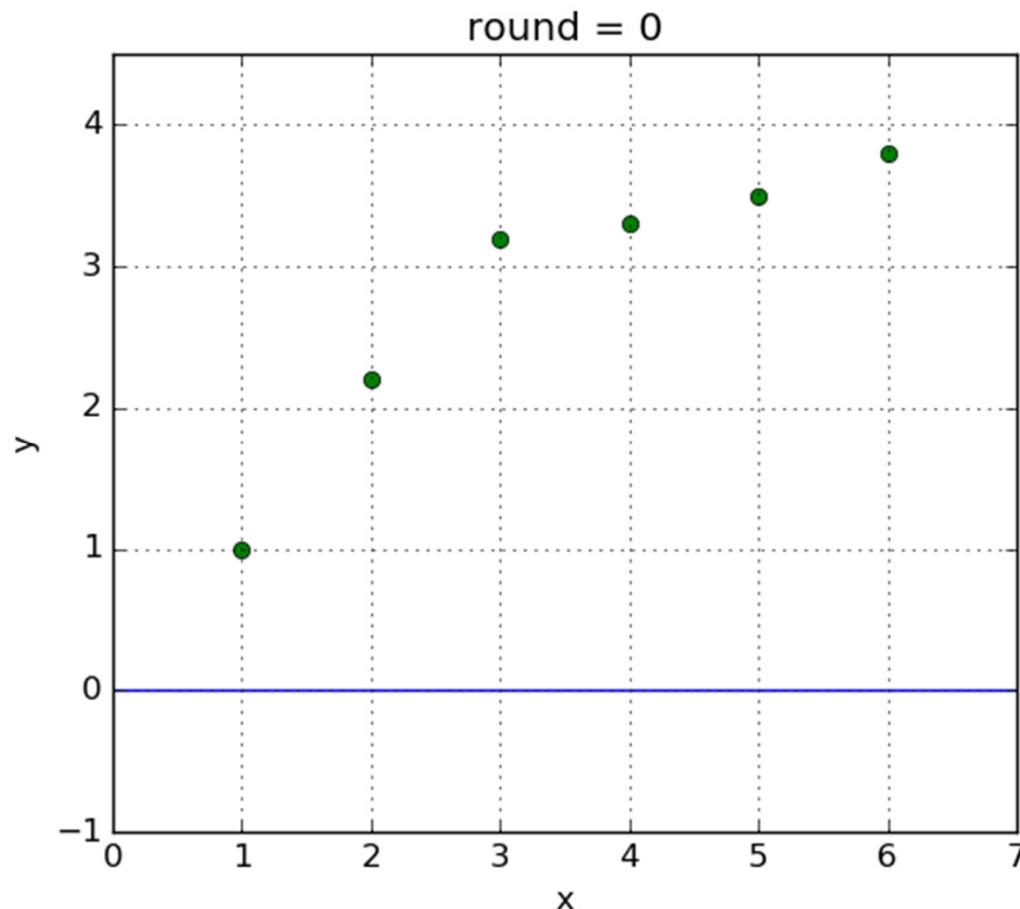
$$f(x) = \theta_0 + \theta_1 x$$



$$f(x) = \theta_0 + \theta_1 x + \theta_2 x^2$$

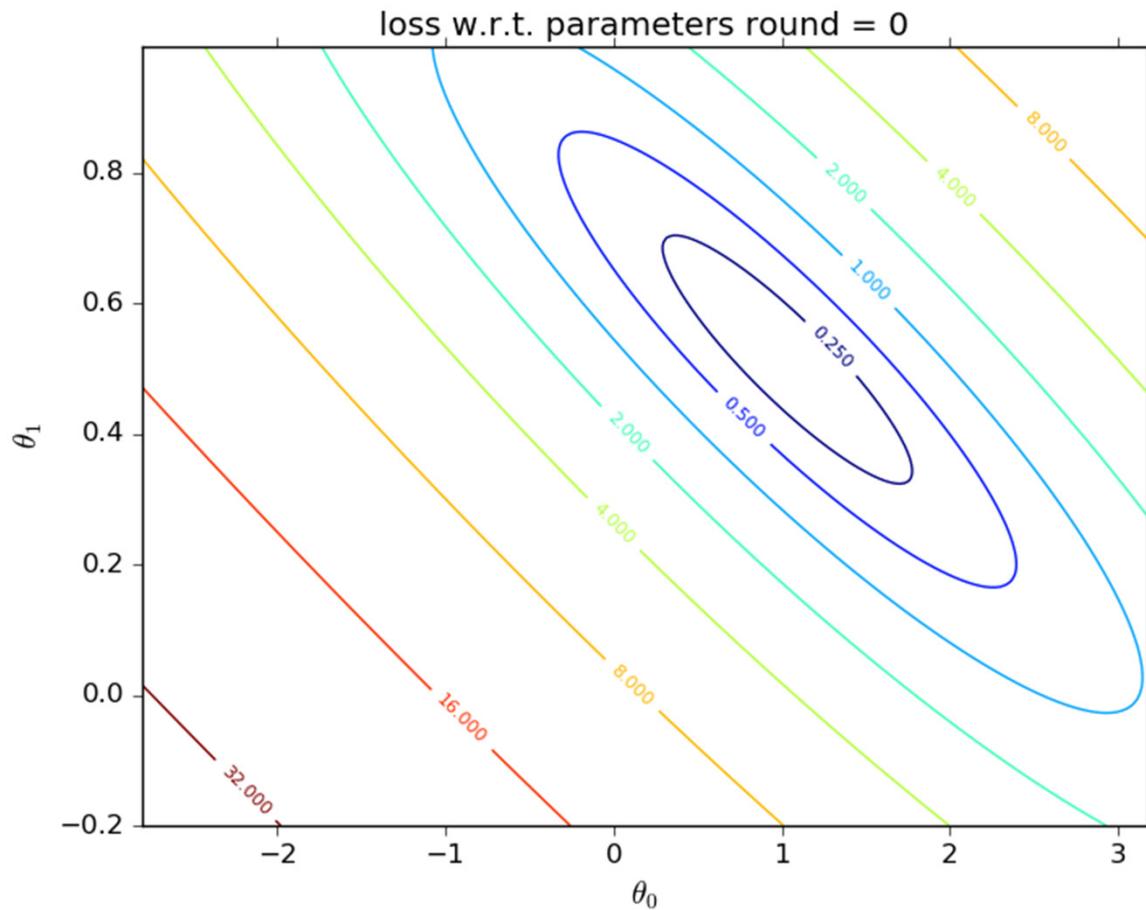
- Observing the data  $\{(x_i, y_i)\}_{i=1,2,\dots,N}$ , we can use different models (hypothesis spaces) to learn
  - First, model selection (linear or quadratic)
  - Then, learn the parameters

# Learning Linear Model - Curve



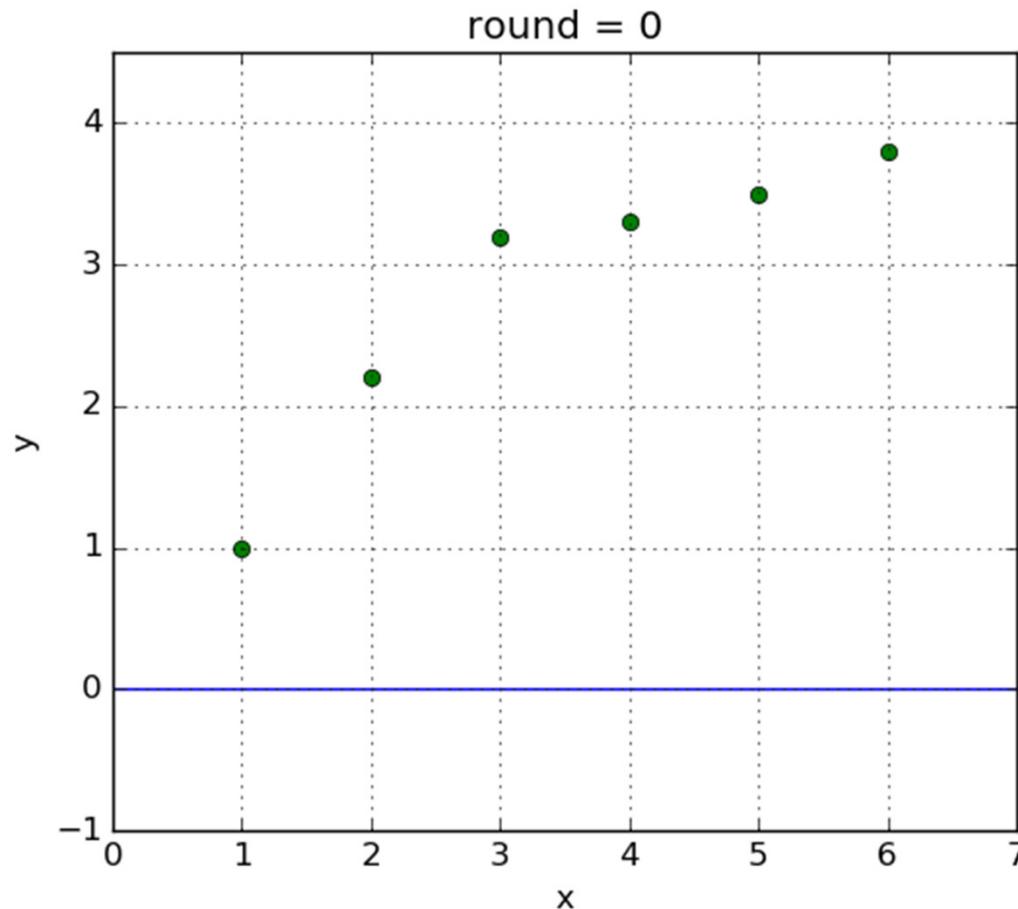
$$f(x) = \theta_0 + \theta_1 x$$

# Learning Linear Model - Weights



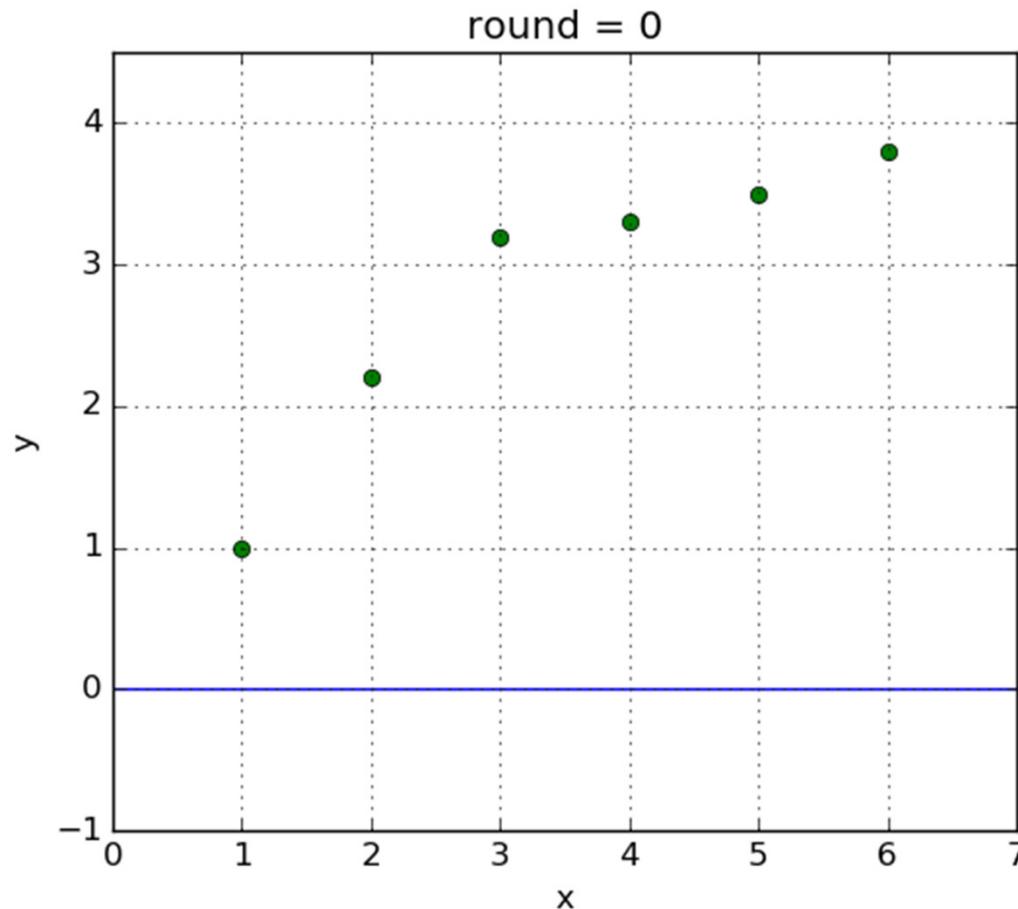
$$f(x) = \theta_0 + \theta_1 x$$

# Learning Quadratic Model



$$f(x) = \theta_0 + \theta_1 x + \theta_2 x^2$$

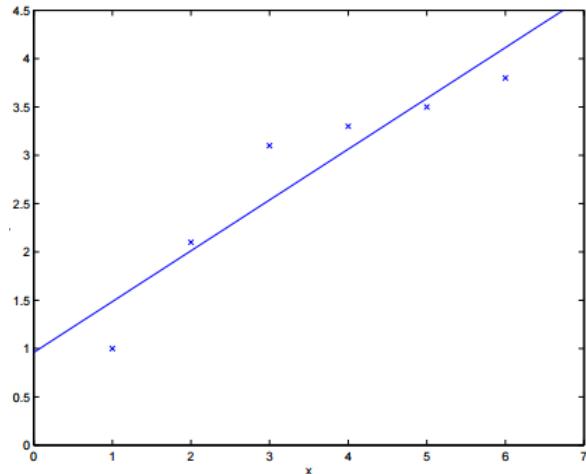
# Learning Cubic Model



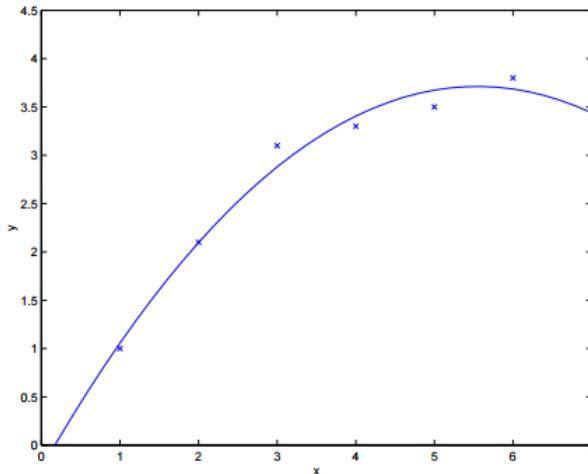
$$f(x) = \theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3$$

# Model Selection

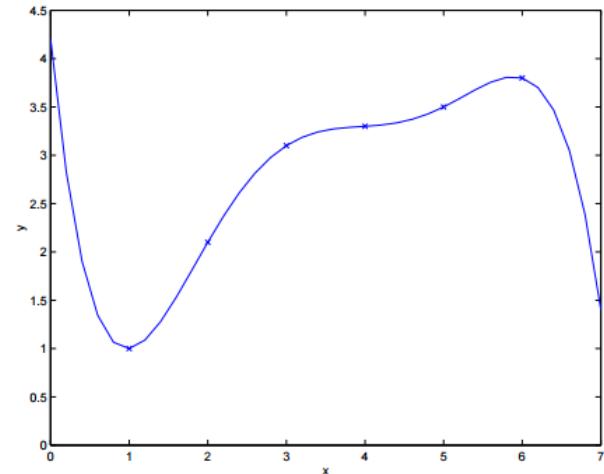
- Which model is the best?



Linear model: underfitting



Quadratic model: well fitting

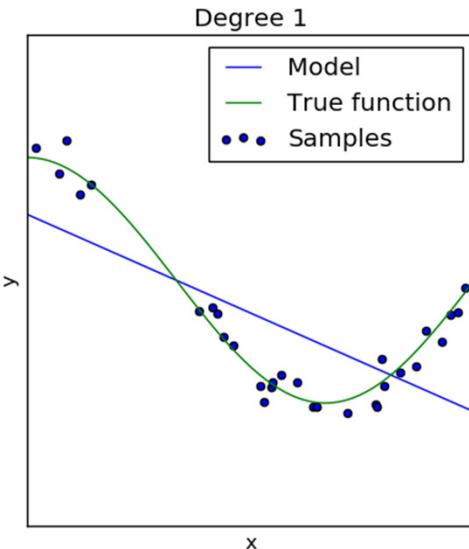


5<sup>th</sup>-order model: overfitting

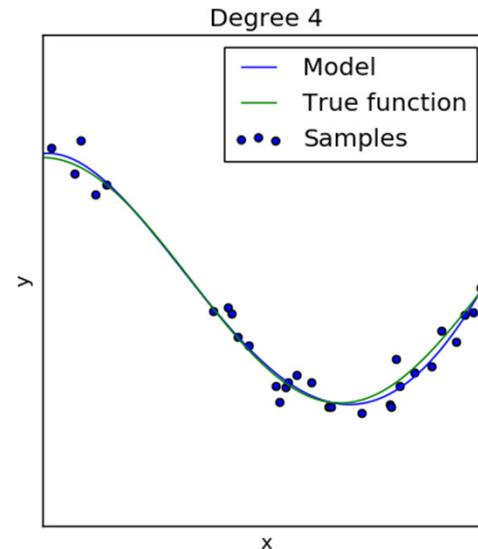
- Underfitting occurs when a statistical model or machine learning algorithm cannot capture the underlying trend of the data.
- Overfitting occurs when a statistical model describes random error or noise instead of the underlying relationship

# Model Selection

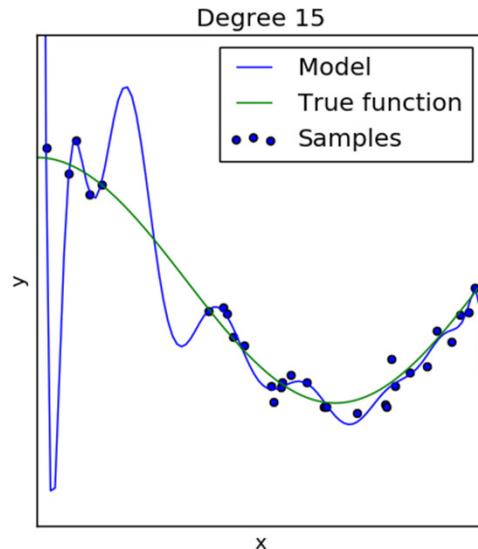
- Which model is the best?



Linear model: underfitting



4<sup>th</sup>-order model: well fitting



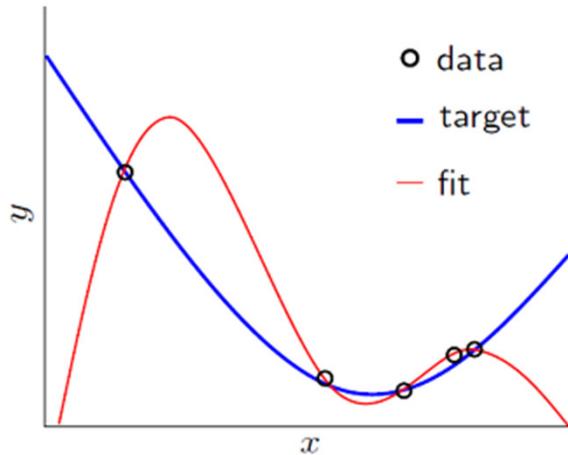
15<sup>th</sup>-order model: overfitting

- Underfitting occurs when a statistical model or machine learning algorithm cannot capture the underlying trend of the data.
- Overfitting occurs when a statistical model describes random error or noise instead of the underlying relationship

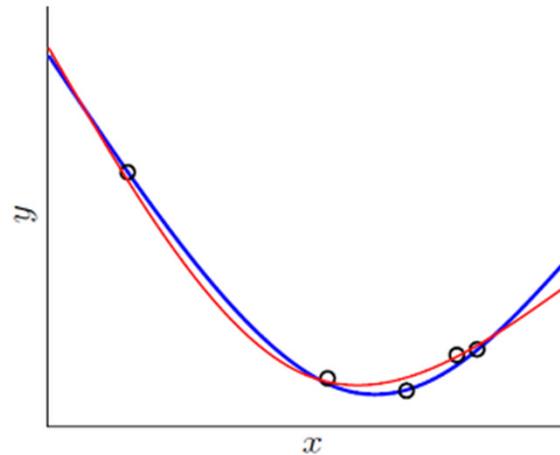
# Regularization

- Add a penalty term of the parameters to prevent the model from overfitting the data

$$\min_{\theta} \frac{1}{N} \sum_{i=1}^N \mathcal{L}(y_i, f_{\theta}(x_i)) + \lambda \Omega(\theta)$$



(a) without regularization



(b) with regularization

# Typical Regularization

- L2-Norm (Ridge)

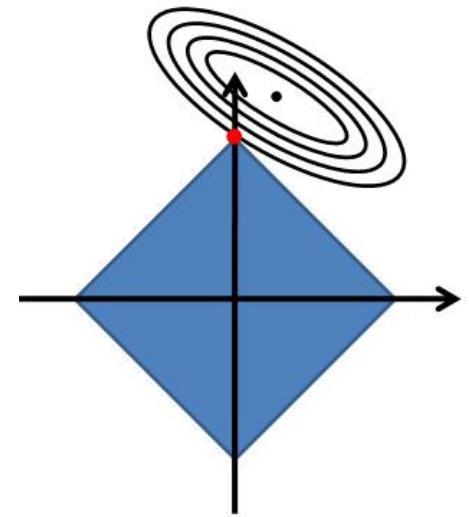
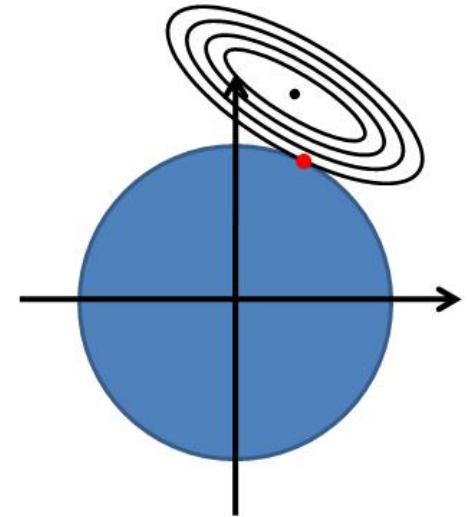
$$\Omega(\theta) = \|\theta\|_2^2 = \sum_{m=1}^M \theta_m^2$$

$$\min_{\theta} \frac{1}{N} \sum_{i=1}^N \mathcal{L}(y_i, f_{\theta}(x_i)) + \lambda \|\theta\|_2^2$$

- L1-Norm (LASSO)

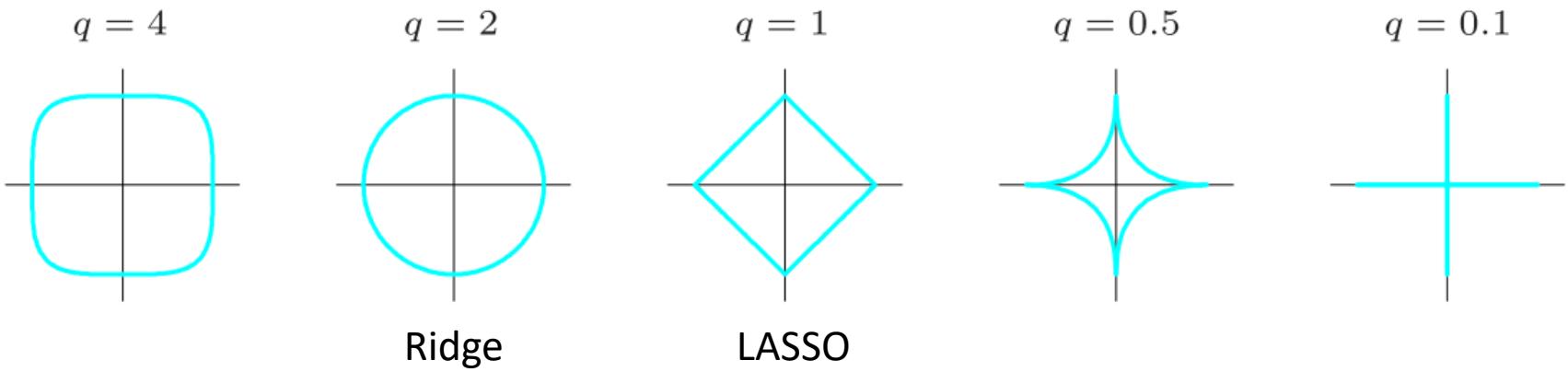
$$\Omega(\theta) = \|\theta\|_1 = \sum_{m=1}^M |\theta_m|$$

$$\min_{\theta} \frac{1}{N} \sum_{i=1}^N \mathcal{L}(y_i, f_{\theta}(x_i)) + \lambda \|\theta\|_1$$



# More Normal-Form Regularization

- Contours of constant value of  $\sum_j |\theta_j|^q$



- Sparse model learning with  $q$  not higher than 1
- Seldom use of  $q > 2$
- Actually, 99% cases use  $q = 1$  or 2

# Principle of Occam's razor

Among competing hypotheses, the one with the fewest assumptions should be selected.

- Recall the function set  $\{f_\theta(\cdot)\}$  is called **hypothesis space**

$$\min_{\theta} \frac{1}{N} \sum_{i=1}^N \mathcal{L}(y_i, f_\theta(x_i)) + \lambda \Omega(\theta)$$

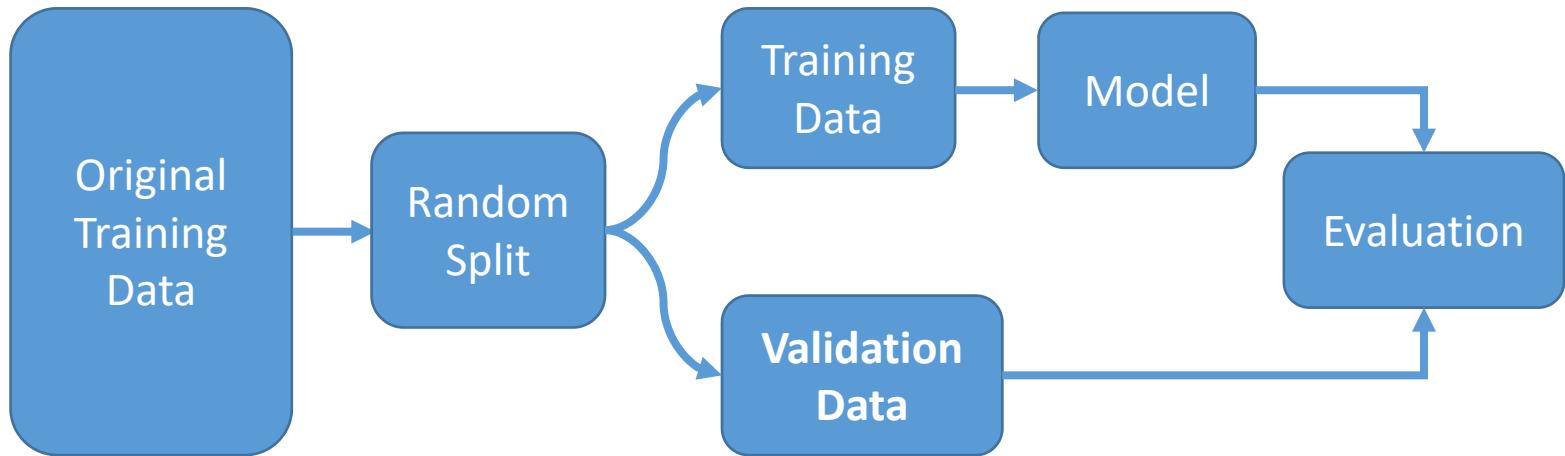
Original loss      Penalty on assumptions

# Model Selection

$$\min_{\theta} \frac{1}{N} \sum_{i=1}^N \mathcal{L}(y_i, f_{\theta}(x_i)) + \lambda \|\theta\|_2^2$$

- An ML solution has model parameters  $\theta$  and optimization **hyperparameter**  $\lambda$
- Hyperparameters
  - Define higher level concepts about the model such as complexity, or capacity to learn.
  - **Cannot be learned directly from the data** in the standard model training process and need to be predefined.
  - Can be decided by setting different values, training different models, and choosing the values that test better
- Model selection (or hyperparameter optimization) cares how to select the optimal hyperparameters.

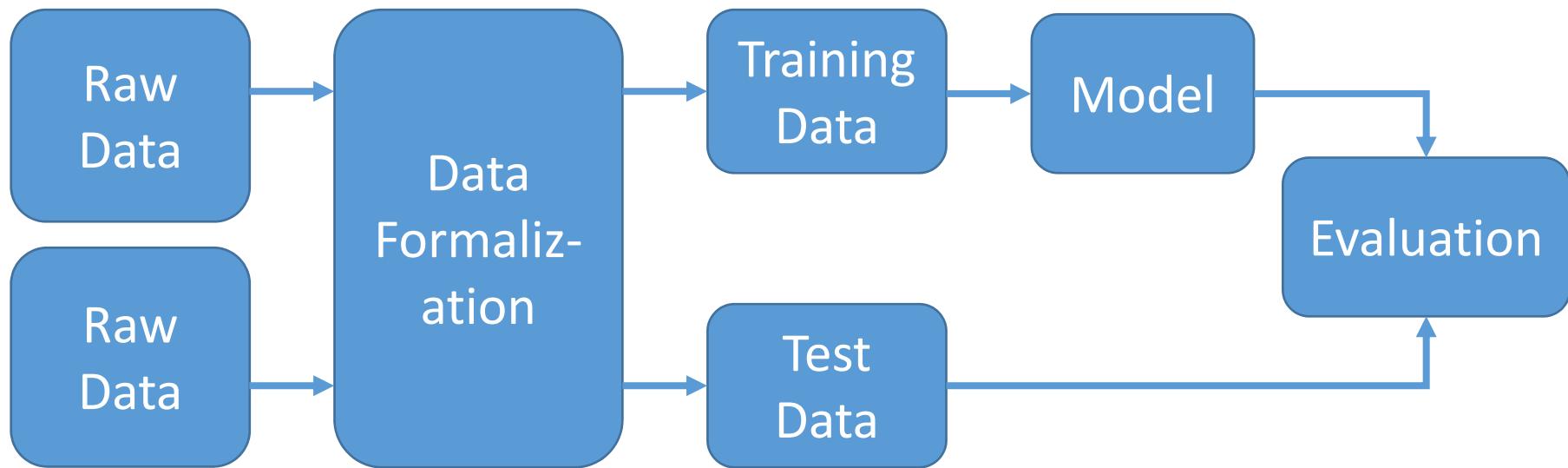
# Cross Validation for Model Selection



## *K*-fold Cross Validation

1. Set hyperparameters
2. For  $K$  times repeat:
  - Randomly split the original training data into training and validation datasets
  - Train the model on training data and evaluate it on validation data, leading to an evaluation score
3. Average the  $K$  evaluation scores as the model performance

# Machine Learning Process



- After selecting ‘good’ hyperparameters, we train the model over the whole training data and the model can be used on test data.

# Generalization Ability

- Generalization Ability is the model prediction capacity on **unobserved** data
  - Can be evaluated by **Generalization Error**, defined by
$$R(f) = \mathbb{E}[\mathcal{L}(Y, f(X))] = \int_{X \times Y} \mathcal{L}(y, f(x)) p(x, y) dx dy$$
  - where  $p(x, y)$  is the underlying (probably unknown) joint data distribution
- Empirical estimation of GA on a training dataset is

$$\hat{R}(f) = \frac{1}{N} \sum_{i=1}^N \mathcal{L}(y_i, f(x_i))$$

# A Simple Case Study on Generalization Error

- Finite hypothesis set  $\mathcal{F} = \{f_1, f_2, \dots, f_d\}$
- Theorem of generalization error bound:  
For any function  $f \in \mathcal{F}$ , with probability no less than  $1 - \delta$ , it satisfies

$$R(f) \leq \hat{R}(f) + \epsilon(d, N, \delta)$$

where

$$\epsilon(d, N, \delta) = \sqrt{\frac{1}{2N} \left( \log d + \log \frac{1}{\delta} \right)}$$

- $N$ : number of training instances
- $d$ : number of functions in the hypothesis set

# Lemma: Hoeffding Inequality

Let  $X_1, X_2, \dots, X_n$  be bounded independent random variables  $X_i \in [a, b]$ , the average variable  $Z$  is

$$Z = \frac{1}{n} \sum_{i=1}^n X_i$$

Then the following inequalities satisfy:

$$P(Z - \mathbb{E}[Z] \geq t) \leq \exp\left(\frac{-2nt^2}{(b-a)^2}\right)$$

$$P(\mathbb{E}[Z] - Z \geq t) \leq \exp\left(\frac{-2nt^2}{(b-a)^2}\right)$$

# Proof of Generalized Error Bound

- Assume the bounded loss function  $L(y, f(x)) \in [0, 1]$
- Based on Hoeffding Inequality, for  $\epsilon > 0$ , we have

$$P(R(f) - \hat{R}(f) \geq \epsilon) \leq \exp(-2N\epsilon^2)$$

- As  $\mathcal{F} = \{f_1, f_2, \dots, f_d\}$  is a finite set, it satisfies

$$\begin{aligned} P(\exists f \in \mathcal{F} : R(f) - \hat{R}(f) \geq \epsilon) &= P\left(\bigcup_{f \in \mathcal{F}} \{R(f) - \hat{R}(f) \geq \epsilon\}\right) \\ &\leq \sum_{f \in \mathcal{F}} P(R(f) - \hat{R}(f) \geq \epsilon) \\ &\leq d \exp(-2N\epsilon^2) \end{aligned}$$

# Proof of Generalized Error Bound

- Equivalence statements

$$P(\exists f \in \mathcal{F} : R(f) - \hat{R}(f) \geq \epsilon) \leq d \exp(-2N\epsilon^2)$$

$\Updownarrow$

$$P(\forall f \in \mathcal{F} : R(f) - \hat{R}(f) < \epsilon) \geq 1 - d \exp(-2N\epsilon^2)$$

- Then setting

$$\delta = d \exp(-2N\epsilon^2) \quad \Leftrightarrow \quad \epsilon = \sqrt{\frac{1}{2N} \log \frac{d}{\delta}}$$

The generalized error is bounded with the probability

$$P(R(f) < \hat{R}(f) + \epsilon) \geq 1 - \delta$$

□

# Discriminative Model and Generative Model

- **Discriminative model**
  - modeling the dependence of unobserved variables on observed ones
  - also called conditional models.
  - Deterministic:  $y = f_\theta(x)$
  - Probabilistic:  $p_\theta(y|x)$
- **Generative model**
  - modeling the joint probabilistic distribution of data
  - given some hidden parameters or variables
$$p_\theta(x, y)$$
  - then do the conditional inference

$$p_\theta(y|x) = \frac{p_\theta(x, y)}{p_\theta(x)} = \frac{p_\theta(x, y)}{\sum_{y'} p_\theta(x, y')}$$

# Discriminative Model and Generative Model

- **Discriminative model**

- modeling the dependence of unobserved variables on observed ones
- also called conditional models.
- Deterministic:  $y = f_\theta(x)$
- Probabilistic:  $p_\theta(y|x)$
- Directly model the dependence for label prediction
- Easy to define dependence-specific features and models
- Practically yielding higher prediction performance
- Linear regression, logistic regression, k nearest neighbor, SVMs, (multi-layer) perceptrons, decision trees, random forest etc.

# Discriminative Model and Generative Model

- Generative model

- modeling the joint probabilistic distribution of data
- given some hidden parameters or variables

$$p_{\theta}(x, y)$$

- then do the conditional inference

$$p_{\theta}(y|x) = \frac{p_{\theta}(x, y)}{p_{\theta}(x)} = \frac{p_{\theta}(x, y)}{\sum_{y'} p_{\theta}(x, y')}$$

- Recover the data distribution [essence of data science]
- Benefit from hidden variables modeling
- Naive Bayes, Hidden Markov Model, Mixture Gaussian, Markov Random Fields, Latent Dirichlet Allocation etc.