Feature Engineering in Machine Learning

Chun-Liang Li (李俊良)
chunlial@cs.cmu.edu
2016/07/17@台灣資料科學年會

Carnegie Mellon University

About Me

Academic

- **NTU CSIE BS/MS (2012/2013)**
  - Advisor: Prof. Hsuan-Tien Lin
- **CMU MLD PhD (2014-)**
  - Advisor: Prof. Jeff Schneider
    Prof. Barnabás Póczos

Competition

- **KDD Cup 2011 Champions**
  KDD Cup 2013 Champions
  - With Prof. Chih-Jen Lin
    Prof. Hsuan-Tien Lin
    Prof. Shou-De Lin
    Many students

Working

Appier (2012 intern)

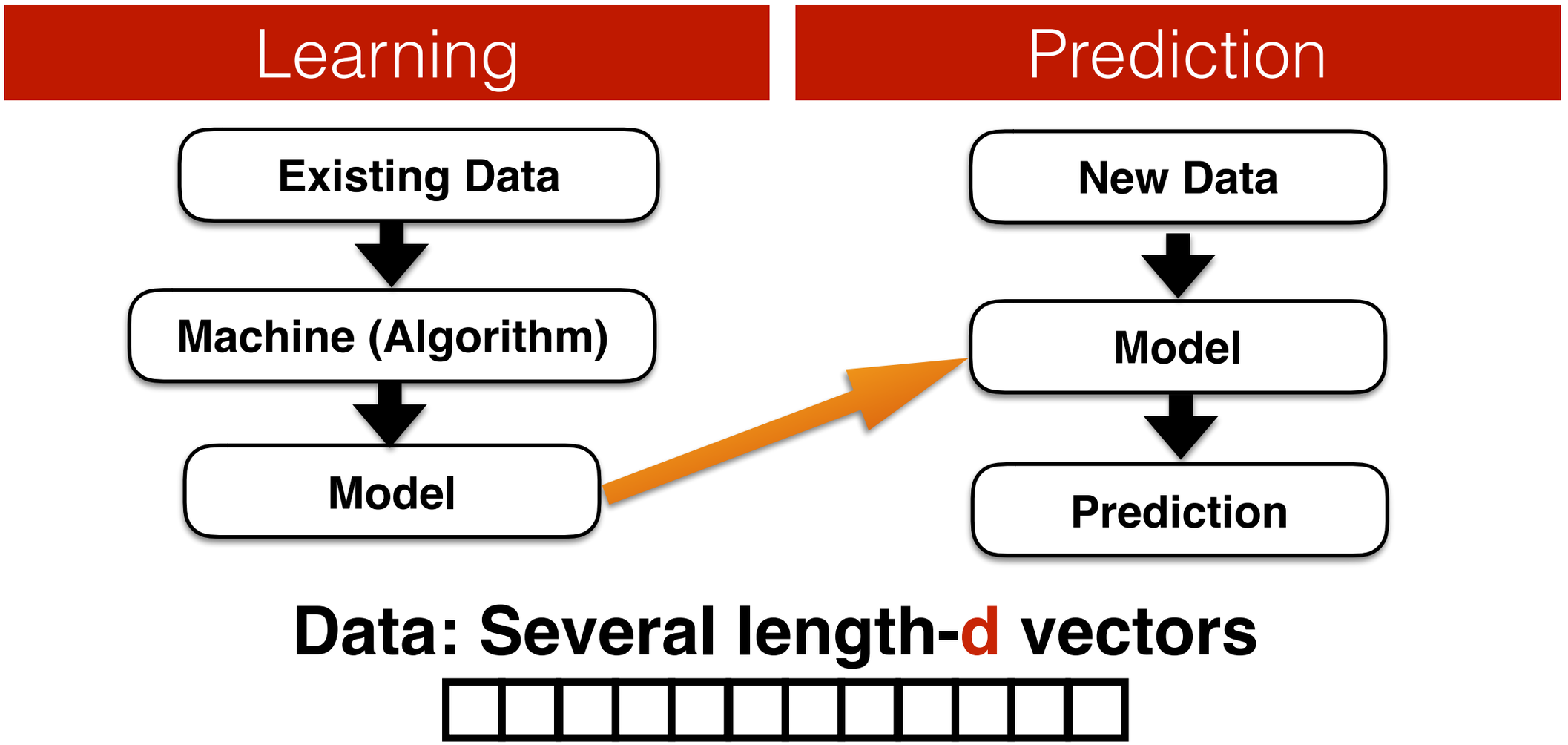
(2015 intern)

Carnegie Mellon University

MACHINE LEARNING DEPARTMENT

What is Machine Learning?

- What is Machine Learning?



Data? Algorithm?

- In academic
  - Assume we are given good enough data (in d-dimensional of course ☺)
  - Focus on designing better algorithms
    Sometimes complicated algorithms imply publications ☹
- In practice
  - Where is your good enough data?
  - Or, how to **transform** your data into a d-dimensional one?

From Zero to One:

Create your features by your observations

Carnegie Mellon University

MACHINE LEARNING DEPARTMENT

An Apple

How to describe this picture?

More Fruits

- **Method I:** Use size of picture

(640, 580)

(640, 580)

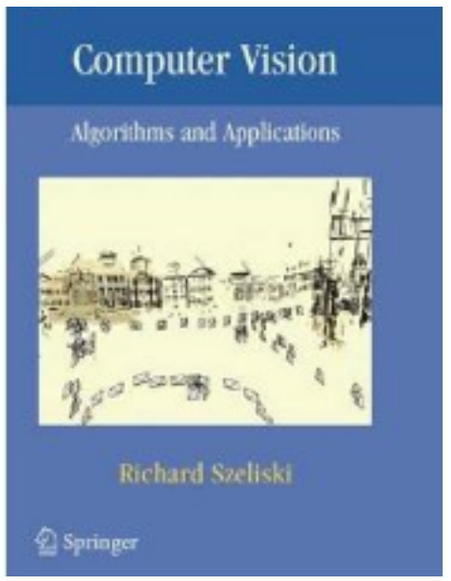
- **Method II:** Use RGB average

(219, 156, 140)

(243, 194, 113)

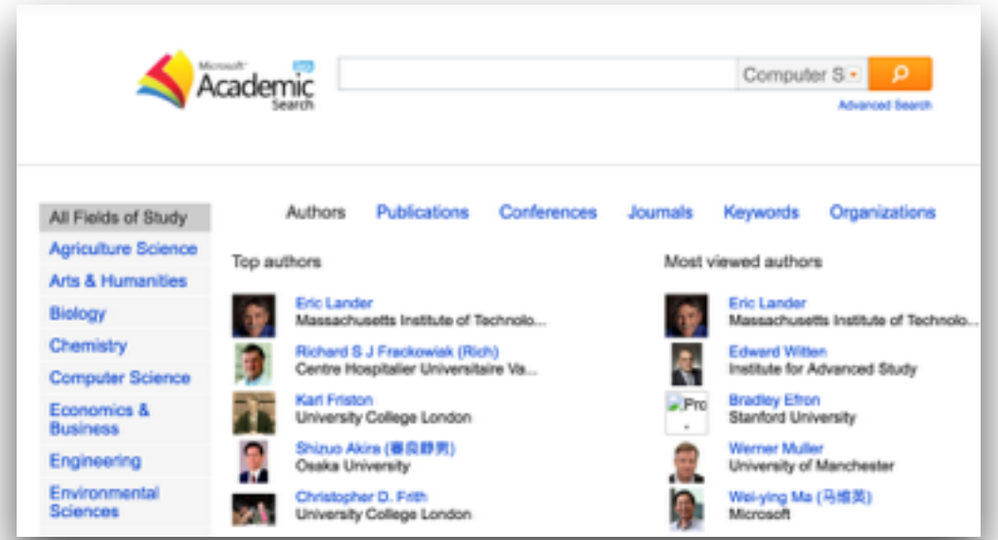
(216, 156, 155)

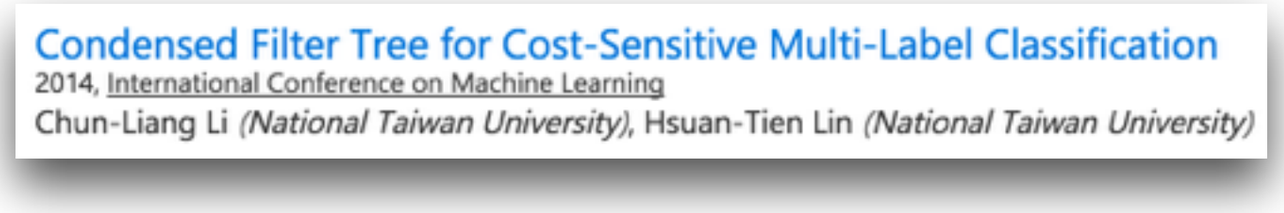
- Many more powerful features developed in computer vision



Case Study (KDD Cup 2013)

- Determine whether a paper is written by a given author





We are given raw text of these

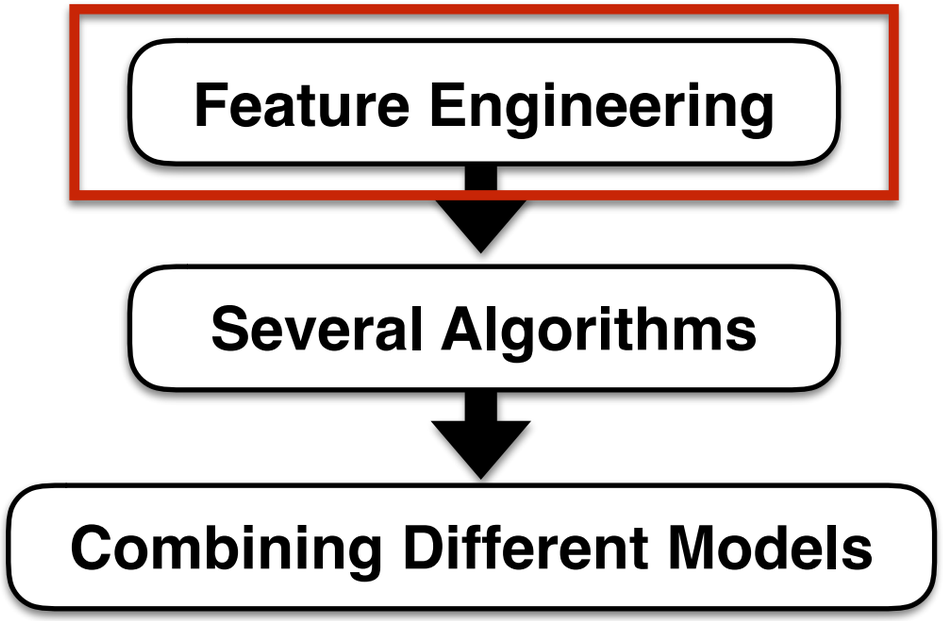
Data: https://www.kaggle.com/c/kdd-cup-2013-author-paper-identification-challenge

Carnegie Mellon University

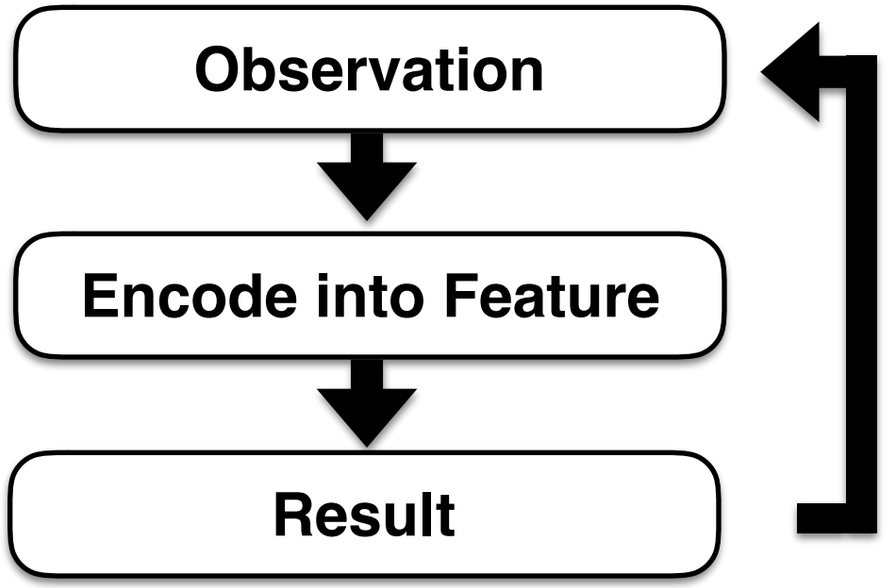
MACHINE LEARNING DEPARTMENT

NTU Approaches

Pipeline



Feature Engineering



First observation: Authors Information

- Are these my (Chun-Liang Li) papers? (Easy! check author names)
  1. **Chun-Liang Li** and Hsuan-Tien Lin. Condensed filter tree for cost-sensitive multi-label classification.
  2. Yao-Nan Chen and Hsuan-Tien Lin. Feature-aware label space dimension reduction for multi-label classification.

  - Encode by name similarities (e.g., how many characters are the same)

- Are **Li, Chun-Liang** and **Chun-Liang Li** the same?
  - Yes! Eastern and Western order
  - How about **Li Chun-Liang**? (Calculate the similarity of the **reverse** order)
- Also take co-authors into account
- 29 features in total

Second Observation: Affiliations

- Are Dr. Chi-Jen Lu and Prof. Chih-Jen Lin the same?
  - **Similar name:** Chi-Jen Lu v.s. Chih-Jen Lin
  - **Shared** co-author (me!)
- Take affiliations into account!
  - Academia Sinica v.s. National Taiwan University
- 13 features in total

Carnegie Mellon University

MACHINE LEARNING DEPARTMENT

Last of KDD Cup 2013

- Many other features, including
  - Can you live for more than 100 years? At least I think I can't do research after 100 years ☺

- More advanced: social network features

Summary

The **97 features** designed by students **won the competition**

Furthermore

- If I can access the content, can I do better? **Definitely**☺

Author: Robert Galbraith

Who is Robert Galbraith?

"I thought it was by a very mature writer, and not a first-timer." — *Peter James*

Writing Style?

- "I was testing things like word length[1], sentence length[2], paragraph length[3], frequency of particular words[4] and the pattern of punctuation[5]"

— *Peter Millican (University of Oxford)*

Game Changing Point:

Deep Learning

Common Type of Data

- **Image**

- **Text**

The New York Times

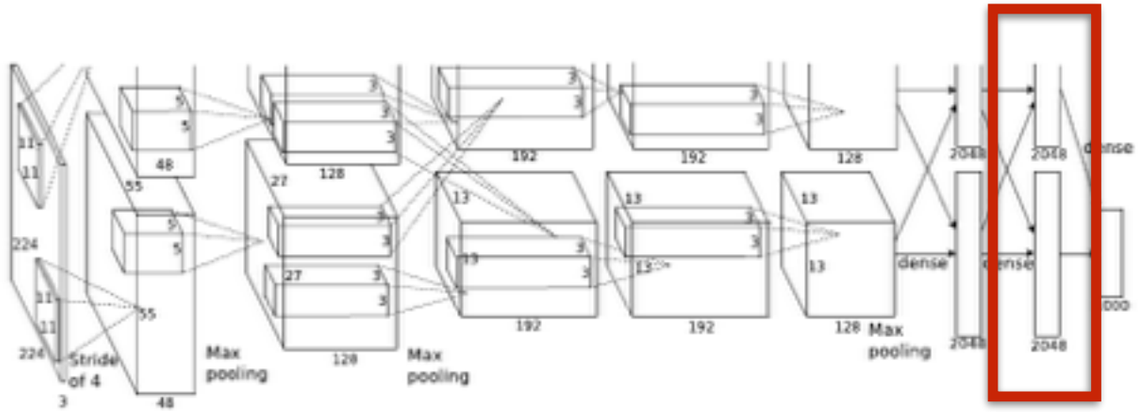
Carnegie Mellon University

MACHINE LEARNING DEPARTMENT

Representation Learning

- Deep Learning as learning hidden representations

Raw data ➡



Use last layer to extract features (Krizhevsky et al., 2012)

(Check Prof. Lee's talk and go to deep learning session later ☺)

- An active research topic in academia and industry

facebook

Google DeepMind

Microsoft Research

NVIDIA

Carnegie Mellon University

MACHINE LEARNING DEPARTMENT

Use Pre-trained Network

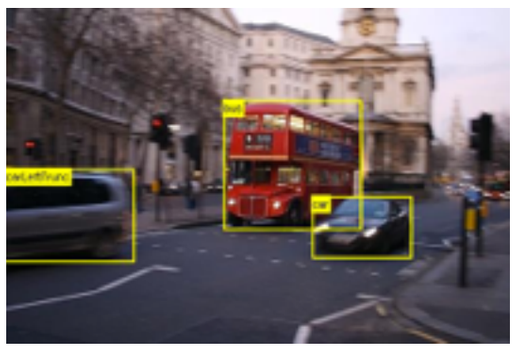
- Yon don’t need to train a network by yourself
- Use existing pre-trained network to extract features
  - AlexNet
  - VGG
  - Word2Vector

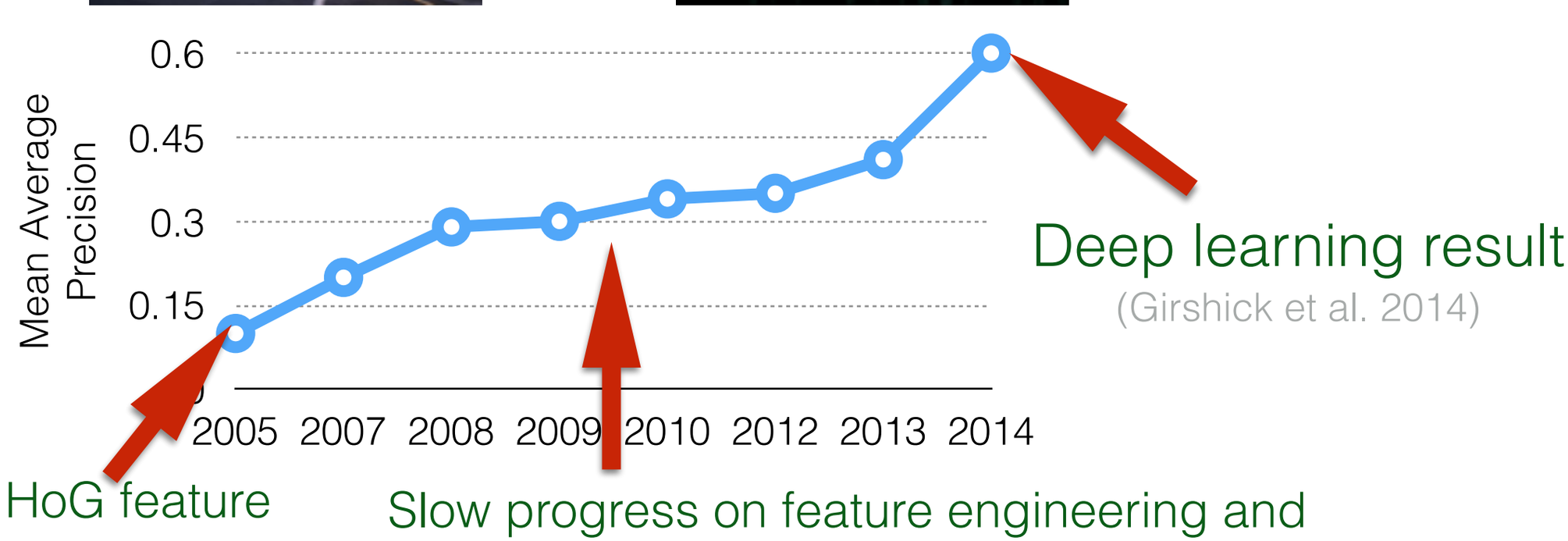
Result

Simply using deep learning features achieves **state-of-the-art** performance in many applications

Successful Example

- The PASCAL Visual Object Classes Challenge





Curse of Dimensionality:

Feature Selection and Dimension Reduction

The more, the better?

Practice

If we have 1,000,000 data with 100,000 dimensions, how much memory do we need?

Ans:

$$\begin{aligned} & 10^6 \times 10^5 \times 8 \\ = \; & 8 \times 10^{11} \text{ (B)} \\ = \; & 800 \text{ (GB)} \end{aligned}$$

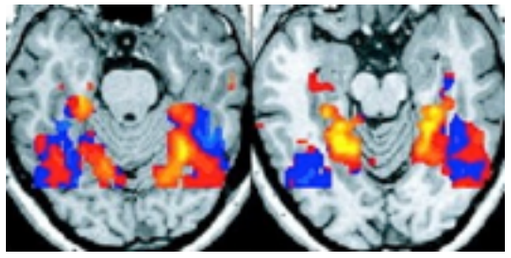
Noisy Feature

Is every feature useful? Redundancy?

Theory

Without any assumption, you need $O(\frac{1}{\epsilon^d})$ data to achieve ϵ error for d-dimensional data

Feature Selection

- Select import features 
  - Reduce dimensions 
- Explainable Results 

Commonly Used Tools

- **LASSO (Sparse Constraint)**
- **Random Forests**
- **Many others**

KDD Cup Again

- In KDD Cup 2013, we actually generated more than 200 features (some secrets you won't see in the paper ☺)

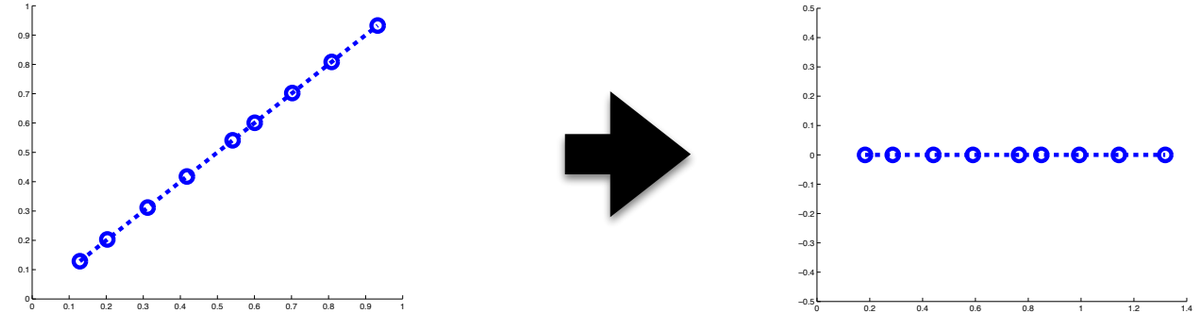
- Use **random forests** to select only 97 features, since many features are unimportant and even harmful, but why?

Non-useful Features

- Duplicated features
  - Example I: Country (Taiwan) v.s. Coordinates (121, 23.5)
  - Example II: Date of birth (1990) v.s. Age (26)
- Noisy features
  - Noisy information (something wrong in your data)
  - Missing values (something missing in your data)
- **What if we still have too many features?**

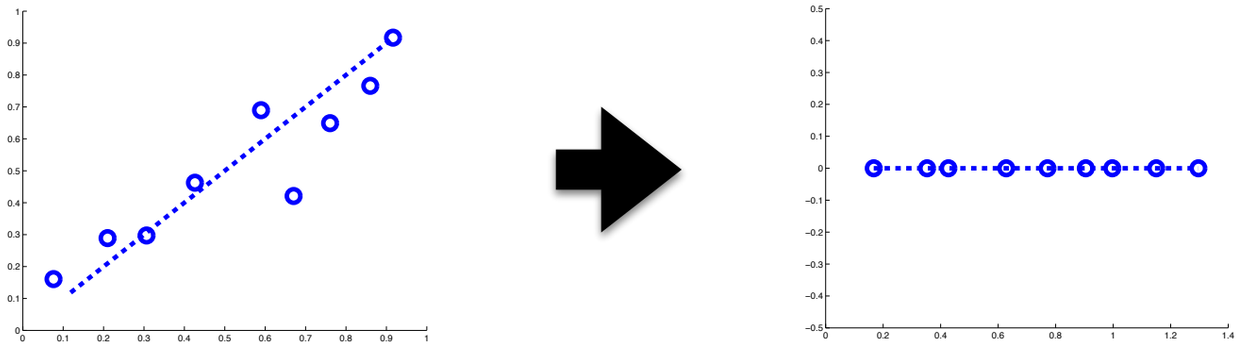
Dimension Reduction

- Let's visualize the data (a perfect example)



One dimension is enough

- Non-perfect example in practice



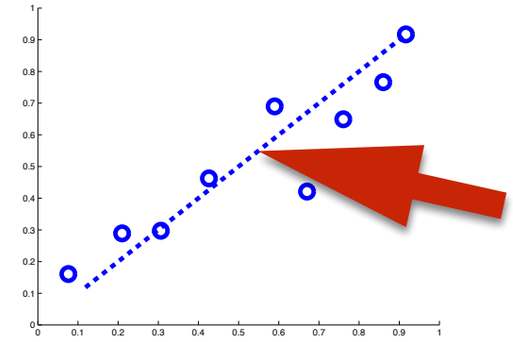
Trade-off between information and space

Commonly Used Tools

- **Principal Component Analysis (PCA)**

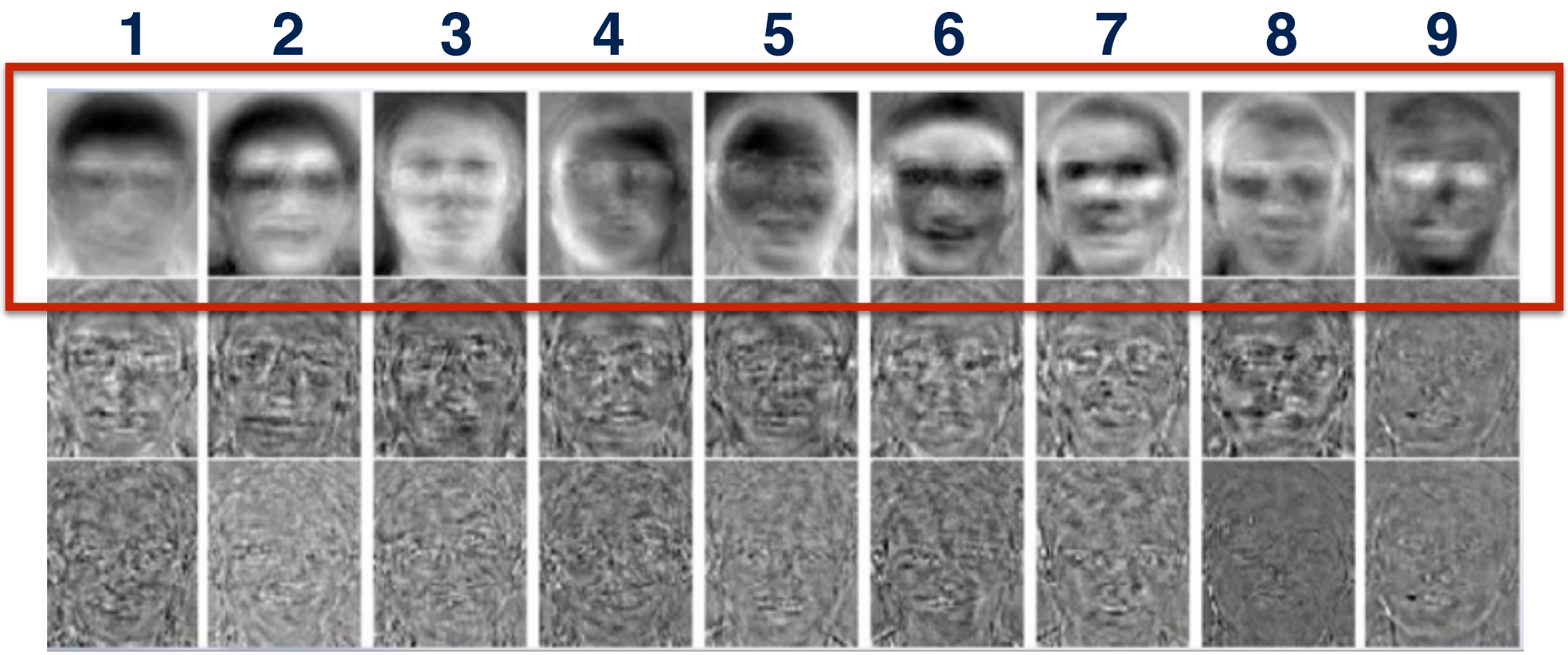
PCA — Intuition

- Let's apply PCA on these faces (raw pixels) and visualize the coordinates



PCA — Intuition (cont.)

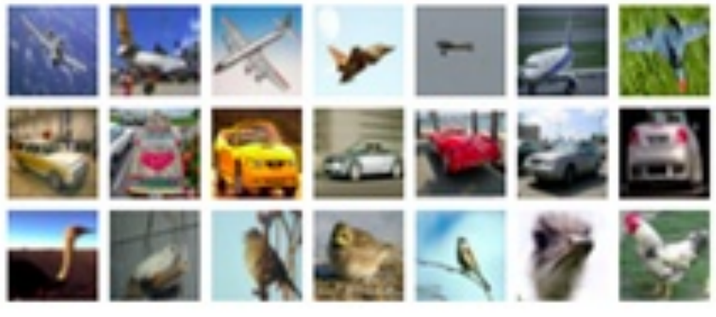
- We can use very few **base faces** to approximate (describe) the original faces



(Sirovich and Kirby, Low-dimensional procedure for the characterization of human faces)

PCA — Case Study

- CIFAR-10 image classification with **raw pixels** as features and using approximated kernel SVM



(Li and Pòczos, Utilize Old Coordinates: Faster Doubly Stochastic Gradients for Kernel Methods, UAI 2016)

Dimensions	Accuracy	Time
3072 (all)	63.1%	**~2 Hrs**
100 (PCA)	59.8%	250 s

Trade-off between information, space and time

PCA in Practice

- **Practical concern:**
  - Time complexity: $O(Nd^2)$
  - Space complexity: $O(d^2)$

Small Problem

PCA takes **<10 seconds** for CIFAR-10 dataset (d=3072) by using 12 cores (E5-2620)

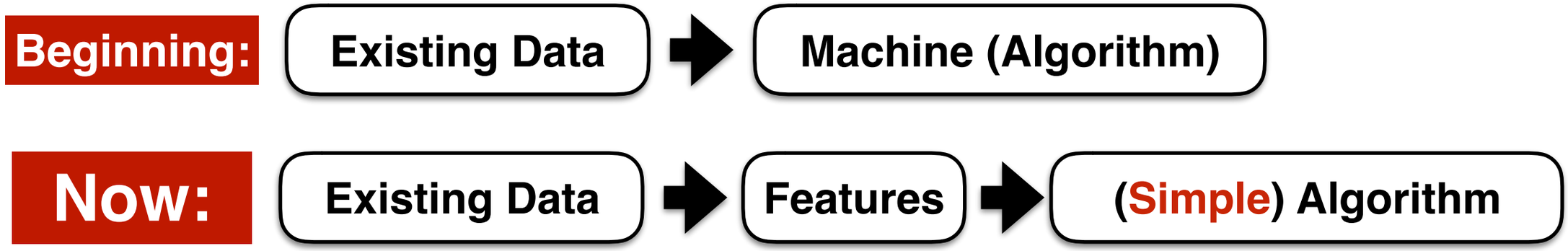
- **Remark: Use fast approximation for large-scale problem (e.g., >100k dimensions)**

1. PCA with random projection (implemented in scikit-learn)
   (Halko et al., Finding Structure with Randomness, 2011)

2. Stochastic algorithms (easy to implement from scratch)
   (Li et al., Rivalry of Two Families of Algorithms for Memory-Restricted Streaming PCA, AISTATS 2016)

Conclusion

- Observe the data and encode them into meaningful features



- Deep learning is a powerful tool to use
- Reduce number of features if necessary
  - Reduce non-useful features
  - Computational concern

Thanks!
Any Question?

References

1. Richard Szeliski. Computer Vision: Algorithms and Applications, 2010.
2. Senjuti Basu Roy, Martine De Cock, Vani Mandava, Swapna Savanna, Brian Dalessandro, Claudia Perlich, William Cukierski, and Ben Hamner. The Microsoft academic search dataset and KDD cup 2013. In *KDD Cup 2013 Workshop, 2013.*
3. Chun-Liang Li, Yu-Chuan Su, Ting-Wei Lin, Cheng-Hao Tsai, Wei-Cheng Chang, Kuan-Hao Huang, Tzu-Ming Kuo, Shan-Wei Lin, Young-San Lin, Yu-Chen Lu, Chun-Pai Yang, Cheng-Xia Chang, Wei-Sheng Chin, Yu-Chin Juan, Hsiao-Yu Tung, Jui-Pin Wang, Cheng-Kuang Wei, Felix Wu, Tu-Chun Yin, Tong Yu, Yong Zhuang, Shou-De Lin, Hsuan-Tien Lin, and Chih-Jen Lin. Combination of feature engineering and ranking models for paper-author identification in KDD Cup 2013. In *JMLR, 2015.*
4. How JK Rowling was unmasked. http://www.bbc.com/news/entertainment-arts-23313074
5. Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. In *IEEE PAMI, 2015.*
6. Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS, 2012.*
7. Karen Simonyan and Andrew Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *ICLR, 2015.*
8. Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient Estimation of Word Representations in Vector Space. *Technical Report, 2013.*

9. Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR, 2014.*

10. Matthew A. Turk, and Alex Peatland. Face Recognition Using Eigenfaces. In *CVPR, 1991.*

11. Chun-Liang Li, and Barnabás Póczos. Utilize Old Coordinates: Faster Doubly Stochastic Gradients for Kernel Methods. In *UAI, 2016.*

12. Nathan Halko, Per-Gunnar Martinsson, Joel A. Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. In SIAM Rev., 2011.

13. Chun-Liang Li, Hsuan-Tien Lin and, Chi-Jen Lu. Rivalry of Two Families of Algorithms for Memory-Restricted Streaming PCA. In *AISTATS, 2016.*