UVA CS 4501: Machine Learning

Lecture 1: Introduction

Dr. Yanjun Qi

University of Virginia Department of Computer Science

Welcome

- CS 4501 Machine Learning
  - TuTh 3:30pm-4:45pm,
  - Rice Hall 130

- Your UVA collab for Assignments:
- Course Website:
  - https://qiyanjun.github.io/2018fUVA-CS4501MachineLearning/

Today

- ❑ Course Logistics
- ❑ Machine Learning Basics
- ❑ Machine Learning History
- ❑ Rough Plan of Course Content

Course Staff

- Instructor: Prof. Yanjun Qi
  - QI: /ch ee/
  - You can call me “professor”, “professor Qi”;
  - I have been teaching Graduate-level and Under-Level Machine Learning course for five years!
  - My research is about machine learning

- TA and Office Hour information @ CourseWeb

Course Logistics

- Q0- Quiz for the minimum background test !!!!

- Course email list has been setup. You should have received emails already !

- Policy, the grade will be calculated as follows:
  - Assignments (60%, **Six** total, each ~10%)
  - Midterm exam (20%)
  - Final exam (20%)

Course Logistics

- Midterm: 75mins
- Final: 75mins

- Six assignments (each 10%)
  - **Three** extension days policy (check course website)
- All late Homework should be submitted to 18f-cs-4501-001-ta@collab.its.virginia.edu

Homework Policy

- Policy,
  - Homework should be submitted electronically through UVaCollab
  - Homework should be finished individually
  - Due at midnight on the due date

  - In order to pass the course, the average of your midterm and final must also be "pass".

Late Homework Policy

- Each student has **three** extension days to be used at his or her own discretion throughout the entire course. Your grades would be discounted by 15% per day when you use these 3 late days. You could use the 3 days in whatever combination you like. For example, all 3 days on 1 assignment (for a maximum grade of 55%) or 1 each day over 3 assignments (for a maximum grade of 85% on each). After you've used all 3 days, you cannot get credit for anything turned in late.

Course Material

- Text books for this class is:
  - NONE

- My slides – **if it is not mentioned in my slides, it is not an official topic of the course**

Course Background Needed

- **Background Needed**
  - Calculus, Basic linear algebra, Basic probability and Basic Algorithm
  - Statistics is recommended.
  - Students should already have good programming skills, i.e. python is required for all programming assignments
  - We will review "algebra" and "probability" in class

Today

- ❑ Course Logistics
- ❑ Machine Learning Basics
- ❑ Machine Learning History
- ❑ Rough Plan of Course Content

OUR DATA-RICH WORLD

- Biomedicine
  - Patient records, brain imaging, MRI & CT scans, …
  - Genomic sequences, bio-structure, drug effect info, …
- Science
  - Historical documents, scanned books, databases from astronomy, environmental data, climate records, …
- Social media
  - Social interactions data, twitter, facebook records, online reviews, …
- Business
  - Stock market transactions, corporate sales, airline traffic, …

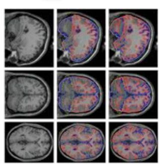
What can we do with the data wealth?

→ REAL-WORLD IMPACT

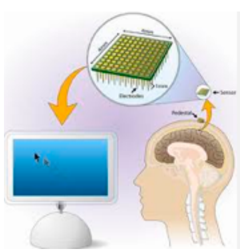
Transportation Data

Genomic Data

Medical Images



Brain computer interaction (BCI)



Device sensor data

- Business efficiencies
- Scientific breakthroughs
- Improve quality-of-life:
  - healthcare,
  - energy saving / generation,
  - environmental disasters,
  - nursing home,
  - transportation,
  - …

BIG DATA CHALLENGES

- Data capturing (sensor, smart devices, medical instruments, et al.)
- Data transmission
- Data storage
- Data management
- High performance data processing
- Data visualization
- Data security & privacy (e.g. multiple individuals)
- ……

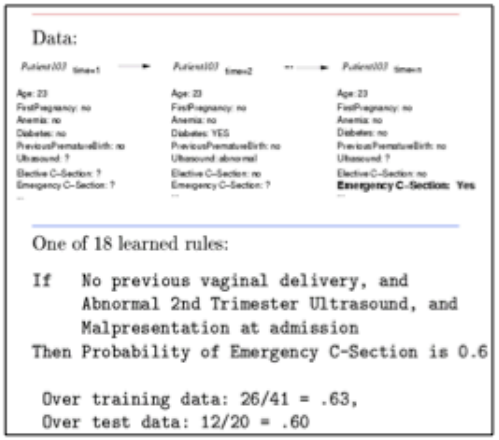
e.g. cloud computing

e.g. HCI

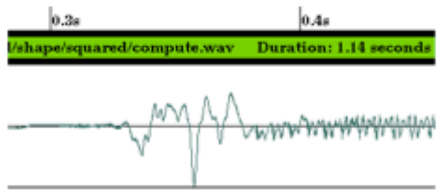
this course

- Data analytics
  - How can we analyze this big data wealth ?
  - E.g. Machine learning and data mining

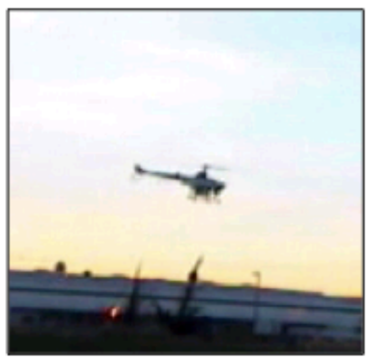
MACHINE LEARNING IS CHANGING THE WORLD



Mining Databases



Speech Recognition



Control learning

Object recognition

Text analysis

Peter H. van Oppen , Chairman of the Board & Chief Executive Officer
Mr. van Oppen has served as chairman of the board and chief executive officer of ADIC since its acquisition by Interpoint in 1994 and a director of ADIC since 1986. Until its acquisition by Crane Co. in October 1996, Mr. van Oppen served as chairman of the board of directors, president and chief executive officer of Interpoint . Prior to 1985, Mr. van Oppen worked as a consulting manager at Price Waterhouse LLP and at Bain & Company in Boston and London. He has additional experience in medical electronics and venture capital. Mr. van Oppen also serves as a director of Seattle FilmWorks Inc. and Spacelabs Medical, Inc.. He holds a B.A. from Whitman College and an M.B.A. from Harvard Business School, where he was a Baker Scholar.

Many more !

BASICS OF MACHINE LEARNING

- "The goal of machine learning is to build computer systems that can learn and adapt from their experience." – Tom Dietterich
- "Experience" in the form of available data examples (also called as instances, samples)
- Available examples are described with properties (data points in feature space X)

e.g. SUPERVISED LEARNING

- Find function to map input space X to output space Y $f : X \longrightarrow Y$
- So that the difference between y and $f(\boldsymbol{x})$ of each example $\boldsymbol{x}$ is small.

e.g.

x	I believe that this book is not at all helpful since it does not explain thoroughly the material . it just provides the reader with tables and calculations that sometimes are not easily understood …

Input X : e.g. a piece of English text

→

y	-1

Output Y: {1 / Yes , -1 / No }
e.g. Is this a positive product review ?

SUPERVISED Linear Binary Classifier

- Now let us check out a VERY SIMPLE case of

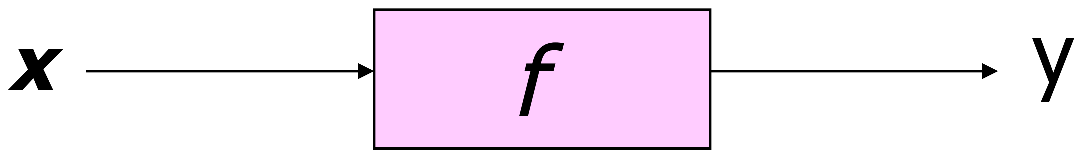
$$\boldsymbol{x} \longrightarrow \boxed{f} \longrightarrow y$$

e.g.: Binary *y* / Linear *f* / X as R^2

$$\boldsymbol{f}(\boldsymbol{x},\boldsymbol{w},b) = sign(\boldsymbol{w}^T\ \boldsymbol{x} + b)$$

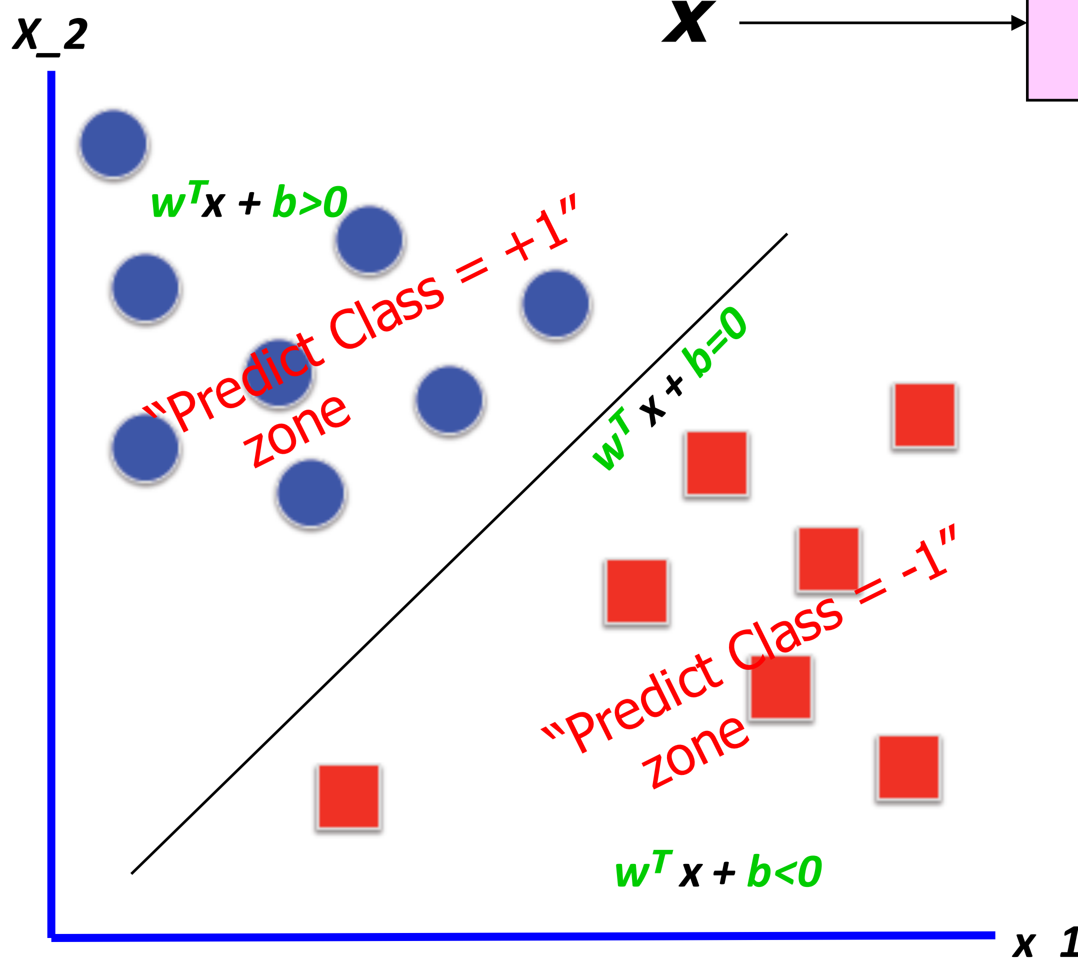
$$\boldsymbol{x} = (x\_1, x\_2)$$

SUPERVISED Linear Binary Classifier



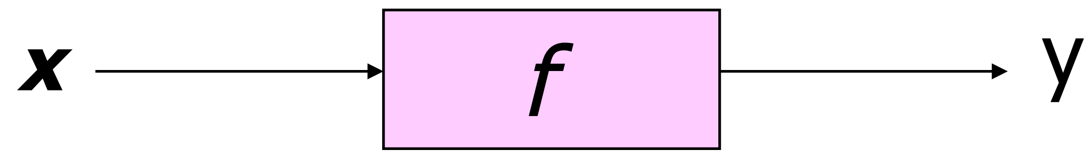
$f(\mathbf{x}, \mathbf{w}, b) = sign(\mathbf{w}^T \mathbf{x} + b)$

$\mathbf{x} = (x\_1, x\_2)$



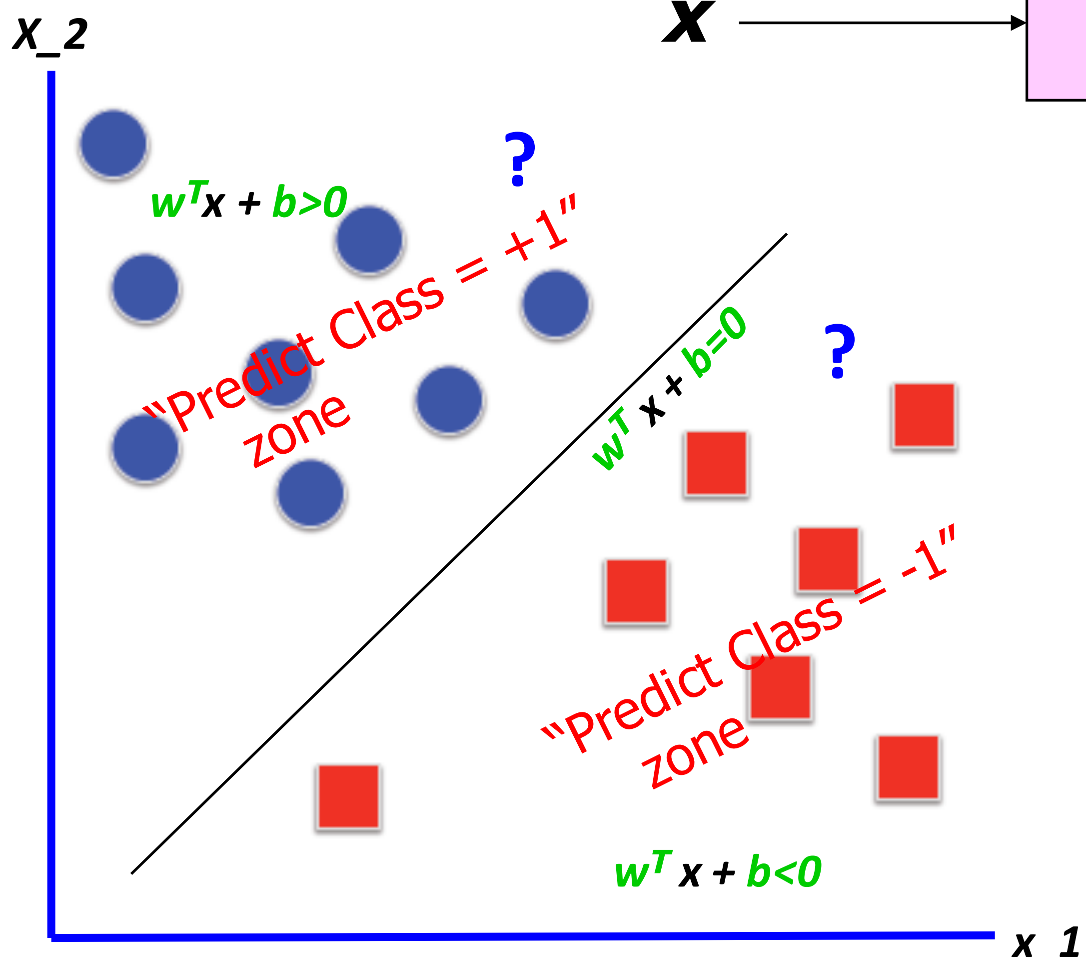
Courtesy slide from Prof. Andrew Moore's tutorial

SUPERVISED Linear Binary Classifier



$$f(\mathbf{x},\mathbf{w},b) = sign(\mathbf{w}^T \mathbf{x} + b)$$

$$\mathbf{x} = (x\_1, x\_2)$$



denotes +1 point

denotes -1 point

? denotes future points

Courtesy slide from Prof. Andrew Moore's tutorial

Basic Concepts

- Training (i.e. learning parameters $\boxed{\boldsymbol{w}, b}$)
  - Training set includes
    - available examples $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_L$
    - available corresponding labels $y_1, \ldots, y_L$
  - Find ($\boldsymbol{w}$,b) by minimizing loss (i.e. difference between y and $f(\boldsymbol{x})$ *on available examples in training set*)

$$(\boldsymbol{W}, b) = \underset{\boldsymbol{w},\, b}{\operatorname{argmin}} \sum_{i=1}^{L} \ell(f(x_i), y_i)$$

Basic Concepts

- Testing (i.e. evaluating performance on "future" points)
  - Difference between true $y_?$ and the predicted $f(x_?)$ on a set of testing examples (i.e. *testing set*)
  - Key: example $x_?$ not in the training set

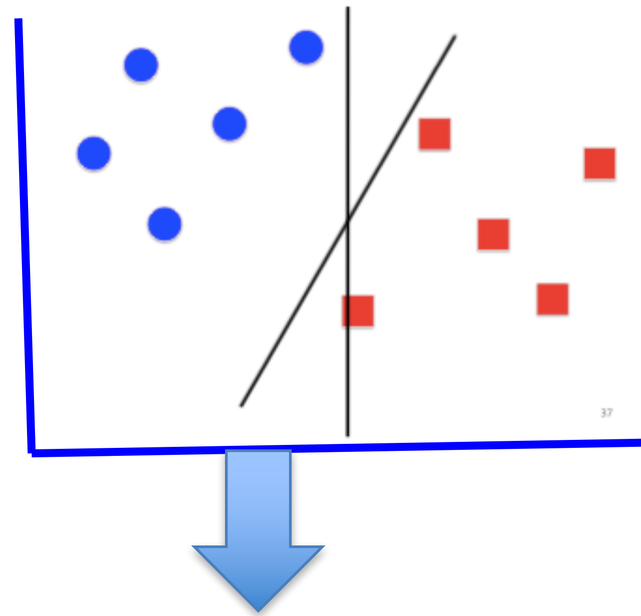
- Generalisation: learn function / hypothesis from past data in order to "explain", "predict", "model" or "control" new data examples

Basic Concepts

- Loss function
  - e.g. hinge loss for binary classification task

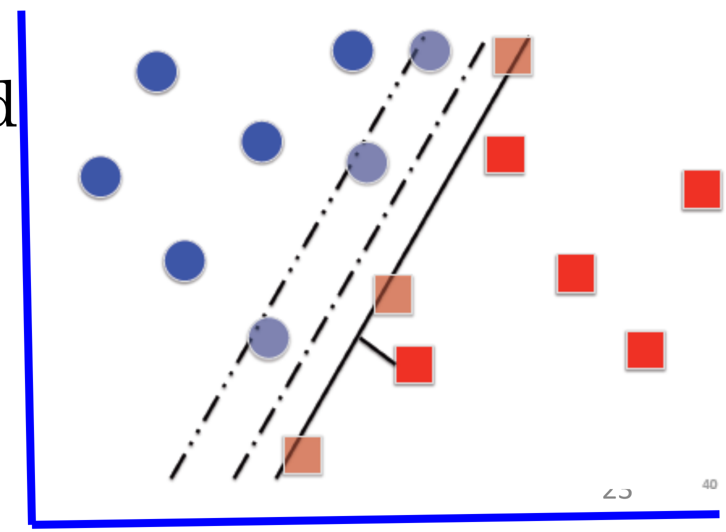
$$\sum_{i=1}^{L} \ell(f(x_i), y_i) = \sum_{i=1}^{L} \max(0, 1 - y_i f(x_i)).$$

  - e.g. pairwise ranking loss for ranking task (i.e. ordering examples by preference)

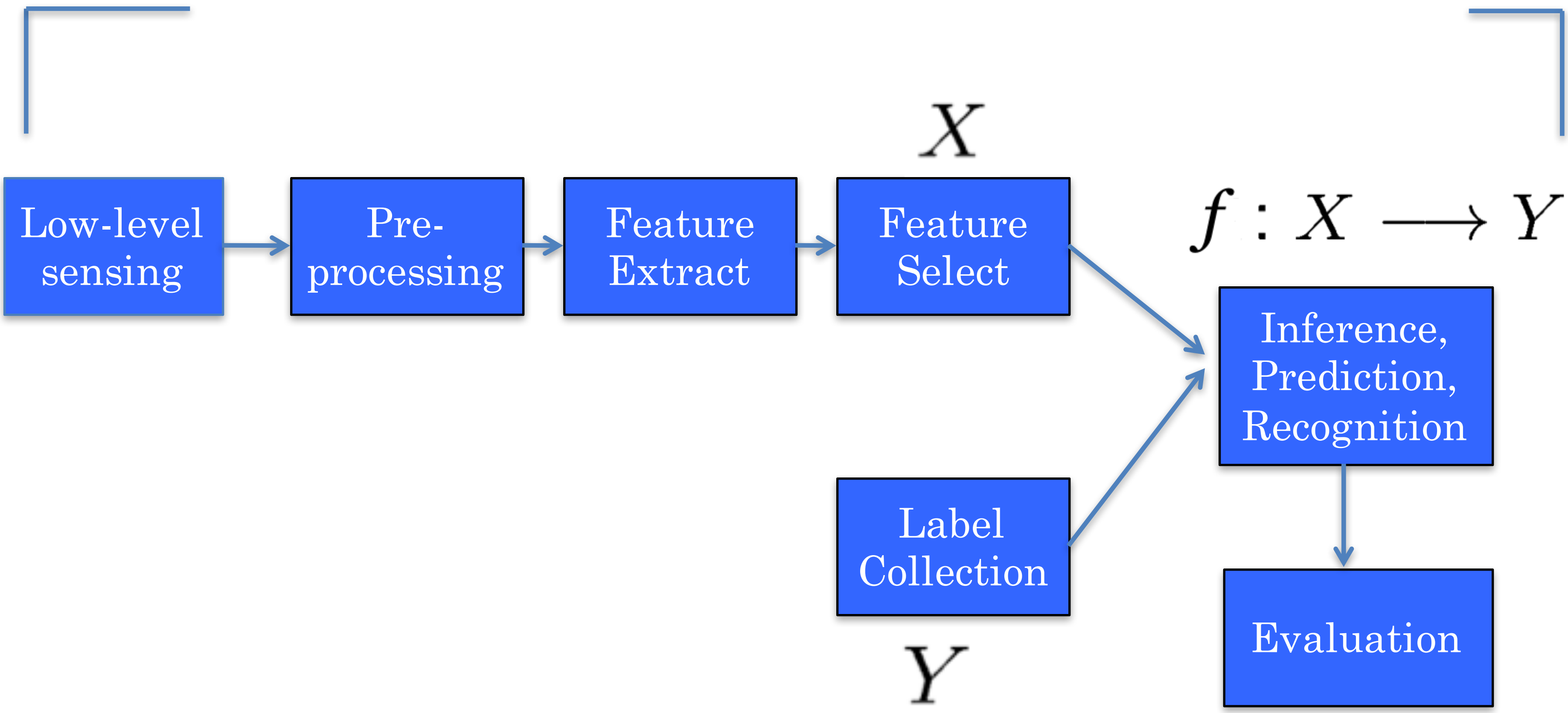


- Regularization
  - E.g. additional information added on loss function to control f

$$C \sum_{i=1}^{L} \ell(f(x_i), y_i) + \frac{1}{2} \|w\|^2.$$



TYPICAL MACHINE LEARNING SYSTEM



"Big Data" Challenges for Machine Learning

- ✓Large size of samples
- ✓High dimensional features

Not the focus, being covered in my advanced-level course

Large-Scale Machine Learning: SIZE MATTERS

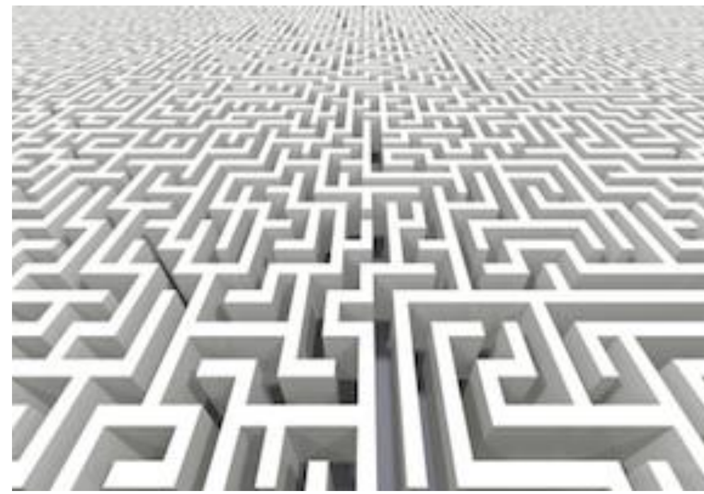
- One thousand data instances
- One million data instances
- One billion data instances
- One trillion data instances

Those are not different numbers, those are different mindsets !!!

BIG DATA CHALLENGES FOR MACHINE LEARNING

LARGE-SCALE

Highly Complex

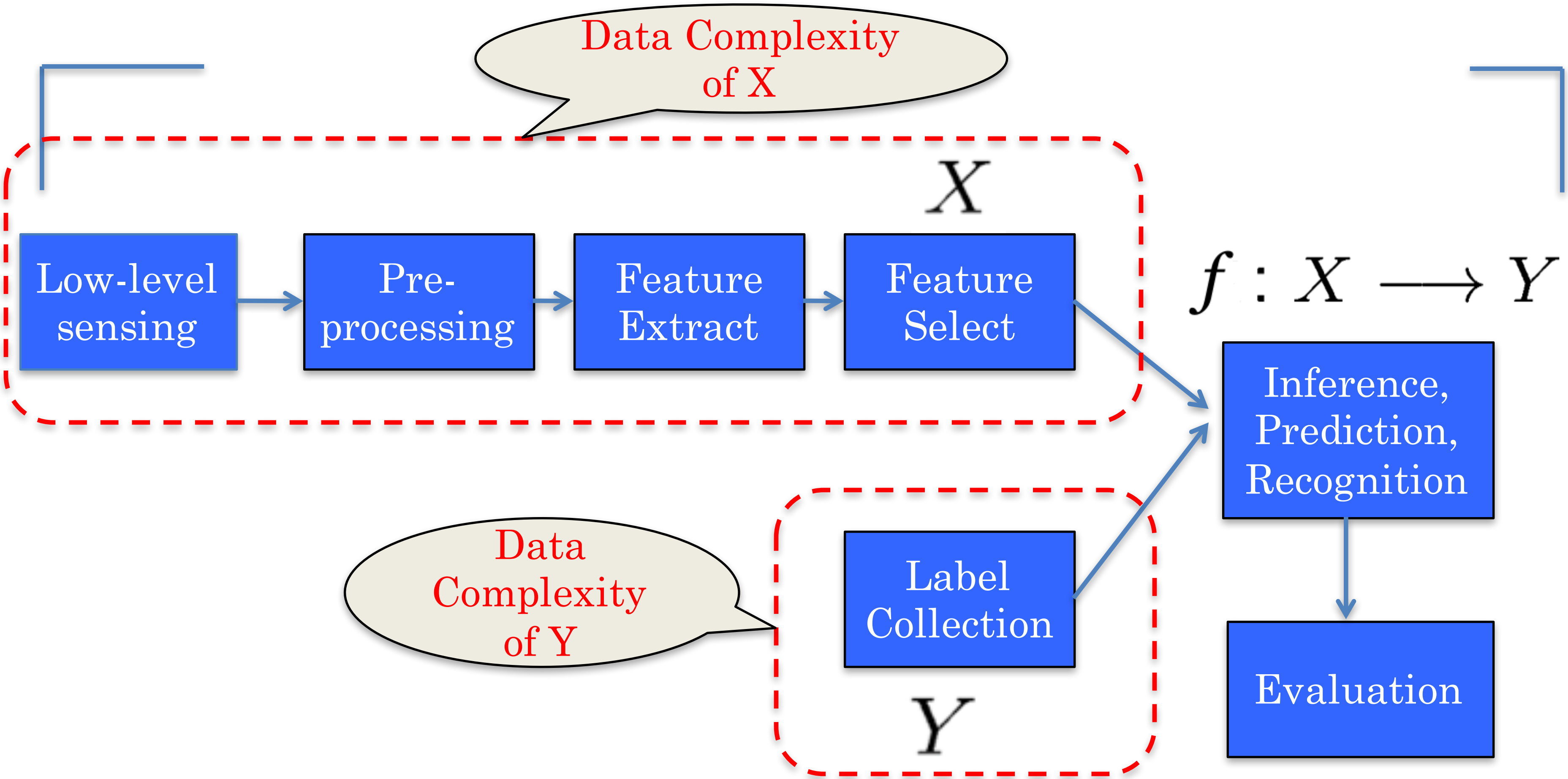


Most of this course

The variations of both X (feature, representation) and Y (labels) are complex !

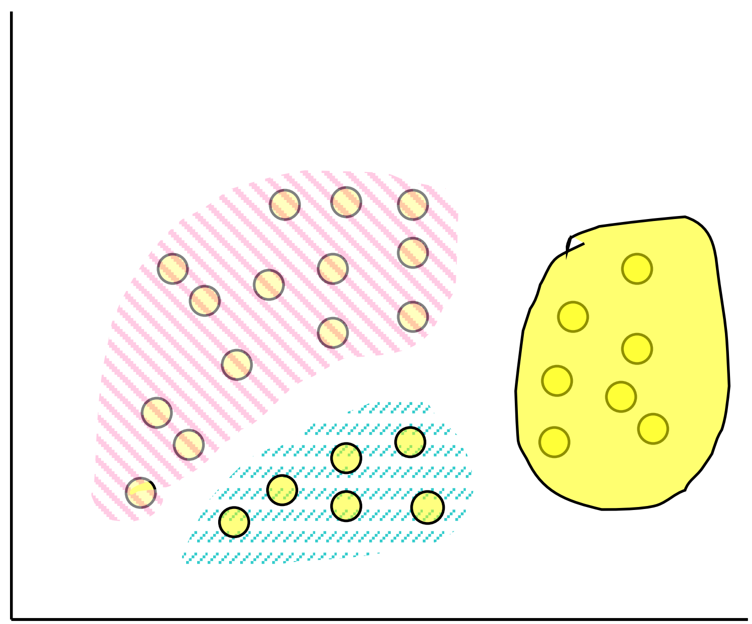
✓Complexity of X
✓Complexity of Y

TYPICAL MACHINE LEARNING SYSTEM



UNSUPERVISED LEARNING : [COMPLEXITY in Y]

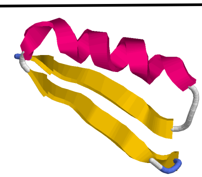
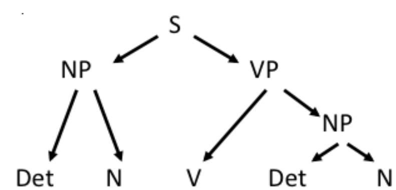
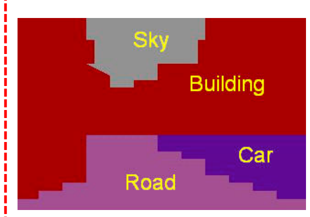
- No labels are provided (e.g. No Y provided)
- Find patterns from unlabeled data, e.g. clustering



e.g. clustering => to find "natural" grouping of instances given un-labeled data

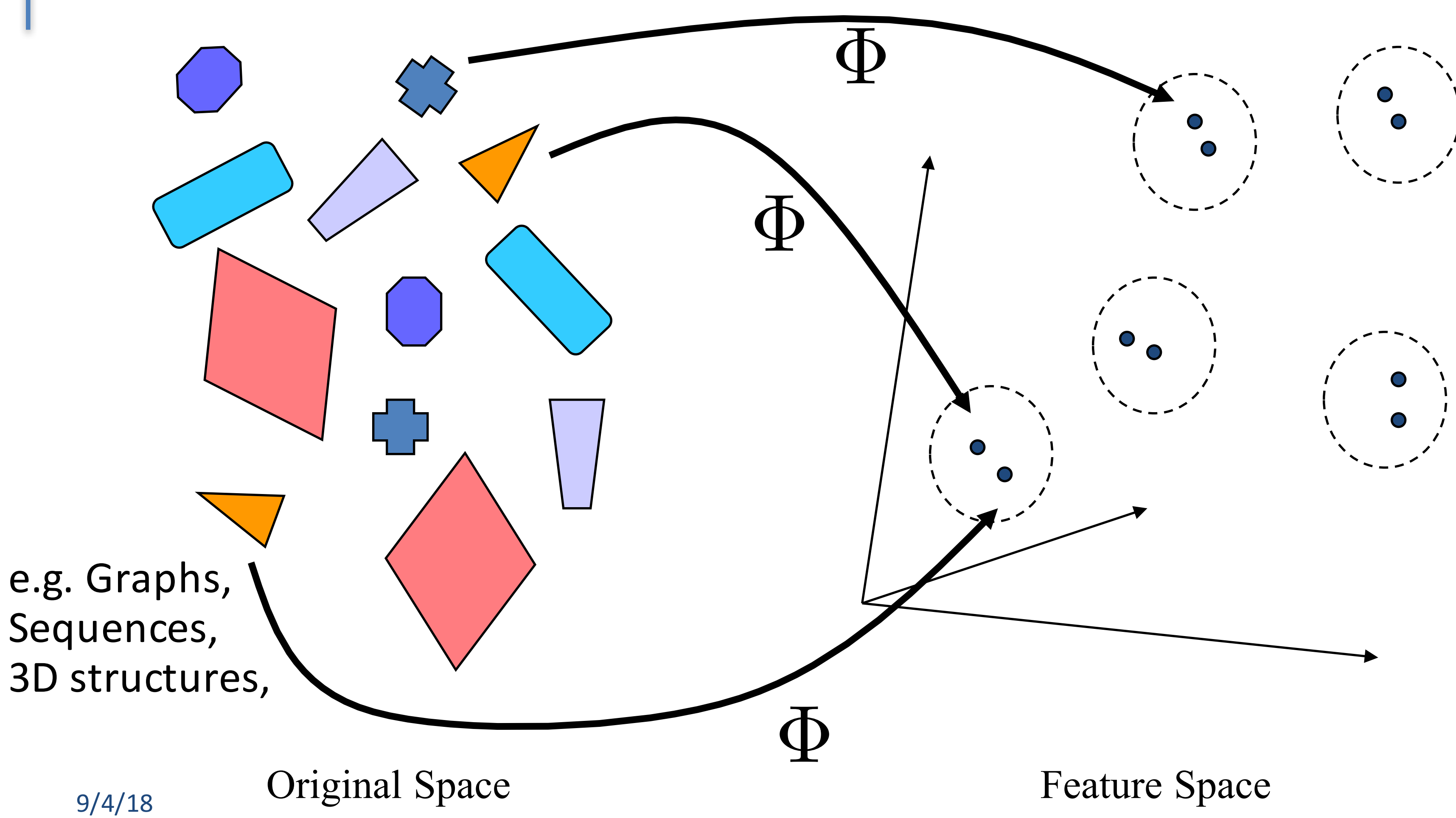
STRUCTURAL OUTPUT LEARNING : [COMPLEXITY OF Y]

- Many prediction tasks involve output labels having structured correlations or constraints among instances

Structured Dependency between Examples' Y	Sequence	Tree	Grid
Input X	APAFSVSPASGACGPECA…	The dog chased the cat	
Output Y	 CCEEEEEECCCCCHHHCCC…		

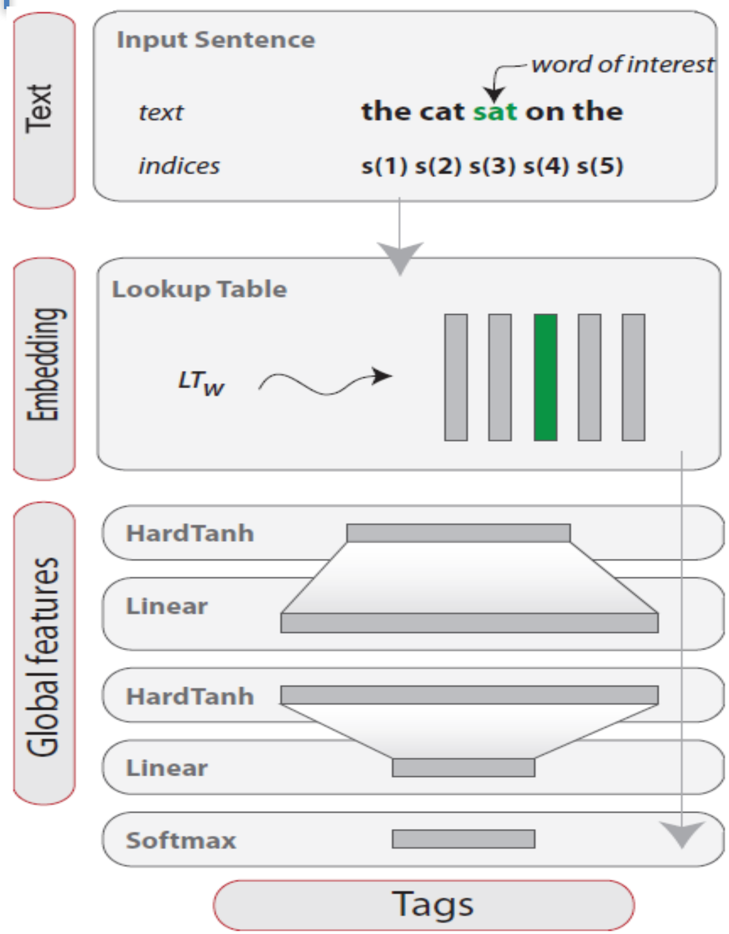
Many more possible structures between y_i , e.g. spatial , temporal, relational …

STRUCTURAL INPUT : Kernel Methods
[COMPLEXITY OF X]

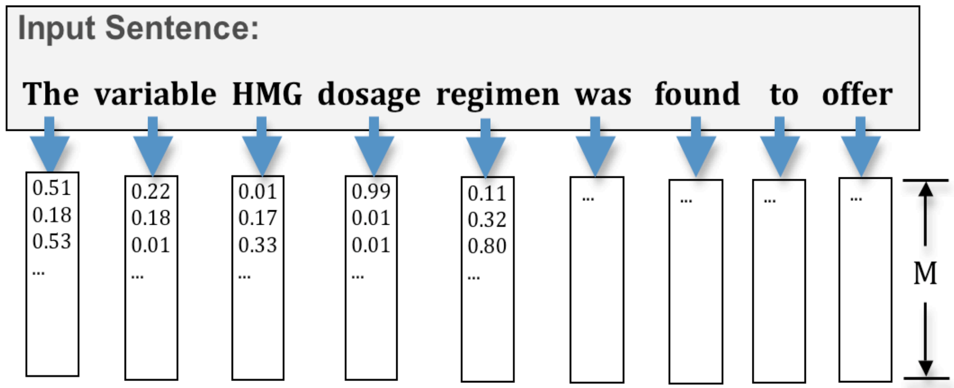


MORE RECENT: FEATURE LEARNING
[COMPLEXITY OF X]

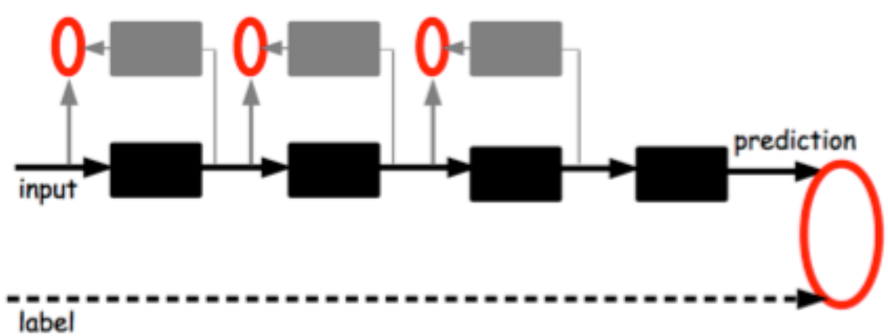
Deep Learning



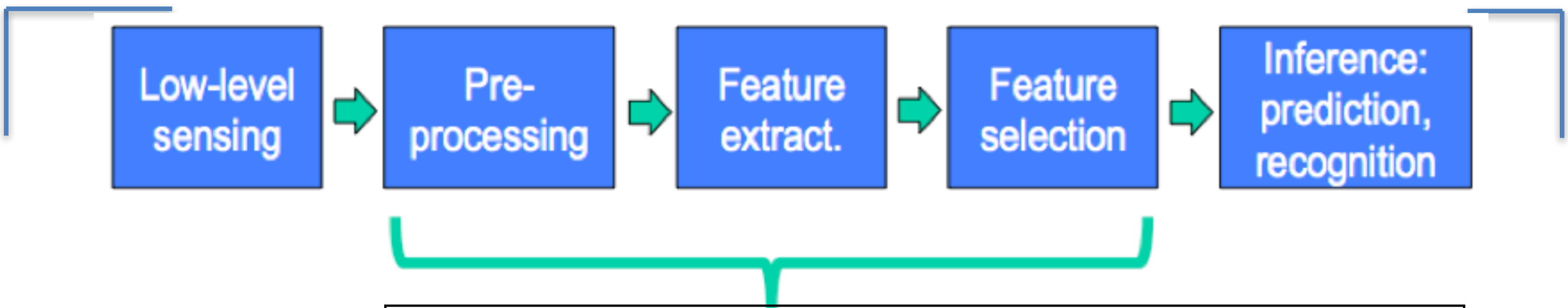
Supervised Embedding



Layer-wise Pretraining



DEEP LEARNING / FEATURE LEARNING : [COMPLEXITY OF X]

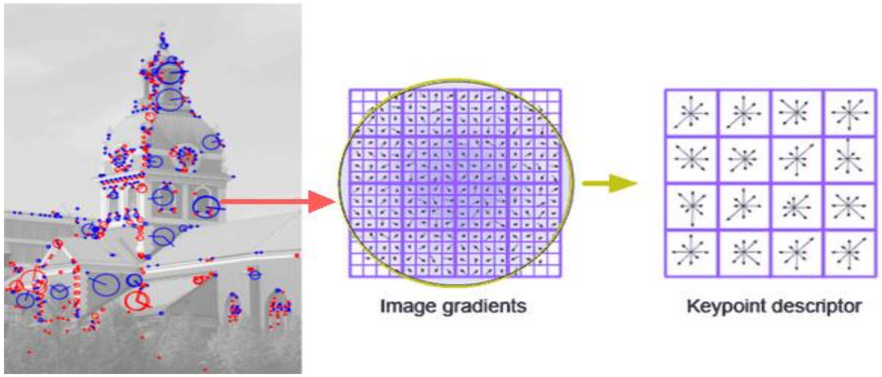


Feature Engineering

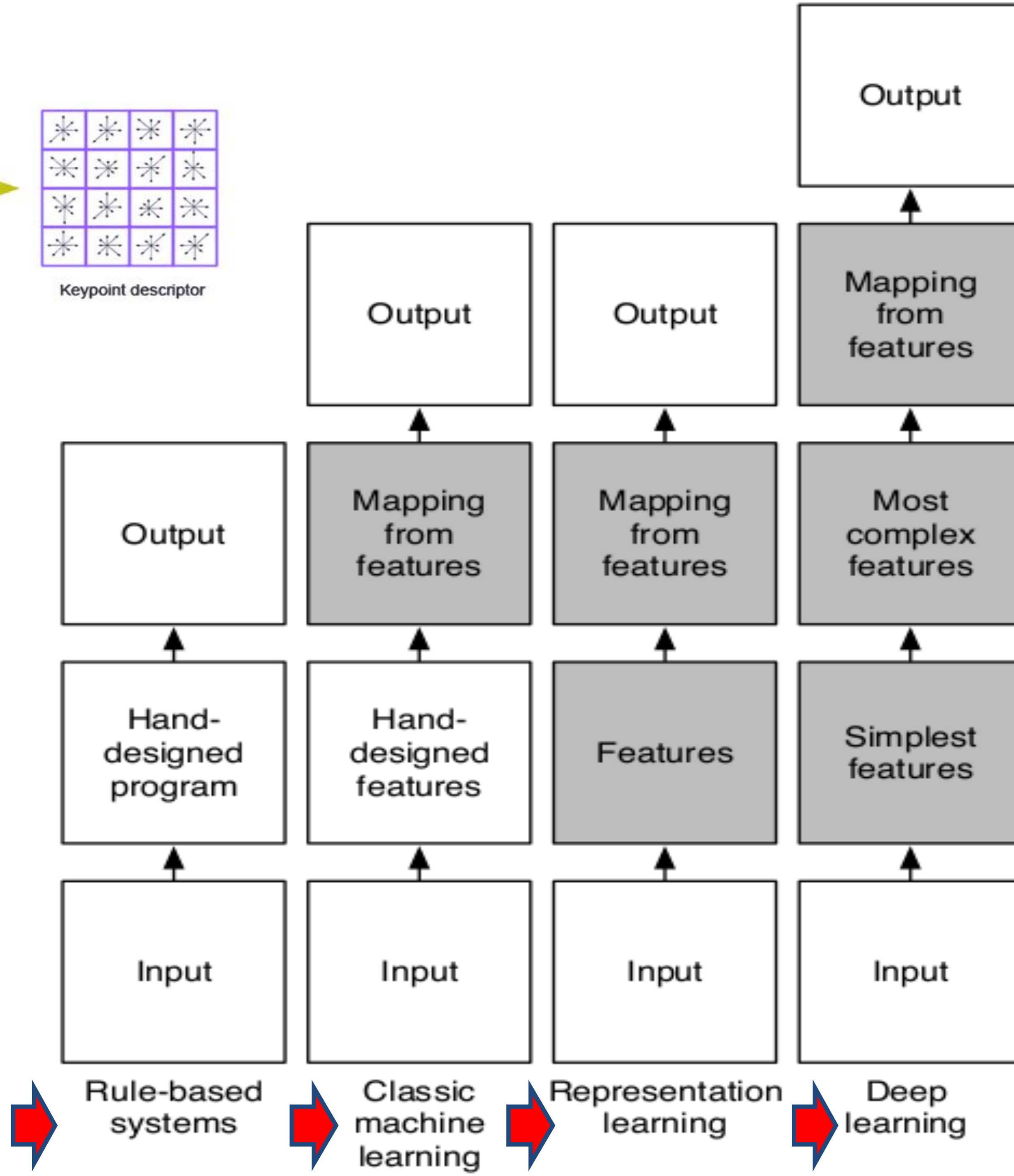
- ✓ Most critical for accuracy
- ✓ Account for most of the computation for testing
- ✓ Most time-consuming in development cycle
- ✓ Often hand-craft and task dependent in practice

Feature Learning

- ✓ Easily adaptable to new similar tasks
- ✓ Layerwise representation
- ✓ Layer-by-layer unsupervised training
- ✓ Layer-by-layer supervised training



Why learn features?



When to use Machine Learning (Adapt to / learn from data) ?

- 1. Extract knowledge from data
  - Relationships and correlations can be hidden within large amounts of data
  - The amount of knowledge available about certain tasks is simply too large for explicit encoding (e.g. rules) by humans

- 2. Learn tasks that are difficult to formalise
  - Hard to be defined well, except by examples, e.g., face recognition

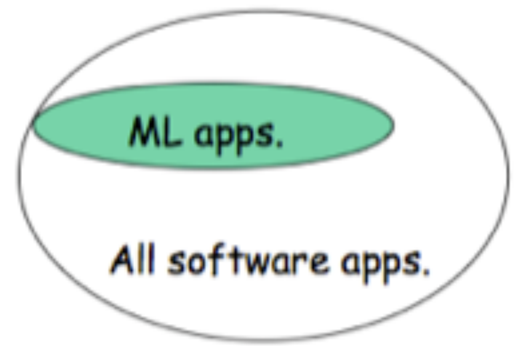
- 3. Create software that improves over time
  - New knowledge is constantly being discovered.
  - Rule or human encoding-based system is difficult to continuously re-design "by hand".

Today

- ❑ Course Logistics
- ❑ Machine Learning Basics
- ❑ Machine Learning History
- ❑ Rough Plan of Course Content

MACHINE LEARNING IN COMPUTER SCIENCE

- Machine learning is already the preferred approach for
  - Speech recognition, natural language processing
  - Computer vision
  - Medical outcome analysis
  - Robot control …



- Why growing ?
  - Improved machine learning algorithms
  - Improved CPU / GPU powers
  - Increased data capture, new sensors, networking
  - Systems/Software too complex to control manually
  - Demand to self-customization for user, environment, ….

HISTORY OF MACHINE LEARNING

- 1950s
  - Samuel's checker player
  - Selfridge's Pandemonium
- 1960s:
  - Neural networks: Perceptron
  - Pattern recognition
  - Learning in the limit theory
  - Minsky and Papert prove limitations of Perceptron
- 1970s:
  - Symbolic concept induction
  - Winston's arch learner
  - Expert systems and the knowledge acquisition bottleneck
  - Quinlan's DT ID3
  - Michalski's AQ and soybean diagnosis
  - Scientific discovery with BACON
  - Mathematical discovery with AM

Adapted From Prof. Raymond J. Mooney's slides

HISTORY OF MACHINE LEARNING (CONT.)

- 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

Adapted From Prof. Raymond J. Mooney's slides

HISTORY OF MACHINE LEARNING (CONT.)

- 2000s
  - Support vector machines
  - Kernel methods
  - Graphical models
  - 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

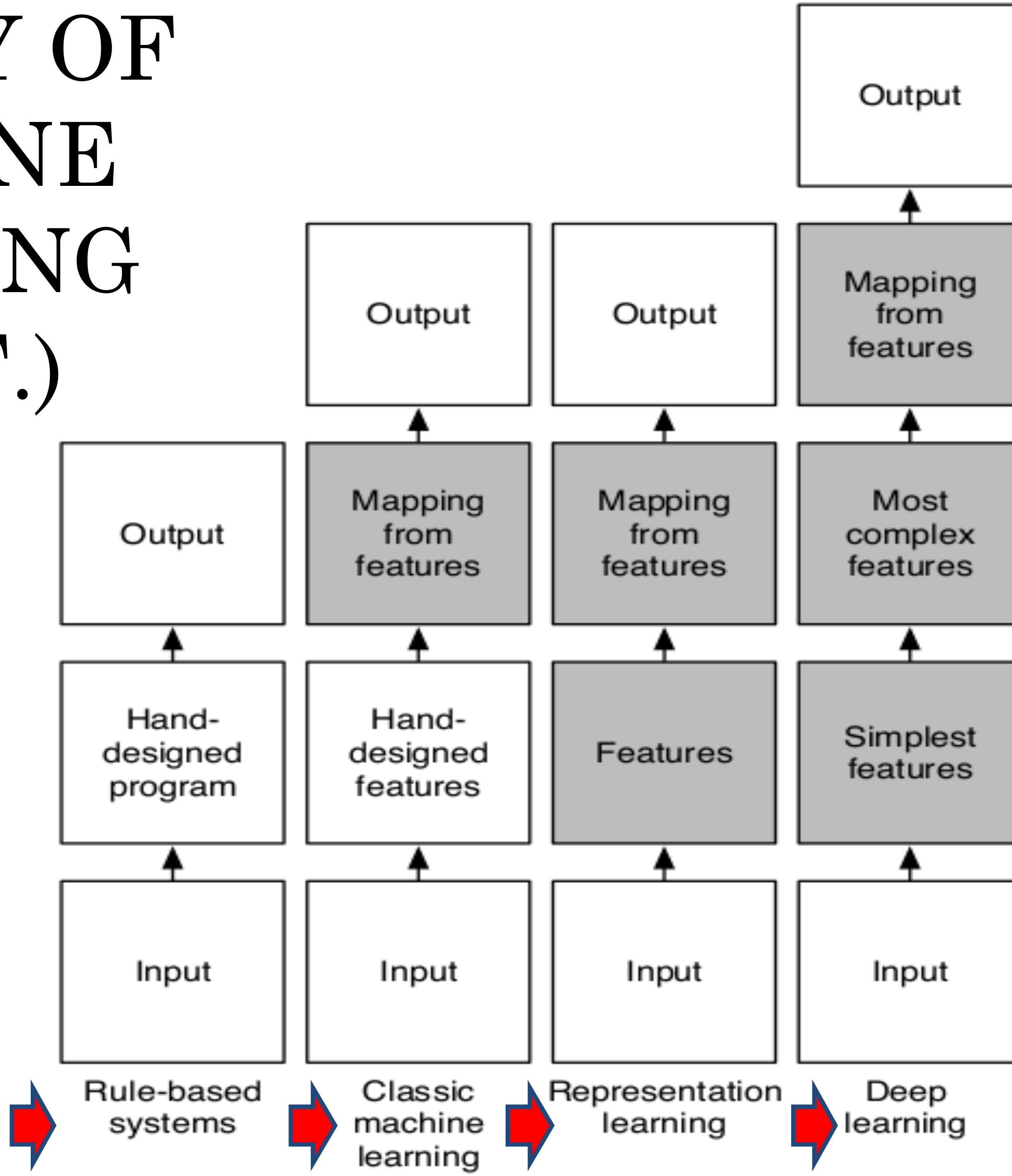
Adapted From Prof. Raymond J. Mooney's slides

HISTORY OF MACHINE LEARNING (CONT.)

- 2010s
  - Speech translation, voice recognition (e.g. SIRI)
  - Google search engine uses numerous machine learning techniques (e.g. grouping news, spelling corrector, improving search ranking, image retrieval, …..)
  - 23 and me (scan sample of person genome, predict likelihood of genetic disease, …)
  - DeepMind, Google Brain, …
  - IBM waston QA system
  - Machine Learning as a service (e.g. google prediction API, bigml.com, cloud autoML .)
  - IBM healthcare analytics
  - ……

Adapted From Prof. Raymond J. Mooney's slides

HISTORY OF MACHINE LEARNING (CONT.)



RELATED DISCIPLINES

- Artificial Intelligence
- Data Mining
- Probability and Statistics
- Information theory
- Numerical optimization
- Computational complexity theory
- Control theory (adaptive)
- Psychology (developmental, cognitive)
- Neurobiology
- Linguistics
- Philosophy

What are the goals of AI research?

Artifacts that THINK like HUMANS	Artifacts that THINK RATIONALLY
Artifacts that ACT like HUMANS	Artifacts that ACT RATIONALLY

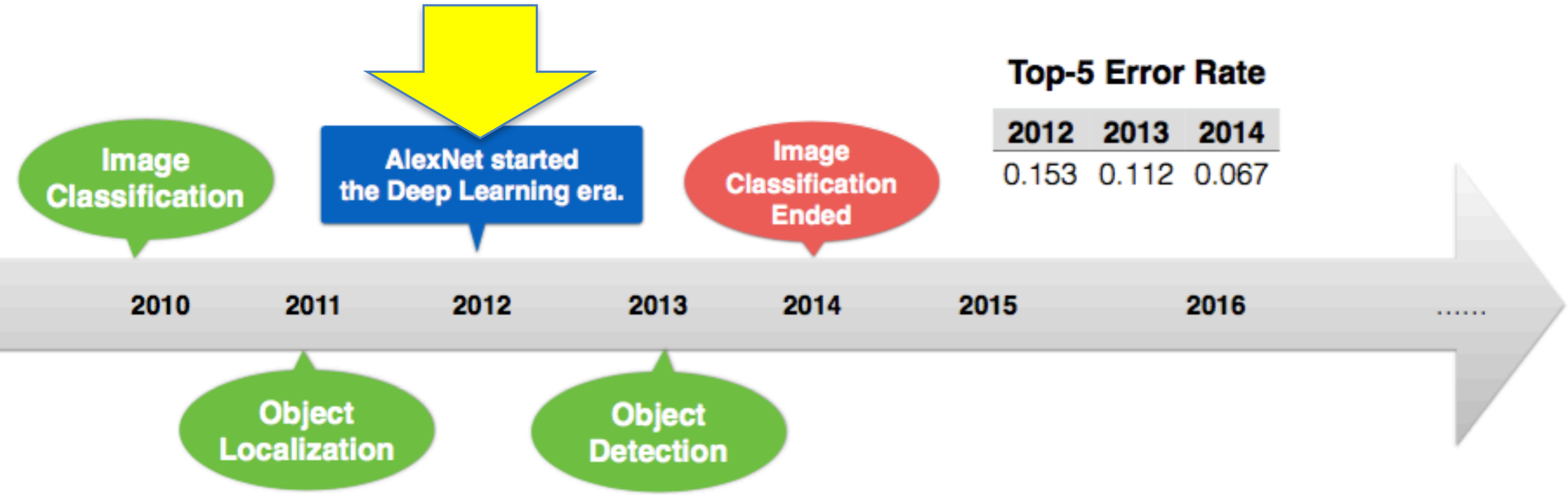
From: M.A. Papalaskar

How can we build more intelligent computer / machine ?

- Able to
  - **perceive the world**
  - **understand the world**
  - **react to the world**

- This needs
  - Basic speech capabilities
  - Basic vision capabilities
  - Language/semantic understanding
  - User behavior / emotion understanding
  - Able to act
  - Able to think ??

How can we build more intelligent computer / machine ? : Milestones in Recent Vision/AI Fields



Top-5 Error Rate

2012	2013	2014
0.153	0.112	0.067

72%, 2010

74%, 2011

85%, 2012

ImageNet Competition:

[Training on 1.2 million images [X] vs. 1000 different word labels [Y]]

Detour: three planned programming assignments about AI tasks

- HW: Semantic language understanding (sentiment classification on movie review text)
- HW: Visual object recognition (labeling images about handwritten digits)
- HW: Audio speech recognition (unsupervised learning based speech recognition task)

Today

- ❑ Course Logistics
- ❑ Machine Learning Basics
- ❑ Machine Learning History
- ❑ Rough Plan of Course Content

Course Content Plan → Five major sections of this course

- ❑ Regression (supervised)
- ❑ Classification (supervised)
- ❑ Unsupervised models
- ❑ Learning theory
- ❑ Graphical models

Summary

- This is not a course about how to use a toolbox
- We focus on learning fundamental principles, mathematical formulation, algorithm design and learning theory.

Some negative comments from last Spring

- Class was boring, …
- The instructor stated that the course was going to be math-heavy which 90% of students did not want, and even the remaining 10% were probably blown away at how intensive it really was…

A FEW SAMPLE SLIDES

L4

$$J(\theta) = (X\theta - y)^T (X\theta - y)\frac{1}{2}$$

$$= ((X\theta)^T - y^T)(X\theta - y)\frac{1}{2}$$

$$= (\theta^T X^T - y^T)(X\theta - y)\frac{1}{2}$$

$$= \left(\theta^T X^T X \theta - \underbrace{\theta^T X^T y - y^T X \theta}_{} + y^T y\right)\frac{1}{2}$$

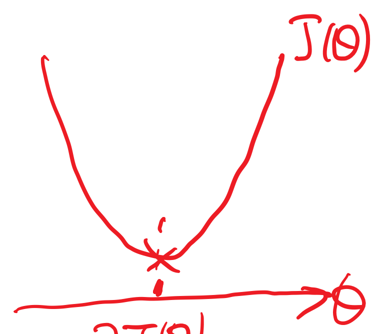
Since $\theta^T X^T y = y^T X \theta$

$\langle X\theta, y \rangle$ $\langle y, X\theta \rangle$

$$= \left(\theta^T X^T X \theta - 2\theta^T X^T y + y^T y\right)\frac{1}{2}$$

$\Rightarrow$ $J(\theta)$ quadratic func of θ;

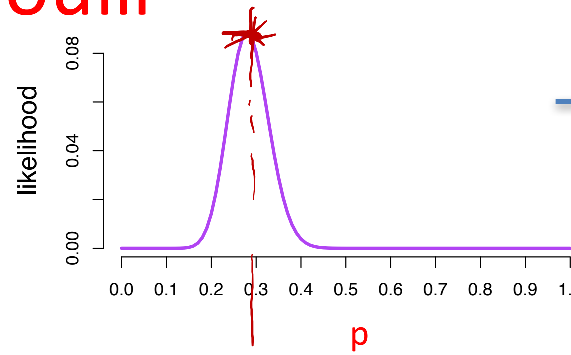
1d case



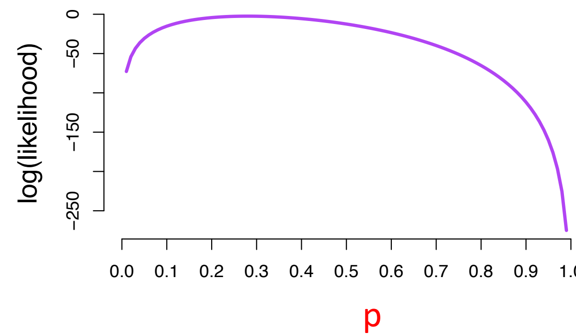
$$\frac{\partial J(\theta)}{\partial \theta} = 0$$

L12: Deriving the Maximum Likelihood Estimate for Bernoulli

$$\overbrace{L(p)}^{\text{maximize}} = p^x(1-p)^{n-x}$$

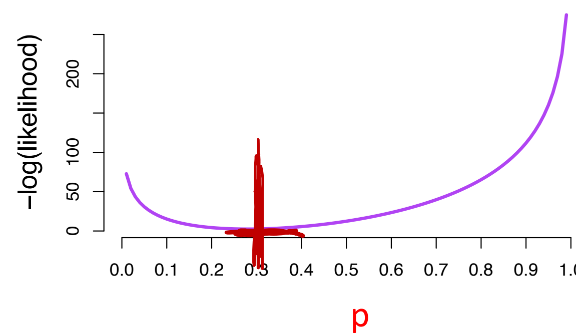


$$\overbrace{\log(L(p)}^{\text{maximize}} = \log\left[p^x(1-p)^{n-x}\right]$$



Minimize the negative log-likelihood

$$\overbrace{-l(p)}^{\text{Minimize the negative log-likelihood}} = -\log\left[p^x(1-p)^{n-x}\right]$$



Next lesson: Machine Learning in a Nutshell

Task: y

↓

Representation: x

↓

Score Function: L()

↓

Search/Optimization : argmin()

↓

Models, Parameters : f, w, b

ML grew out of work in AI

Optimize a performance criterion using example data or past experience,

Aiming to generalize to unseen data

Next lesson: Review of linear algebra and basic calculus

References

- Prof. Andrew Moore's tutorials
- Prof. Raymond J. Mooney's slides
- Prof. Alexander Gray's slides
- Prof. Eric Xing's slides
- http://scikit-learn.org/
- Hastie, Trevor, et al. The elements of statistical learning. Vol. 2. No. 1. New York: Springer, 2009.
- Prof. M.A. Papalaskar's slides