Overview of Machine Learning

Dong Wang

PART I: Introduction

We hold the belief that poetry generation (and other artistic activities) is a pragmatic process and can be largely learned from past experience...

Incomplete understanding you may have

- Machine learning is a set of tools
- Machine learning is a bunch of algorithms
- Machine leraning is pattern recognition
- Machine learning is artificial intelligence

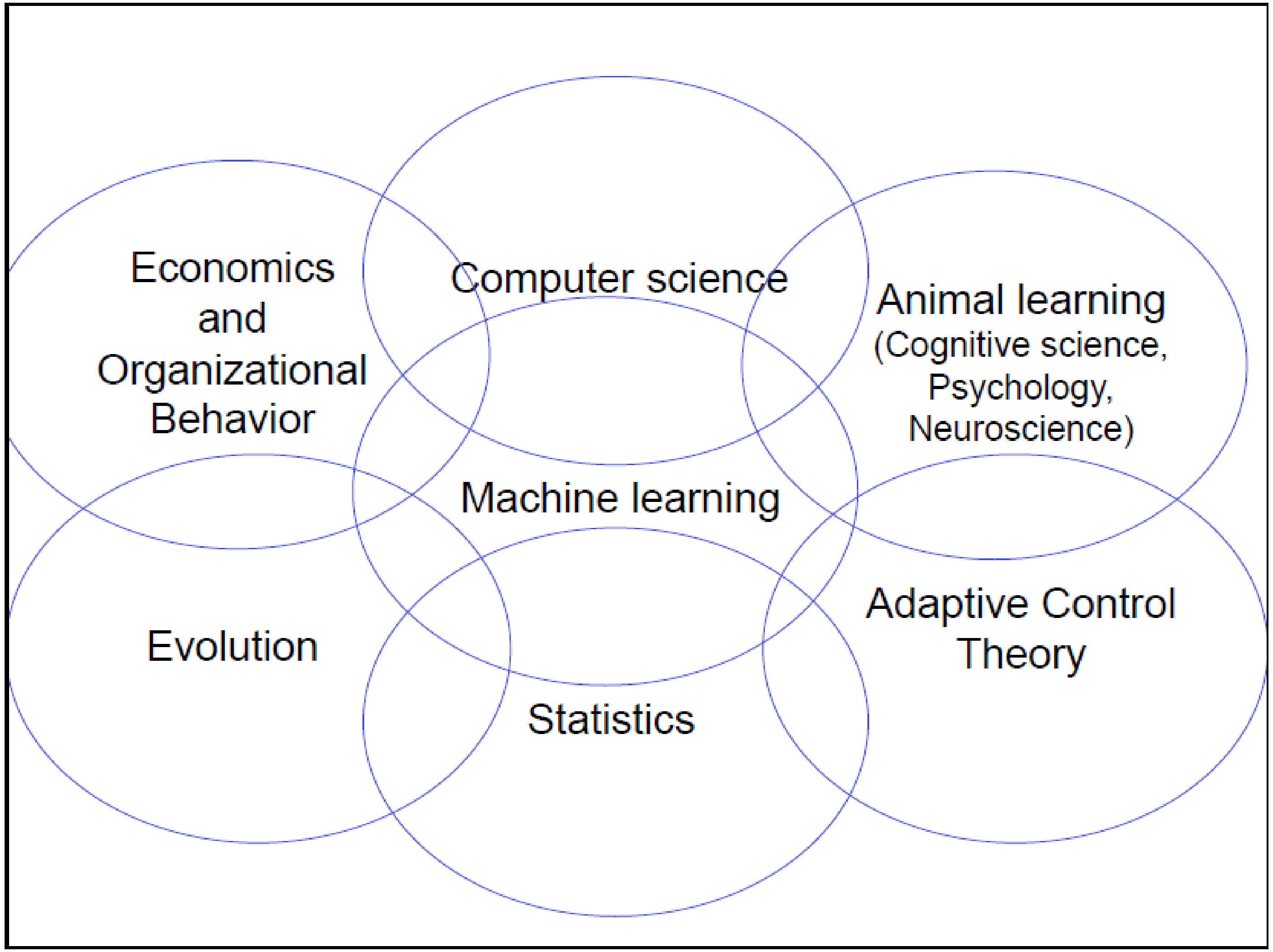
What is machine learning?

- Machine learning is a “Field of study that gives computers the ability to learn without being explicitly programmed” --1959, Arthur Samuel
- A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P if its performance at tasks in T, as measured by P, improves with experience E. -Tom M. Mitchell

Contributors to ML

- Statistics
- Brain models
- Adaptive control theory
- Psychological models
- Artificial intelligence
- Evolutionray models

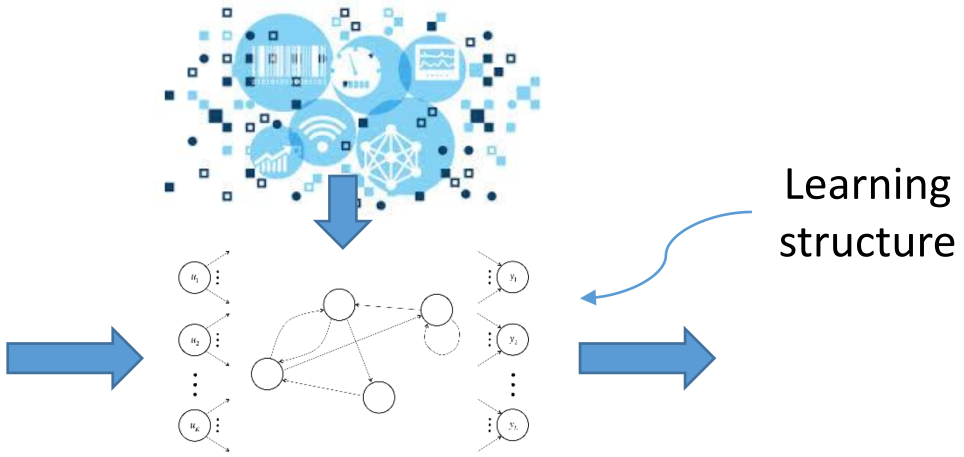
INTRODUCTION TO MACHINE LEARNING , Nils J. Nilsson, Standford, 1998



Tom M. Mitchell, Machine Learning Department
Carnegie Mellon University, Machine Learning 10-701, January 11, 2011.
http://www.cs.cmu.edu/~tom/10701_sp11/slides/DTreesAndOverfitting-1-11-2011_final.pdf

What is machine learning?

- Machine learning is a computing framework that integrates human knowlege and empirical evidence. It is a way of conceptual design: give some model structure (prior), then the algorithm learns within it by experiencing.
- Knowledge (prior) and empirial evidence (samples) are two ends of the spectrum of resources in ML. Different approaches reside in different trade-off positions.
- Ingredients
  - Task
  - Data
  - Learning structure
  - Learning algorithm



Task

- Category
  - From AI perspective
    - Perception
    - Induction
    - Generation
  - From technical perspective
    - Predictive
      - Regression
      - Classification
    - Descriptive
      - Clustering
      - Density estimation
- Objective function
  - xEnt,MSE,Fisher score,sparsity, information
  - Task-dependent (MPE in ASR, e.g.)

Data

- Complexity of data
  - Binary, category, continuous, scale, vector, graph, natural object
  - Dependent or independent
  - Complete or incomplete
  - Dyanics
- Data representation
  - Feature extraction
  - Dimension reduction
  - Data selection

Learning structure

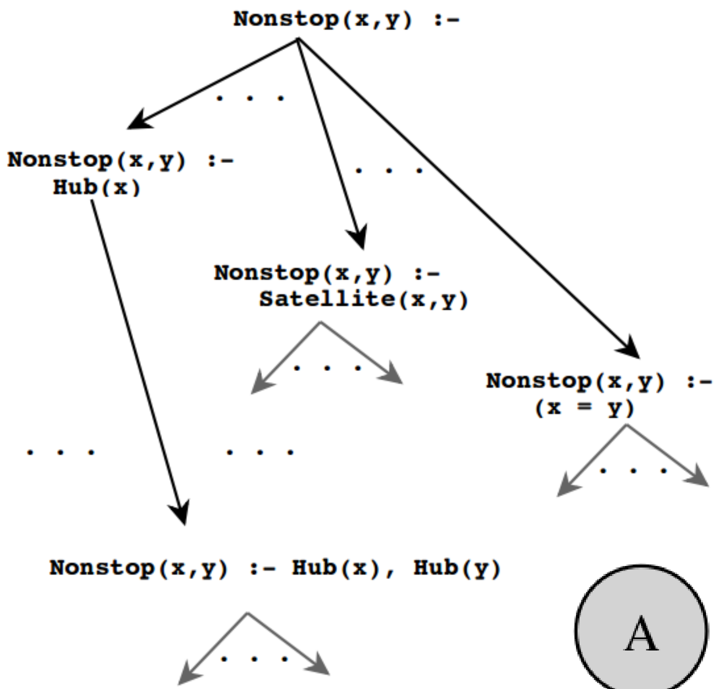
- Functions
- Networks (NN, graph)
- Logic programs and rule sets
- Finite-state machines
- Grammars
- Problem solving systems

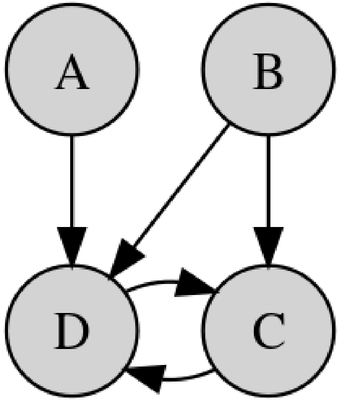
Learning algorithms

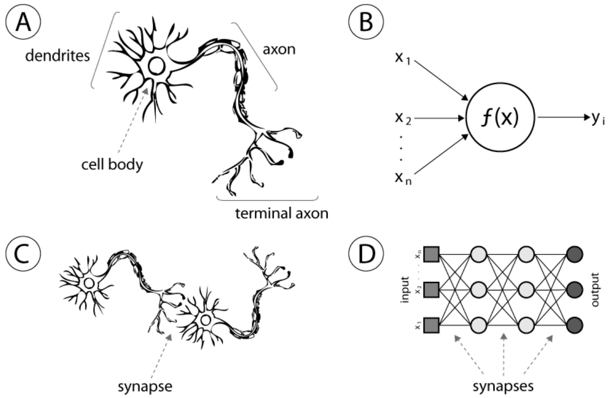
- Supervision
  - Supervised learning, unsupervised learning, semisupervised learning, reinforcement learning
- Model
  - Probabilistic model (GMM,HMM,pLDA,...)
  - Neural model (MLP,LSTM,...)
  - Distance-based model (kNN, metric learning, SVM)
  - Information-based model (ME, CART,...)
  - Other criteria (LDA,...)
- Learning approach
  - Direct solution
  - Neumerical optimization
  - Evolution

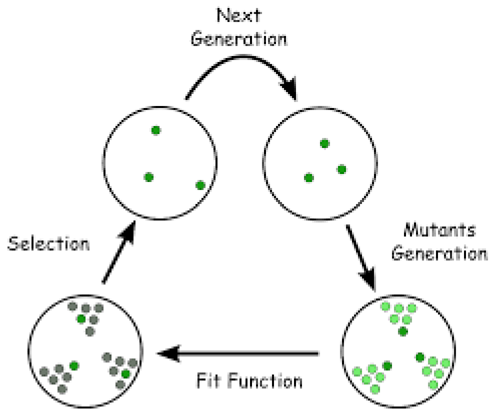
Four paradigms

- Symbolic learning
  - Inductive Logic Programming
- Bayesian learning
- Neural learning
- Evolutionary learning









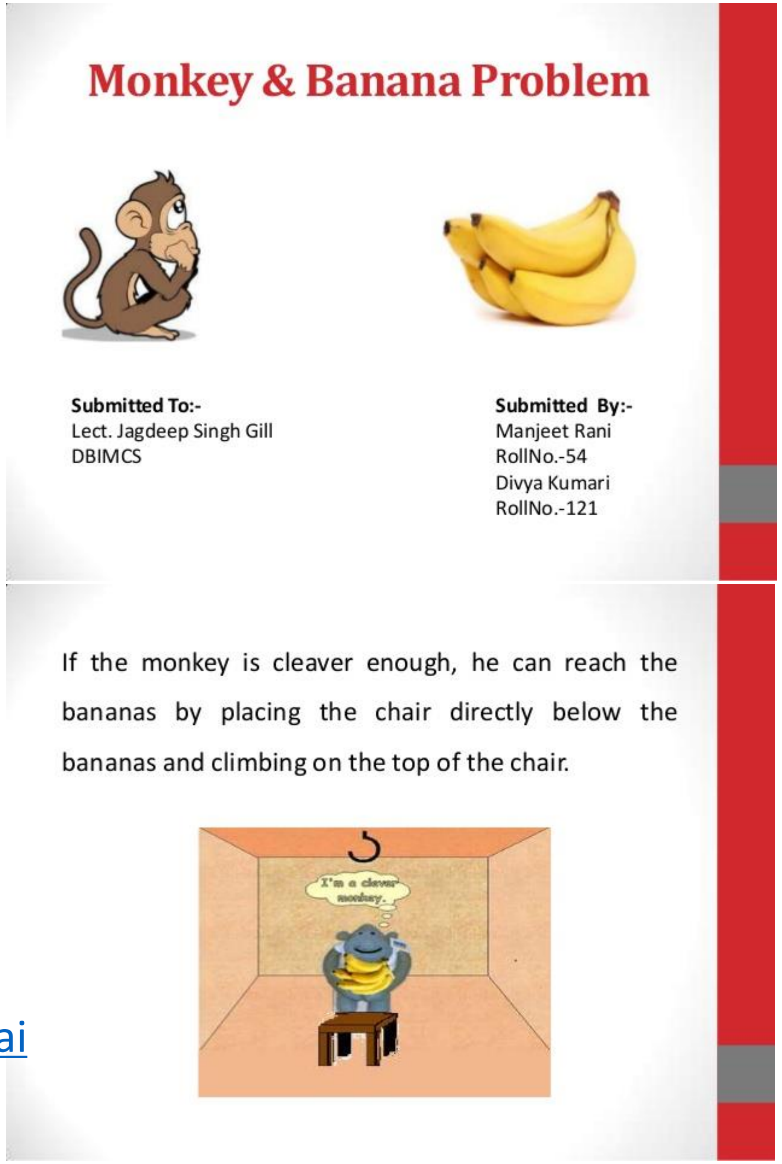
Why learn?

- We can not design much
  - It is hard to design everything (we don’t know the exact process)
  - It is hard to design even one thing (limited knowledge,dynamics, inaccuracy)
  - Trade-off between “explicit design with ASSUMPTION” and “conceptual design with approximation“
  - It is a black box with unknown process, but could be better than a white box with presumebly known (but in fact wrong or inaccurate) process.
  - Let data tells you more!

- Intelligence is from experience.
- It is the way we human do every day.

Example 1: Monkey master

- Task: get the banana
- Symbolic approach
  - design feature, knowlege structure, induction rules, and do search
- Learning approach
  - Let the monkey try many times...
  - Can play games very well

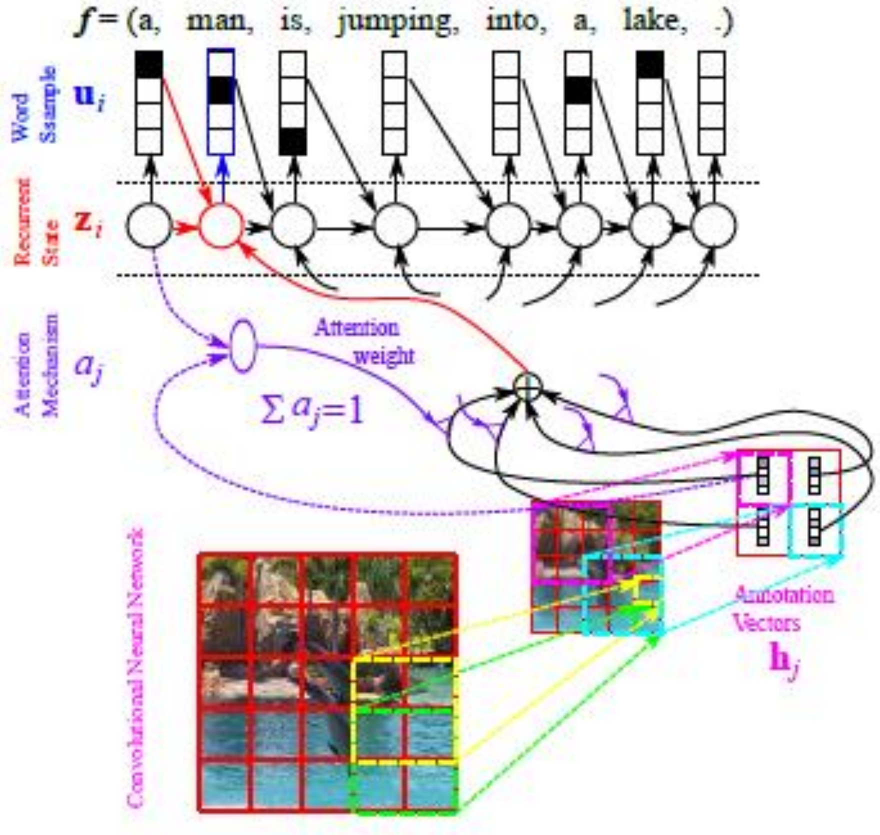


http://www.slideshare.net/ManjeetKamboj/monkey-banana-problem-in-ai

Human-level control through deep reinforcement learning, Nature

Example 2: Image semantics

- CV approach:
  - edge detection, image participation, pattern matching
- Learning approach:
  - End to end



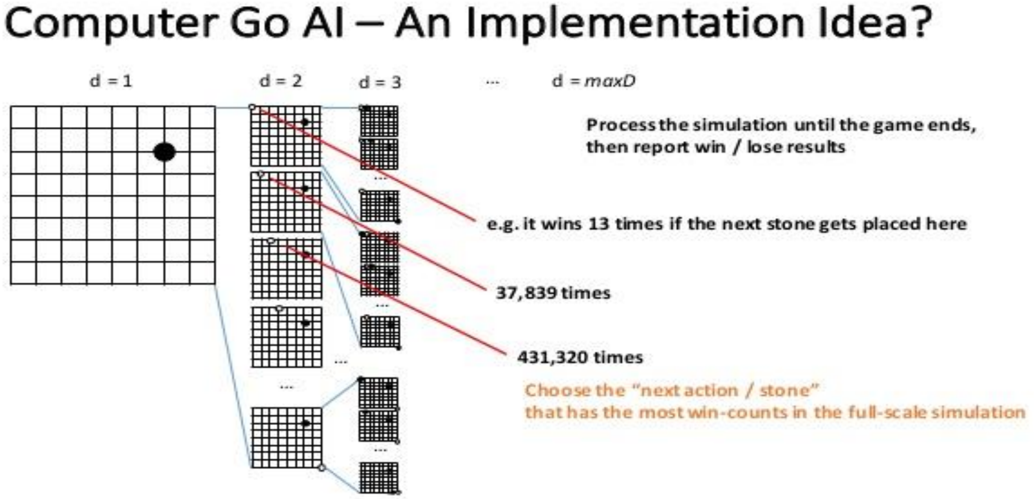
A woman is throwing a frisbee in a park.

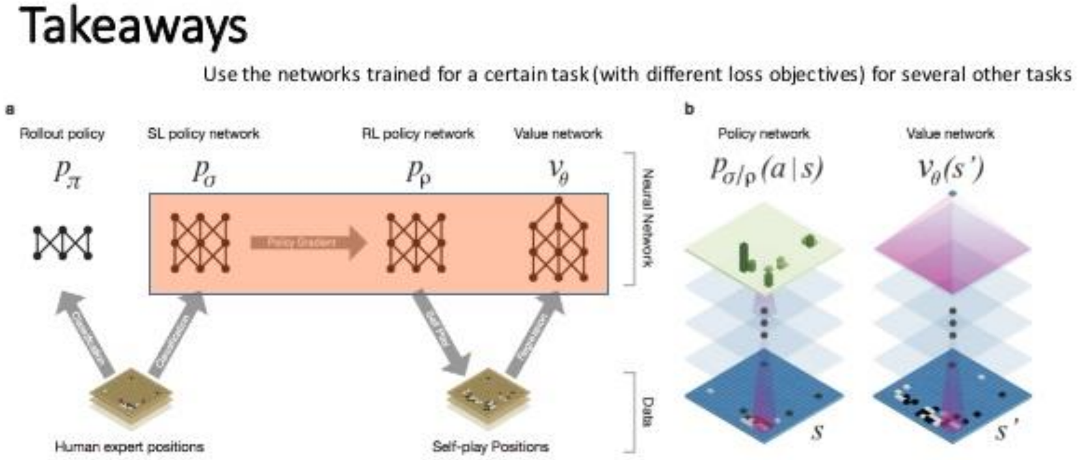
A little girl sitting on a bed with a teddy bear.

Kyunghyun Cho, Aaron Courville and Yoshua Bengio?, Describing Multimedia Content using Attention-based Encoder–Decoder Networks, arxIV

Example 3: Alapha Go

- AI approach: heuristic path search.
- Learning approach: not know exact things, but know more inexact things.



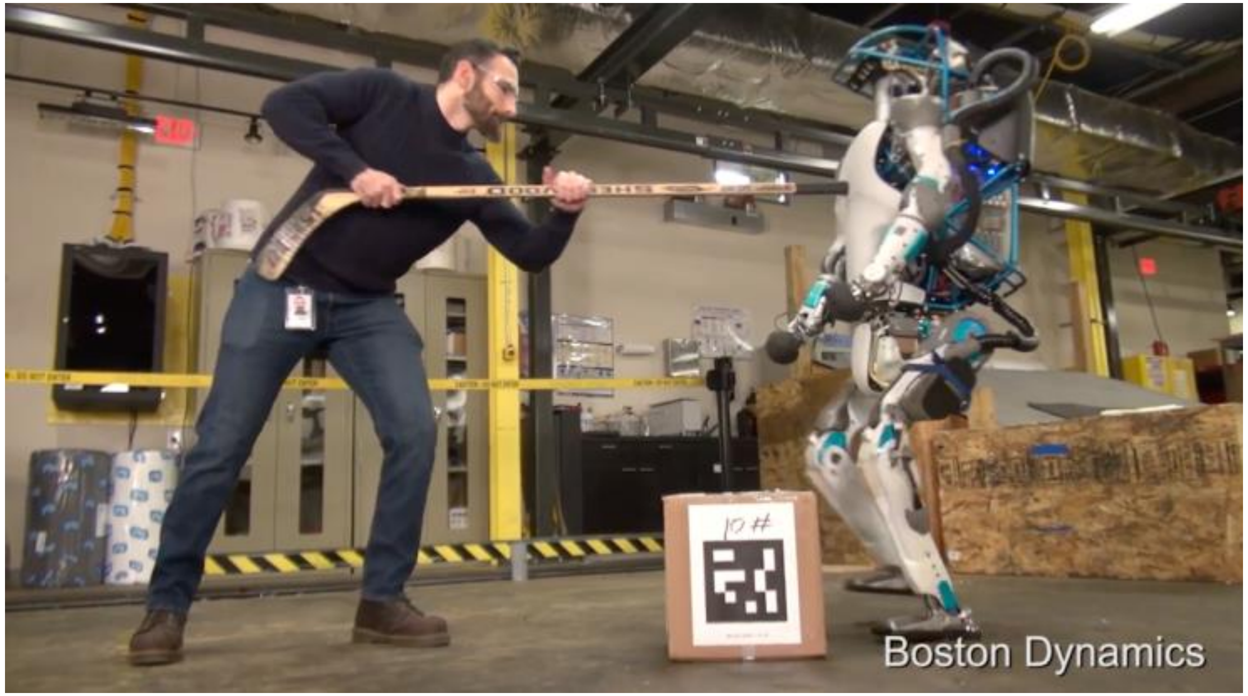


http://www.slideshare.net/ShaneSeungwhanMoon/how-alphago-works

Example 4: Robot

- Human design approach:
  - Compute gravity, arm angle, force, velocity, make decision
  - Do we do that?
- Learning from experience

http://www.inf.ed.ac.uk/teaching/courses/mlsc/



Example 5: Financial trading

- Financial approach:
  - Design model, select parameters, predict, game theory..
- ML approach:
  - Learn trader’s operation
  - Learn time series
  - Reinforcement learning

http://mp.weixin.qq.com/s?__biz=MzA4NTk5MjY0MA==&mid=2659053801&idx=1&sn=38cf0117fd6b8b2b5060a834c8a162e5&scene=1&srcid=0615AXObul44JTg2V2zbCfnV#rd

Where is the frontend?

- Deep and complex learning with big data and computing graph
- Human-like learning (one-shot, collaborative,transfer,...)
- Creativity (motivation,emotion,artist)

Can Turing machine be curious about its Turing test results? Three informal lectures on physics of intelligence
https://arxiv.org/abs/1606.08109

Change your mind

- If you are from engineering
  - Pay more attention on theory
  - Don’t try

- If you are from mathematics
  - Refrain from rigorous equation design
  - But pay attention to rigorous statistics equation design
  - Pay more attention on data, randomness
  - Do try

FAQ

- Is ML hard?
  - Yes, many many algorithms, theories, change quickly, all confusing
  - No, in most cases the algorithms follow similar threads and easy to understand
  - And it is fascinating!
- What you need to prepare?
  - Algebra, particlarly matrix operations and eigen analysis
  - Statistics, particularly Gaussian
  - Prepare to thinking, global thinking
  - Focus, and agile to new things
  - Hard work

- QA

PART II: Basic concepts

Learning is a set of trade off : data and model, complexity and efficiency, memory and time, fitting and generalization....

Machine learning

$$y = f(\mathbf{x})$$

output

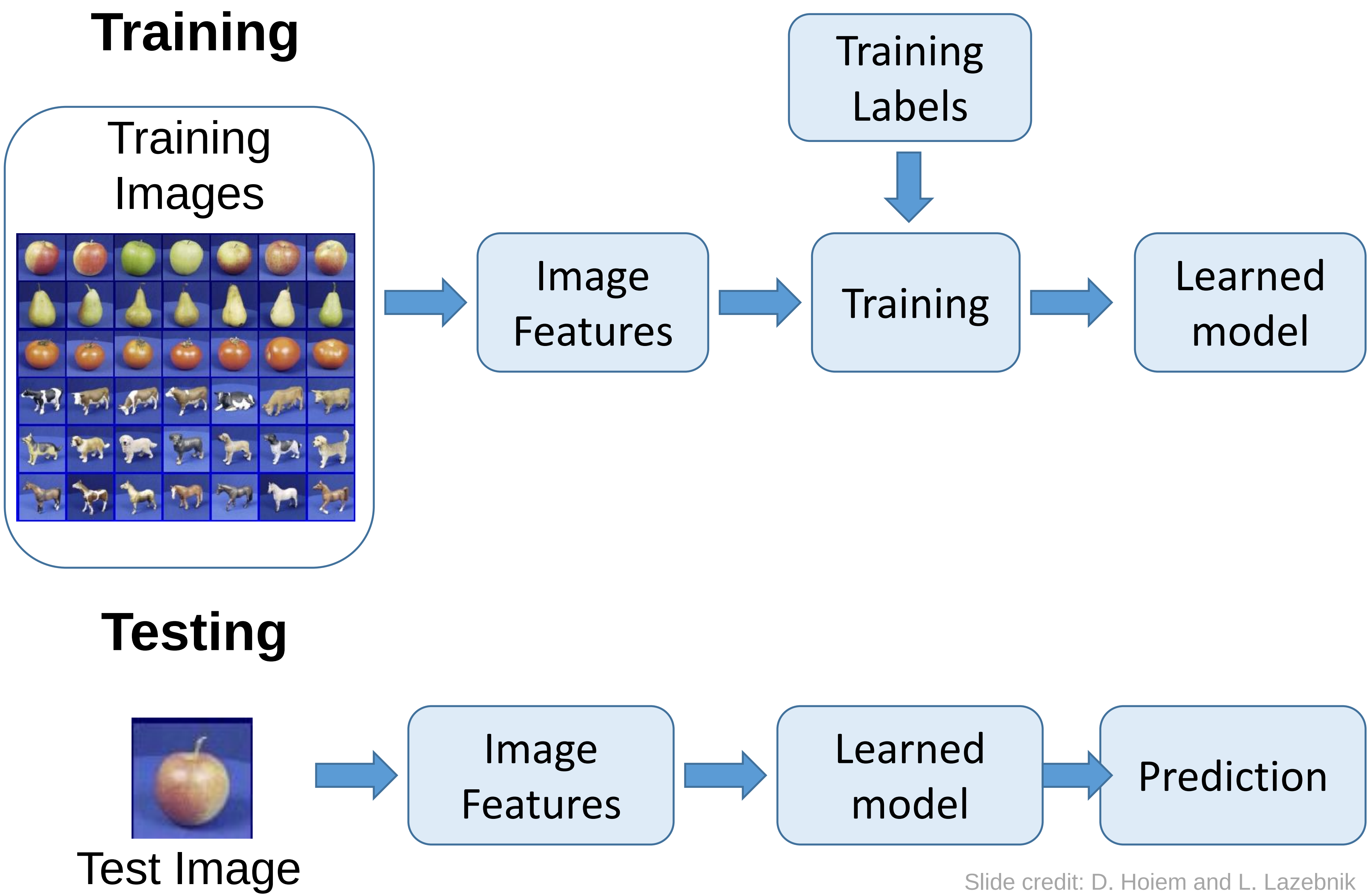
prediction function

Image feature

- **Training:** given a *training set* of labeled examples $\{(\mathbf{x}_1,y_1), \ldots, (\mathbf{x}_N,y_N)\}$, estimate the prediction function f by minimizing the prediction error on the training set
- **Testing:** apply f to a never before seen *test example* $\mathbf{x}$ and output the predicted value $y = f(\mathbf{x})$

Slide credit: L. Lazebnik

Basic steps



Slide credit: D. Hoiem and L. Lazebnik

Fitting and generalization

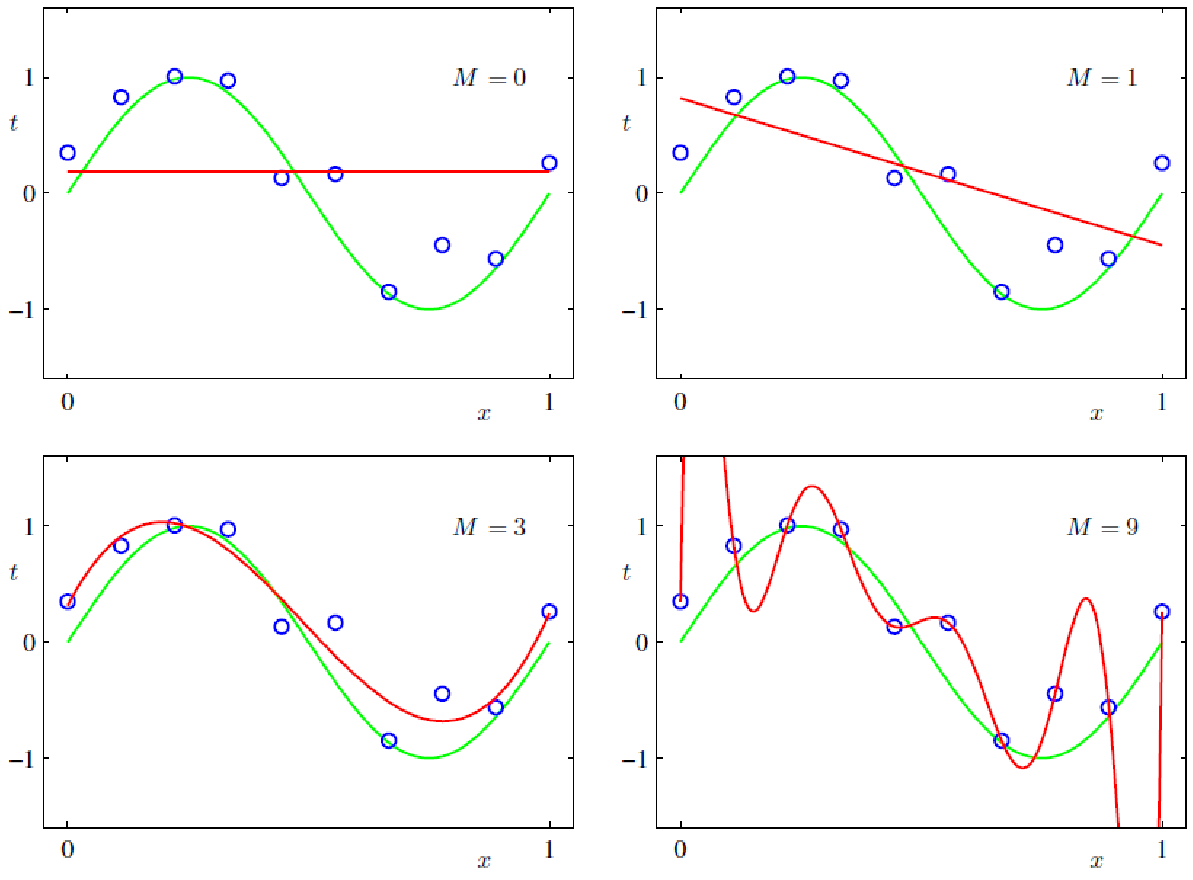
Training set (labels known)

Test set (labels unknown)

- How well does a learned model generalize from the data it was trained on to a new test set?

Slide credit: L. Lazebnik

What model to use?



PRML, Bishop, Fig.1.4

Bias-Variance Trade-off

- Let cost function

Prediction error

Noise on t

$$\mathbb{E}[L] = \int \{y(\mathbf{x}) - h(\mathbf{x})\}^2 \, p(\mathbf{x}) \, \mathrm{d}\mathbf{x} + \int \{h(\mathbf{x}) - t\}^2 p(\mathbf{x}, t) \, \mathrm{d}\mathbf{x} \, \mathrm{d}t.$$

$$\begin{aligned}
&\{y(\mathbf{x};\mathcal{D}) - \mathbb{E}_{\mathcal{D}}[y(\mathbf{x};\mathcal{D})] + \mathbb{E}_{\mathcal{D}}[y(\mathbf{x};\mathcal{D})] - h(\mathbf{x})\}^2 \\
&\quad = \{y(\mathbf{x};\mathcal{D}) - \mathbb{E}_{\mathcal{D}}[y(\mathbf{x};\mathcal{D})]\}^2 + \{\mathbb{E}_{\mathcal{D}}[y(\mathbf{x};\mathcal{D})] - h(\mathbf{x})\}^2 \\
&\quad\quad +2\{y(\mathbf{x};\mathcal{D}) - \mathbb{E}_{\mathcal{D}}[y(\mathbf{x};\mathcal{D})]\}\{\mathbb{E}_{\mathcal{D}}[y(\mathbf{x};\mathcal{D})] - h(\mathbf{x})\}.
\end{aligned}$$

$$\begin{aligned}
&\mathbb{E}_{\mathcal{D}}\left[\{y(\mathbf{x};\mathcal{D}) - h(\mathbf{x})\}^2\right] \\
&\quad = \underbrace{\{\mathbb{E}_{\mathcal{D}}[y(\mathbf{x};\mathcal{D})] - h(\mathbf{x})\}^2}_{(\text{bias})^2} + \underbrace{\mathbb{E}_{\mathcal{D}}\left[\{y(\mathbf{x};\mathcal{D}) - \mathbb{E}_{\mathcal{D}}[y(\mathbf{x};\mathcal{D})]\}^2\right]}_{\text{variance}}.
\end{aligned}$$

PRML, Bishop, Eq 3.37,3.39, 3.40

Bias-Variance Trade-off

$$E(MSE) = noise^2 + bias^2 + variance$$

Unavoidable error on target

Error due to incorrect assumptions

Error due to variance of training samples

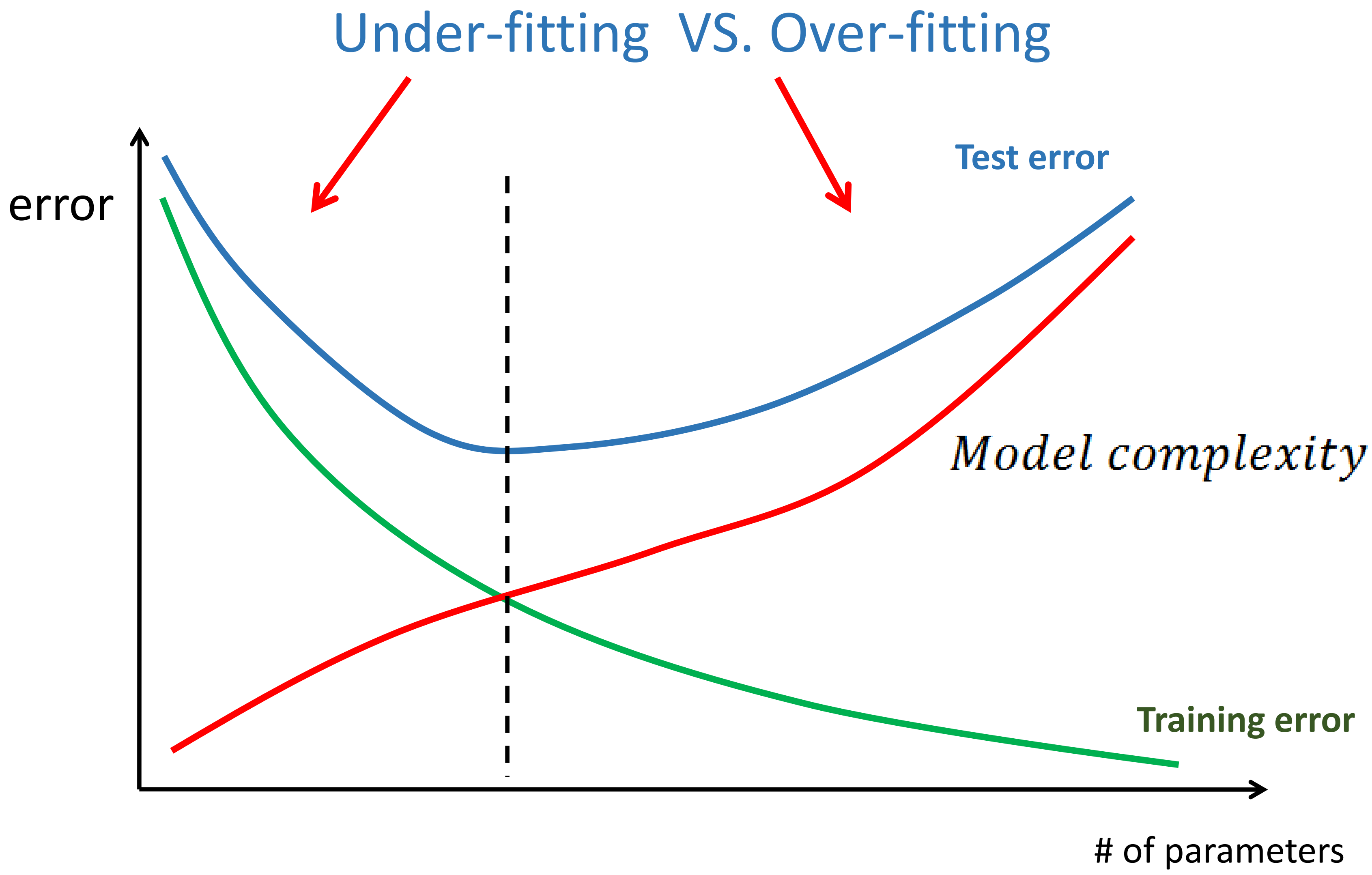
Slide credit: D. Hoiem

Training and generalization

- Components of training error
  - **Bias:** how much the average model over all training sets differ from the true model?
    - Error due to inaccurate assumptions/simplifications made by the model
  - **Noise:** error due to the target randomness, e.g., measure inaccuracy or incorrect labels
- Additional componet of generalization error
  - **Variance:** how much models estimated from different training sets differ from each other

Slide credit: L. Lazebnik

Training and generalization



Underfitting: model is too "simple" to represent all the relevant class characteristics

- High bias and low variance
- High training error and high test error

Overfitting: model is too "complex" and fits irrelevant characteristics (noise) in the data

- Low bias and high variance
- Low training error and high test error

Occam’s razor

- Prefer simplest hypothesis that fits the data
- Various regularizations to enforce the data simpler
  - Constraints on task
  - Easier training
  - Better statistics

No free lunch…

- No classifier is inherently better than any other: you need to make assumptions to generalize
- The better the assumption fits the data, the better the model.

Slide credit D. Hoiem

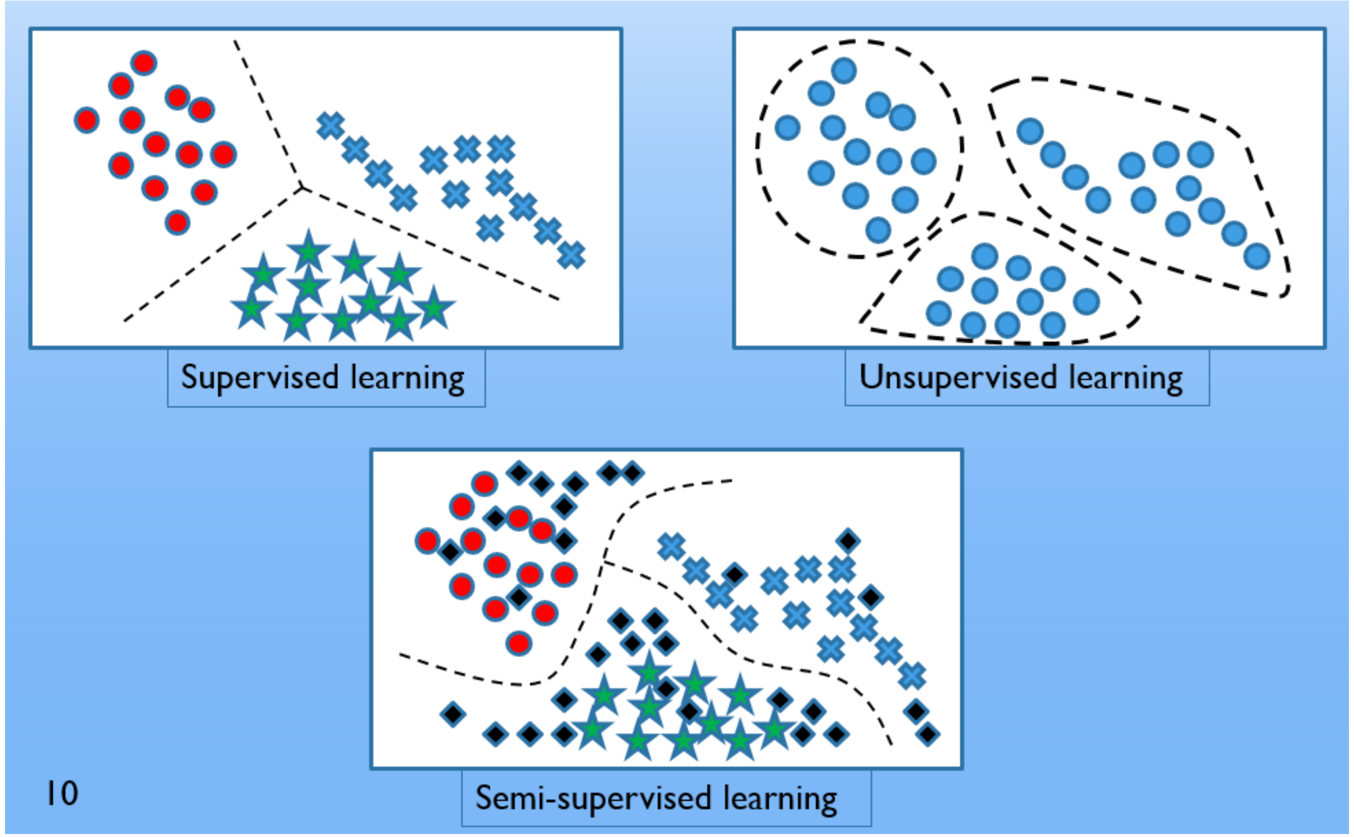
How to deal with a given task?

- Set objective function: encodes the right loss for the problem
- Set model structure: makes assumptions that fit the problem
- Set regularization: right level of regularization
- Set training algorithm: can find parameters that maximize objective on the training set
- Set inference algorithm: can solve for objective function in evaluation

Slide credit: D. Hoiem

Some arguments

- Linear & nonlinear
- Supervised & unsupervised
- Genreative model & discriminative model
- Bayeisan & neural



Some typical models

- Supervised learning categories and techniques
  - **Linear** (linear regression, logistic regression)
  - **Nolinear** (SVM, NN)
  - **Parametric** (Probabilistic functions)
    - Naïve Bayes, Gaussian discriminant analysis (GDA), Hidden Markov models (HMM), Probabilistic graphical models
  - **Non-parametric** (Instance-based functions)
    - *K*-nearest neighbors, Kernel regression, Kernel density estimation, Local regression
    - Classification and regression tree (CART), decision tree
  - **Aggregation**
    - Bagging (bootstrap + aggregation), Adaboost, Random forest

Some typical models

- Unsupervised learning categories and techniques
  - **Clustering**
    - K-means clustering
    - Spectral clustering
  - **Density Estimation**
    - Gaussian mixture model (GMM)
    - Graphical models
  - **Dimensionality reduction**
    - Principal component analysis (PCA)
    - Factor analysis

Some resources

- New member reading list
  - http://cslt.riit.tsinghua.edu.cn/mediawiki/index.php/New_member_reading_list
- Research tools
  - http://cslt.riit.tsinghua.edu.cn/mediawiki/index.php/Public_Research_Tools
- Free data
  - http://cslt.riit.tsinghua.edu.cn/mediawiki/index.php/Data_resources

- Q&A