CSE 847: Machine Learning

Introduction

Jiayu Zhou
Computer Science & Engineering
Michigan State University

Outline of lecture

- Course information
- Project
- Course schedule
- Introduction

Course Information

- Instructor: Dr. Jiayu Zhou
  - Office: EB 2134
  - Office hours: TTh 4:00pm-5:00pm
  - Email: jiayuz@msu.edu

- Lecture
  - Meeting Time: TTh 2:40pm—4:00pm
  - Location: EB 1300

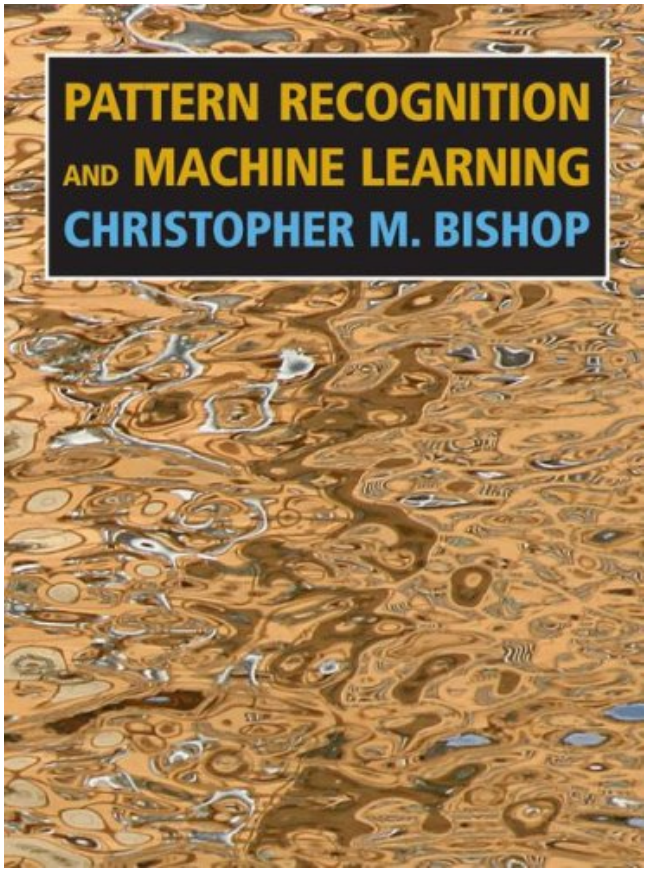
Course information (Cont'd)

- Objective: An in-depth understanding of machine learning and statistical pattern recognition techniques and their applications in biomedical informatics, computer vision, and other domains.

- Topics: Probability distributions, regression, classification, kernel methods, clustering, semi-supervised learning, mixture models, graphical models, dimensionality reduction, manifold learning, sparse learning, multi-task learning, transfer learning, and Hidden Markov Models.

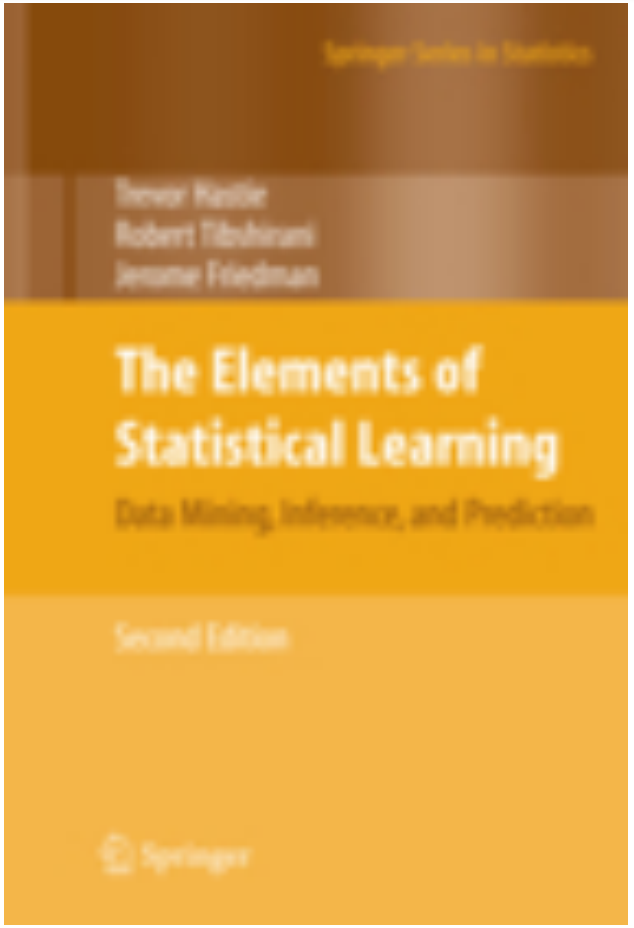
Course information (Cont'd)

- Prerequisite: Basics of linear algebra, probability, algorithm design and analysis.
- Course textbook: Pattern Recognition and Machine Learning, Christopher M. Bishop, 2006.



Course information (Cont'd)

- Reference book: The Elements of Statistical Learning: Data Mining, Inference, and Prediction (Second Edition) by Trevor Hastie, Robert Tibshirani and Jerome Friedman (2009)



http://www-stat.stanford.edu/~hastie/Papers/ESLII.pdf

Grading

- Homework (6): 40%
- Project: 25%. One to three students form a group to carry out a research project.
- Exam (2): 30%.
- Class participation: 5%. Students are required to attend the lecture and participate in the class discussion.

Project

- Project proposal (one page) is due on **2/19/16**
  - (1) project title
  - (2) team members
  - (3) description of the problem you try to address
  - (4) preliminary plan (milestones)
  - (5) paper list

- The intermediate project report (3-5 pages) is due on **3/18/16**
  - (1) a high quality introduction and problem description
  - (2) description of the data used in the project
  - (3) what have you done so far
  - (4) what remains to be done

Project (Cont'd)

- The final project report (10-15 pages) is due on **4/28/16**
  - (1) Introduction, including a summary of the problem, previous work, methods, and results
  - (2) Problem description, including a detailed description of the problem you try to address
  - (3) Methodology, including a detailed description of methods used
  - (4) Results, including a detailed description of your observations from the experiments
  - (5) Conclusions and future work, including a brief summary of the main contributions of the project and the lessons you learn from the project, as well as a list of some potential future work.

Programming language

- Matlab
  - Tutorials
    - http://www.math.ufl.edu/help/matlab-tutorial/
    - http://www.math.mtu.edu/~msgocken/intro/node1.html

LaTeX

- You are strongly encouraged to use *LaTeX* for the project proposal and reports.

Theorem 1 (Residue Theorem). Let f be analytic in the region G except for the isolated singularities $a_1, a_2, \ldots, a_m$. If γ is a closed rectifiable curve in G which does not pass through any of the points a_k and if $\gamma \approx 0$ in G then

$$\frac{1}{2\pi i}\int_\gamma f = \sum_{k=1}^{m} n(\gamma; a_k)\mathrm{Res}(f; a_k).$$

Theorem 2 (Maximum Modulus). *Let G be a bounded open set in $\mathbb{C}$ and suppose that f is a continuous function on G^- which is analytic in G. Then*

$$\max\{|f(z)| : z \in G^-\} = \max\{|f(z)| : z \in \partial G\}.$$

$\mathrm{A\Lambda\Delta\nabla BCD\Sigma EF\Gamma GHIJKLMNO\Theta\Omega\mho P\Phi\Pi\Xi QRSTUVWXY\Upsilon\Psi Z}$ 1234567890

$a\alpha b\beta c\partial d\delta e\epsilon\varepsilon f\zeta\xi g\gamma h\hbar\hbar i\imath j\jmath k\kappa\varkappa l\ell\lambda mn\eta\theta\vartheta o\sigma\varsigma\phi\varphi p\wp p\rho\varrho qrst\tau\pi u\mu v\upsilon vw\omega\varpi x\chi y\psi z \infty \propto \varnothing \mathrm{d}\eth \ni$

GitHub

- https://guides.github.com/activities/hello-world/
- A centralized place to store and track all code related to your project.
- A principled way to collaborate with team members.

Tentative Class Schedule

Week	Topic
1	Introduction
2	Basics: Probability Theory and Linear Algebra
3	Linear Models for Regression
4	Linear Models for Classification
5	Support Vector Machines
6	Kernel Methods
7	Graphical Models
8	Sparse Learning
9	Spring Break
10	Clustering
11	Dimensionality Reduction
12	Manifold Learning
13	Multi-Task Learning
14	Transfer Learning
15	Other topics

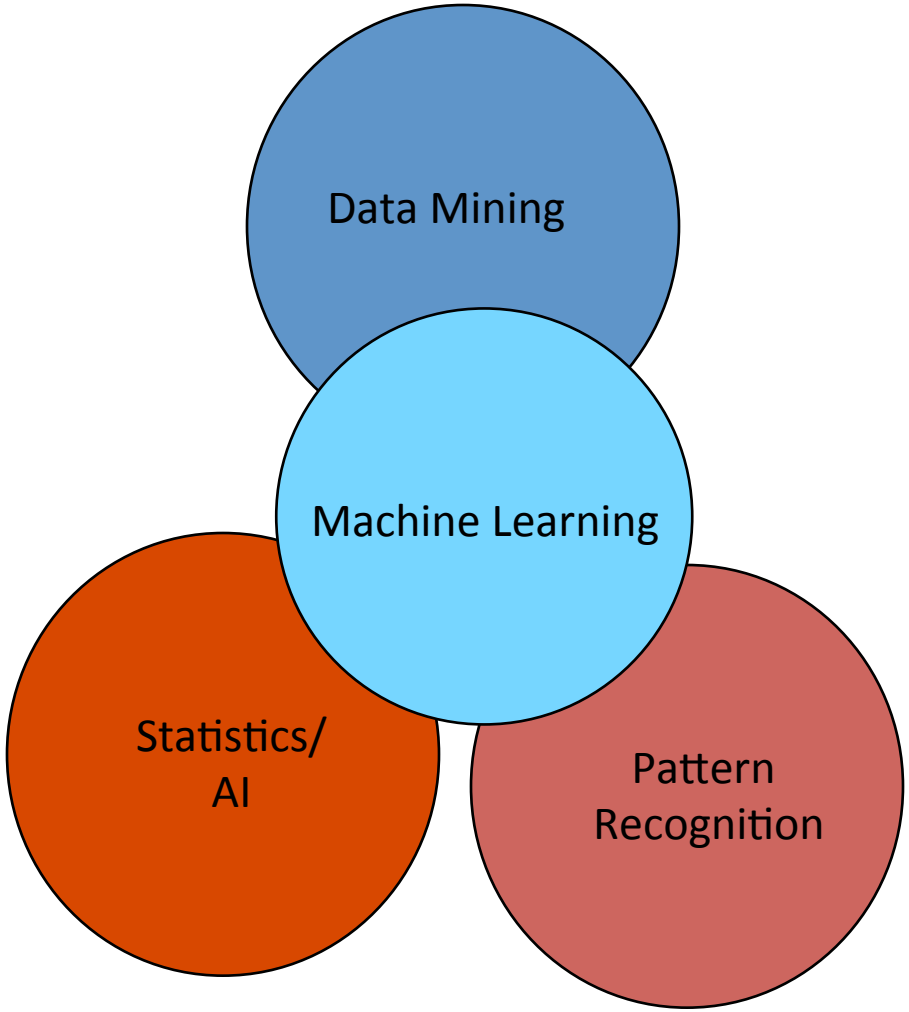
What is Machine Learning?

- The goal of the field of machine learning is to build computer systems that learn from experience and that are capable to adapt to their environments.
  - Learning techniques and methods developed in this field have been successfully applied to a variety of applications, including text classification, gene/protein function prediction, financial forecasting, credit card fraud detection, collaborative filtering, digit/face recognition.

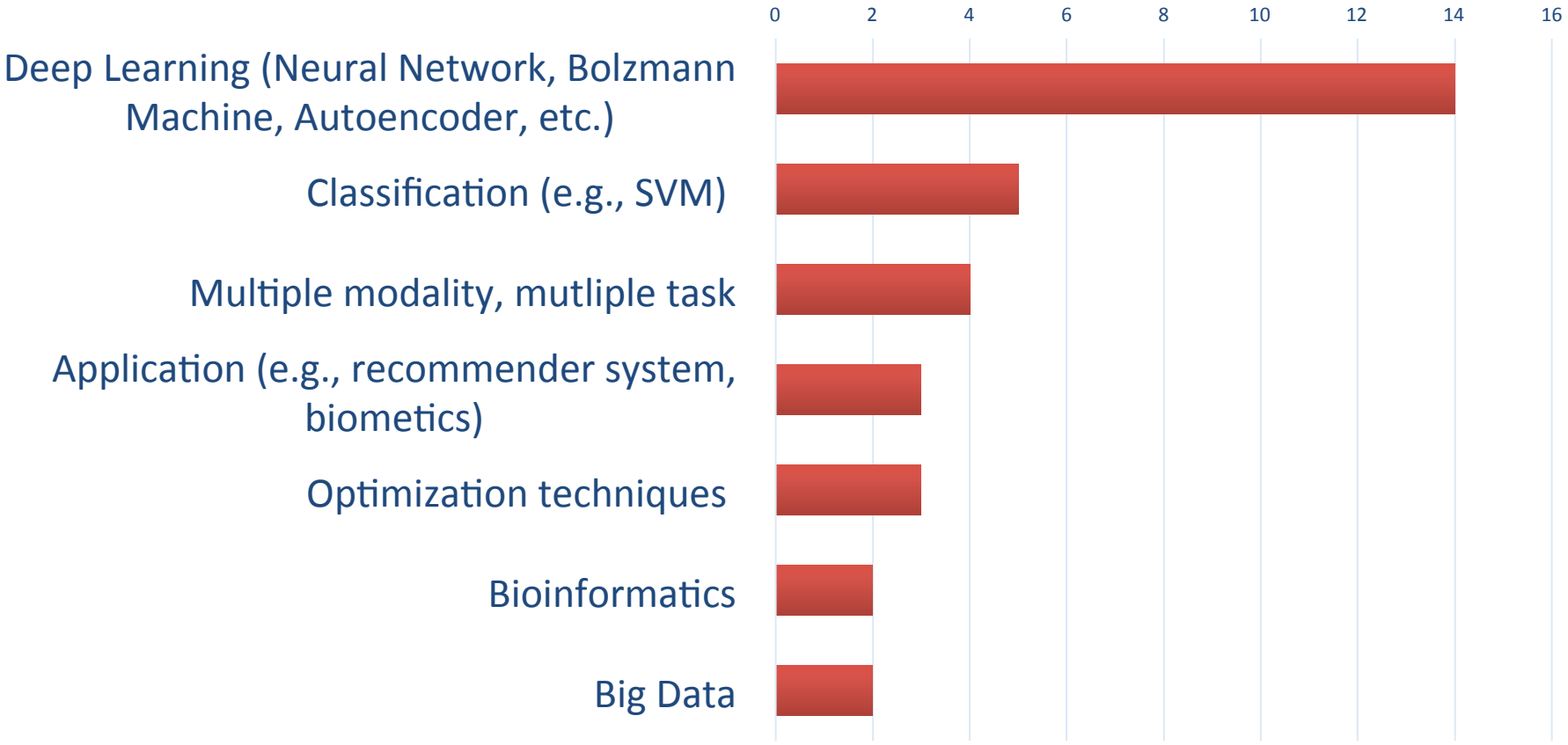
Machine Learning Applications

- Computer vision, natural language processing, search engines, medical diagnosis, bioinformatics, chemoinformatics, detecting credit card fraud, stock market analysis, speech and handwriting recognition, software engineering, robot locomotion.

Related Fields



Voting Result



Machine Learning Tasks

- Supervised learning: (x, y)
  - Classification
  - Regression
- Unsupervised learning
  - Clustering
  - Density estimation
  - Visualization (dimensionality reduction)
- Semi-supervised learning
- …

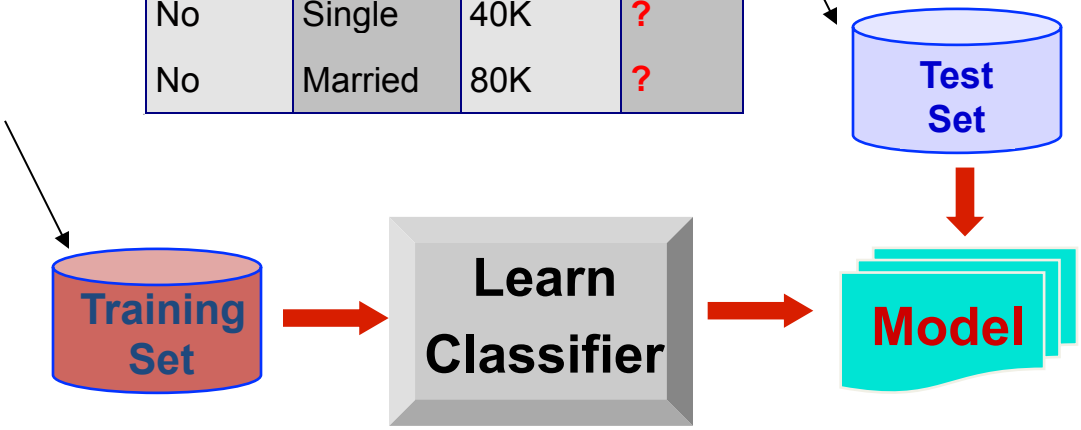
Classification: Definition

- Given a collection of records (*training set*)
  - Each record contains a set of *attributes*, one of the attributes is the *class*.
- Find a *model* for class attribute as a function of the values of other attributes.
- Goal: previously unseen records should be assigned a class as accurately as possible.
  - A *test set* is used to determine the accuracy of the model. Usually, the given data set is divided into training and test sets, with training set used to build the model and test set used to validate it.

Classification Example

Tid	Refund (categorical)	Marital Status (categorical)	Taxable Income (continuous)	Cheat (class)
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Refund	Marital Status	Taxable Income	Cheat
No	Single	75K	?
Yes	Married	50K	?
No	Married	150K	?
Yes	Divorced	90K	?
No	Single	40K	?
No	Married	80K	?



Example I

❑Fraud Detection

- Goal: Predict fraudulent cases in credit card transactions.
- Approach:
  - ➢ Use credit card transactions and the information on its account-holder as attributes.

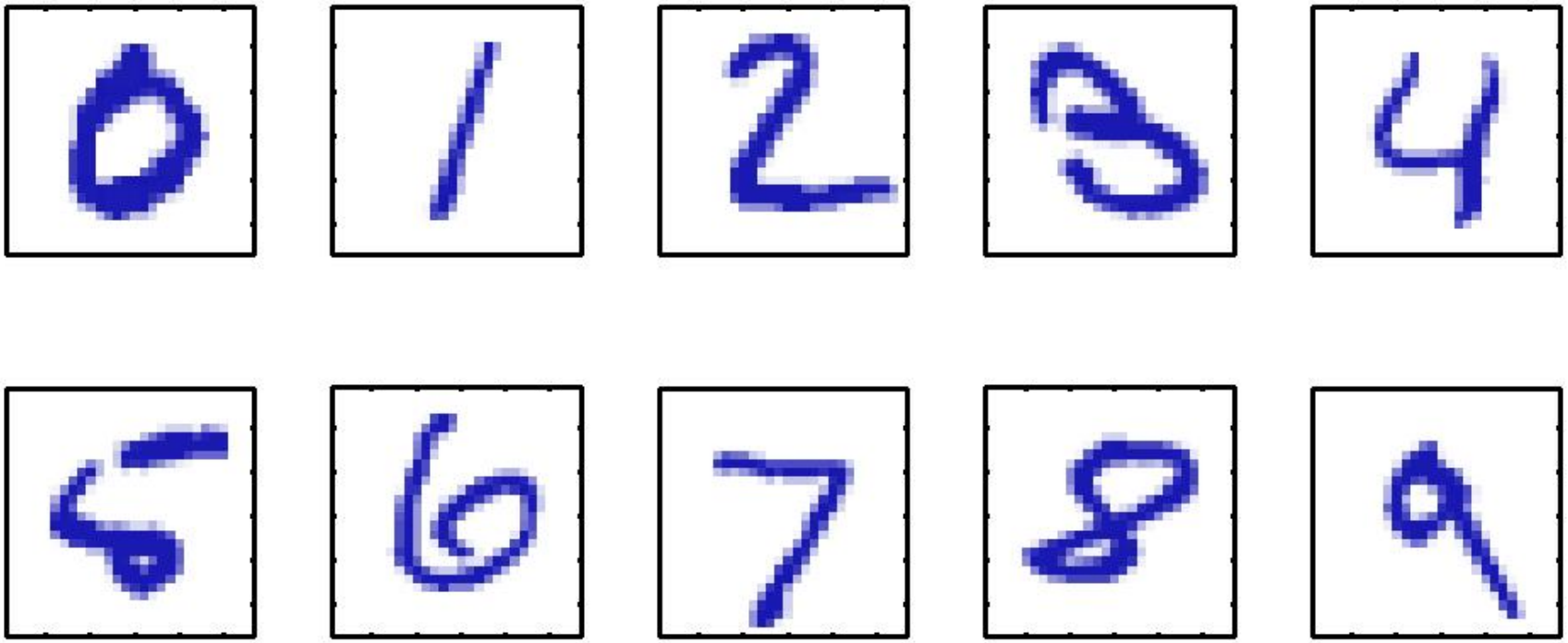
    When does a customer buy, what does he buy, how often he pays on time, etc
  - ➢ Label past transactions as fraud or fair transactions. This forms the class attribute.
  - ➢ Learn a model for the class of the transactions.
  - ➢ Use this model to detect fraud by observing credit card transactions on an account.

Example II

Face Recognition

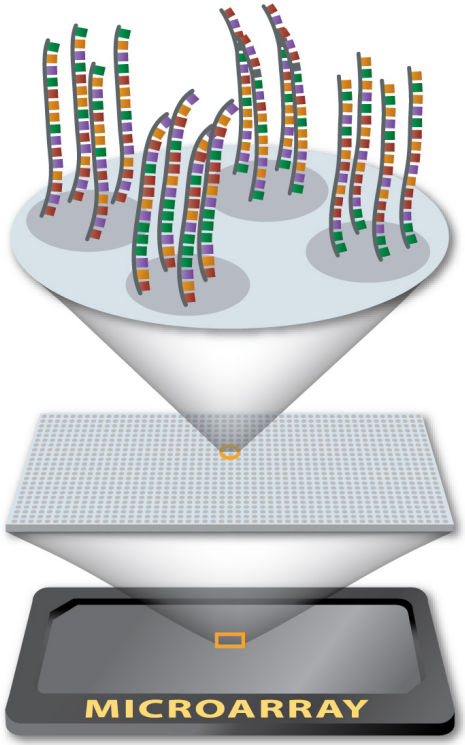
Example III

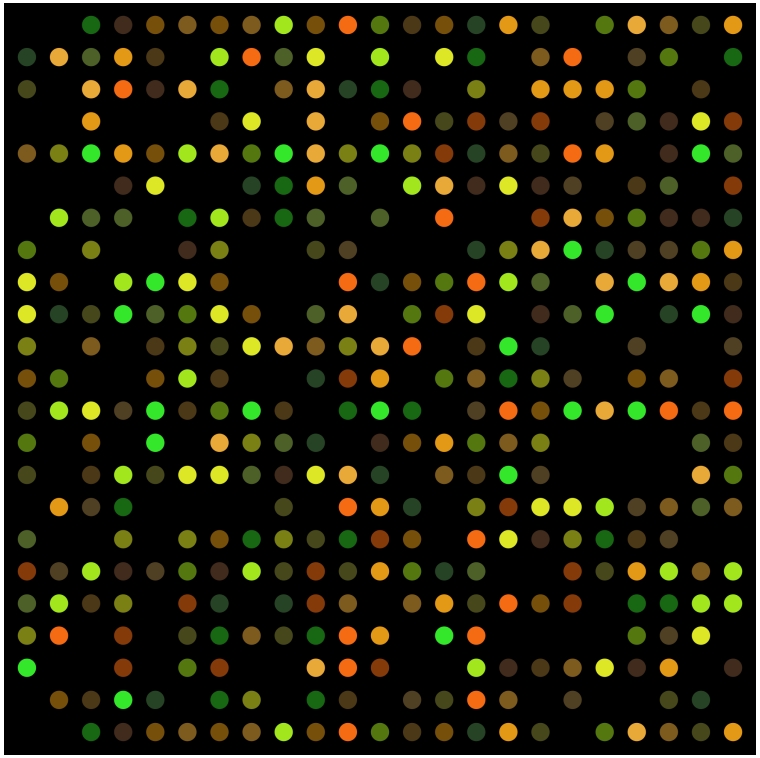
Handwritten Digit Recognition



Example IV

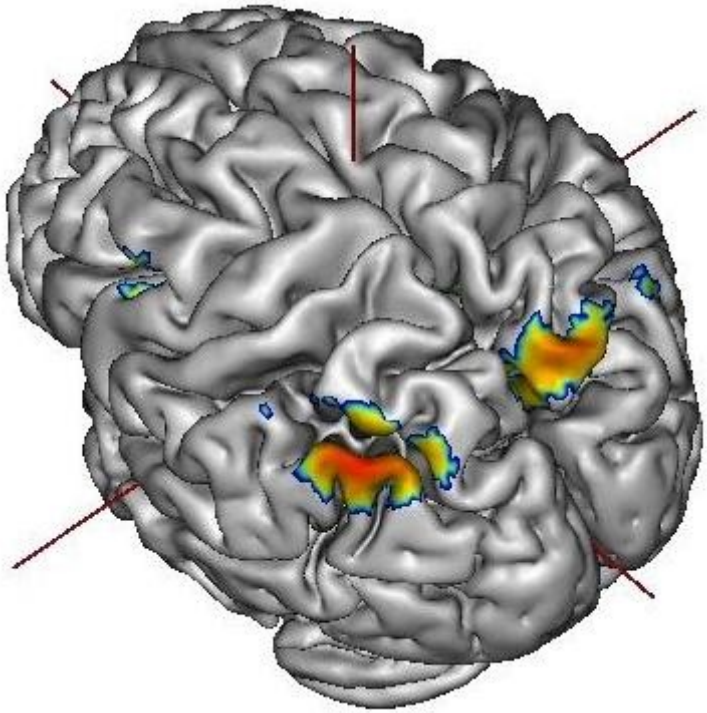
- Cancer Detection
  - Goal: To predict class (cancer or normal) of a sample (person), based on the microarray gene expression data





Example V

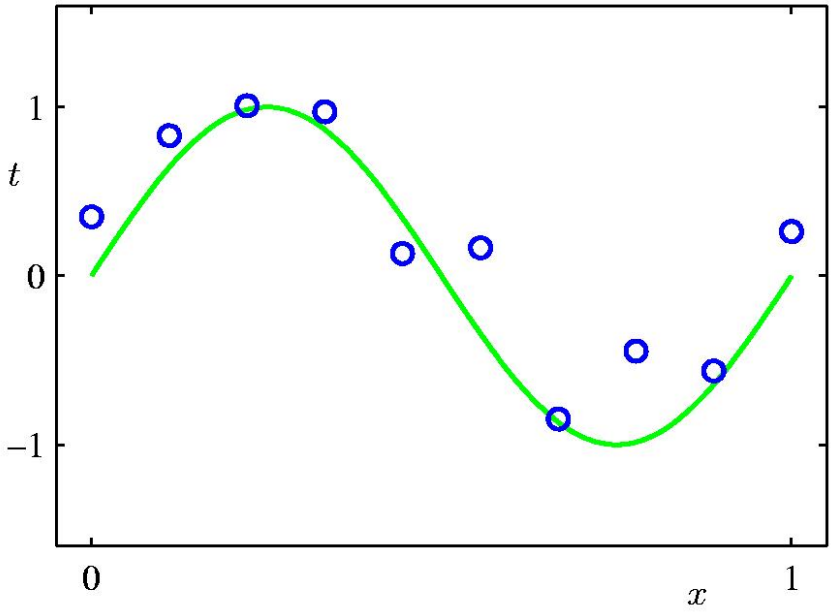
- Alzheimer's Disease Detection
  - Goal: To predict class (AD or normal) of a sample (person), based on neuroimaging data such as MRI and PET



Reduced gray matter volume (colored areas) detected by MRI voxel-based morphometry in AD patients compared to normal healthy controls.

Polynomial Curve Fitting

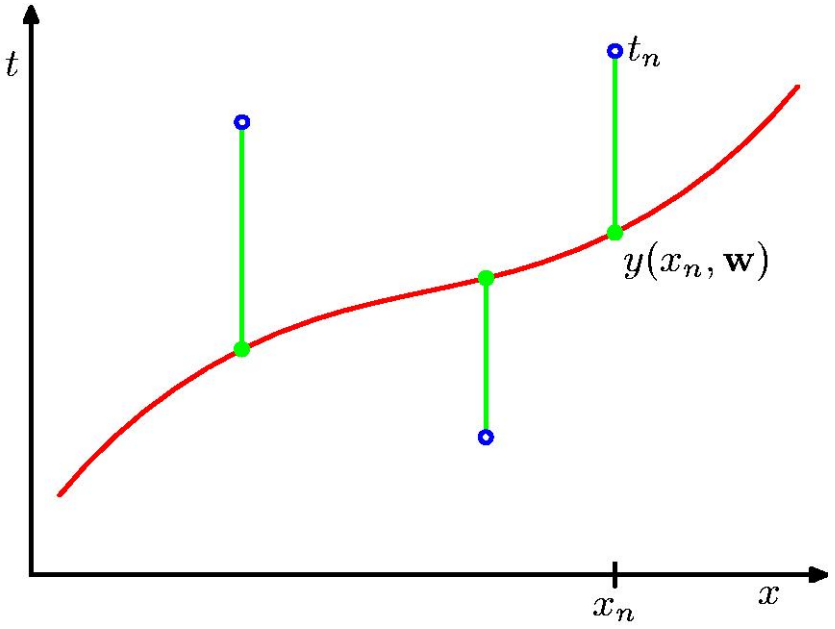
Suppose we observe a real-valued input variable **x** and we wish to use this observation to predict the value of a real-valued target variable t.



polynomial function $y(x, \mathbf{w}) = w_0 + w_1 x + w_2 x^2 + \ldots + w_M x^M = \sum_{j=0}^{M} w_j x^j$

Sum-of-Squares Error Function

The values of the coefficients will be determined by fitting the polynomial to the training data. This can be done by minimizing an error function that measures the misfit between the function y(x,w), for any given value of w, and the training set data points.

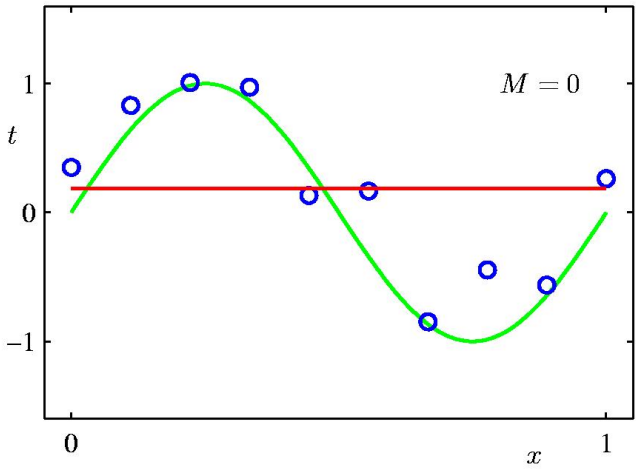


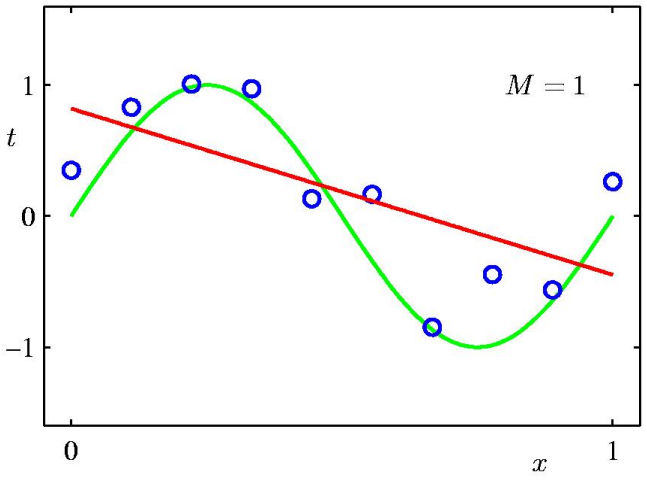
The sum of the squares error function:

$$E(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^{N} \{y(x_n, \mathbf{w}) - t_n\}^2$$

How to choose the order M?

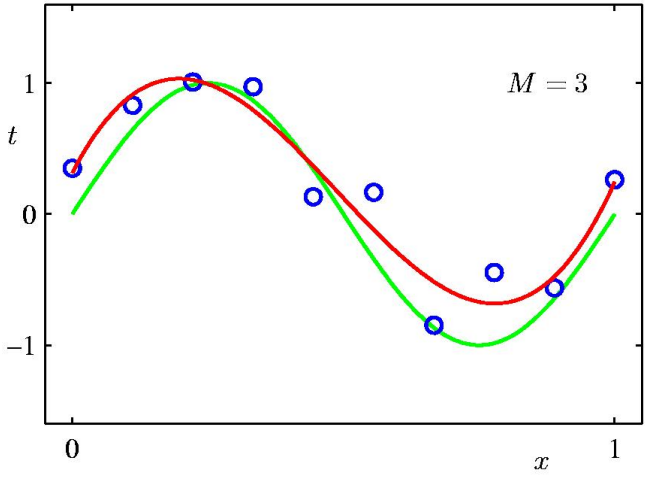
M=0

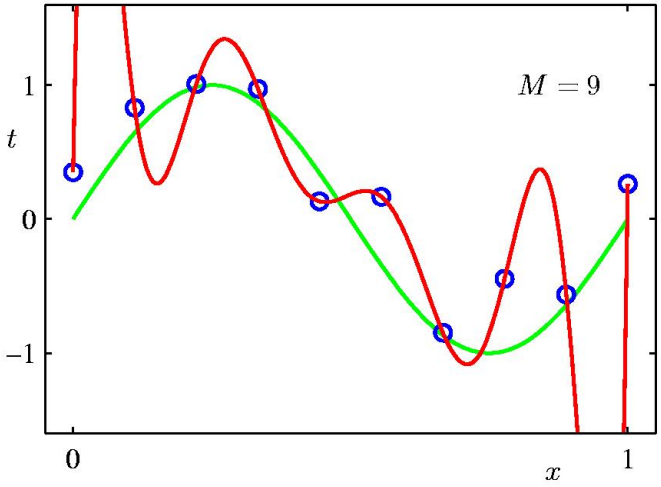




M=1

M=3



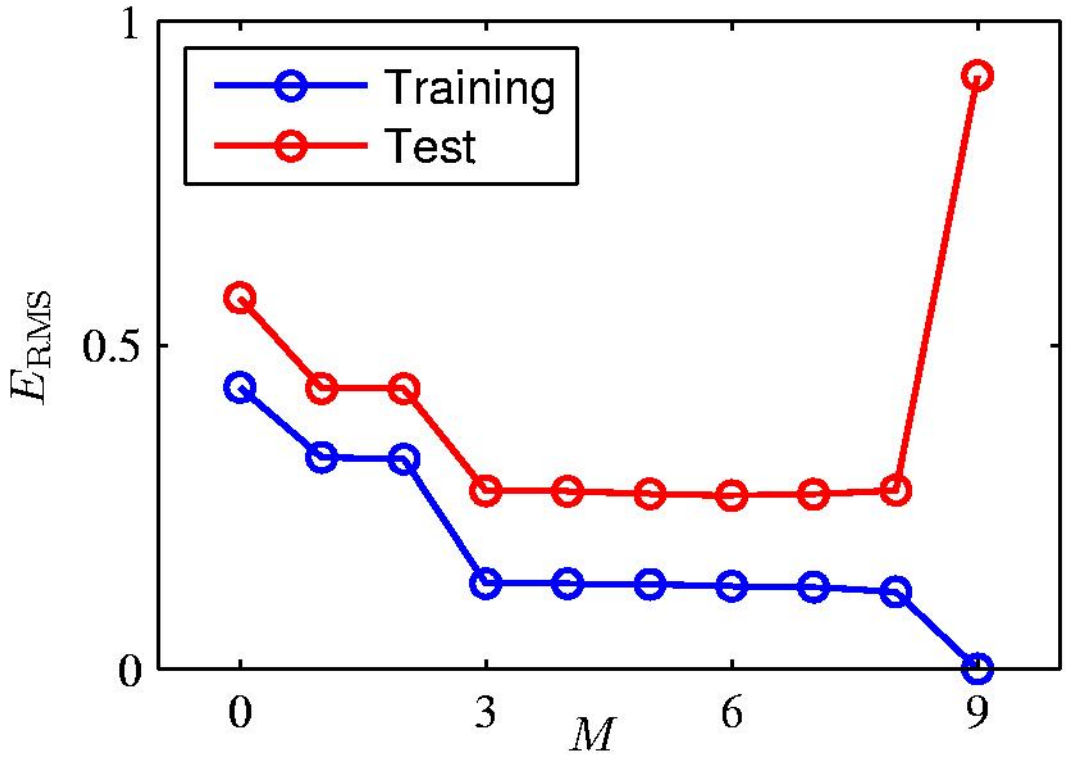


M=9

Observations

- The constant (M = 0) and first order (M = 1) polynomials give rather poor fits to the data.
- The third order (M = 3) polynomial seems to give the best fit to the data.
- Using a much higher order polynomial (M = 9), we obtain an excellent fit to the training data. However, the fitted curve oscillates wildly and gives a very poor representation. This leads to over-fitting.

Over-fitting



poor generalization

Root-Mean-Square (RMS) Error: $E_{\mathrm{RMS}} = \sqrt{2E(\mathbf{w}^\star)/N}$

Polynomial Coefficients

	$M = 0$	$M = 1$	$M = 3$	$M = 9$
$w_0^\star$	0.19	0.82	0.31	0.35
$w_1^\star$		-1.27	7.99	232.37
$w_2^\star$			-25.43	-5321.83
$w_3^\star$			17.37	48568.31
$w_4^\star$				-231639.30
$w_5^\star$				640042.26
$w_6^\star$				-1061800.52
$w_7^\star$				1042400.18
$w_8^\star$				-557682.99
$w_9^\star$				125201.43

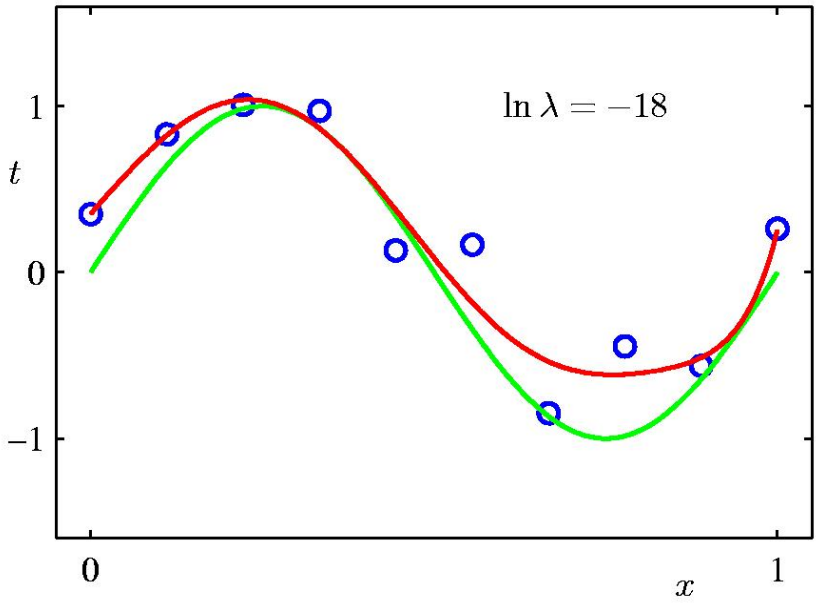
Regularization

- One technique that is often used to control the over-fitting phenomenon in such cases is that of regularization, which involves adding a penalty term to the error function in order to discourage the coefficients from reaching large values.

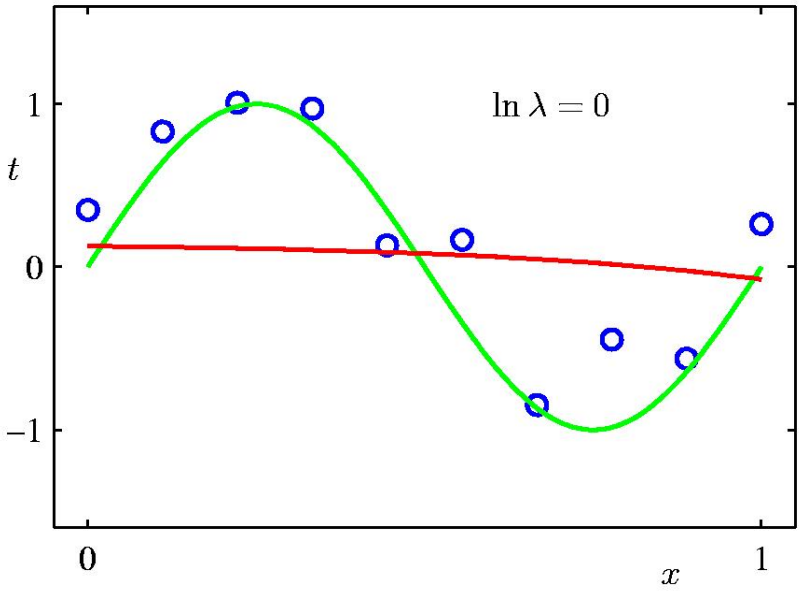
$$\widetilde{E}(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\{y(x_n, \mathbf{w}) - t_n\}^2 + \frac{\lambda}{2}\|\mathbf{w}\|^2$$

- The coefficient λ governs the relative importance of the regularization term compared with the sum-of-squares error term.

Effect of Regularization Parameter



$\ln \lambda = -18$

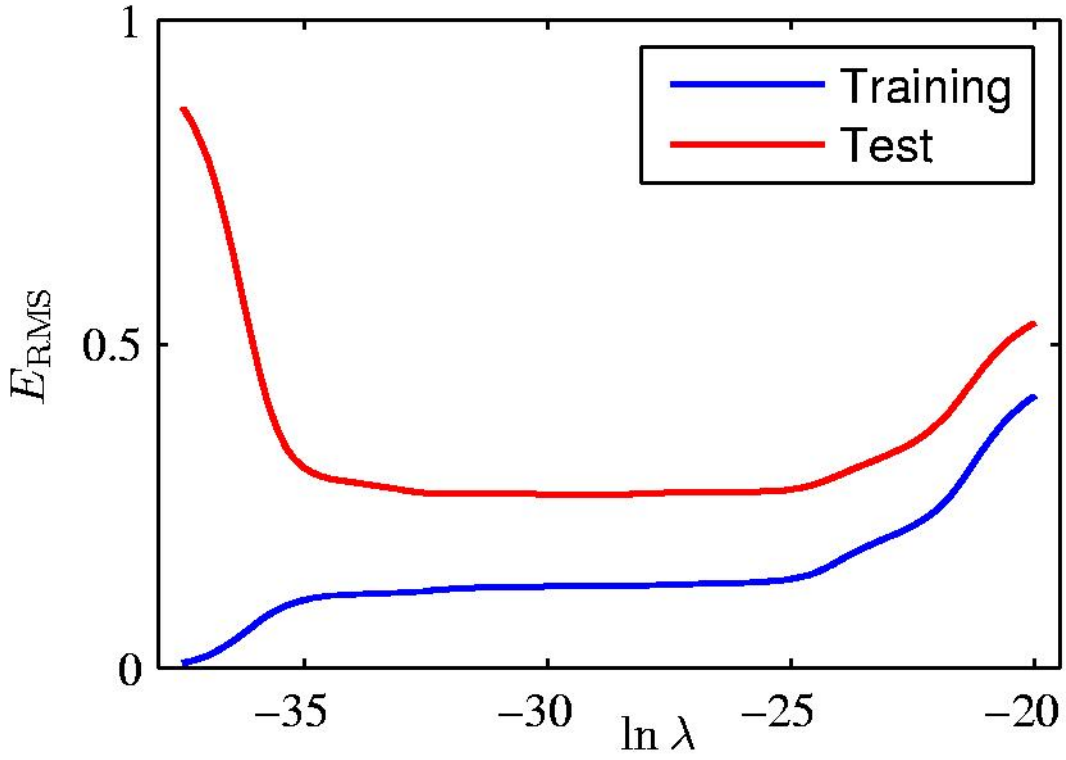


$\ln \lambda = 0$

Polynomial Coefficients

	$\ln \lambda = -\infty$	$\ln \lambda = -18$	$\ln \lambda = 0$
$w_0^\star$	0.35	0.35	0.13
$w_1^\star$	232.37	4.74	-0.05
$w_2^\star$	-5321.83	-0.77	-0.06
$w_3^\star$	48568.31	-31.97	-0.05
$w_4^\star$	-231639.30	-3.89	-0.03
$w_5^\star$	640042.26	55.28	-0.02
$w_6^\star$	-1061800.52	41.32	-0.01
$w_7^\star$	1042400.18	-45.95	-0.00
$w_8^\star$	-557682.99	-91.53	0.00
$w_9^\star$	125201.43	72.68	0.01

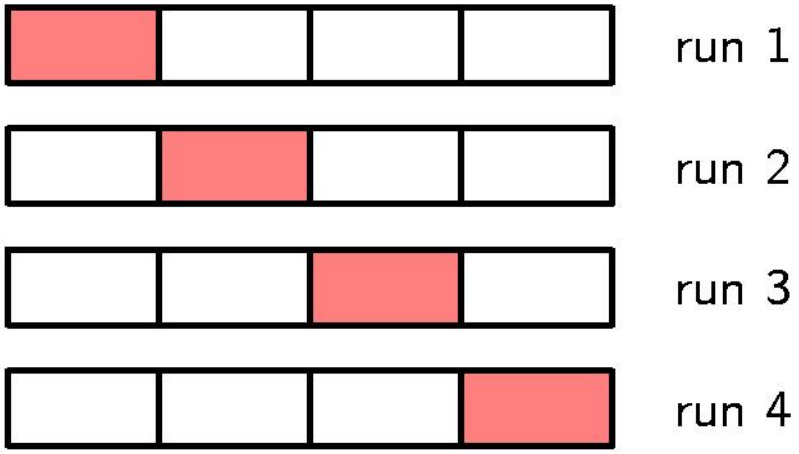
Regularization: E_{RMS} vs. $\ln \lambda$



Model selection: Estimation of the optimal value of the regularization parameter. In practice, cross validation is commonly applied for model selection.

Model Selection

Cross-Validation



***S-fold cross-validation**:*
Partition the data into S groups of equal size. Then S – 1 of the groups are used to train a set of models that are then evaluated on the remaining group. This procedure is repeated for all S possible choices for the held-out group (red blocks), and the performance scores from the S runs are then averaged.

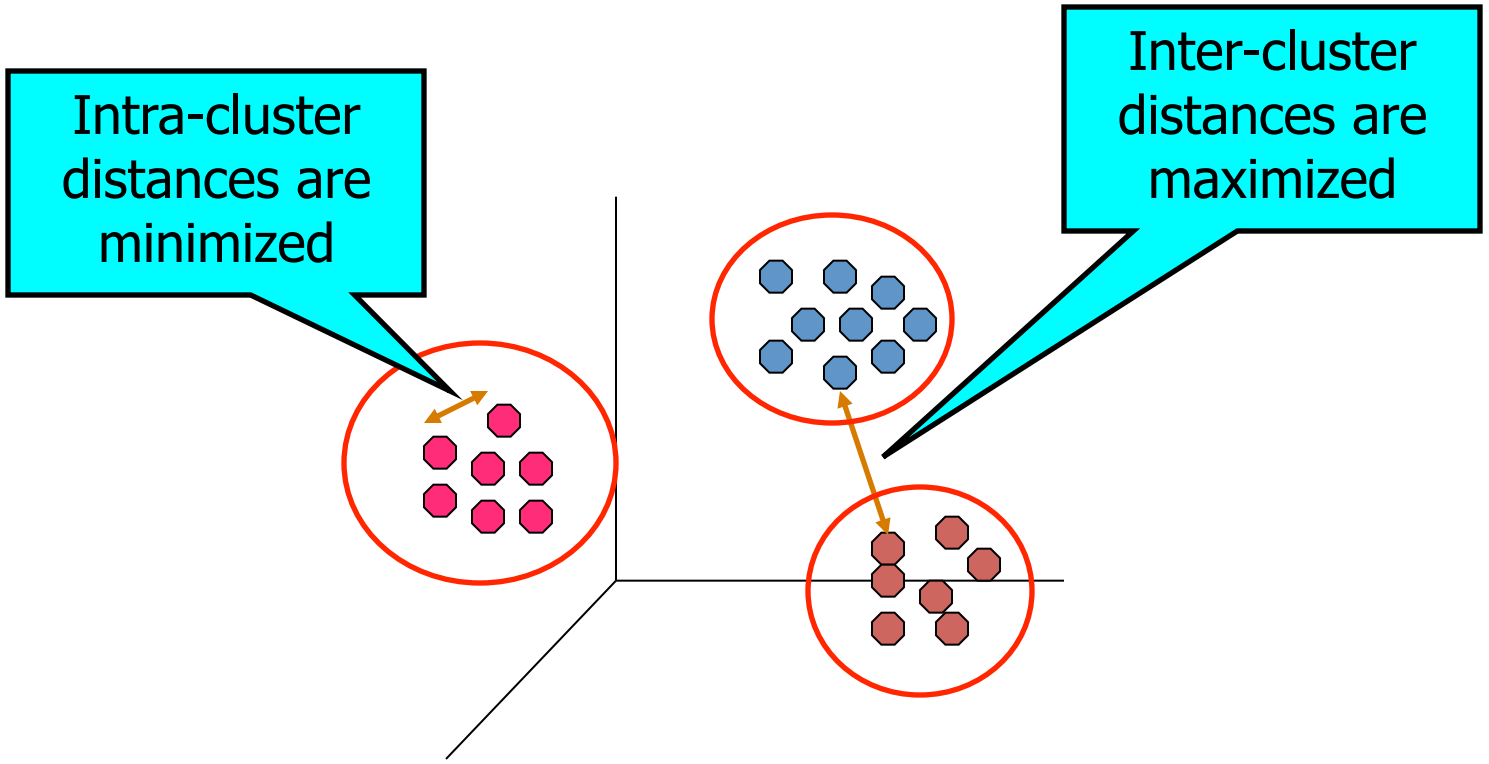
Clustering Definition

- Given a set of data points, each having a set of attributes, and a similarity measure among them, find clusters such that
  - Data points in one cluster are more similar to one another.
  - Data points in separate clusters are less similar to one another.
- Similarity Measures:

  Euclidean Distance if attributes are continuous.

  Other Problem-specific Measures.

Illustrating Clustering



Clustering: Application I

Market Segmentation

Goal: subdivide a market into distinct subsets of customers where any subset may conceivably be selected as a market target to be reached with a distinct marketing mix.

Approach:

- Collect different attributes of customers (income, lifestyle, age, etc.)
- Find clusters of similar customers.
- Measure the clustering quality by observing buying patterns of customers in same cluster vs. those from different clusters.

Clustering: Application II

- Document Clustering:
  - Goal: To find groups of documents that are similar to each other based on the important terms appearing in them.
  - Approach:
    1. To identify frequently occurring terms in each document.
    2. Form a similarity measure based on the frequencies of different terms.
    3. Use it to cluster.
  - Gain: Information Retrieval can utilize the clusters to relate a new document or search term to clustered documents.

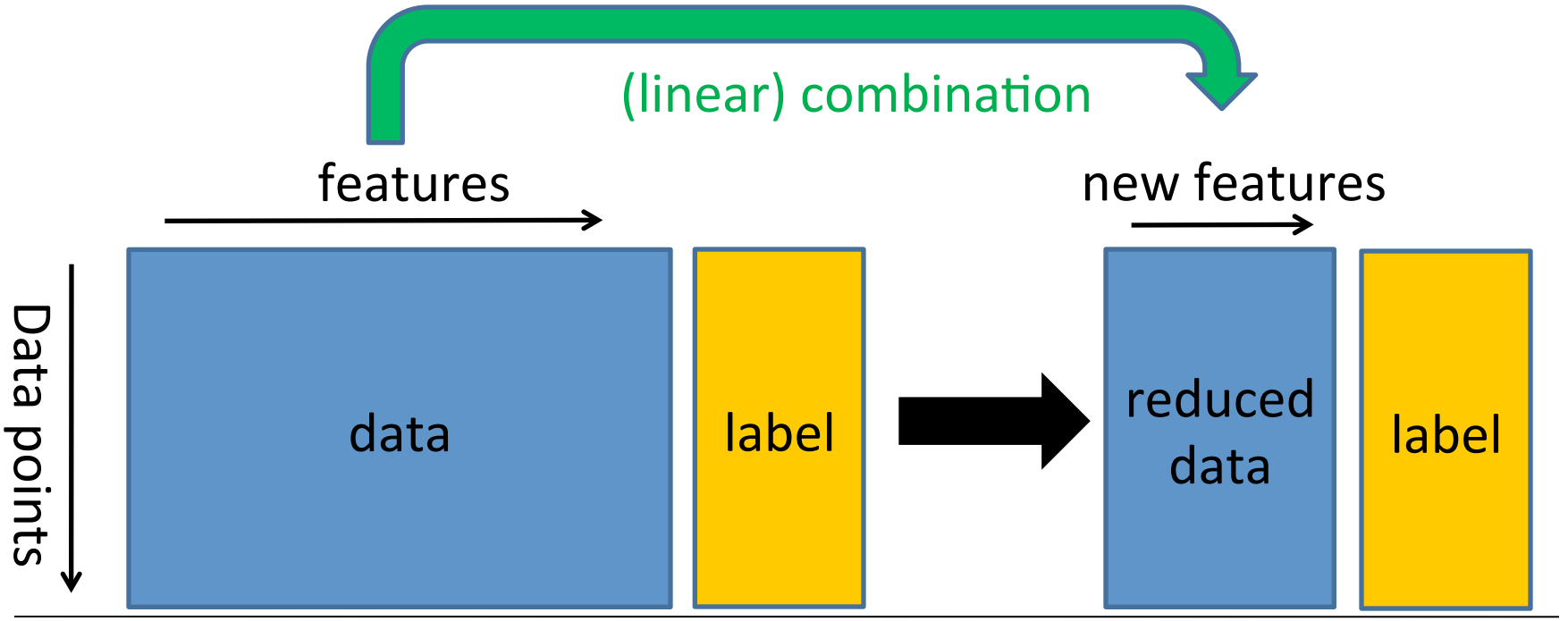
Illustrating Document Clustering

- Clustering Points: 3204 Articles of Los Angeles Times.
- Similarity Measure: How many words are common in these documents (after some word filtering).

Category	Total Articles	Correctly Placed	
Financial	555	364	65.6%
Foreign	341	260	76.2%
National	273	36	13.2%
Metro	943	746	79.1%
Sports	738	573	77.6%
Entertainment	354	278	78.5%

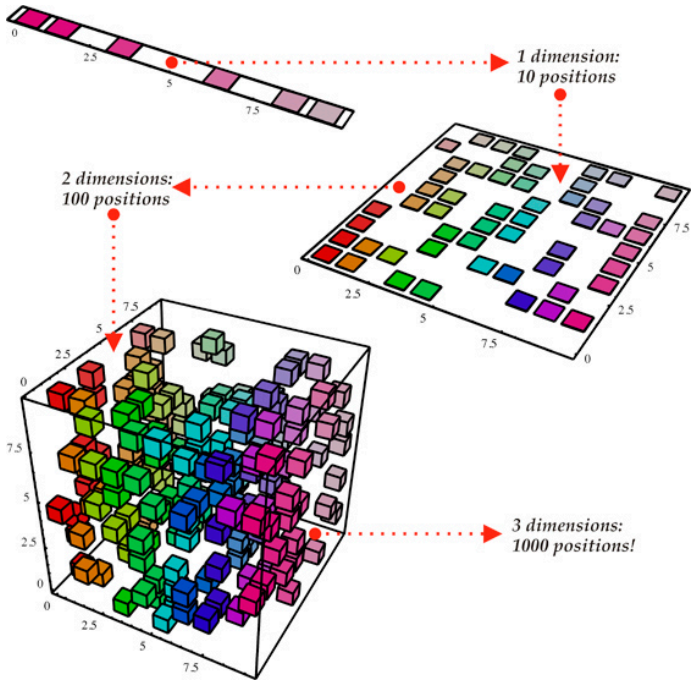
Dimensionality Reduction

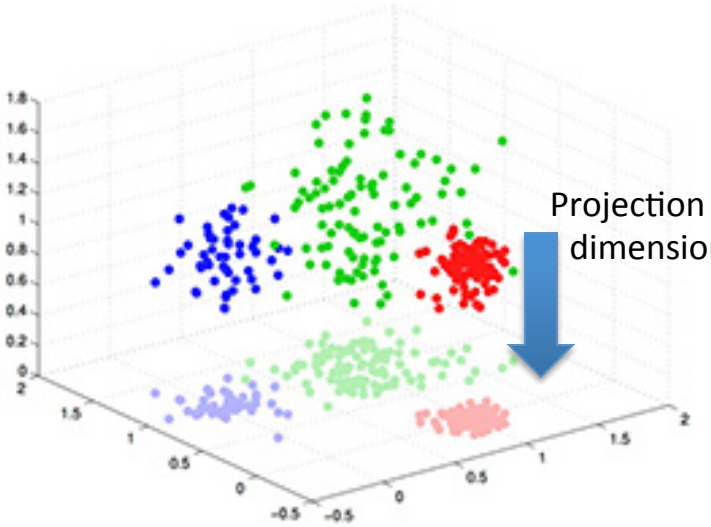
- Dimensionality reduction extracts a small number of features by removing irrelevant, redundant, and noisy information
- Different from feature selection



Why Dimensionality Reduction

- High-dimensional space is sparse and data points are far apart from each other
- Exponential number of data points are required



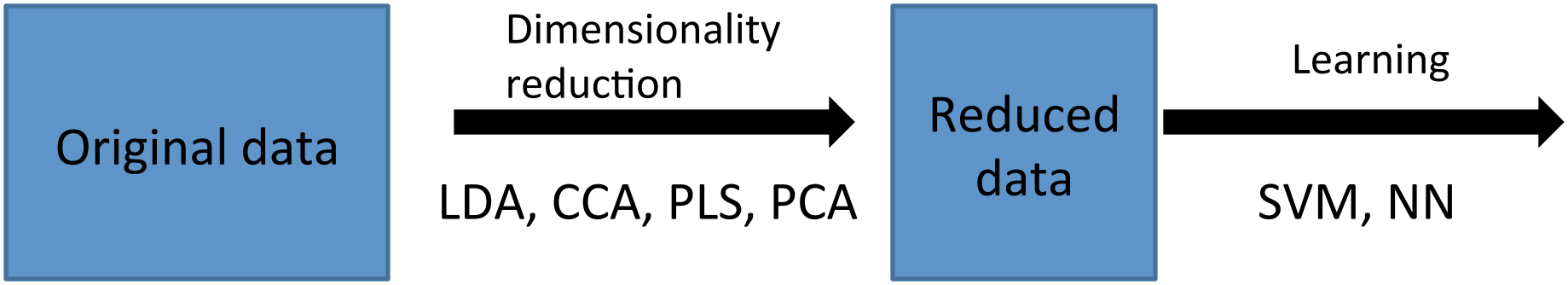


Projection into lower dimensional space

http://bigdata.csail.mit.edu/node/277
http://www.iro.umontreal.ca/~bengioy/yoshua_en/research_files/CurseDimensionality.jpg

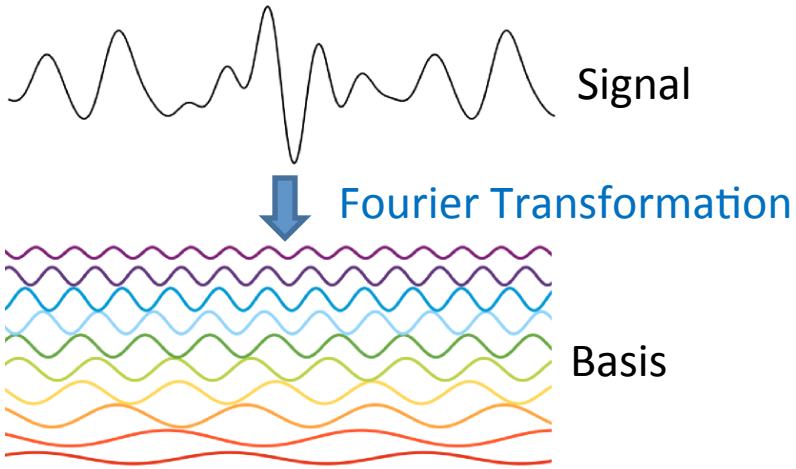
Dimensionality Reduction Algorithms

- Supervised:
  - Linear discriminant analysis (LDA)
  - Canonical correlation analysis (CCA)
  - Partial least squares (PLS)
- Unsupervised:
  - Principal component analysis (PCA)



Sparse Learning I

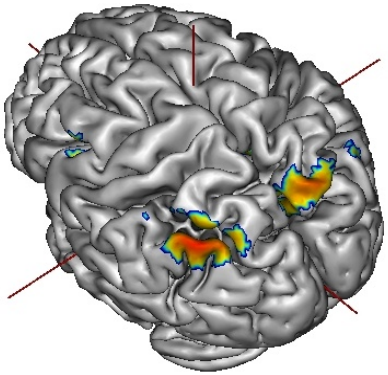
- The underlying representations of many real-world processes are often sparse.
  - Disease diagnosis
  - Neural representation of sounds
  - Signals represented under a proper basis

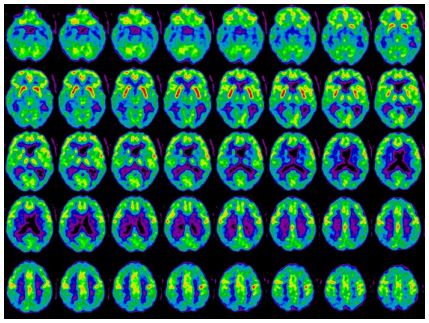


Sparse Learning II

- Most existing work on sparse learning are based on a variant of the *L_1 norm regularization*
  - Sparsity inducing property
  - convexity (efficient solvers)
  - Strong theoretical guarantees
  - Great empirical success

Application I: Feature Selection





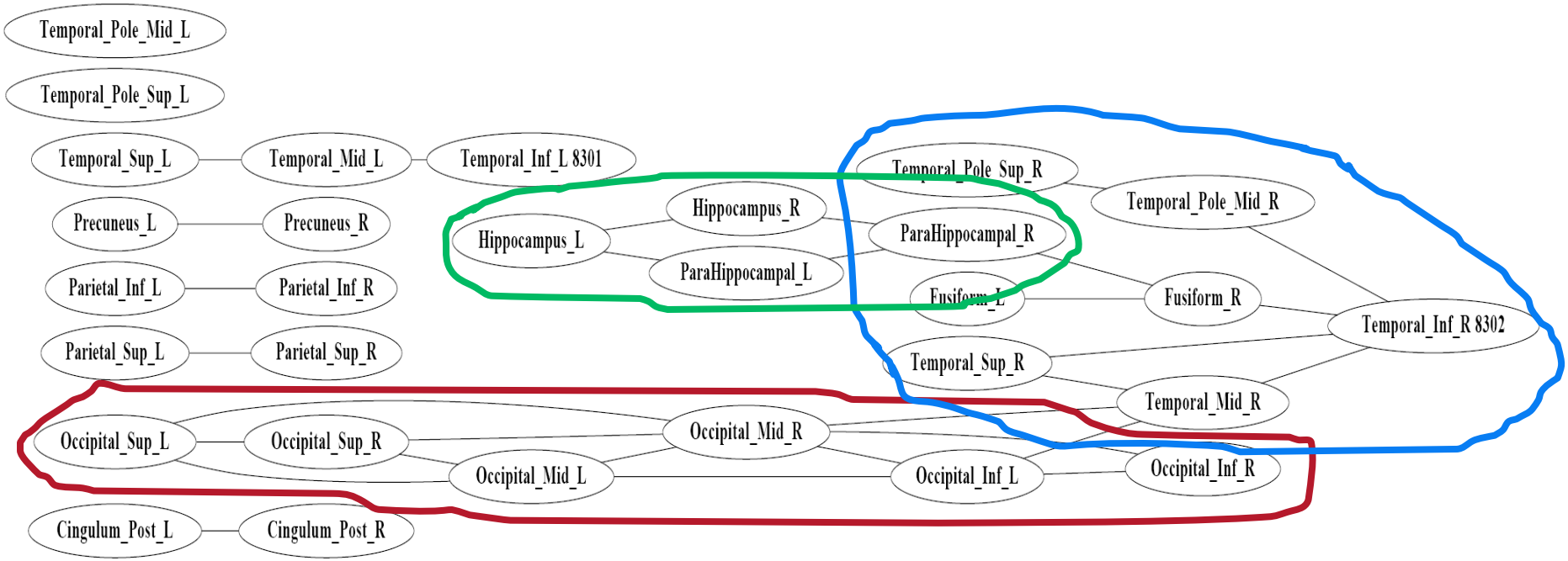
Demographic, genetic, cognitive measures

Identify the optimal combination of features for distinguishing Alzheimer's Disease patients from Normal Controls

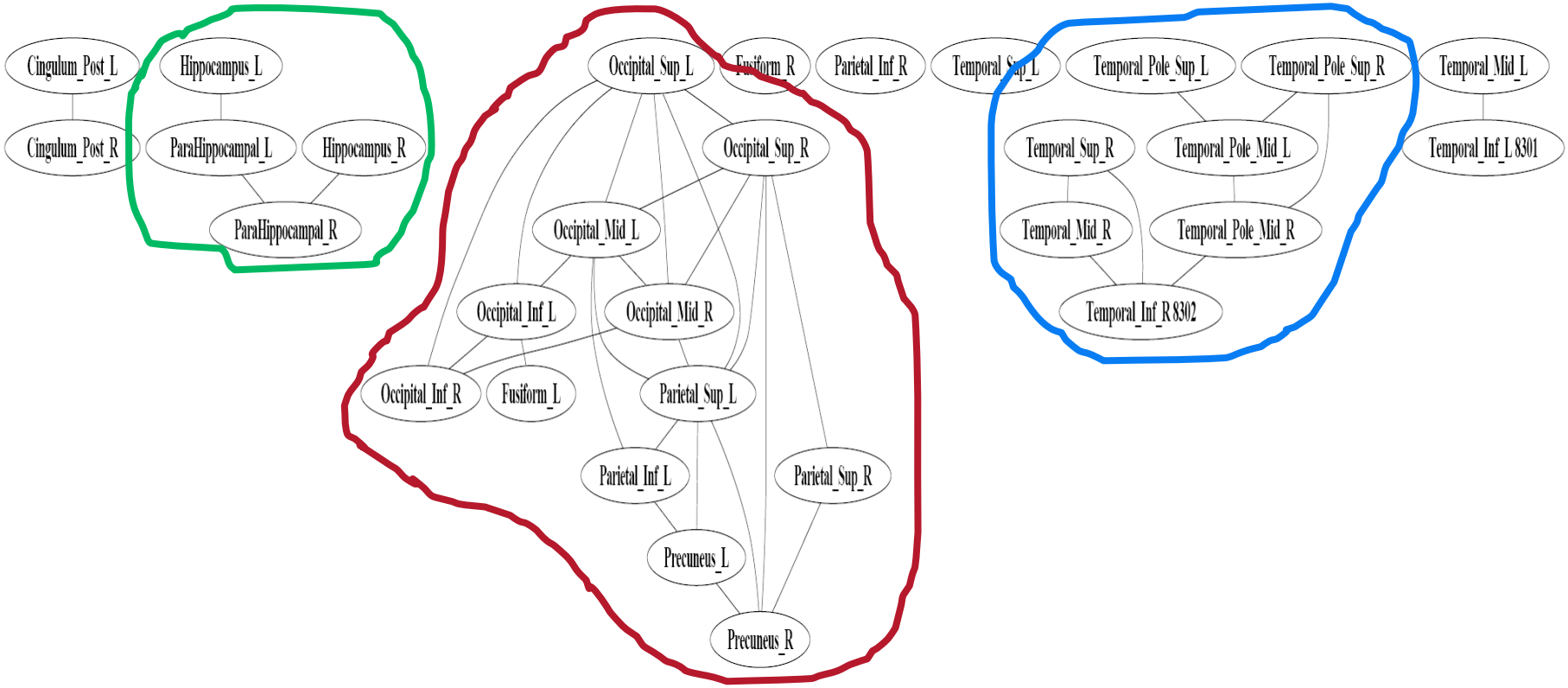
Application II: AD Brain Region Connectivity

- Cognition is a result of different brain regions interacting with each other, rather than individual regions working independently.
  - AD, with major symptoms being dramatic cognitive decline, may have abnormal brain connectivity patterns.
- Hypotheses
  - There is significant, quantifiable difference in brain connectivity between AD and normal brains.

Brain Connectivity for Normal Controls

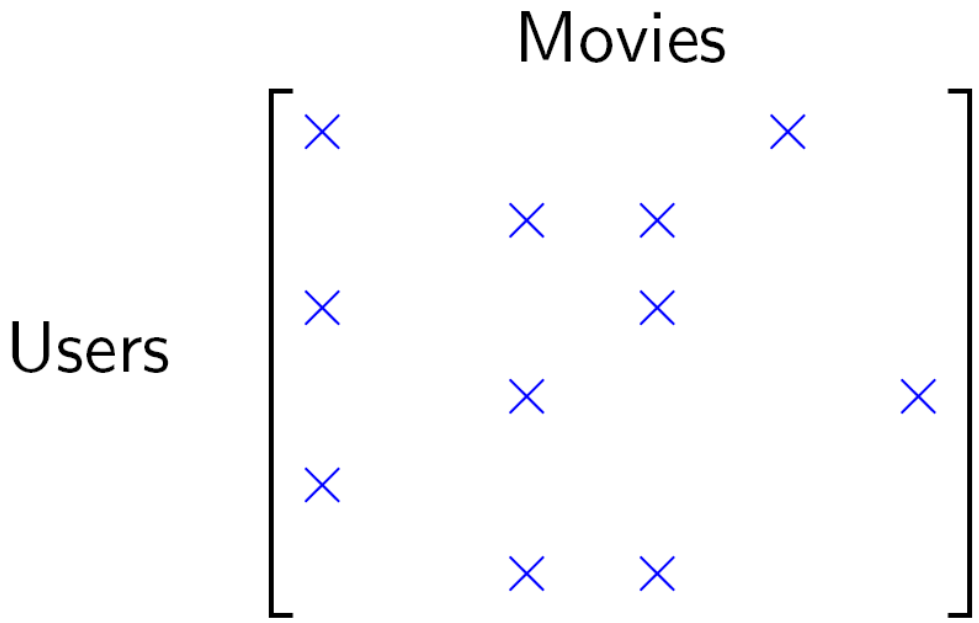


Brain Connectivity for AD Patients



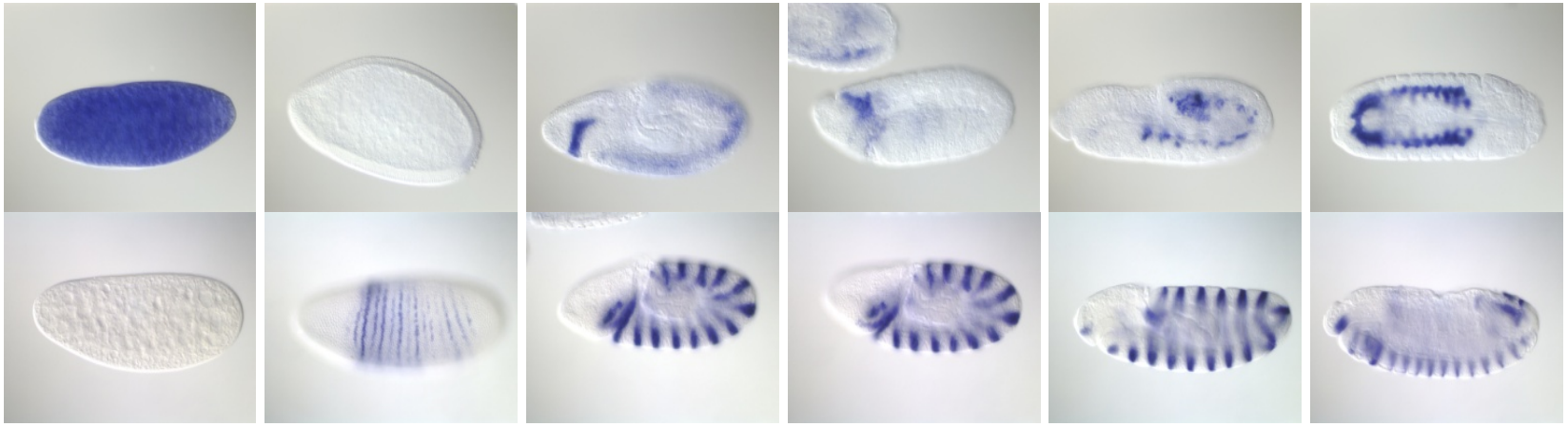
Application III: Collaborative Filtering

- Netix database
  - About a million users
  - About 25,000 movies
- People rate movies
- Sparsely sampled entries

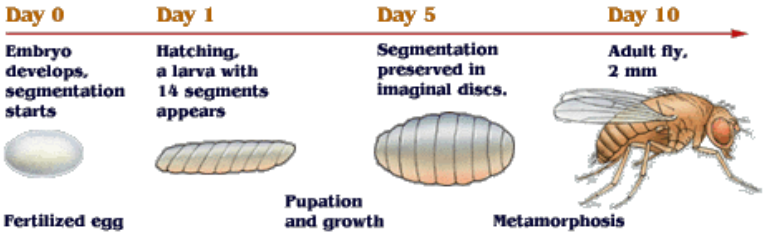


Example VI

- Drosophila Embryonic Developmental Stage Prediction

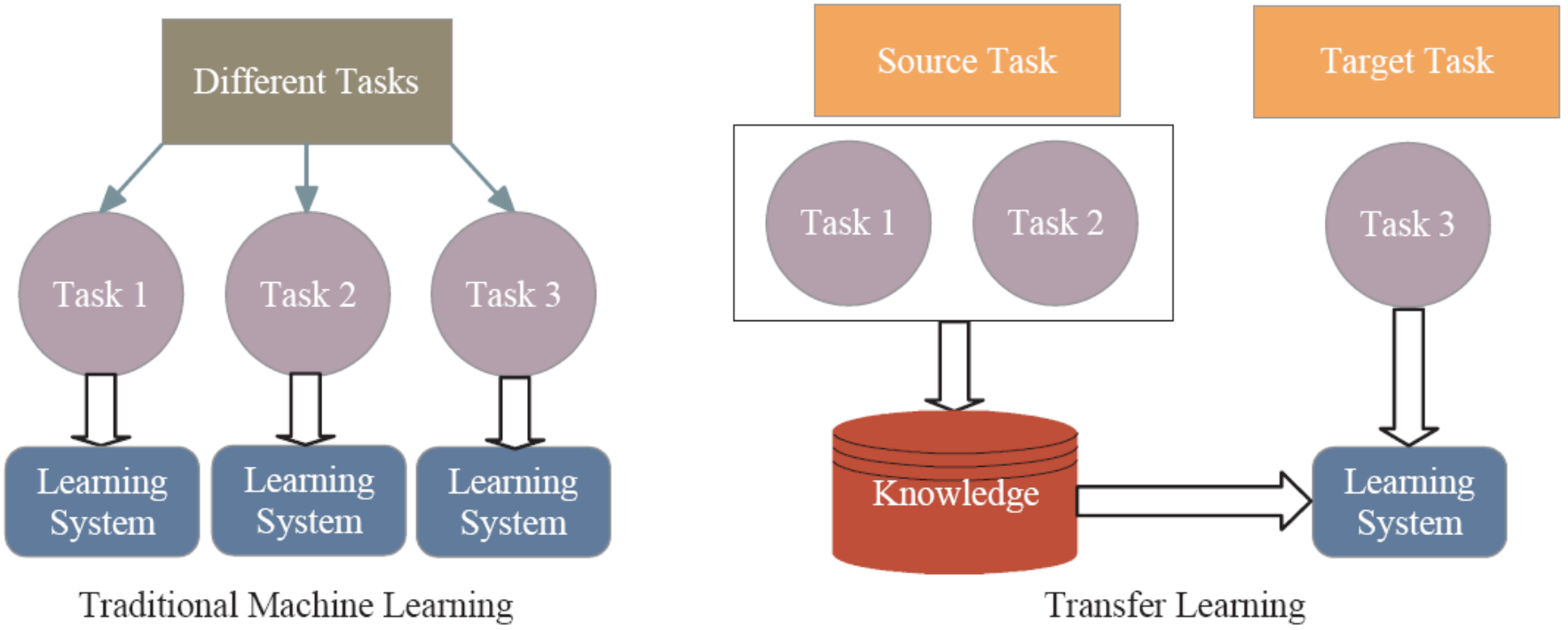


Stage 1-3 Stage 4-6 Stage 7-8 Stage 9-10 Stage 11-12 Stage 13-



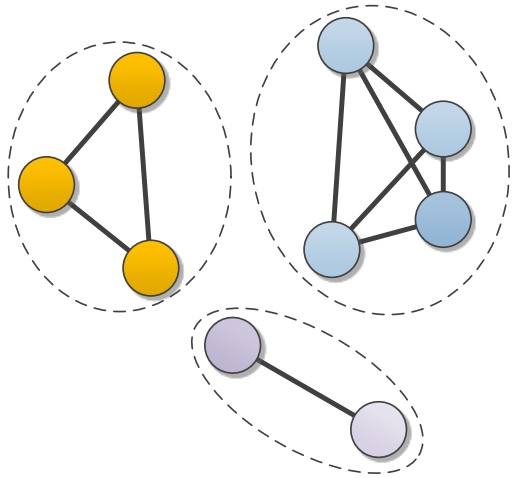
Transfer learning

Transfer knowledge from the source to the target



Multi-Task Learning

- A set of related machine learning tasks (regression, classification, clustering, etc.)
- Use the relatedness to improve the machine learning.



Multi-Task Learning Application

HIV Therapy Screening

[Bickel, ICML'08]

Collaborative ordinal regression

[Yu et. al. NIPS'06]

Disease progression modeling

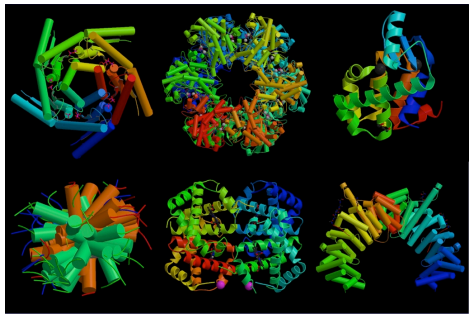
[Zhou et. al. KDD'11, 12]

Web image and video search

[Wang et. al. CVPR'09]

Disease prediction

[Zhang et. al. NeuroImage 12]

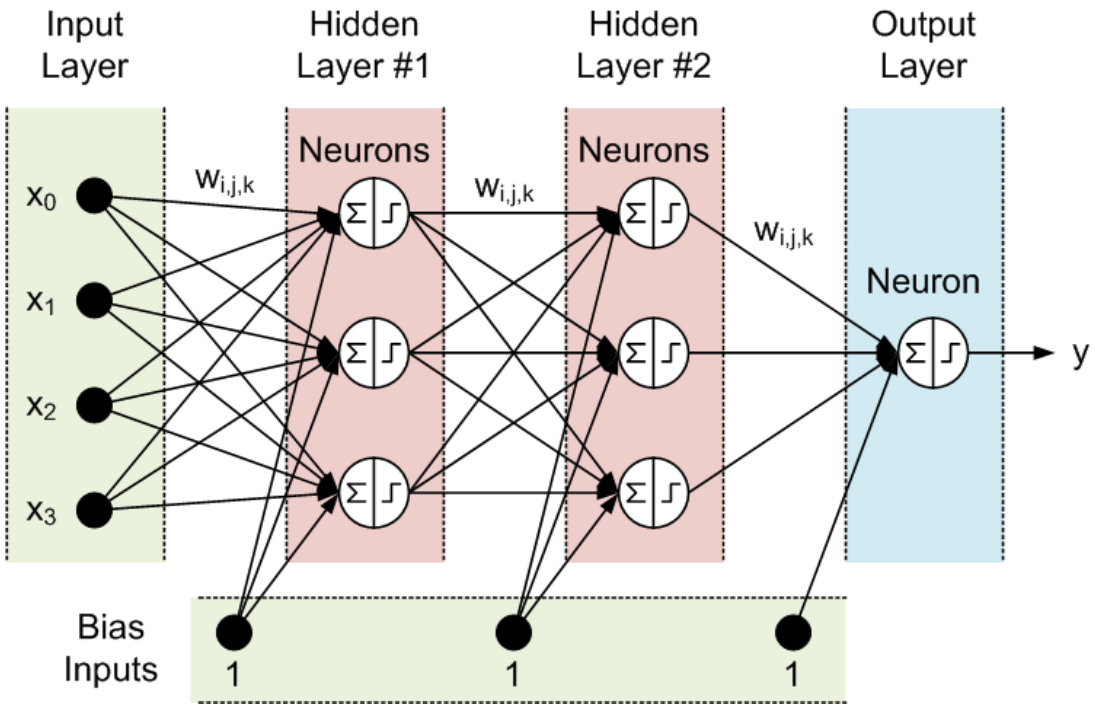


Protein classification

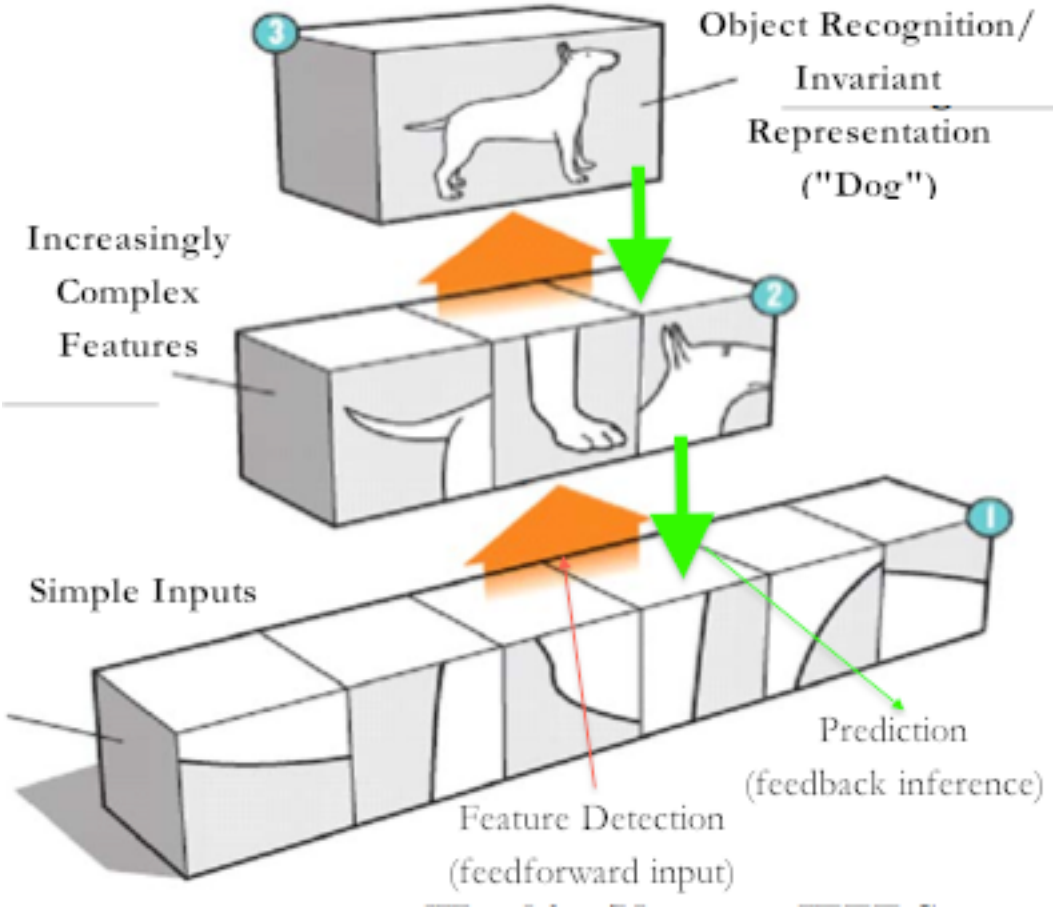
[Charuvaka et. al. ICDM'12]

Neural Network

- Soared in popularity in the 1980s,
- peaked in the early 1990s, and slowly declined after that.

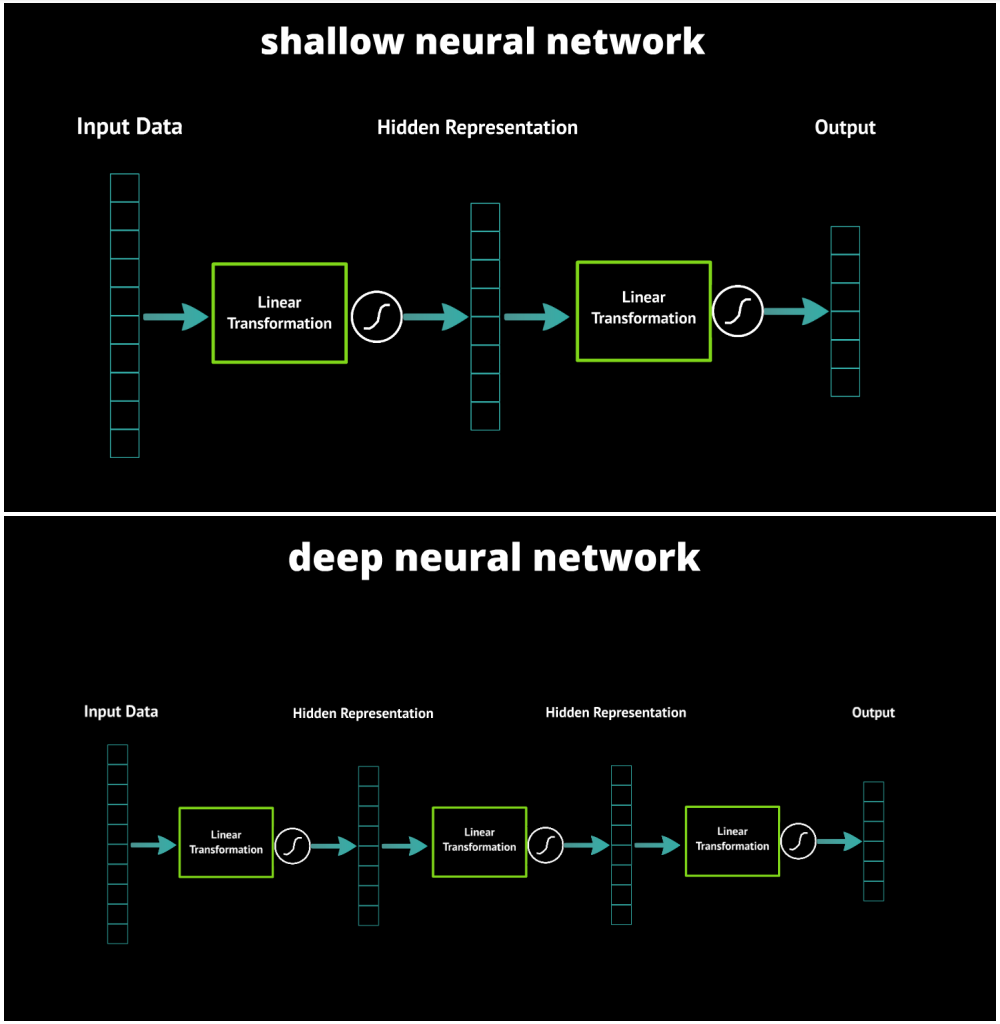


Why Layers?



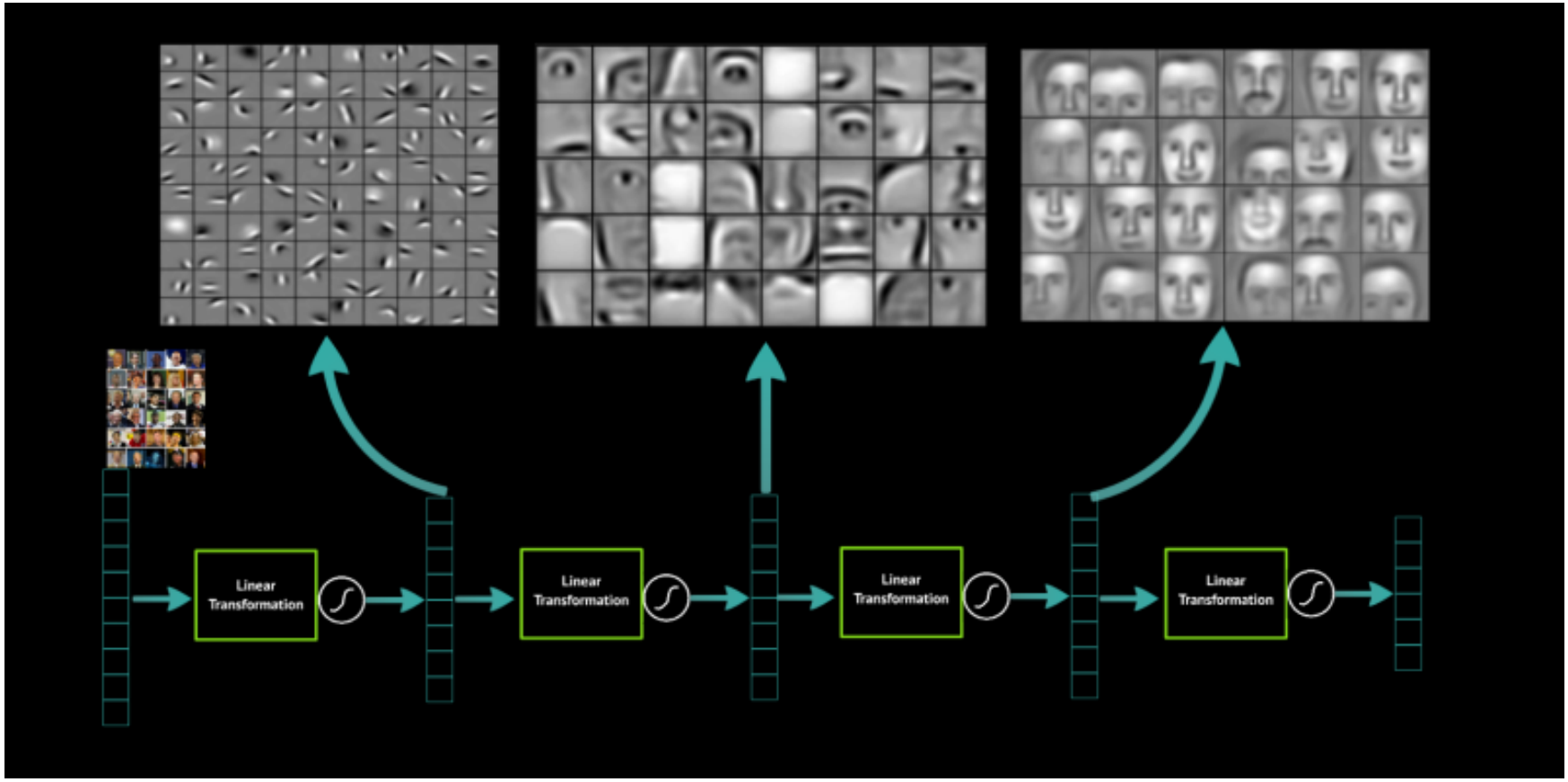
http://mshs.unice.fr/wp-content/uploads/2014/12/Precioso.pdf

Deep Learning



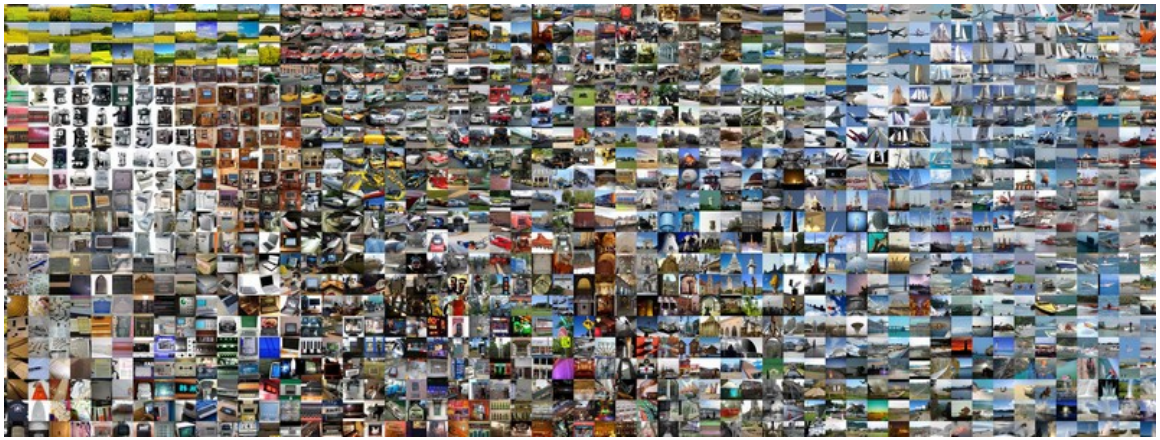
http://www.datarobot.com/blog/a-primer-on-deep-learning/

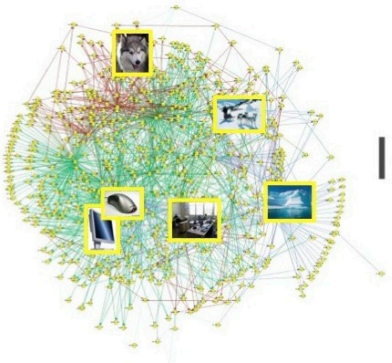
Deep Learning Features



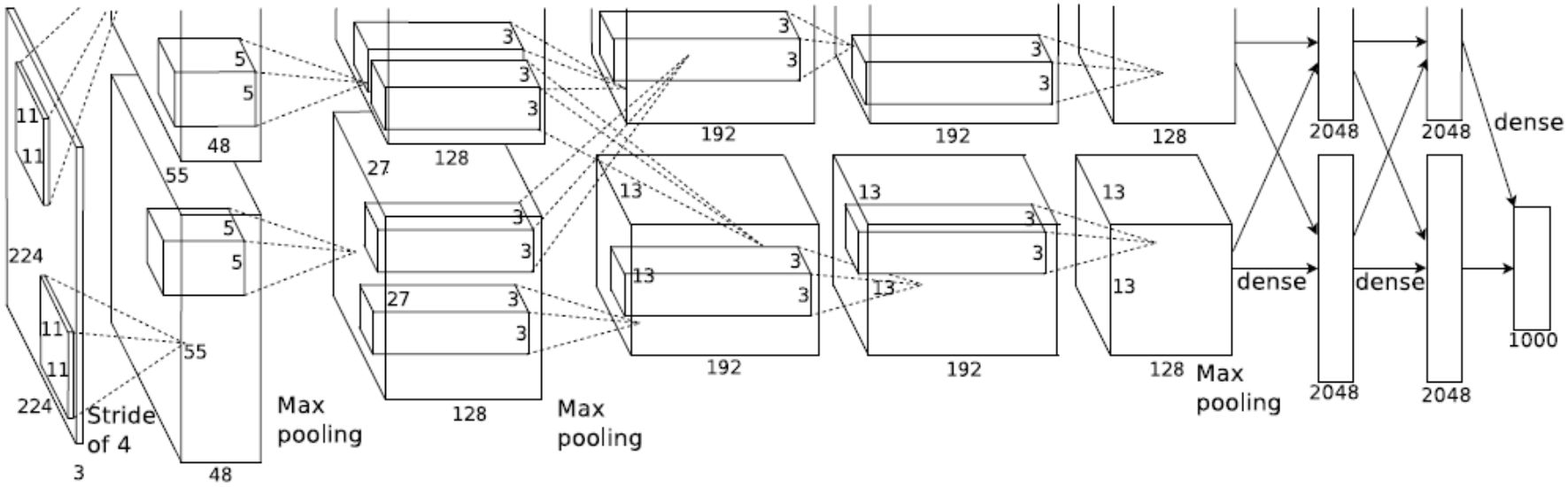
http://www.datarobot.com/blog/a-primer-on-deep-learning/

Deep Learning Example: CNN



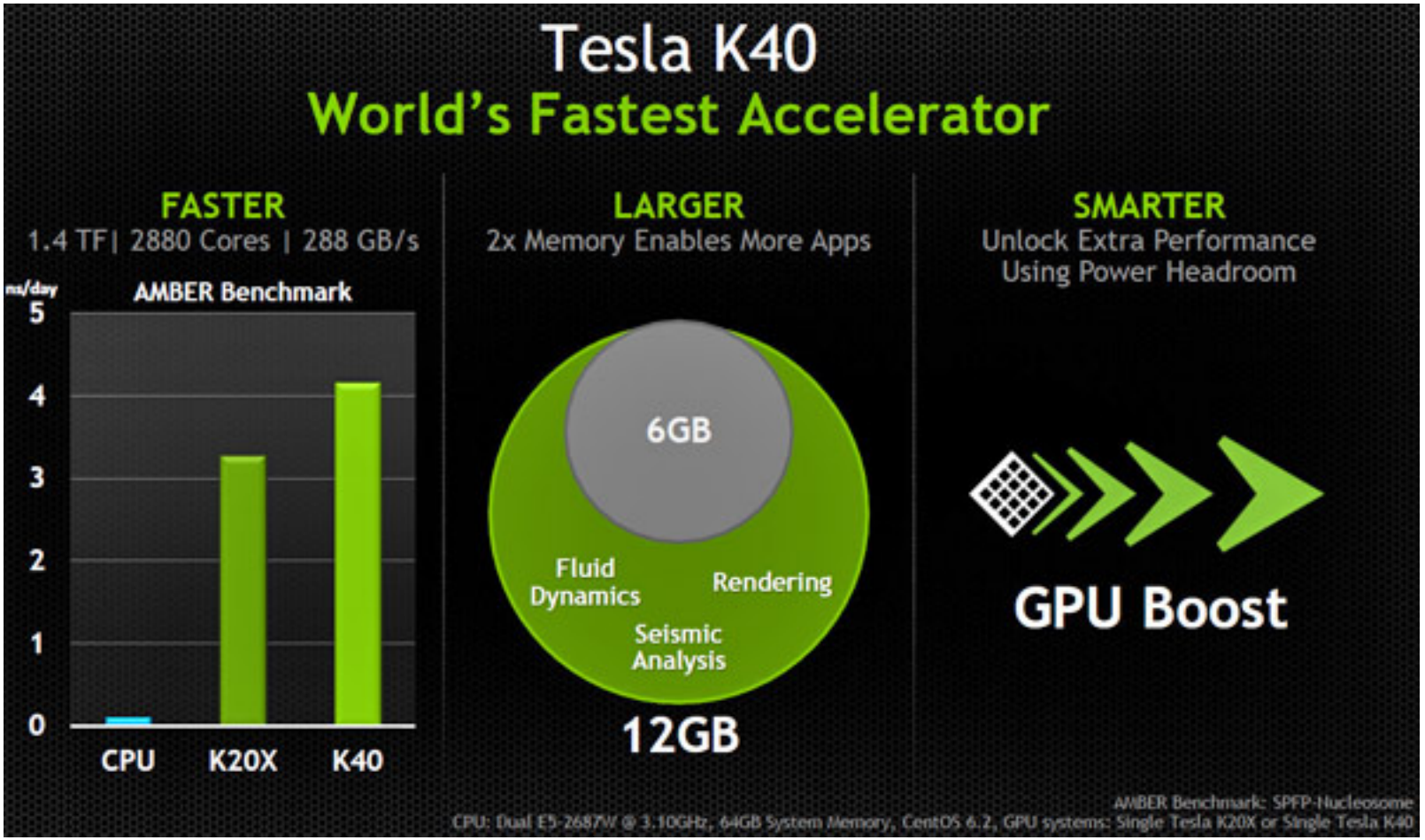


IMAGENET



Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. "Imagenet classification with deep convolutional neural networks." *Advances in neural information processing systems*. 2012.

Why Now?


Tesla K40
World's Fastest Accelerator
FASTER
1.4 TF| 2880 Cores | 288 GB/s
ns/day
AMBER Benchmark
5
4
3
2
1
0
CPU
K20X
K40
LARGER
2x Memory Enables More Apps
6GB
Fluid Dynamics
Rendering
Seismic Analysis
12GB
SMARTER
Unlock Extra Performance
Using Power Headroom
GPU Boost
AMBER Benchmark: SPFP-Nucleosome
CPU: Dual E5-2687W @ 3.10GHz, 64GB System Memory, CentOS 6.2, GPU systems: Single Tesla K20X or Single Tesla K40

Next Class

- Topics
  - Probability Basics

- Reading
  - Sections 1.2, 1.5