Machine Learning

CUNY Graduate Center, Spring 2013

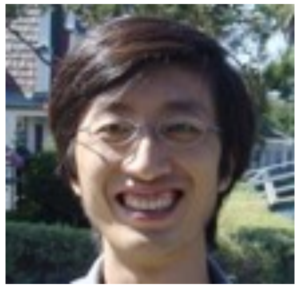
Professor Liang Huang

huang@cs.qc.cuny.edu

http://acl.cs.qc.edu/~lhuang/teaching/machine-learning

Logistics

- Lectures M 9:30-11:30 am Room 4419
- Personnel
  - Instructor Prof. Liang Huang huang@cs
  - TA Kai Zhao z.kaayy@gmail
- Office Hours
  - LH -- CS Lab M 11:30-12 pm (and occasionally on Fridays)
  - KZ -- CS Lab M TBD
  - additional office hours available before quizzes/exams



Logistics - cont'd

- Course Homepage
  - schedule, syllabus, homework, handouts, etc.
- Newsgroup
  - questions and discussions => post your Qs here first!
  - part of class participation (5% of grades)
  - we'll monitor newsgroup
- Announcements will be emailed to you
- Blackboard -- the [2nd] worst software I ever used!
  - grades and electronic submissions

Grades (tentative)

- Homework: 15x4 = 60%.
  - programming exercises in Python + numpy
  - late penalty: you can submit only one HW late for 48 hours.
- Paper Presentation: 10%
- Final Project: 25%
- Class Participation: 5%
  - asking/answering questions in class and newsgroup
  - catching/fixing bugs in slides/exams/hw & other suggestions

Resources

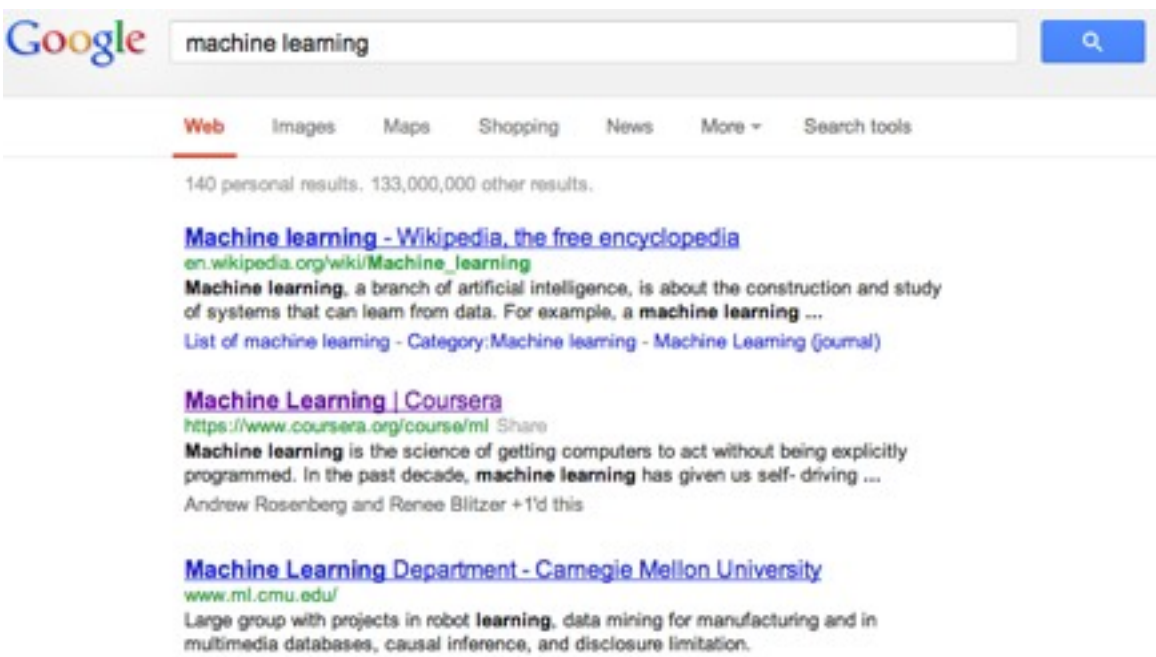
we will not follow any textbook.

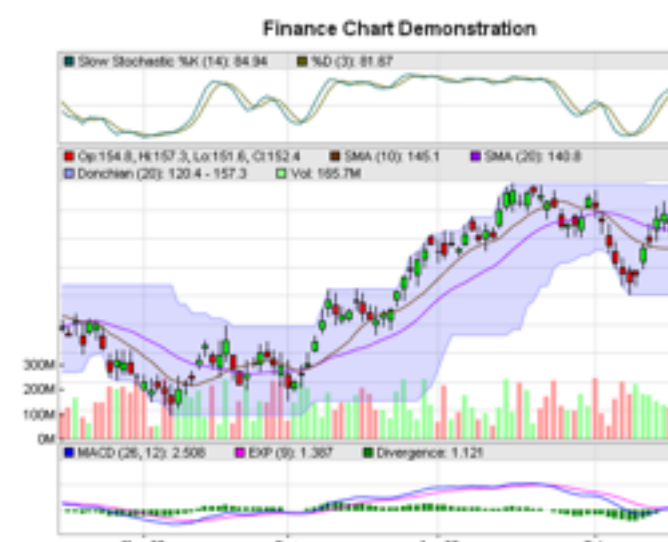
- Textbooks
  - ✓ Mitchell (1997). Machine Learning. *classical text; CS flavor. doesn't cover new stuff*
  - Duda et al (2001, 2nd). Pattern Classification. *stat/numeric-heavy; no CS perspective.*
  - Bishop (2006). PRML. *not for intro. stat-heavy. little CS perspective.*
  - Marsland (09). ML: An Algorithmic Perspective. too simple; bad figures; python code
  - Murphy (2012). ML: A Probabilistic Perspective. *little overlap with this course*
  - ✓ Mohri et al (2012). Foundations of ML. *rigorous CS perspective.*

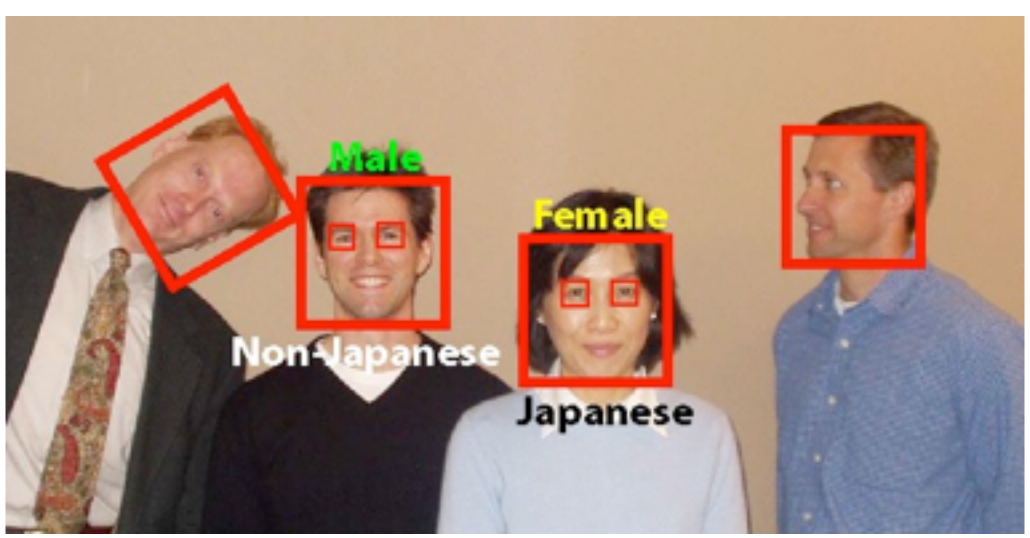
- Online Courses and Course Videos
  - Pedro Domingos (UW) on coursera. :)
  - Andrew Ng (Stanford) on coursera and youtube

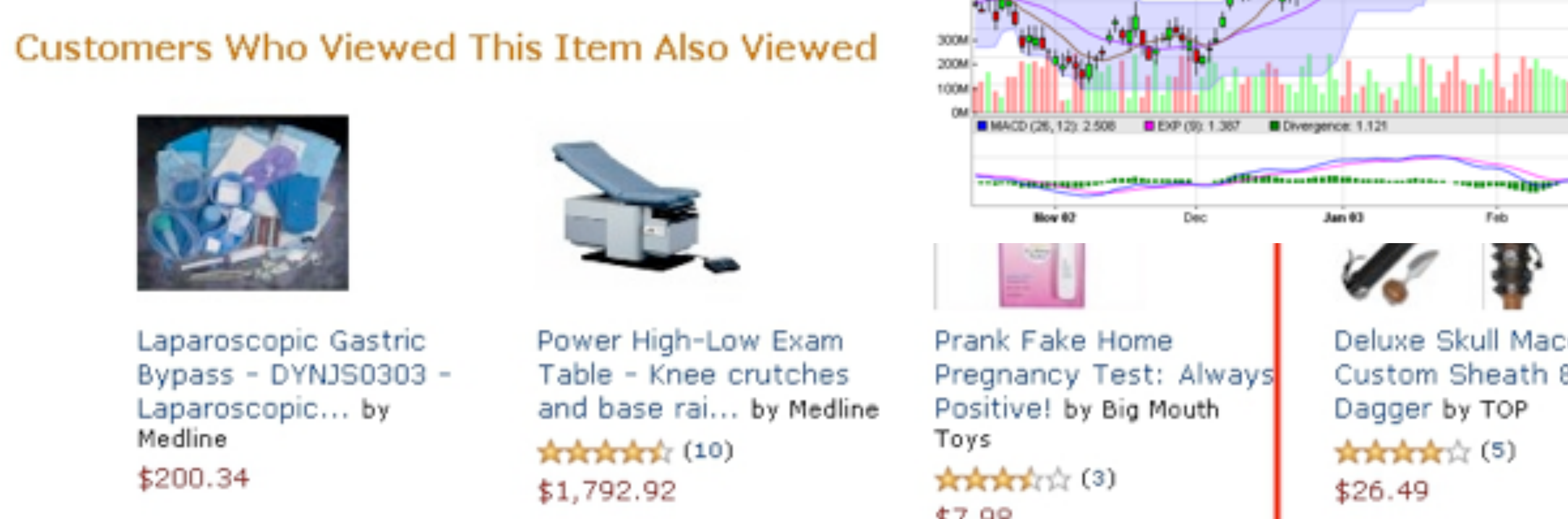
Machine Learning is Everywhere

- “A breakthrough in machine learning would be worth ten Microsofts” (Bill Gates)
- Machine learning is the hot new thing” (John Hennessy, President, Stanford)
- “Web rankings today are mostly a matter of machine learning” (Prabhakar Raghavan, Dir. Research, Yahoo)





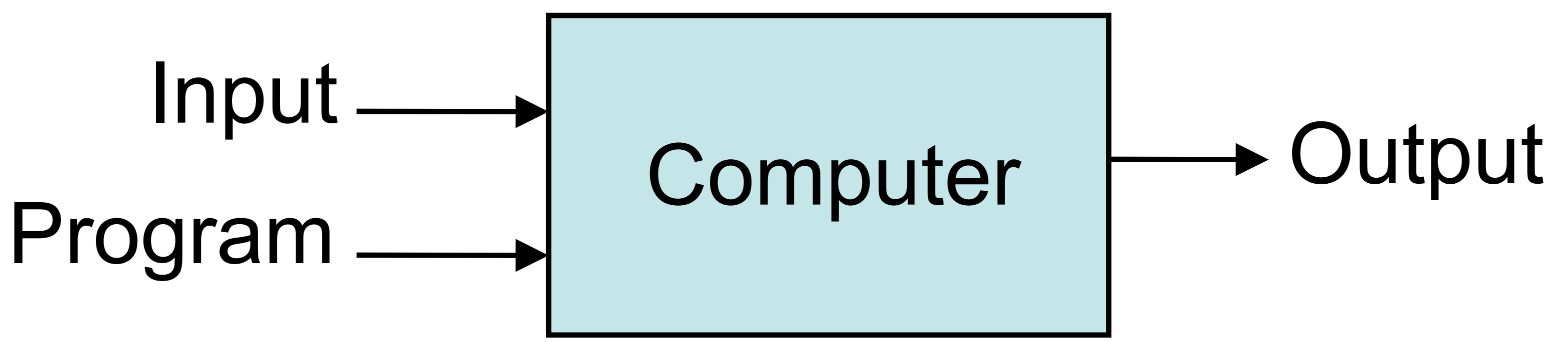




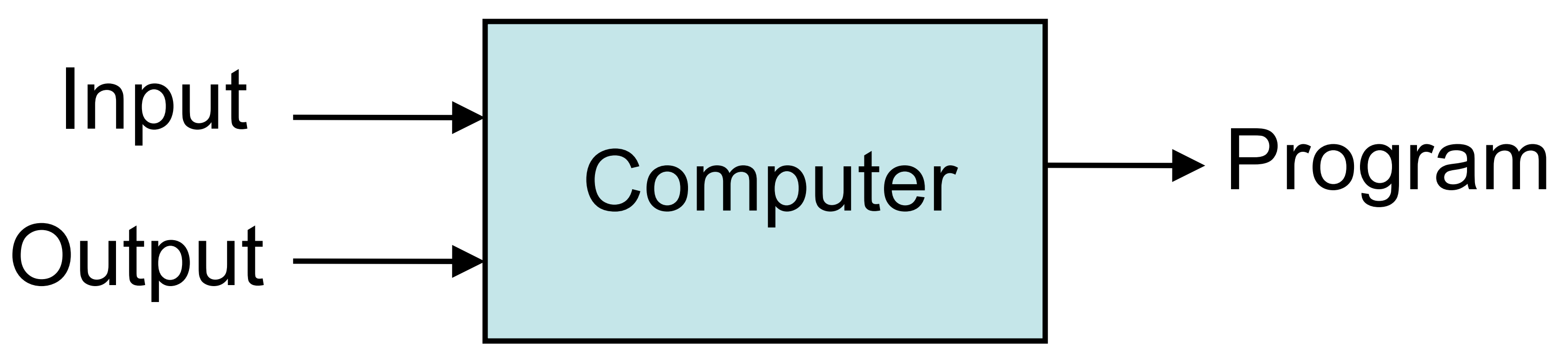
What is Machine Learning

- Machine Learning = Automating Automation
  - Getting computers to program themselves
  - Let the data do the work instead!

Traditional Programming



Machine Learning



Magic?

No, more like gardening

- **Seeds** = Algorithms
- **Nutrients** = Data
- **Gardener** = You
- **Plants** = Programs

ML in a Nutshell

- Tens of thousands of machine learning algorithms
- Hundreds new every year
- Every machine learning algorithm has three components:
  - **Representation**
  - **Evaluation**
  - **Optimization**

Representation

- Separating Hyperplanes
- Support vectors
- Decision trees
- Sets of rules / Logic programs
- Instances (Nearest Neighbor)
- Graphical models (Bayes/Markov nets)
- Neural networks
- Model ensembles
- Etc.

Evaluation

- Accuracy
- Precision and recall
- Squared error
- Likelihood
- Posterior probability
- Cost / Utility
- Margin
- Entropy
- K-L divergence
- Etc.

Optimization

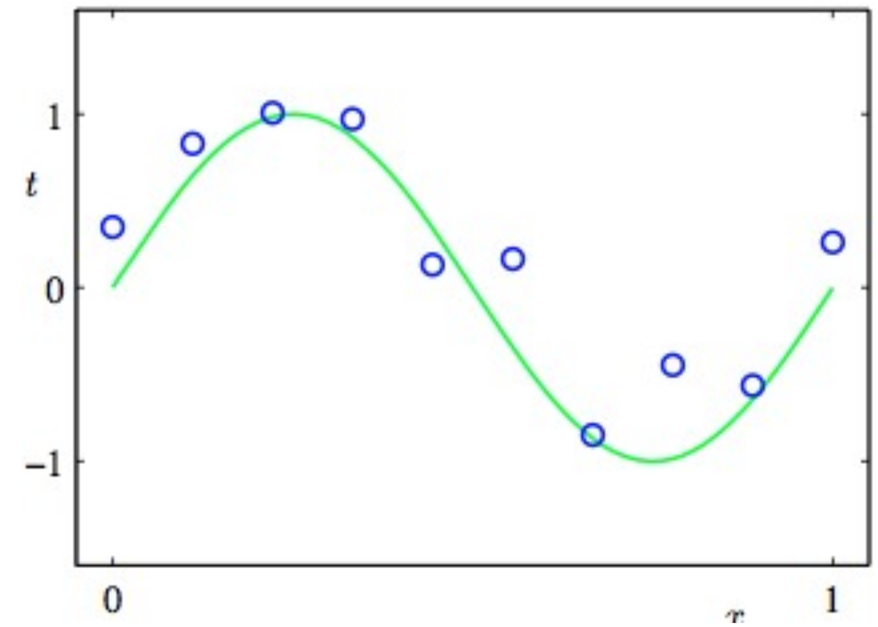
- Combinatorial optimization
  - E.g.: Greedy search, Dynamic programming
- Convex optimization
  - E.g.: Gradient descent, Coordinate descent
- Constrained optimization
  - E.g.: Linear programming, Quadratic programming

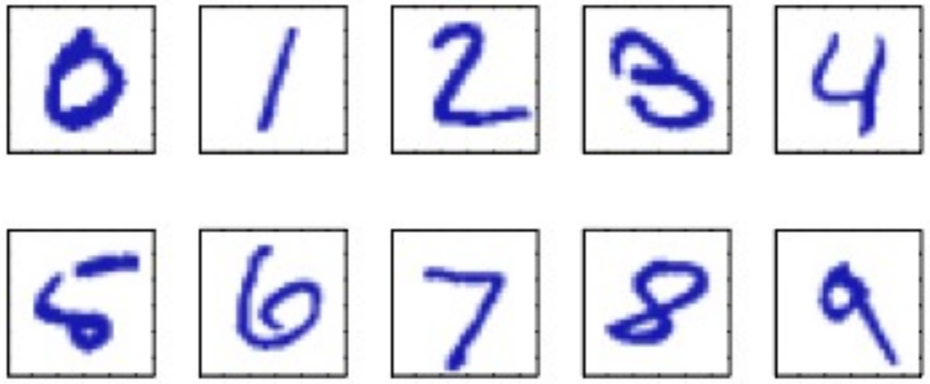
Types of Learning

- **Supervised (inductive) learning**
  - Training data includes desired outputs
- **Unsupervised learning**
  - Training data does not include desired outputs
- **Semi-supervised learning**
  - Training data includes a few desired outputs
- **Reinforcement learning**
  - Rewards from sequence of actions

Supervised Learning

- **Given** examples (X, $f(X)$) for an unknown function f
- **Find** a good approximation of function f
  - Discrete $f(X)$: Classification (binary, multiclass, structured)
  - Continuous $f(X)$: Regression



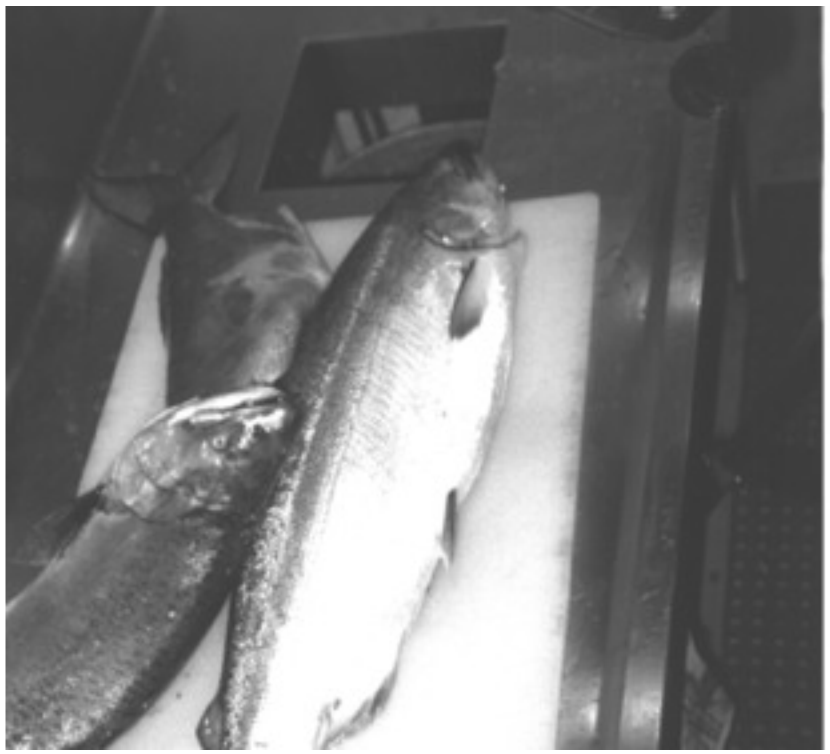


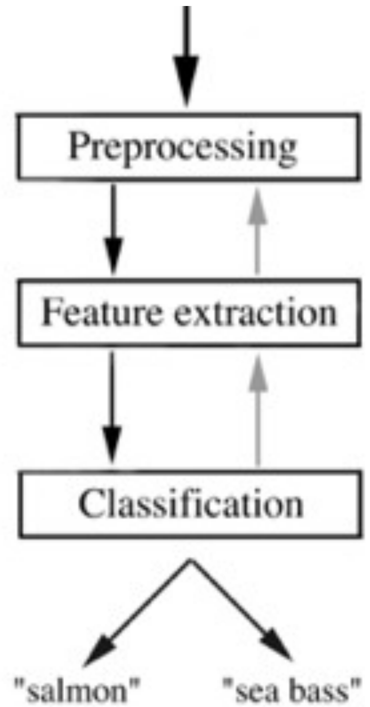
When is Supervised Learning useful

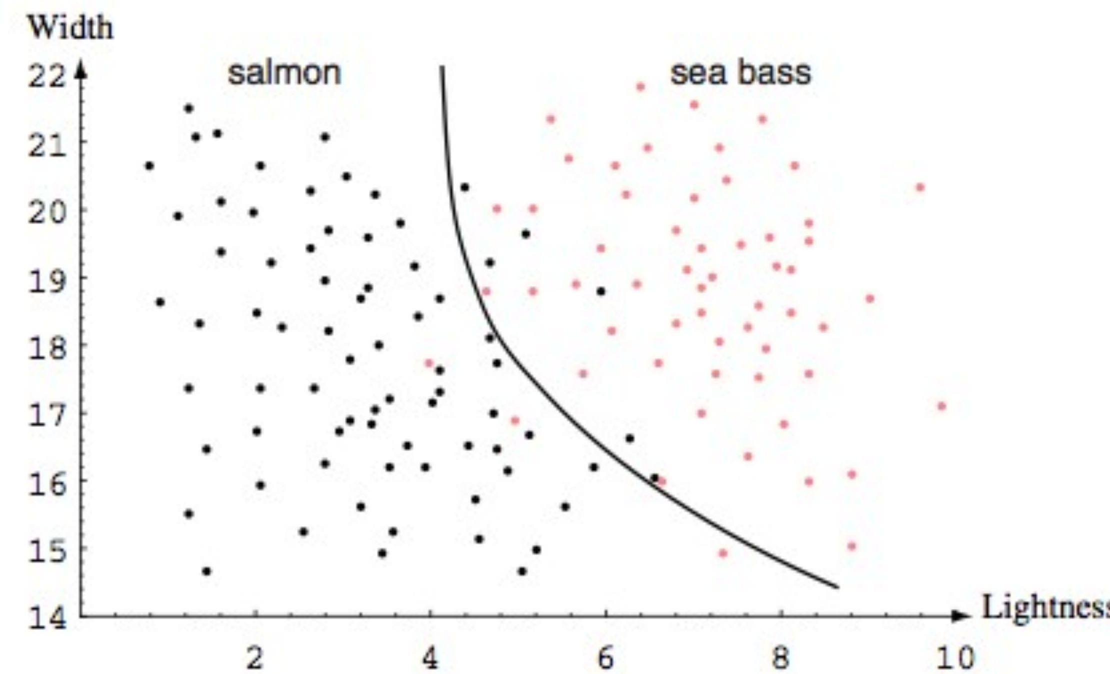
- when there is no human expert
  - input x: bond graph for a new molecule
  - output $f(x)$: predicted binding strength to AIDS protease
- when humans can perform the task but can't describe it
  - computer vision: face recognition, OCR
- where the desired function changes frequently
  - stock price prediction, spam filtering
- where each user needs a customized function
  - speech recognition, spam filtering

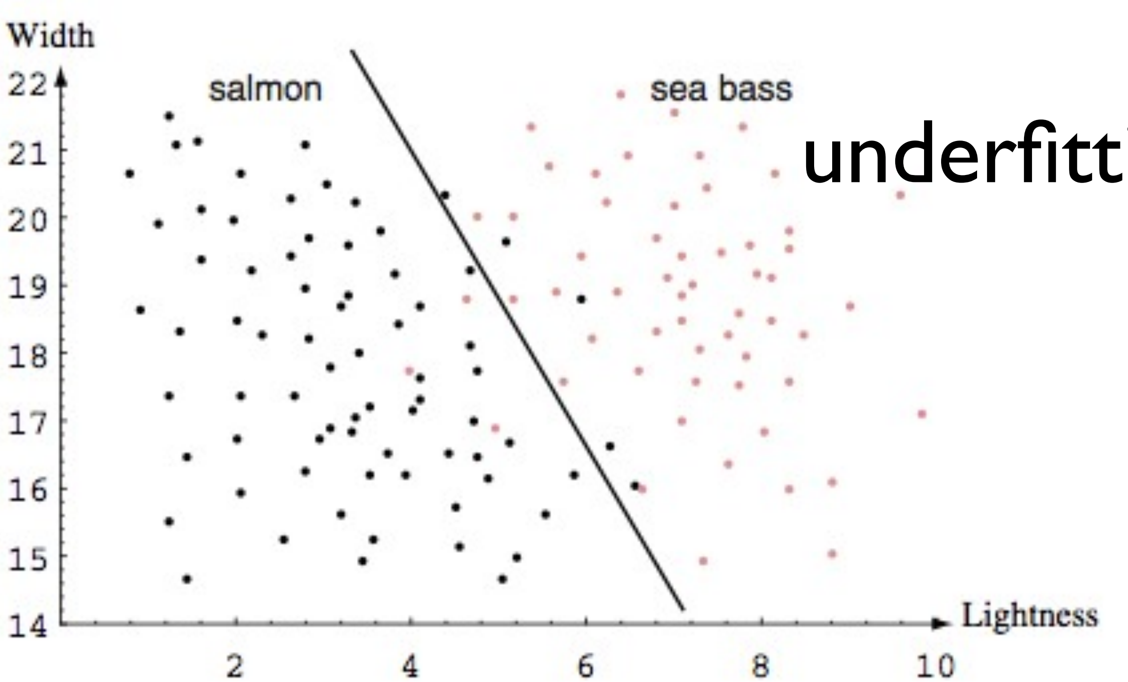
Classification

- input *X*: feature representation ("observation")

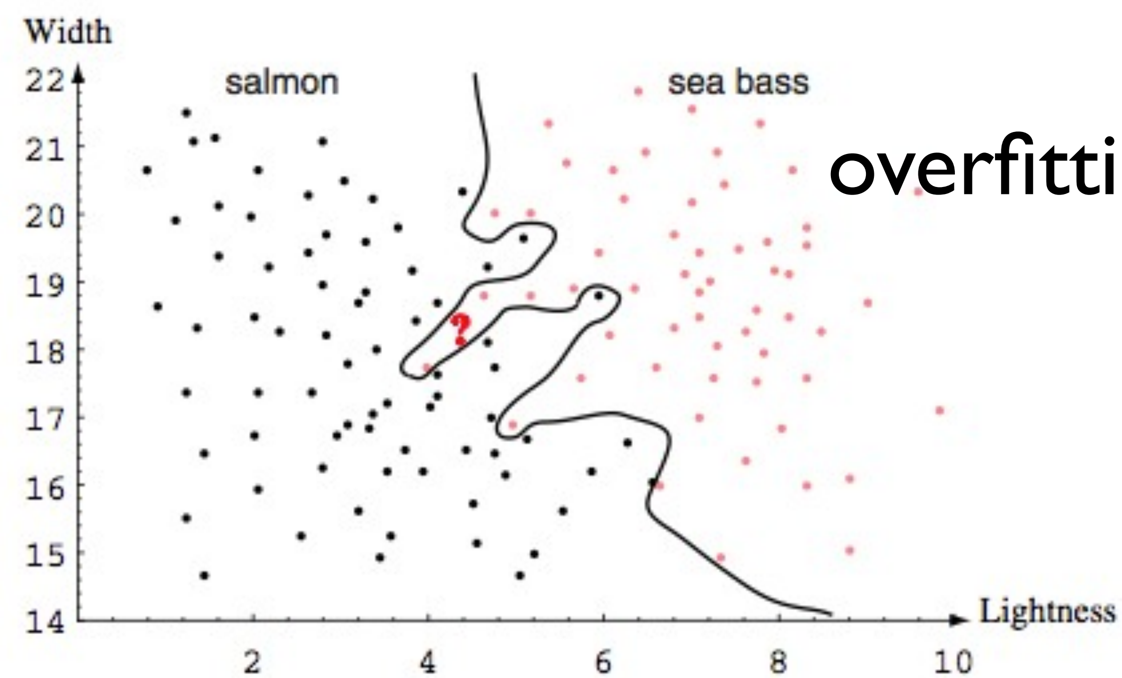








underfitting



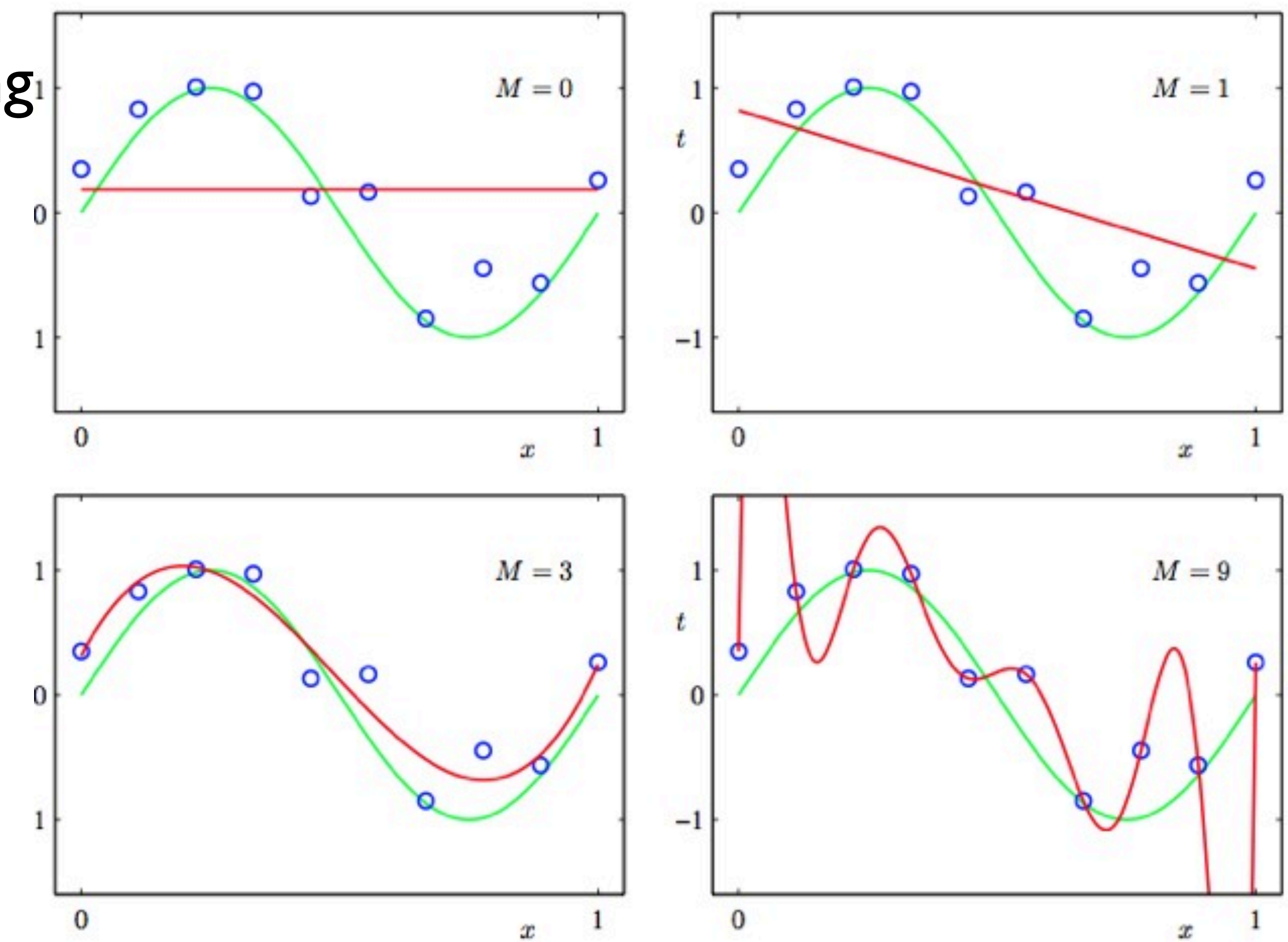
overfitting

Regression

- linear and non-linear regression
- overfitting and underfitting (same as in classification)
  - how to choose the optimal model complexity?

underfitting

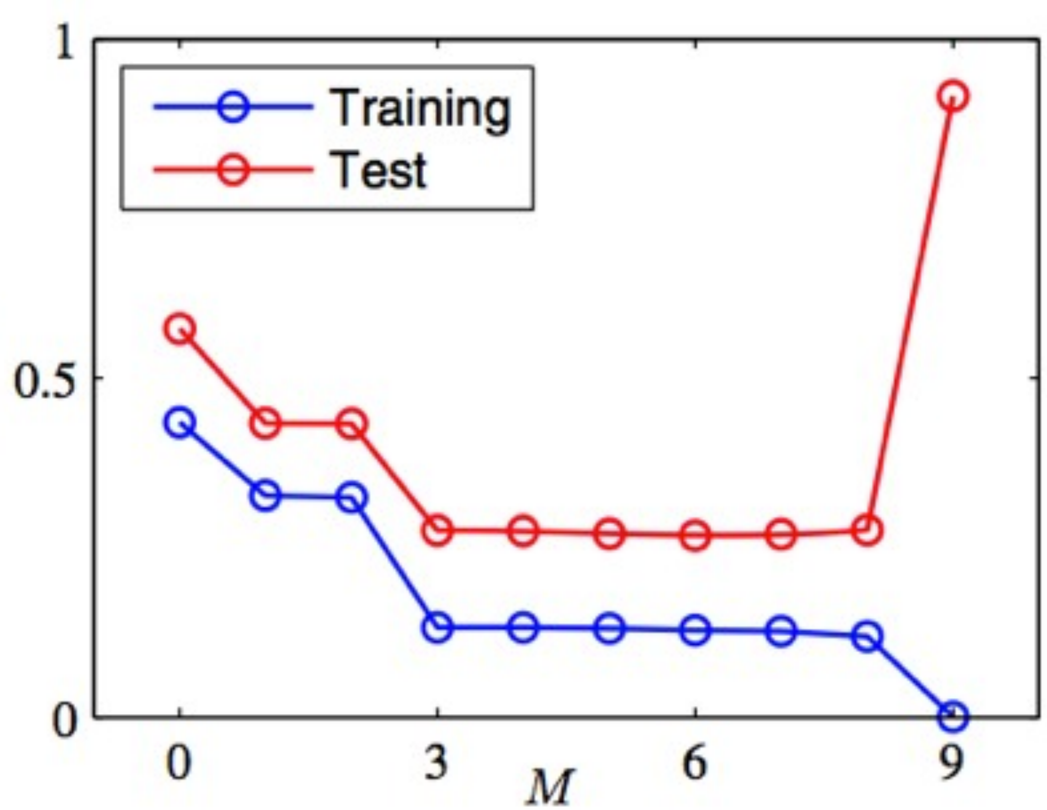
underfitting



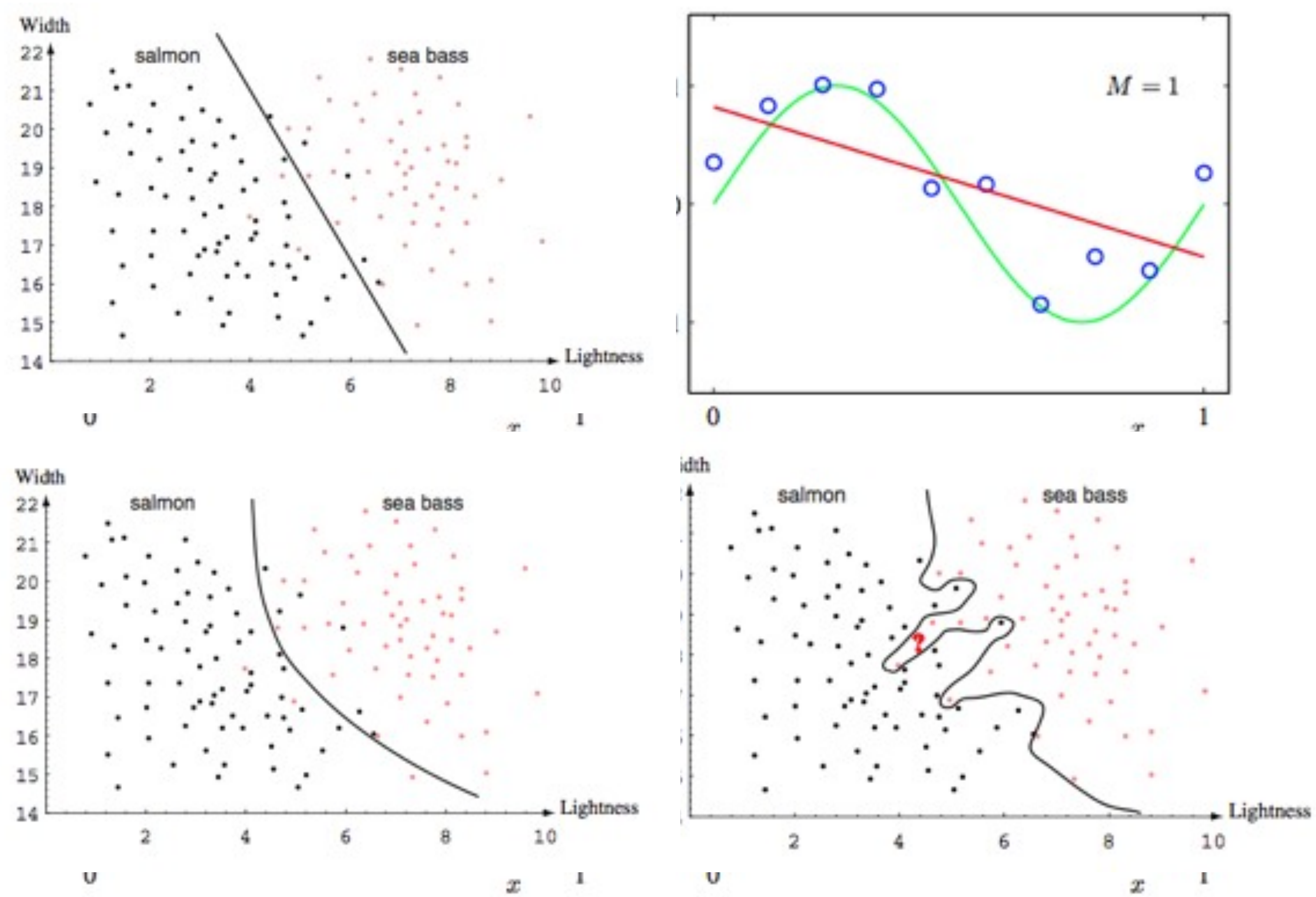
overfitting

Training, Test, & Generalization Error

- but you don't know test data a priori
  - generalization error: prob. of error on possible test data
- use held-out training data to "simulate" test-data

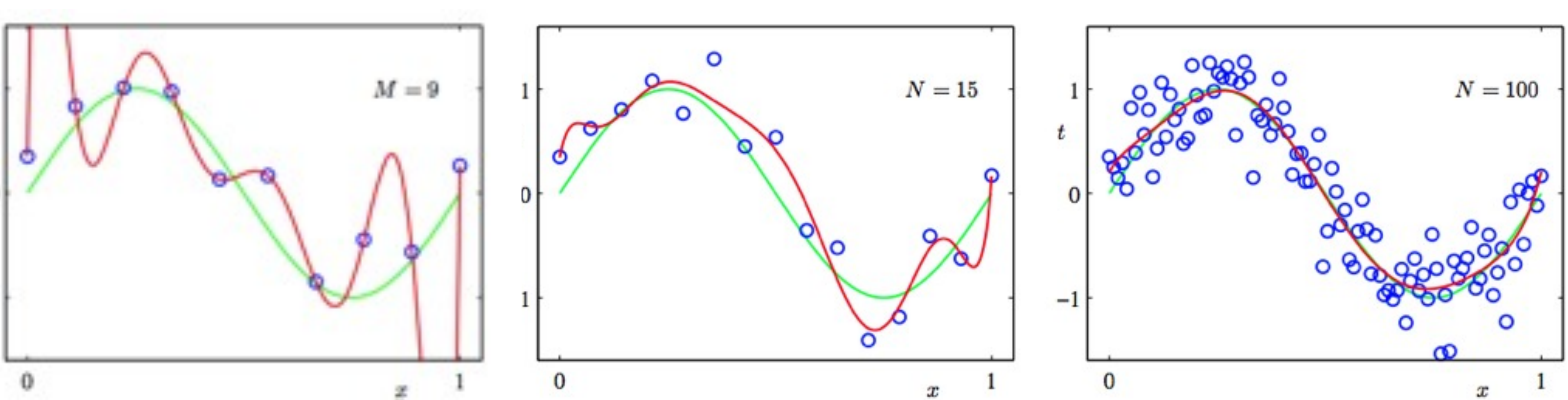


underfitting overfitting



Ways to Prevent Overfitting

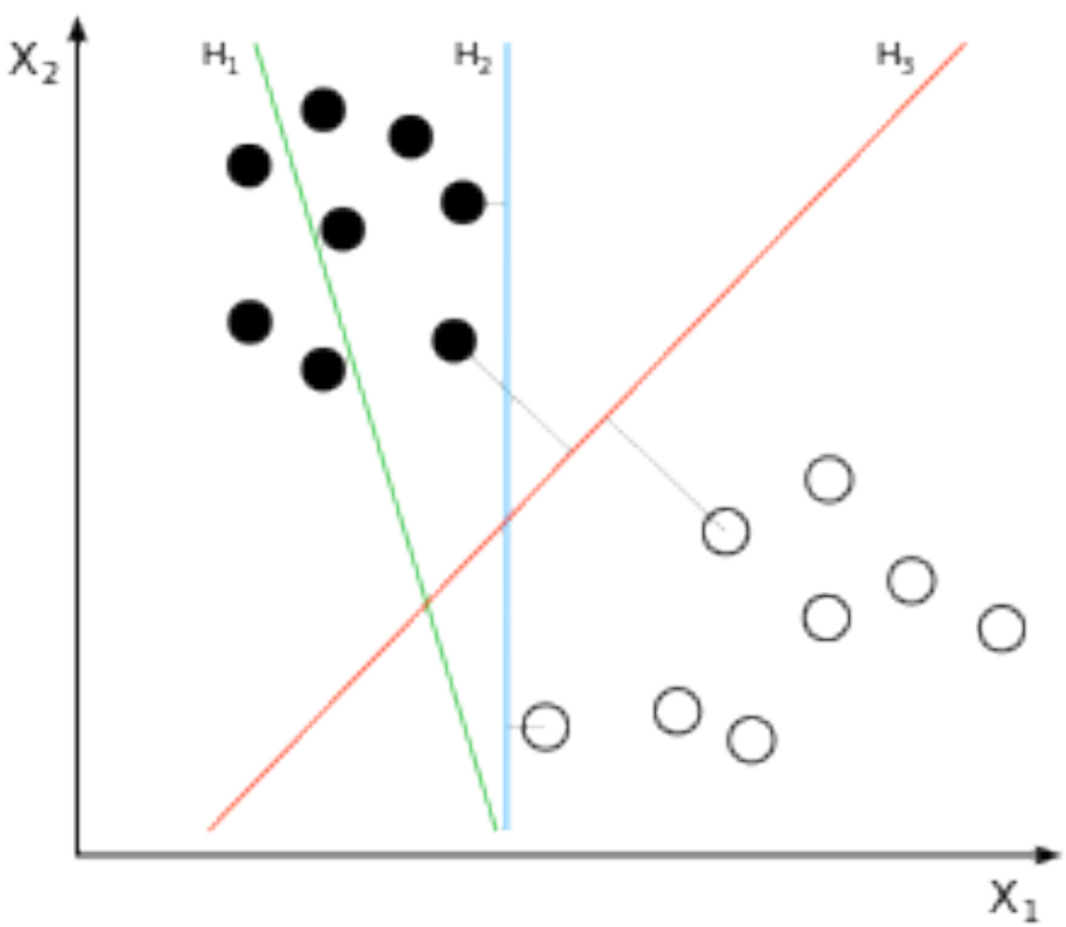
- held-out data to simulate generalization error
- more data points (overfitting is more likely on small data)
    - assuming same model complexity
- regularization (explicit control of model complexity)

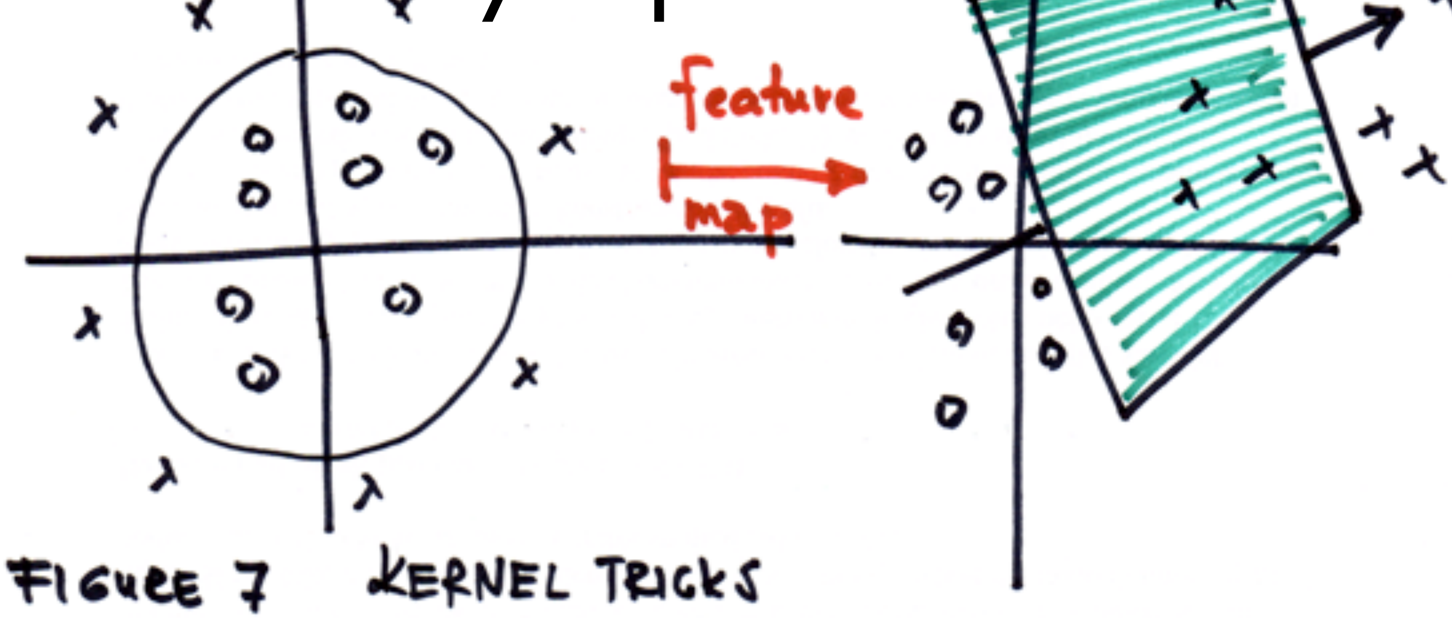


polynomials of degree 9

Linear Classification

- Q1: how to learn a separating hyperplane
- Q2: how to learn the optimal separating hyperplane
- Q3: what if the data is NOT linearly separable

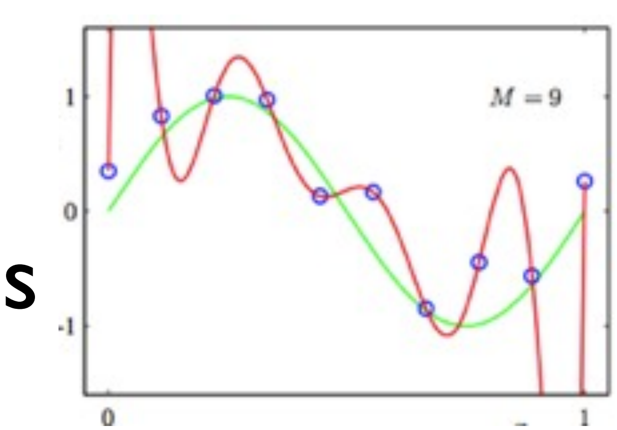




low-dimension

high-dimension

potential overfitting
in very high dimensions



What We'll Cover

- **Supervised learning**
  - Linear Classification and Online Learning (Perceptron)
  - Support vector machines
  - Decision tree induction
  - Rule induction
  - Instance-based learning (e.g. Nearest Neighbors)
  - Learning theory
- **Unsupervised learning**
  - Clustering
  - Dimensionality reduction

Gradient Descent

- if learning rate is too big, it'll diverge
- if learning rate is too small, it'll converge very slowly

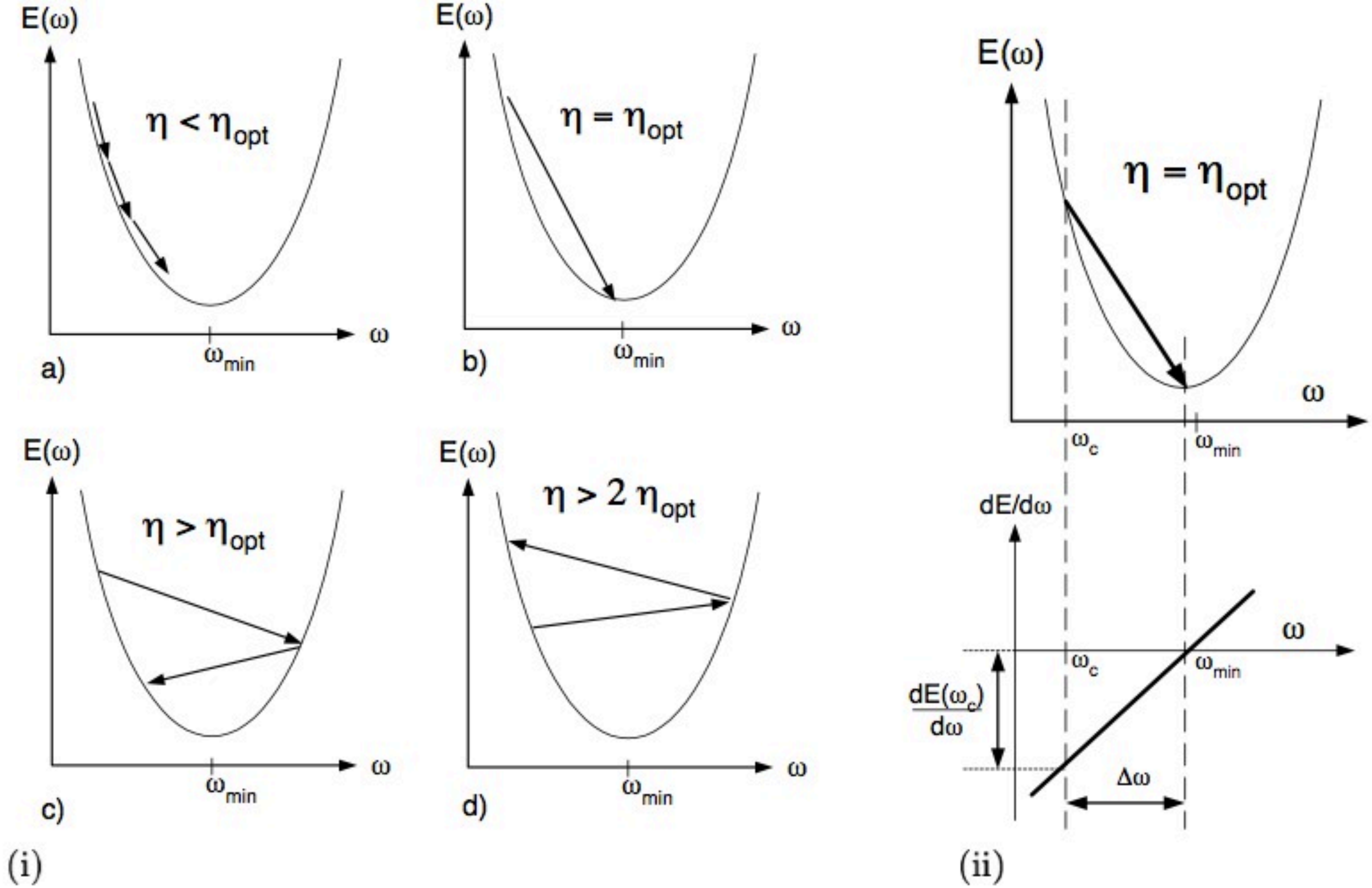
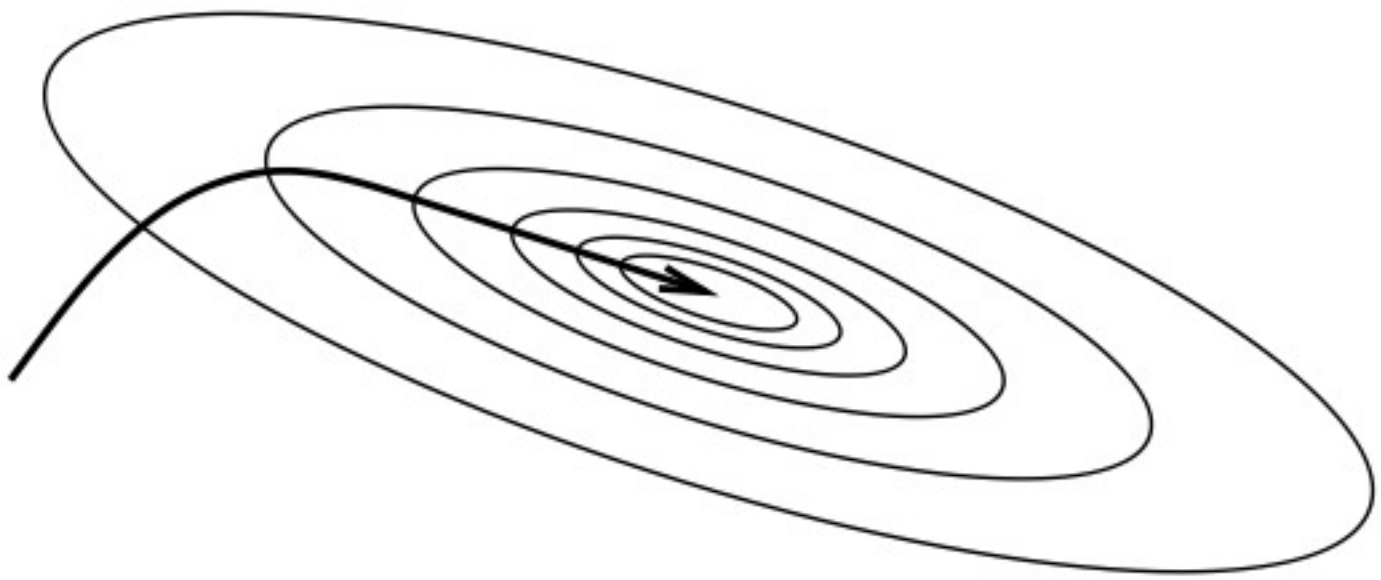
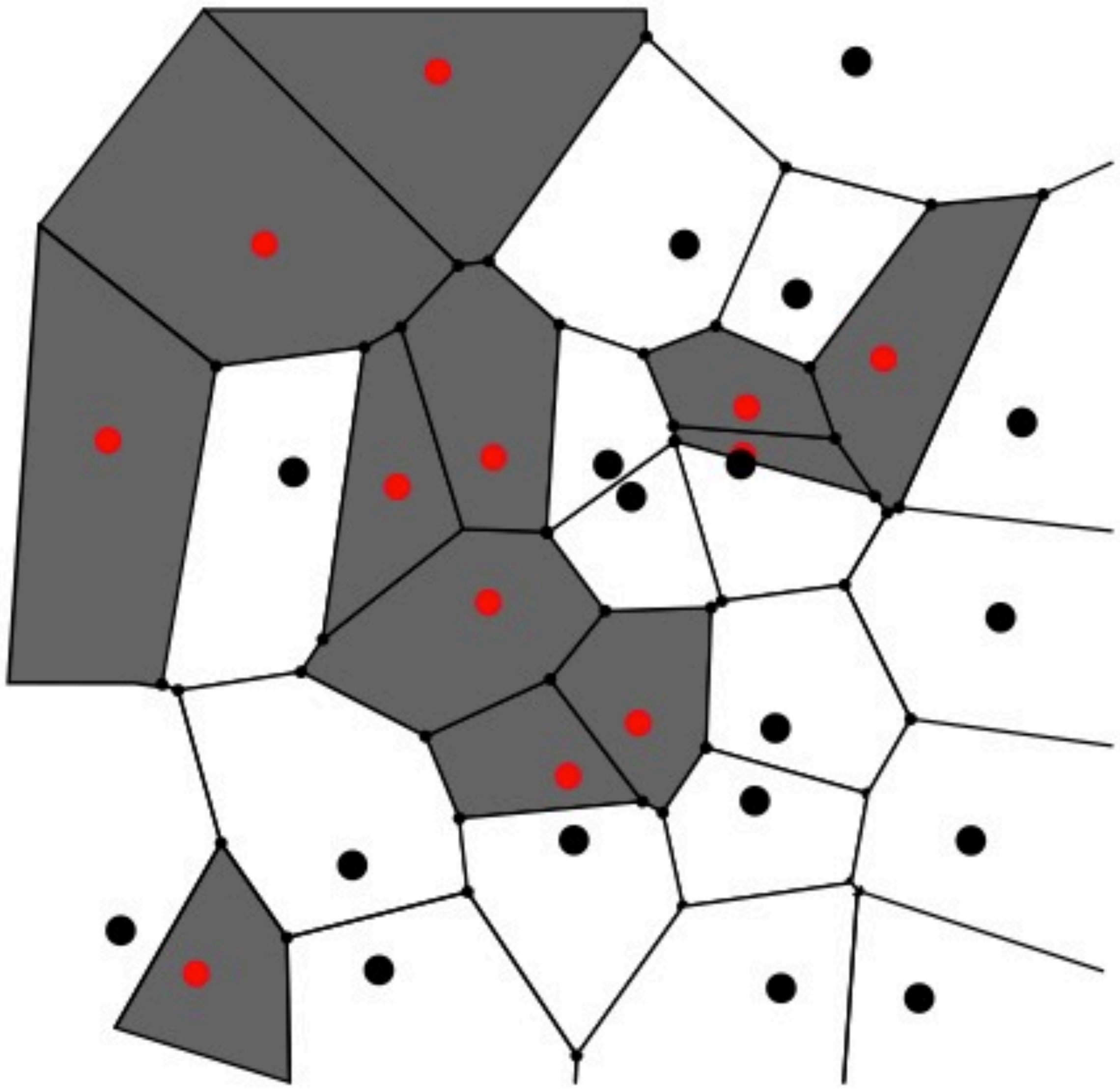


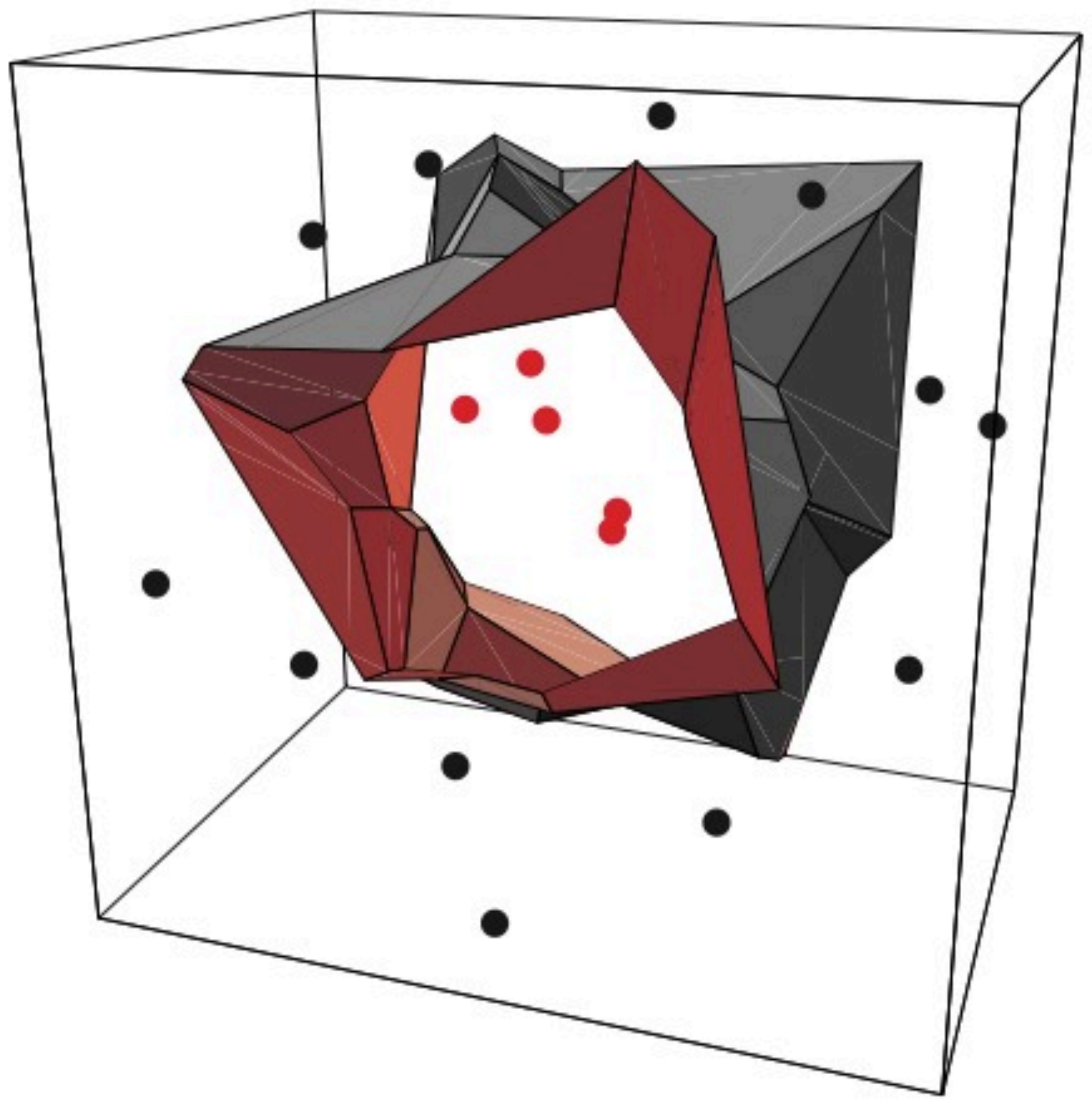
Fig. 6. Gradient descent for different learning rates.

Gradient Descent



NN Voronoi in 2D and 3D





Voronoi for Manhattan Distance