

Introduction to Machine Learning

1. Overview

Alex Smola
Carnegie Mellon University

<http://alex.smola.org/teaching/cmu2013-10-701>
10-701



MAGIC Etch A Sketch[®] SCREEN

Administrative stuff

Horizontal
Lid

OHIO ART The World of Toys[®]

Vertical
Lid

MAGIC SCREEN IS GLASS SET IN CURVED PLASTIC FRAME.
USE WITH CARE.

Important Stuff

- Lectures Monday and Wednesday 12:00-1:20pm
- Recitation Tuesday 5-6pm
- Office hours Tuesday 2-4pm (Alex), TBA (Barnabas)
- Grading policy (best 3 out of 4, final exam is mandatory)
- Project (33%)
Mid project report due after midterm
- Exams: Midterm (33%) and Final (34%)
The exams without technology. You can bring a paper notebook.
- Homework (33%)
Best 4 out of 5 homeworks. To receive points you must submit on due date in class.
No exceptions.
- Google Group <https://groups.google.com/forum/#!forum/10-701-spring-2013-cmu>
(questions, discussions, announcements)
- Homepage <http://alex.smola.org/teaching/cmu2013-10-701/>
(videos, problems, slides, timing, extra resources)

Projects & Homework

- Don't copy. You won't learn anything if you do.
 - Teamwork is OK (encouraged) for discussions.
 - For projects 3 is a good number. 2-4 are OK.
 - Each member gets the same score.
 - Start your projects early.
 - Ask for comments and feedback on projects
- Can we beat the Stanford class?

<http://cs229.stanford.edu/projects2012.html>

Color Coding

- Really important stuff
- Important stuff
- Regular stuff



Feedback please

- Let Barnabas and me (or the TAs) know if you have comments, concerns, suggestions!

This is our FIRST class at CMU.

Outline

- Basics
Problems, Statistics, Applications
- Standard algorithms
Naive Bayes, Nearest Neighbors, Decision Trees, Neural Networks, Perceptron
- (Generalized) Linear Models
Support Vector Classification, Regression, Novelty Detection, Kernel PCA
- Theoretical Tools
Risk Minimization, Convergence Bounds, Information Theory
- Probabilistic Methods
Exponential Families, Graphical Models, Dynamic Programming, Latent Variables, Sampling
- Interacting with the environment
Online Learning, Bandits, Reinforcement Learning
- Scalability

Outline

- Basics

Problems, Statistics, Applications

for the internet

- all you need for a startup

Support Vector Classification, Decision Trees, Neural Networks, Perceptron

- Support Vector Classification, Regression

Kernel PCA

- Theoretical Tools

Risk Minimization, Convergence B

for your PhD

- Replicability Methods

for Wall Street

Graphical Models, Dynamic Programming, Latent

- Interacting with the environment

Online Learning, Reinforcement Learning

biology

- Scalability

energy



MAGIC Etch A Sketch[®] SCREEN

Programming
with data

Horizontal
Lid

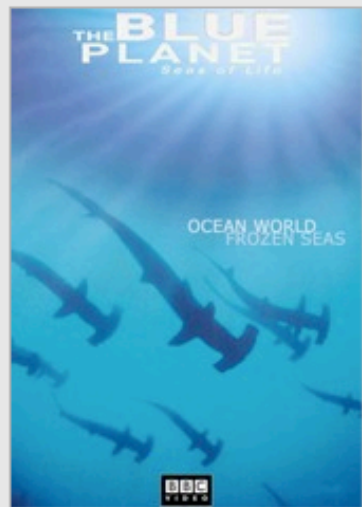
OHIO ART The World of Toys[®]

Vertical
Lid

MAGIC SCREEN IS GLASS SET IN CURVED PLASTIC FRAME.
USE WITH CARE.

Collaborative Filtering

Recently Watched

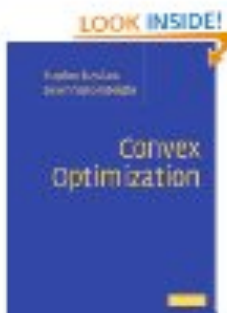


Top 10 for Alexander



Don't mix preferences on Netflix!

Customers Who Bought This Item Also Bought



Convex Optimization by Stephen Boyd
★★★★☆ (11)
\$65.78



Point Processes
(Chapman & Hall / CRC Monographs on S... by D.R. Cox
\$125.47



Probabilistic Graphical Models: Principles and Techniques by Daphne Koller
★★★★☆ (5)
\$71.52

Amazon books

Imitation Learning in Games



Avatar learns from
your behavior

Black & White
Lionsgate Studios

Imitation Learning



Drivatar in Forza

FORZA MOTORSPORT | 4

Spam Filtering

+Alex Search Images Maps Play YouTube News Gmail Drive Calendar More

Google Alex Smola 0 + Share

Gmail 1-50 of 15,803

COMPOSE

Inbox (7,180)
Important
Sent Mail
Drafts (61)

<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Southwest Airlines	Your trip is around the corner! - You're all set for your San Jose trip! My Account View My Itinerary Online	2:12 pm
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	DiscountMags.com	\$3.99 Business & Finance Sale.. starts now! - Trouble Seeing This Email? View as Webpage STOP these e-r	12:03 pm
<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	support, Alex (3)	Your order has shipped... - please send to the address below for an exchange remoteresremotes.com(exchange)	7:22 am
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	American Airlines AAdvan.	AAdvantage eSummary - January 2013 - VIEW IN WEB BROWSER >> http://americanairlines.ed10.net/r/JC	1:17 am
<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	Taesup, Alex, Taesup (3)	Happy new year! - Hi Alex, Thanks for your condolence. I will arrive at Berkeley on 16th (wed) night. So, I car	Jan 11

+Alex Search Images Maps Play YouTube News Gmail Drive Calendar More

Google Alex Smola 0 + Share

Gmail 1-50 of 244

COMPOSE

Inbox (7,180)
Important
Sent Mail
Drafts (61)
All Mail

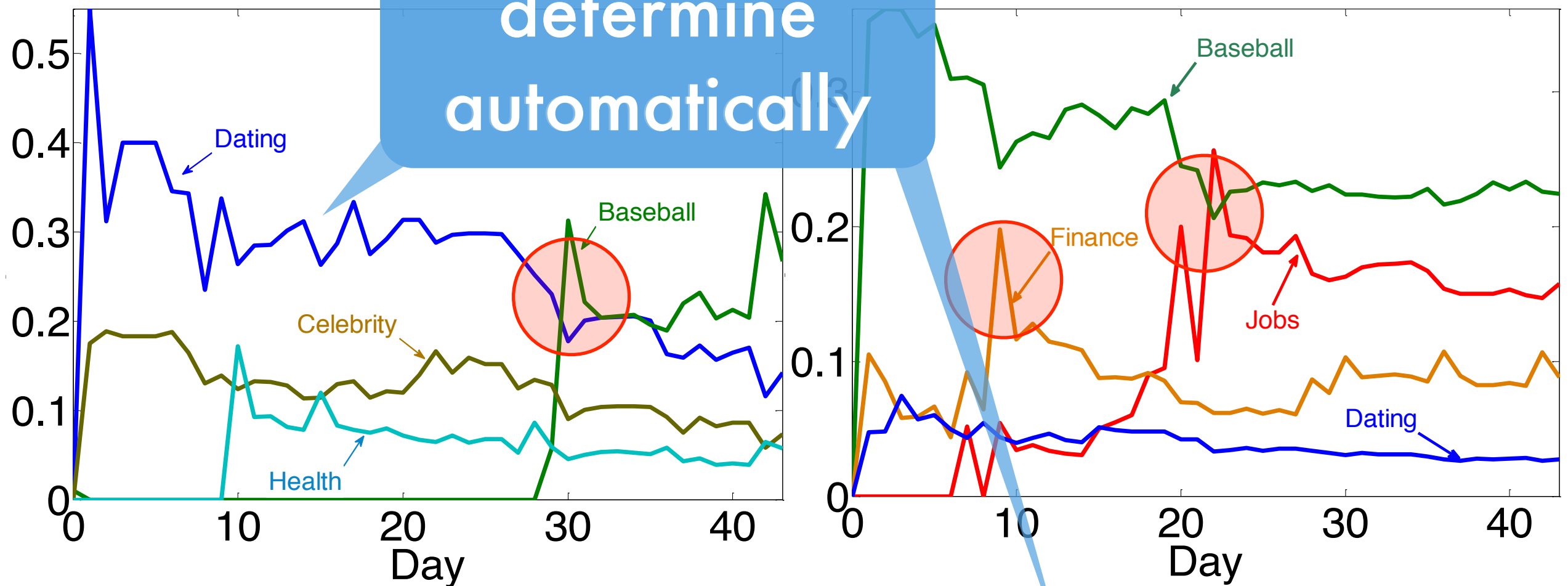
▶ Circles
▼ [Gmail]
Done (1,006)
[Imap]/Drafts
[Imap]/Sent
alex.smola@yah...

Delete all spam messages now (messages that have been in Spam more than 30 days will be automatically deleted)

<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	maee	(Ei&ISTP Index)2013机械与自动化工程国际会议征文: [alex.smola@gmail.com] - 尊敬的老师, 您好: 机械与	Jan 11
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Dear Valued Customers,	Low Interest Rate Loan - Dear Valued Customers, Do you need a loan or funding for any of the following reas	Jan 11
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	garjeti	Call for Research Papers - GLOBAL ADVANCED RESEARCH JOURNAL OF ENGINEERING, TECHNOLOG	Jan 11
<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	Steven Cooke	Congratulations Alex, \$150 awaits you - Alex: IMPORTANT - NOTICE OF WINNINGS Please make sure yo	Jan 11
<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	paper18	【2013-1-15截稿】 【2013年机电与控制工程亚太地区学术研讨会APCMCE 2013】 【EI】 【香港】 【不参-不要.	Jan 10
<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	First-Class Mail Service	Tracking ID (G)BGD35 849 603 4893 4550 - Fed Ex Order: JN-3339-28981768 Order Date: Thursday, 3 Janua	Jan 10
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	garjeti	Call for Research Papers - GLOBAL ADVANCED RESEARCH JOURNAL OF ENGINEERING, TECHNOLOG	Jan 10
<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	Candy.Li	中层,不只当老板的代言人	Jan 9
<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	Ronan Morgan	Ronan Morgan just sent you a personal message. - LinkedIn Ronan Morgan just sent you a private messag	Jan 9
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	RE/MAX®	2013 Valueable Offer! - Hello Friend, RE/MAX® has issued 2013 valuable property offer in your resident from	Jan 9
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	newsletter	newsletter WWW2013 - Newsletter 6 - See the Portuguese and Spanish version right after the English versior	Jan 9
<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	CJCR editor	Chinese Journal of Cancer Research (CJCR) has been indexed by Pubmed and PMC - Click here if this e-mail	Jan 9
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	garjeti (2)	Call for Research Papers - GLOBAL ADVANCED RESEARCH JOURNAL OF ENGINEERING, TECHNOLOG	Jan 9

User profiling

determine automatically



Dating

women
men
dating
singles
personals
seeking
match

Baseball

League
baseball
basketball,
doublehead
Bergesen
Griffey
bullpen
Greinke

Celebrity

Snooki
Tom
Cruise
Katie
Holmes
Pinkett
Kudrow
Hollywood

Health

skin
body
fingers
cells
toes
wrinkle
layers

Jobs

job
career
business
assistant
hiring
part-time
receptionist

Finance

financial
Thomson
chart
real
Stock
Trading
currency

Cheque reading

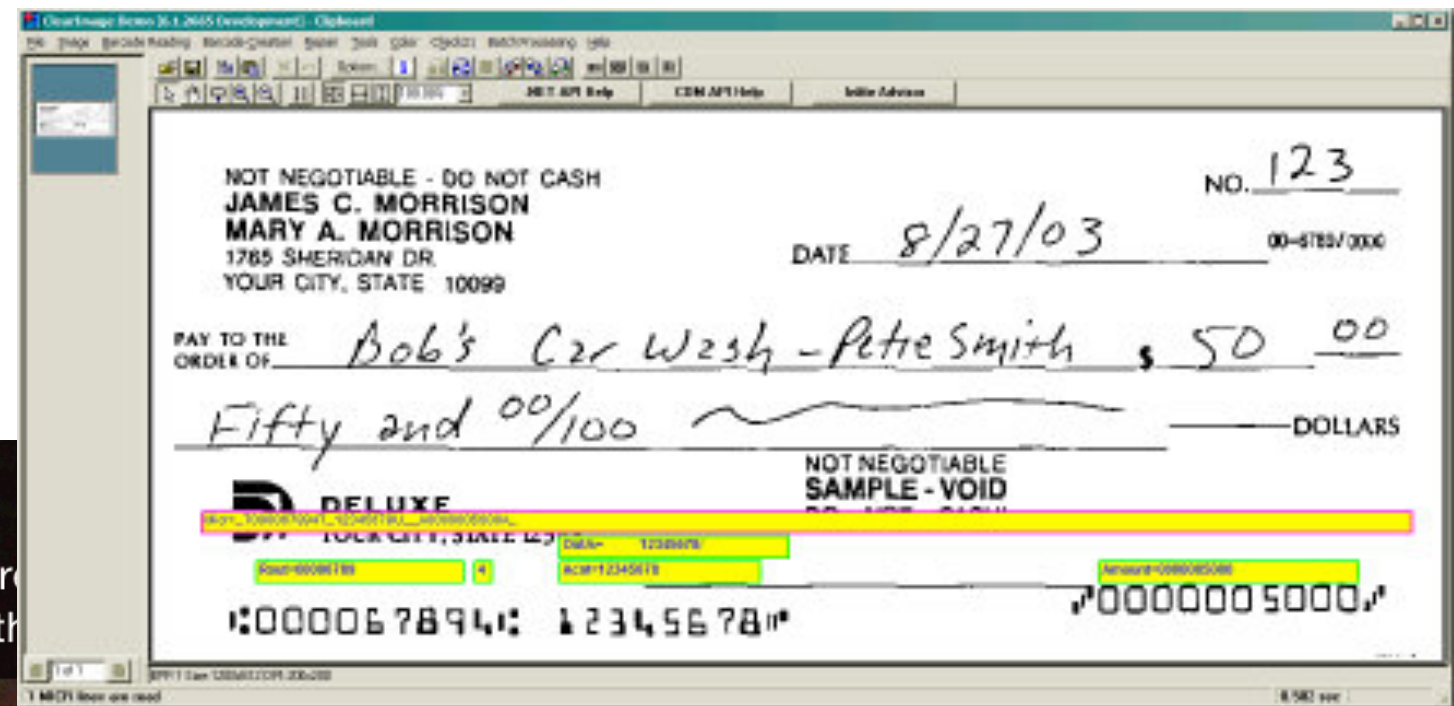
segment image

Photograph Front of Check

Place the check on a dark background in a well-lit area, hold the camera steady and align the check's edges with the frame.



Note: Fidelity cannot act on any written instructions



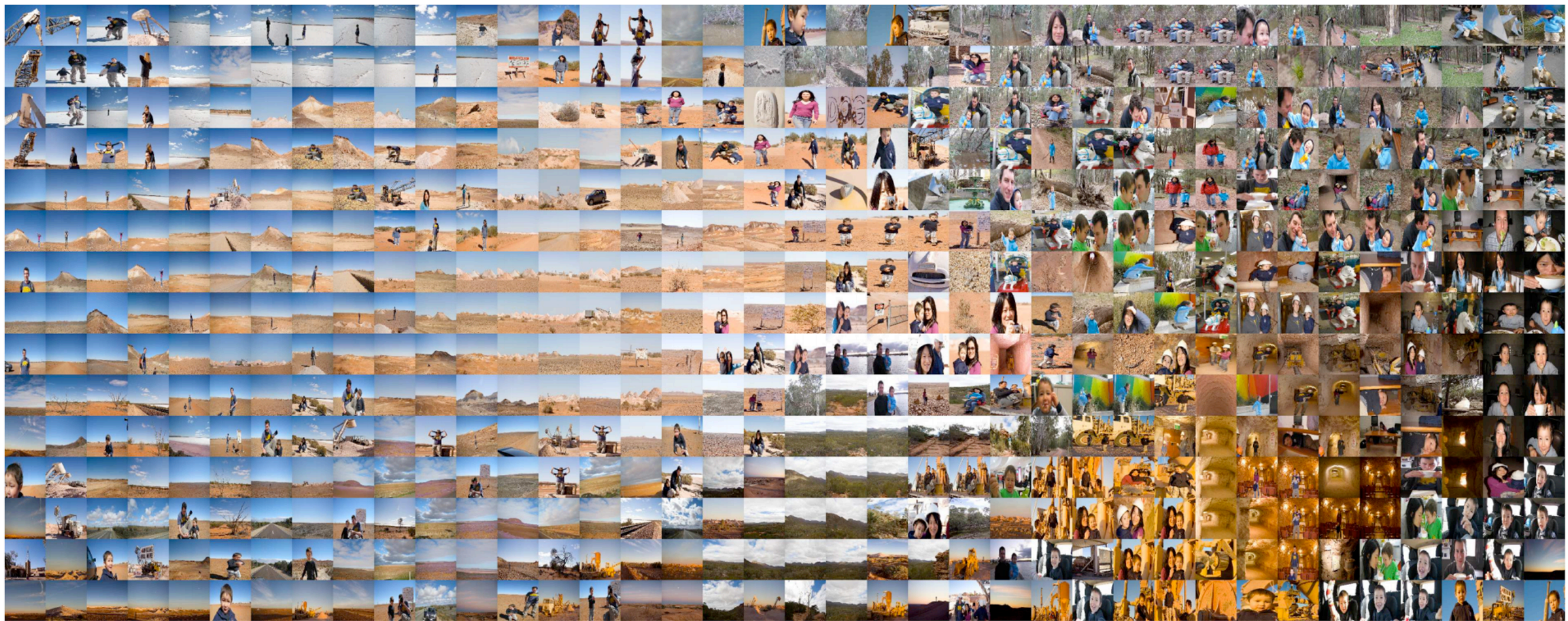
recognize
handwriting

Autonomous Helicopter

<http://heli.stanford.edu>

Carnegie Mellon University

Image Layout



- Raw set of images from several cameras
- Joint layout based on image similarity

Search ads

Google mesothelioma Alex

Web Images Maps Shopping News More Search tools

About 10,600,000 results (0.25 seconds)

Ads related to mesothelioma ⓘ

Mesothelioma Symptoms - Lung cancer from Asbestos.
www.mesothelioma-lung-cancer.org/
It can take 20-30 years to develop
What Is It? Symptoms
Portal Entrance Treatments

Mesothelioma Symptoms - 101 Facts about Mesothelioma.
www.mesothelioma-answer.org/
By Anna Kaplan, M.D.
Free Mesothelioma Book - Nutrition Book - Free Mesothelioma DVDs - Asbestos

Mesothelioma Diagnosis? - Get the money you deserve fast
www.mesotheliomaclaimscenter.info/
File with **Mesothelioma** Claim Center
Mesothelioma Compensation Amounts - File a Mesothelioma Claim

Mesothelioma - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/Mesothelioma
Mesothelioma (or, more precisely, malignant **mesothelioma**) is a rare form of cancer that develops from transformed cells originating in the mesothelium, the ...
Signs and symptoms - Cause - Diagnosis - Screening

Mesothelioma Cancer Alliance | The Authority on Asbestos Cancer
www.mesothelioma.com/
Mesothelioma treatment, diagnosis and related information for patients and families. Legal options for those diagnosed with malignant **mesothelioma**.

Ads ⓘ

Mesothelioma compensation
www.simmonsfirm.com/888-360-4189
Free Consultation with Lawyers that Focus on **Mesothelioma** Cases.

Mesothelioma Compensation
www.sokolovelaw.com/Call_Now
Mesothelioma Diagnosis? Get the Money You Deserve! [800-581-8243](tel:800-581-8243)

Mesothelioma 800-582-0706

You Don't Have To Sue Anyone.
\$30 Billion Asbestos Trust Fund

Mesothelioma & Asbestos
www.navy-veterans-mesothelioma.org/
Important info for Navy Vets.
Learn About **Mesothelioma** Claims

Asbestos Exposure?
www.mesotheliomalawfirm.com/
Mesothelioma victims are entitled

why these ads?

True startup story

- Startup builds exchange for ads on webpages
- Clients bid on opportunities, market takes a cut

- System gets popular
- Stuff works better if ads and pages are matched
 - Programmer adds a few IF ... THEN ... ELSE clauses (system improves)
 - Programmer adds even more clauses (system sort-of improves, ruleset is a mess)
 - Programmer discovers decision trees (lots of rules, but they work better)
 - Programmer discovers boosting (combining many trees, works even better)
- Startup is bought ... (machine learning system is replaced entirely)

Programming with Data

- Want adaptive robust and fault tolerant systems
- Rule-based implementation is (often)
 - difficult (for the programmer)
 - brittle (can miss many edge-cases)
 - becomes a nightmare to maintain explicitly
 - often doesn't work too well (e.g. OCR)
- Usually easy to obtain examples of what we want
- IF x THEN DO y
- Collect many pairs (x_i, y_i)
- Estimate function f such that $f(x_i) = y_i$ (supervised learning)
- Detect patterns in data (unsupervised learning)



MAGIC Etch A Sketch[®] SCREEN

Problem
Prototypes

Horizontal
Lid

OHIO ART The World of Toys[®]

Vertical
Lid

MAGIC SCREEN IS GLASS SET IN CURVED PLASTIC FRAME.
USE WITH CARE.

Supervised Learning $y = f(x)$

- Binary classification
Given x find y in $\{-1, 1\}$
- Multicategory classification
Given x find y in $\{1, \dots, k\}$
- Regression
Given x find y in \mathbb{R} (or \mathbb{R}^d)
- Sequence annotation
Given sequence $x_1 \dots x_l$ find $y_1 \dots y_l$
- Hierarchical Categorization (Ontology)
Given x find a point in the hierarchy of y (e.g. a tree)
- Prediction
Given x_t and $y_{t-1} \dots y_1$ find y_t

often with loss
 $l(y, f(x))$

Binary Classification

+Alex Search Images Maps Play YouTube News

Google

Gmail

COMPOSE

Inbox (7,180)

Important

Sent Mail

Drafts (61)

- Southwest Airlines
- DiscountMags.com
- support, Alex (3)
- American Airlines AAdv...
- Taesup, Alex, Taesup (3)

+Alex Search Images Maps Play YouTube News

Google

in:spam

Gmail

COMPOSE

Inbox (7,180)

Important

Sent Mail

Drafts (61)

All Mail

Circles

[Gmail]

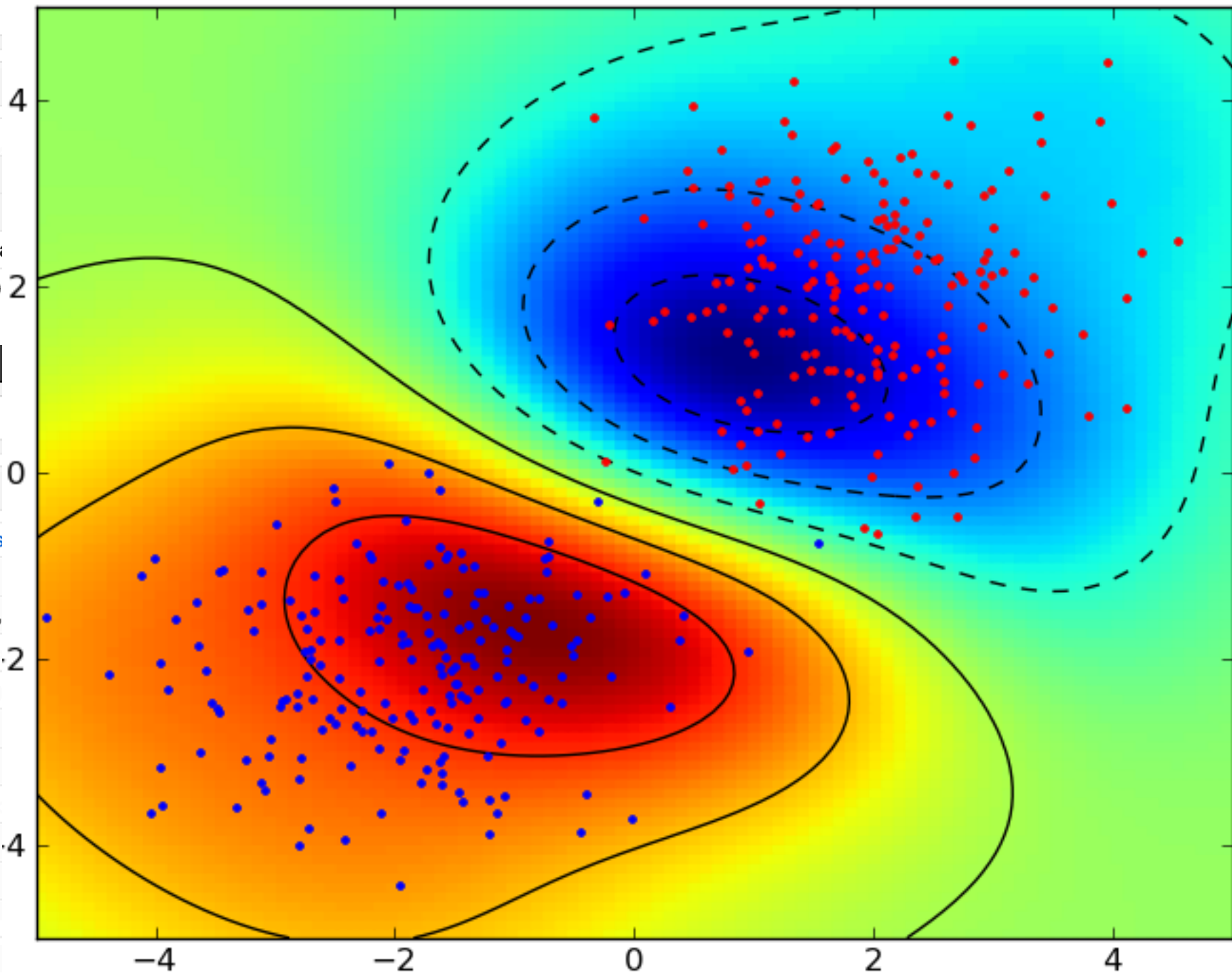
Done (1,006)

[Imap]/Drafts

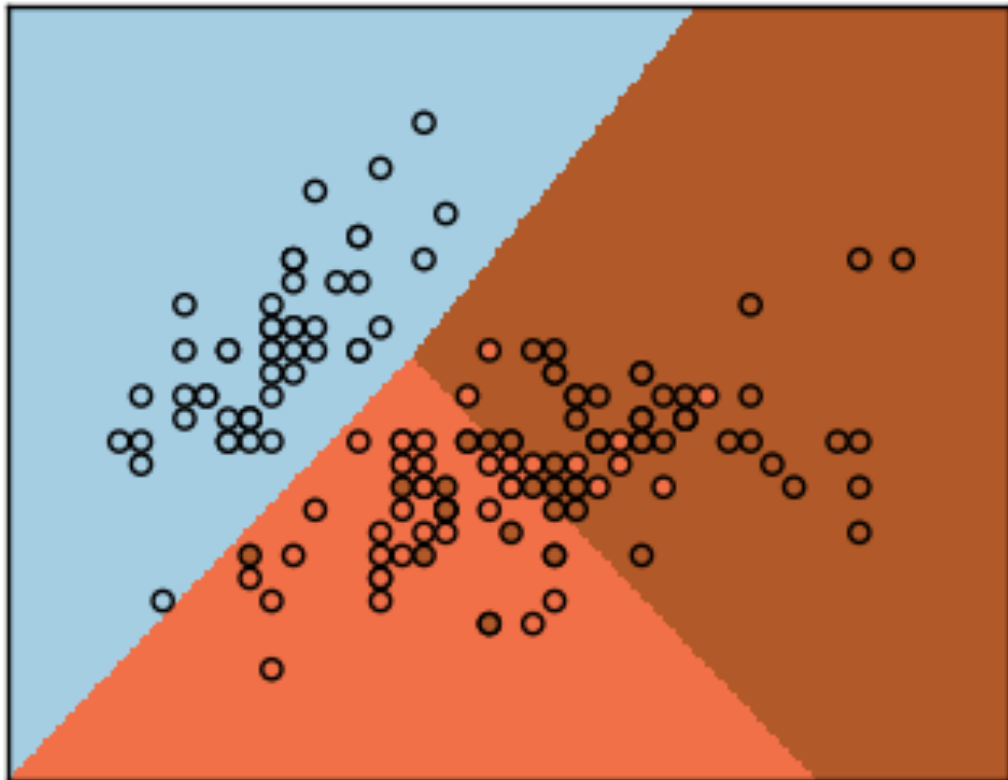
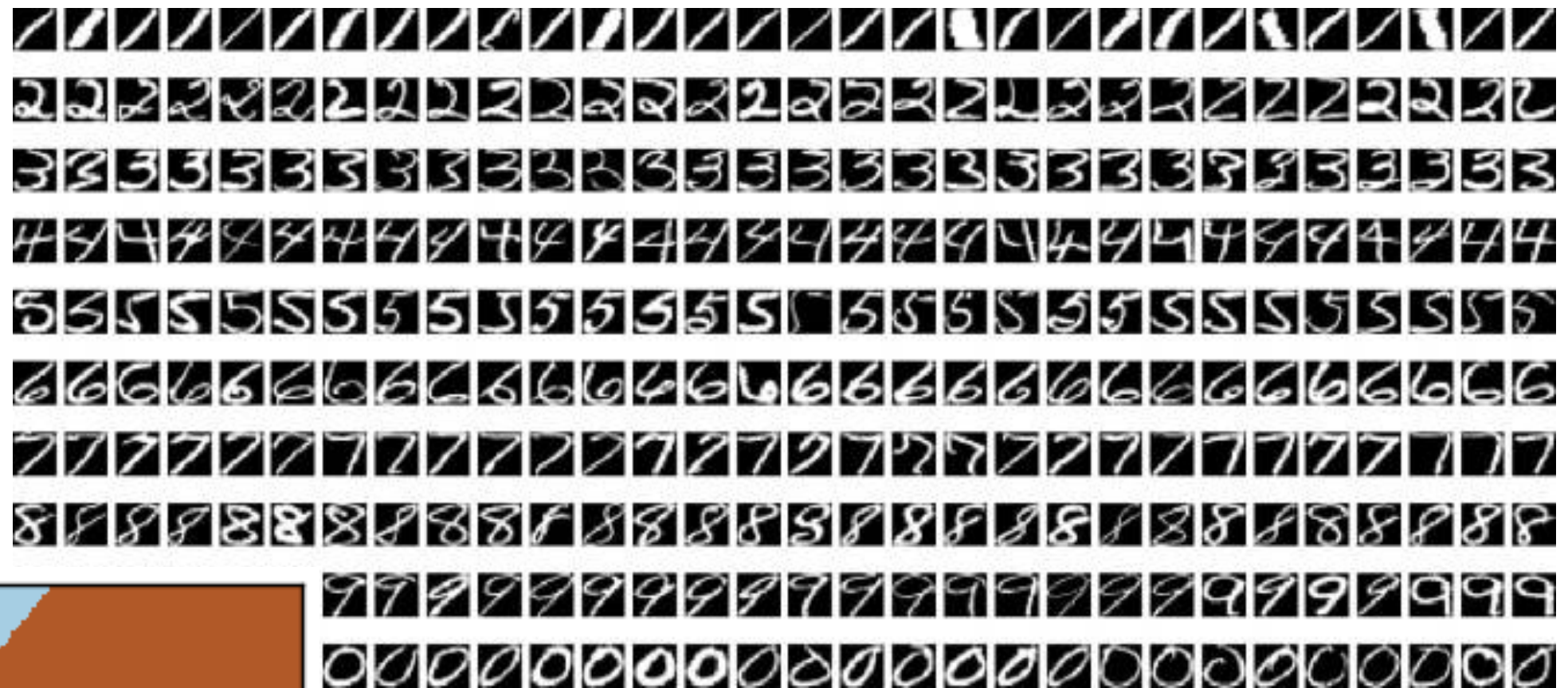
[Imap]/Sent

alex.smola@yah...

- maee
- Dear Valued Customers,
- garjeti
- Steven Cooke
- paper18
- First-Class Mail Service
- garjeti
- Candy.Li
- Ronan Morgan
- RE/MAX®
- newsletter
- CJCR editor
- garieti (2)

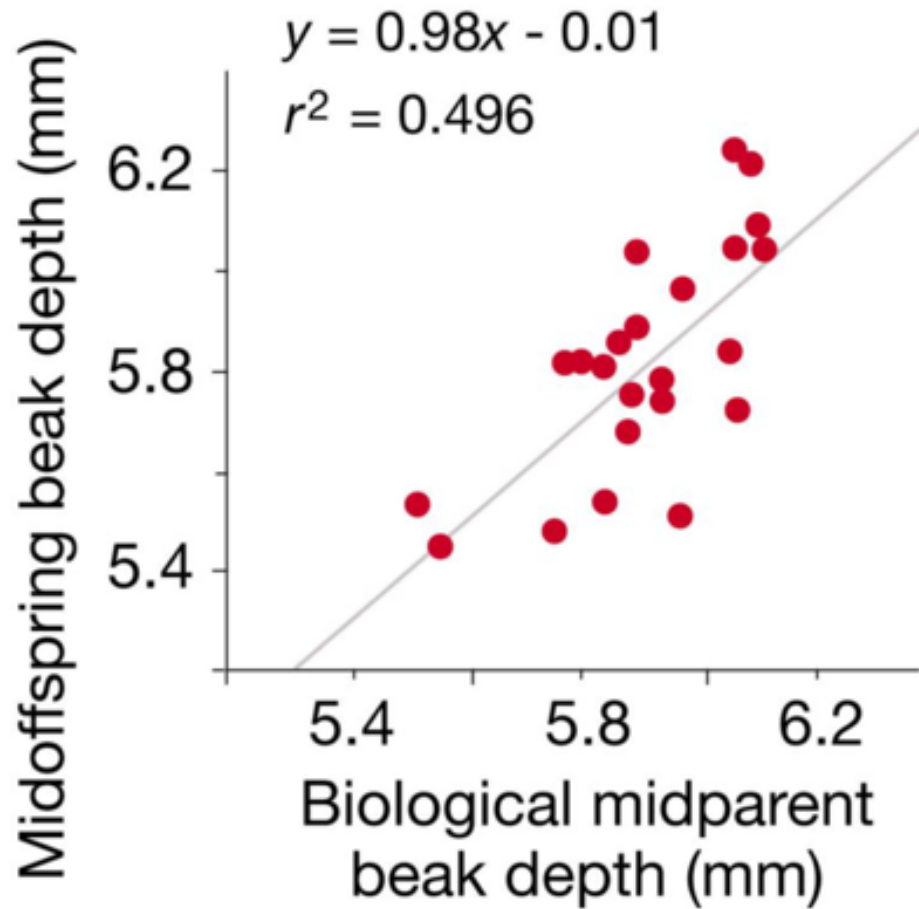


Multiclass Classification



map image x to digit y

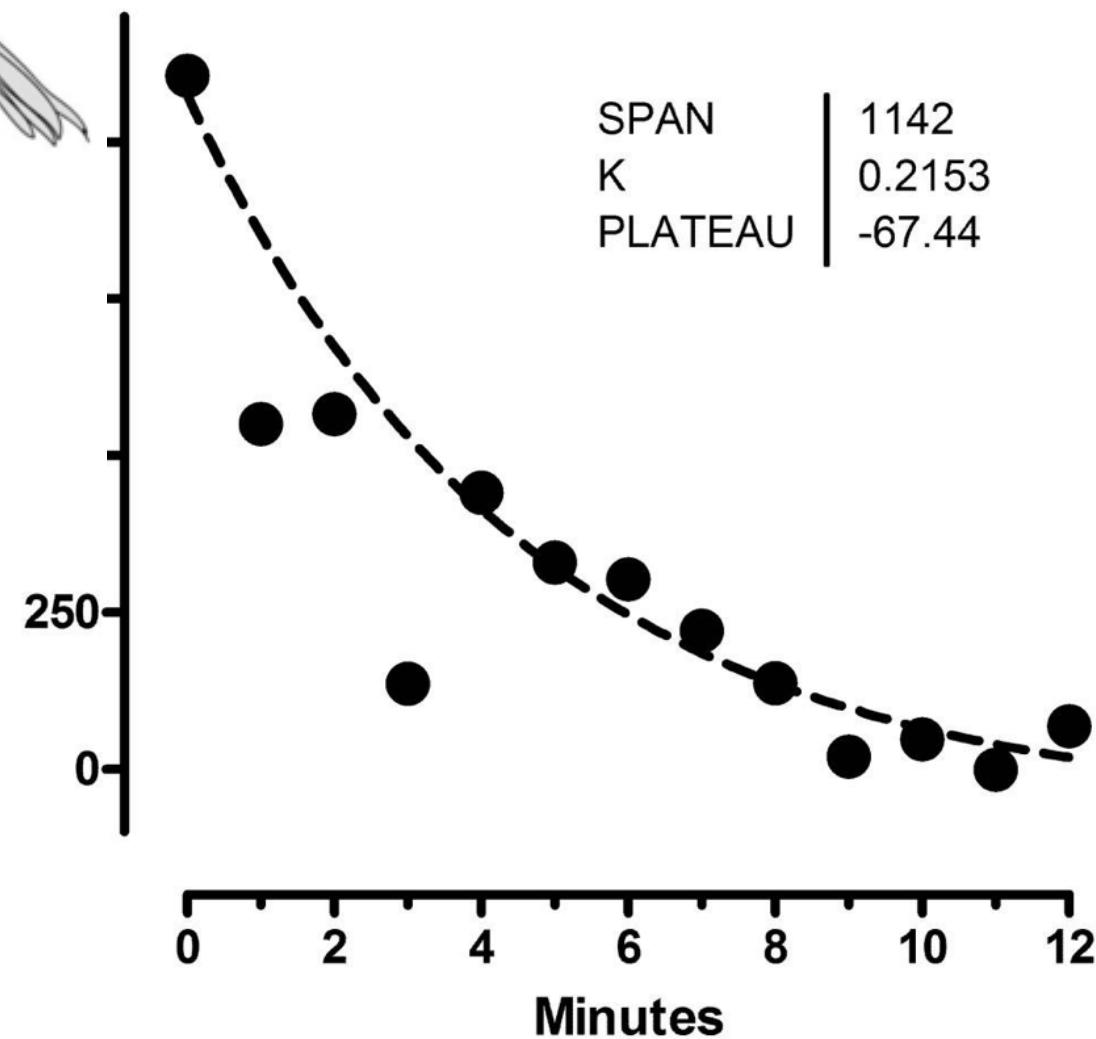
Regression



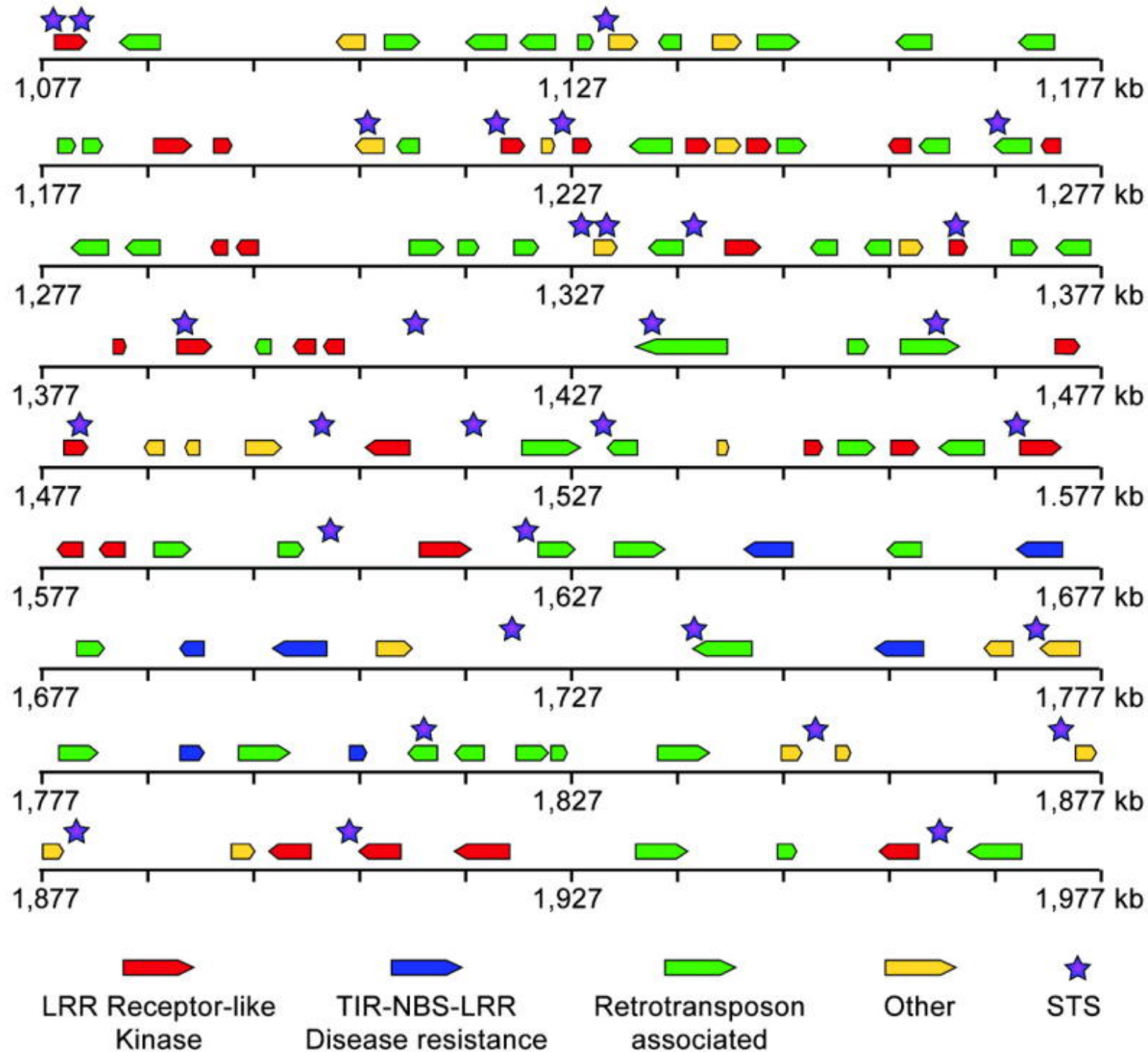
Copyright © 2004 Pearson Prentice Hall, Inc.

linear

nonlinear



Sequence Annotation



given sequence

gene finding
speech recognition
activity segmentation
named entities

Ontology

dmoz open directory project

In partnership with **Aol Search.**

[about dmoz](#) | [dmoz blog](#) | [suggest URL](#) | [help](#) | [link](#) | [editor login](#)

webpages

Search [advanced](#)

Arts

[Movies](#), [Television](#), [Music](#)...

Games

[Video Games](#), [RPGs](#), [Gambling](#)...

Kids and Teens

[Arts](#), [School Time](#), [Teen Life](#)...

Reference

[Maps](#), [Education](#), [Libraries](#)...

Shopping

[Clothing](#), [Food](#), [Gifts](#)...

World

[Català](#), [Dansk](#), [Deutsch](#), [Español](#), [Français](#), [Italiano](#), [日本語](#), [Nederlands](#), [Polski](#), [Русский](#), [Svenska](#)...

Business

[Jobs](#), [Real Estate](#), [Investing](#)...

Health

[Fitness](#), [Medicine](#), [Alternative](#)...

News

[Media](#), [Newspapers](#), [Weather](#)...

Regional

[US](#), [Canada](#), [UK](#), [Europe](#)...

Society

[People](#), [Religion](#), [Issues](#)...

Computers

[Internet](#), [Software](#), [Hardware](#)...

Home

[Family](#), [Consumers](#), [Cooking](#)...

Recreation

[Travel](#), [Food](#), [Outdoors](#), [Humor](#)...

Science

[Biology](#), [Psychology](#), [Physics](#)...

Sports

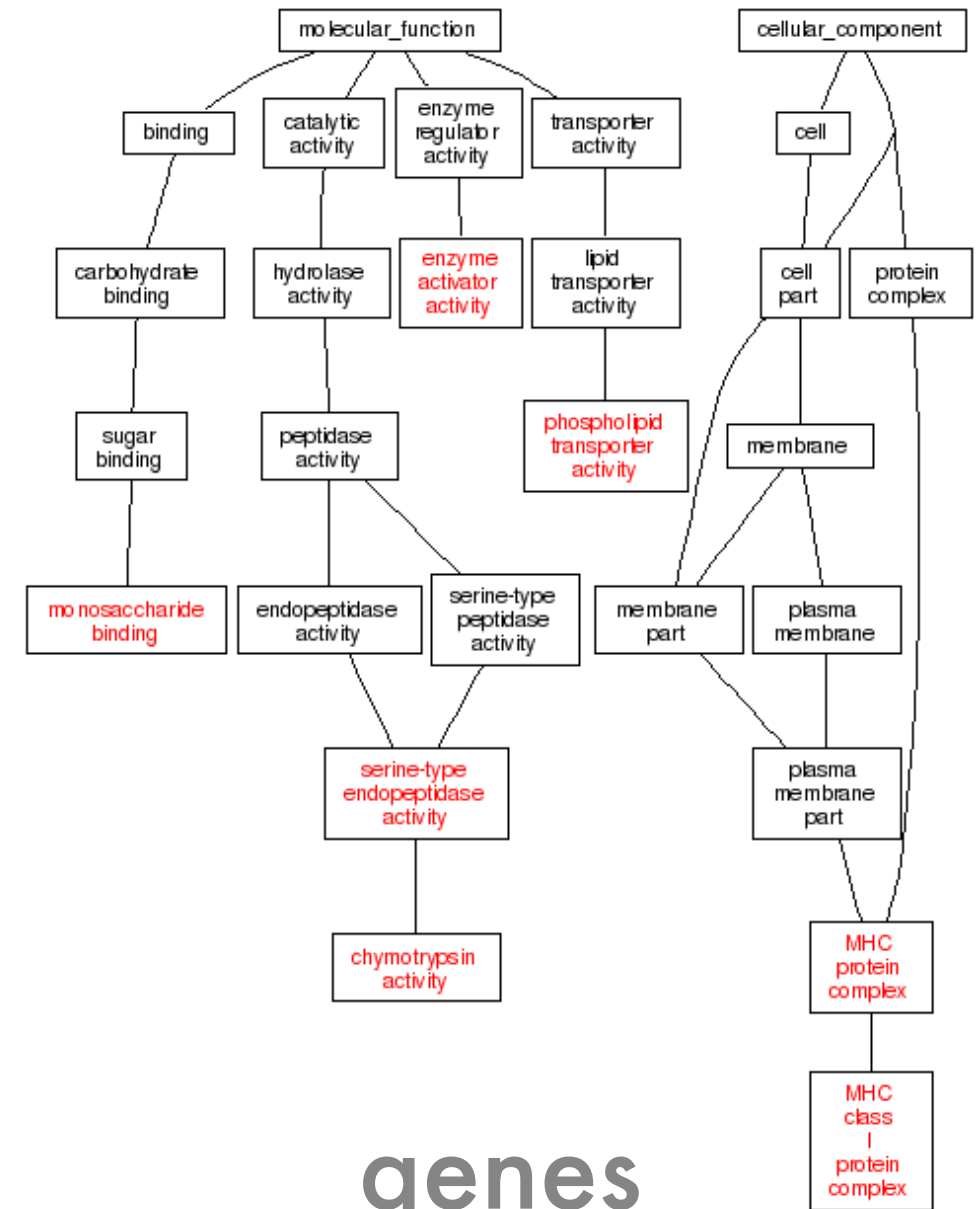
[Baseball](#), [Soccer](#), [Basketball](#)...

Become an Editor Help build the largest human-edited directory of the web



Copyright © 2013 Netscape

5,114,083 sites - 96,877 editors - over 1,014,849 categories



Prediction



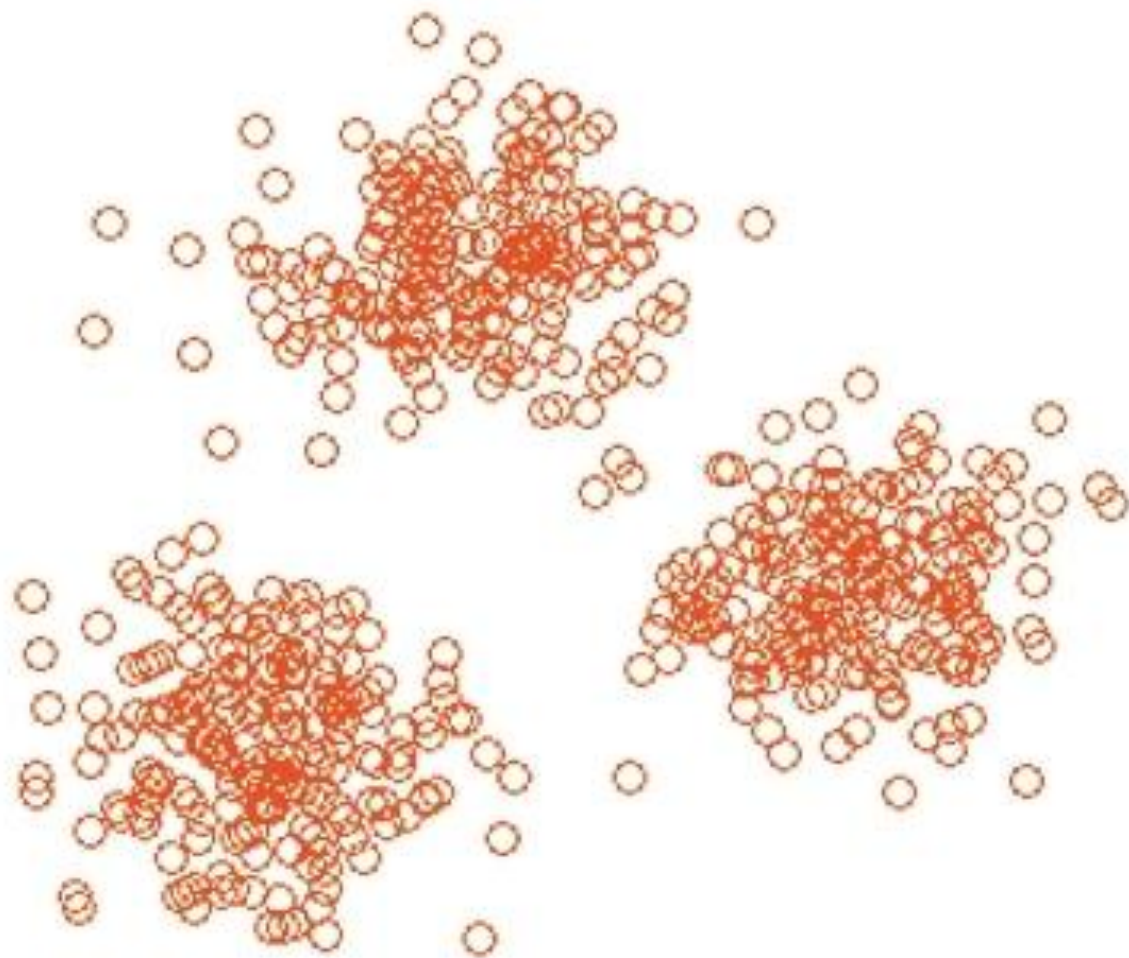
tomorrow's stock price

Carnegie Mellon University

Unsupervised Learning

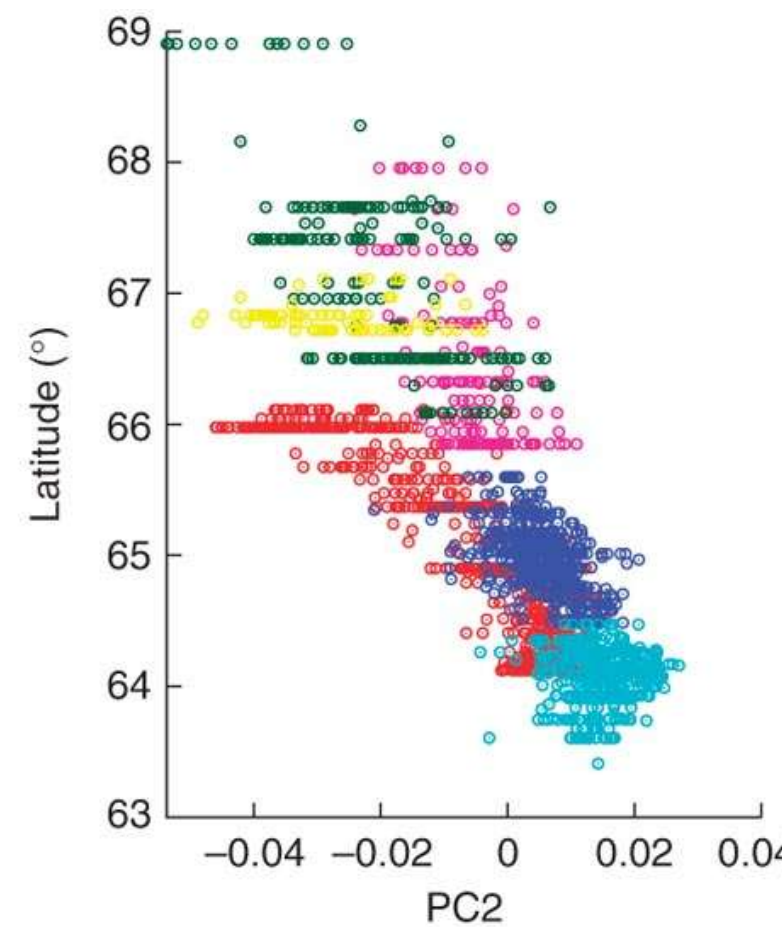
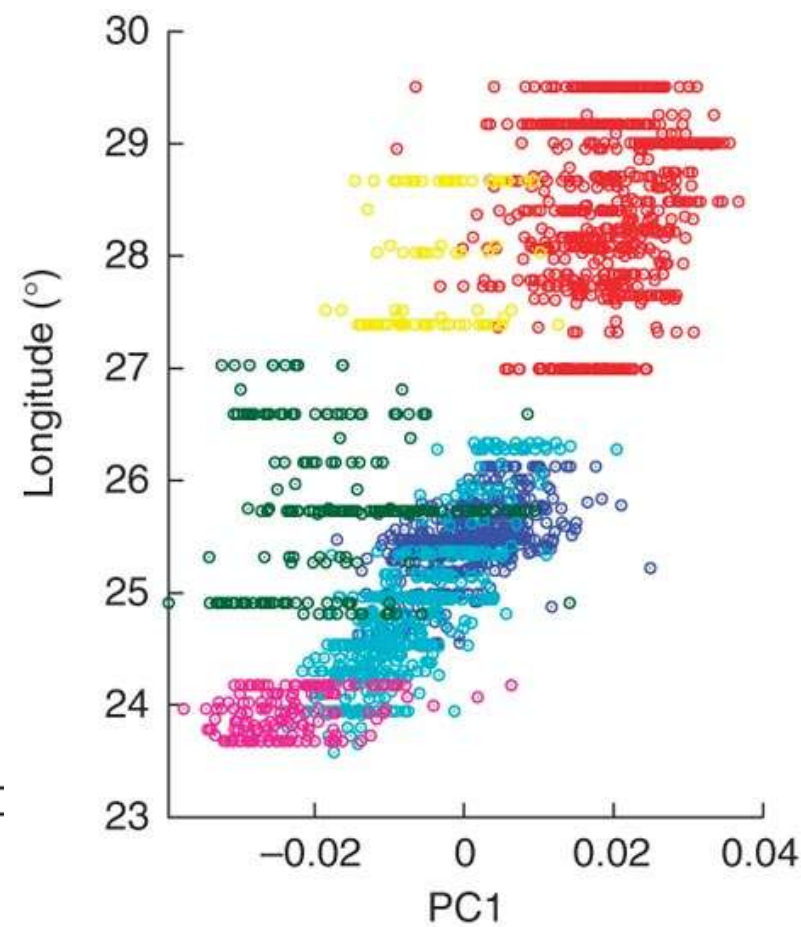
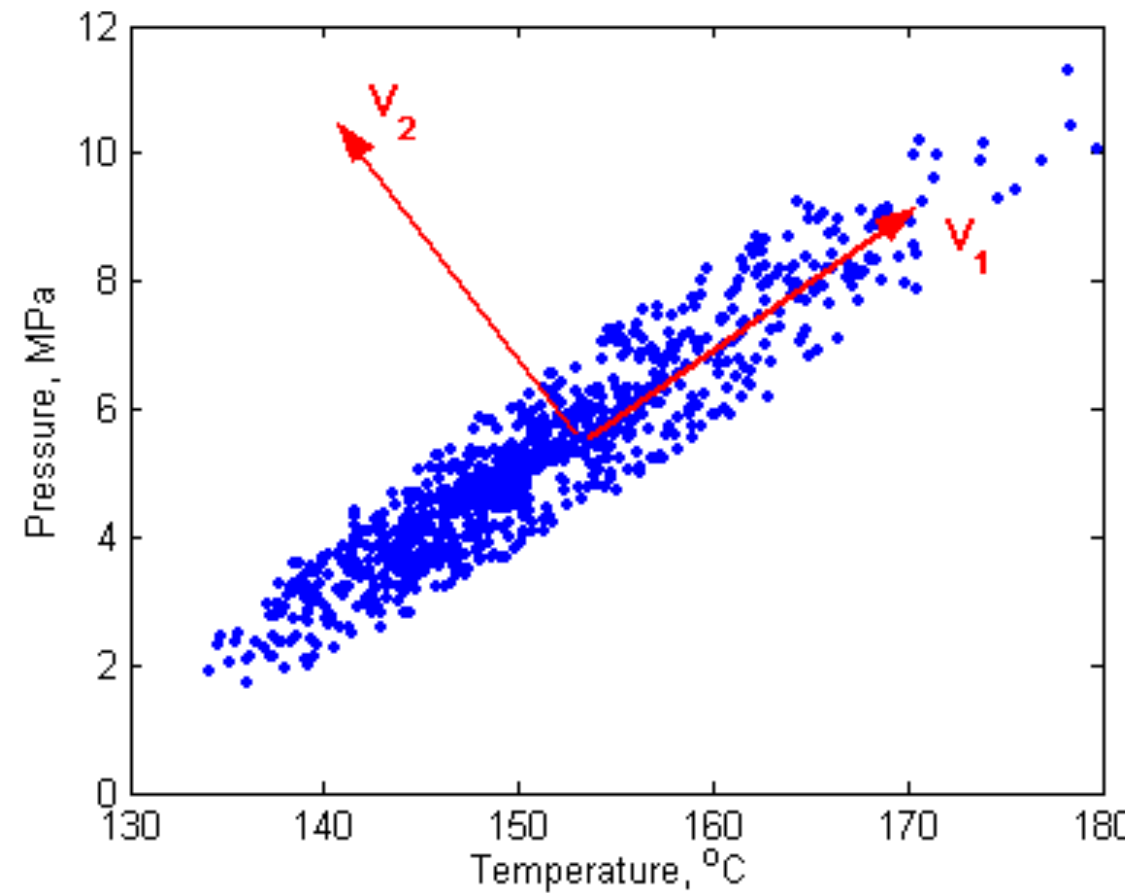
- Given data x , ask a good question ... about x or about model for x
- Clustering
Find a set of prototypes representing the data
- Principal Components
Find a subspace representing the data
- Sequence Analysis
Find a latent causal sequence for observations
 - Sequence Segmentation
 - Hidden Markov Model (discrete state)
 - Kalman Filter (continuous state)
- Hierarchical representations
- Independent components / dictionary learning
Find (small) set of factors for observation
- Novelty detection
Find the odd one out

Clustering



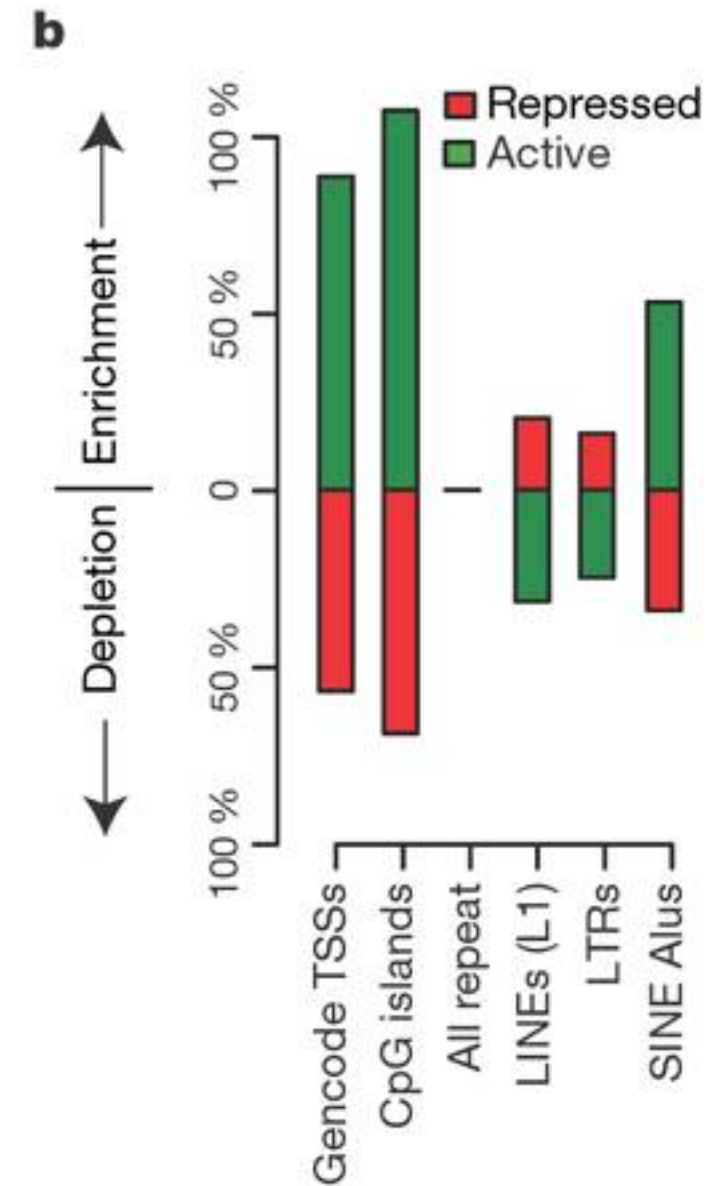
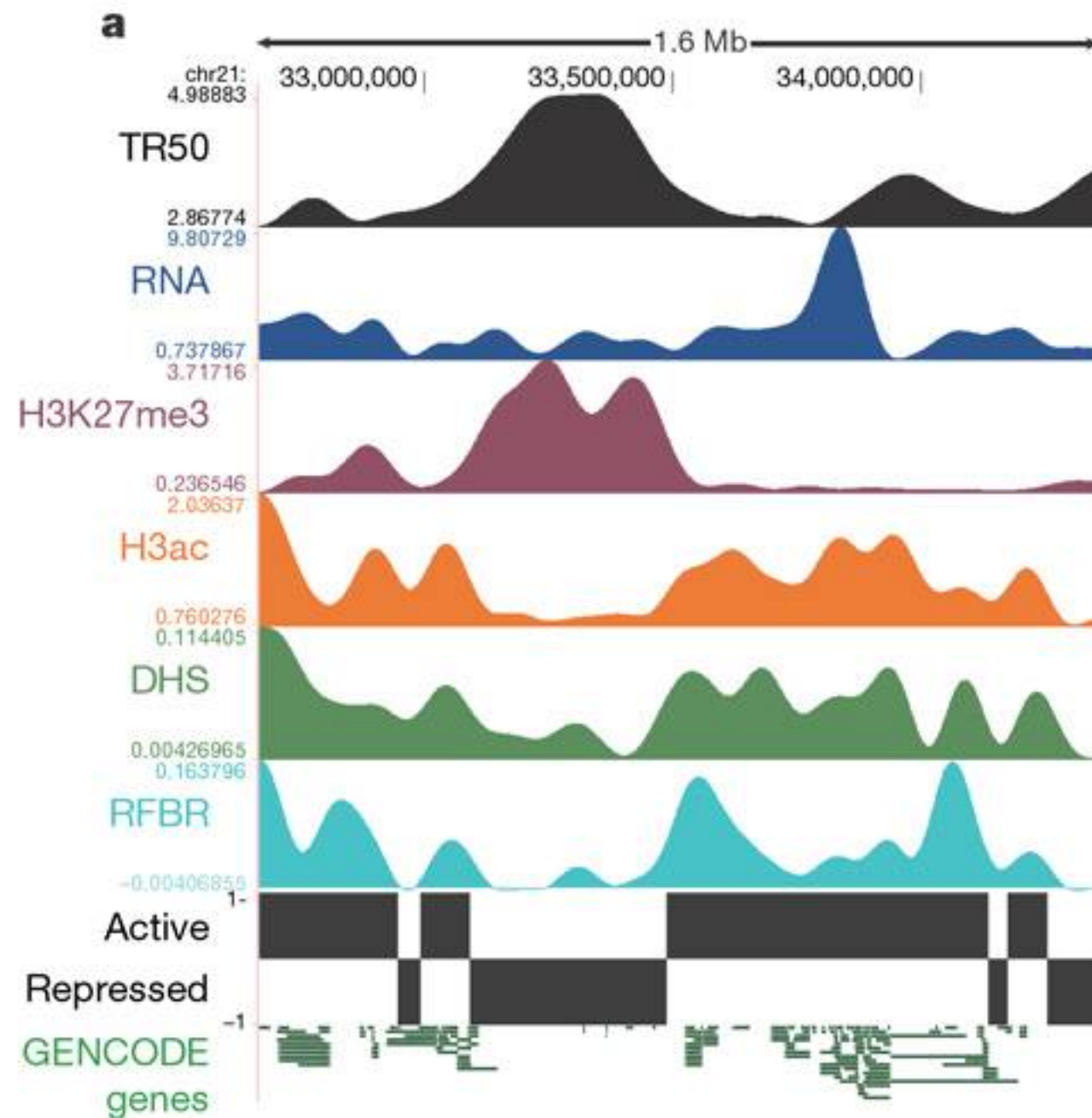
- Documents
- Users
- Webpages
- Diseases
- Pictures
- Vehicles
- ...

Principal Components



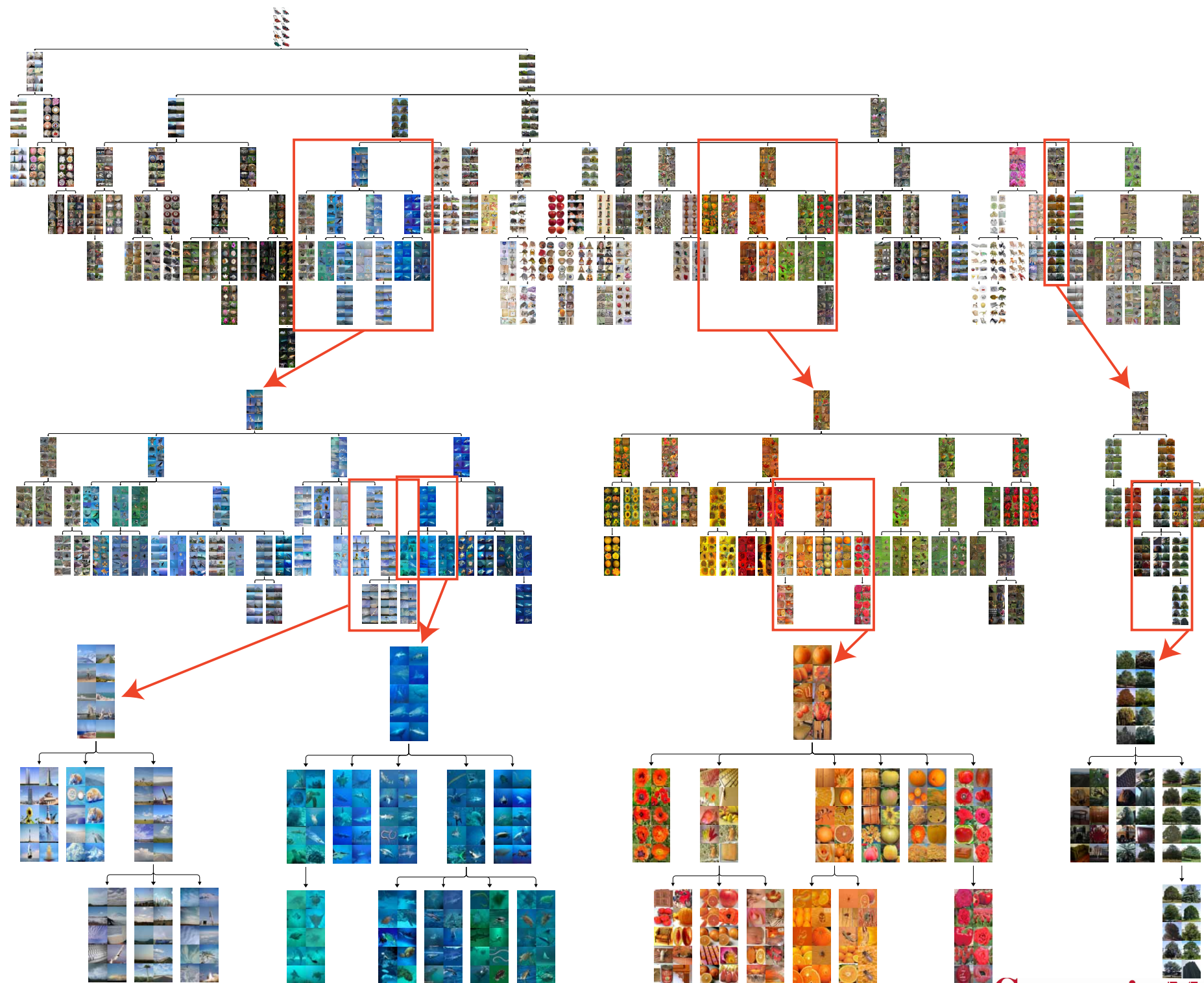
Variance component model to account for sample structure in genome-wide association studies, Nature Genetics 2010

Sequence Analysis

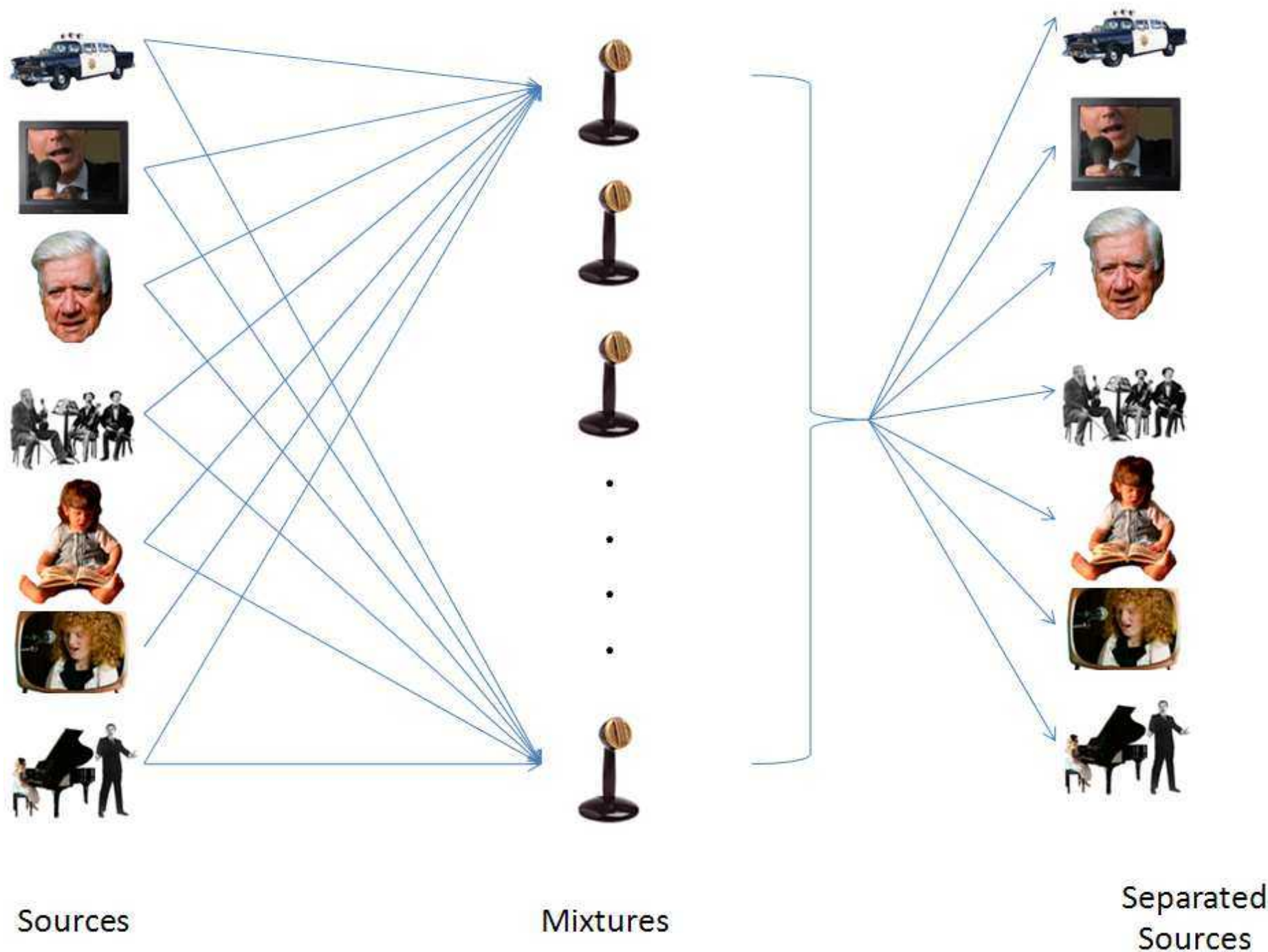


Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project, Nature 2007

Hierarchical Grouping

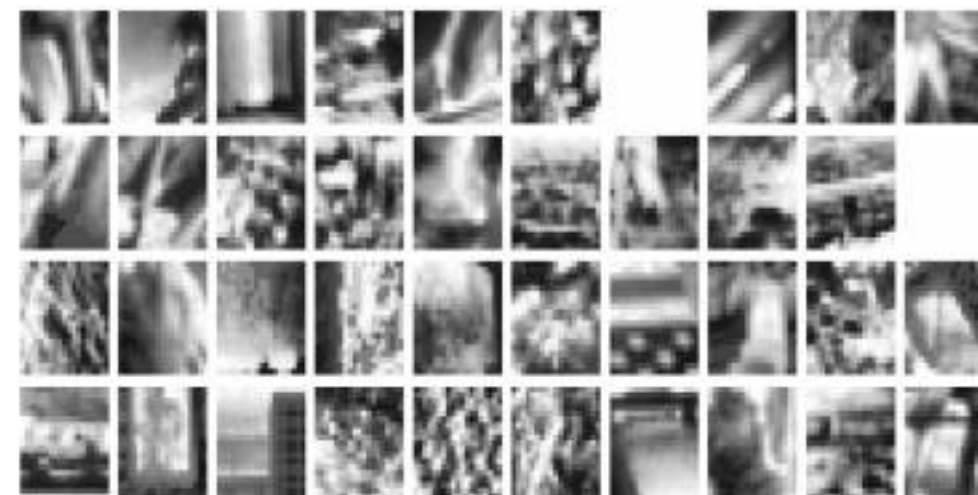
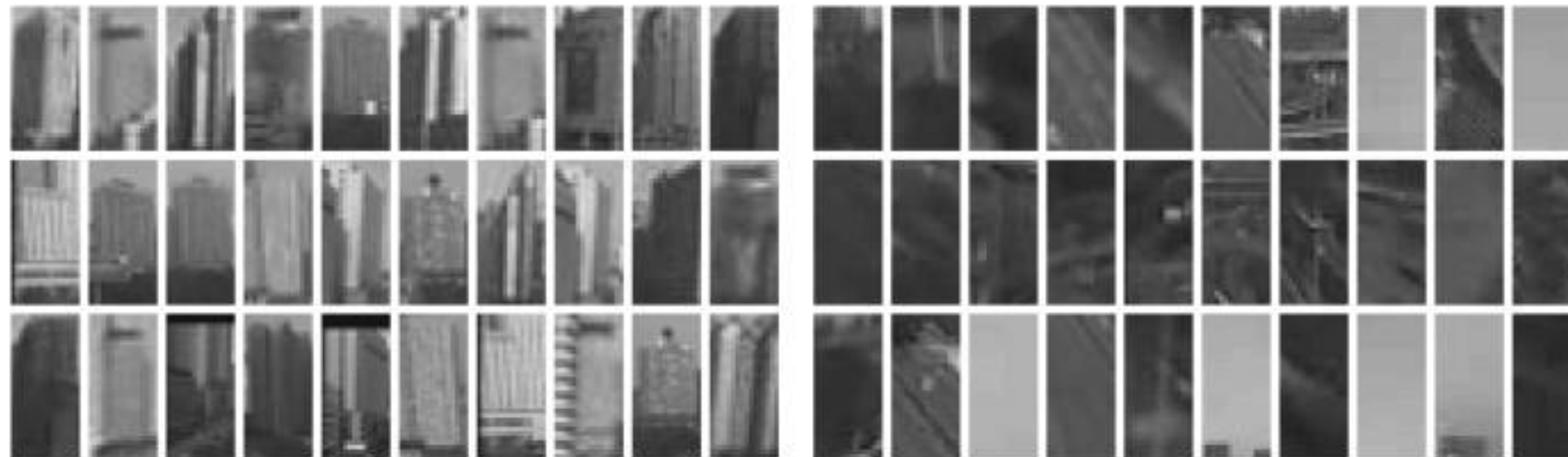


Independent Components



find them
automatically

Novelty detection



typical

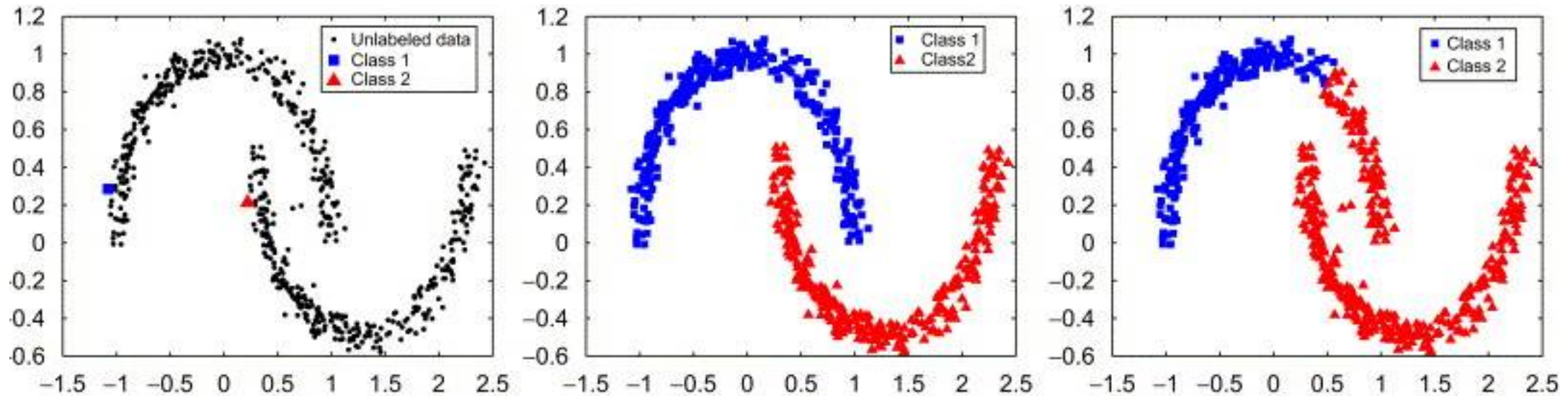
atypical

Some Problem types

iid = Independently Identically Distributed

- Induction
 - Training data (x,y) drawn iid
 - Test data x drawn iid from same distribution (not available at training time)
- Transduction
Test data x available at training time (you see the exam questions early)
- Semi-supervised learning
Lots of unlabeled data available at training time (past exam questions)
- Covariate shift
 - Training data (x,y) drawn iid from q (lecturer sets homework)
 - Test data x drawn iid from p (TAs set exams)
- Cotraining
Observe a number of similar problems at once

Induction - Transduction



- **Induction**
We only have training set. Do the best with it.
- **Transduction**
We have lots more problems that need to be solved with the same method.

Covariate Shift

- Problem (true story)
 - Biotech startup wants to detect prostate cancer.
 - Easy to get blood samples from sick patients.
 - Hard to get blood samples from healthy ones.
- Solution?
 - Get blood samples from male university students.
 - Use them as healthy reference.
 - Classifier gets 100% accuracy
- **What's wrong?**

Cotraining and Multitask

- **Multitask Learning**
Use correlation between tasks for better result
 - Task 1 - Detect spammy webpages
 - Task 2 - Detect people's homepages
 - Task 3 - Detect adult content
- **Cotraining**
For many cases both sets of covariates are available
 - Detect spammy webpages based on page content
 - Detect spammy webpages based on user viewing behavior

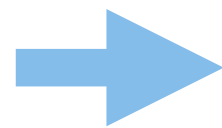
Interaction with Environment

- Batch (download a book)
Observe training data $(x_1, y_1) \dots (x_l, y_l)$ then deploy
- Online (follow the class)
Observe x , predict $f(x)$, observe y (stock market, homework)
- Active learning (ask questions in class)
Query y for x , improve model, pick new x
- Bandits (do well at homework)
Pick arm, get reward, pick new arm (also with context)
- Reinforcement Learning (play chess, drive a car)
Take action, environment responds, take new action

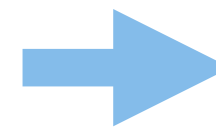
Batch

training
data

6	5	5	4	1	0
7	4	0	8	4	3
3	4	2	8	1	0
0	0	\	6	5	5
1	1	1	6	7	1
8	6	4	5	3	8
1	7	2	8	4	7
5	2	8	0	4	8
3	3	7	0	5	3
4	8	9	4	0	4



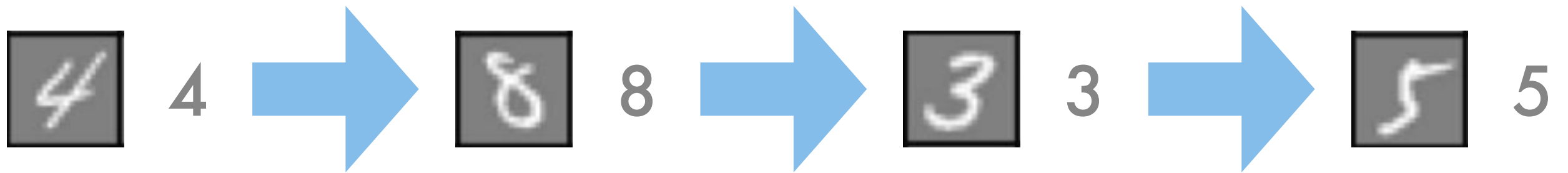
build
model



test

4	9	1	7
6	4	5	6
7	5	9	7
1	1	5	7
4	1	3	\
7	2	9	1
6	8	9	3
3	7	+	6
1	1	0	3
5	0	5	0

Online



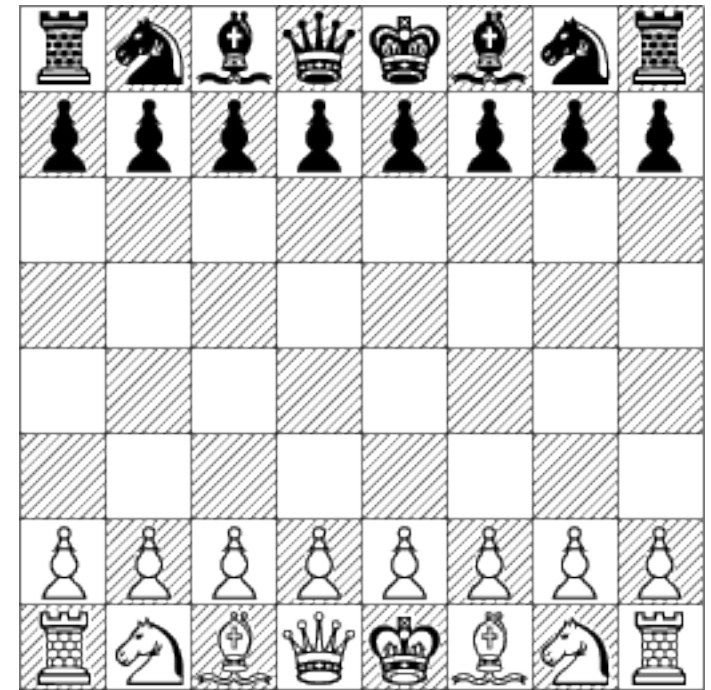
Bandits

- Choose an option
- See what happens (get reward)
- Update model
- Choose next option



Reinforcement Learning

- Take action
- Environment reacts
- Observe stuff
- Update model
- Repeat

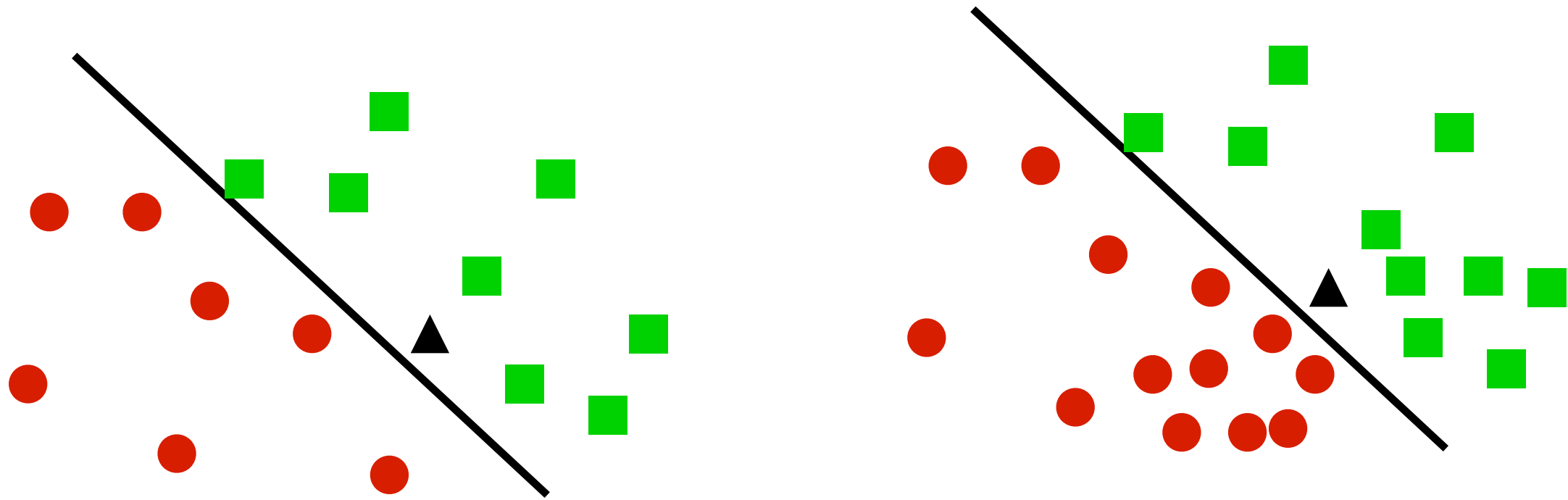


environment (cooperative, adversary, doesn't care)
memory (goldfish, elephant)
state space (tic tac toe, chess, car)

Discriminative vs. Generative (mainly relevant for supervised models)

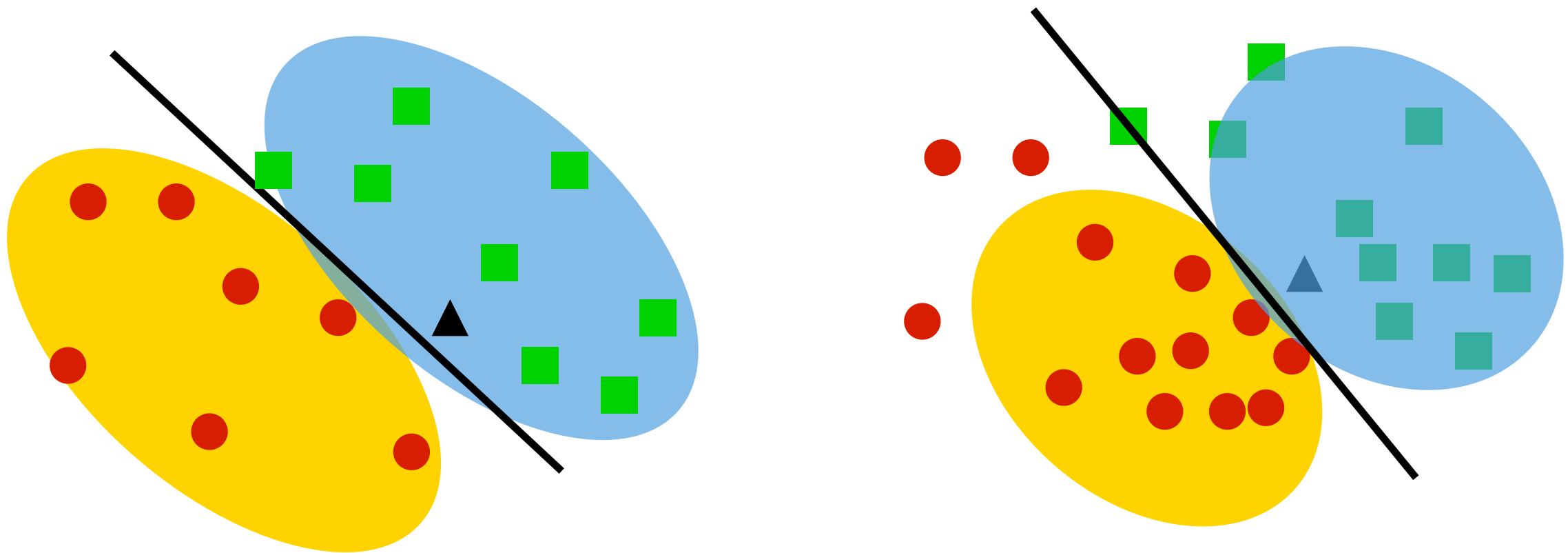
- **Discriminative Models**
 - Estimate $y | x$ directly
 - Often better convergence + simpler solutions
- **Generative models**
 - Estimate joint distribution over (x,y)
 - Use conditional probability to infer $y | x$
 - Often more intuitive
 - Easier to add prior knowledge

Discriminative



- Only care about estimating the conditional probabilities
- Very good when underlying distribution of data is really complicated (e.g. texts, images, movies)

Generative



- Model observations (x,y) first
- Then infer $p(y|x)$
- Good for missing variables, better diagnostics
- Easy to add prior knowledge about data



MAGIC Etch A Sketch[®] SCREEN

Very
Basic Tools

Horizontal
Lid

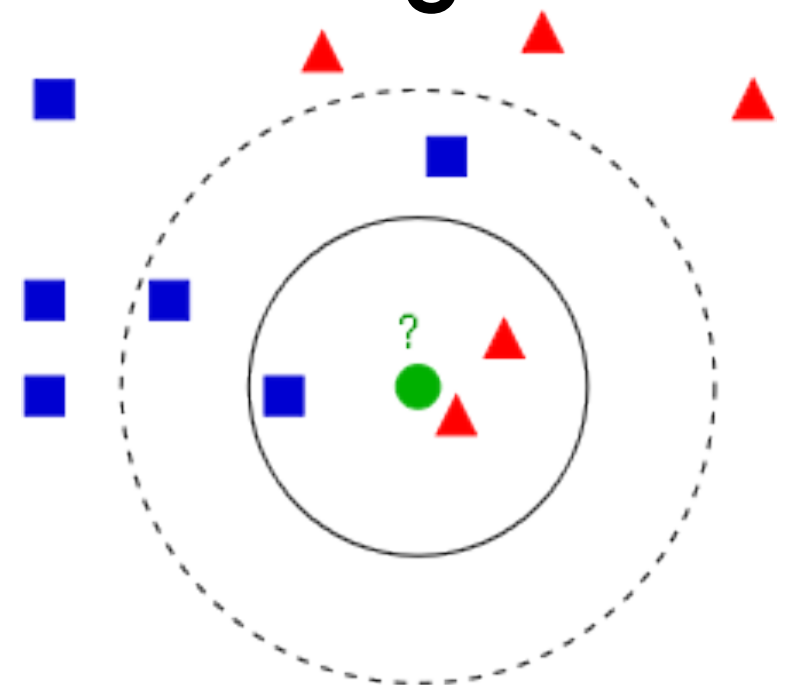
OHIO ART The World of Toys[®]

Vertical
Lid

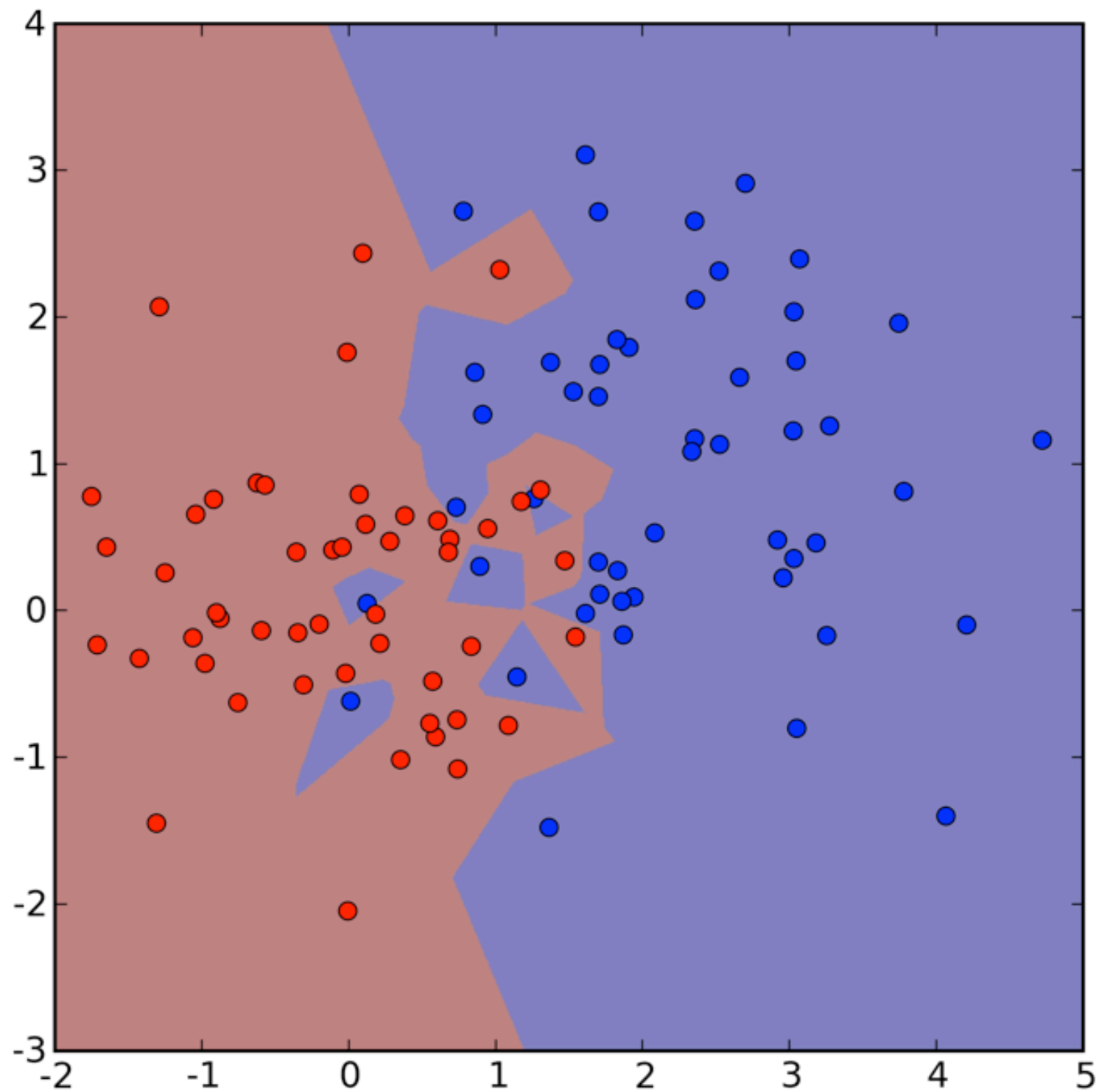
MAGIC SCREEN IS GLASS SET IN CURVED PLASTIC FRAME.
USE WITH CARE.

Nearest Neighbors

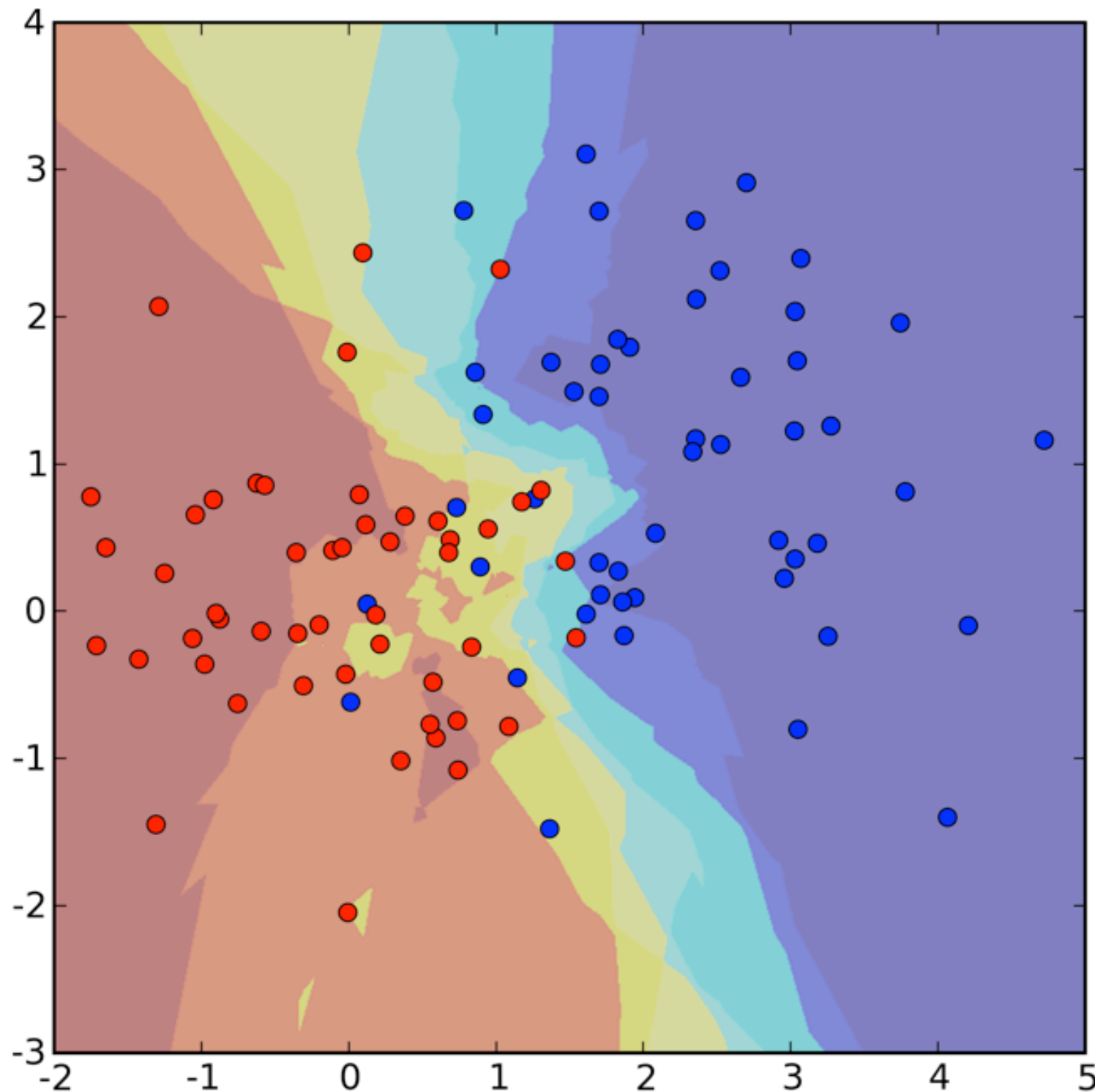
- Table lookup
For previously seen instance remember label
- Nearest neighbor
 - Pick label of most similar neighbor
 - Slight improvement - use k-nearest neighbors
 - For regression average
 - Really useful baseline!
 - Easy to implement for small amounts of data. Why?



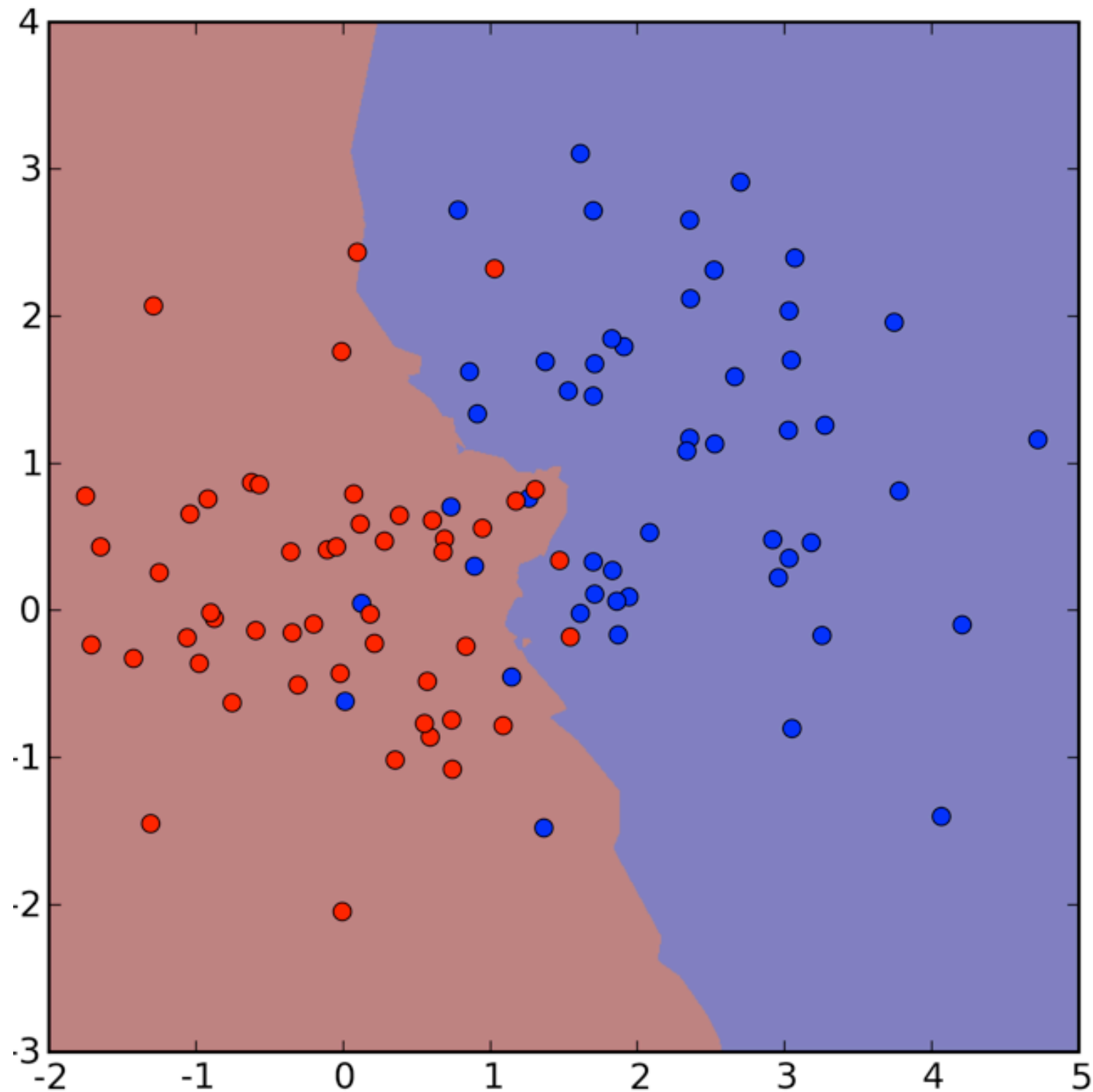
1-Nearest Neighbor



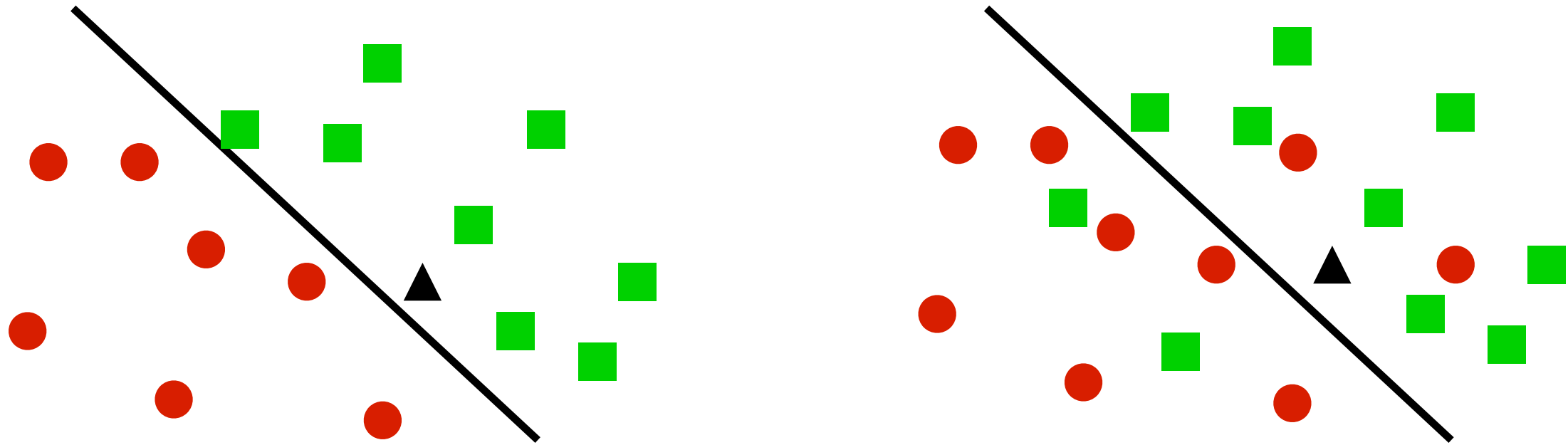
4-Nearest Neighbors



4-Nearest Neighbors Sign

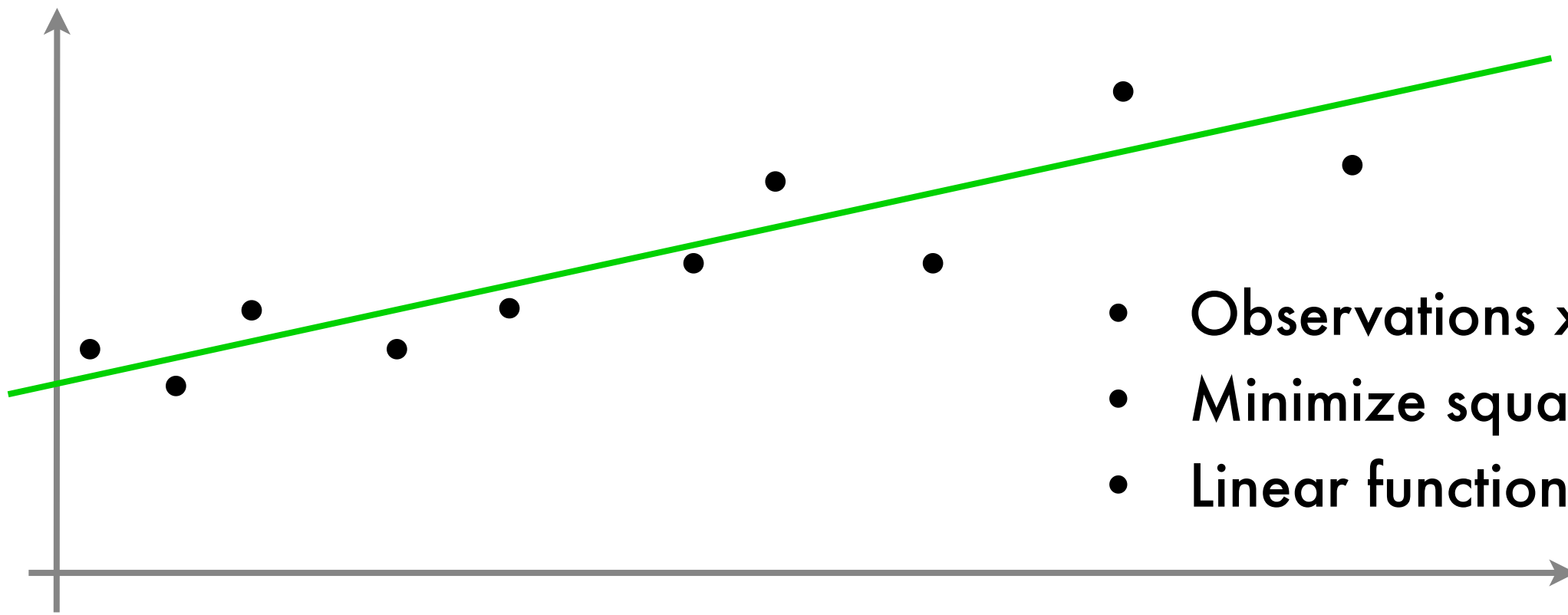


If we get more data



- 1 Nearest Neighbor
 - Converges to perfect solution if clear separation
 - Twice the minimal error rate $2p(1-p)$ for noisy problems
- k-Nearest Neighbor
 - Converges to perfect solution if clear separation (**but needs more data**)
 - Converges to minimal error $\min(p, 1-p)$ for noisy problems if k increases

Linear Regression



- Observations x , labels y
- Minimize squared distance
- Linear function

$$f(x) = ax + b$$

$$\underset{a,b}{\text{minimize}} \sum_{i=1}^m \frac{1}{2} (ax_i + b - y_i)^2$$

$$\partial_a [\dots] = 0 = \sum_{i=1}^m x_i (ax_i + b - y_i)$$

$$\partial_b [\dots] = 0 = \sum_{i=1}^m (ax_i + b - y_i)$$

Linear Regression

- Optimization Problem

$$f(x) = \langle a, x \rangle + b = \langle w, (x, 1) \rangle$$

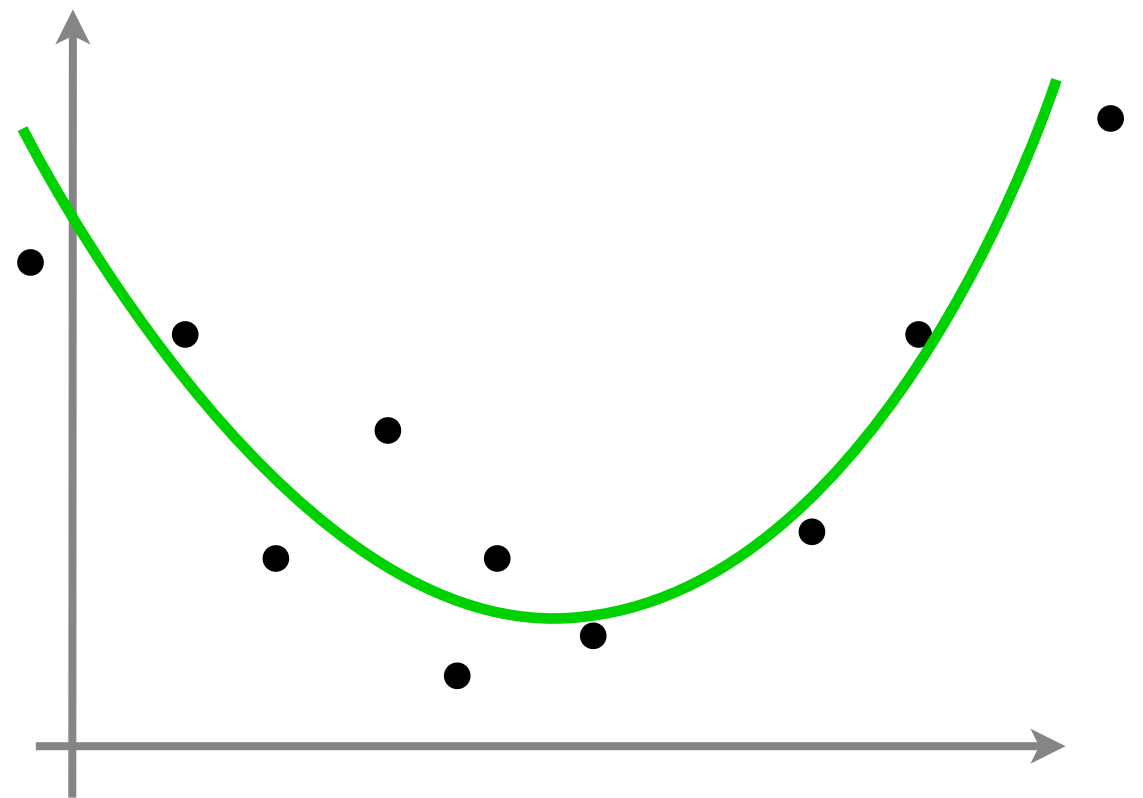
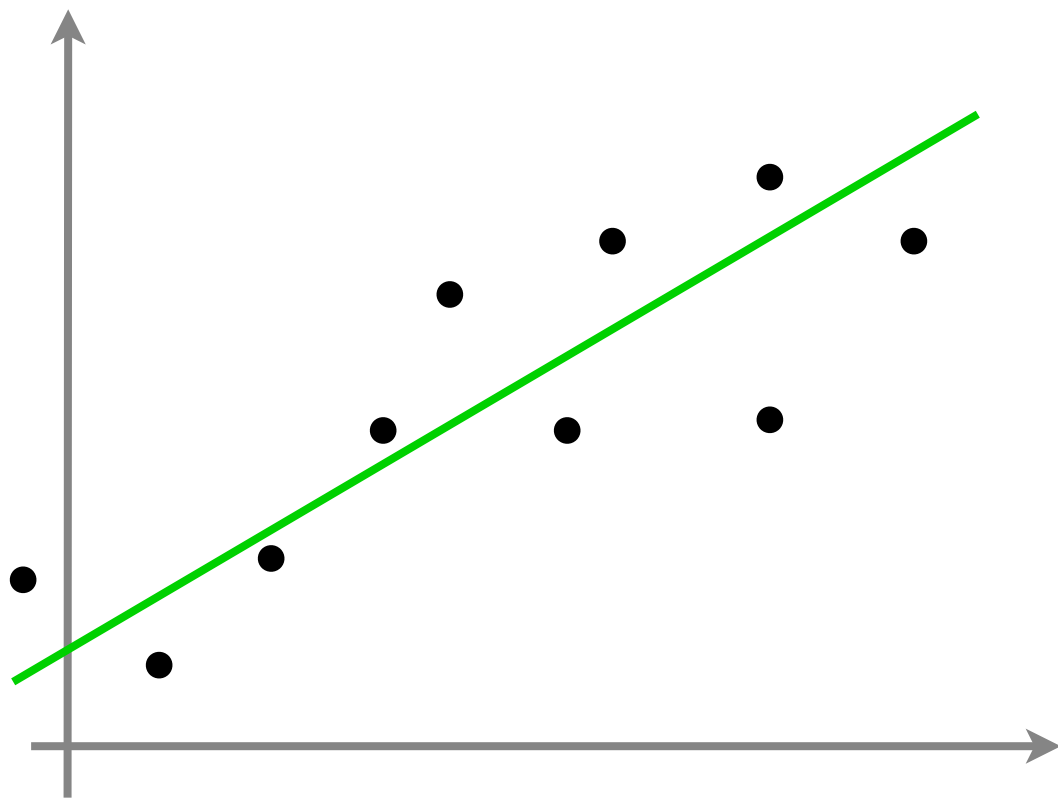
$$\underset{w}{\text{minimize}} \sum_{i=1}^m \frac{1}{2} (\langle w, \bar{x}_i \rangle - y_i)^2$$

- Solving it

$$0 = \sum_{i=1}^m \bar{x}_i (\langle w, \bar{x}_i \rangle - y_i) \iff \left[\sum_{i=1}^m \bar{x}_i \bar{x}_i^\top \right] w = \sum_{i=1}^m y_i \bar{x}_i$$

only requires a matrix inversion.

Nonlinear Regression



- **Linear model** $f(x) = \langle w, (1, x) \rangle$
- **Quadratic model** $f(x) = \langle w, (1, x, x^2) \rangle$
- **Cubic model** $f(x) = \langle w, (1, x, x^2, x^3) \rangle$
- **Nonlinear model** $f(x) = \langle w, \phi(x) \rangle$

Linear Regression

- Optimization Problem

$$f(x) = \langle a, x \rangle + b = \langle w, (x, 1) \rangle$$

$$\underset{w}{\text{minimize}} \sum_{i=1}^m \frac{1}{2} (\langle w, \bar{x}_i \rangle - y_i)^2$$

- Solving it

$$0 = \sum_{i=1}^m \bar{x}_i (\langle w, \bar{x}_i \rangle - y_i) \iff \left[\sum_{i=1}^m \bar{x}_i \bar{x}_i^\top \right] w = \sum_{i=1}^m y_i \bar{x}_i$$

only requires a matrix inversion.

Nonlinear Regression

- Optimization Problem

$$f(x) = \langle w, \phi(x) \rangle$$

$$\underset{w}{\text{minimize}} \sum_{i=1}^m \frac{1}{2} (\langle w, \phi(x_i) \rangle - y_i)^2$$

- Solving it

$$\sum_{i=1}^m \phi(x_i) (\langle w, \phi(x_i) \rangle - y_i) \iff \left[\sum_{i=1}^m \phi(x_i) \phi(x_i)^\top \right] w = \sum_{i=1}^m y_i \phi(x_i)$$

only requires a matrix inversion.

Pseudocode (degree 4)

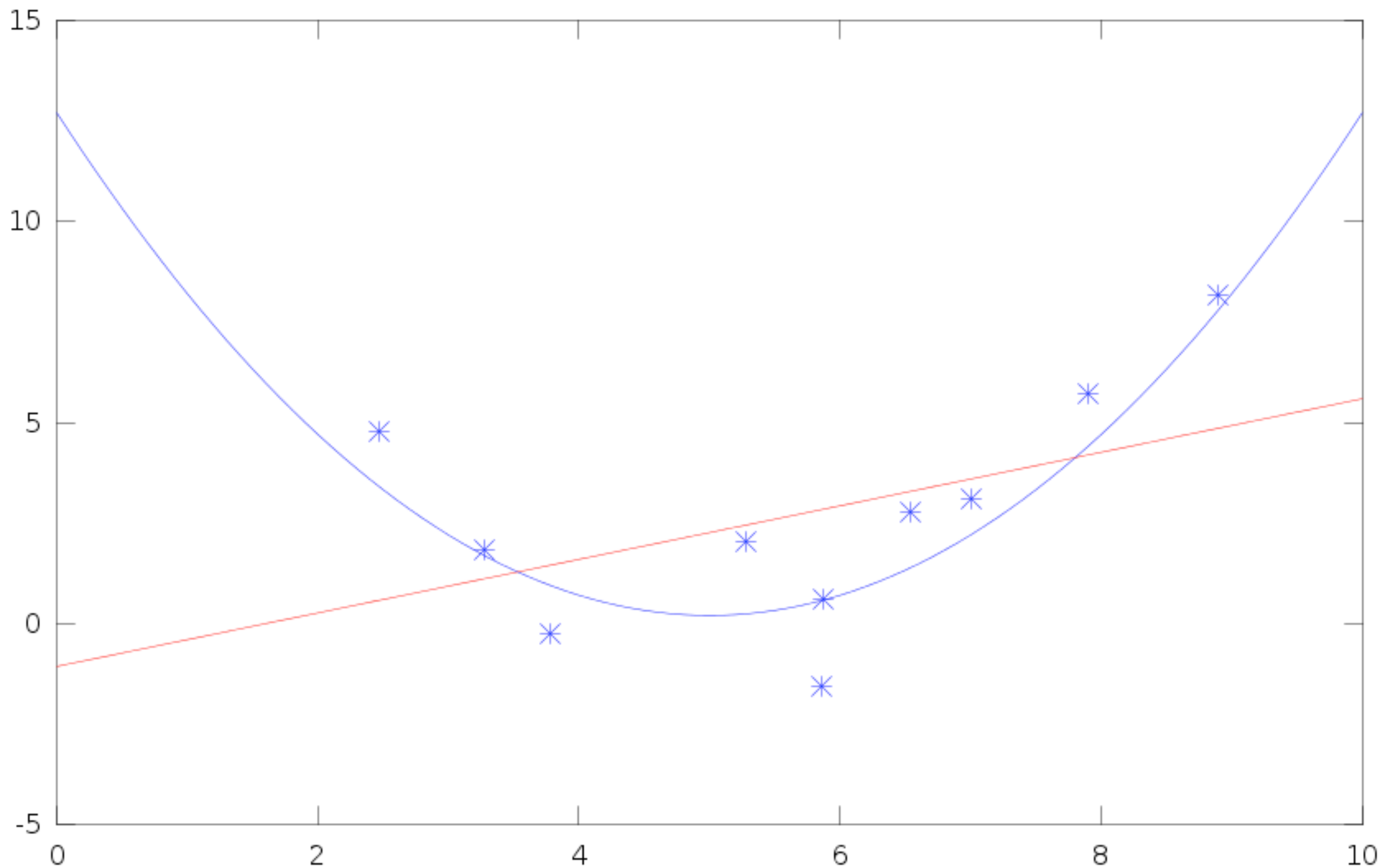
Training

```
phi_xx = [xx.^4, xx.^3, xx.^2, xx, 1.0 + 0.0 * xx];  
w = (yy' * phi_xx) / (phi_xx' * phi_xx);
```

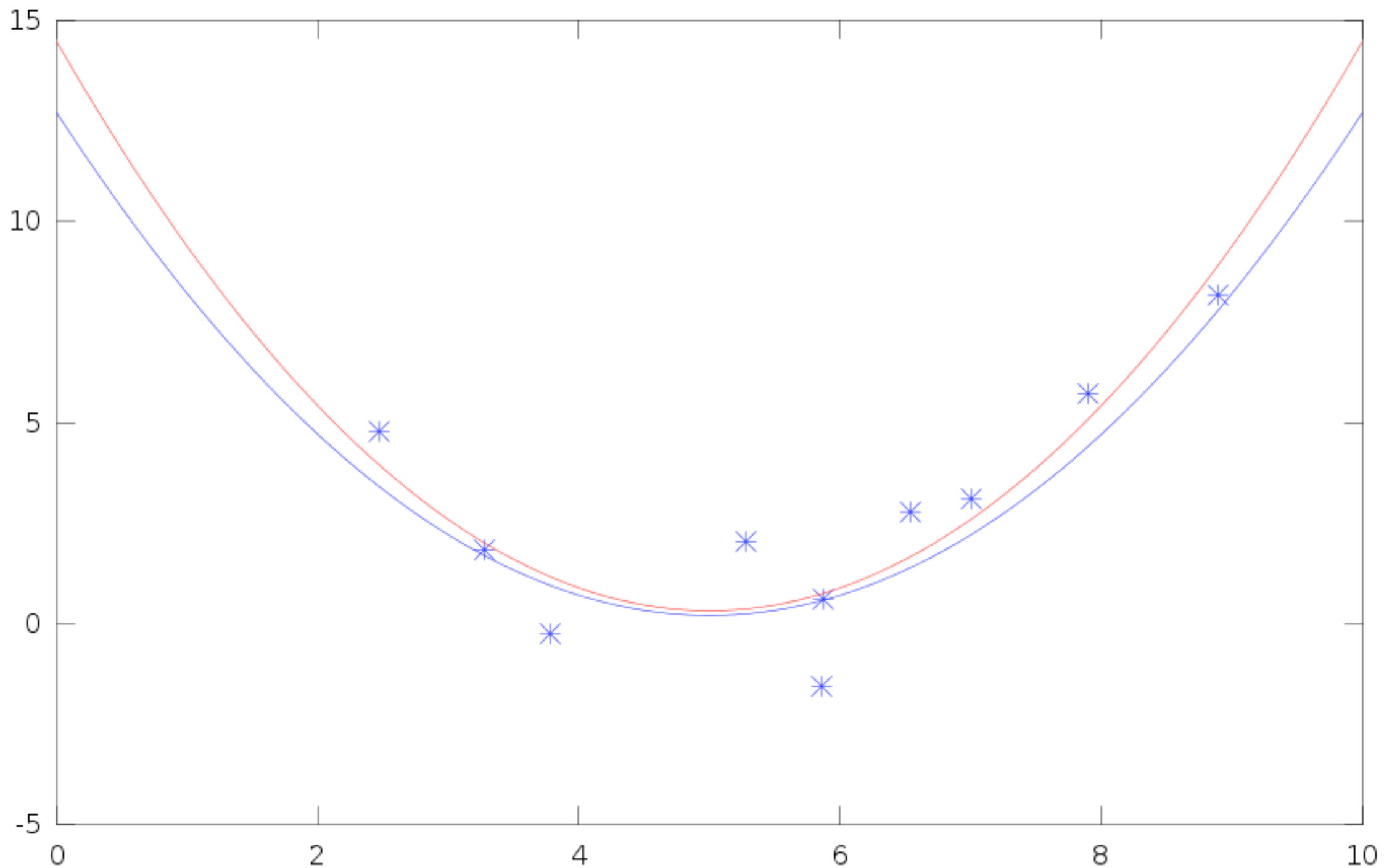
Testing

```
phi_x = [x.^4, x.^3, x.^2, x, 1.0 + 0.0 * x];  
y = phi_x * w';
```

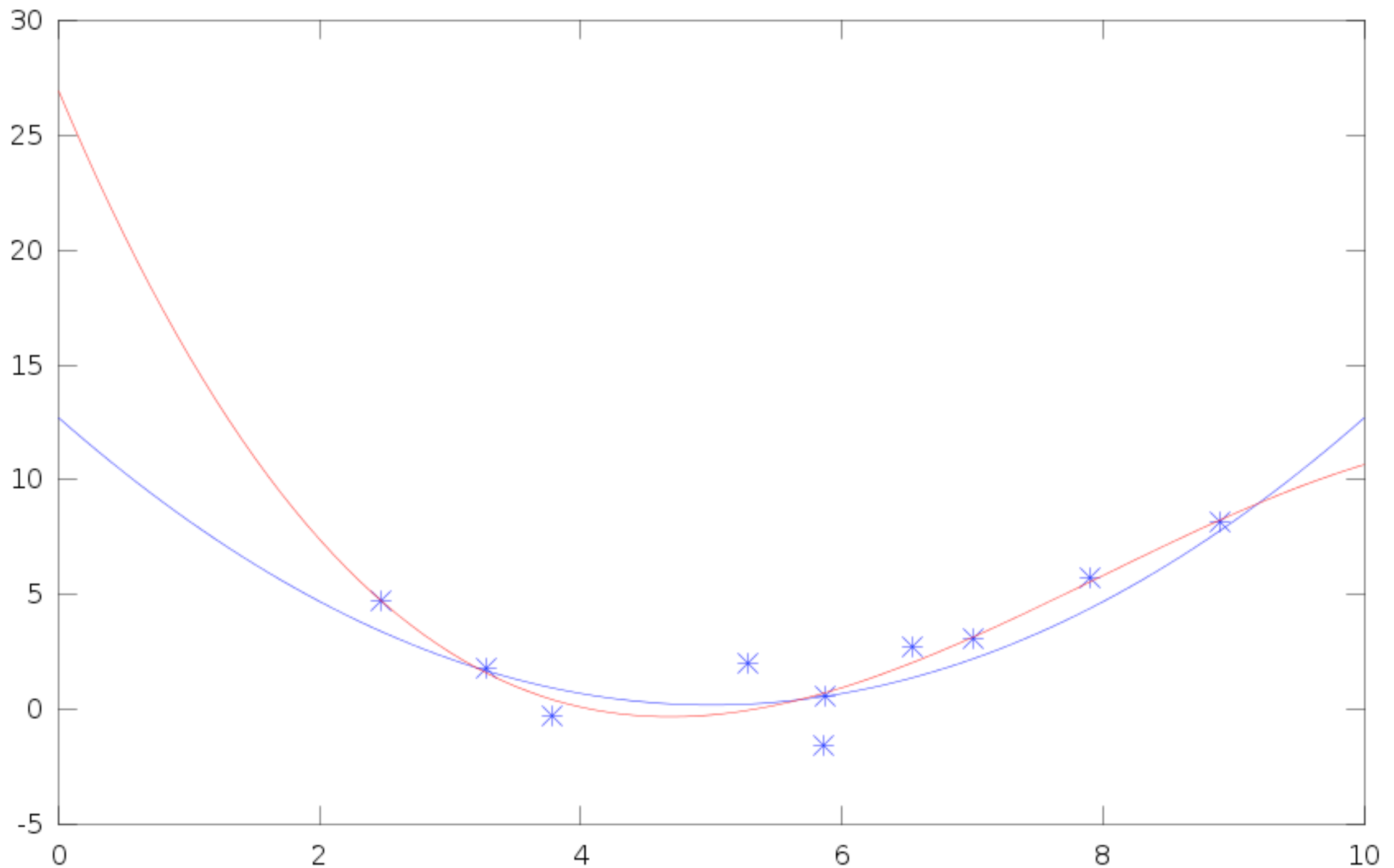
Regression (d=1)



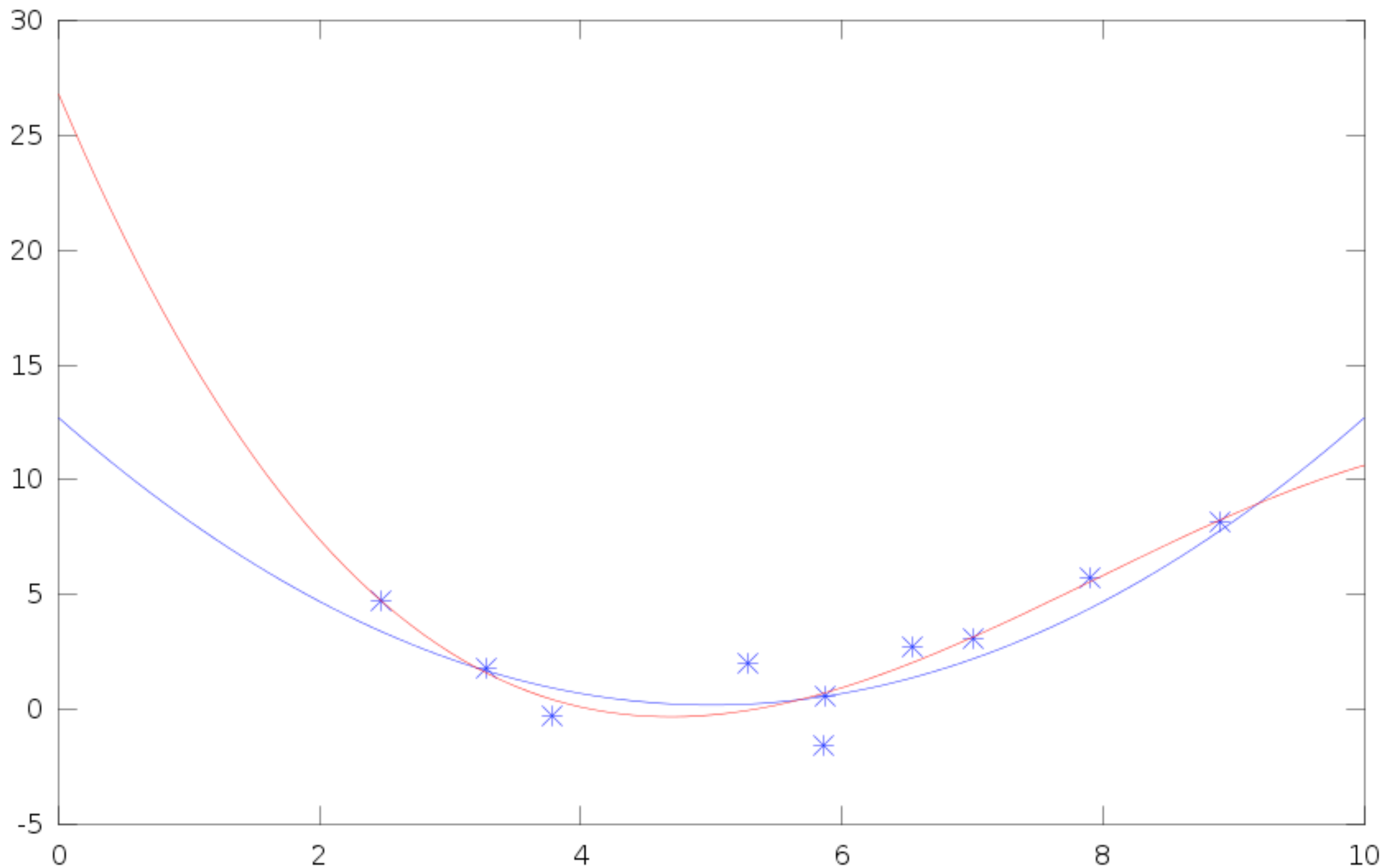
Regression (d=2)



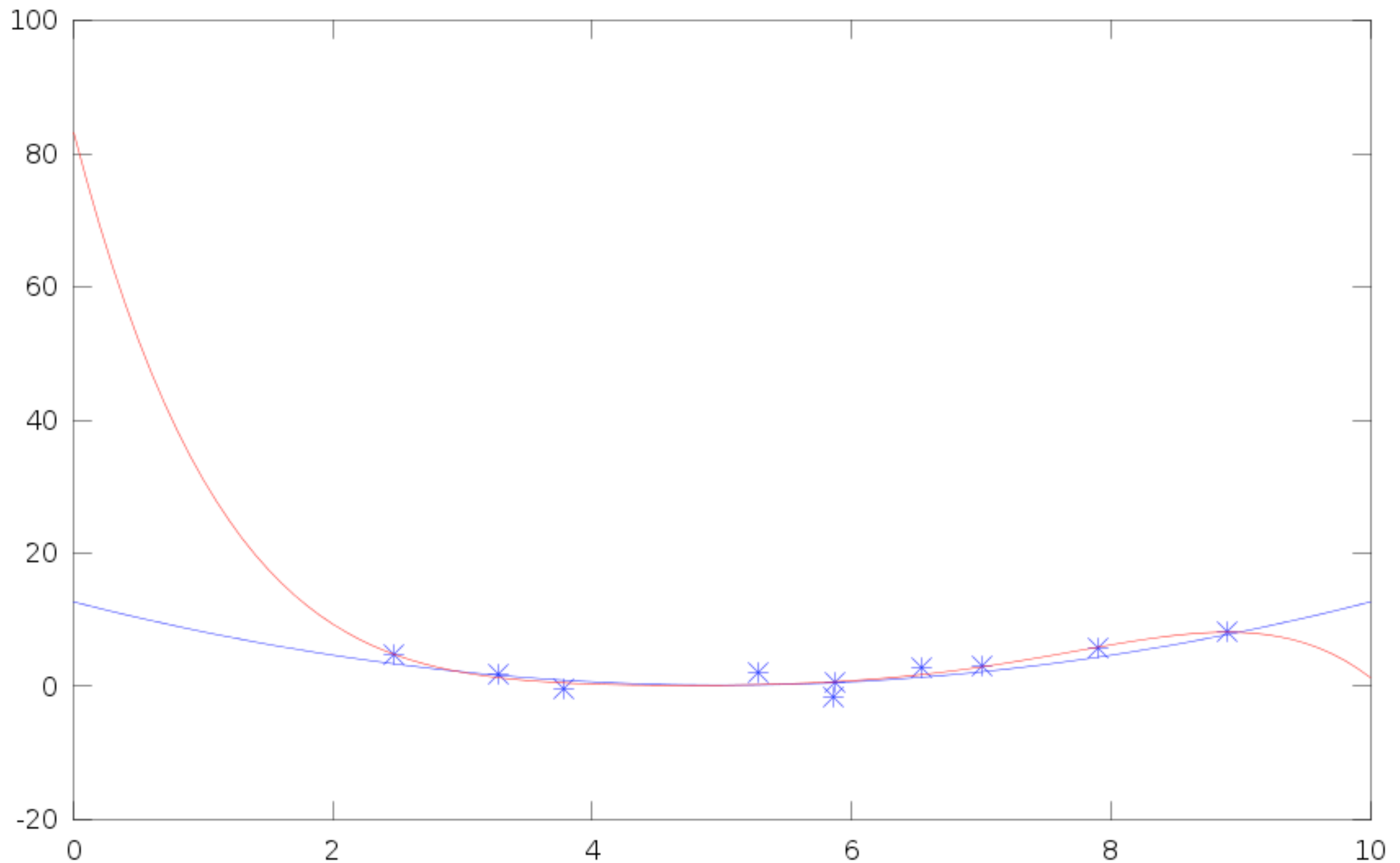
Regression (d=3)



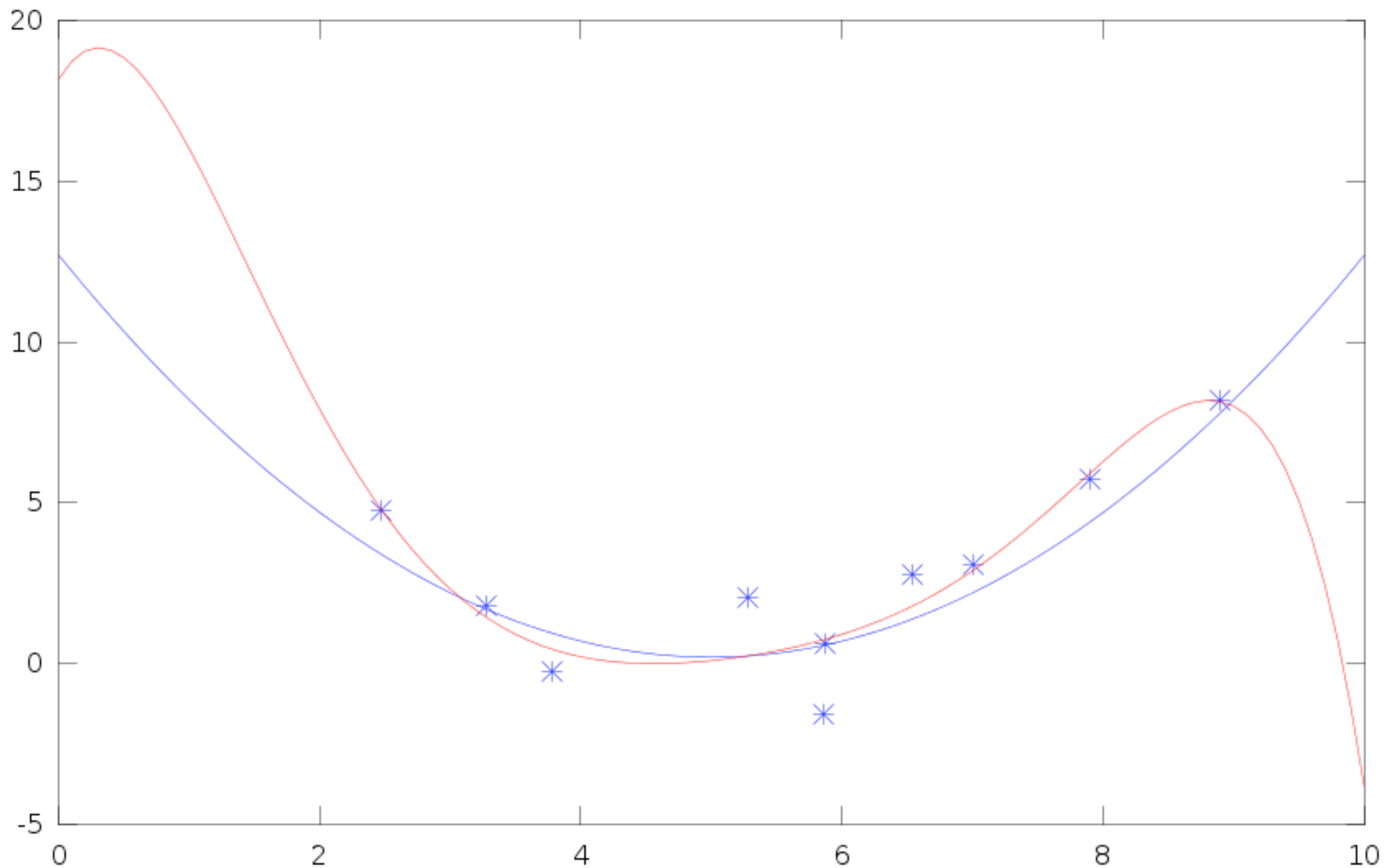
Regression (d=4)



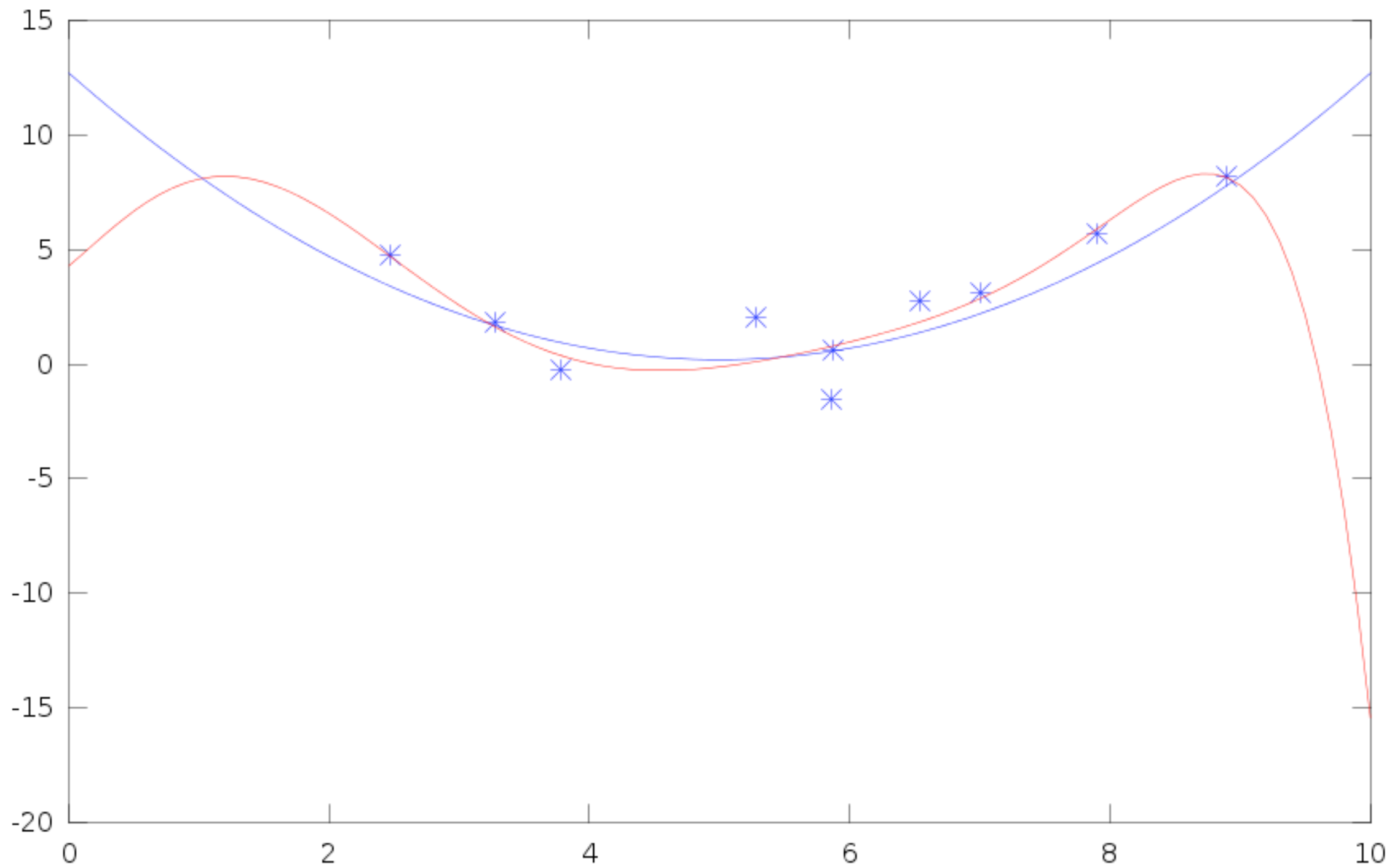
Regression (d=5)



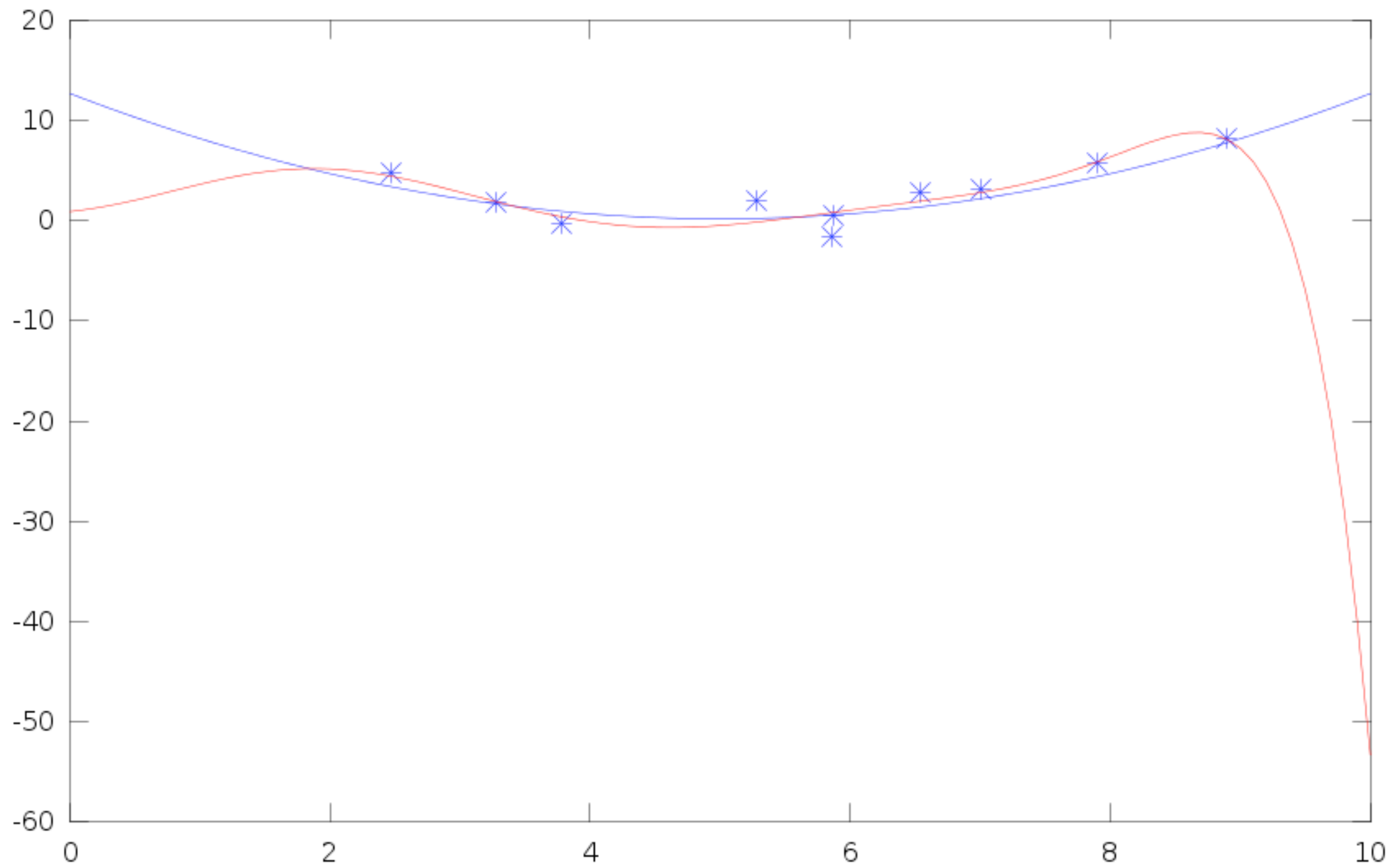
Regression (d=6)



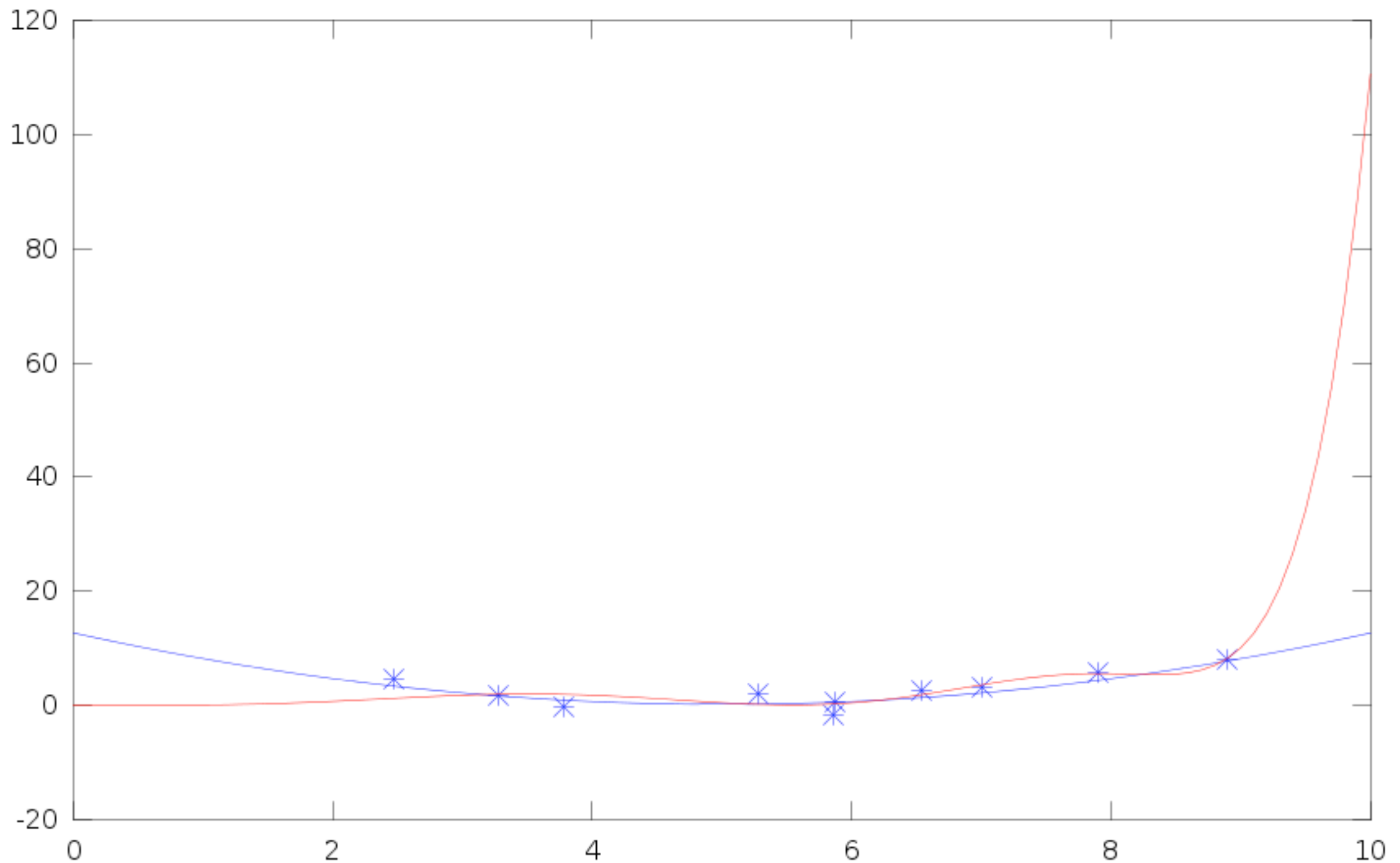
Regression (d=7)



Regression (d=8)



Regression (d=9)



Nonlinear Regression

```
warning: matrix singular to machine precision, rcond = 5.8676e-19
warning: attempting to find minimum norm solution
warning: matrix singular to machine precision, rcond = 5.86761e-19
warning: attempting to find minimum norm solution
warning: dgelsd: rank deficient 8x8 matrix, rank = 7
warning: matrix singular to machine precision, rcond = 1.10156e-21
warning: attempting to find minimum norm solution
warning: matrix singular to machine precision, rcond = 1.10145e-21
warning: attempting to find minimum norm solution
warning: dgelsd: rank deficient 9x9 matrix, rank = 6
warning: matrix singular to machine precision, rcond = 2.16217e-26
warning: attempting to find minimum norm solution
warning: matrix singular to machine precision, rcond = 1.66008e-26
warning: attempting to find minimum norm solution
warning: dgelsd: rank deficient 10x10 matrix, rank = 5
```

Nonlinear Regression

```
warning: matrix singular to machine precision, rcond = 5.8676e-19
warning: attempting to find minimum norm solution
warning: matrix singular to machine precision, rcond = 5.86761e-19
warning: attempting to find minimum norm solution
warning: dgelsd: rank deficient 8x8 matrix, rank = 7
warning: matrix singular to machine precision, rcond = 1.10156e-21
warning: attempting to find minimum norm solution
warning: matrix singular to machine precision, rcond = 1.0145e-21
warning: attempting to find minimum norm solution
warning: dgelsd: rank deficient 9x9 matrix, rank = 6
warning: matrix singular to machine precision, rcond = 2.16217e-26
warning: attempting to find minimum norm solution
warning: matrix singular to machine precision, rcond = 1.66008e-26
warning: attempting to find minimum norm solution
warning: dgelsd: rank deficient 10x10 matrix, rank = 5
```

Why does it fail?

Model Selection

- Underfitting
(model is too simple to explain data)
- Overfitting
(model is too complicated to learn from data)
 - E.g. too many parameters
 - Insufficient confidence to estimate parameter
(failed matrix inverse)
 - Often training error decreases nonetheless
- Model selection
Need to quantify model complexity vs. data
- This course - algorithms, model selection, questions



MAGIC Etch A Sketch[®] SCREEN

Big Data
on the
Internet

Horizontal
Lid

OHIO ART The World of Toys[®]

Vertical
Lid

MAGIC SCREEN IS GLASS SET IN CURVED PLASTIC FRAME.
USE WITH CARE.

Data - User generated content

- Webpages (content, graph)
- Clicks (ad, page, social)
- Users (OpenID, FB Connect)
- e-mails (Hotmail, Y!Mail, Gmail)
- Photos, Movies (Flickr, YouTube, Vimeo ...)
- Cookies / tracking info (see Ghostery)
- Installed apps (Android market etc.)
- Location (Latitude, Loopt, Foursquared)
- User generated content (Wikipedia & co)
- Ads (display, text, DoubleClick, Yahoo)
- Comments (Disqus, Facebook)
- Reviews (Yelp, Y!Local)
- Third party features (e.g. Experian)
- Social connections (LinkedIn, Facebook)
- Purchase decisions (Netflix, Amazon)
- Instant Messages (YIM, Skype, Gtalk)
- Search terms (Google, Bing)
- Timestamp (everything)
- News articles (BBC, NYTimes, Y!News)
- Blog posts (Tumblr, Wordpress)
- Microblogs (Twitter, Jaiku, Meme)

flickr™



DISQUS



You Tube

yelp. 

> 1 B images, 40h video/minute

Data - User generated content

- Webpages (content, graph)
- Clicks (ad, page, social)
- Users (OpenID, FB Connect)
- e-mails (Hotmail, Y!Mail, Gmail)
- Photos, Movies (Flickr, YouTube, Vimeo ...)
- Cookies / tracking info (see Ghostery)
- Installed apps (Android market etc.)
- Location (Latitude, Loopt, Foursquared)
- User generated content (Wikipedia & co)
- Ads (display, text, DoubleClick, Yahoo)
- Comments (Disqus, Facebook)
- Reviews (Yelp, Y!Local)
- Third party features (e.g. Experian)
- Social connections (LinkedIn, Facebook)
- Purchase decisions (Netflix, Amazon)
- Instant Messages (YIM, Skype, Gtalk)
- Search terms (Google, Bing)
- Timestamp (everything)
- News articles (BBC, NYTimes, Y!News)
- Blog posts (Tumblr, Wordpress)
- Microblogs (Twitter, Jaiku, Meme)

crawl it

flickr™



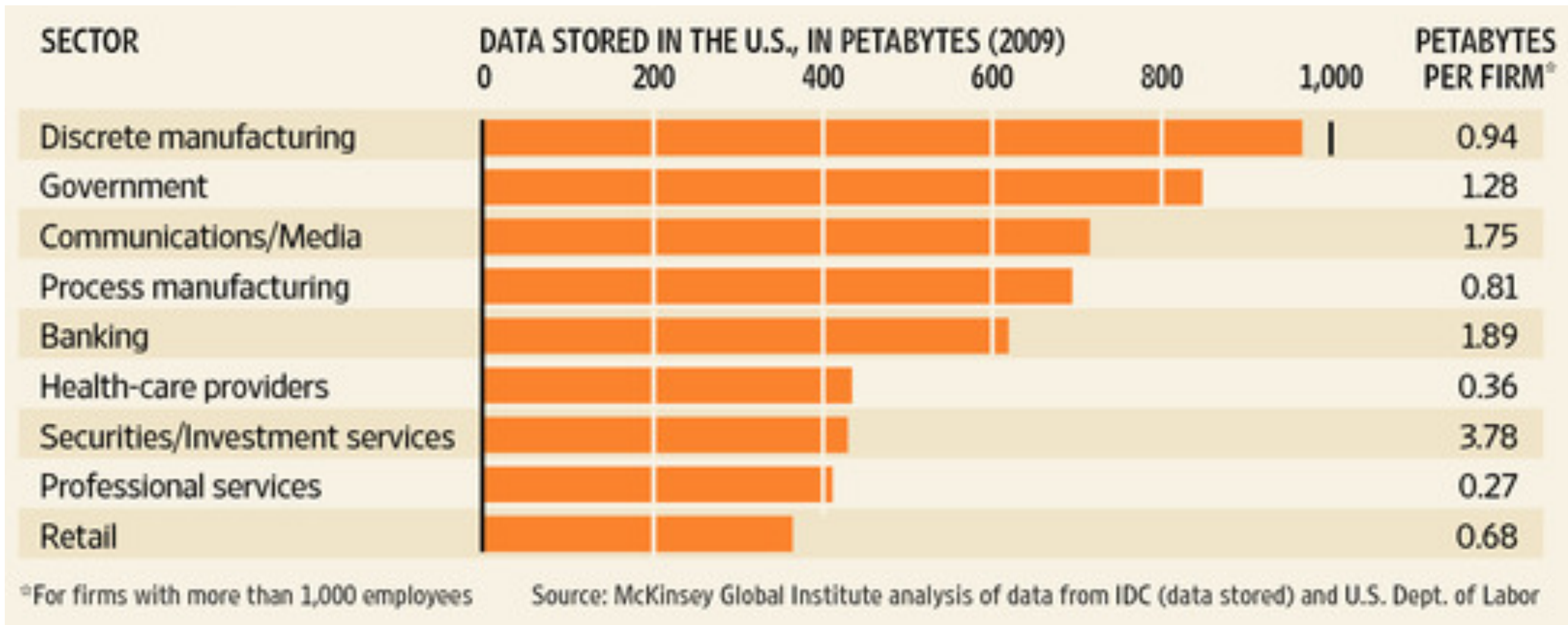
You Tube

DISQUS

yelp. 

> 1 B images, 40h video/minute

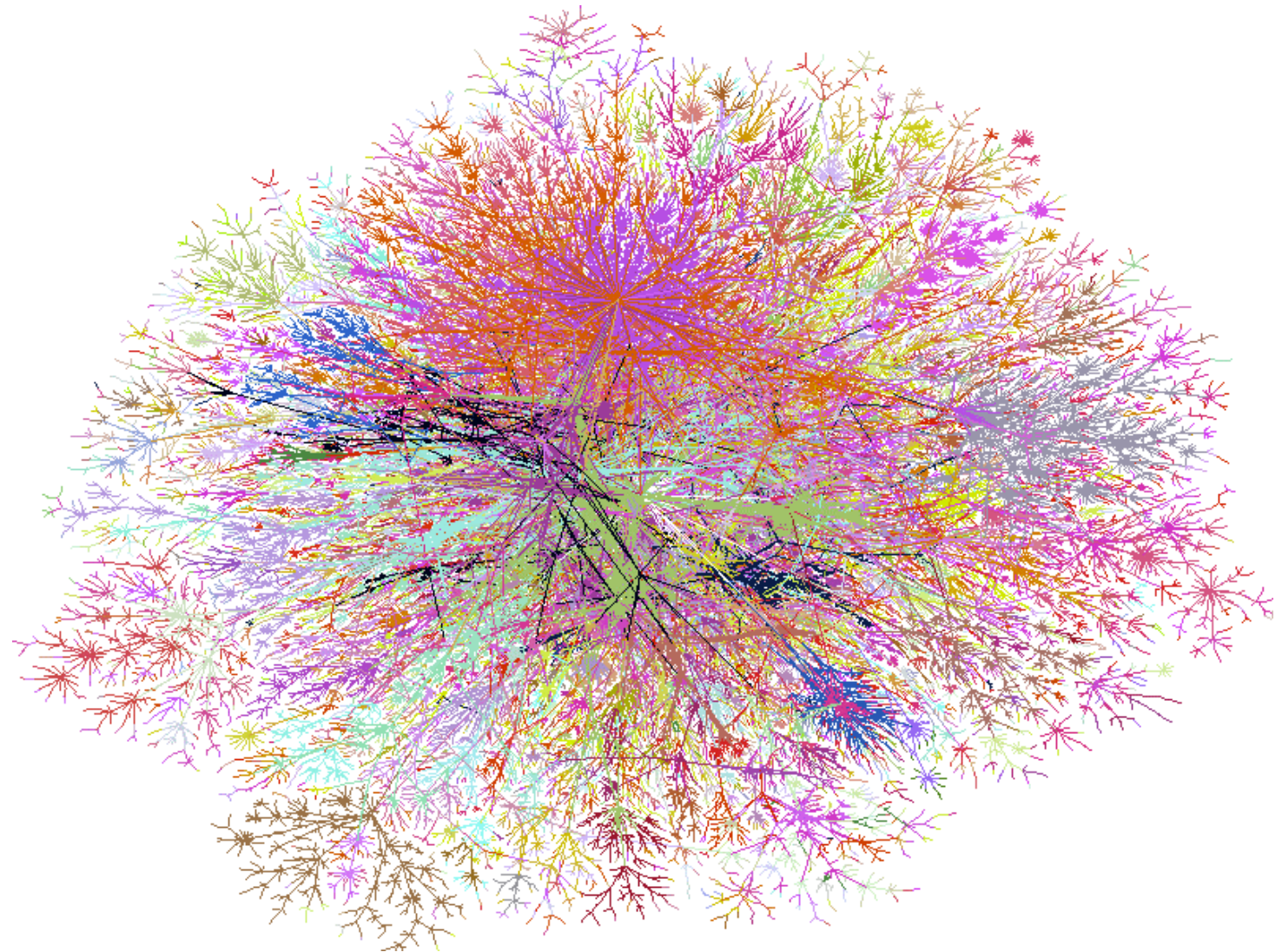
Big Data



we need Big Learning

Data

- Webpages (content, graph)
- Clicks (ad, page, social)
- Users (OpenID, FB Connect)
- e-mails (Hotmail, Y!Mail, Gmail)
- Photos, Movies (Flickr, YouTube, Vimeo ...)
- Cookies / tracking info (see Ghostery)
- Installed apps (Android market etc.)
- Location (Latitude, Loopt, Foursquared)
- User generated content (Wikipedia & co)
- Ads (display, text, DoubleClick, Yahoo)
- Comments (Disqus, Facebook)
- Reviews (Yelp, Y!Local)
- Third party features (e.g. Experian)
- Social connections (LinkedIn, Facebook)
- Purchase decisions (Netflix, Amazon)
- Instant Messages (YIM, Skype, Gtalk)
- Search terms (Google, Bing)
- Timestamp (everything)
- News articles (BBC, NYTimes, Y!News)
- Blog posts (Tumblr, Wordpress)
- Microblogs (Twitter, Jaiku, Meme)



>10B useful webpages

The Web for \$100k/month

- Webpages (content, graph)
- Clicks (ad, page, social)
- Users (OpenID, FB Connect)
- e-mails (Hotmail, Y!Mail, Gmail)
- Photos, Movies (Flickr, YouTube, Vimeo ...)
- Cookies / tracking info (see Ghostery)
- Installed apps (Android market etc.)
- Location (Latitude, Loopt, Foursquared)
- User generated content (Wikipedia & co)
- Ads (display, text, DoubleClick, Yahoo)
- Comments (Disqus, Facebook)
- Reviews (Yelp, Y!Local)
- Third party features (e.g. Experian)
- Social connections (LinkedIn, Facebook)
- Purchase decisions (Netflix, Amazon)
- Instant Messages (YIM, Skype, Gtalk)
- Search terms (Google, Bing)
- Timestamp (everything)
- News articles (BBC, NYTimes, Y!News)
- Blog posts (Tumblr, Wordpress)
- Microblogs (Twitter, Jaiku, Meme)

- **10 billion pages**
(this is a small subset, maybe 10%)
10k/page = 100TB
(\$10k for disks or EBS 1 month)
- **1000 machines**
10ms/page = 1 day
afford 1-10 MIP/page
(\$20k on EC2 for 0.68\$/h)
- **10 Gbit link**
(\$10k/month via ISP or EC2)
 - 1 day for raw data
 - 300ms/page roundtrip
 - **1000 servers for 1 month**
(\$70k on EC2 for 0.085\$/h)

Data - Identity & Graph

- Webpages (content, graph)
- Clicks (ad, page, social)
- Users (OpenID, FB Connect)
- e-mails (Hotmail, Y!Mail, Gmail)
- Photos, Movies (Flickr, YouTube, Vimeo ...)
- Cookies / tracking info (see Ghostery)
- Installed apps (Android market etc.)
- Location (Latitude, Loopt, Foursquared)
- User generated content (Wikipedia & co)
- Ads (display, text, DoubleClick, Yahoo)
- Comments (Disqus, Facebook)
- Reviews (Yelp, Y!Local)
- Third party features (e.g. Experian)
- Social connections (LinkedIn, Facebook)
- Purchase decisions (Netflix, Amazon)
- Instant Messages (YIM, Skype, Gtalk)
- Search terms (Google, Bing)
- Timestamp (everything)
- News articles (BBC, NYTimes, Y!News)
- Blog posts (Tumblr, Wordpress)
- Microblogs (Twitter, Jaiku, Meme)



100M-1 B vertices
Carnegie Mellon University

Crawling Twitter for \$10k

- Webpages (content, graph)
- Clicks (ad, page, social)
- Users (OpenID, FB Connect)
- e-mails (Hotmail, Y!Mail, Gmail)
- Photos, Movies (Flickr, YouTube, Vimeo ...)
- Cookies / tracking info (see Ghostery)
- Installed apps (Android market etc.)
- Location (Latitude, Loopt, Foursquared)
- User generated content (Wikipedia & co)
- Ads (display, text, DoubleClick, Yahoo)
- Comments (Disqus, Facebook)
- Reviews (Yelp, Y!Local)
- Third party features (e.g. Experian)
- Social connections (LinkedIn, Facebook)
- Purchase decisions (Netflix, Amazon)
- Instant Messages (YIM, Skype, Gtalk)
- Search terms (Google, Bing)
- Timestamp (everything)
- News articles (BBC, NYTimes, Y!News)
- Blog posts (Tumblr, Wordpress)
- Microblogs (Twitter, Jaiku, Meme)
- 300M users
- Per user 300 queries/h
- 100 edges/query
- 100 edges/account
- Need 100 machines for 2 weeks (crawl it at 10 queries/s)
 - Tweets
 - Inlinks
 - Outlinks
- Cost
 - \$3k for computers on EC2
 - Similar for network & storage
 - **Need 10k user keys**

Data - Messages

- Webpages (content, graph)
- Clicks (ad, page, social)
- Users (OpenID, FB Connect)
- e-mails (Hotmail, Y!Mail, Gmail)
- Photos, Movies (Flickr, YouTube, Vimeo ...)
- Cookies / tracking info (see Ghostery)
- Installed apps (Android market etc.)
- Location (Latitude, Loopt, Foursquared)
- User generated content (Wikipedia & co)
- Ads (display, text, DoubleClick, Yahoo)
- Comments (Disqus, Facebook)
- Reviews (Yelp, Y!Local)
- Third party features (e.g. Experian)
- Social connections (LinkedIn, Facebook)
- Purchase decisions (Netflix, Amazon)
- Instant Messages (YIM, Skype, Gtalk)
- Search terms (Google, Bing)
- Timestamp (everything)
- News articles (BBC, NYTimes, Y!News)
- Blog posts (Tumblr, Wordpress)
- Microblogs (Twitter, Jaiku, Meme)



> 1 B texts

Data - Messages

- Webpages (content, graph)
- Clicks (ad, page, social)
- Users (OpenID, FB Connect)
- e-mails (Hotmail, Y!Mail, Gmail)
- Photos, Movies (Flickr, YouTube, Vimeo ...)
- Cookies / tracking info (see Ghostery)
- Installed apps (Android market etc.)
- Location (Latitude, Loopt, Foursquared)
- User generated content (Wikipedia & co)
- Ads (display, text, DoubleClick, Yahoo)
- Comments (Disqus, Facebook)
- Reviews (Yelp, Y!Local)
- Third party features (e.g. Experian)
- Social connections (LinkedIn, Facebook)
- Purchase decisions (Netflix, Amazon)
- Instant Messages (YIM, Skype, Gtalk)
- Search terms (Google, Bing)
- Timestamp (everything)
- News articles (BBC, NYTimes, Y!News)
- Blog posts (Tumblr, Wordpress)
- Microblogs (Twitter, Jaiku, Meme)

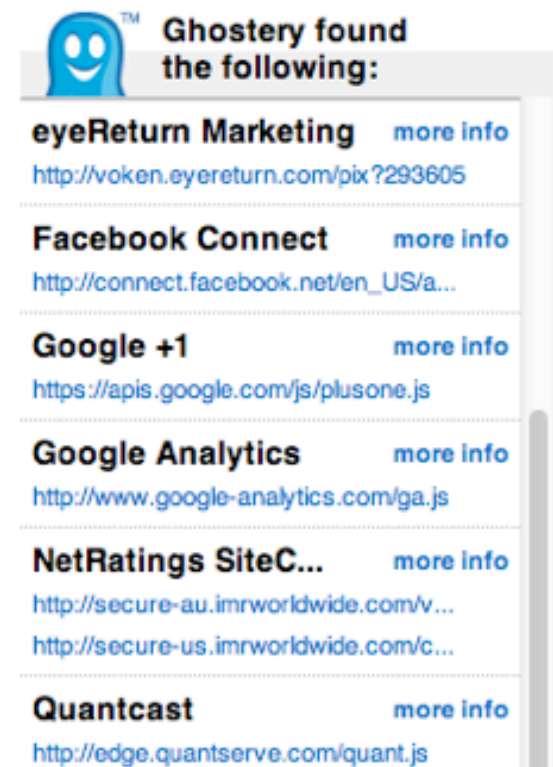


> 1 B texts

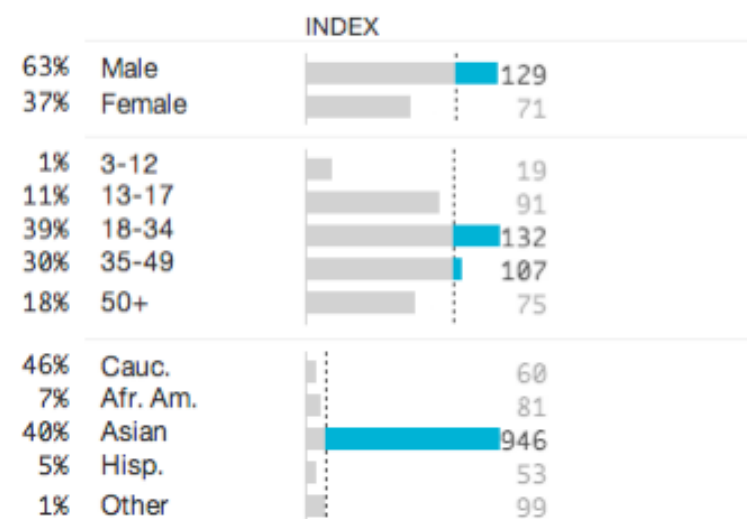
impossible without NDA

Data - User Tracking

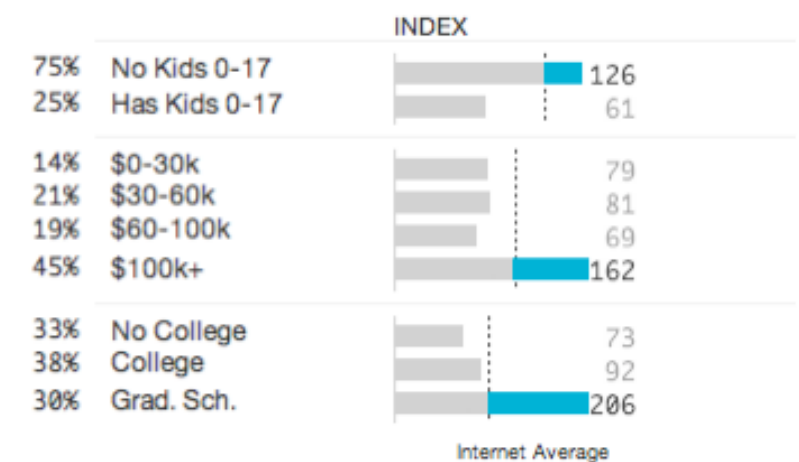
- Webpages (content, graph)
- Clicks (ad, page, social)
- Users (OpenID, FB Connect)
- e-mails (Hotmail, Y!Mail, Gmail)
- Photos, Movies (Flickr, YouTube, Vimeo ...)
- Cookies / tracking info (see Ghostery)
- Installed apps (Android market etc.)
- Location (Latitude, Loopt, Foursquared)
- User generated content (Wikipedia & co)
- Ads (display, text, DoubleClick, Yahoo)
- Comments (Disqus, Facebook)
- Reviews (Yelp, Y!Local)
- Third party features (e.g. Experian)
- Social connections (LinkedIn, Facebook)
- Purchase decisions (Netflix, Amazon)
- Instant Messages (YIM, Skype, Gtalk)
- Search terms (Google, Bing)
- Timestamp (everything)
- News articles (BBC, NYTimes, Y!News)
- Blog posts (Tumblr, Wordpress)
- Microblogs (Twitter, Jaiku, Meme)



US Demographics



Updated Sep 10, 2011 • Next: Sep 21, 2011 by 9AM PDT



alex.smola.org

> 1 B 'identities'

Carnegie Mellon University

Data - User Tracking

- Webpages (content, graph)
- Clicks (ad, page, social)
- Users (OpenID, FB Connect)
- e-mails (Hotmail, Y!Mail, Gmail)
- Photos, Movies (Flickr, YouTube, Vimeo ...)
- Cookies / tracking info (see Ghostery)
- Installed apps (Android market etc.)
- Location (Latitude, Loopt, Foursquared)
- User generated content (Wikipedia & co)
- Ads (display, text, DoubleClick, Yahoo)
- Comments (Disqus, Facebook)
- Reviews (Yelp, Y!Local)
- Third party features (e.g. Experian)
- Social connections (LinkedIn, Facebook)
- Purchase decisions (Netflix, Amazon)
- Instant Messages (YIM, Skype, Gtalk)
- Search terms (Google, Bing)
- Timestamp (everything)
- News articles (BBC, NYTimes, Y!News)
- Blog posts (Tumblr, Wordpress)
- Microblogs (Twitter, Jaiku, Meme)

Privacy Information

Privacy Policy:

<http://www.facebook.com/policy.php>

Data Collected:

Anonymous (browser type, location, page views), Pseudonymous (IP address, "actions taken")

Data Sharing:

Data is shared with third parties. 

Data Retention:

Data is deleted from backup storage after 90 days. 

Privacy Information

Privacy Policy:

<http://www.google.com/intl/en/priv...>

Data Collected:

Anonymous (ad serving domains, browser type, demographics, language settings, page views, time/date), Pseudonymous (IP address)

Data Sharing:

Anonymous data is shared with third parties. 

Data Retention:

Undisclosed 

(implicit) Labels

no Labels

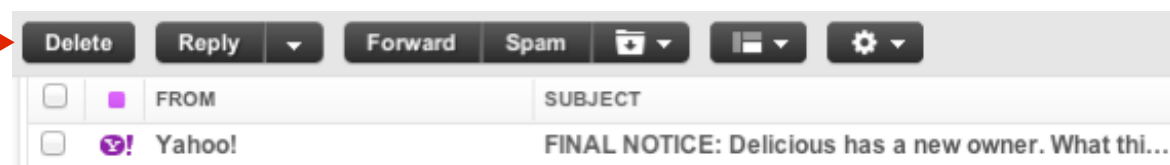
- Ads



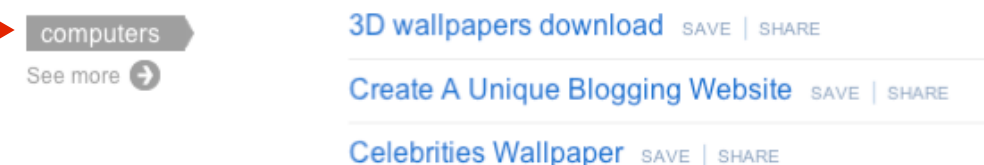
- Click feedback



- Emails

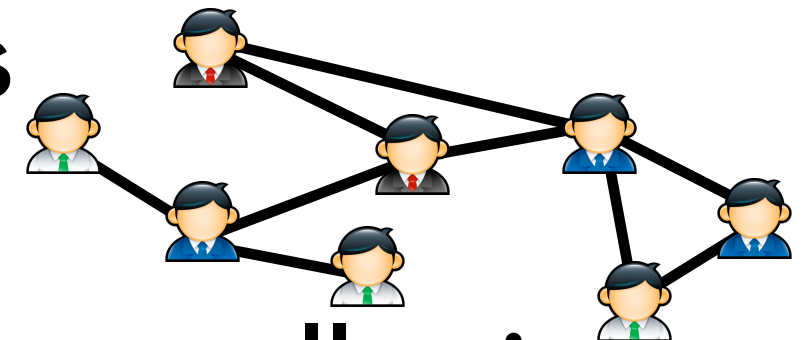


- Tags



- Editorial data is very expensive! Do not use!

- Graphs



- Document collections



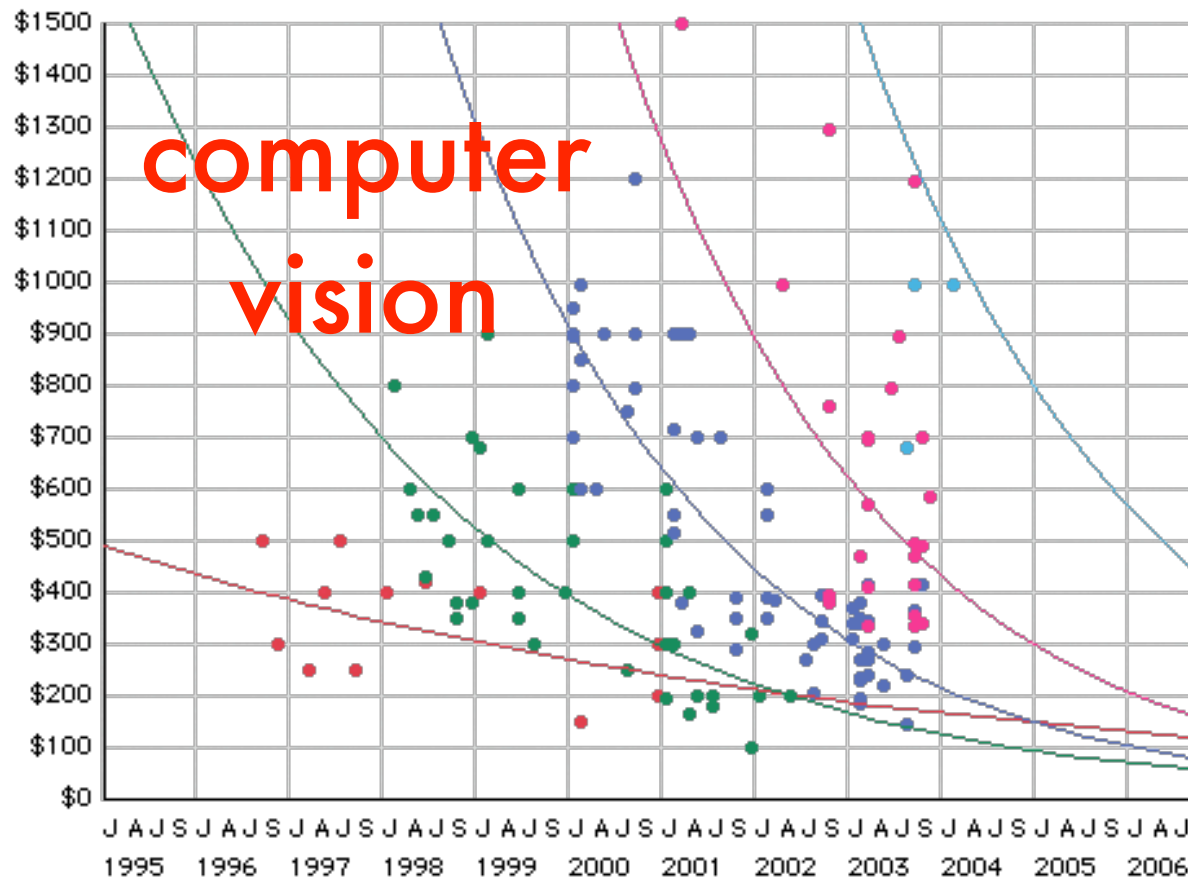
- Email/IM/Discussions



- Query stream



Many more sources



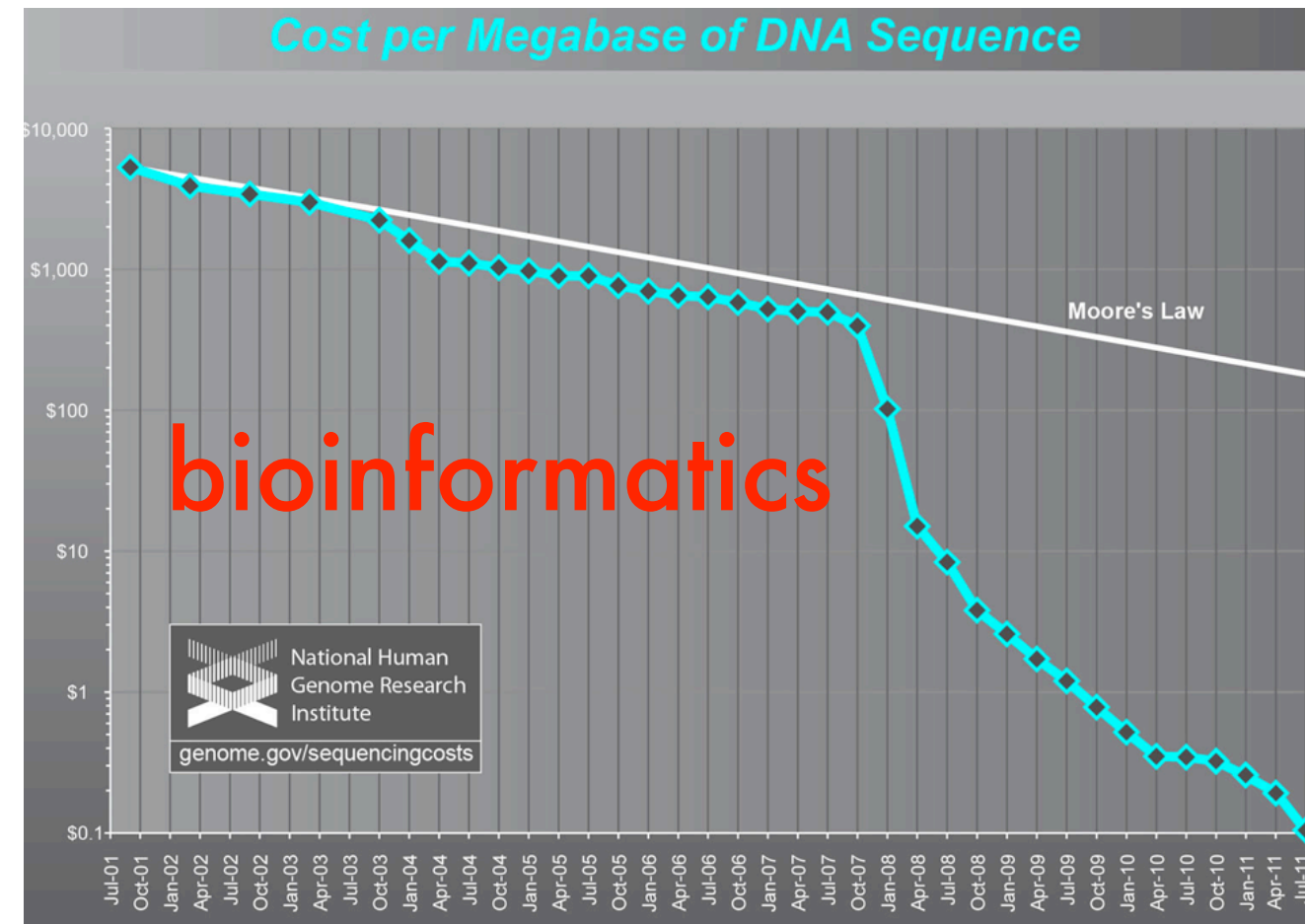
computer vision

6.0 - 6.29 megapixels
 4.92 - 5.1 megapixels
 3.14 megapixels
 1.2 megapixels
 .3 megapixels

<http://keithwiley.com/mindRamblings/digitalCameras.shtml>



personalized sensors

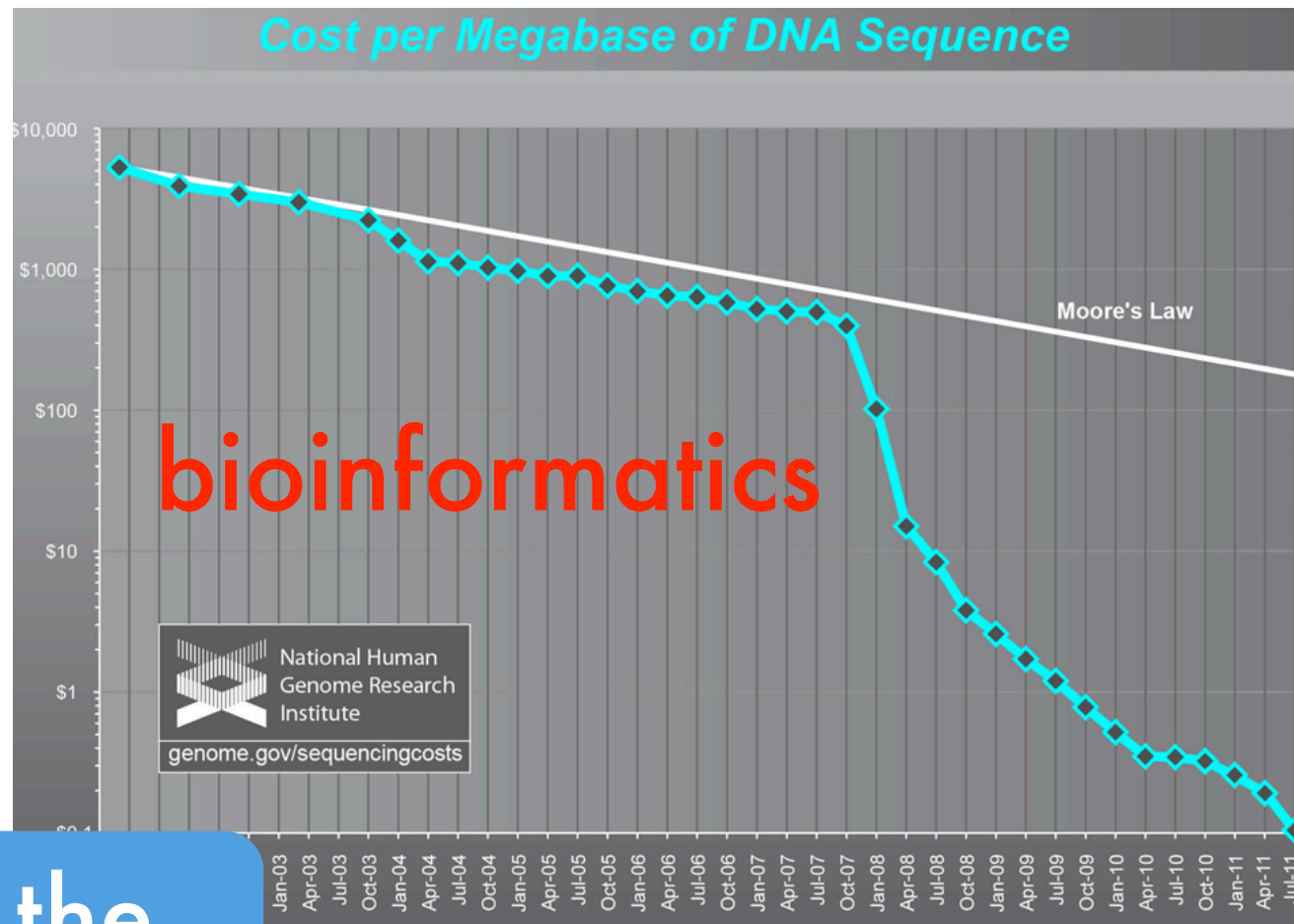
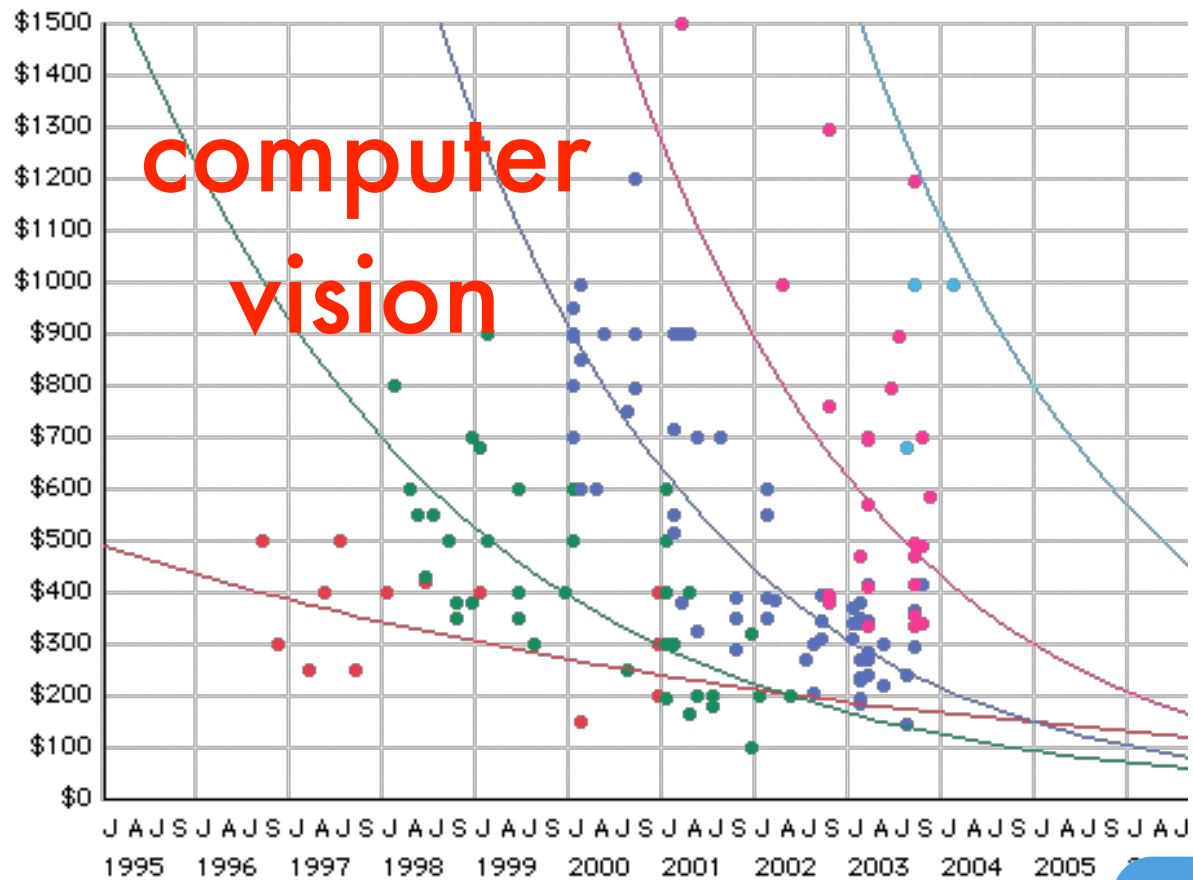


bioinformatics



ubiquitous control

Many more sources



in the cloud



personalized sensors



ubiquitous control

Further material

- Machine learning tutorial
http://alex.smola.org/teaching/cmu2013-10-701/papers/intro_chapter.pdf
- Machine Learning (Tom Mitchell's book)
- Machine Learning Summer Schools
<http://mlss.cc> (lots of videos there)
- Coursera ML intro (more like the 601 class)
<https://www.coursera.org/course/ml>