

1.1 Administrative Stuff

1 Introduction

Alexander Smola

Introduction to Machine Learning 10-701

<http://alex.smola.org/teaching/10-701-15>

<https://piazza.com/cmu/spring2015/10701/>

ADMINISTRATION

Times

- Lectures
9:00 to 10:20 Monday and Wednesday
- Recitations
TBD
- Office Hours
10:20 until 11:00 outside the lecture hall
- TAs
Jay-Yoon Lee, Jin Sun, Shen Wu,
Di Xu, Zhou Yu

Resources

- **Website**

<http://alex.smola.org/teaching/10-701-15>
Slides, links to papers, homework, etc.

- **Piazza**

<https://piazza.com/cmu/spring2015/10701>

- Post all questions and discussions there
- A TA will manage the forums

- **Videos** (Live on YouTube in 4K resolution)

- Machine Learning (Tom Mitchell's book)

- Machine Learning Summer Schools <http://mlss.cc>



**KEEP
CALM
AND
DO YOUR
HOMEWORK**

Homework

- **Homework is due weekly**
(nobody starts earlier than days before anyway)
Monday 9:00am by e-mail to TA
- **You can be late by 1 week for 2 homeworks**
 - But not for 2 weeks for 1 homework
 - It does not matter whether you're 1 hour late or 1 week late. Late is late.
- **It's OK to collaborate.** In fact, you should discuss with others. This will help you understand stuff.
- **But you must not copy. You will get 0 points!**

Projects

- **Teams of 3 students per project**
Teams of 2 are OK but not encouraged (what if someone drops out), solo is definitely not OK
- **Team formation complete by January 28**
Email to TAs to register. Post on Piazza today!
- **Project proposal due on February 9**
Email to TAs with proposal
 - Title, abstract, brief sketch of the idea
 - 2 pages, double column on ACM Template
<http://www.acm.org/sigs/publications/proceedings-templates>
- **Project presentations on April 27 and 29**

Projects - Heilmeier's Catechism

- What are you trying to do? Articulate your objectives using absolutely no jargon.
- How is it done today, and what are the limits of current practice?
- What's new in your approach and why do you think it will be successful?
- Who cares? If you're successful, what difference will it make?
- What are the risks and the payoffs?
- How much will it cost? How long will it take?
- What are the midterm and final "exams" to check for success?

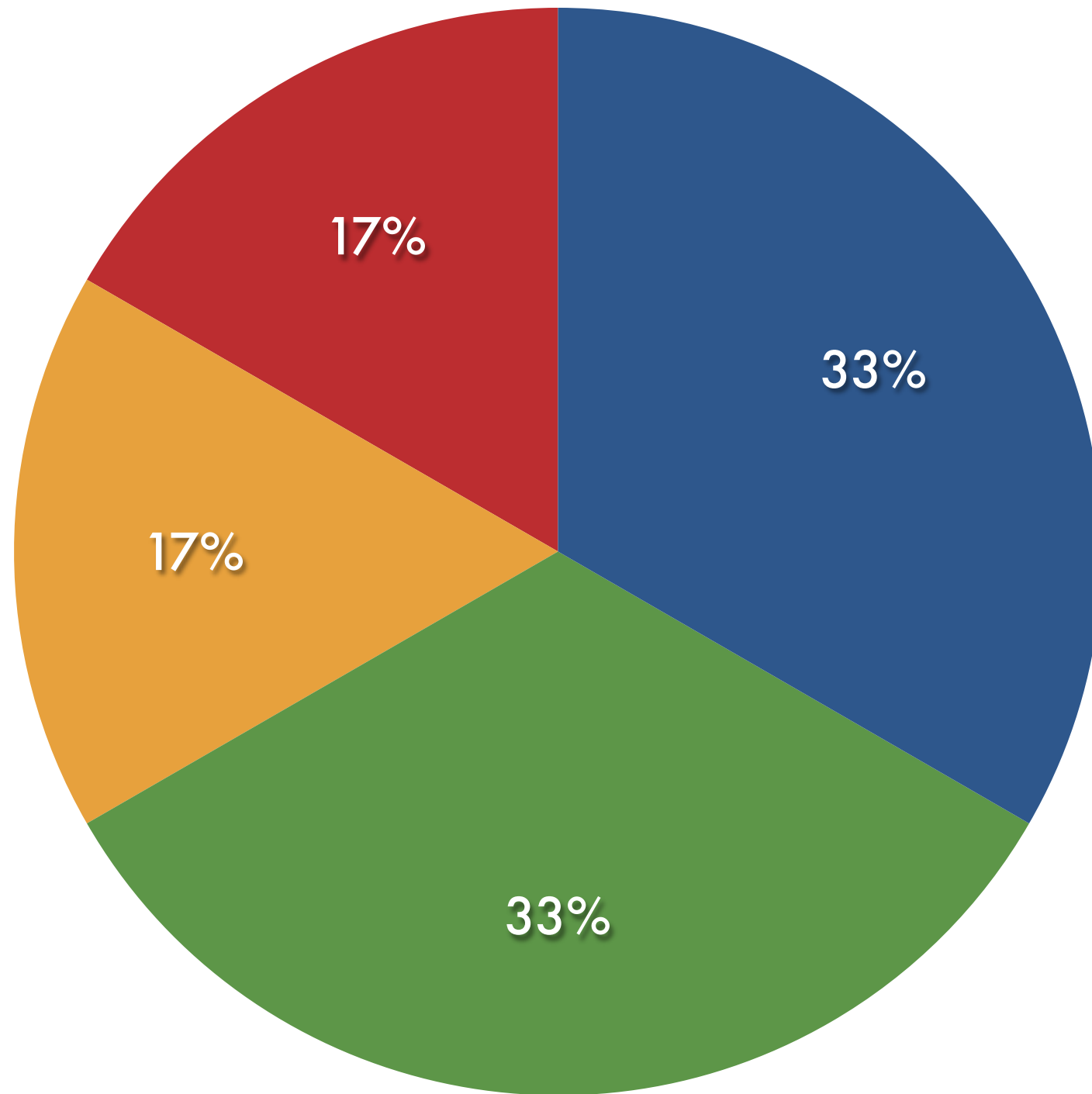
<http://cseweb.ucsd.edu/~ddahlstr/misc/heilmeier.html>

Exams

- Midterm exam
most likely around March 2, 2015
- Final exam
most likely around May 11, 2015
- Luddite exam conditions
 - No computers, tablets, smartphones
 - Big stack of paper less than 10" tall
- You must pass both exams



● Homework ● Project ● Midterm ● Final



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
- **all you need for a startup**
Nearest Neighbors, Decision Trees, Neural Networks, **for the internet**
Fuzzy Logic
- (Generalized) Linear Models
Support Vector Classification, Relevance Vector Machines, Kernel PCA
- Theoretical Tools
Risk Minimization, Convergence Bounds, Information Theory **for your PhD**
- **for Wall Street**
Bayesian Networks, Graphical Models, Dynamic Programming, Latent Markov Models
- Interacting with the environment
Online Learning, Reinforcement Learning
- Scalability **energy** **biology**

1.2 Programming with Data

1 Introduction

Alexander Smola

Introduction to Machine Learning 10-701

<http://alex.smola.org/teaching/10-701-15>



Machine Learning Problems

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
T... 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

Reinforcement Learning

```
Game will be controlled through named FIFO pipes.  
Size 160-210  
OK  
<type 'str'> 67200  
<type 'numpy.ndarray'> 84  
S: 1 A: 0 R: 0 D: 0  
Start
```

```
action: 1  
S: 2 A: 1 R: 1 D: 0  
Reward 0
```

```
action: 1  
S: 3 A: 2 R: 2 D: 0  
Reward 0
```

```
action: 1  
S: 4 A: 3 R: 3 D: 0  
Reward 0
```

```
action NEURALNET: 3  
S: 5 A: 4 R: 4 D: 1  
Reward 0
```

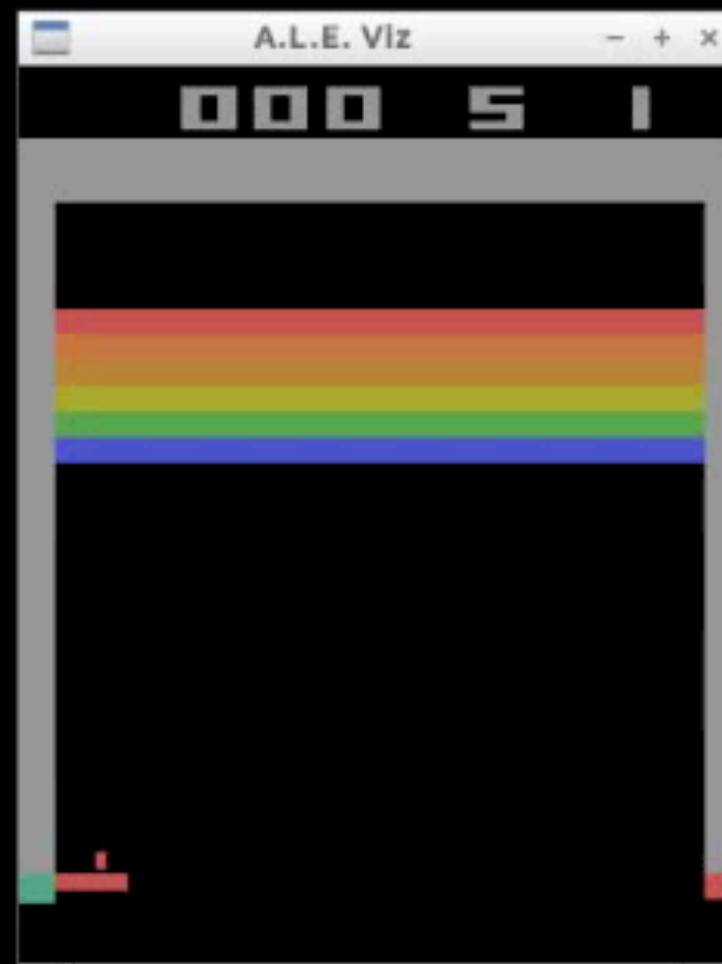
```
action NEURALNET: 3  
S: 6 A: 5 R: 5 D: 2  
Reward 0
```

```
action NEURALNET: 0  
S: 7 A: 6 R: 6 D: 3  
Reward 0
```

```
action NEURALNET: 3  
S: 8 A: 7 R: 7 D: 4  
Reward 0
```

```
action NEURALNET: 0  
S: 9 A: 8 R: 8 D: 5  
Reward 0
```

```
action NEURALNET: 3
```



Reinforcement Learning

```
Game will be controlled through named FIFO pipes.  
Size 160-210  
OK  
<type 'str'> 67200  
<type 'numpy.ndarray'> 84  
S: 1 A: 0 R: 0 D: 0  
Start
```

```
action: 1  
S: 2 A: 1 R: 1 D: 0  
Reward 0
```

```
action: 1  
S: 3 A: 2 R: 2 D: 0  
Reward 0
```

```
action: 1  
S: 4 A: 3 R: 3 D: 0  
Reward 0
```

```
action NEURALNET: 3  
S: 5 A: 4 R: 4 D: 1  
Reward 0
```

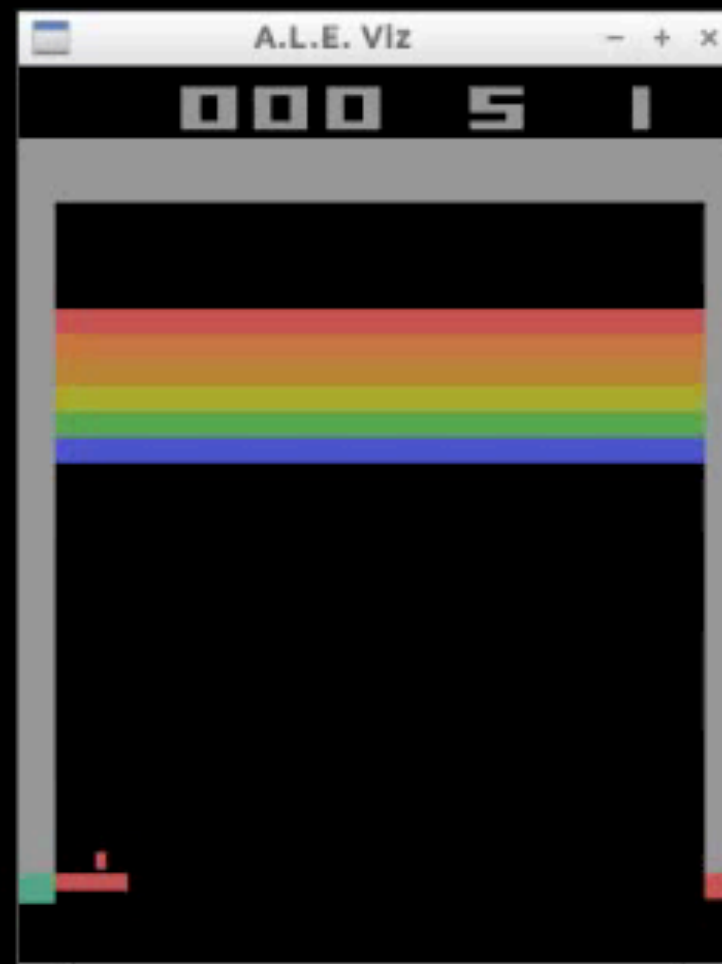
```
action NEURALNET: 3  
S: 6 A: 5 R: 5 D: 2  
Reward 0
```

```
action NEURALNET: 0  
S: 7 A: 6 R: 6 D: 3  
Reward 0
```

```
action NEURALNET: 3  
S: 8 A: 7 R: 7 D: 4  
Reward 0
```


```
action NEURALNET: 0  
S: 9 A: 8 R: 8 D: 5  
Reward 0
```


```
action NEURALNET: 3
```



Spam Filtering

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

Google 

Alex Smola 0 + Share 

Gmail ▾

☐ ▾  More ▾

ham

1–50 of 15,803

< > 

COMPOSE

Inbox (7,180)


Important


Sent Mail

Drafts (61)

<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"/>			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"/>			support, Alex (3)	Your order has shipped... - please send to the address below for an exchange remotesremotes.com(exchange,	7:22 am
<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"/>			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 ▾

☐ ▾  More ▾

spam

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...

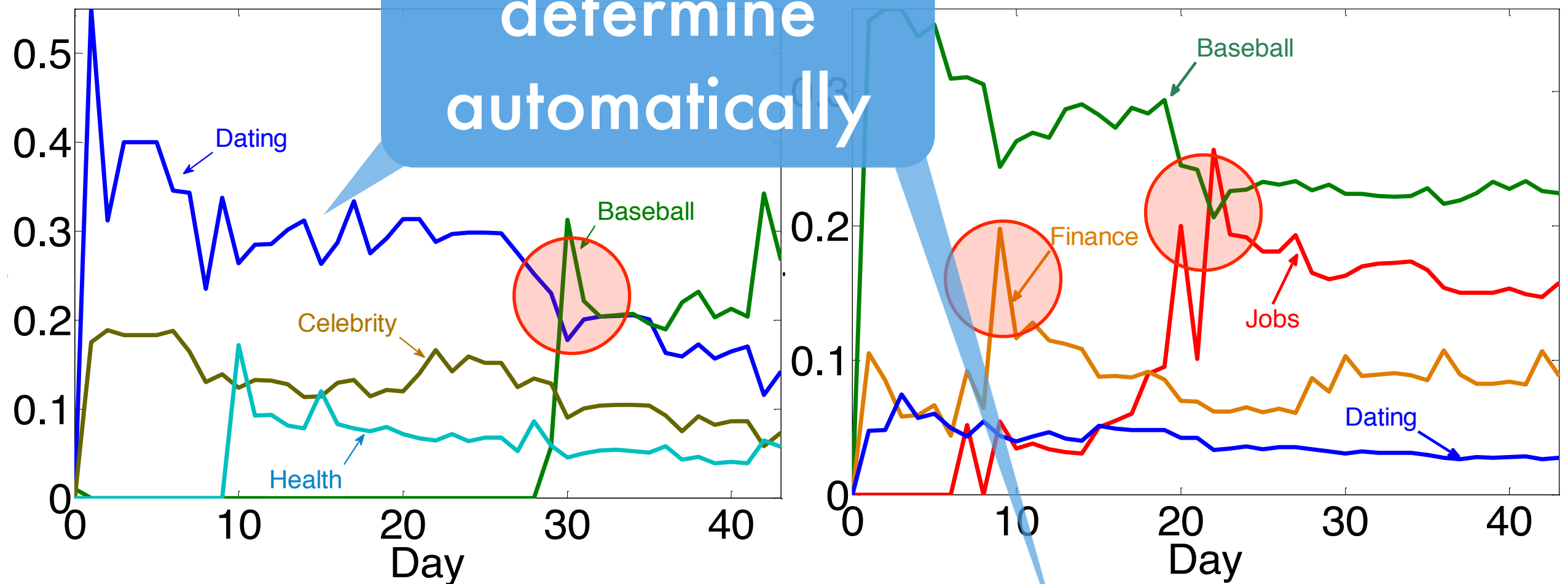
Search people...

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

<input 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"/>			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"/>			garjeti	Call for Research Papers - GLOBAL ADVANCED RESEARCH JOURNAL OF ENGINEERING, TECHNOLOG	Jan 11
<input type="checkbox"/>			Steven Cooke	Congratulations Alex, \$150 awaits you - Alex: IMPORTANT - NOTICE OF WINNINGS Please make sure yo	Jan 11
<input 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"/>			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"/>			garjeti	Call for Research Papers - GLOBAL ADVANCED RESEARCH JOURNAL OF ENGINEERING, TECHNOLOG	Jan 10
<input type="checkbox"/>			Candy.Li	中层,不只当老板的代言人	Jan 9
<input 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"/>			RE/MAX®	2013 Valueable Offer! - Hello Friend, RE/MAX® has issued 2013 valuable property offer in your resident from	Jan 9
<input type="checkbox"/>			newsletter	newsletter WWW2013 - Newsletter 6 - See the Portuguese and Spanish version right after the English versio	Jan 9
<input 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"/>			garieti (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

NOT NEGOTIABLE - DO NOT CASH
JAMES C. MORRISON
MARY A. MORRISON
1785 SHERIDAN DR.
YOUR CITY, STATE 10099

NO. 123
DATE 8/27/03 00-5789/0000

PAY TO THE ORDER OF Bob's Car Wash - Peter Smith \$ 50 00
Fifty and 00/100 DOLLARS

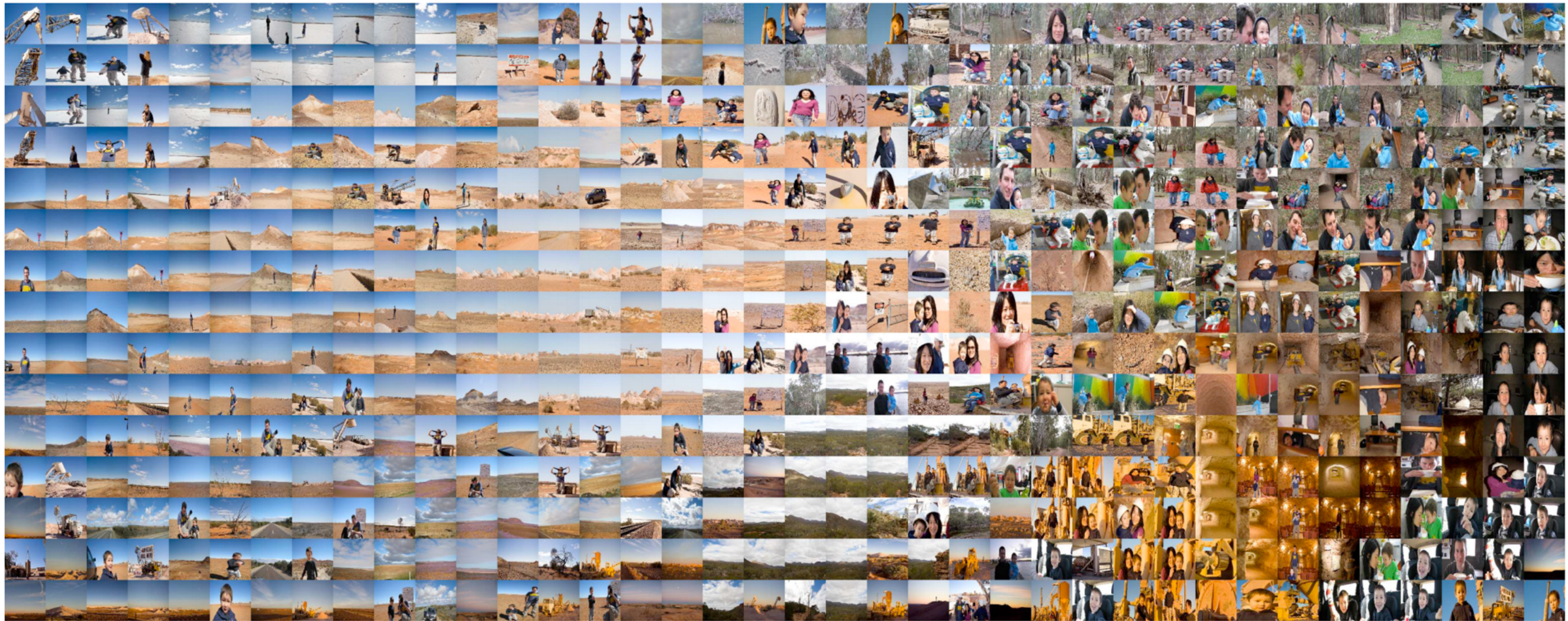
NOT NEGOTIABLE
SAMPLE - VOID

DELUXE
YOUR CITY, STATE 10099

00000067894 12345678 00000005000

recognize
handwriting

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

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?

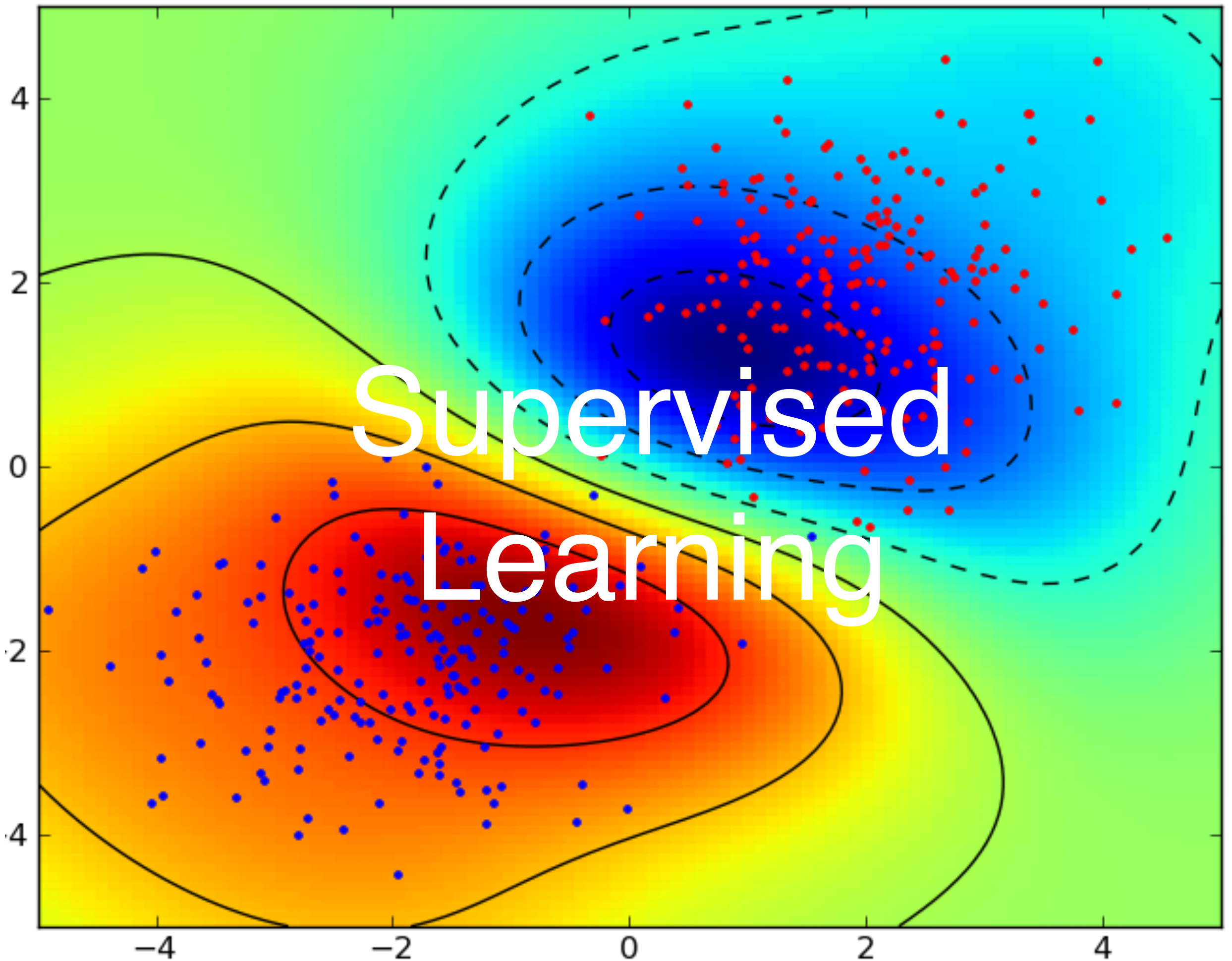
True startup story

- 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)

Supervised Learning

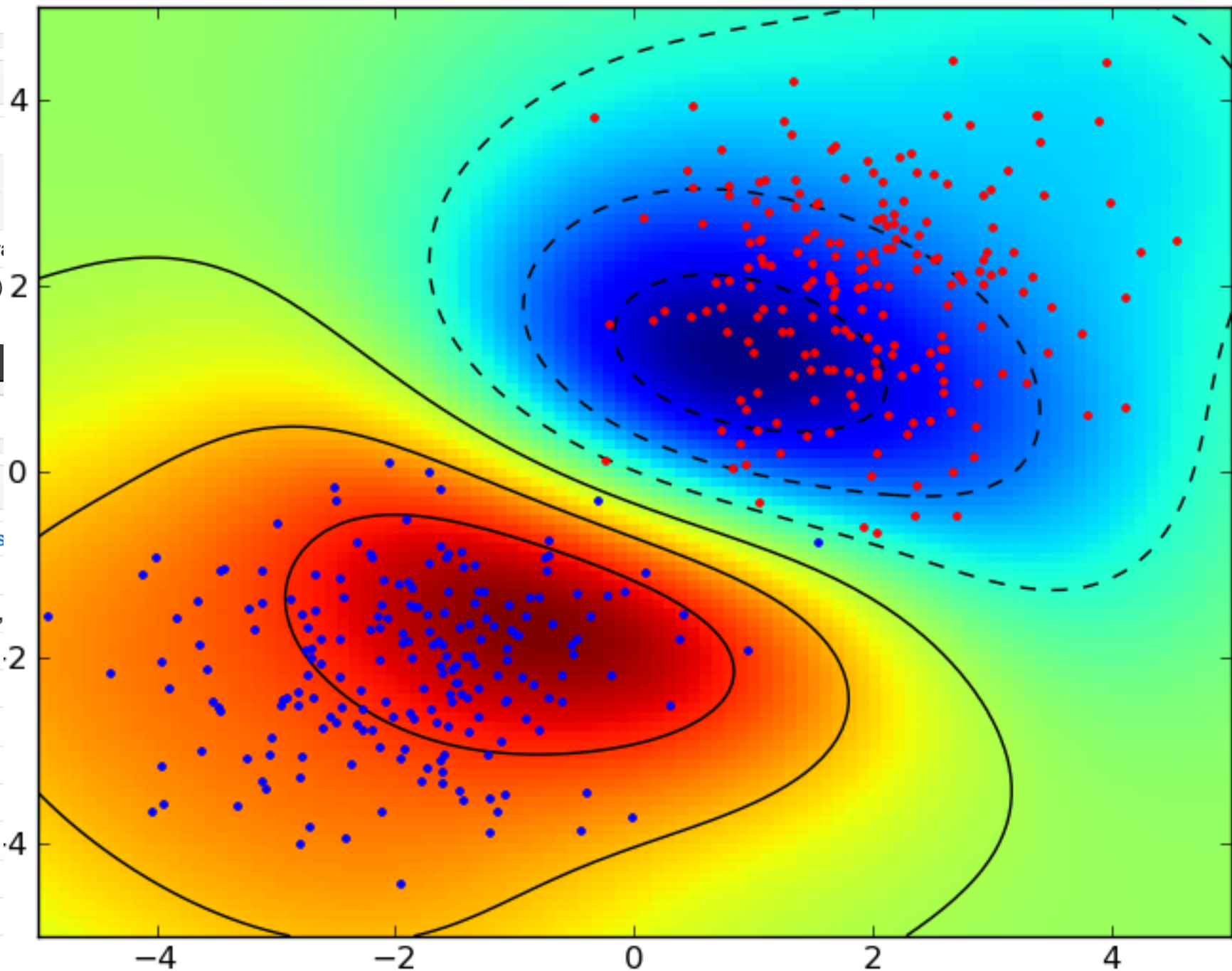
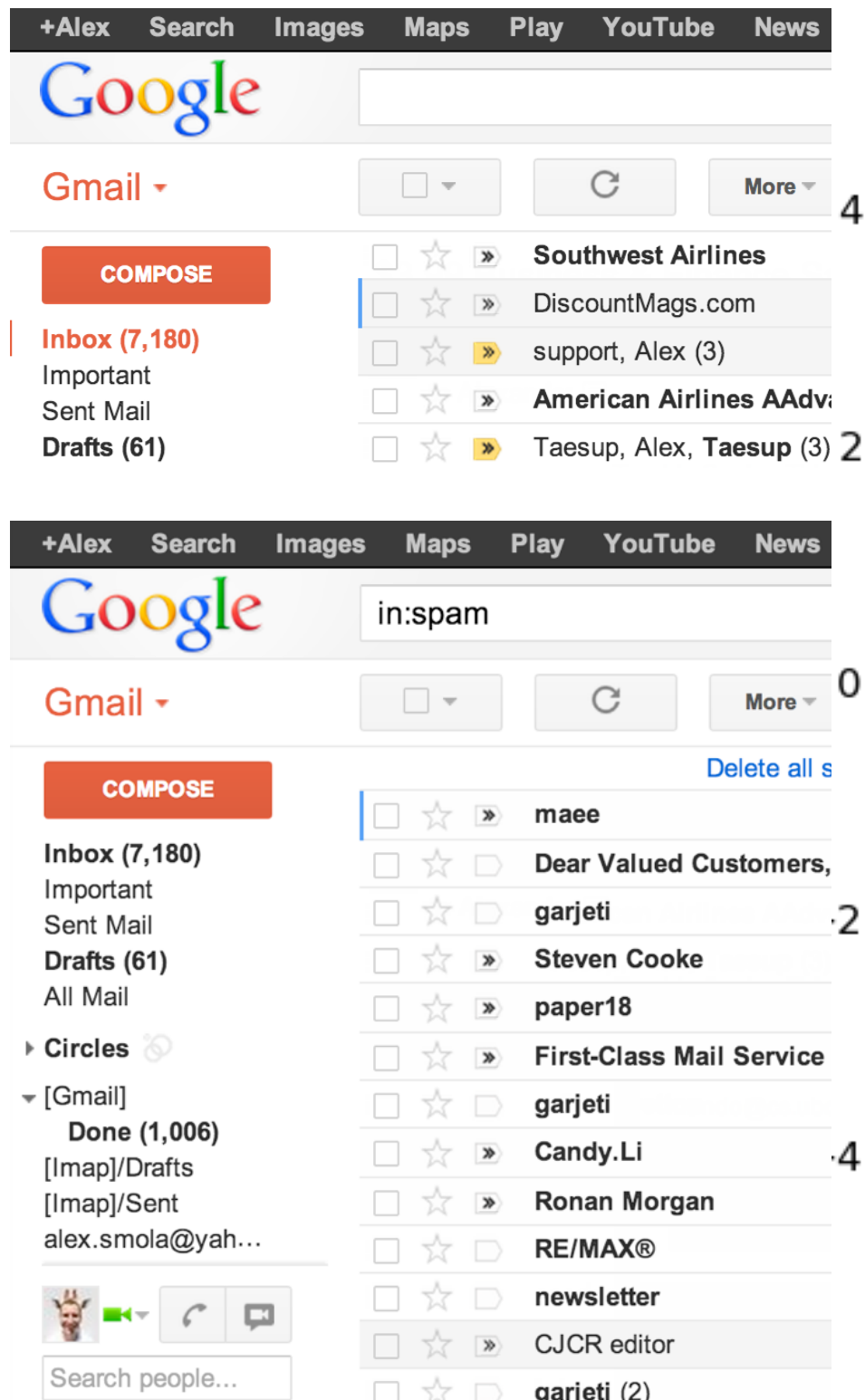


Supervised Learning $y = f(x)$

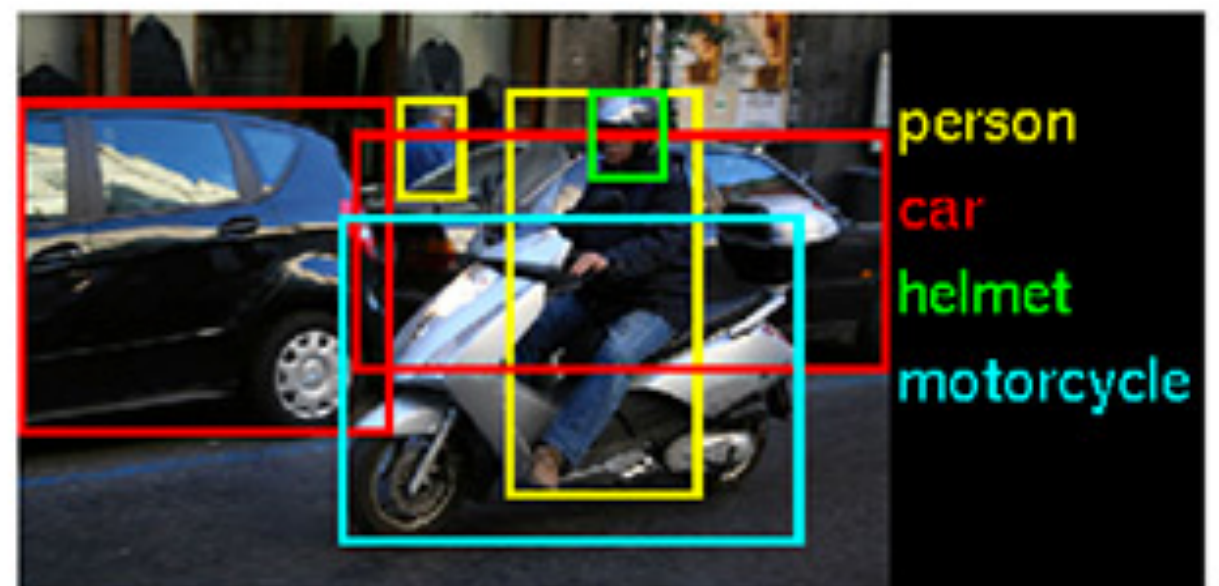
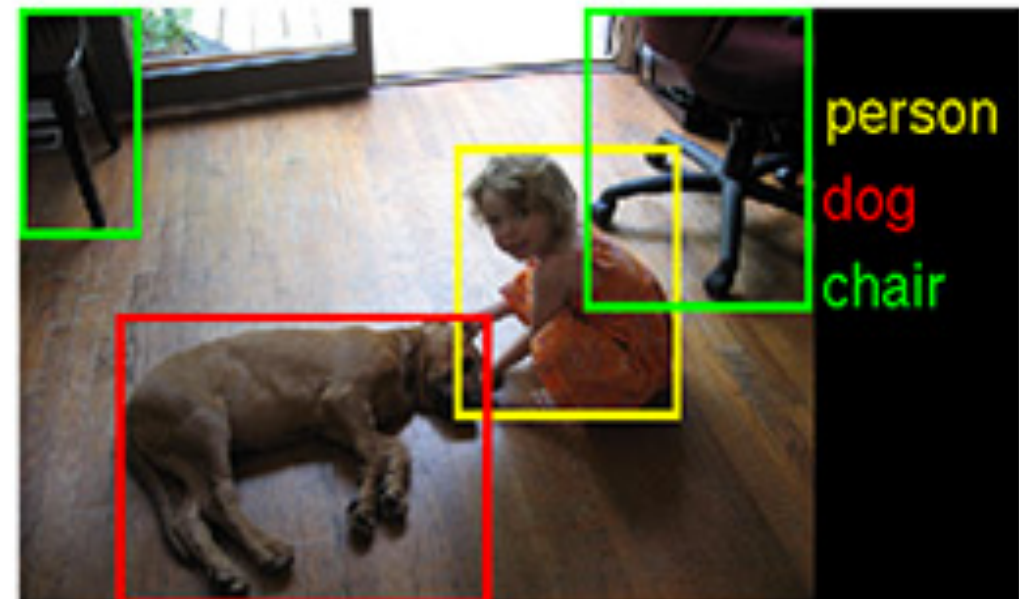
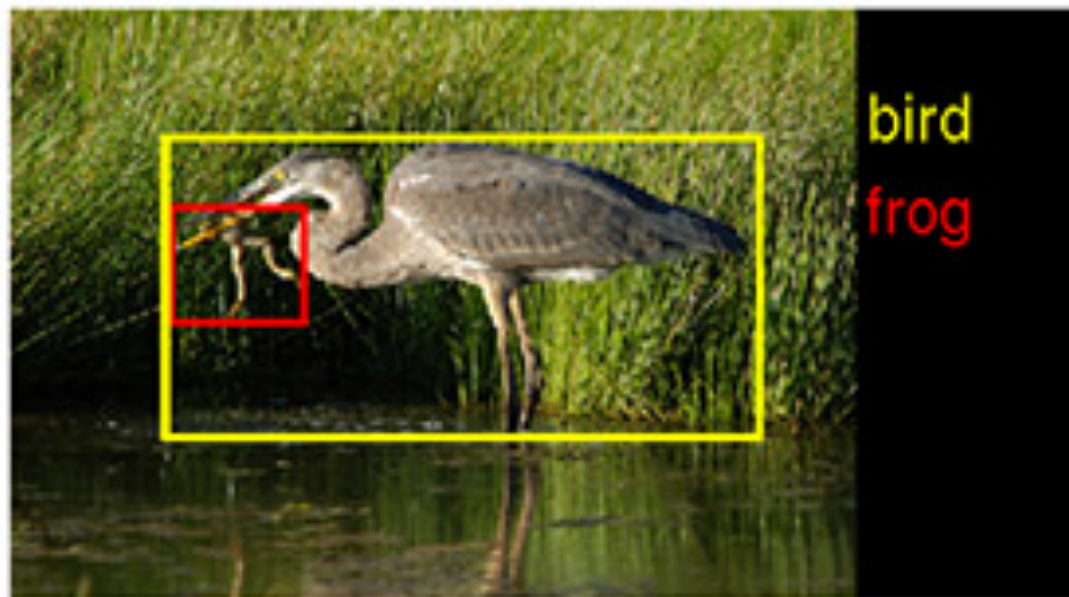
- Binary classification
Given x find y in $\{-1, 1\}$
- Multicategory classification
Given x find y in $\{1, \dots, k\}$
- Regression
Given x find y in \mathbb{R} (or \mathbb{R}^d)
- Sequence annotation
Given sequence $x_1 \dots x_l$ find $y_1 \dots y_l$
- Hierarchical Categorization (Ontology)
Given 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
 $l(y, f(x))$

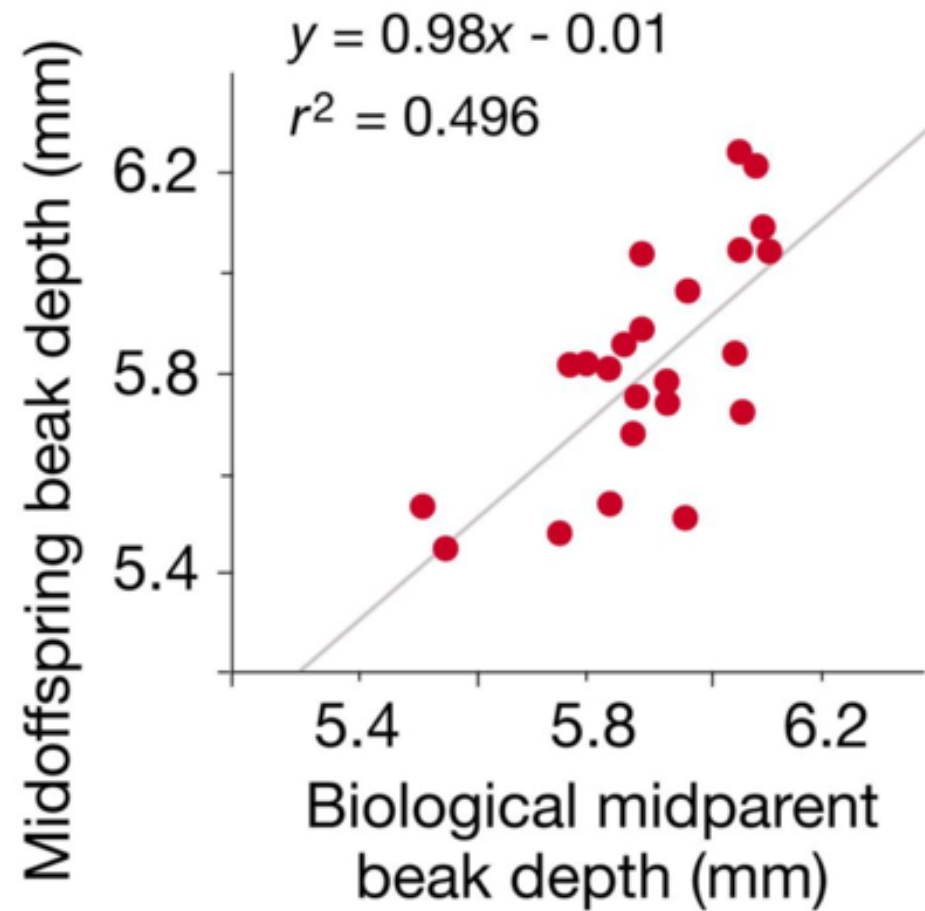
Binary Classification



Multiclass Classification + Annotation



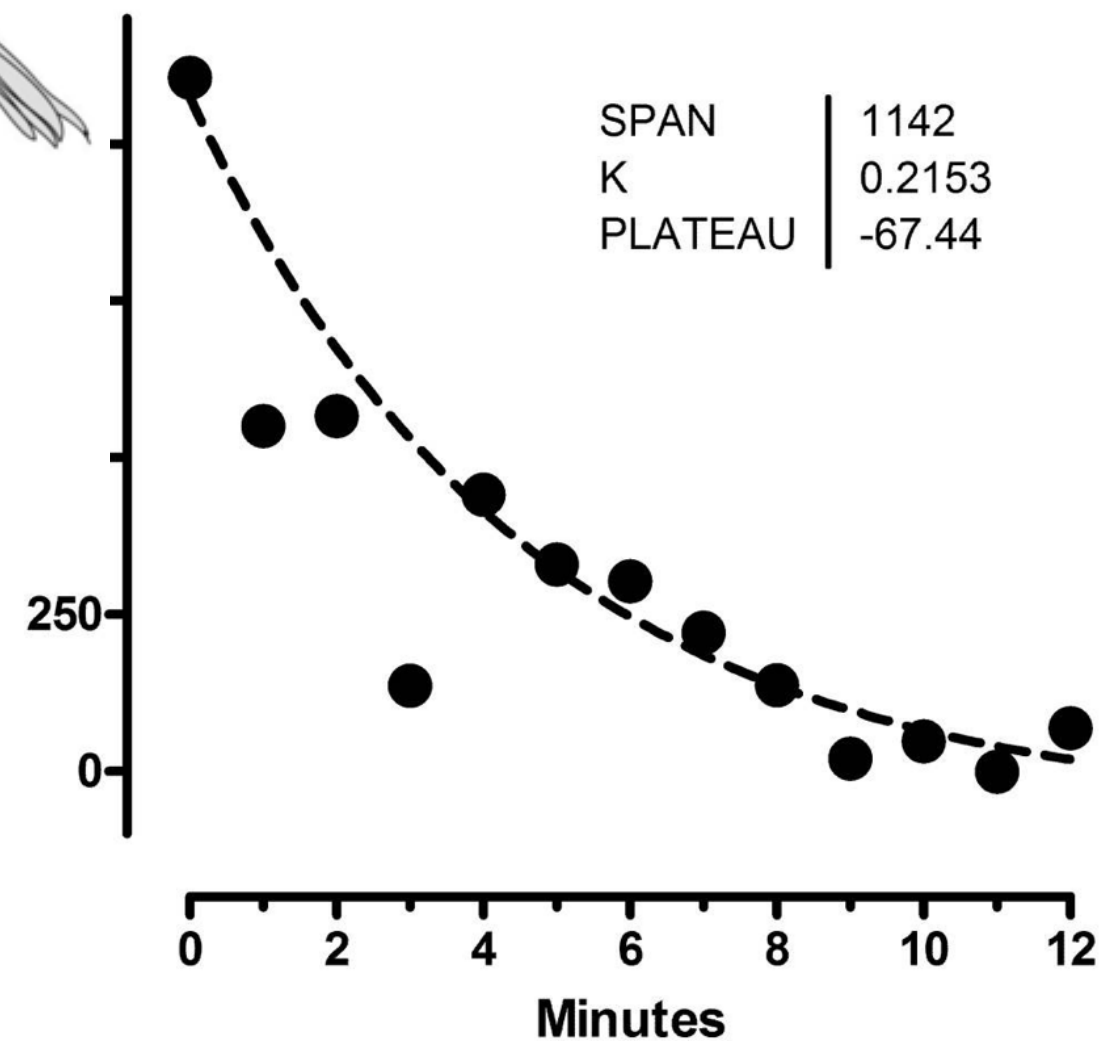
Regression



Copyright © 2004 Pearson Prentice Hall, Inc.

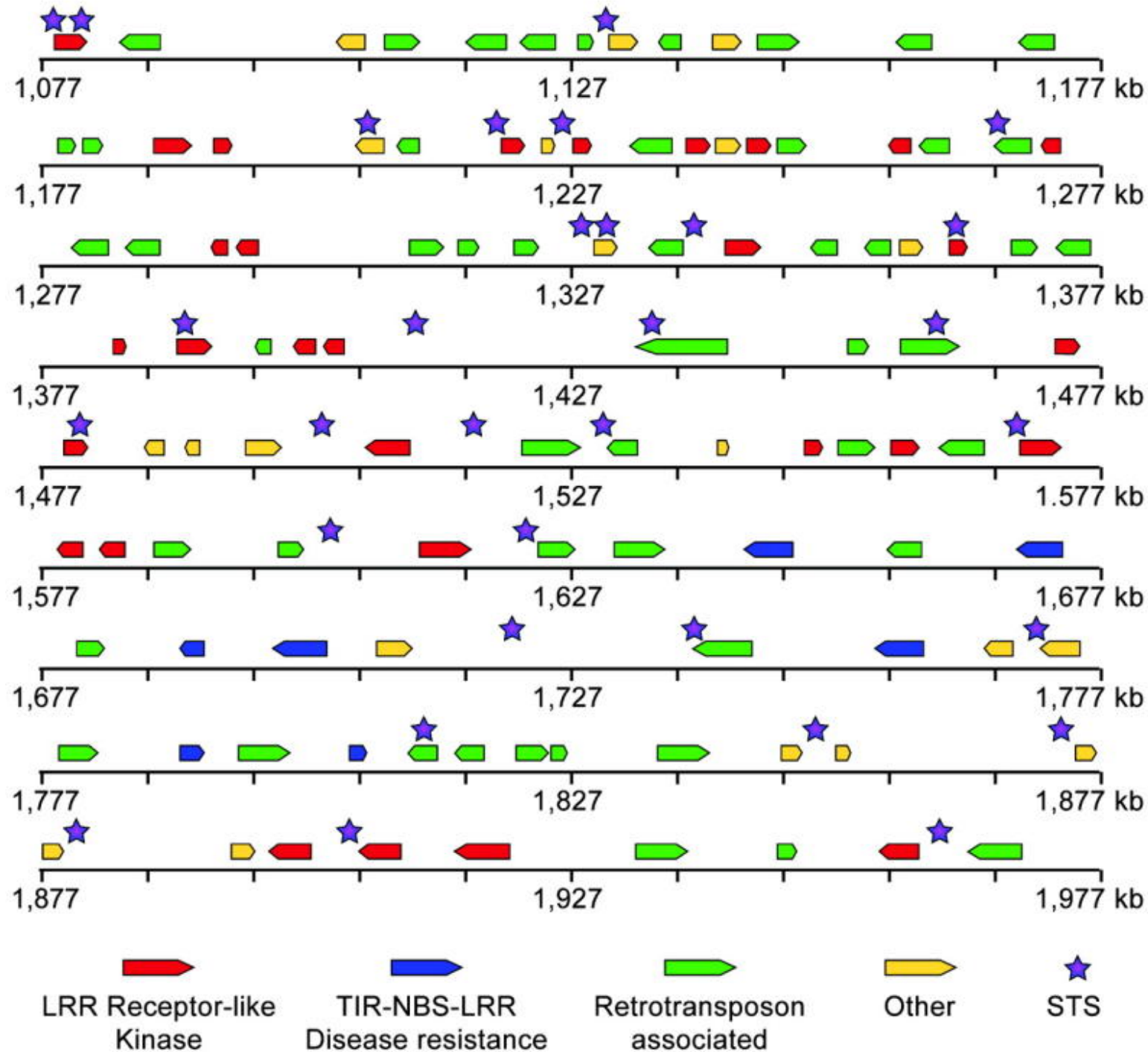
linear

nonlinear



SPAN	1142
K	0.2153
PLATEAU	-67.44

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](#)...

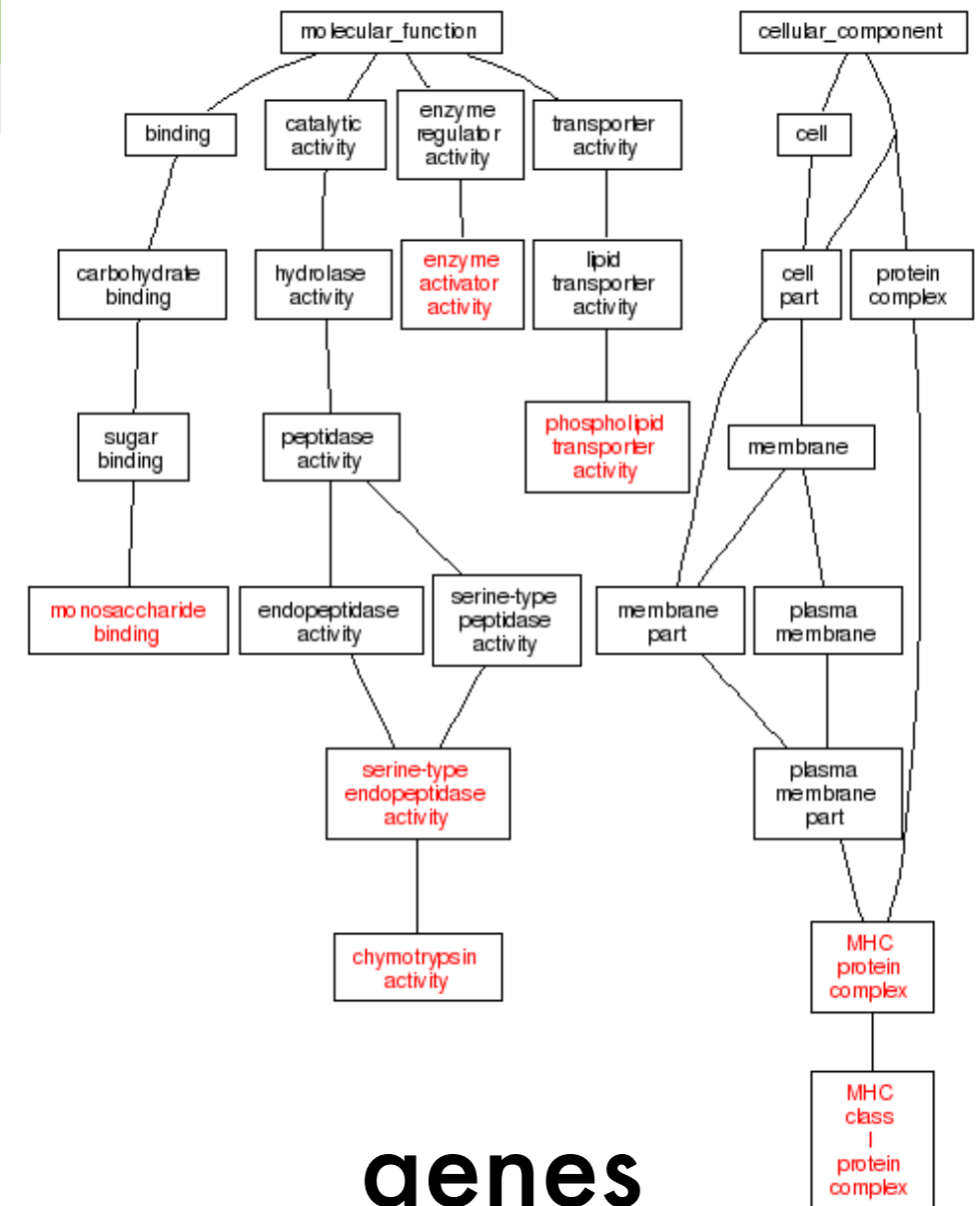
Sports

[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



genes

Prediction



tomorrow's stock price

Carnegie Mellon University

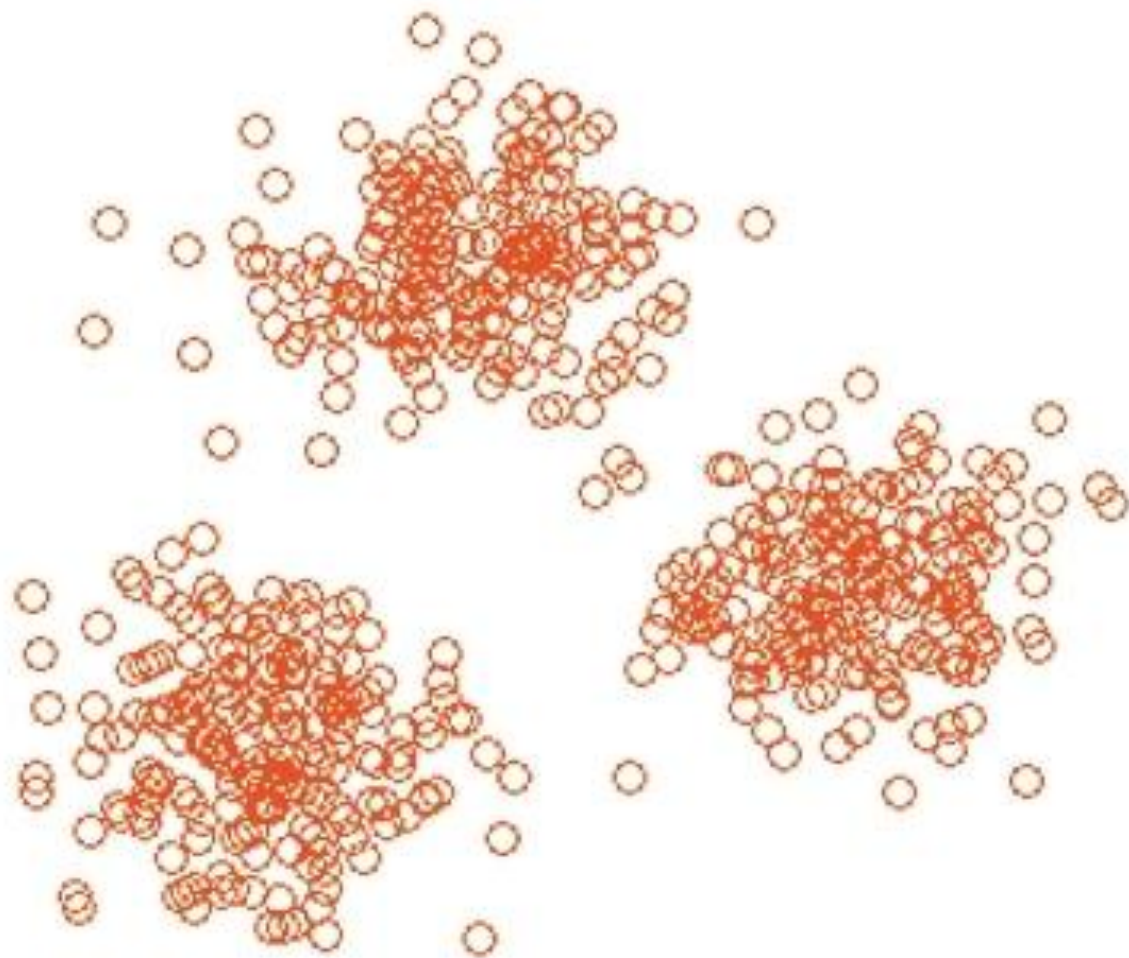
Unsupervised Learning



Unsupervised Learning

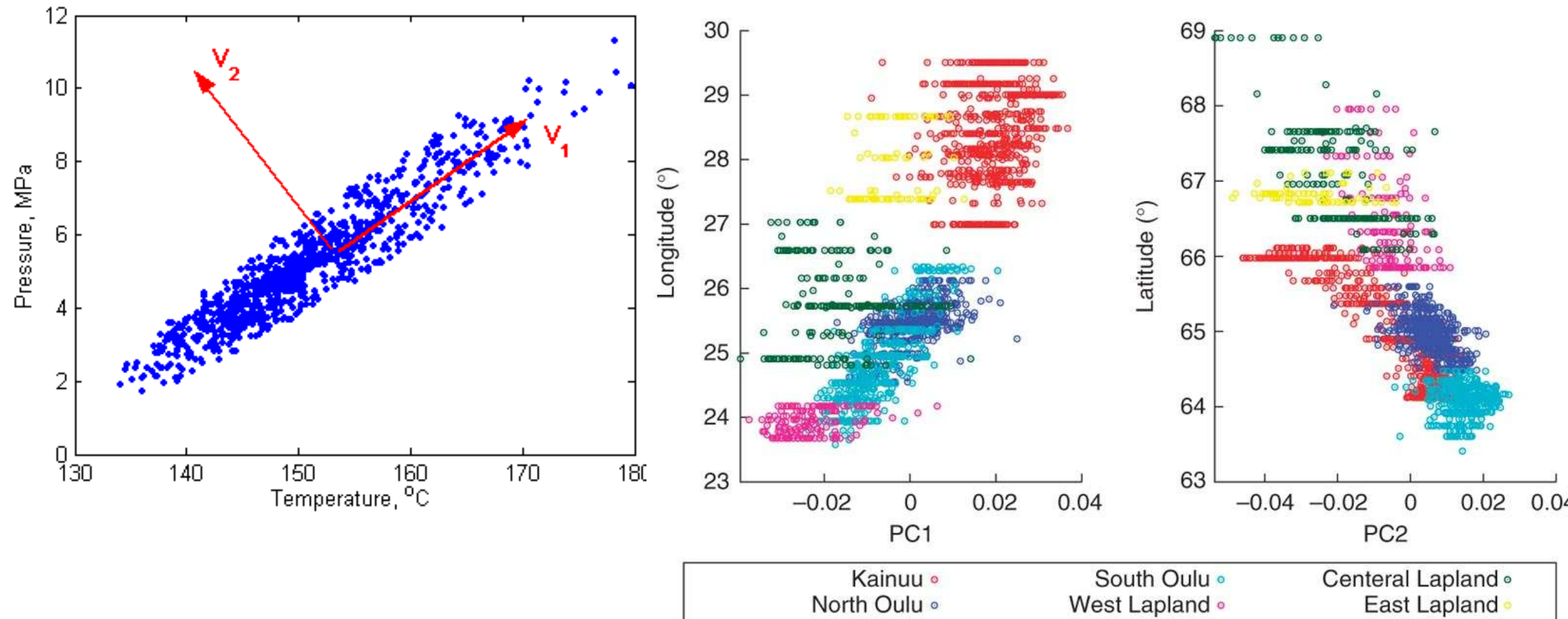
- Given data x , ask a good question ... about x or about model for x
- **Clustering**
Find a set of prototypes representing the data
- **Principal Components**
Find a subspace representing the data
- **Sequence Analysis**
Find a latent causal sequence for observations
 - Sequence Segmentation
 - Hidden 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



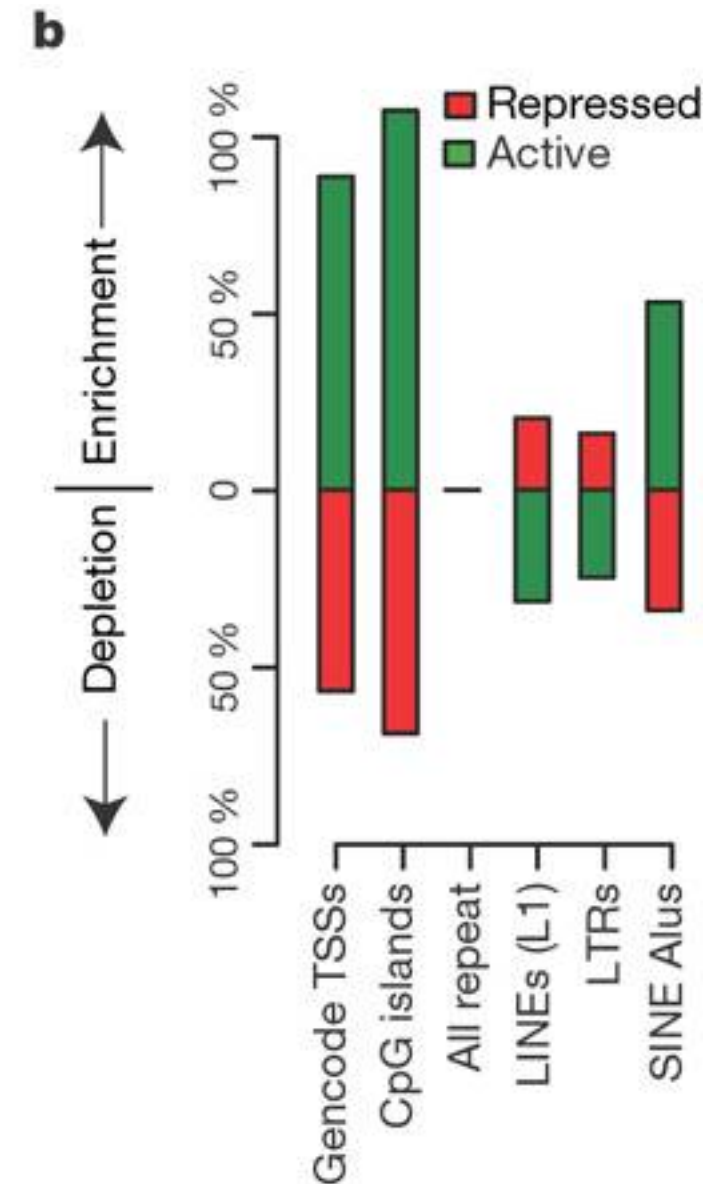
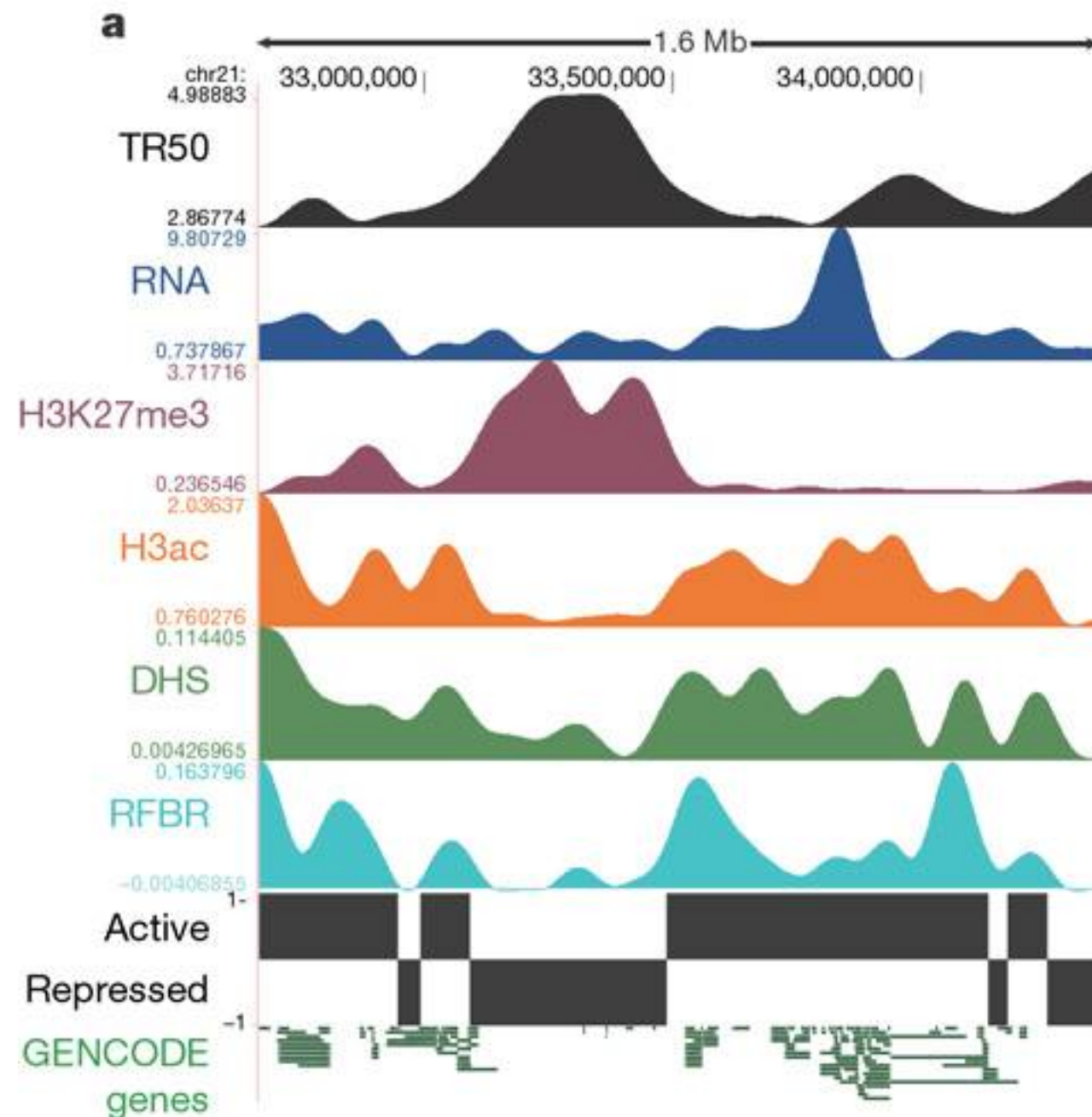
- Documents
- Users
- Webpages
- Diseases
- Pictures
- Vehicles
- ...

Principal Components



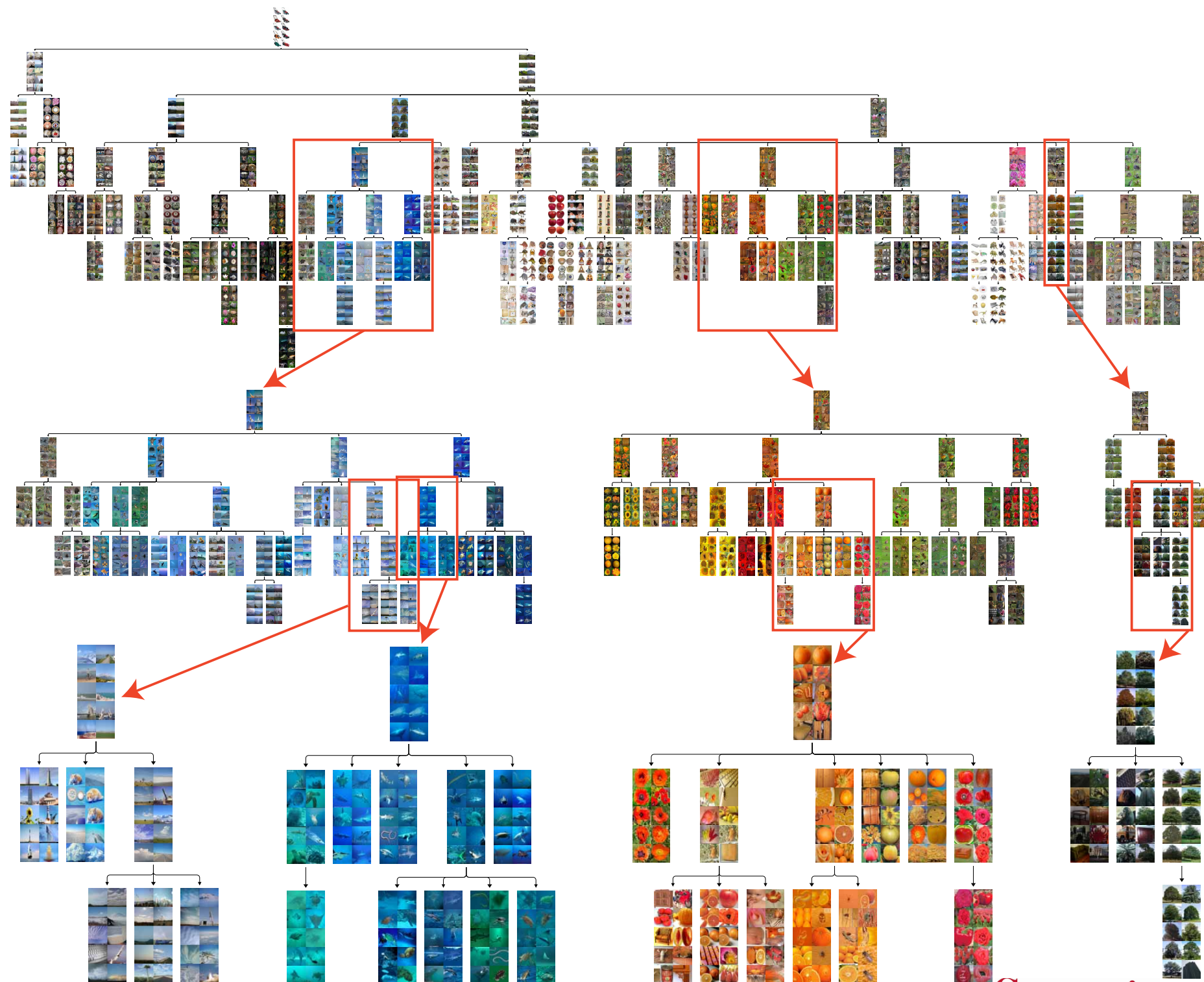
Variance component model to account for sample structure in genome-wide association studies, Nature Genetics 2010

Sequence Analysis

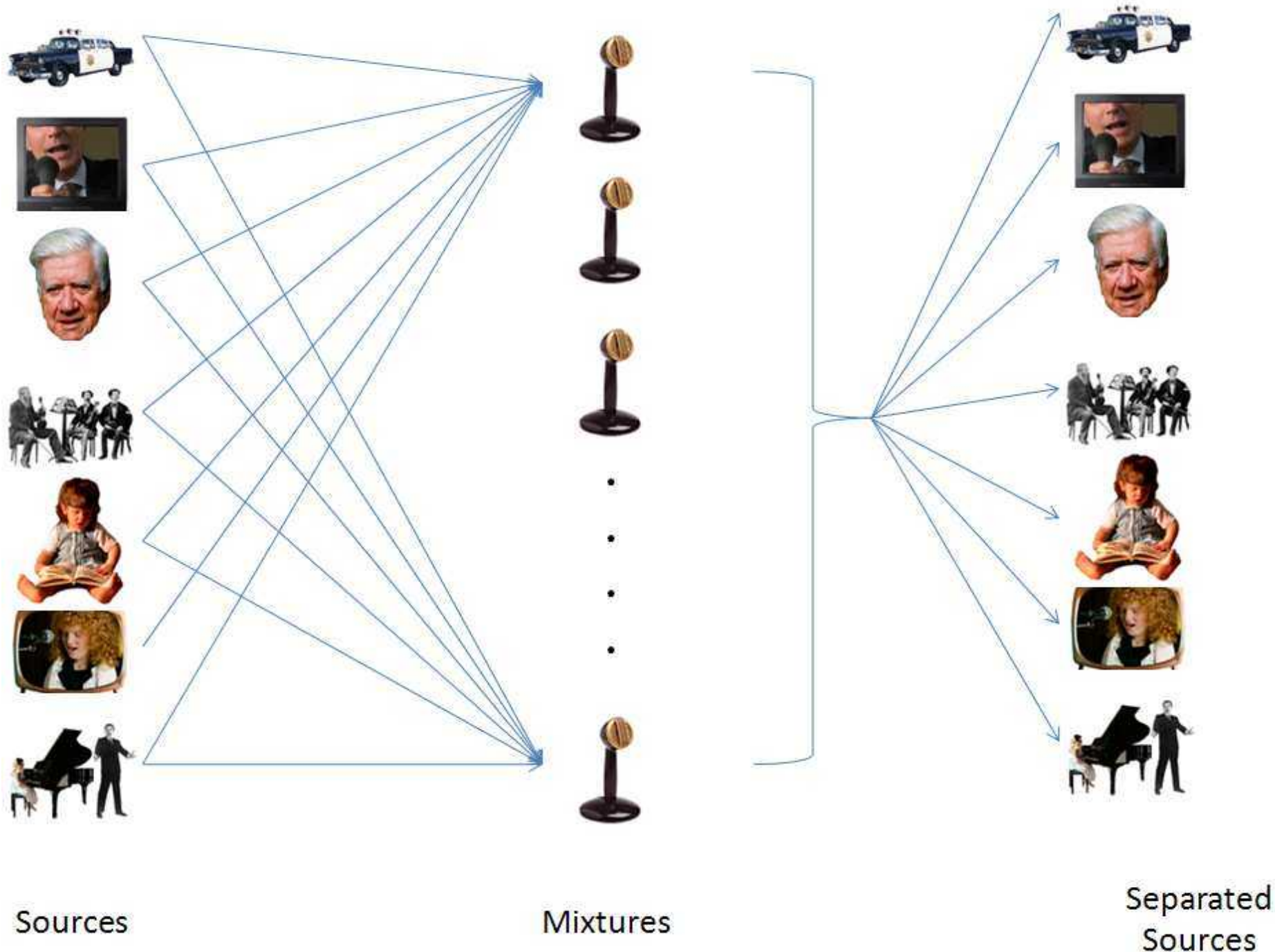


Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project, Nature 2007

Hierarchical Grouping

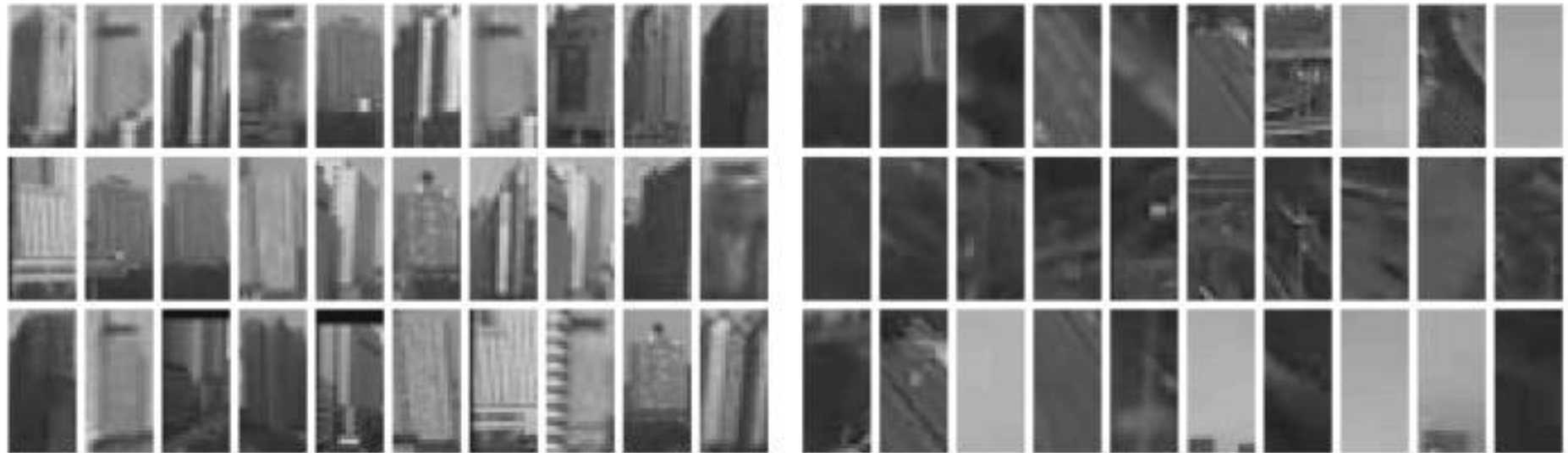


Independent Components



**find them
automatically**

Novelty detection



typical

atypical

1.3 Problem Settings

1 Introduction

Alexander Smola

Introduction to Machine Learning 10-701

<http://alex.smola.org/teaching/10-701-15>



Non-responsive Environment

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)

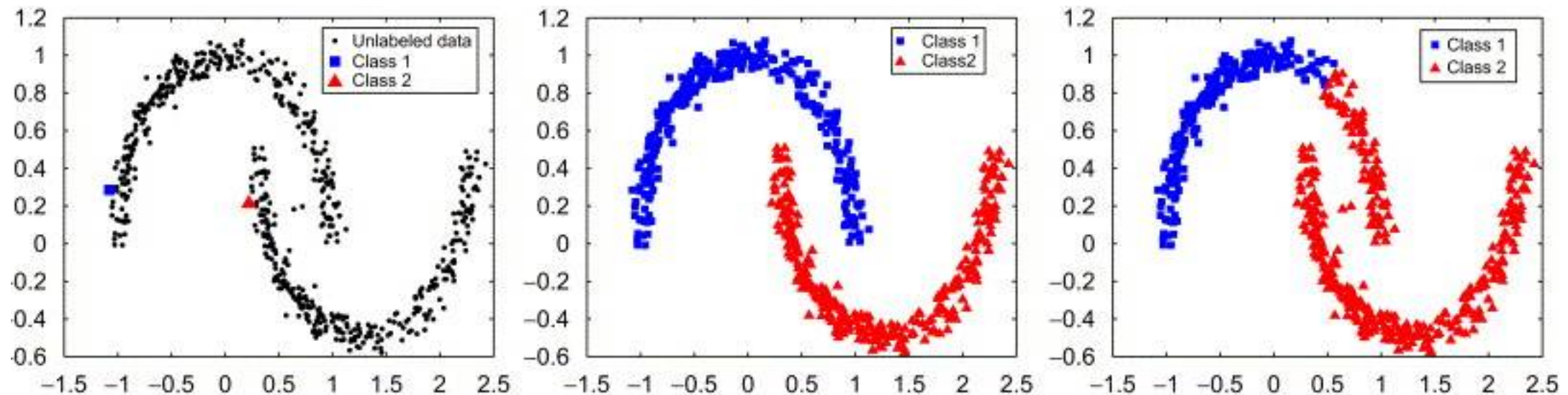
- **Covariate shift**

- 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?**
 - Blood samples from male university students.
 - Use them as healthy reference.
 - Classifier gets 100% accuracy
- What could possibly go 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



Responsive Environment

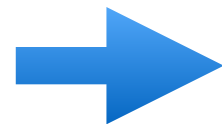
Interaction with Environment

- **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 (stocks, 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** (do your PhD)
Take action, environment responds, take new action
(play chess, drive a car)

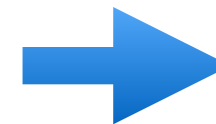
Batch

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	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



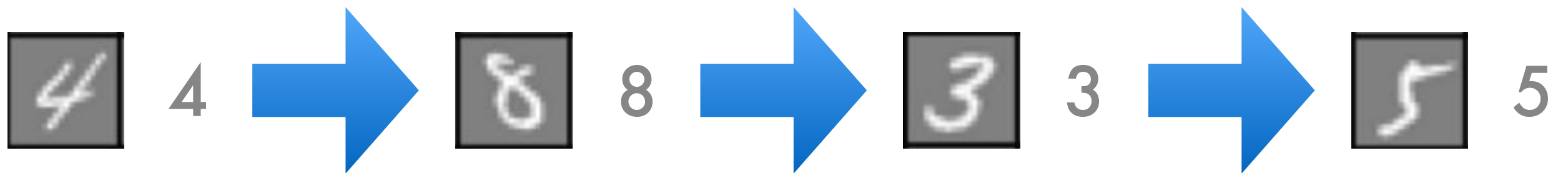
build
model



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



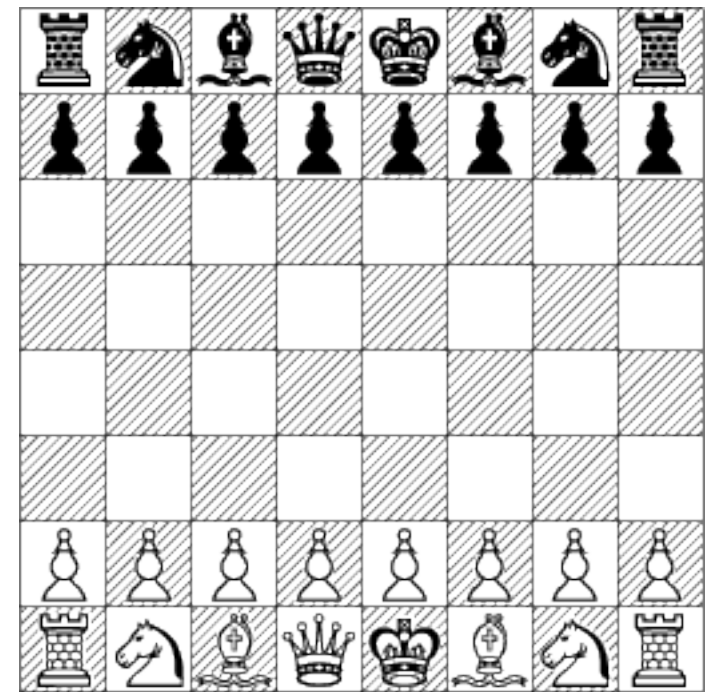
Bandits

- 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 and Generative Models



Discriminative vs. Generative (mainly relevant for supervised models)

- **Discriminative Models**

- Estimate $p(y|x)$ directly
- Often better convergence + simpler solutions

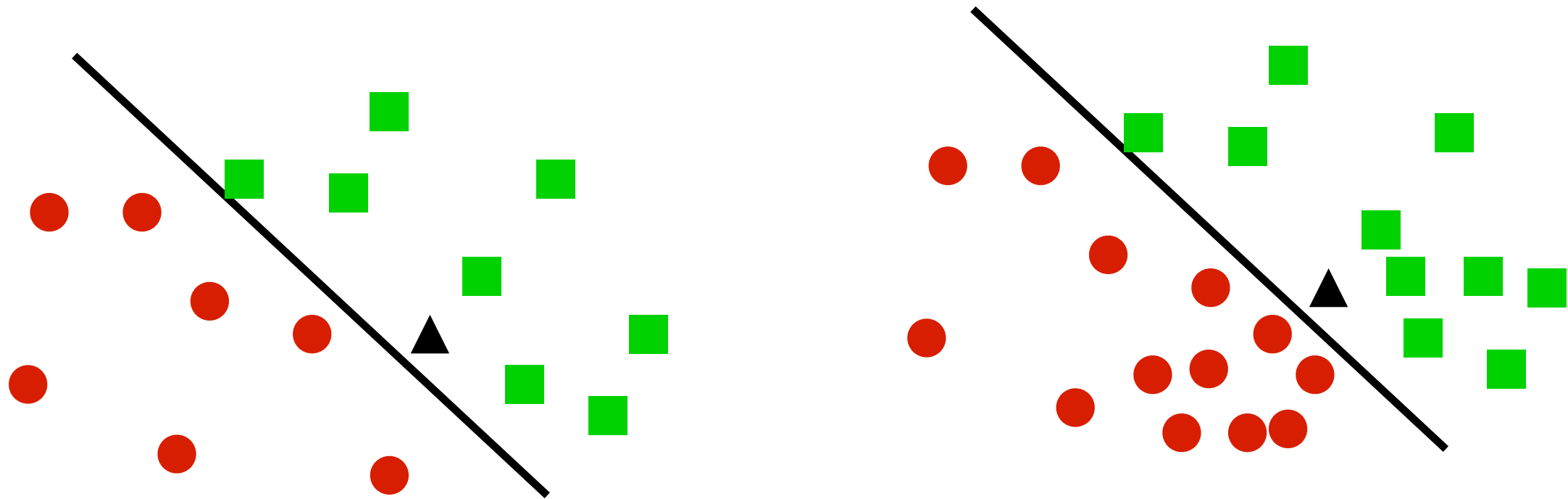
- **Generative models**

- Estimate joint distribution over $p(y, x)$
- Use conditional probability to infer
- Often more intuitive

$$p(y|x) = \frac{p(y, x)}{p(x)}$$

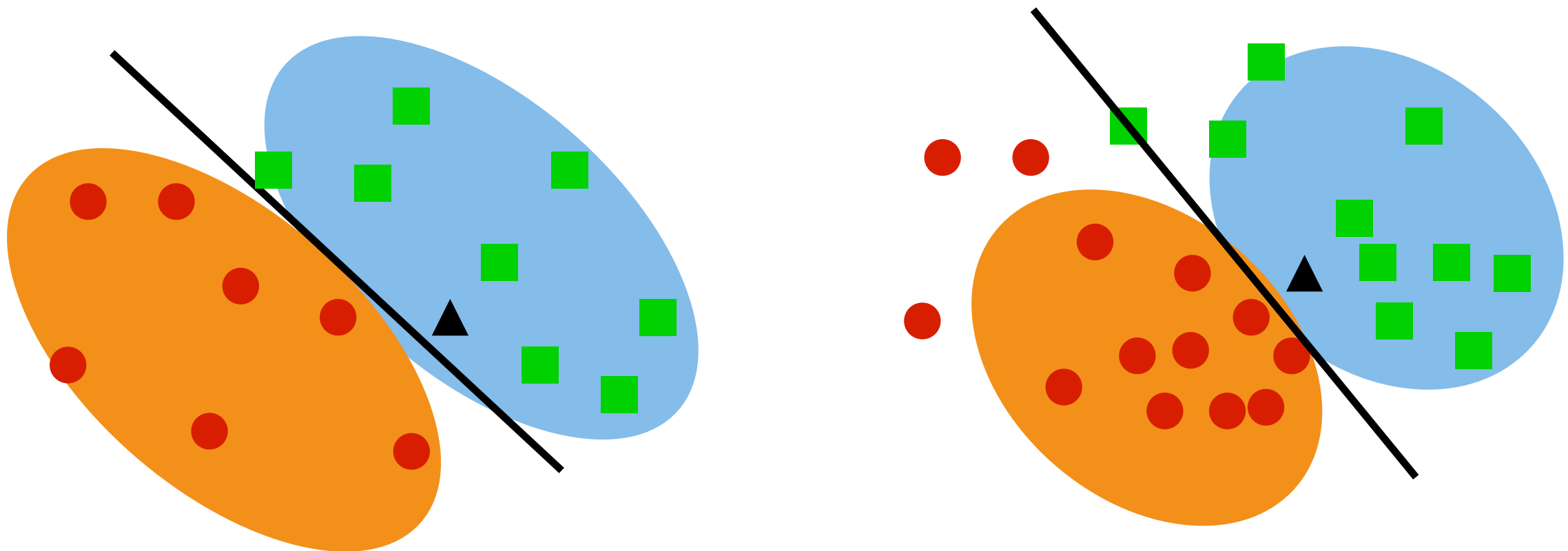
- Easier to add prior knowledge
- Sensitive to mis-specification

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

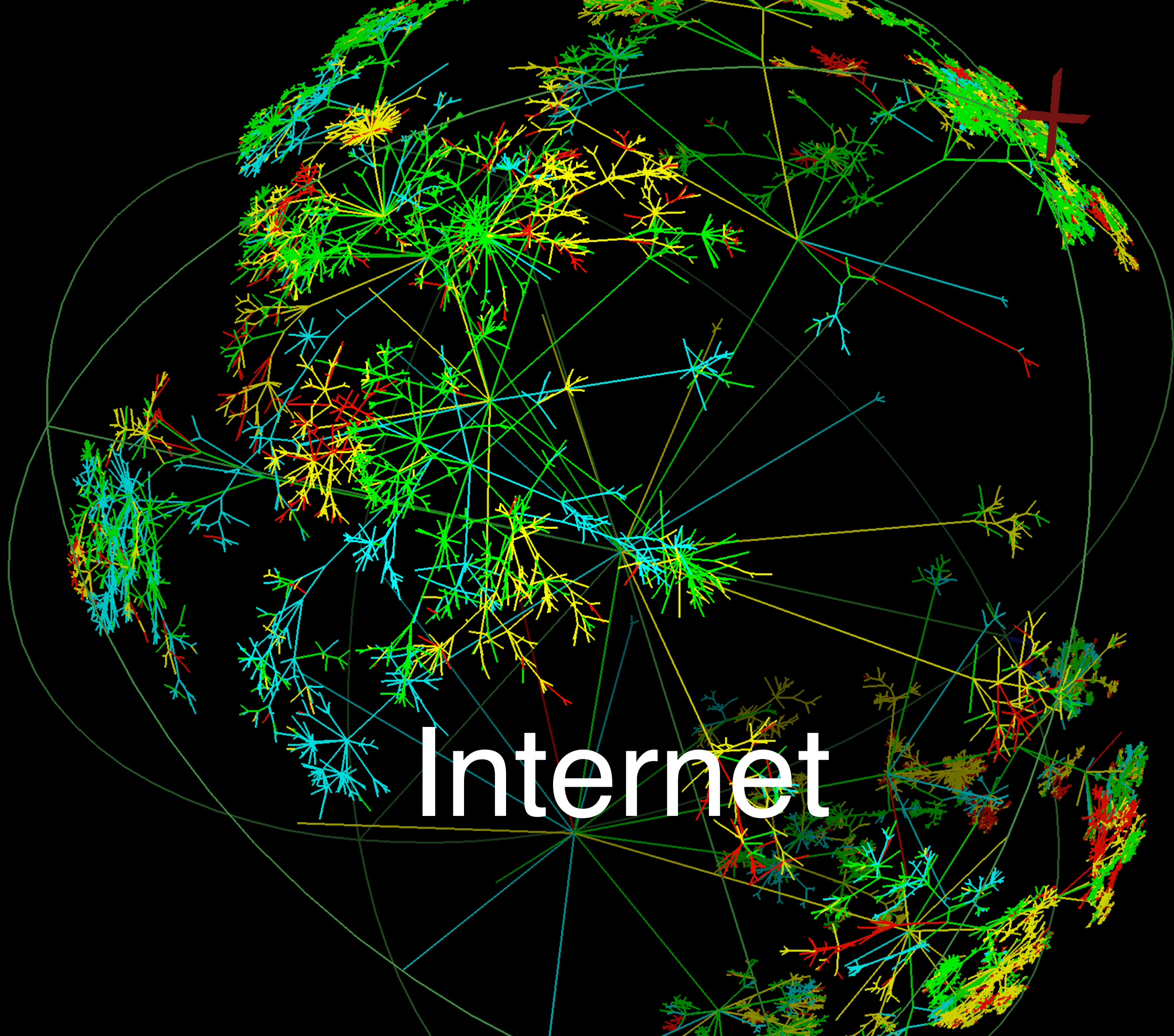
1.4 Data

1 Introduction

Alexander Smola

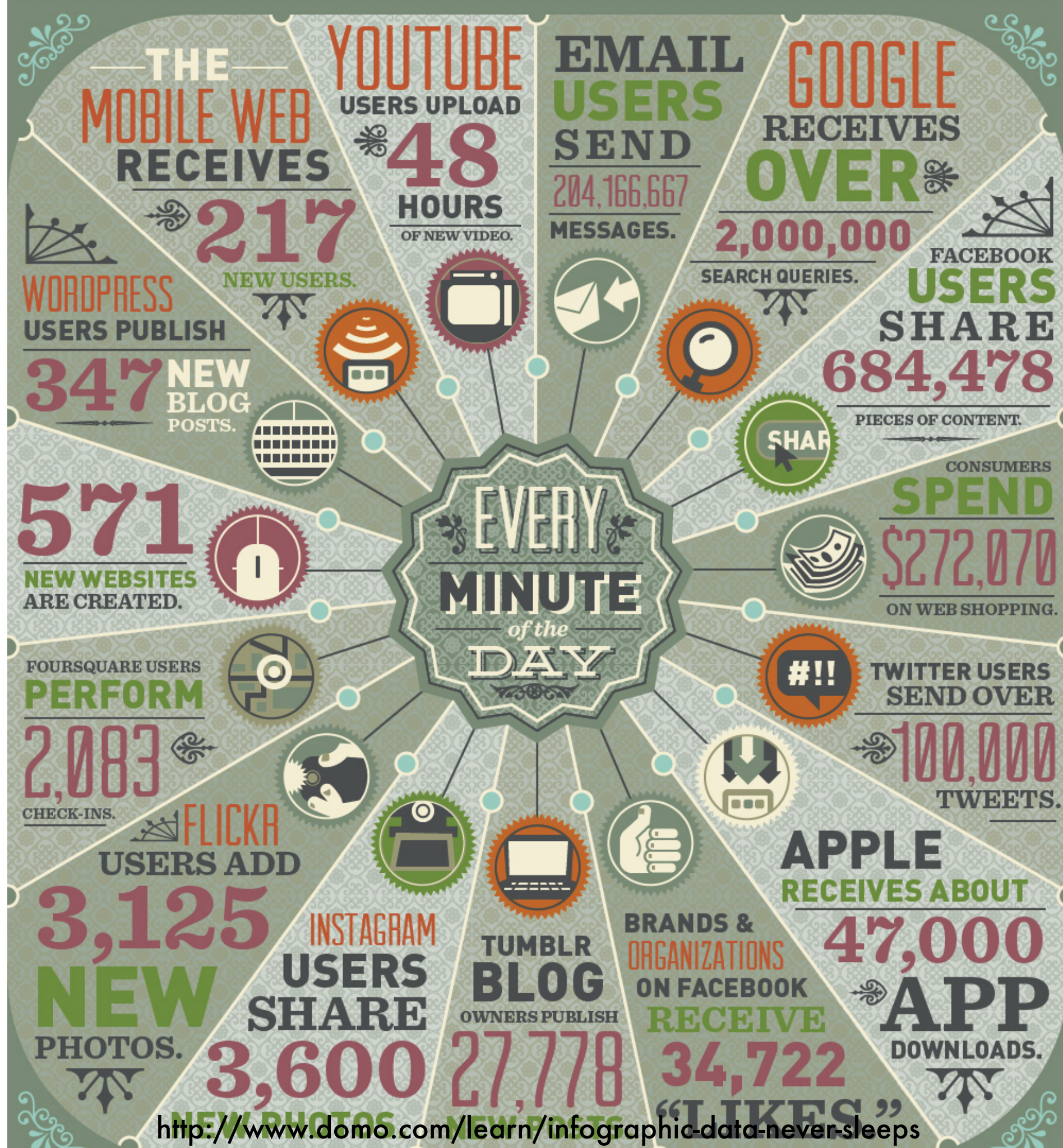
Introduction to Machine Learning 10-701

<http://alex.smola.org/teaching/10-701-15>

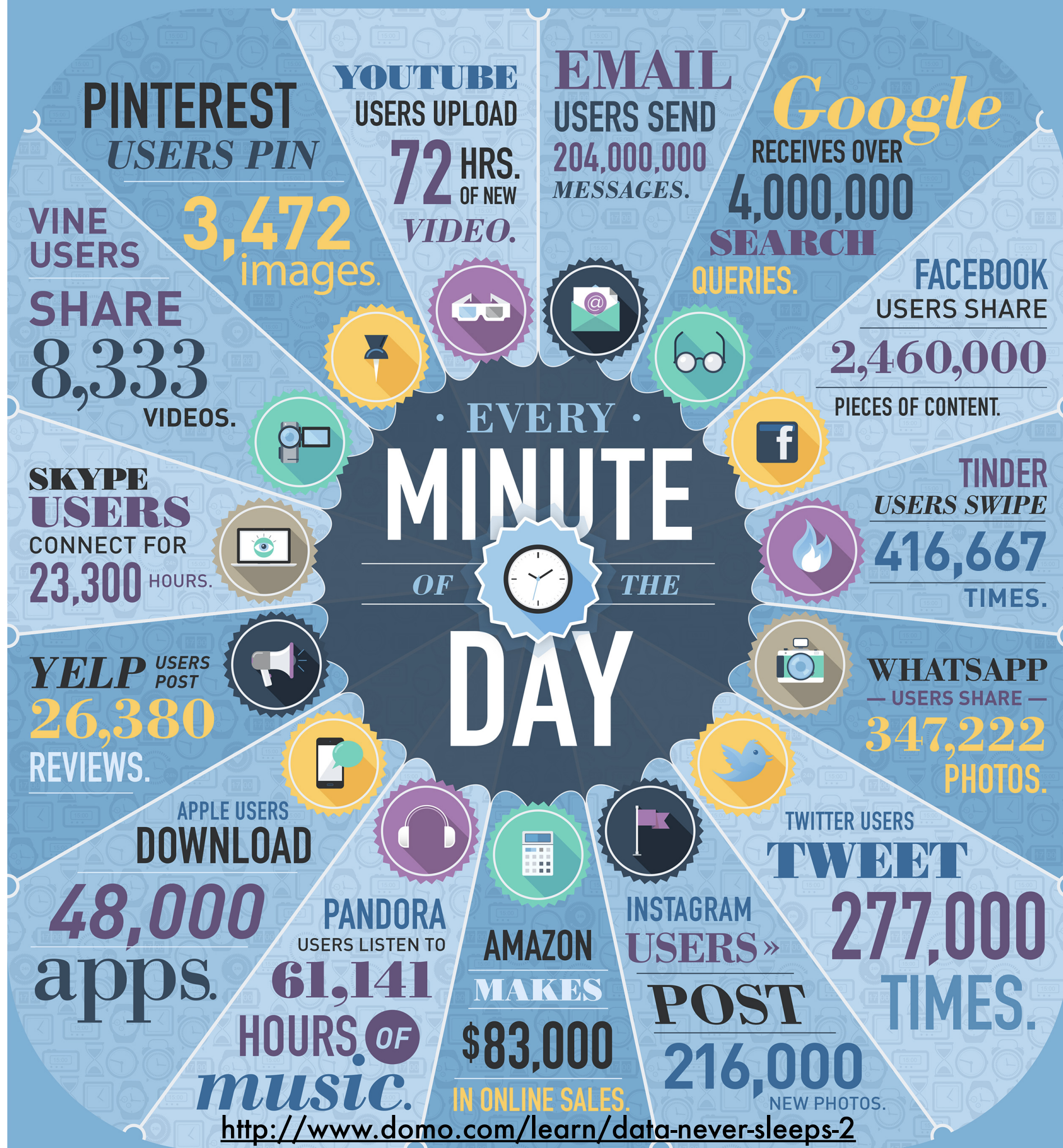


Internet

Data per minute 2012



Data per minute 2014



Computational Advertising

mesothelioma

Web Videos News Images Books More Search tools

About 2,970,000 results (0.21 seconds)

Mesothelioma Compensation
Ad www.nationalmesotheliomaclaims.com/
The Money's Already There. \$30 Billion Asbestos Trust Fund
What Is Mesothelioma? - National Claims Center - Mesothelioma Claims

Mesothelioma Symptoms - Mesothelioma-Answers.org
Ad www.mesothelioma-answers.org/
By Anna Kaplan, M.D. 101 Facts about Mesothelioma.
Asbestos - Treatments - Top Doctors - Free Mesothelioma Book

CA Mesothelioma Resource - californiamesothelioma.com
Ad www.californiamesothelioma.com/ (800) 259-9249
Learn about mesothelioma & receive a free book of helpful answers.
What is Mesothelioma? - Asbestos Exposure in CA - California Legal Rights

Mesothelioma Cancer - Mesothelioma.com
www.mesothelioma.com/mesothelioma/
by Dr. Howard Jack West - Apr 2, 2014 - Mesothelioma is an aggressive cancer affecting the membrane lining ... Between 50 and 70% of all mesotheliomas are of the epithelial variety.
Mesothelioma Symptoms - Mesothelioma Prognosis - Mesothelioma Survival Rate

Mesothelioma - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/Mesothelioma Wikipedia
Mesothelioma (or, more precisely, malignant mesothelioma) is a rare form of cancer that develops from cells of the mesothelium, the protective lining that covers ...
Asbestos - Mesothelium - Paul Kraus - Category:Mesothelioma

Ads

Mesothelioma
www.mesothelioma-attorney-locators.com/
Easily Find Mesothelioma Attorneys.
Locations Across The United States

CA Mesothelioma
www.mesotheliomatreatmentcenters.org/
Mesothelioma? Get the Money you Deserve Fast-Help Filing your Claim

Mesothelioma Compensation
www.mesotheliomaclaimscenter.info/
(877) 456-3935
Mesothelioma? Get Money You Deserve Fast! Get Help with Filing a Claim.

California Mesothelioma
www.mesotheliomaattorney-usa.com/Legal
(888) 707-4525
100% Free Mesothelioma Legal Help!
\$30 Billion Trust Fund Available.

Mesothelioma
meso.lawyers.local.alotresults.com/
Seasoned Lawyers in your Area.
In your Local Lawyer Listings!

sponsored
search picks
position of
ad using

$$p(\text{click}|\text{ad}) \cdot \text{bid}(\text{ad})$$

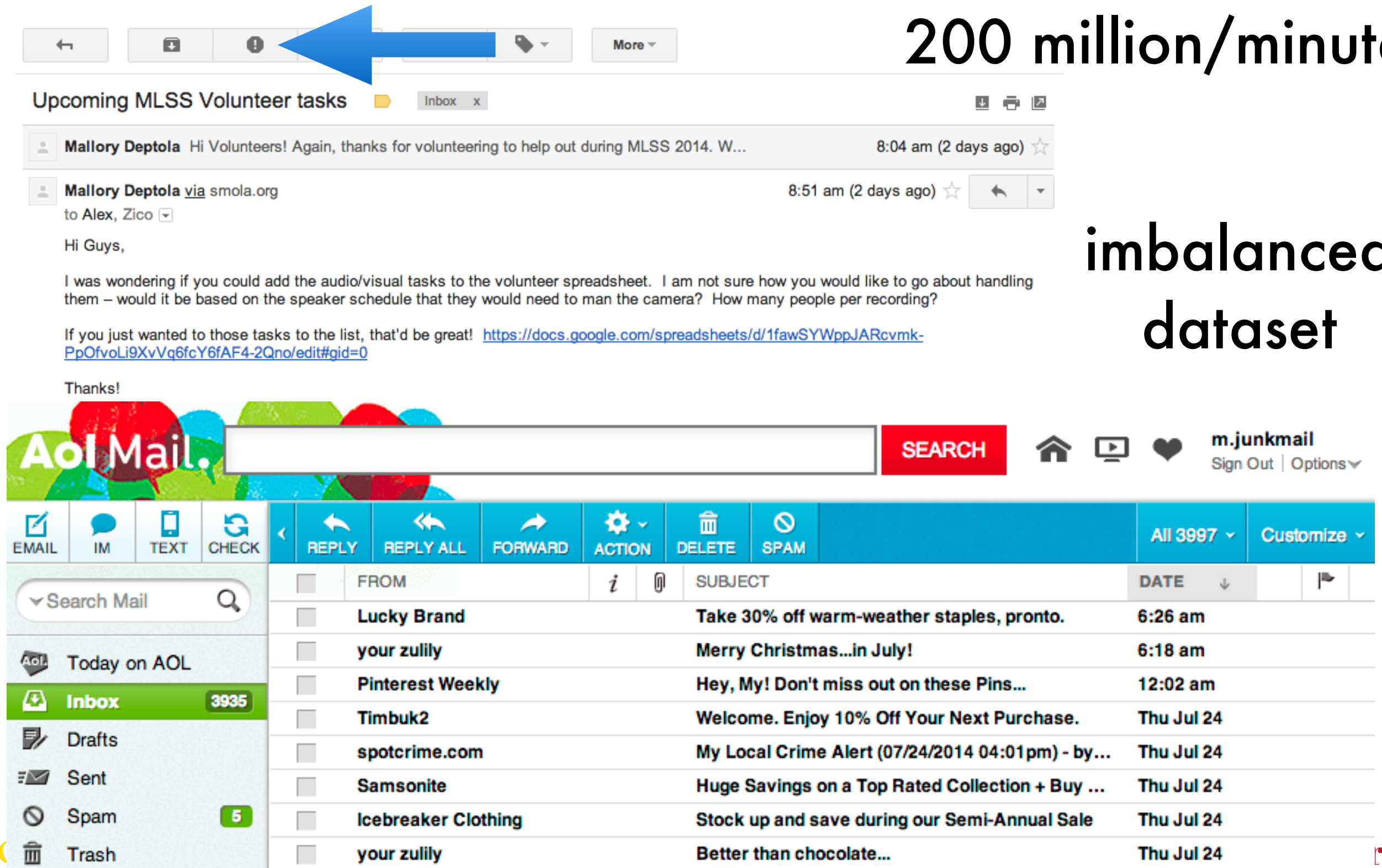
estimate it

4 million/minute

Spam filtering

200 million/minute

imbalanced
dataset



The screenshot shows an AOL Mail interface. At the top, a blue arrow points to the 'Spam' button in the toolbar. The inbox list contains several emails, including legitimate ones from 'Mallory Deptola' and several spam emails from 'Lucky Brand', 'your zullily', 'Pinterest Weekly', 'Timbuk2', 'spotcrime.com', 'Samsonite', 'Icebreaker Clothing', and 'your zullily'. The 'Spam' button is highlighted in the toolbar.

Upcoming MLSS Volunteer tasks

Mallory Deptola Hi Volunteers! Again, thanks for volunteering to help out during MLSS 2014. W... 8:04 am (2 days ago) ☆

Mallory Deptola via smola.org 8:51 am (2 days ago) ☆

to Alex, Zico ▾

Hi Guys,

I was wondering if you could add the audio/visual tasks to the volunteer spreadsheet. I am not sure how you would like to go about handling them – would it be based on the speaker schedule that they would need to man the camera? How many people per recording?

If you just wanted to those tasks to the list, that'd be great! <https://docs.google.com/spreadsheets/d/1fawSYWppJARcvmk-PpOfvoLi9XvVq6fcY6fAF4-2Qno/edit#gid=0>

Thanks!

AOL Mail. **SEARCH** **m.junkmail** Sign Out | Options ▾

EMAIL IM TEXT CHECK < REPLY REPLY ALL FORWARD ACTION DELETE SPAM All 3997 ▾ Customize ▾

<input type="checkbox"/>	FROM	SUBJECT	DATE
<input type="checkbox"/>	Lucky Brand	Take 30% off warm-weather staples, pronto.	6:26 am
<input type="checkbox"/>	your zullily	Merry Christmas...in July!	6:18 am
<input type="checkbox"/>	Pinterest Weekly	Hey, My! Don't miss out on these Pins...	12:02 am
<input type="checkbox"/>	Timbuk2	Welcome. Enjoy 10% Off Your Next Purchase.	Thu Jul 24
<input type="checkbox"/>	spotcrime.com	My Local Crime Alert (07/24/2014 04:01 pm) - by...	Thu Jul 24
<input type="checkbox"/>	Samsonite	Huge Savings on a Top Rated Collection + Buy ...	Thu Jul 24
<input type="checkbox"/>	Icebreaker Clothing	Stock up and save during our Semi-Annual Sale	Thu Jul 24
<input type="checkbox"/>	your zullily	Better than chocolate...	Thu Jul 24

Today on AOL

- Inbox** 3935
- Drafts
- Sent
- Spam 5
- Trash

Spam filtering

200 million/minute

imbalanced
dataset

The screenshot shows an AOL Mail interface. At the top, a blue arrow points to a toolbar containing icons for back, forward, and other actions. Below this, an email from Mallory Deptola is displayed. A red arrow points from the 'SEARCH' button to the 'm.junkmail' link in the top right corner. On the left sidebar, the 'Inbox' is highlighted with 3935 items, and the 'Spam' folder is highlighted with 5 items. Blue arrows point from the 'Inbox' and 'Spam' folders to the email list. The email list contains several promotional and commercial messages, such as 'Lucky Brand', 'your zully', and 'Interest Weekly'. A red arrow points from the 'SEARCH' button to the 'm.junkmail' link.

FROM	SUBJECT	DATE
Lucky Brand	Take 30% off warm-weather staples, pronto.	6:26 am
your zully	Merry Christmas...in July!	6:18 am
Interest Weekly	Hey, My! Don't miss out on these Pins...	12:02 am
mbuk2	Welcome. Enjoy 10% Off Your Next Purchase.	Thu Jul 24
spotcrime.com	My Local Crime Alert (07/24/2014 04:01 pm) - by...	Thu Jul 24
amsonite	Huge Savings on a Top Rated Collection + Buy ...	Thu Jul 24
ing	Stock up and save during our Semi-Annual Sale	Thu Jul 24
your zully	Better than chocolate...	Thu Jul 24

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

Carnegie Mellon University

Data - User generated content

- Webpages (content, graph)
- Clicks (ad, page, social)
- Users (OpenID, FB Connect)
- e-mails (Hotmail, Y!Mail, Gm...)
- 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



DISQUS



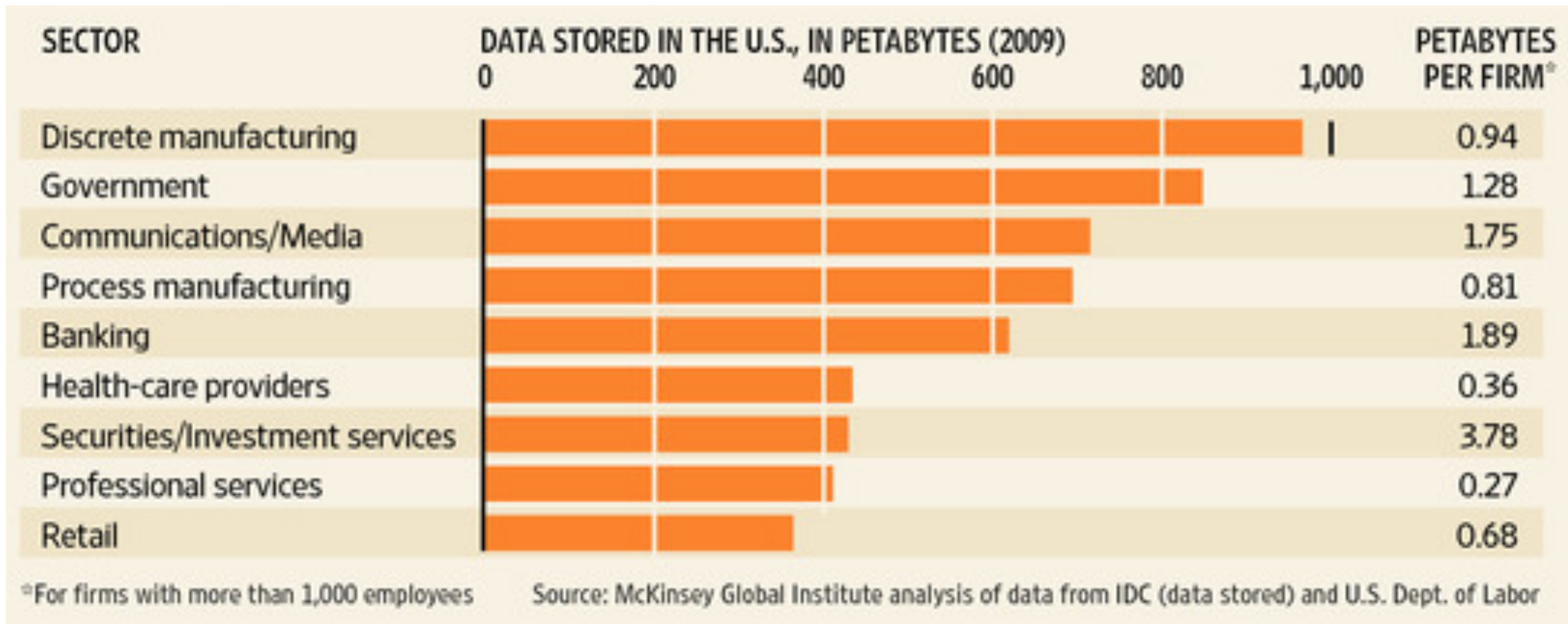
You Tube

yelp

> 1 B images, 40h video/minute

Carnegie Mellon University

Big Data



we need Big Learning

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

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)

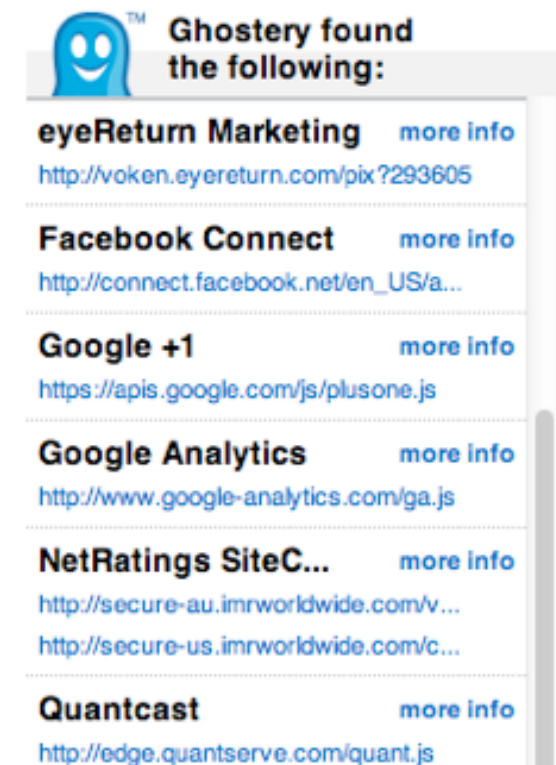


> 1 B texts

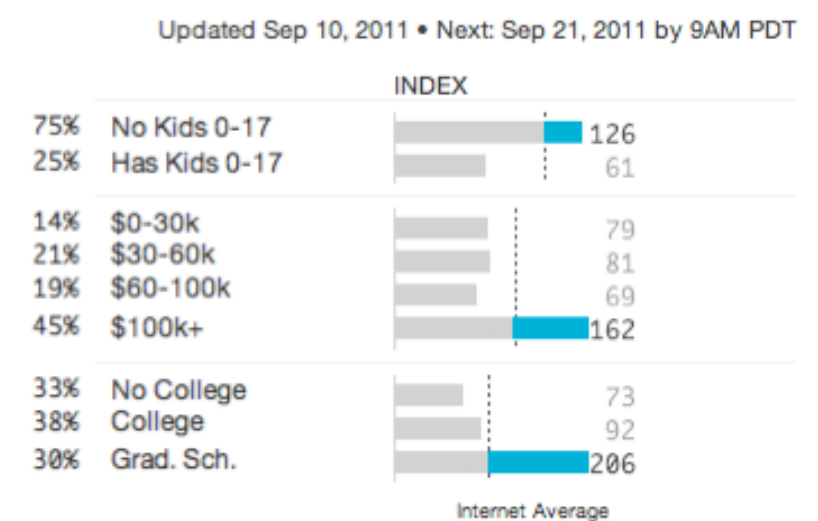
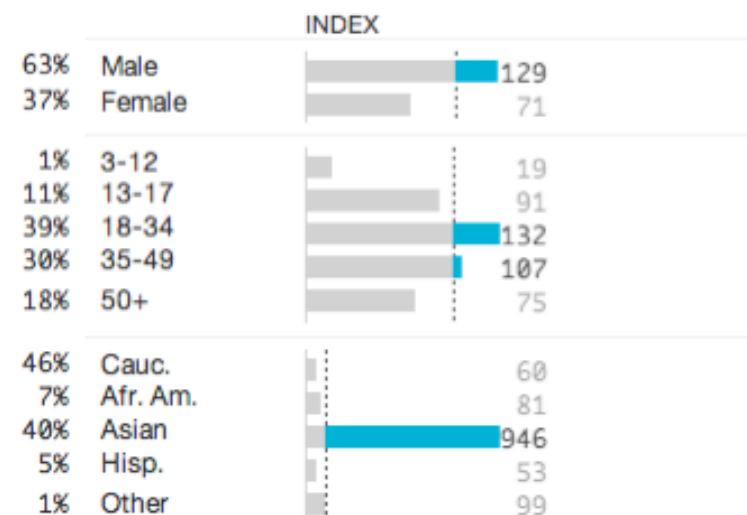
impossible without NDA

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 ?



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

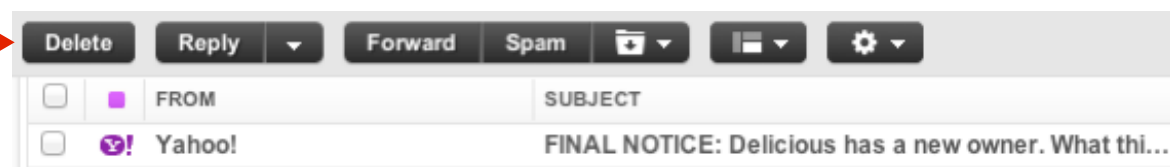
- Ads



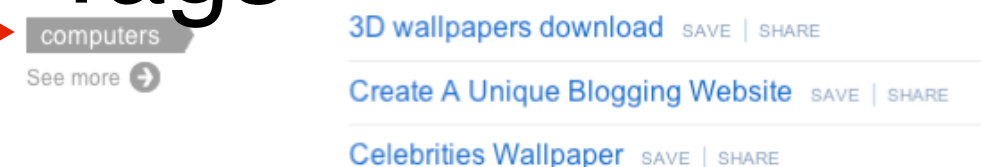
- Click feedback



- Emails



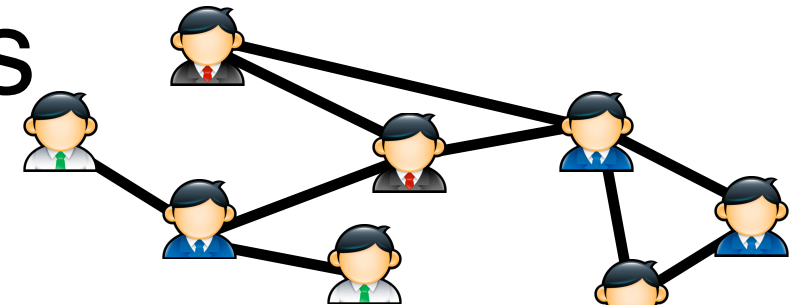
- Tags



- Editorial data is very expensive! Do not use!

no Labels

- Graphs



- Document collections



- Email/IM/Discussions



- Query stream





Medicine

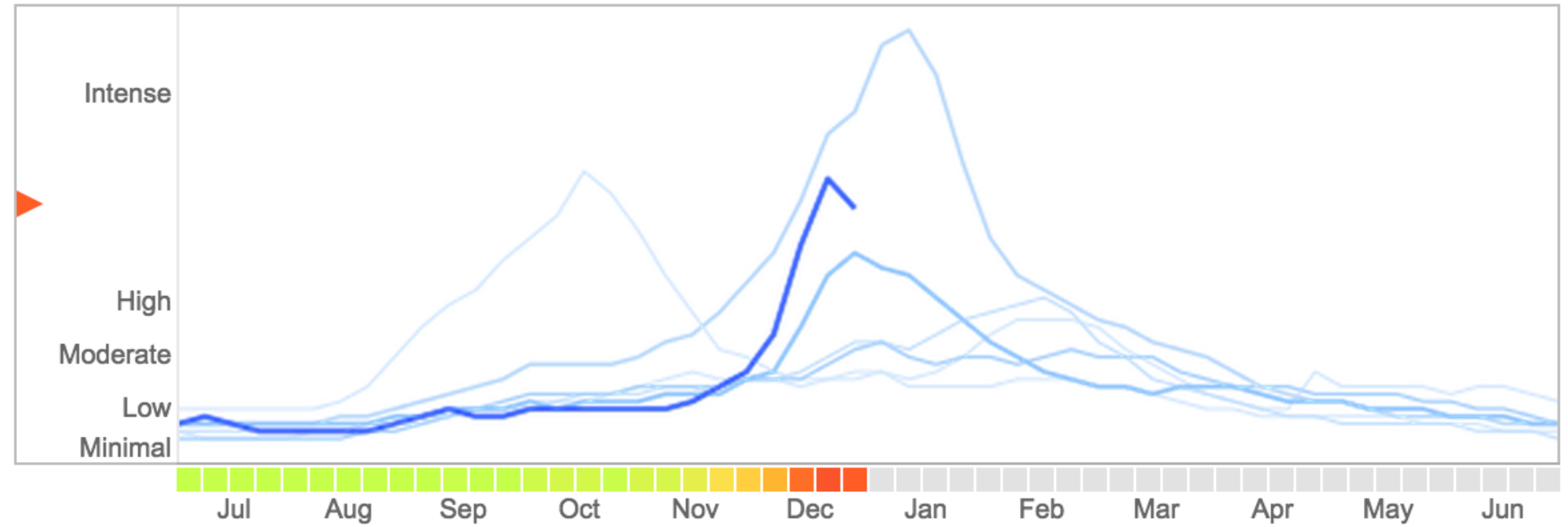
Healthcare

- Many hospital records
 - Messy and incomplete data
 - Different schemas, merging can be difficult
- Side effects of drugs
 - Actually observed (for approved drugs)
 - Interactions with other drugs
- Diagnosis / survival prediction
 - Personalized cancer medication

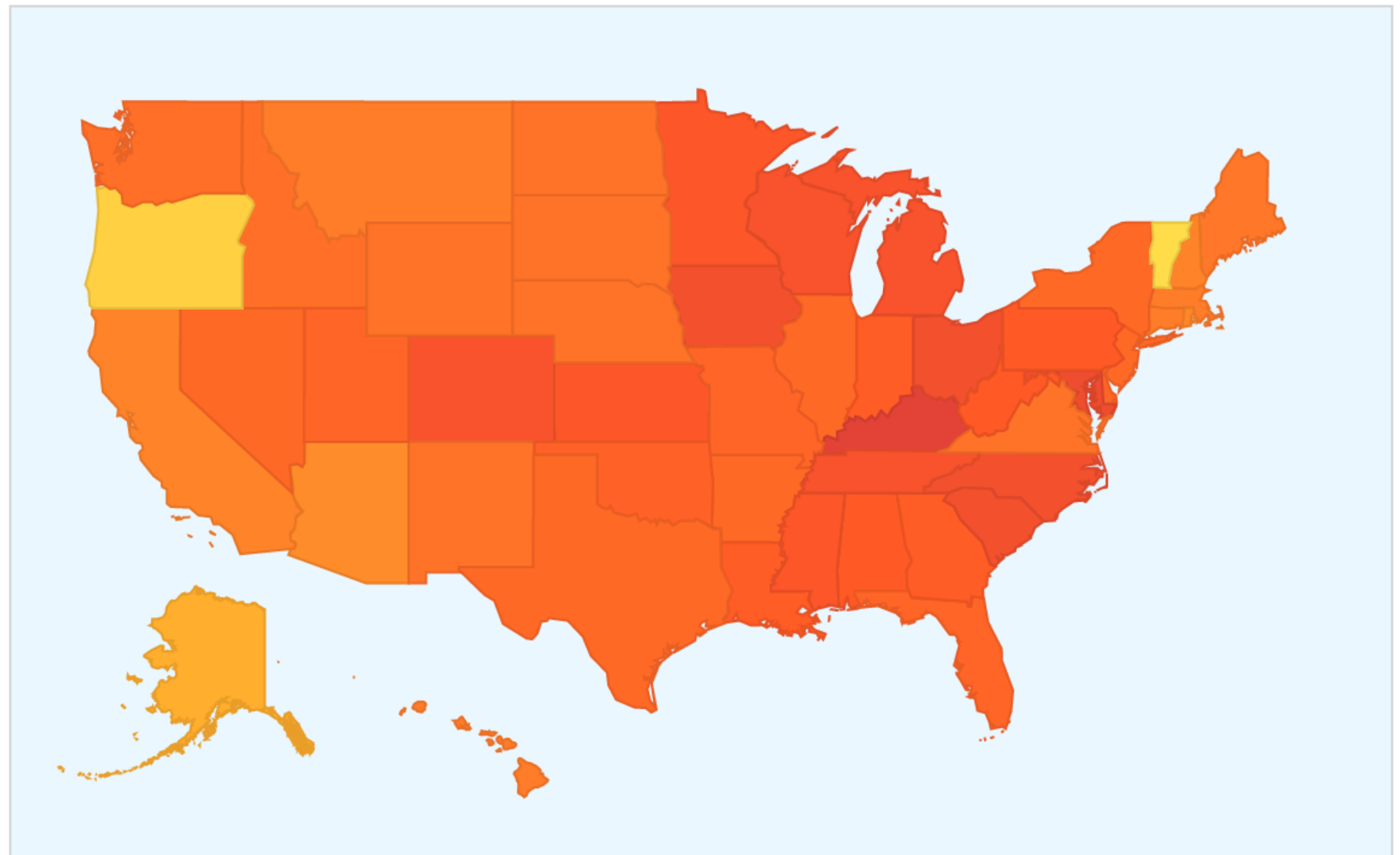
Flu trends (from search queries)

National

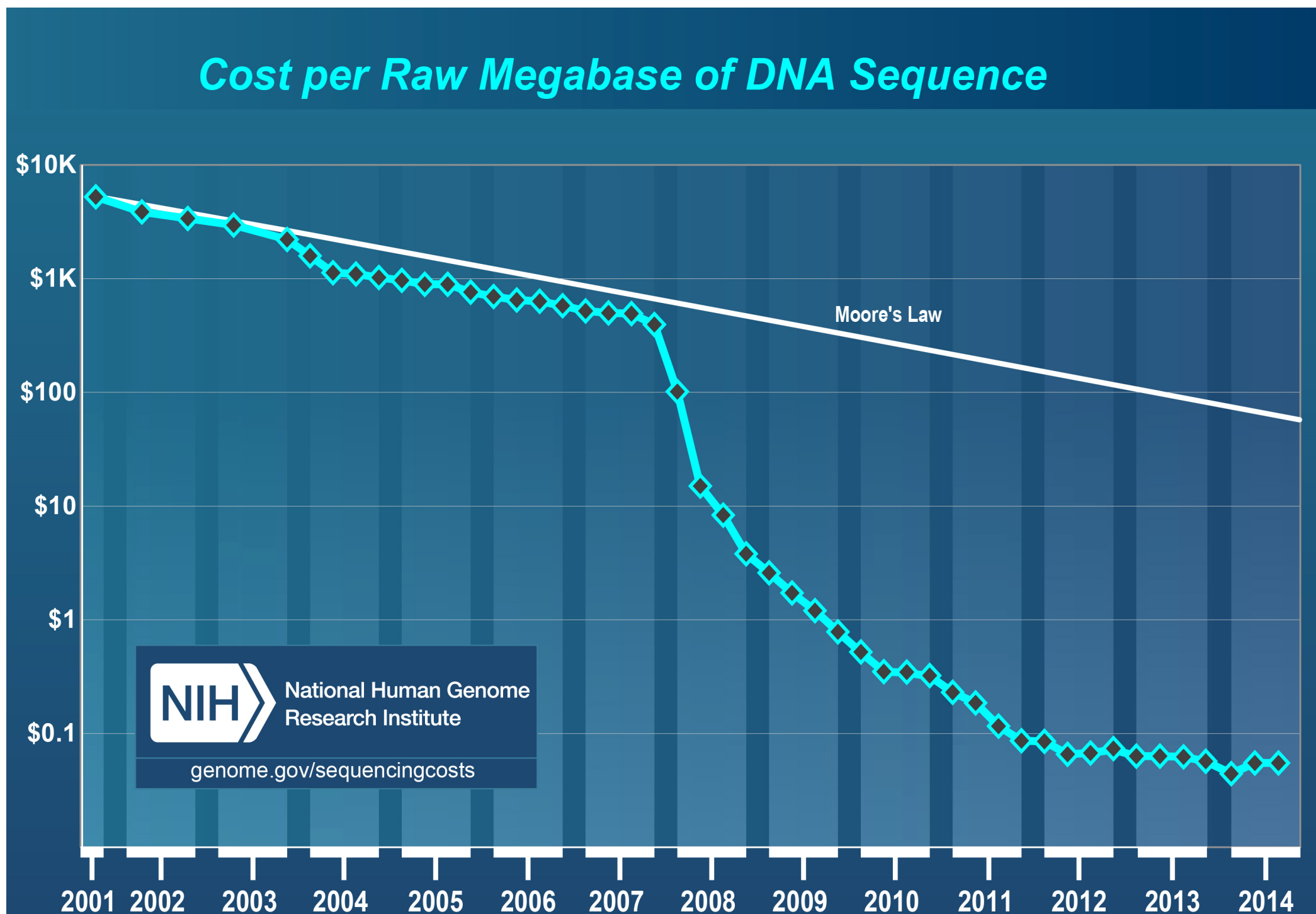
● 2014-2015 ● Past years ▼



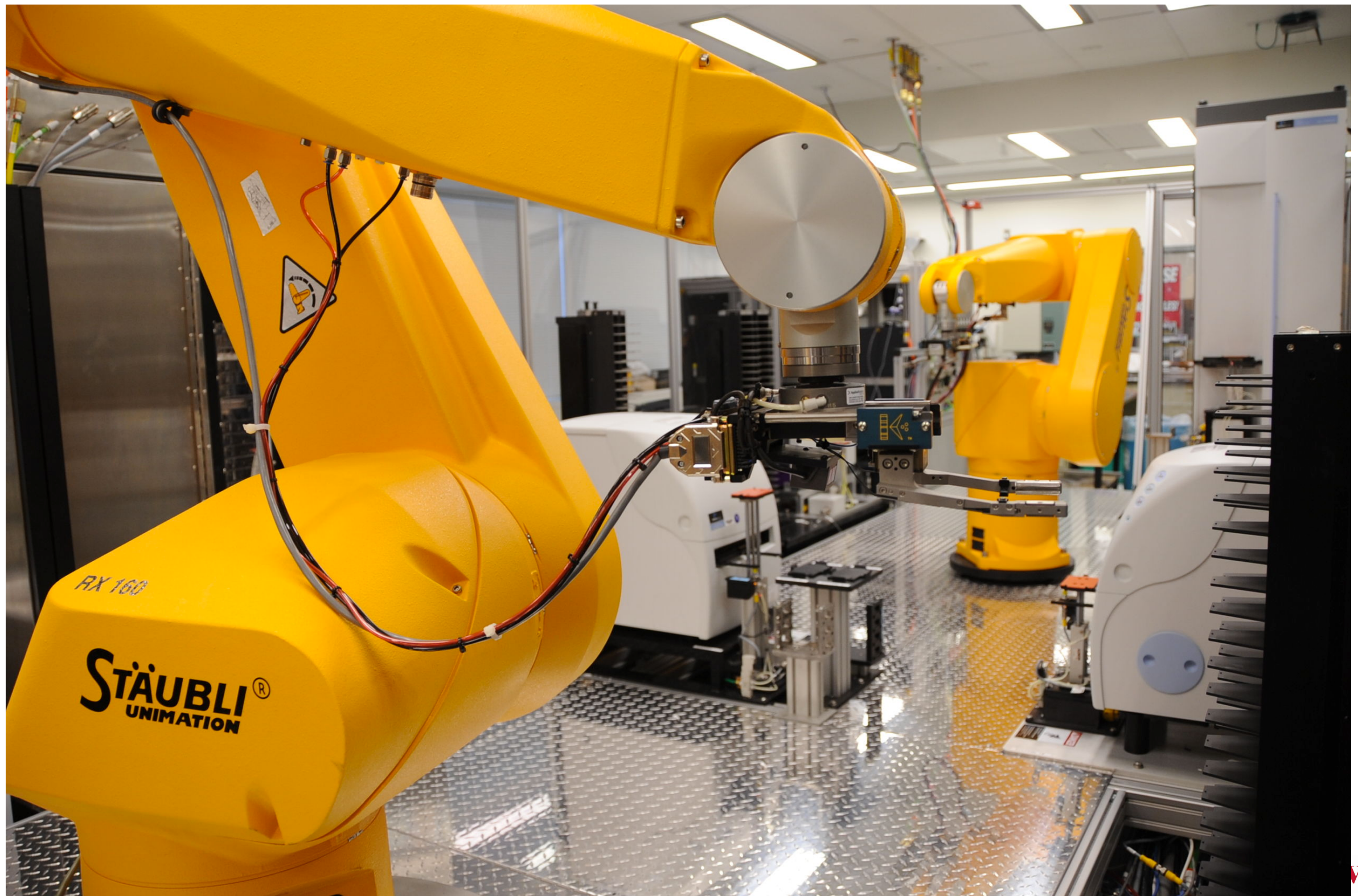
States | [Cities](#) (Experimental)



DNA Sequencing



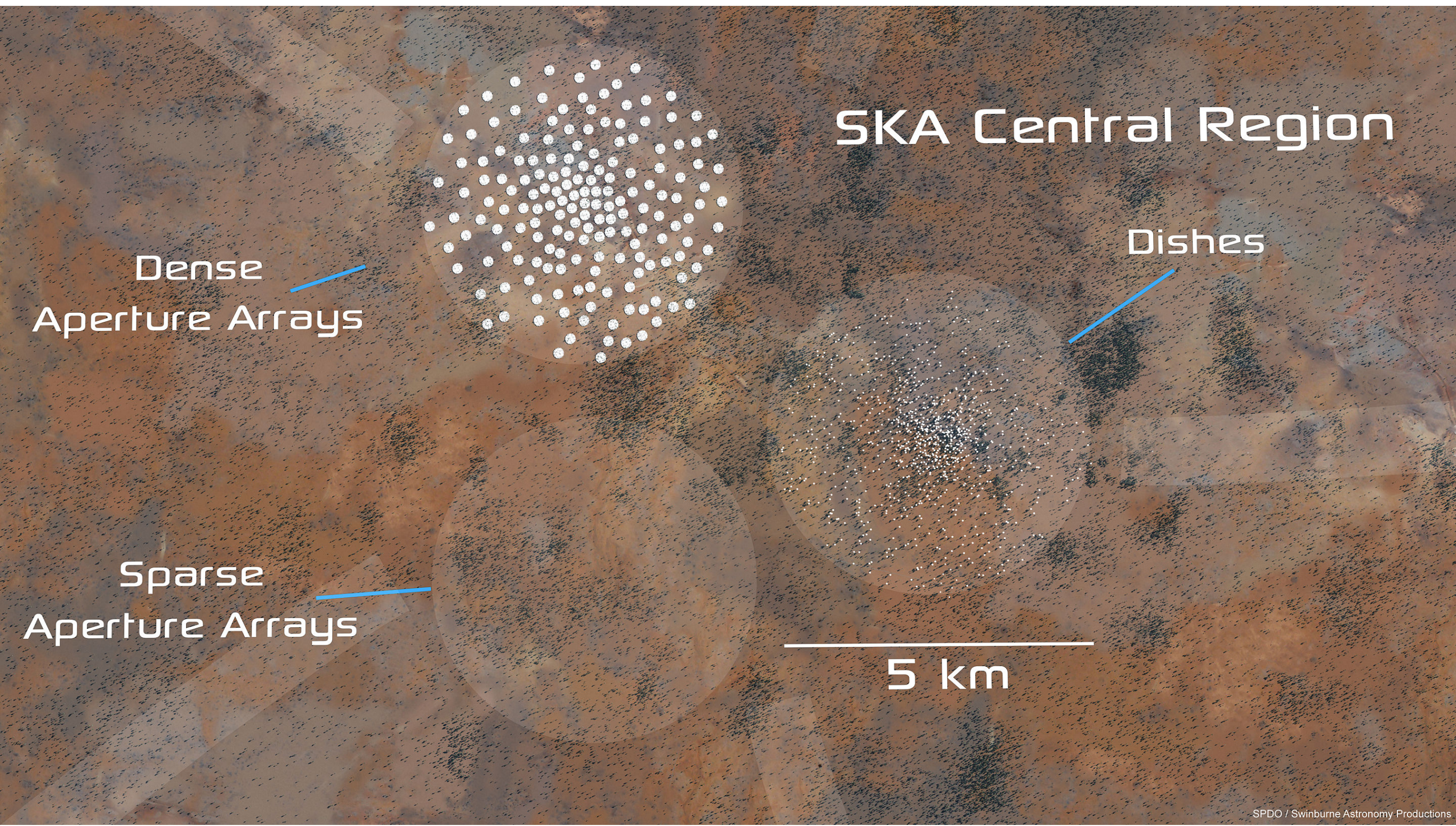
High throughput screening





Physics

Square Kilometer Array



SKA Central Region

Dense
Aperture Arrays

Dishes

Sparse
Aperture Arrays

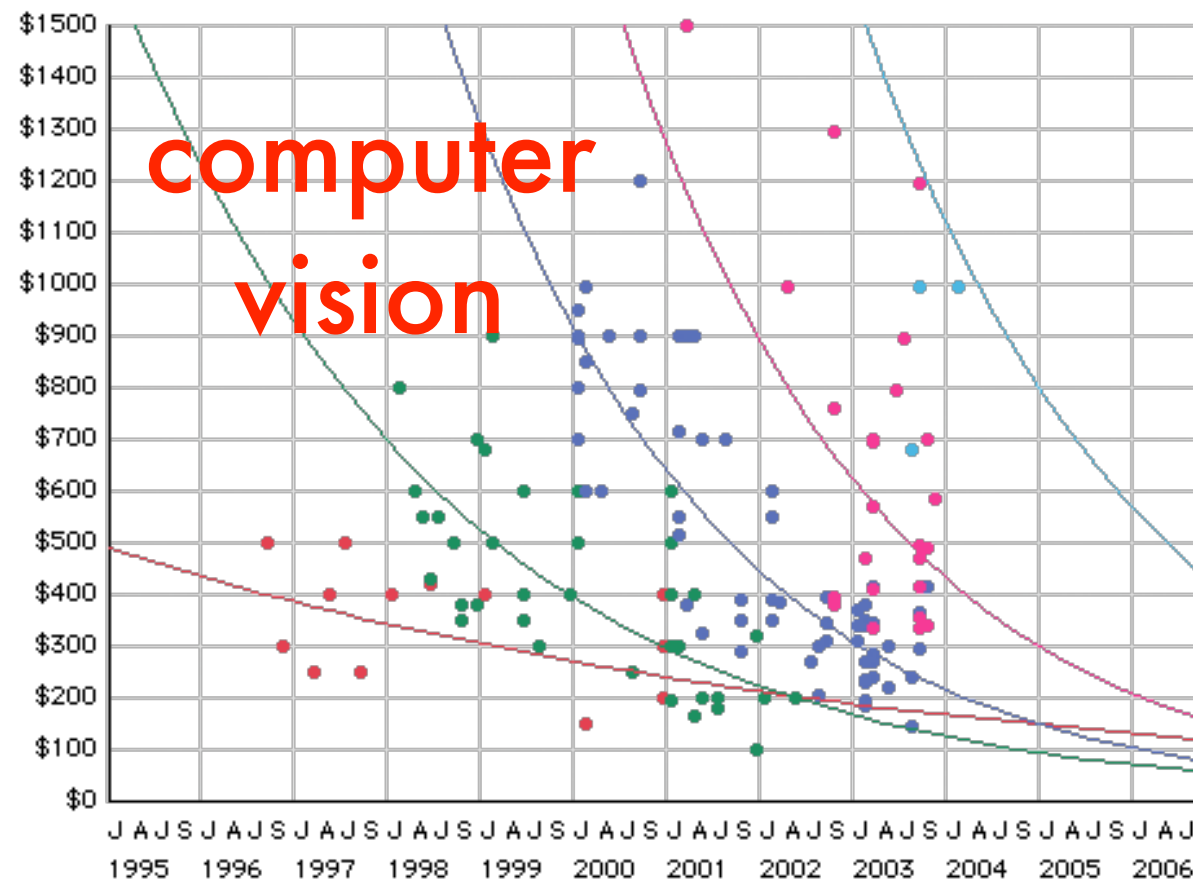
5 km

More sources

- Satellites
- High energy physics
CERN is essentially a giant sensor array
- Geophysics
Find locations of oil based on seismic readings
(talk to me if you want to work on this)
- Semiconductor fabs
100s of steps but want small failure rate
$$1 - \delta = (1 - \epsilon)^n \text{ hence } \epsilon = 1 - (1 - \delta)^{\frac{1}{n}}$$

1% error for 100 steps yields 1-1/e success

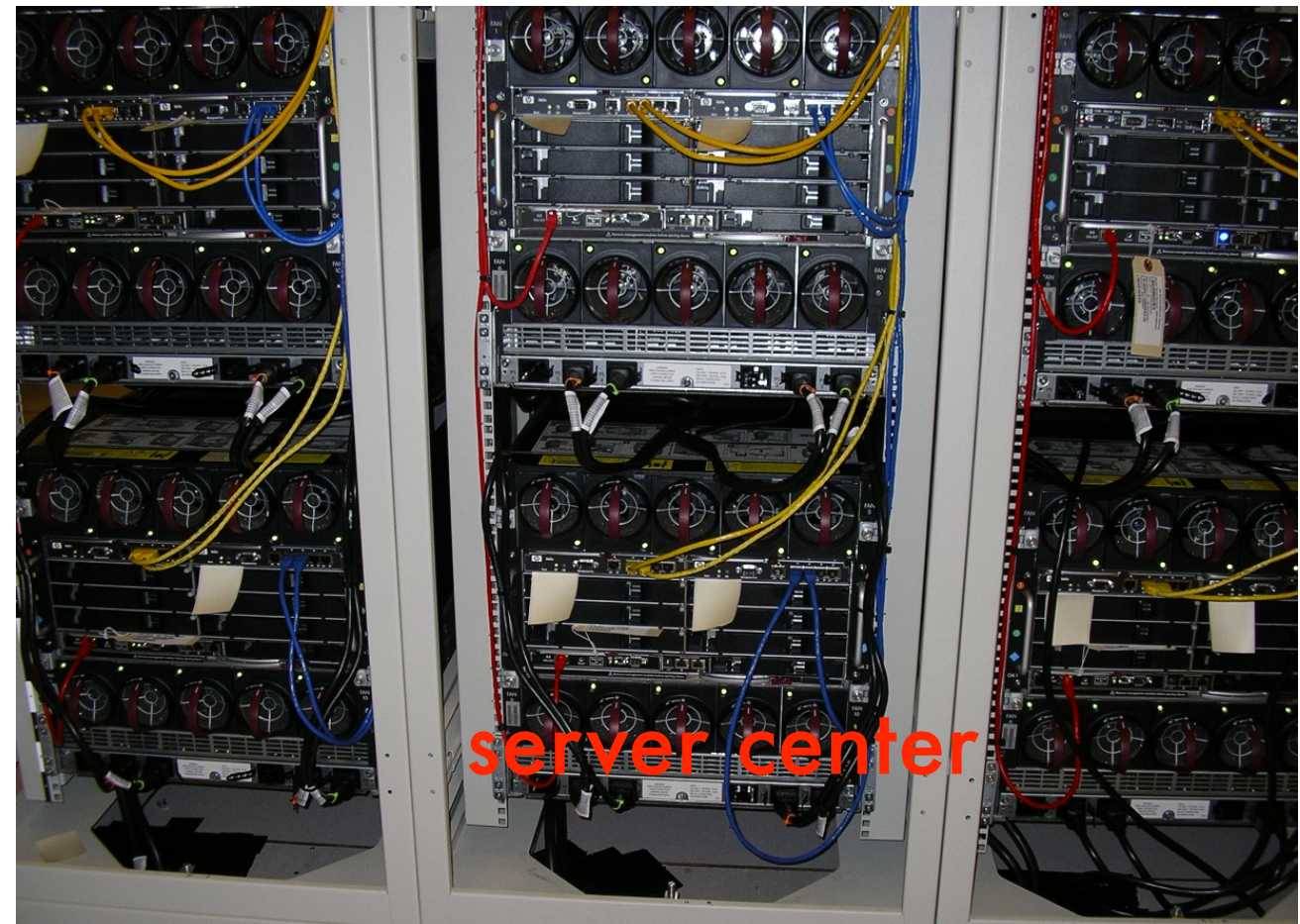
Many more sources



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



personalized sensors



server center

power grid



ubiquitous control

1.5 Basic Tools

1 Introduction

Alexander Smola

Introduction to Machine Learning 10-701

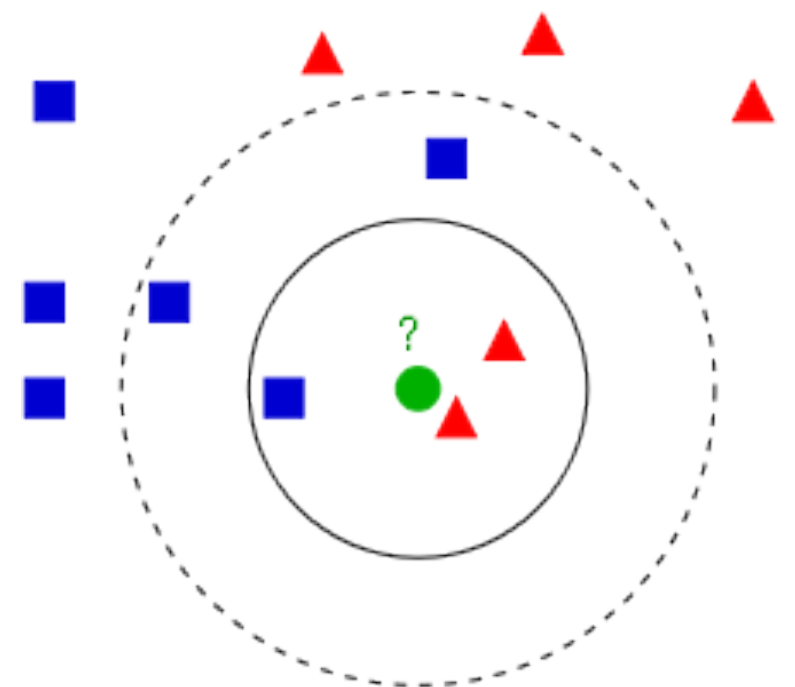
<http://alex.smola.org/teaching/10-701-15>

The background of the image is a complex Voronoi diagram. It consists of numerous irregular, non-overlapping polygons that fill the entire frame. Each polygon is a different color, with a wide variety of hues including greens, blues, purples, pinks, oranges, and browns. Some colors are repeated in different parts of the image. Small, solid black dots are scattered throughout the diagram, with one dot located in the center of each colored polygon. The text "Nearest Neighbor" is superimposed over the center of the image in a large, white, sans-serif font.

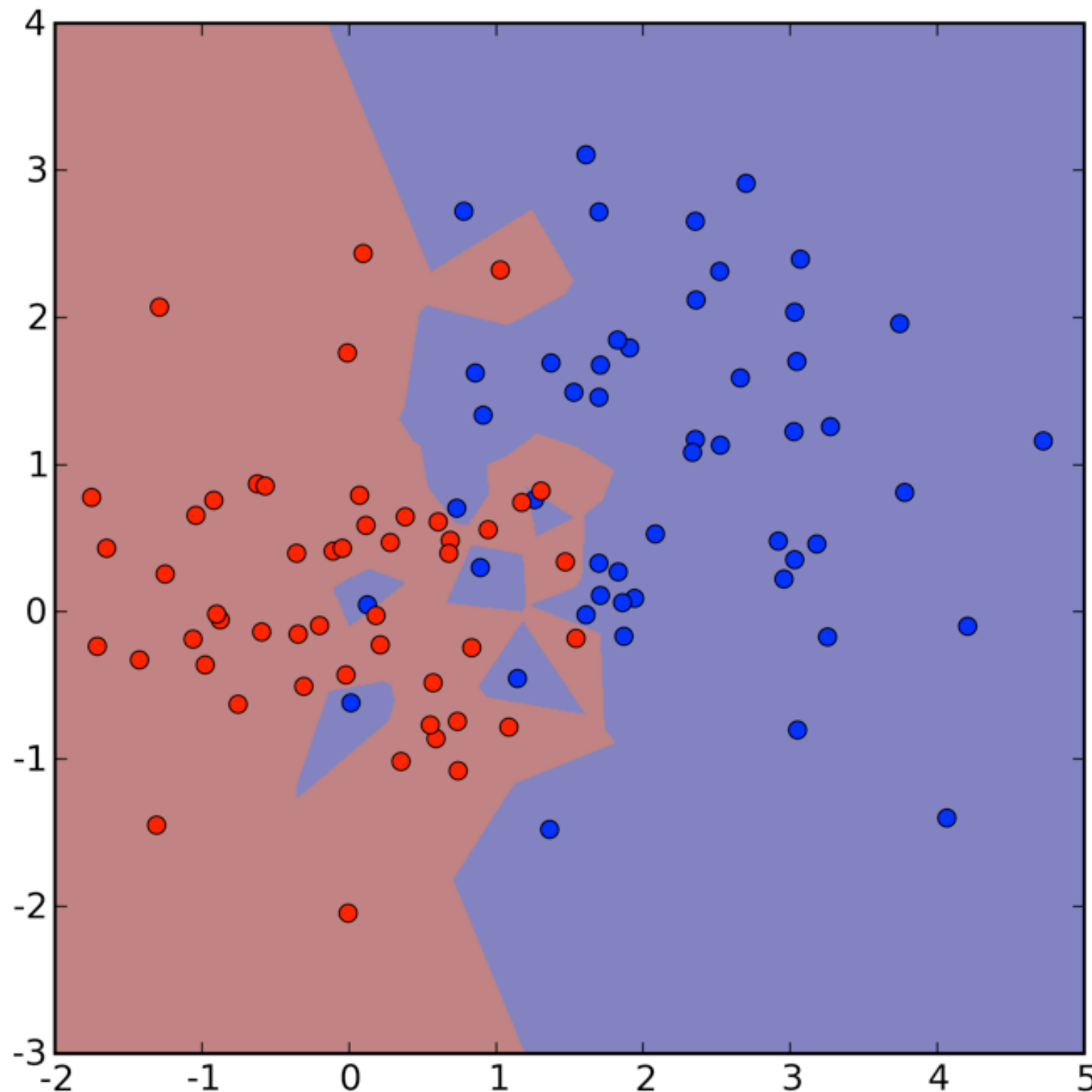
Nearest Neighbor

Nearest Neighbors

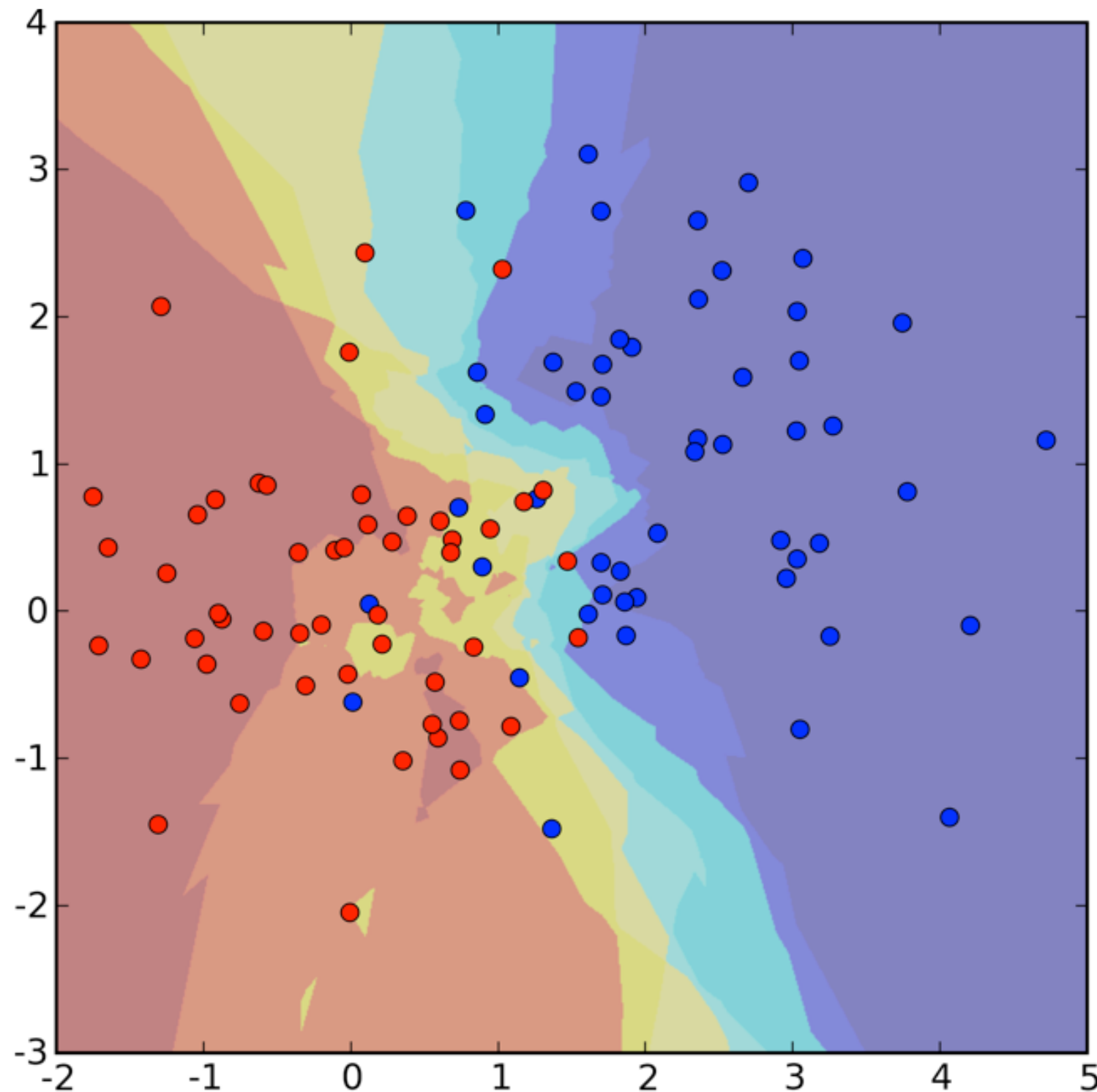
- Table lookup
For previously seen instance remember label
- Nearest neighbor
 - Pick label of most similar neighbor
 - Slight improvement - use k-nearest neighbors
 - For regression average
 - Really useful baseline!
 - Easy to implement for small amounts of data. Why?



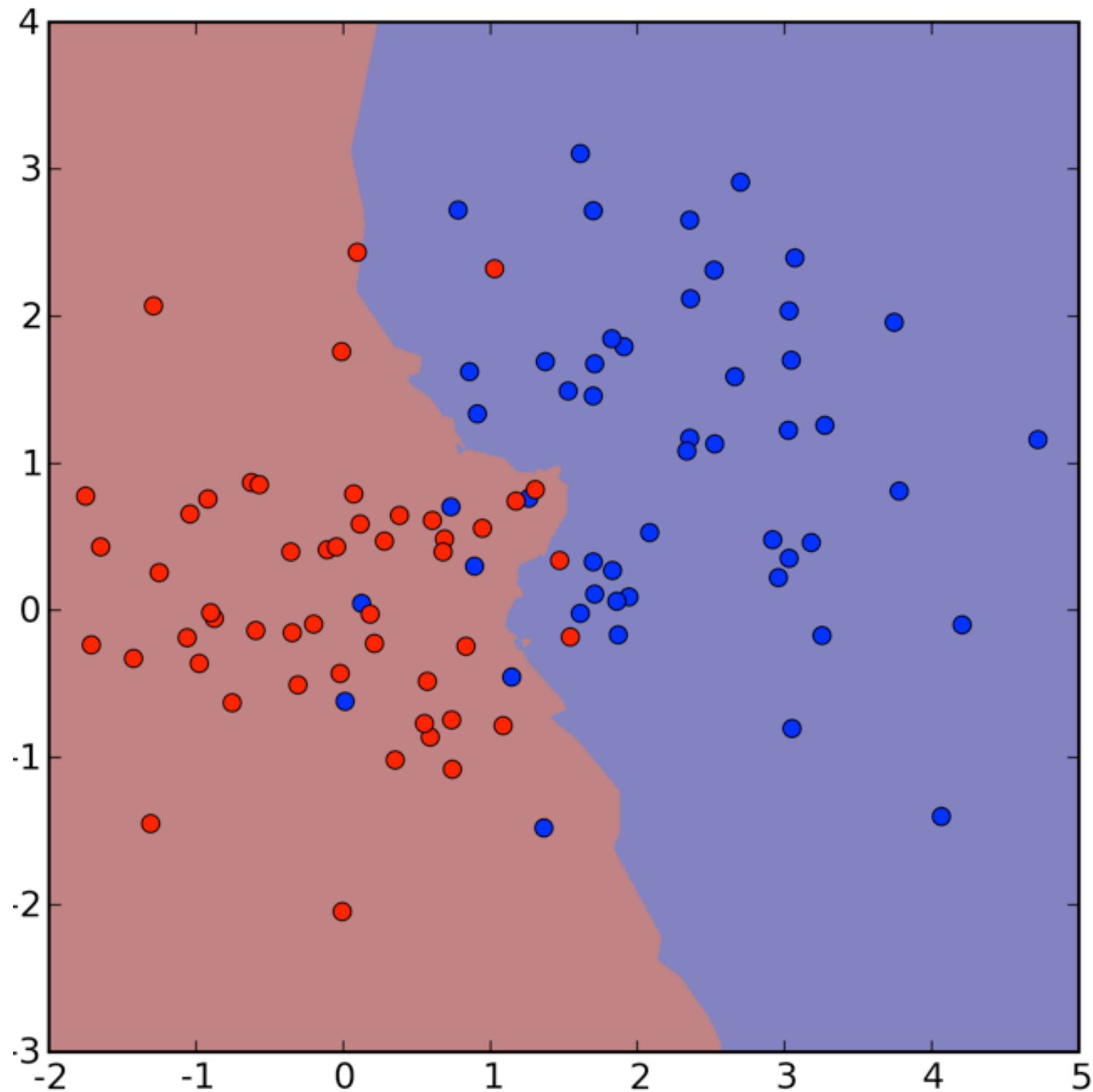
1-Nearest Neighbor



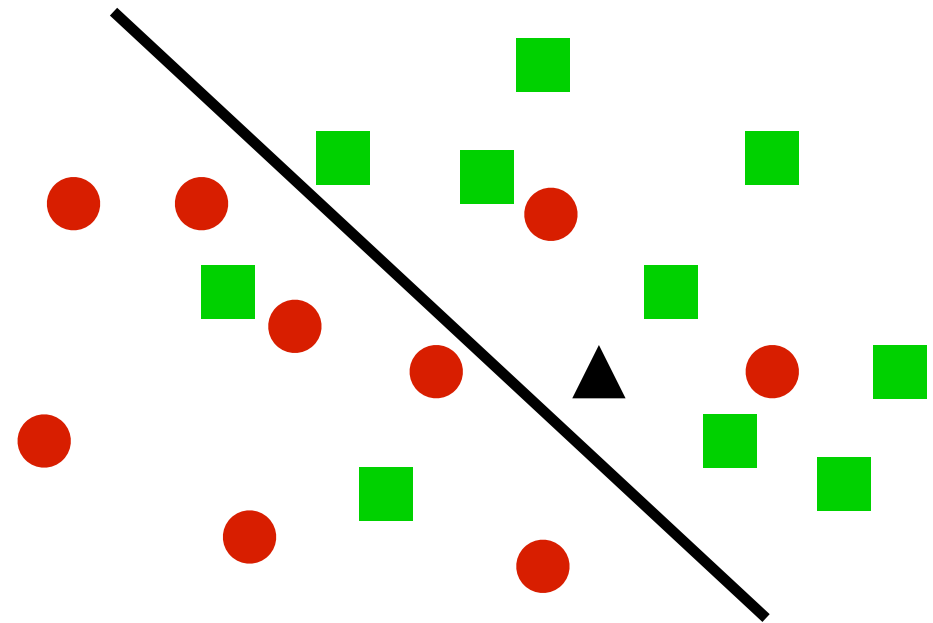
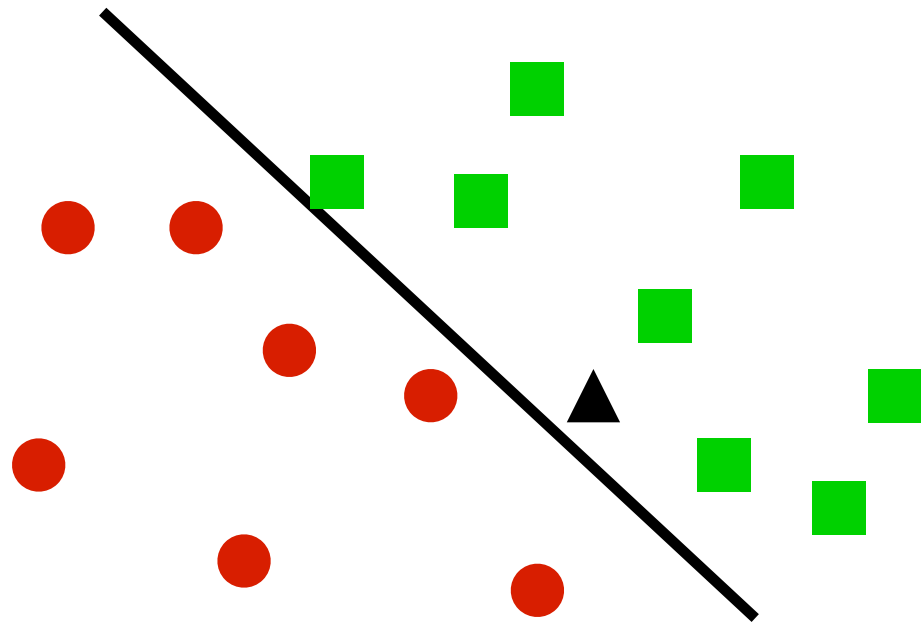
7-Nearest Neighbors



7-Nearest Neighbors Sign

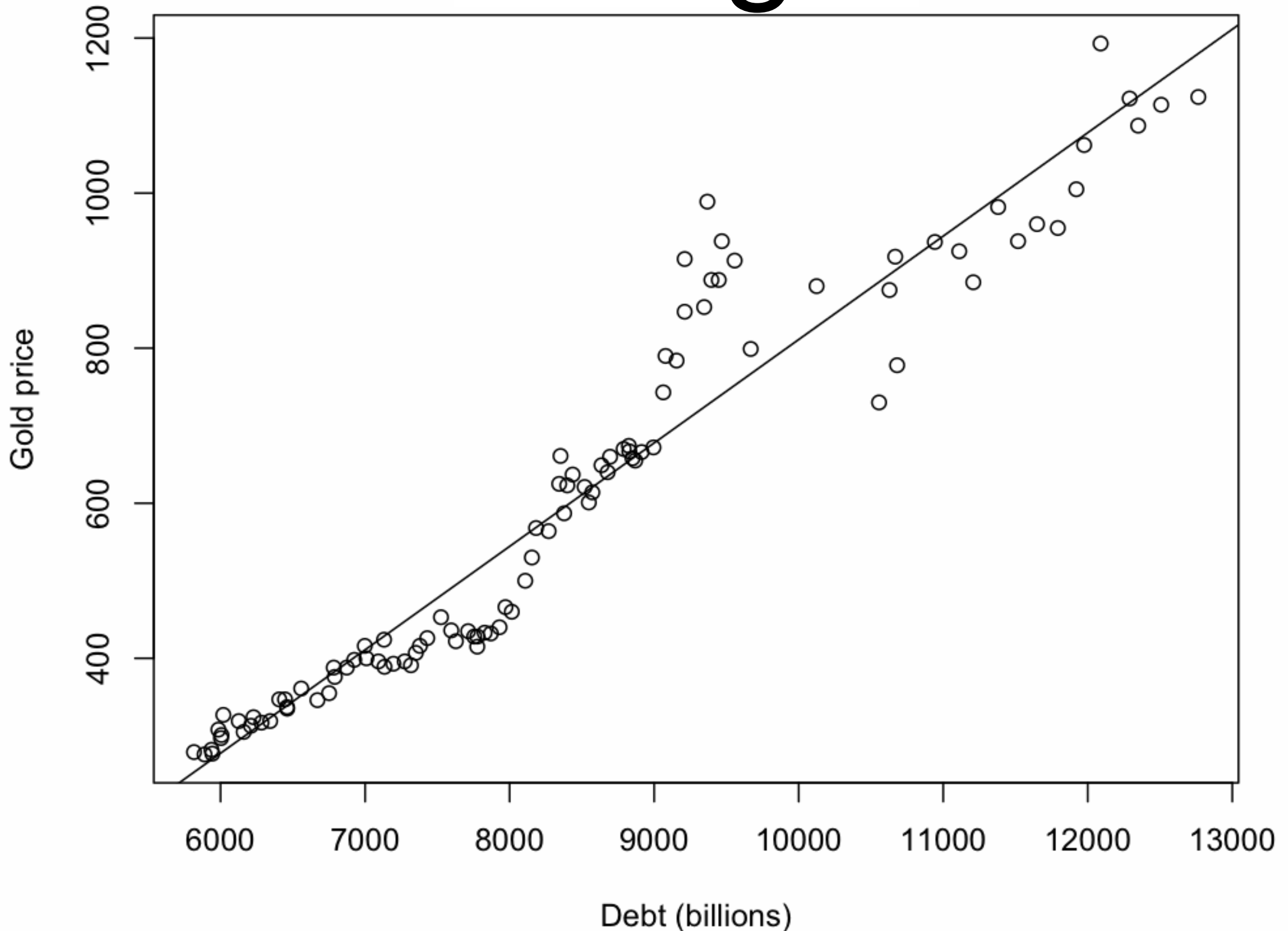


If we get more data

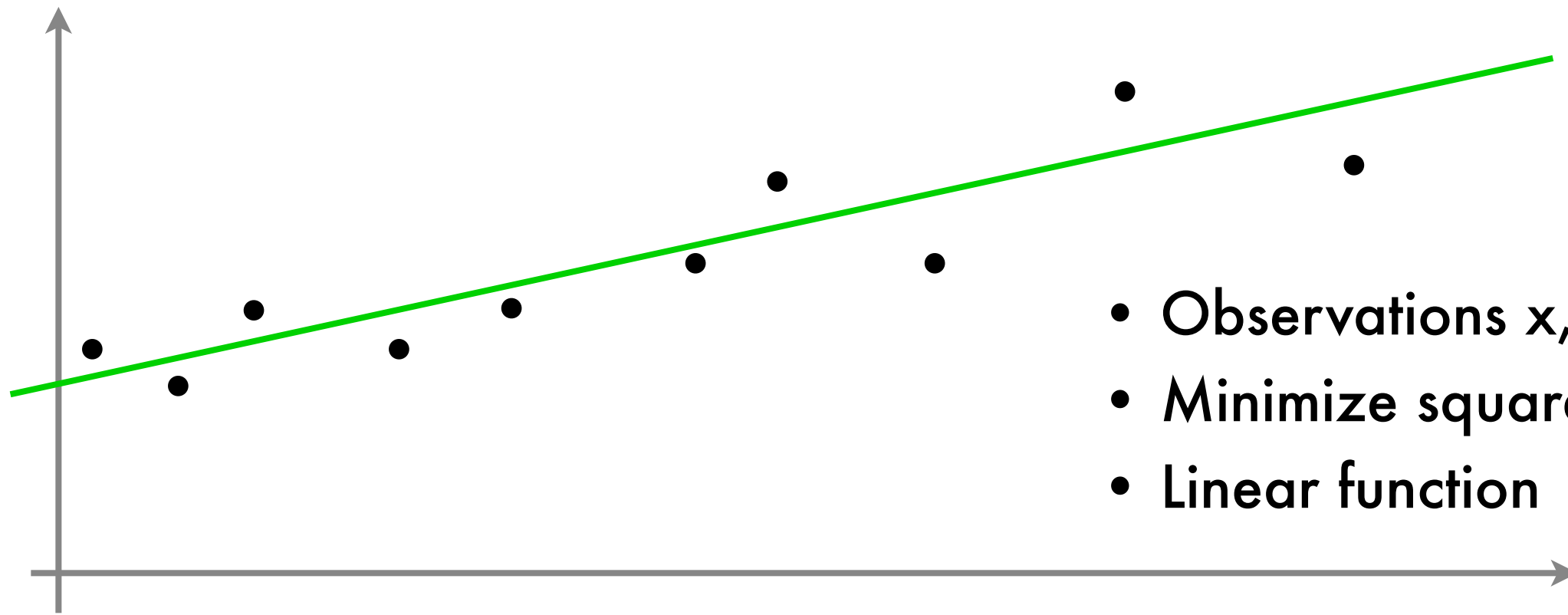


- 1 Nearest Neighbor
 - Converges to perfect solution if clear separation
 - Twice the minimal error rate $2p(1-p)$ for noisy problems
- k-Nearest Neighbor
 - Converges to perfect solution if clear separation (**but needs more data**)
 - Converges to minimal error $\min(p, 1-p)$ for noisy problems if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

Linear Regression



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

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

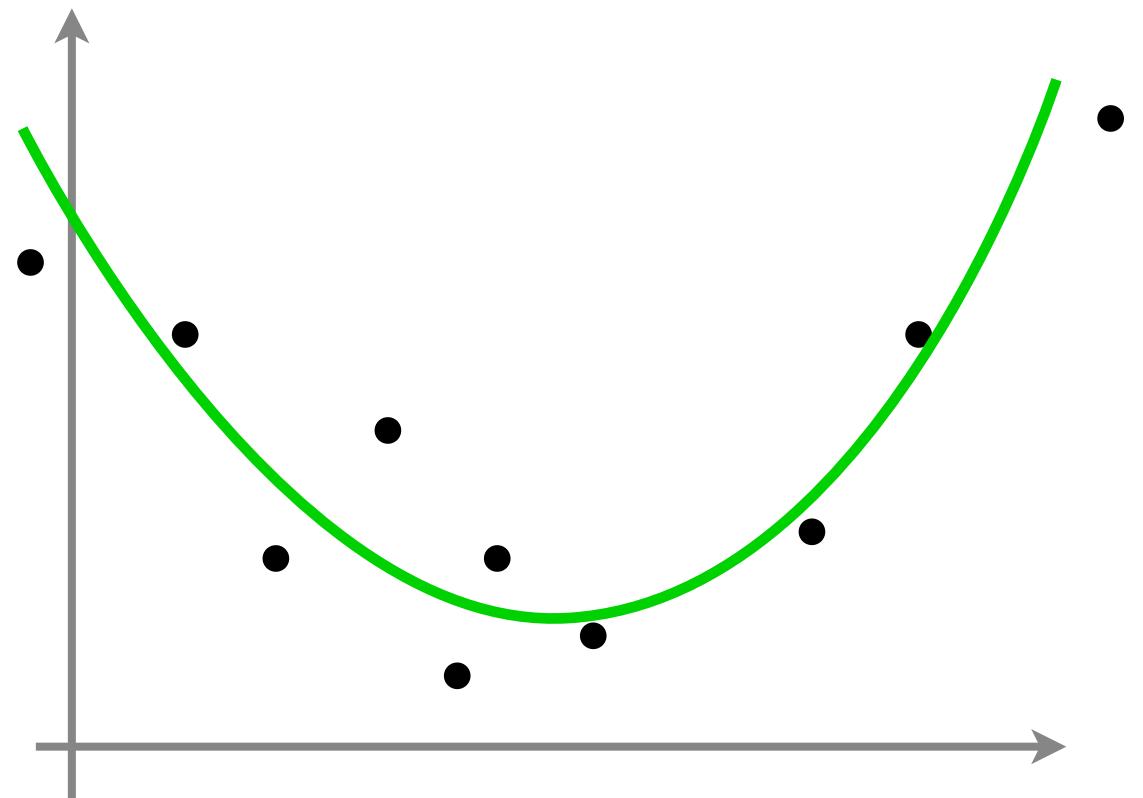
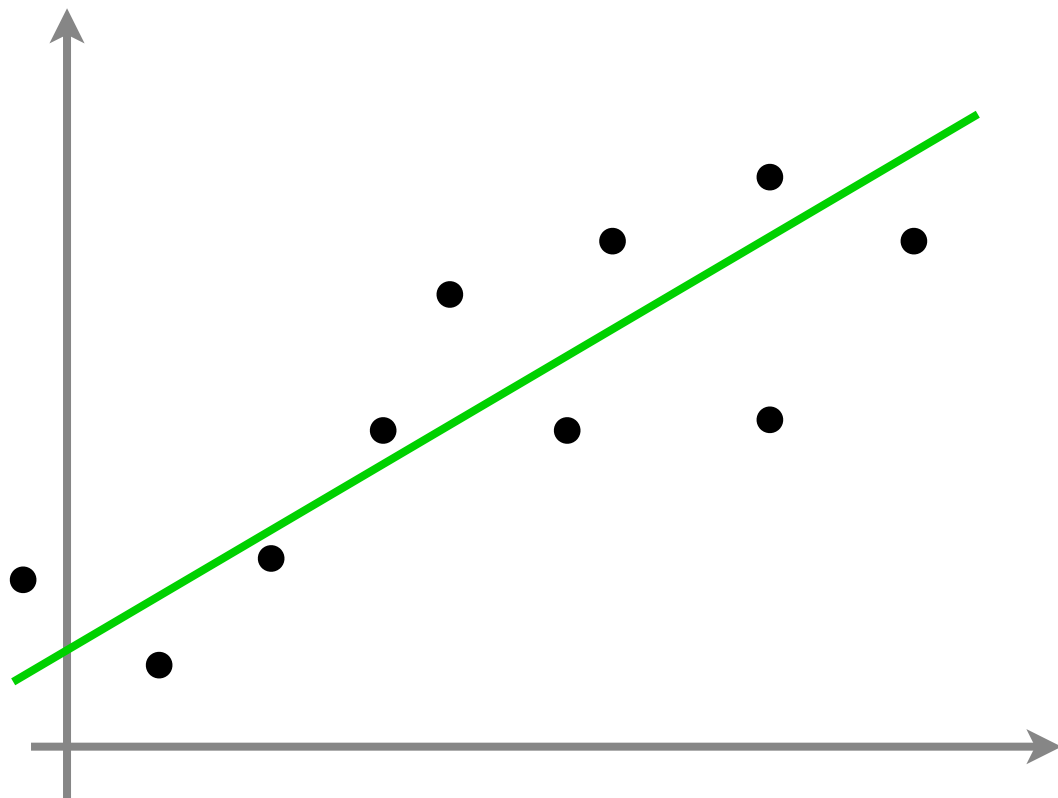
$$\underset{w}{\text{minimize}} \sum_{i=1}^m \frac{1}{2} (\langle w, \bar{x}_i \rangle - y_i)^2$$

- 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
- Quadratic model
- Cubic model
- Nonlinear model

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

$$f(x) = \langle w, (1, x, x^2) \rangle$$

$$f(x) = \langle w, (1, x, x^2, x^3) \rangle$$

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

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)

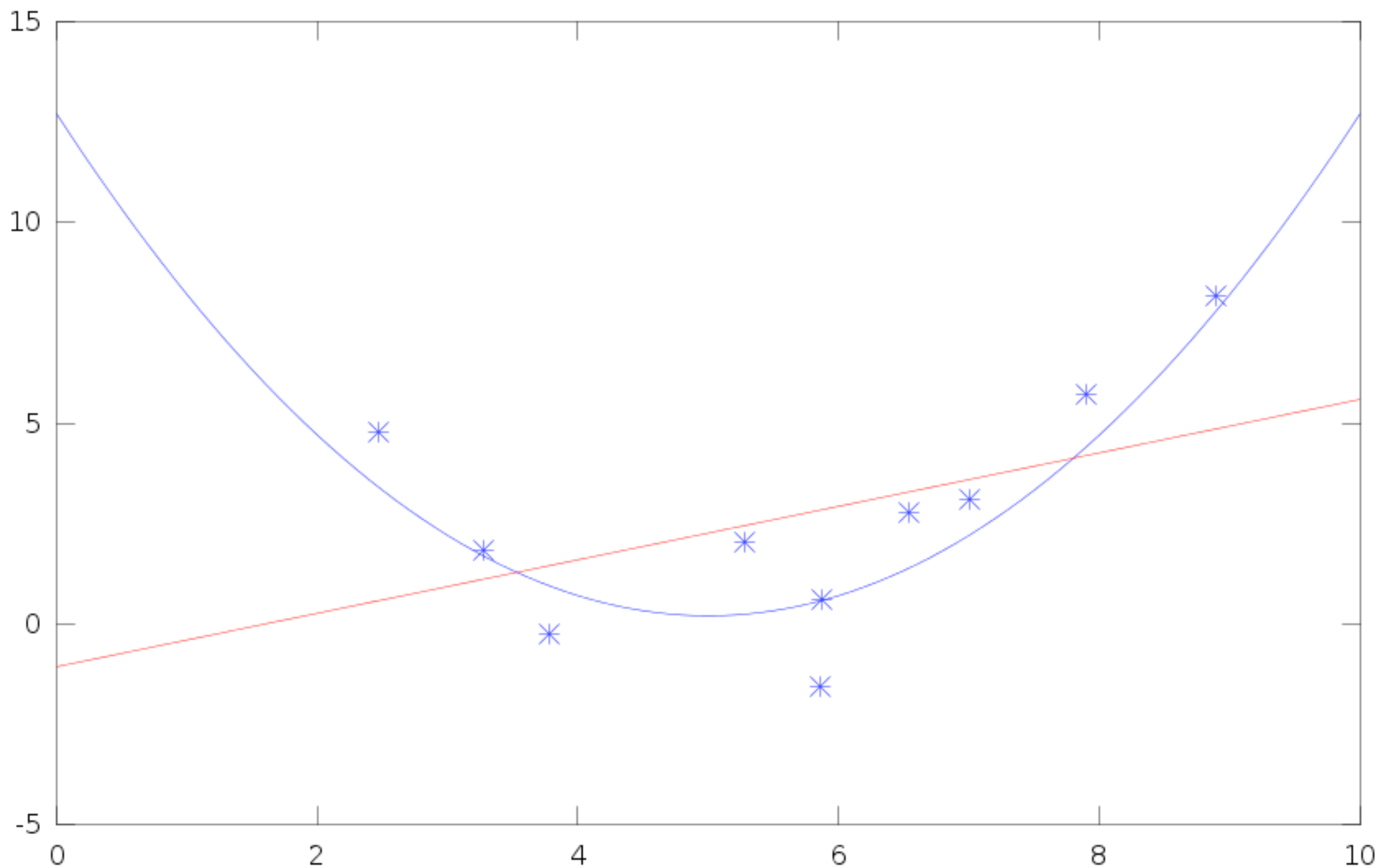
Training

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

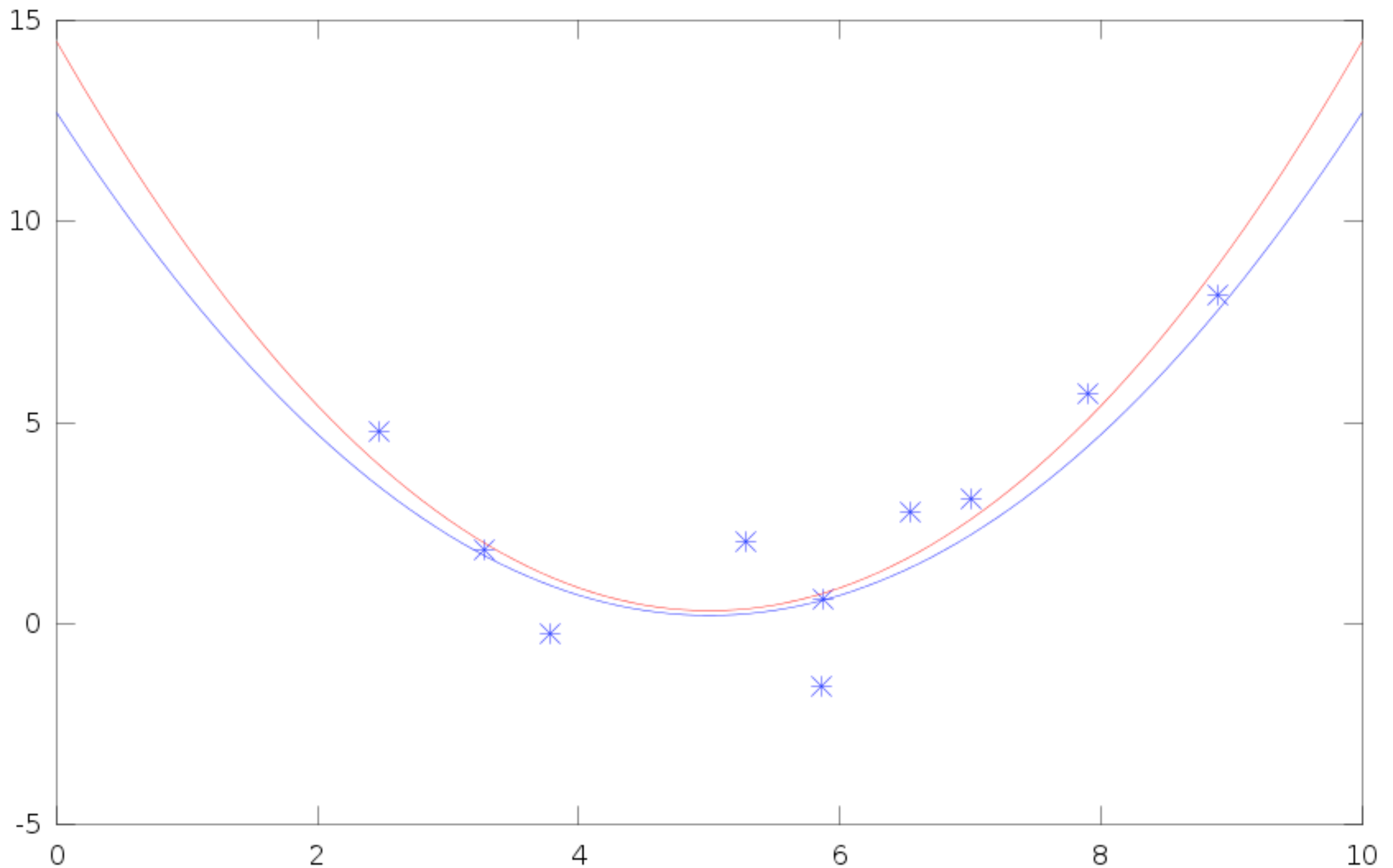
Testing

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

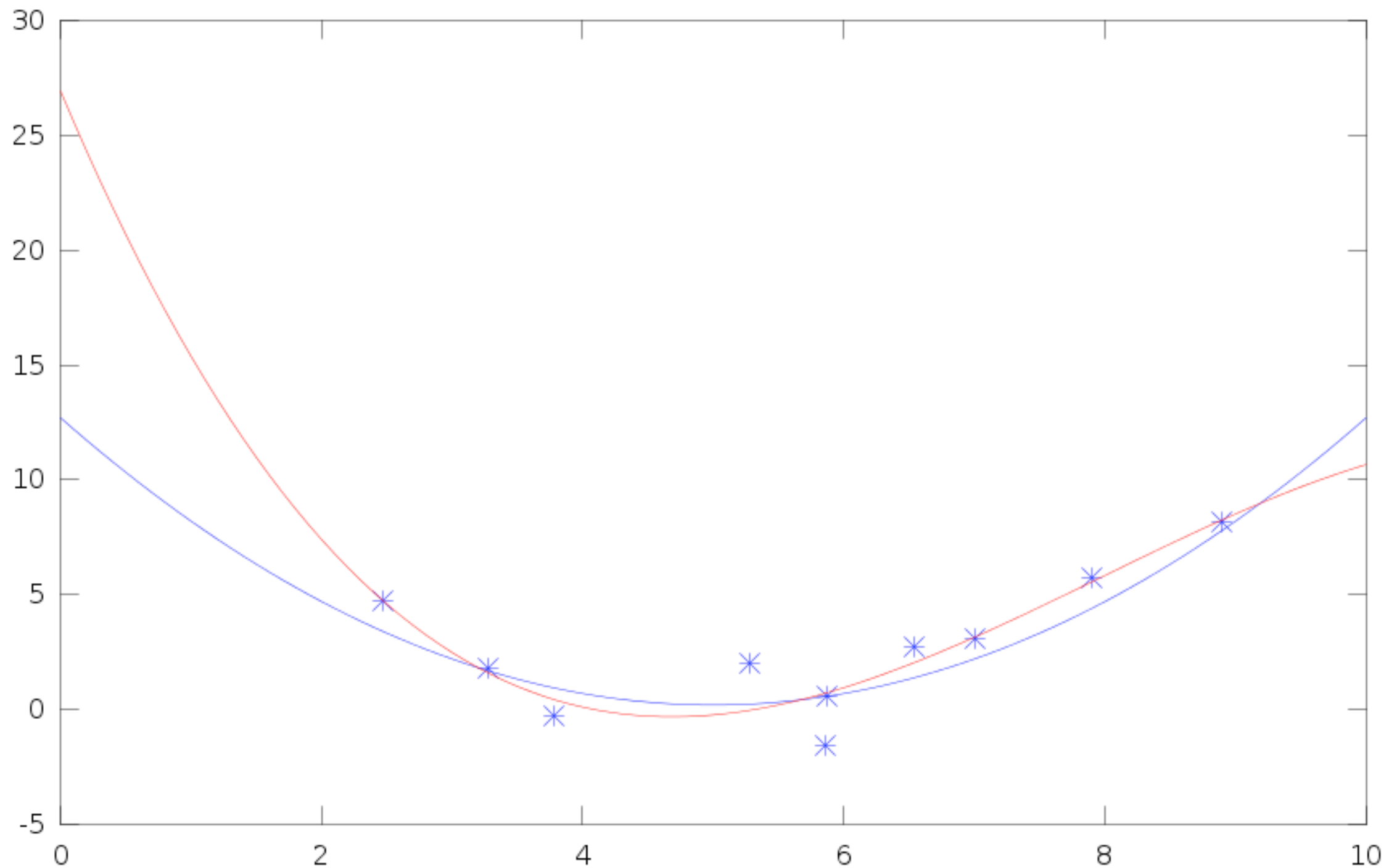
Regression (d=1)



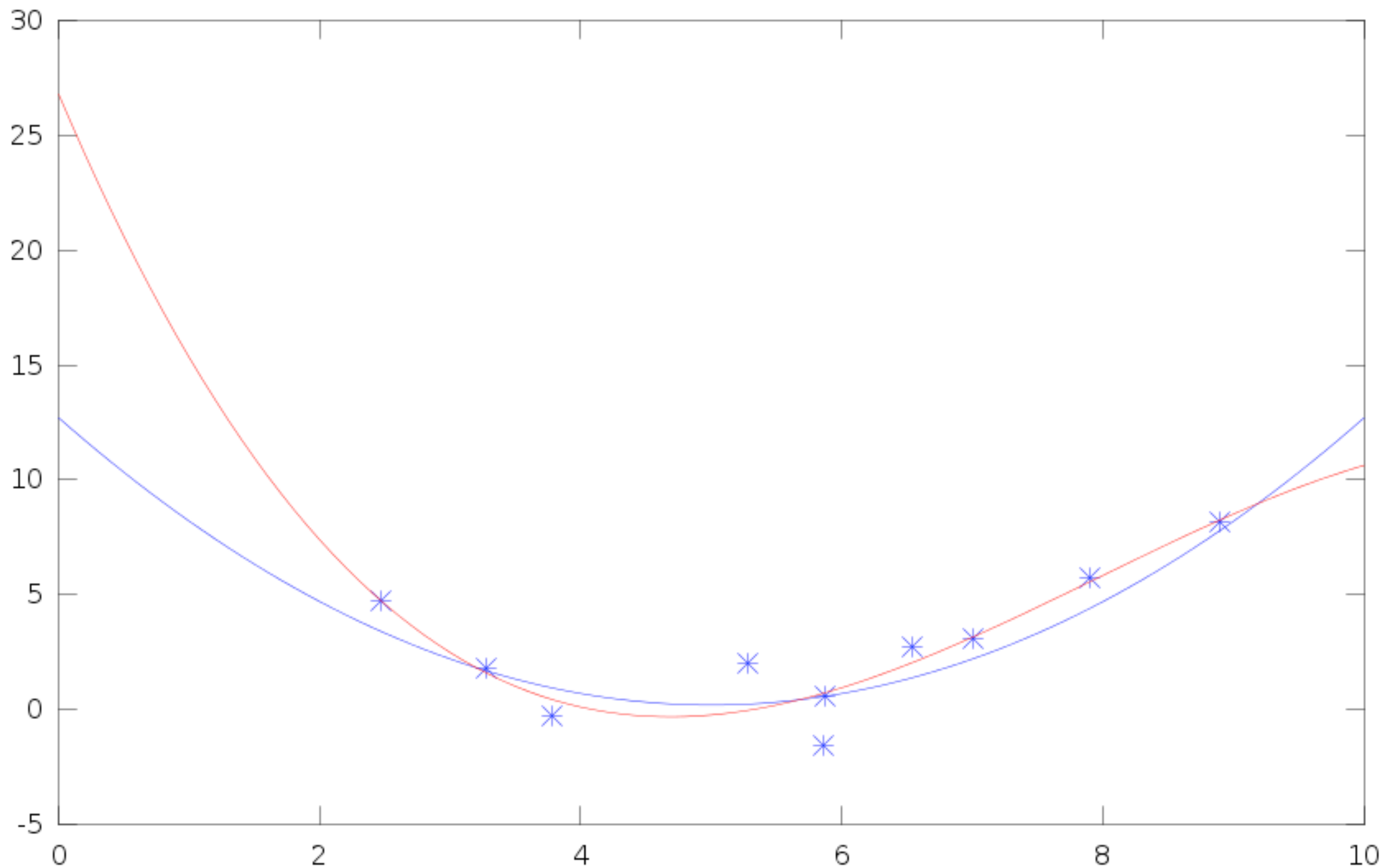
Regression (d=2)



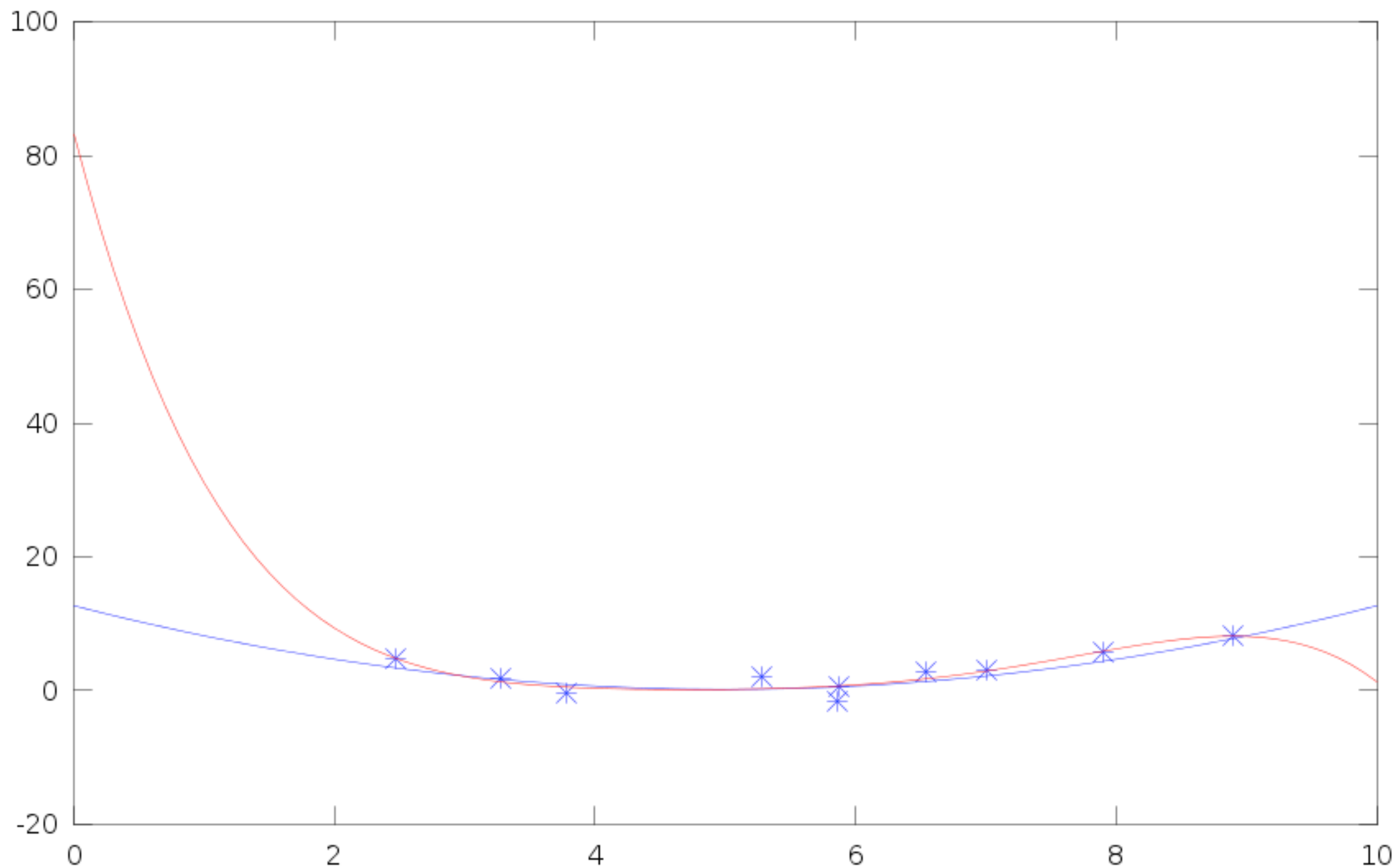
Regression (d=3)



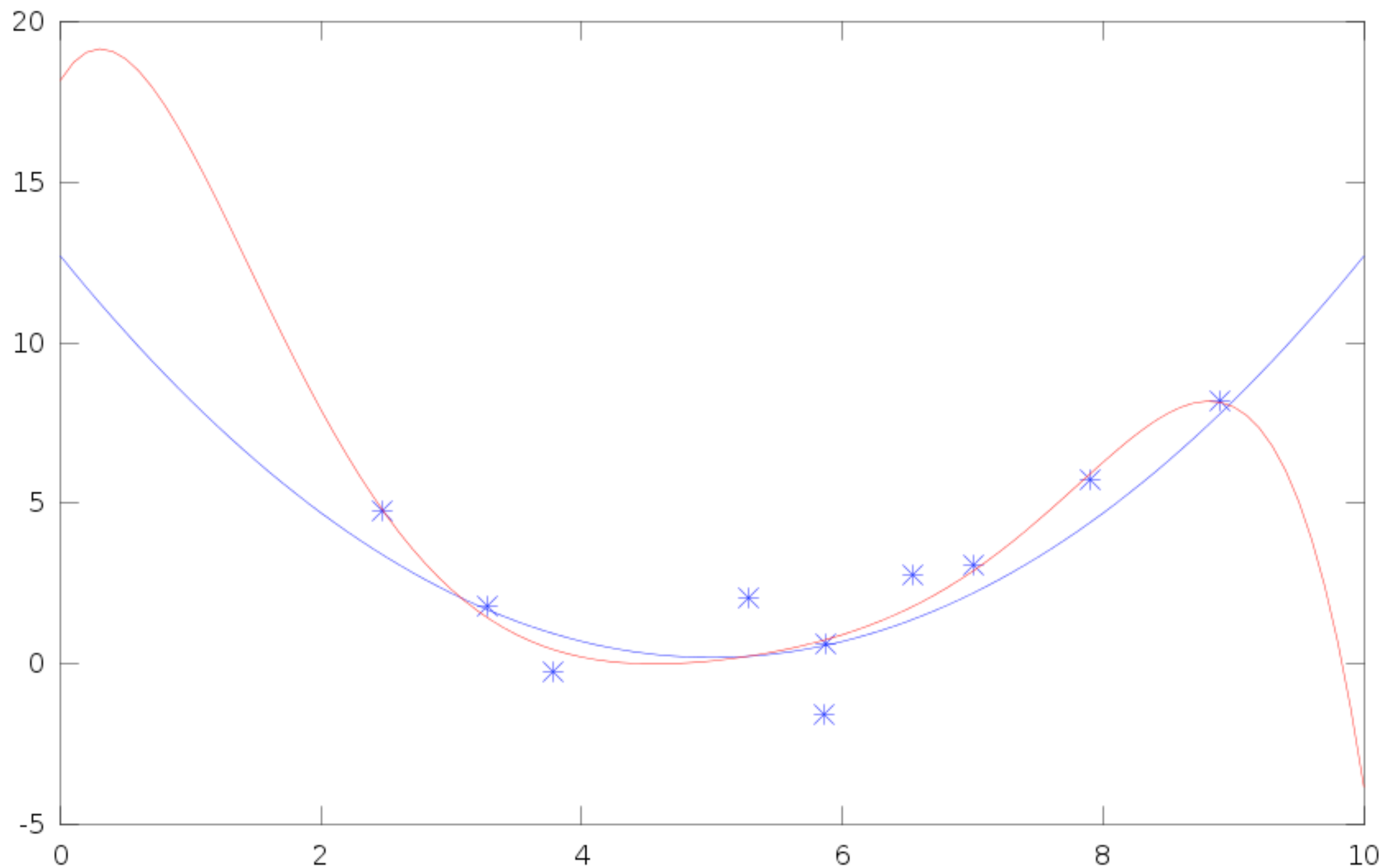
Regression (d=4)



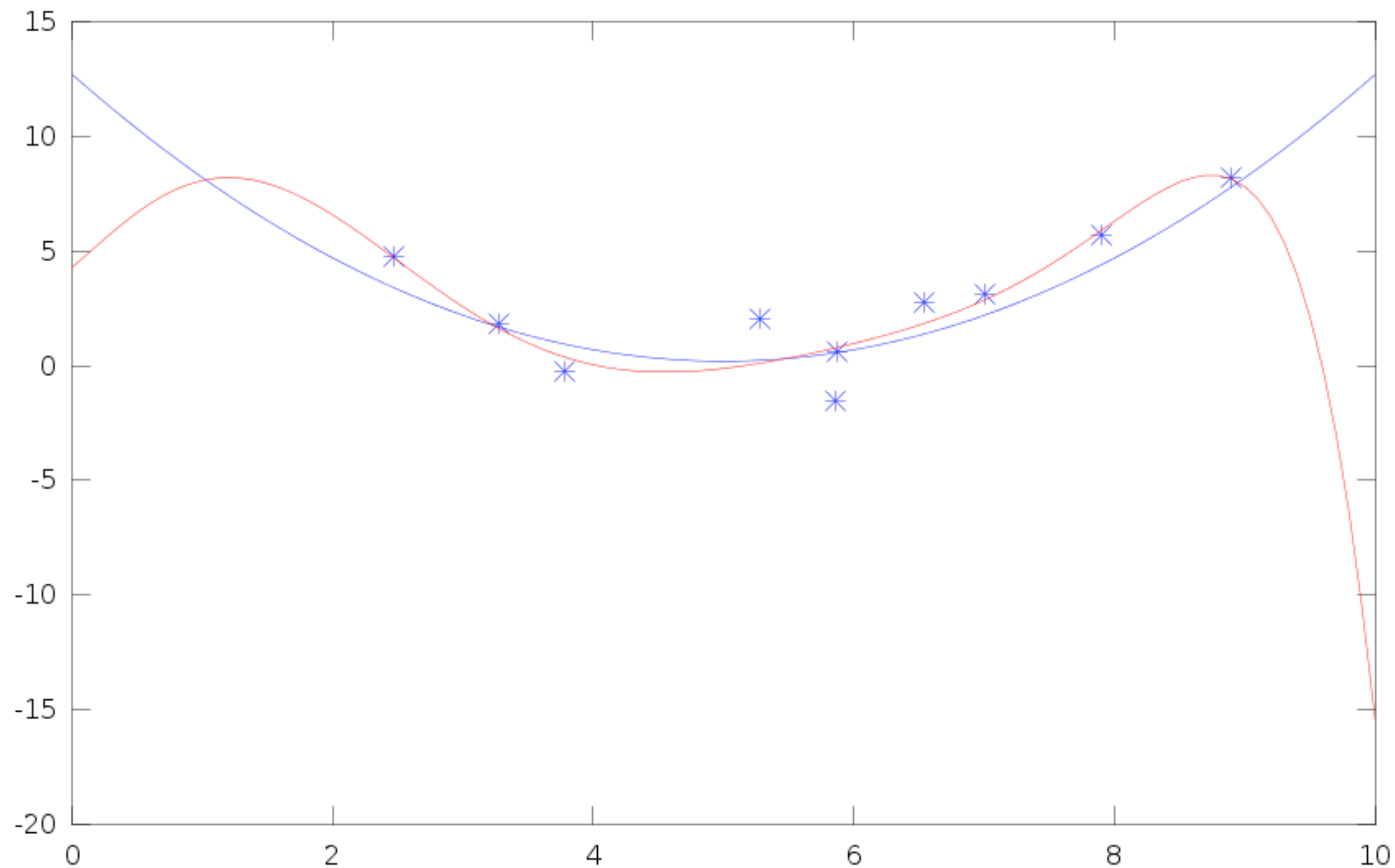
Regression (d=5)



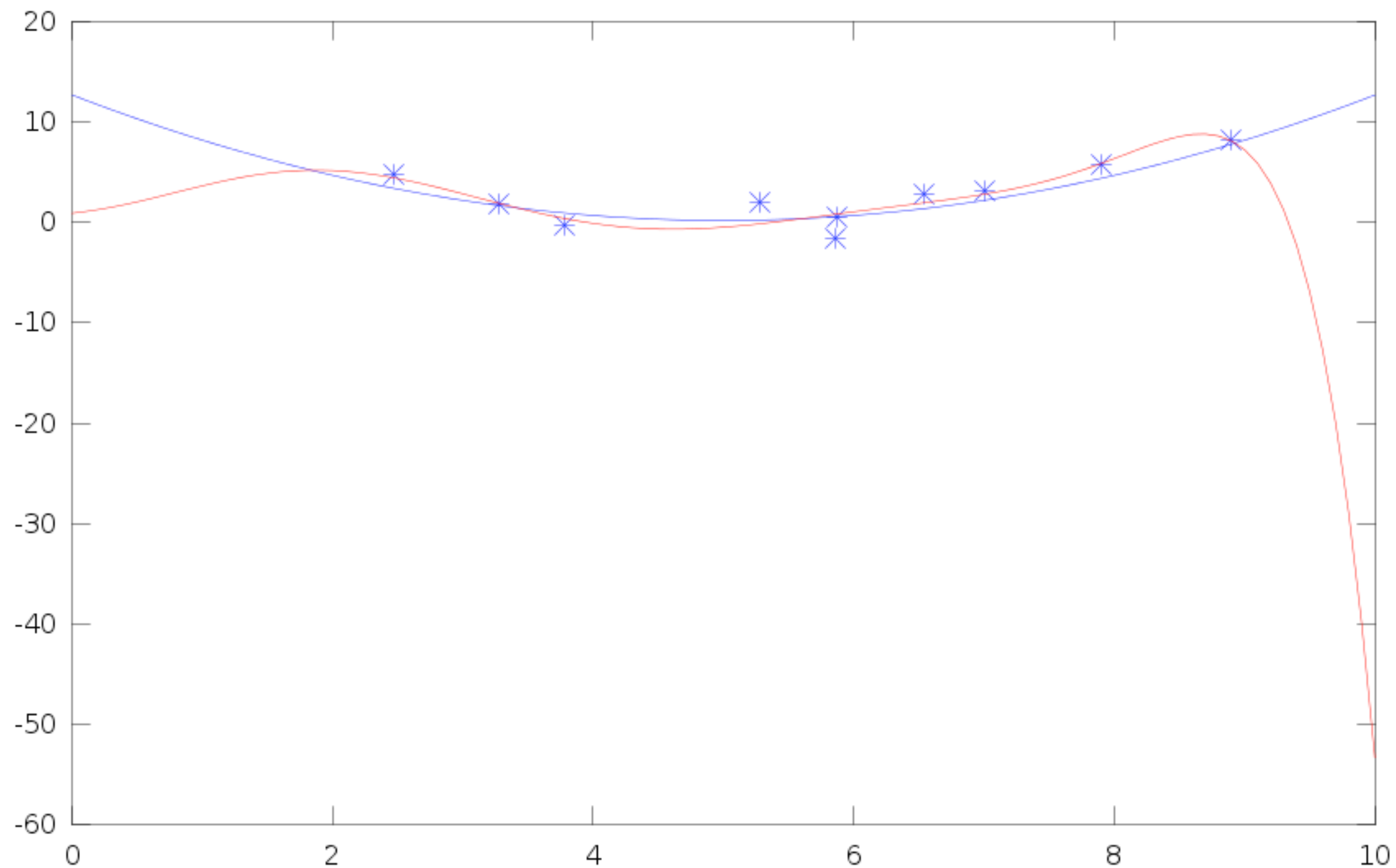
Regression (d=6)



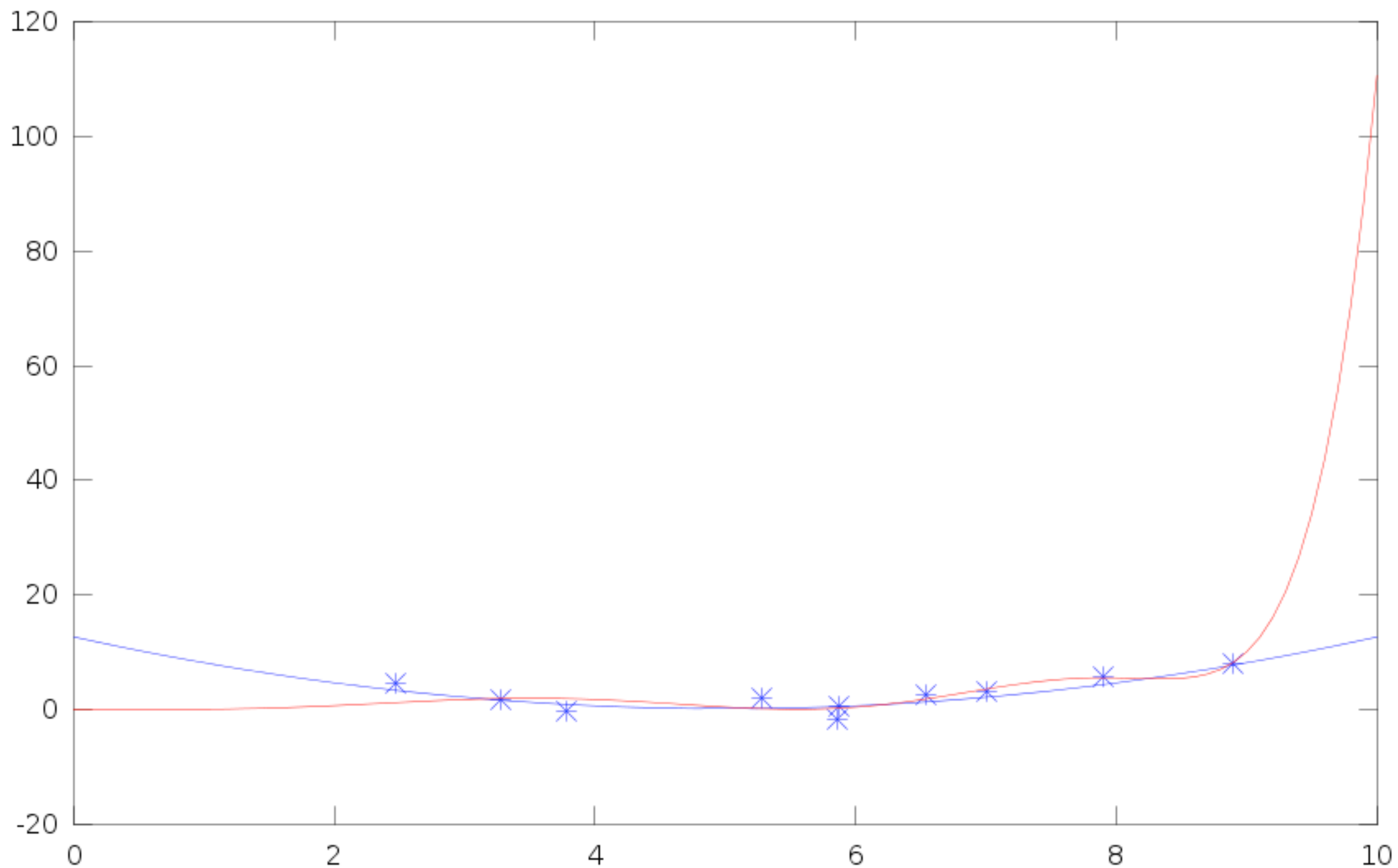
Regression (d=7)



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 = 1.10156e-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