MACHINE LEARNING
DEPARTMENT

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/

Carnegie Mellon University


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

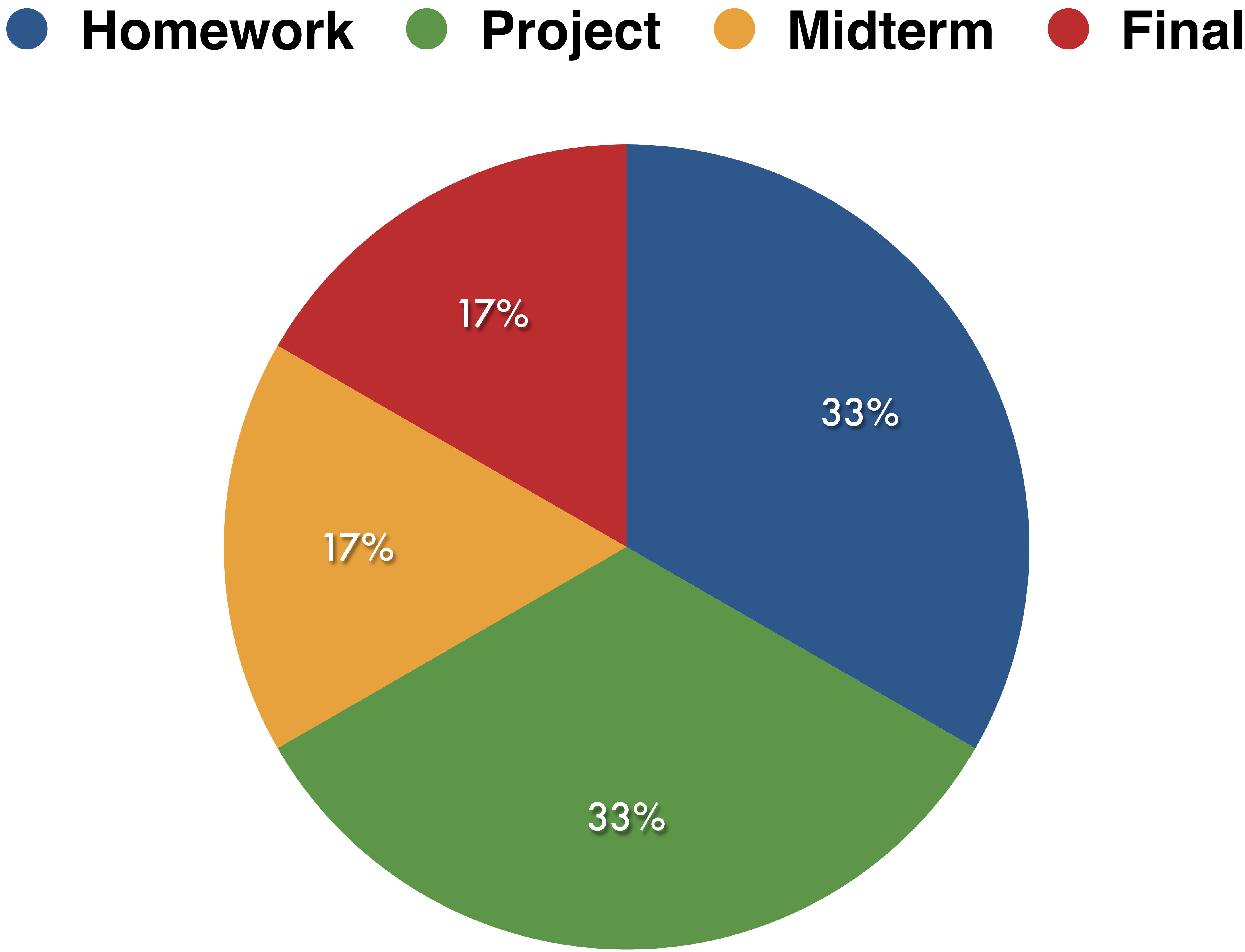
- What are you trying to do? Articulate your objectives using absolutely no jargon.
- How is it done today, and what are the limits of current practice?
- What's new in your approach and why do you think it will be successful?
- Who cares? If you're successful, what difference will it make?
- What are 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
17%
33%
17%
33%

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, Application
- S
  N bors, Decision Trees, Neural Networks,
  P
- (Generalized) Linear Models
  Support Vector Classification, Re n, Kernel PCA
- Theoretical Tools
  Risk Minimization, Convergence Bounds, Information Theory
- raphical Models, Dynamic Programming, Latent
- Interacting with the environment
  Online Le mer
- Scalabilit

for the internet

all you need for a startup

for your PhD

for Wall Street

biology

energy

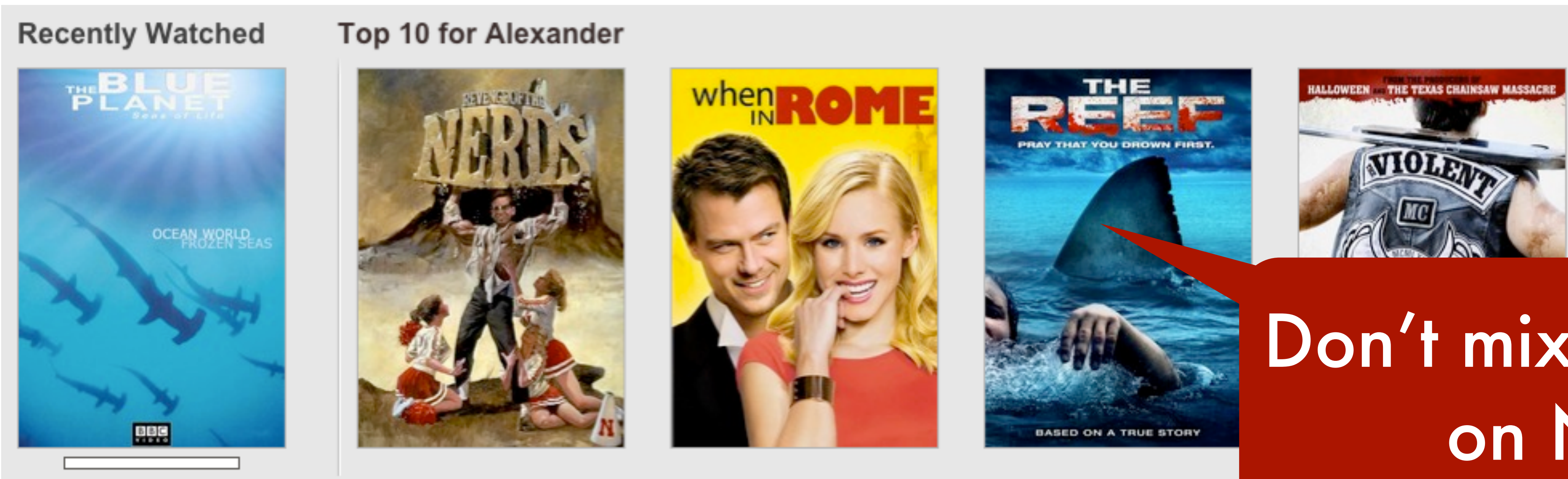
1.2 Programming with Data

1 Introduction

Alexander Smola
Introduction to Machine Learning 10-701
http://alex.smola.org/teaching/10-701-15

Machine Learning Problems

Collaborative Filtering



Don’t mix preferences on Netflix!

Amazon books

Imitation Learning in Games

Black & White
Lionsgate Studios

Imitation Learning

Reinforcement Learning

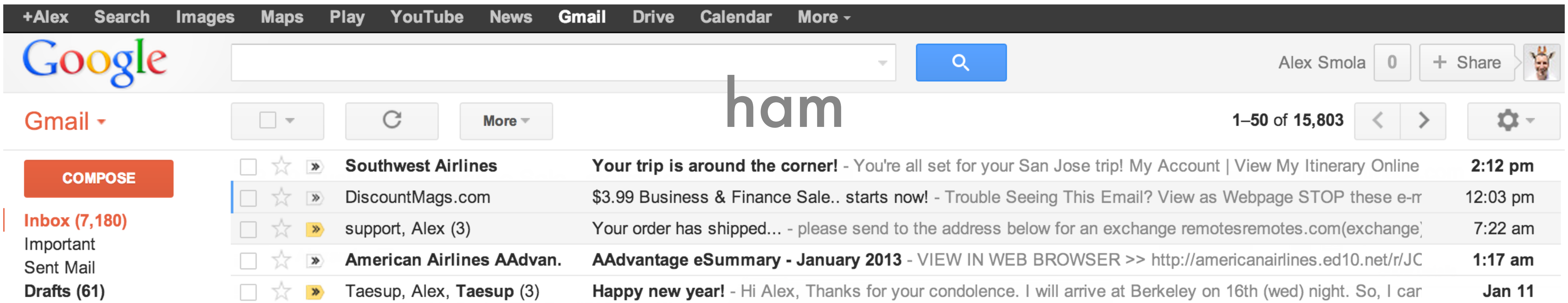
https://www.youtube.com/watch?v=lleRKHsJBJ0

Reinforcement Learning

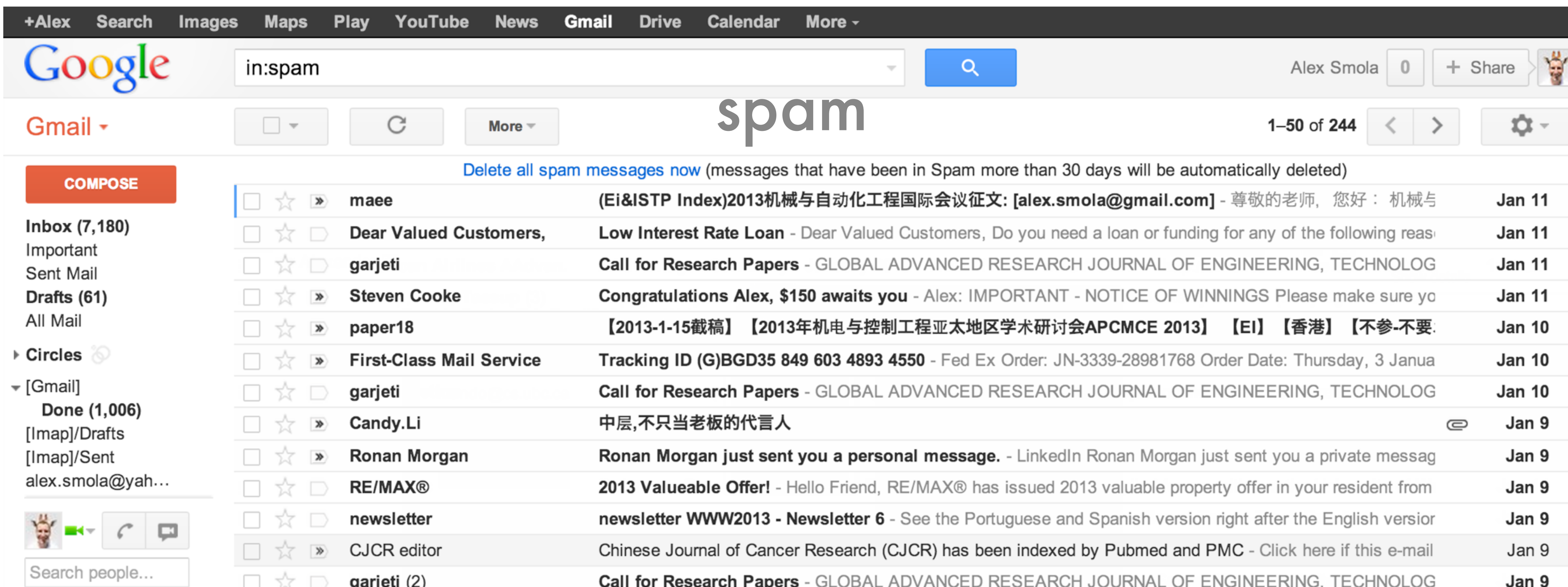
https://www.youtube.com/watch?v=IleRKHsJBJ0

Spam Filtering

ham

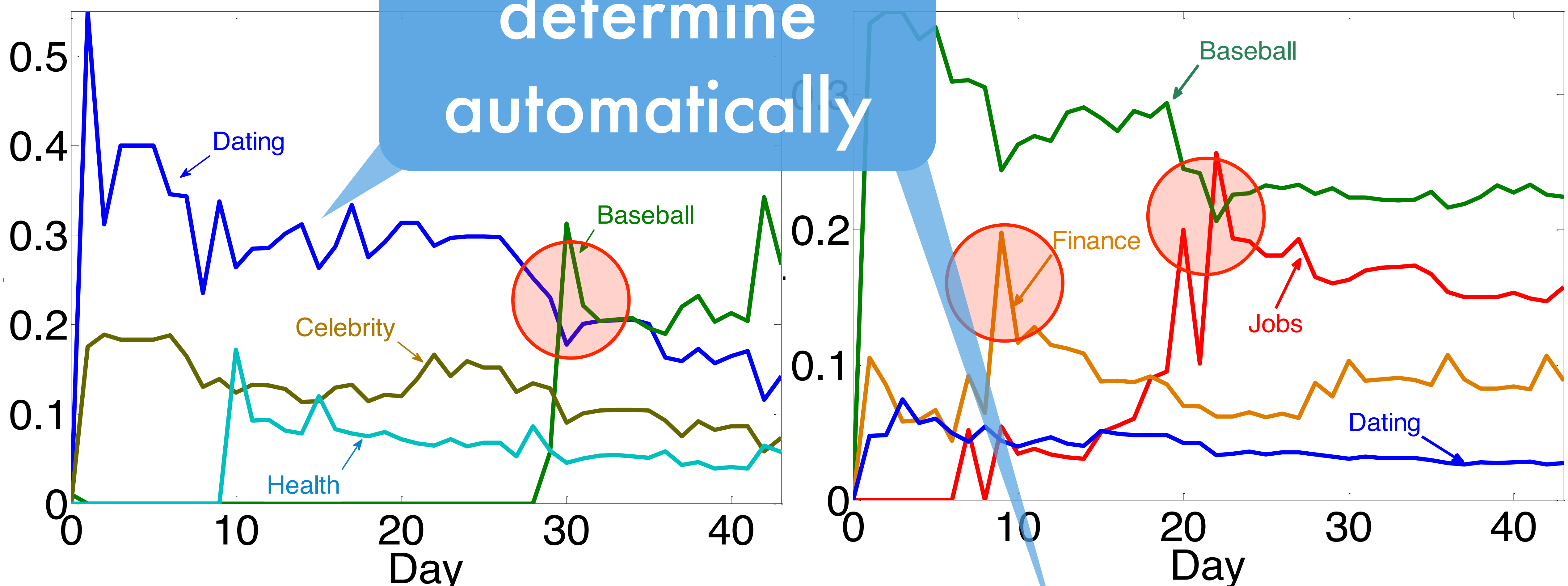


spam



User profiling

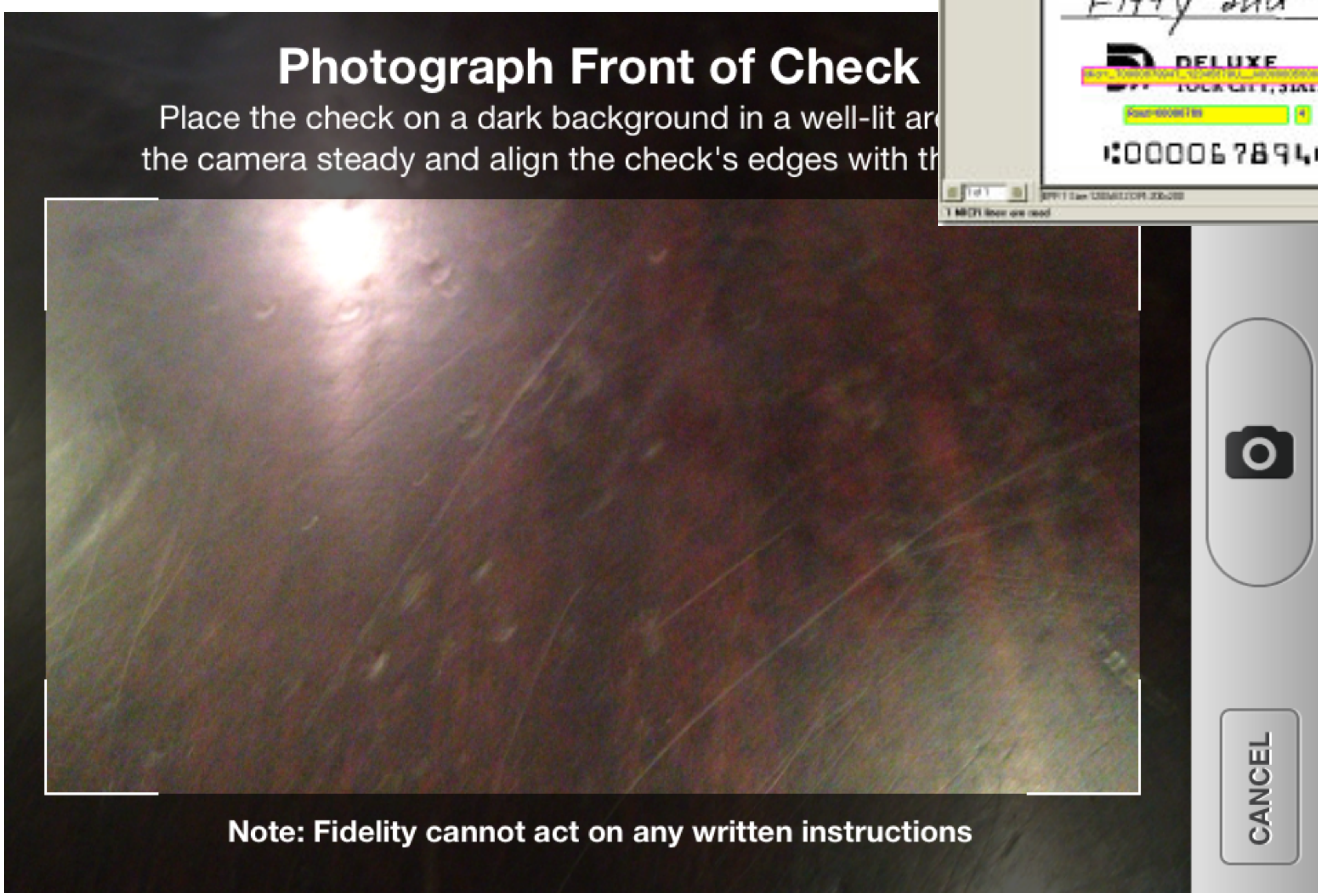
determine automatically

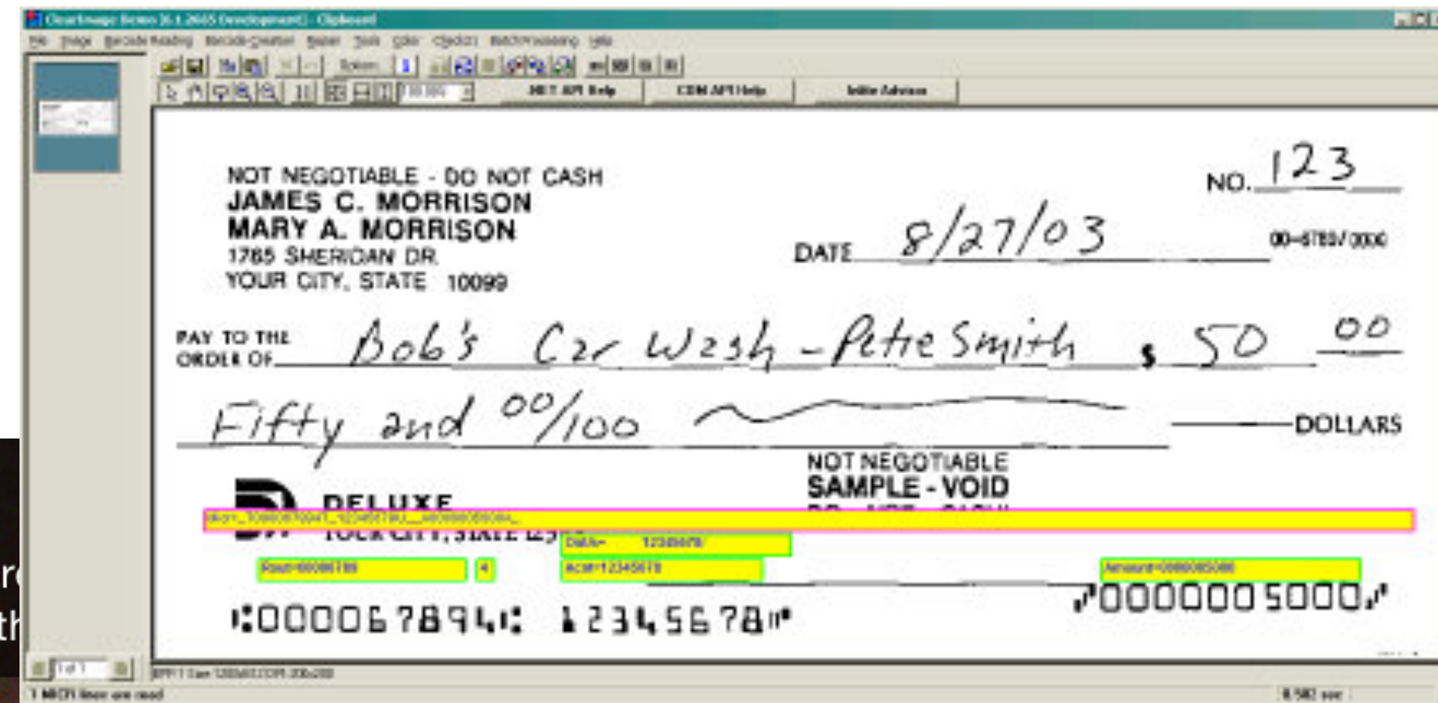


Dating	Baseball	Celebrity	Health	Jobs	Finance
women	League	Snooki	skin	job	financial
men	baseball	Tom	body	career	Thomson
dating	basketball,	Cruise	fingers	business	chart
singles	doublehead	Katie	cells	assistant	real
personals	Bergesen	Holmes	toes	hiring	Stock
seeking	Griffey	Pinkett	wrinkle	part-time	Trading
match	bullpen	Kudrow	layers	receptionist	currency
	Greinke	Hollywood			

Cheque reading

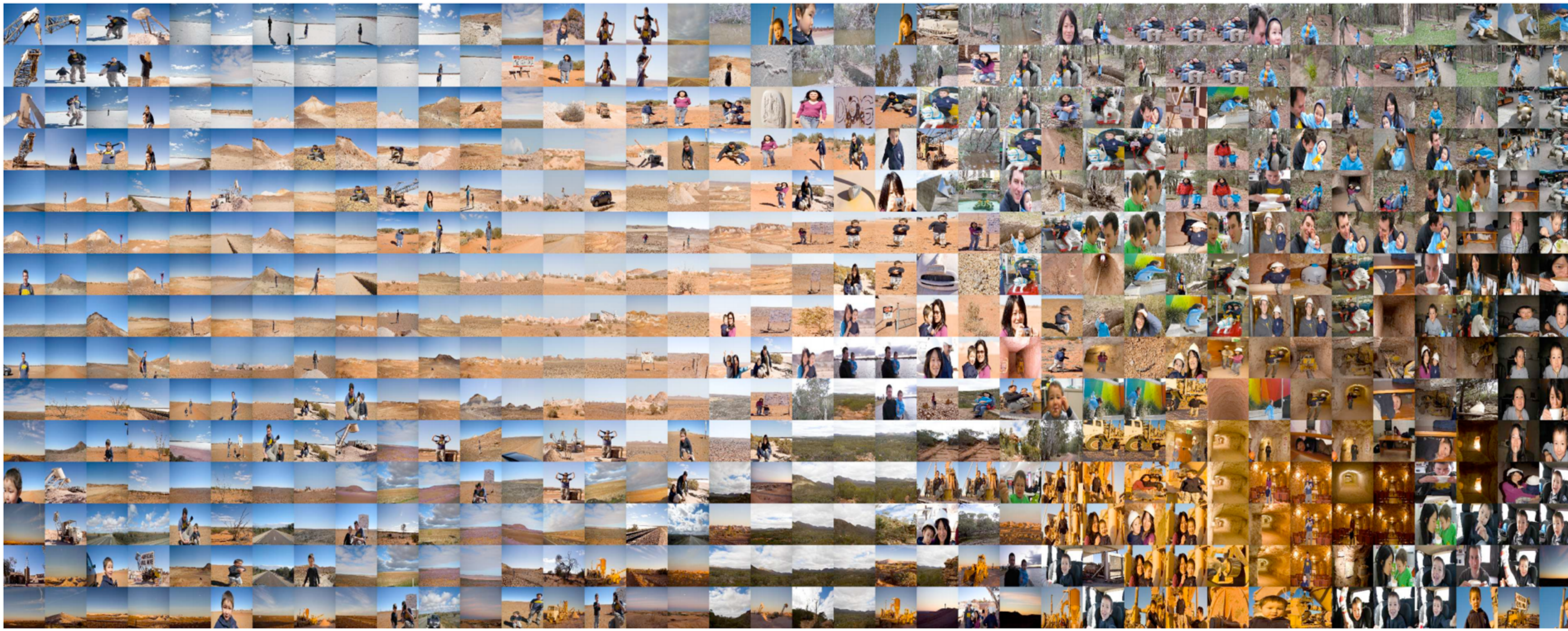
segment image





recognize handwriting

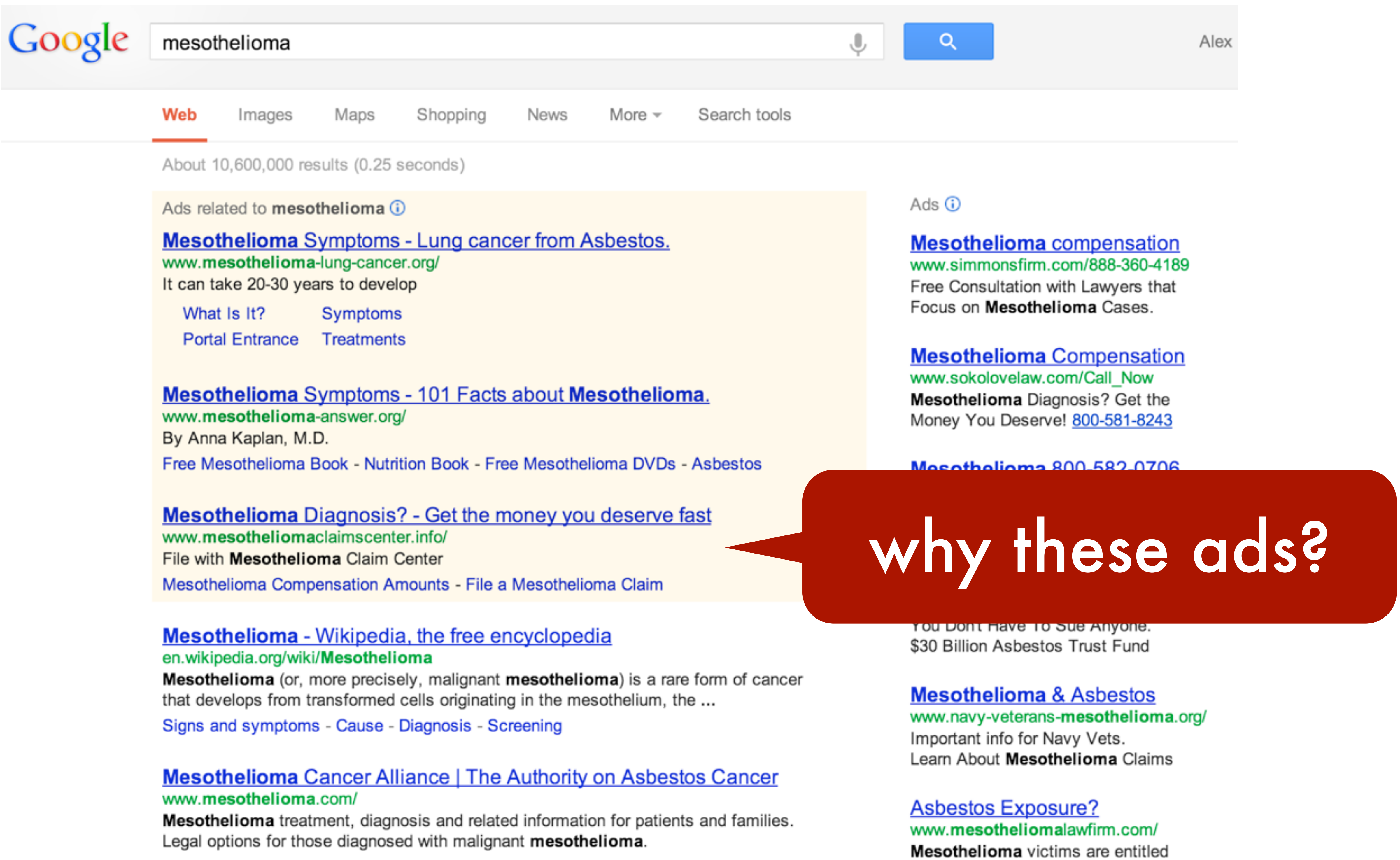
Image Layout



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

Carnegie Mellon University

Search ads



Carnegie Mellon University

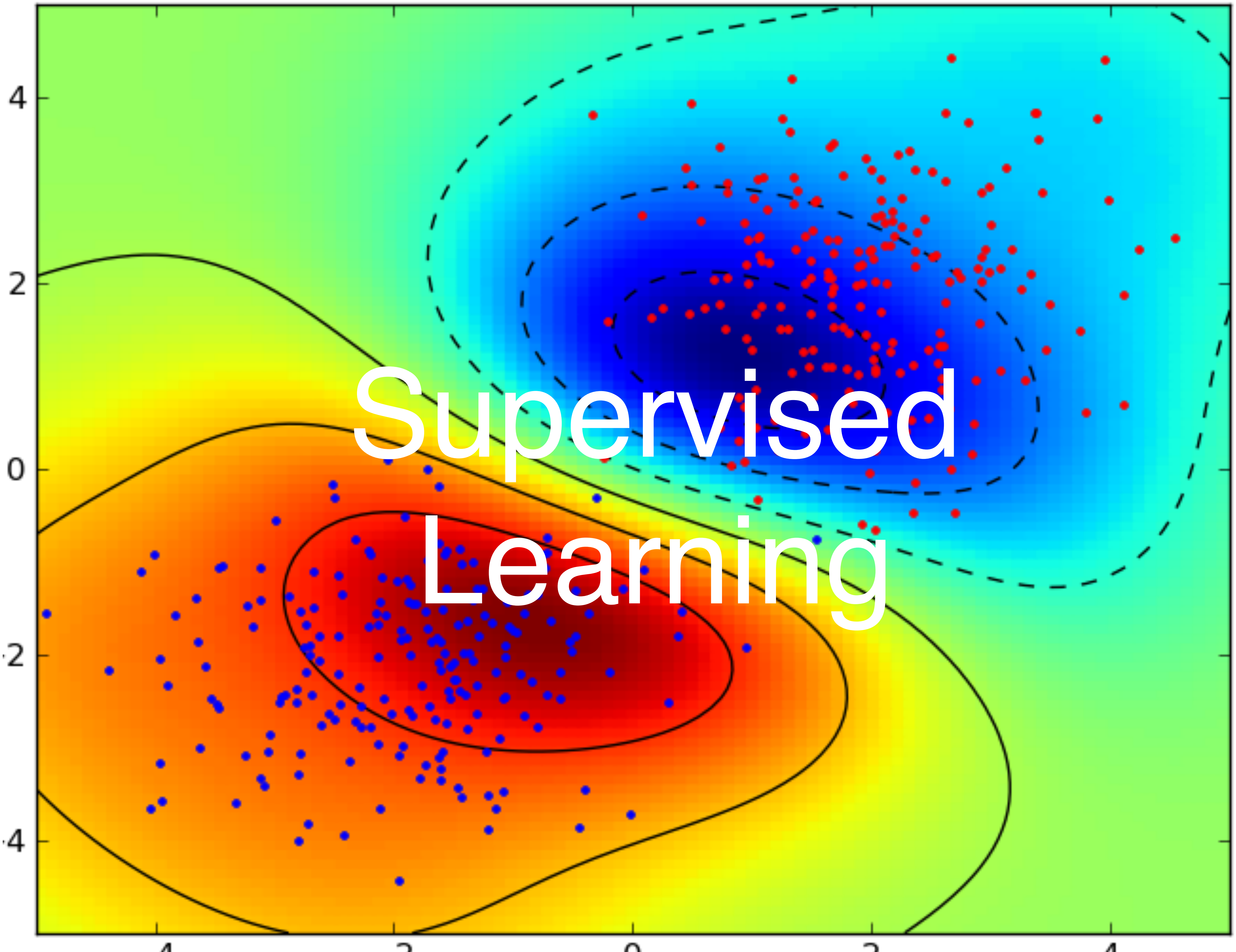
True startup story

- 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)



ty

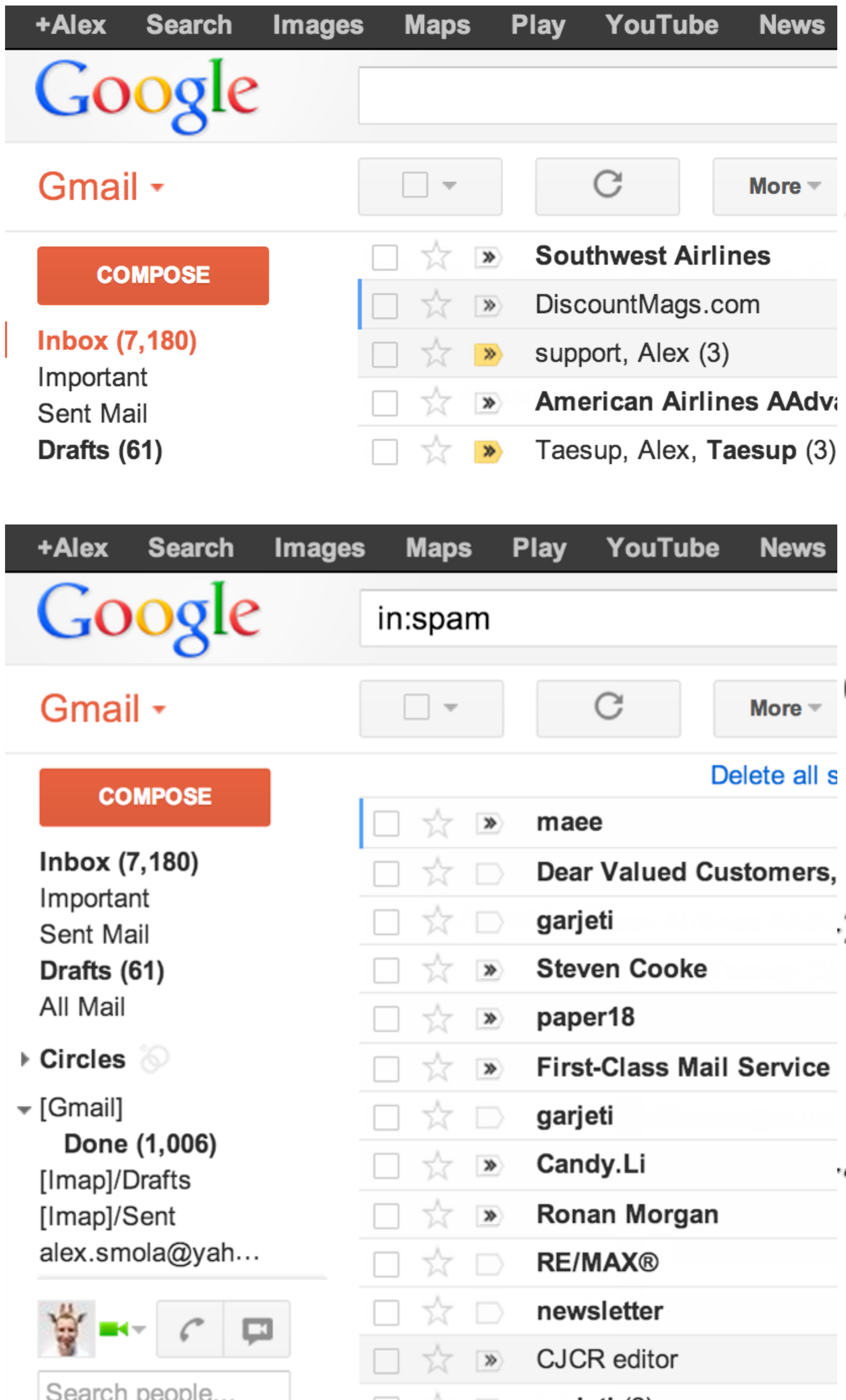
Supervised Learning $y = f(x)$

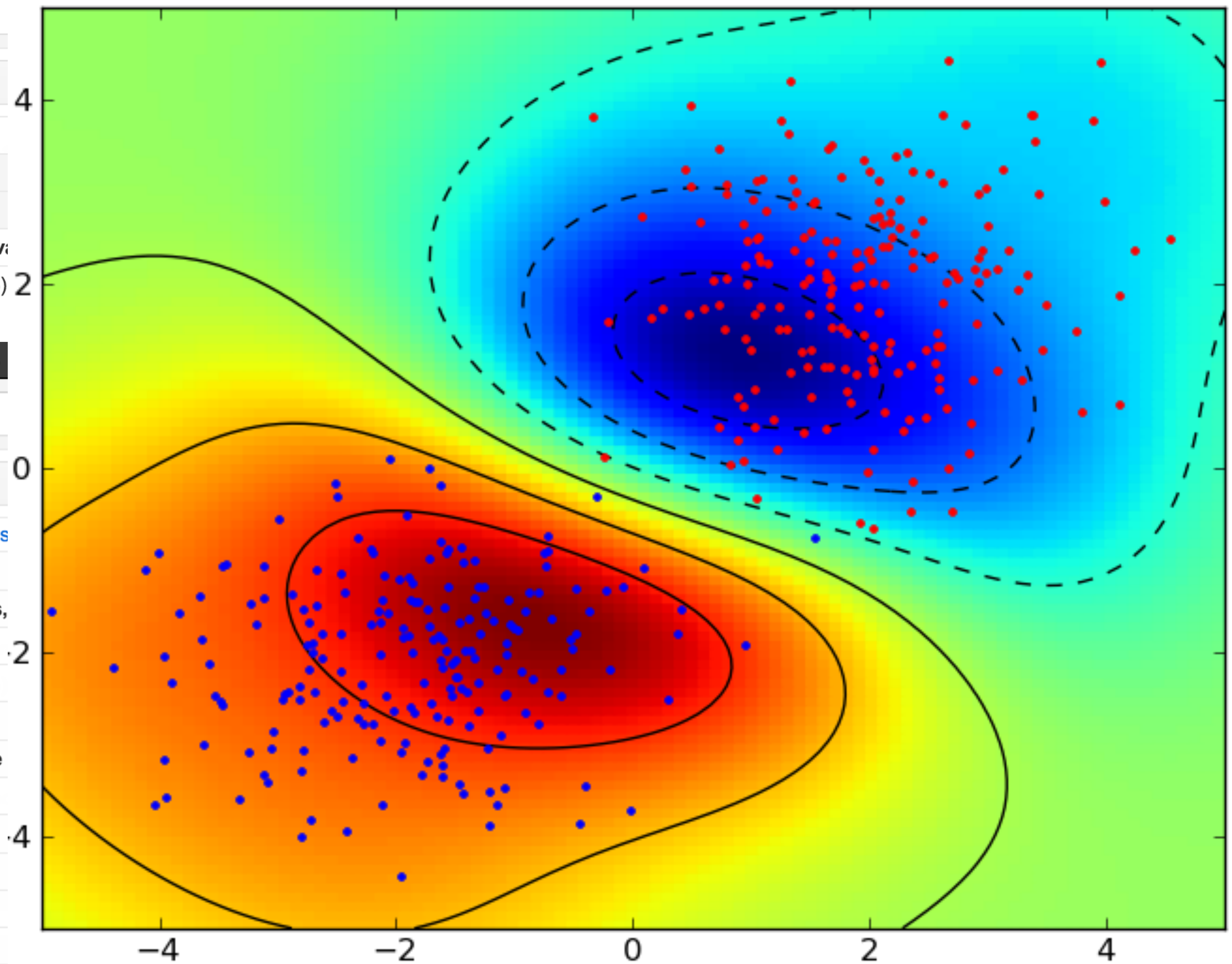
- Binary classification
  Given x find y in {-1, 1}
- Multicategory classification
  Given x find y in {1, ... k}
- Regression
  Given x find y in R (or R^d)
- Sequence annotation
  Given sequence x_1 ... x_l find y_1 ... y_l
- Hierarchical Categorization (Ontology)
  Given x find a point in the hierarchy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

often with loss

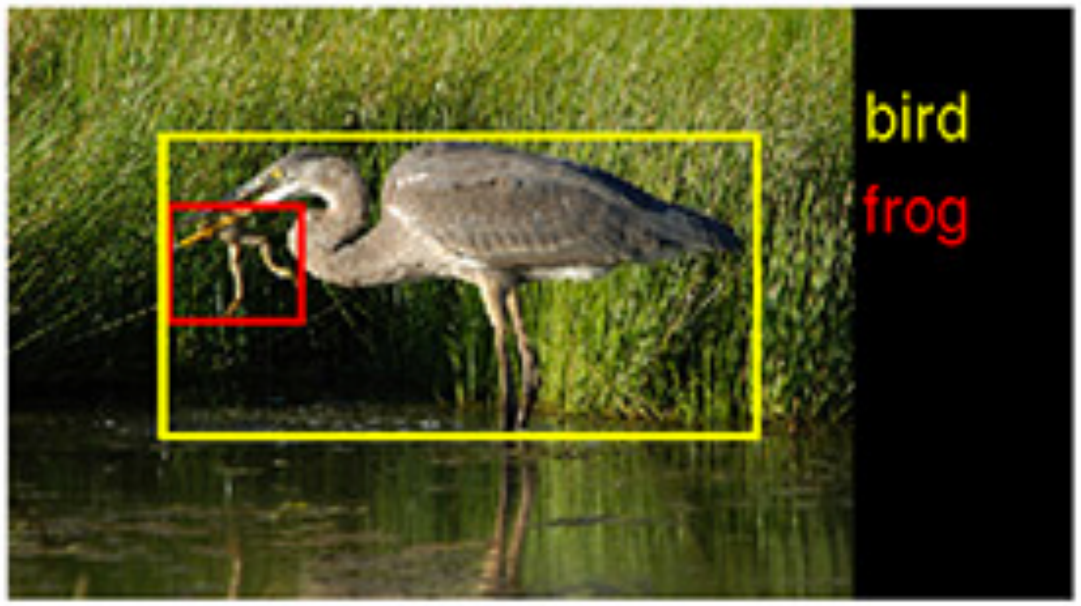
$$l(y, f(x))$$

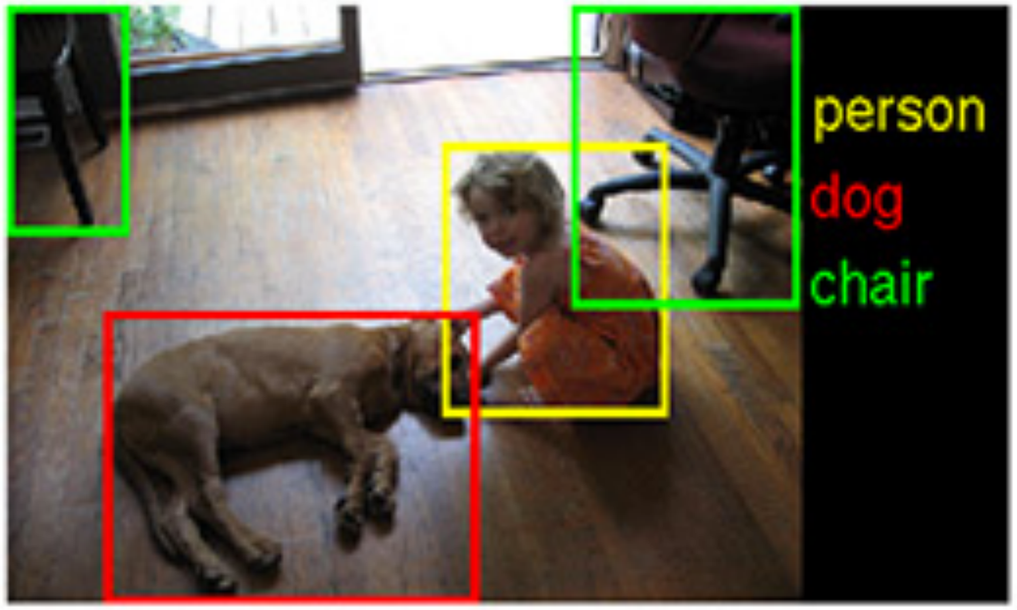
Binary Classification

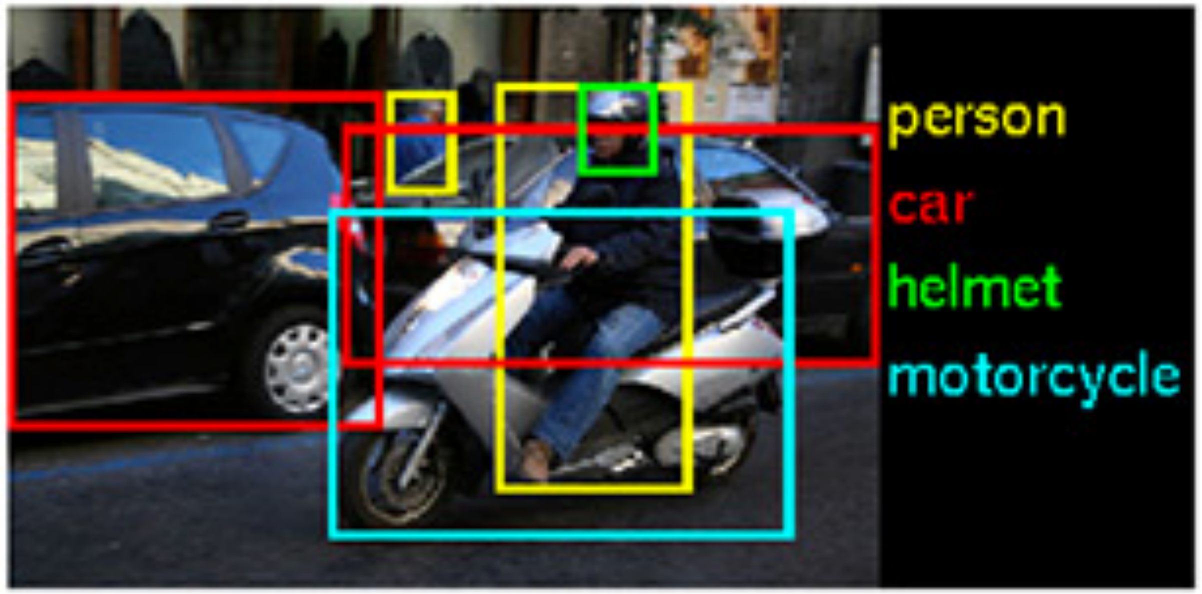




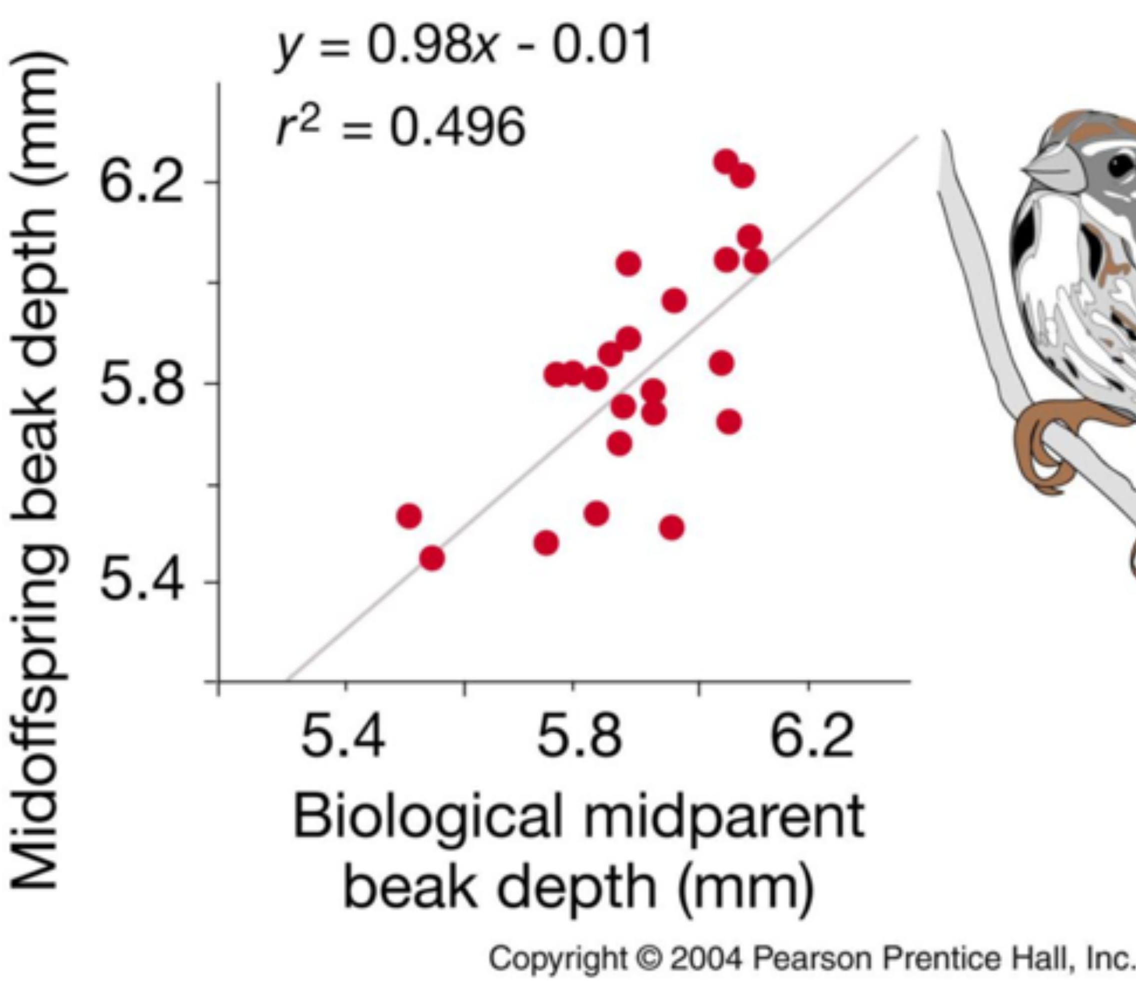
Multiclass Classification + Annotation



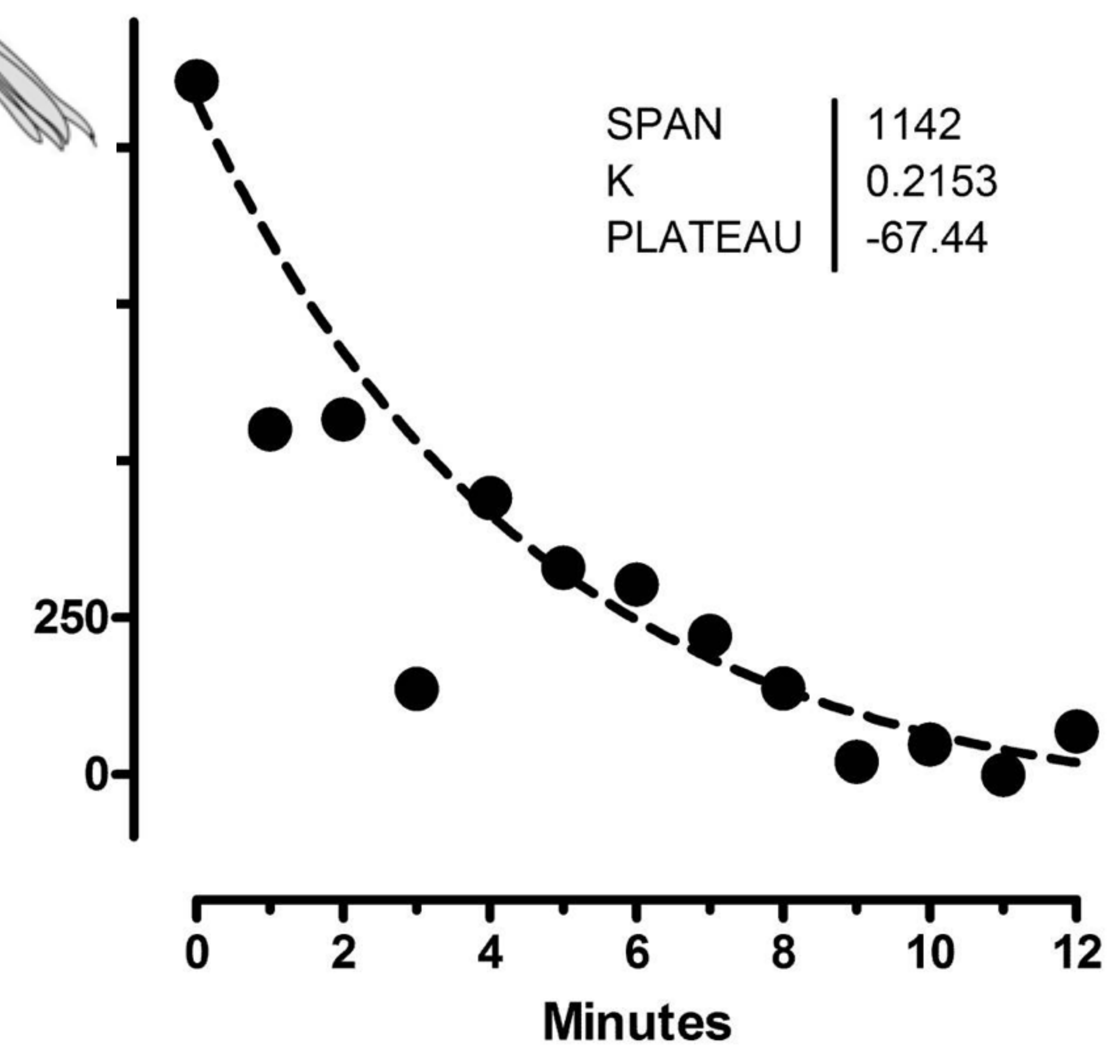




Regression

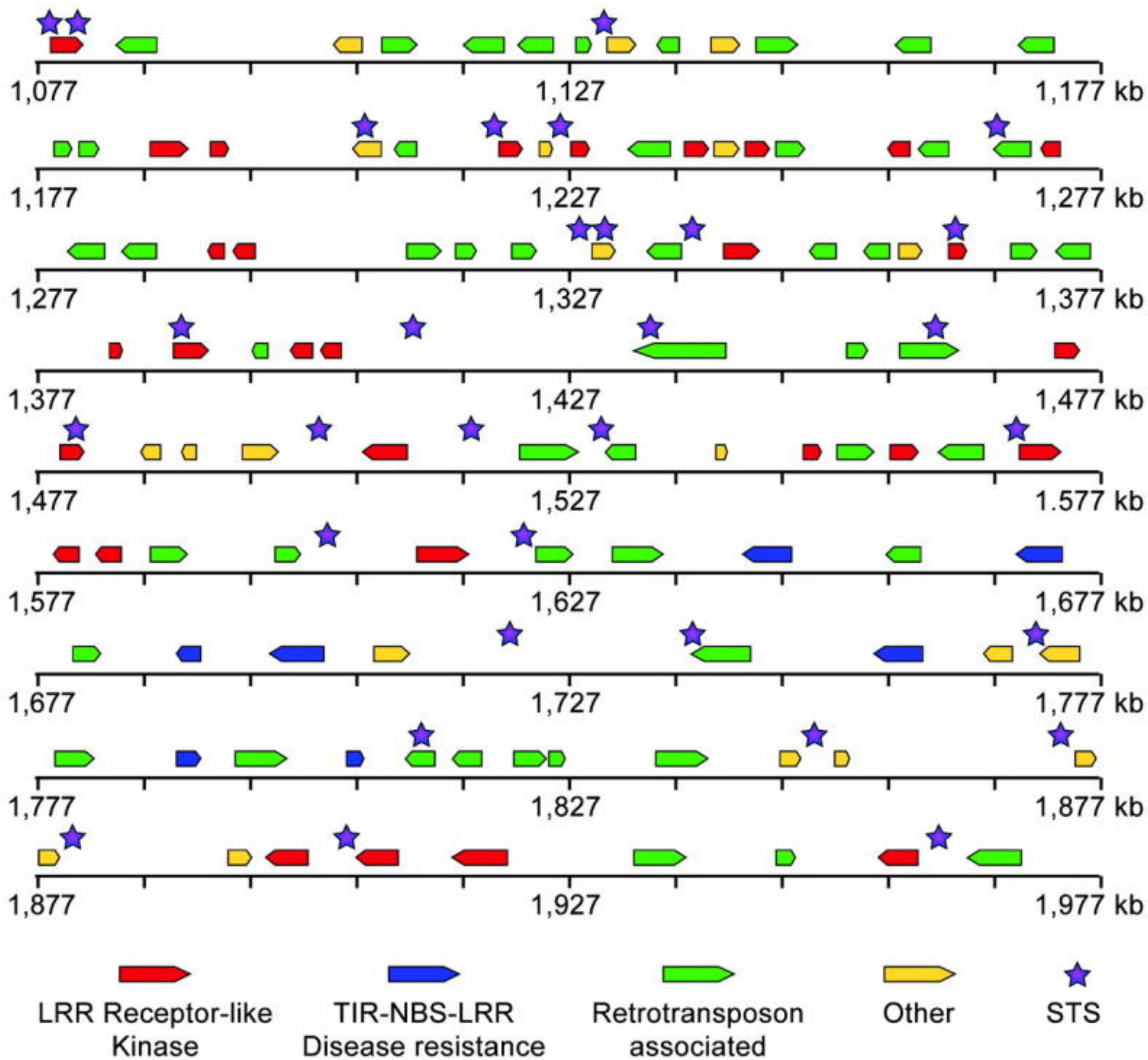


nonlinear



linear

Sequence Annotation

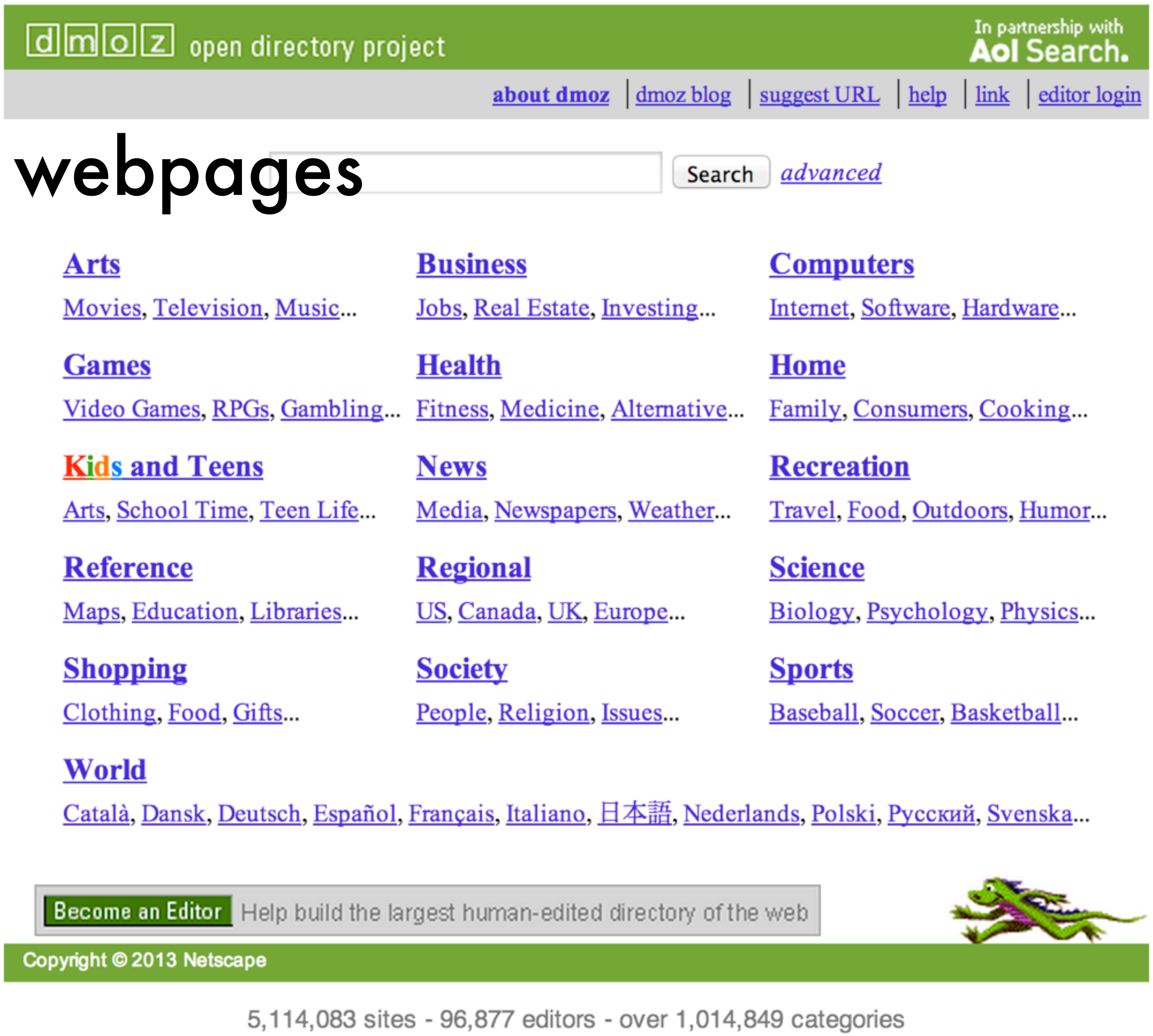


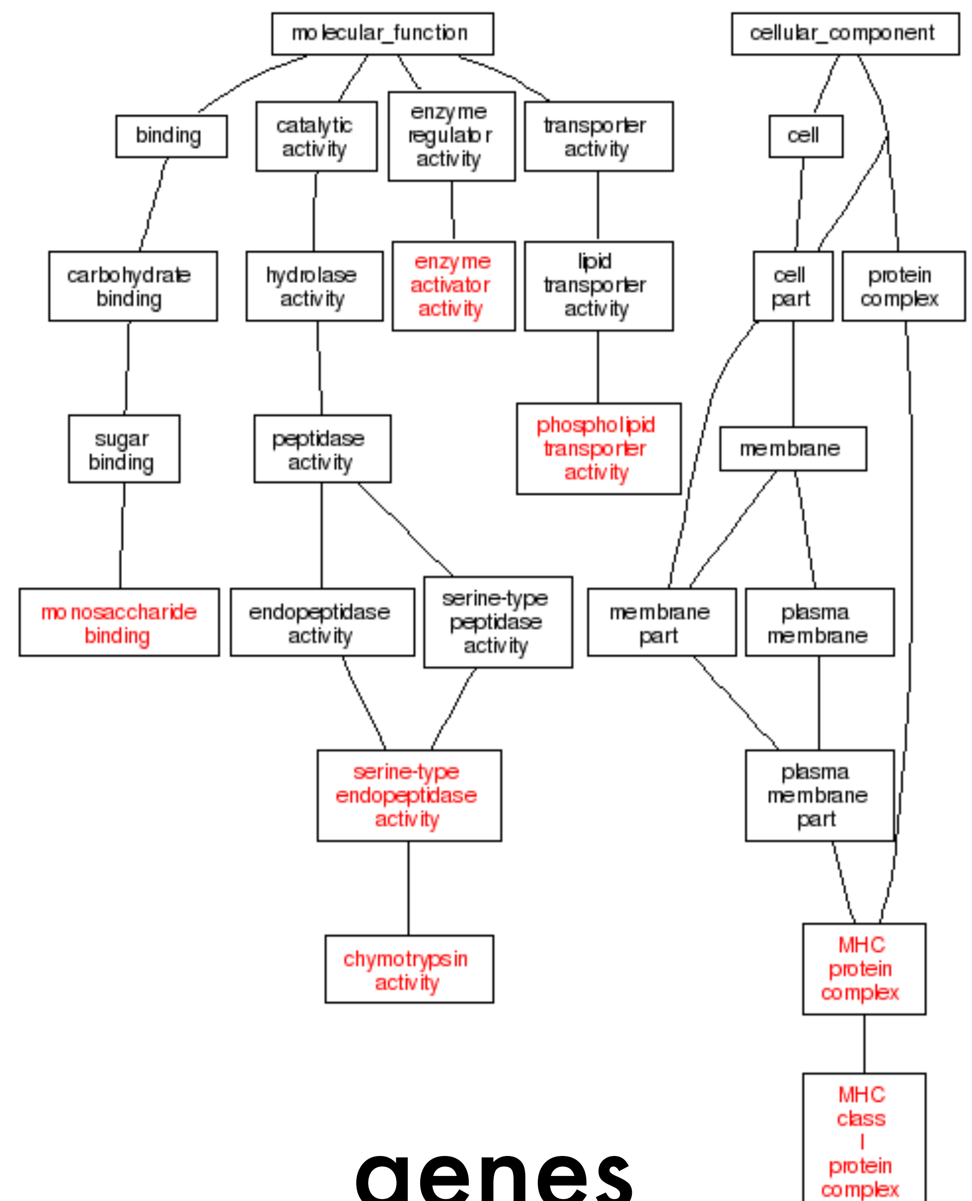
given sequence

gene finding
speech recognition
activity segmentation
named entities

Ontology

webpages





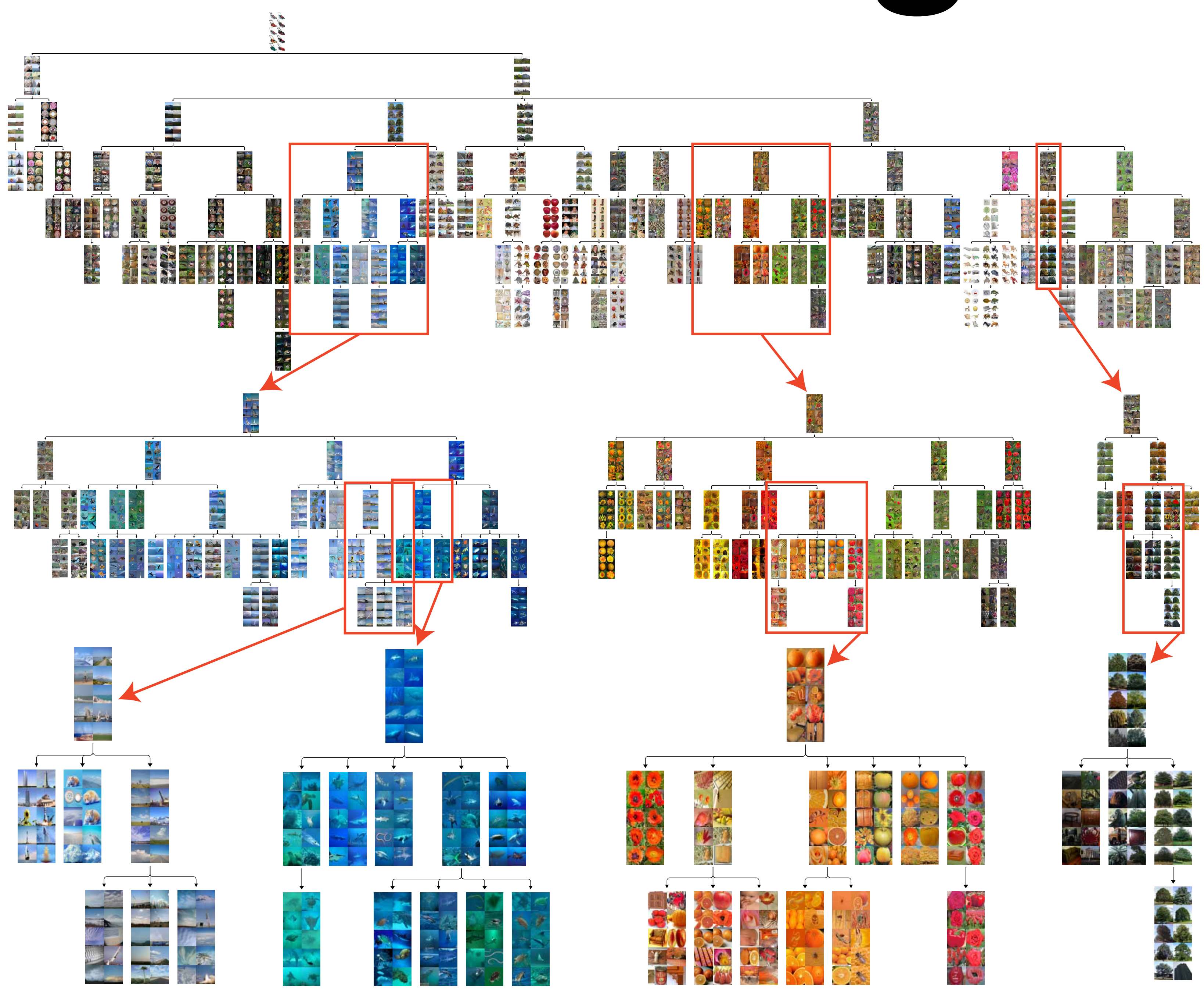
genes

Prediction

tomorrow's stock price

Carnegie Mellon University

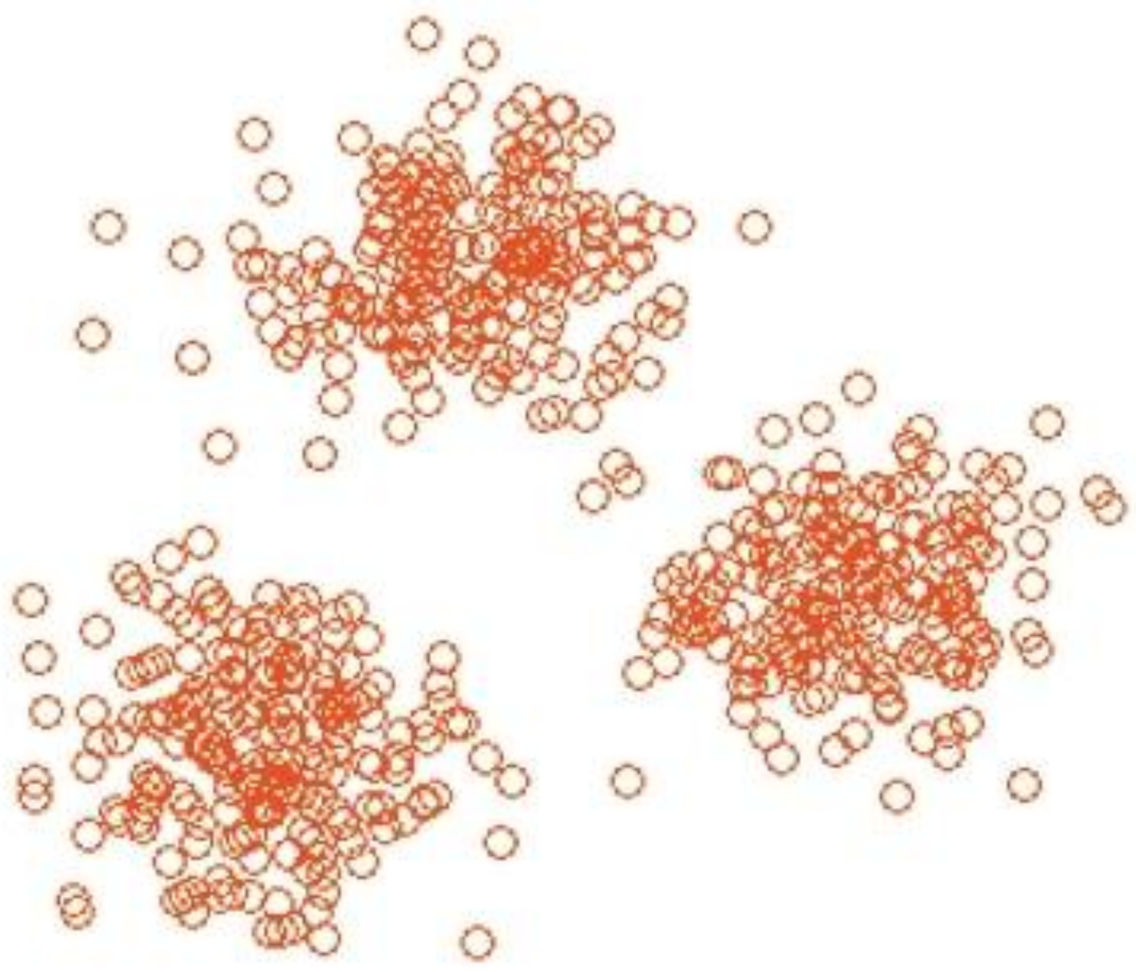
Unsupervised Learning



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

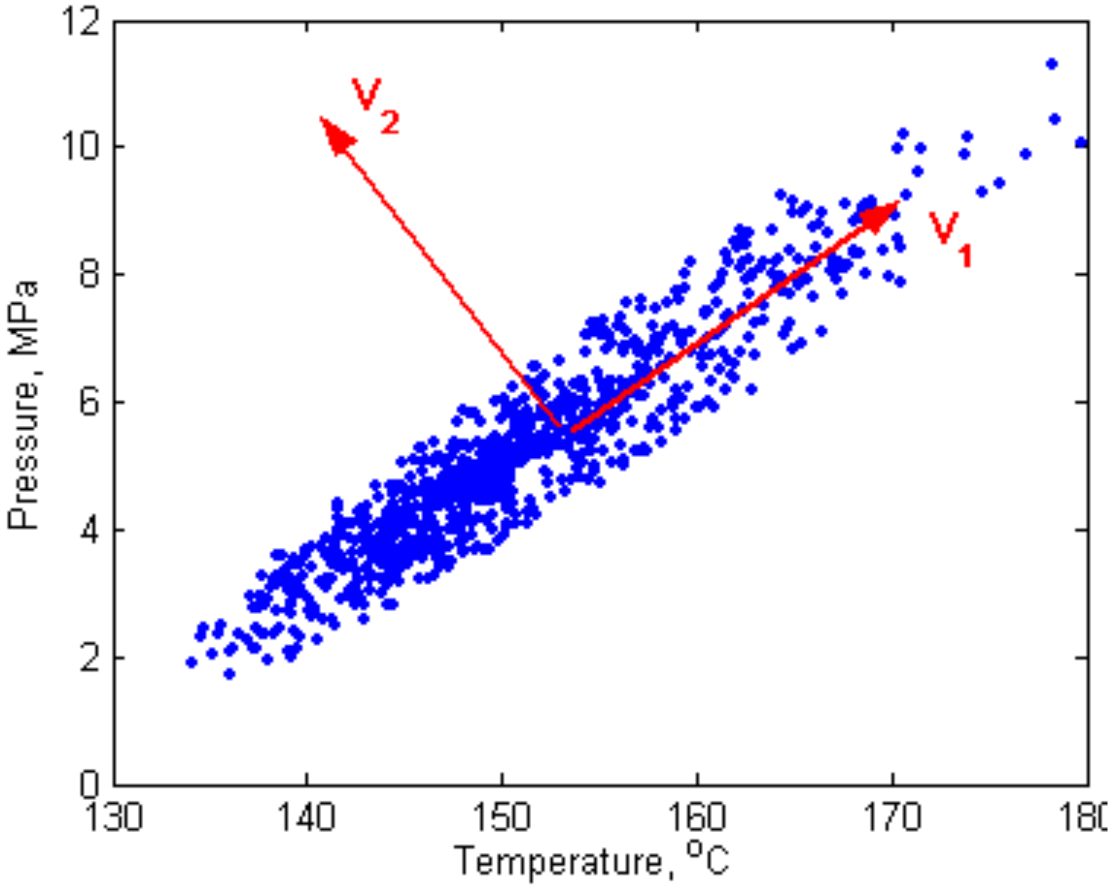
Clustering

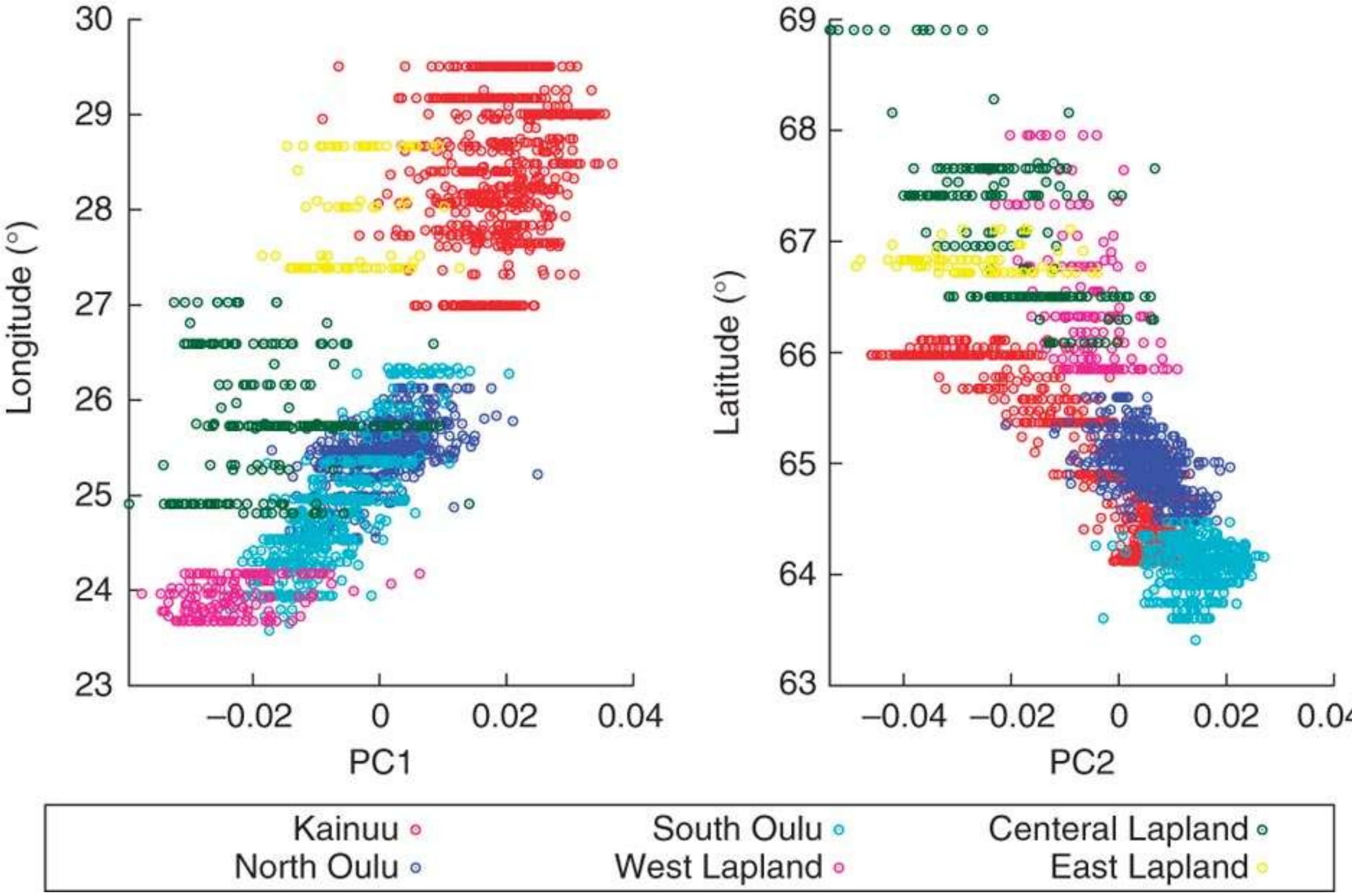


- Documents
- Users
- Webpages
- Diseases
- Pictures
- Vehicles

...

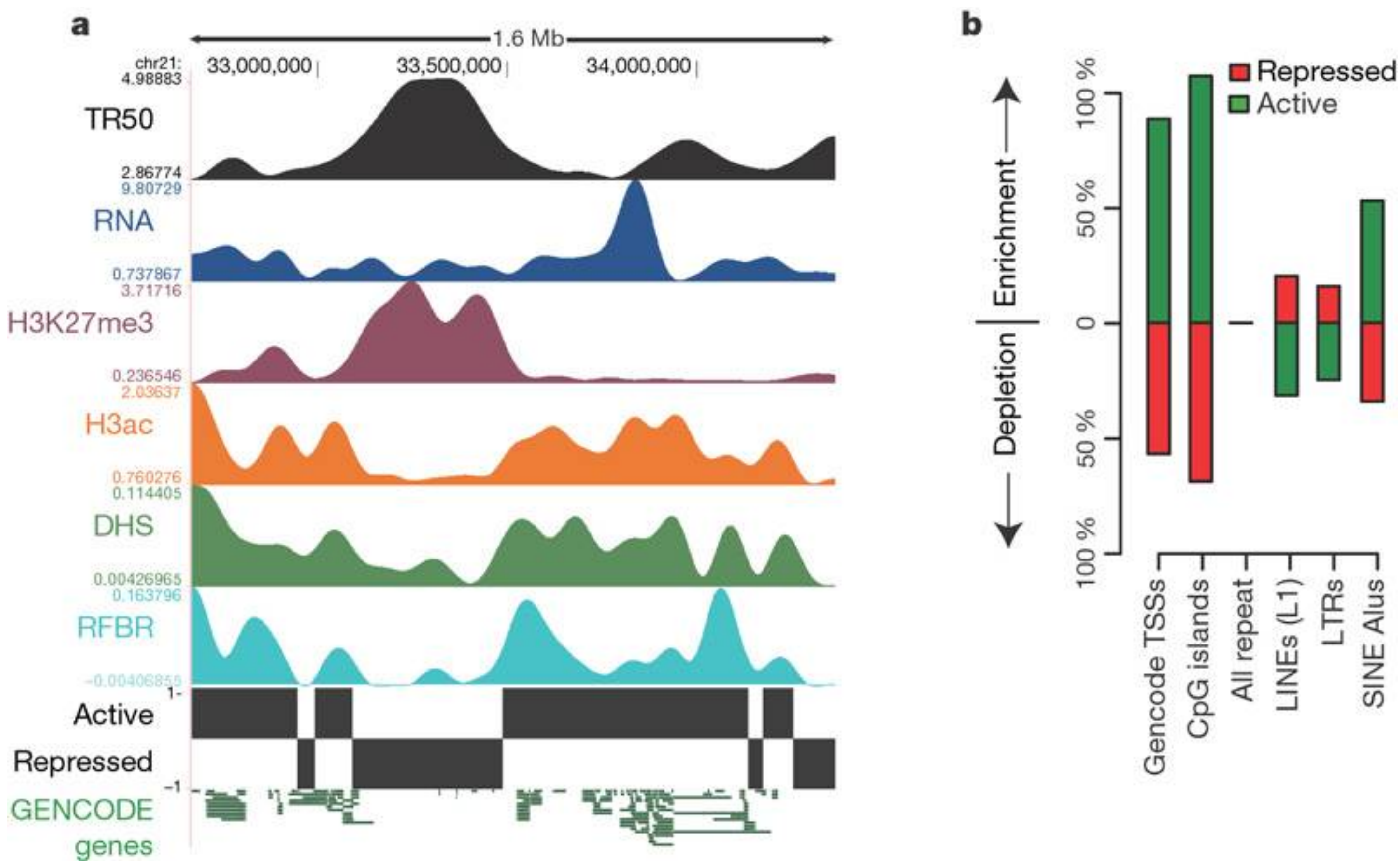
Principal Components





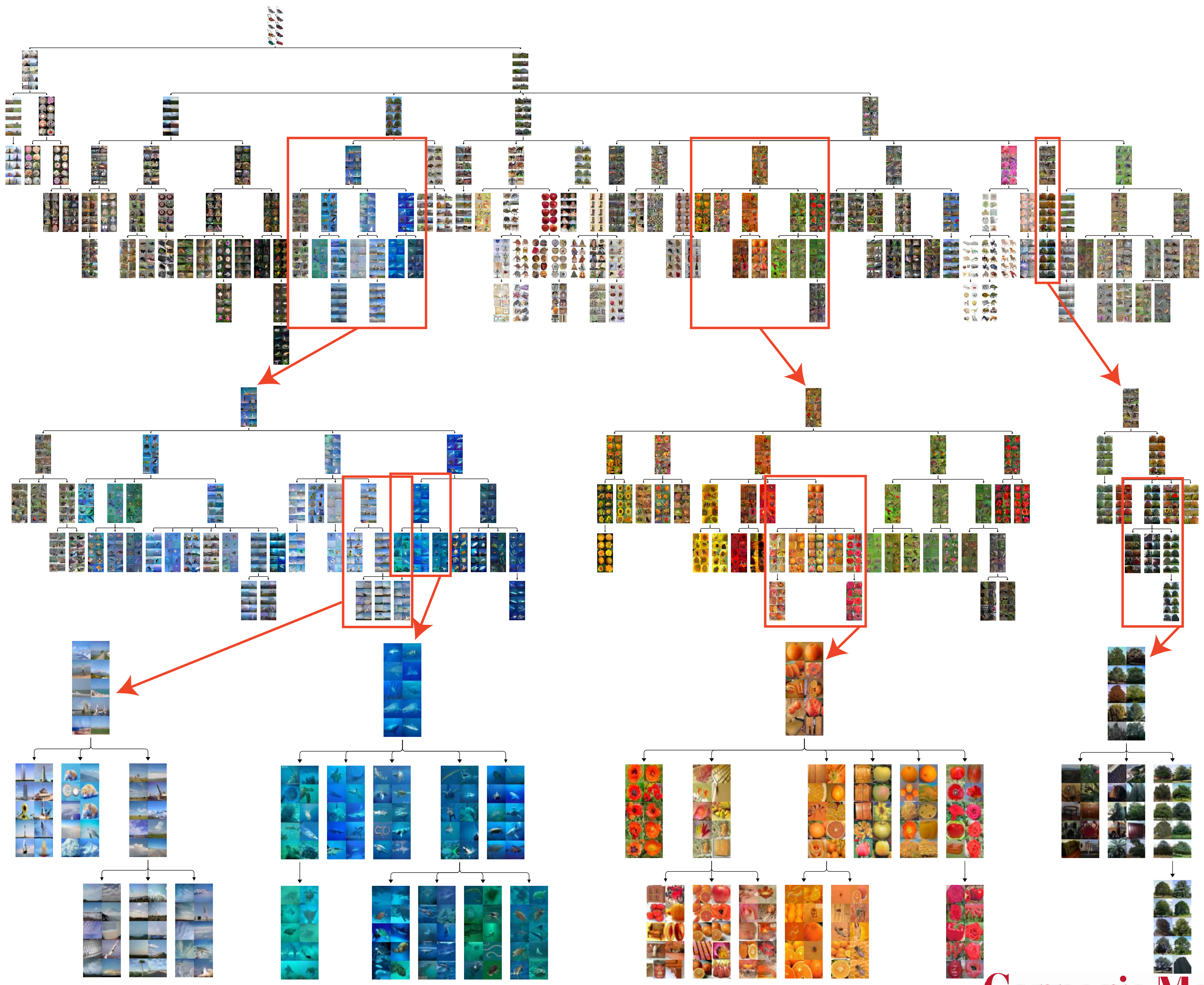
[Variance component model to account for sample structure in genome-wide association studies](), Nature Genetics 2010

Sequence Analysis



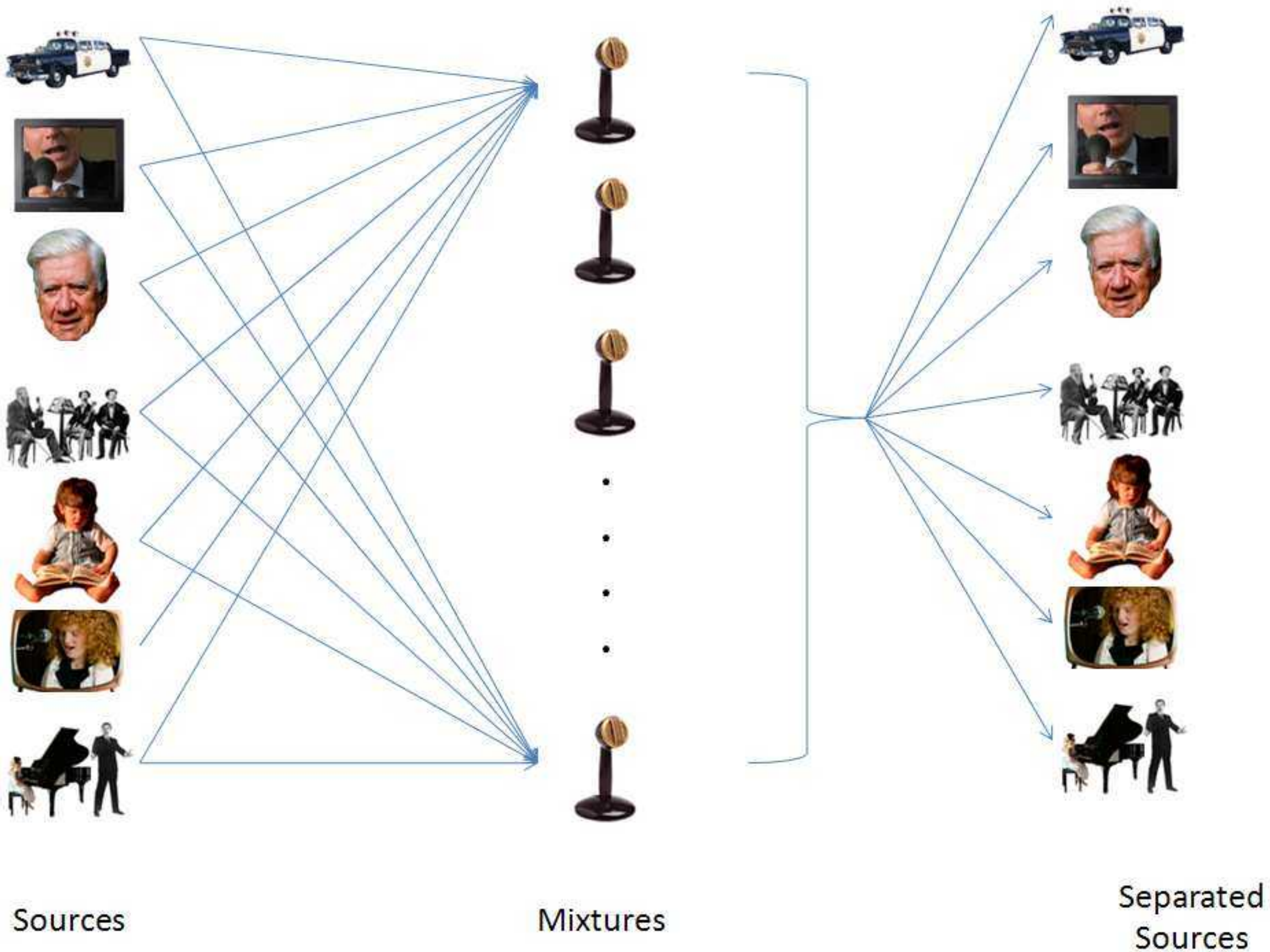
Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project, Nature 2007

Hierarchical Grouping



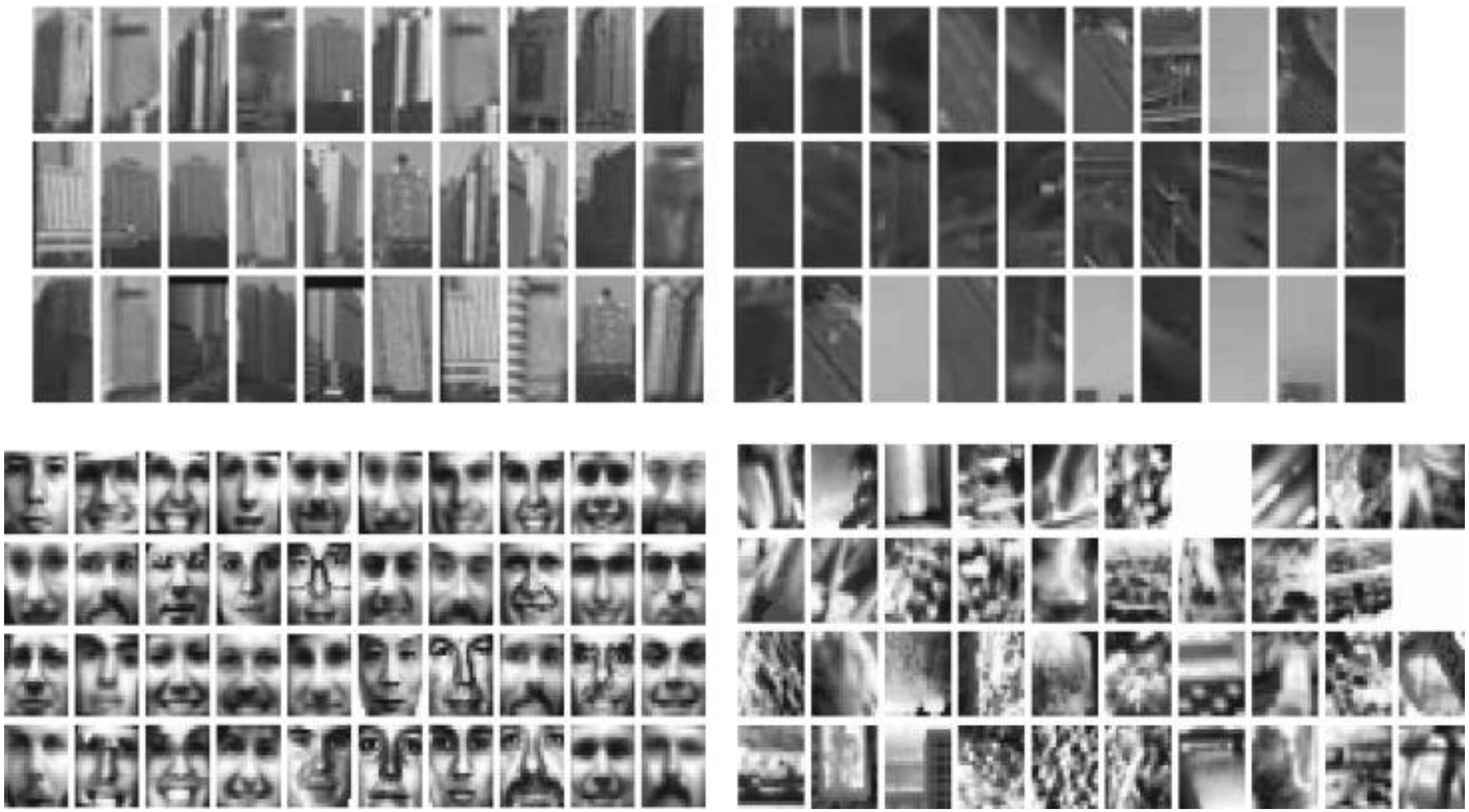
Carnegie Mellon University

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

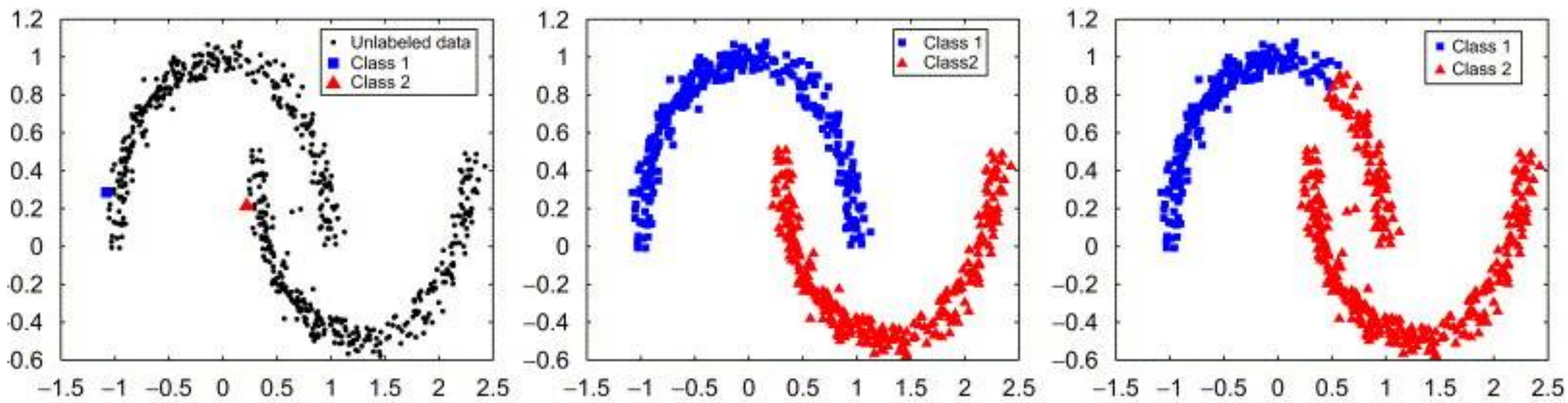
Non-responsive Environment

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

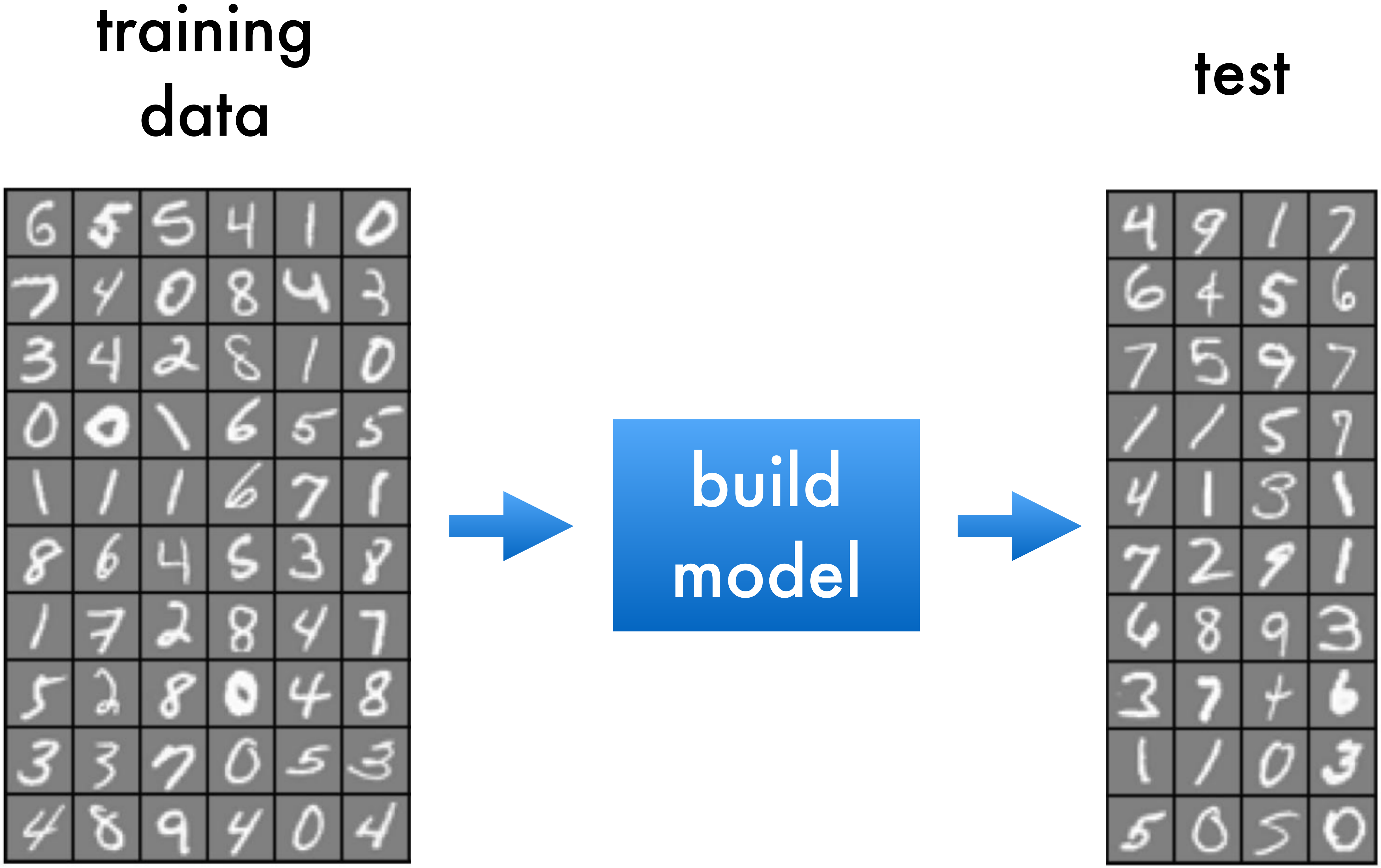
- **Multitask Learning**
  Use correlation between tasks for better result
  - Task 1 - Detect spammy webpages
  - Task 2 - Detect people’s homepages
  - Task 3 - Detect adult content
- **Cotraining**
  For many cases both sets of covariates are available
  - Detect spammy webpages based on page content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


Responsive
Environment

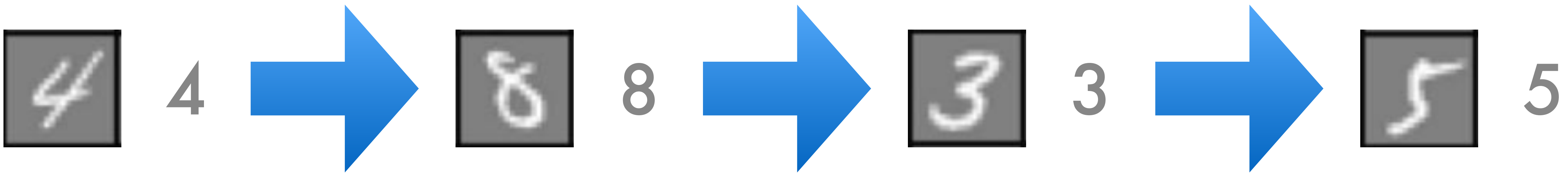
Interaction with Environment

- **Batch** (download a book)
  Observe training data (x_1,y_1) ... (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



Online



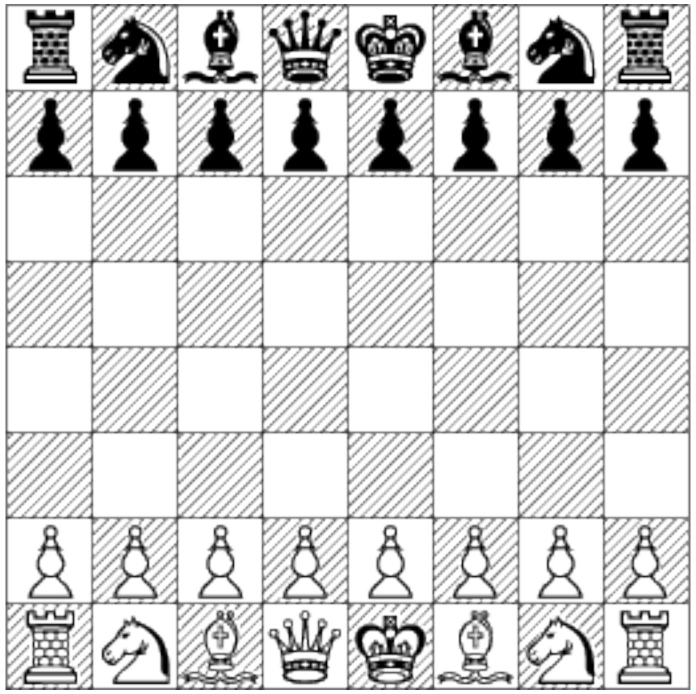
Bandits

- 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)

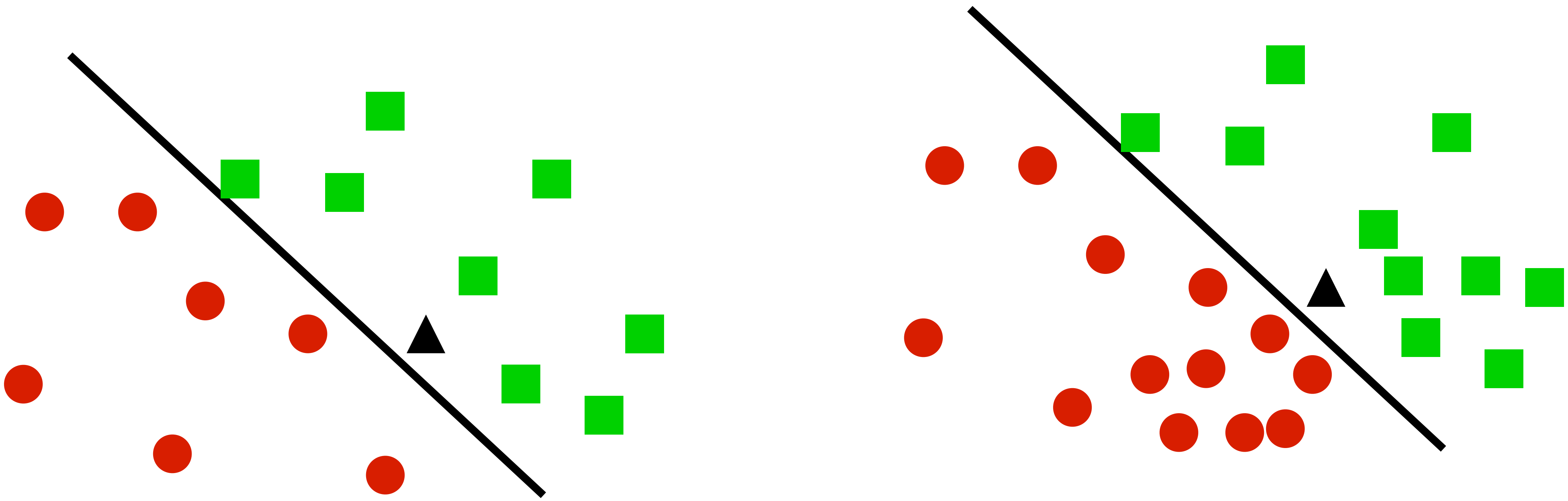
Discriminative and Generative Models

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

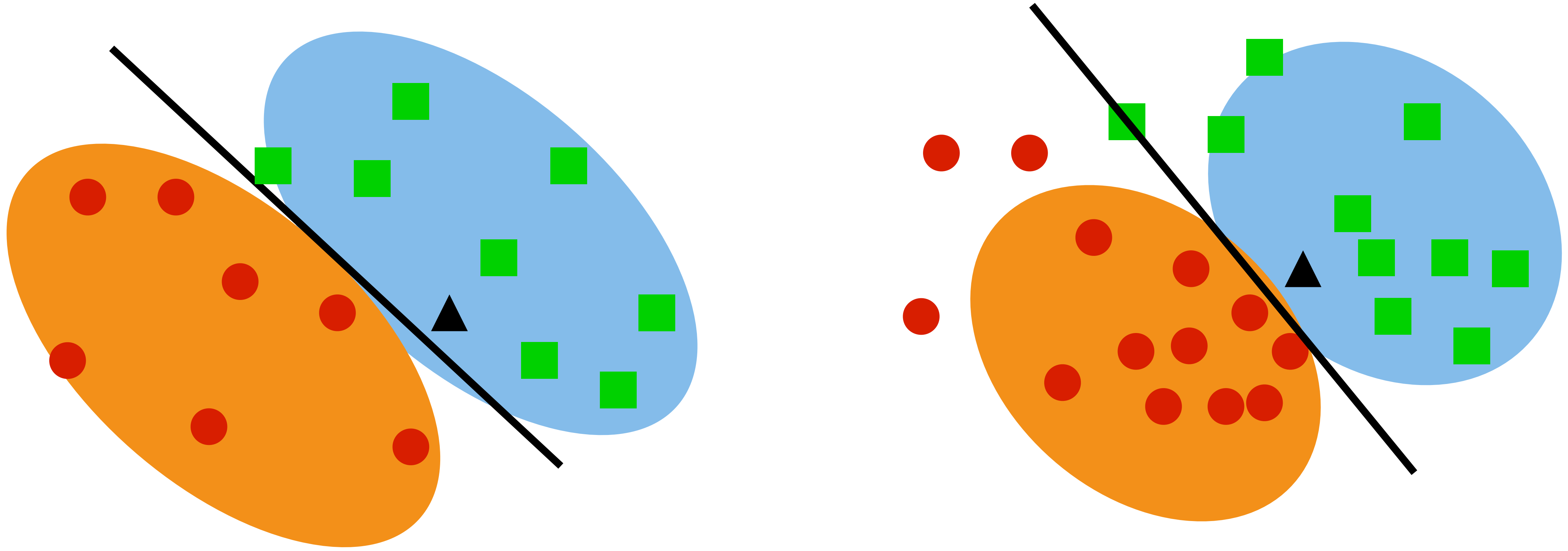
$$p(y|x) = \frac{p(y, x)}{p(x)}$$

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

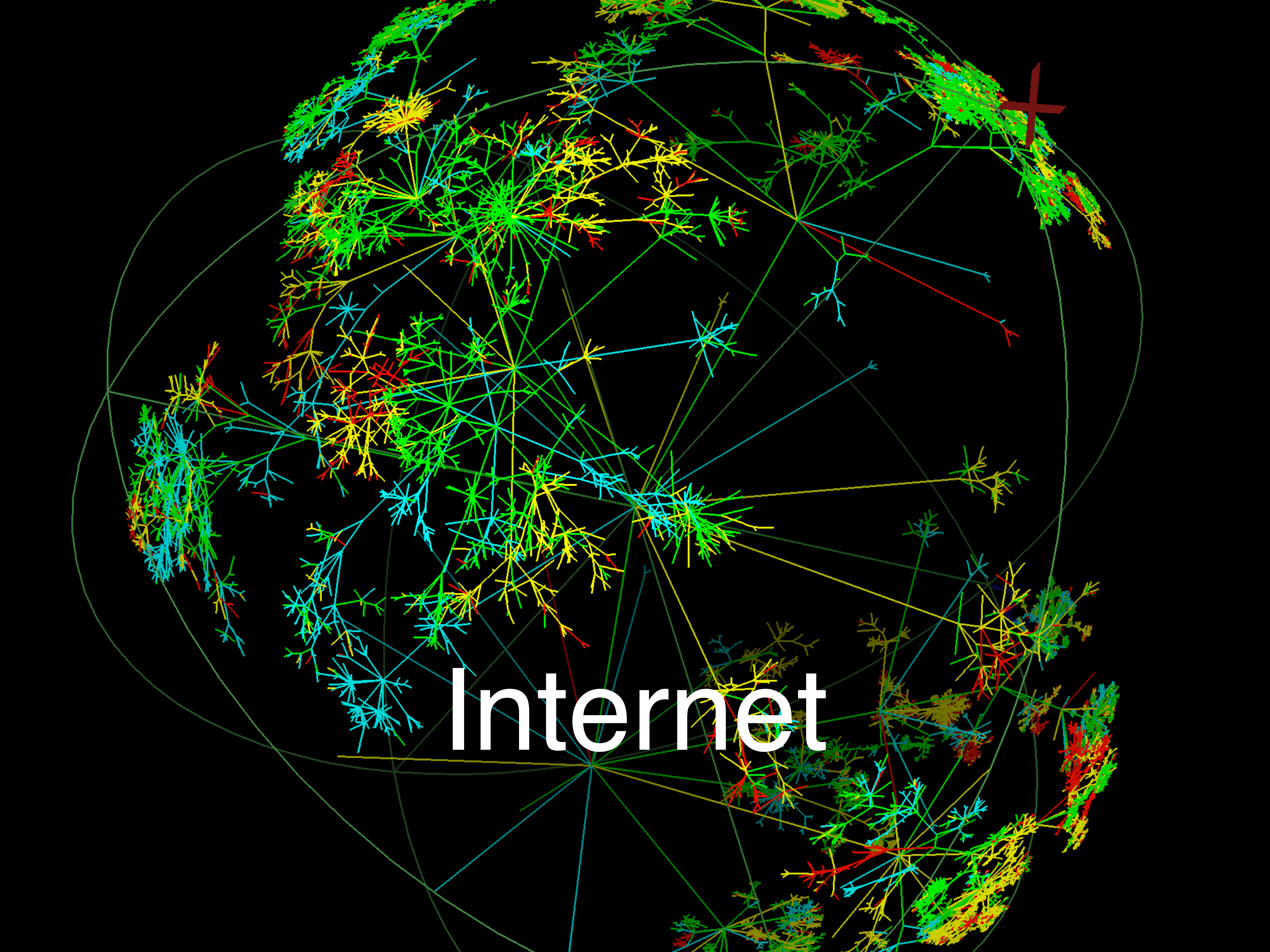
1.4 Data

1 Introduction

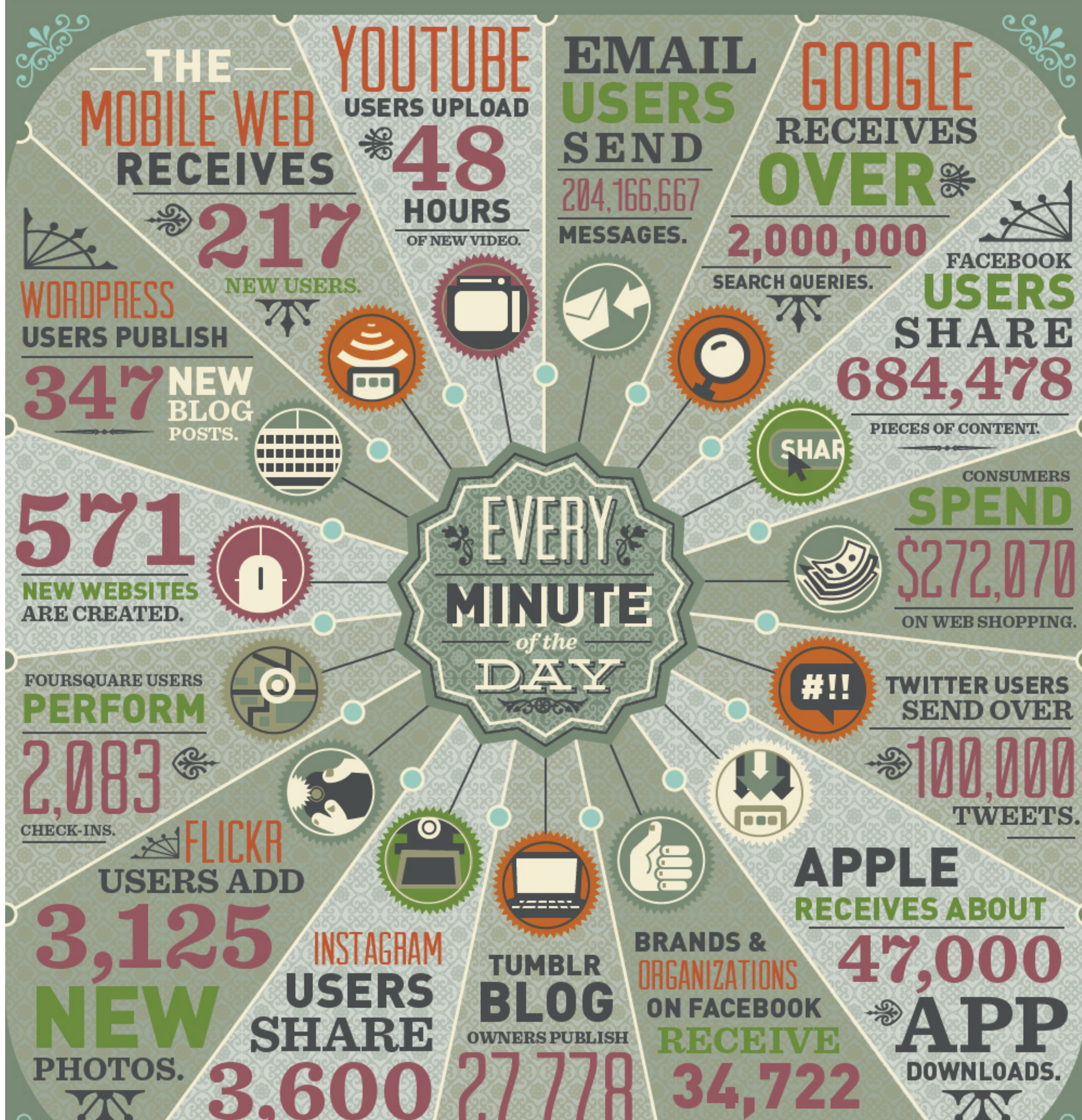
Alexander Smola
Introduction to Machine Learning 10-701
http://alex.smola.org/teaching/10-701-15

MACHINE LEARNING
DEPARTMENT

Carnegie Mellon University

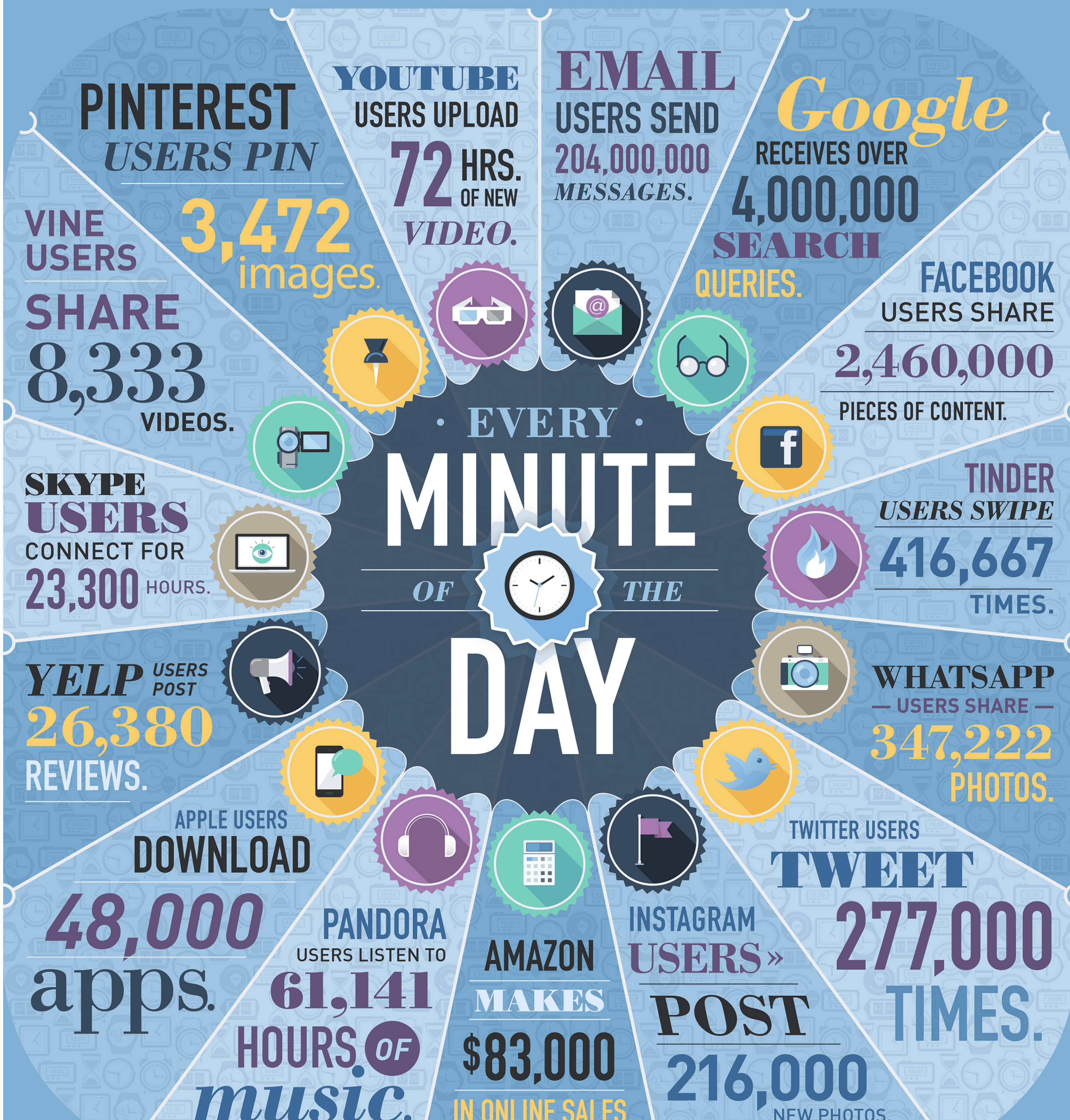

Internet

Data per minute 2012



http://www.domo.com/learn/infographic-data-never-sleeps

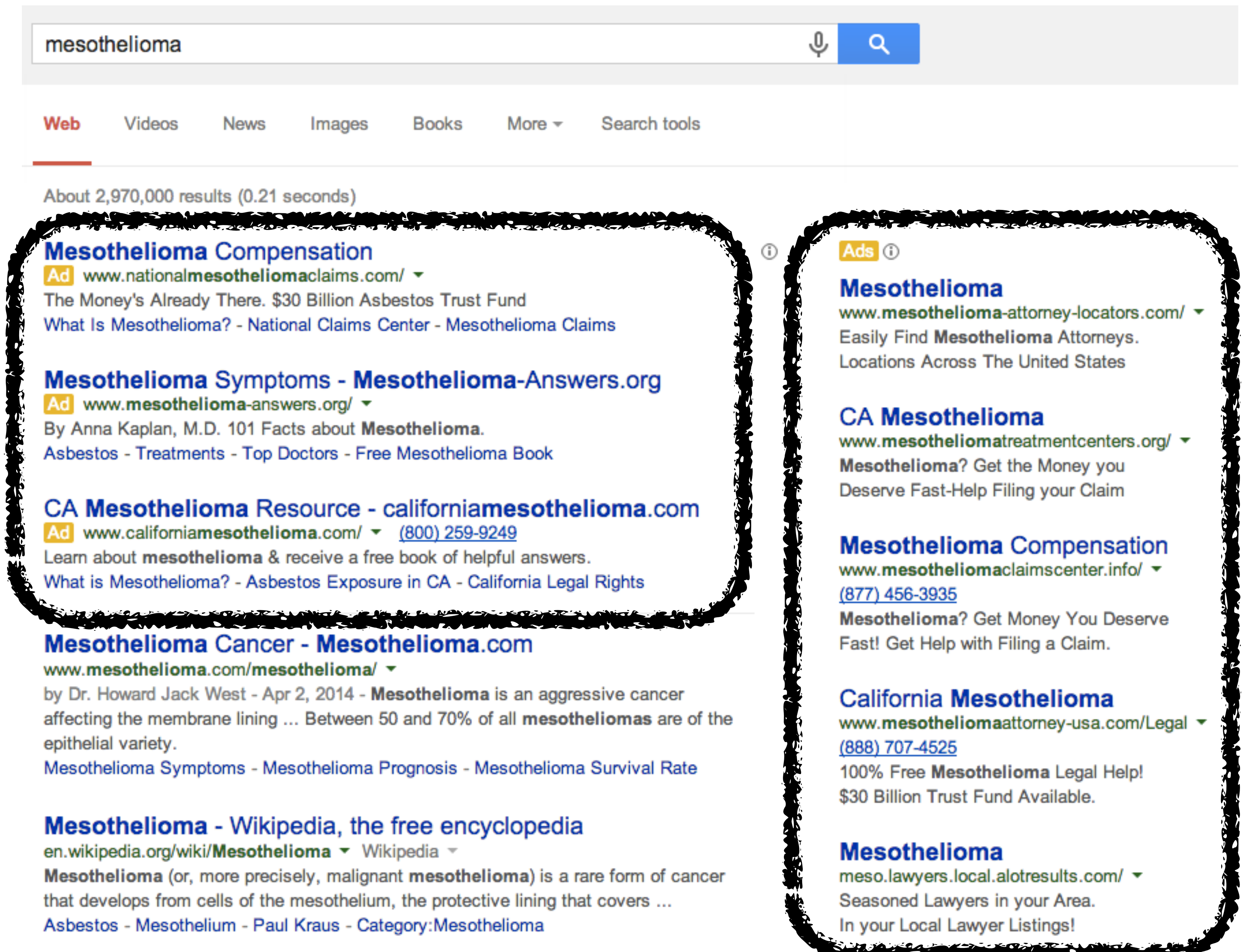
Data per minute 2014



http://www.domo.com/learn/data-never-sleeps-2

Google

Computational Advertising



Google

sponsored search picks position of ad using

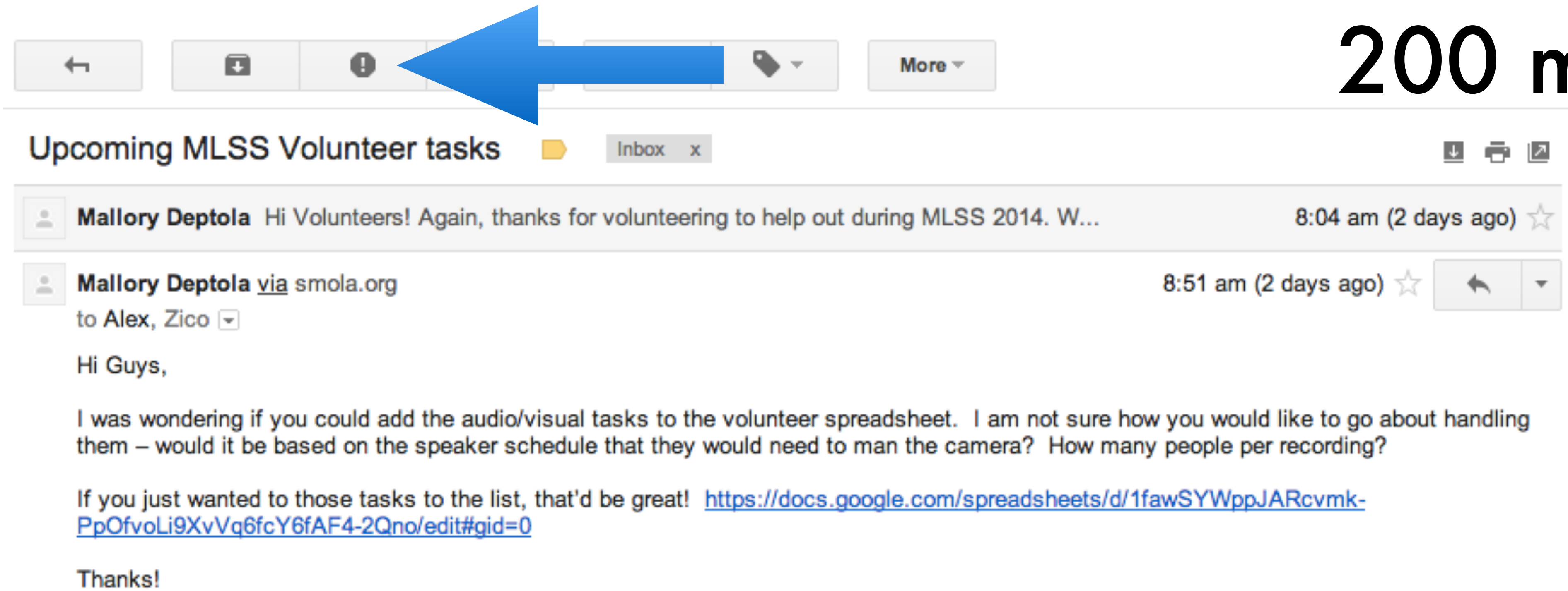
$$p(\text{click}|\text{ad}) \cdot \text{bid}(\text{ad})$$

estimate it

4 million/minute

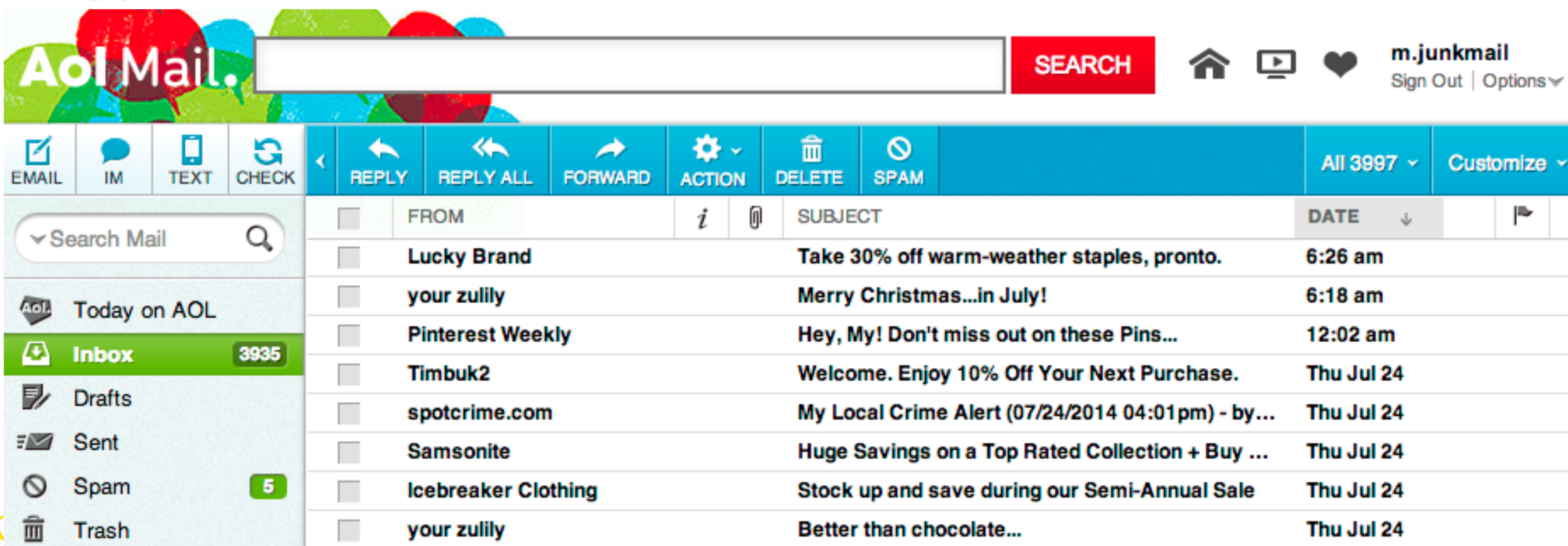
Carnegie Mellon University

Spam filtering

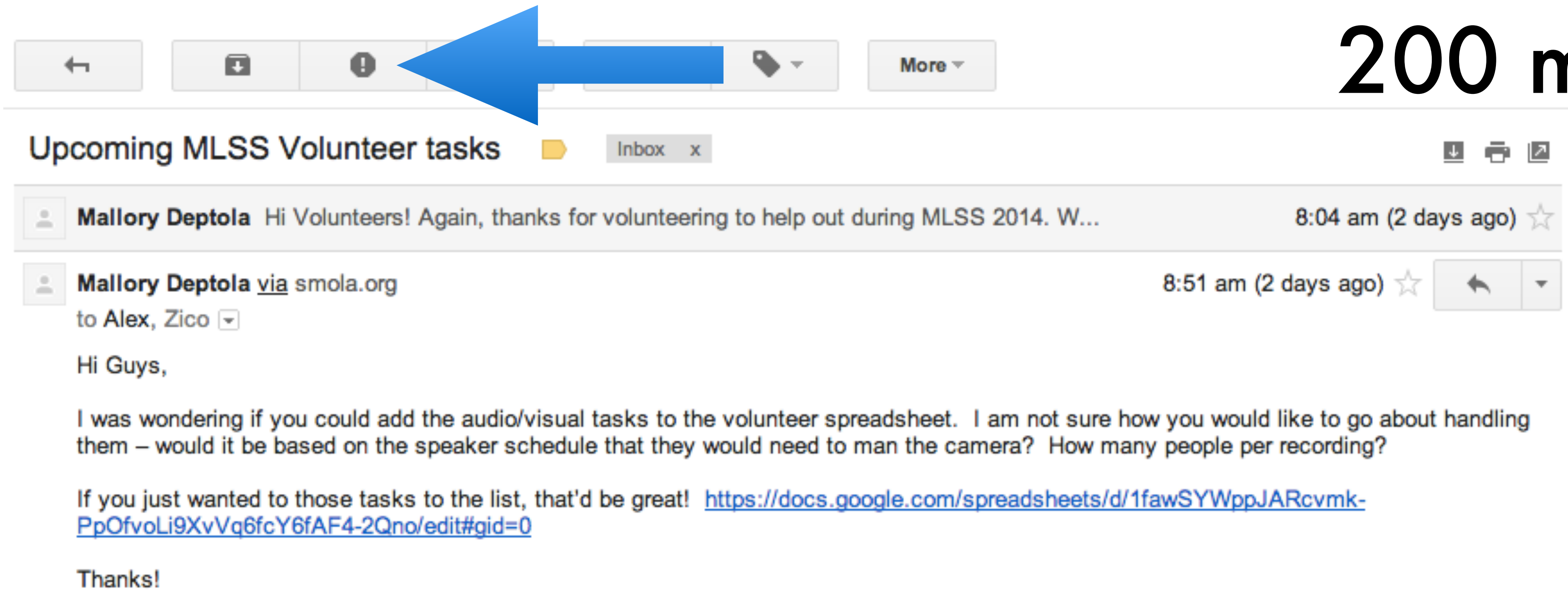


200 million/minute

imbalanced dataset

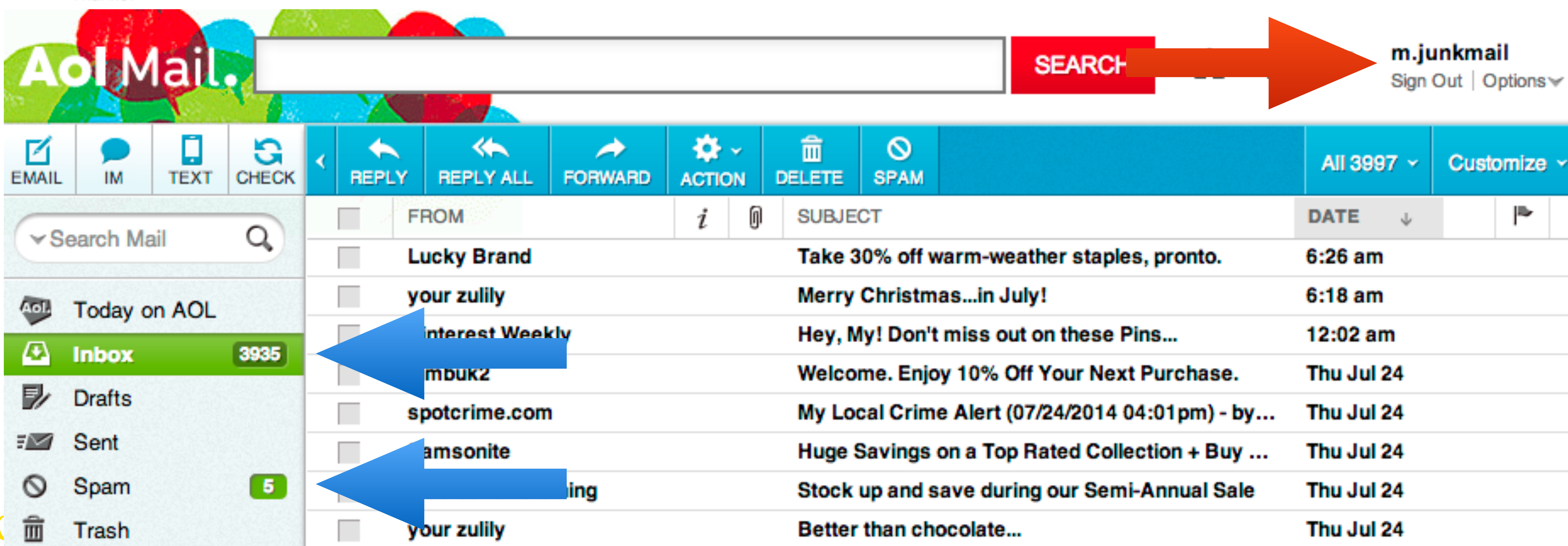


Spam filtering



200 million/minute

imbalanced dataset



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)

>1B images, 40h video/minute

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

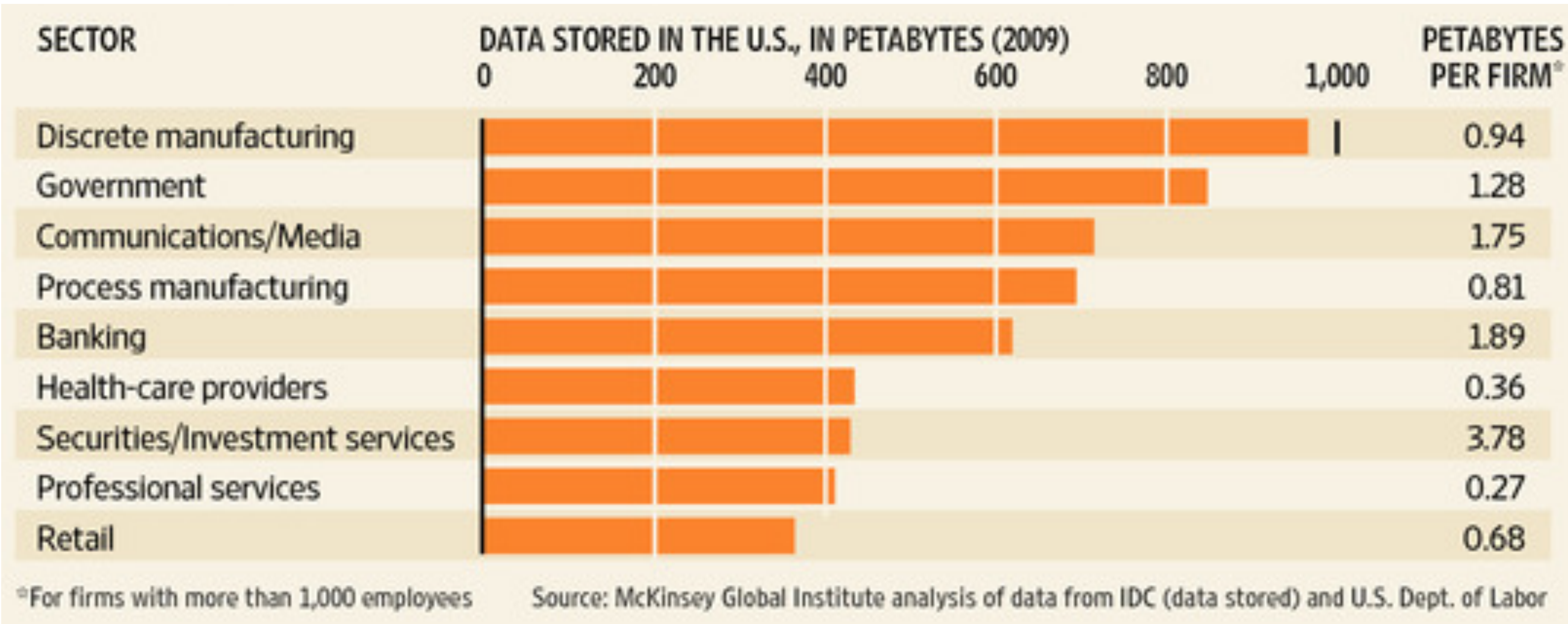
DISQUS

YouTube

yelp

>1B images, 40h video/minute

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-1B vertices

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)

>1B texts

Data - Messages

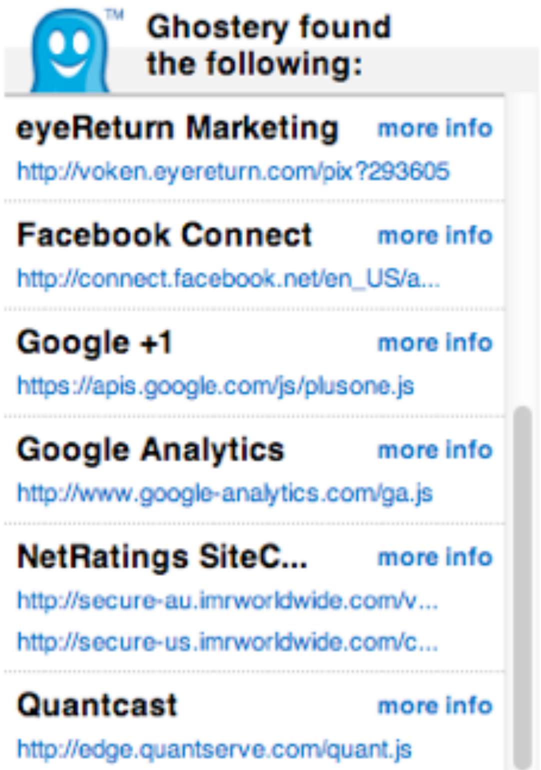
- Webpages (content, graph)
- Clicks (ad, page, social)
- Users (OpenID, FB Connect)
- e-mails (Hotmail, Y!Mail, Gmail)
- Photos, Movies (Flickr, YouTube, Vimeo ...)
- Cookies / tracking info (see Ghostery)
- Installed apps (Android market etc.)
- Location (Latitude, Loopt, Foursquared)
- User generated 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)

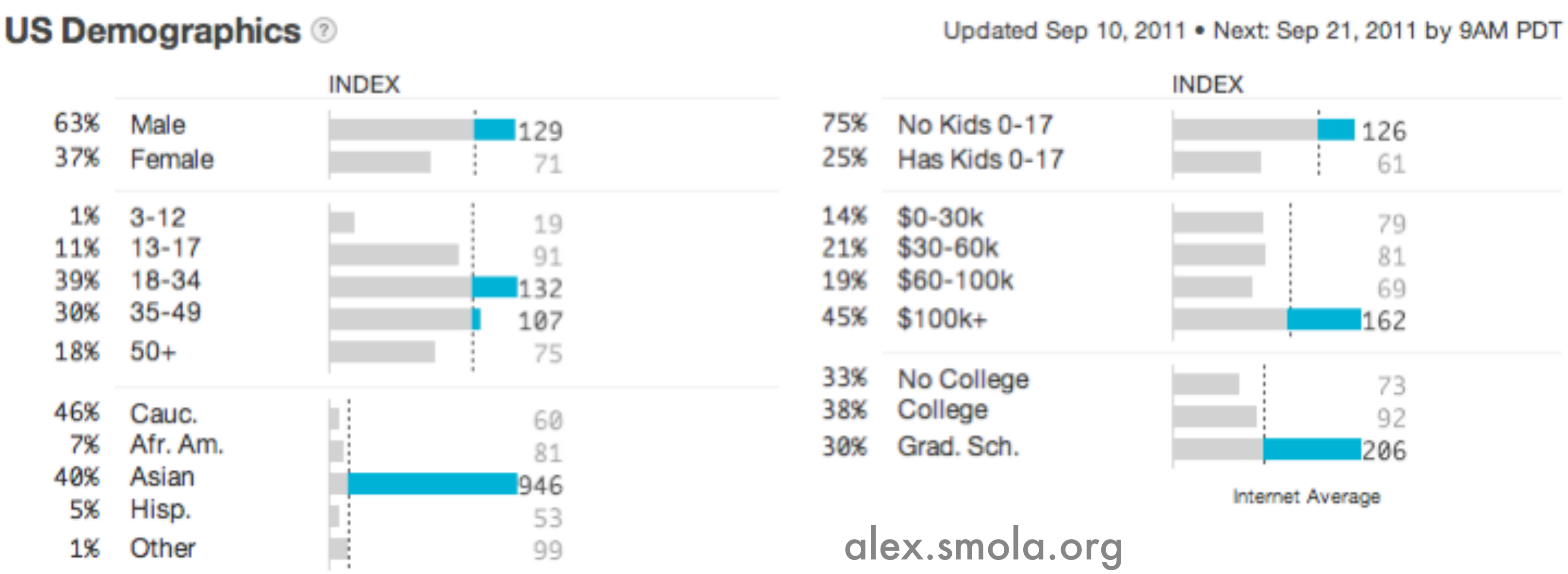
>1B texts

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)





alex.smola.org

>1B '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.

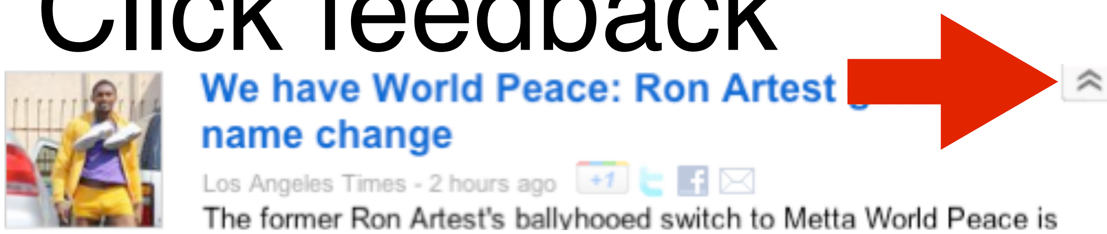
Data Retention:

Undisclosed

(implicit) Labels

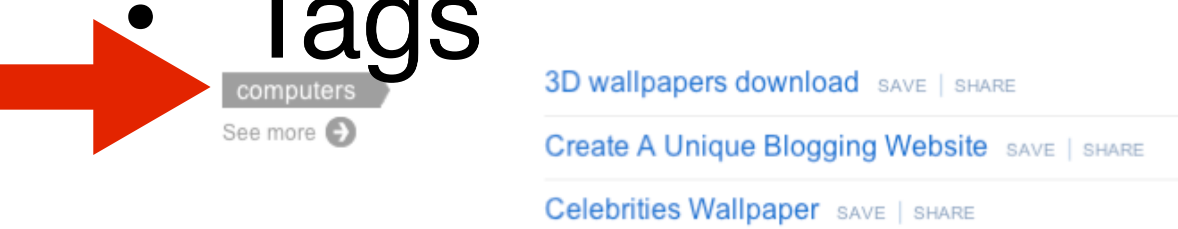
- Ads

- Click feedback



- Emails

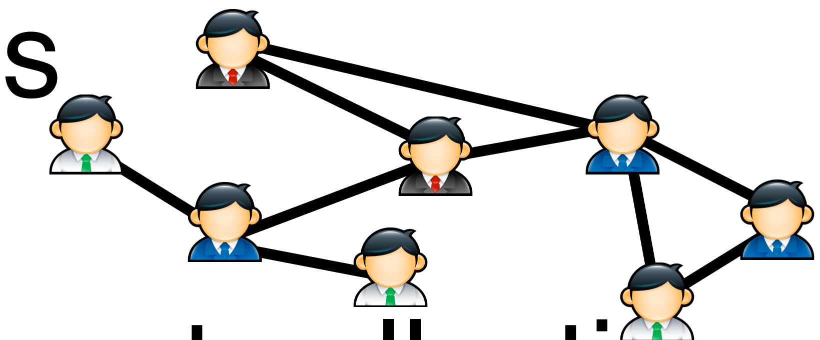
- Tags



- Editorial data is very expensive! Do not use!

no Labels

- Graphs



- Document collections

- Email/IM/Discussions

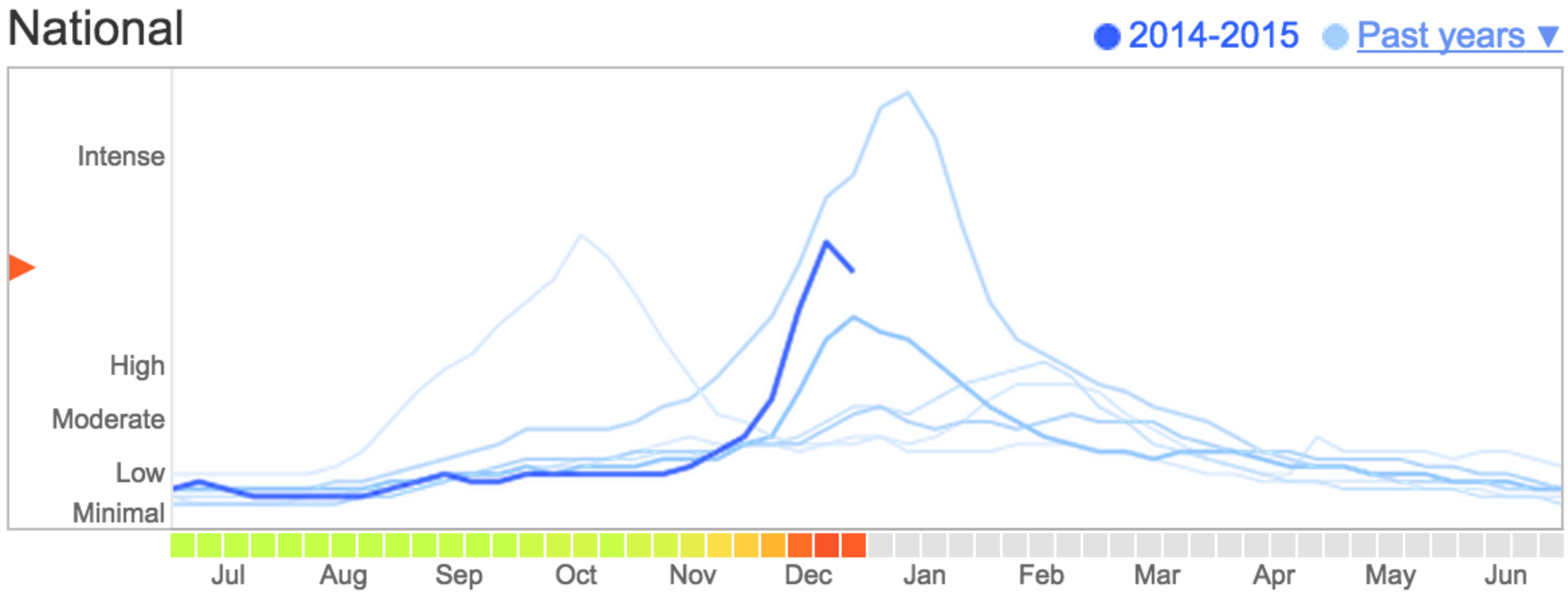
- Query stream


Medicine

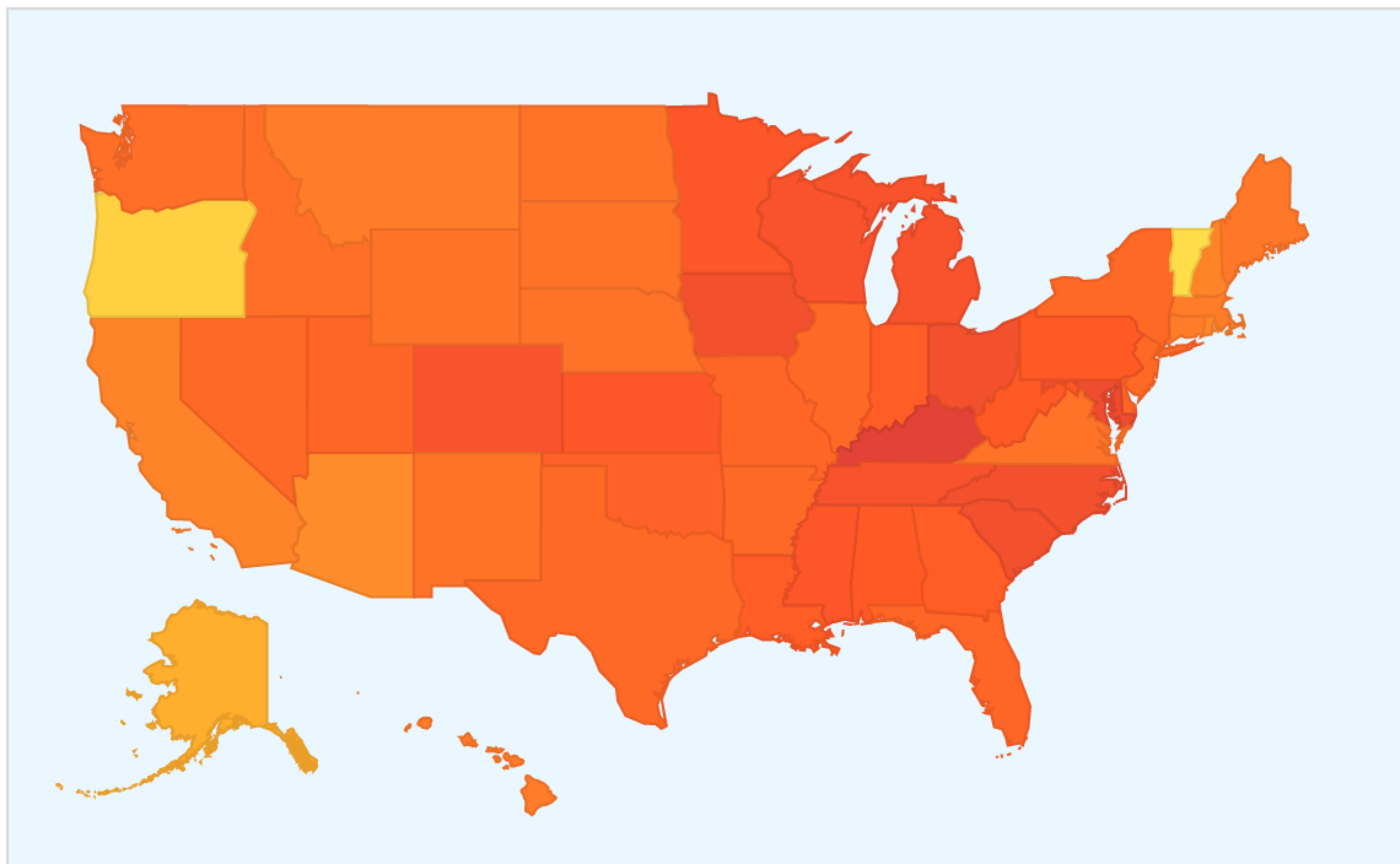
Healthcare

- 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

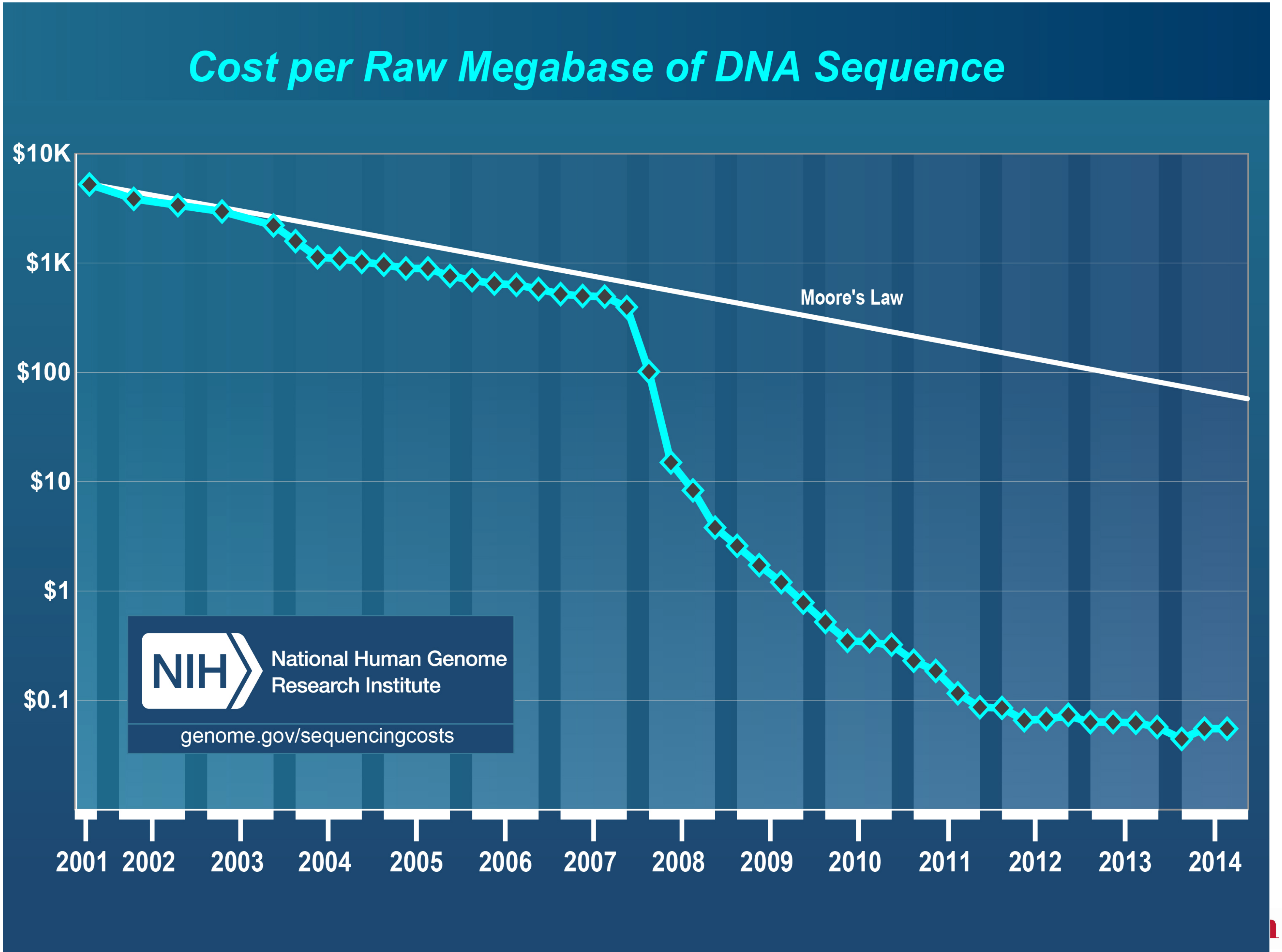
Flu trends (from search queries)



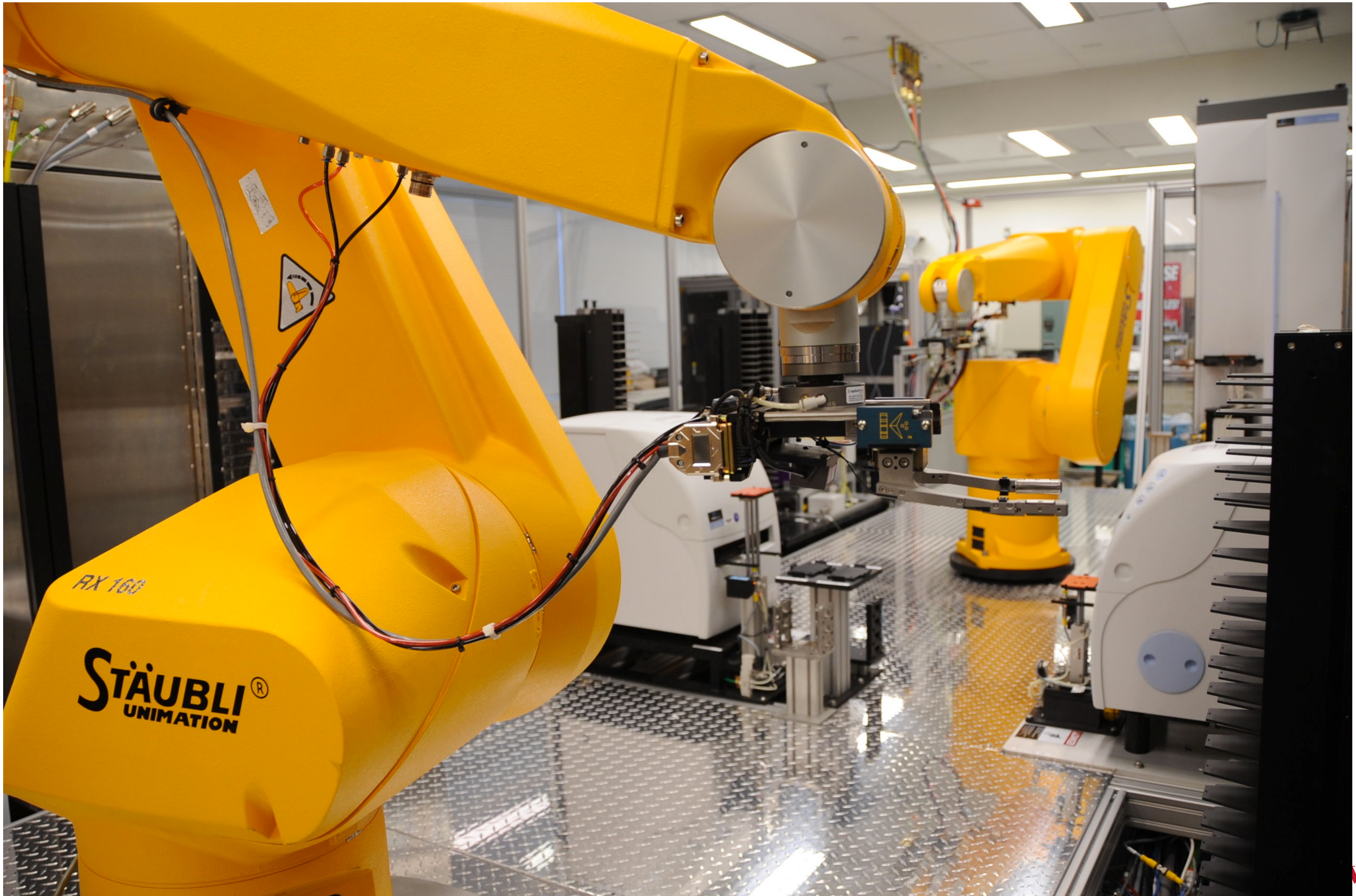
States | Cities (Experimental)



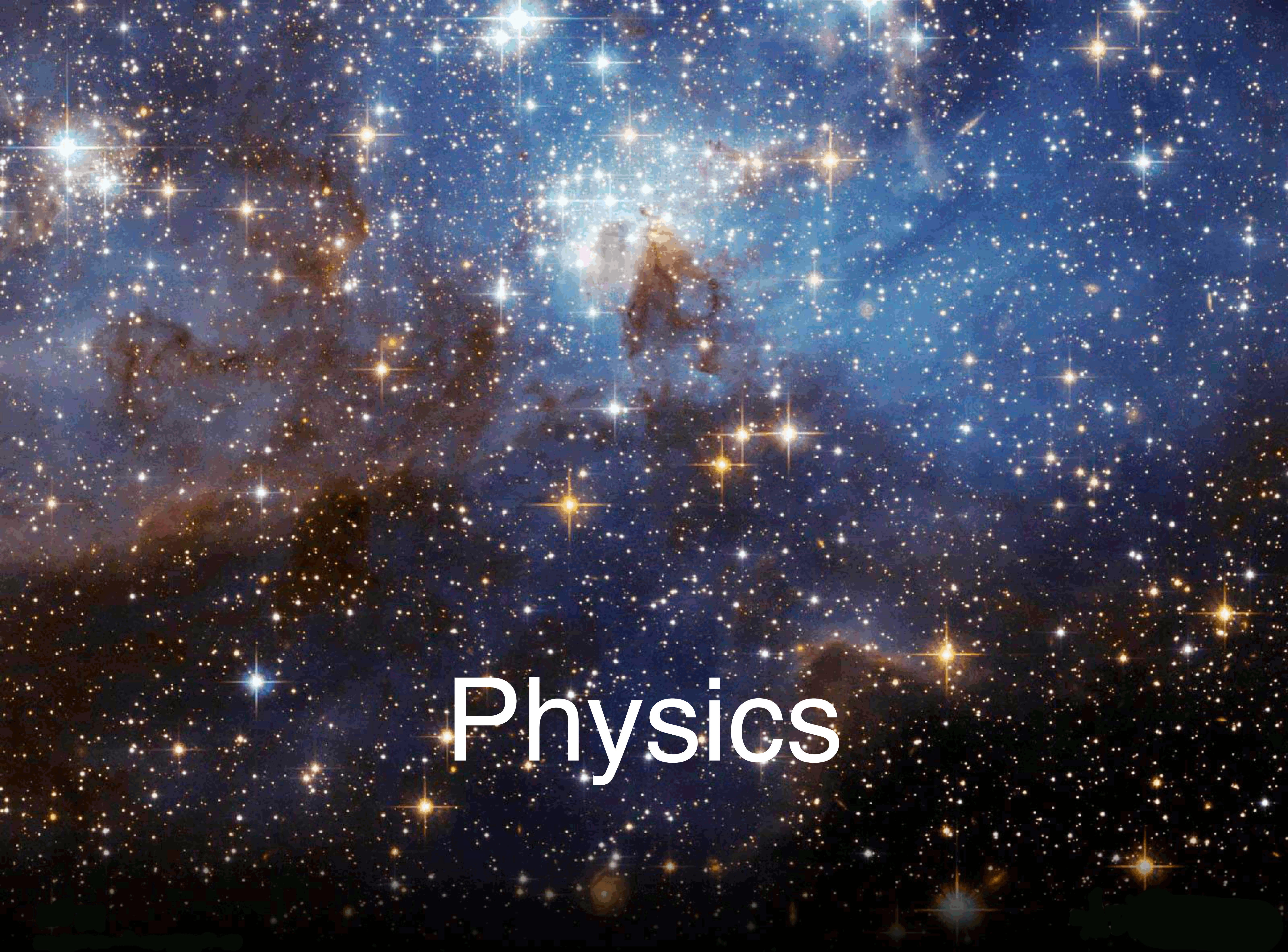
DNA Sequencing



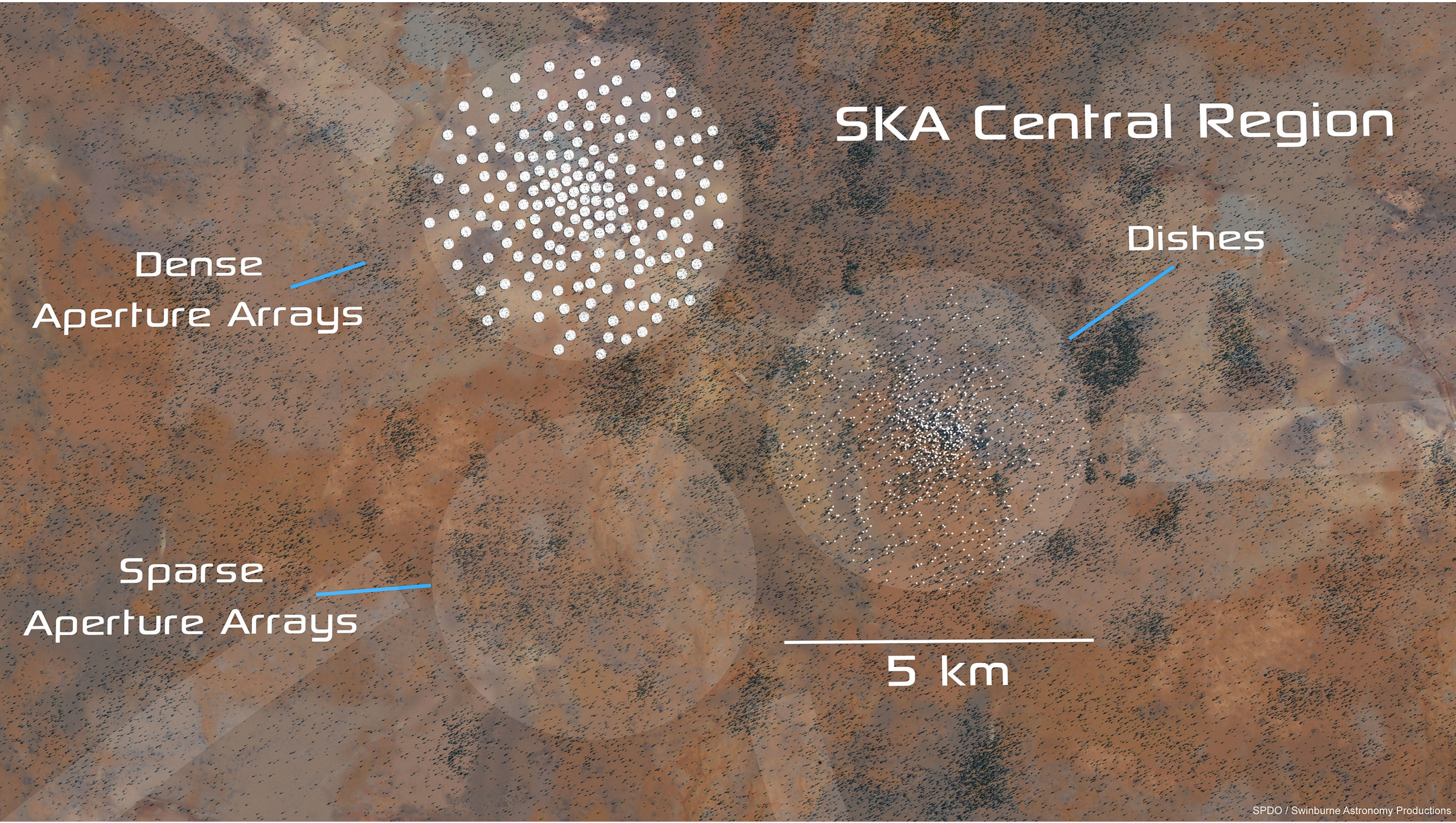
High throughput screening



versity


Physics

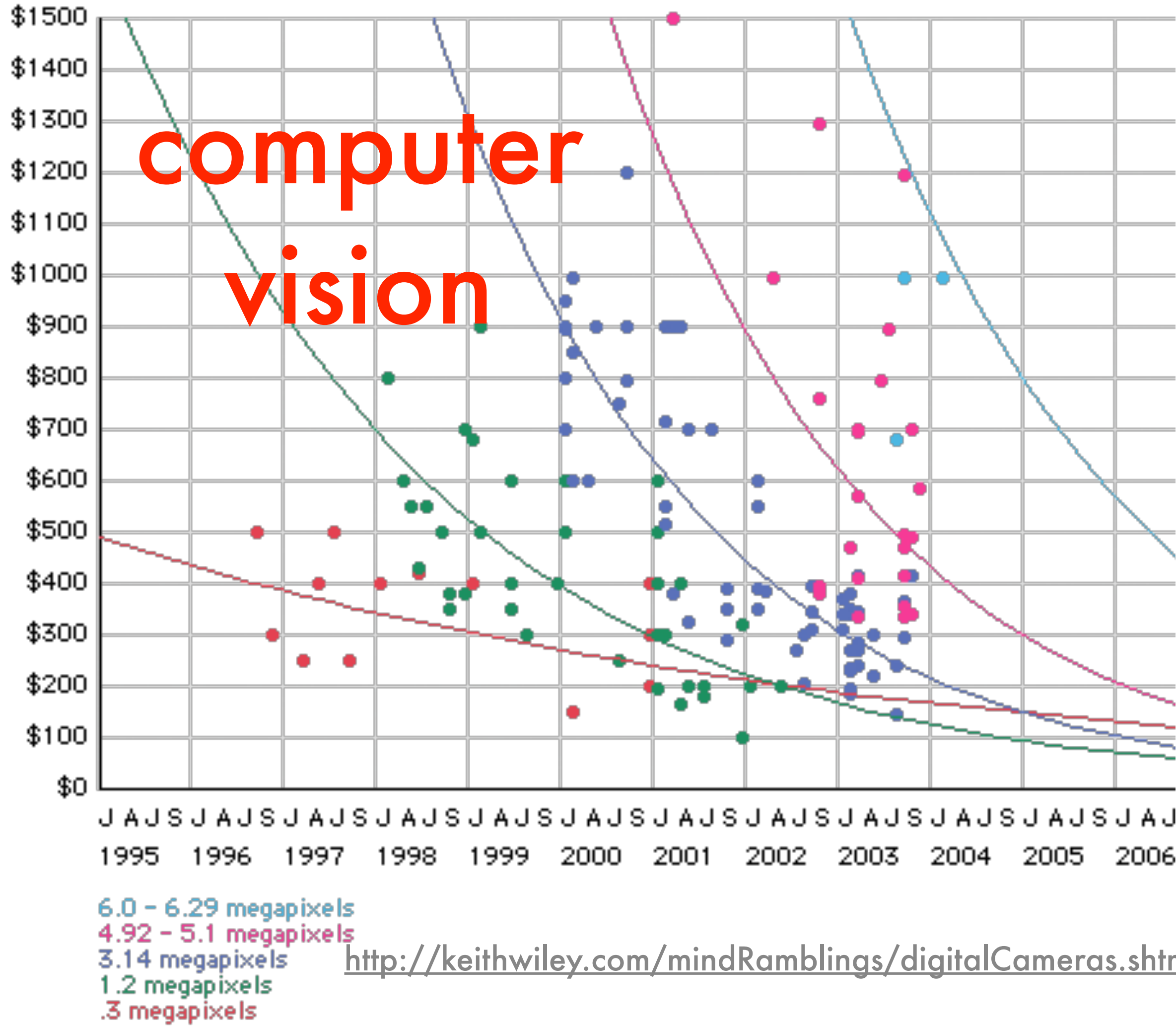
Square Kilometer Array



More sources

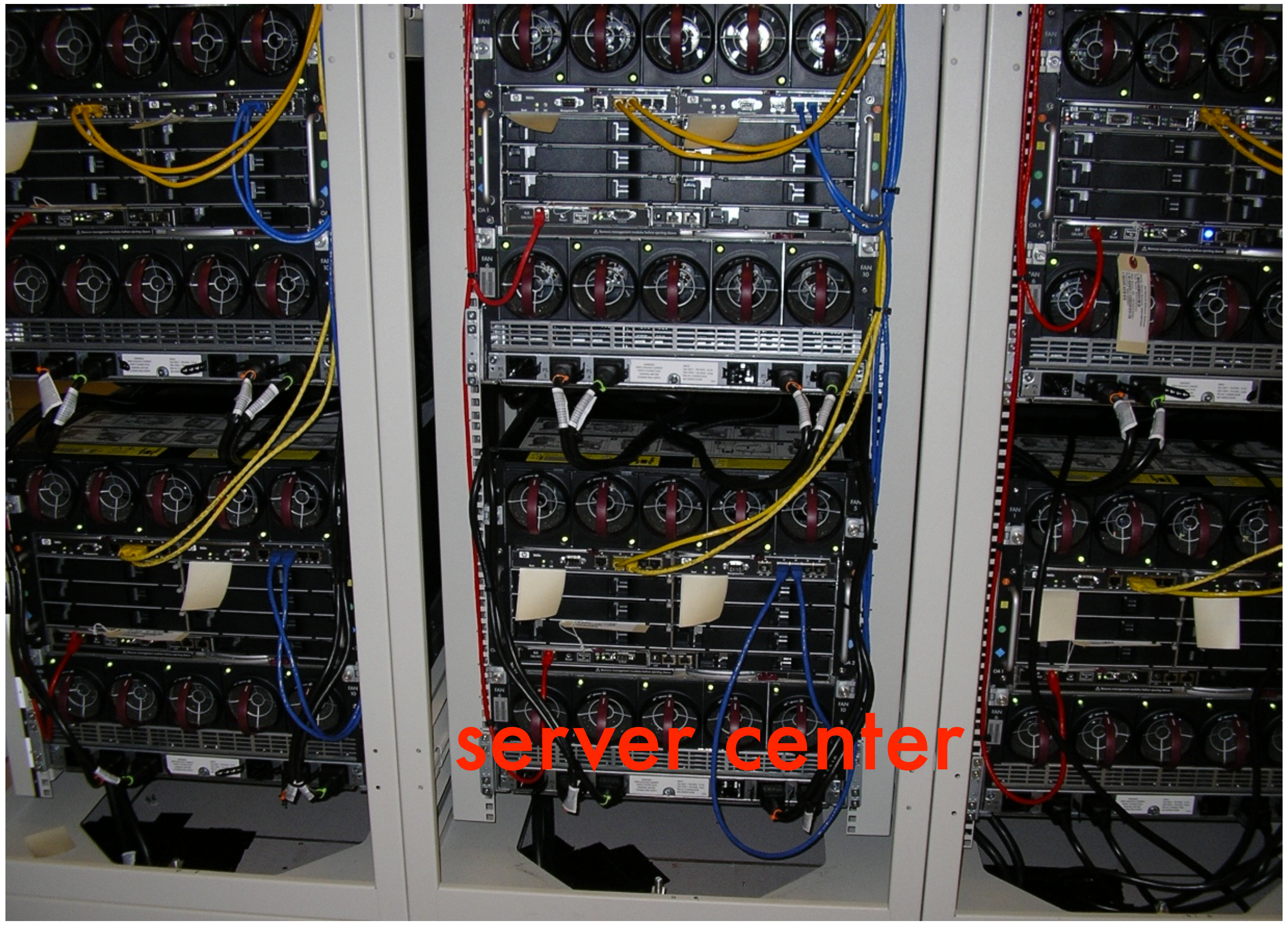
- Satellites
- High energy physics
  CERN is essentially a giant sensor array
- Geophysics
  Find locations of oil based on seismic readings
  (talk to me if you want to work on this)
- Semiconductor fabs
  100s of steps but want small failure rate
  $$1 - \delta = (1 - \epsilon)^n \text{ hence } \epsilon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

Many more sources



computer vision

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



server center

power grid

personalized sensors

ubiquitous control

1.5 Basic Tools

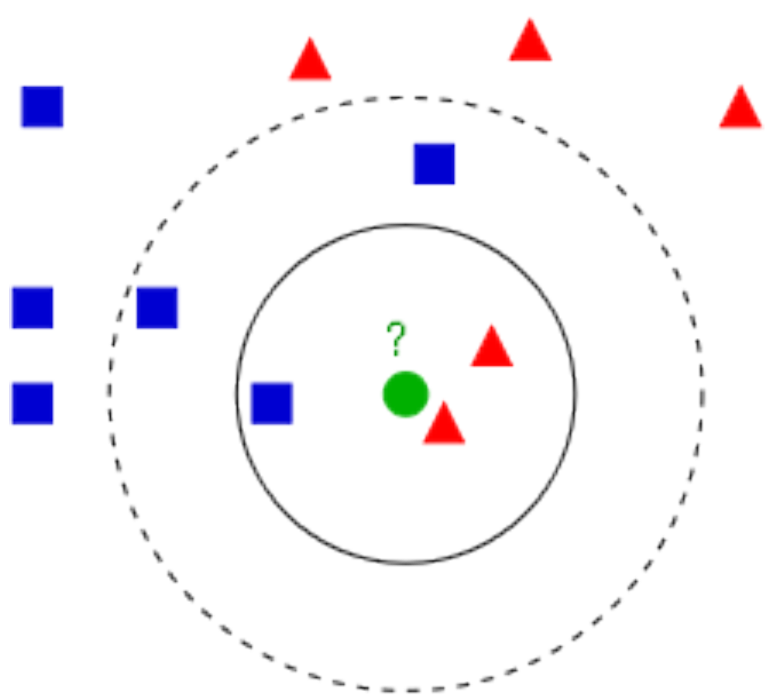
1 Introduction

Alexander Smola
Introduction to Machine Learning 10-701
http://alex.smola.org/teaching/10-701-15

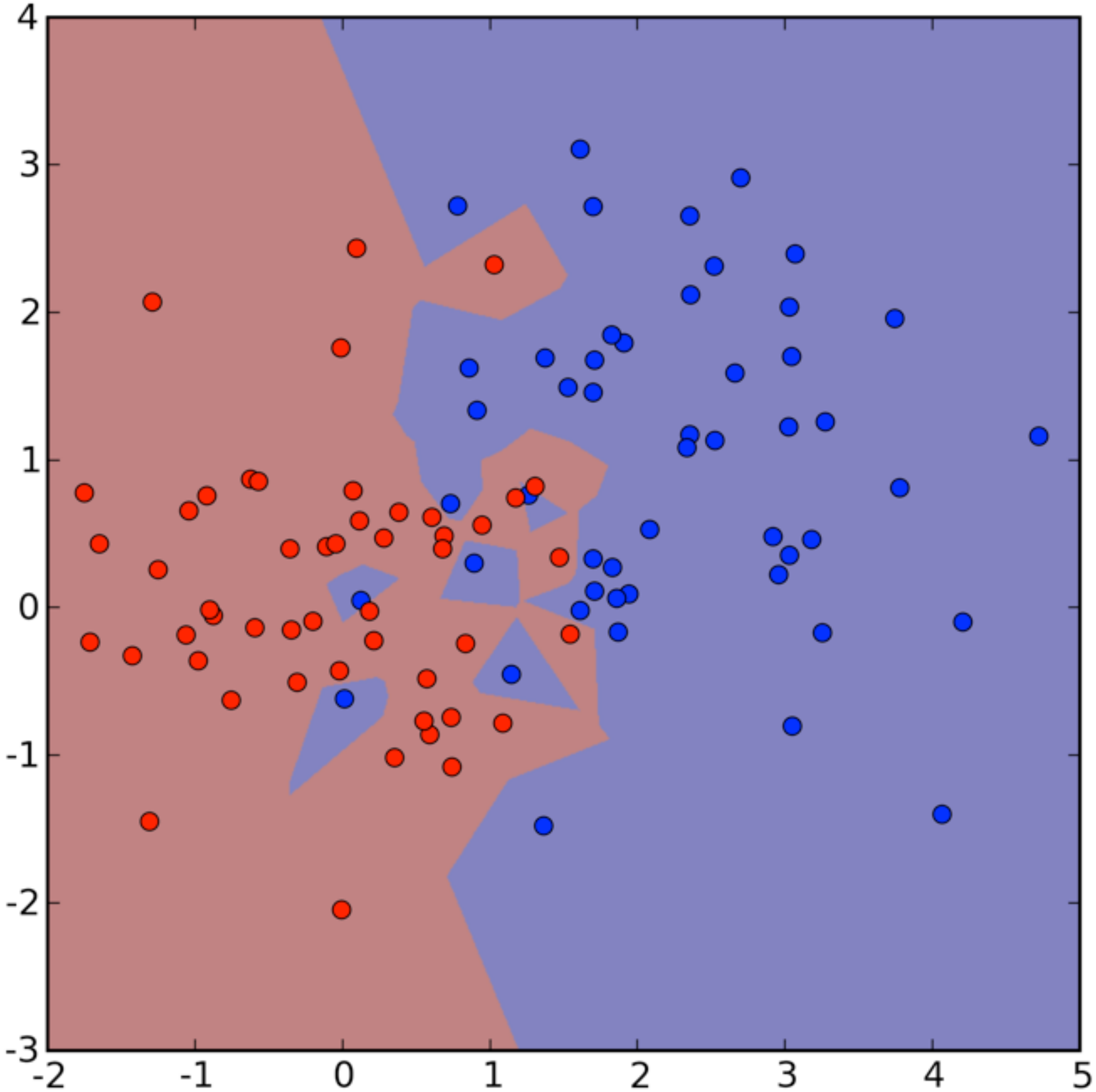

Nearest Neighbor

Nearest Neighbors

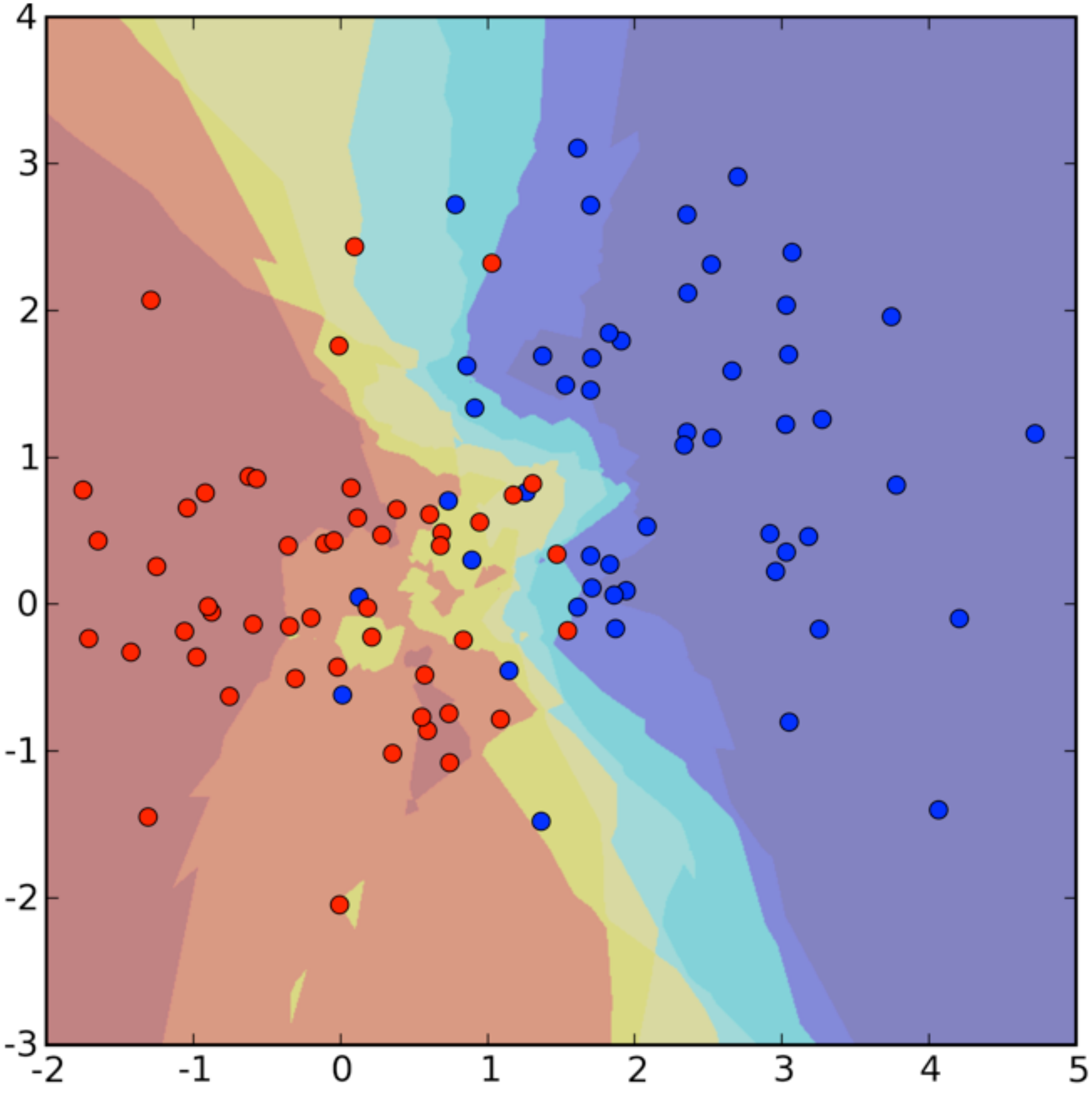
- Table lookup
  For previously seen instance remember label
- Nearest neighbor
  - Pick label of most similar neighbor
  - Slight improvement - use k-nearest neighbors
  - For regression average
  - Really useful baseline!
  - Easy to implement for small amounts of data. Why?



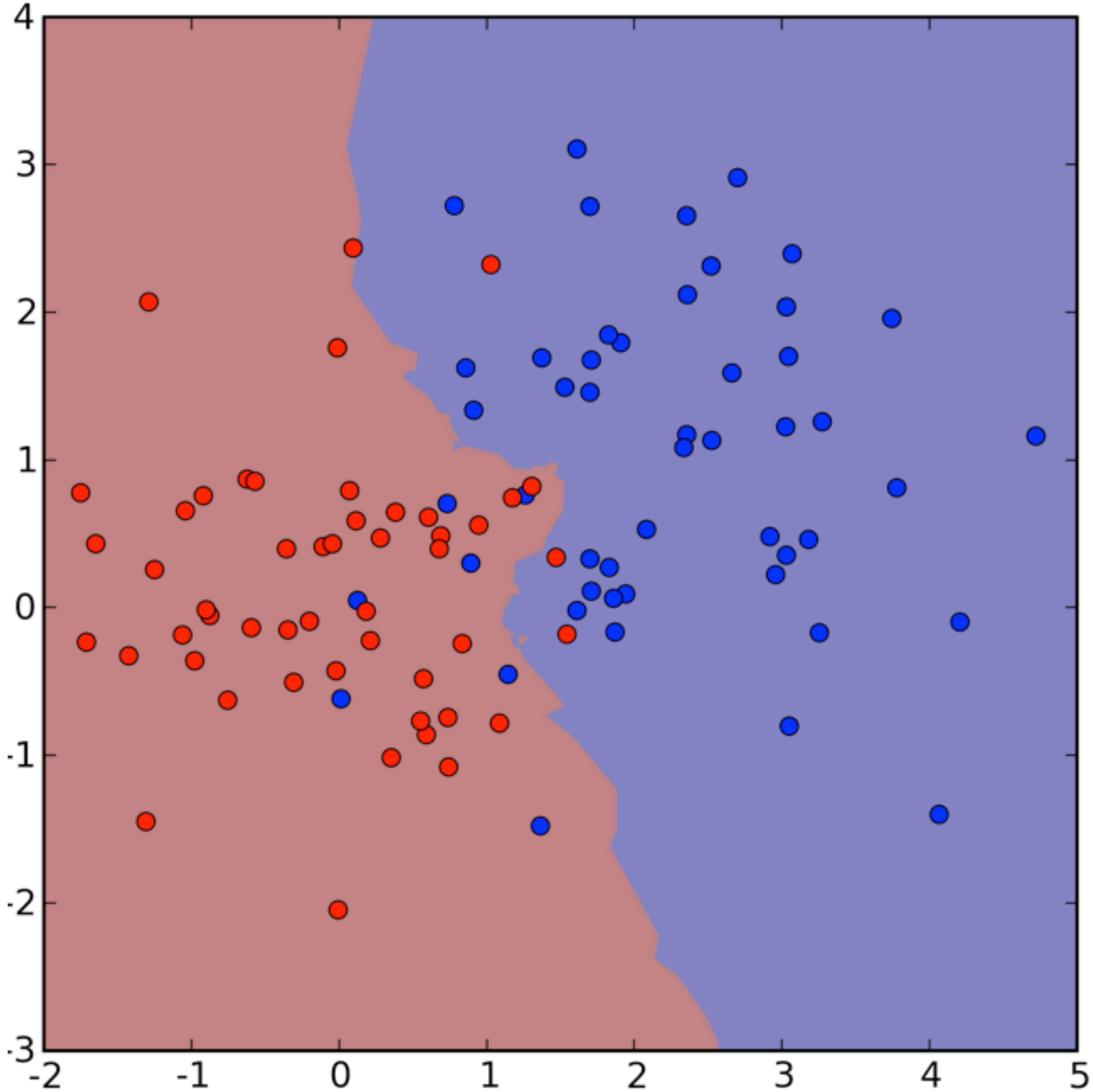
1-Nearest Neighbor



7-Nearest Neighbors

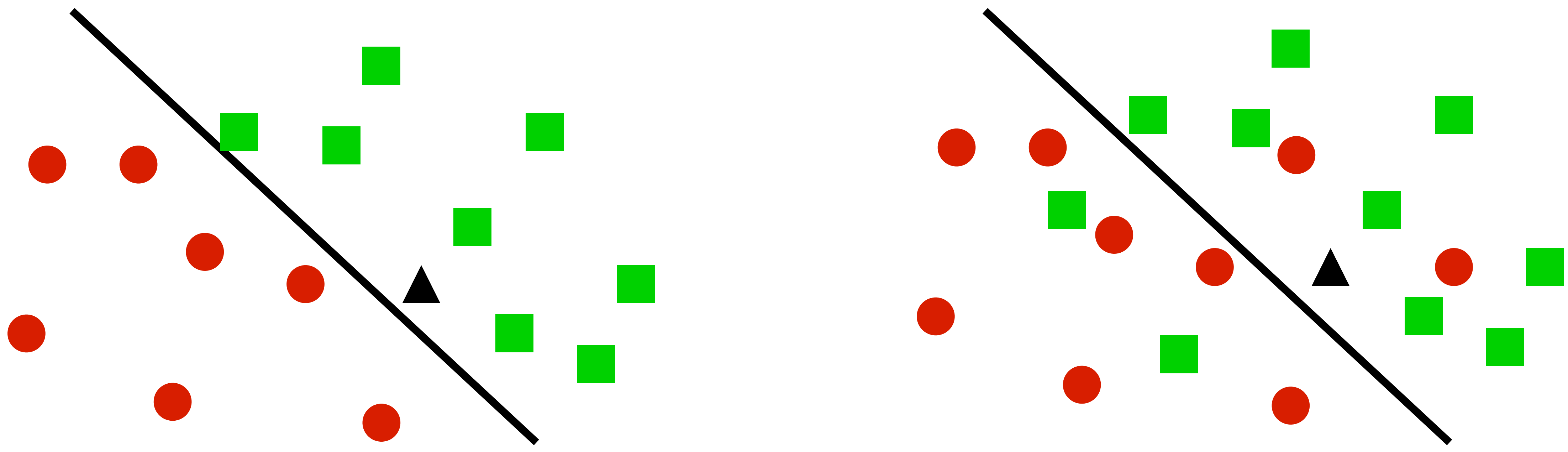


7-Nearest Neighbors Sign



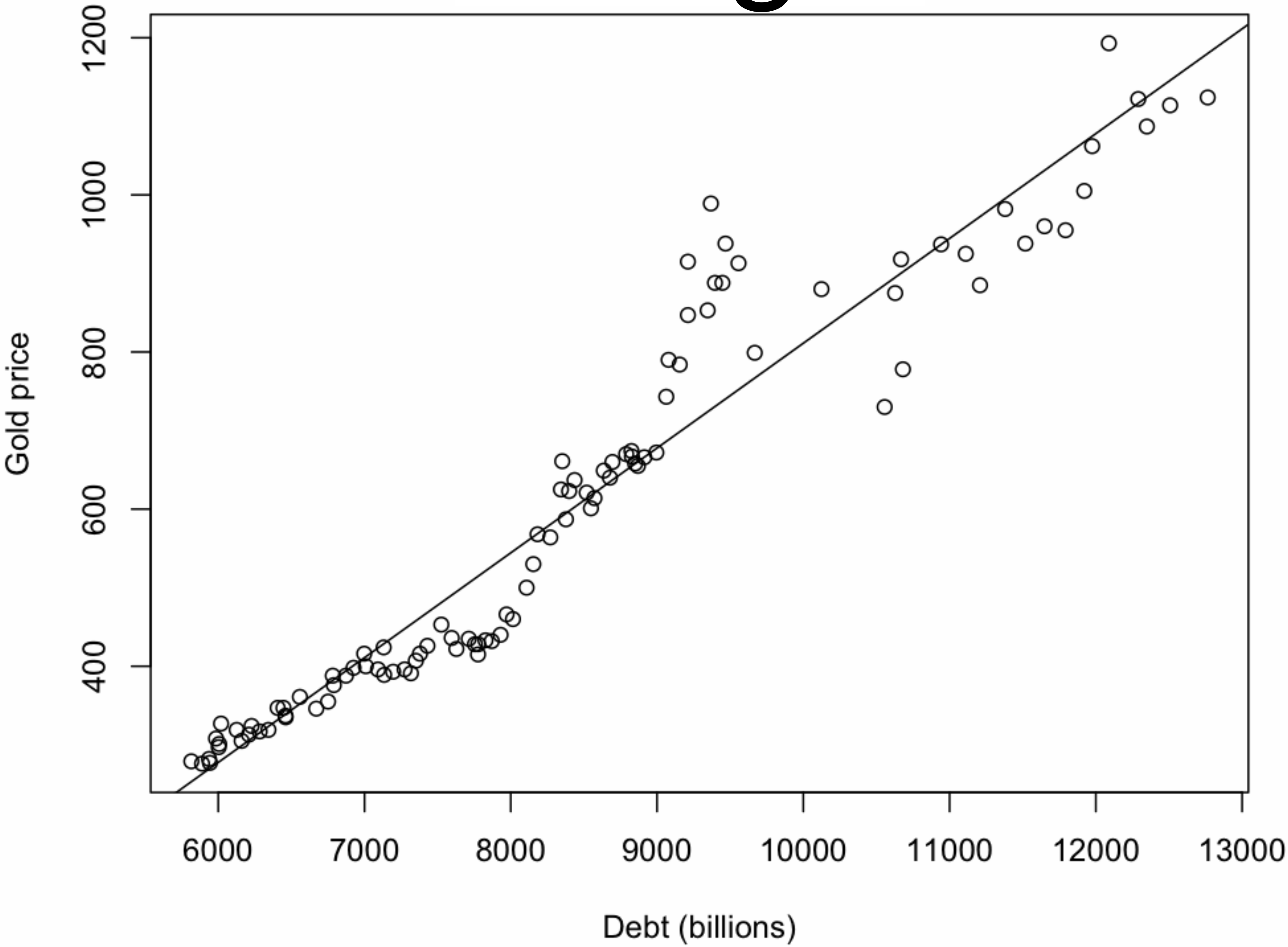
Carnegie Mellon University

If we get more data



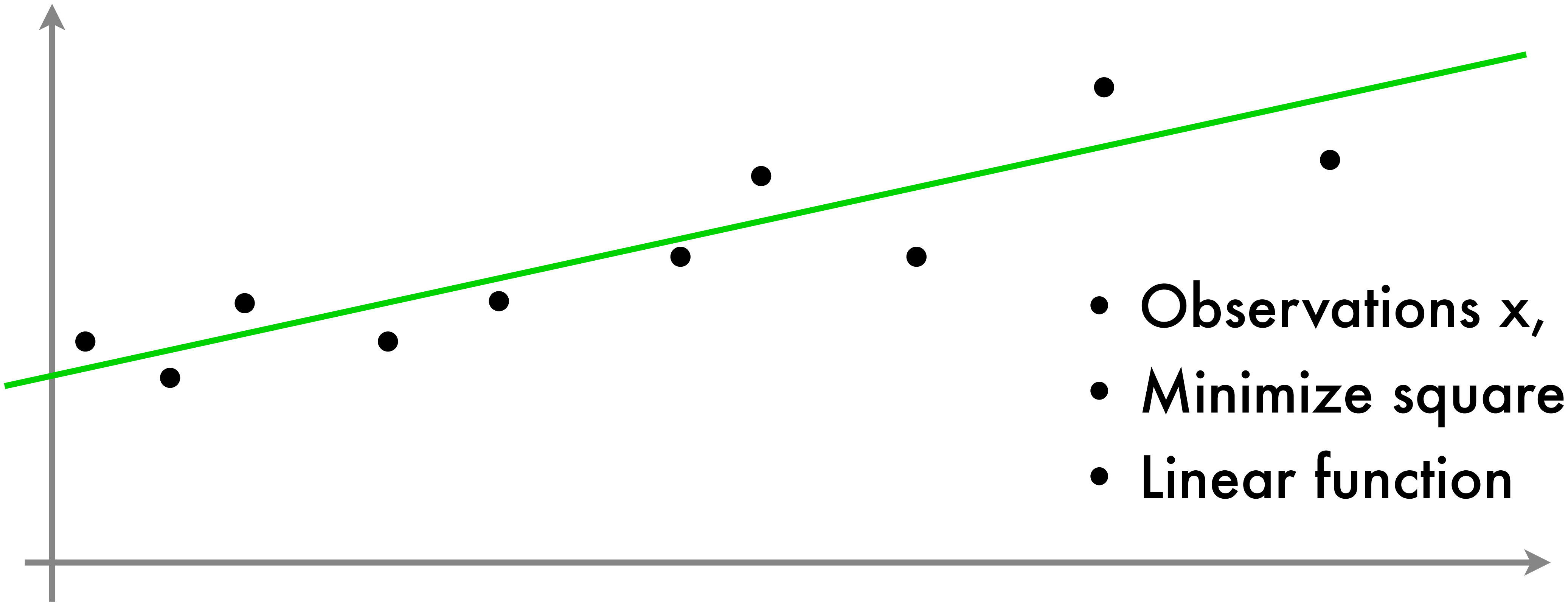
- 1 Nearest Neighbor
  - Converges to perfect solution if clear separation
  - Twice the minimal error rate 2p(1-p) for noisy problems
- k-Nearest Neighbor
  - Converges to perfect solution if clear separation (but needs more data)
  - Converges to minimal error min(p, 1-p) for noisy problems if k increases

Linear Regression



ersity

Linear Regression



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

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

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

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

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

Carnegie Mellon University

Linear Regression

- Optimization Problem

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

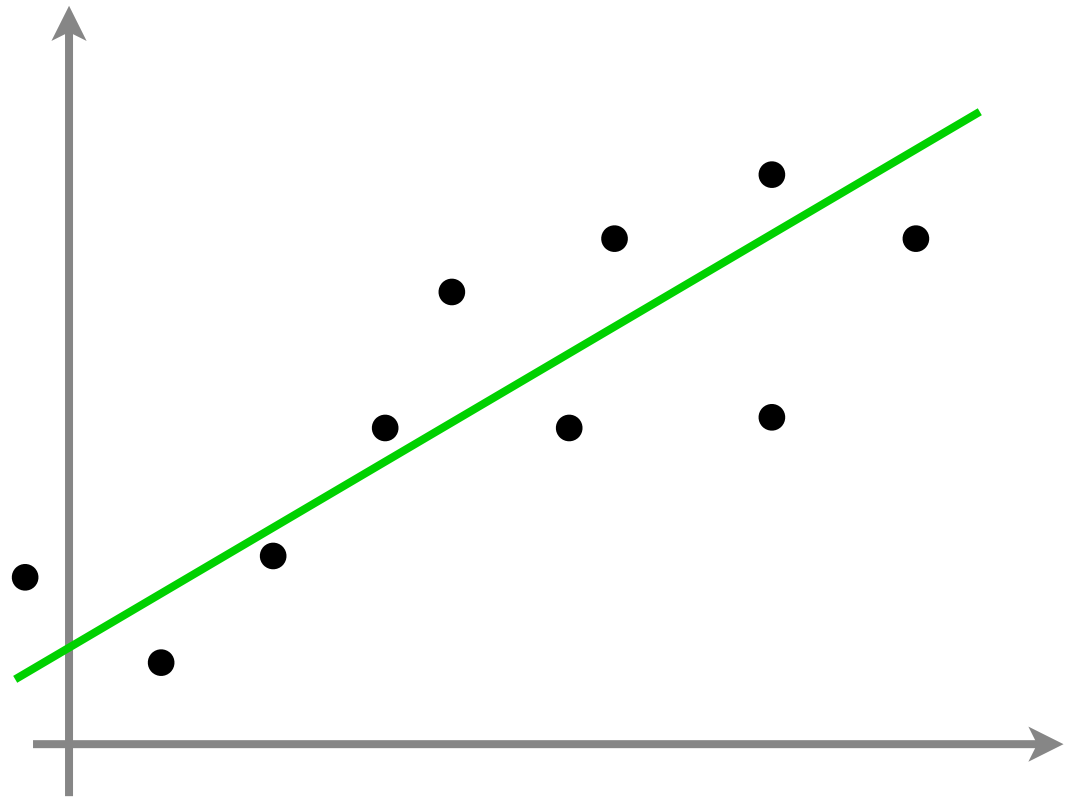
$$\underset{w}{\text{minimize}} \sum_{i=1}^{m} \frac{1}{2} (\langle w, \bar{x}_i \rangle - y_i)^2$$

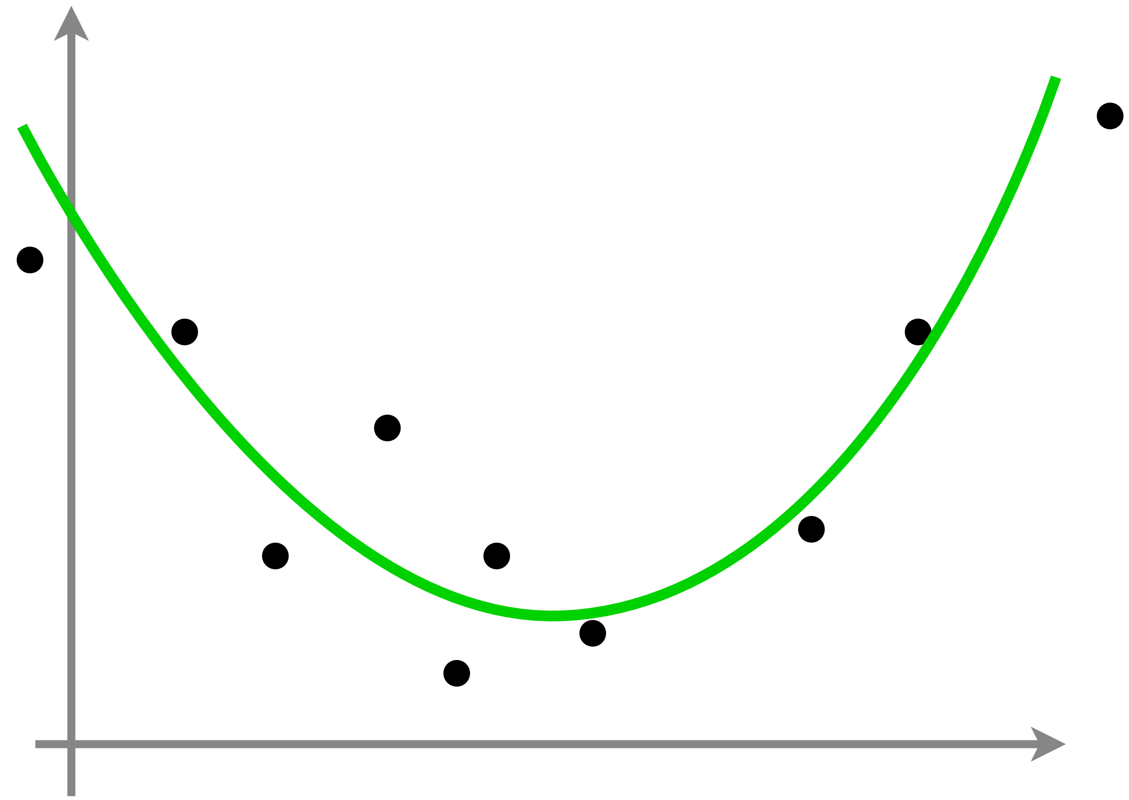
- Solving it

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

only requires a matrix inversion.

Nonlinear Regression





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

Linear Regression

- Optimization Problem

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

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

- Solving it

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

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

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

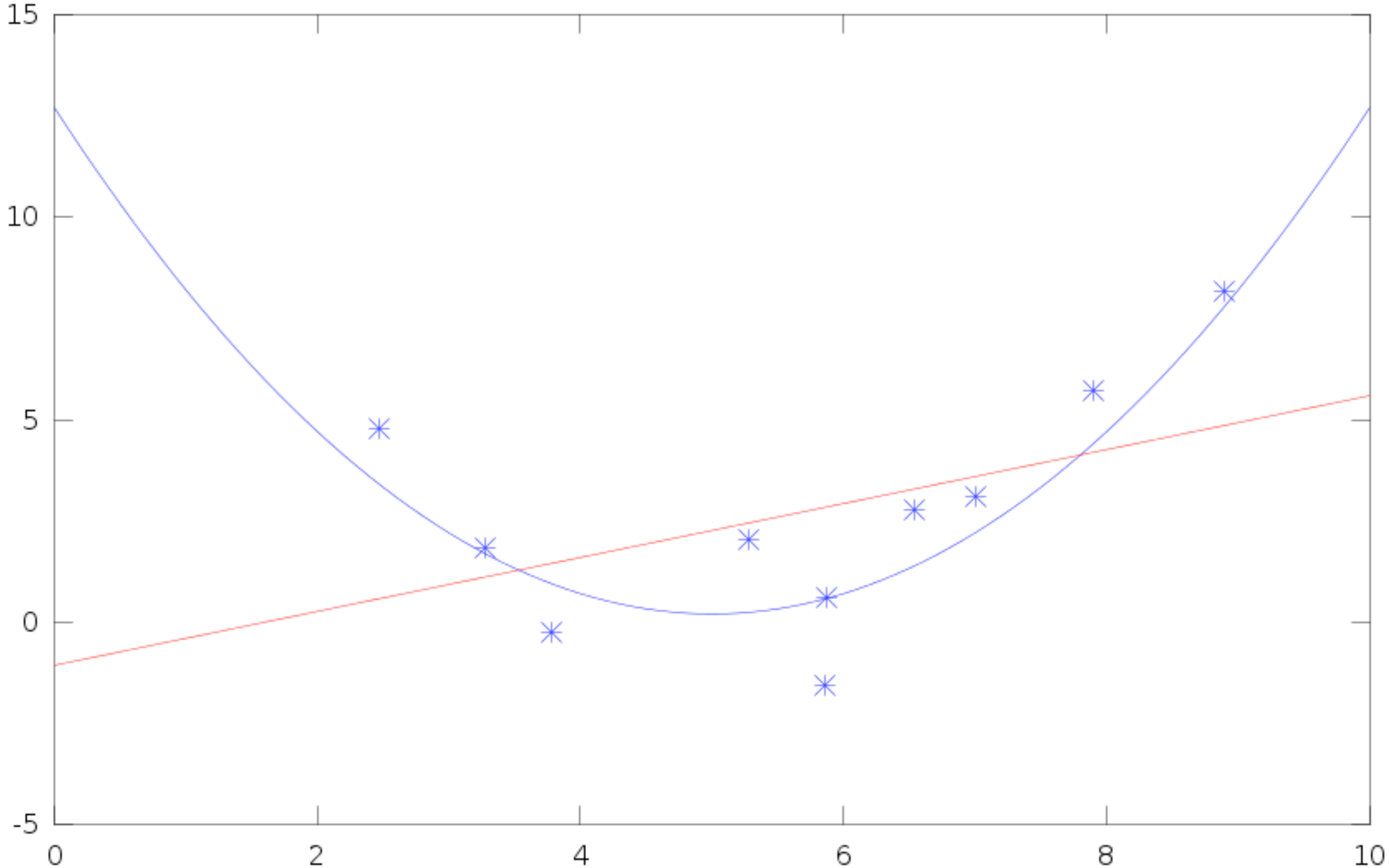
only requires a matrix inversion.

Pseudocode (degree 4)

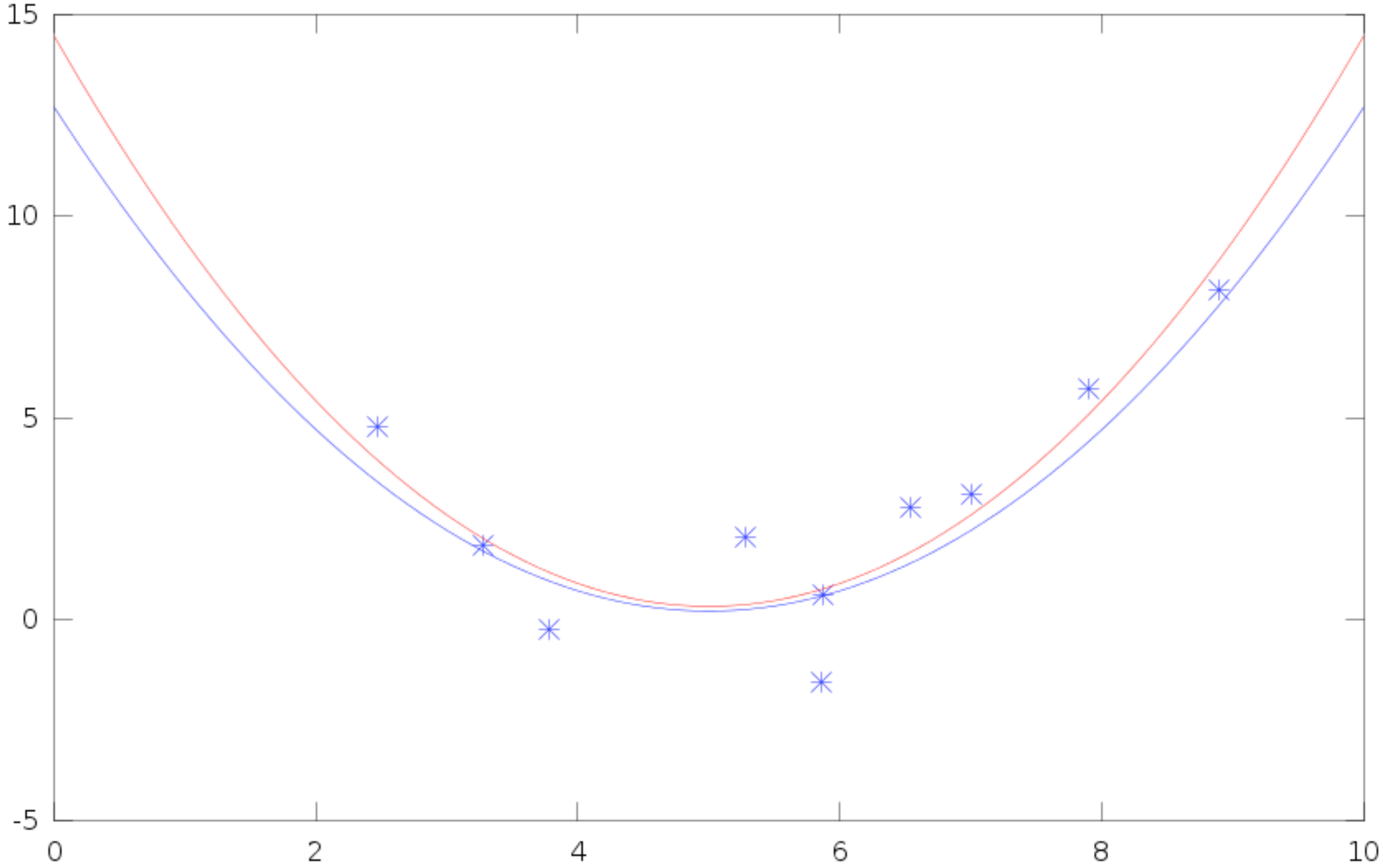
```
Training
phi_xx = [xx.^4, xx.^3, xx.^2, xx, 1.0 + 0.0 * xx];
w = (yy' * phi_xx) / (phi_xx' * phi_xx);

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

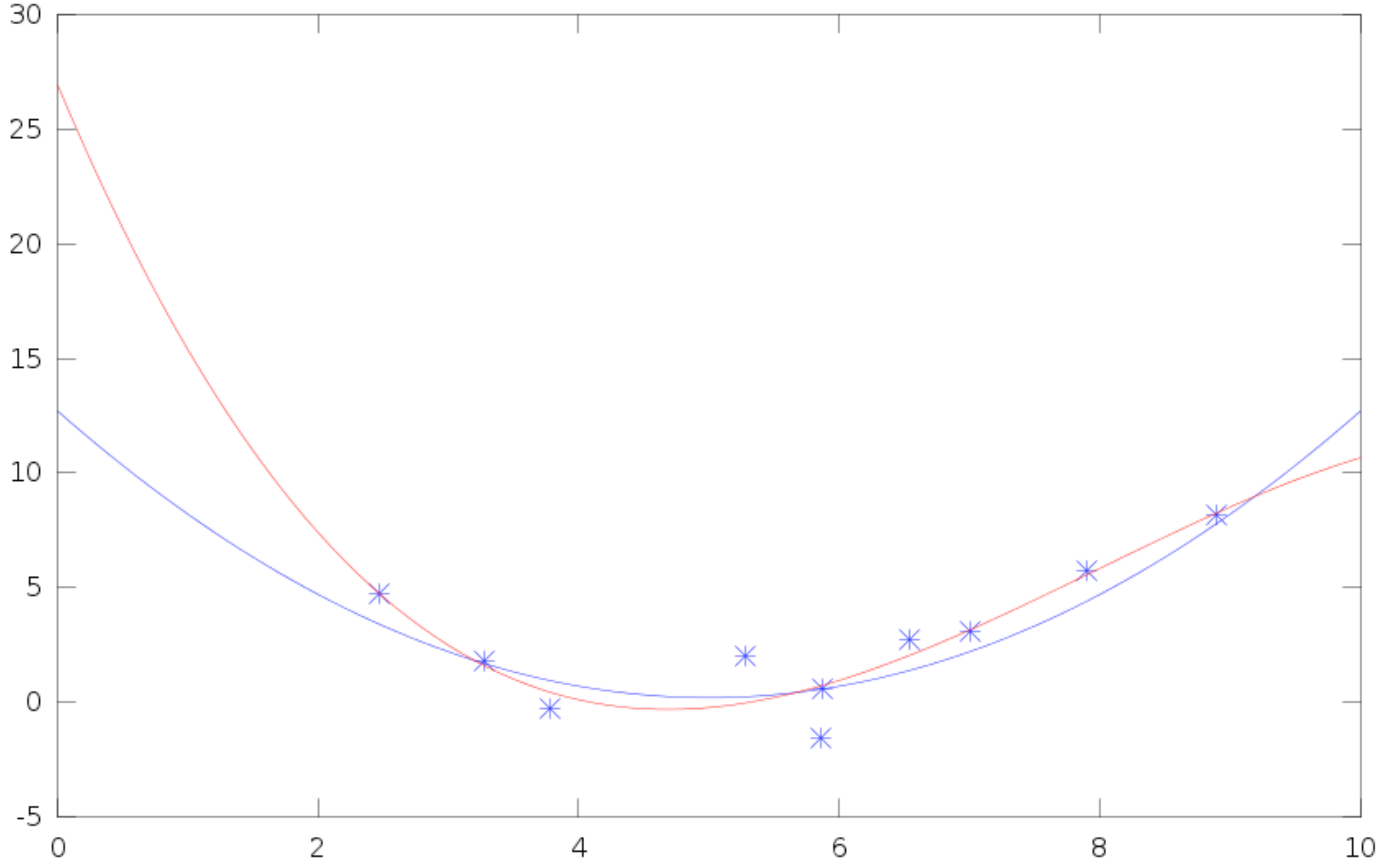
Regression (d=1)



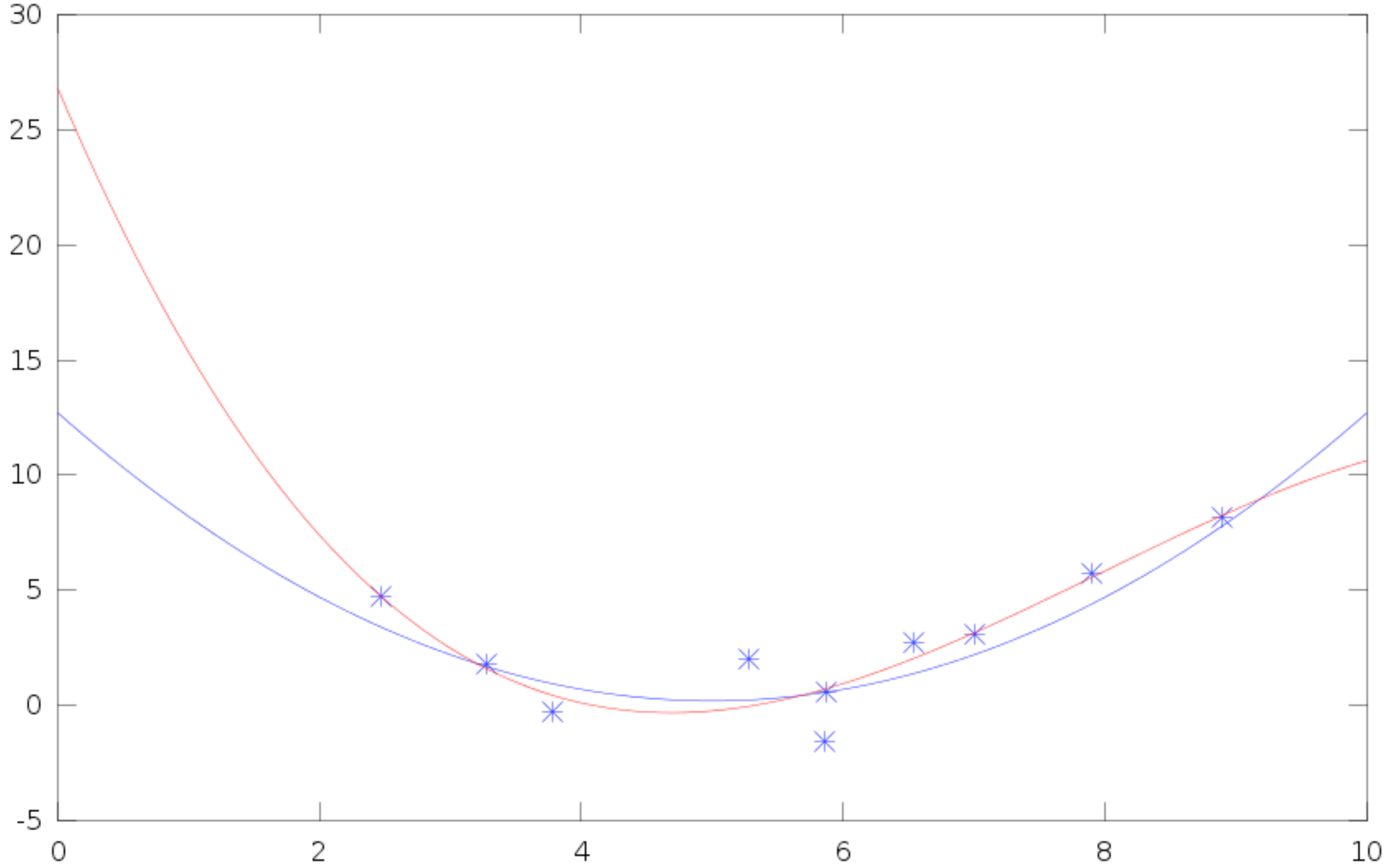
Regression (d=2)



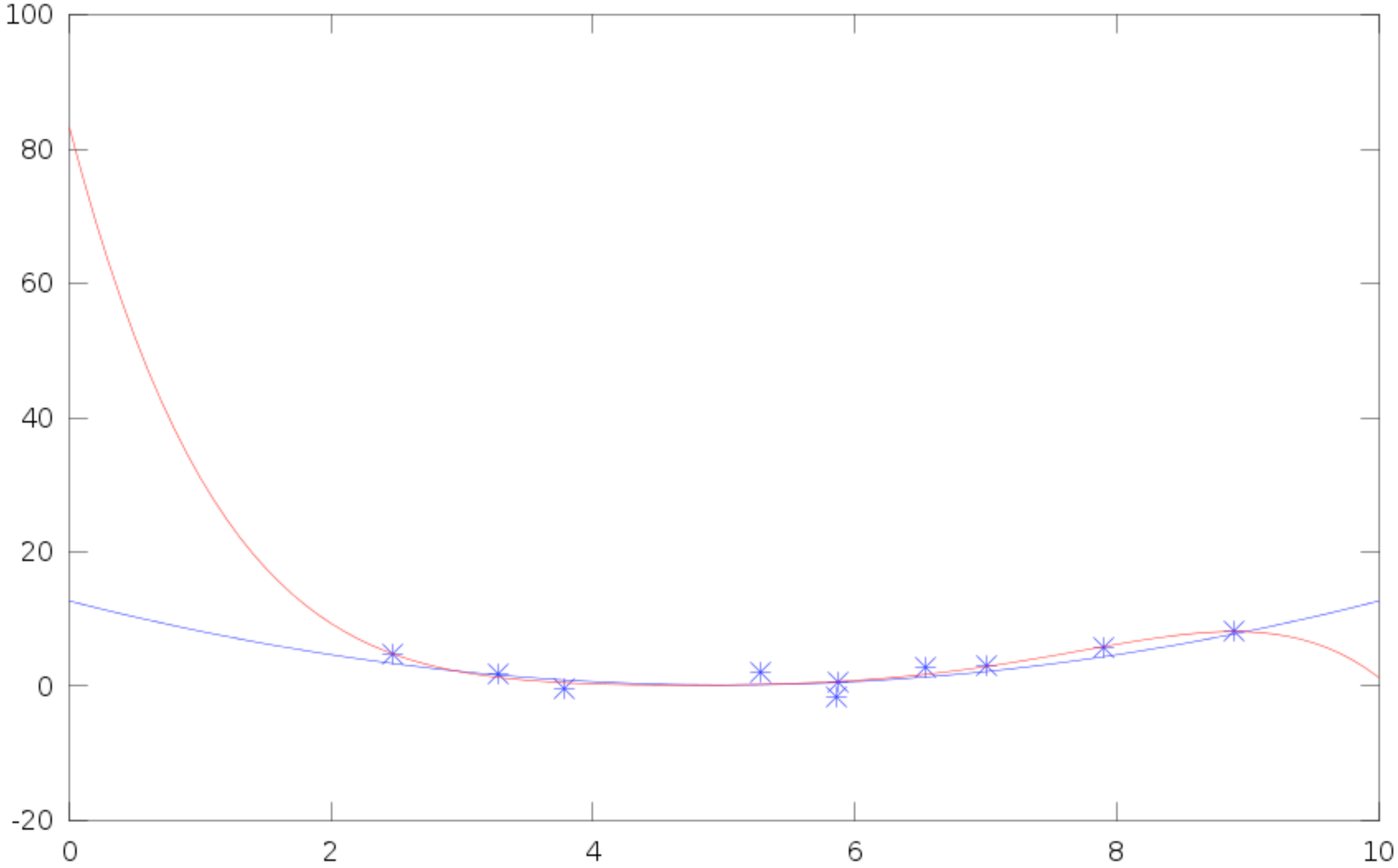
Regression (d=3)



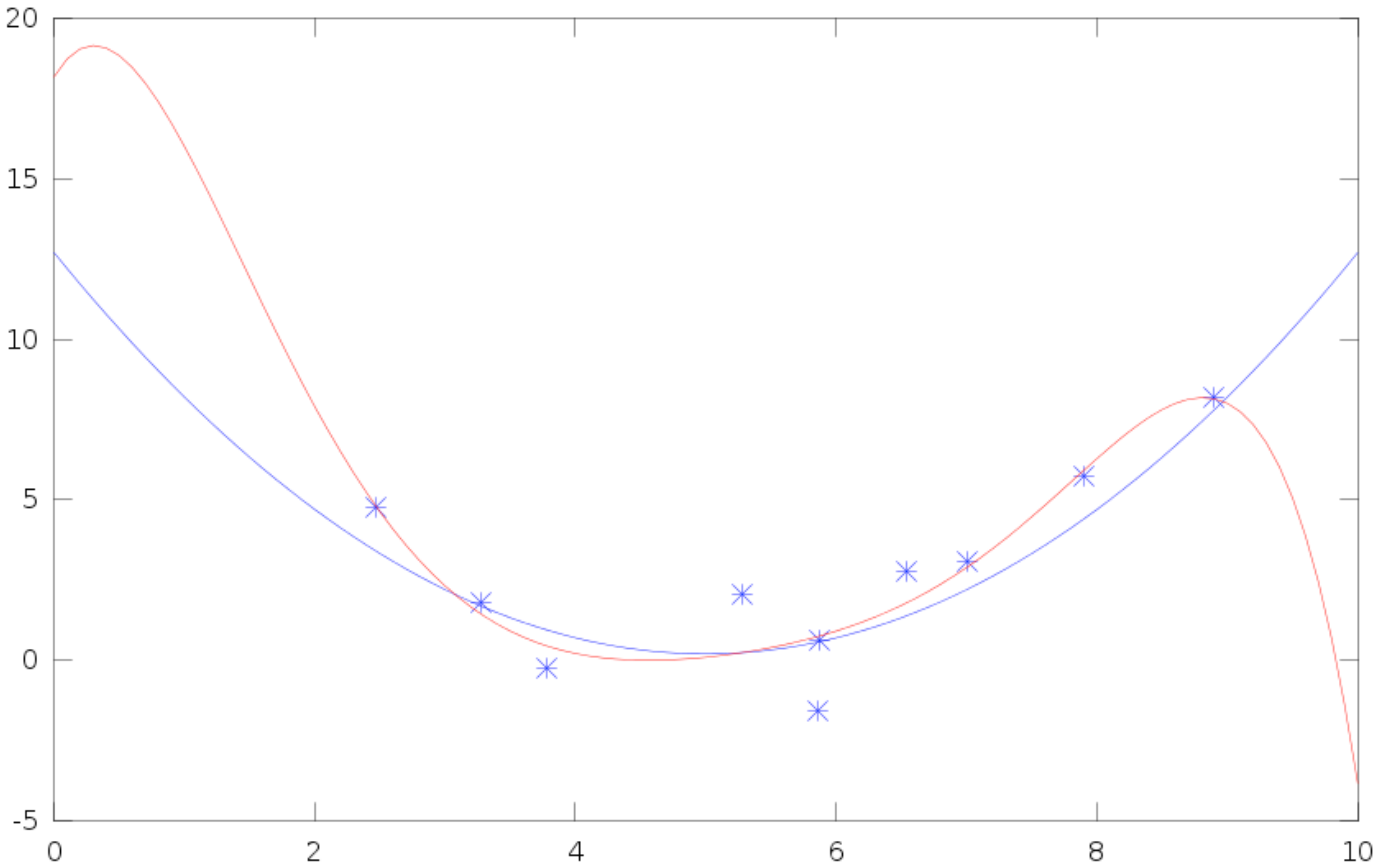
Regression (d=4)



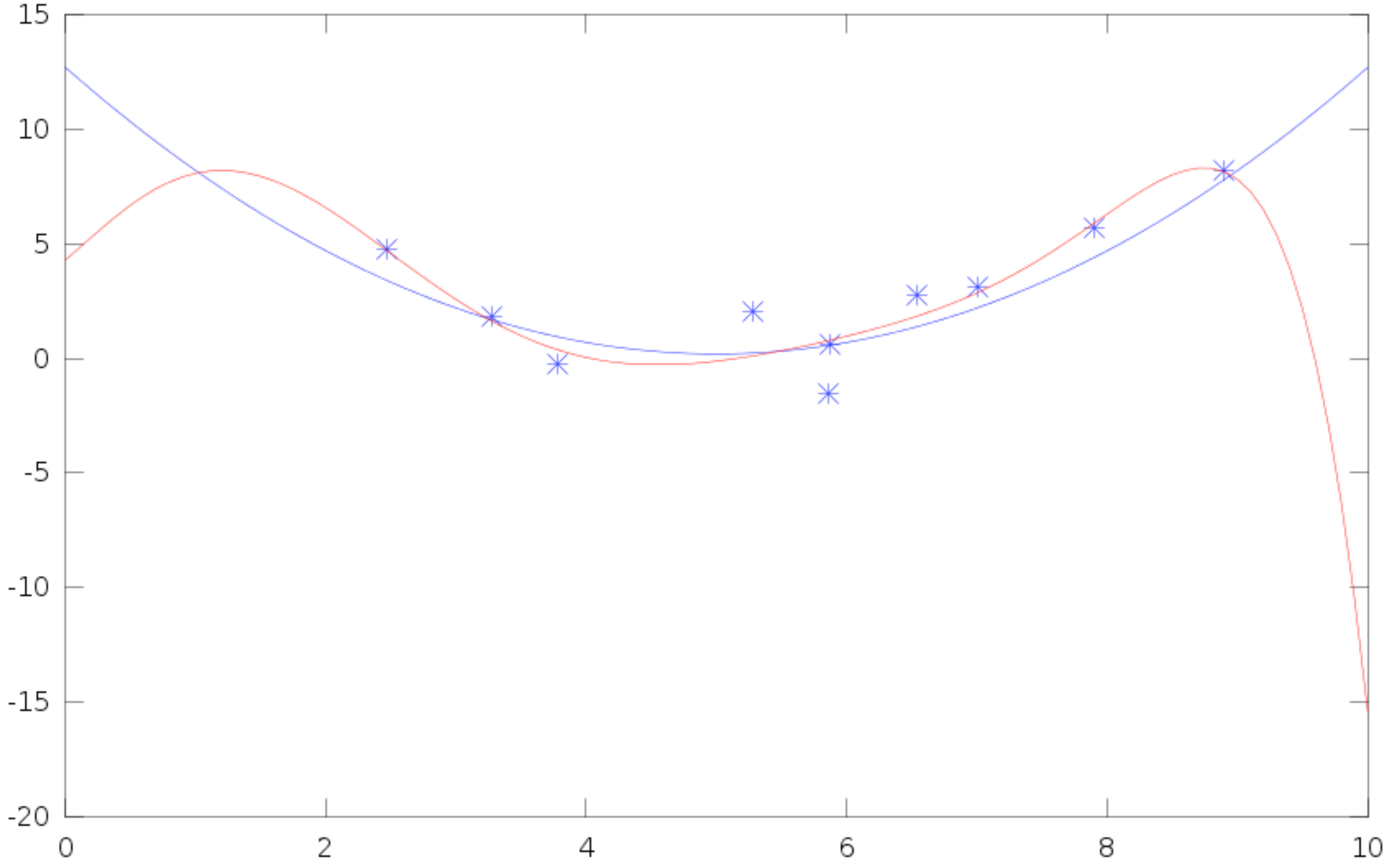
Regression (d=5)



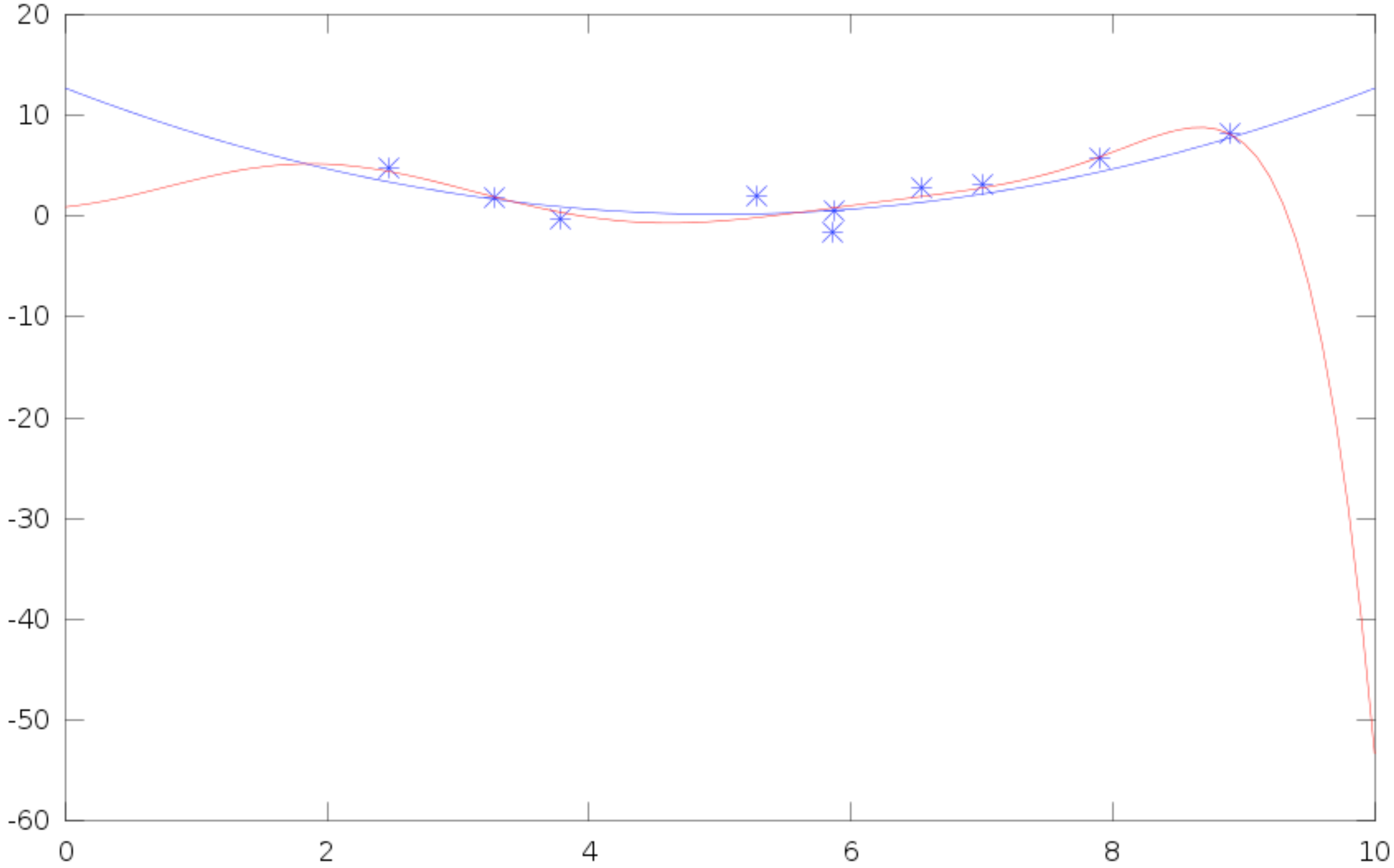
Regression (d=6)



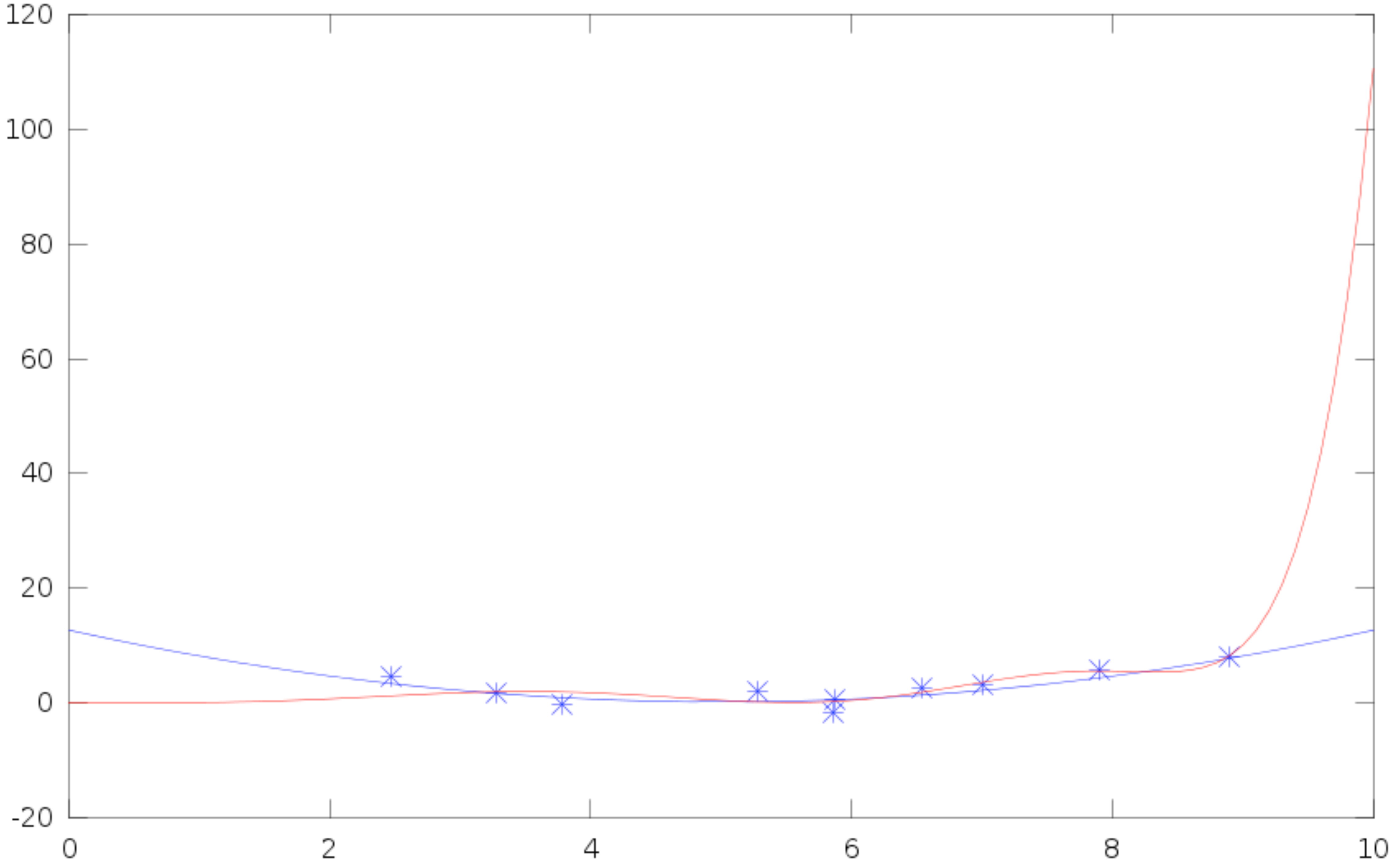
Regression (d=7)



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 = 1.10145e-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
```

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