Introduction to Machine Learning

1. Overview

Alex Smola
Carnegie Mellon University

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

Administrative Stuff

Important Stuff

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

Projects & Homework

- Don’t copy. You won’t learn anything if you do.
- Teamwork is OK (encouraged) for discussions.
- For projects 3 is a good number. 2-4 are OK.
- Each member gets the same score.
- Start your projects early.
- Ask for comments and feedback on projects
  Can we beat the Stanford class?
  http://cs229.stanford.edu/projects2012.html

Color Coding

- Really important stuff
- Important stuff
- Regular stuff

Feedback please

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

This is our FIRST class at CMU.

Outline

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

Outline

- Basics
  Problems, Statistics, Applications
- ors, Decision Trees, Neural Networks, Perceptron
- Support Vector Classification, Reg ... rnel PCA
- Theoretical Tools
  Risk Minimization, Convergence B
- aphical Models, Dynamic Programming, Latent
- Interacting with the environment
  Online L ... nt L
- Scalabili

for the internet

all you need
for a startup

for your PhD

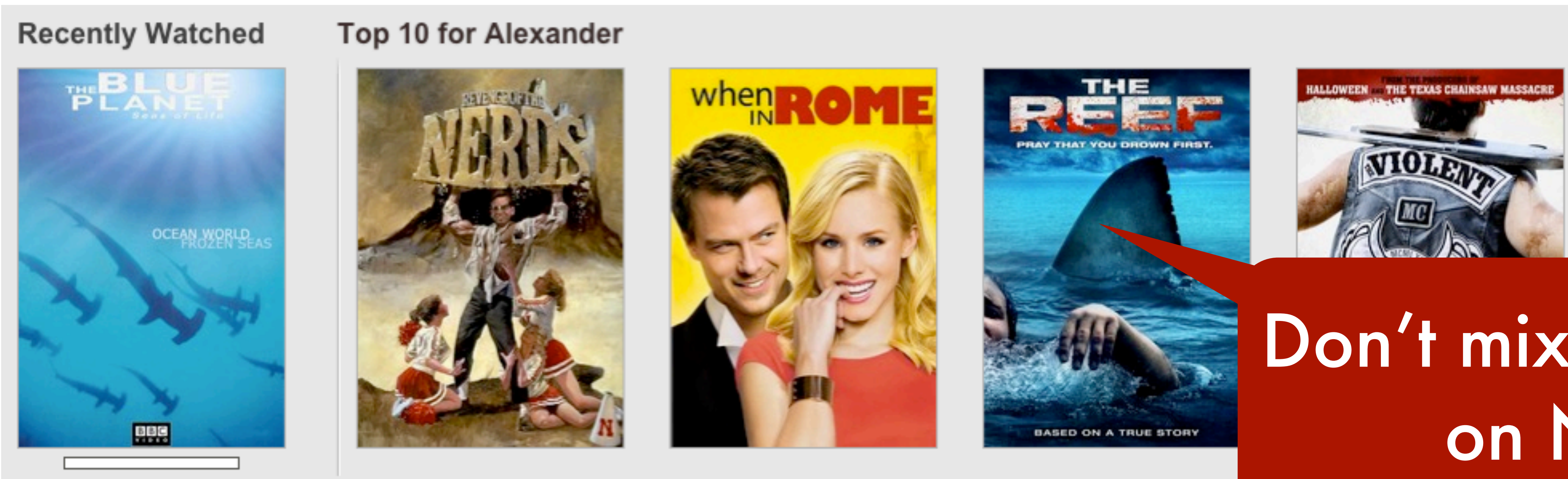
for Wall Street

biology

energy

Programming with data

Collaborative Filtering



Don't mix preferences on Netflix!

Amazon books

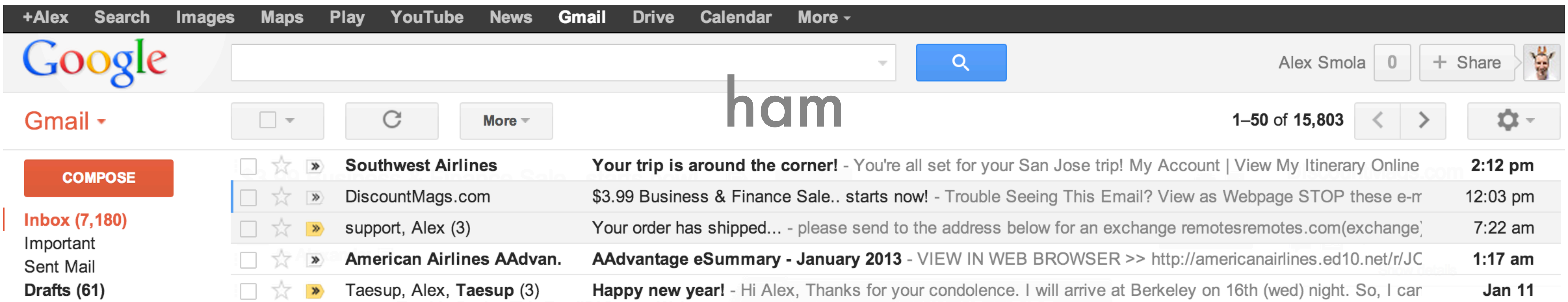
Imitation Learning in Games

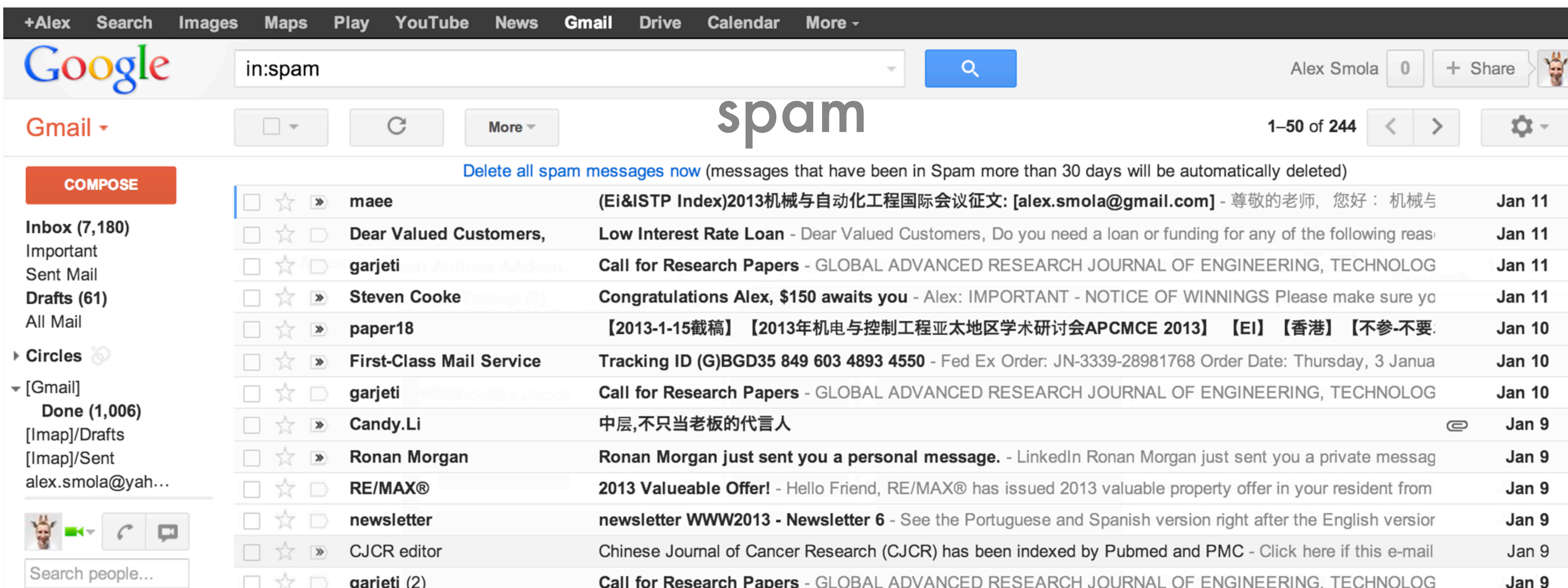
Avatar learns from your behavior

Black & White
Lionsgate Studios

Imitation Learning

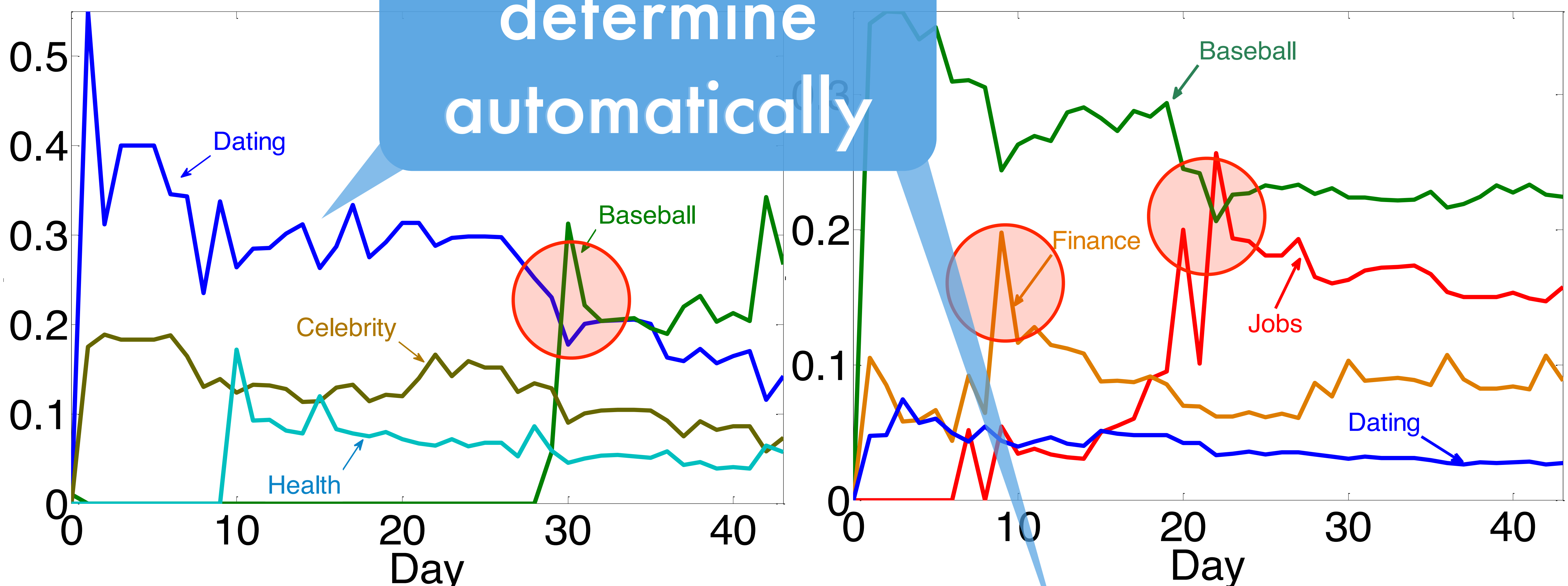
Spam Filtering





User profiling

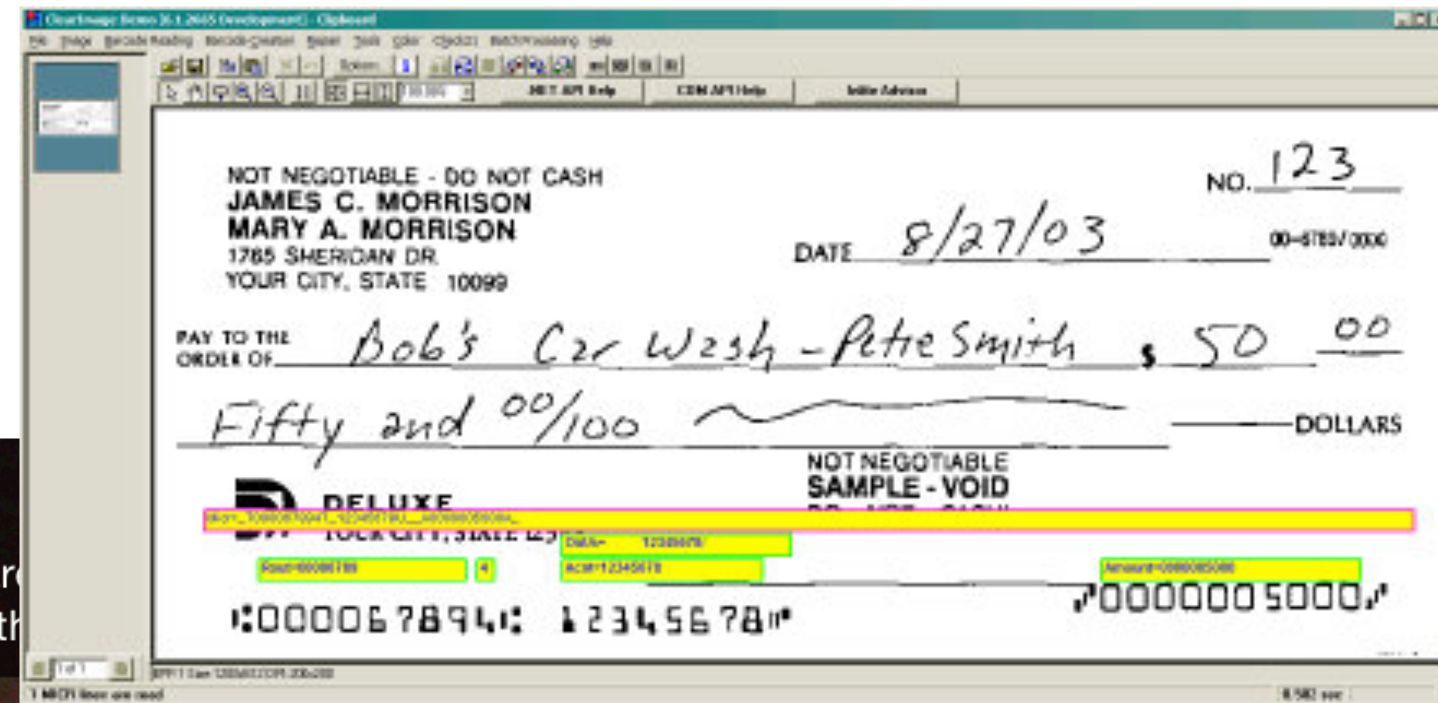
determine automatically

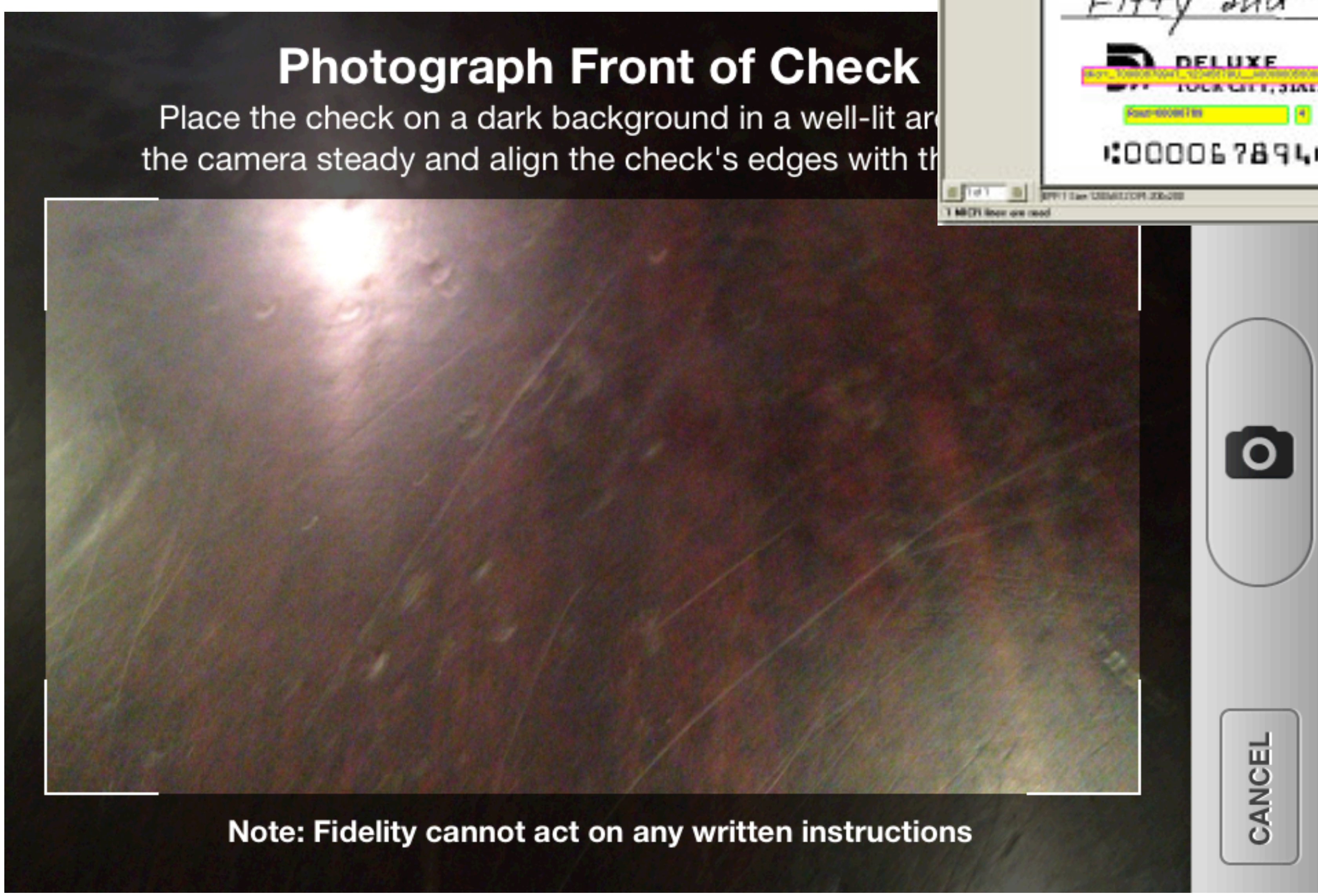


Dating	Baseball	Celebrity	Health	Jobs	Finance
women	League	Snooki	skin	job	financial
men	baseball	Tom	body	career	Thomson
dating	basketball,	Cruise	fingers	business	chart
singles	doublehead	Katie	cells	assistant	real
personals	Bergesen	Holmes	toes	hiring	Stock
seeking	Griffey	Pinkett	wrinkle	part-time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

Cheque reading

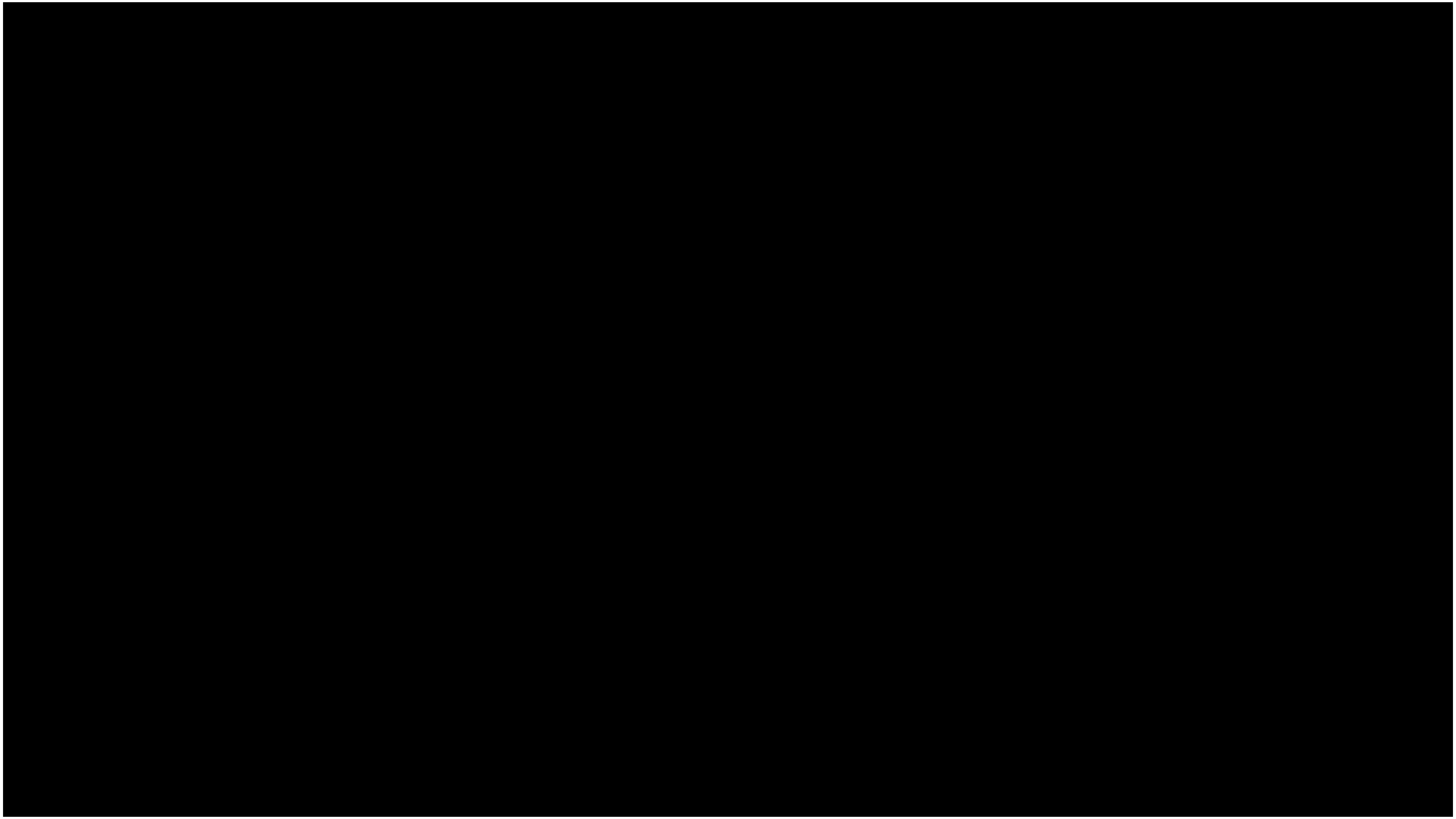
segment image





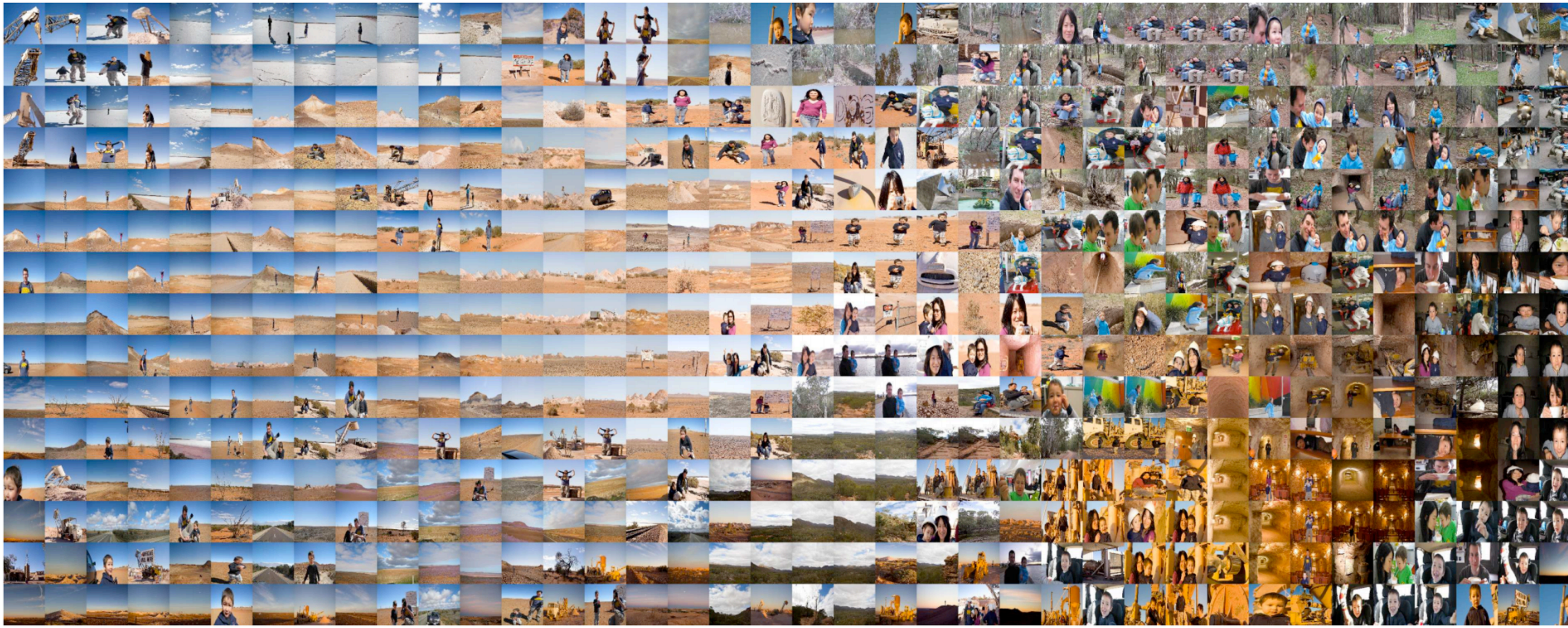
recognize handwriting

Autonomous Helicopter



http://heli.stanford.edu

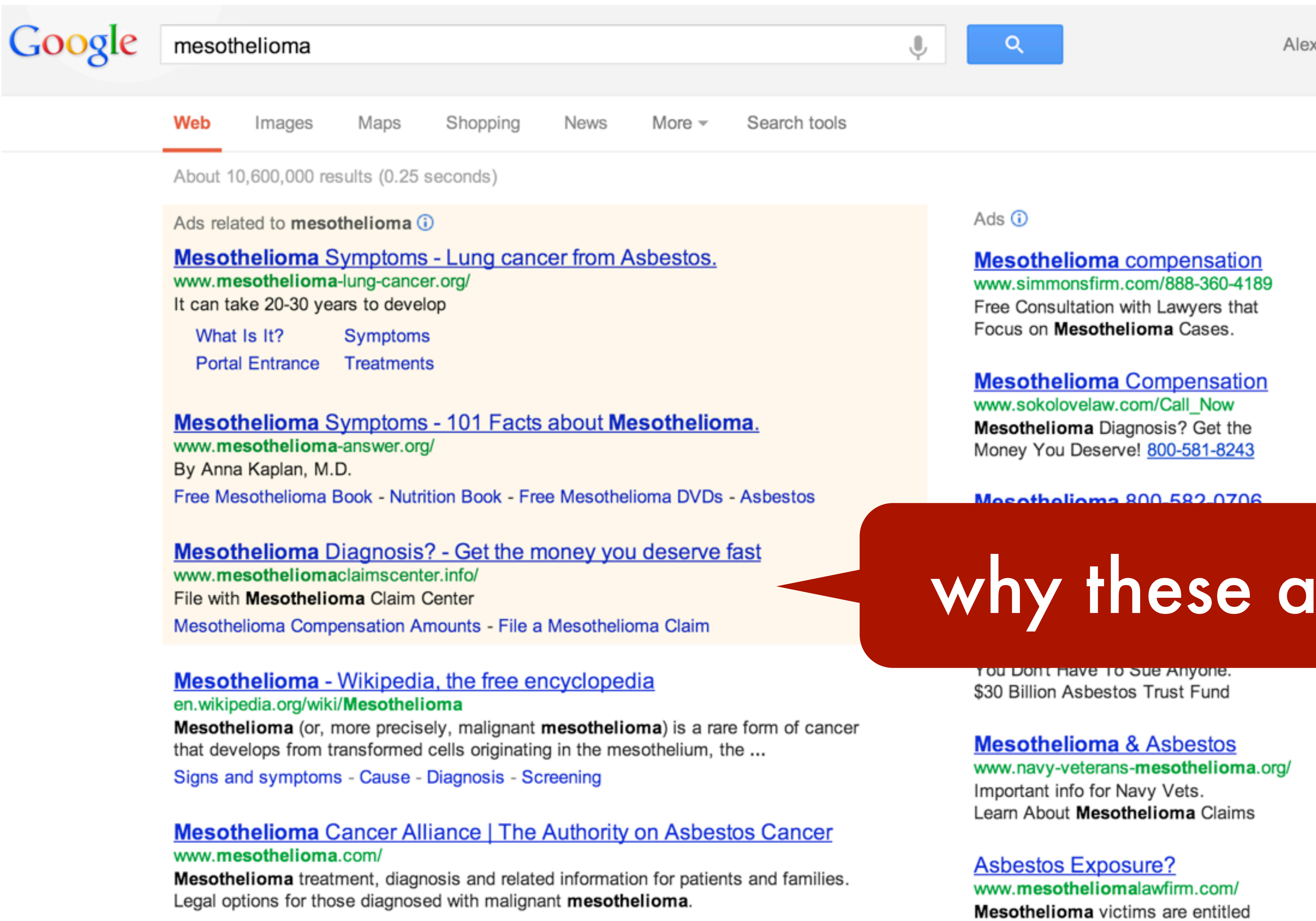
Image Layout



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

Carnegie Mellon University

Search ads



why these ads?

True startup story

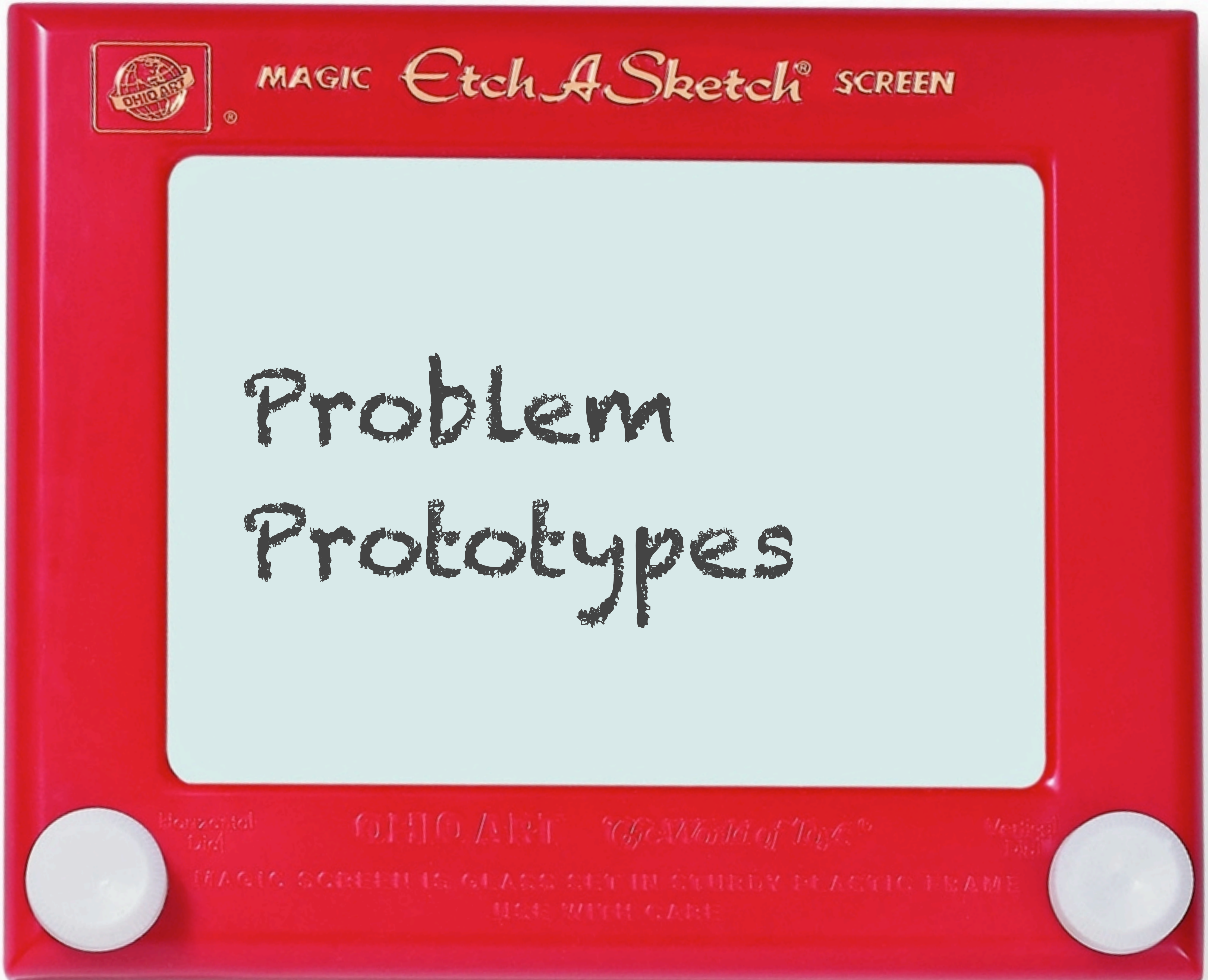
- Startup builds exchange for ads on webpages
- Clients bid on opportunities, market takes a cut

- 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)

Problem Prototypes



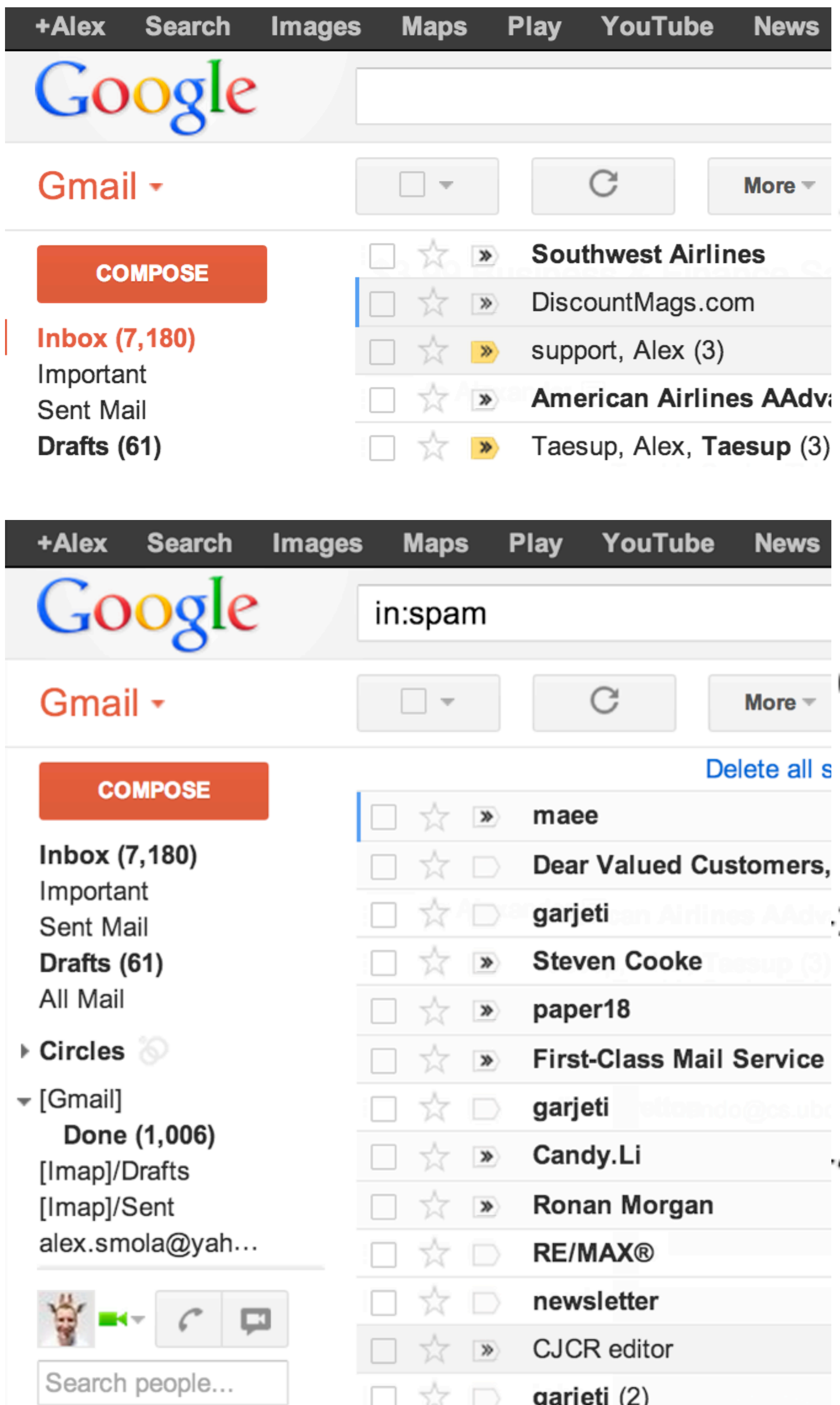
Supervised Learning $y = f(x)$

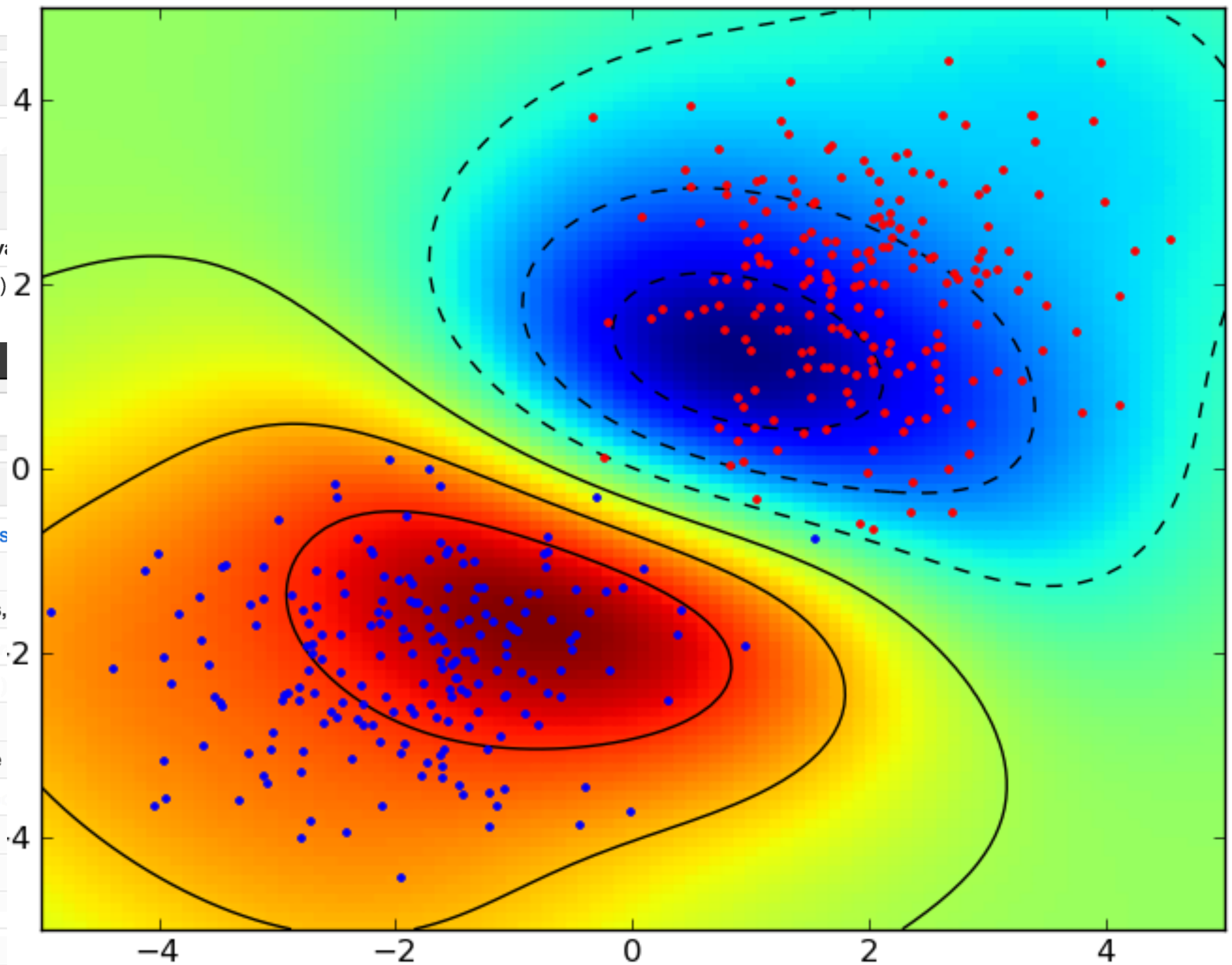
- Binary classification
  Given x find y in {-1, 1}
- Multicategory classification
  Given x find y in {1, ... k}
- Regression
  Given x find y in R (or R^d)
- Sequence annotation
  Given sequence x_1 ... x_l find y_1 ... y_l
- Hierarchical Categorization (Ontology)
  Given x find a point in the hierarchy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

often with loss

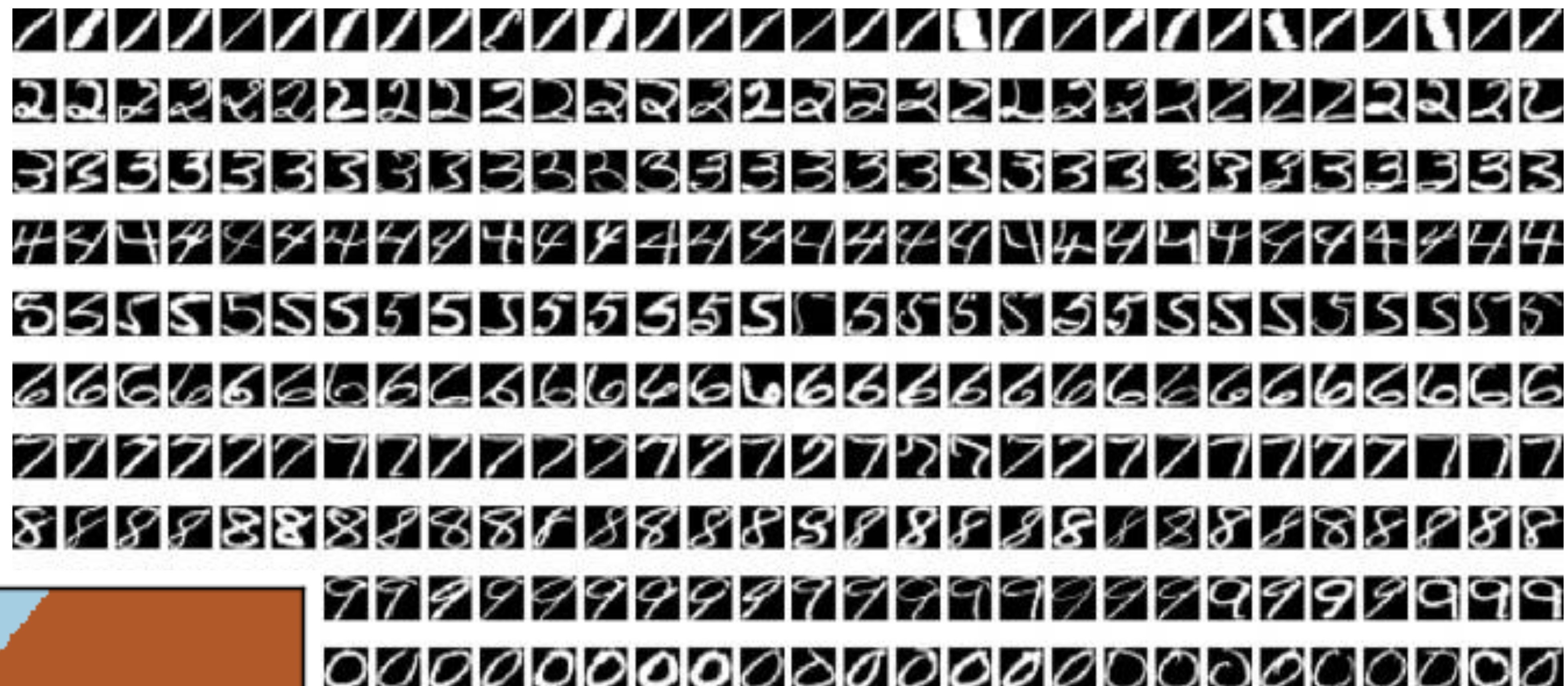
$$l(y, f(x))$$

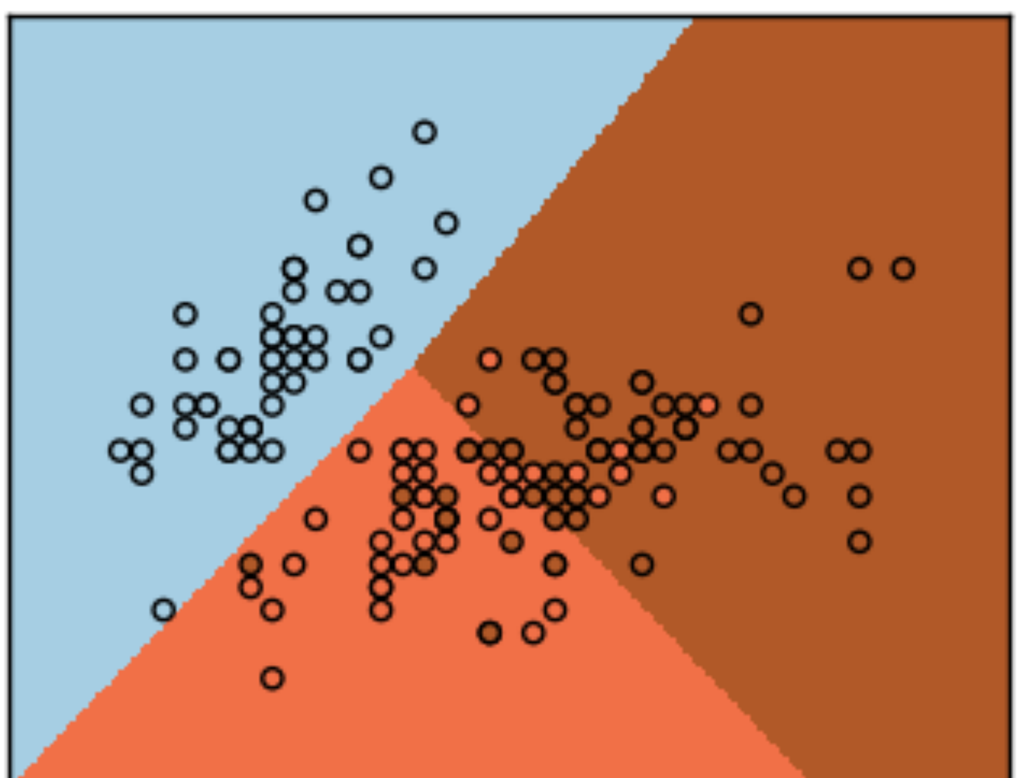
Binary Classification





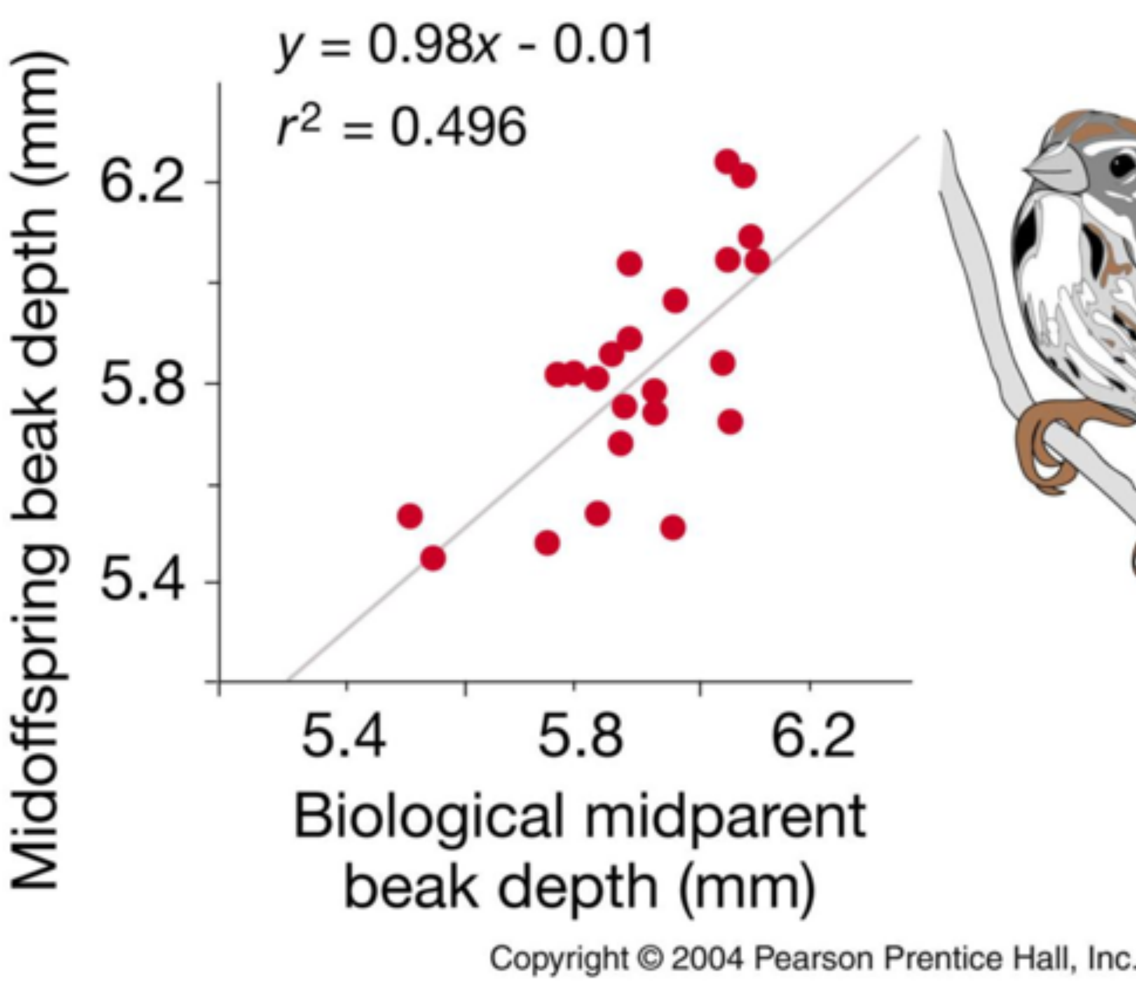
Multiclass Classification



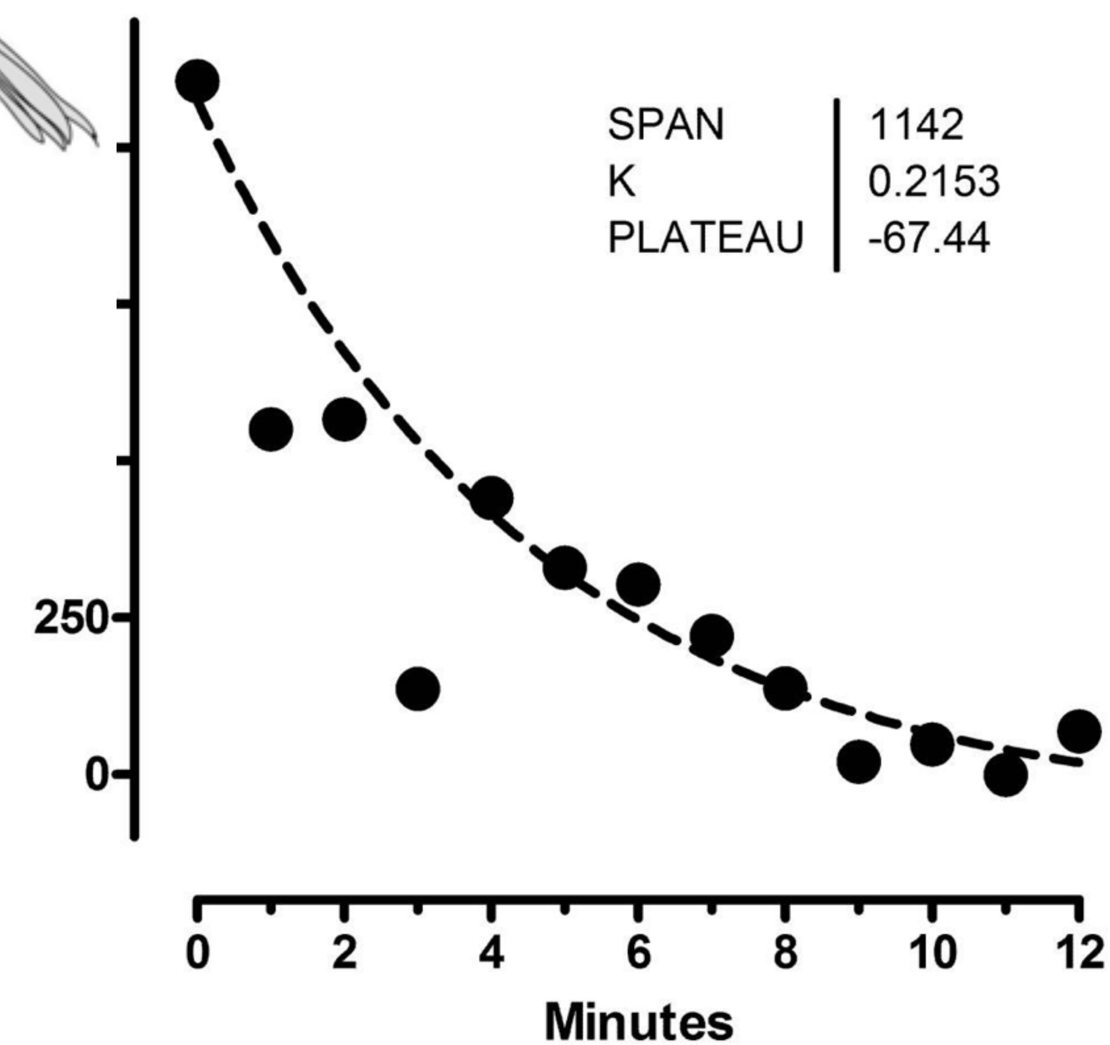


map image x to digit y

Regression

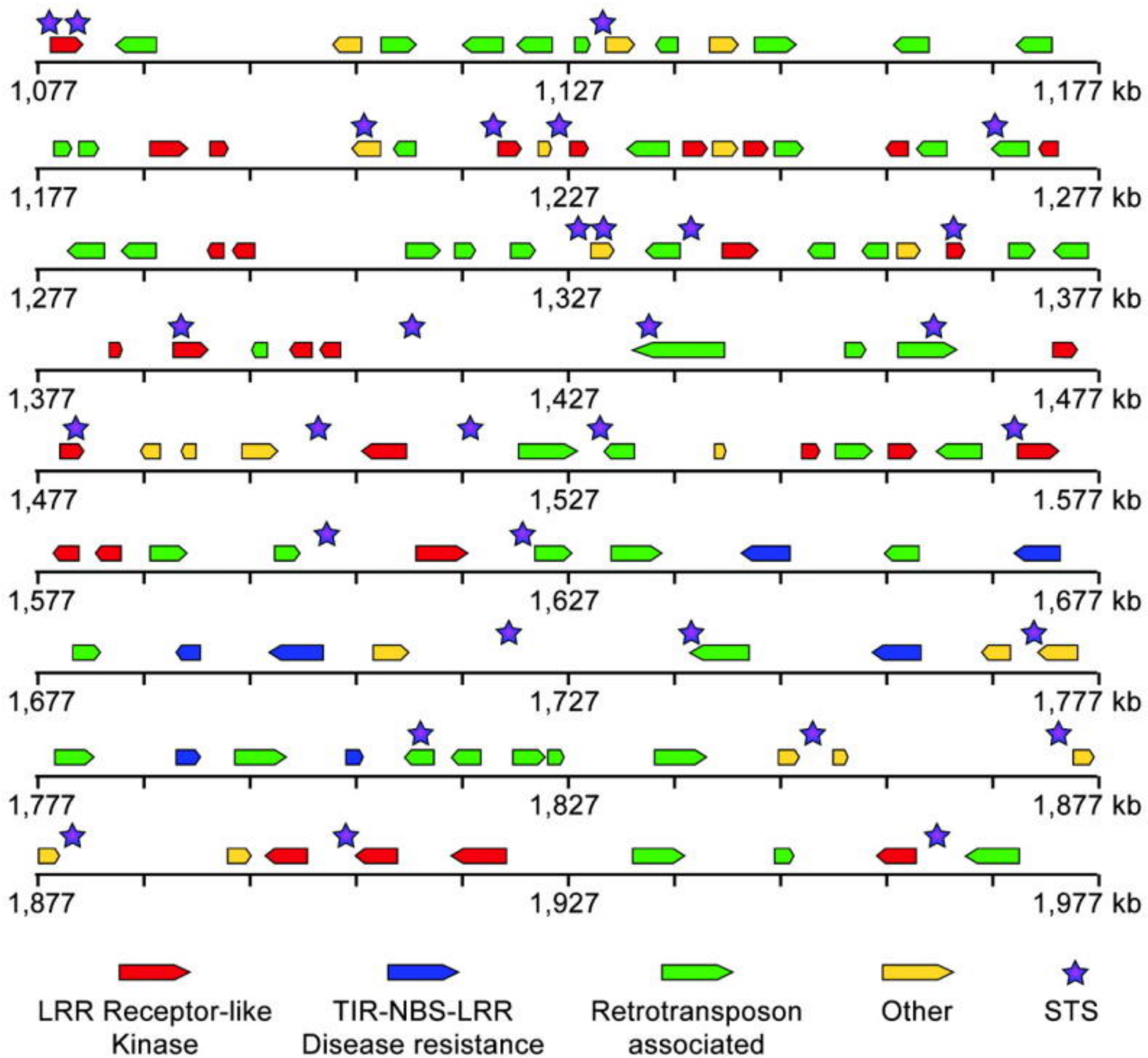


nonlinear



linear

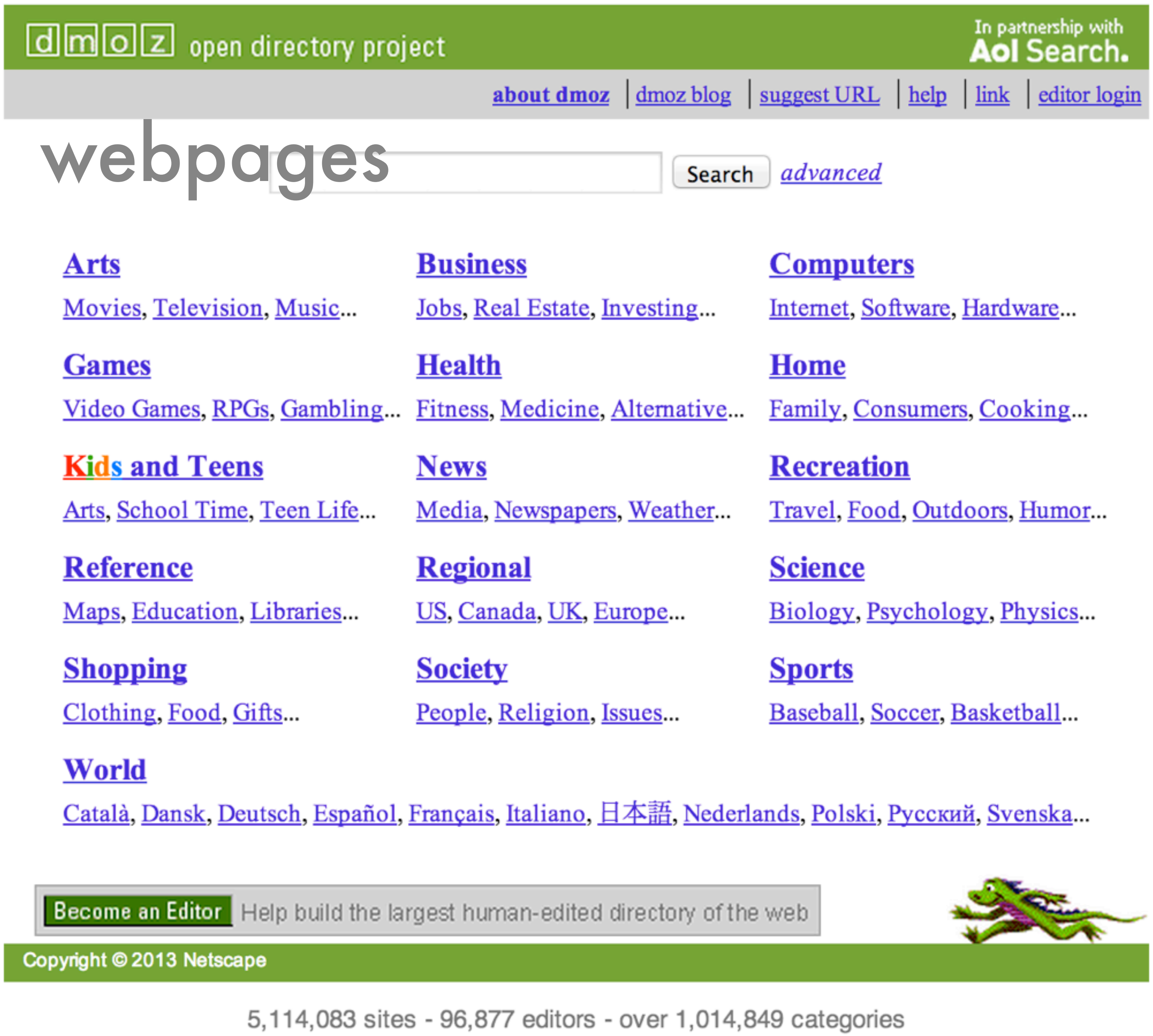
Sequence Annotation



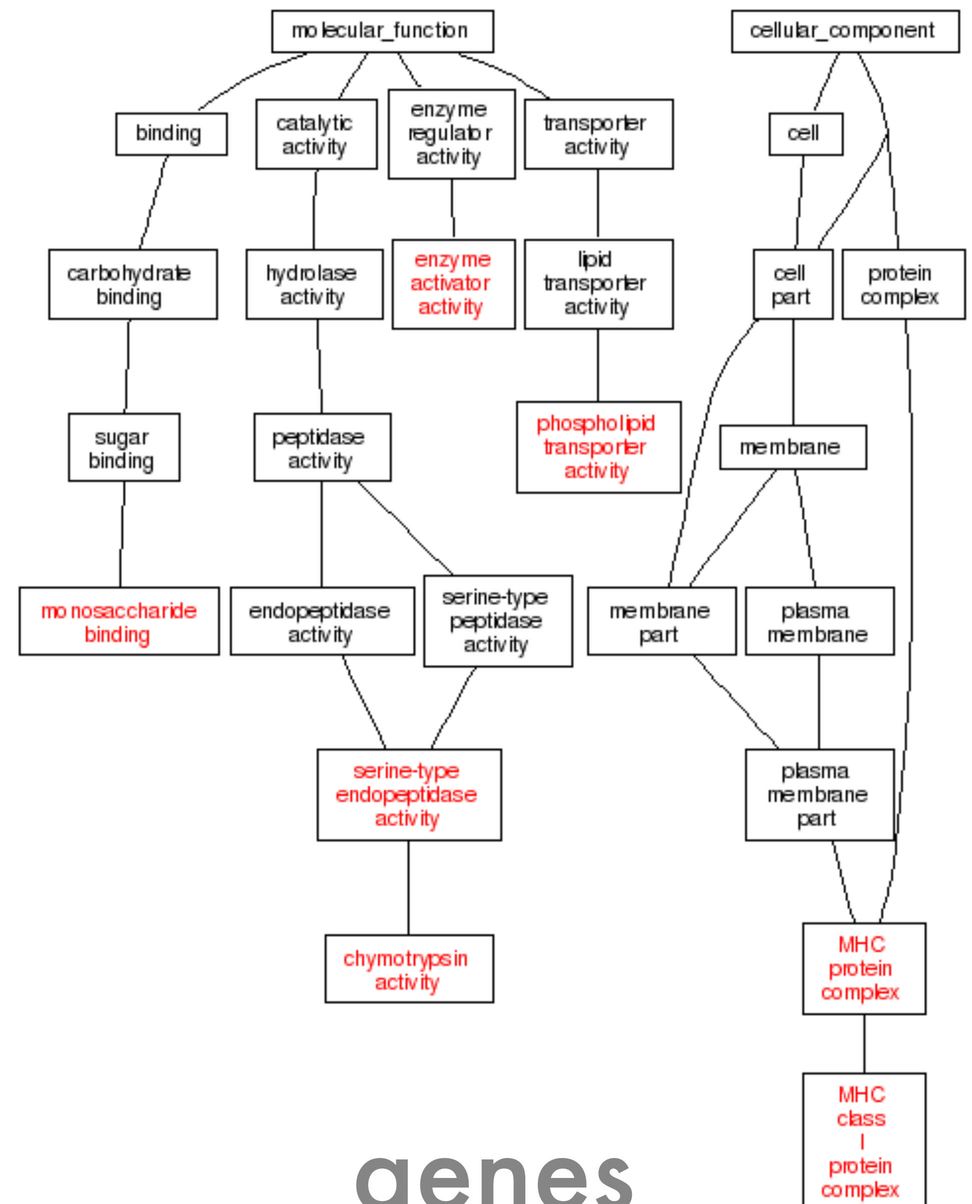
given sequence

gene finding
speech recognition
activity segmentation
named entities

Ontology



webpages



genes

Prediction

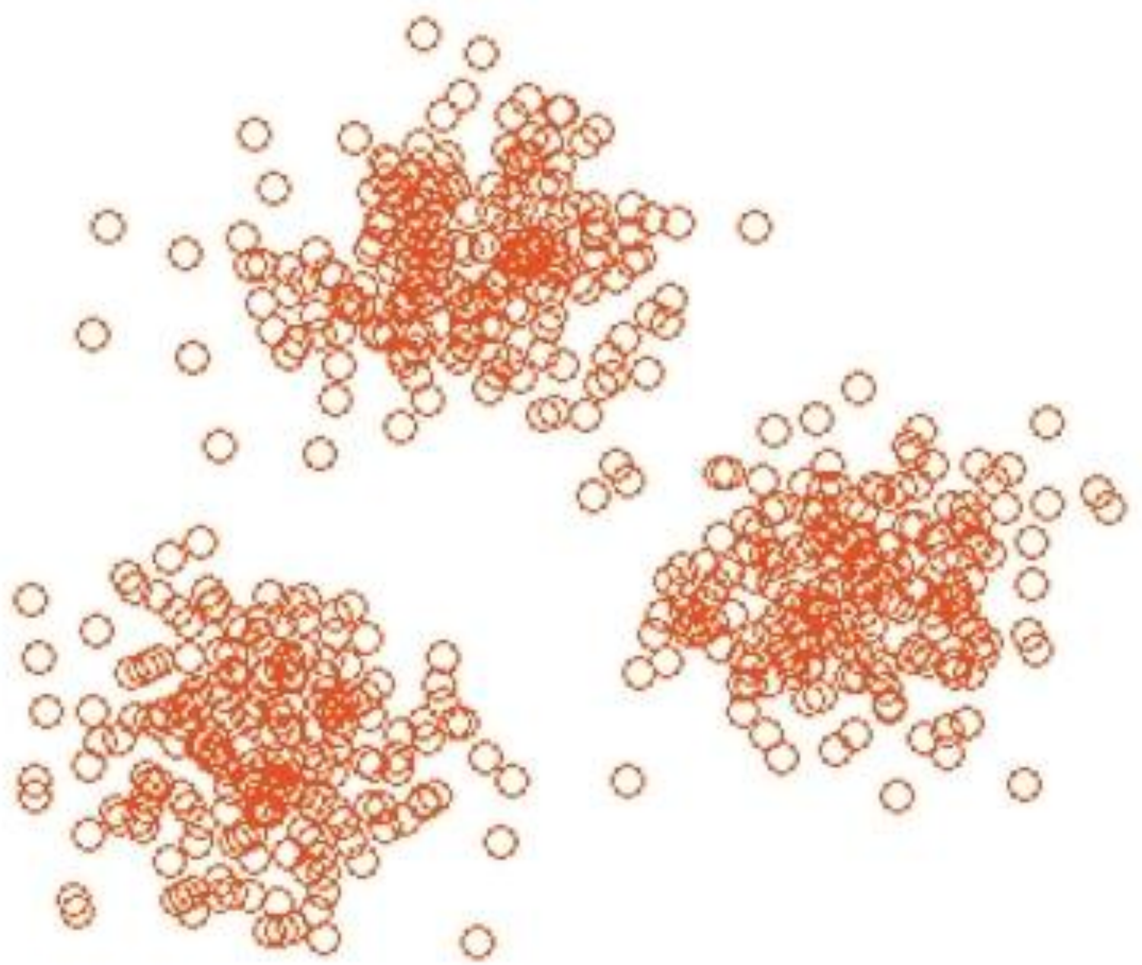
tomorrow's stock price

Carnegie Mellon University

Unsupervised Learning

- 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

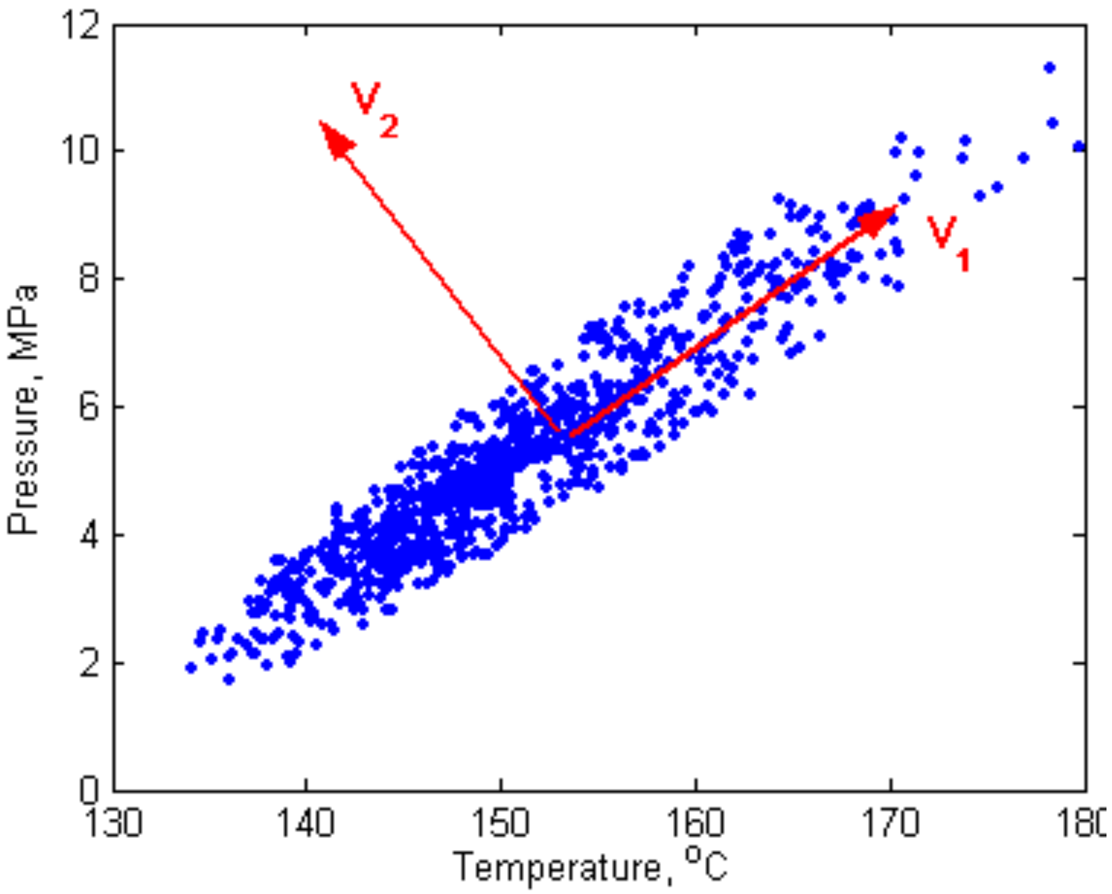
Clustering

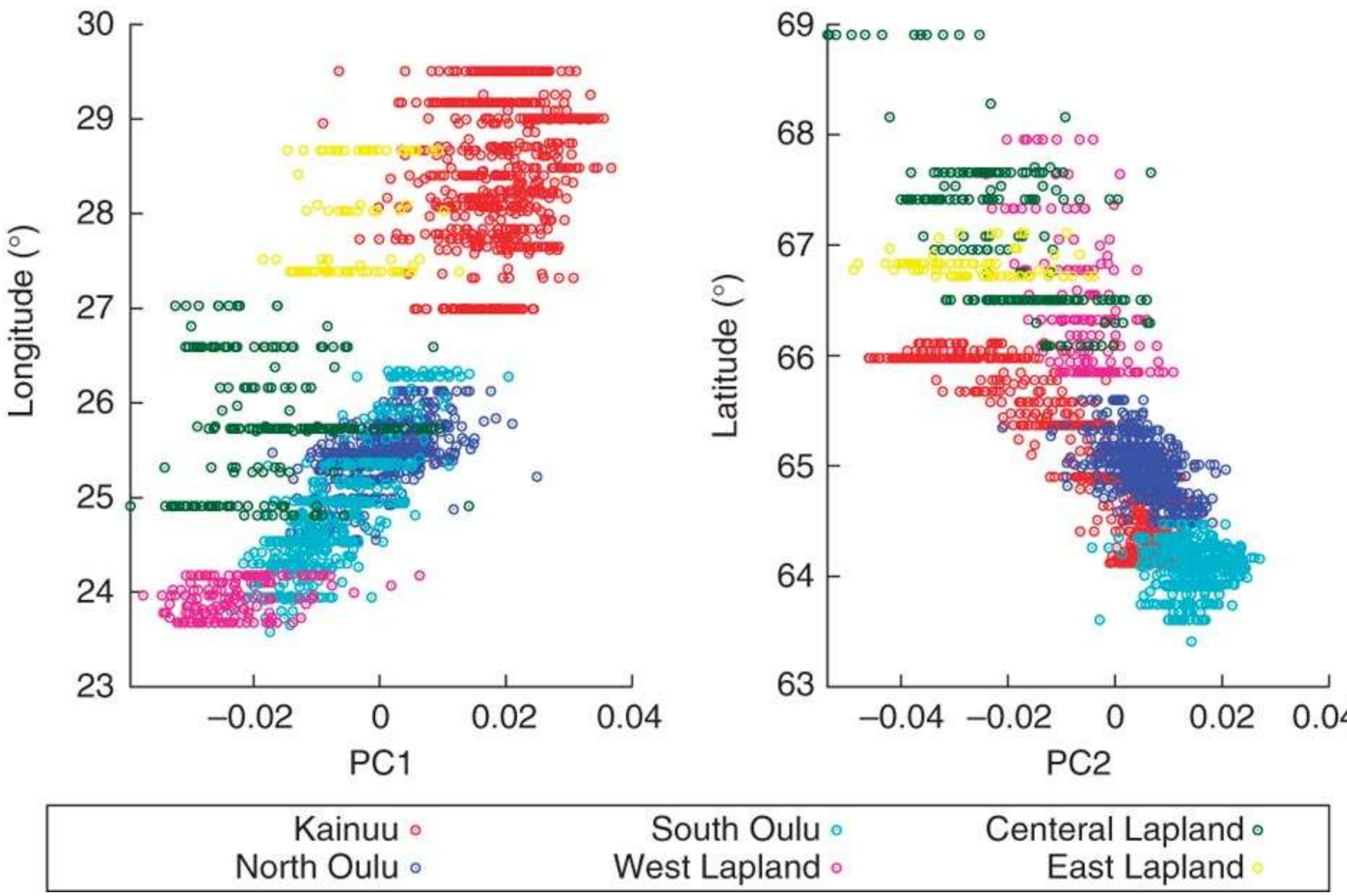


- Documents
- Users
- Webpages
- Diseases
- Pictures
- Vehicles

…

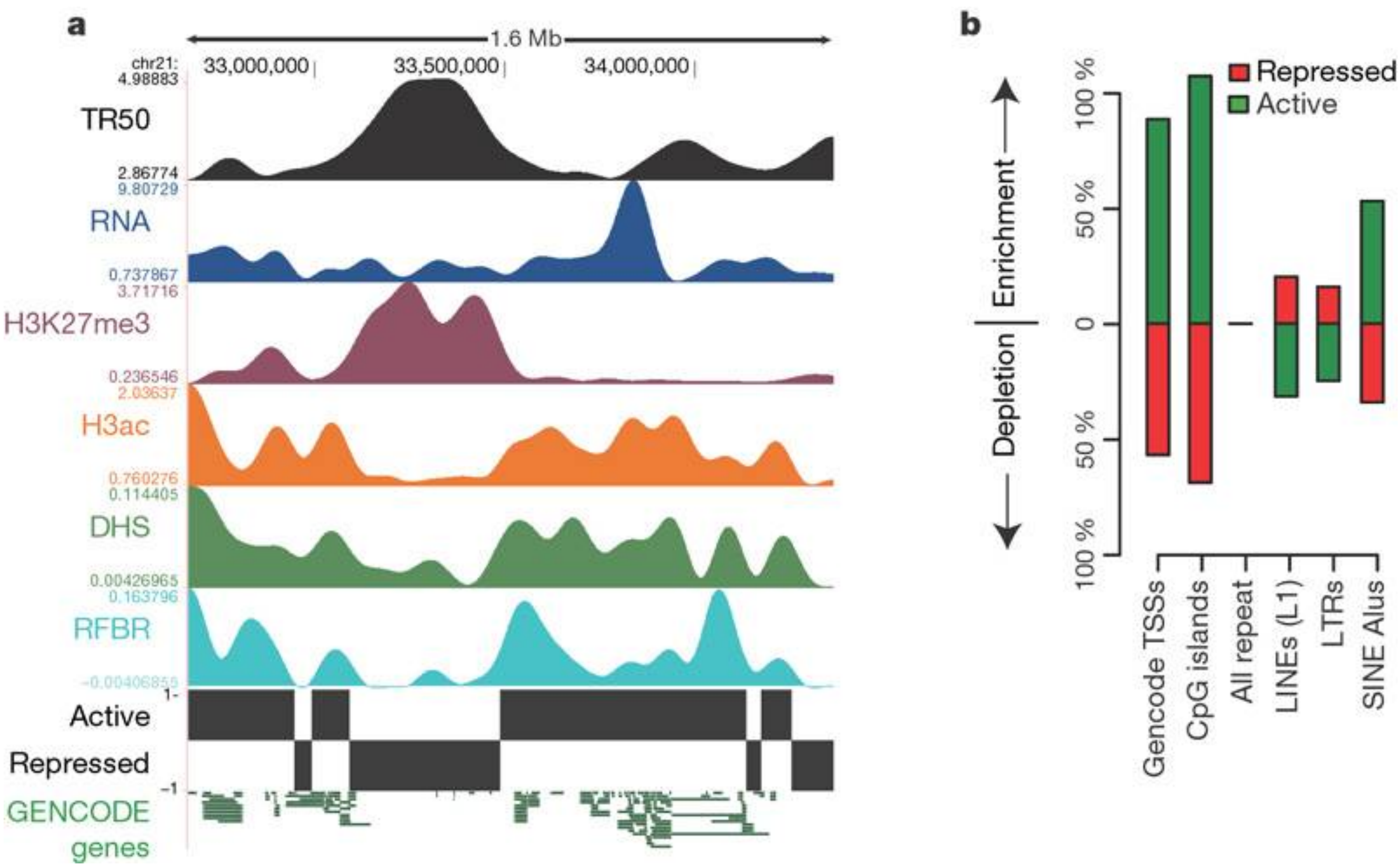
Principal Components





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

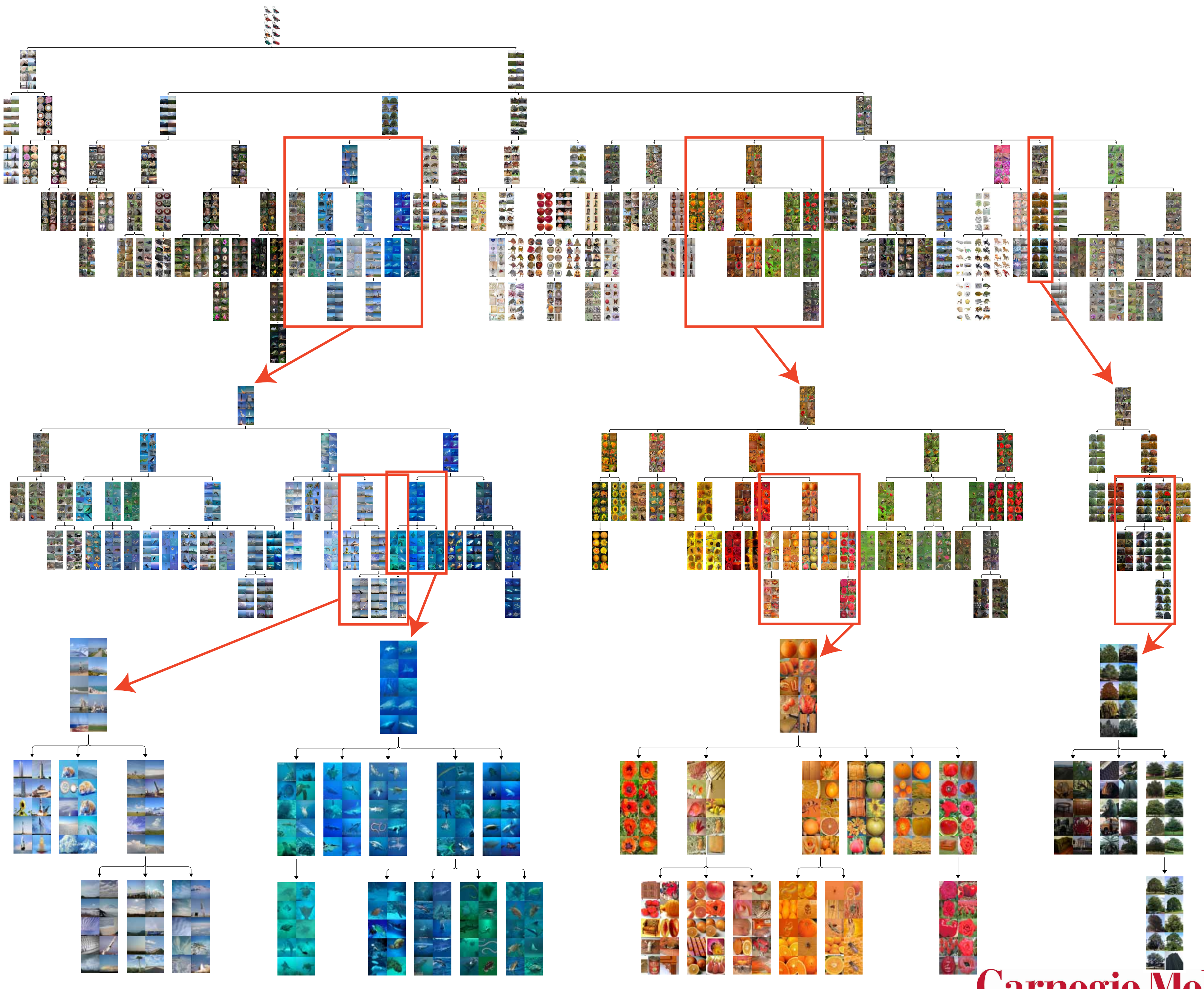
Sequence Analysis



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

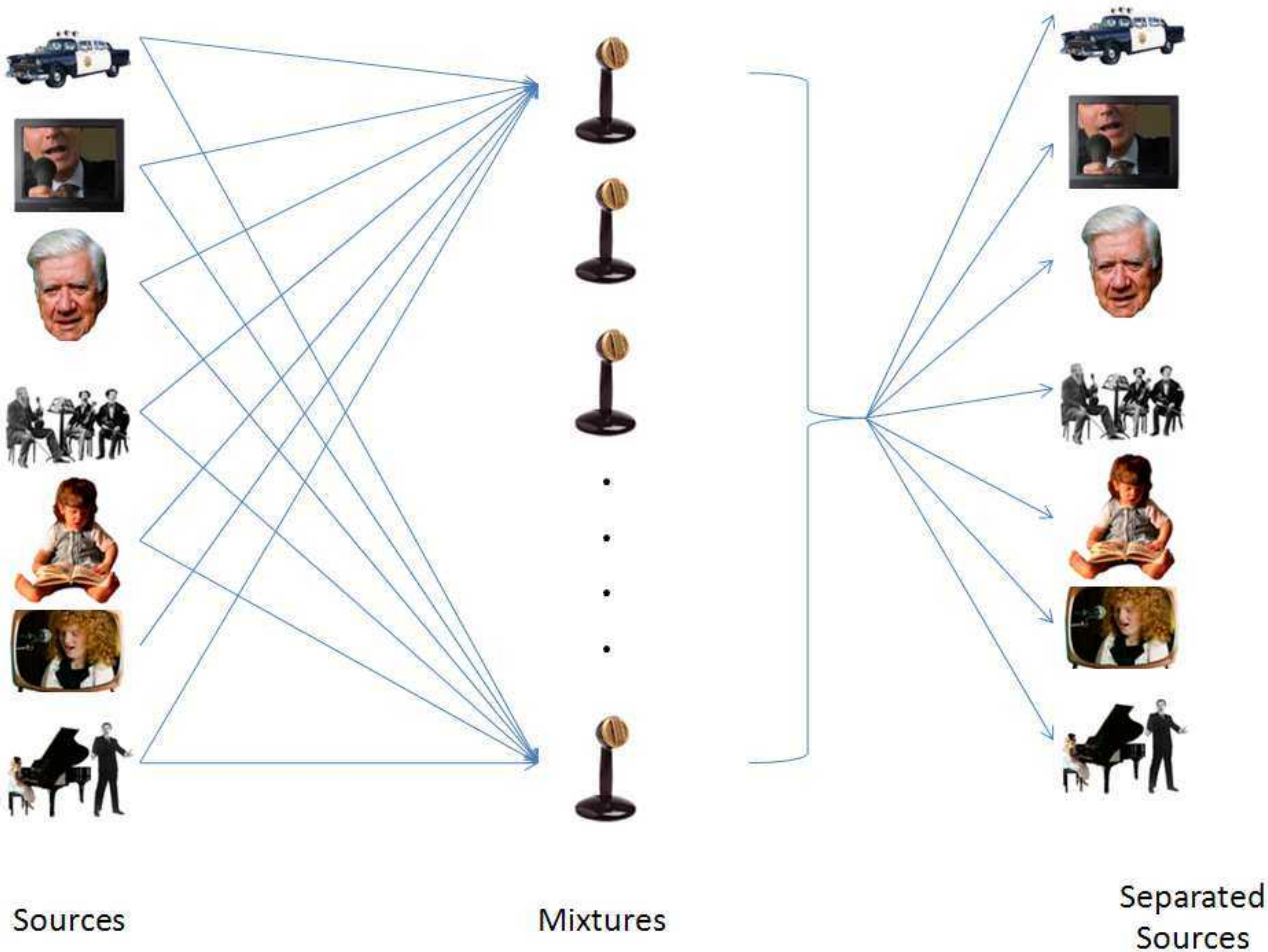
Carnegie Mellon University

Hierarchical Grouping



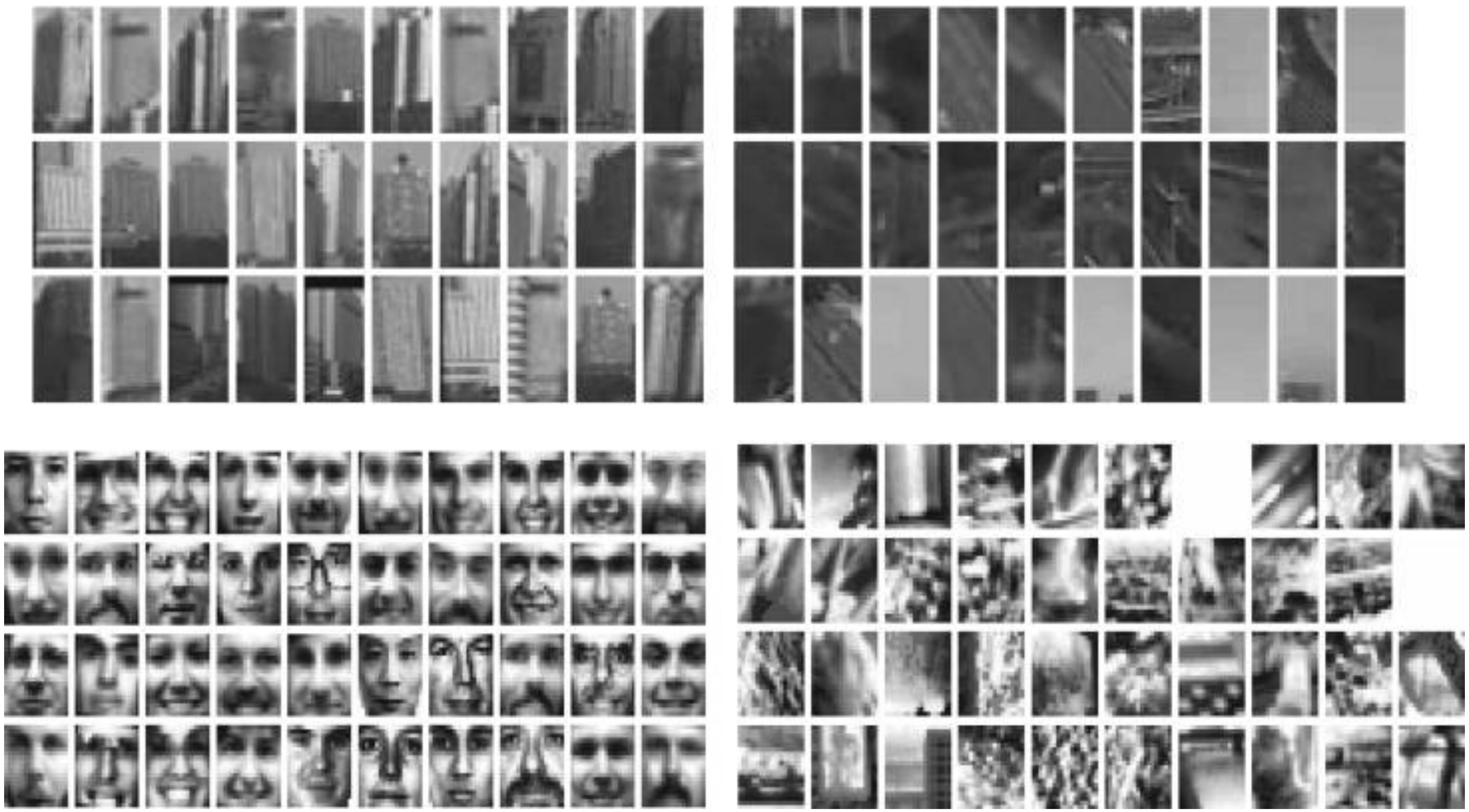
Carnegie Mellon University

Independent Components



find them automatically

Novelty detection



typical | atypical

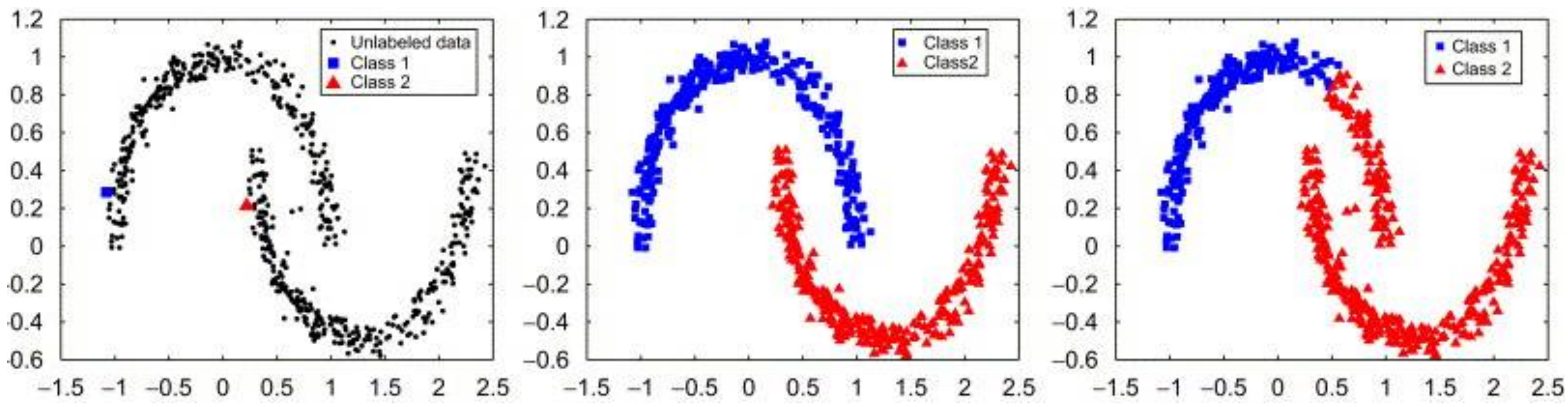
Carnegie Mellon University

Some Problem types

iid = Independently Identically Distributed

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

Induction - Transduction



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

Covariate Shift

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

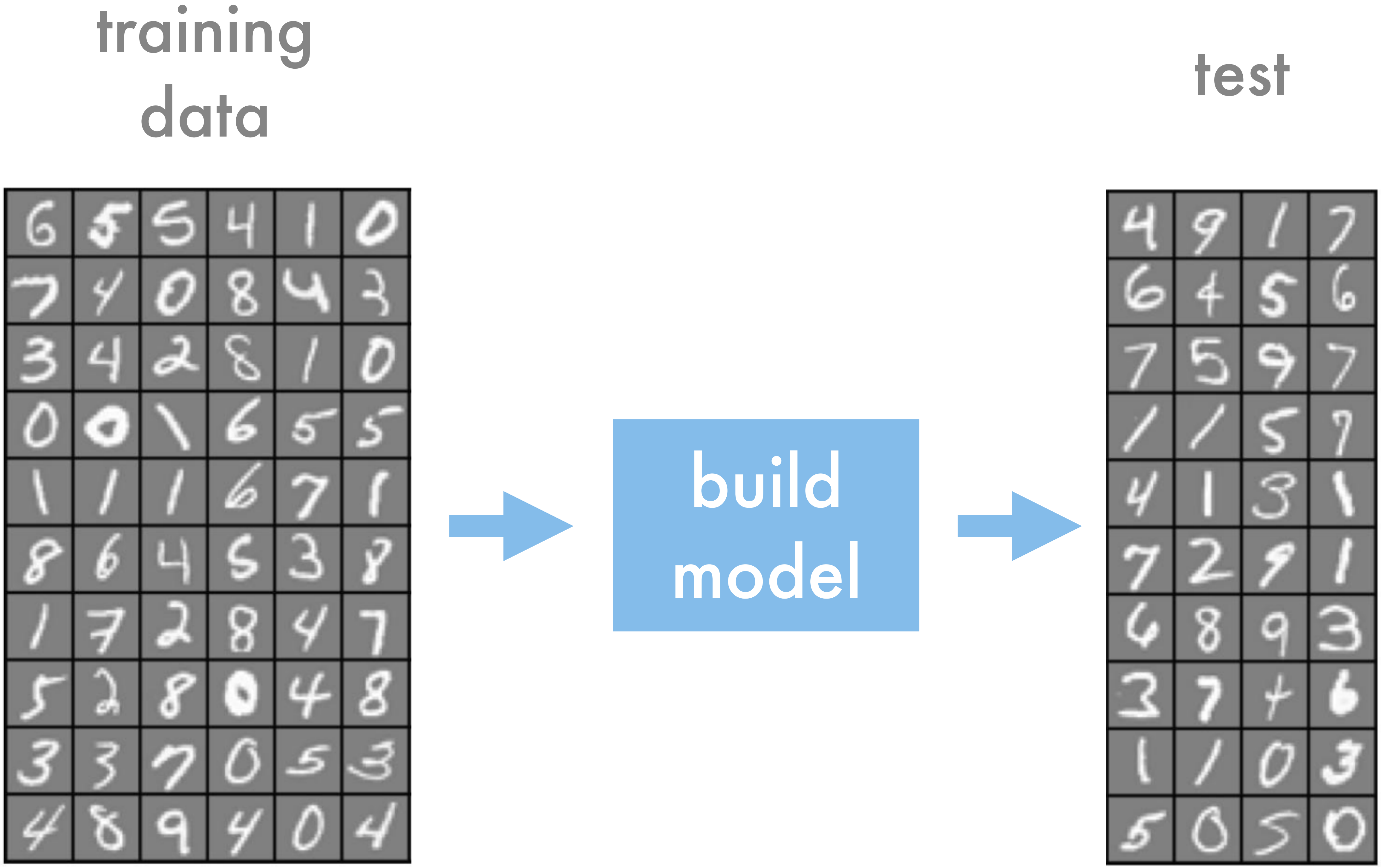
Cotraining and Multitask

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

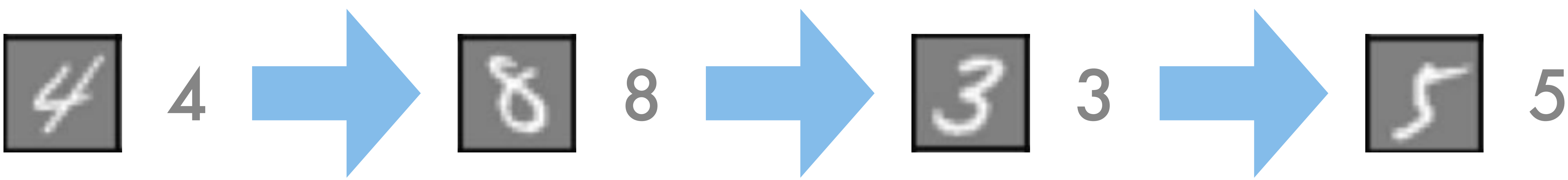
Interaction with Environment

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

Batch



Online



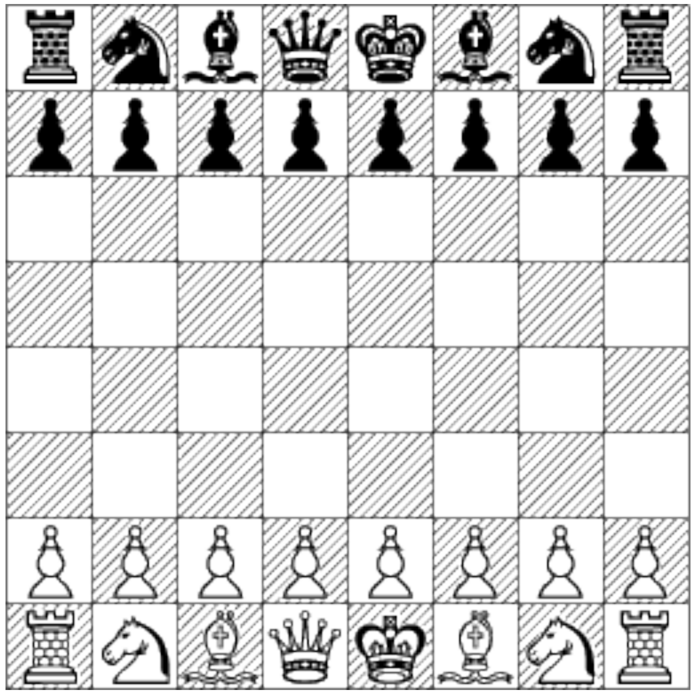
Bandits

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




Reinforcement Learning

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

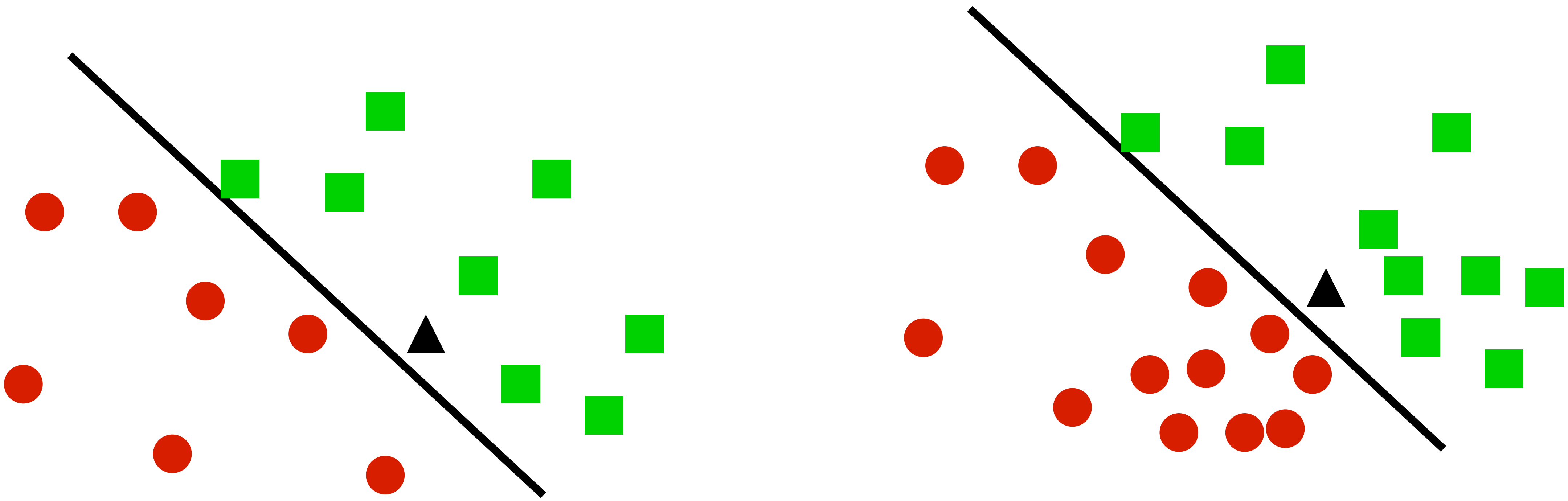


environment (cooperative, adversary, doesn’t care)
memory (goldfish, elephant)
state space (tic tac toe, chess, car)

Discriminative vs. Generative (mainly relevant for supervised models)

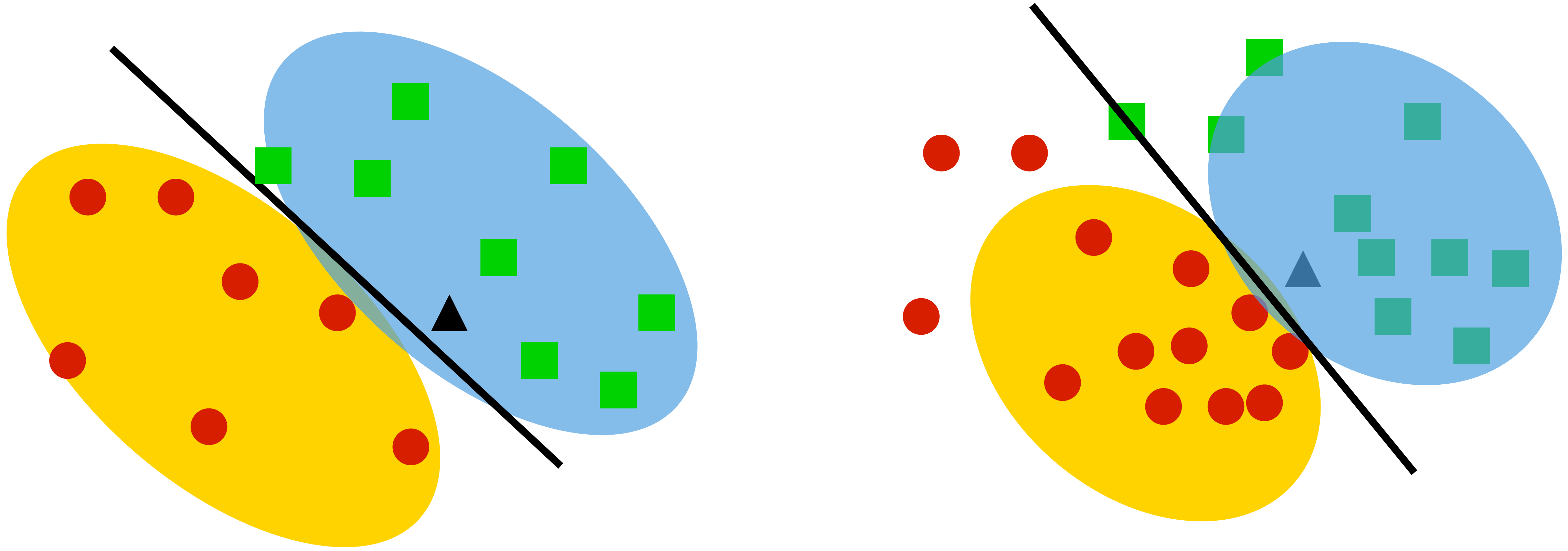
- Discriminative Models
  - Estimate y|x directly
  - Often better convergence + simpler solutions
- Generative models
  - Estimate joint distribution over (x,y)
  - Use conditional probability to infer y|x
  - Often more intuitive
  - Easier to add prior knowledge

Discriminative



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

Generative

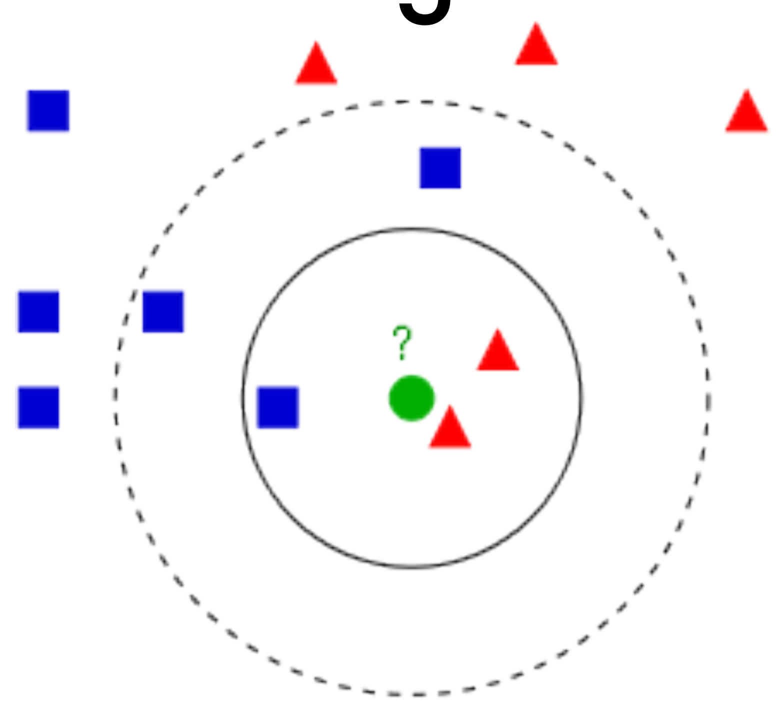


- Model observations (x,y) first
- Then infer p(y|x)
- Good for missing variables, better diagnostics
- Easy to add prior knowledge about data

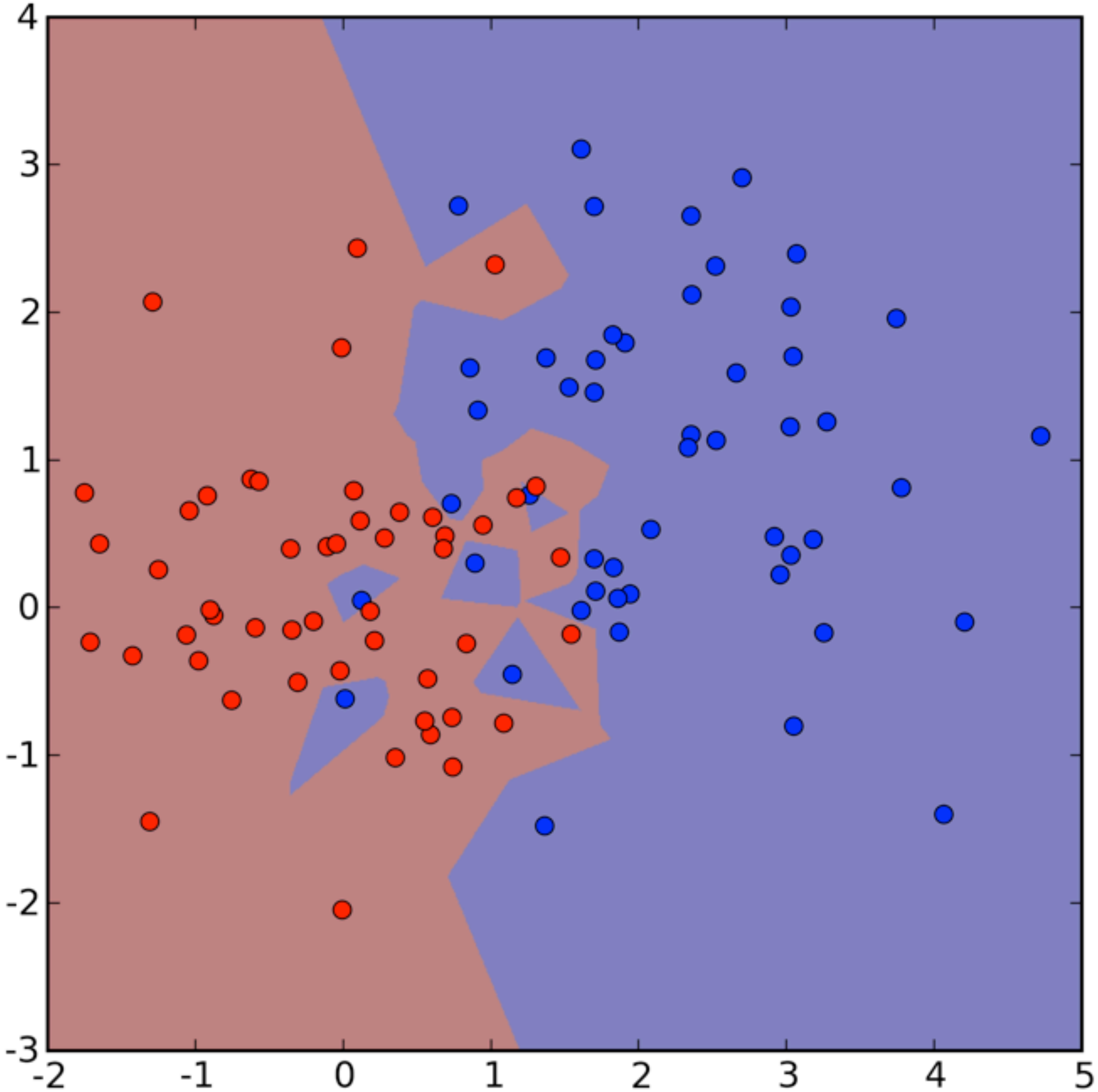
Very Basic Tools

Nearest Neighbors

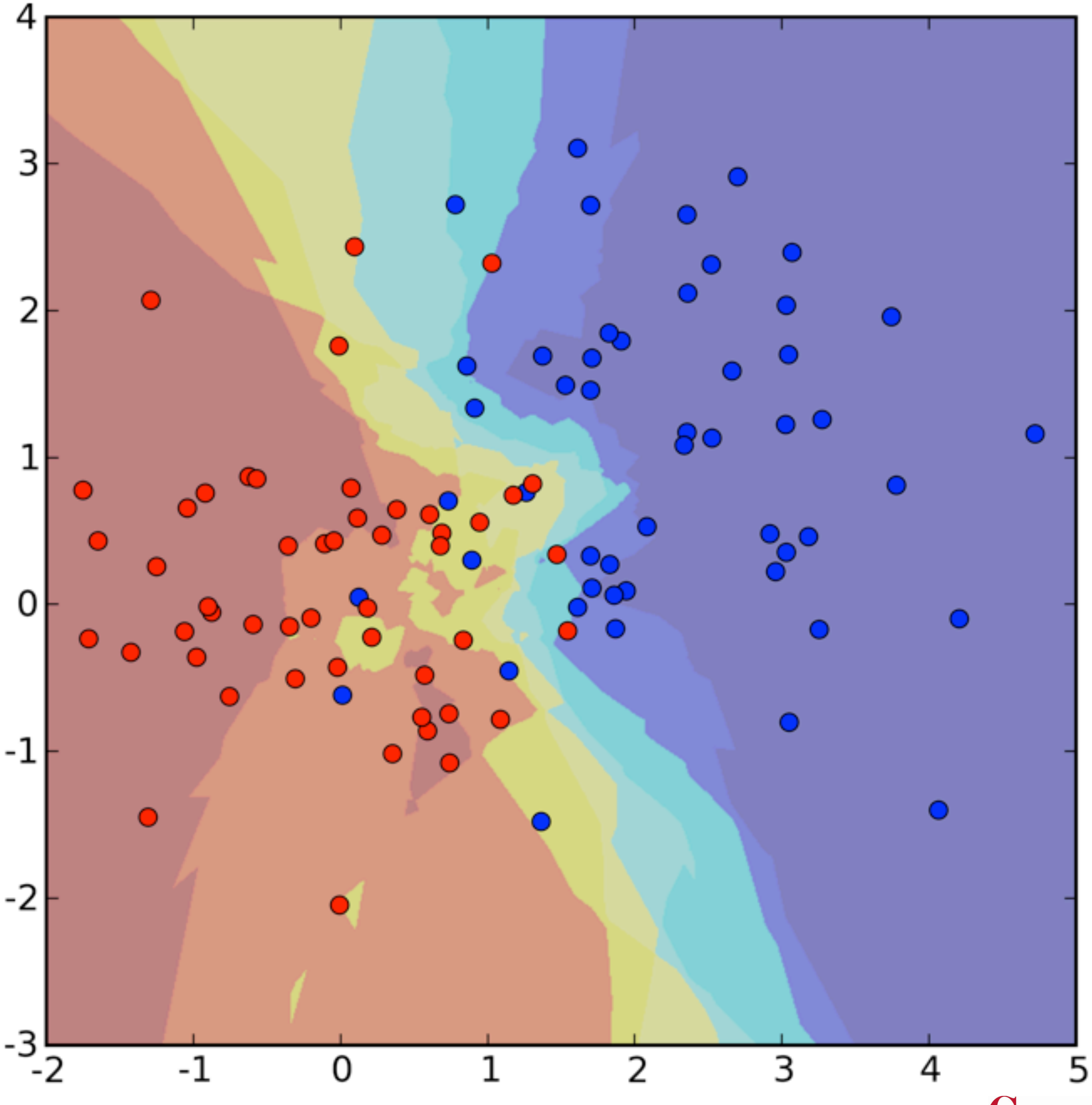
- Table lookup
  For previously seen instance remember label
- Nearest neighbor
  - Pick label of most similar neighbor
  - Slight improvement - use k-nearest neighbors
  - For regression average
  - Really useful baseline!
  - Easy to implement for small amounts of data. Why?



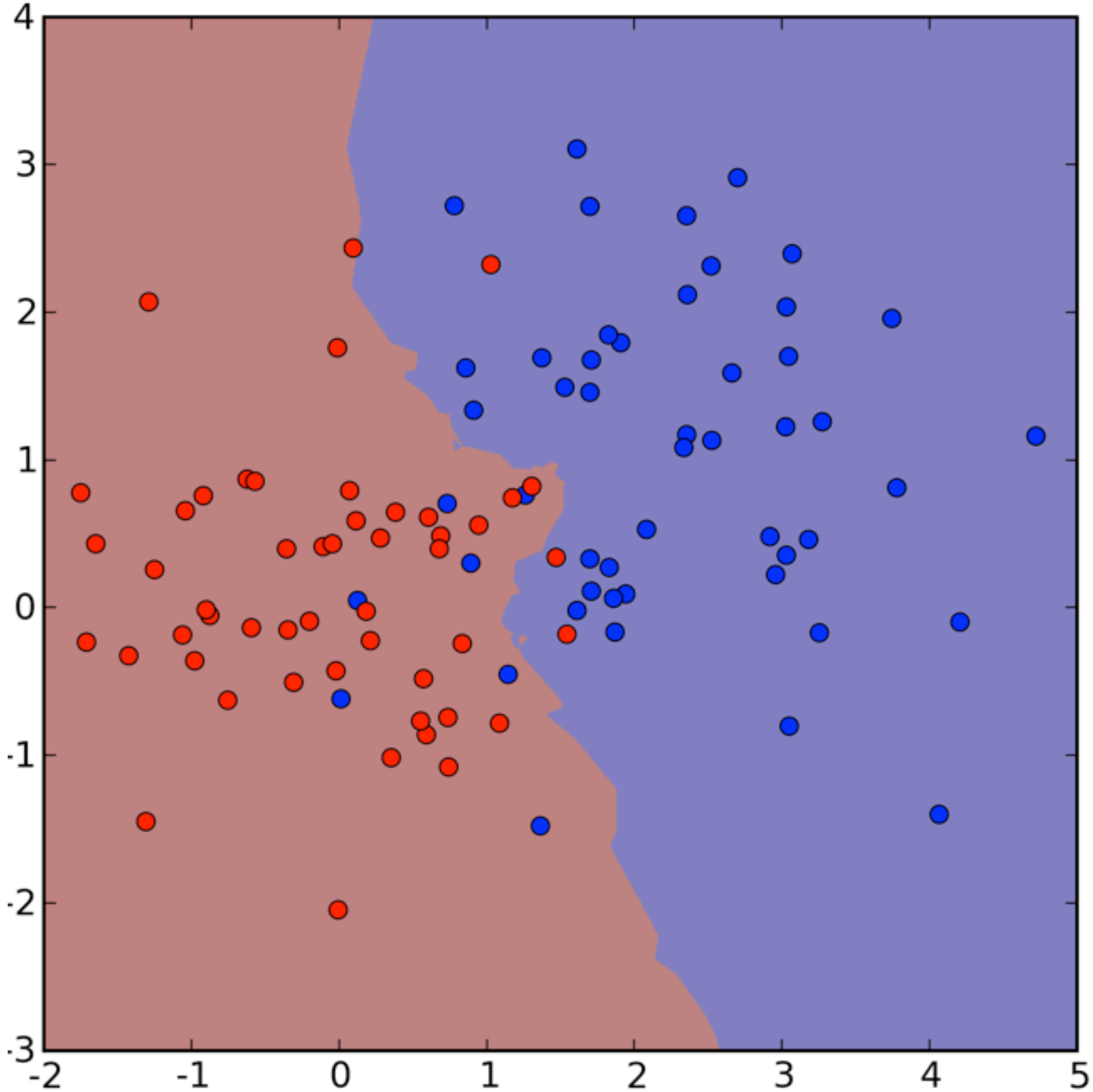
1-Nearest Neighbor



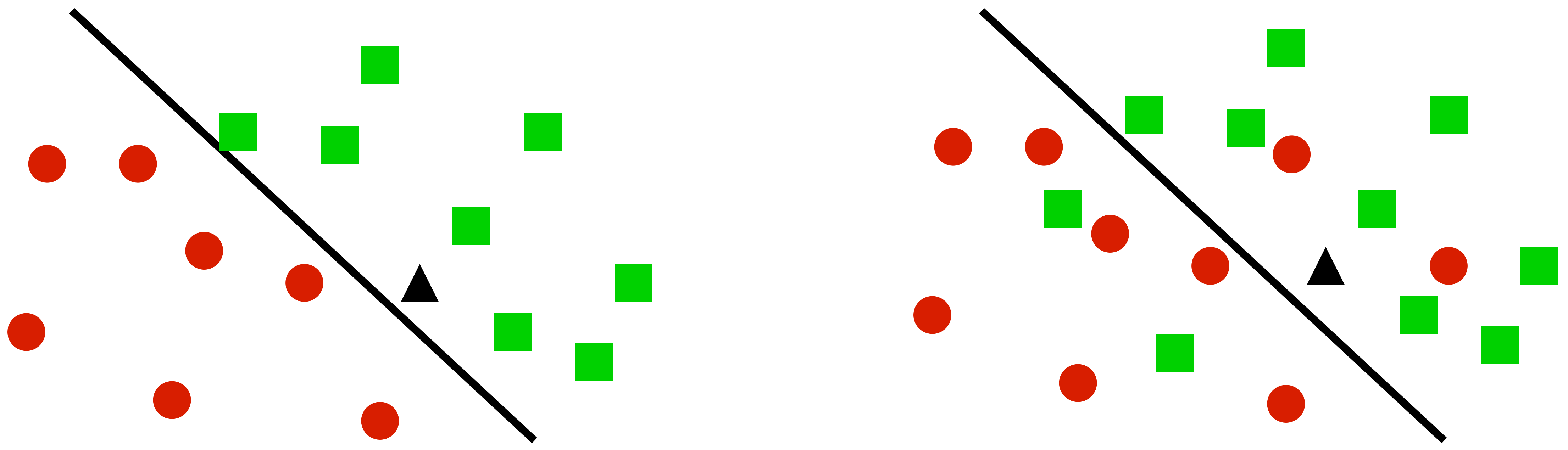
4-Nearest Neighbors



4-Nearest Neighbors Sign

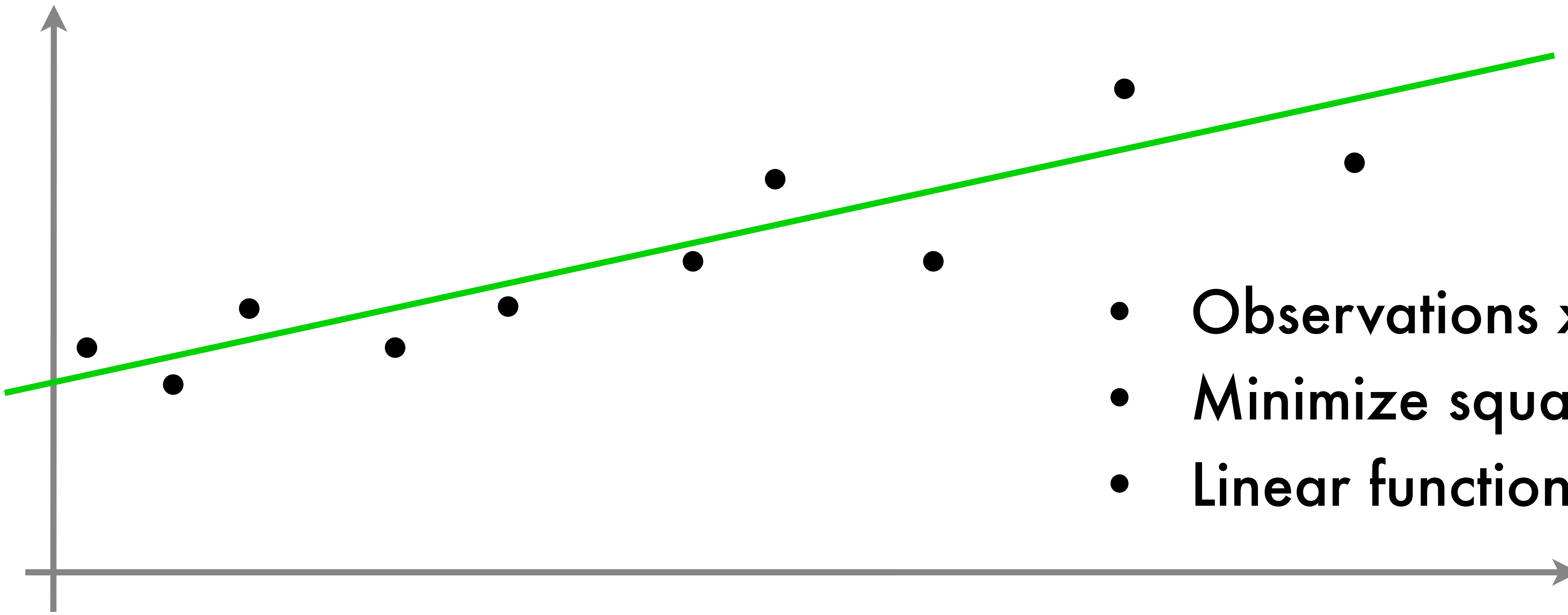


If we get more data



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

Linear Regression



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

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

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

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

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

Linear Regression

- Optimization Problem

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

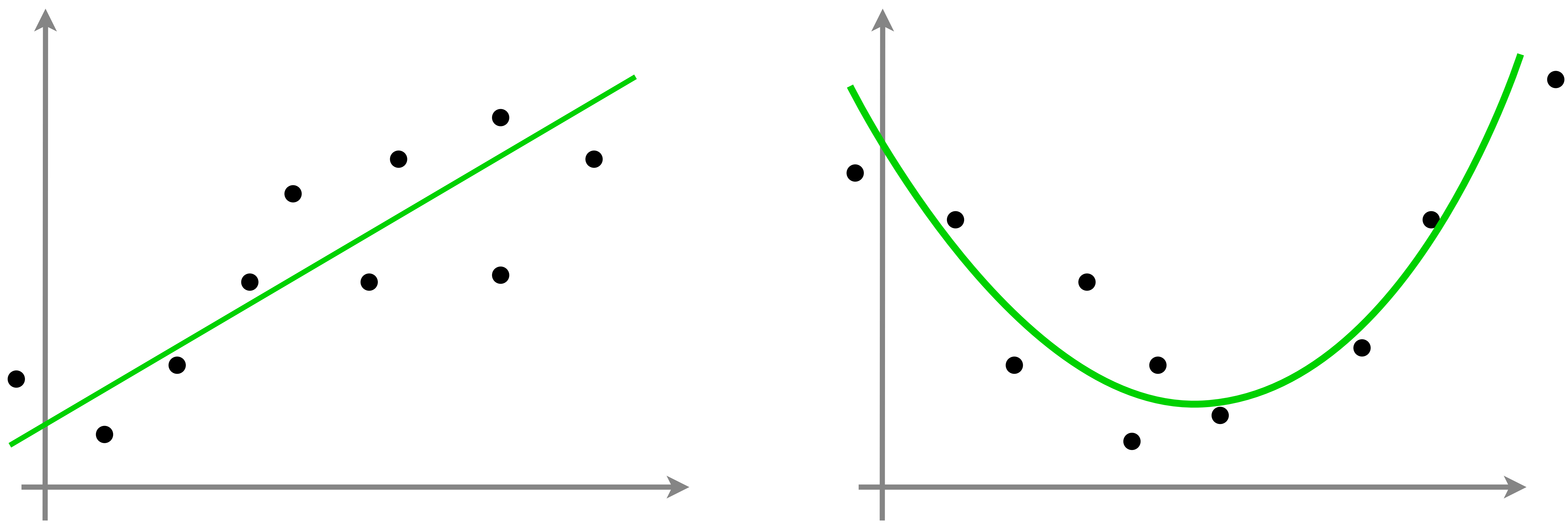
$$\underset{w}{\text{minimize}} \sum_{i=1}^{m} \frac{1}{2} (\langle w, \bar{x}_i \rangle - y_i)^2$$

- Solving it

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

only requires a matrix inversion.

Nonlinear Regression



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

Linear Regression

- Optimization Problem

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

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

- Solving it

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

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

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

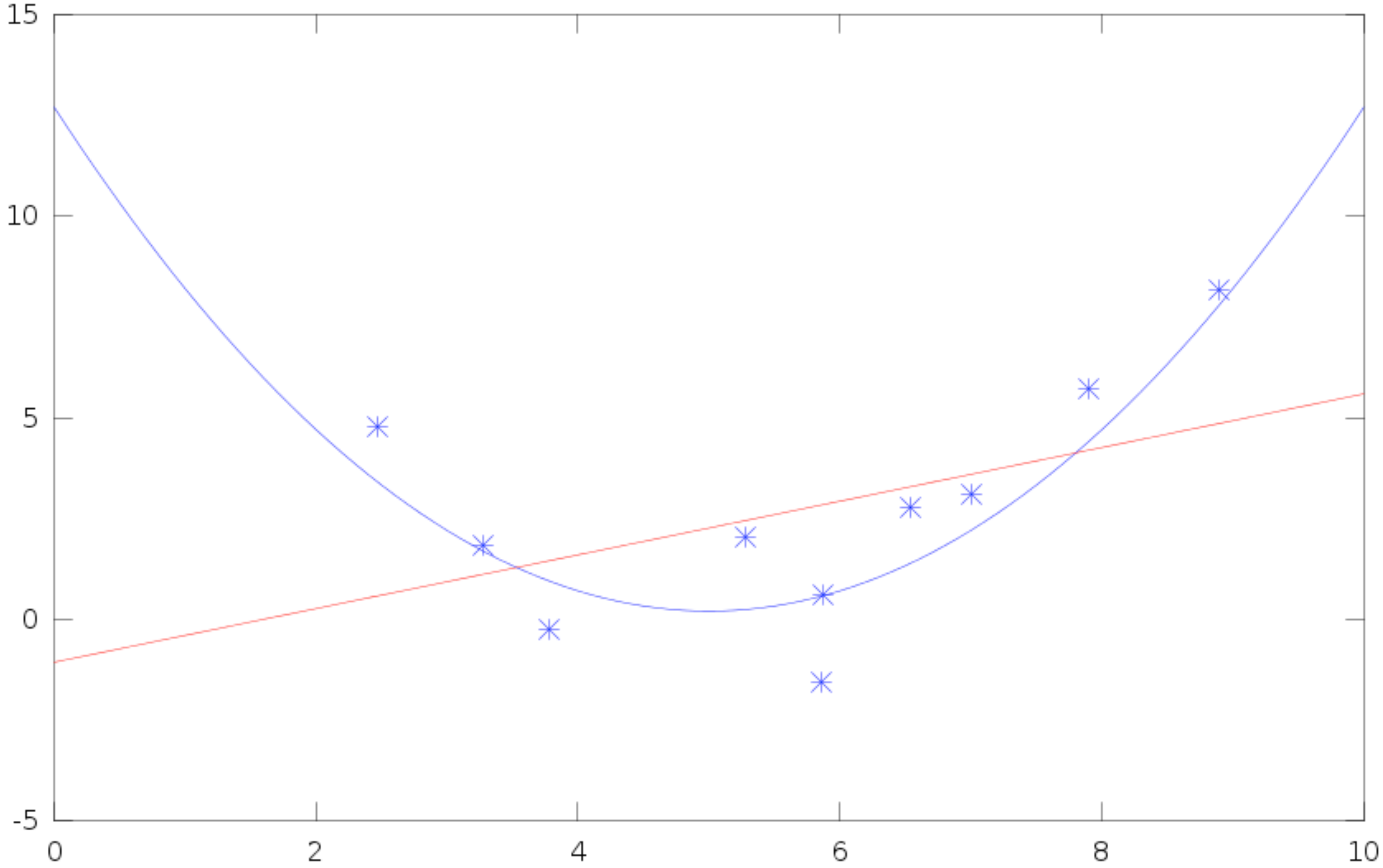
only requires a matrix inversion.

Pseudocode (degree 4)

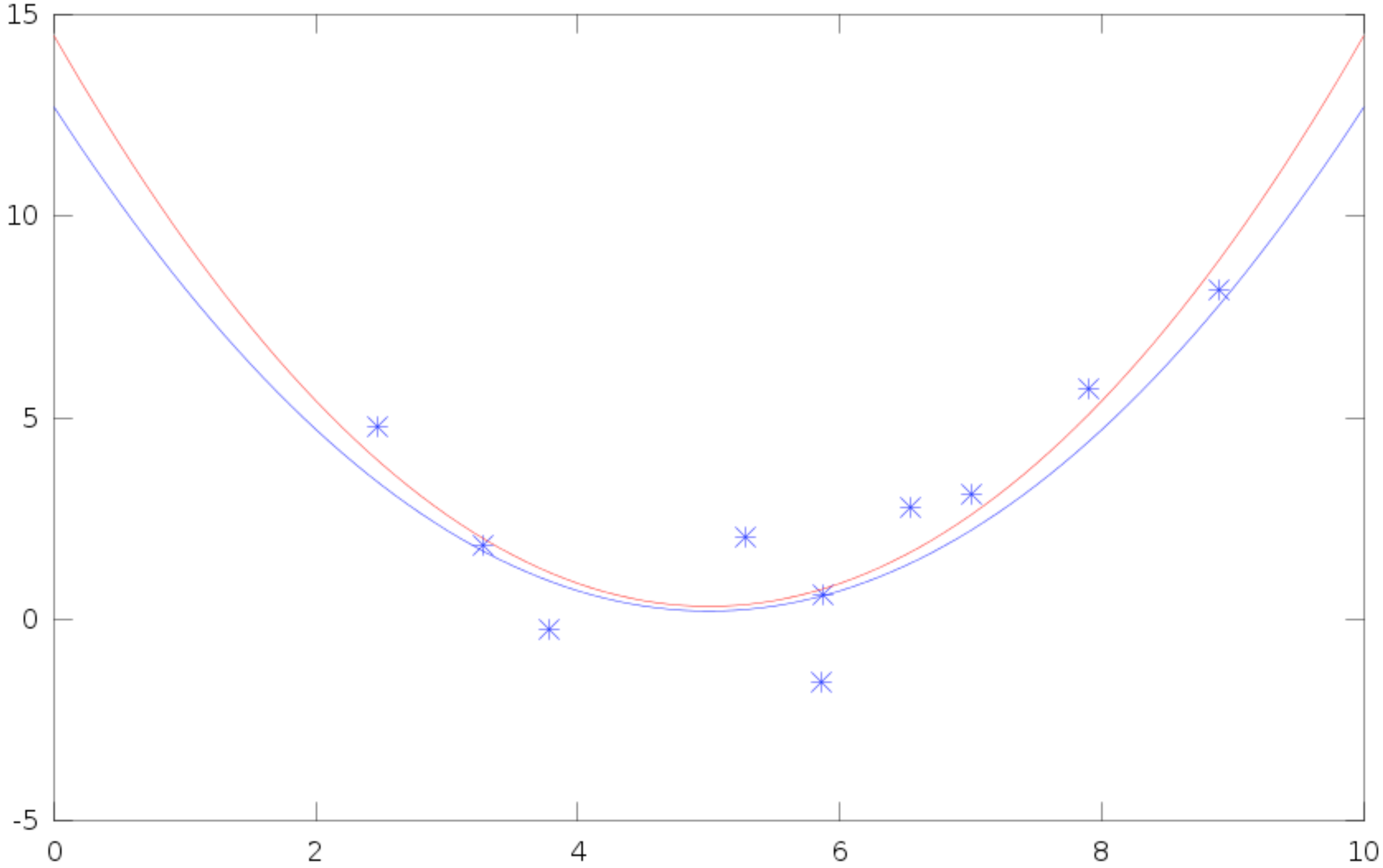
```
Training
phi_xx = [xx.^4, xx.^3, xx.^2, xx, 1.0 + 0.0 * xx];
w = (yy' * phi_xx) / (phi_xx' * phi_xx);

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

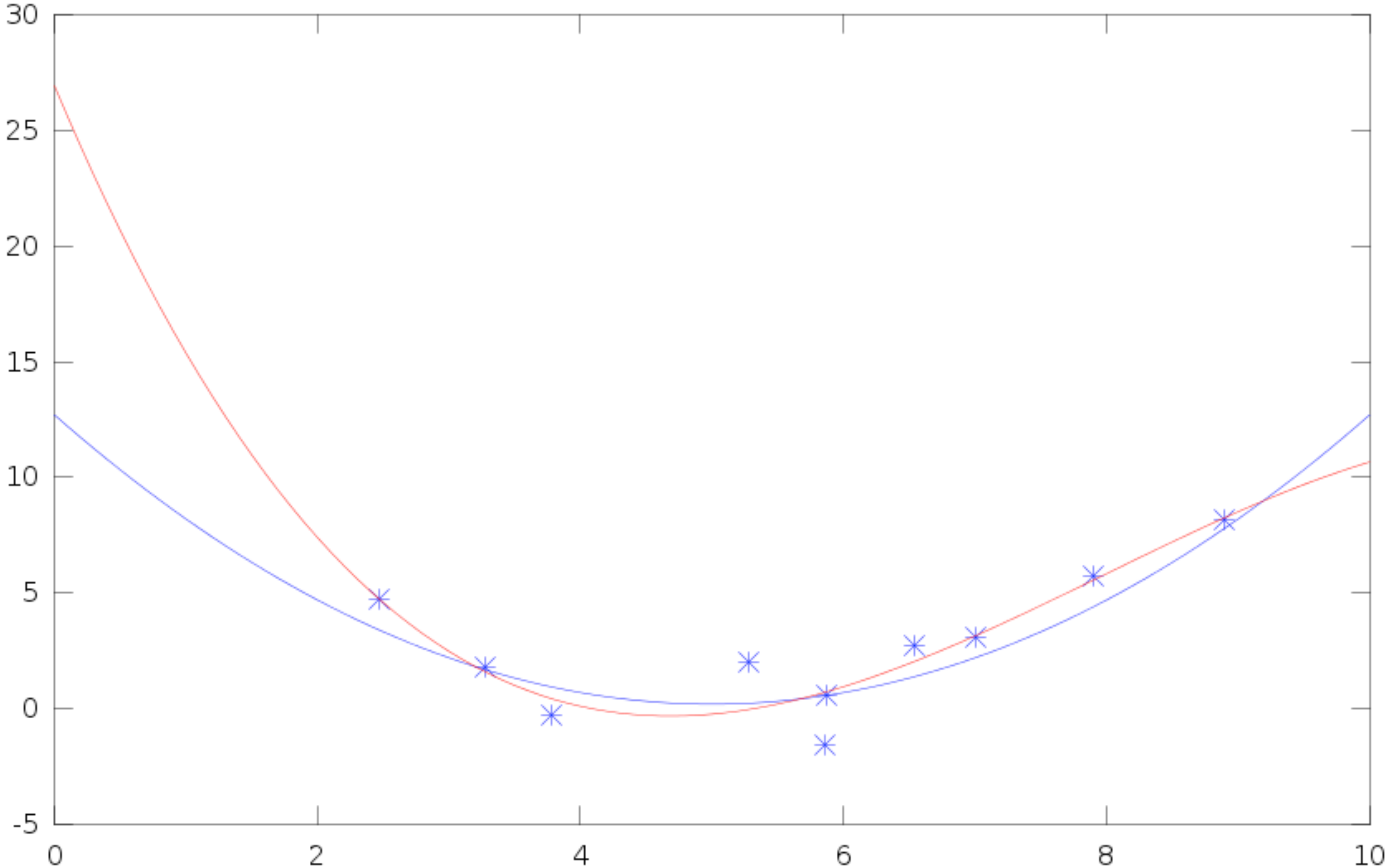
Regression (d=1)



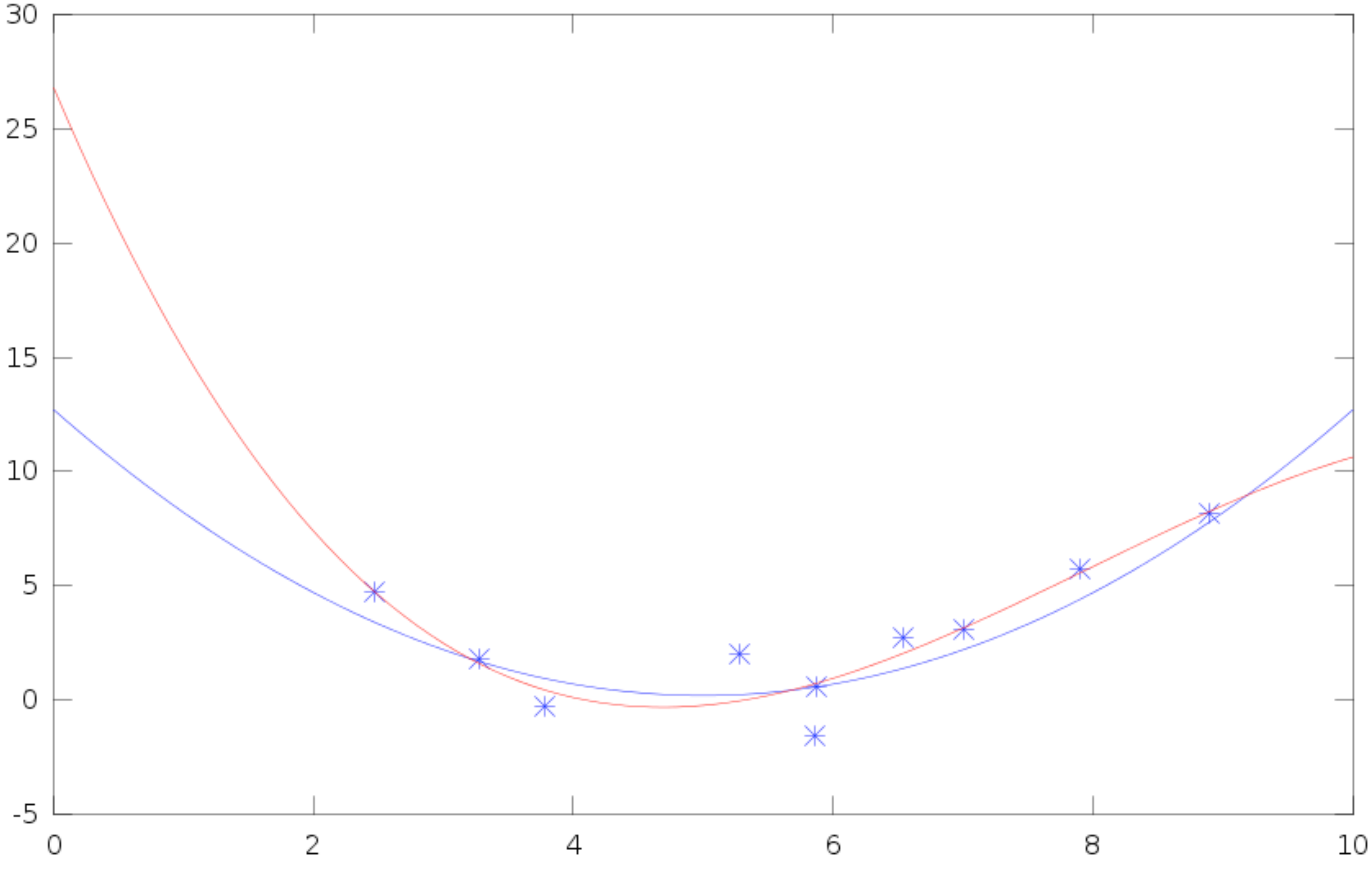
Regression (d=2)



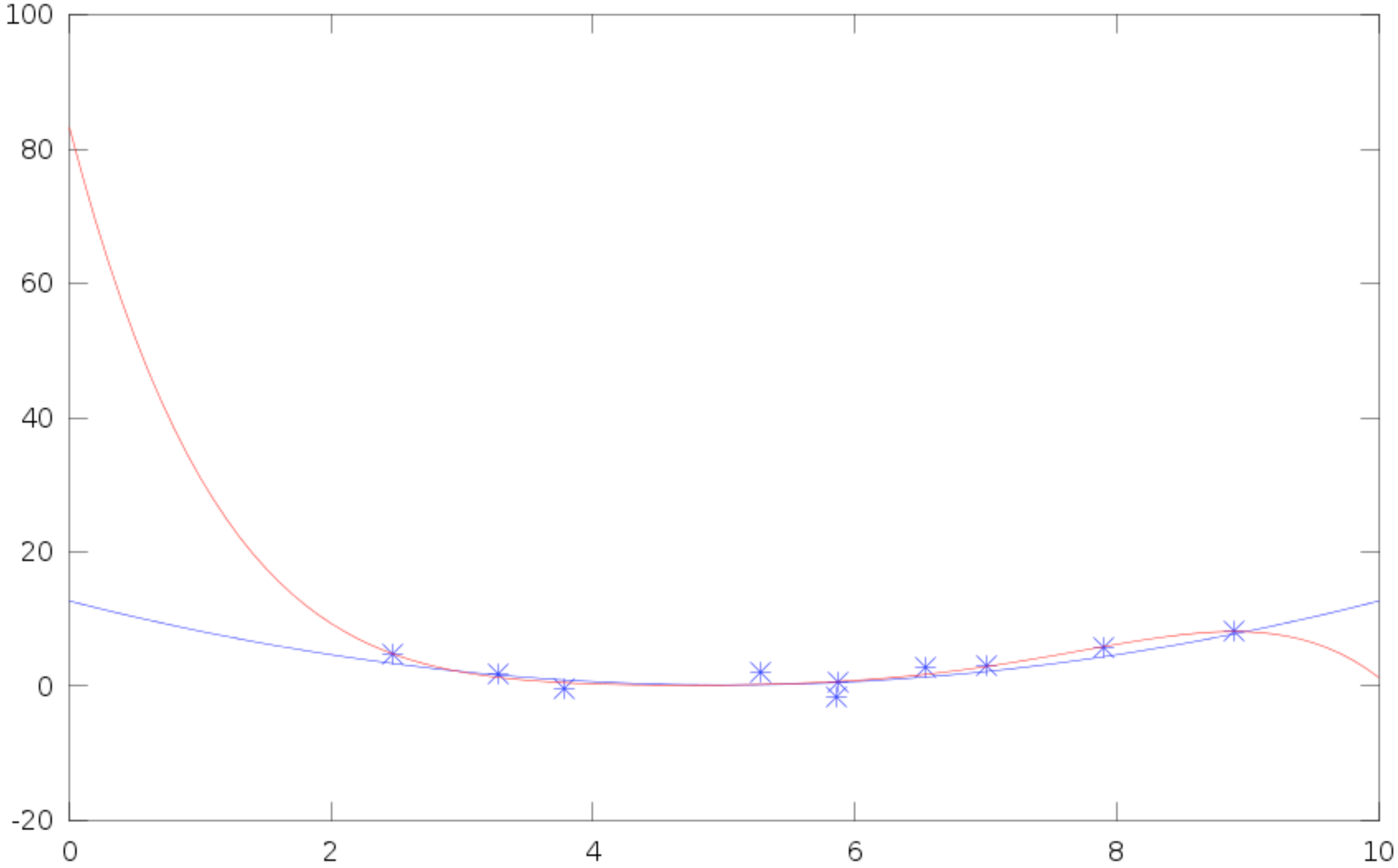
Regression (d=3)



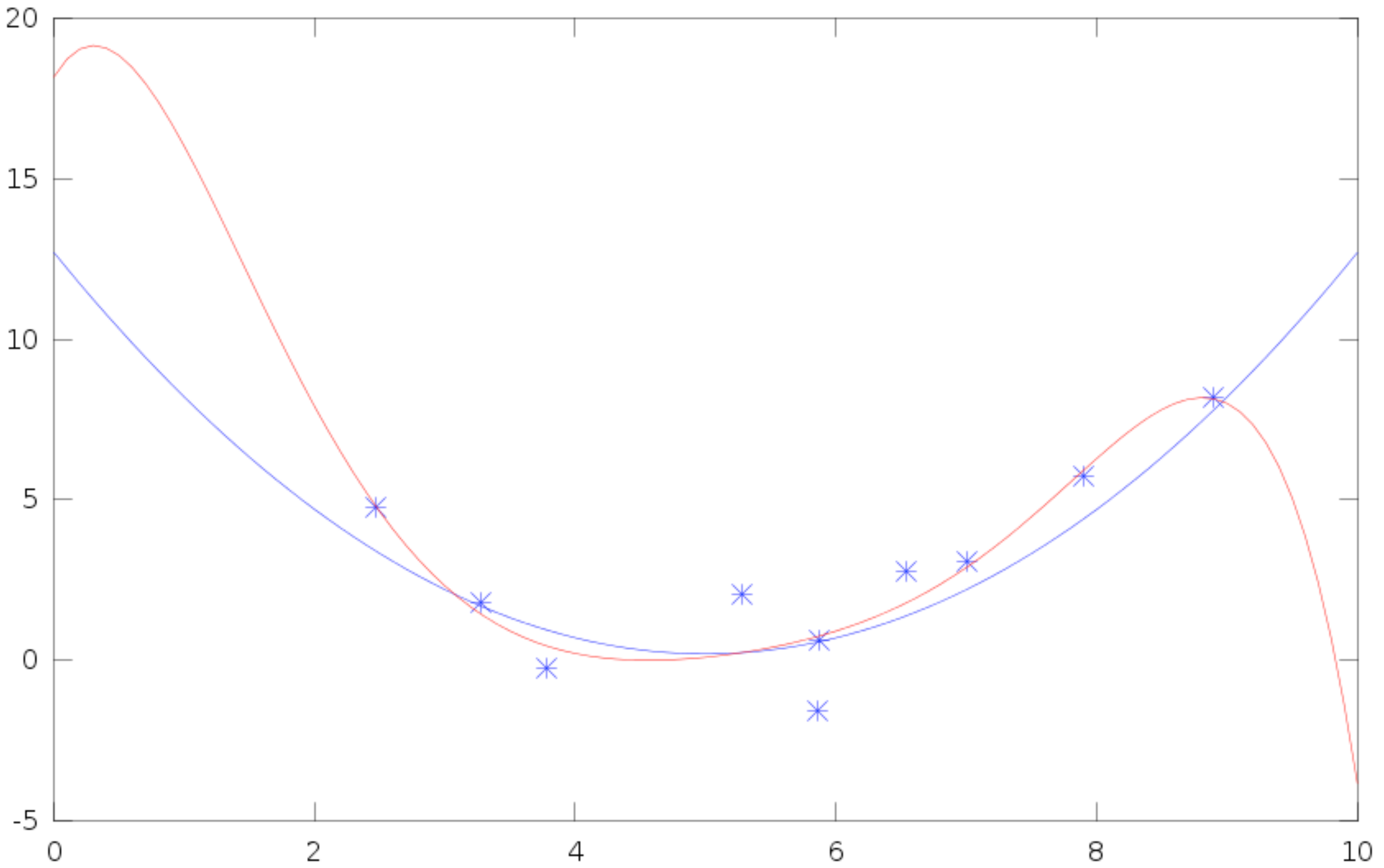
Regression (d=4)



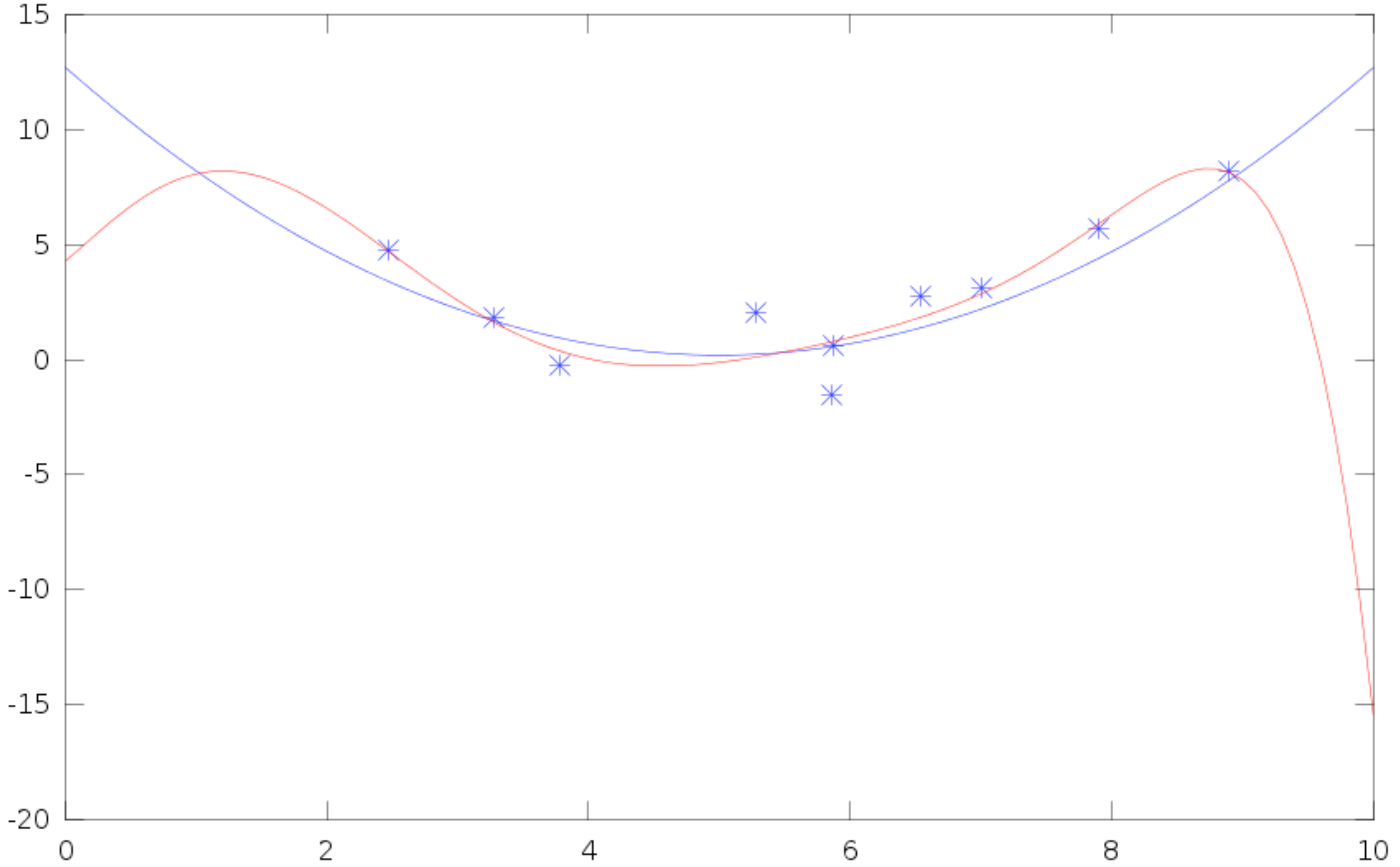
Regression (d=5)



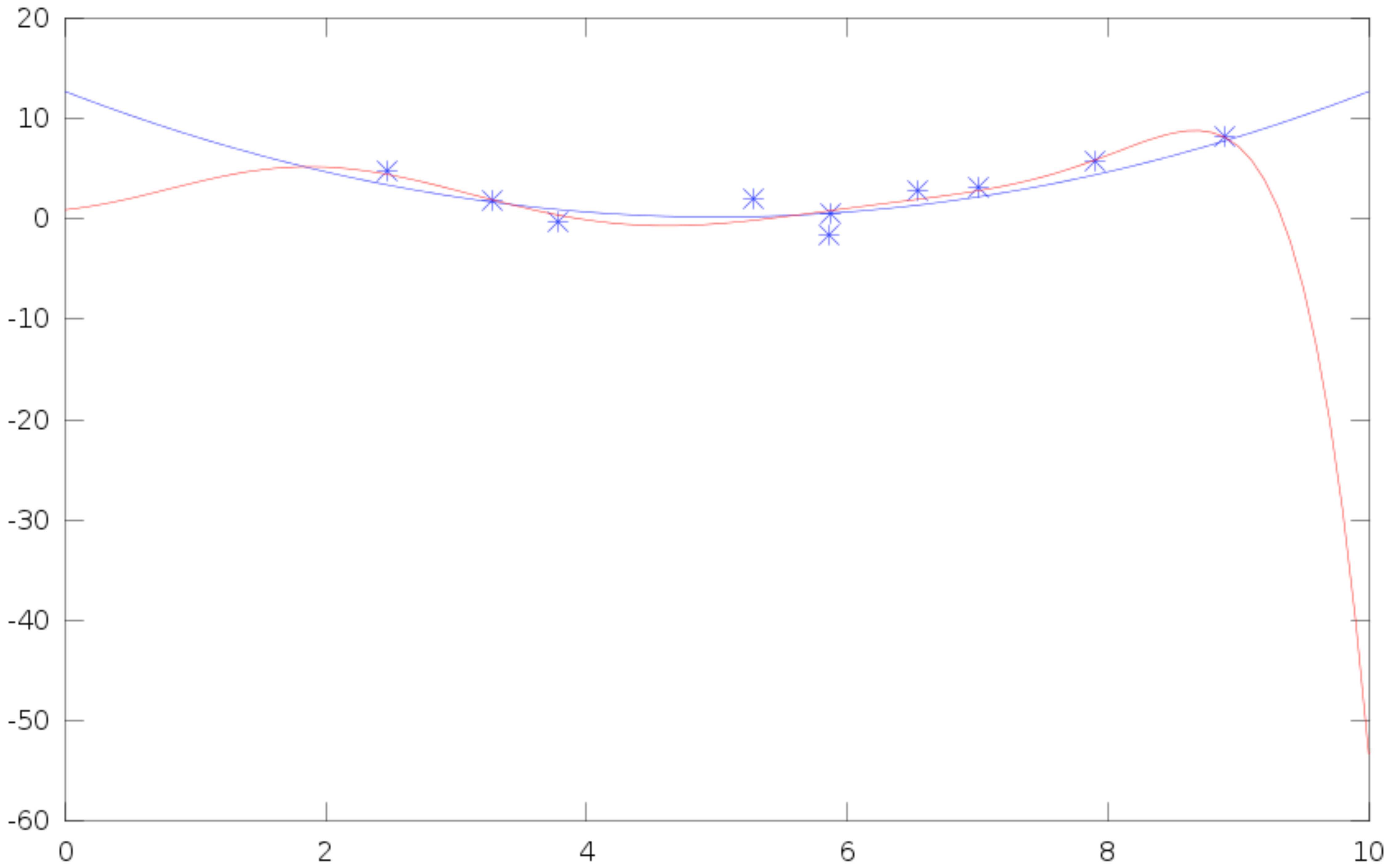
Regression (d=6)



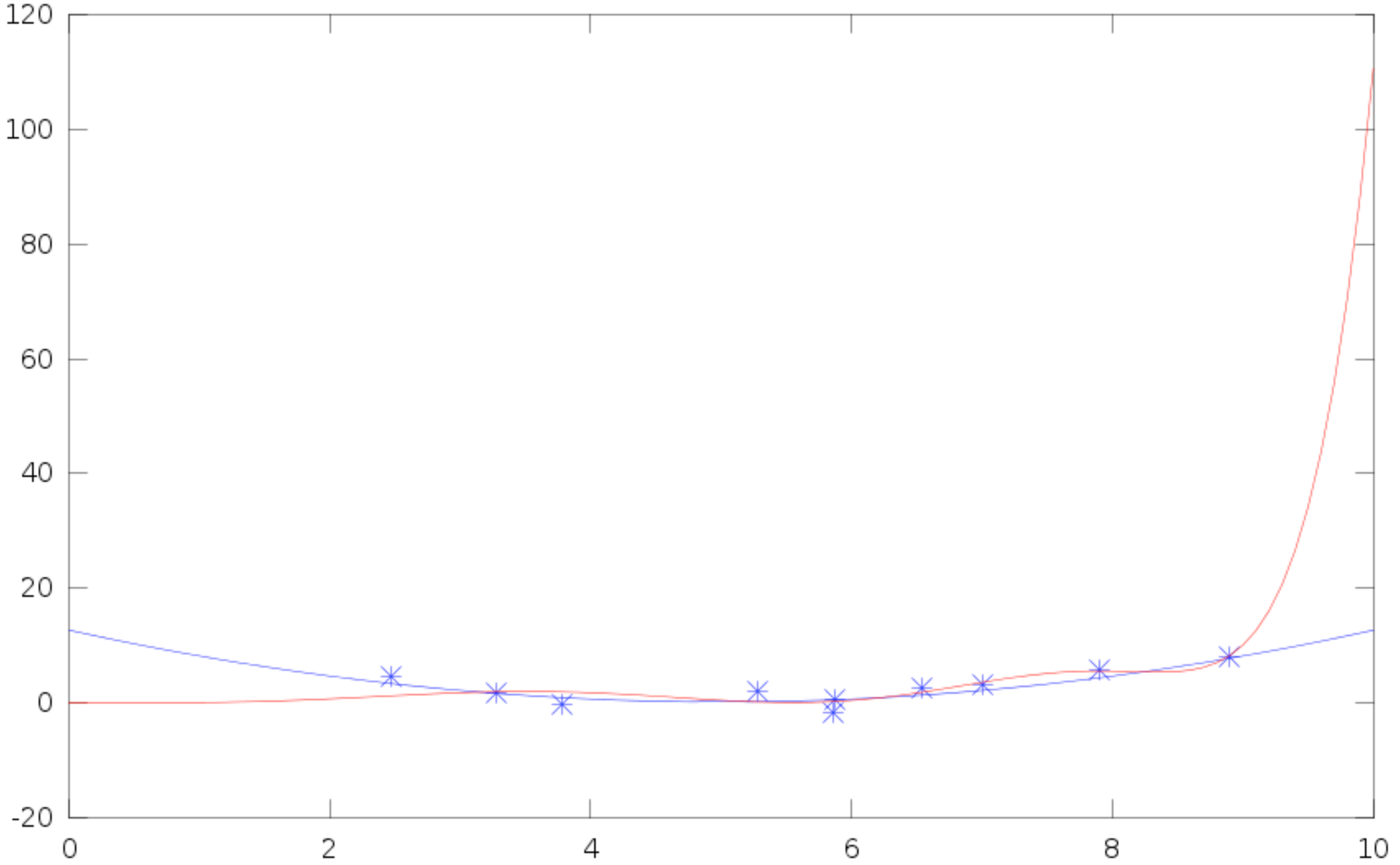
Regression (d=7)



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


MAGIC Etch A Sketch® SCREEN
OHIO ART
Big Data
on the
Internet

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

crawl it

flickr

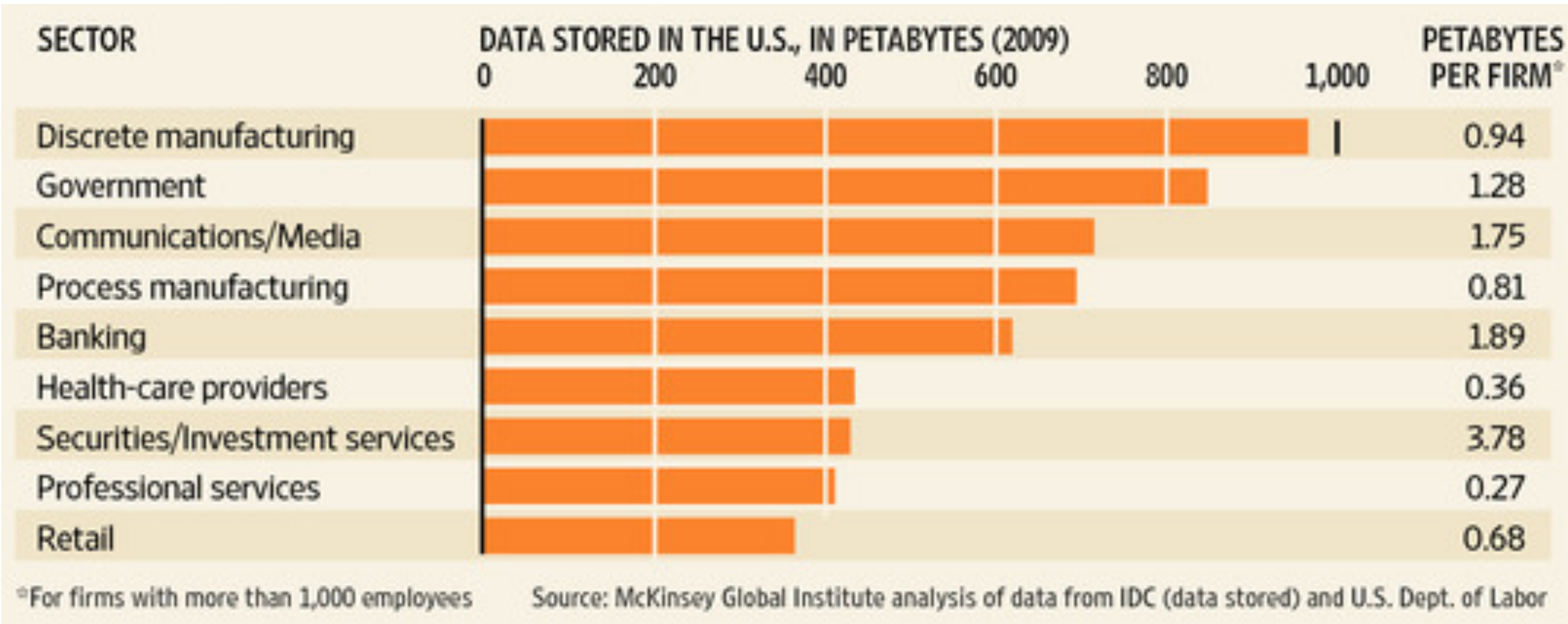
DISQUS

YouTube

yelp

>1B images, 40h video/minute

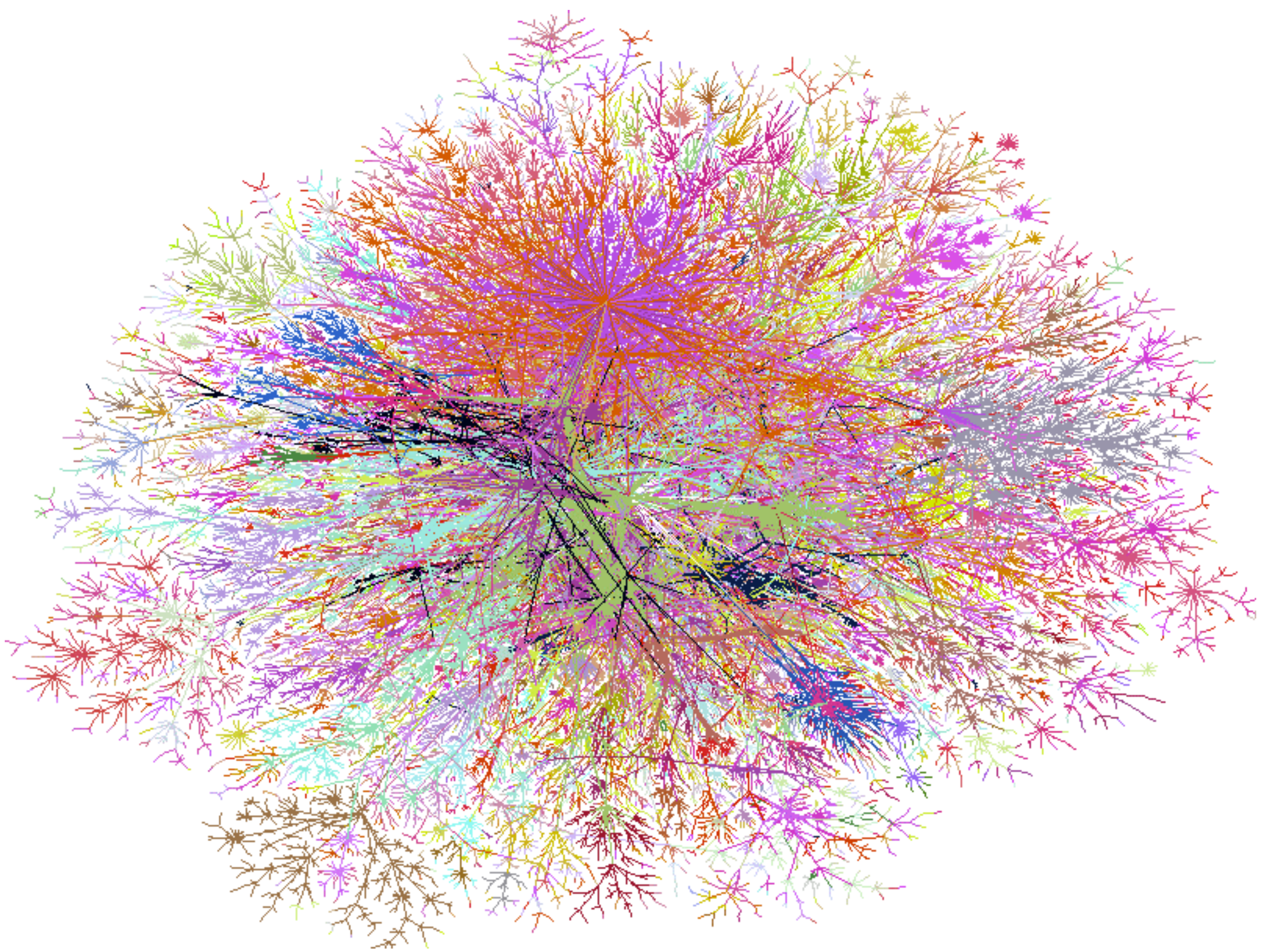
Big Data



we need Big Learning

Data

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



>10B useful webpages

The Web for $100k/month

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

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

Data - Identity & Graph

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

100M-1B vertices

Crawling Twitter for $10k

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

- 300M users
- Per user 300 queries/h
- 100 edges/query
- 100 edges/account
- Need 100 machines for 2 weeks (crawl it at 10 queries/s)
  - Tweets
  - Inlinks
  - Outlinks
- Cost
  - $3k for computers on EC2
  - Similar for network & storage
  - Need 10k user keys

Data - Messages

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

Windows Live Hotmail

YAHOO! Mail

AOL

meebo

>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)

Windows Live Hotmail

YAHOO! Mail

AOL

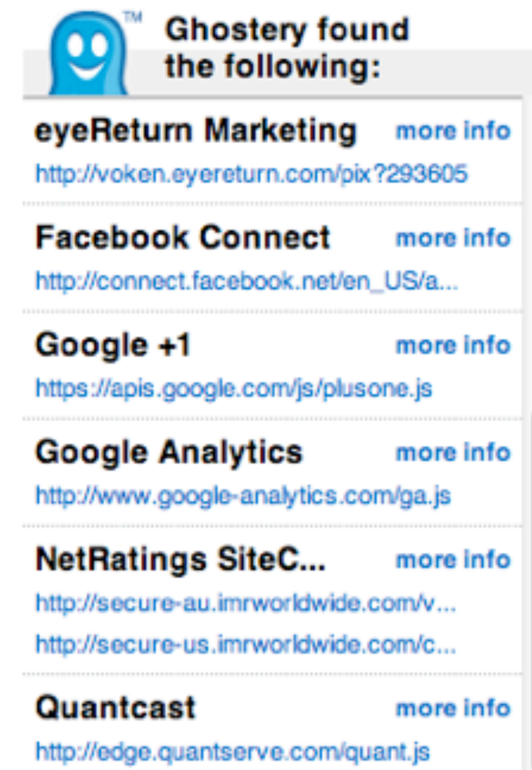
meebo

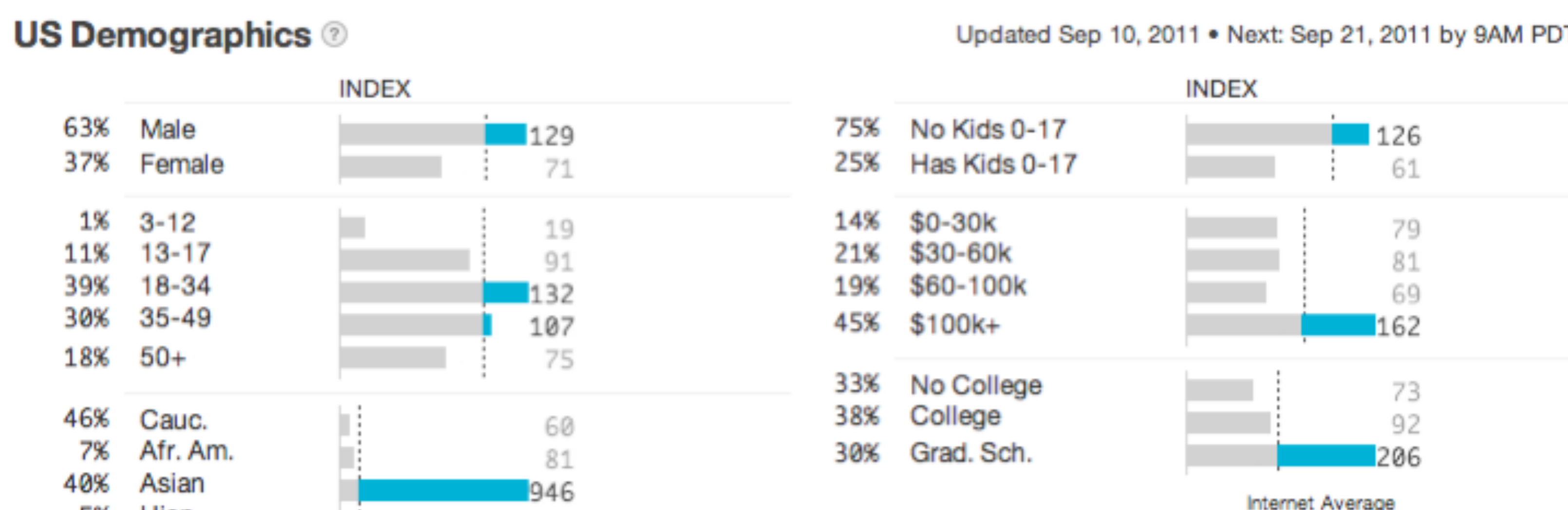
>1B texts

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.

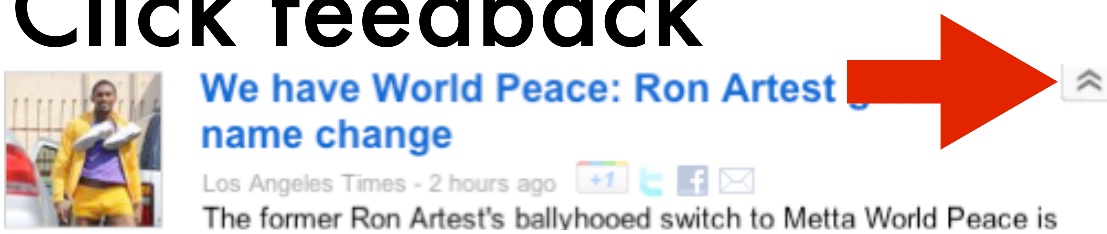
Data Retention:

Undisclosed

(implicit) Labels   no Labels

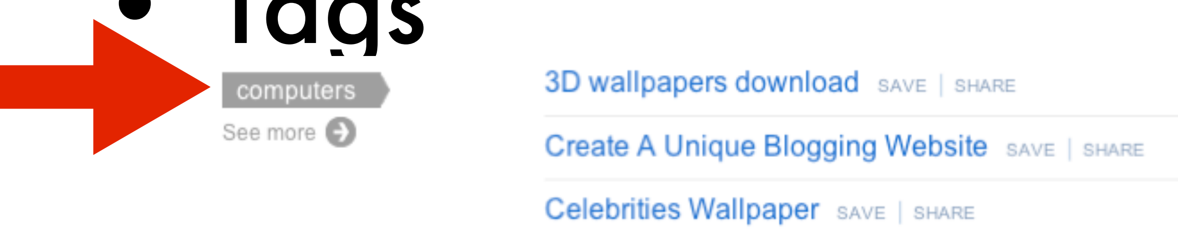
- Ads

- Click feedback



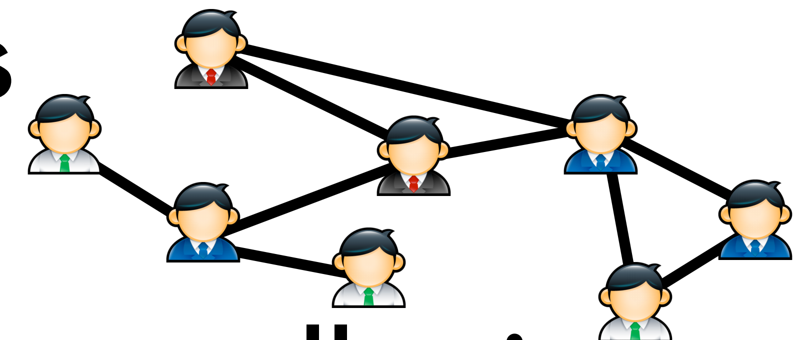
- Emails

- Tags



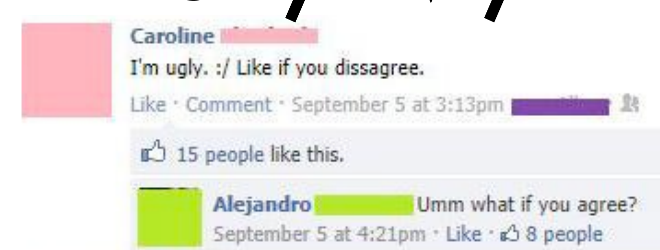
- Editorial data is very expensive! Do not use!

- Graphs



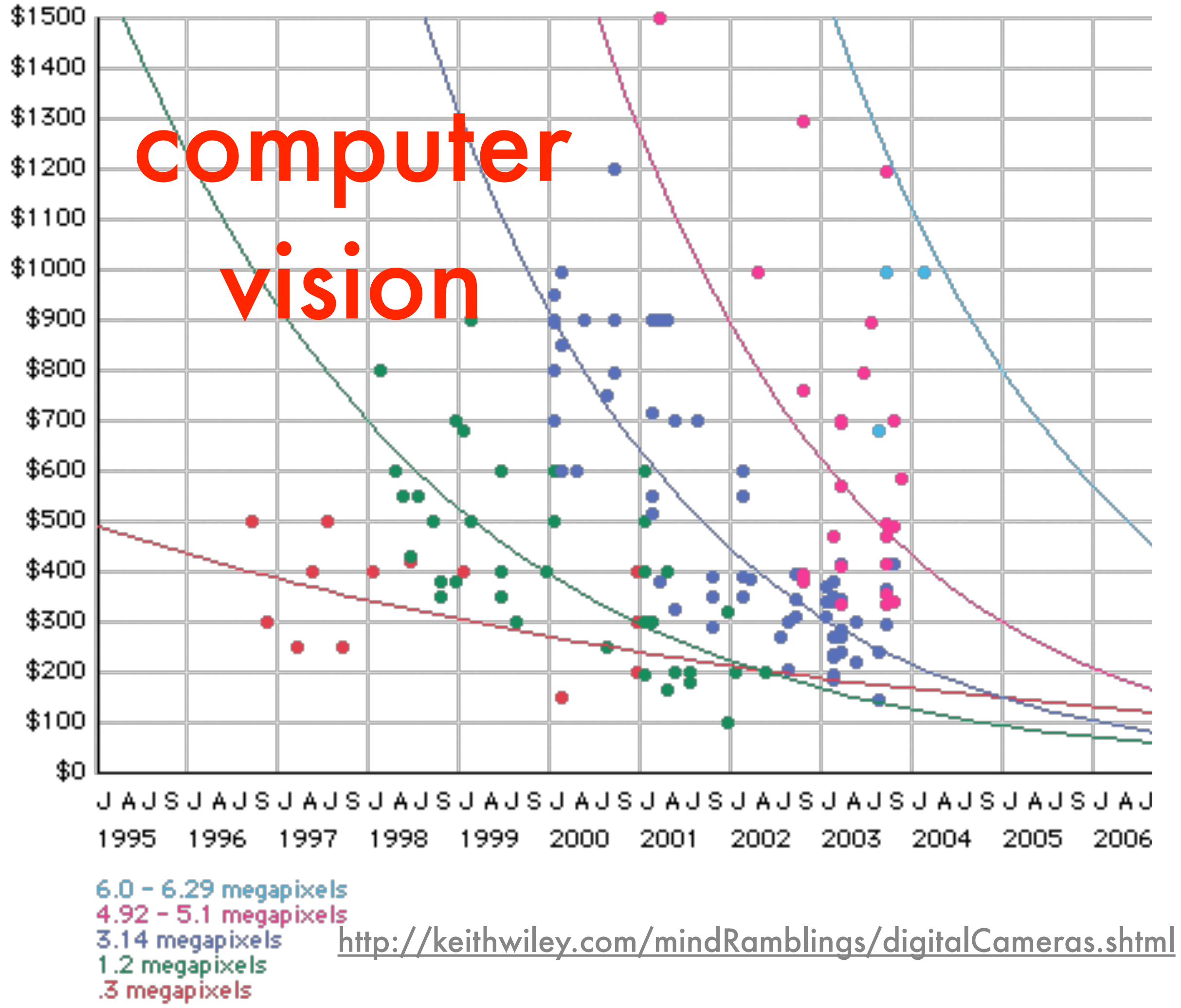
- Document collections

- Email/IM/Discussions

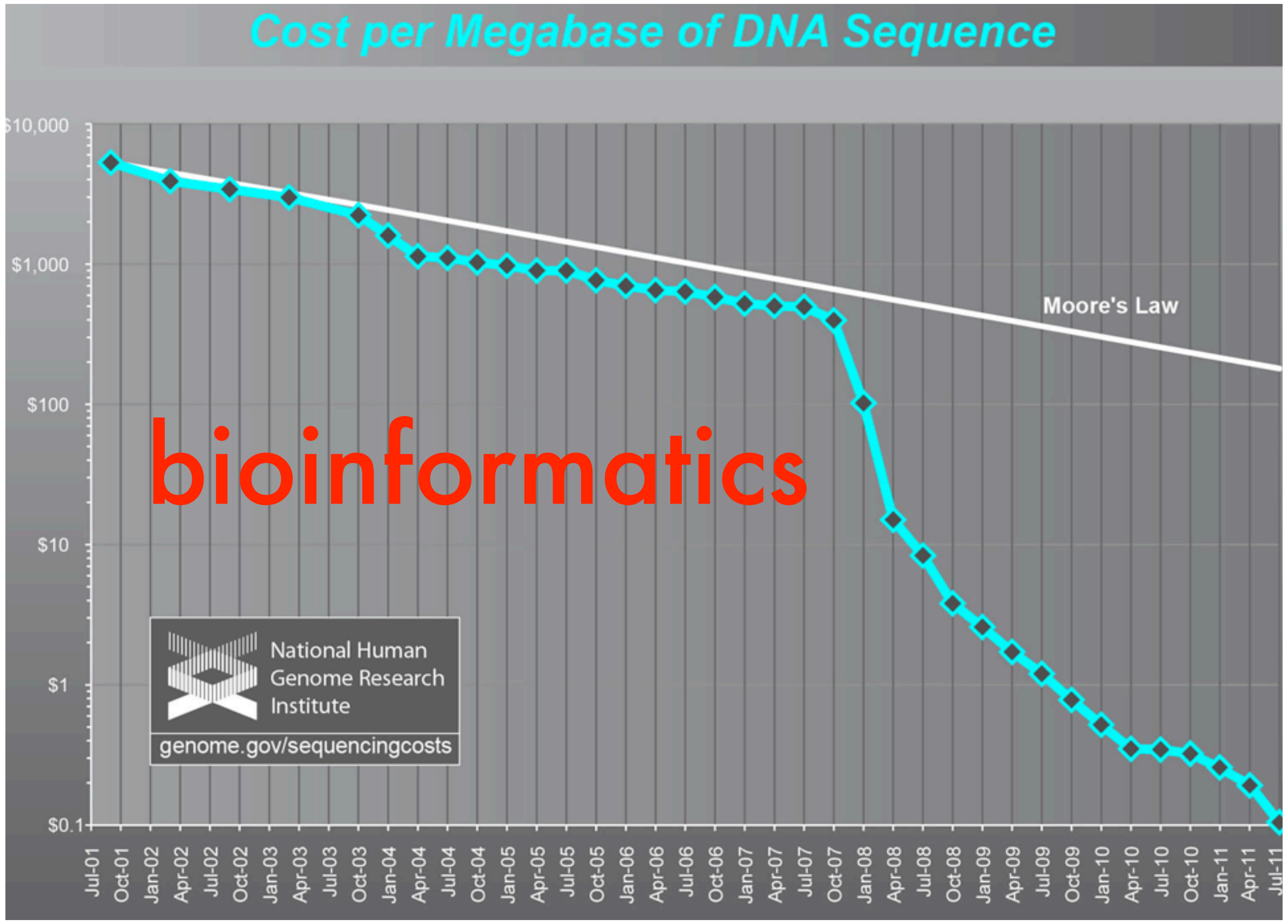


- Query stream

Many more sources



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

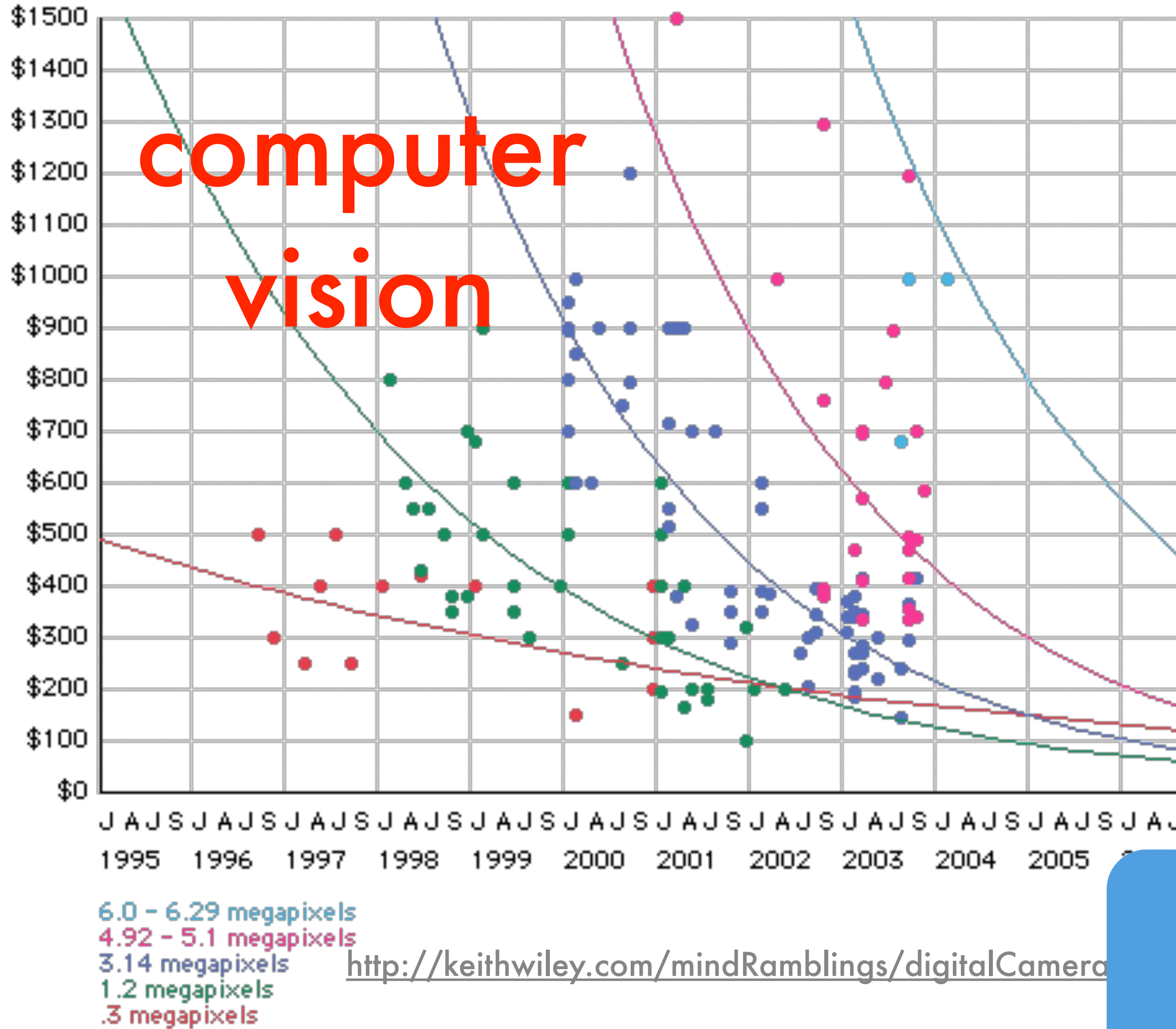


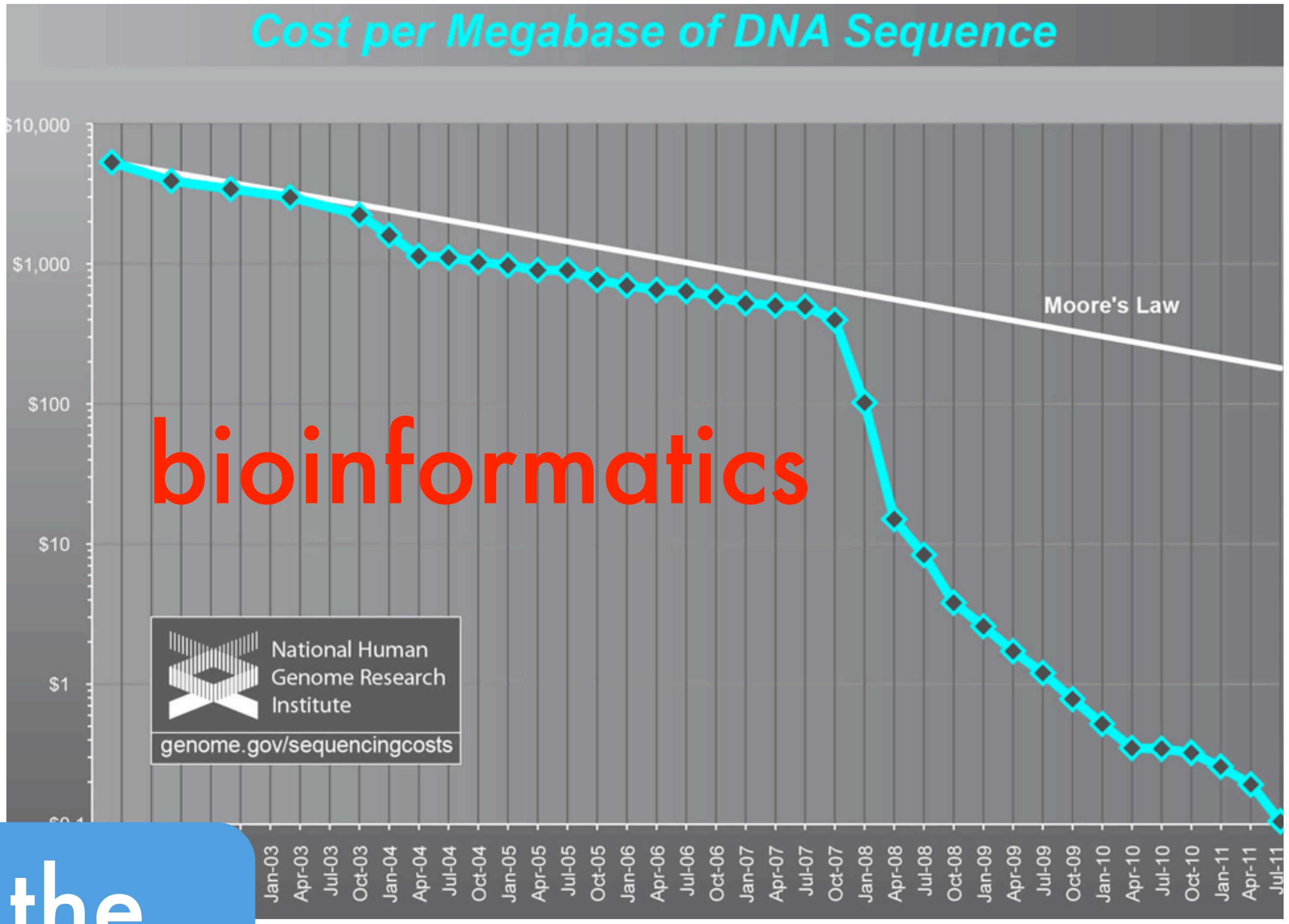
personalized sensors

ubiquitous control

University

Many more sources





in the cloud

Further material

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