MACHINE LEARNING
DEPARTMENT

Introduction to Machine Learning

1. Overview

Alex Smola & Geoff Gordon
Carnegie Mellon University

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

MAGIC Etch A Sketch® SCREEN

Administrative Stuff

Important Stuff

- Lectures Monday and Wednesday 10:30-11:50am, Wean Hall 7500
- Recitation Tuesday 5-6:30pm, Wean Hall 7500
- Office hours Monday 1-3pm (Alex), Wednesday (Geoff)
- Grading policy
  - Project (34%) Mid project report due after midterm
  - Exams: Midterm (33%) Exam is without technology
  - Homework (33%) Best (n-1) out of n.
    To receive points you must submit on due date. No exceptions.
- Google Group https://groups.google.com/forum/#!forum/10-701-fall-2013 (questions, discussions, announcements)
- Homepage http://alex.smola.org/teaching/cmu2013-10-701x/ (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
- Pitch the project to Geoff or me before you decide

Color Coding

- Really important stuff
- Important stuff
- Regular stuff

Feedback please

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

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

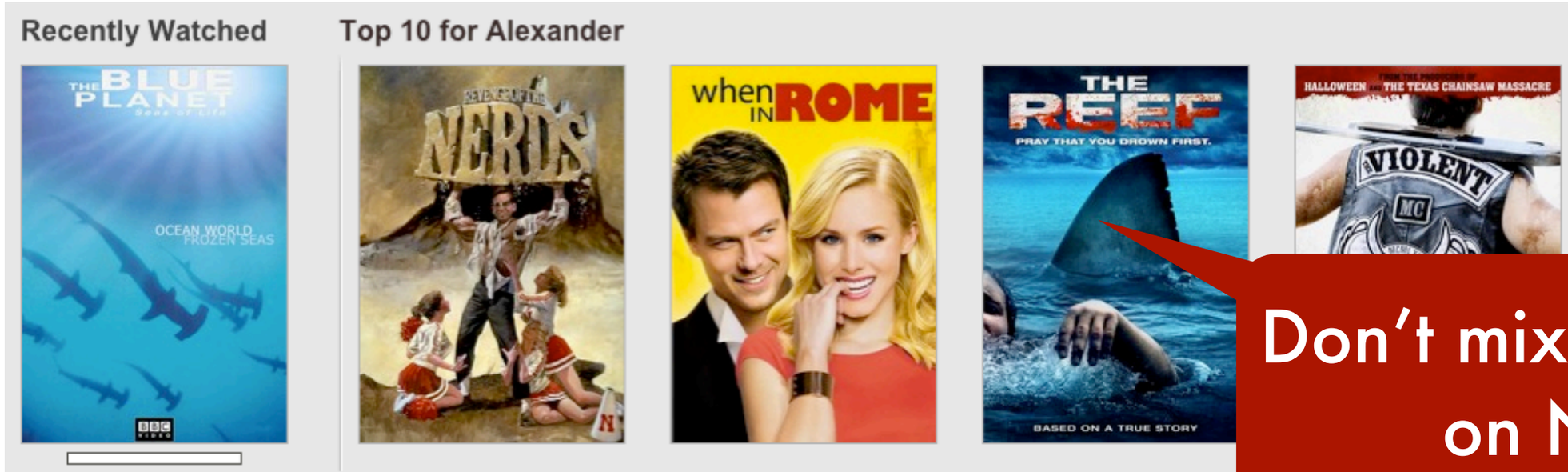
for Wall Street

biology

energy

Carnegie Mellon University

Collaborative Filtering



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

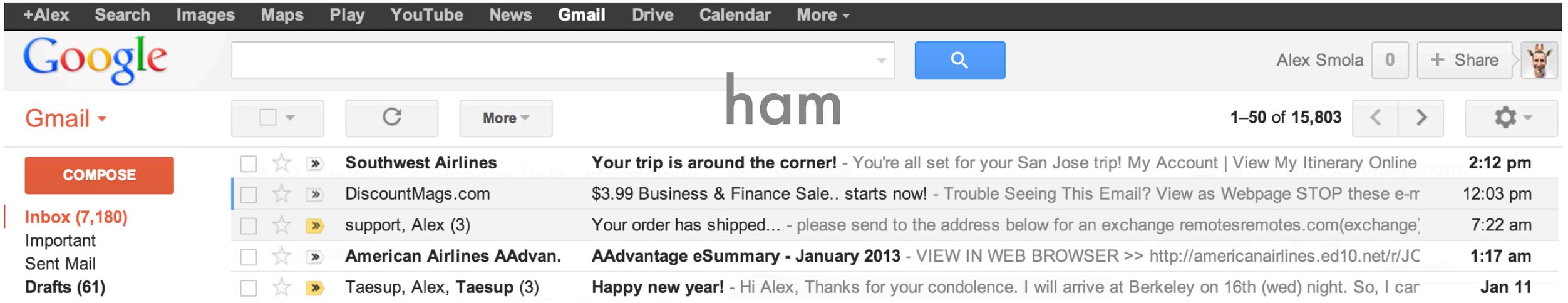
Avatar learns from your behavior

Black & White
Lionsgate Studios

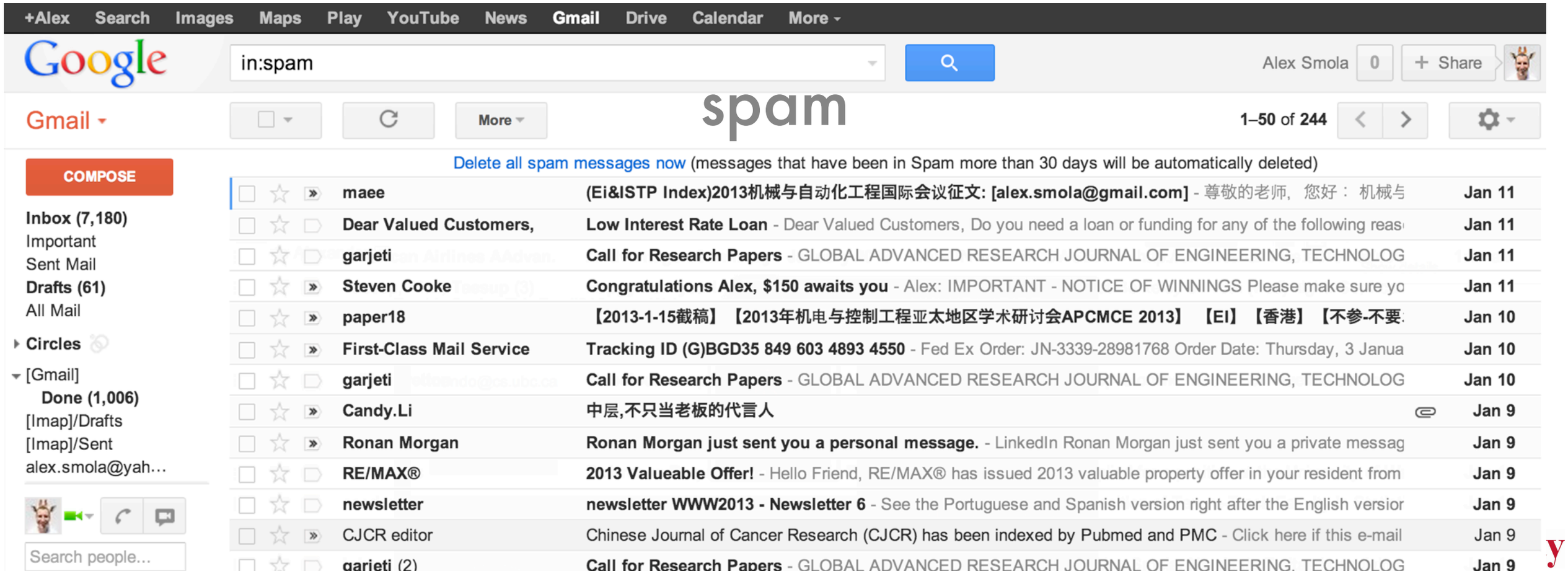
Imitation Learning

Carnegie Mellon University

Spam Filtering



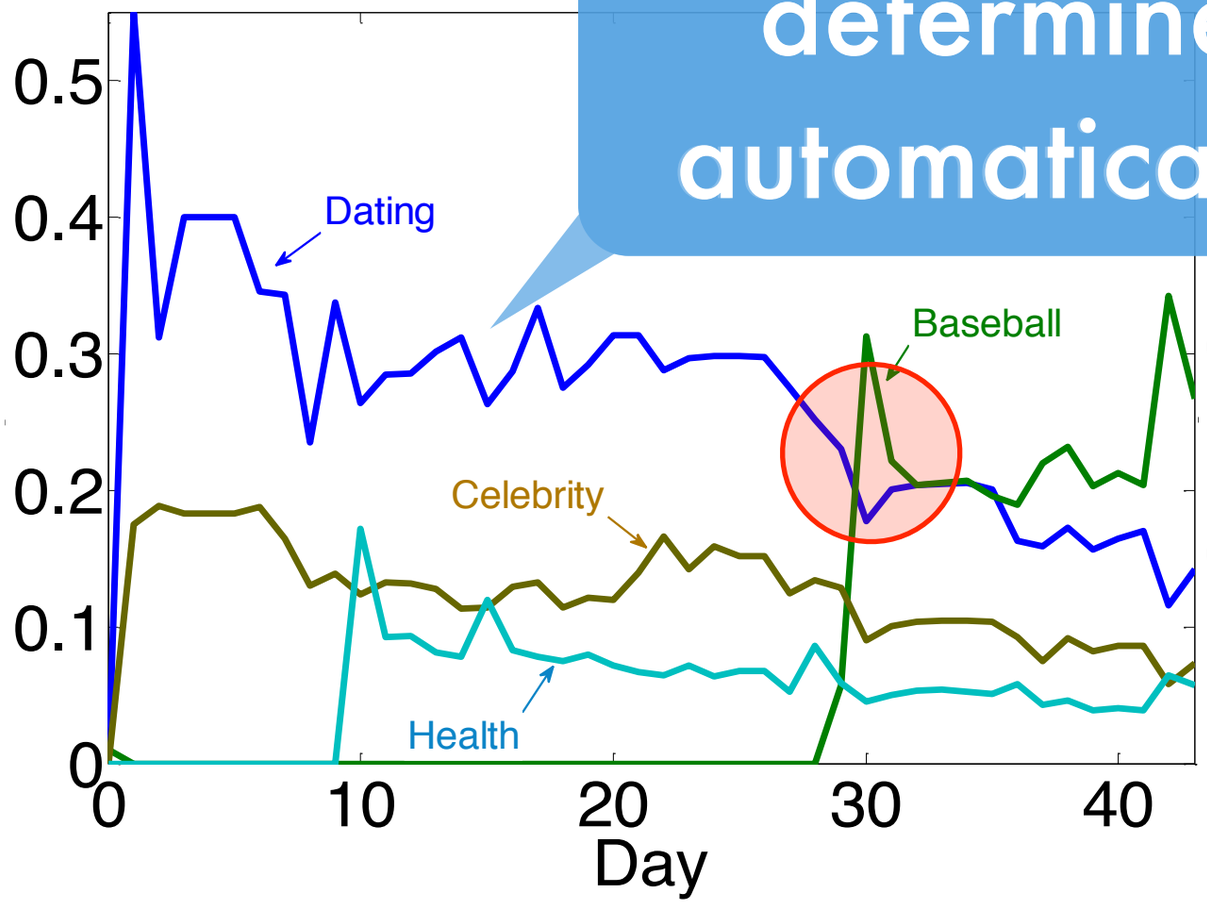
ham

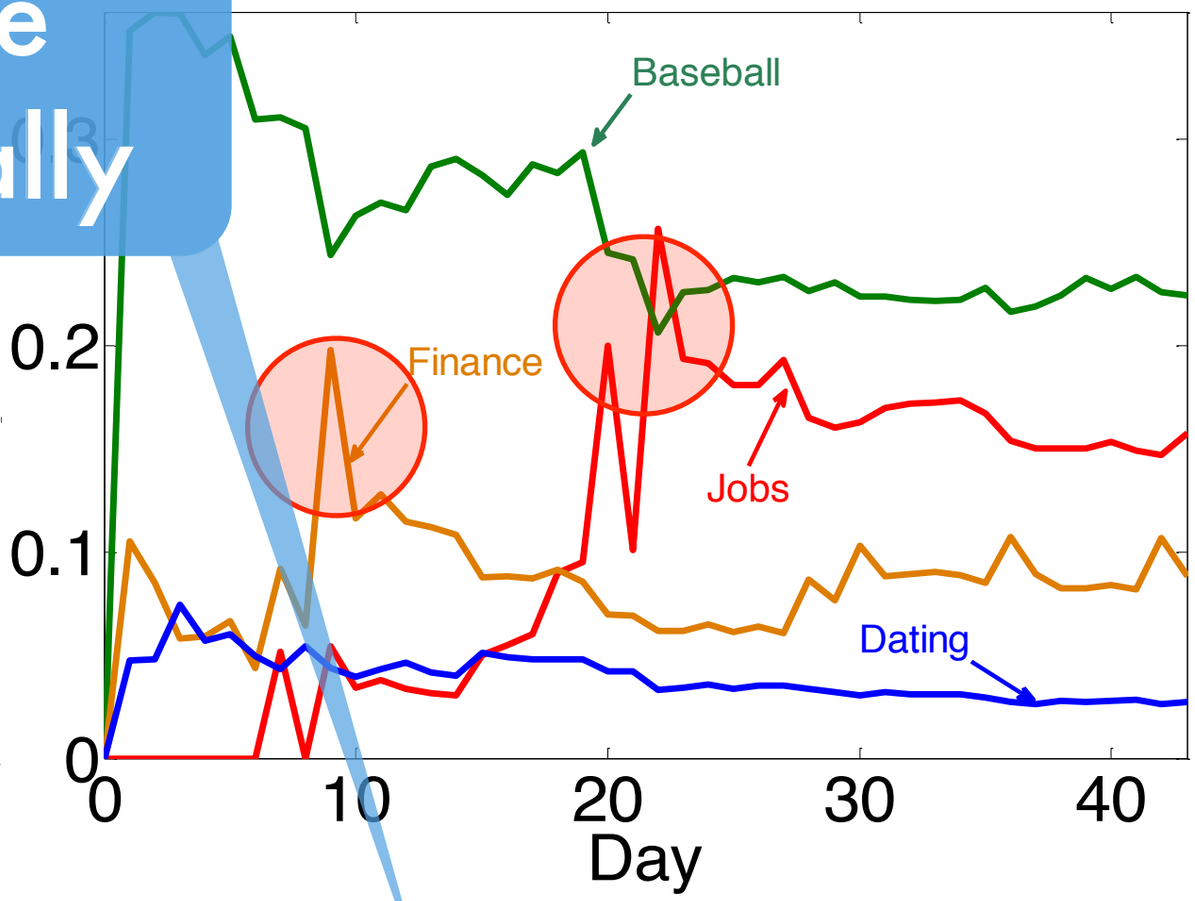


spam

User profiling

determine automatically

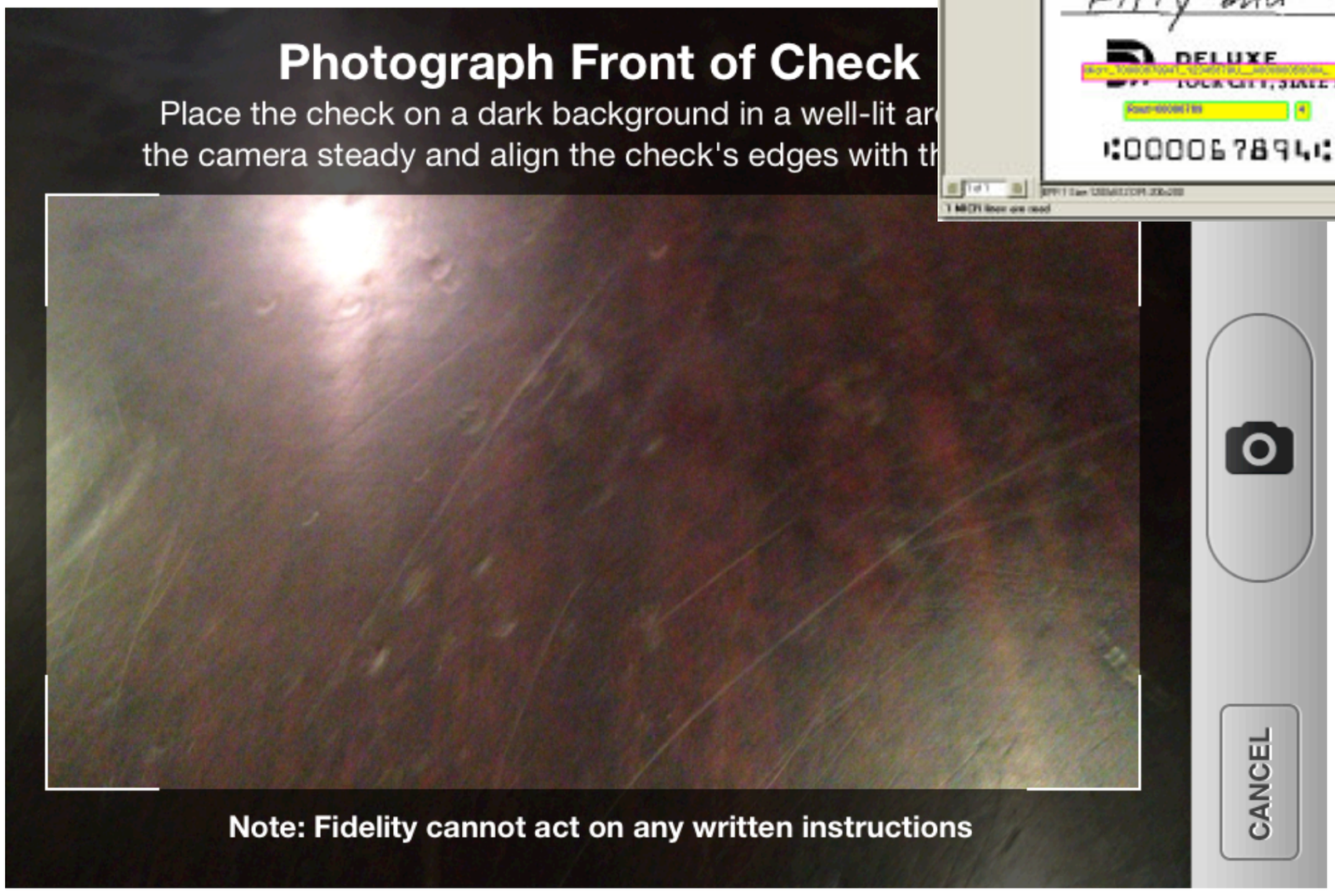


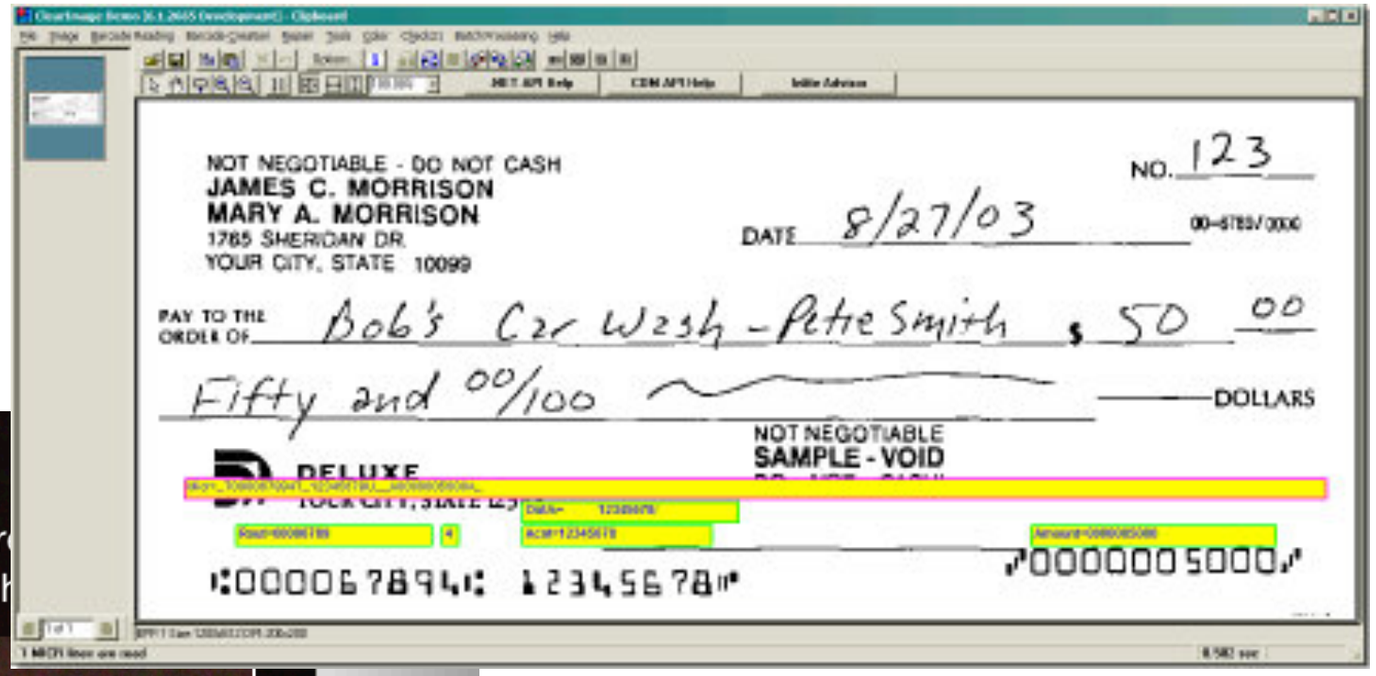


Dating	Baseball	Celebrity	Health	Jobs	Finance
women	League	Snooki	skin	job	financial
men	baseball	Tom	body	career	Thomson
dating	basketball,	Cruise	fingers	business	chart
singles	doublehead	Katie	cells	assistant	real
personals	Bergesen	Holmes	toes	hiring	Stock
seeking	Griffey	Pinkett	wrinkle	part-time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

Cheque reading

segment image





recognize handwriting

Carnegie Mellon University

Autonomous Helicopter

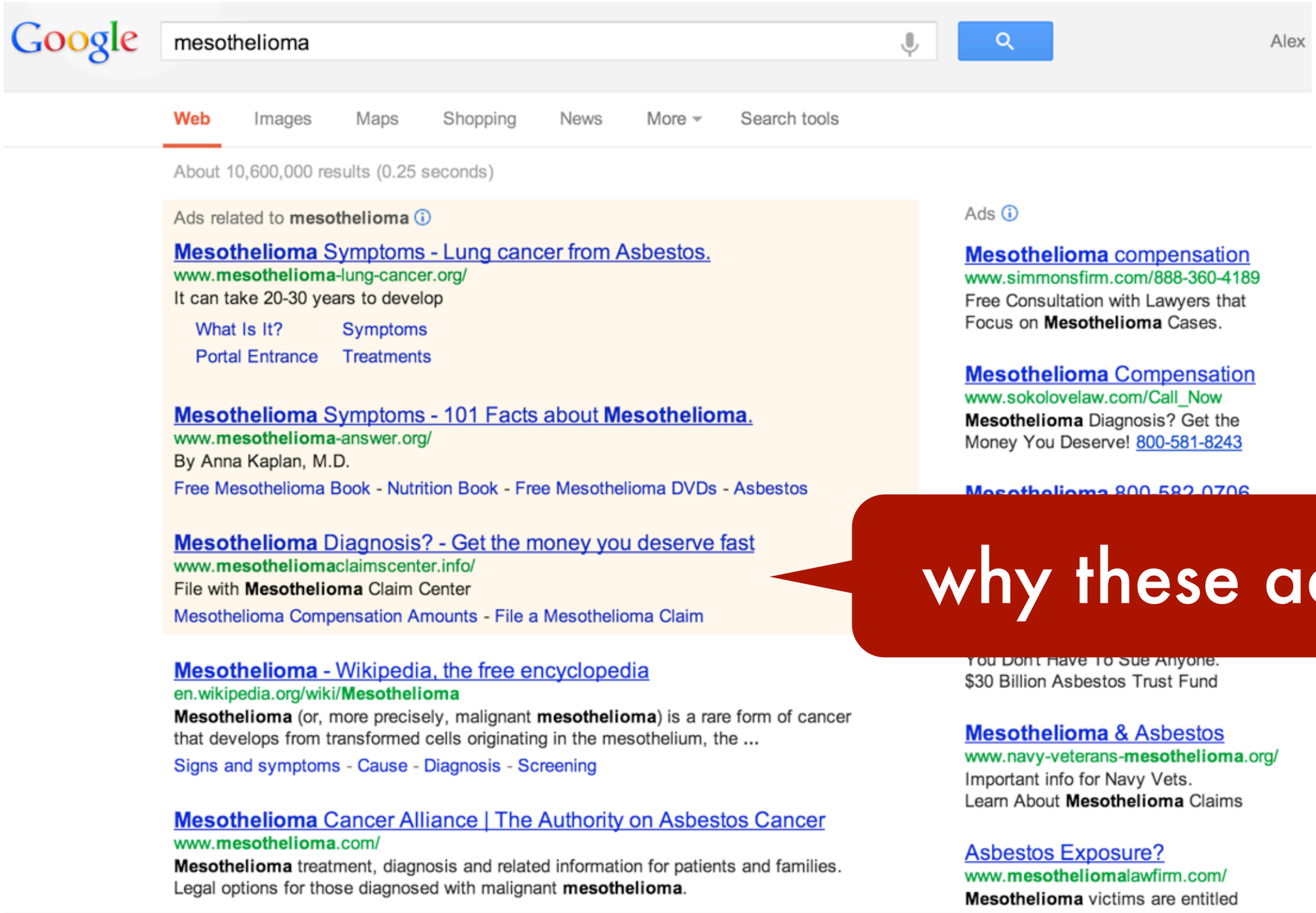
http://heli.stanford.edu

Carnegie Mellon University

Image Layout

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

Search ads



why these ads?

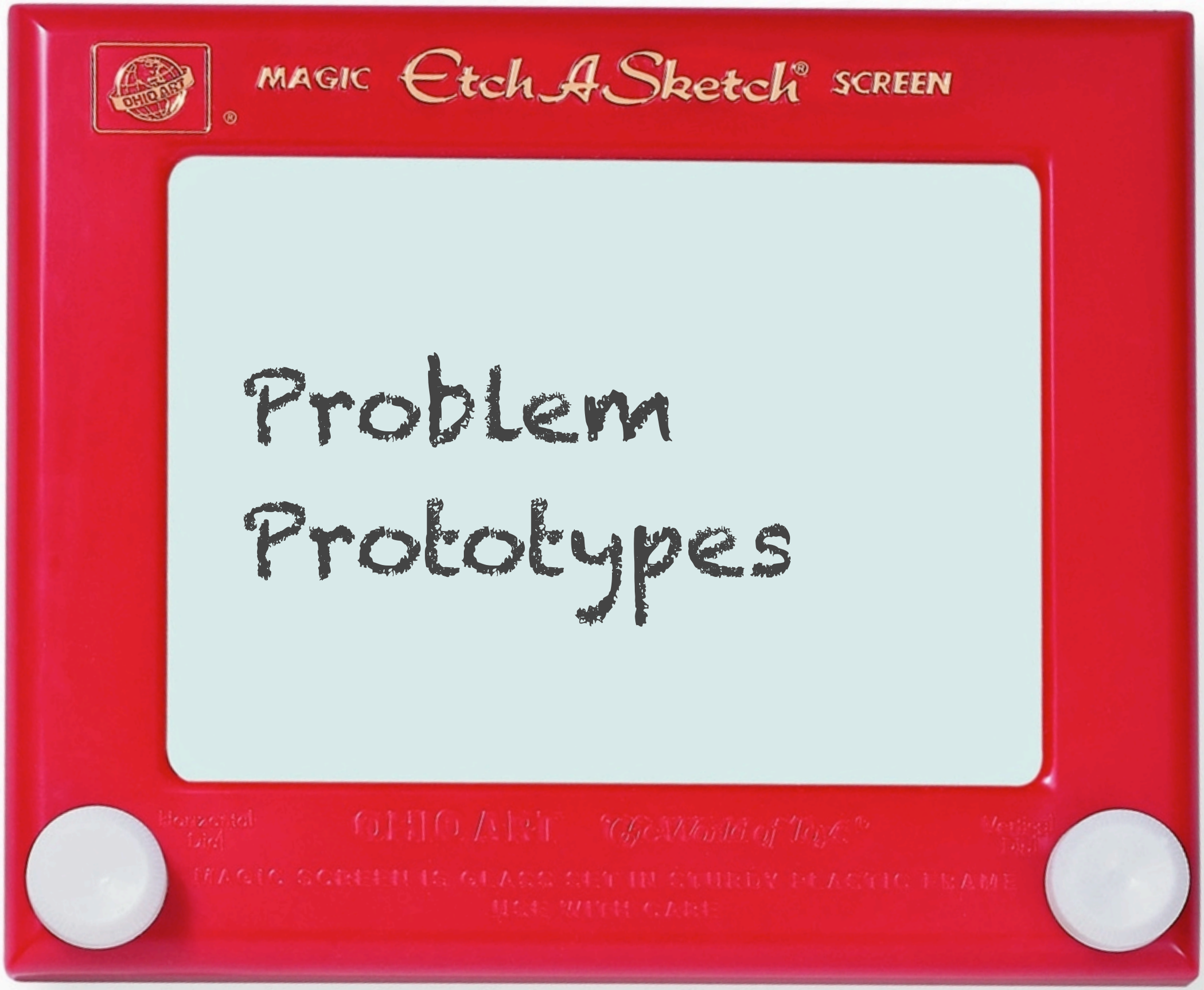
True startup story

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

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

Programming with Data

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


MAGIC Etch A Sketch® SCREEN
Problem
Prototypes

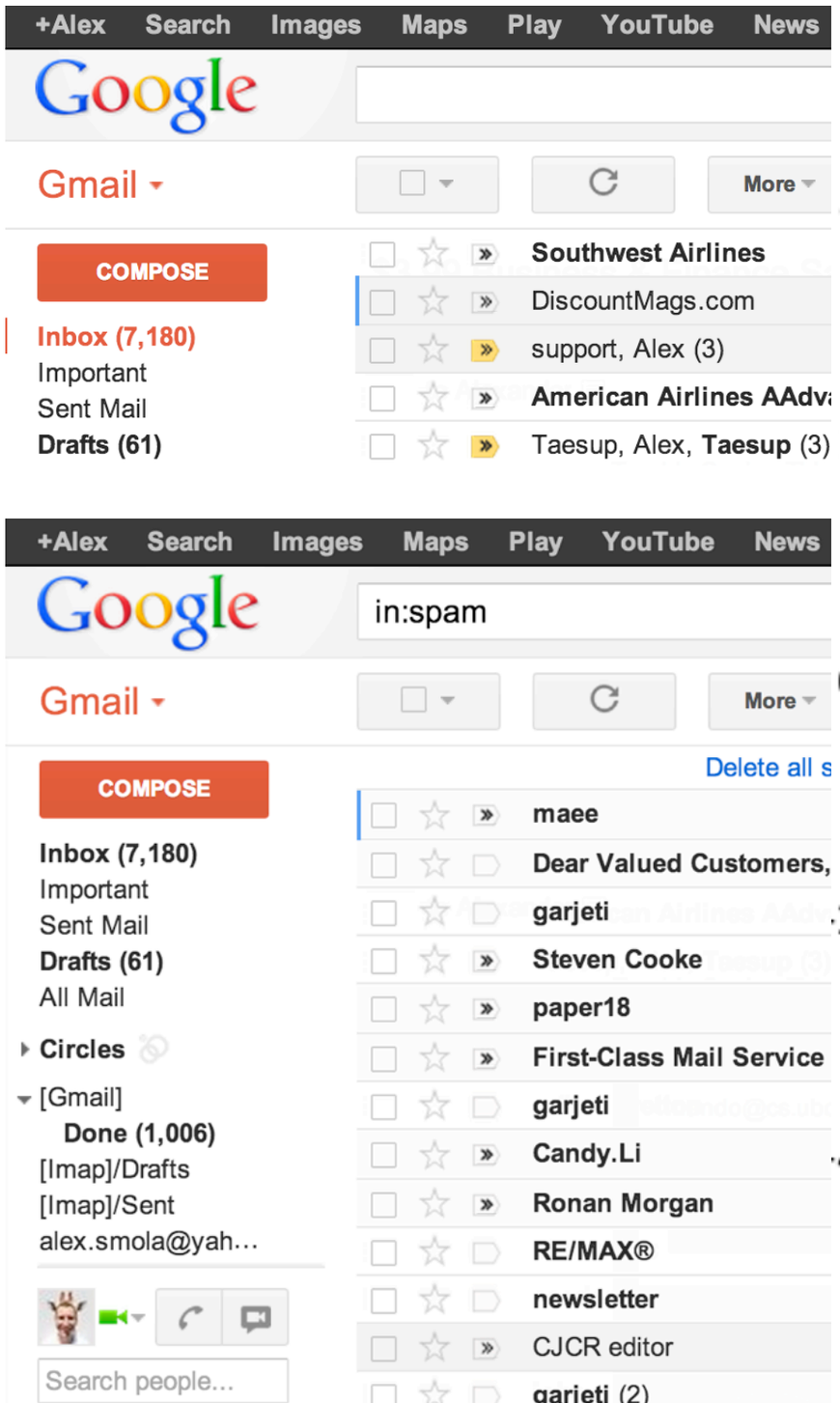
Supervised Learning $y = f(x)$

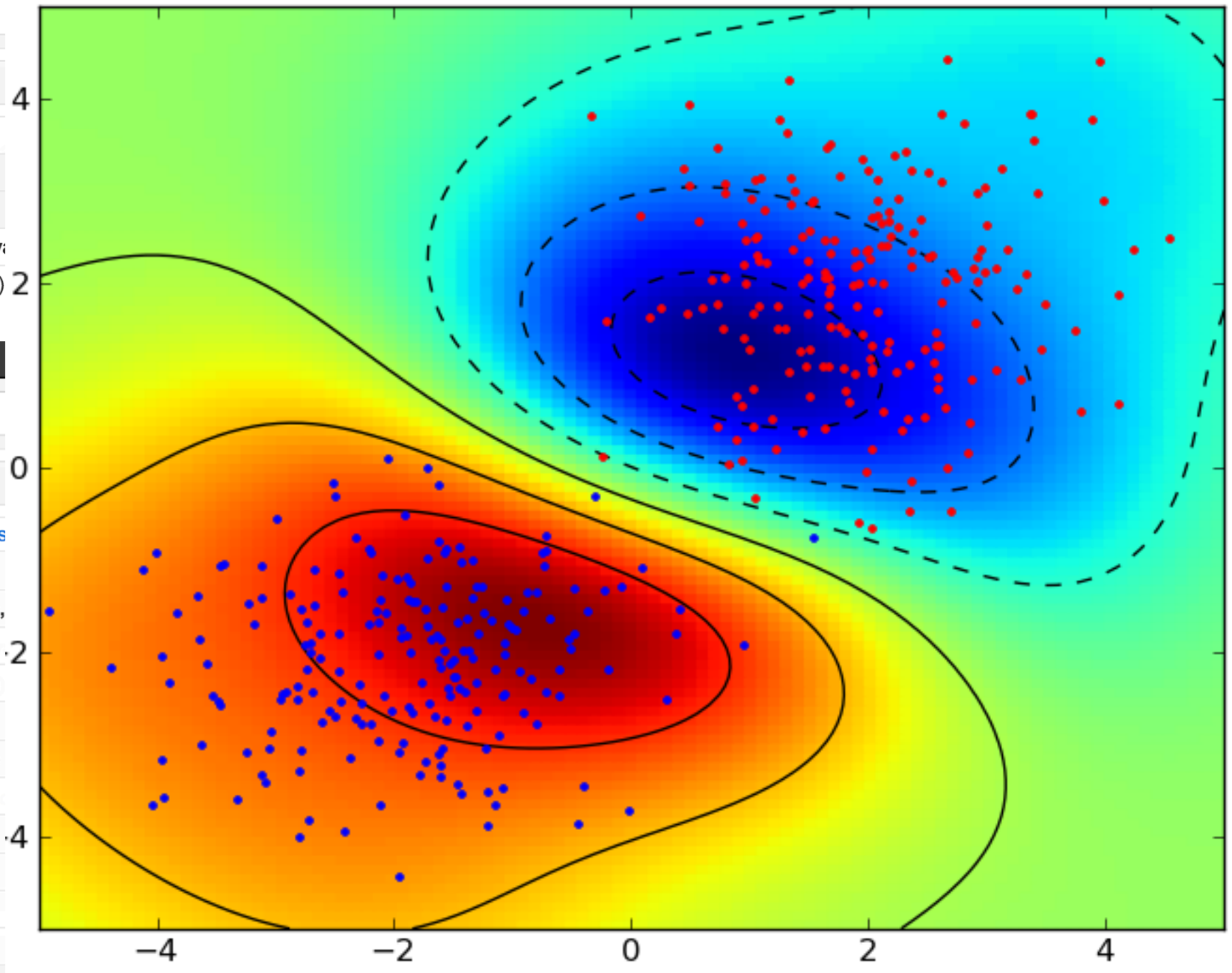
- Binary classification
  Given x find y in {-1, 1}
- Multicategory classification
  Given x find y in {1, … k}
- Regression
  Given x find y in R (or R^d)
- Sequence annotation
  Given sequence x_1 … x_l find y_1 … y_l
- Hierarchical Categorization (Ontology)
  Given x find a point in the hierarchy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

often with loss

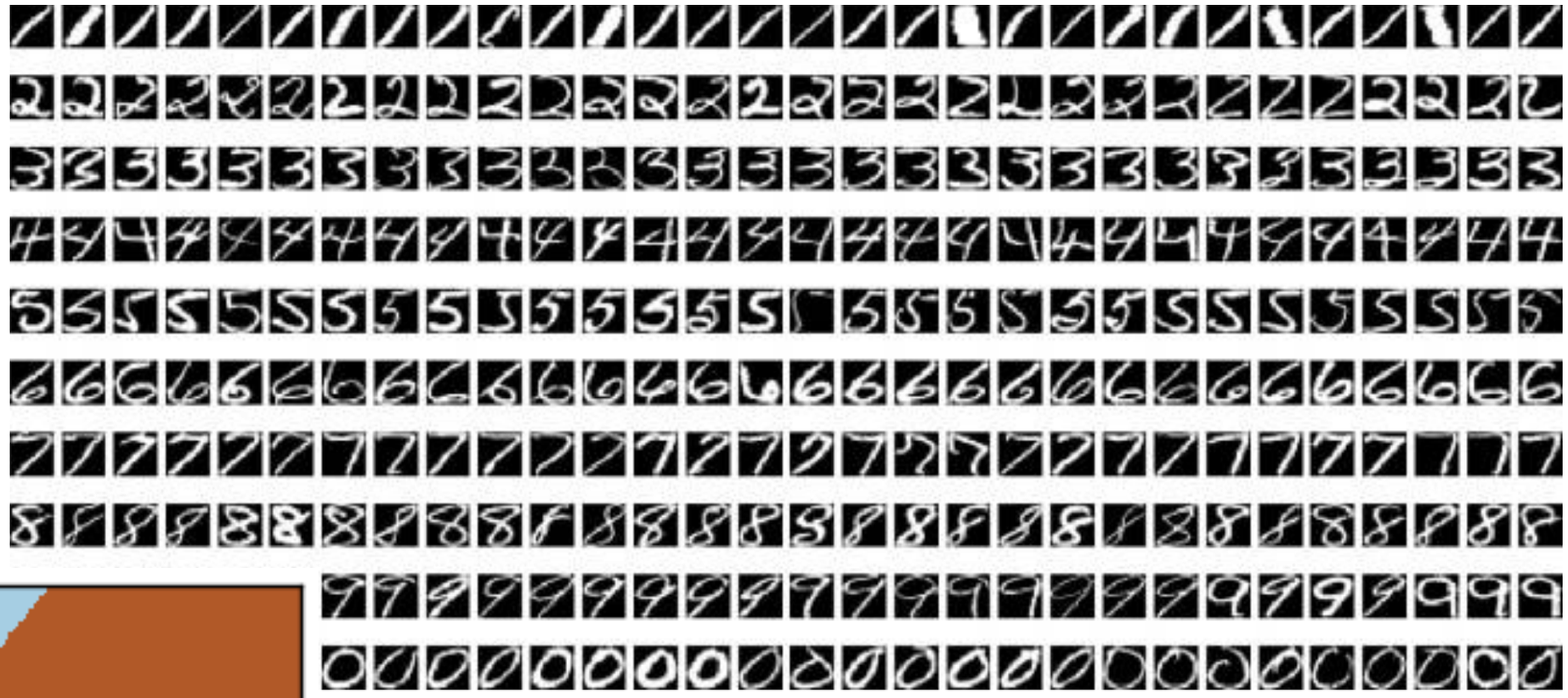
$$l(y, f(x))$$

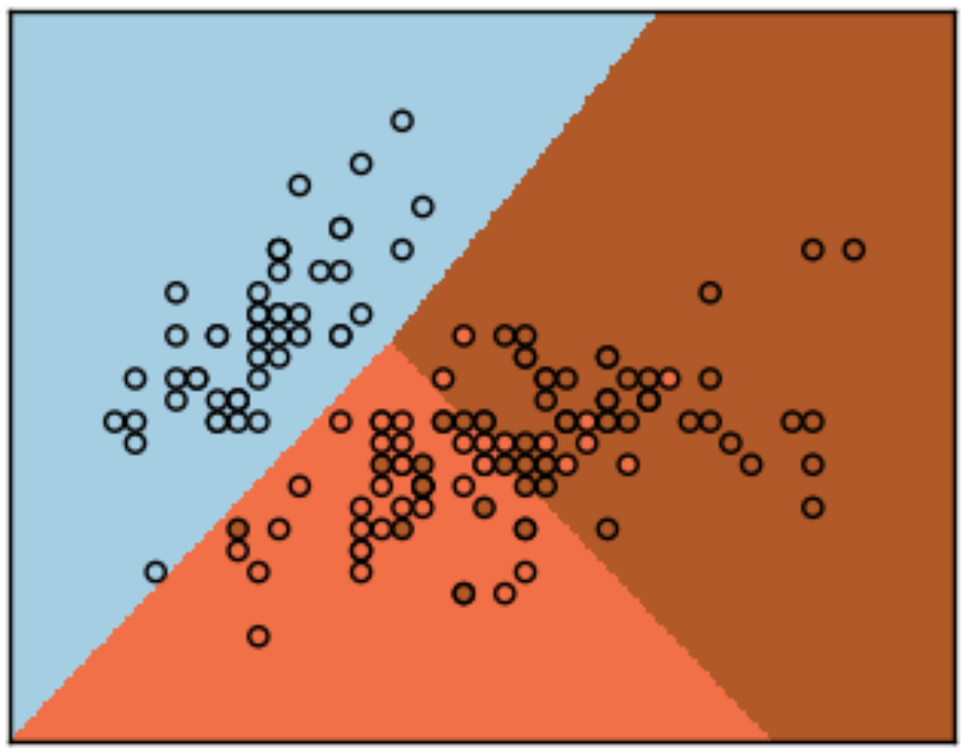
Binary Classification





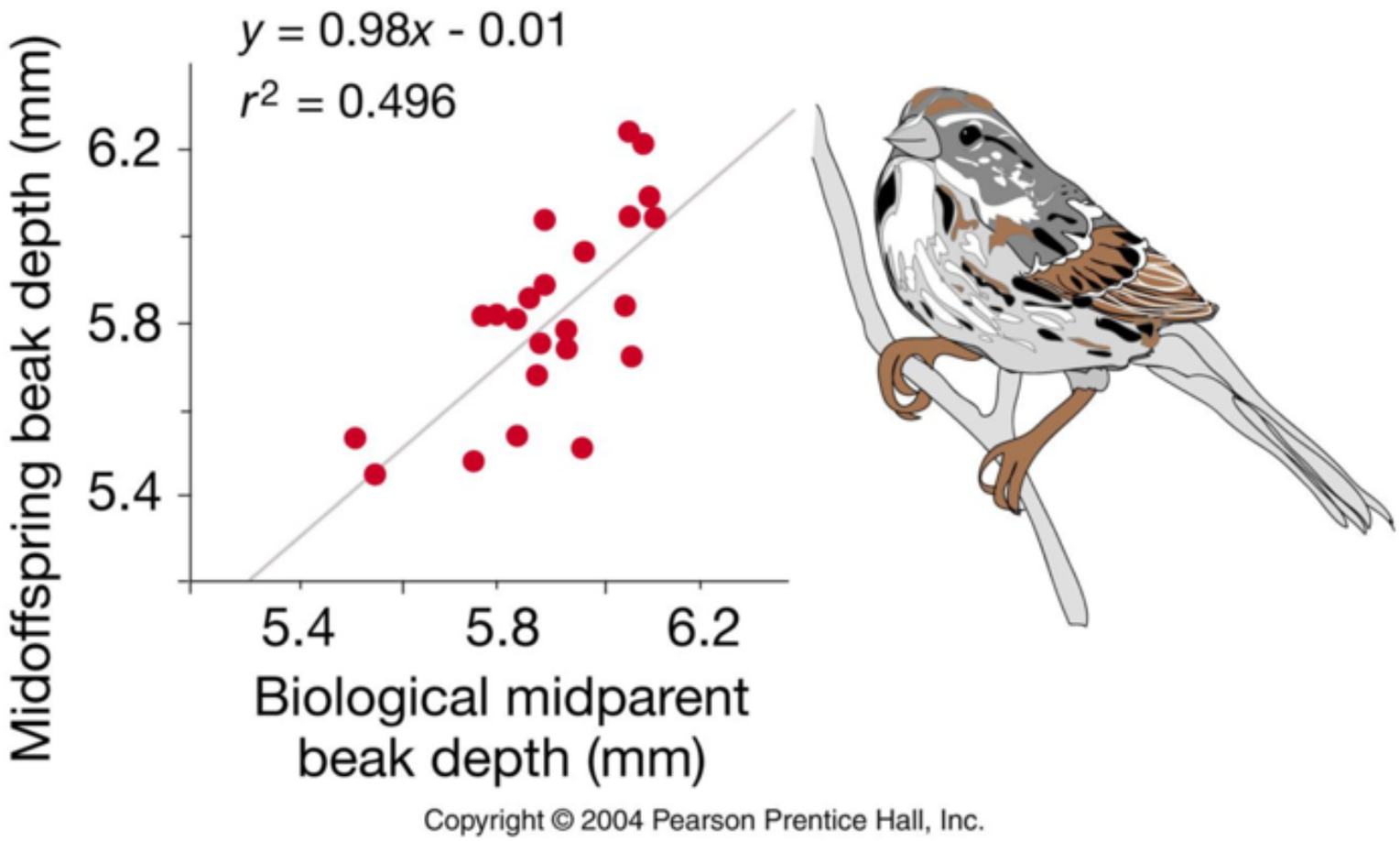
Multiclass Classification





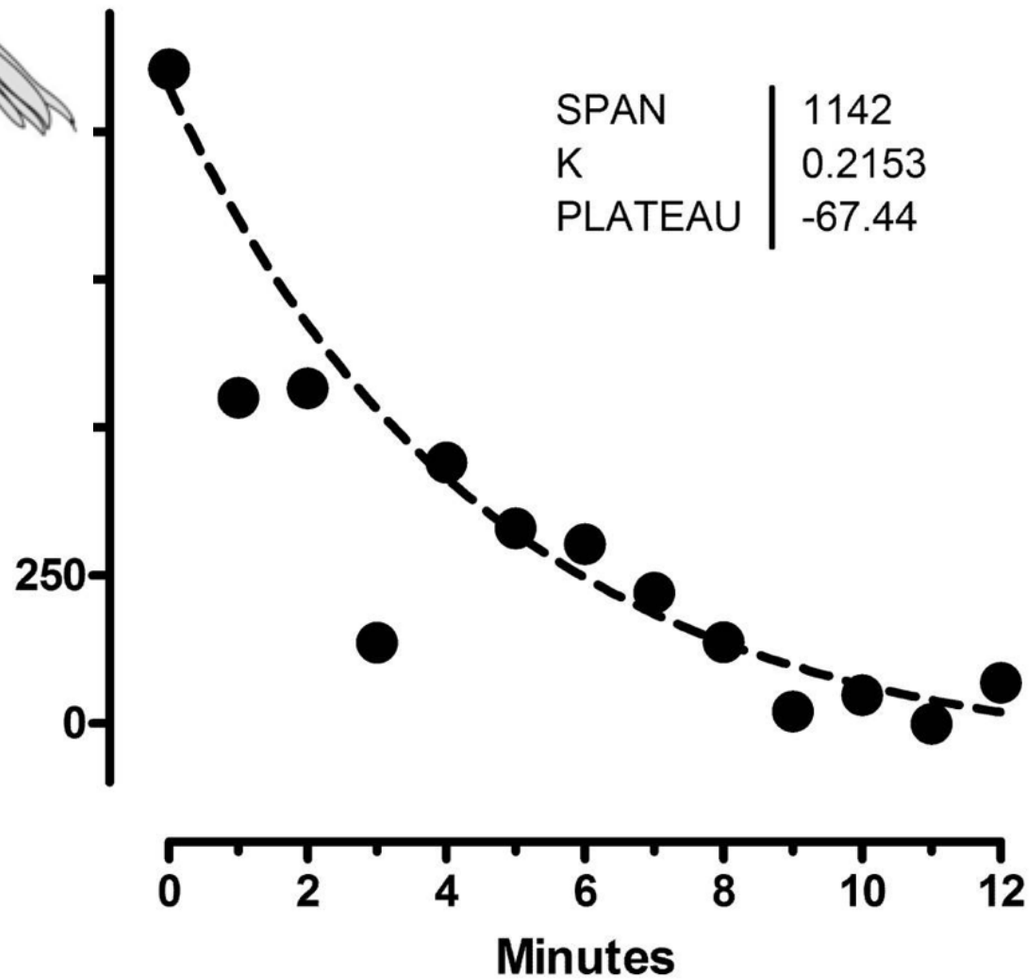
map image x to digit y

Regression

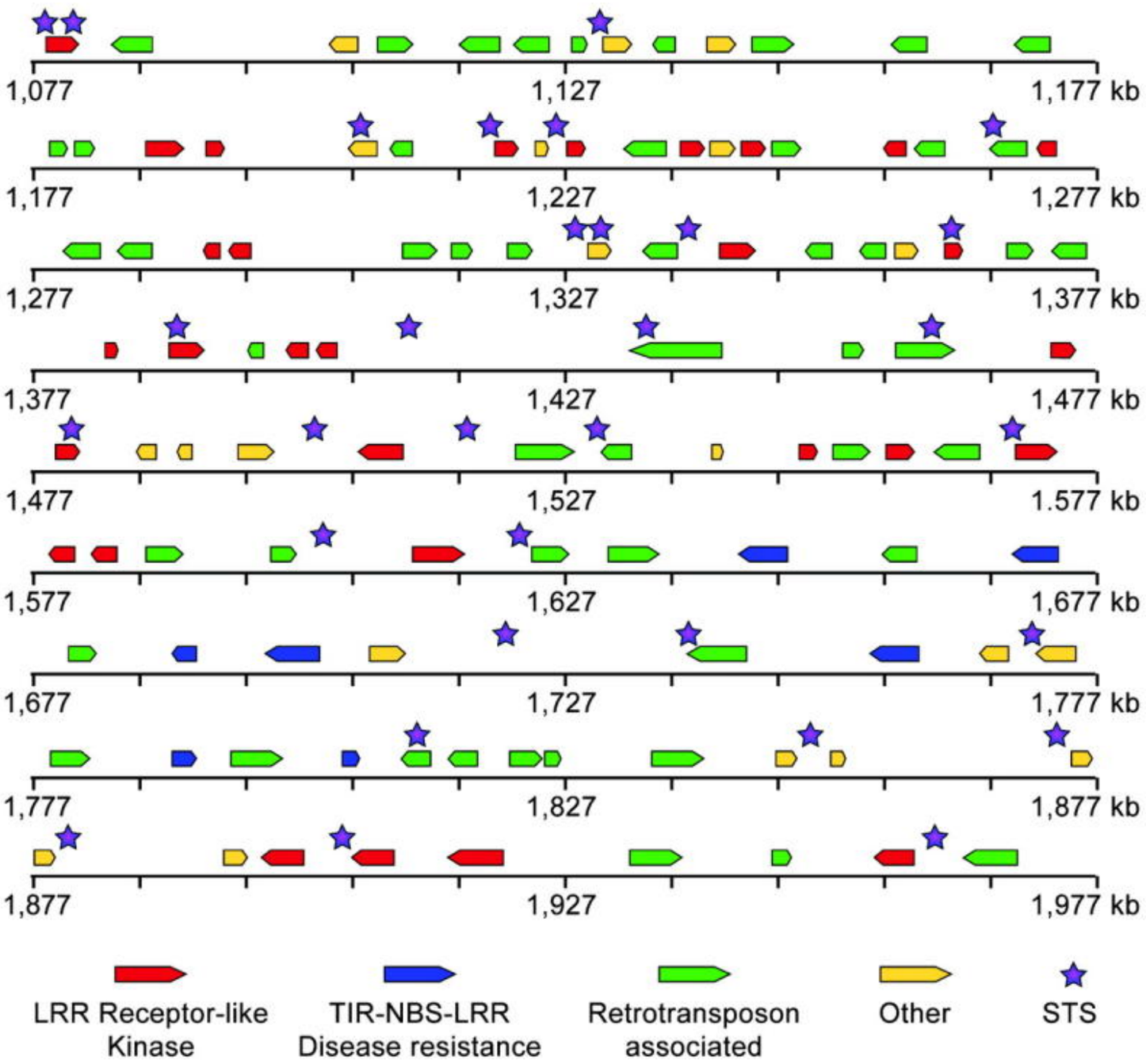


linear

nonlinear



Sequence Annotation



given sequence

gene finding
speech recognition
activity segmentation
named entities

Ontology

dmoz open directory project — In partnership with Aol Search.

about dmoz | dmoz blog | suggest URL | help | link | editor login

webpages

Search *advanced*

Arts
Movies, Television, Music...

Business
Jobs, Real Estate, Investing...

Computers
Internet, Software, Hardware...

Games
Video Games, RPGs, Gambling...

Health
Fitness, Medicine, Alternative...

Home
Family, Consumers, Cooking...

Kids and Teens
Arts, School Time, Teen Life...

News
Media, Newspapers, Weather...

Recreation
Travel, Food, Outdoors, Humor...

Reference
Maps, Education, Libraries...

Regional
US, Canada, UK, Europe...

Science
Biology, Psychology, Physics...

Shopping
Clothing, Food, Gifts...

Society
People, Religion, Issues...

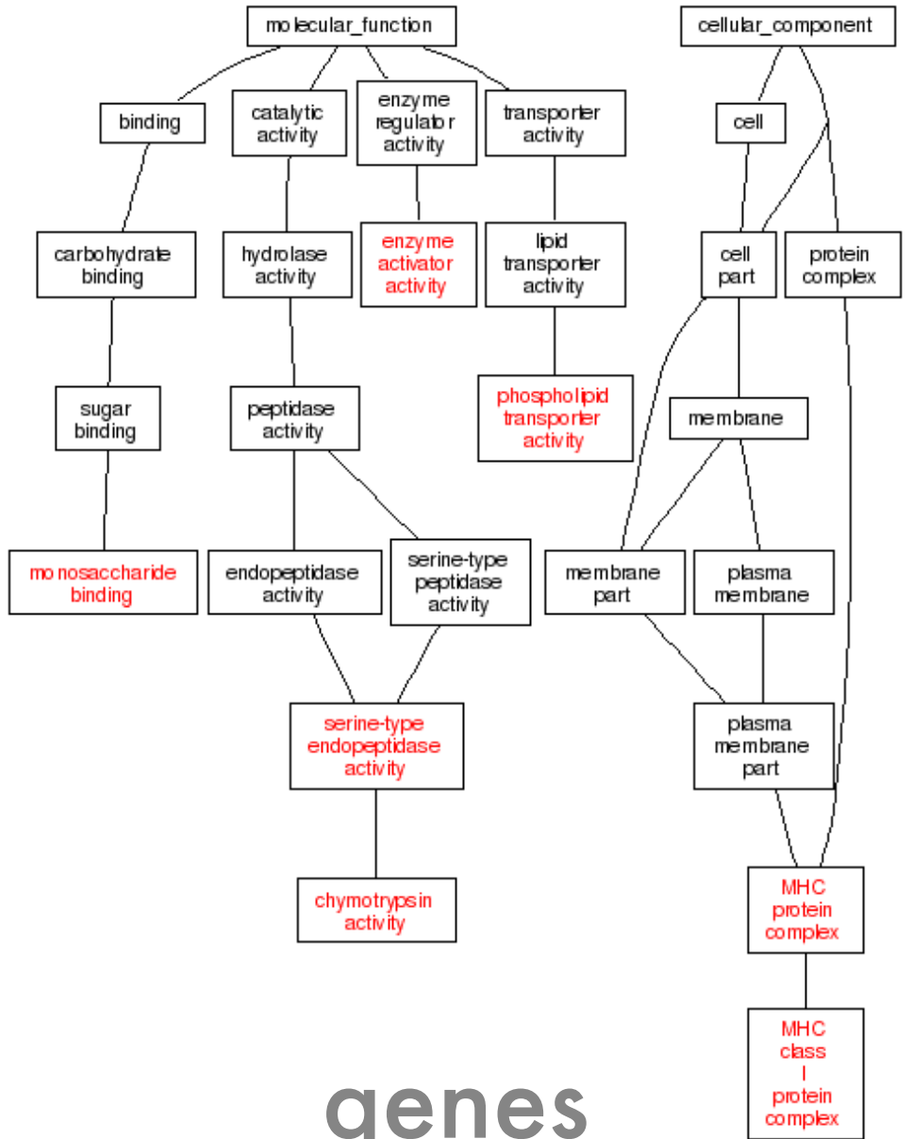
Sports
Baseball, Soccer, Basketball...

World
Català, Dansk, Deutsch, Español, Français, Italiano, 日本語, Nederlands, Polski, Русский, Svenska...

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

Carnegie Mellon University

Prediction

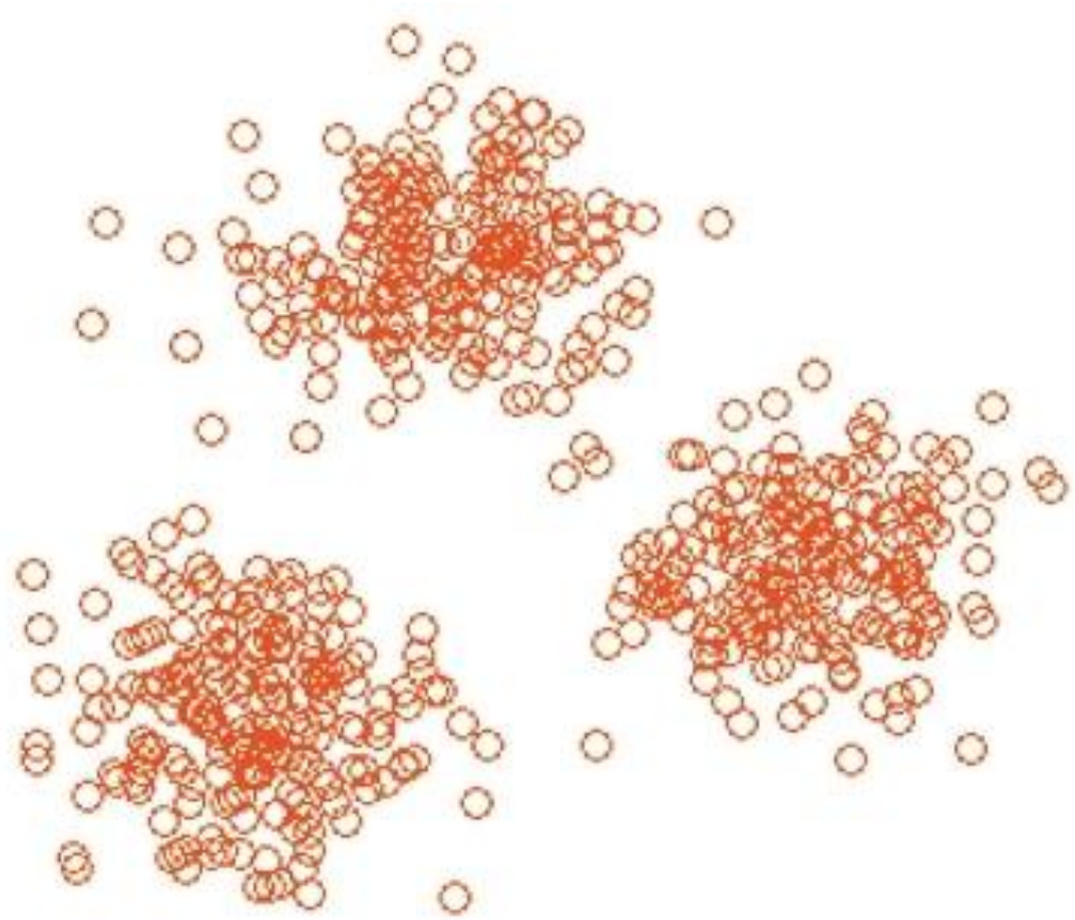
tomorrow's stock price

Carnegie Mellon University

Unsupervised Learning

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

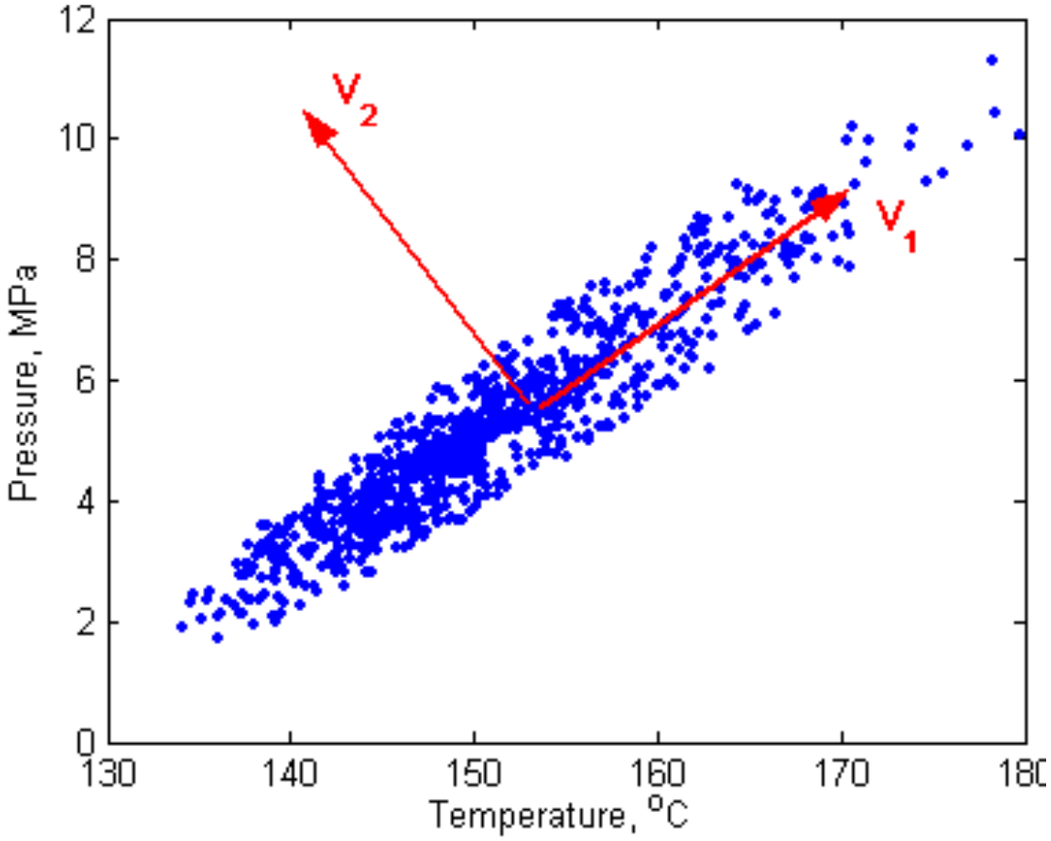
Clustering

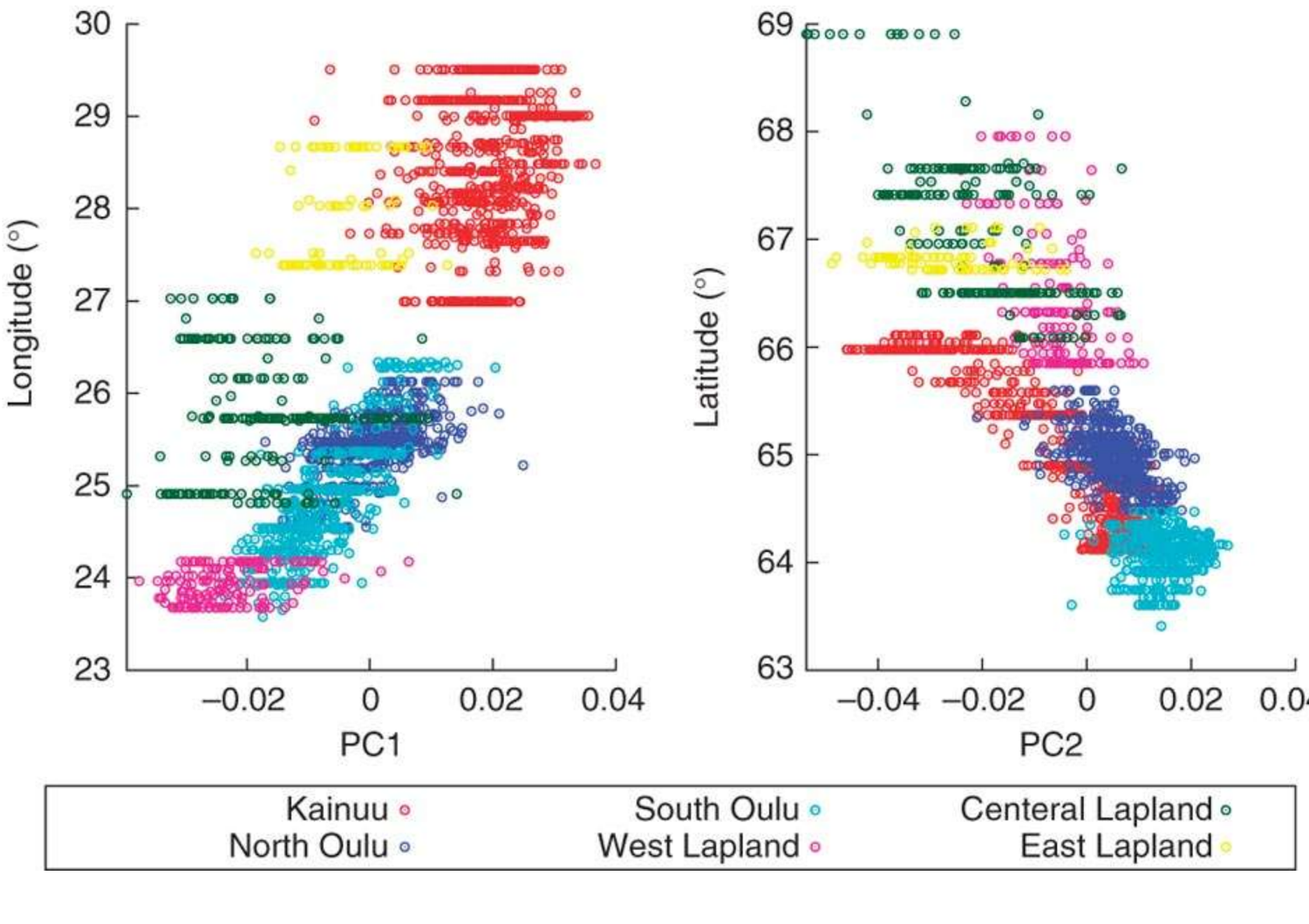


- Documents
- Users
- Webpages
- Diseases
- Pictures
- Vehicles

…

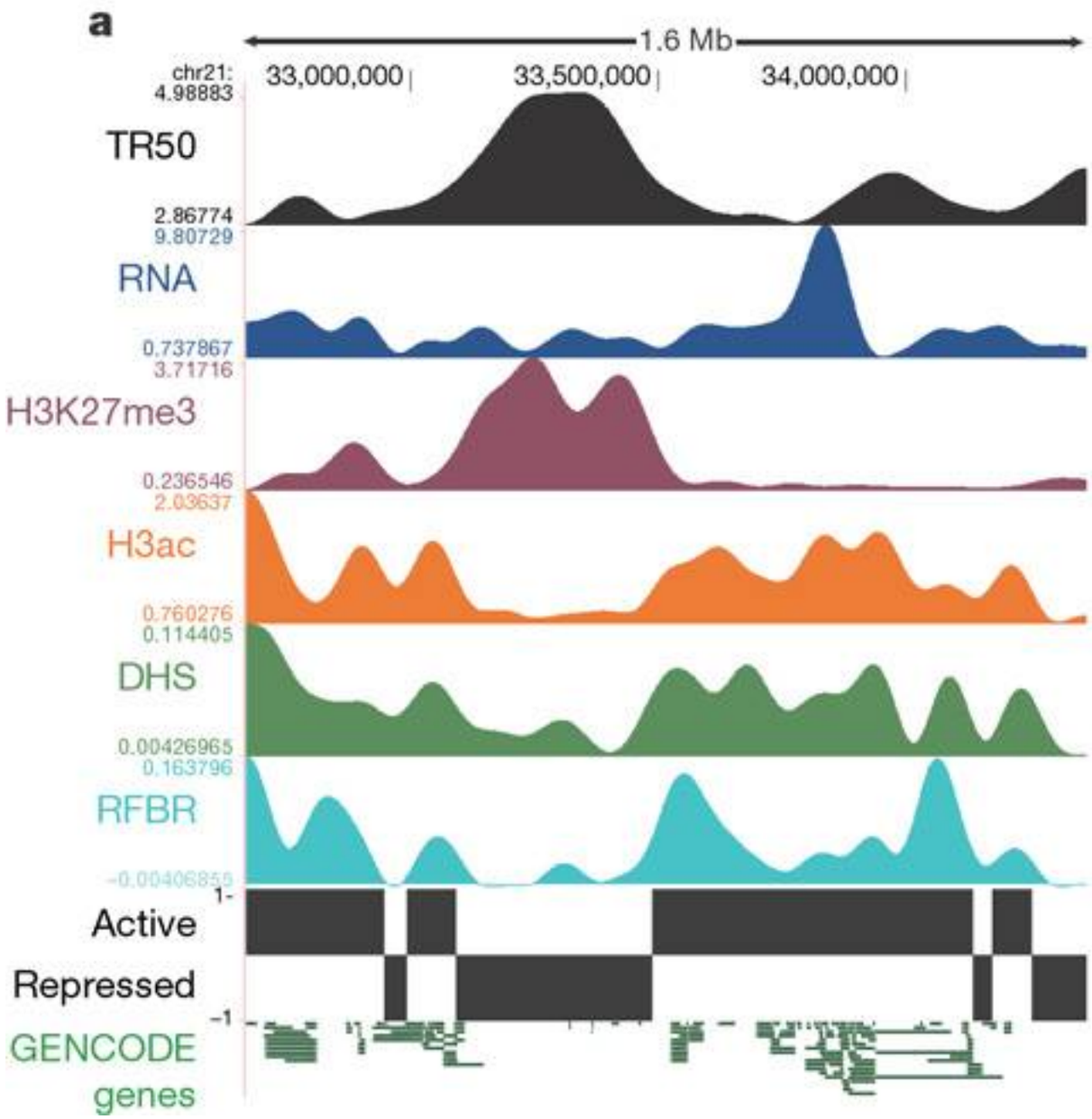
Principal Components

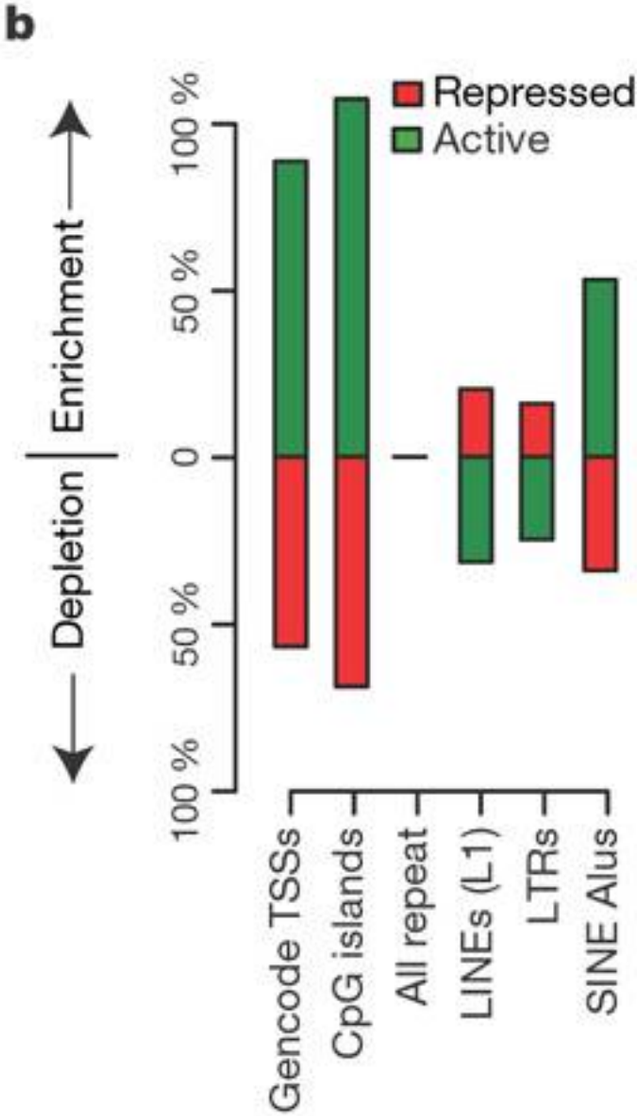




[Variance component model to account for sample structure in genome-wide association studies](), **Nature Genetics 2010**

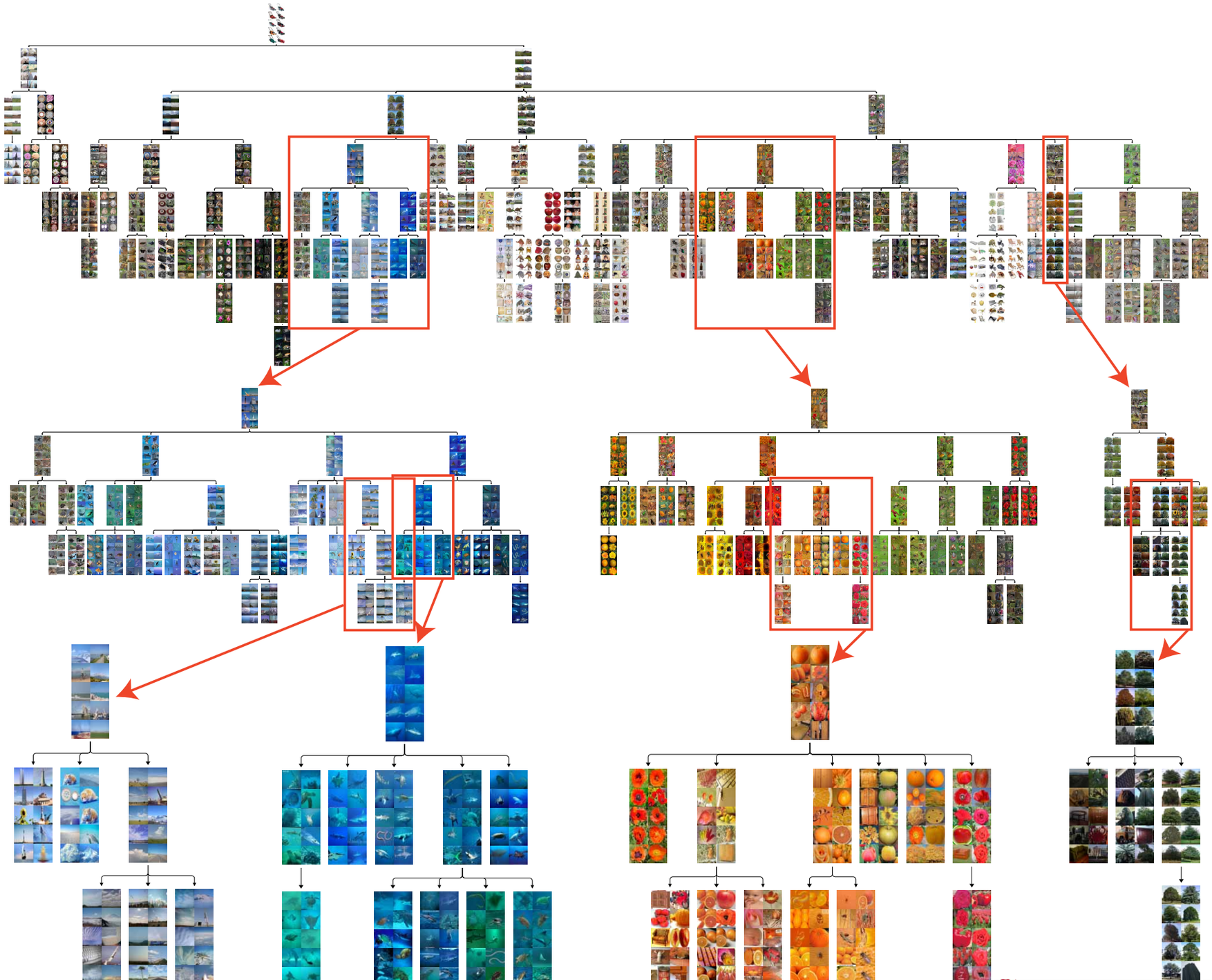
Sequence Analysis



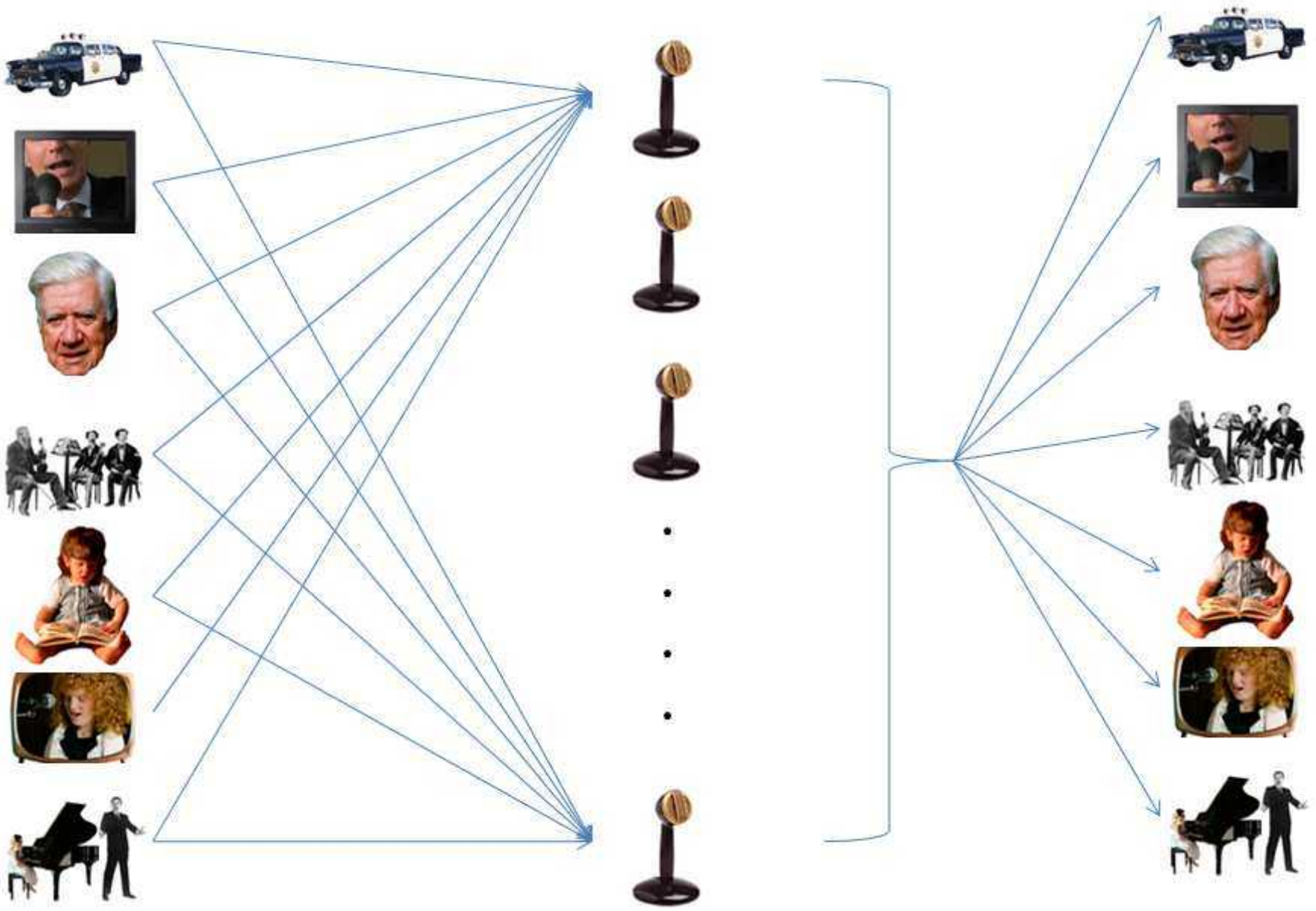


Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project, Nature 2007

Hierarchical Grouping

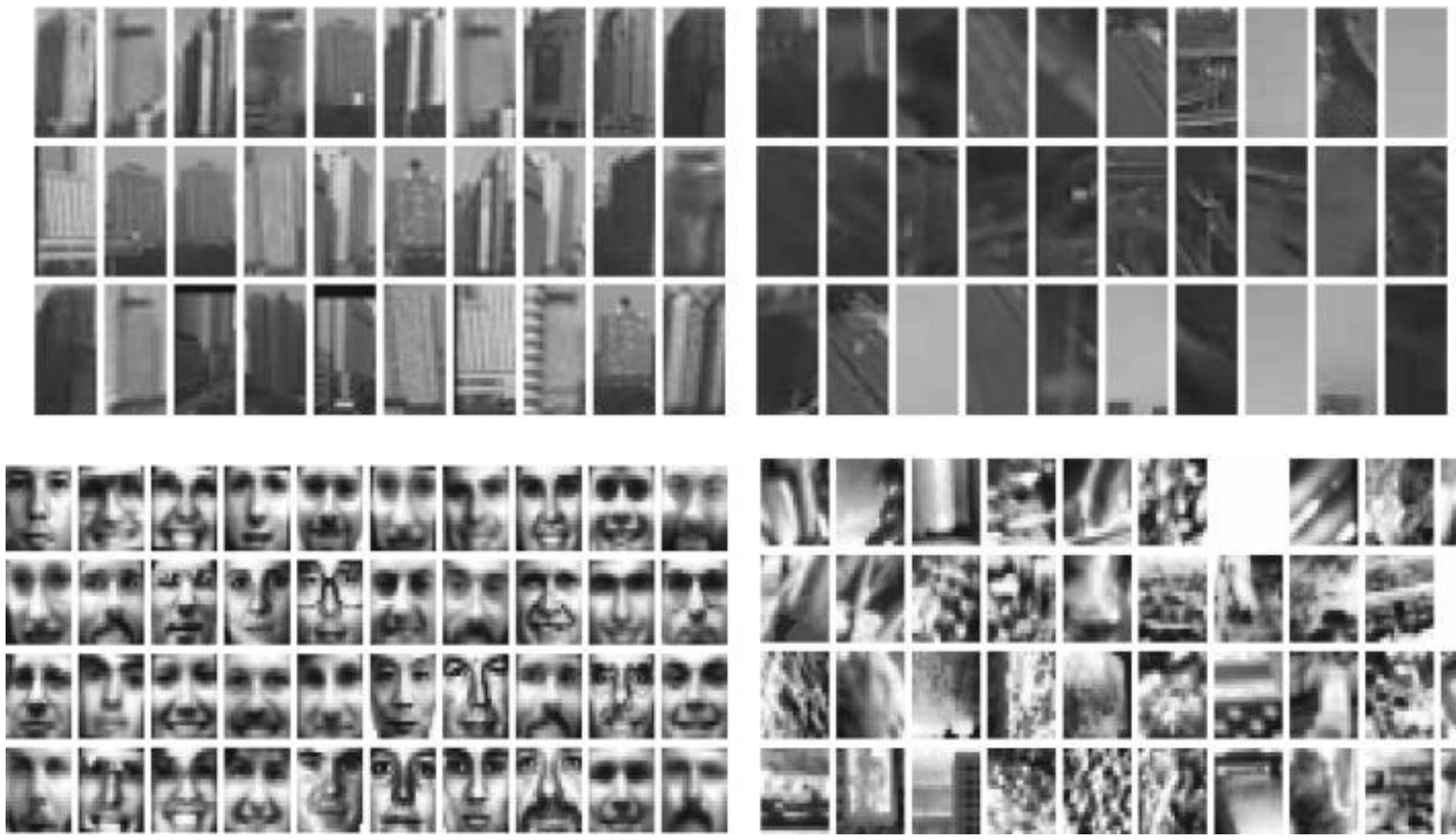


Independent Components



find them automatically

Novelty detection



typical

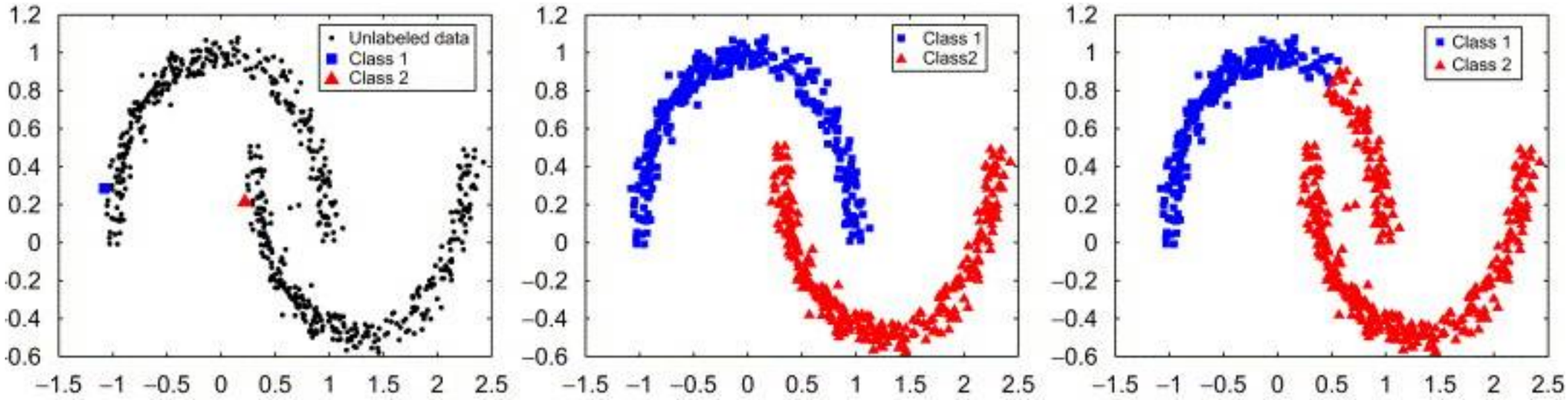
atypical

Some Problem types

iid = Independently Identically Distributed

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

Induction - Transduction



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

Covariate Shift

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

Cotraining and Multitask

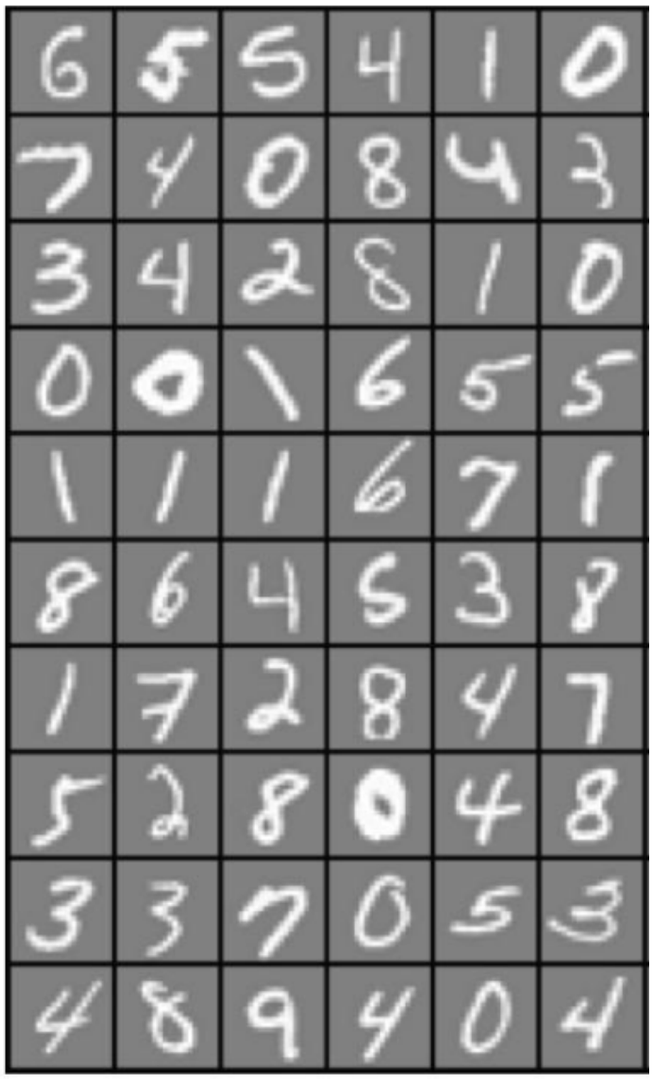
- Multitask Learning
  Use correlation between tasks for better result
  - Task 1 - Detect spammy webpages
  - Task 2 - Detect people's homepages
  - Task 3 - Detect adult content
- Cotraining
  For many cases both sets of covariates are available
  - Detect spammy webpages based on page content
  -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

Interaction with Environment

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

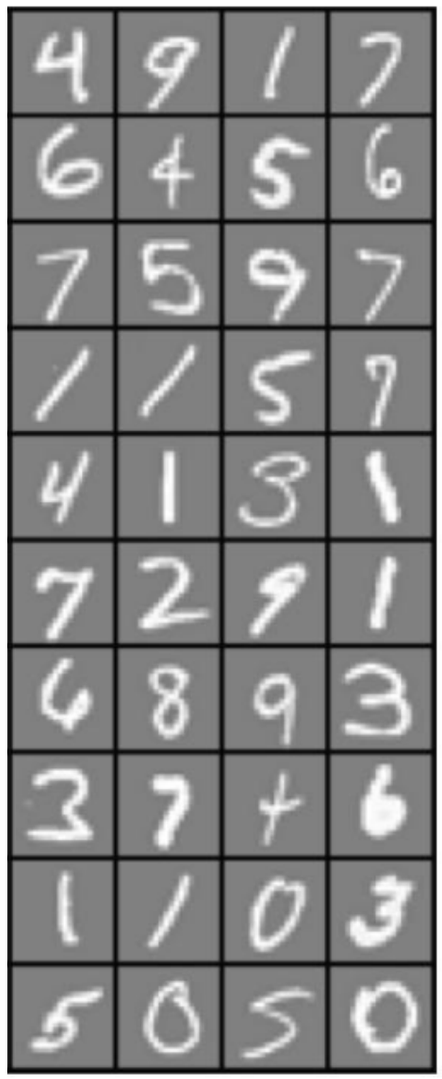
Batch

training data

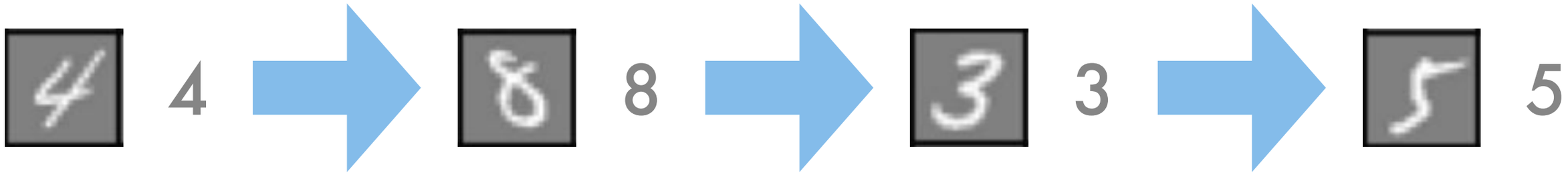


build model

test



Online



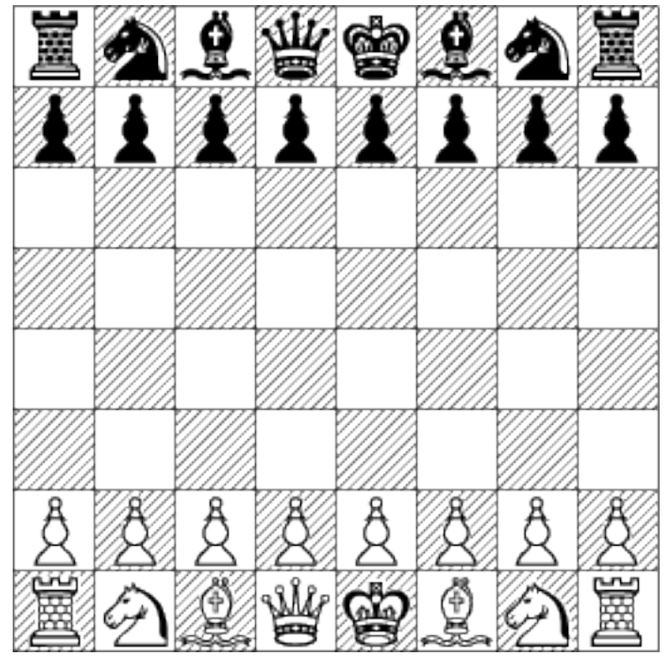
Bandits

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

Reinforcement Learning

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

- Repeat

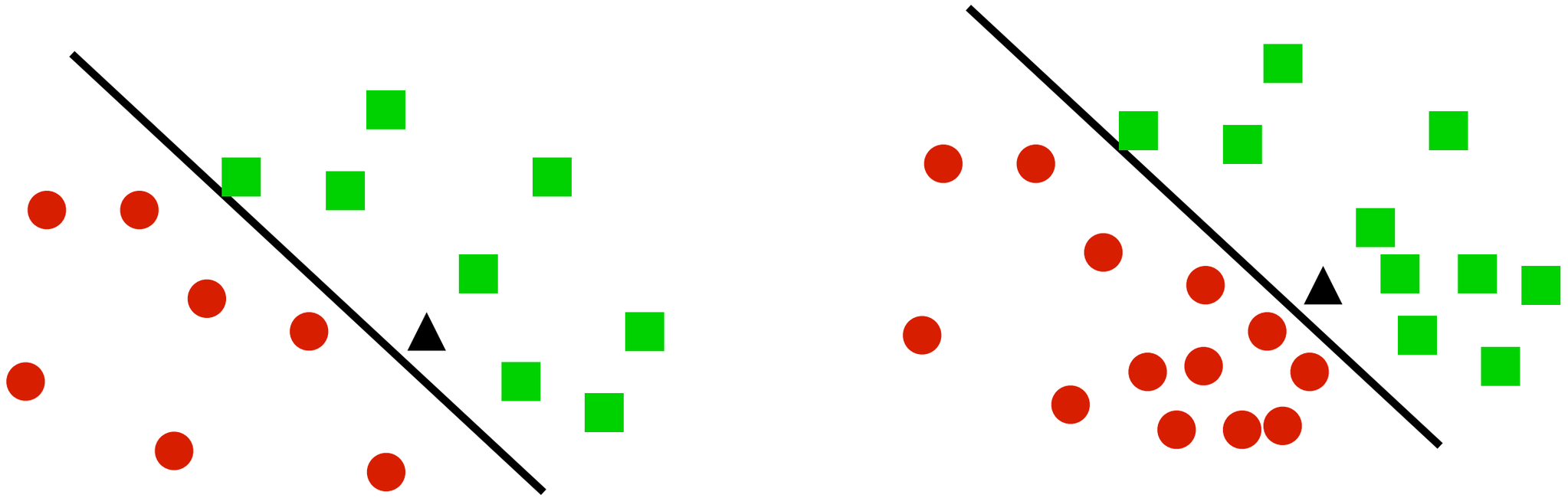


environment (cooperative, adversary, doesn't care)
memory (goldfish, elephant)
state space (tic tac toe, chess, car)

Discriminative vs. Generative (mainly relevant for supervised models)

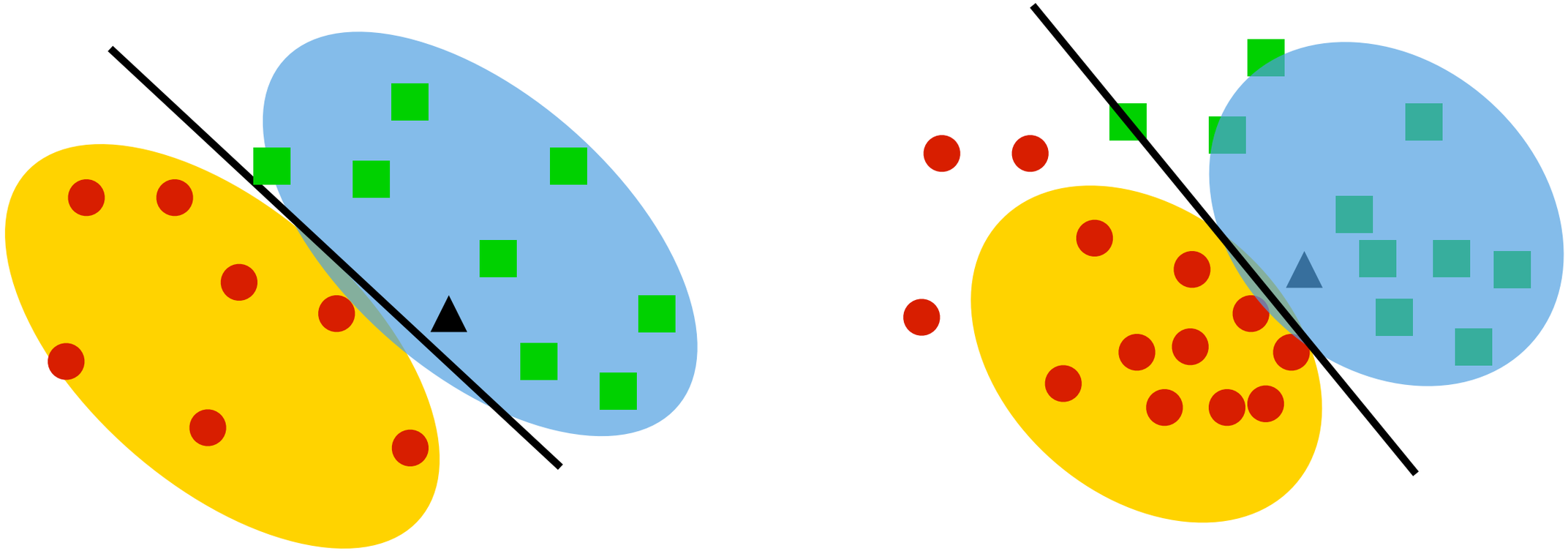
- Discriminative Models
  - Estimate y|x directly
  - Often better convergence + simpler solutions
- Generative models
  - Estimate joint distribution over (x,y)
  - Use conditional probability to infer y|x
  - Often more intuitive
  - Easier to add prior knowledge

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

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