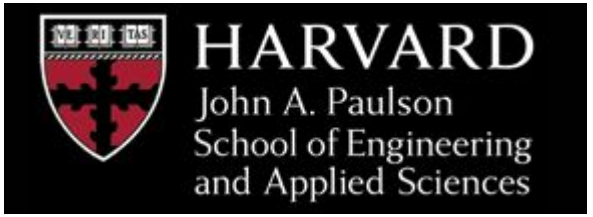
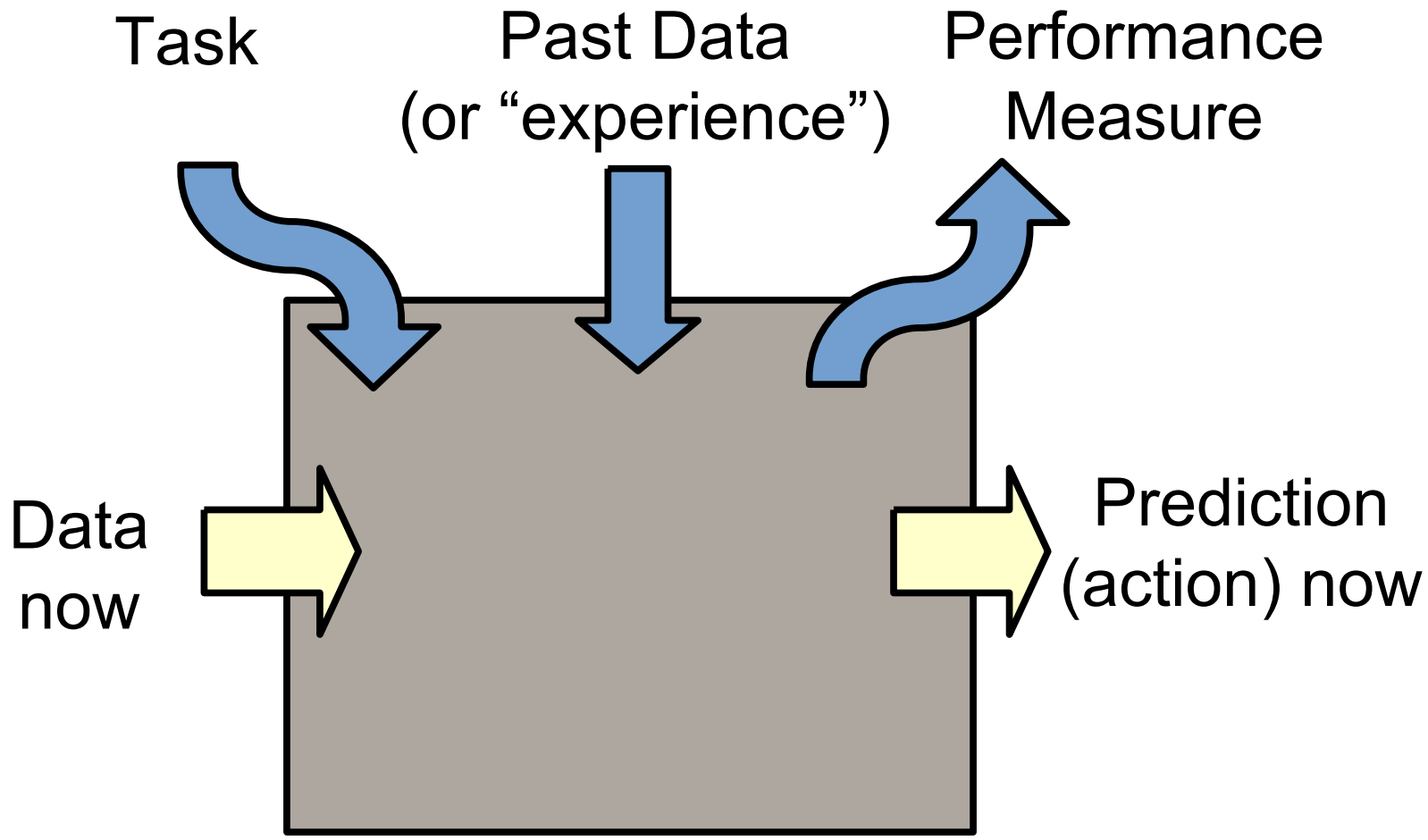
Machine Learning (CS 181)
Spring 2017

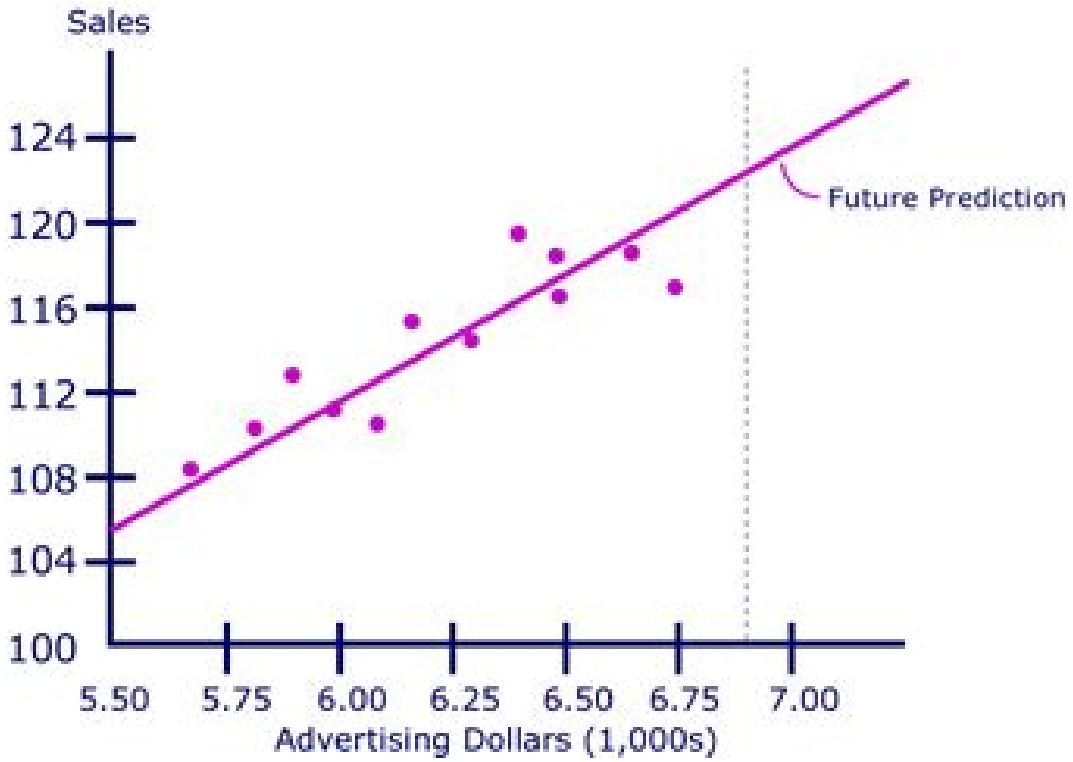
David Parkes
Sasha Rush



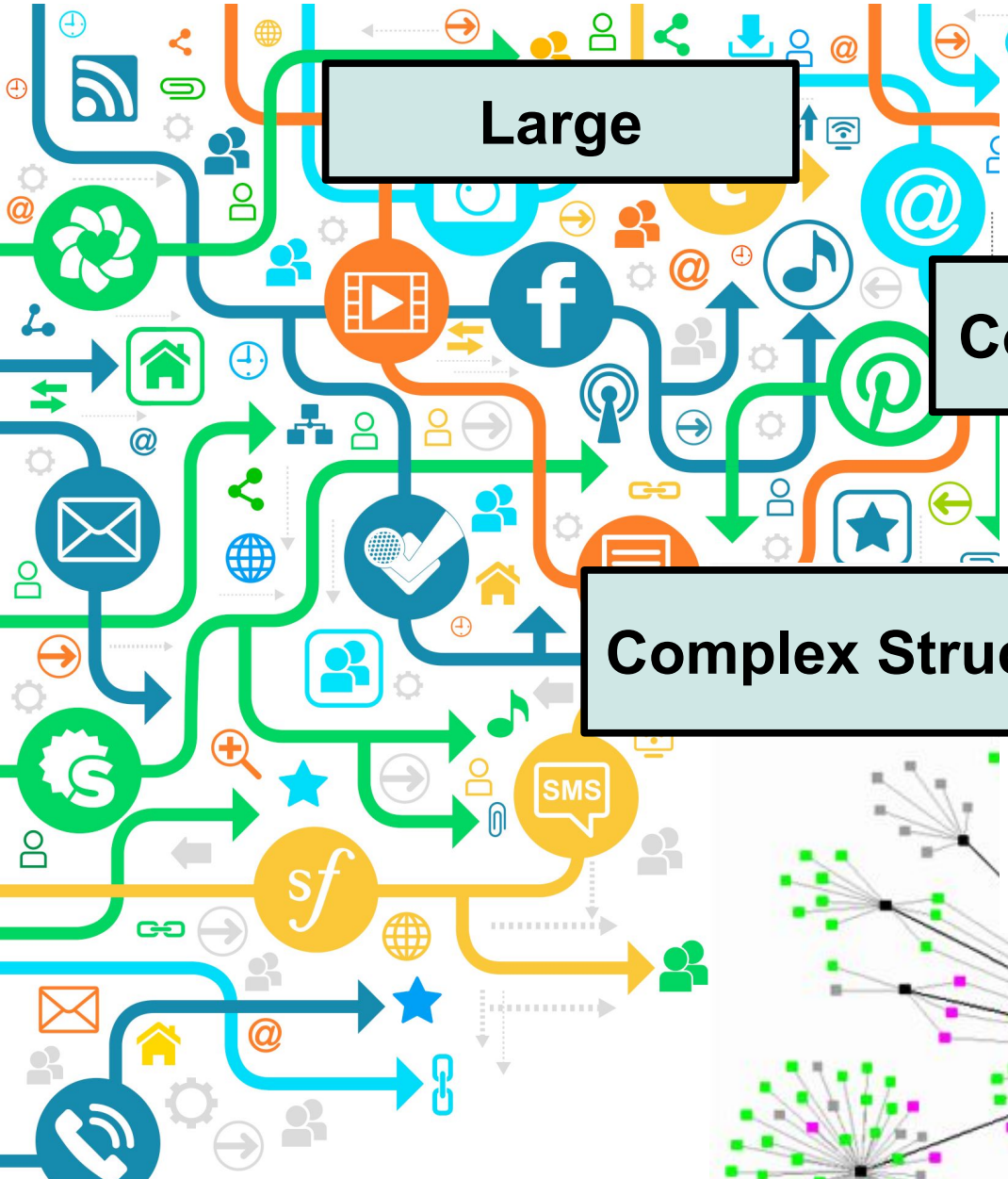
What is machine learning?



The starting point...



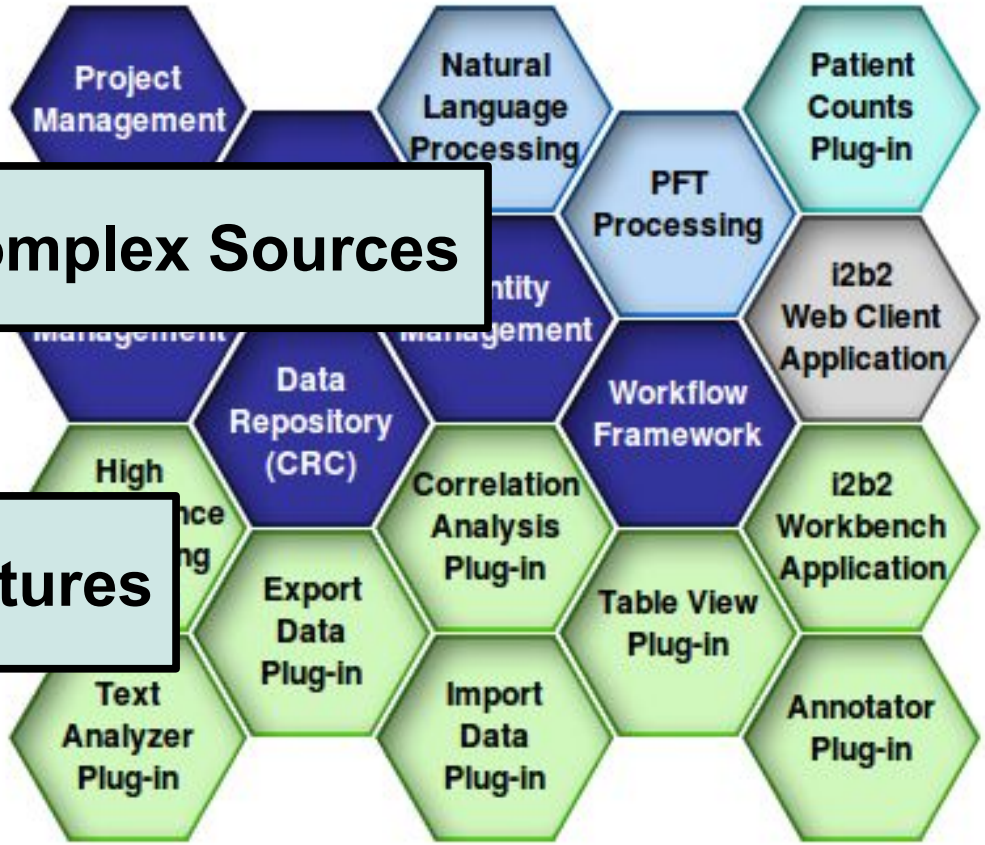
… where we are now...

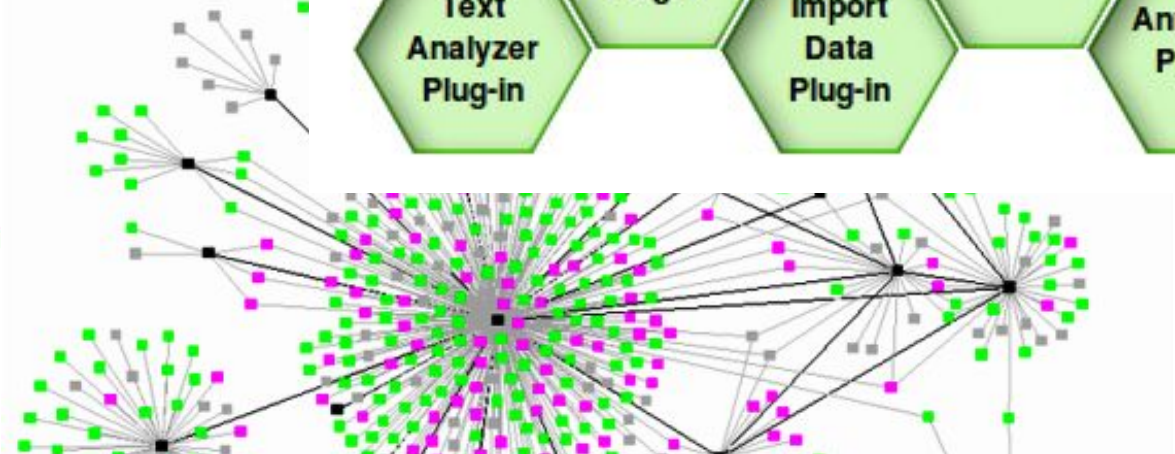


Large

Complex Sources

Complex Structures





Example: Collaborative marketing

~70,000 examples, each has ~3000 features

Predict “conversion rate”

[5] "REDEEMED"

[10] "PROVIDED_ZIP"

[13] "GIFT_PRICE"

[14] "DISTANCE"

[33] "WALKING_DISTANCE_FROM_SELECTED"

[38] "GMAIL"
[42] "HOTMAIL"
[46] "EDU"

[55] "MAC"
[56] "WINDOWS"

[66] "MOZILLA"
[67] "FIREFOX"

[85] "FEATURED_CATEGORY_EYEWEAR"
[96] "FEATURED_CATEGORY_HAIR_SALON_SPA"
[99] "FEATURED_MULTIPLE_CATEGORIES_BEAUTY_SPAS.SKIN_CARE"

[128] "FEATURED_MERCHANT_COOLIDGE_CORNER_YOGA"
[130] "FEATURED_MERCHANT_THE_CLAYROOM"
[133] "FEATURED_MERCHANT_MELLIE_HAIR_SALON"

[862] "FEATURED_MERCHANT_EASTERN_CLOTHING_CO"

[986] "person.interests.interestgroups.Photography"

[1016] "person.interests.pets.Cat owner"

[1128] "person.mediaConsumption.mediaUsage.likely.Internet"

[1322] "person.basicDemographics.occupation.Architect"

[2010] "person.productPurchases.gifts.flowers"

[2042] "person.productPurchases.hobbies.Hunting"

[2262] "person.personicxLifestage.lifestageGroup.18M"

[2322] "person.vehicleOwnership.vehicles.make.Dodge"

[2460] "household.basicDemographics.adoptionPlans.Very likely"

[2504] "household.basicDemographics.Unknown gender 35-44"

[2538] "household.basicDemographics.babyPlans.Very unlikely"

[2602] "household.basicDemographics.countryOfOrigin.Czech"

[2716] "household.basicDemographics.engagementPlans.Unlikely"

[2792] "household.basicDemographics.estimatedIncomeMax.49999"

[2804] "household.basicDemographics.ethnicity.Hispanic"

[2828] "household.basicDemographics.maritalStatus.Married"
[2832] "household.basicDemographics.maritalStatus.Inferred Married"

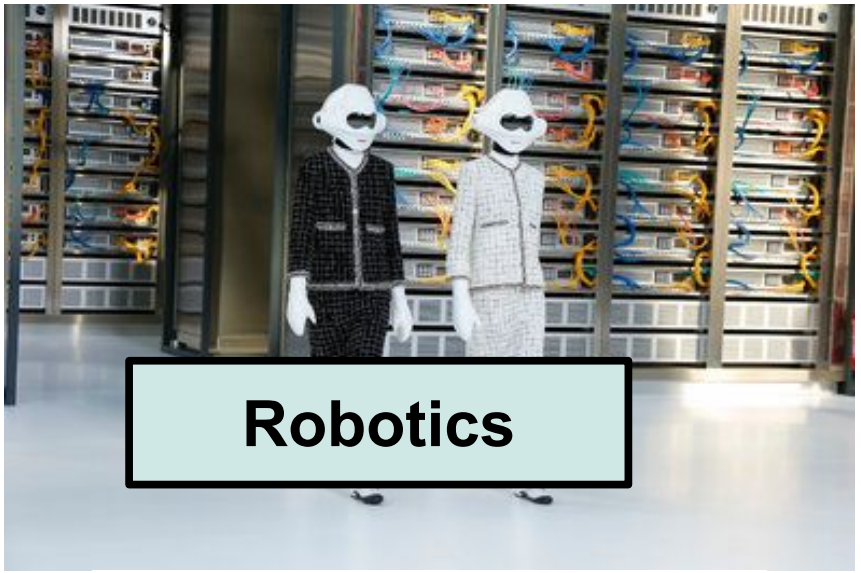
[2846] "household.basicDemographics.ownerRenter.Owner"

[2872] "household.basicDemographics.religiousAffiliation.Christian"

[2884] "household.basicDemographics.youngAdult"

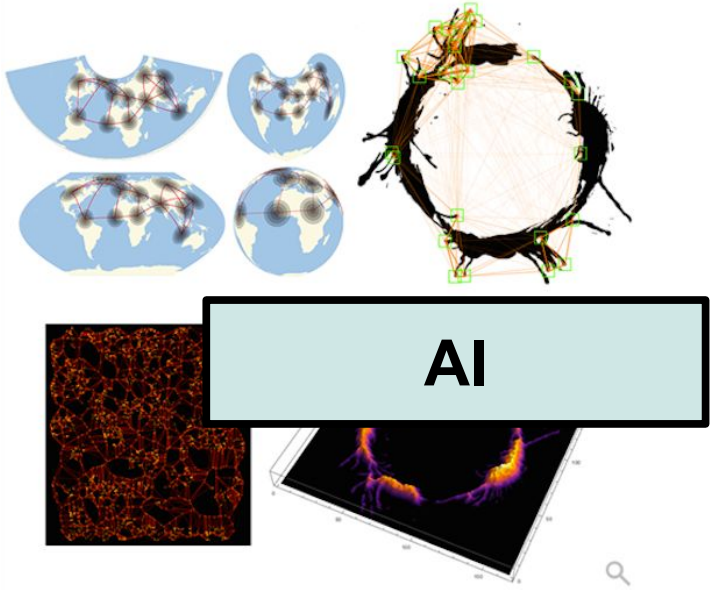
… where we're going

Language/Vision



Robotics

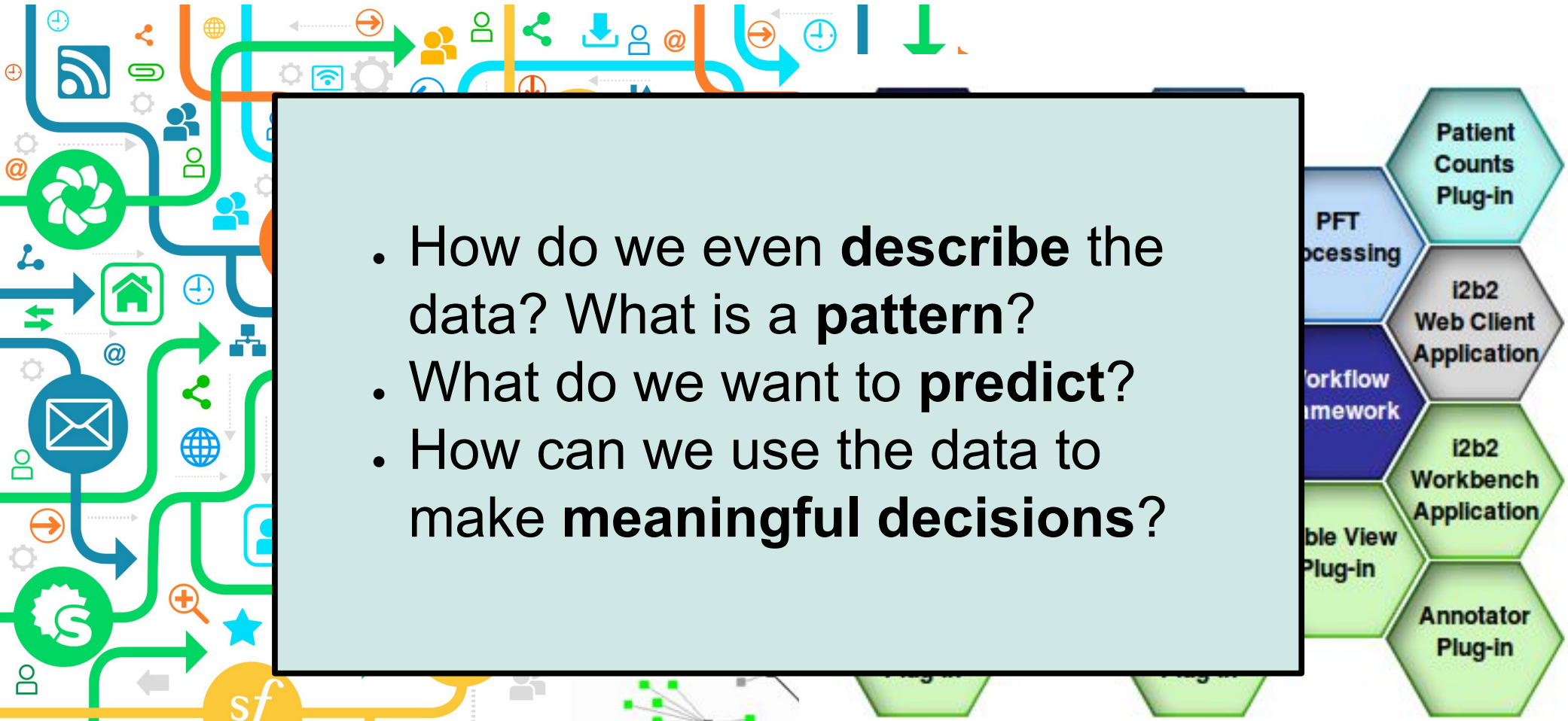
Decision Making



AI

Fundamentals

- How do we even **describe** the data? What is a **pattern**?
- What do we want to **predict**?
- How can we use the data to make **meaningful decisions**?



Example: Disease Subtyping

(From the research of Finale Doshi-Velez)

Many “single diseases” have varied manifestations.

- Scientific importance: different etiologies
- Clinical importance: different treatment plans

2012 CDC: 1/88 kids have autism spectrum disorder (ASD)

Autism is incurable, but **appropriate early intervention** can boost a child's odds of becoming **an independent adult**.

Case: Autism Spectrum Disorders

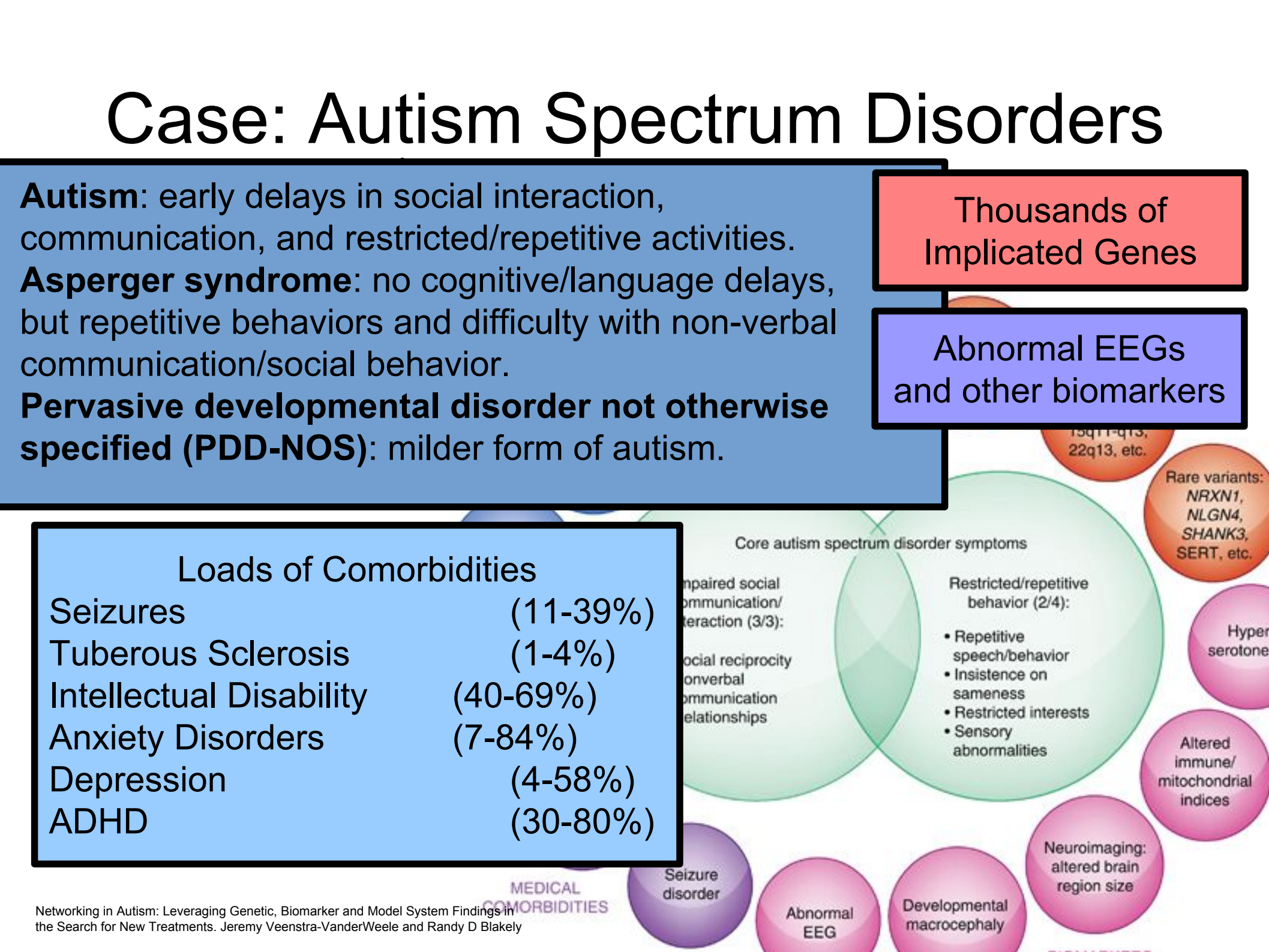
Autism: early delays in social interaction, communication, and restricted/repetitive activities.
Asperger syndrome: no cognitive/language delays, but repetitive behaviors and difficulty with non-verbal communication/social behavior.
Pervasive developmental disorder not otherwise specified (PDD-NOS): milder form of autism.

Thousands of Implicated Genes

Abnormal EEGs and other biomarkers

Loads of Comorbidities

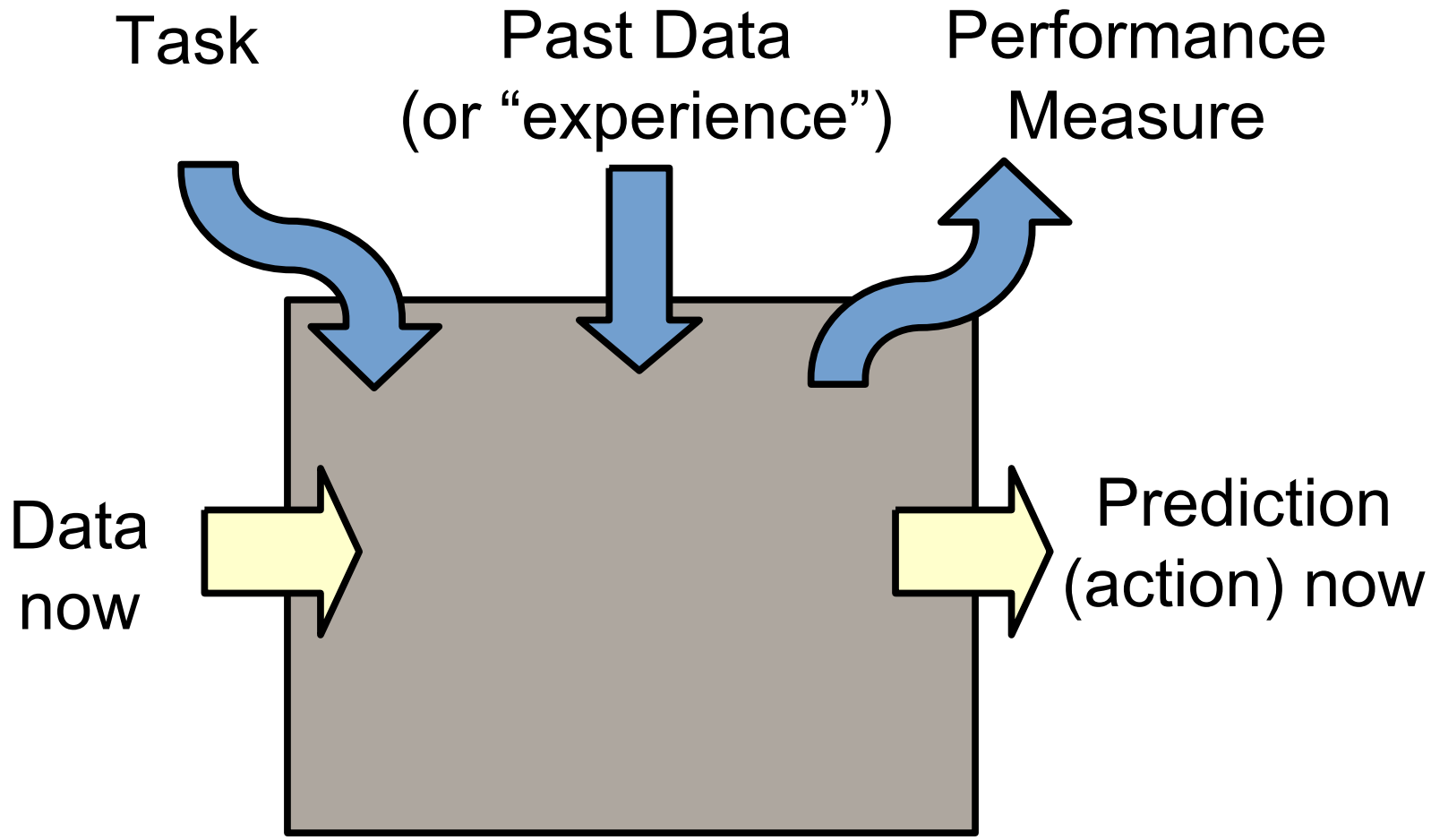
Seizures	(11-39%)
Tuberous Sclerosis	(1-4%)
Intellectual Disability	(40-69%)
Anxiety Disorders	(7-84%)
Depression	(4-58%)
ADHD	(30-80%)



Networking in Autism: Leveraging Genetic, Biomarker and Model System Findings in the Search for New Treatments. Jeremy Veenstra-VanderWeele and Randy D Blakely

“If you've met a kid with autism, you've met **a** kid with autism.”

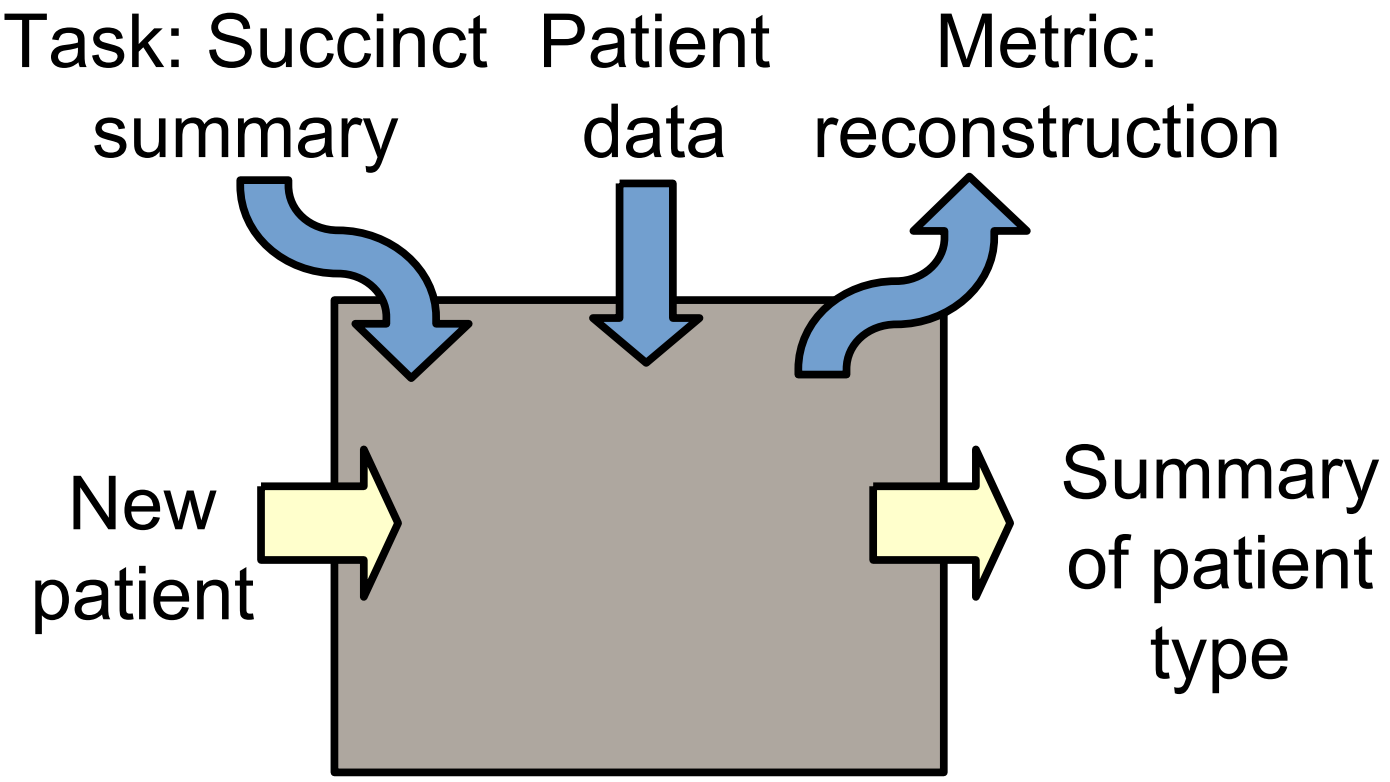
What is machine learning?



How can Machine Learning help?

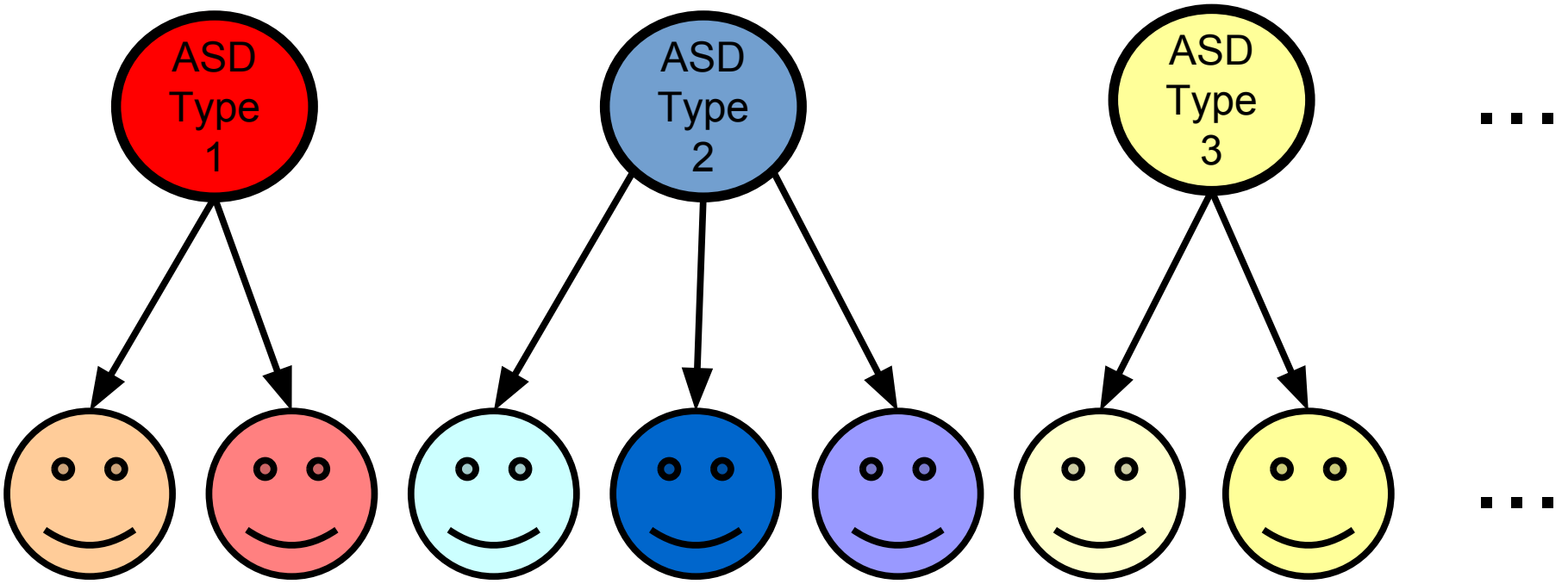
Discovering Subtypes

- **Scientific Question:** Do the patterns of co-occurring diagnoses suggest a few subtypes in autism?
- **Technical Question**: Can we succinctly describe the patterns of co-occurring diagnoses with a few *hidden variables*?



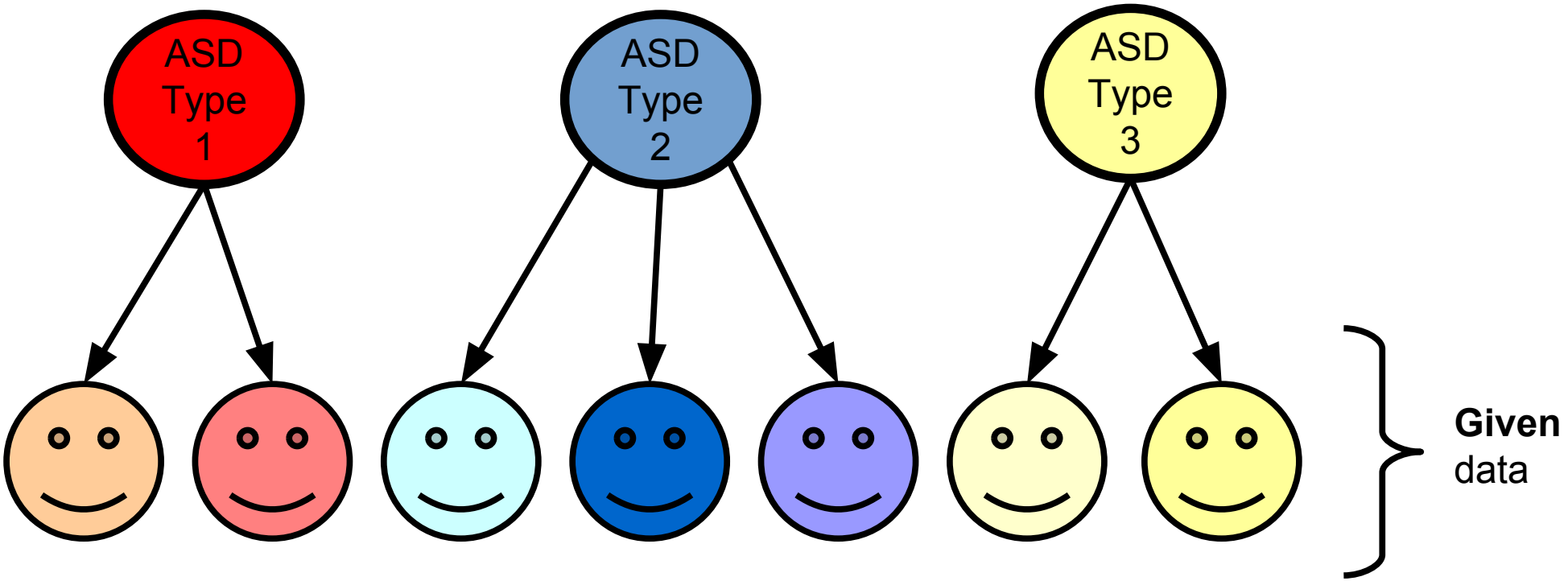
Modeling Effort, Part I

Each **patient** belongs to an Autism (**ASD) type**.
Each **ASD type** is characterized by a **set of comorbidities**.



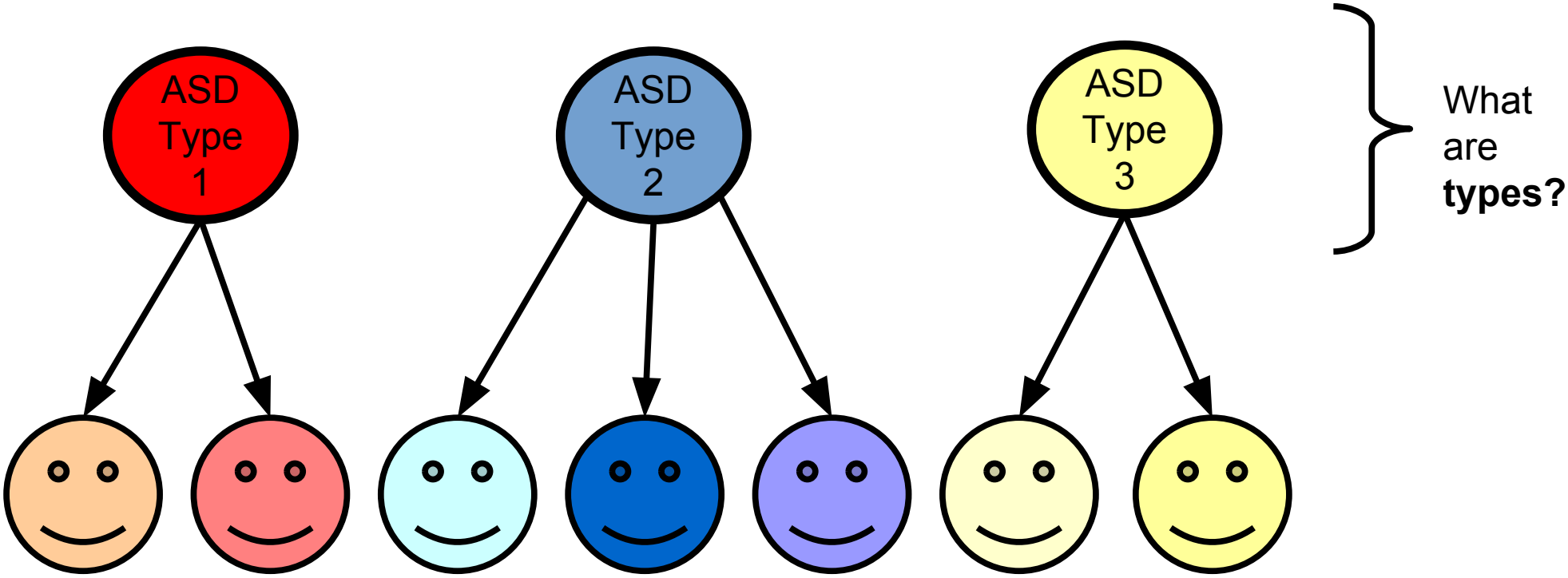
Modeling Effort, Part I

Each **patient** belongs to an Autism (**ASD) type**.
Each **ASD type** is characterized by a **set of comorbidities**.



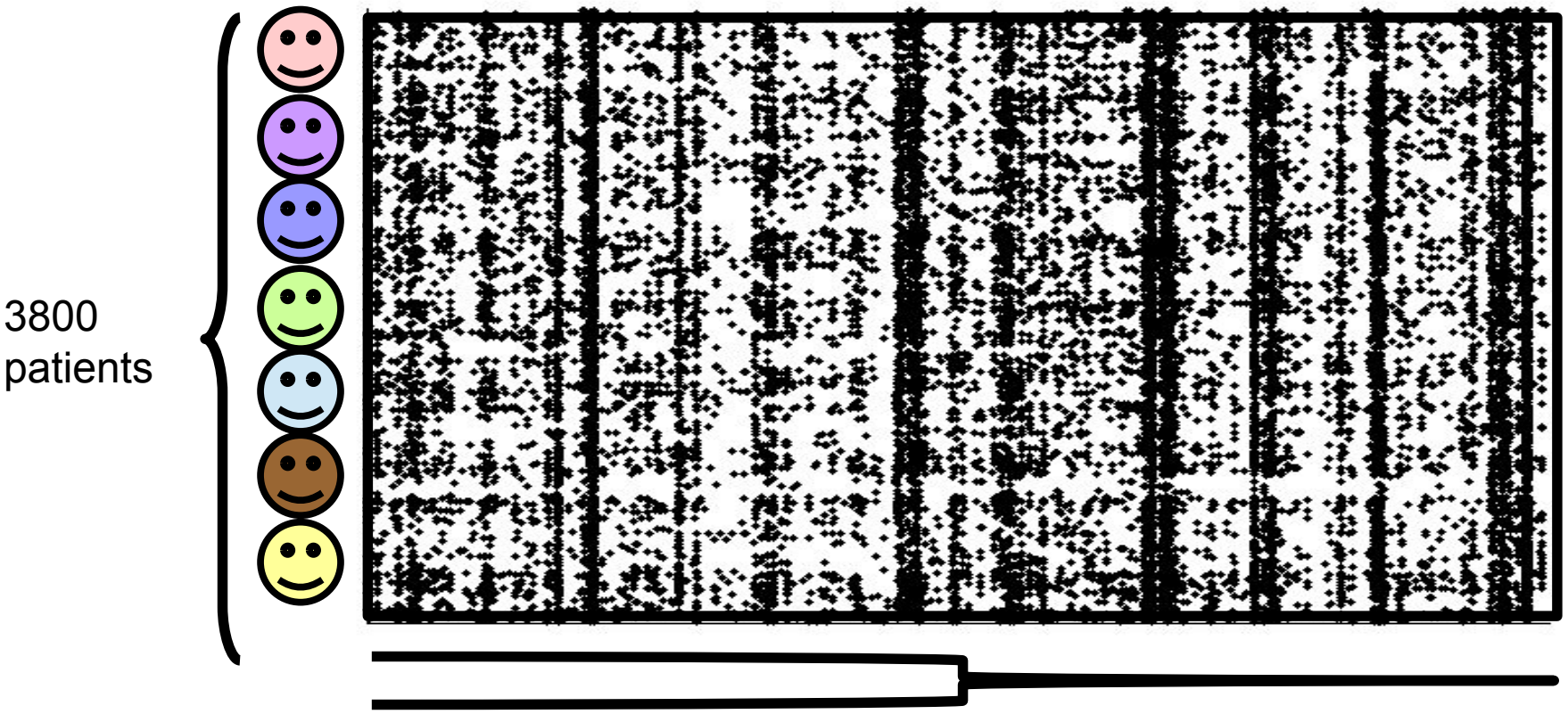
Modeling Effort, Part I

Each **patient** belongs to an Autism (**ASD) type**.
Each **ASD type** is characterized by a **set of comorbidities**.



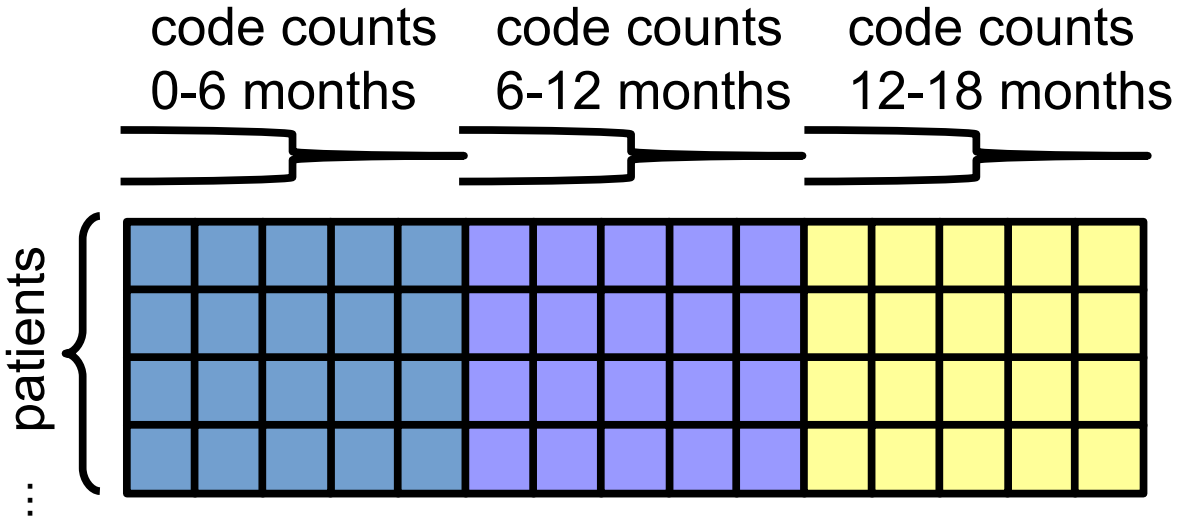
The Data

Diagnoses extracted from the electronic health records of patients with austism at Boston Children's Hospital. (Processed to ~3800 patients and ~3600 unique diagnoses).

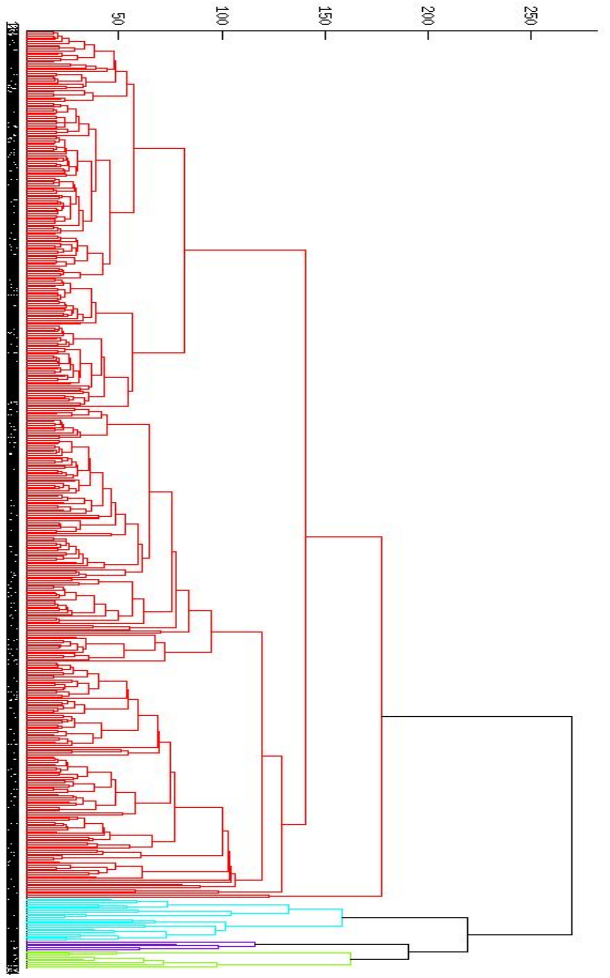


Discover the ASD Types (Clustering)

Count occurrences of popular codes in 6-month time windows from age 0-15.

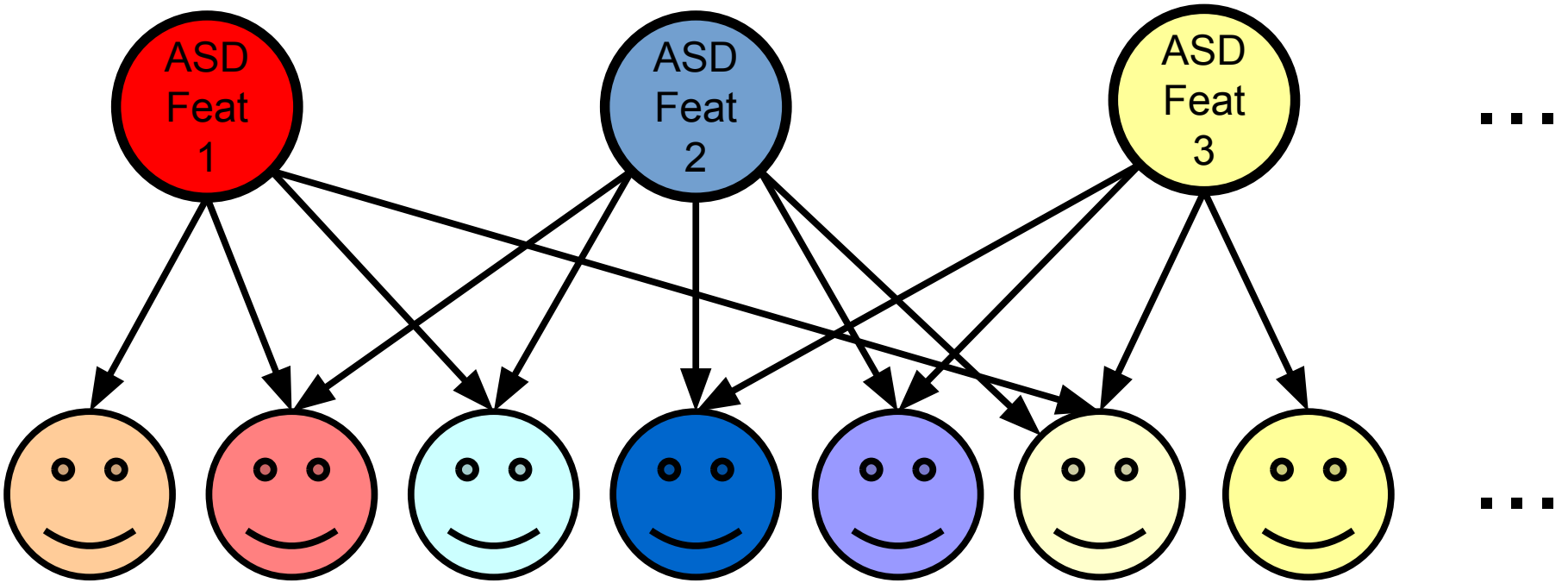


Apply clustering algorithm.



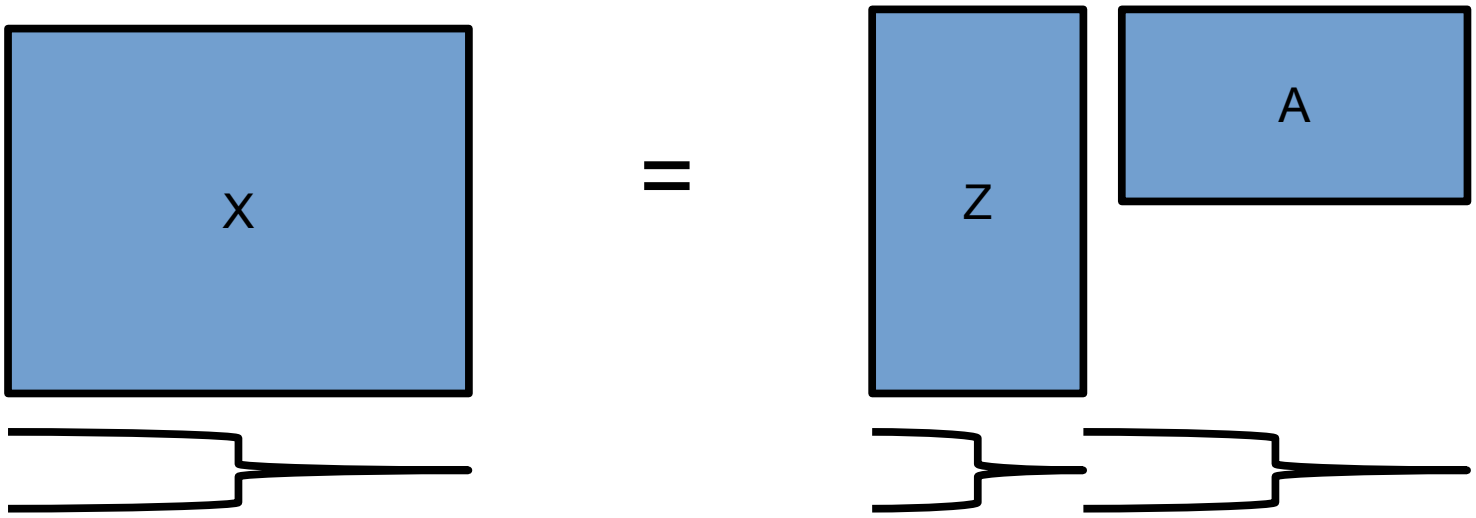
Modeling Effort, Part II (Embedding)

Each **ASD feature** is characterized by a **set of comorbidities**.
Each **patient** has a **combination** of **ASD features**.



Modeling Effort, Part II

Each **ASD feature** is characterized by a **set of comorbidities**.
Each **patient** has a **combination** of **ASD features**.



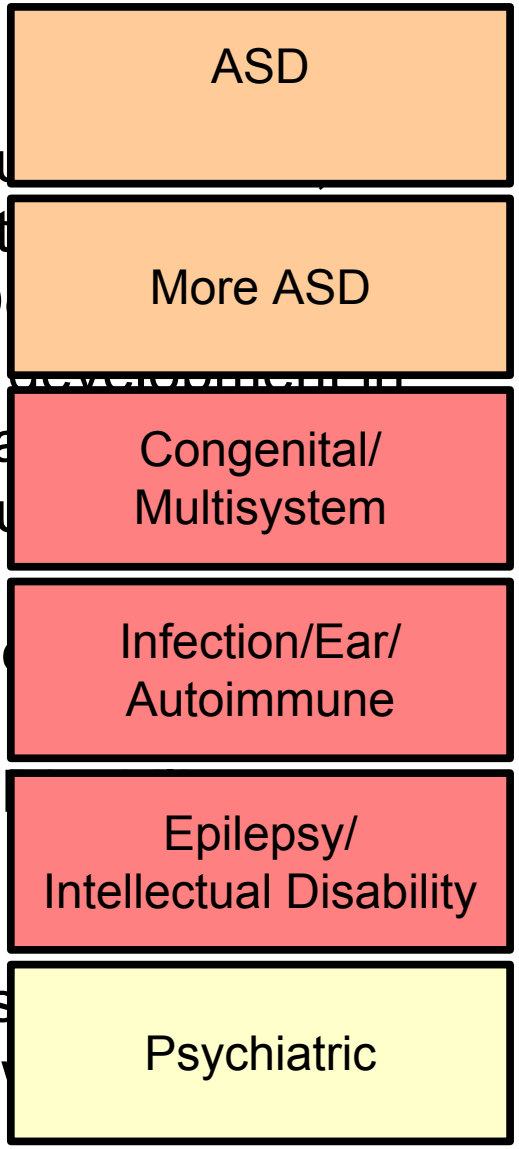
Data: patient by code matrix of counts

Patient factors: how much of feature A does patient n have?

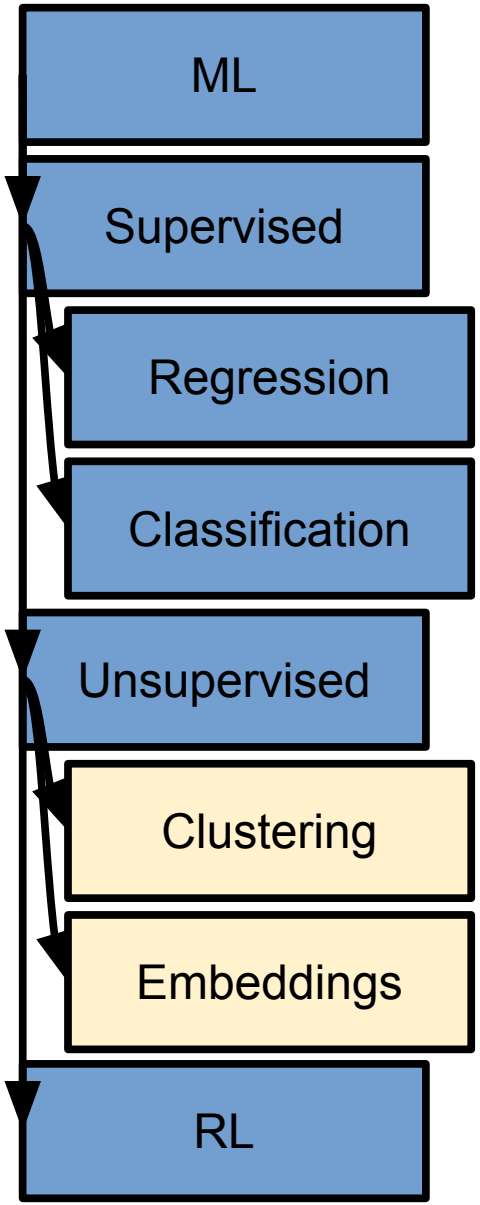
Global features: what codes are in each feature?

Features Discovered

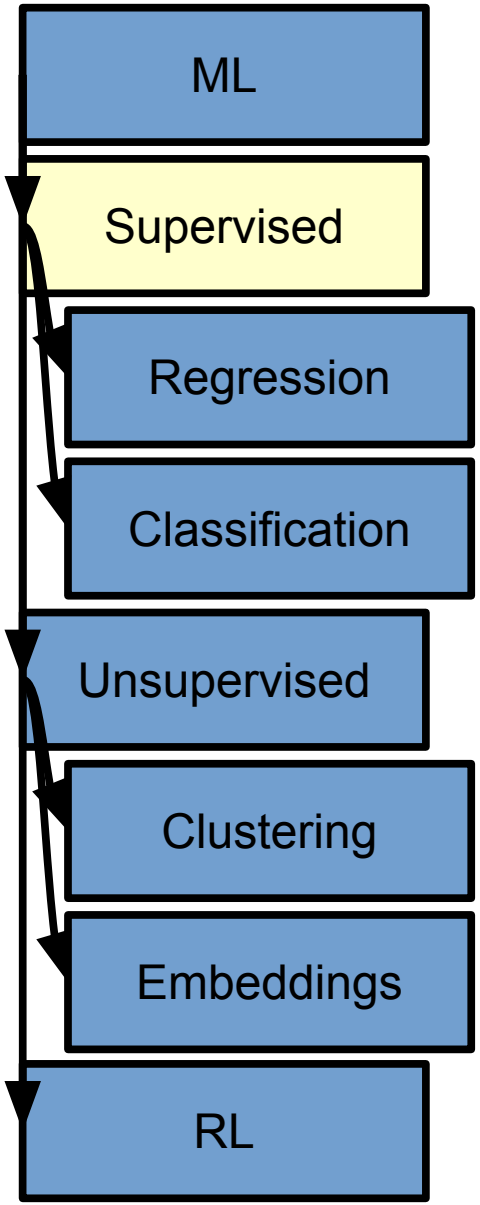
- Autism Spectrum Disorder
- Specific Delays in Development, Autism Spectru
  Symptoms of Head and Neck, Speech Dysfunct
- Conduct Disorder with Specific, Physiological D
  Delays, Lack of expected normal physiological
  childhood, Physiological Delays with Genetic Ca
  Congenital Anomalies, Physiological Delays, Mu
  Dystrophy
- Otitis Media, Hearing Loss, Congenital Anomali
  Viral Infection, Diseases of the Ear, Asthma
- Seizures, Cerebral Palsy, Epilepsy, Intellectual
  Abnormal Movements
- Depressive Disorder, Not Elsewhere Classified,
  Emotions Specific to Childhood, Specific Delays
  Development, Disturbance of Conduct Not Elsew
  Classified, Acute Reaction to Stress

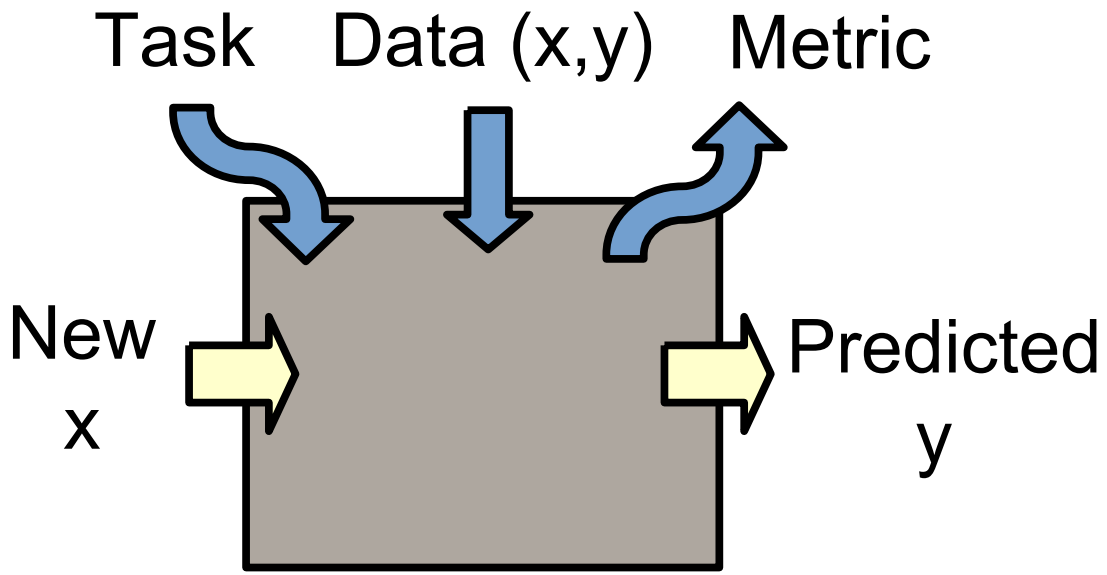


Machine Learning Taxonomy

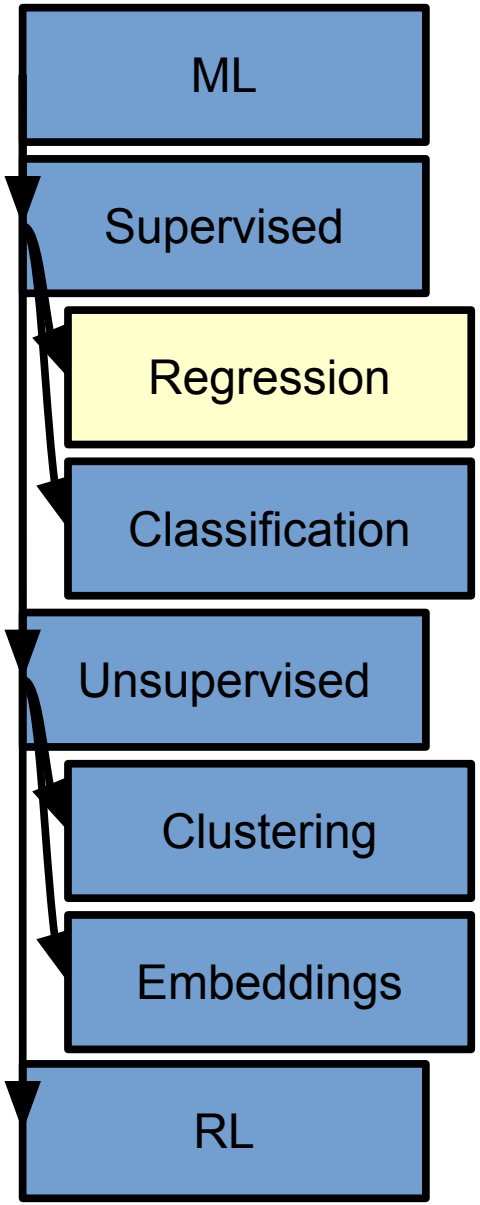


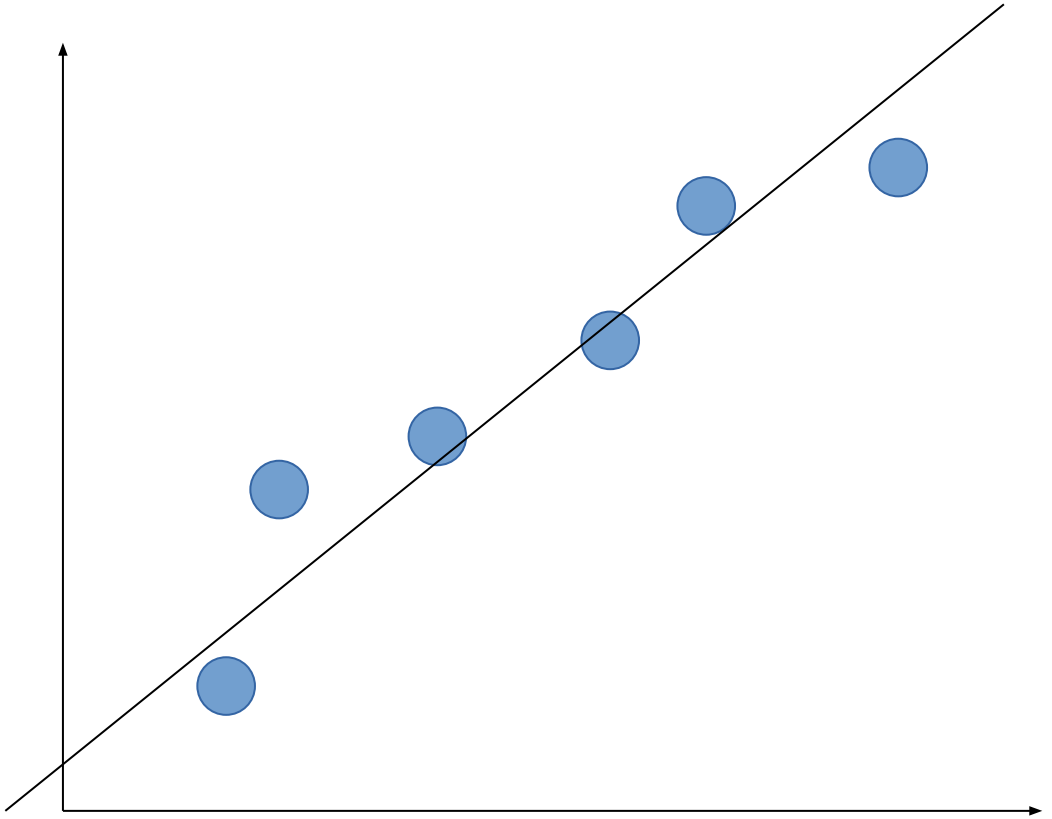
Machine Learning Taxonomy





Terminology: Regression





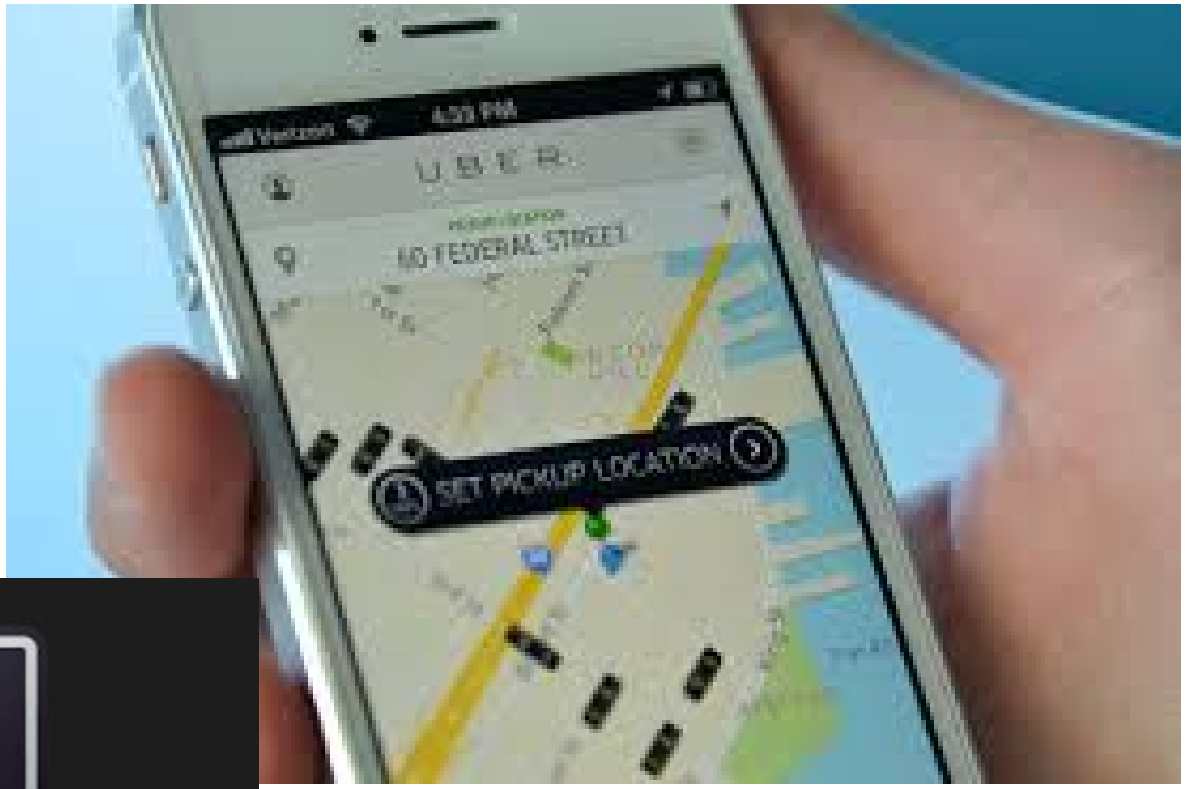
Example: Virtu

VIRTU FINANCIAL

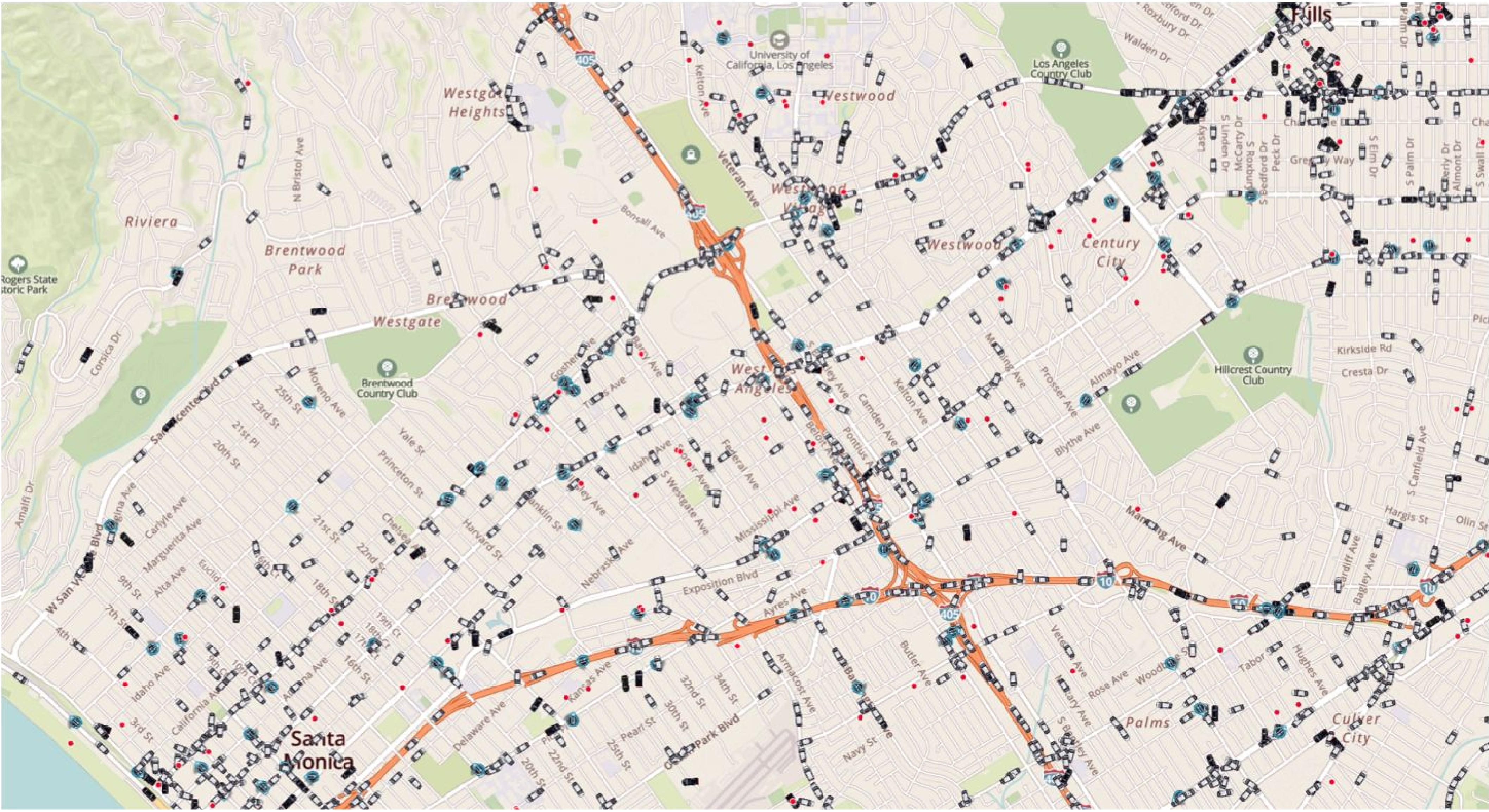
Core technology: Choosing what to trade, and when

Example: Uber

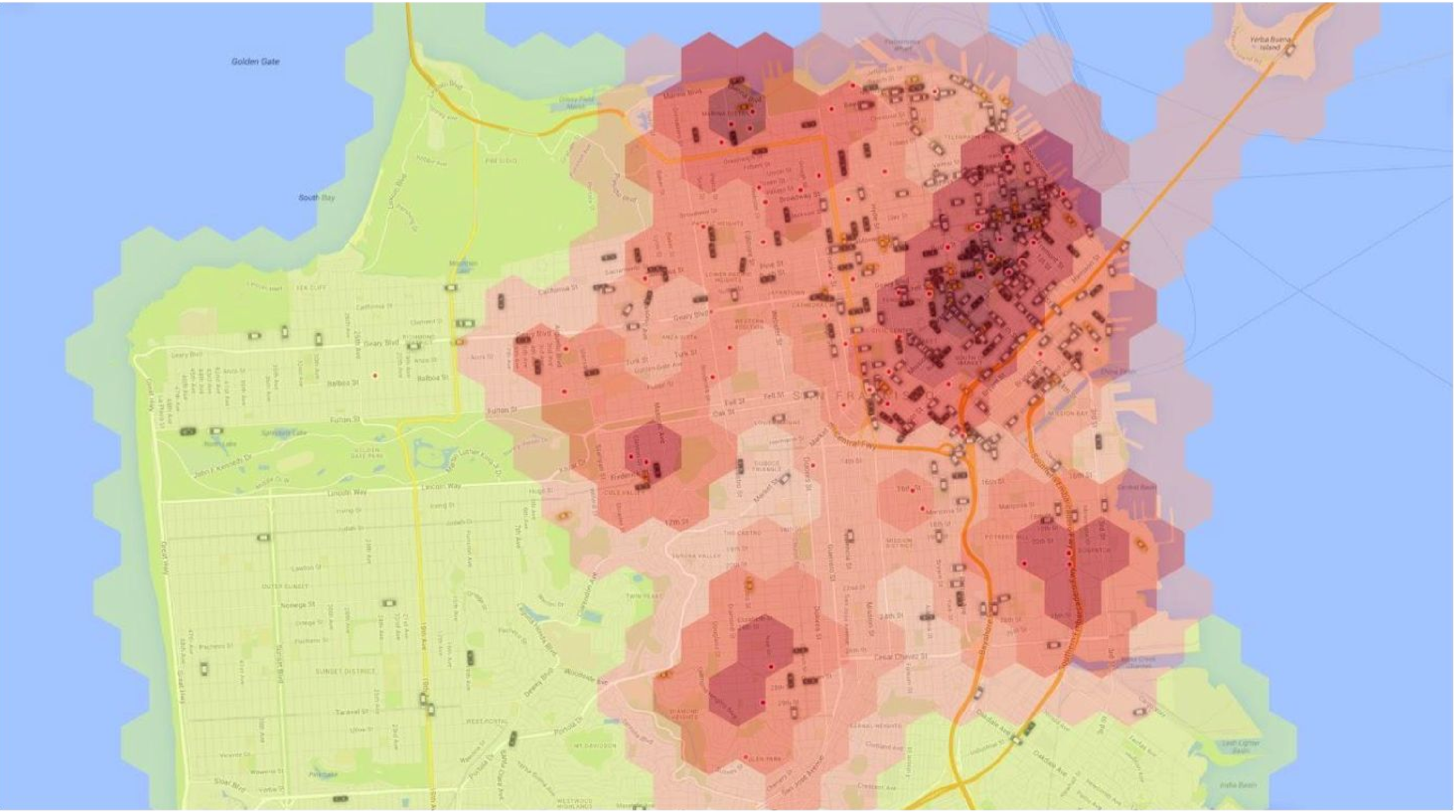
Predictions of travel time, price, supply, demand



UBER

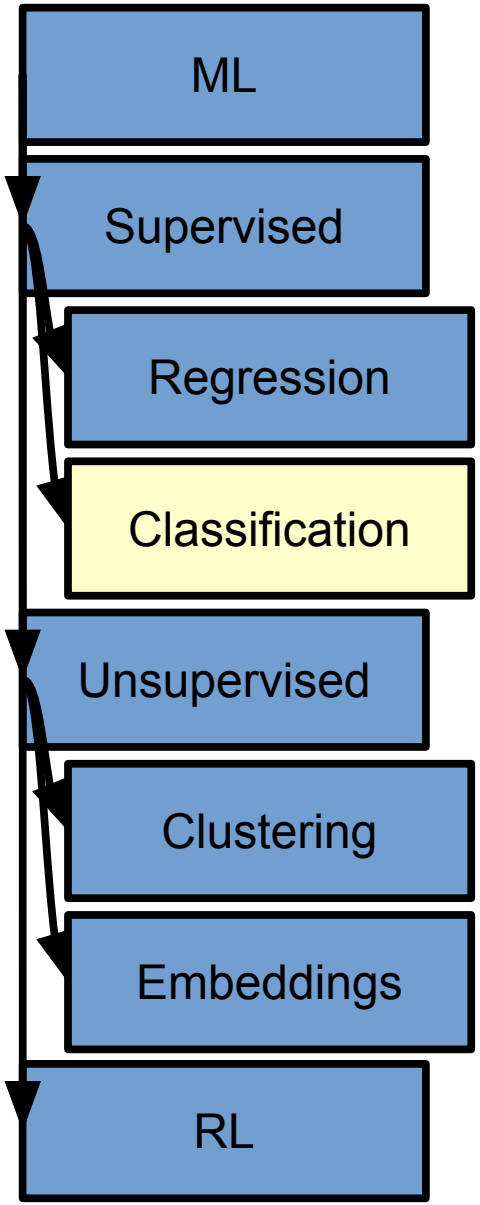

Westgate Heights
Westwood
Los Angeles Country Club
Walden Dr
Riviera
Brentwood Park
Westgate
Brentwood Country Club
N Bristol Ave
Bonsall Ave
Veteran Ave
Kelton Ave
Century City
Hillcrest Country Club
Kirkside Rd
Cresta Dr
Hargis St
Olin St
S Canfield Ave
Manning Ave
Prosser Ave
Almayo Ave
Blythe Ave
Camden Ave
Pontius Ave
Federal Ave
Mississippi Ave
Nebraska Ave
Exposition Blvd
Ayres Ave
Armacost Ave
Butler Ave
Navy St
Rose Ave
Tabor St
Hughes Ave
Palms
Culver City
Pearl St
Delaware Ave
Kansas Ave
Moreno Ave
Yale St
Princeton St
Harvard St
Chelsea Ave
Carlyle Ave
Marguerita Ave
Alta Ave
Euclid St
Idaho Ave
California Ave
Corsica Dr
Amalfi Dr
Cheviot Dr
S Palm Dr
Greenway
405
10

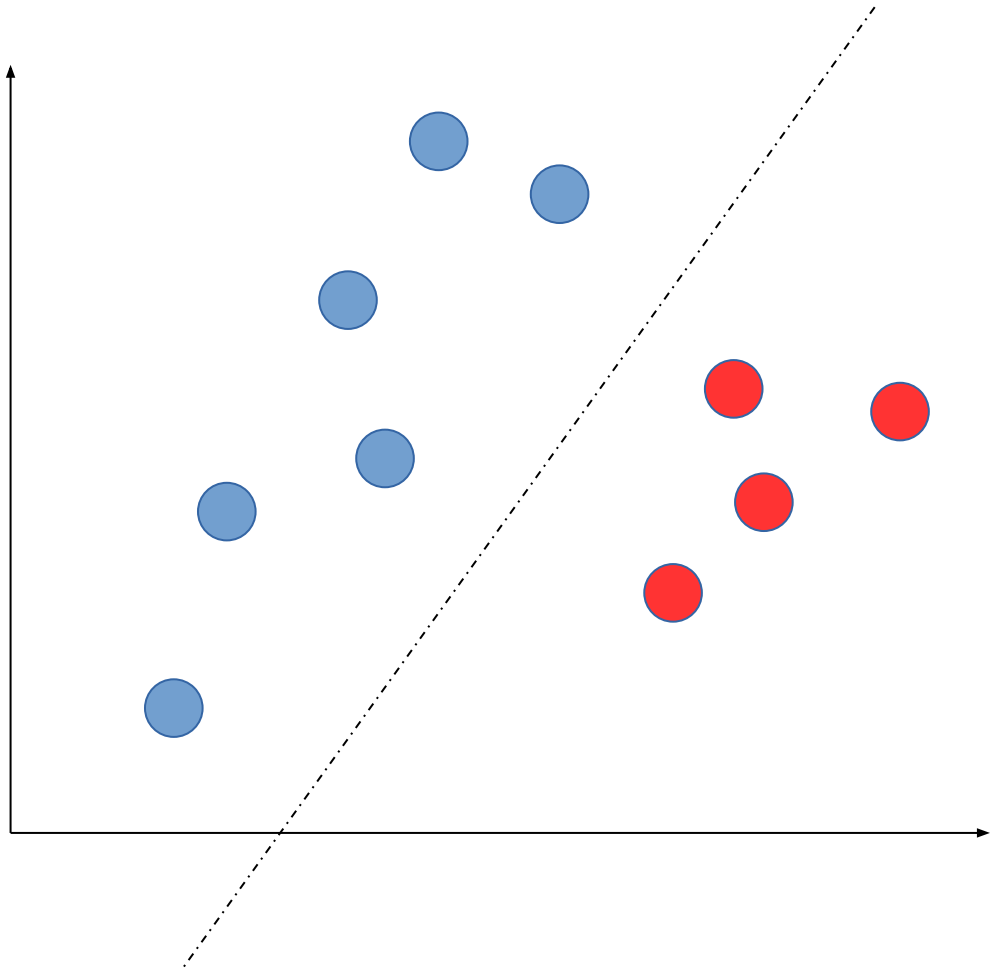
(Keith Chen)



(Keith Chen)

Terminology: Classification

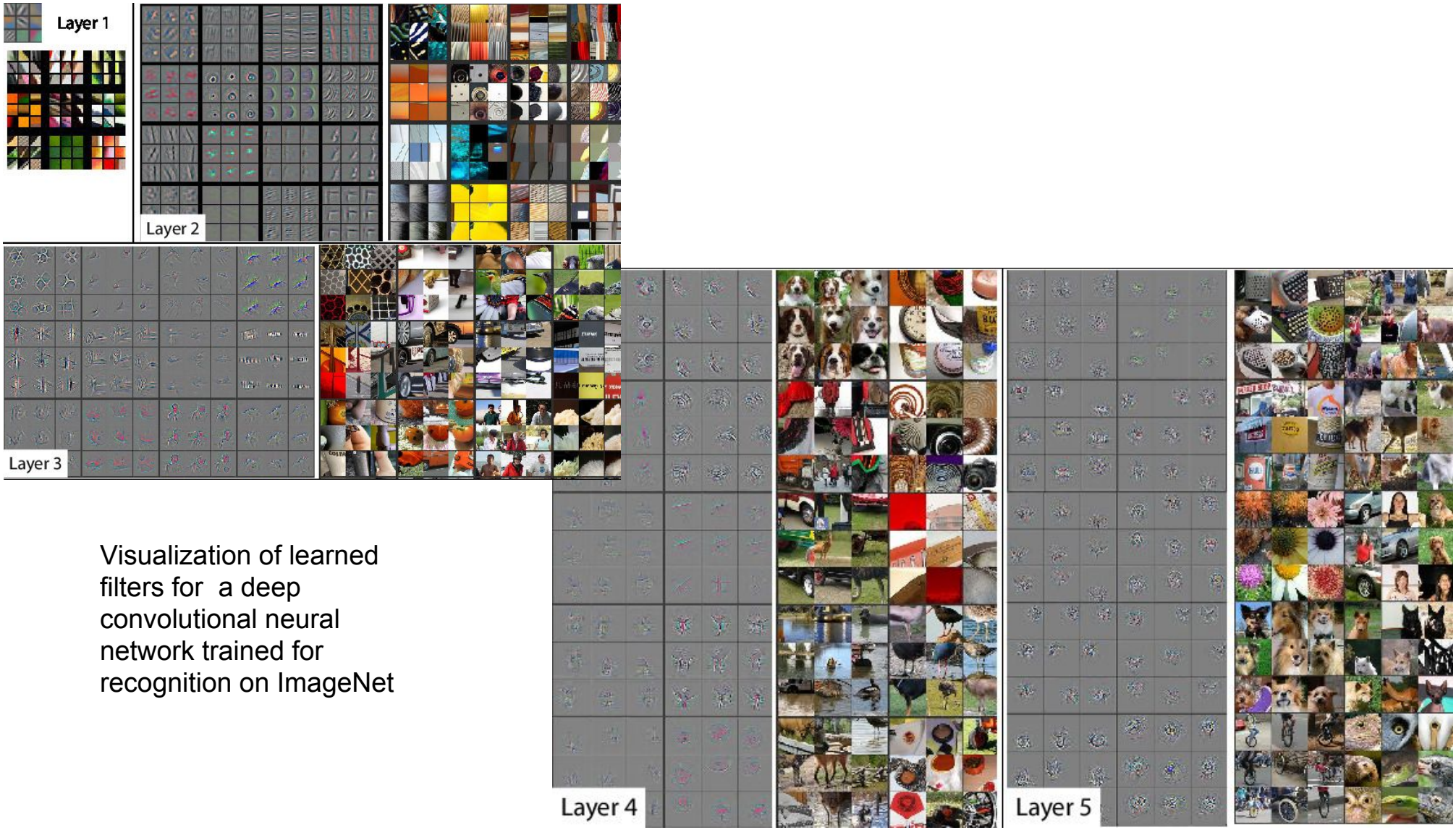




Example: Digit Classification

- Data: Handwritten US zip codes

Example: Image Recognition



Visualization of learned filters for a deep convolutional neural network trained for recognition on ImageNet

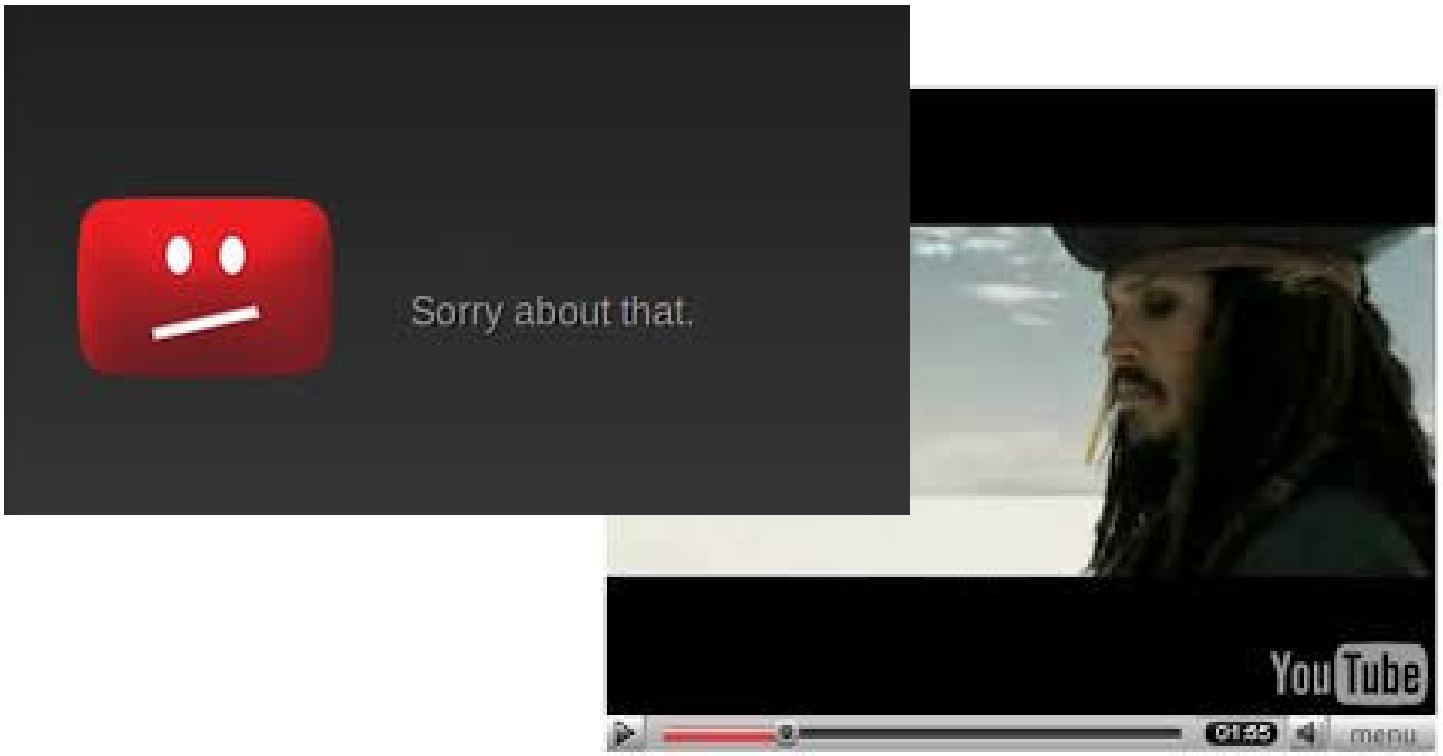
Example: Swype

Predict words from keyboard trajectories

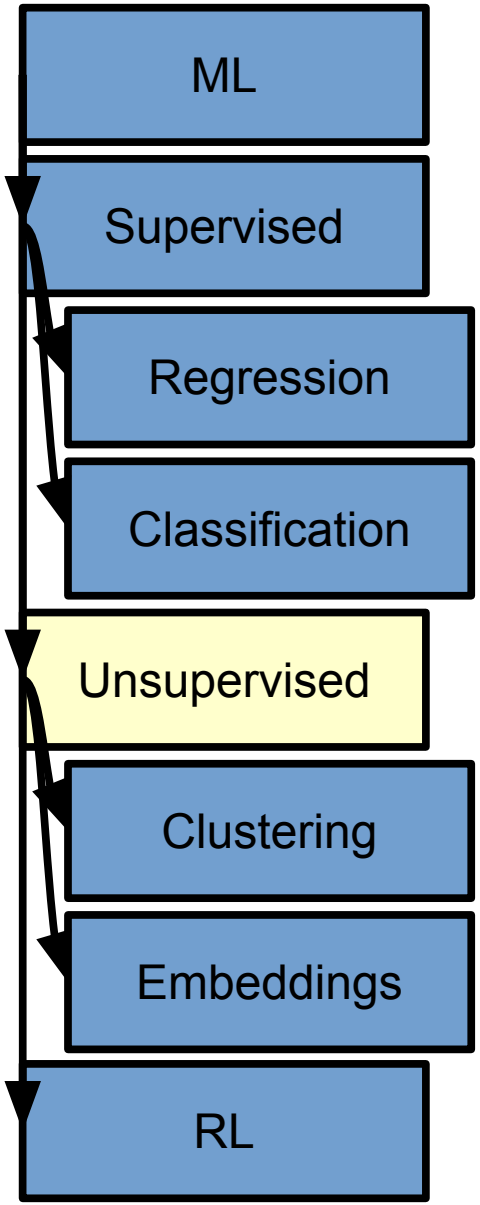
Novel Product:
An easier way
to input text on
mobile devices

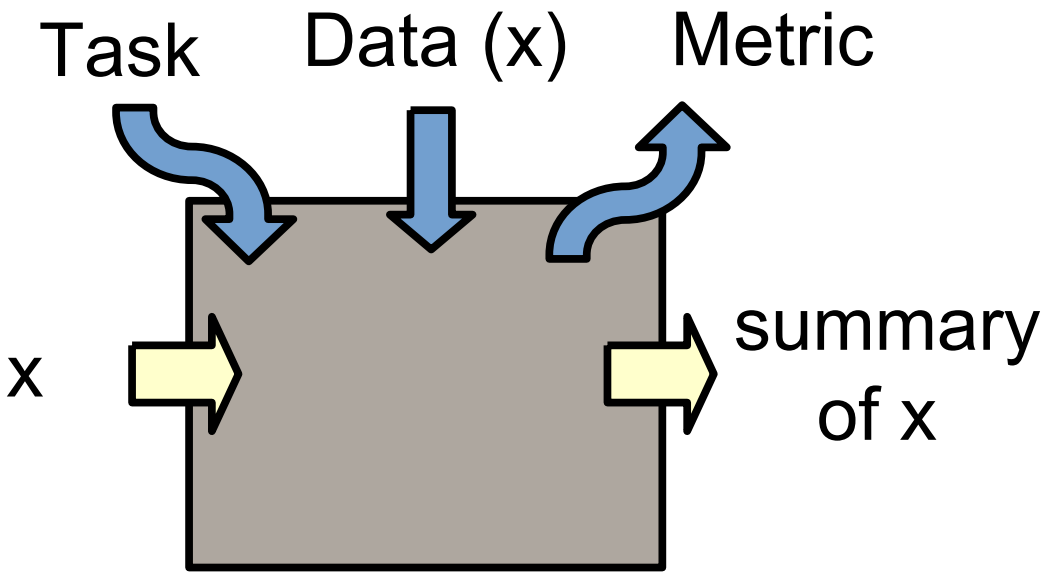
Detecting Copyright Violations on Youtube...

(~300 hours of new video per minute)

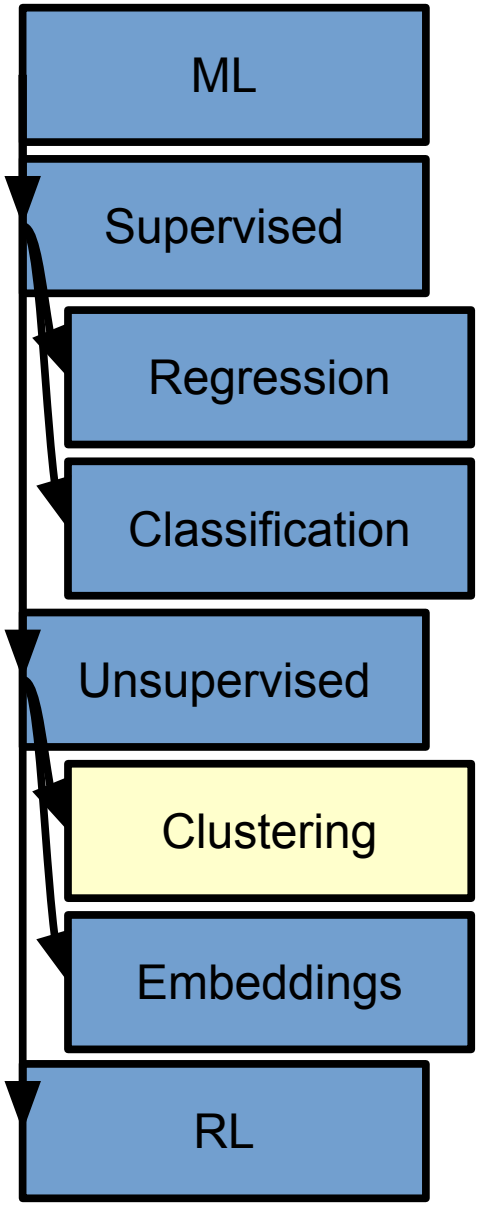


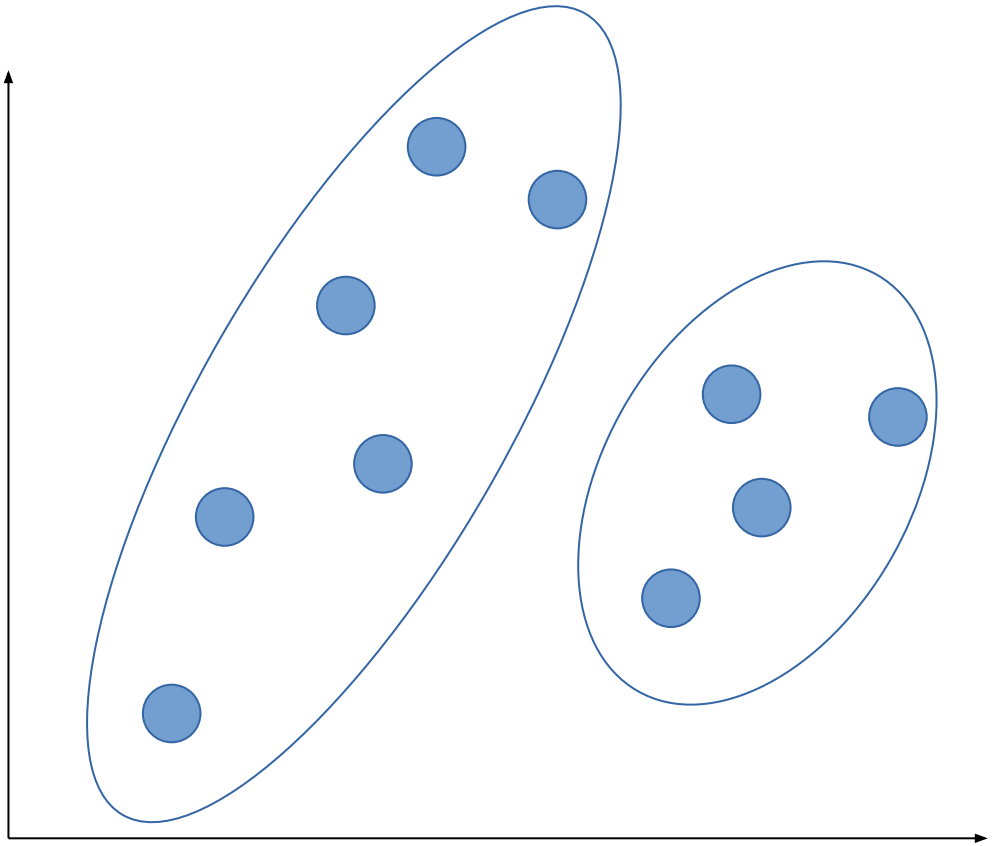
Machine Learning Taxonomy





Terminology: Clustering





Example: News Clustering

Top Stories

Fox News

See realtime coverage

Intensive manhunt underway after daring jail escape in California

Fox News - 37 minutes ago

An intensive manhunt was underway Monday for three inmates who pulled a "Shawshank"-style escape through a hole in their California jail cell -- and, who may have ties to notorious Vietnamese street gangs and Iran.

Manhunt Expands For 'Dangerous' Trio After Daring Jailbreak NBCNews.com

Orange County manhunt: Officials suggest violent jail escapees could be hiding nearby Los Angeles Times

Related
California »

In Depth: Authorities struggling to piece together daring jail escape Washington Post


OCRegister


City News Los A...


WOODTV.com


Los Angel...


Philly.com


OCRe

Huffington Post

See realtime coverage

HUFFPOLLSTER: Trump And Clinton Lead, But Iowa Polling Remains Volatile With A Week To Go

Huffington Post - 5 hours ago

Donald Trump has regained the lead in Iowa but things can still change. On the Democratic side, young voters could tip the caucus toward Bernie Sanders, but only if they turn out.

Here's Bernie Sanders's best closing argument against Hillary Clinton in Iowa Washington Post

Bernie Sanders' One Answer on How He Would Get Anything Done ABC News

Related
Hillary Rodham Clinton »
Bernie Sanders »

Opinion: Democratic Iowa Forum: How to Watch the Live Stream Online Daily Beast

Wikipedia: Statewide opinion polling for the Democratic Party presidential primaries, 2016


CNN


Kansas City Star

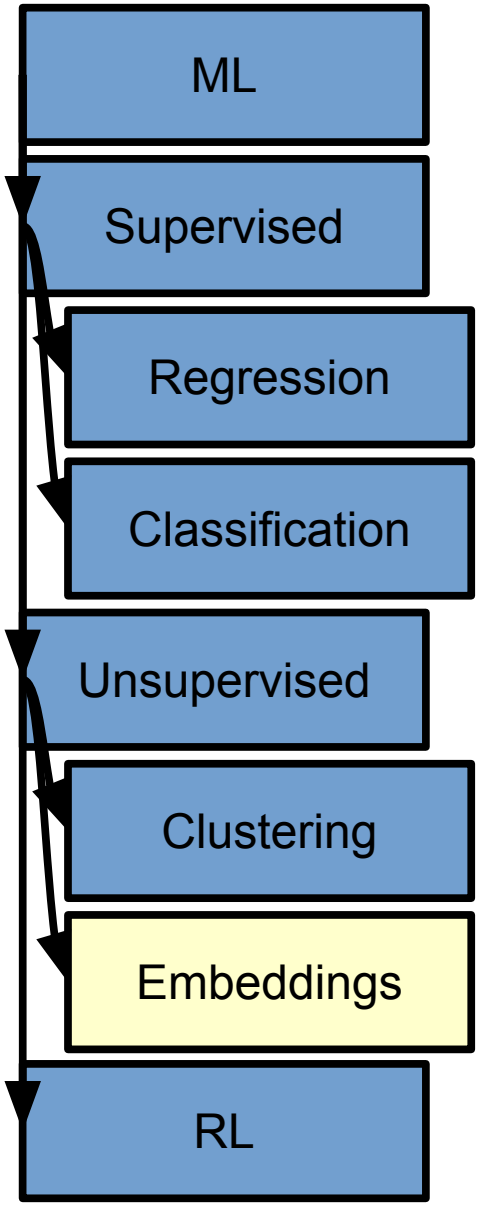

CNN

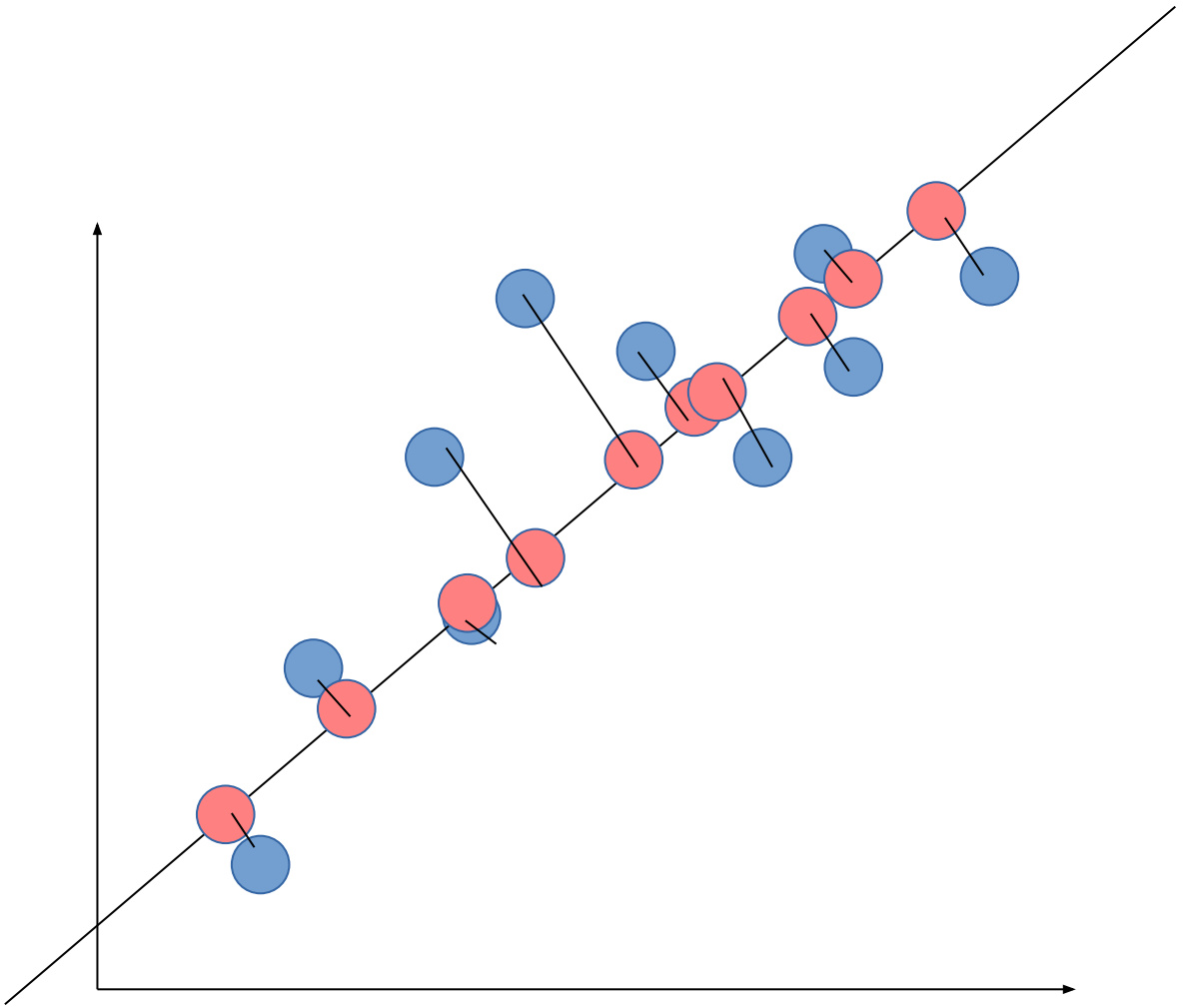

Reuters


Washingto...

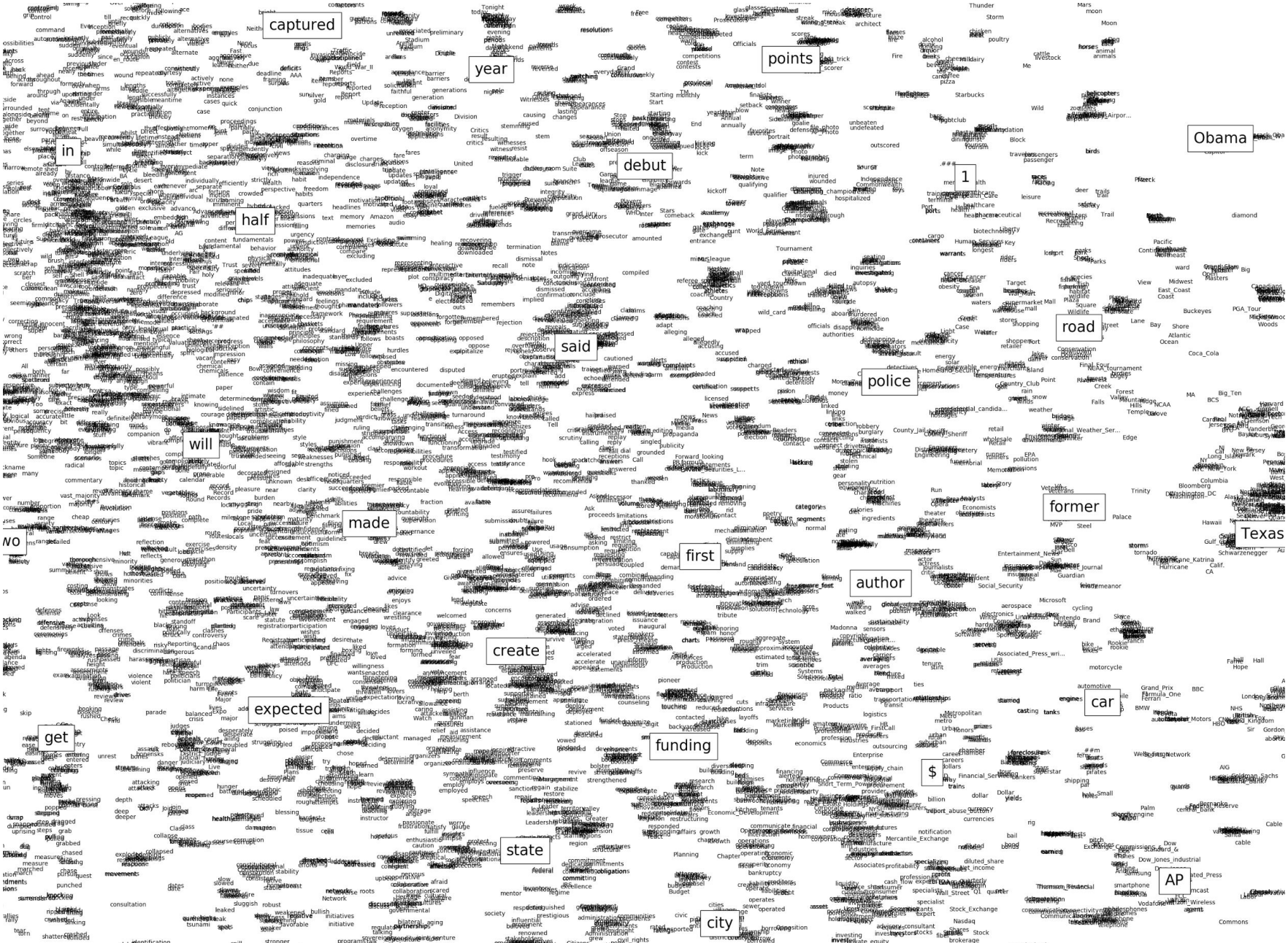

New Yc

Terminology: Embedding

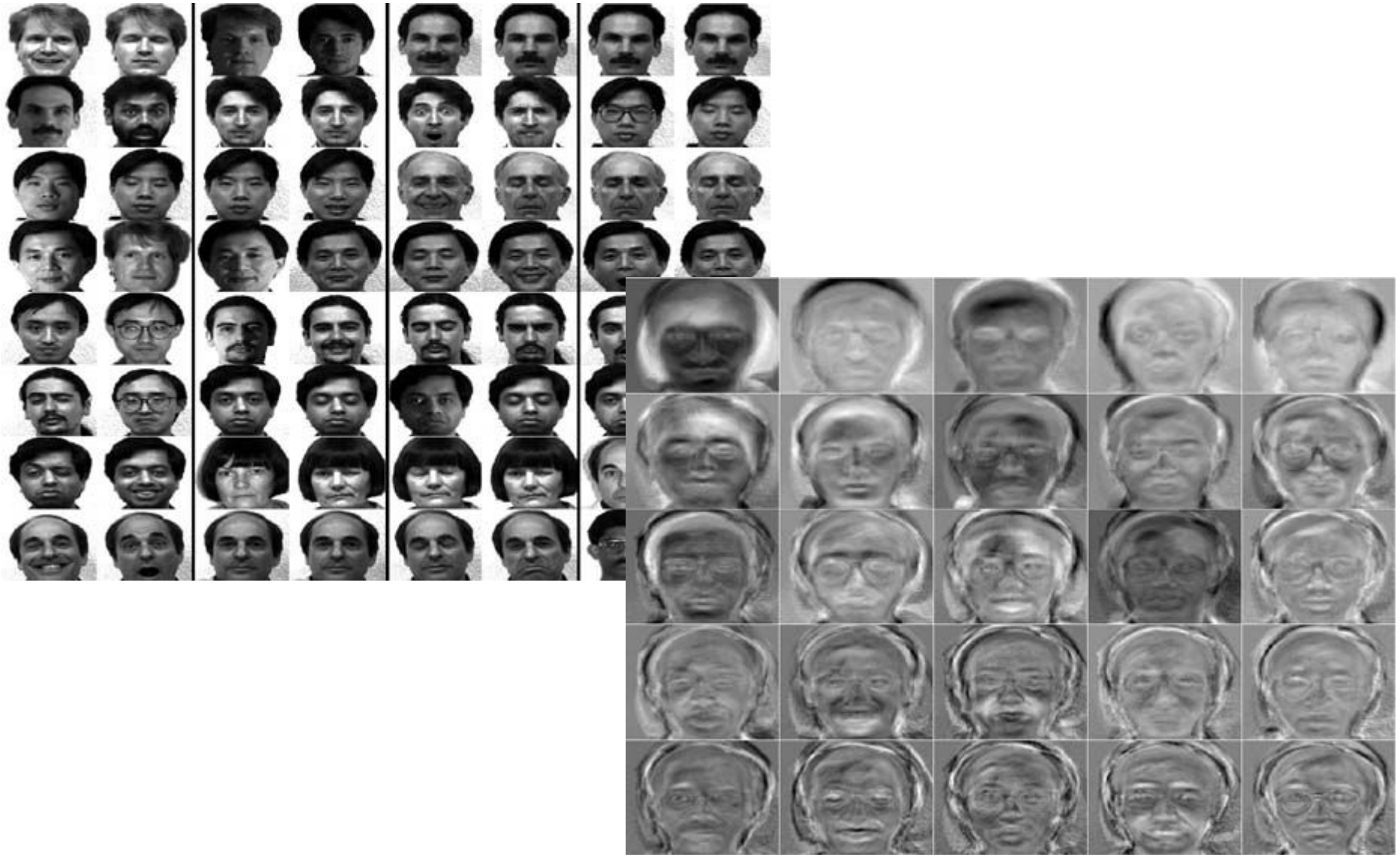




Example: Word Vectors



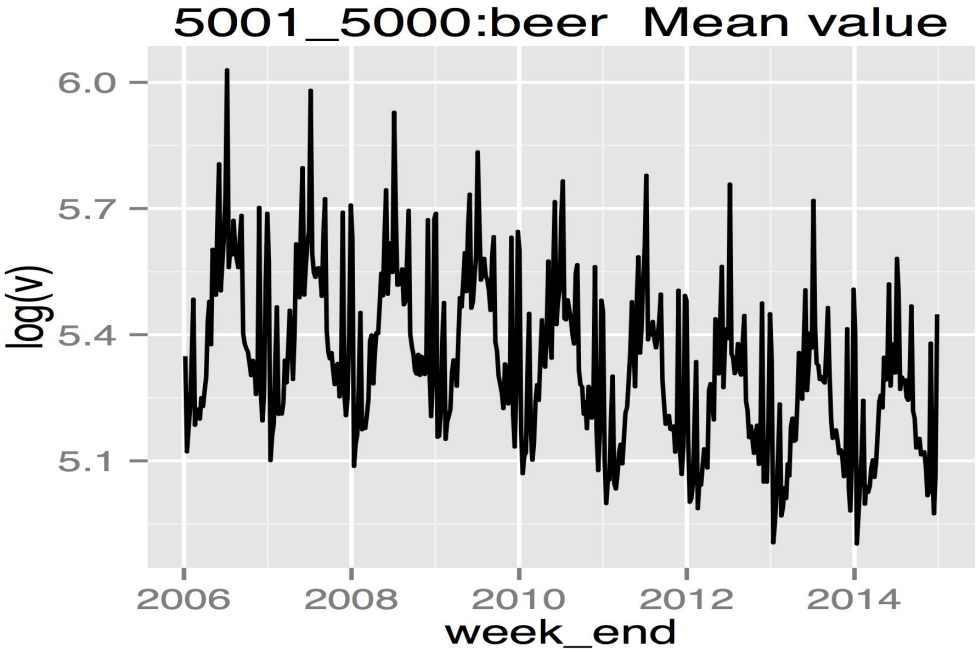
Example: Eigenfaces

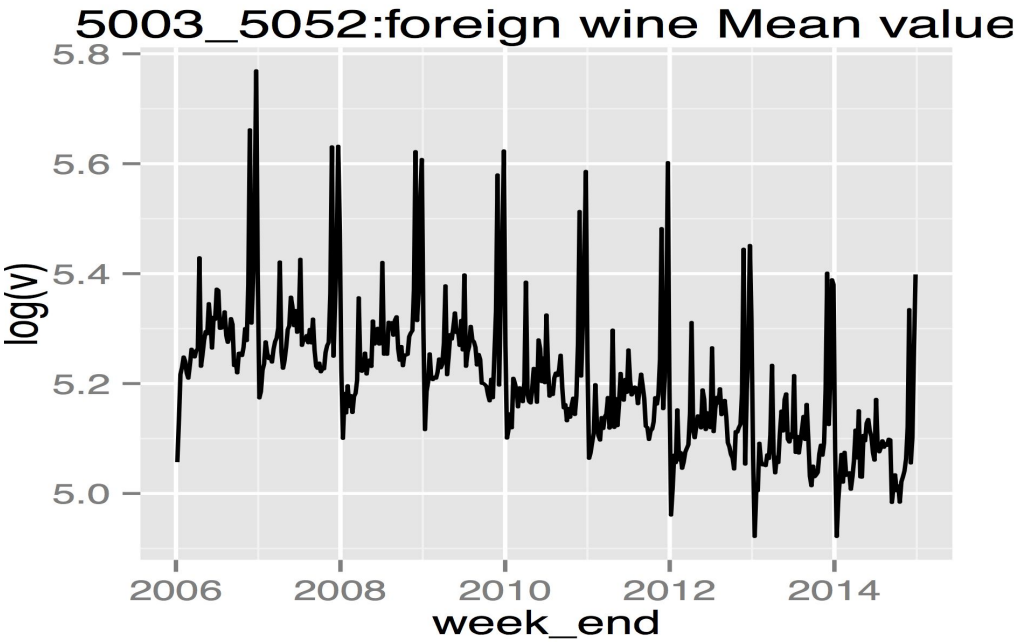


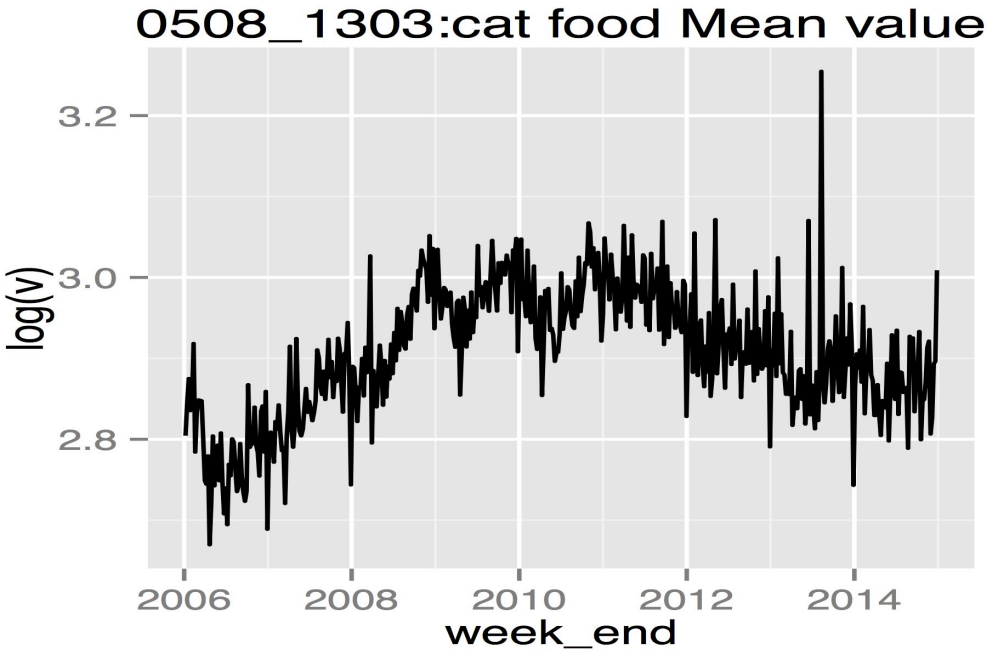
https://www.cs.princeton.edu/~cdecoro/eigenfaces/

Example: Scanner Data

(S. Ng 2016)





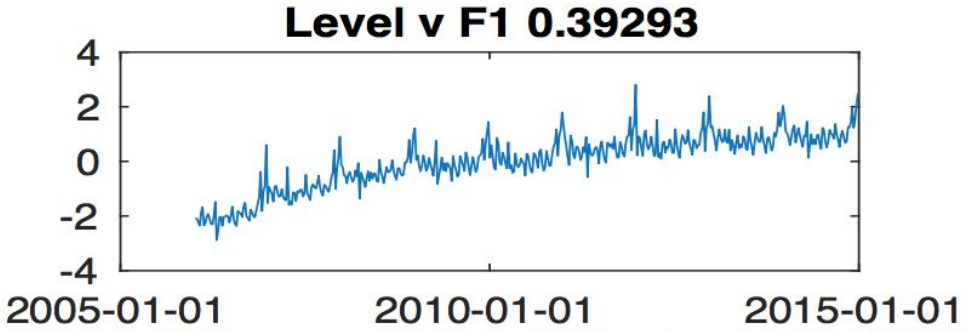


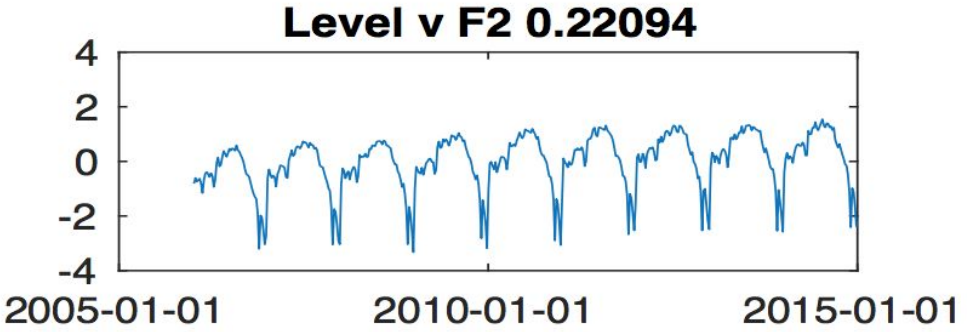
Nielsen scanner data, 2006-2014 on 1073 products in different locations ~6 TB

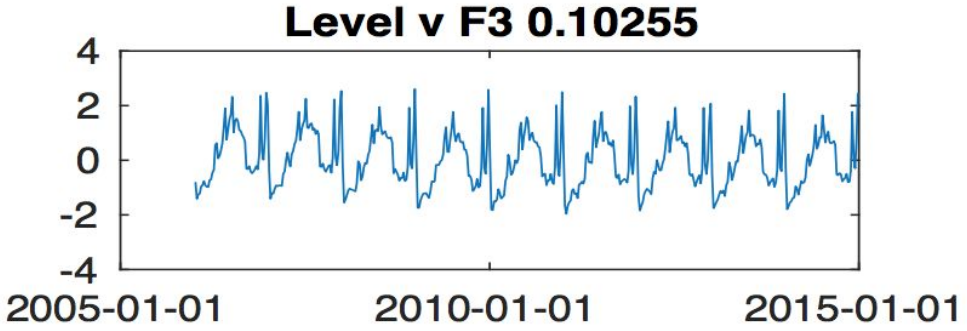
Can we see the effect on demand of the 2008-09 financial crisis?

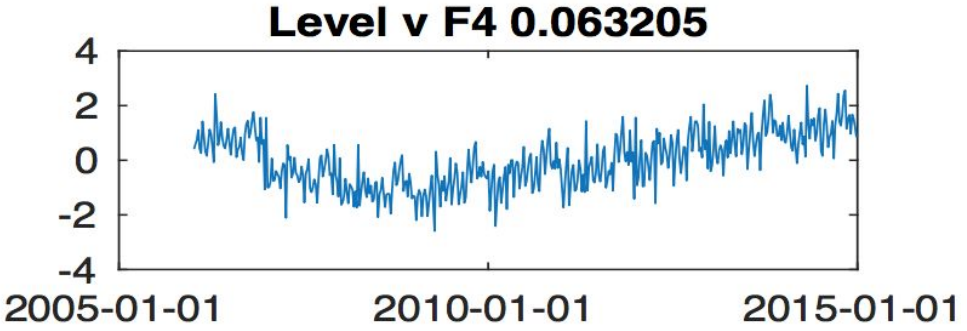
Eight Components (PCA)

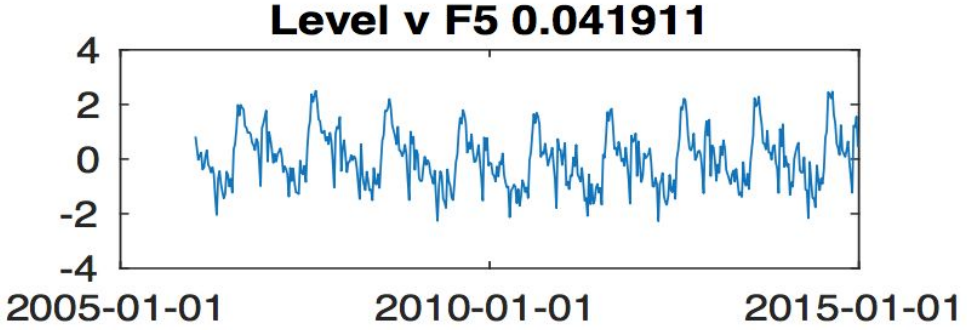
(S. Ng 2016)

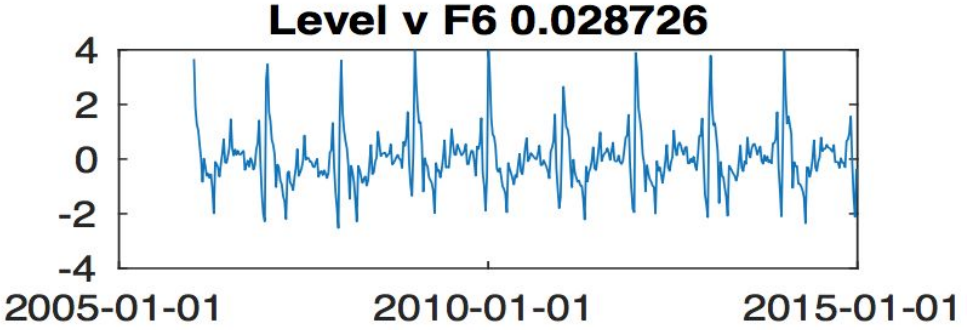


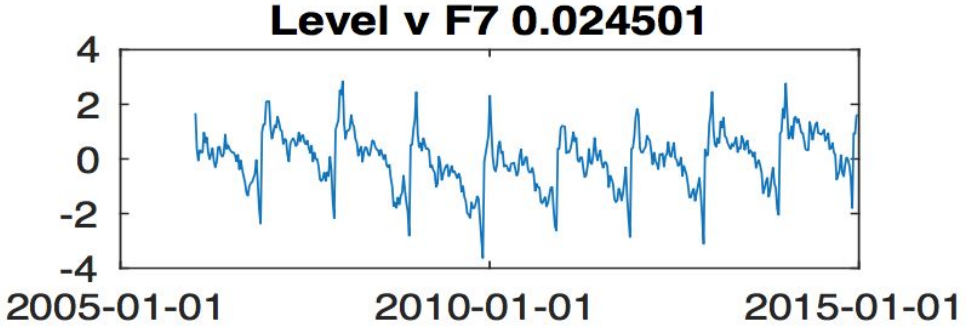


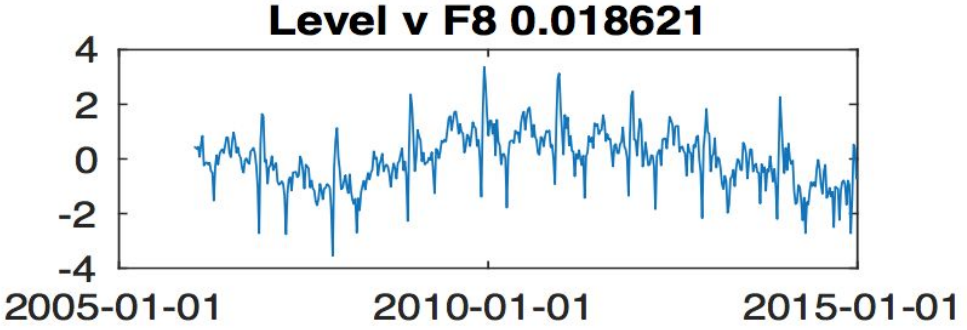




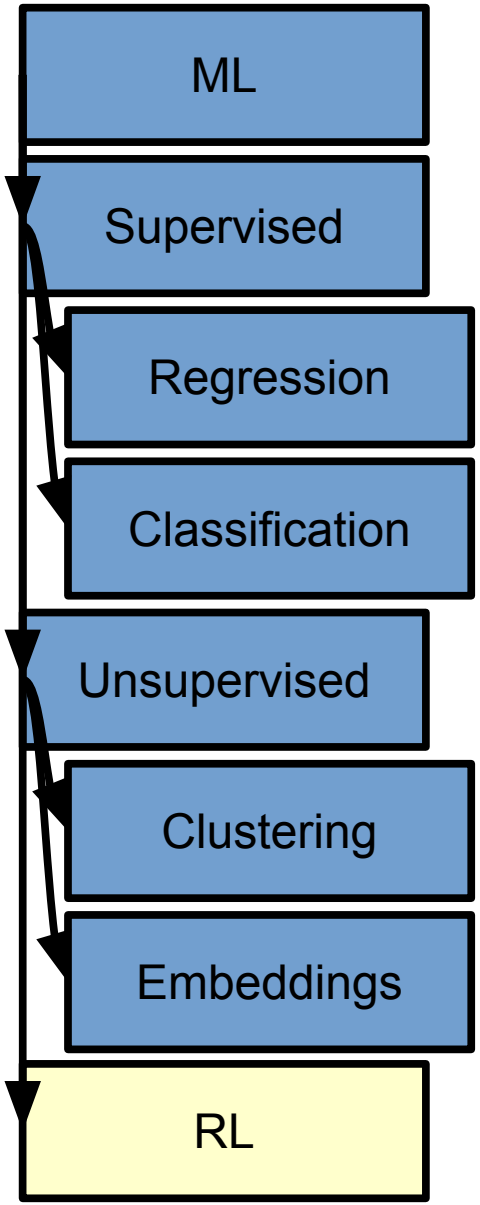


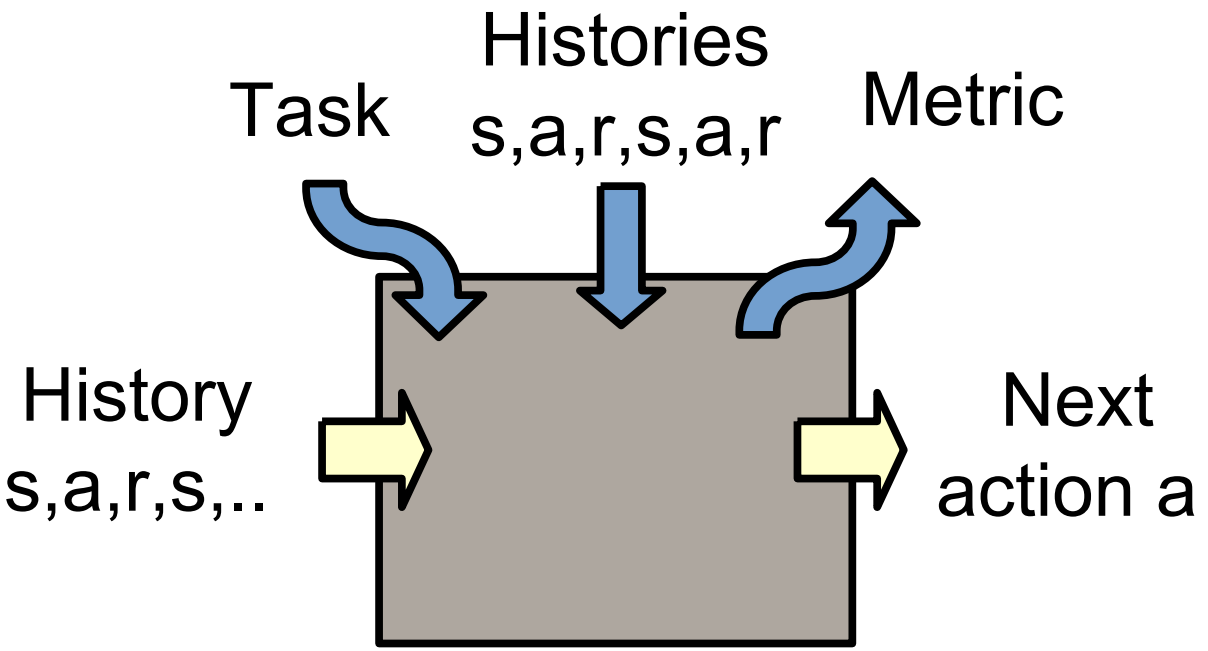




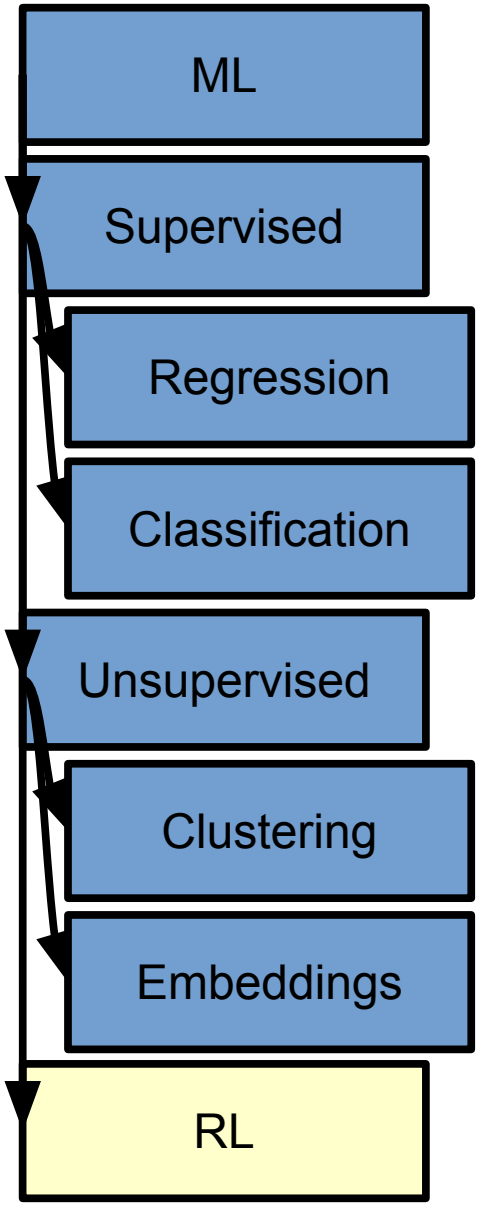


Terminology: Reinforcement Learning



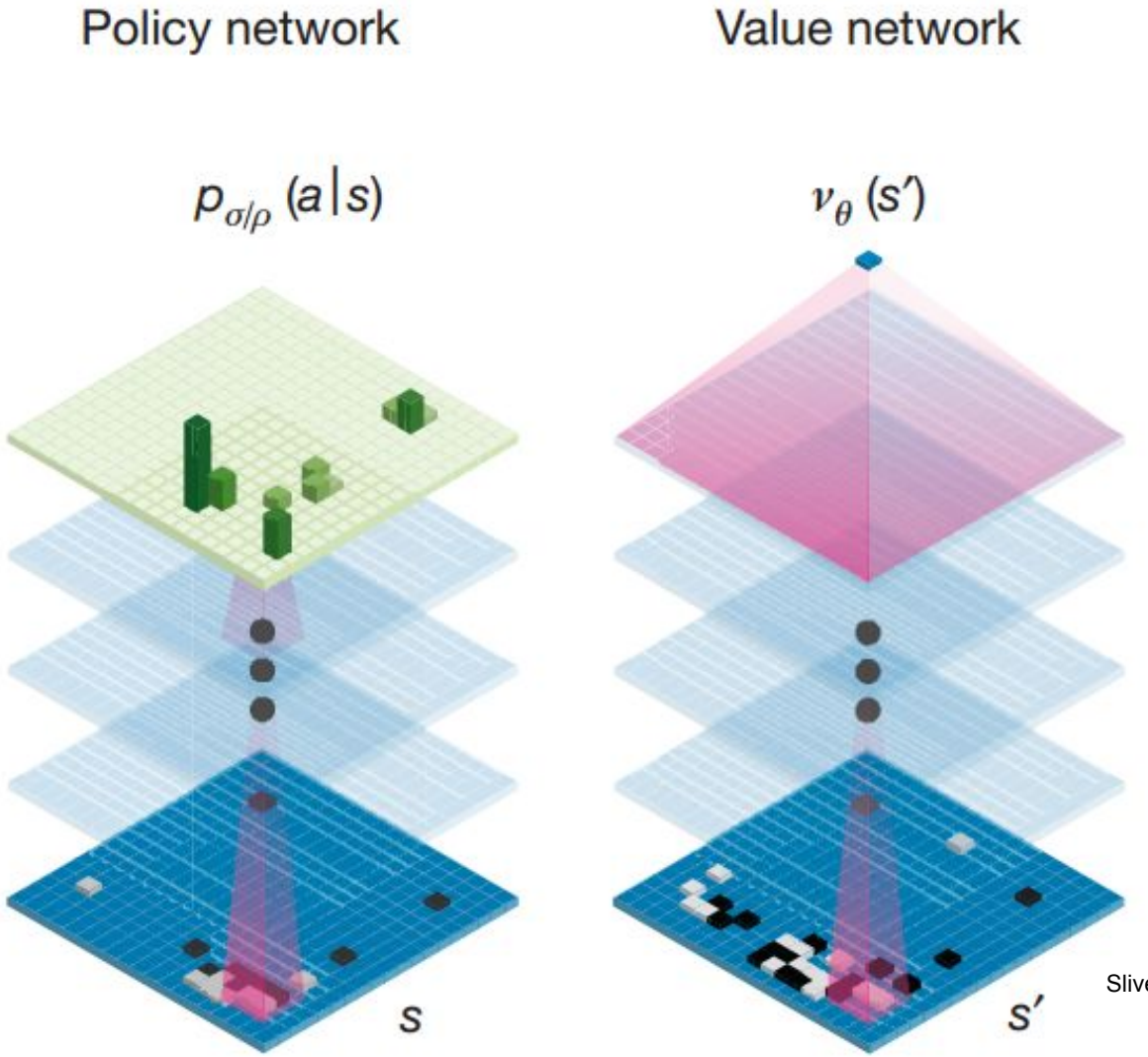


Terminology: Reinforcement Learning



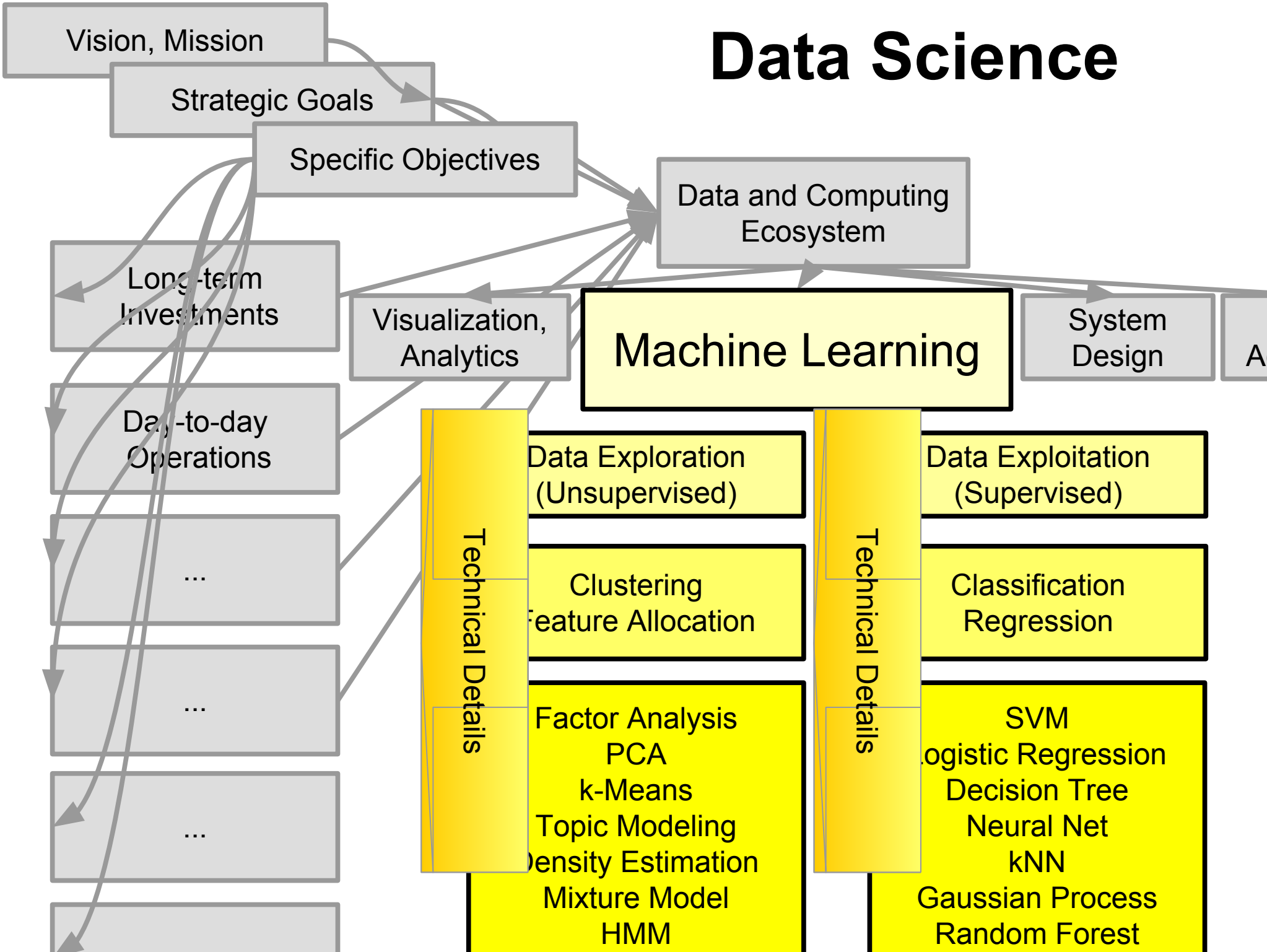
Example: AlphaGo (Hybrid Model)

Combines regression (value network), classification (move prediction), embedding (visualization), reinforcement learning (self-play), and game tree search (AI, CS182)



Sliver et al., *Science* **529** 2016

The Bigger Picture


Data Science
Vision, Mission
Strategic Goals
Specific Objectives
Data and Computing Ecosystem
Long-term Investments
Visualization, Analytics
Machine Learning
System Design
Day-to-day Operations
...
...
...
Technical Details
Data Exploration (Unsupervised)
Clustering
Feature Allocation
Factor Analysis
PCA
k-Means
Topic Modeling
Density Estimation
Mixture Model
HMM
Technical Details
Data Exploitation (Supervised)
Classification
Regression
SVM
Logistic Regression
Decision Tree
Neural Net
kNN
Gaussian Process
Random Forest

Data Science

Vision, Mission

Strategic Goals

Specific Objectives

Data and Computing Ecosystem

Long-term Inve

Visualizatio

System Design

A

Da

Op

ploitation

rvised)

fication

ession

VM

Regression

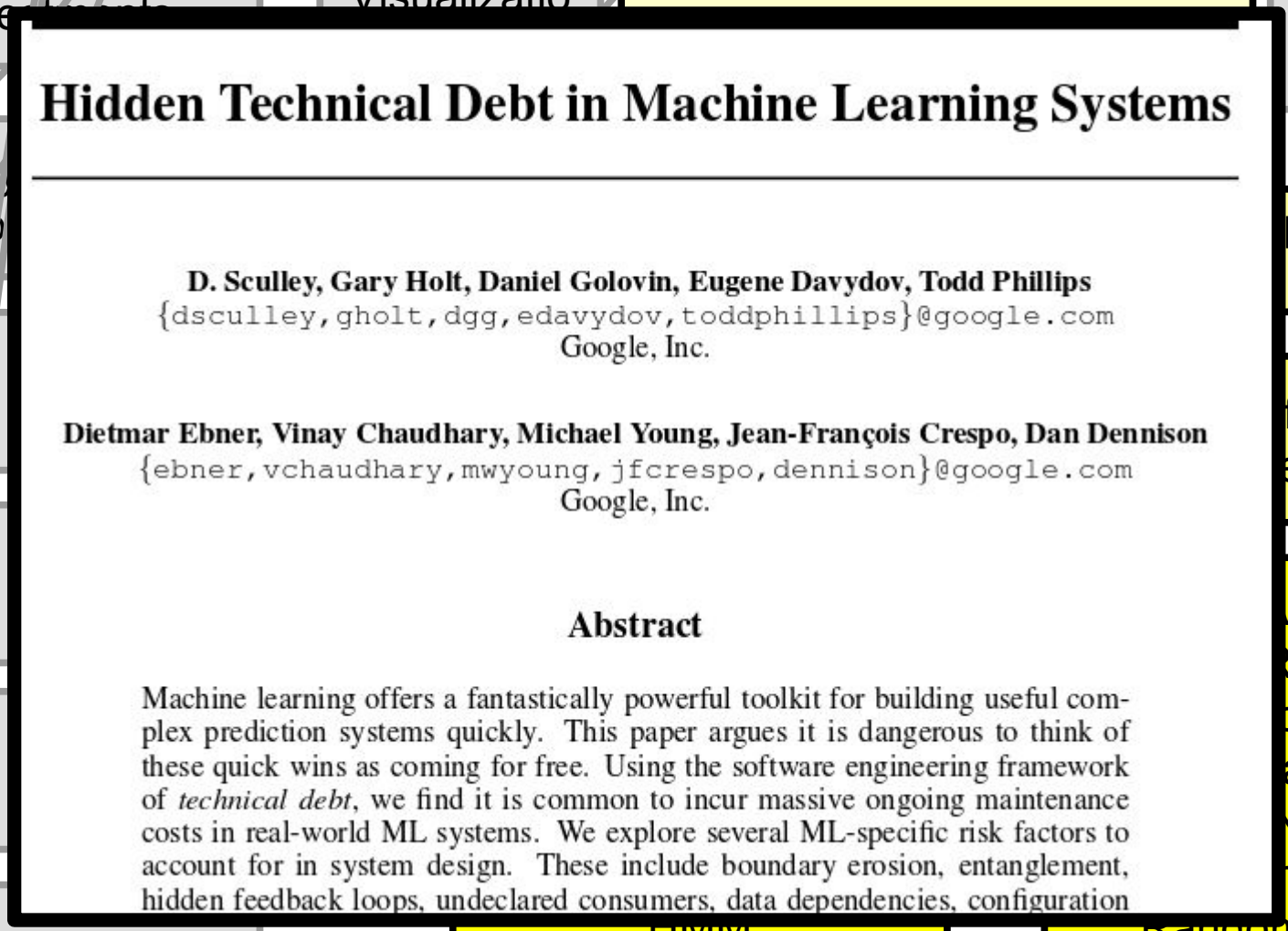
on Tree

al Net

NN

n Process

Random Forest



Hidden Technical Debt in Machine Learning Systems

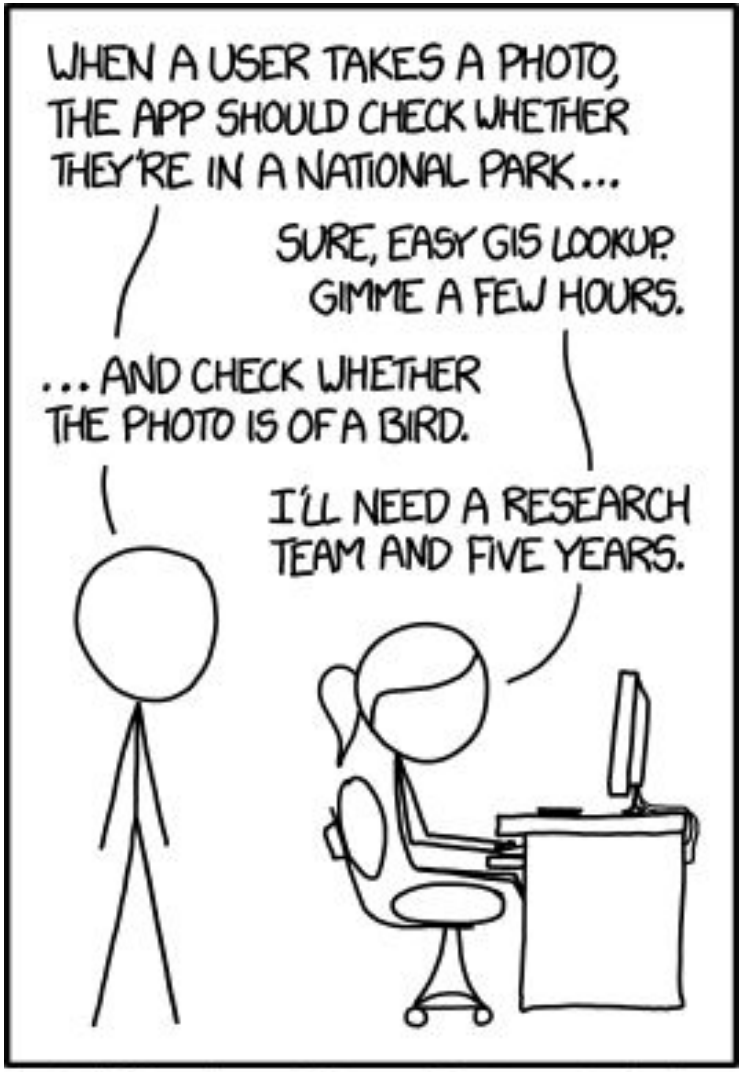
D. Sculley, Gary Holt, Daniel Golovin, Eugene Davydov, Todd Phillips
{dsculley,gholt,dgg,edavydov,toddphillips}@google.com
Google, Inc.

Dietmar Ebner, Vinay Chaudhary, Michael Young, Jean-François Crespo, Dan Dennison
{ebner,vchaudhary,mwyoung,jfcrespo,dennison}@google.com
Google, Inc.

Abstract

Machine learning offers a fantastically powerful toolkit for building useful complex prediction systems quickly. This paper argues it is dangerous to think of these quick wins as coming for free. Using the software engineering framework of *technical debt*, we find it is common to incur massive ongoing maintenance costs in real-world ML systems. We explore several ML-specific risk factors to account for in system design. These include boundary erosion, entanglement, hidden feedback loops, undeclared consumers, data dependencies, configuration

Small changes can make a problem a lot harder

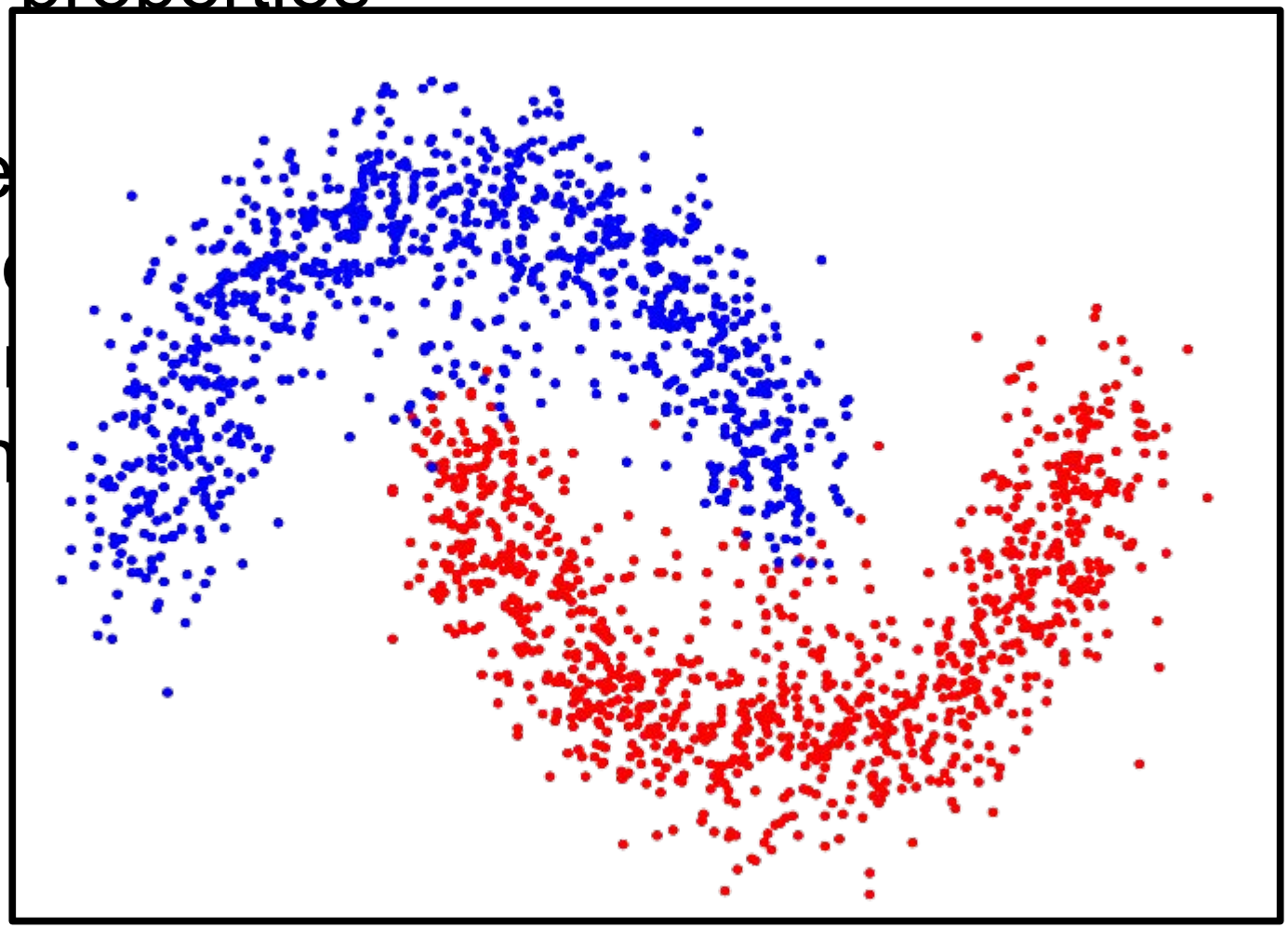


Hidden Assumptions

- Parameters, hyperparameters
- Distributional properties
- Cluster sizes and proportions
- Number of iterations
- Presence of local optima
- Approximations to distributions
- Conditional independence, Markov property

Hidden Assumptions

- Parameters, hyperparameters
- Distributional properties
- Cluster sizes
- Number of ite
- Presence of l
- Approximatio
- Conditional in



Solution: Evaluate Carefully

- What data should be collected?
- How should it be processed?
- How should we separate a test set?
- What algorithms and parameters are used?
- How do we define success?

Fancy algorithms are no substitute for good science!!

CS 181 Overview

Technical Resources

- Website/Schedule: cs181.fas.harvard.edu
- Piazza: Questions, clarifications
- Kaggle: Practical competitions
- Github: Code distribution
- Canvas: Gradebook only

Assessment

- Theory Homeworks (5): concepts, self-graded
- Practicals (4): compete in teams, Kaggle + write
- Midterms (2)
- 3 late days
- Grading:
  - Four Practicals [P] (30%)

    Five Homeworks [T] (30%)

    Midterm Exam 1 (20%)

    Midterm Exam 2 (20%)

Prerequisites

- CS 50-51: For programming competency
- Math 21ab: For lin. algebra and multivar. calc
- Stat 110: For disc. and continuous distributions
- Familiarity with Python and LaTeX

Disclaimer: these are the fundamentals, class will go beyond these basics.

Theoretical Homeworks

- Two sections:
  - Mathematical Problems (in LaTeX)
    - Prob, linalg, optimization, derivations
  - Implementations (in Python)
    - Data processing, plots, experiments
- Individually completed
- Self-graded using solution sets

Practicals

- Team-based ML on real problems
- Primarily using Python / Scikit-Learn

- Practical 1: Chemistry Photovoltaics
- Practical 2: Malware Detection
- Practical 3: Music Clustering
- Practical 4: RL Game Playing

Teaching Staff

- Shai Szulanski
- Jeffrey Ling
- Samuel Cheng
- Ankit Gupta
- Aidi Zhang
- Lily Zhang
- Frances Ding
- Mark Goldstein
- Charles Liu
- Jeffrey Chang
- Fanney Zhu
- Christine Hwang
- Carl Denton

Office Hours and Sections

Office Hours

- Tue 8-10pm: Quincy DH
- Wed 6-8pm: MD 119
- Thu 8-10pm: Currier DH
- Thu 8-10pm: Eliot DH
- Fri 10-Noon: MD First Floor Lounge

Parkes: Th 2.30-3.30, 5.15-5.45 (+ *right after class this week*)
Rush: Wed 2.30-4

Sections: math review, then flipped classroom. Optional:

- Mon, Wed 4-5,5-6p: MD 119
- Section 0 for prob/linear algebra on Fri Jan 27 (time TBD)

Learning to Play Atari

Next Class

Linear Regression