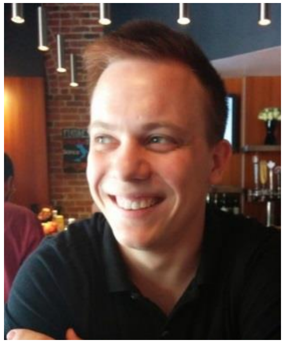
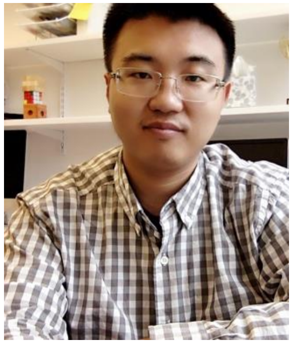
From Worst-Case to Realistic-Case Analysis for Large Scale Machine Learning Algorithms

Maria-Florina Balcan, PI
Avrim Blum, Co-PI
Tom M Mitchell, Co-PI

Students

Travis Dick
Nika Haghtalab
Hongyang Zhang





Motivation

- Machine learning increasingly in use everywhere

Google amazon UBER PayPal

- Significant advances in theory and application

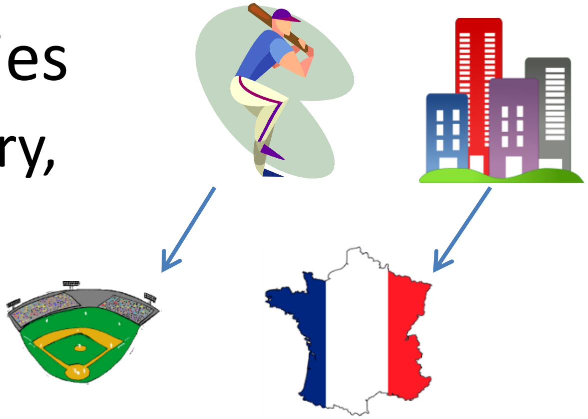
- Yet large gap between the two
  - Practical success on theoretically-intractable problems

    “it may work in practice but it will never work in theory”?

  - Theory focused on learning single targets. Large-scale systems aim to learn many tasks, and to use synergies among them to learn faster and better

Example: NELL system [Mitchell et al.]
(Never-Ending Language Learner)

- Learns many (thousands) of categories
  - river, city, athlete, sports team, country, attraction,…
- And relations
  - athletePlaysSport, cityInCountry, drugHasSideEffect,…
- From mostly unlabeled data (reading the web)



- ➢ ford makes the automobile escape
- ➢ camden_yards is the home venue for the sports team baltimore_orioles
- ➢ christopher_nolan directed the movie inception

High level goals: address the gaps

- Machine learning increasingly in use everywhere

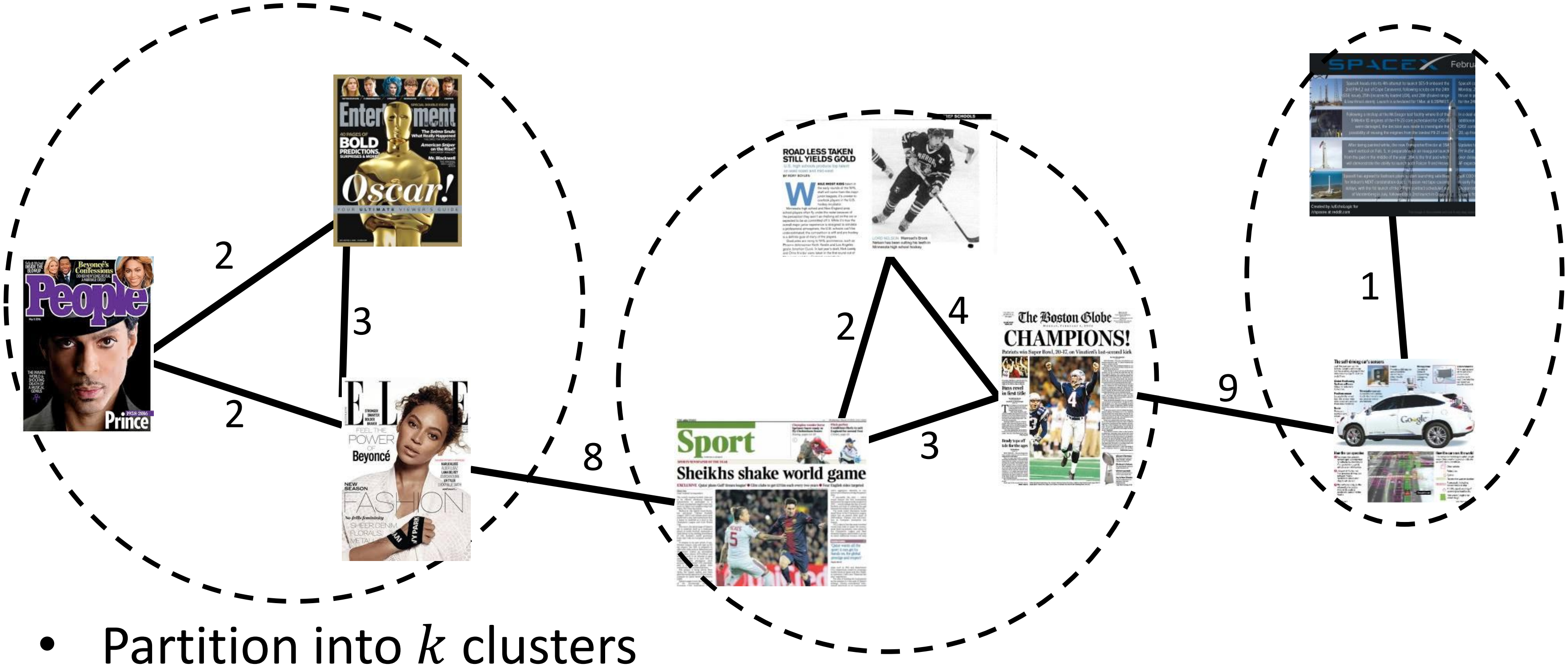
- Significant advances in theory and application

- Yet large gap between the two
  - Practical success on theoretically-intractable problems
  - Theory focused on learning single targets. Large-scale systems aim to learn many tasks, and to use synergies among them to learn faster and better

Clustering

Core problem in making sense of data, including in NELL

Given a set of elements, with distances



- Partition into k clusters
- Minimize distances within each cluster
- Objective function: k-means, k-median, k-center

Maria-Florina Balcan, Nika Haghtalab, and Colin White. *k-Center Clustering under Perturbation Resilience*. Int. Colloquium on Automata, Languages, and Programming (ICALP), 2016.

k-Center Clustering

Minimize maximum radius of each cluster
Known theoretical results:

- NP-hard
- 2-approx for symmetric distances, tight [Gonzalez 1985]
- $O(\log^* n)$-approx for asymmetric distances [Vishwanathan 1996]
- $\Omega(\log^* n)$-hardness for asymmetric [Chuzhoy et al. 2005]

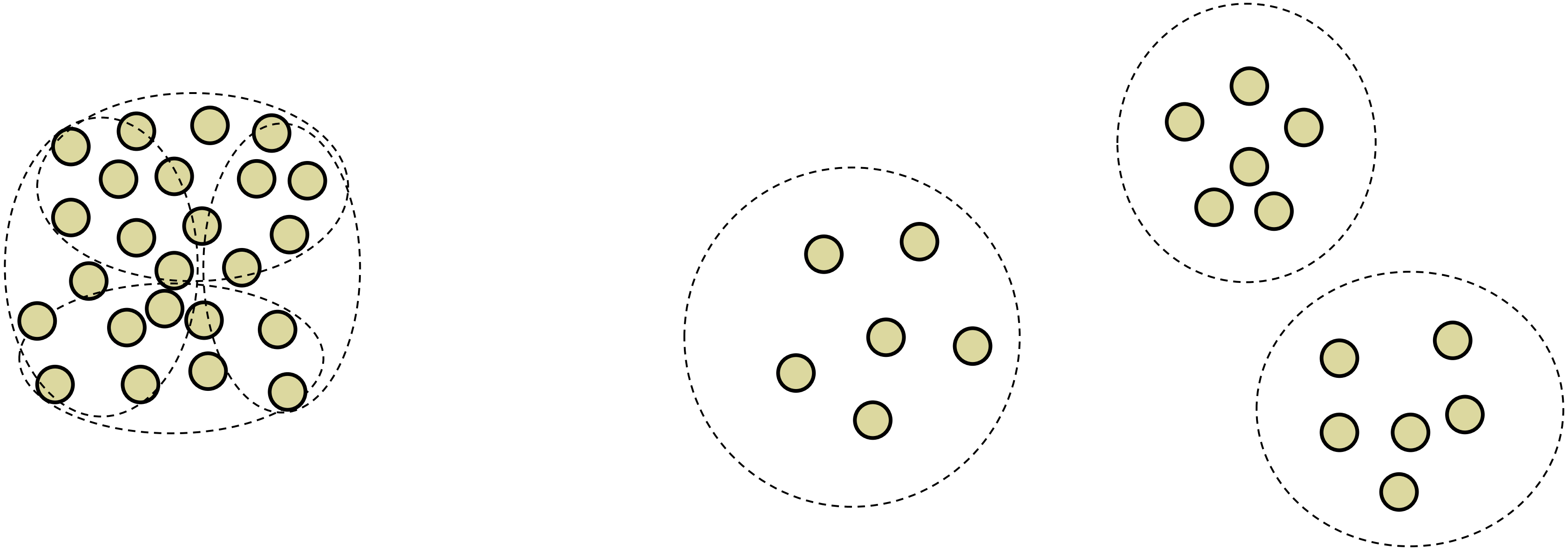
Issue: even if k-center is the "right" objective in that the optimal solution partitions data correctly, it's not clear that a 2-apx or $\mathrm{O}(\log^* n)$-apx will.

To address, assume data has some reasonable non-worst-case properties. In particular, perturbation-resilience [Bilu-Linial 2010]

Maria-Florina Balcan, Nika Haghtalab, and Colin White. *k-Center Clustering under Perturbation Resilience*. Int. Colloquium on Automata, Languages, and Programming (ICALP), 2016.

k-Center Clustering

Assumption: perturbing distances by up to a factor of 2 doesn’t change how the optimal k-center solution partitions the data.

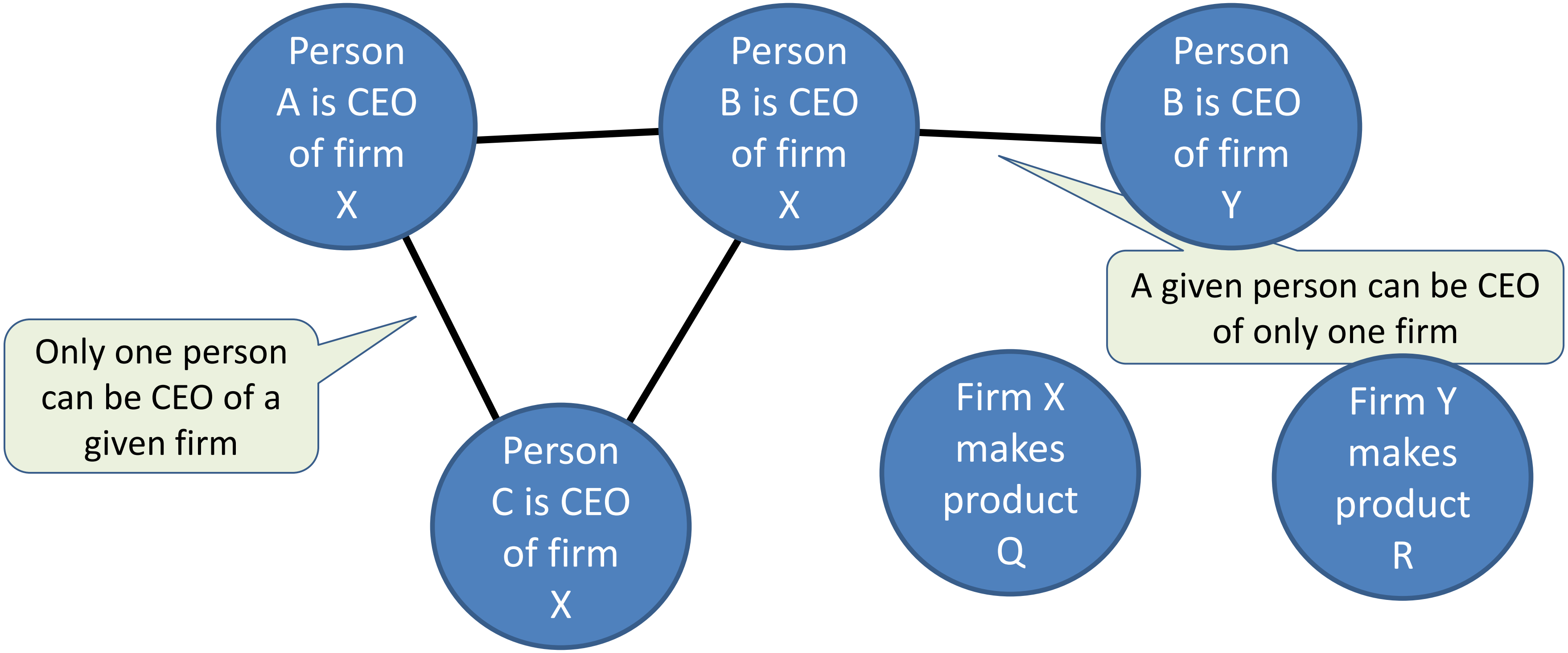


Results: under stability to factor-2 perturbations, can efficiently solve for optimal solution in both the symmetric and asymmetric case.

Maria-Florina Balcan, Nika Haghtalab, and Colin White. *k-Center Clustering under Perturbation Resilience*. Int. Colloquium on Automata, Languages, and Programming (ICALP), 2016.

Inference from Data given Constraints

NELL combines what it sees on the web with logical constraints that it knows about categories and relations



Person A is CEO of firm X

Person B is CEO of firm X

Person B is CEO of firm Y

Person C is CEO of firm X

Firm X makes product Q

Firm Y makes product R

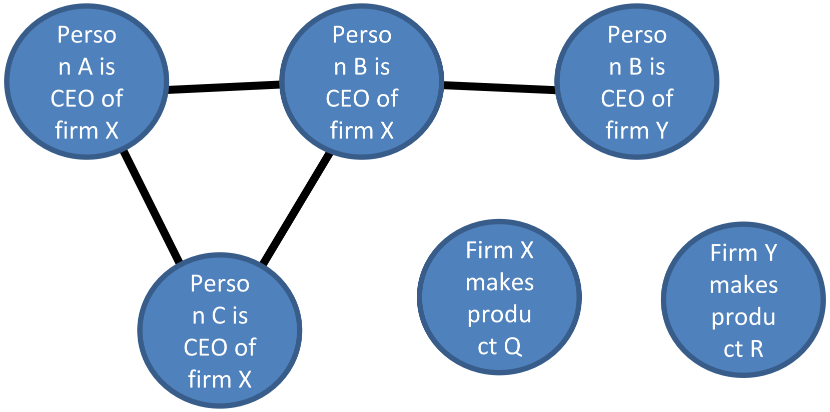
Only one person can be CEO of a given firm

A given person can be CEO of only one firm

In case of “not both” constraints, the max log-likelihood set of consistent beliefs = Max Weighted Independent Set

Pranjal Awasthi, Avrim Blum, Chen Dan. In preparation.

Max Weighted Independent Set


Person A is CEO of firm X
Person B is CEO of firm X
Person B is CEO of firm Y
Person C is CEO of firm X
Firm X makes product Q
Firm Y makes product R

Very hard to approximate in worst case

But, under some reasonable conditions:

- Low degree
- Instance is stable to bounded perturbations in vertex weights

Can show that natural heuristics will find correct solution

Pranjal Awasthi, Avrim Blum, Chen Dan. In preparation.

High level goals: address the gaps

- Machine learning increasingly in use everywhere

Google amazon UBER PayPal

- Significant advances in theory and application
- Yet large gap between the two
  - Practical success on theoretically-intractable problems
  - Theory focused on learning single targets. Large-scale systems aim to learn many tasks, and to use synergies among them to learn faster and better

Multitask and Lifelong Learning

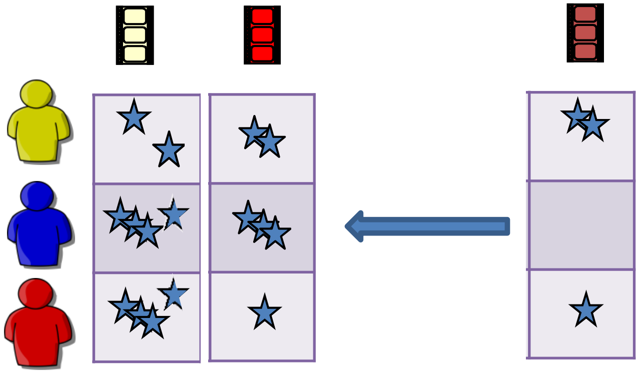
Modern applications often involve learning many things either in parallel, in sequence, or both.

E.g., want to:

- Personalize an app to many concurrent users (recommendation system, calendar manager, …)
- Quickly identify the best treatment for new disease being studied, by levaraging experience studying related diseases.
- Use relations among tasks to learn with much less supervision than would be needed for learning a task in isolation

Lifelong Matrix Completion

Consider a recommendation system where items (e.g., movies) arrive online over time



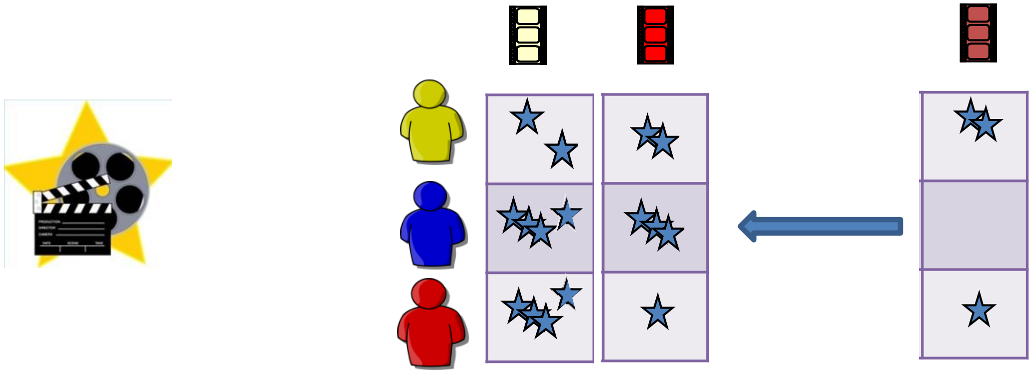
- From a few entries in the new column, want to predict a good approximation to the remainder

- Traditionally studied in offline setting. Goal is to solve in online, noisy setting

Maria-Florina Balcan and Hongyang Zhang. *Noise-Tolerant Life-Long Matrix Completion via Adaptive Sampling*. NIPS 2016.

Lifelong Matrix Completion

Assumptions: Underlying clean matrix is low rank & incoherent column space. Corrupted by bounded worst-case noise or sparse random noise.



Sampling model: can see a few random entries (cheap) or pay to get entire column (expensive).

Ideas: build a basis to use for prediction, but need to be careful to control error propagation!

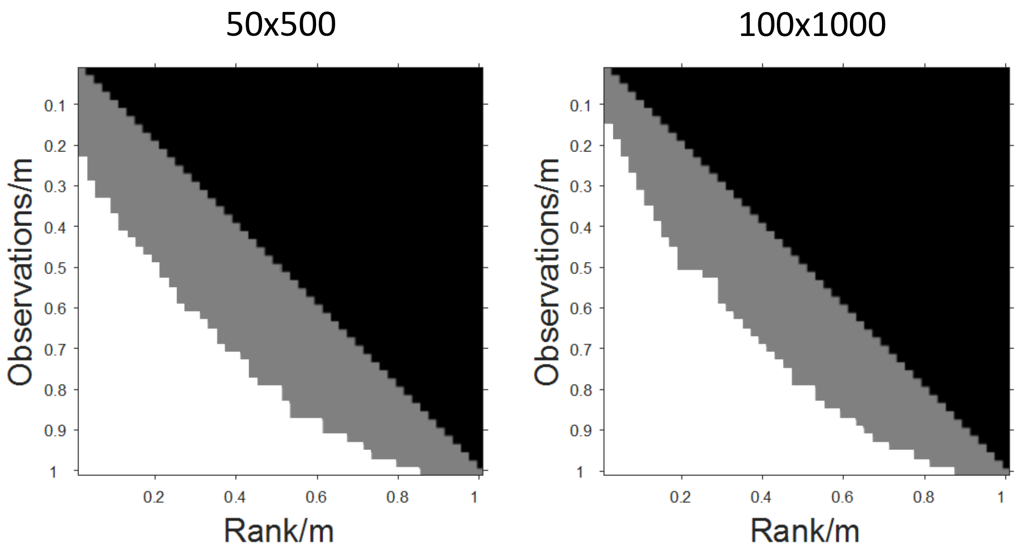
Extensions: low rank → mixture of low dim'l subspaces

Maria-Florina Balcan and Hongyang Zhang. *Noise-Tolerant Life-Long Matrix Completion via Adaptive Sampling*. NIPS 2016.

Lifelong Matrix Completion

Theorems: algs with strong guarantees on output error from limited observations under two noise models

Experiments: Synthetic data with sparse random noise



White Region: Nuclear norm minimization succeeds.
White and Gray Regions: Our algorithm succeeds.
Black Region: Our algorithm fails.

Maria-Florina Balcan and Hongyang Zhang. *Noise-Tolerant Life-Long Matrix Completion via Adaptive Sampling*. NIPS 2016.

Lifelong Matrix Completion

Theorems: algs with strong guarantees on output error from limited observations under two noise models

Experiments: Real data, using mixture of subspaces

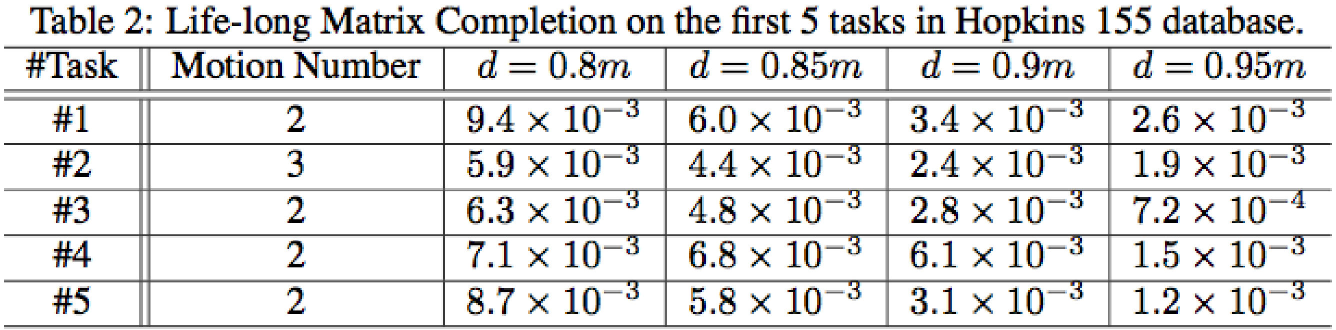


Table 2: Life-long Matrix Completion on the first 5 tasks in Hopkins 155 database.

#Task	Motion Number	$d = 0.8m$	$d = 0.85m$	$d = 0.9m$	$d = 0.95m$
#1	2	9.4×10^{-3}	6.0×10^{-3}	3.4×10^{-3}	2.6×10^{-3}
#2	3	5.9×10^{-3}	4.4×10^{-3}	2.4×10^{-3}	1.9×10^{-3}
#3	2	6.3×10^{-3}	4.8×10^{-3}	2.8×10^{-3}	7.2×10^{-4}
#4	2	7.1×10^{-3}	6.8×10^{-3}	6.1×10^{-3}	1.5×10^{-3}
#5	2	8.7×10^{-3}	5.8×10^{-3}	3.1×10^{-3}	1.2×10^{-3}

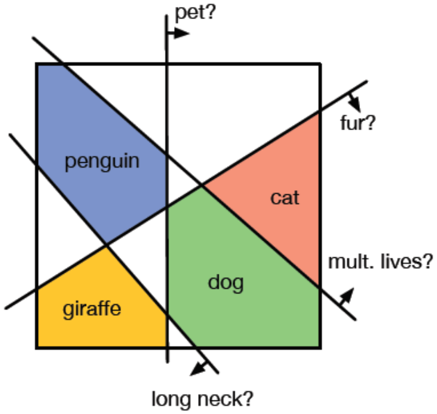
average relative error over 10 trials

Maria-Florina Balcan and Hongyang Zhang. *Noise-Tolerant Life-Long Matrix Completion via Adaptive Sampling*. NIPS 2016.

Multiclass unsupervised learning

Error-Correcting Output Codes [Dietterich & Bakiri '95]: method for multiclass learning from **labeled data**.

$$\begin{array}{c|cccc} & \text{pet?} & \text{fur?} & \text{long neck?} & \text{multiple lives?} \\ \hline \text{dog} & +1 & +1 & -1 & -1 \\ \text{cat} & +1 & +1 & -1 & +1 \\ \text{penguin} & -1 & -1 & -1 & -1 \\ \text{giraffe} & -1 & +1 & +1 & -1 \end{array}$$



What if you only have unlabeled data?

Idea: Separability + ECOC assumption implies structure that we can hope to use, even without labels!

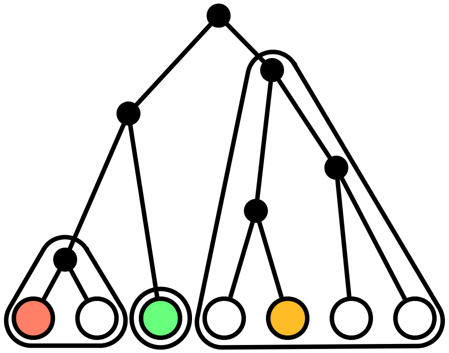
Thm: Learn from unlabeled data (plus very small labeled sample) when data comes from natural distributions

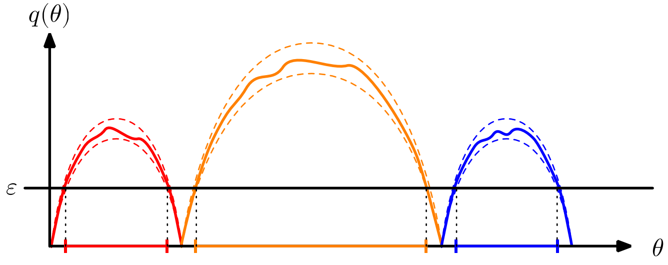
Maria-Florina Balcan, Travis Dick, and Yishay Mansour. *Label Efficient Learning by Exploiting Multi-class Output Codes*. AAAI 2017

Multiclass unsupervised learning

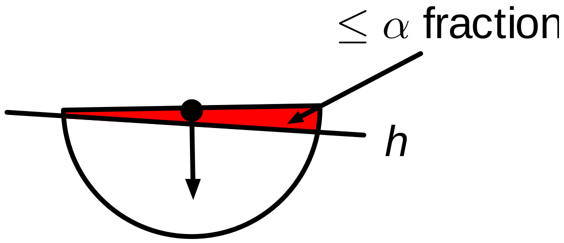
A taste of the techniques:

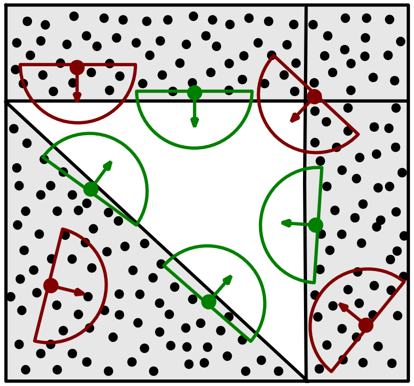
Robust Linkage Clustering





Hyperplane Detection

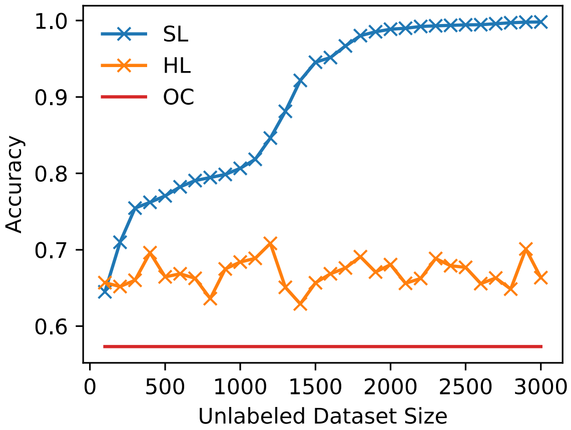




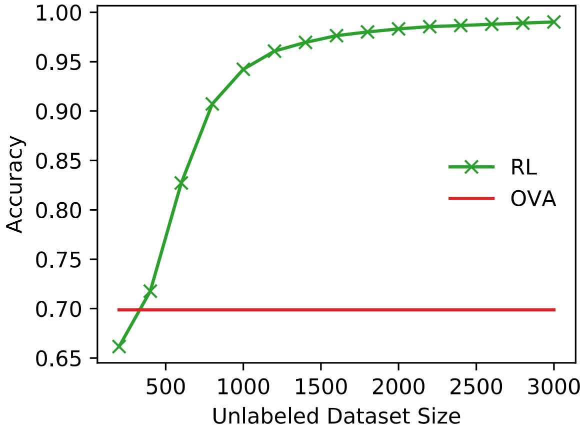
Maria-Florina Balcan, Travis Dick, and Yishay Mansour. *Label Efficient Learning by Exploiting Multi-class Output Codes*. AAAI 2017

Experiments

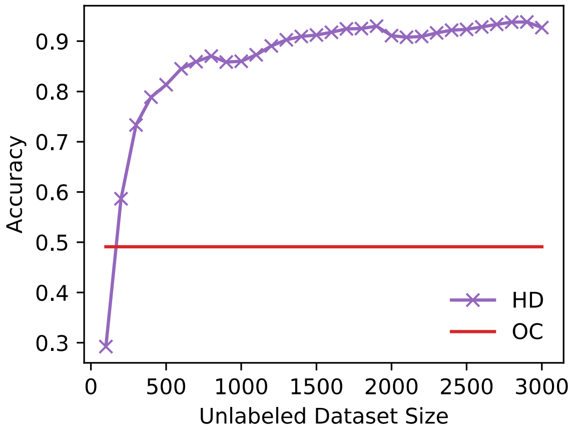
Synthetic Datasets:



Error Correcting

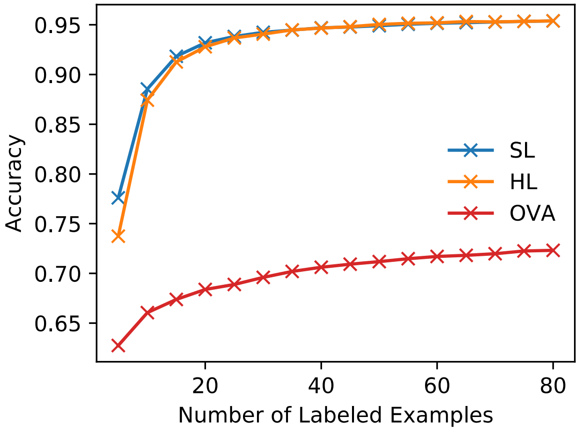


One-vs-all

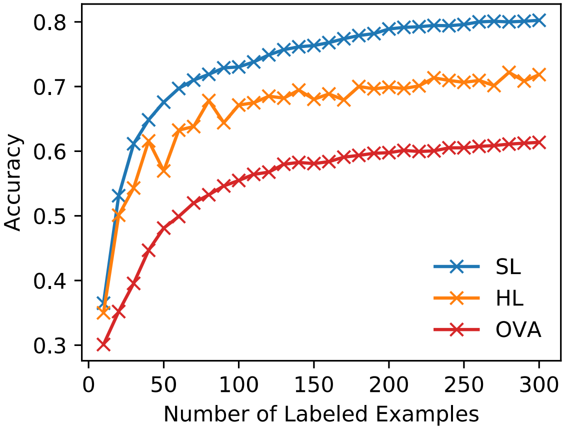


Boundary Features

Real-world Datasets



Iris



MNIST

- 🟥 output code
- 🟦 single linkage
- 🟧 hierarchical single linkage
- 🟩 robust single linkage
- 🟪 hyperplane detection

Maria-Florina Balcan, Travis Dick, and Yishay Mansour. *Label Efficient Learning by Exploiting Multi-class Output Codes*. AAAI 2017

Results in progress / under submission

Maria-Florina Balcan, Avrim Blum, and Vaishnavh Nagarajan. *Lifelong Learning in Costly Feature Spaces.*

- Given a series of related learning tasks, want to extract commonalities to learn new tasks more efficiently
- E.g., decision trees (often used in medical diagnosis) that share common substructures

- Focus: using learned commonalities to reduce number of features that need to be examined in training data

Avrim Blum and Nika Haghtalab. *Generalized Topic Modeling.*

- Generalize co-training approach for semi/un-supervised learning to the case that objects can belong to a mixture of classes

Maria-Florina Balcan, Travis Dick, Yingyu Liang, Wenlong Mou, and Hongyang Zhang. *Differentially Private Clustering in High-Dimensional Euclidean Spaces*

Staged Curricular Learning

Maria-Florina Balcan, Avrim Blum, and Tom Mitchell. In progress

Recall the setting of NELL

- Learns many (thousands) of categories
  - river, city, athlete, sports team, country, attraction,…
- And relations
  - athletePlaysSport, cityInCountry, drugHasSideEffect,…
- From mostly unlabeled data (reading the web)

- ➢ ford makes the automobile escape
- ➢ camden_yards is the home venue for the sports team baltimore_orioles
- ➢ christopher_nolan directed the movie inception

Staged Curricular Learning

Maria-Florina Balcan, Avrim Blum, and Tom Mitchell. In progress

NELL is aided by a given ontology, which helps it bootstrap from unlabeled data

- E.g., $athlete(x) \Rightarrow person(x)$, $city(x) \Rightarrow place(x)$, $\neg(city(x) \wedge country(x))$

Q: can you learn new implications as you learn categories and relations, in a self-improving way?

- E.g., $playsOnTeam(p,t) \wedge teamPlaysSport(t,s) \Rightarrow playsSport(p,s)$
- and $playsOnTeam(p,t) \wedge playsSport(p,s) \Rightarrow teamPlaysSport(t,s)$

In this work we have been understanding conditions under which these can be effectively learned and used to improve the overall learning process