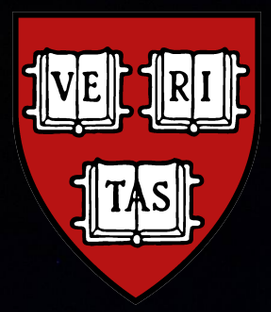
LAMBDA MEANS CLUSTERING

AUTOMATIC PARAMETER SEARCH AND DISTRIBUTED COMPUTING IMPLEMENTATION

MARCUS COMITER, MIRIAM CHA, HT KUNG, SURAT TEERAPITTAYANON
HARVARD UNIVERSITY

ICPR 2016

DECEMBER 6, 2016



TALK OUTLINE

- **Motivation and Introduction**
- Background
- Lambda Means
- Benefits of Lambda Means
- Results
- Extension to Distributed Framework

MACHINE LEARNING: VISION VS. REALITY

MACHINE LEARNING: VISION VS. REALITY

Vision

MACHINE LEARNING: VISION VS. REALITY

Vision

Reality

CLUSTERING

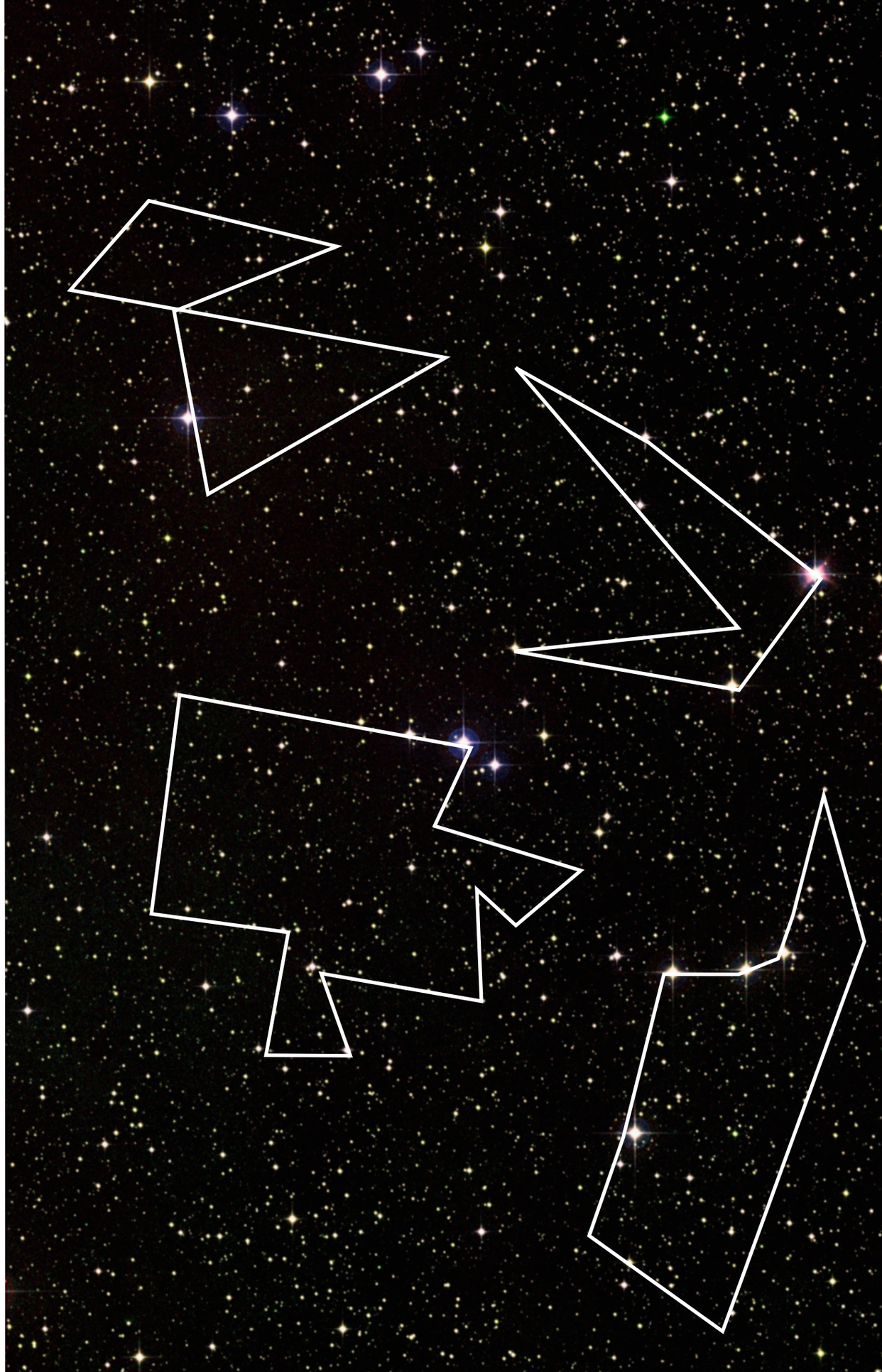
- Clustering is one of the most basic yet most powerful and fundamental of machine learning algorithms

- But even in this simple setting, the choice of parameters are both **difficult** and greatly **impact performance**

CLUSTERING

- Clustering is one of the most basic yet most powerful and fundamental of machine learning algorithms

- But even in this simple setting, the choice of parameters are both **difficult** and greatly **impact performance**



If machine learning is fundamentally a *data driven science*, shouldn't the use of machine learning itself follow a data driven methodology?

INTRODUCTION

- We present Lambda Means, a meta algorithm for the newly popular clustering algorithm DP-means

- Lambda Means automatically finds DP-means' main parameter (**λ**) automatically

- It finds **λ** using **the data itself** on which the clustering is being performed

TALK OUTLINE

- Motivation and Introduction
- **Background**
- Lambda Means
- Benefits of Lambda Means
- Results
- Extension to Distributed Framework

DP-MEANS

- DP-means forms clusters of superior quality using a distance parameter **λ** to ensure minimum separation between cluster centroids rather than specifying **k** in advance
  - B. Kulis and M. I. Jordan (the authors of DP-means) show that this new algorithm outperforms the traditional k-means algorithm!

- The algorithm forms a new cluster when a data point is found to be more than **λ** distance away from all existing cluster centroids

DIRICHLET PROCESS

- Under an assumption that a sequence of data is drawn from a Dirichlet Process Mixture Model, B. Kulis and M. I. Jordan (the authors of DP-means) prove that there exists a lambda value such that when used by DP-means, the algorithm will discover the ground truth number of clusters k.

$$
\begin{aligned}
\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_k &\sim G_0 \\
\boldsymbol{\pi} &\sim \text{Dir}(k, \boldsymbol{\pi}_0) \\
\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n &\sim \text{Discrete}(\boldsymbol{\pi}) \\
\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n &\sim \mathcal{N}(\boldsymbol{\mu}_{\boldsymbol{z}_i}, \sigma I)
\end{aligned}
$$

- μ corresponds to the mean of each of the clusters, drawn from some base distribution G0, which is the prior distribution over the means
- $\pi=(\pi_1, \pi_2\ldots)$ corresponds to the vector of probabilities of being in a cluster (k → infinity)
- z_i is an indicator of cluster assignment
- x_i is a data point

DP-MEANS

- In practice, without knowing the parameters of the distribution from which the data is drawn, it is unclear how to find the appropriate value of **λ** for use with DP-means

- To solve this problem, a Farthest-first Heuristic requiring a user-provided approximation of **k** can be used

  - However, it is **not easy** to set **k**

  - The choice of **k** has a **marked impact** on the resulting value of **λ**

TALK OUTLINE

- Motivation and Introduction
- Background
- **Lambda Means**
- Benefits of Lambda Means
- Results
- Extension to Distributed Framework

LAMBDA MEANS

- As a solution for **automatically** finding the **λ** parameter for use with DP-means, we present Lambda Means
- It finds **λ** using **the data itself** on which the clustering is being performed
- Under an assumption that the data is generated by a Dirichlet Process Mixture Model, we formally prove that the **λ** value found by Lambda Means is the same **λ** used in generating the data (see Section III.D in our paper)

LAMBDA MEANS

- The algorithm's main mechanism is to decrease **λ** at each iteration, automatically terminating at the proper **λ** value

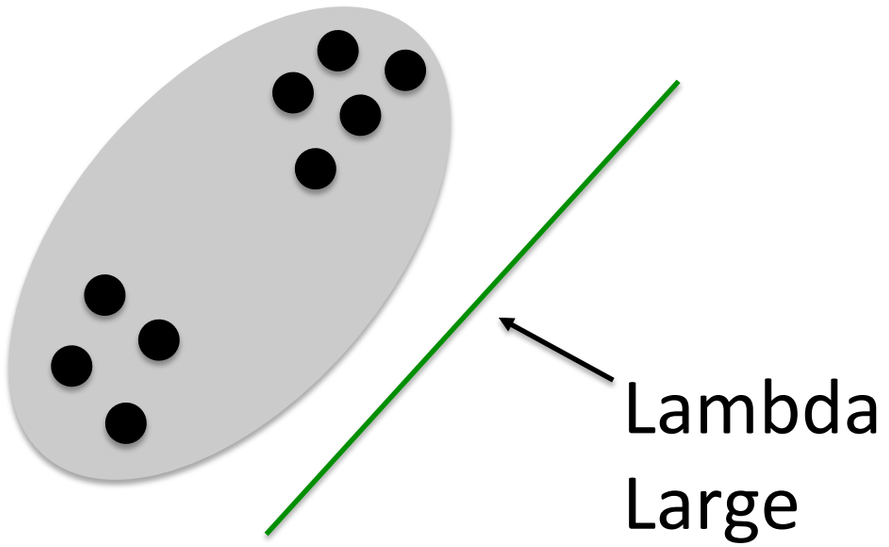
- This has the effect of precipitating clusters at each iteration up to the point at which **all clusters have been identified**, but before the point at which true clusters are broken up into individual points

ILLUSTRATION OF EFFECT OF DECREASING λ

Iteration: T

Lambda: Large

A large value of lambda causes the two sets of points to be clustered together



Iteration: T + ΔT

Lambda: Small

A small value of lambda causes the two sets of points to be clustered separately

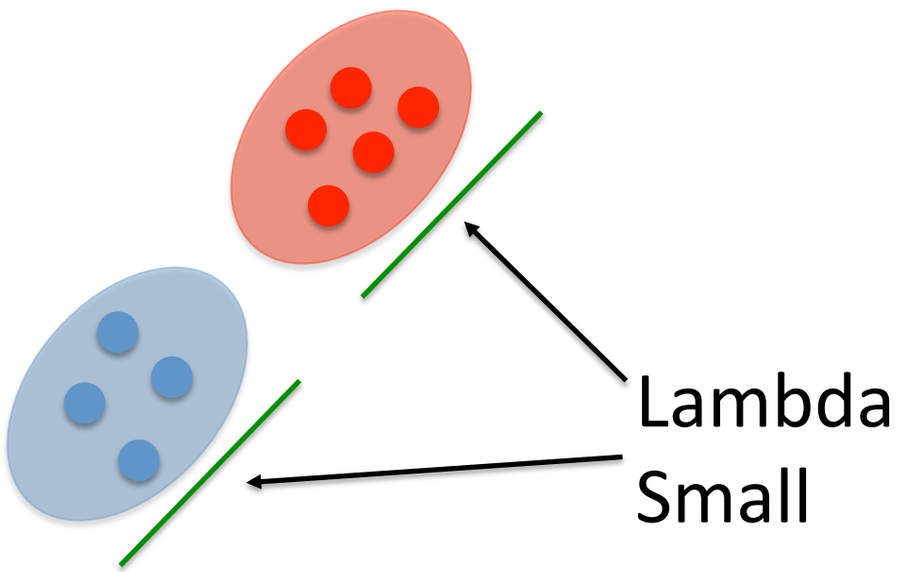
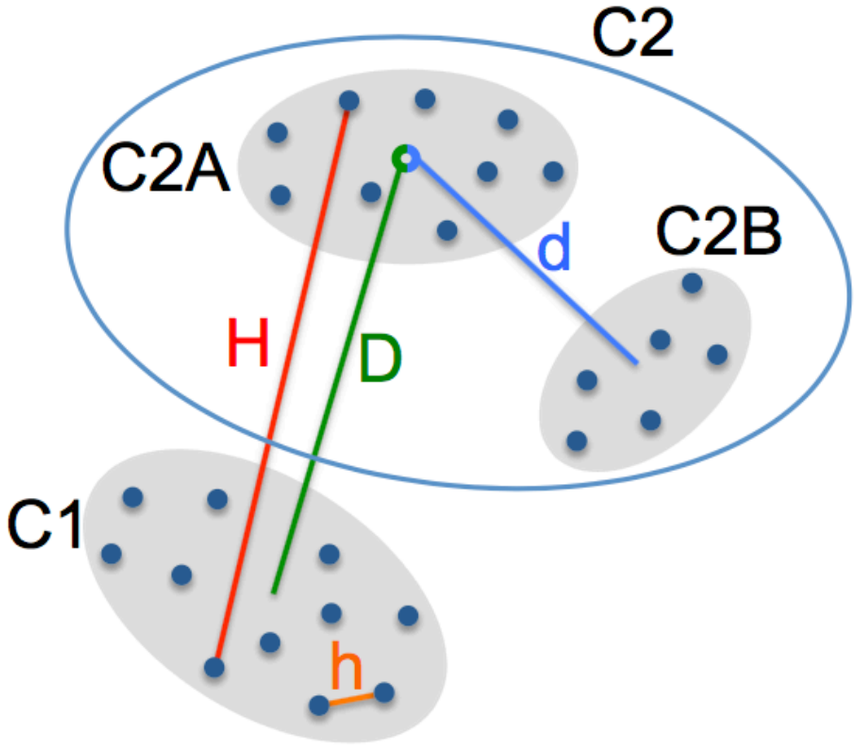
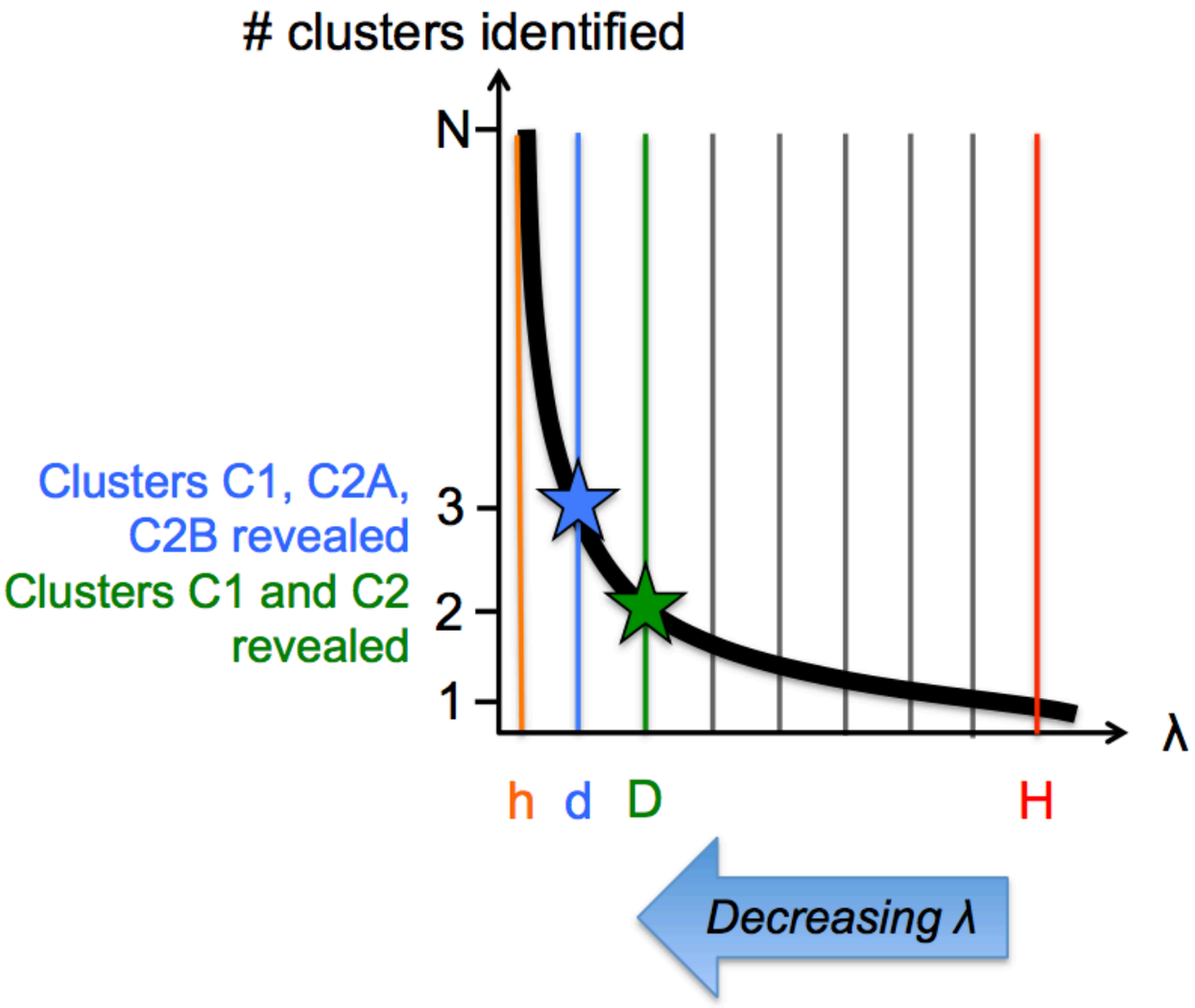


ILLUSTRATION OF EFFECT OF DECREASING λ





D, d: maximum and minimum distance between cluster centroids

H, h: maximum and minimum distance between data points

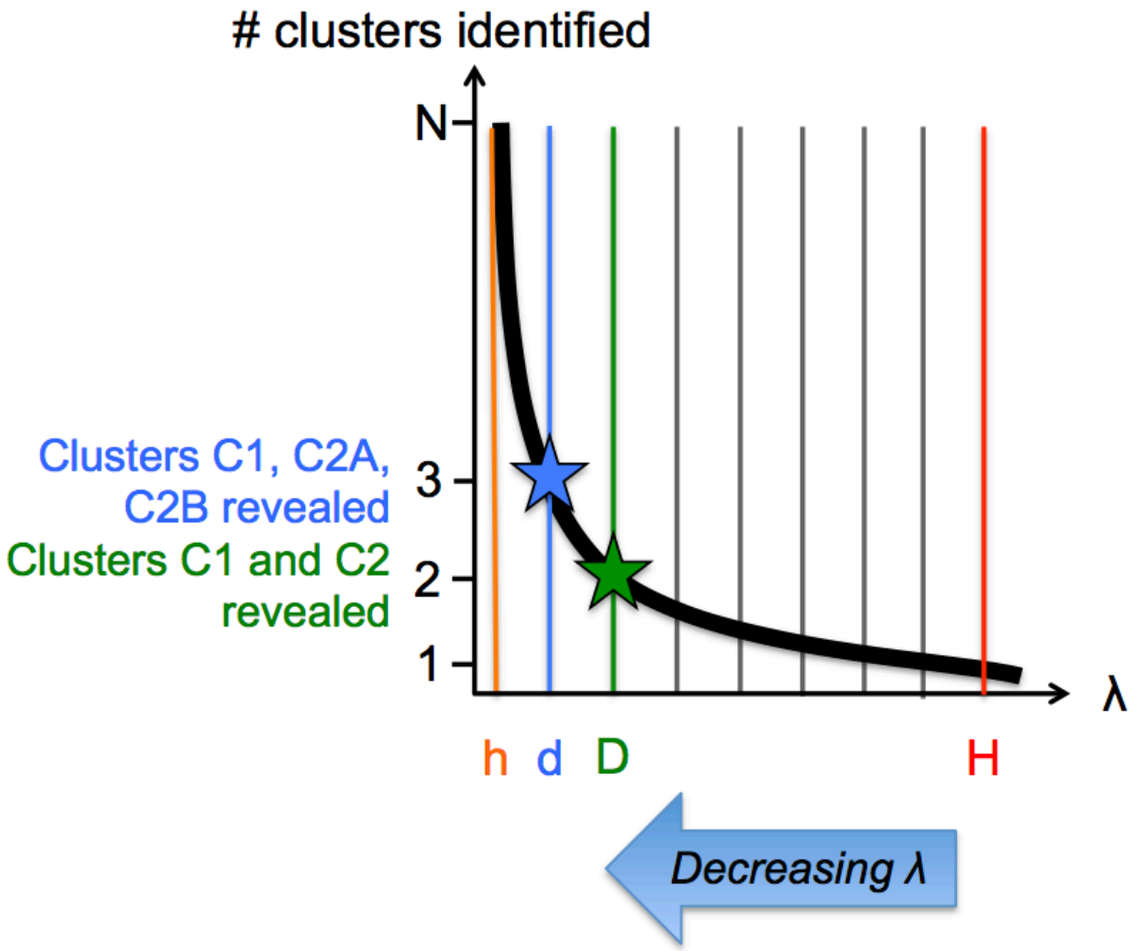
N: total # of data points

LAMBDA MEANS

- Note that a naive implementation would generate the entire curve and then search for the elbow

- Lambda Means replaces the need for this **exhaustive** search for the elbow of the curve

- The algorithm uses the cumulative number of clusters formed as a signaling mechanism, continuing to iterate with smaller values of **λ** until the stopping criteria is met



TALK OUTLINE

- Motivation and Introduction
- Background
- Lambda Means
- **Benefits of Lambda Means**
- Results
- Extension to Distributed Framework

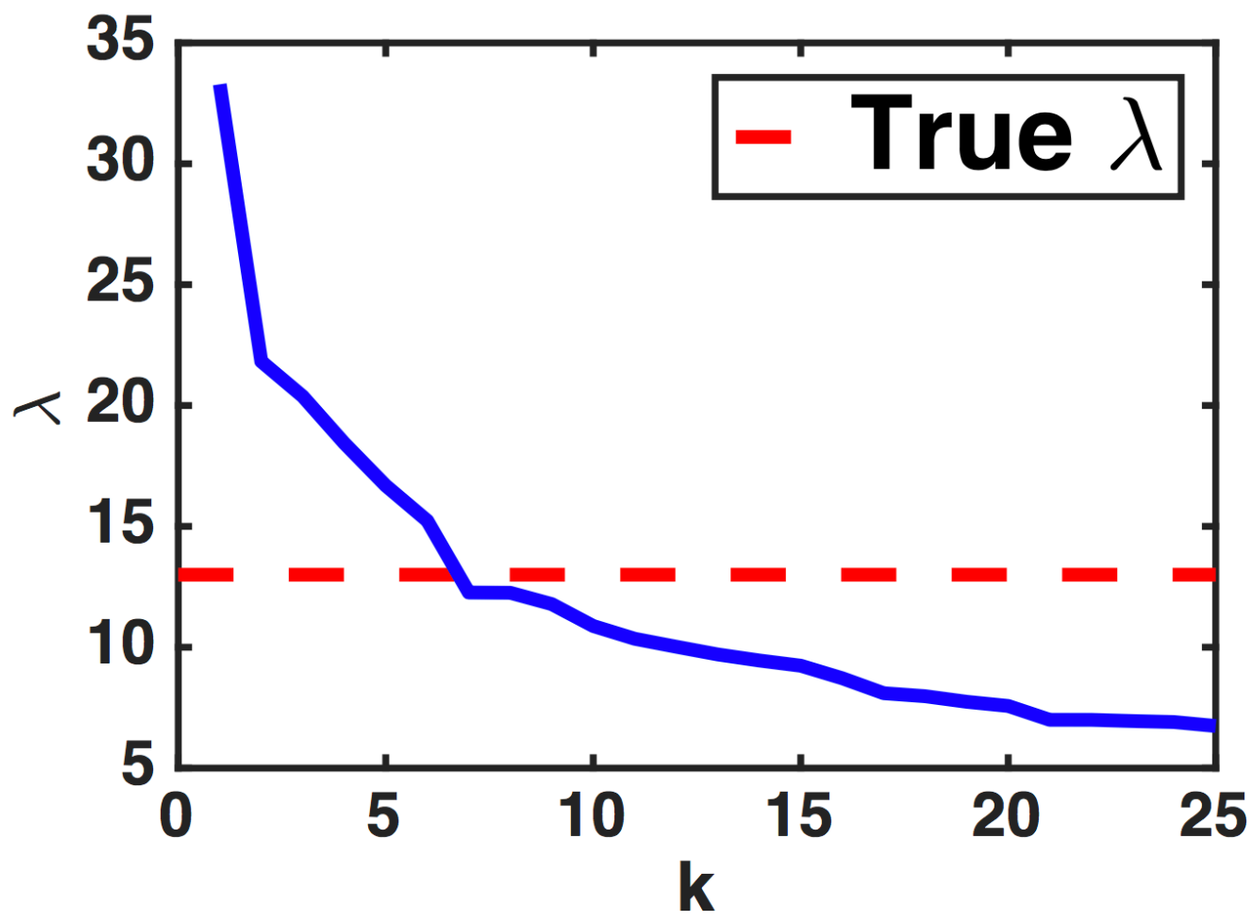
BENEFITS

- Lambda means is more robust then using a Farthest-first Heuristic, which requires a user-defined **k**
  - Reason 1: Setting this **k** can be very difficult
  - Reason 2: If the initial approximation to **k** is wrong, it negatively affects finding the correct **λ**

BENEFITS

- To show the effect of an incorrect **k,** we generate a dataset and then use the Farthest-first Heuristic with a number of different values of k to derive **λ**

- We find that **λ** varies greatly based on the initial **k** used



BENEFITS

- The drawbacks of the farthest-first heuristic are clear:
  - The method is **brittle** to small changes in the approximation of **k**
  - The method has a **large impact** on the derived value of **λ** as well as potentially on the **resulting cluster quality**
- In contrast, Lambda Means automatically finds the **λ** value without an initial approximation for **k**

TALK OUTLINE

- Motivation and Introduction
- Background
- Lambda Means
- Benefits of Lambda Means
- **Results**
- Extension to Distributed Framework

RESULTS

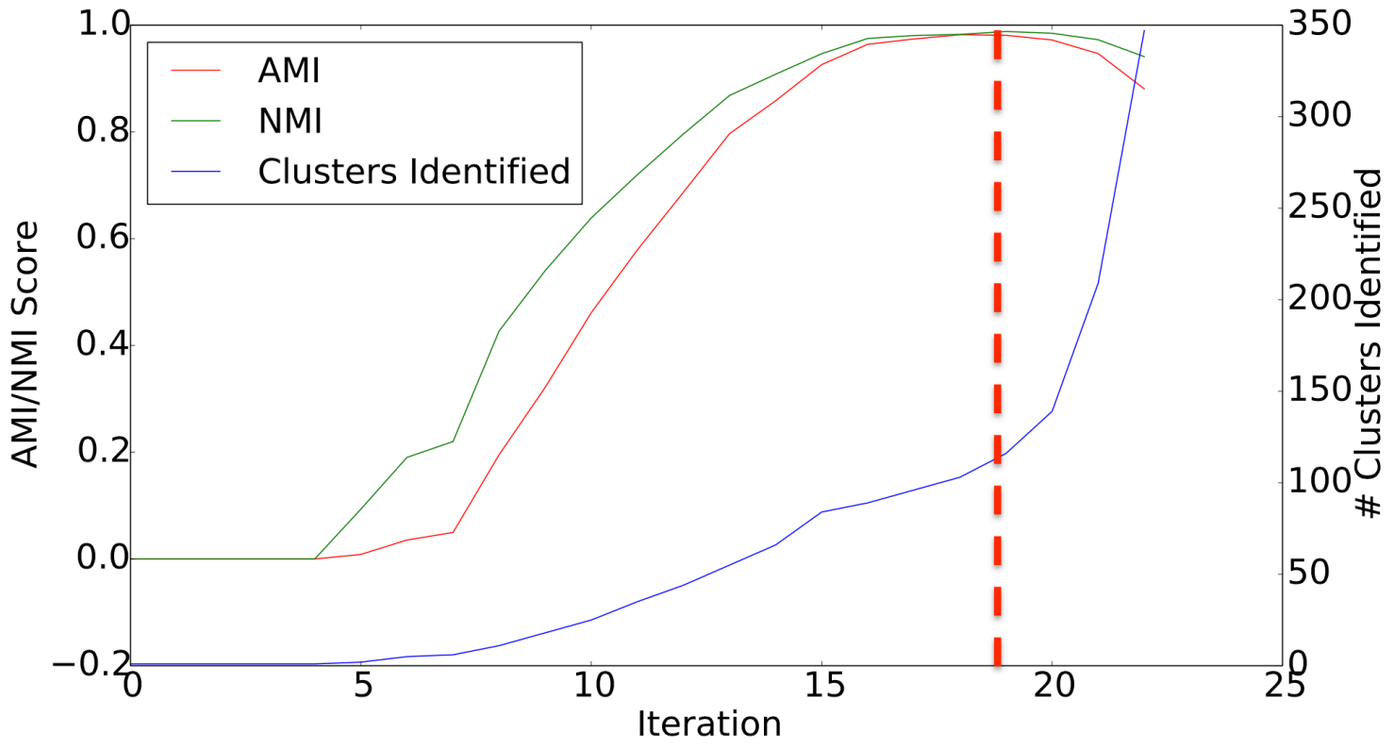
- We provide experimental evaluation of **λ**-means on both synthetic and real world data

- For synthetic data, we generate data with different values of inter-cluster **variance ρ** and the intra-cluster **variance σ**

- For real-world data, we use the MNIST hand written digit dataset

RESULTS

- This figure shows that for synthetic data with a high value of **ρ/σ**, Lambda Means is able to automatically find the **λ** value that maximizes AMI and NMI scores
    - NMI measures the amount of mutual information normalizing for number of clusters, and AMI measures the amount of mutual information accounting for chance
- We can also judge Lambda Means by its ability to identify the correct number of clusters, which it does (as shown by the blue line)



RESULTS

- We now compare the AMI and NMI scores for Lambda Means and DP-means in Table I for additional values of ρ/σ, as well as for the MNIST dataset

- Lambda Means outperforms DP-means where λ is set via the Farthest-first heuristic

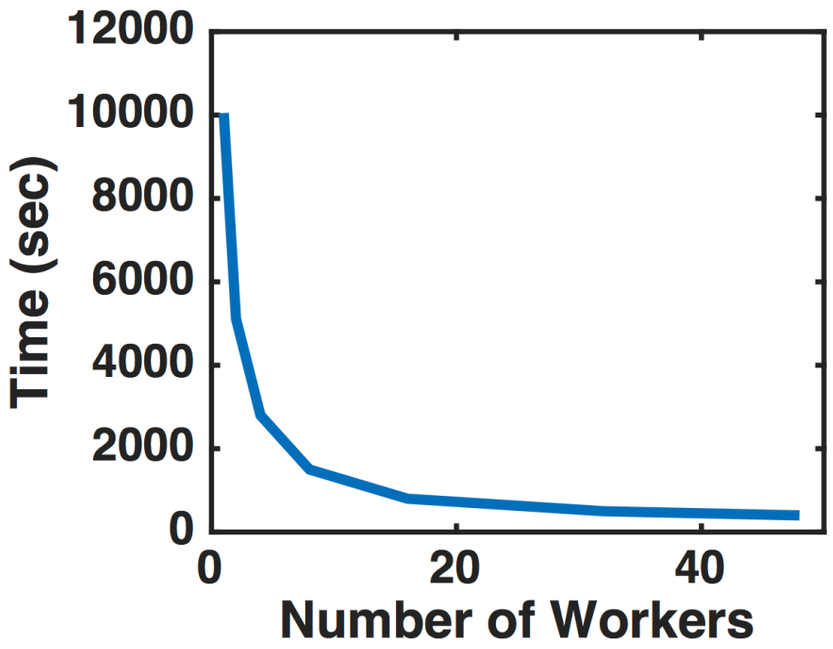
	Dataset					
	Syn. $\frac{\rho}{\sigma} = 15$		Syn. $\frac{\rho}{\sigma} = 5$		MNIST	
Algorithm	AMI	NMI	AMI	NMI	AMI	NMI
λ-means	0.97	0.98	0.77	0.82	0.43	0.53
DP-means	0.87	0.92	0.52	0.78	0.32	0.38

TALK OUTLINE

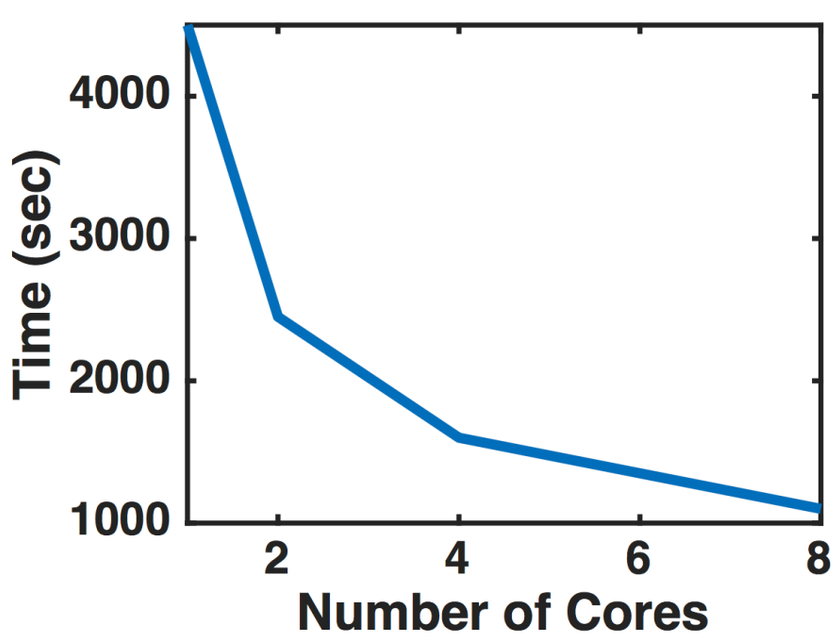
- Motivation and Introduction
- Background
- Lambda Means
- Benefits of Lambda Means
- Results
- **Extension to Distributed Framework**

DISTRIBUTED RESULTS

- Lambda Means easily extends to the distributed framework under the optimistic concurrency control framework
- We achieve within a factor of two away from a perfect speed-up in both the multicore and multi-processor distributed settings



(a) Distributed cluster in the cloud



(b) Multicore

THANK YOU

MARCUS COMITER, MIRIAM CHA, HT KUNG, SURAT TEERAPITTAYANON
HARVARD UNIVERSITY