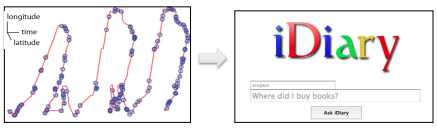
Tackling the Challenges of Big Data
Big Data Analytics

Daniela Rus
Professor
Director of CSAIL
Massachusetts Institute of Technology

Tackling the Challenges of Big Data
Big Data Analytics
Data Compression

Daniela Rus
Professor
Director of CSAIL
Massachusetts Institute of Technology

Learning Big Data Patterns from Tiny Core-sets



Data Challenge

- 2.5 quintillion ($2.5 * 10^{18}$) bytes per day in 2012 (source: ibm.com)
- Per capita information capacity ~ doubled every 40 months since 1980
- Cameras, phones, sensors enable life-logging capabilities

Photos removed due to copyright restrictions.
We are sorry for any inconvenience this may cause.

Photos removed due to copyright restrictions.
We are sorry for any inconvenience this may cause.

Data
Data
Data
Data
= Data
Data
Data

How much data?

Photo removed due to copyright restrictions.
We are sorry for any inconvenience this may cause.

1 GPS Packet
=
100 bytes

(latitude, longitude, time)

1 GPS Packet
=
100 bytes

every 10 seconds

~ 0.4 Mb / hour
or
~100Mb / day

~0.1 Gb / day
per device

~300 million
smart phones
sold in 2010

http://mobithinking.com/mobile-marketing-tools/latest-mobile-stats

For
100 million devices

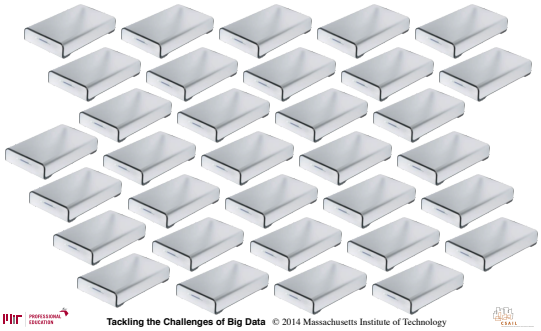
For
100 million devices

~ 10 petabytes
per day

~ 10 thousand
terabytes
per day

2 terabytes each

x5000 / day



That's a lot of data.

Example: Activities

Photo removed due to copyright restrictions.
We are sorry for any inconvenience this may cause.

Travel Mode

Blue: Predicted as if on Wheels

Red: Predicted as if on Foot

Travel Mode

Recall: fraction of relevant instances classified correctly
Precision: fraction of instances correctly classified

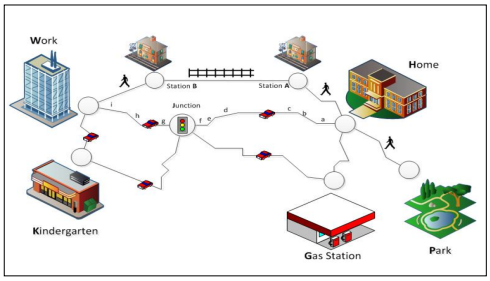
On Foot Prediction:
93% Recall, 86% Precision

Travel Mode

Recall: fraction of relevant instances classified correctly
Precision: fraction of instances correctly classified

On Wheels Prediction:
96% Recall, 98% Precision

Example: Activities



abcdefghihgfedcba…

From Signals to Activities

time	latitude	longitude
8:44:57	1.295783	103.7816
8:44:59	1.295785	103.7816
8:45:00	1.295782	103.7816
8:45:01	1.295782	103.7816
8:45:04	1.29579	103.7817
8:45:05	1.295802	103.7817
8:45:08	1.295915	103.7818
8:45:09	1.29598	103.7819
8:45:10	1.296015	103.7819
8:45:11	1.296057	103.782
…	…	…

Time	Mode	Location
8:40	Wheel	Central Expy, Singapore
9:10	Wheel	Ayer Rajah Expy, Singapore
9:35	Wheel	Chin Swee Rd, Singapore
9:40	Wheel	261 Outram Rd, Singapore 169057
9:41	Wheel	1 St Andrew's Rd, Singapore 178957
9:43	Wheel	290A Havelock Rd, Singapore 169664
9:48	Foot	9A Raffles Ave, Singapore 039801
9:49	Foot	2 Raffles Blvd, Singapore 039593
9:49	Foot	3 Bunna Vista Rd, Singapore
9:50	Foot	5 Lower Kent Ridge Rd, Singapore
…	Foot	8 Medical Dr, Singapore 117594
…	Foot	20 Lorrie Hill, Singapore
…	Foot	133 Devonshire Rd, Singapore 239878
…	Wheel	121 Devonshire Rd, Singapore 239682
…	Wheel	55 Grange Rd, Singapore
…	Wheel	57 Grange Rd, Singapore 239700
…	Wheel	Natl Youth Council, Singapore
…	Foot	50K Paterson Rd, Singapore 238517
…	Wrng	Boston

Time	Activity
8:40 – 9:30	Driving to the university of Singapore
9:30 – 13:10	Meetings in the university
13:10-14:00	Eating Lunch
14:00-18:00	More meetings
18:00-19:00	Taking a walk
19:00-19:30	At a coffee shop
19:30–20:30	Taking Taxi to airport
20:30–23:30	Flying home

semantic compression

Challenges

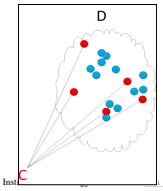
- Storing data on smart phone is expensive
- Transmission data is expensive
- Hard to interpret raw data
- Dynamic real-time streaming data

Outline

- Motivation
- Coreset definition, computation model
- Coresets for k-means
- Use case: life logging systems
- Coreset for k-segments
- From coresets to text

Challenge: Find RIGHT data from Big Data

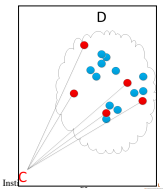
Given data *D* and Algorithm *A* with *A(D)* intractable, can we efficiently reduce *D* to *C* so that *A(C)* fast and *A(C)* ~ *A(D)*?



Challenge:
Find RIGHT data from Big Data

Given data *D* and Algorithm *A* with *A(D)* intractable, can we efficiently reduce *D* to C so that *A(C)* fast and *A(C)* ~ *A(D)*?

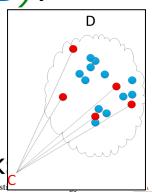
C should be fast to compute
A(C) ~ A(D) should be provable



Challenge:
Find RIGHT data from Big Data

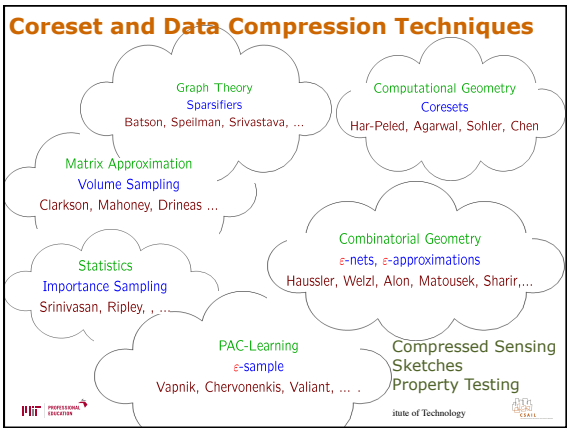
Given data *D* and Algorithm *A* with *A(D)* intractable, can we efficiently reduce *D* to C so that *A(C)* fast and *A(C)* *A(D)*?

Provable guarantees for:
- Generalization
- Debugging
- Removing bottleneck



Big Data Computation Model

- Input: Infinite stream of vectors
- n vectors seen so far
- ~ log n memory
- M processors
- ~log(n) / M insertion time per point



Tackling the Challenges of Big Data
Big Data Analytics
Data Compression

THANK YOU

Tackling the Challenges of Big Data
Big Data Analytics

Daniela Rus
Professor
Director of CSAIL
Massachusetts Institute of Technology

Tackling the Challenges of Big Data
Big Data Analytics
Data Compression

Daniela Rus
Professor
Director of CSAIL
Massachusetts Institute of Technology

References for Compression

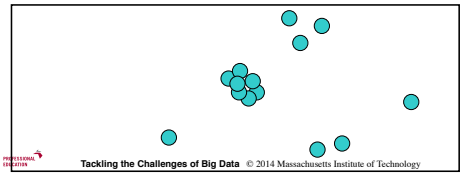
- k-Means [Feldman, Langberg, STOC'11]
- k-Segments [Feldman, Sung, Rus, ACM-GIS'12]
- Text Mining (LSA/PCA) [Feldman, Sohler, SODA'13]
- k-Lines [Feldman, Fiat, Sharir, FOCS'09]
- Mixture of Gaussians [Feldman, Krause, NIPS'11]
- Google Pagerank [Feldman, Yahoo! Research]
- Image Compression
  [Feldman, Sochen, J. of Math. Image & Vision]
- Video Compression
  [Feldman, Newman, submitted]

Outline

- Motivation
- Coreset definition, computation model
- Coresets for k-means
- Use case: life logging systems
- Coreset for k-segments
- From coresets to text

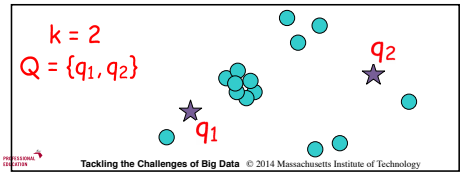
Coreset Example: K-Median Queries

- Input: $P \subseteq R^d$



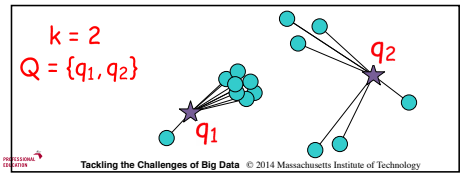
K-Median Queries

- Input: $P \subseteq R^d$
- Query: A set Q of k points



k-Median Queries

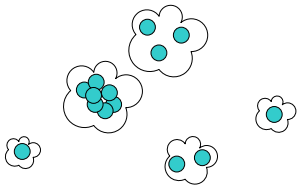
- Input: $P \subseteq R^d$
- Query: A set Q of k points
- Output: $\sum_{p \in P} \text{dist}(p, Q) = \sum_{p \in P} \min_{q \in Q} \|p - q\|$



(k,ε)-Median Coreset

Answer k-median queries in sub-linear time

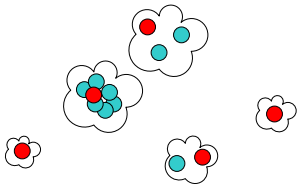
Key Idea: Replace many points by one weighted representative:



(k,ε)-Median Coreset

Answer k-median queries in sub-linear time

Key Idea: Replace many points by one weighted representative:

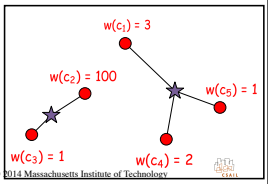


Definition: (k,ε)-Median Coreset

C is a (k, ε)-coreset for P, if $\forall Q, |Q| = k$:

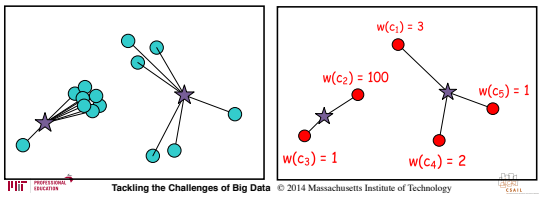
$$\sum_{p \in P} \text{dist}(p, Q) \sim \sum_{c \in C} w(c) \cdot \text{dist}(c, Q)$$

Multiplicative error $\leq 1 + \varepsilon$

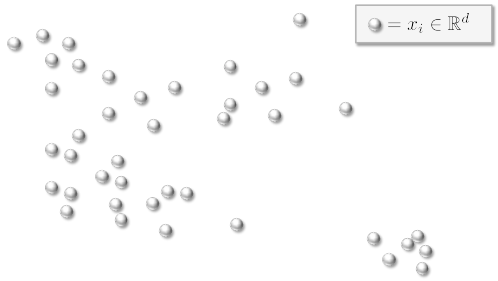


(k,ε)-Median Coreset

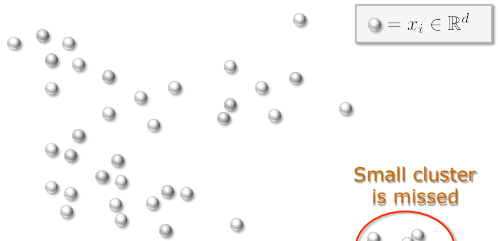
$$\sum_{p \in P} \text{dist}(p, Q) \sim \sum_{c \in C} w(c) \cdot \text{dist}(c, Q)$$



Can we use Naïve Uniform Sampling?



Naïve Uniform Sampling

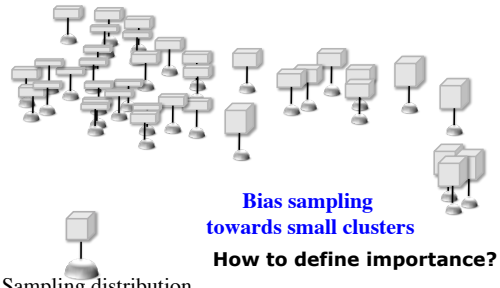


Small cluster is missed

Sample a set *U* of *m* points uniformly

← High variance

Sampling Distribution

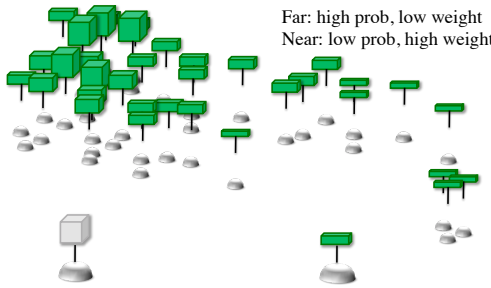


Bias sampling towards small clusters

How to define importance?

Sampling distribution

How to Define Distribution? Weights



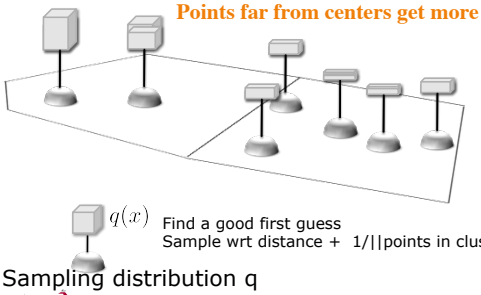
Far: high prob, low weight
Near: low prob, high weight

Sampling distribution

Weights

Creating a Sampling Distribution

Points in sparse cells get more mass
Points far from centers get more mass



$q(x)$

Find a good first guess
Sample wrt distance + 1/||points in cluster||

Sampling distribution q

Theorem: general claim for coresets

[Feldman, Langberg, STOC'11]

Let P, Q and $dist: P \times Q \rightarrow R^+$.

A sample $C \subseteq P$ from the distribution

$$\text{sensitivity(p)} = \max_{q \in Q} \frac{dist(p,q)}{\sum_{p'} dist(p',q)}$$

is a coreset if $|C| \geq \frac{\text{dimension of } Q}{\epsilon^2} \cdot \sum_p \text{sensitivity}(p)$

Coreset for k-means

[Feldman, Sohler, Monemizadeh, SoCG'07]

The coreset for k-means can be computed by choosing points from the distribution:

$$\text{sensitivity}(p) = \frac{dist(p,q^*)}{\sum_{p'} dist(p',q^*)} + \frac{1}{n_p}$$

q^* = k-means of P

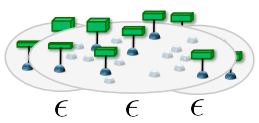
n_p = number of points in the cluster of p

$|C| = \frac{k \cdot d}{\epsilon^2}$

Composition of Coresets

Merge The union of two (k, ϵ)-coresets is a (k, ϵ)-coreset.

- [c.f. Har-Peled, Mazumdar 04]



Composition of Coresets

Merge The union of two (k, ϵ)-coresets is a (k, ϵ)-coreset.

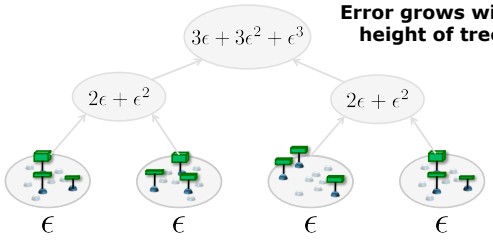
Compress A (k, δ)-coreset of a (k, ϵ)-coreset is a $(k, \epsilon + \delta + \epsilon\delta)$-coreset

ϵ

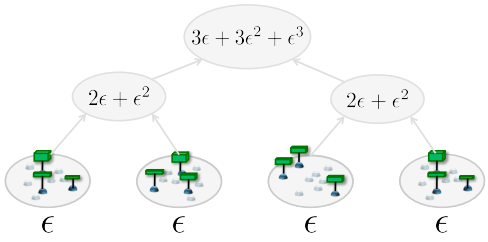
$\epsilon + \delta + \epsilon\delta$

Coresets on Streams

Error grows with height of tree



Coresets in Parallel



Tackling the Challenges of Big Data
Big Data Analytics
Data Compression

THANK YOU

Tackling the Challenges of Big Data
Big Data Analytics

Daniela Rus
Professor
Director of CSAIL
Massachusetts Institute of Technology

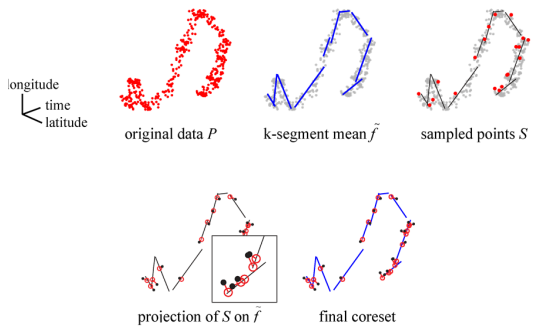
Tackling the Challenges of Big Data
Big Data Analytics
Data Compression

Daniela Rus
Professor
Director of CSAIL
Massachusetts Institute of Technology

Outline

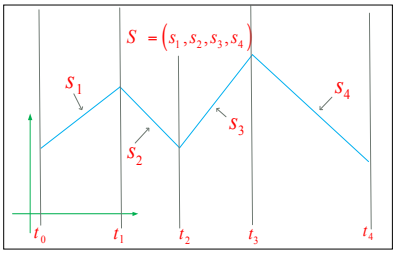
- Motivation
- Coreset definition, computation model
- Coresets for k-means
- Use case: life logging systems
- Coreset for k-segments
- From coresets to text

Coresets for Travel: k-segments



Definition: k-spline

A k-spline is a sequence of k connected segments in R^d

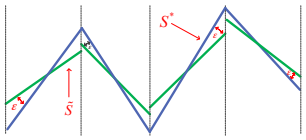


k=4

d=2

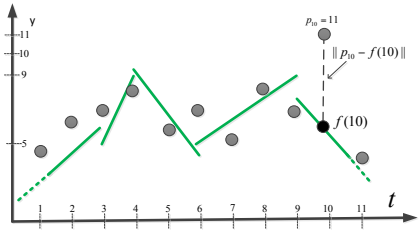
ϵ-approximation for $\mathrm{Opt}(P, k)$

A set $\tilde{S}$ of k segments is an ϵ-approximation for $\mathrm{Opt}(P, k)$ if rotating every segment of $\tilde{S}$ by angle of at most $\epsilon > 0$ yields an optimal k-spline S^* of P



k-segment Mean

The k-segment that minimizes the fitting costs from points to a d-dimensional signal

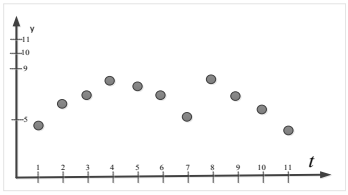


k=5

d=1

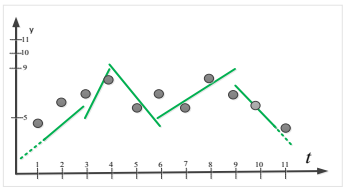
k-segment Queries

Input: *d*-dimensional signal *P* over time



k-segment Queries

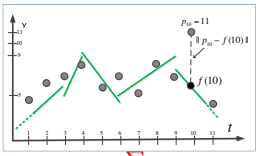
Input: *d*-dimensional signal *P* over time
Query: *k* segments over time



k-Piecewise linear function *f* over *t*

k-segment Queries

Input: *d*-dimensional signal *P* over time
Query: *k* segments over time
Output: Sum of squared distances from *P*

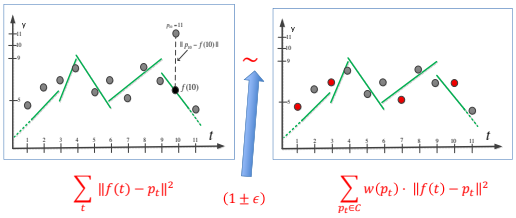


$$\text{cost}(P, f) := \sum_t \|f(t) - p_t\|^2$$

Coreset for k-segments

- A weighted set of segments C such as for every k-segment f

$$\text{cost}(P, f) \sim \text{cost}_w(C, f)$$



$$\sum_t \|f(t) - p_t\|^2 \quad (1 \pm \epsilon) \quad \sum_{p_t \in C} w(p_t) \cdot \|f(t) - p_t\|^2$$

Coreset Compression Theorem

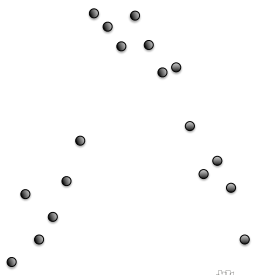
[ACM GIS'12, Feldman, Sung, and Rus]

For every discrete signal of n points in R^d there is a

- coreset of space $O(k/\varepsilon^2)$
- that can be computed in the big data model

Coreset Construction

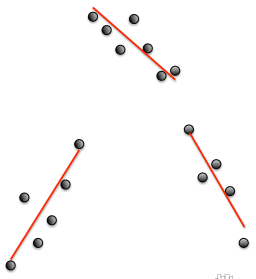
Input: signal of ***n*** points,
constants ***k***, ***ε***



Coreset Construction

Input: signal of n points,
constants k, ε

Compute k-segment mean

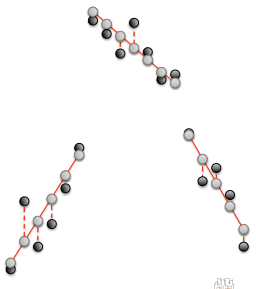


Coreset Construction

Input: signal of n points,
constants k, ε

Compute k-segment mean

Project points onto segments



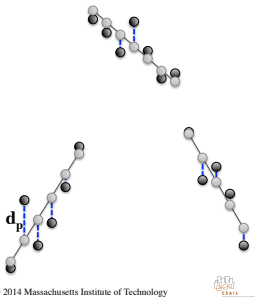
Coreset Construction

Input: signal of n points,
constants k, ε

Compute k-segment mean

Project points onto segments

Assign probability $\sigma_p = \frac{d_p}{\sum_p d_p}$



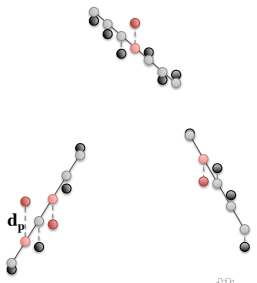
Coreset Construction

Input: signal of n points, constants k, ε

Compute k-segment mean

Project points onto segments

Sample $|S| \sim \frac{k}{\varepsilon^2}$ points with probability $\sigma_p = \frac{d_p}{\sum_p d_p}$



Coreset Construction

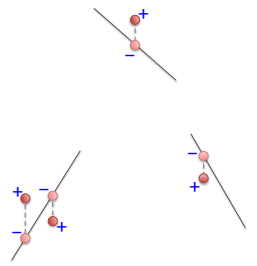
Input: signal of n points, constants k, ε

Compute k-segment mean

Project points onto segments

Sample $|S| \sim \frac{k}{\varepsilon^2}$ points with probability $\sigma_p = \frac{d_p}{\sum_p d_p}$

Assign weights $\pm \frac{1}{|S|\sigma_p}$



Coreset Construction

Input: signal of n points, constants k, ε, δ

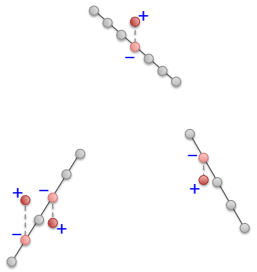
Compute k-segment mean

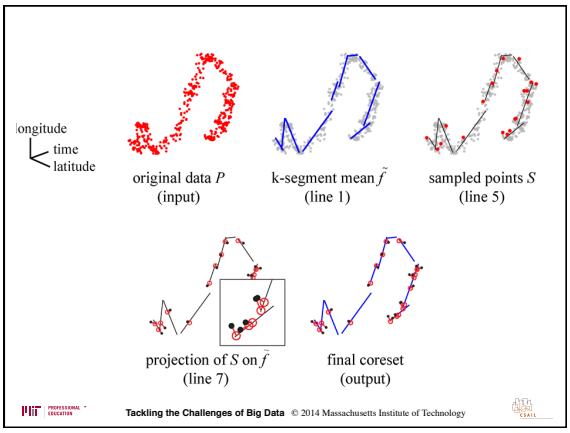
Project points onto segments

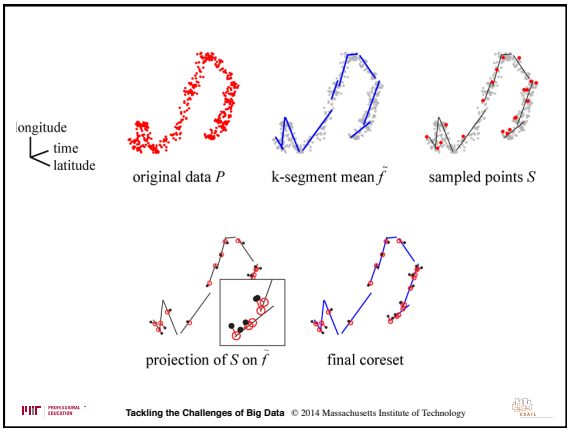
Sample $|S| \sim k/\varepsilon^2 (\log n + \log 1/\delta)$ points with probability $\sigma_p = d_p$

Assign weights $\pm 1/|S|\sigma_p$

Output: k segments and $|S|$ weighted points pairs







Tackling the Challenges of Big Data
Big Data Analytics
Data Compression

THANK YOU

PROFESSIONAL EDUCATION

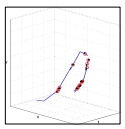
CSAIL

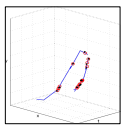
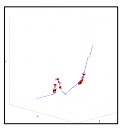
Tackling the Challenges of Big Data
Big Data Analytics

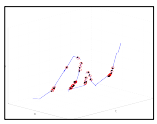
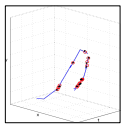
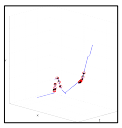
Daniela Rus
Professor
Director of CSAIL
Massachusetts Institute of Technology

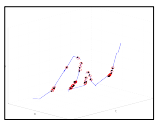
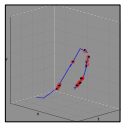
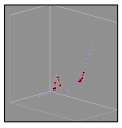
Tackling the Challenges of Big Data
Big Data Analytics
Data Compression

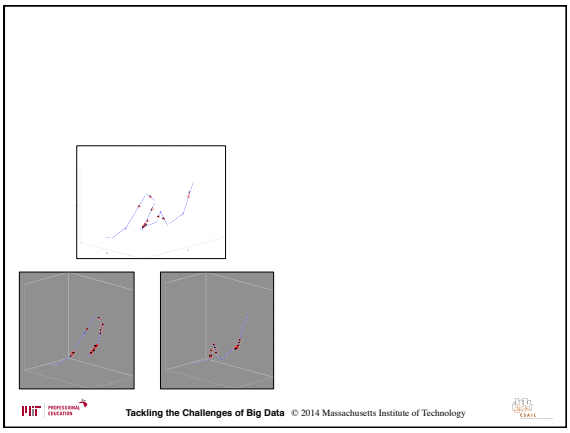
Daniela Rus
Professor
Director of CSAIL
Massachusetts Institute of Technology

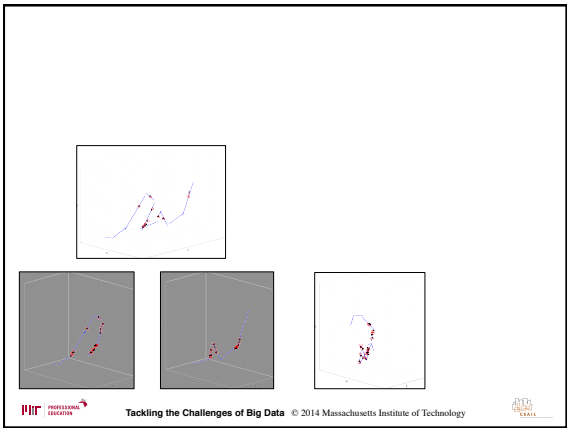


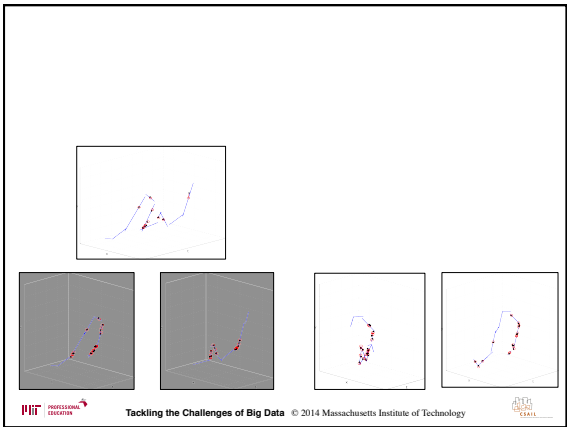



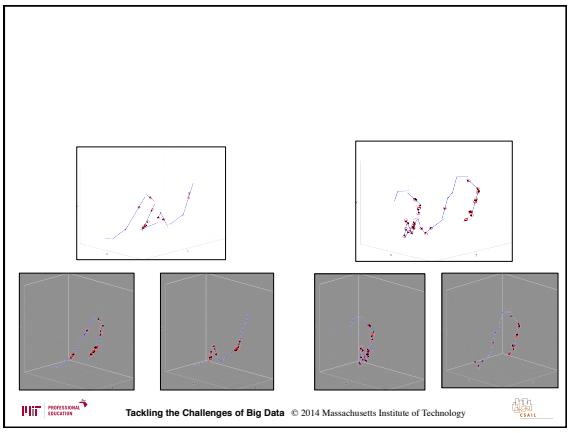




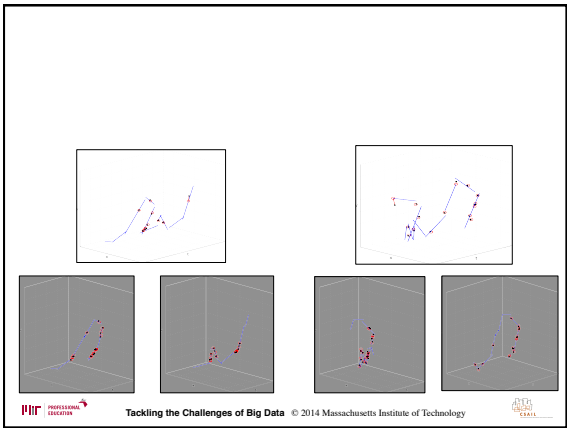




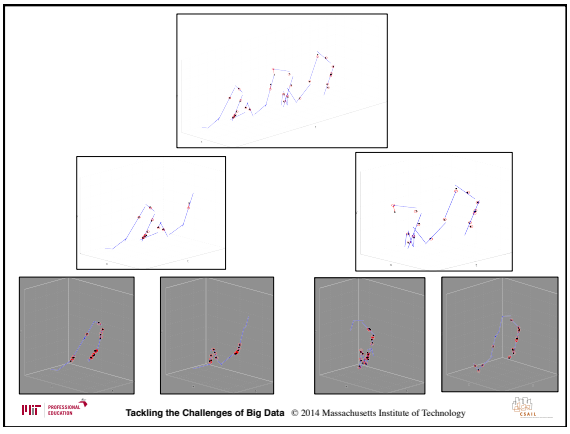


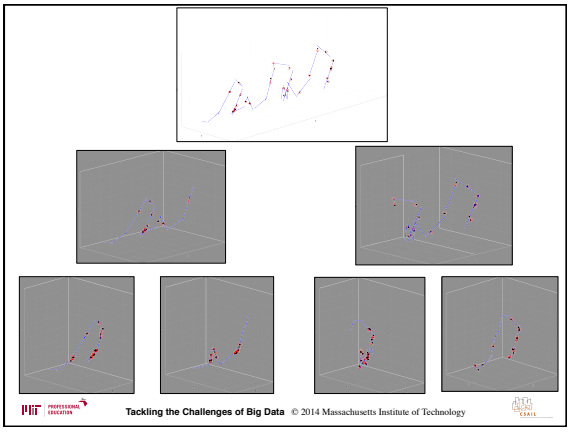


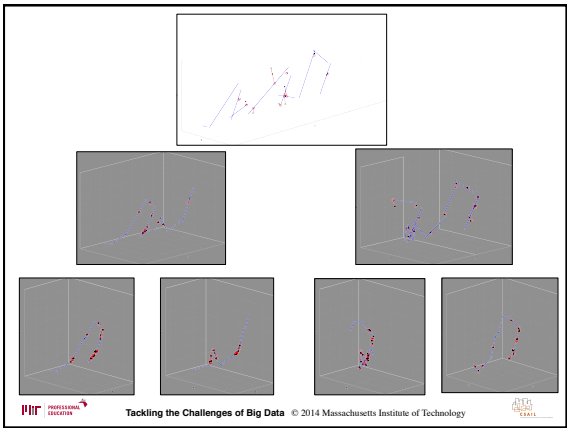



Tackling the Challenges of Big Data © 2014 Massachusetts Institute of Technology

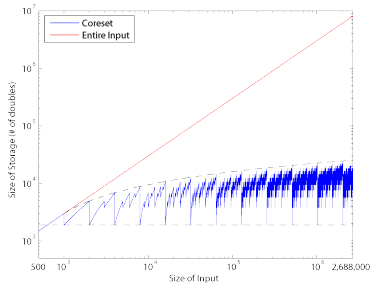

Tackling the Challenges of Big Data © 2014 Massachusetts Institute of Technology


Tackling the Challenges of Big Data © 2014 Massachusetts Institute of Technology





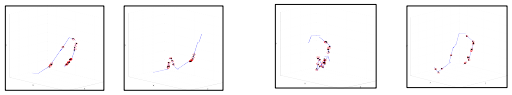
Coreset Storage vs Input Storage



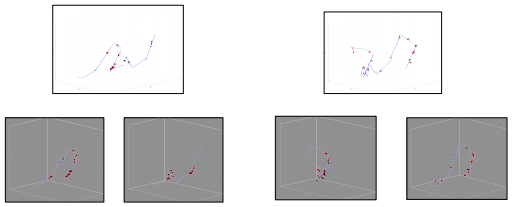
Computation

- We need to apply the coreset only on some points
- The construction time is $O(k^3)$ instead of $O(n^3)$
- If k^3 is still too long, or we don't know how to compute optimal solution
  - Use bi-criteria approximations
  - or heuristics

Parallel Computation

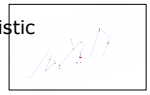


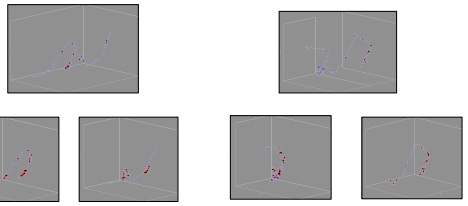
Parallel Computation



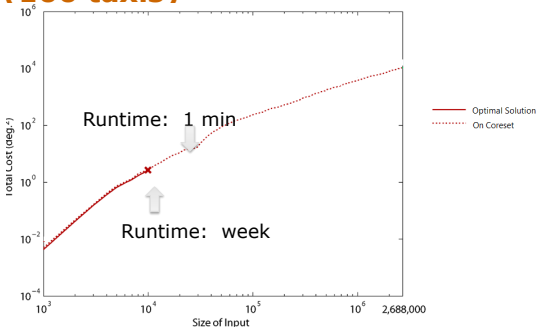
Parallel Computation

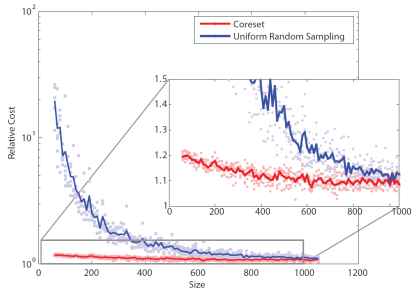
Run off-line heuristic on coreset





100-segment mean on GPS traces (100 taxis)





30-segment mean of SIGSPATIAL'12 data

Tackling the Challenges of Big Data
Big Data Analytics
Data Compression

THANK YOU

Tackling the Challenges of Big Data
Big Data Analytics

Daniela Rus
Professor
Director of CSAIL
Massachusetts Institute of Technology

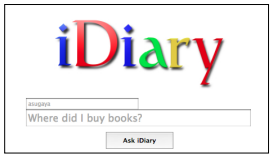
Tackling the Challenges of Big Data
Big Data Analytics
Data Compression

Daniela Rus
Professor
Director of CSAIL
Massachusetts Institute of Technology

Outline

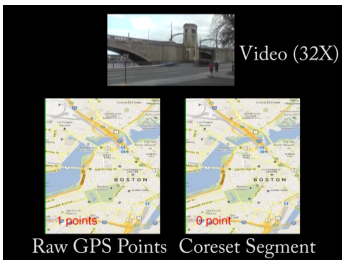
- Motivation
- Coreset definition, computation model
- Coresets for k-means
- Use case: life logging systems
- Coreset for k-segments
- From coresets to text

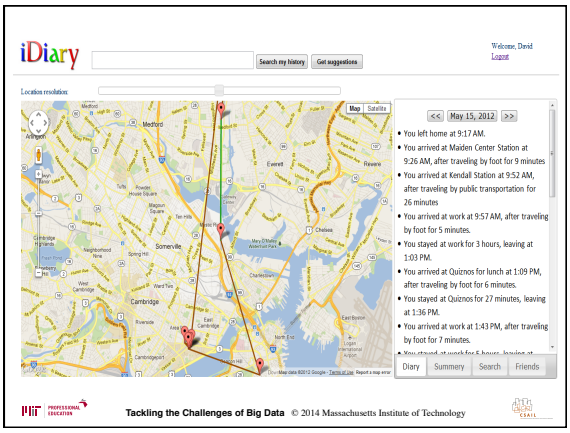
Example: Coresets for Life Logging

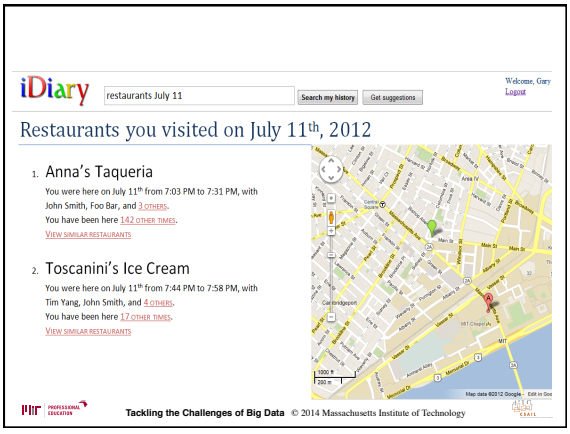


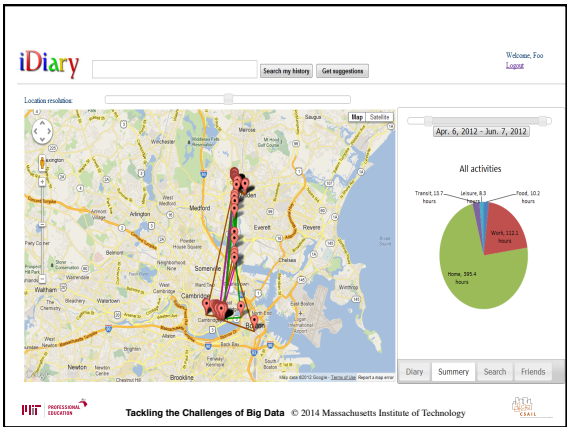
- Text search application on GPS data
- Other GPS managers:
  - Foursquare: manual check-ins
  - Google Latitude: no text search

Example Coreset Computation




iDiary
Search my history
Get suggestions
Welcome, David
Logout
Location resolution:
Map
Satellite
<< May 15, 2012 >>
• You left home at 9:17 AM.
• You arrived at Maiden Center Station at 9:26 AM, after traveling by foot for 9 minutes
• You arrived at Kendall Station at 9:52 AM, after traveling by public transportation for 26 minutes
• You arrived at work at 9:57 AM, after traveling by foot for 5 minutes.
• You stayed at work for 3 hours, leaving at 1:03 PM.
• You arrived at Quiznos for lunch at 1:09 PM, after traveling by foot for 6 minutes.
• You stayed at Quiznos for 27 minutes, leaving at 1:36 PM.
• You arrived at work at 1:43 PM, after traveling by foot for 7 minutes.
Diary
Summery
Search
Friends
Tackling the Challenges of Big Data © 2014 Massachusetts Institute of Technology


iDiary
restaurants July 11
Search my history
Get suggestions
Welcome, Gary
Logout
Restaurants you visited on July 11th, 2012
1. Anna's Taqueria
You were here on July 11th from 7:03 PM to 7:31 PM, with John Smith, Foo Bar, and 3 OTHERS.
You have been here 142 OTHER TIMES.
VIEW SIMILAR RESTAURANTS
2. Toscanini's Ice Cream
You were here on July 11th from 7:44 PM to 7:58 PM, with Tim Yang, John Smith, and 4 OTHERS.
You have been here 17 OTHER TIMES.
VIEW SIMILAR RESTAURANTS
Tackling the Challenges of Big Data © 2014 Massachusetts Institute of Technology


iDiary
Search my history
Get suggestions
Welcome, Foo
Logout
Location resolution:
Map
Satellite
Apr. 6, 2012 - Jun. 7, 2012
All activities
Diary
Summery
Search
Friends
Tackling the Challenges of Big Data © 2014 Massachusetts Institute of Technology

iDiary

user:timyang | Search my history | Get suggestions

Welcome, Gary
Logout

You and Tim Yang

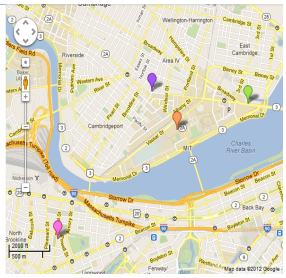
Recent encounters:

1. Toscanini's Ice Cream – July 11th
2. Home – July 7th
3. Home – July 1st
4. MIT Student Center – June 25th
5. 123 Main Street – June 24th
6. Home – June 23rd

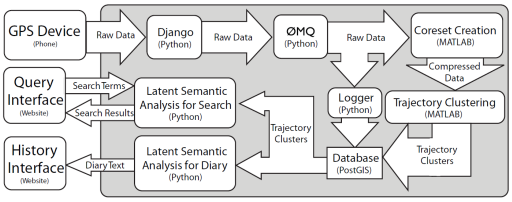
See more...

Time analysis

1. Home – 5 hours/week
   - Weekend evenings
2. Leisure – 3 hours/week
   - Weekday afternoons



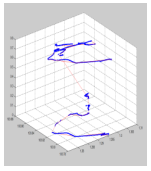
System Overview



Coreset Computation

- Approximate the input GPS-points by connected segments using a k-spline
- Output the text description of the endpoints (e.g., using Google Maps)

time	latitude	longitude
8:44:57	1.295783	103.7816
8:44:59	1.295785	103.7816
8:45:00	1.295782	103.7816
8:45:01	1.295782	103.7816
8:45:04	1.29579	103.7817
8:45:05	1.295802	103.7817
8:45:08	1.295915	103.7818
8:45:09	1.29598	103.7819
8:45:10	1.296015	103.7819
8:45:11	1.296057	103.782
...	...	...



$k = 20$

Central Expy, Singapore
Ayer Rajah Expy, Singapore
Chin Swee Rd, Singapore
261 Outram Rd, Singapore 169057
1 St Andrew's Rd, Singapore 178957
100A Havelock Rd, Singapore 169664
6A Raffles Ave, Singapore 039801
7 Raffles Blvd, Singapore 039595
8 Buona Vista Rd, Singapore
1 Lower Kent Ridge Rd, Singapore
8 Mallicol Dr, Singapore 117504
20 Layone Hill, Singapore
113 Dorsetshire Rd, Singapore 239078
121 Devonshire Rd, Singapore 239882
53 Grange Rd, Singapore
52 Grange Rd, Singapore 239700
Natl Youth Council, Singapore
62K Paterson Rd, Singapore 238517
521 Orchard Rd, Singapore 238866
520 Orchard Rd, Singapore 238852

$k+1 = 21$

From Signals to Text

m locations

Reverse Geo-coding

Location ID	Begin Point	End Point
	(42.374,-71.120)	(42.374,-71.120)
2	(42.386, -71.130)	(42.386,-71.130)
3	(42.391,-71.128)	(42.391,-71.128)
4	(42.393,-71.130)	(42.394,-71.129)
5	(42.385,-71.132)	(42.384,-71.130)
6	(42.358, -71.091)	(42.358,-71.098)
...	...	...

Google maps, mapQuest …

Starbucks, Harvard Square
225 Walden St.
Peabody Preschool
55-92 Rice St.
128-157 Garden St.
130-169 Vassar St, Cambridge
...

Tackling the Challenges of Big Data
Big Data Analytics
Data Compression

THANK YOU

PROFESSIONAL EDUCATION

CSAIL

Tackling the Challenges of Big Data
Big Data Analytics

Daniela Rus

Professor

Director of CSAIL

Massachusetts Institute of Technology

PROFESSIONAL EDUCATION

CSAIL

Tackling the Challenges of Big Data
Big Data Analytics
Data Compression

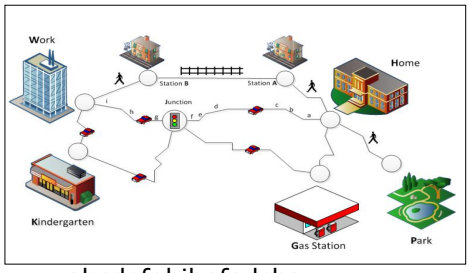
Daniela Rus
Professor
Director of CSAIL
Massachusetts Institute of Technology

Outline

- Motivation
- Coreset definition, computation model
- Coresets for k-means
- Use case: life logging systems
- Coreset for k-segments
- From coresets to text

Application: Travel Patterns

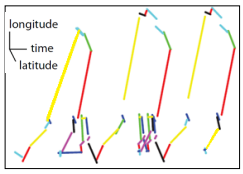
(k,m)- Hidden Markov Model.
Chain of length *k*, between *m* states

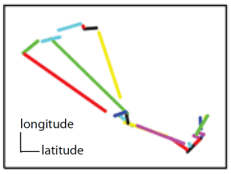


abcdefghihgfedcba…

(k,m) Formulation

Minimizes cost over every case-segment combination whose projection is only *m* segments

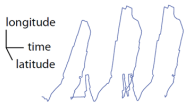




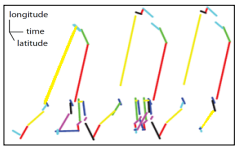
Travel Pattern Input

- GPS-point = (latitude, longitude, time)

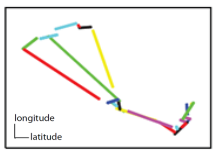
latitude	longitude	time
1.295783	103.7816	8:44:57
1.295785	103.7816	8:44:59
1.295787	103.7816	8:45:00
1.295787	103.7816	8:45:01
1.29577	103.7817	8:45:04
1.295883	103.7817	8:45:05
1.295915	103.7818	8:45:08
1.2958	103.7819	8:45:09
1.296015	103.7819	8:45:10
1.296085	103.782	8:45:11
...	...	...



Travel Pattern Output



k trajectories



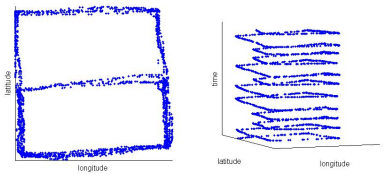
m locations

Begin time	End time	Location ID	Speed
8:44:57	8:48:57	c	30
8:49:59	8:51:59	d	24
8:52:00	8:54:00	g	24
8:54:01	8:55:01	q	11
8:56:57	8:57:57	r	120
8:58:57	8:59:57	m	55
...	...	...	65

Location ID	Begin Point	End Point
a	(42.374,-71.120)	(42.374,-71.120)
b	(42.386, -71.130)	(42.386,-71.130)
c	(42.391,-71.128)	(42.391,-71.128)
d	(42.393,-71.130)	(42.394,-71.129)
...	...	...

A GPS signal: a sequence of geographic points progressing through time

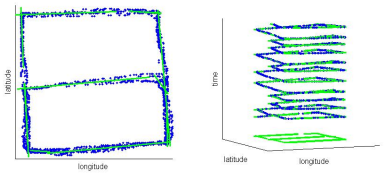
Derived from an entity's repeated traversal of a set of distinct geographic paths



But where are those paths?

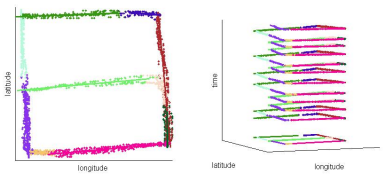
The human eye can make a good guess

Can an algorithm do the same?

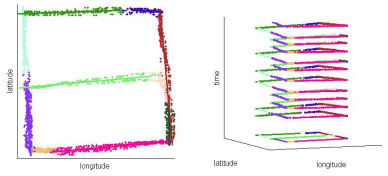


A (k, m)-segment trajectory:

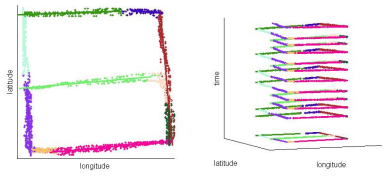
- Partition the signal into k sections
- Group the sections into m clusters
- Assign a line segment to each cluster



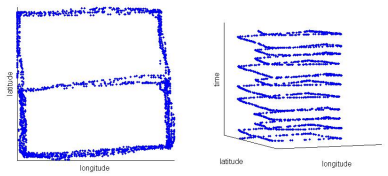
The (k, m)-segment mean:
- A (k, m)-segment trajectory
- Minimizes the point signal's fitting cost



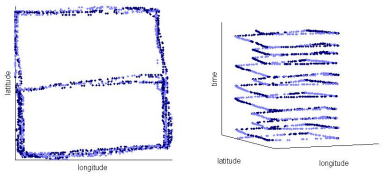
Our (k, m)-segment mean algorithm:
- Approximates the true mean
- Uses EM to find local minimum of fitting cost
- Efficient in runtime and space



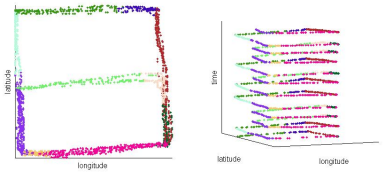
Start with the GPS point signal
No external environment data (e.g. road maps)



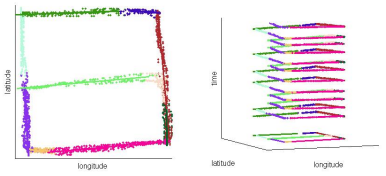
Partition the signal into k distinct sections
Use Ramer-Douglas-Peucker for a good first guess (it targets high-cost points)



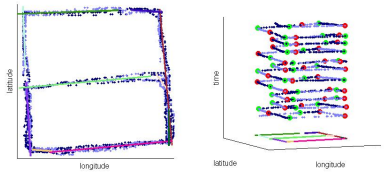
Group k sections into m clusters (colors)
Use k-means for a good first guess (clustering according to endpoint locations)



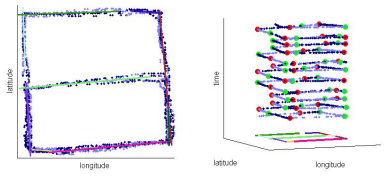
Calculate the segment mean of each cluster
Using linear algebra makes it exact (non-approximate)



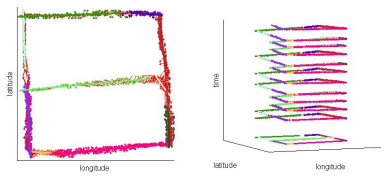
Hold fixed the odd partition boundaries (red)
Given the m segments, optimally adjust the even boundaries (green) and the section cluster assignments



Hold fixed the even partition boundaries (red)
Given the m segments, optimally adjust the odd boundaries (green) and the section cluster assignments



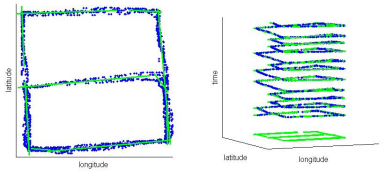
With the updated partition and clusters, recalculate each cluster's segment mean, and the total fitting cost (red lines)
If cost is reduced repeat; else terminate



Upon termination, the locally optimal (k, m)-segment trajectory has been found

Line segments give the map

Clustering gives trajectory on map



Two domains of algorithmic effectiveness

- Accuracy (minimal error)
- Speed (minimal runtime)

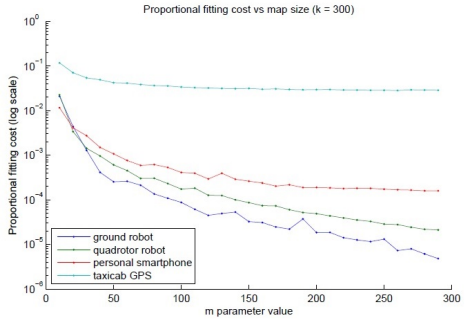
Measure of accuracy: proportional fitting cost

- Average squared error distance per input point, relative to points' variance (squared error from mean)
- Independent of input's set size and geographic magnitude
- In range [0, 1] for exact (k, m)-segment mean

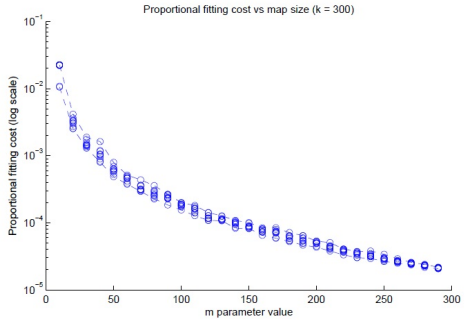
Measure of speed: runtime per point

- Average runtime per point
- Independent of input's set size and geographic magnitude

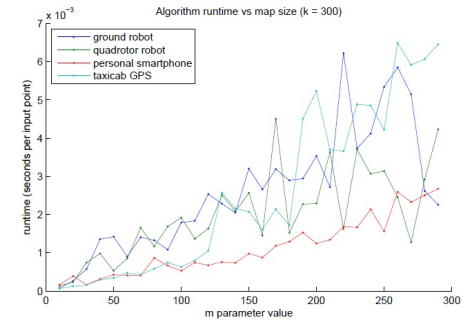
Accuracy across four sample data sets



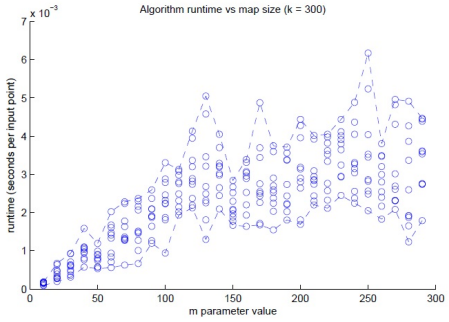
Accuracy across ten runs of a data set



Speed across four sample data sets

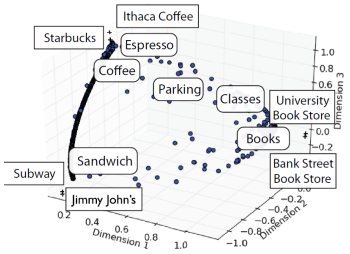


Speed across ten runs of a data set



Coresets for Latent Semantic Analysis

[Feldman, Sohler, SODA'13]



Wrapping Up

- Instead of Speeding Algorithms, Find Right Data
- Coresets provide semantic compression for data
- Sample coreset algorithms
- Applications for personal travel management, transportation

Tackling the Challenges of Big Data
Big Data Analytics
Data Compression

THANK YOU