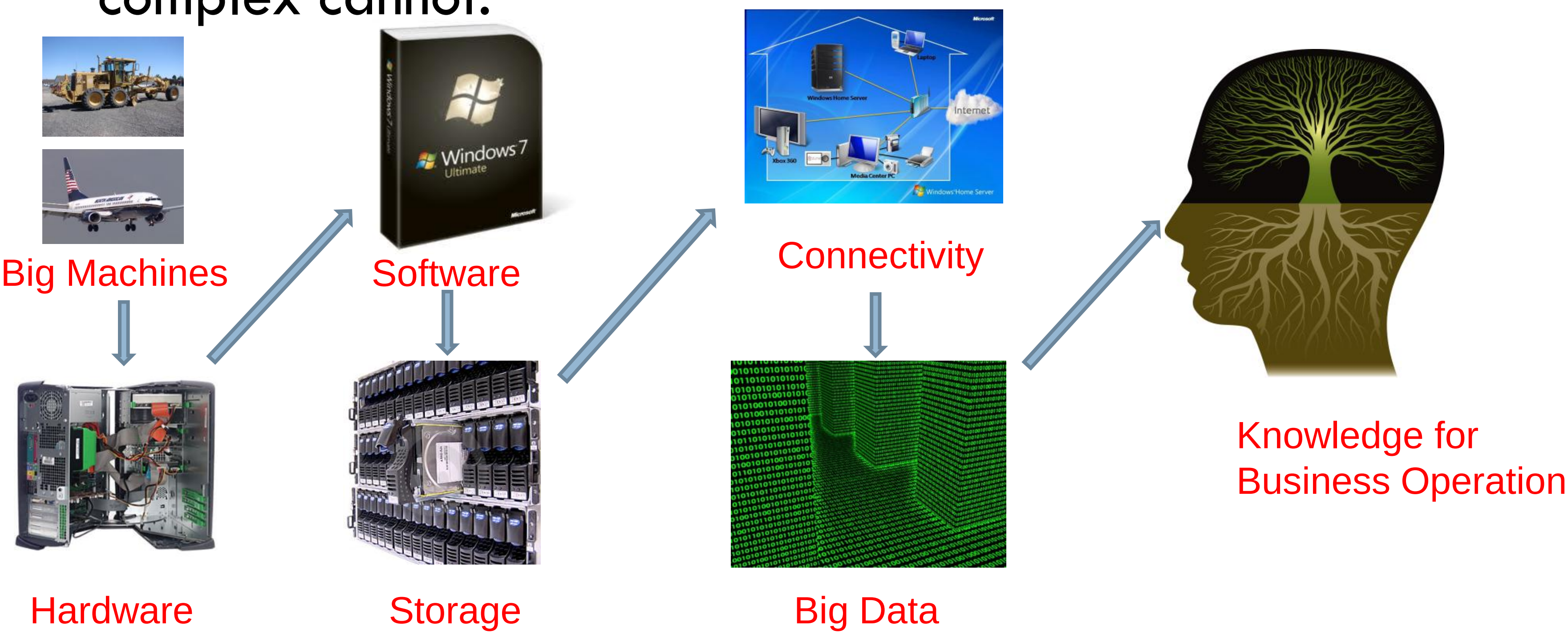
Big Data Analytics in Mobile Environments

熊辉 教授
罗格斯-新泽西州立大学

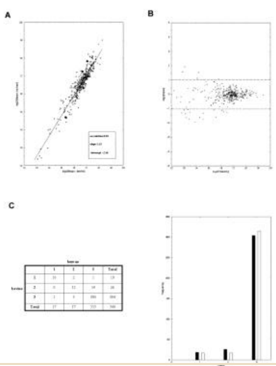
Why big data: historical view?

- Productivity versus Complexity (interrelatedness, ambiguity)
- Complex versus Complicated
  - While the complicated can be unfolded for analysis, the complex cannot.

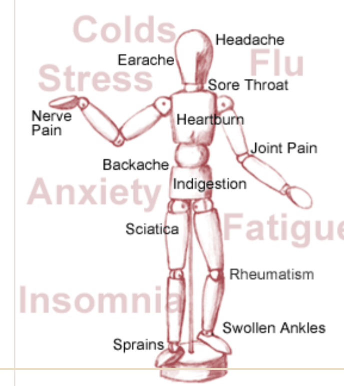


Similarities Between Data Miners and Doctors

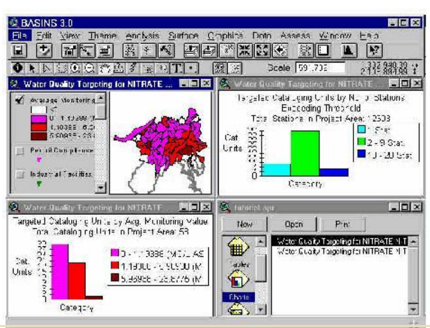
Very Often, No Standardized Solutions



Data Characteristics



Your Symptoms?



Data Mining Techniques

Medical Devices

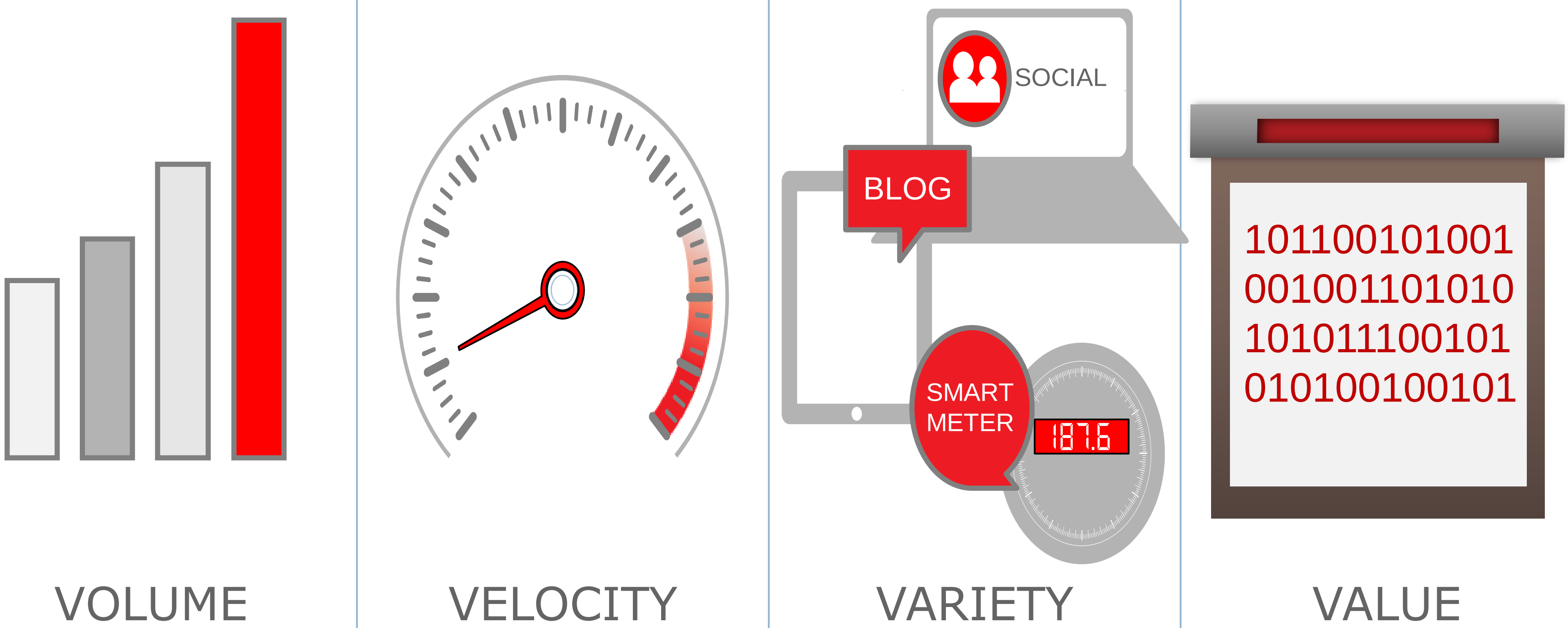
So What is Big Data?

THE STATE UNIVERSITY OF NEW JERSEY
RUTGERS

Big Data refers to datasets that grow so large that it is difficult to capture, store, manage, share, analyze and visualize with the typical database software tools.

What Makes it Big Data?

THE STATE UNIVERSITY OF NEW JERSEY
RUTGERS



"Big" is also a relative concept.

Data Size / Solution-Time-Window >= Computing Capacity Per Time Unit

Big Data Use Cases

THE STATE UNIVERSITY OF NEW JERSEY
RUTGERS

Today's Challenge	New Data	What's Possible
Healthcare Expensive office visits Hospital Dynamics	Remote patient monitoring, Hospital Sensors	Preventive care, reduced hospitalization, reduced human mistakes
Manufacturing In-person support	Product sensors	Automated diagnosis, customized support
Location-Based Services Based on home zip code	Real time location data	Geo-advertising, urban computing, mobile recommendation
Finance Fast-paced, Variety	Social Media, High-frequency Trading Data	Sentiment analysis Finance engineering
Retail One size fits all marketing	Market basket data, user behavior logs	Personalized Recommendation, Segmentation

10 Ways Mobile Tech Is Changing Our World

THE STATE UNIVERSITY OF NEW JERSEY
RUTGERS

1. Elections Will Never Be The Same

1. Doing Good By Texting

2. Bye-Bye, Wallets

3. The Phone Knows All

4. Your Life Is Fully Mobile

5. The Grid Is Winning

6. A Camera Goes Anywhere

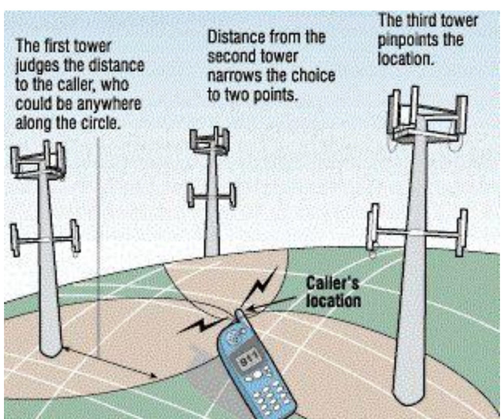
7. Toys Get Unplugged

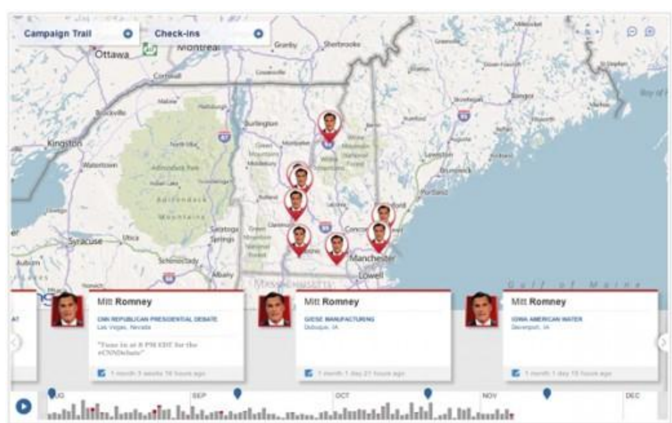
8. Gadgets Go To Class

9. Disease Can't Hide

Human Mobility

- Human mobility is people's movement trajectories which can be
  - Phone traces or trajectories of driving routes
  - a sequences of posts (like geo-tweets, geo-tagged photos, or check-ins)
- Indoor Traces and Outdoor Traces.

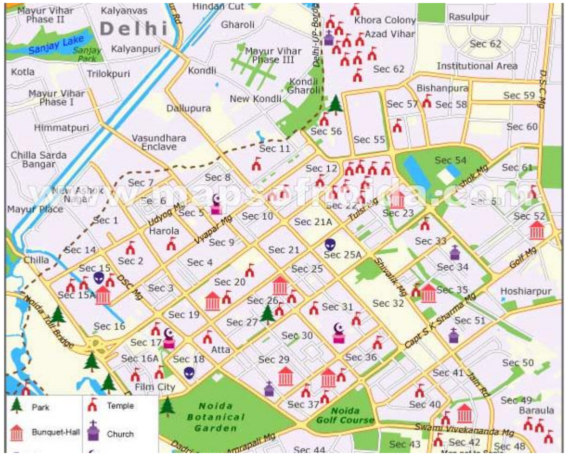




Urban Geography

THE STATE UNIVERSITY OF NEW JERSEY
RUTGERS

- Urban geography is a set of geographic characteristics of a city including
  - road networks, public transportation
  - places of interest (POIs), regional functions



Public transportation data

THE STATE UNIVERSITY OF NEW JERSEY
RUTGERS

Table 2: Statistics of Transportation Data

Dataset	Year	2011
Bus Stop	num of bus stop	9810
	num of buses	28,343
	num of lines	948
	length of lines (km)	187,453
	total kms travelled (km)	1,753,000,000
	total passengers traffic	4,888,380,000
Subway	num of subway station	215
	num of lines	15
	length of lines(km)	339.5
	average traffic/day(million)	5.1
Road Network	num of road segments	162,246
	percentage of major roads	0.189
	num of formal regions	554

Point of Interests (POI)

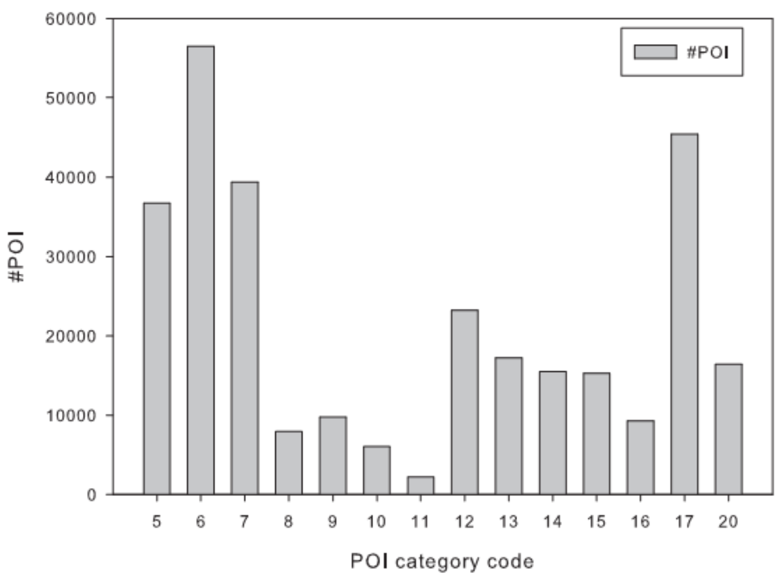


Figure 5: POI distribution over different category

POI code	POI category	POI code	POI category
1	car service	16	banking and insurance service
2	car sales	17	corporate business
3	car repair	18	street furniture
4	motorcycle service	19	entrance/bridge
5	café/tea Bar	20	public utilities
6	sports/stationery shop	21	chinese restaurant
7	living service	22	foreign restaurant
8	sports	23	fastfood restaurant
9	hospital	24	shopping mall
10	hotel	25	convenience store
11	scenic spot	26	electronic products store
12	residence	27	supermarket
13	governmental agencies and public organizations	28	furniture building materials market
14	science and education	29	pub/bar
15	transportation facilities	30	theaters

Outdoor Location Traces

- Taxi GPS trajectories

Table 3: Statistics of Mobility Data

Dataset	Year	2011
Taxi trajectories	num of taxi	13,597
	num of occupied trips	8,202,012
	num of effective days	92
	average trip distance(km)	7.47
	average trip duration(min)	16.1
	average sampling interval (sec)	70.45

Data Miners in Big Data Analytics

THE STATE UNIVERSITY OF NEW JERSEY
RUTGERS

Big Data Analytics
Understand goals of business
Collaborate in interdisciplinary teams
Integrate large volumes of structured and unstructured data
Formulate problems, develop solutions
Blend statistical modeling, data mining, forecasting, optimization
Develop/run integrated software solutions
Gain higher visibility

Change business operation

Big Data Application Requirements

- Timely observation

- Timely analysis

$$\alpha+\lambda\beta$$

- Timely solution

Big Data Application Trends

Personalized Recommendation

Urban Computing

Micro-Scope

Classic

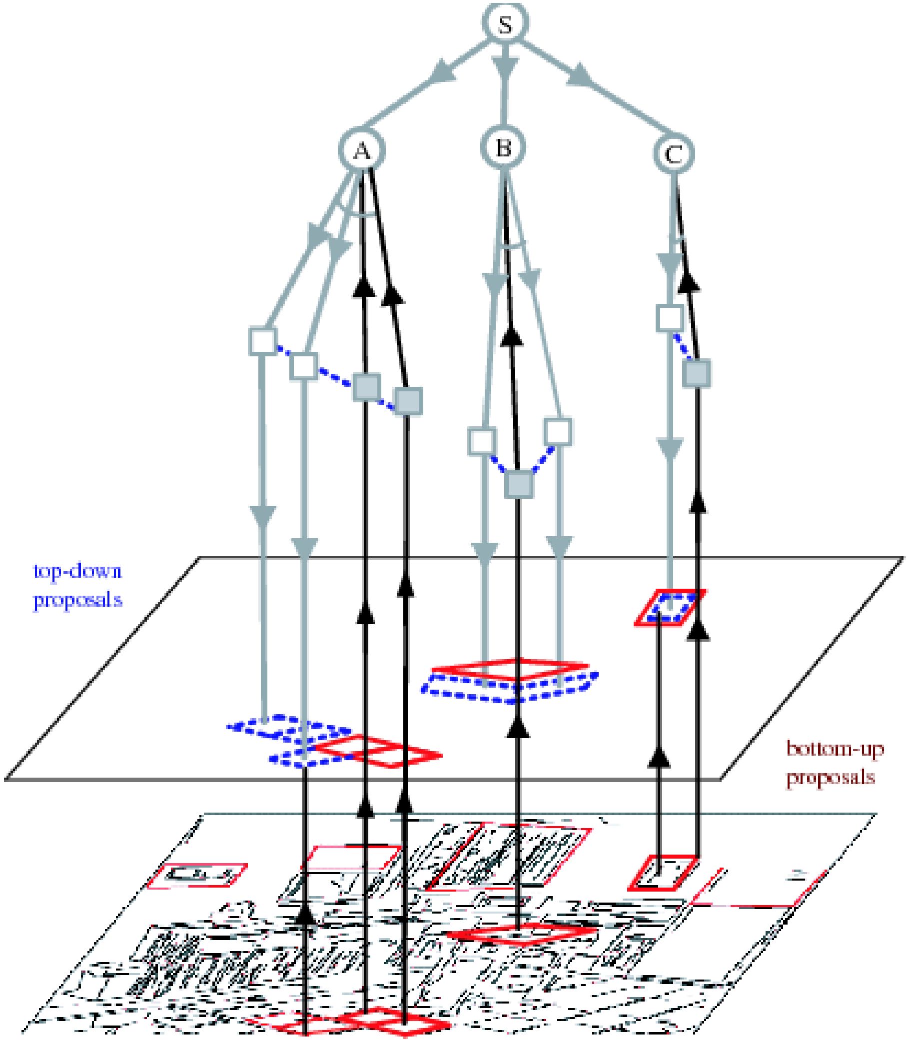
Macro-Scope

Applications/Techniques

Data Driven Solutions

THE STATE UNIVERSITY OF NEW JERSEY RUTGERS

- Theoretical top-down solutions
- Data driven bottom-up solutions



Big Data Application Requirements

Big Data Experiences

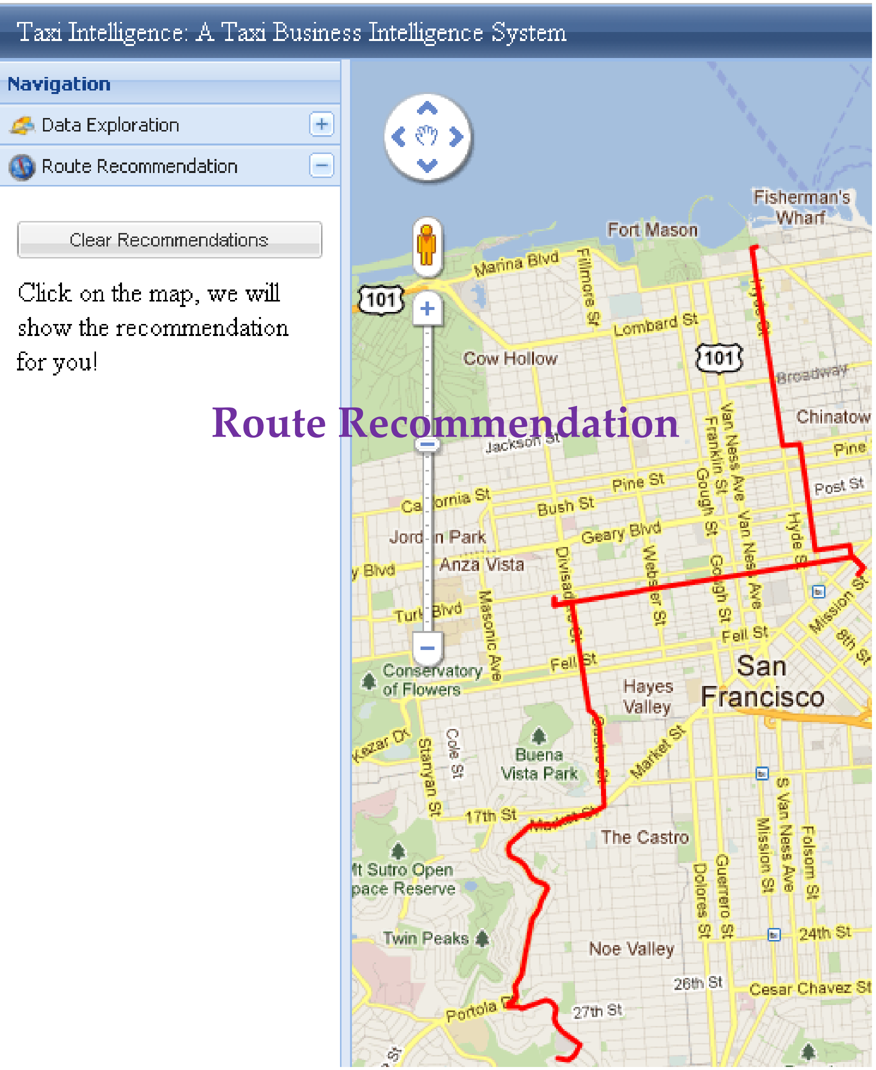
THE STATE UNIVERSITY OF NEW JERSEY RUTGERS

- Understanding Data Characteristics
  - Data Distribution, Data Quality etc.
- Feature Engineering
  - Feature engineering is one of the key strategy for the success of big data analytics.
  - The goal is to explicitly reveal important information to the model by feature selection or feature generation
  - Original features → different encoding of the features → combined features
- Instance Selection (particularly mobile environment)
  - The goal is to select the right instances/objects for the underlying data analytics

- Background
  - Revolution in Mobile Devices
    - GPS
    - WiFi
    - Mobile phone
  - The Urgent Demand for Better Service
    - Driving route suggestion
    - Mobile tourist guides
- Definition
  - Mobile pervasive recommendation is promised to provide mobile users access to personalized recommendations anytime, anywhere.

Mobile Recommender Systems

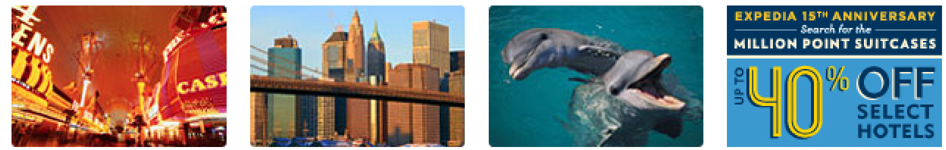
THE STATE UNIVERSITY OF NEW JERSEY
RUTGERS



Route Recommendation

Expedia

FEATURED VACATION PACKAGE DEALS



Last Minute Las Vegas — from $369 — Flight + Hotel

Last Minute NY — from $503 — Flight + Hotel

Great Savings in Orlando — from $283 — Flight + Hotel

Anniversary Sale Deals — from $312 — Flight + Hotel

Travel Recommendation

TOP VACATION PACKAGE DEALS

Top Destinations | Last-minute Deals | Current Promotions

Choose your departure: Atlanta

Airport	Destination	Travel Dates	Nights	Rating	Flight + Hotel Per Person
ATL	Las Vegas	03 Nov - 07 Nov	4	★★★☆☆	$498
ATL	Cancun	03 Nov - 07 Nov	4	★★★☆☆	$485
ATL	Montego Bay	03 Nov - 07 Nov	4	★★★☆☆	$592
ATL	Manhattan	03 Nov - 07 Nov	4	★★★½☆	$832
ATL	Miami	03 Nov - 07 Nov	4	★★★☆☆	$297
ATL	Riviera Maya	03 Nov - 07 Nov	4	★★★☆☆	$514
ATL	Orlando	03 Nov - 07 Nov	4	★★★☆☆	$342
ATL	Los Angeles	03 Nov - 07 Nov	4	★★★☆☆	$453
ATL	Honolulu	03 Nov - 07 Nov	4	★★★☆☆	$938
ATL	Chicago	03 Nov - 07 Nov	4	★★★☆☆	$431

Challenges for Mobile Recommendation (I)

- Complexity of the Mobile Data
  - Heterogeneous
  - Spatial and temporal auto-correlation
  - Noisy
- The Validation Problem
  - No Ratings
- The Generality Problem
  - Different application domains with different recommendation techniques

Challenges for Mobile Recommendation (II)

THE STATE UNIVERSITY OF NEW JERSEY
RUTGERS

- The Cost Constraints
  - Time
  - Price

- The Life Cycle Problem
- The Transplantation Problem
  - Difficult to apply traditional Recommendation techniques for mobile recommendation

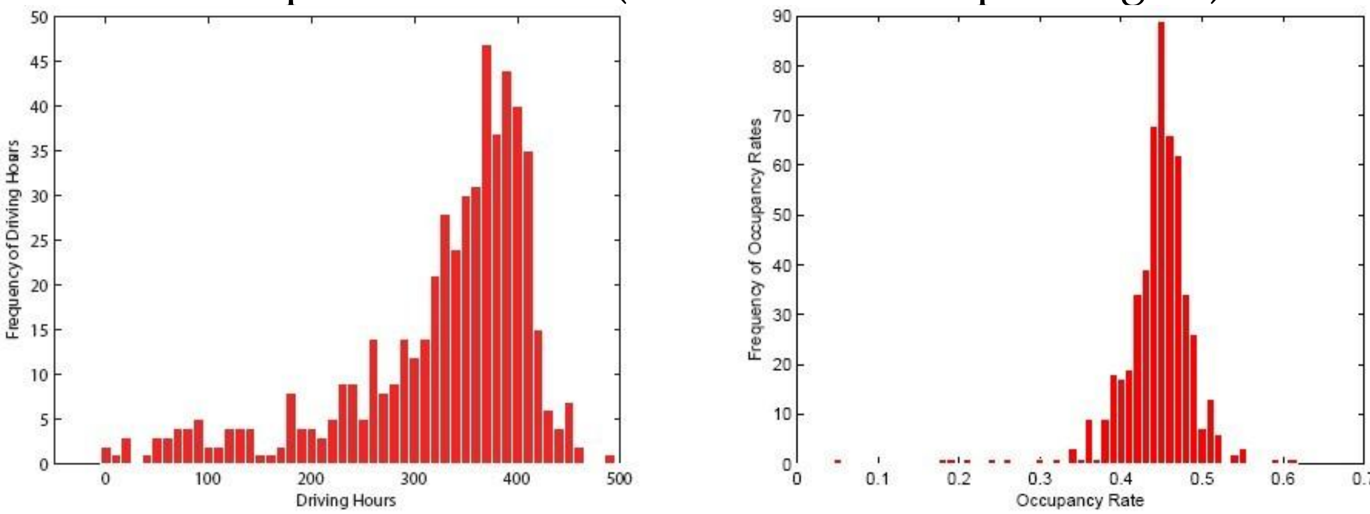
The Characteristics of Mobile Data

THE STATE UNIVERSITY OF NEW JERSEY
RUTGERS

- Two Cases
  - Case 1. Location trace by taxi drivers
  - Case2. The tourism data
- Why?
  - A good coverage of unique characteristics of mobile data
  - Can be naturally exploited for developing mobile recommender systems
  - They are the real-world data

The Characteristics of Mobile Data

- Case 1. Location trace by taxi drivers
  - Data Description
    - GPS traces
      - Location information (Longitude, Latitude), timestamp
      - The operation status (with or without passengers)

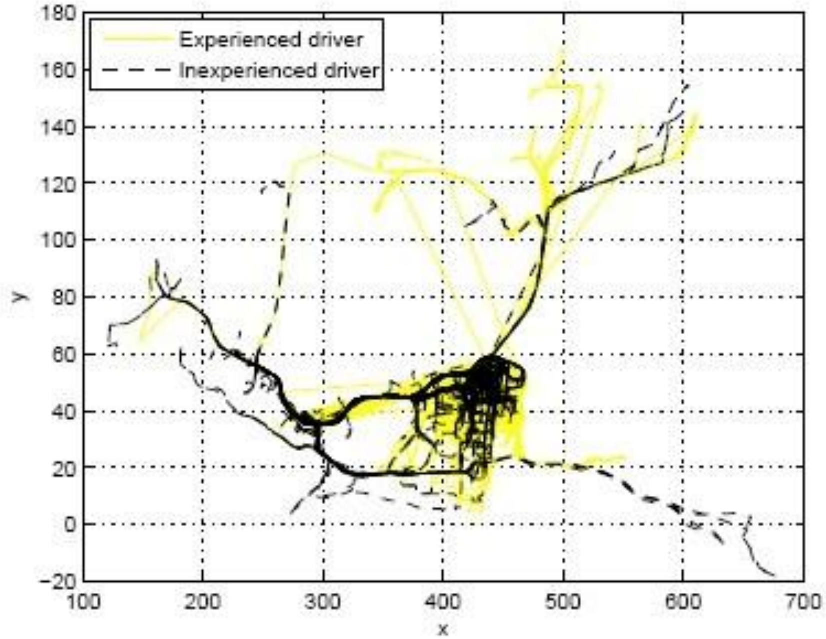


(a) Driving Hours (b) Occupancy Rates

- Experienced drivers can usually have more driving hours and high occupancy rates
- Inexperienced drivers tend to have less driving hours and low occupancy rates

The Characteristics of Mobile Data

- Case 1. Location trace by taxi drivers
  - Driving pattern comparison



A Comparison of Trajectories between an Experienced Driver and an Inexperienced Driver.

- The experienced drivers have a wider operation area.
- The experienced drivers know the roads as well the traffic patterns better.

The Characteristics of Mobile Data

- Case 1. Location trace by taxi drivers
  - Develop a mobile recommender system
    - Users ~ Taxi drivers
    - Items ~ Potential pick-up points

- What did we learn?
  - The difference between Mobile RS and traditional RS
    - The items are application-dependent
    - There is some cost to extract items
    - The items are not i.i.d while spatial auto-correlation

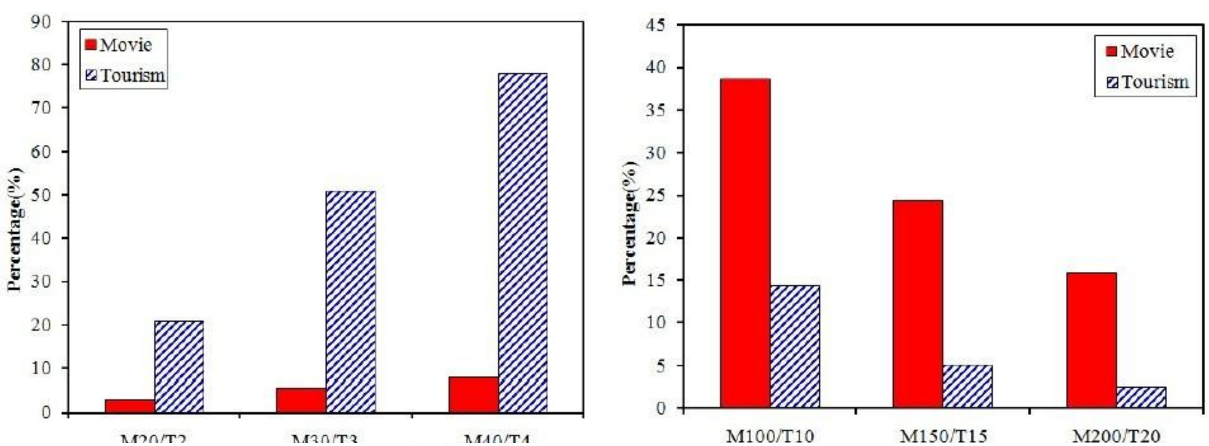
An Illustration of Pick-up Points.

The Characteristics of Mobile Data

- Case2. The travel data
  - Data Description
    - Expense records
      - Tourists: ID, travel time
      - Package: ID, name, landscapes, price, travel days
      - Duration: 2000—2010
  - Recommender System
    - Users ~ Tourists
    - Items ~ Packages

The Characteristics of Mobile Data

- Case2. The travel data
  - Characteristics of Tourism Data (I)
    - Spatial auto correlation of packages
      - For example, the 1-day Niagara Falls Tour
    - The Sparseness

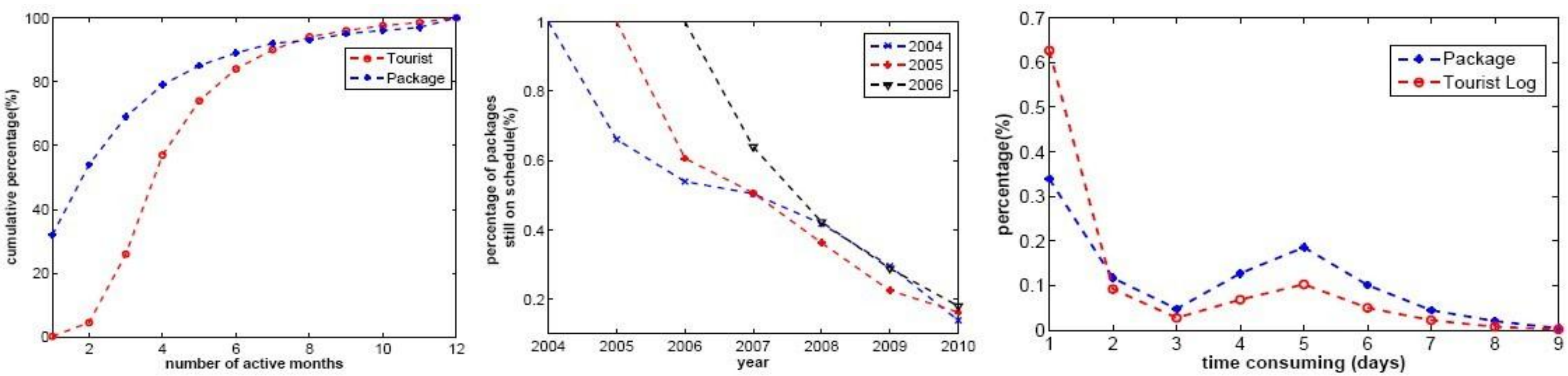


A comparison of the data sparseness between the movie data and the tourism data. (a) The percentage of users/tourists whose co-rating movies/co-traveling packages with their nearest neighbors are no more than 20, (30, 40 for the movie users)/(2, 3, 4 for the tourists). (b) The percentage of users/tourists whose rated movies are more than 100, 150, 200 in all movie users/whose traveling logs are 10, 15, 20 in all tourists, respectively.

The Characteristics of Mobile Data

THE STATE UNIVERSITY OF NEW JERSEY
RUTGERS

- Case2. The travel data
  - Characteristics of Tourism Data(II)
    - The time dependence
      - Packages and tourists have seasonal tendency
      - Packages have a life cycle



(a) (b) (c)

The illustration of the time-dependence of the tourism data. (a) The distribution of cumulative percentages of packages/tourists by the number of their active months in a year; (b) The percentage of remaining packages in the following several years after they have been introduced; (c) The percentage of different packages and tourist logs according to their travel days.

Theoretical Abstraction

- Given: a set of objects $O=\{O_1, O_2, \ldots, O_n\}$
- Find:

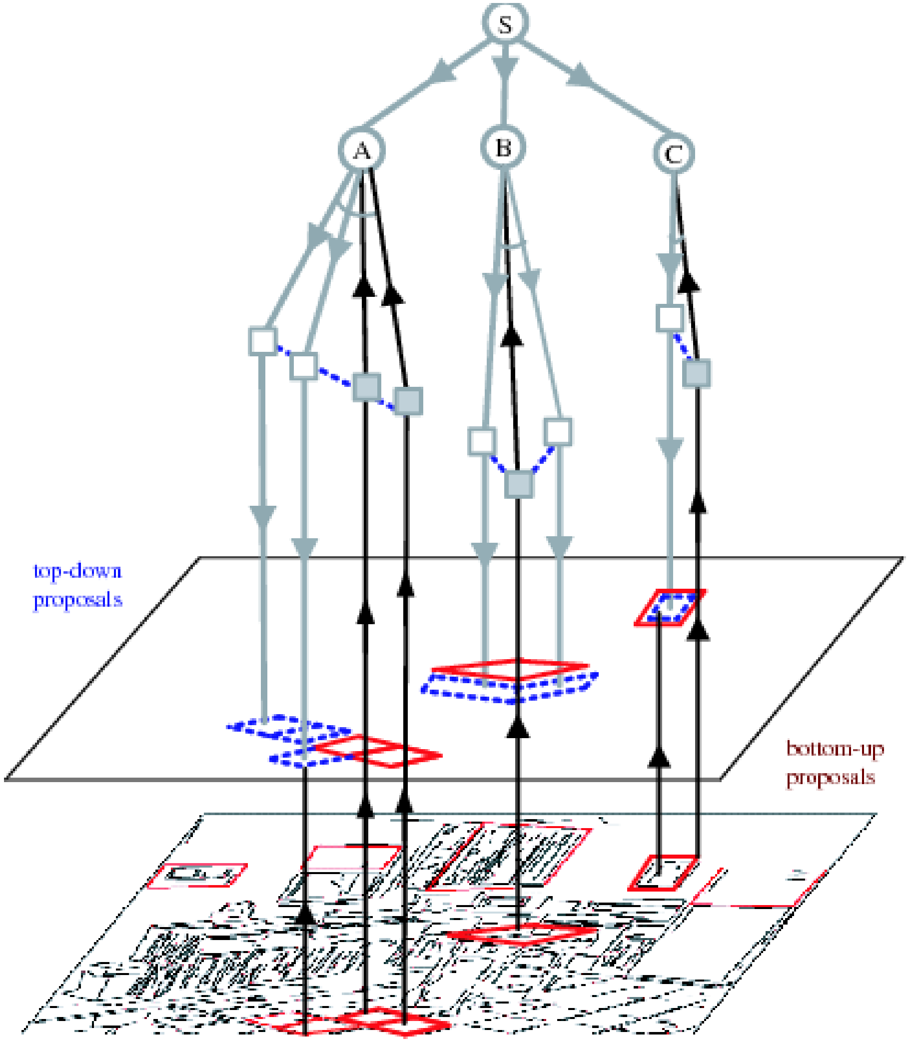
An ordered subset $S=\{S_1, S_2, \ldots, S_k\} \subseteq O$

The order of $S_1, S_2, \ldots, S_k$ is optimized subject to certain constraints.

- For taxi driver recommendation, the set O is a set of potential pick-up points
- For travel package recommendation, the set O is a set of landscapes.

Data Driven Solutions

- Theoretical top-down solutions
- Data driven bottom-up solutions

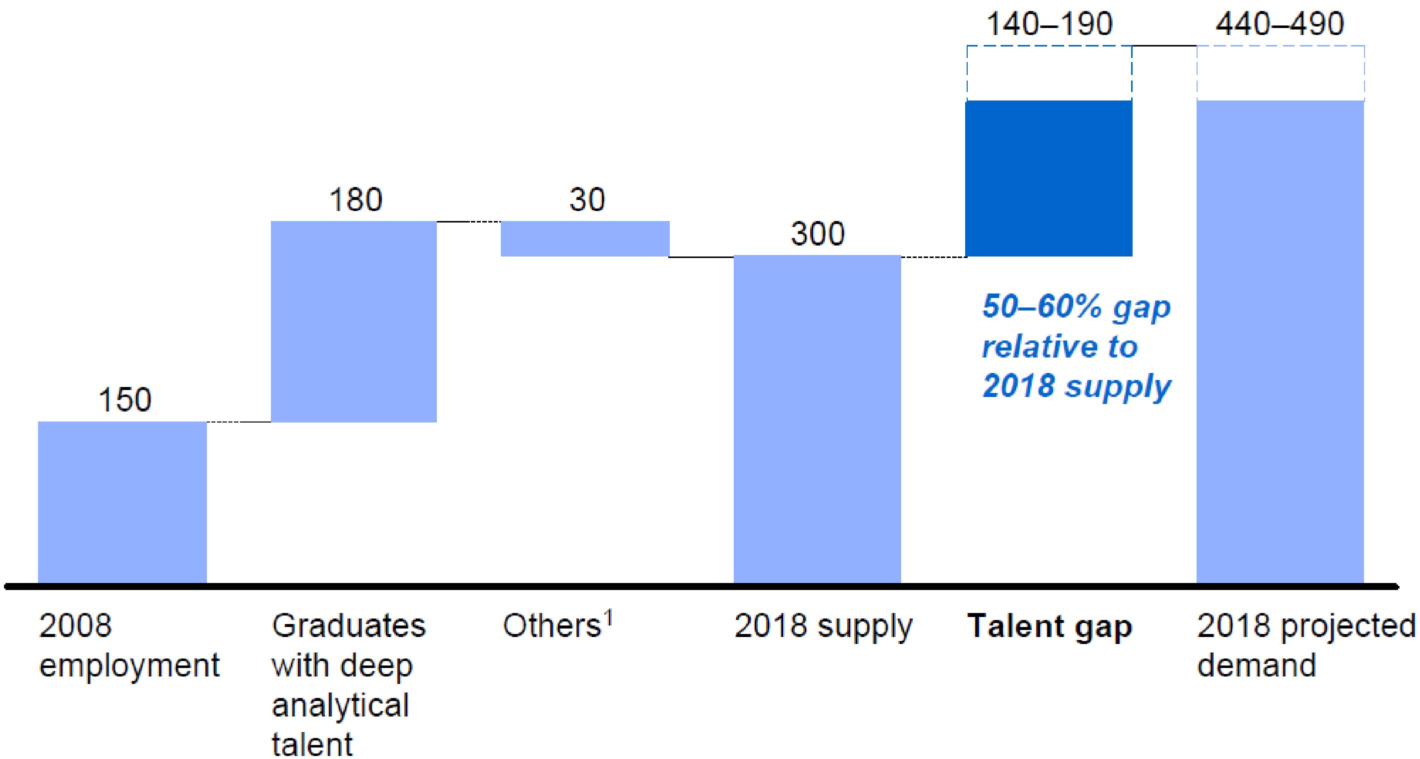


JOBS: Projected shortage of 140,000-190,000 people with deep analytical talent in the US by the year 2018.

Demand for deep analytical talent in the United States could be 50 to 60 percent greater than its projected supply by 2018

Supply and demand of deep analytical talent by 2018

Thousand people



1 Other supply drivers include attrition (-), immigration (+), and reemploying previously unemployed deep analytical talent (+).

SOURCE: US Bureau of Labor Statistics; US Census; Dun & Bradstreet; company interviews; McKinsey Global Institute analysis

Source: *"Big data: The next frontier for innovation, competition, and productivity," McKinsey Global Institute, May 2011.*

Data Miners in Big Data Analytics

Big Data Analytics
Understand goals of business
Collaborate in interdisciplinary teams
Integrate large volumes of structured and unstructured data
Formulate problems, develop solutions
Blend statistical modeling, data mining, forecasting, optimization
Develop/run integrated software solutions
Gain higher visibility

Change business operation

Thank You!

My WEB site: http://datamining.rutgers.edu

Reference

- Yong Ge, Hui Xiong, Alexander Tuzhilin, Keli Xiao, Marco Gruteser, Michael J. Pazzani, An Energy-Efficient Mobile Recommender System , the 16th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining (**KDD 2010**), pp. 899 - 908, 2010.
- Yong Ge, Qi Liu, Hui Xiong, Alexander Tuzhilin, Jian Chen, Cost-aware Travel Tour Recommendation, the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (**KDD 2011**), to appear, 2011.
- Qi Liu, Yong Ge, Zhongmou Li, Enhong Chen, Hui Xiong, Personalized Travel Package Recommendation, the 11th IEEE International Conference on Data Mining (**ICDM 2011**) (ICDM 2011), Best Research Paper Award, 2011.
- Yong Ge, Chuanren Liu, Hui Xiong, A Taxi Business Intelligence System, the 17th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining **(KDD 2011**), to appear,2011.

Reference

- Chuanren Liu, Hui Xiong, Yong Ge, Wei Geng, Matt Perkins. A Stochastic Model for Context-Aware Anomaly Detection in Indoor Location Traces. the 12th IEEE Conference on Data Mining (ICDM 2012), to appear, 2012.
- Baik Hoh, Marco Gruteser, Hui Xiong, Ansaf Alrabady, Preserving Privacy in GPS Traces via Uncertainty-Aware Path Cloaking, the 14th ACM Conference on Computer and Communication Security, (**ACM CCS**), pp. 161 - 171, 2007.
- Jing Yuan, Yu Zheng, Xing Xie: Discovering regions of different functions in a city using human mobility and POIs. the 18th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining (KDD 2012), pp. 186-194, 2012.