Tackling the Challenges of Big Data
Introduction and Use Cases

Daniela Rus
Professor
Director of CSAIL
Massachusetts Institute of Technology

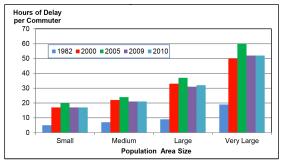
Tackling the Challenges of Big Data
Introduction and Use Cases
Case Study: Transportation

Daniela Rus
Professor
Director, CSAIL
Massachusetts Institute of Technology

Big Data Meets the Big City

[Image Source: T-mobile]

Transportation: A Crisis in the Making



[2011 Urban Mobility Report, Texas Transportation Institute]

Cost
- Time
- Fuel
- Air pollution
- Frustration, etc.

Challenges
- Uncertainty
- Congestion
- Estimation
- Computational cost

Future Urban Mobility

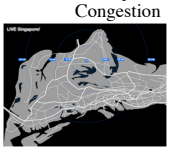
- Data-Driven Transportation
  - Improve Level of Service
  - Enhance sustainability
  - Modeling, Predicting, Controlling
  - Personalization
  - Optimization

[Future of Urban Mobility, SMART]

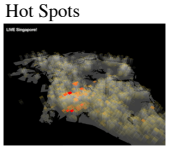
Data-Driven Transportation Goals

- Algorithmic and policy solutions for improving traffic
  - Visualizing traffic
  - Analysis of travel patterns
  - Congestion-aware routing
  - Urban planning
  - Road pricing

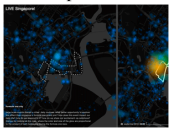
Congestion



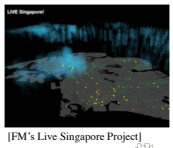
Hot Spots



Special Events

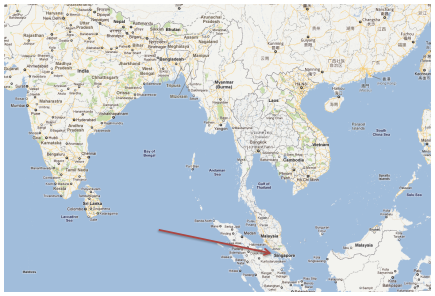


Weather



[FM's Live Singapore Project]

Case Study in Singapore



Singapore

30 x 16 miles, 5M people, high GDP per capita, *lots of taxis*…

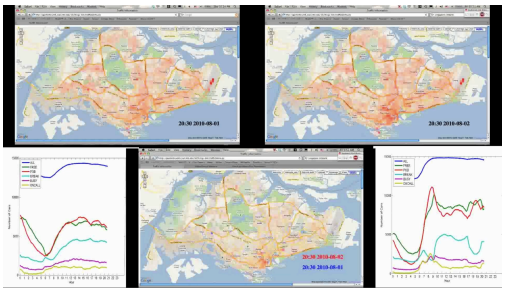
Singapore Taxi Data

- 16k unique taxis
- ID, GPS location, speed, status, timestamp
- ≈ 0.5-1 min intervals
- One month: 512M data points, 33 GB of data

In-car Data Device

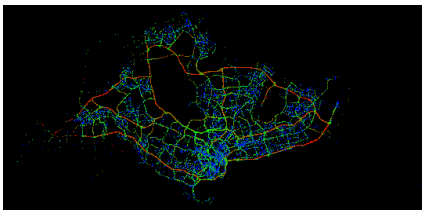
- Data cleaning, map matching, …

Traffic Visualization: Volume



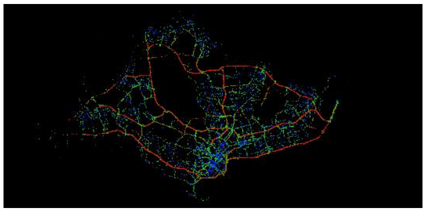
Traffic Visualization Example

Traffic Visualization: Speed/ Congestion



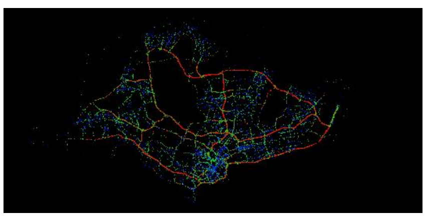
Slow Medium Fast

Traffic Visualization: Congestion Sweep



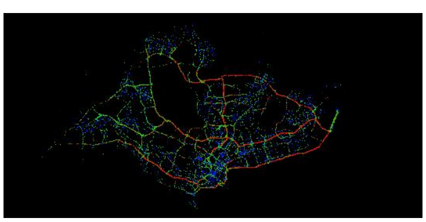
Slow Medium Fast

Traffic Visualization: Congestion Sweep



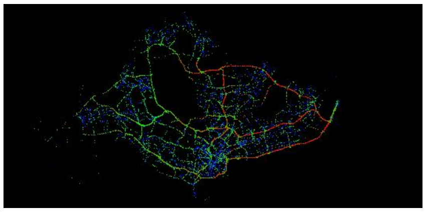
Slow Medium Fast

Traffic Visualization: Congestion Sweep



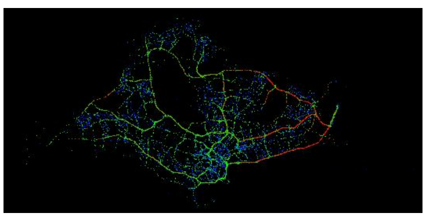
Slow Medium Fast

Traffic Visualization: Congestion Sweep



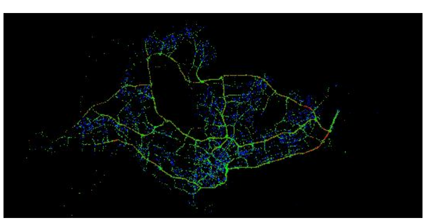
Slow Medium Fast

Traffic Visualization: Congestion Sweep



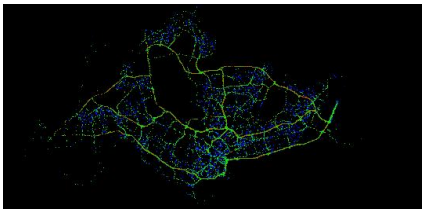
Slow Medium Fast

Traffic Visualization: Congestion Sweep



Slow Medium Fast

Traffic Visualization: Congestion Sweep



Slow Medium Fast

Case Study: Transportation in Singapore

Source: Live Singapore, Future of Urban Mobility, Singapore-MIT Alliance for Research and Technology

Tackling the Challenges of Big Data
Introduction and Use Cases
Case Study: Transportation

THANK YOU

PROFESSIONAL EDUCATION

CSAIL

Tackling the Challenges of Big Data
Introduction and Use Cases

Daniela Rus
Professor
Director of CSAIL
Massachusetts Institute of Technology

Tackling the Challenges of Big Data
Introduction and Use Cases
Case Study: Transportation

Daniela Rus
Professor
Director, CSAIL
Massachusetts Institute of Technology

Outline

- The traffic problem
- Using taxi probes to analyze traffic
- Why are taxis good probes?
- Applications

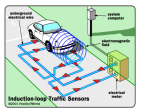
Travel Pattern Analysis

- How does traffic flow ?
  - workdays / non-workdays
  - rush hour / non- rush hour
- Where does traffic originate?
- What is its destination?
- What routes are taken?
- How does this affect planning, resource allocation, etc.?

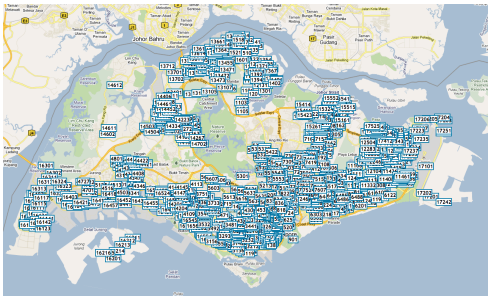
Data-Driven Traffic Modeling

- GPS sensors on vehicles
  - Measures traffic speed or delay
  - Easy to deploy
- Inductive loop detectors
  - Measures traffic flow
  - Costly to install

CarTel sensor node:
[Hull, Madden, Balakrishnan 2006]



Full Study

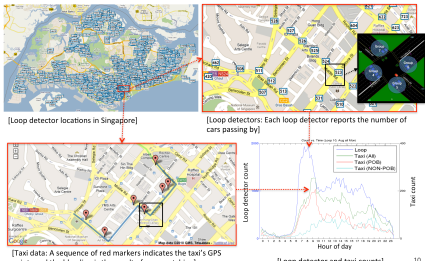


How Representative is Taxi Traffic?

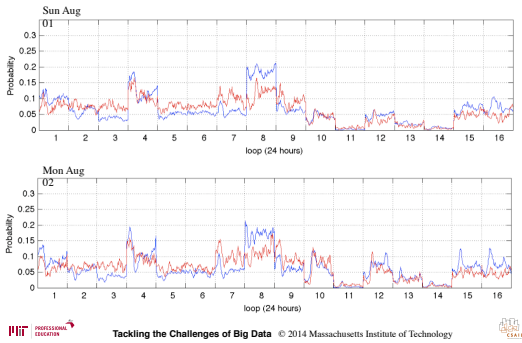
- Taxis are a *sample* of overall traffic…
- Is it an *unbiased* sample?
  - If not, can we systematically *correct* for the bias?

Do Taxis Measure General Traffic?

loop detector count per 15 minutes: ground-truth of traffic flow
taxi count per 15 minutes: sampled version of general traffic flow

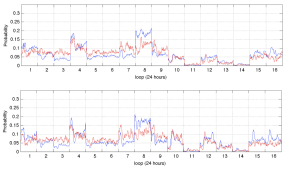


Taxis vs. Loop Distribution

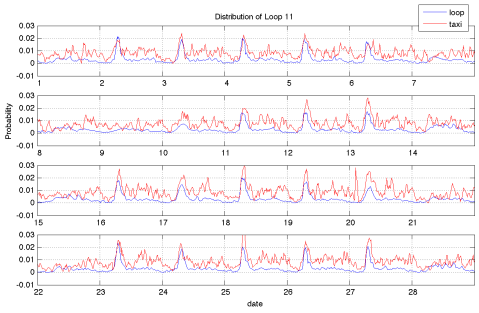


Biased Sample

- Conclusion…
  - taxi sample is biased,
  - but is it consistently biased and thus correctable?



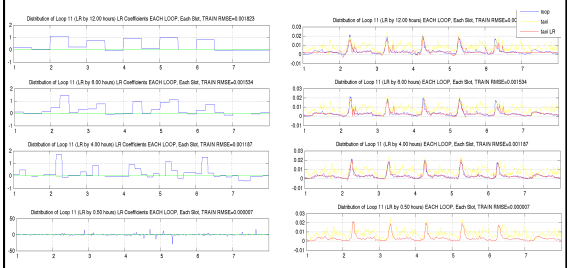
One Month, Loop 11



Traffic Flow Estimation

- Goal:
  - Predict the general traffic flow using taxi data and available set of loop detector data
- Approach:
  - Learn linear regression coefficients
    - trafficFlow= **m** * taxiFlow + **b**
  - Different time models for linear regression:
    - Day category
    - Hour category
  - Find the best time model with the most predictive power

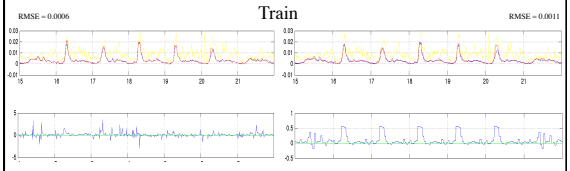
Correction Factors: Regression



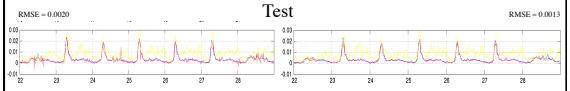
Cross-validation: Regression

Day-of-week, 15 minutes | Workday/Non-workday, 1 hour

Train

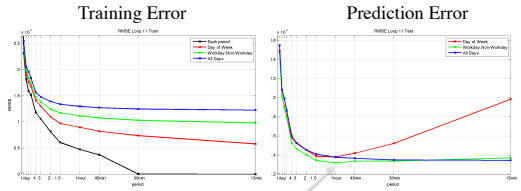


Test



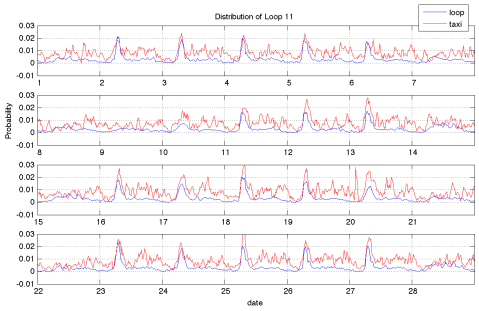
Cross Validation, Model Selection

Training Error | Prediction Error

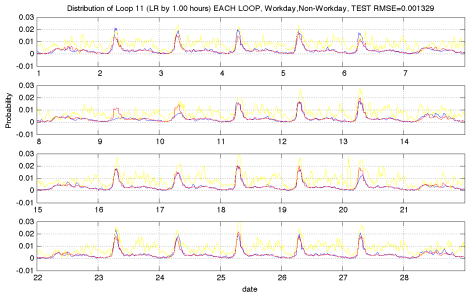


Workday/Non-workday, 1 hour

Before…

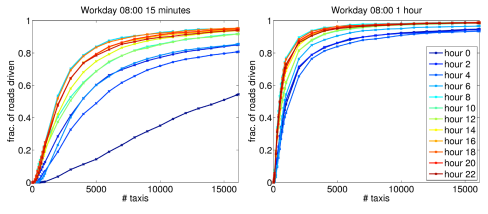


… After Regression



How Many Taxis are Needed?

- What fraction of roads are covered at least 30 times within 15 minutes / 1 hour?

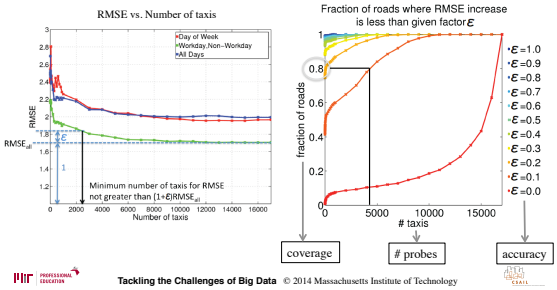


Taxi Traffic Conclusions

- Taxi traffic is a biased sample of overall traffic
- Bias is consistent and easily correctable
  - even at the granularity of individual intersections
  - using just workday/non-workday + hour of sample
- Can infer overall traffic from taxi traffic…
  - but why does this work?

Number of Probe Vehicles Needed

- Identify the number of probe vehicles required to predict the general traffic flow with defined accuracy and coverage



Travel Analysis from Personal Phones

- Slide was removed due to copyright restrictions.
- We are sorry for any inconvenience this may cause.

Tackling the Challenges of Big Data
Introduction and Use Cases
Case Study: Transportation

THANK YOU

Tackling the Challenges of Big Data
Introduction and Use Cases

Daniela Rus
Professor
Director of CSAIL
Massachusetts Institute of Technology

Tackling the Challenges of Big Data
Introduction and Use Cases
Case Study: Transportation

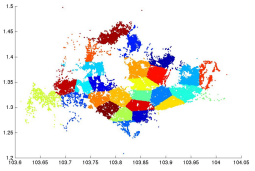
Daniela Rus
Professor
Director, CSAIL
Massachusetts Institute of Technology

Outline

- The traffic problem
- Using taxi probes to analyze traffic
- Why are taxis good probes?
- Applications

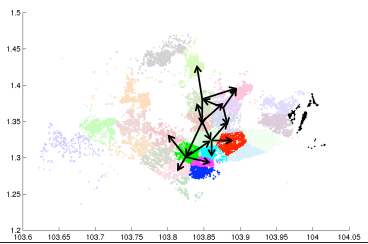
Bias Assessment Methodology

- Election forecasting analogy:
  - distribution of sample vs. distribution of voters
  - different distributions imply biased sample
  - need *ground truth* to compare against
  - loop detectors…



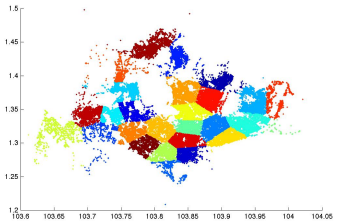
Why Are Taxis Good Probes?

- Stochastic taxi movement due to randomness of passenger
-



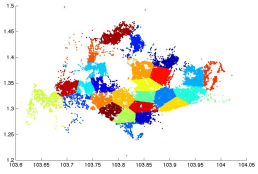
Traffic Analysis Setup

- What are origins/destinations?
  - GPS coordinates, street addresses, *regions*…

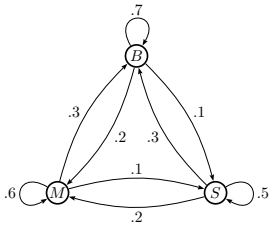


Traffic Analysis Setup

- Analysis criteria?
  - Micro-level: o/d prob, cond prob, …
  - Macro-level: uniformity, randomness
- Macro-level uniformity measures:
  - entropy (bits)
  - *perplexity* (regions)

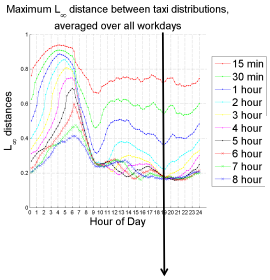


Markov Chains

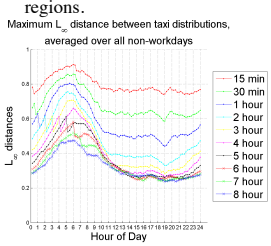


Pr[B] = 1/2, Pr[M] = 1/3, Pr[S] = 1/6

Rapidly Mixing Taxi Distributions



Upper bound on divergence between distributions of taxis starting from different regions.



Distributions converge quickly.
Taxi probes provide good coverage regardless of initial distribution.

RMMC Summary

- At a *coarse* granularity, taxis behave like a rapidly mixing Markov chain.
- Consequence: taxis are good for *sampling*
- *Future work:*
  - granularity vs. fleet size vs. mixing time vs. coverage probability
  - sensing and monitoring applications
    - road conditions, air quality, noise pollution, potholes…

Tackling the Challenges of Big Data
Introduction and Use Cases
Case Study: Transportation

THANK YOU

PROFESSIONAL EDUCATION

CSAIL

Tackling the Challenges of Big Data
Introduction and Use Cases

Daniela Rus
Professor
Director of CSAIL
Massachusetts Institute of Technology

Tackling the Challenges of Big Data
Introduction and Use Cases
Case Study: Transportation

Daniela Rus
Professor
Director, CSAIL
Massachusetts Institute of Technology

Outline

- The traffic problem
- Using taxi probes to analyze traffic
- Why are taxis good probes?
- Applications

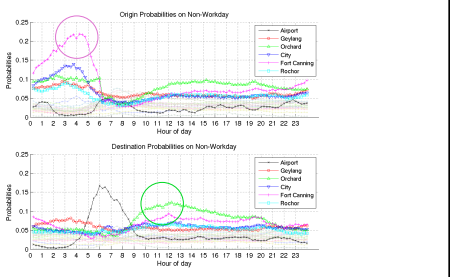
Application: Hotspots

- Detect regions of high traffic volume and potential congestion

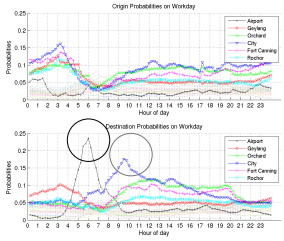
Application: Hotspots

- Detect regions of high traffic volume and potential congestion

Application: Origin-Destination Analysis

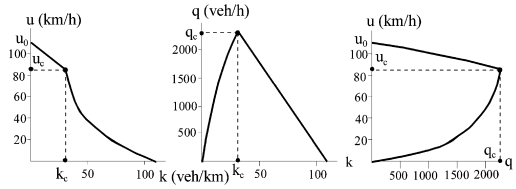


Application: Origin-Destination Analysis



Fundamental Traffic Flow Diagram

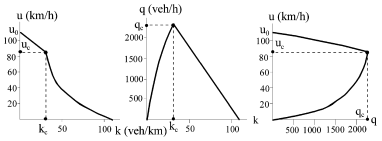
- Road performance: speed vs. flow vs. density



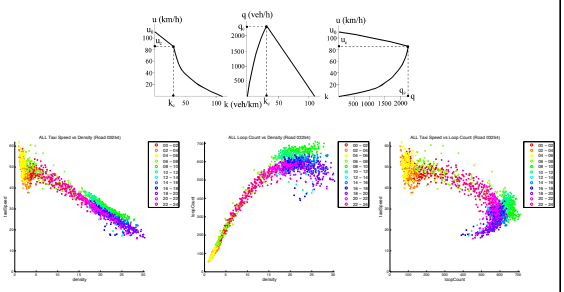
flow (veh/hr) = speed (km/hr) * density (veh/km)

Fundamental Traffic Flow Diagram

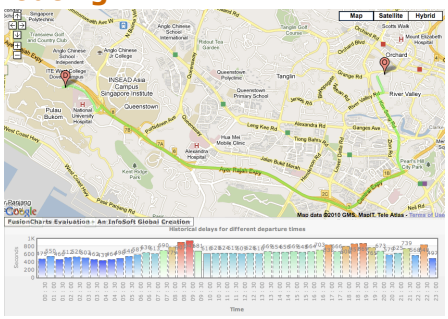
- Can be inferred from taxi and loop data
- Given earlier results…
  - diagrams can be inferred from taxi data alone
  - performance can be inferred in *real time*



Church Street



Application: Congestion-Aware Routing



Wrapping Up

- Roving sensor network of taxi probes
- Predict general traffic flow using taxi flow
- Intuition: Rapidly mixing taxi distributions
- Identify the number of probe vehicles required to predict the general traffic flow
- Demonstrate use of dynamic traffic probes at city scale

Tackling the Challenges of Big Data
Introduction and Use Cases
Case Study: Transportation

THANK YOU