TMIP
Better Methods. Better Outcomes.

Webinar Series

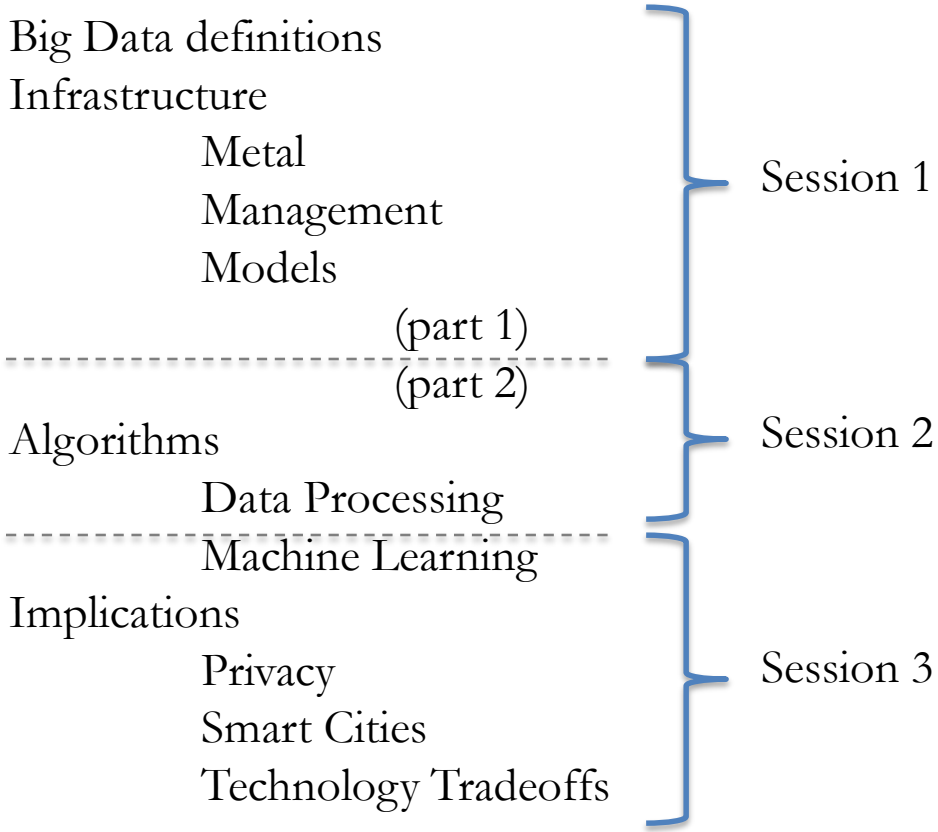
TMIP VISION

TMIP provides technical support and promotes knowledge and information exchange in the transportation planning and modeling community.

DISCLAIMER

The views and opinions expressed during this webinar are those of the presenters and do not represent the official policy or position of FHWA and do not constitute an endorsement, recommendation or specification by FHWA. The webinar is based solely on the professional opinions and experience of the presenters and is made available for information and experience sharing purposes only.

Webinar Roadmap



Webinar Roadmap

Big Data definitions
Infrastructure
- Metal
- Management
- Models

(part 1)

(part 2)

Algorithms
- Data Processing

Machine Learning

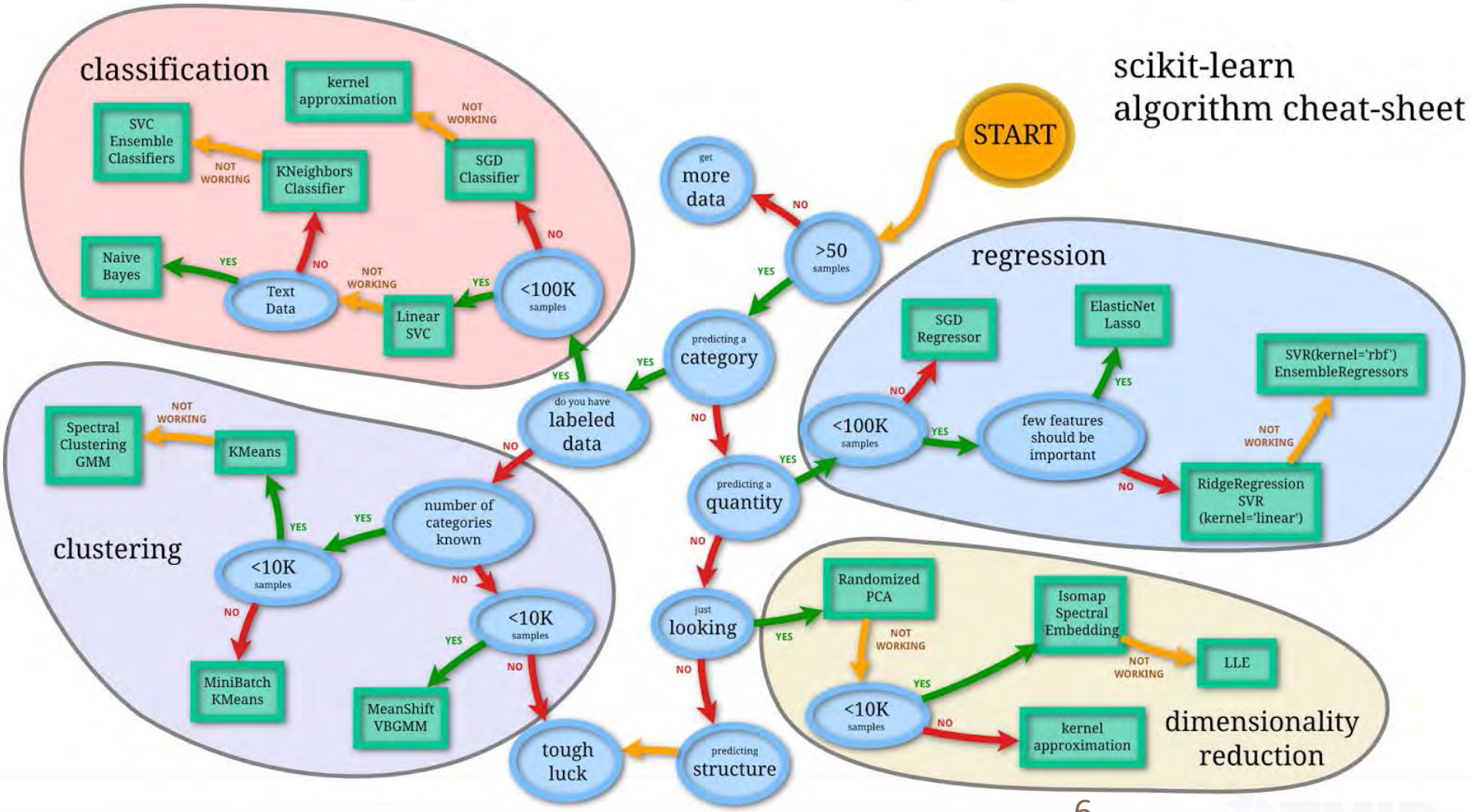
Implications
- Privacy
- Smart Cities
- Technology Tradeoffs

Session 3

Big Data Algorithms - Continued

MACHINE LEARNING

An Algorithm for Choosing Algorithms



Machine Learning Algorithms

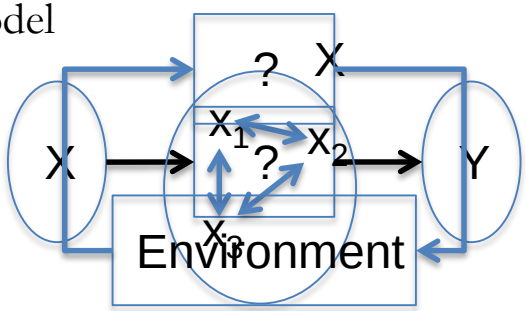
What is Machine Learning (ML)?

- Algorithms that attempt to learn based on N samples of known data and then try to predict properties of unknown data
  - Samples are characterized by *features*
  - Patterns of features are learned
- More informally, machine learning is:
  - Automated Pattern Matching
  - Programming by Example

Types of Machine Learning

Based on Learning Styles

- Supervised Learning
  - Sample data is labeled (x,y) and used to train the model
  - Detects patterns between input/output sets
  - Important Types: Regression, Classification
- Unsupervised Learning
  - Sample data is unlabeled (x)
  - Detects patterns within input sets
  - Important Types: Clustering, Association Rules
- Reinforcement Learning
  - A reward maximization goal is set
  - Detects patterns between actions/rewards



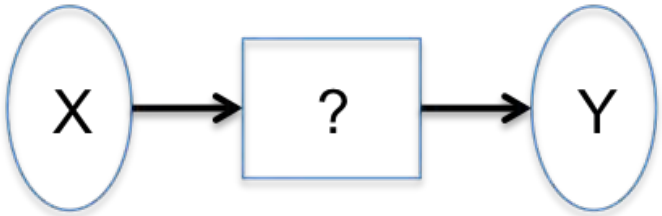
Supervised Learning Algorithms

Classification

- Model a relationship between input data and discrete (categorical) output data
- Classifier: An algorithm implementing classification
  - Enables prediction
- Example Inputs/Outputs
  - Email → Spam/Ham
  - Tumor → Malignant/Benign

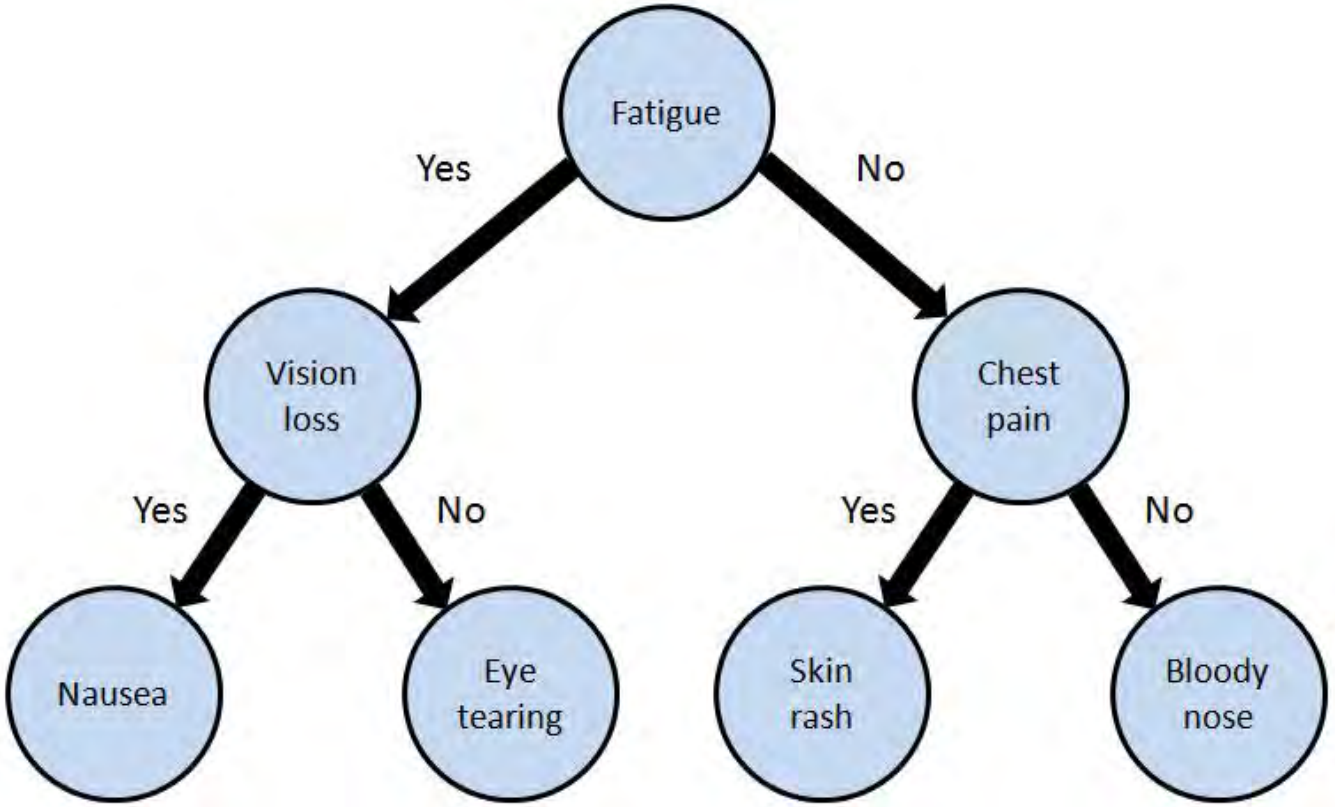
Many Types of Classification Algorithms

- Decision Trees
- Bayesian
- Neural Networks
- …



TMIP

Decision Tree Classifiers



Big Data Thoughts on Decision Tree Classification

Complexity

- Training Time: O(dnlog(n))
- Operation Time: O(log(n))
- Fairly efficient but can still lead to memory issues

Risks

- Over-fitting
- Training an optimal tree

Parameters
- n: Number of training instances
- d: Number of dimensions (features)

Bayesian Classification

Bayes Rule

- As a Bayes Classifier
- As a Naïve Bayes Classifier

Example: Bayesian Spam Filter

$$P(\text{spam}|\text{word}) = \frac{P(\text{word}|\text{spam})\,P(\text{spam})}{P(\text{word})}$$

$$P(\text{spam}|\text{"money"}) = \frac{P(\text{"money"}|\text{spam})\,P(\text{spam})}{P(\text{"money"})}$$

Big Data Thoughts on Bayesian Classification

Bayesian Classification

- Complexity:
  - Training Time: O(nd + cd)
  - Training Space: O(cd)
  - Operation Time: O(d)
- Fairly efficient (and controllable) but
  - Computing Naïve Bayes in R or Scikit-learn.org
    - Limited by memory
  - Alternatives
    - Revolution Analytics
    - Scalding (Scala on Hadoop)

Parameters
- n: Number of training instances
- d: Number of dimensions (features)
- c: Number of classes

TMIP

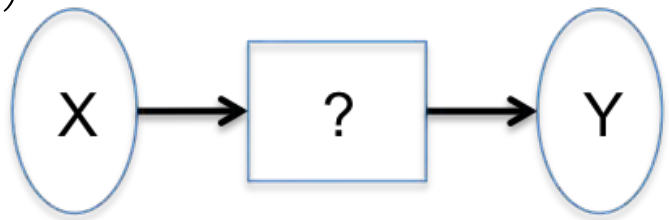
Supervised Learning Algorithms

Regression

- Model a relationship between input data and continuous output data
- Inputs: Predictor/Explanatory Variables
- Outputs: Dependent Variable
- Iterative Improvement Based on Error Measurement

Many Types of Regression Algorithms

- Linear Regression (Ordinary Least Squares)
- Logistic Regression
- Generalized Linear Model



Regression

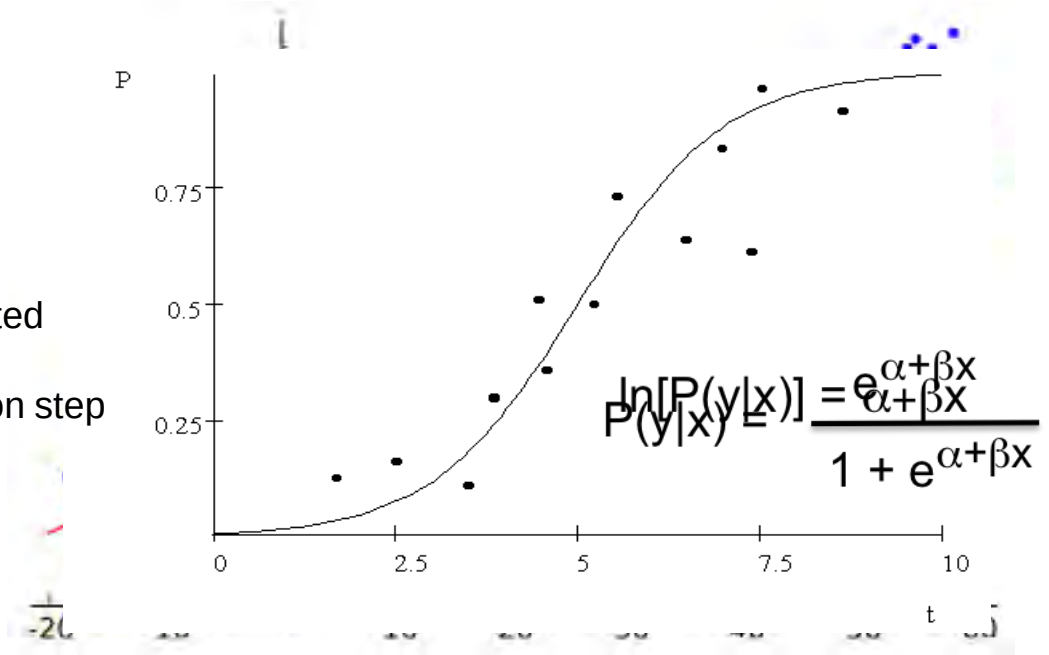
Model Numeric Data as a Function

- Curve Fitting w/Error Measurement

Error is Normally Distributed

Error is Binomially Distributed (supports classification)

- Requires an optimization step to calculate β



Big Data Thoughts on Regression

Linear Regression (OLS)

- Time Complexity: $O(d^2n)$

Logistic Regression

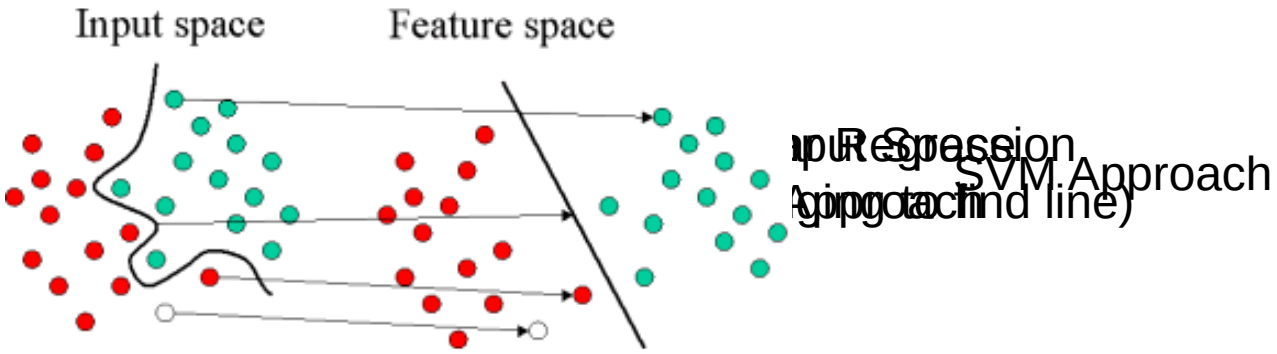
- Time Complexity: $O(n)$
  - But depends on calculation of β
  - There are distributed calculations of β

Parameters
- n: Number of samples
- d: Number of dimensions (features)

Support Vector Machines

Support Vector Machines (SVM)

- Decision Planes:
  - Planes within feature space that separate data
- Kernels:
  - Functions that transform the data input space into feature space



Big Data Thoughts on Support Vector Machines

SVM Complexity

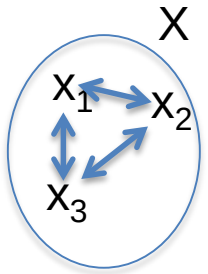
- Time Complexity: O(nd)
  - Depends on type of SVM, but the process can be distributed
    - Hadoop!!!

Parameters
- n: Number of samples
- d: Number of dimensions (features)

Unsupervised Learning Algorithms

Clustering

- Measure the similarity of data elements
  - Cosine Similarity
  - Jaccard Similarity
  - Hamming Distance



Common Clustering Algorithm

- k-NN (K Nearest Neighbor)

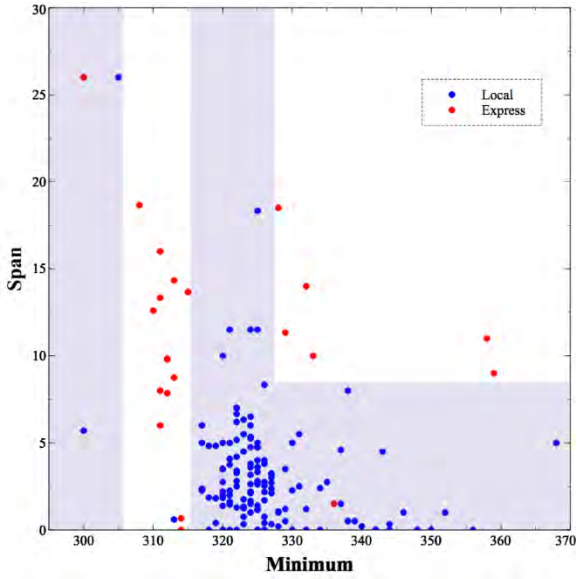
Comparing Algorithms
Train Speed Prediction

Example taken from
Silicon Valley Data Science
http://svds.com/post/listening-caltrain

Goal: Predict train speed based on horn audio detection

- Output: Local vs Express Train

Decision Tree

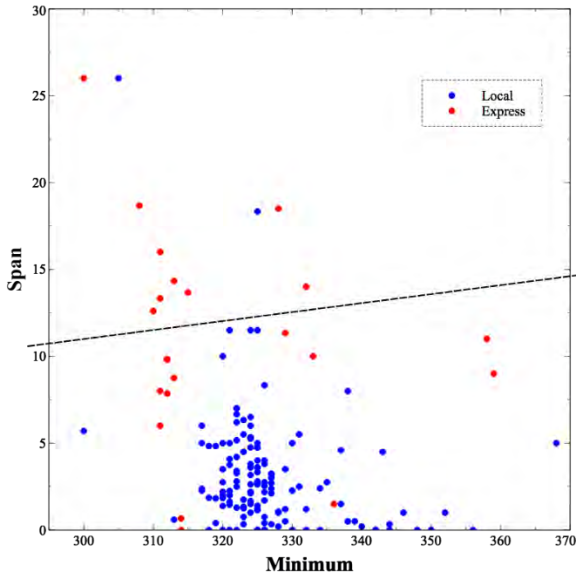


Two Plots show actual data
Blue dots: Local
Red dots: Express

Decision Tree
Blue region predicts local
White region predicts express

Logistic Regression
Above Line: Express
Below Line: Local

Logistic Regression



Webinar Roadmap

Big Data definitions
Infrastructure
- Metal
- Management
- Models

(part 1)

(part 2)

Algorithms
- Data Processing
- Machine Learning

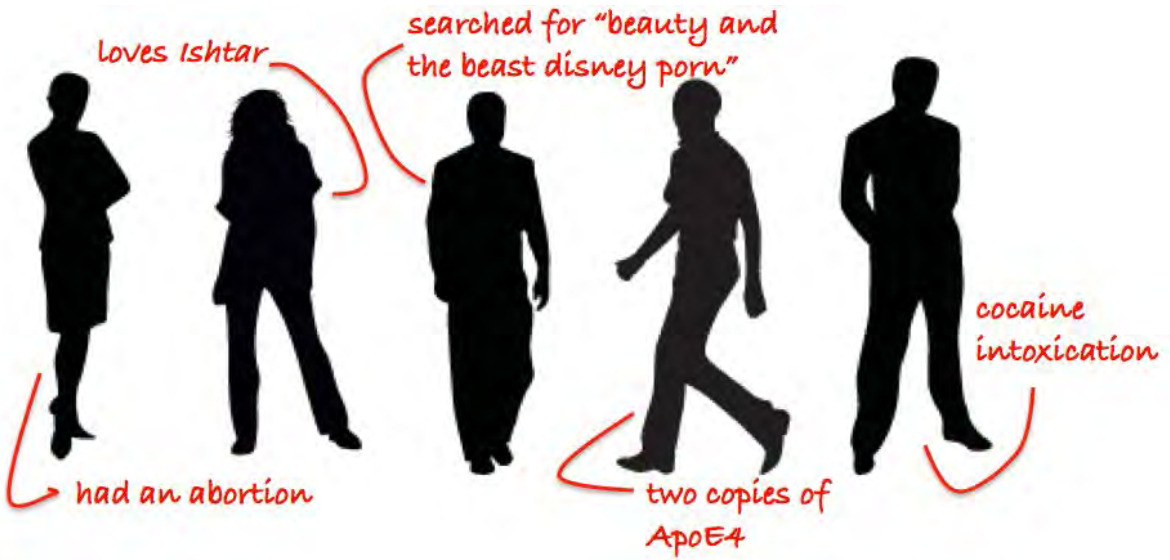
Implications
- Privacy
- Smart Cities
- Tradeoffs

Session 3

WHAT ARE THE IMPLICATIONS OF BIG DATA?

To Consider:

- Public Policy
- Ethics
- Economics
- Ethical Inquiry
- Privacy Policies
- Policy Problems
- Policy Options



Webinar Roadmap

Big Data definitions
Infrastructure
- Metal
- Management
- Models

(part 1)

(part 2)

Algorithms
- Data Processing
- Machine Learning

Implications
- Privacy
- Smart Cities
- Tradeoffs

Session 3

In the US, State and Federal Law is Mostly Based on a Sectoral Approach to Privacy

Example: Fair Credit Reporting Act (FCRA) 1970

- Goals: accuracy, fairness, privacy
- Target: consumer reporting agencies
- Domains: credit and insurance reports, background checks, tenant screenings
- Protections: consumer access to reports and ability to correct, company reports on accuracy, limitations on use, notice when adverse action is taken based on report

Others:

- Family Educational Rights and Privacy Act (1974)
- Electronic Communications Privacy Act (1986)
- Health Insurance Portability and Accountability Act (1996)
- Financial Sector – Gramm-Leach-Bliley Act (1999)
- Genetic Information Nondiscrimination Act (2008)
- …

But, the default position for data privacy is usually "off"

http://www.whitehouse.gov/sites/default/files/docs/big_data_privacy_report_may_1_2014.pdf

Transportation Data Can Reveal Private Information

Uber's "Rides of Glory"

Tracking user rides:

- Friday or Saturday night to somewhere other than home
- Leaving the next day from the same location

Concerns raised in NYTimes article: "We Can't Trust Uber" 12/7/14

- Blackmailing journalists
- Affairs
- Planned Parenthood trips
- Interviews at rival companies

Current legislation doesn't really protect this kind of information

Privacy Concerns in Other Domains are Also Presenting Challenges to the Sectoral Approach…

Powerful connection between lifestyle and health outcomes

- Distinction between personal data and health care data has begun to blur

Big data revolution in education - how best to protect student privacy as technology reaches further into the classroom?

- States and local communities have played the dominant role in providing education
- Software that supports online learning tools and courses is provided by for-profit firms
- Who owns data streams coming off online education platforms and how they can be used?

Information revealed when using a digital education platform can be very personal

- Aptitude for particular types of learning and performance
- Whether students have learning disabilities or have trouble concentrating for long periods
- What time of day and for how long students are logged in reveals lifestyle habits
- How should educational institutions use data to improve learning opportunities for students?
- How can students who use these platforms be confident that their data is safe?

Correlations to protected data are not necessarily protected

…While Data Brokers/Services are Building More Detailed Profiles for Commercial and Government Use

Three type of 'Data Services':

- Consumer reporting (under FCRA): data, analysis, reporting in a separate subject to compliance rules
- Risk mitigation: identity verification, fraud detection and people-search or look up services
- Marketing services: identify potential customers, target ads, and other advertising-related services

Fair Credit Reporting Act provides affirmative rights to consumers.

These statutory rights do not exist for risk mitigation or marketing services

- Develop profiles on hundreds of millions of consumers
- Activities monitored to pinpoint a message and send it at the right moment
- Exceptionally detailed (thousands of pieces of data)
- Segment customers into precise categories:
    - "Ethnic Second-City Strugglers"
    - "Retiring on Empty: Singles"
    - "Tough Start: Young Single Parents"
    - "Credit Crunched: City Families"
    - "Rural and Barely Making It."

What protections are there for this data?

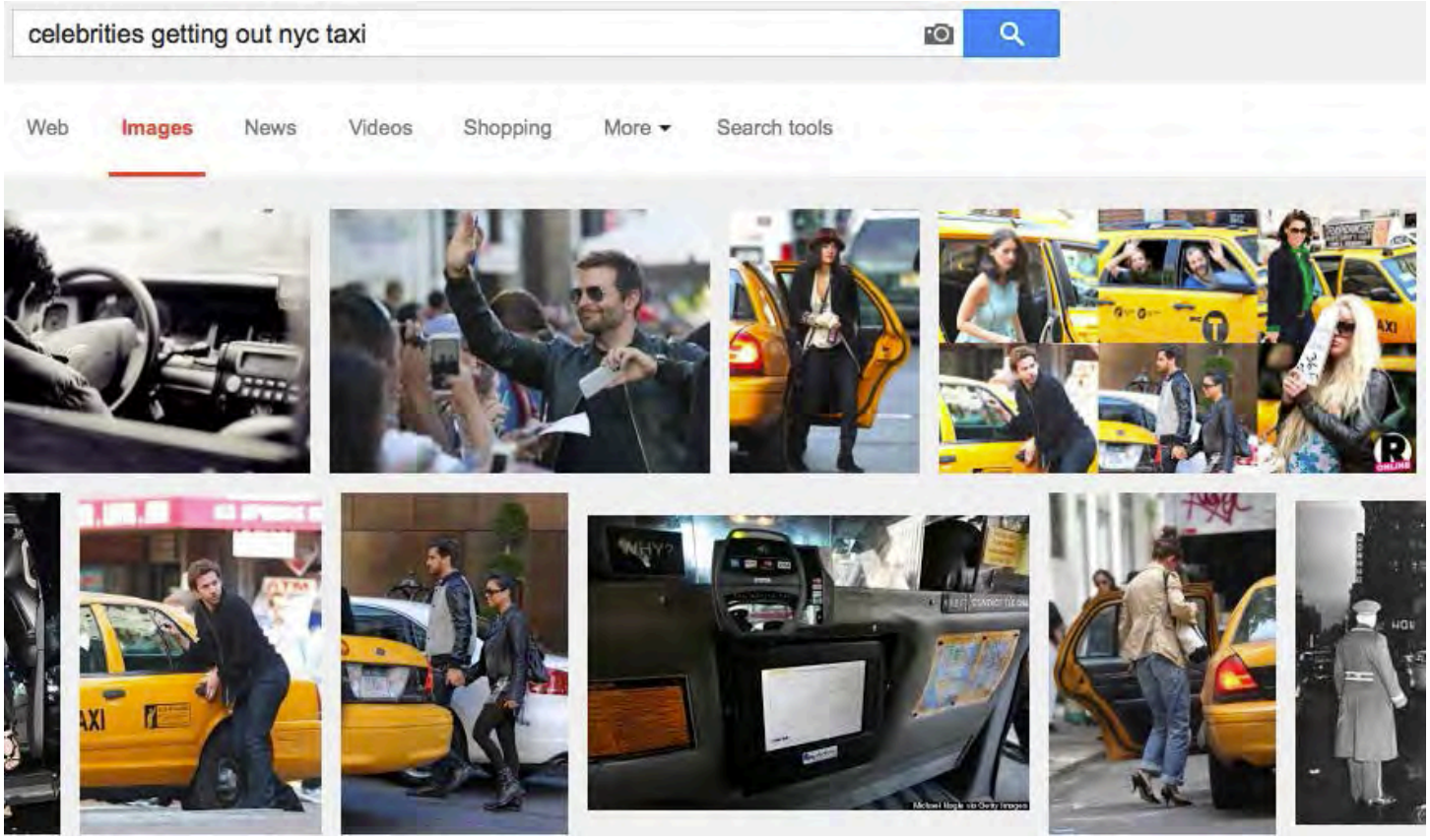
Reviewing Private Sector Policies: Top 50 Fortune 500 Companies

Issue	Findings
Data Sold without consent	34 would not, 0 said they would, 16 said nothing
Targeted advertising	23 said yes, 1 ruled out, 26 said nothing
Customer control over data	33 stated that a user could control, 31 explained how to opt out
Data purchases	11 would buy or obtain, 0 said they would not
Data sharing / aggregation	40 would share with 3rd parties (suppliers/shippers), 8 said nothing, 2 would not share 24 said user data would be aggregated with other sources
Stated reasons	16 stated a reason for protecting 14 "you care and we want your business" 2 "respect for individual or fair information practices"
PII	47 made a distinction between PII and anonymized data 10 said anonymized not treated as protected others didn't say it wouldn't be released

Nothing to worry about if it's anonymized…

Davis, Kord, and Doug Patterson. 2012. *Ethics of Big Data.* Sebastopol, CA: O'Reilly

Anonymization of Data is Challenging and a High Profile Failure Already Exists in the Transportation World



Will there ever be a privacy backlash/major regulatory overhaul? Implications?

https://Medium.com/@vijayp/of-taxis-and-rainbows-f6bc289679a1
http://www.fastcompancy.com/3036573/fast-feed/nyc-taxi-data-blunder-reveals-which-celebs-dont-tip-and-who-frequents-strip-clubs

Webinar Roadmap

Big Data definitions
Infrastructure
- Metal
- Management
- Models

(part 1)

(part 2)

Algorithms
- Data Processing
- Machine Learning

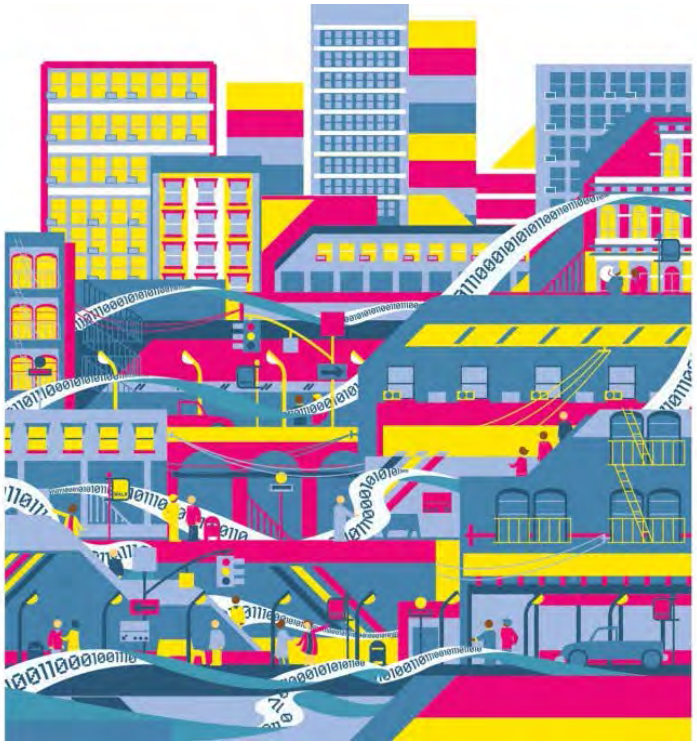
Implications
- Privacy
- Smart Cities
- Tradeoffs

Session 3

SMART CITIES

To Consider:

- Smart Cities
- Examples
- Marketing
- Criticism
- For/Against?



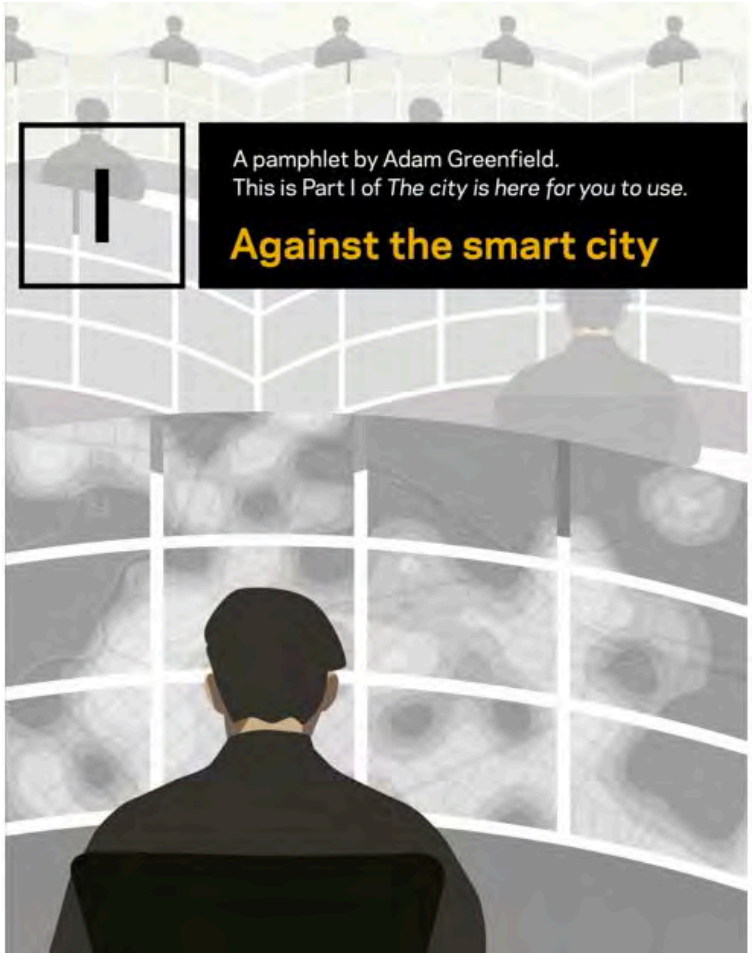
What is most important to you about the cities you have lived in?

TMIP

Consider Two Visions for What A Smart City Looks Like:

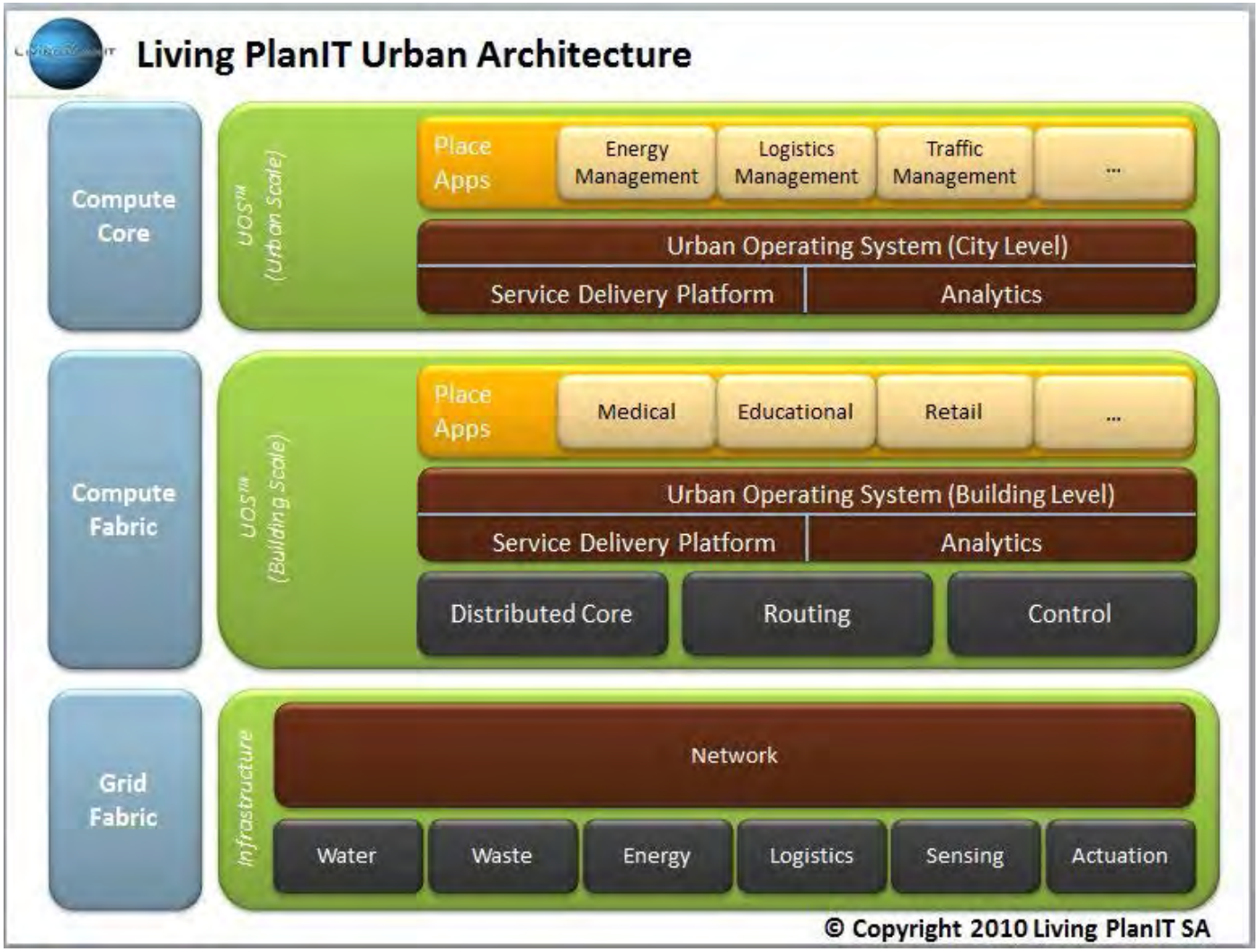
A data driven utopia that improves the lives of citizens

An Orwellian surveillance state



An example: PlanIT Valley – Utopia Delayed by Financial Troubles

Traffic Management is a Key Element of this Architecture



City designed like a technology stack

The Importance of Traffic Management Is Tied to Continuing Trends in Rapid Urbanization

Growth in global population

- More vehicles, traffic density, road accidents
- Challenges: safety, productivity

Smart cities represent an opportunity for traffic management systems that use cloud services to provide necessary information

Another Example: Bucheon, Korea

- From manual review of video feeds to intelligent video analytics
  - Provide real time data to drivers
  - Traffic surveillance
  - Improve city roads
- Increase accuracy and speed of data collection, reduce costs on new systems/labor

$41 Trillion in spending on global upgrades to urban infrastructure in next two decades

Utopia/Surveillance dichotomy = data to drivers vs traffic surveillance

Connected Cars are Becoming a Ubiquitous Sensor Network that Feeds Traffic Management Systems

Old Model

- Rely on expensive sensors installed in a few roads to understand real time traffic conditions
- Expensive to install and maintain, sparse coverage

New Model

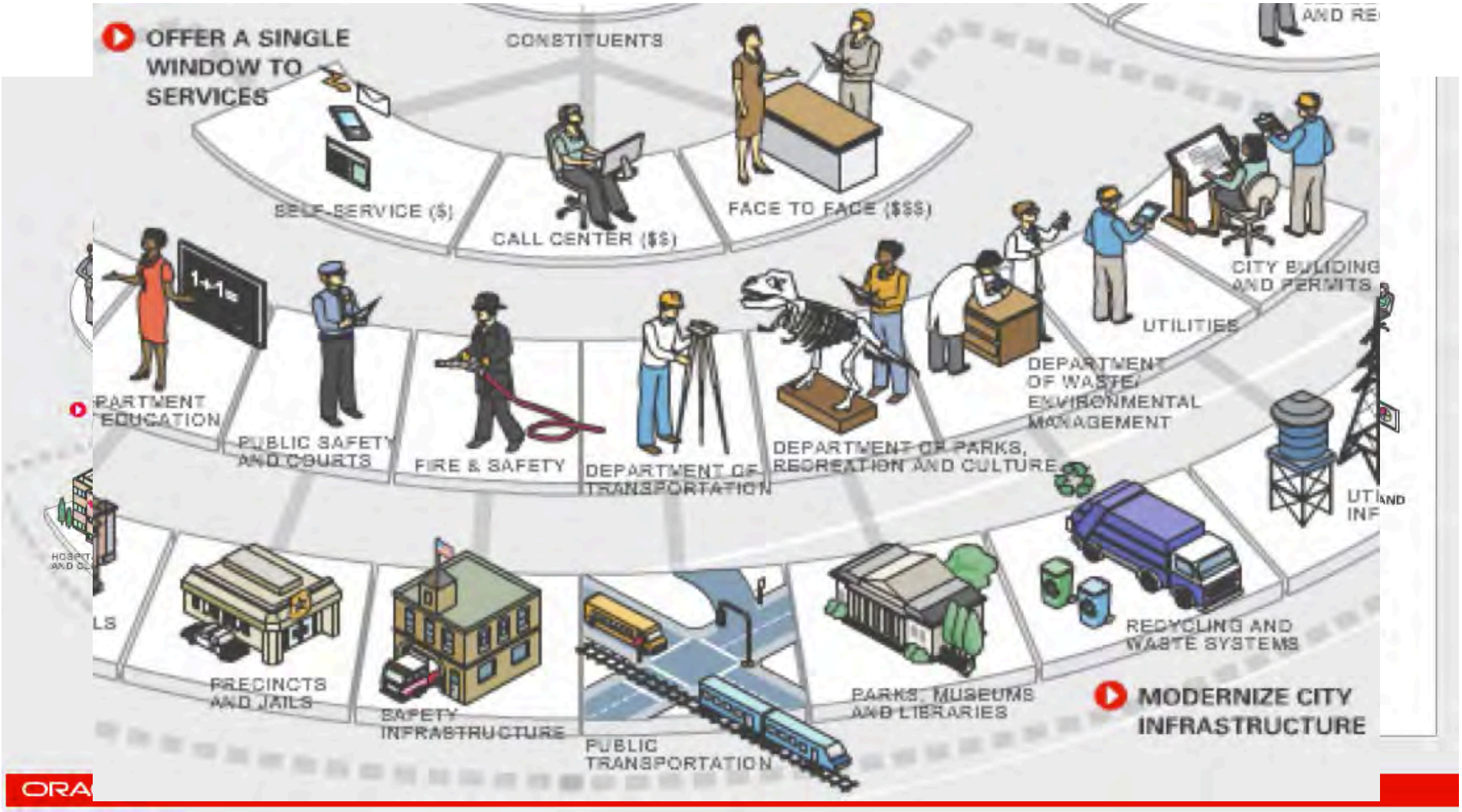
- Analyze real time data from vehicles themselves
- Understand traffic conditions everywhere
- Crowdsource sensor platform
- Similar to smart phone revolution

Example: Inrix

- Operates a network of 175 million sensors (vehicles, smartphones, cameras)
- Covers 4 million miles of road, ramp and interchange in 40 countries
    - Connected Car (parking, fuel, weather, routing)
    - Analytics (including historical data, population movement)
    - Real time and predictive traffic analysis

Tradeoffs for outsourcing? Privacy of INRIX data?

THERE ARE MANY VISIONS FOR SMART CITIES



http://www.oracle.com/us/media/interactivemaps/smartcities/smartcitiesmap-198204.html

http://www.ibm.com/smarterplanet/global/images/us__en_us__cities__smarter_cities_eco_chart_460x350.gif

"People first"… perhaps in response to initial backlash against smart cities

http://www.microsoft.com/global/apac/news/publishingimages/MicrosoftCityNext_Thumb.png

In a Smart City World, Transportation Data Has Many Peers

Oracle	IBM	Microsoft
Health – hospitals and clinics	Healthcare	Health and Social Services
Social services – service buildings	Social programs	
Education – schools	Education	Education
Public safety and courts – precincts and jails	Public safety	Public Safety and Justice
Fire and safety – safety infrastructure		
Parks recreation and culture – parks museums and libraries		Tourism, Recreation and Culture
Waste and environmental management – recycling and waste systems	Environmental	
Utilities – utilities infrastructure	Energy and Water	Energy and Water
Civil building and permits – construction sites	Smarter buildings and urban planning	Buildings Infrastructure and Planning

Foreshadows the potential need for sophisticated data blending

Webinar Roadmap

Big Data definitions
Infrastructure
- Metal
- Management
- Models

(part 1)

(part 2)

Algorithms
- Data Processing
- Machine Learning

Implications
- Privacy
- Smart Cities
- Tradeoffs

Session 3

IMPLEMENTATIONS

Traffic Management Principles Are Being Combined With Big Data in Other Domains

Problem: heavy rains + outdated combined sewer system that mixes rainwater and sewage in the same pipes

- Waste isn't easily diverted away from the river and to the sewage treatment plant
- Backups occur when large streams of water hit bottlenecks in the system

Fixing the problem:

- Spending $120 million on rebuilding the sewer system to add capacity
- Spending $6 million on several dozens sensors and a new software service to make sense of the constant data they're putting off

City chose sensors

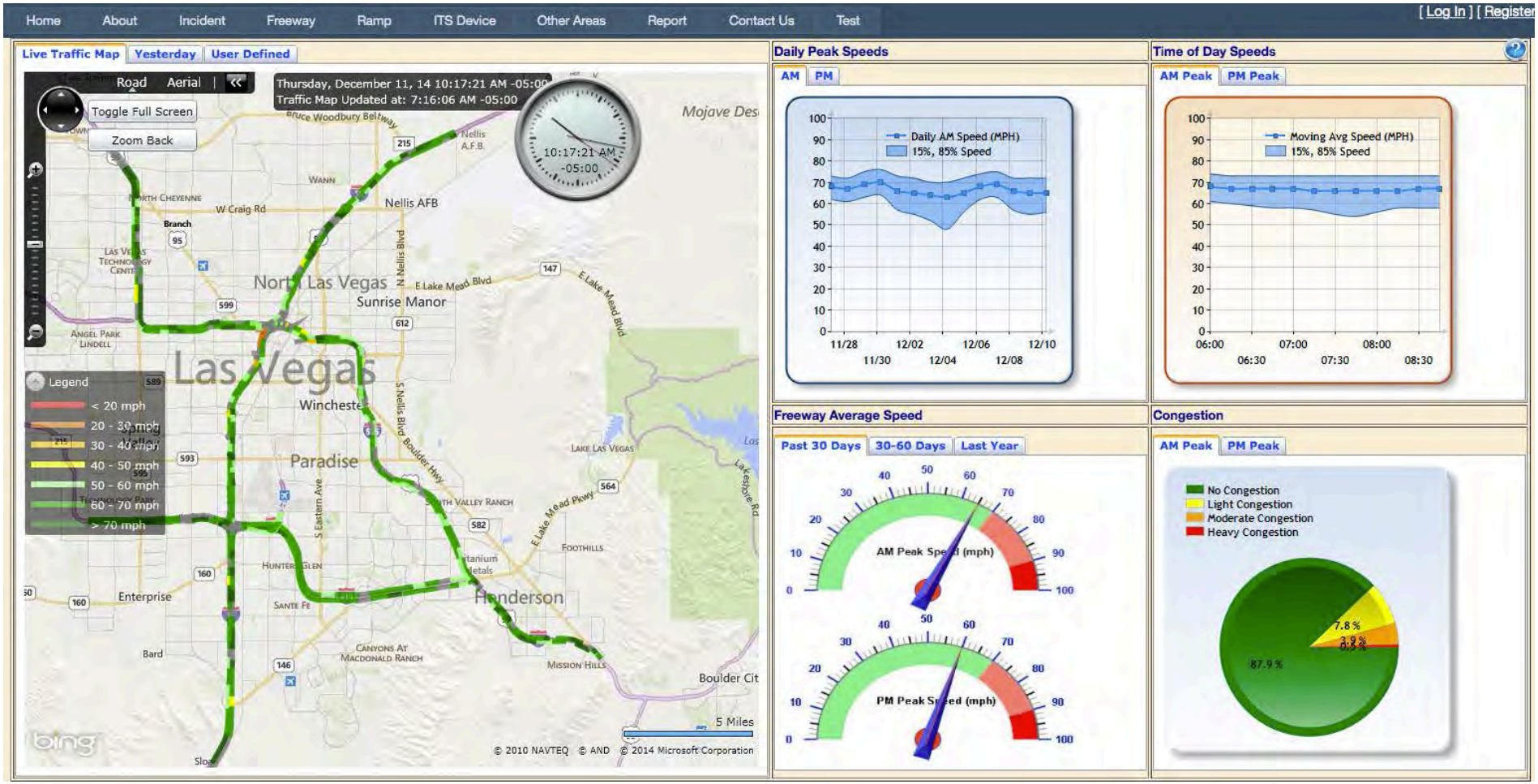
- Can see in real time where problems are arising, as well as where there's excess capacity
- City employees can intelligently divert sewage flows to ensure as much as possible makes its way to the treatment plant and as little as possible finds its way into people's homes

South Bend's sewer evolution analogous to change from timed traffic signals to today's signals that alter the timing of lights based on the flow of traffic

- "We're essentially taking something that's been used for decades [in traffic management] and applying it to sewage instead of the flows of cars and trucks."

Infrastructure improvements: repair, rebuild, or instrument with sensors?

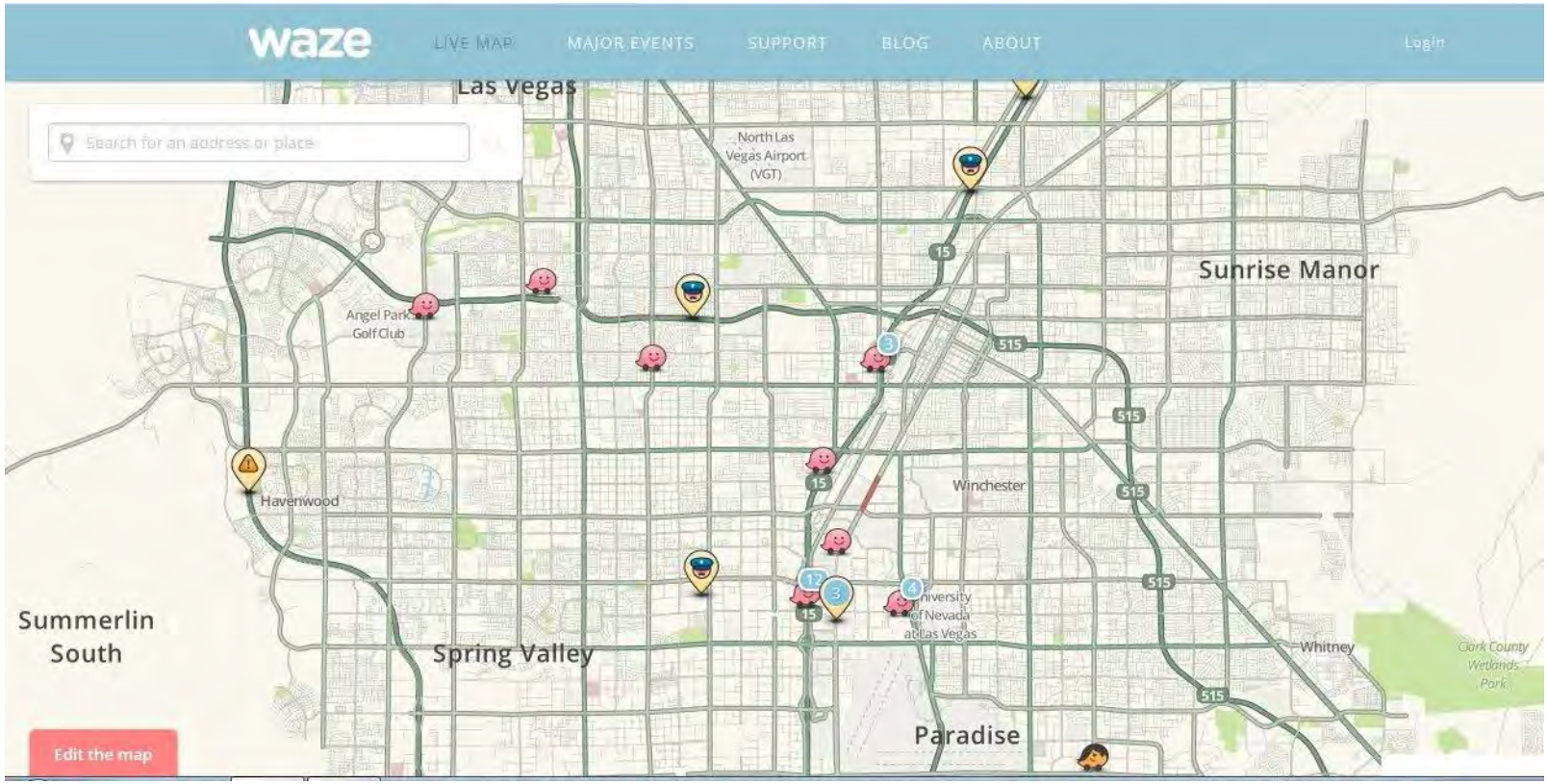
Provoking discussion: BUGATTI…



Consider data sources used here…

…vs WAZE

Use the "WAZE" app to avoid speed traps and police checkpoints



…vs data sources used here – purposes and levels of participation are key

http://www.copblock.org/64992/use-the-waze-app-to-avoid-speed-traps-and-police-checkpoints/

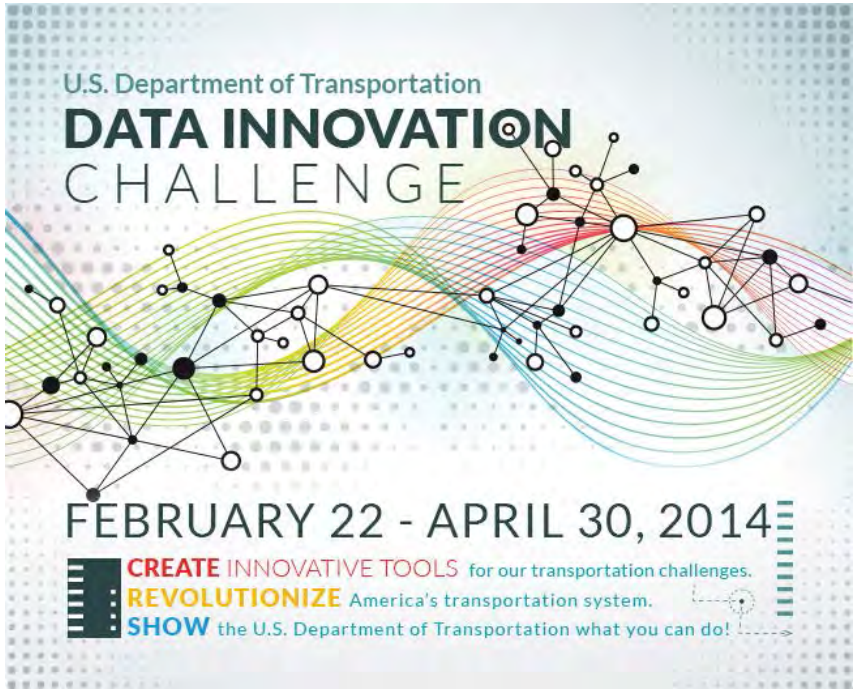
The Future

DOT turns to you with Data Innovation Challenge

> "Here at DOT, we think it's time to take transportation data to the next level, and we need your help to do it."

US Department of Transportation Secretary *Anthony Foxx*

- ***Safety***: how can we address safety concerns and challenges? What communities have the safest roads and transit, and why?
- ***Transportation Access***: how can planners improve the way transportation connects people to jobs, school, housing, and community resources?
- ***Traffic Management and Congestion***: how can we better understand and reduce traffic, congestion, and emissions?



2/15/2014

The Future

U.S. Transportation Secretary Anthony Foxx Unveils President's Bold $94.7 Billion Investment in America's Infrastructure

"U.S. Transportation Secretary Anthony Foxx today announced President Obama's $94.7 billion Fiscal Year 2016 Budget for the U.S. Department of Transportation. The proposal makes critical investments in infrastructure needed to promote long-term economic growth, enhance safety and efficiency, and support jobs for the 21st century. Speaking at a town hall at Google headquarters in Mountain View, California, Secretary Foxx today highlighted the President's budget proposal, which notably includes funding to advance research and autonomous vehicles, while announcing his report "Beyond Traffic," a look at future trends and choices that will impact America's transportation system over the next three decades."

-2/2/2015

The Future

- Urban Insights

  - Big Data + Predictive Analytics – help transportation industry improve operations, reduce costs, better serve travelers

- Industry produces volumes of data from variety of systems

  -Web applications/mobile phones, On board sensors (passenger counting systems), Vehicle location systems, Ticketing/tolling/parking/fare collection back office systems, Scheduling and asset management systems

  -Potential for insights to inform planning activities and help manage and expand transportation networks

    -Predict impact of disruptions (planned/unplanned)

    -Predict effect of transportation network changes on the local economy

    -Meeting sustainability goals

- Example applications for multi modal transport management

  -Transport operator discovers commuters connecting in unexpected locations and times

    under/over utilization => new routes, align timetables to reduce wait times

  -Simulate effect of adding new routes

  -Predict utilization of known connections to anticipate and quantify impact of proposed changes

http://www.urban-insights.com/Resources/All/articleType/CategoryView/categoryId/3/Videos

Conclusions

Privacy is an important consideration for big data in the transportation domain

- Anonymization is difficult

Population growth and urbanization are driving demand for better transportation management

- Different visions for smart cities represent alternative futures for how this might be accomplished
- Data integration/blending will play a key role

Big data solutions are not free

- Worth considering the cost tradeoffs
- South Bend spent $6M – received no improvements to actual infrastructure
- Is that $114M of savings? Or a temporary fix?

Tension between utopia and surveillance state

- Utopia could be amazing…
- Who owns the data/systems?
- What if there is a privacy driven backlash?
- Boundaries between public/private partnerships