J. Mack Robinson College of Business
GEORGIA STATE UNIVERSITY

Center for CEPRIN Process Innovation

Georgia State University

Big Data and Analytics: Emergent Trends and Opportunities for IS Scholarship

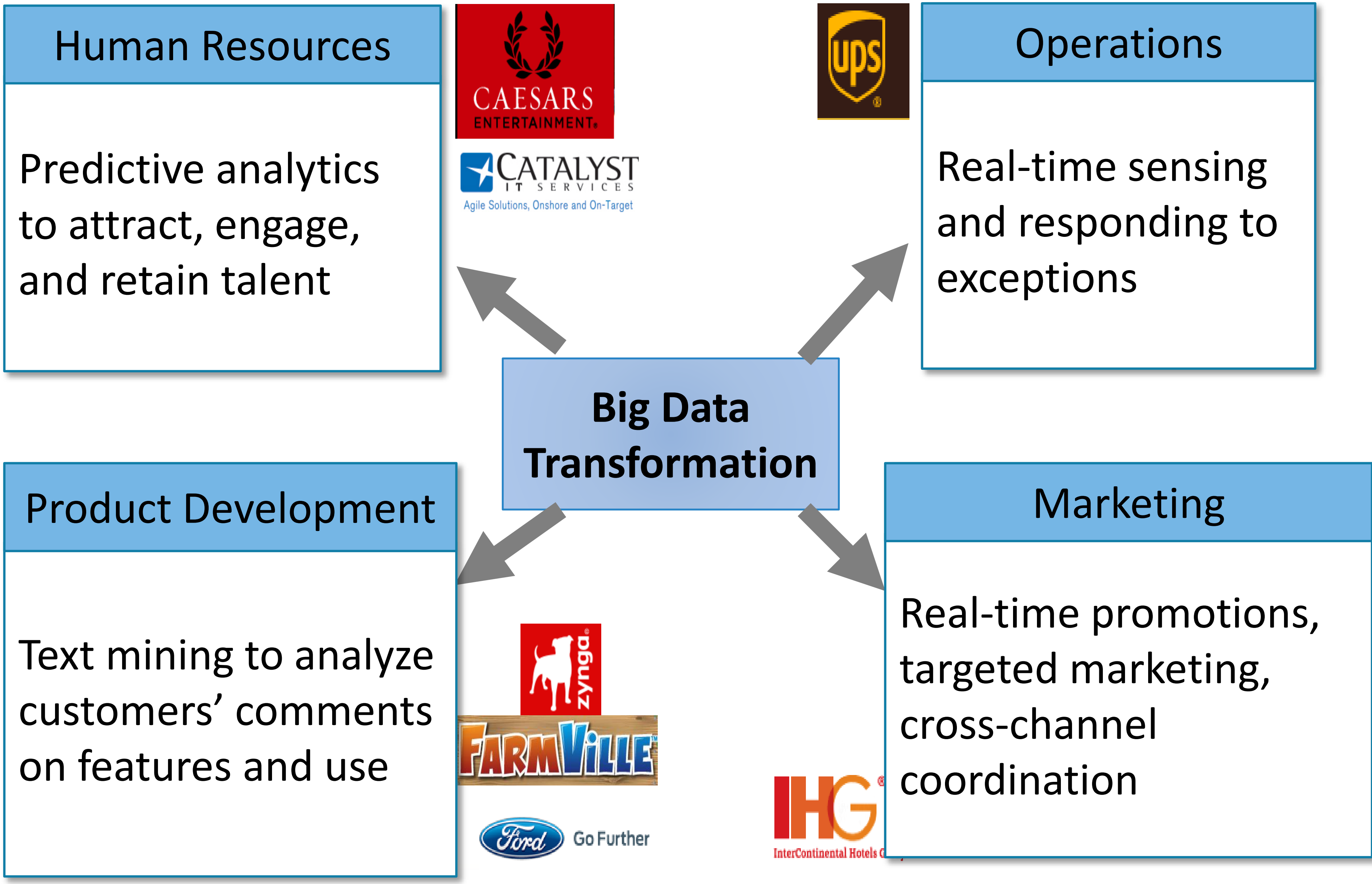
Arun Rai
Regents' Professor of the University System of Georgia &
Harkins Chair of Information Systems
Center for Process Innovation & CIS Department
Robinson College of Business
Georgia State University
Atlanta, GA 30303
arunrai@gsu.edu
Web site: arunrai.us

Presented at WeB
December 12, 2014

Big Hype or Big Change?

The Big Data Phenomenon

Expanding Big Data Applications in Business



Big Data Transformation

Human Resources

Predictive analytics to attract, engage, and retain talent

Operations

Real-time sensing and responding to exceptions

Product Development

Text mining to analyze customers' comments on features and use

Marketing

Real-time promotions, targeted marketing, cross-channel coordination

Source: How Big Data is changing the whole Equation for Business: WSJ

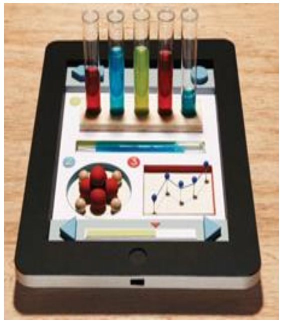
Expanding Big Data Applications in Society

Big Data Transformation

Energy

Predictive analytics to predict granular demand and adjust supply

San Francisco crimespotting

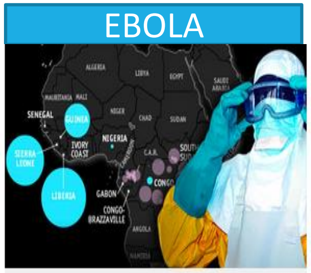


Schools

Adaptive learning platforms to predict student progress & interventions

Government

Predictive analytics to detect taxpayer fraud

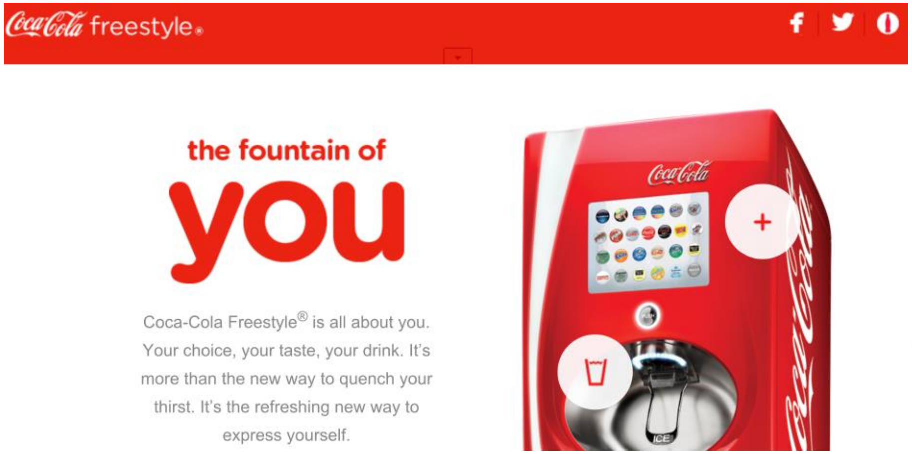


Healthcare

Collect and analyze hefty datasets to depict early infection signs

Source: Big Data Makes Big Inroads into Schools: Scientific American, How Big Data and Analytics will Change Society, FACT SHEET: Big Data and Privacy Working Group Review

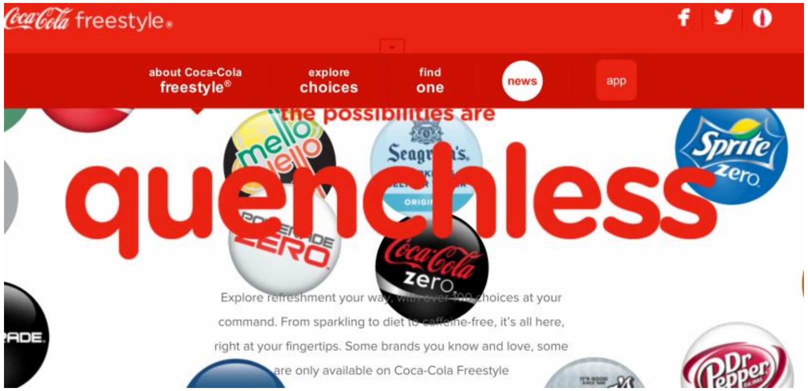
Combining IOT x Mobile x Social x Cloud x Analytics for Prediction and Action: The Case of Coca-Cola Freestyle



the fountain of

you

Coca-Cola Freestyle® is all about you. Your choice, your taste, your drink. It's more than the new way to quench your thirst. It's the refreshing new way to express yourself.



the possibilities are

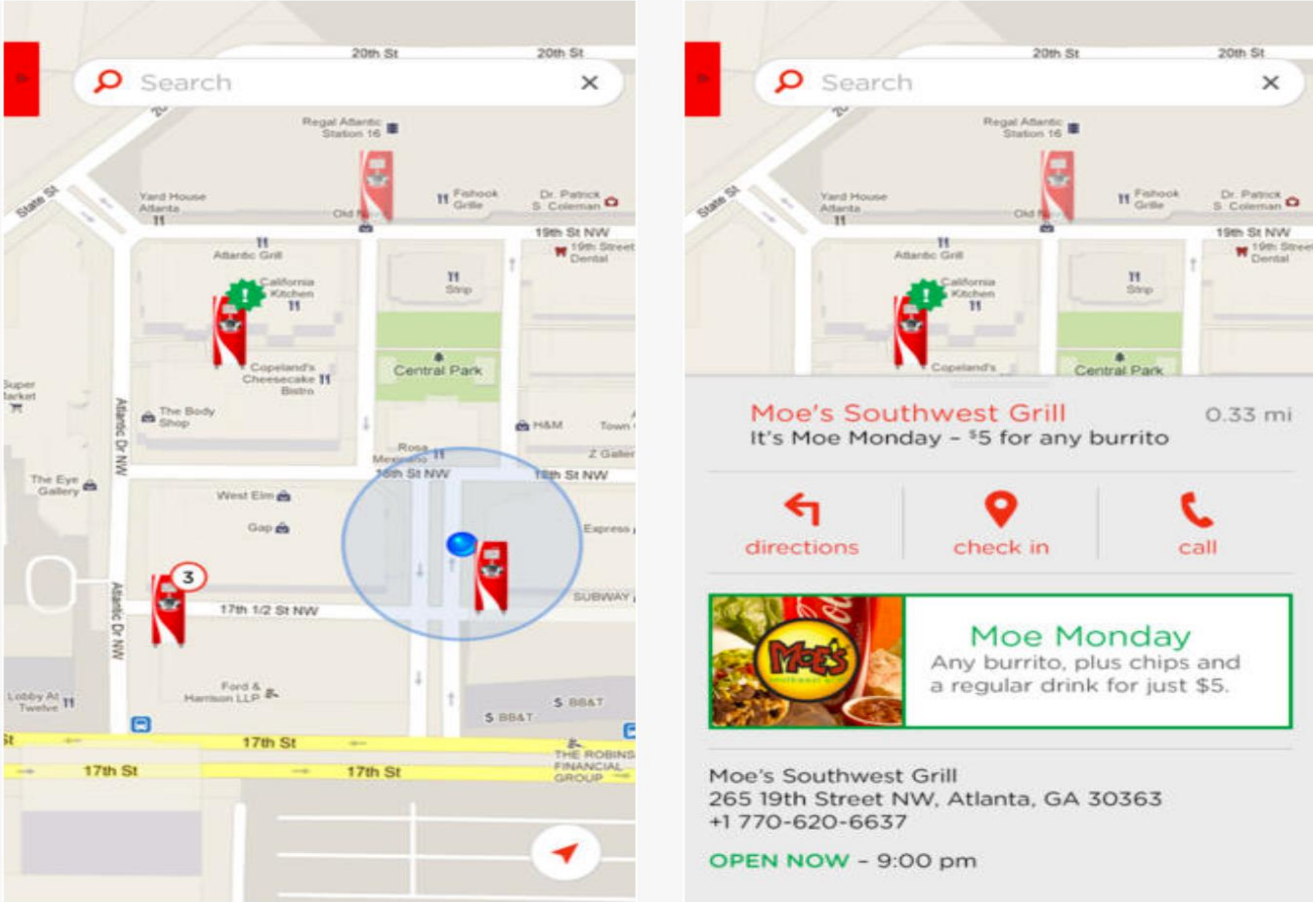
quenchless

Explore refreshment your way, with over 100 choices at your command. From sparkling to diet to caffeine-free, it's all here, right at your fingertips. Some brands you know and love, some are only available on Coca-Cola Freestyle

“Biggest invention from Coca-Cola in 20 years”— Senior Coca-Cola IT Executive

Coca-Cola Freestyle—A Machine That Tweets!

iPhone Screenshots



Moe's Southwest Grill 0.33 mi

It's Moe Monday – $5 for any burrito

Moe Monday

Any burrito, plus chips and a regular drink for just $5.

Moe's Southwest Grill
265 19th Street NW, Atlanta, GA 30363
+1 770-620-6637

OPEN NOW – 9:00 pm

Coca-Cola Freestyle for Supply Chain Execution

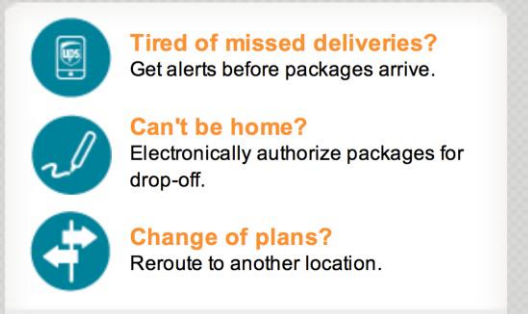
Coca-Cola freestyle

the fountain of you

Coca-Cola Freestyle® is all about you. Your choice, your taste, your drink. It's more than the new way to quench your thirst. It's the refreshing new way to express yourself.

RFID

Nurturing Consumer Experience & Enhancing Operational Excellence

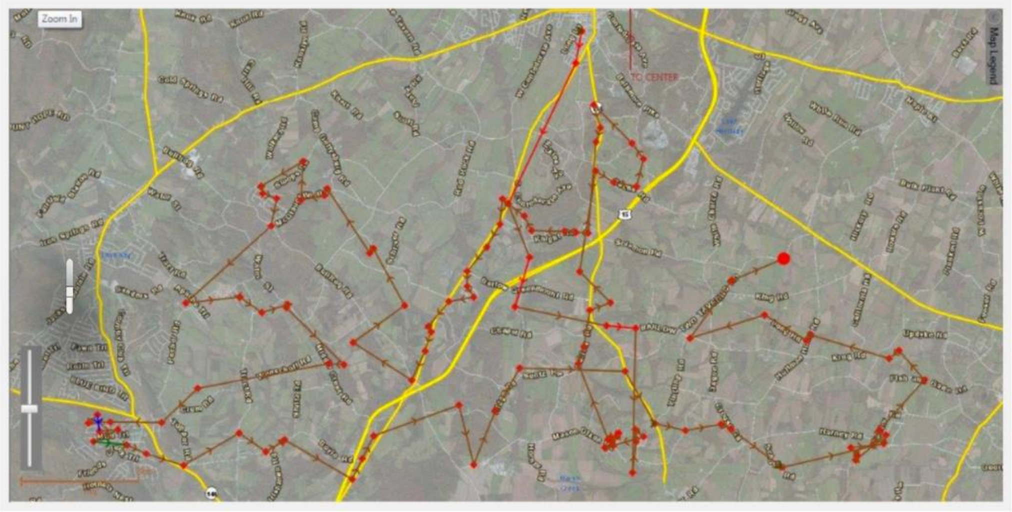


Tired of missed deliveries?
Get alerts before packages arrive.

Can't be home?
Electronically authorize packages for drop-off.

Change of plans?
Reroute to another location.

Integrated with FB to leverage the social nature of FB



An optimized route map using ORION. (Credit: UPS)

On-Road Integrated Optimization and Navigation (ORION) uses connected car-like telematics with data crunching about package info, user preferences and routes.

Customer Listening

- Phone, E-mail, Chat
- Social Media Team
- IVR Surveys
- Quality Monitoring of Call Centers

Secure and Private

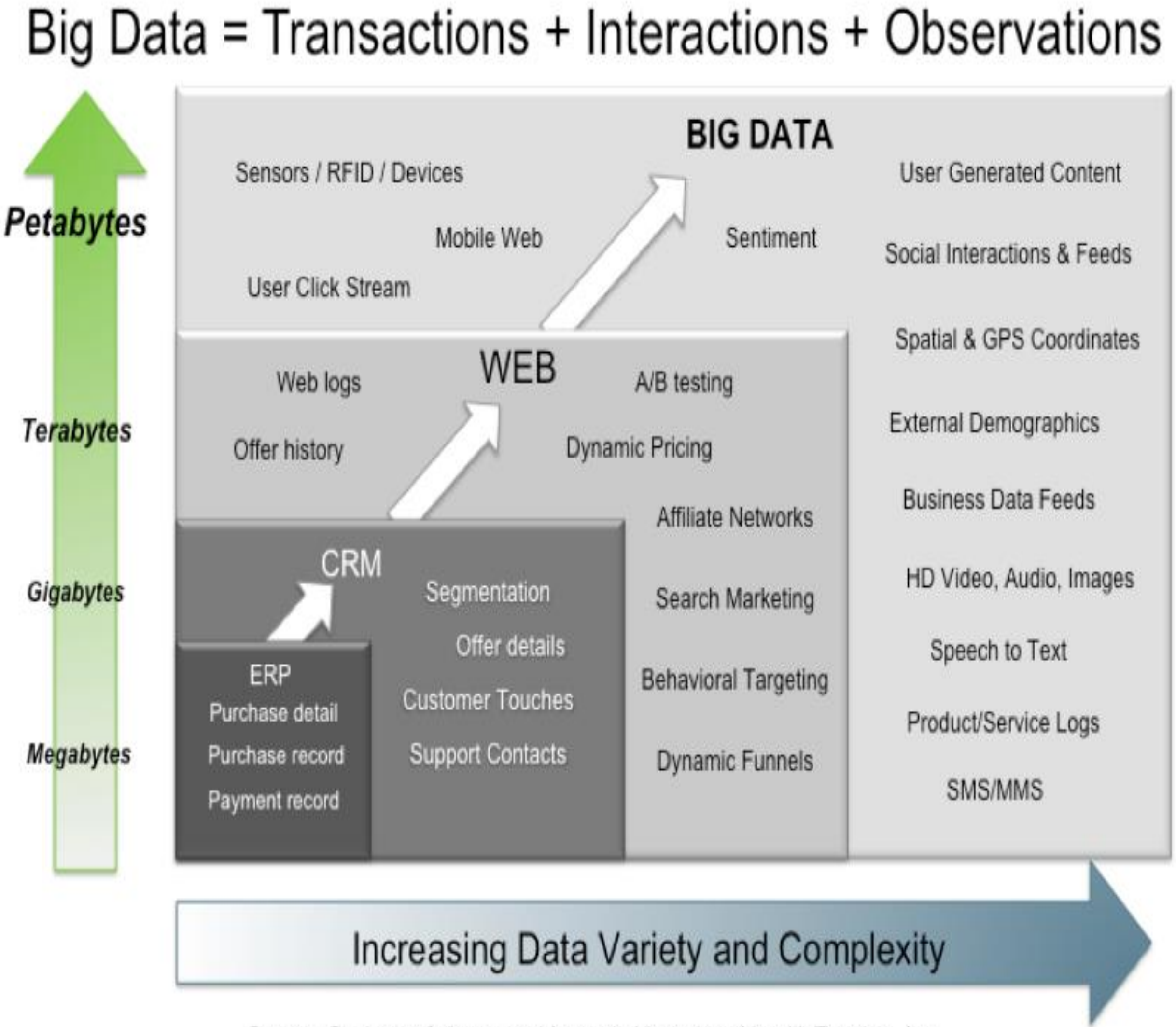
- Capability to interrogate every transaction
- Listen to social media for potential hackers

The Major Big Data Shifts

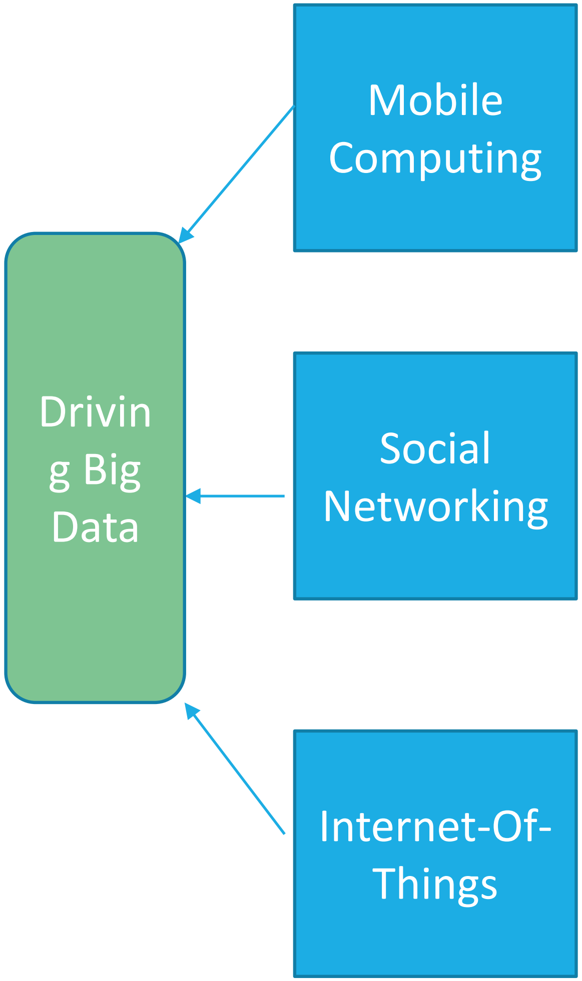
	Pre Big Data	Big Data
Culture	Highest Paid Person's Opinion (HIPPO) driving decisions	Evidence-based management, discovery and learning orientation
Data	Transactions	Transactions + Interactions + Observations
Analytics	Descriptive and Explanatory	Cycle of description, explanation, prediction, and prescription
Human-Machine Relationship	Machine substitutes human for transaction automation	Machine substitutes *and* complements human for learning and discovery

The Big Data Platform

Scope of Data Generation



Source: *Contents of above graphic created in partnership with Teradata, Inc.*



Mobile Computing

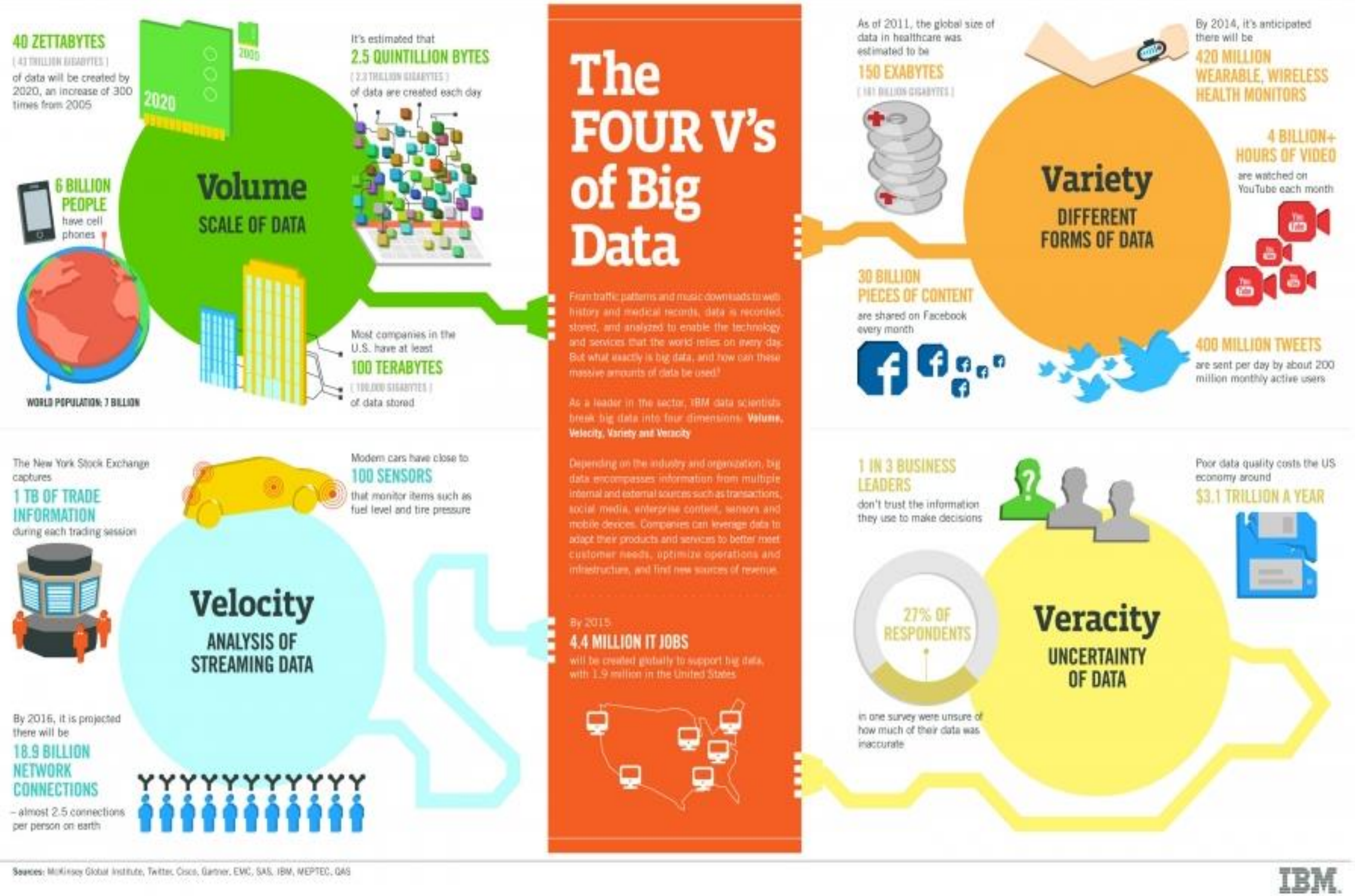
Social Networking

Internet-Of-Things

Drivin g Big Data

Source: 7 Key Drivers for the Big Data Market

The 4V's of Big Data



The FOUR V's of Big Data

From traffic patterns and music downloads to web history and medical records, data is recorded, stored, and analyzed to enable the technology and services that the world relies on every day. But what exactly is big data, and how can these massive amounts of data be used?

As a leader in the sector, IBM data scientists break big data into four dimensions: **Volume, Velocity, Variety and Veracity**

Depending on the industry and organization, big data encompasses information from multiple internal and external sources such as transactions, social media, enterprise content, sensors and mobile devices. Companies can leverage data to adapt their products and services to better meet customer needs, optimize operations and infrastructure, and find new sources of revenue.

By 2015 **4.4 MILLION IT JOBS** will be created globally to support big data, with 1.9 million in the United States

Volume
SCALE OF DATA

40 ZETTABYTES [43 TRILLION GIGABYTES] of data will be created by 2020, an increase of 300 times from 2005

It's estimated that **2.5 QUINTILLION BYTES** [2.3 TRILLION GIGABYTES] of data are created each day

6 BILLION PEOPLE have cell phones

WORLD POPULATION: 7 BILLION

Most companies in the U.S. have at least **100 TERABYTES** [100,000 GIGABYTES] of data stored

Velocity
ANALYSIS OF STREAMING DATA

The New York Stock Exchange captures **1 TB OF TRADE INFORMATION** during each trading session

Modern cars have close to **100 SENSORS** that monitor items such as fuel level and tire pressure

By 2016, it is projected there will be **18.9 BILLION NETWORK CONNECTIONS** – almost 2.5 connections per person on earth

Variety
DIFFERENT FORMS OF DATA

As of 2011, the global size of data in healthcare was estimated to be **150 EXABYTES** [161 BILLION GIGABYTES]

By 2014, it's anticipated there will be **420 MILLION WEARABLE, WIRELESS HEALTH MONITORS**

4 BILLION+ HOURS OF VIDEO are watched on YouTube each month

30 BILLION PIECES OF CONTENT are shared on Facebook every month

400 MILLION TWEETS are sent per day by about 200 million monthly active users

Veracity
UNCERTAINTY OF DATA

1 IN 3 BUSINESS LEADERS don't trust the information they use to make decisions

Poor data quality costs the US economy around **$3.1 TRILLION A YEAR**

27% OF RESPONDENTS in one survey were unsure of how much of their data was inaccurate

Sources: McKinsey Global Institute, Twitter, Cisco, Gartner, EMC, SAS, IBM, MEPTEC, QAS

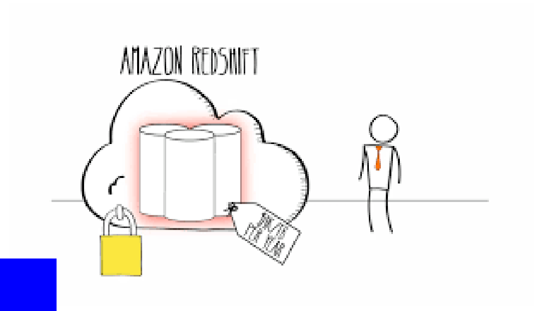
IBM

Source: The Four V's of Big Data

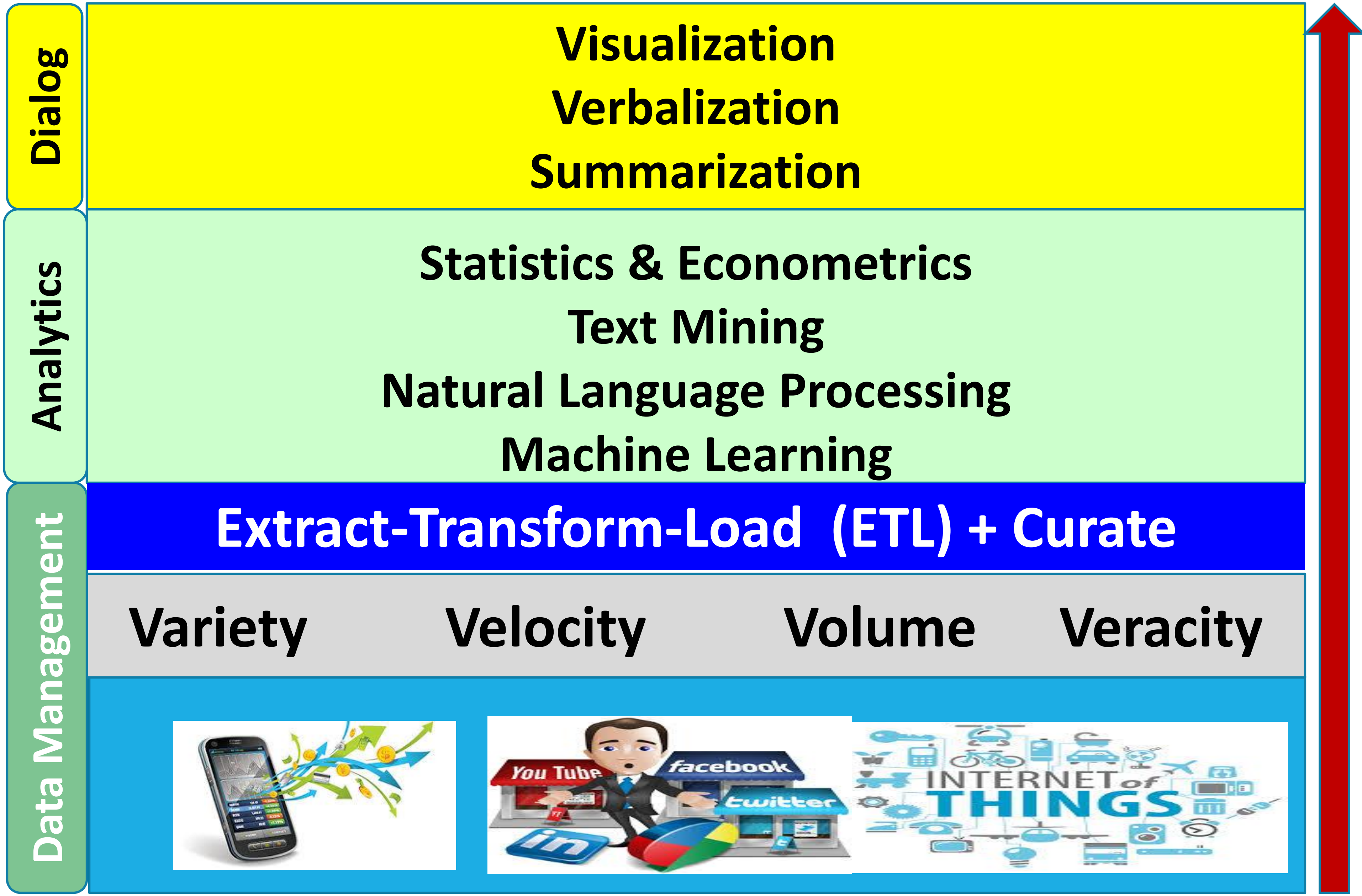
Big Data Platform

Emergent Requirements

- High throughput data
- Monitoring access control
- Encrypting
- Securing the data
- Tracing the lineage of data from source to destination
- Scaling computations
- Executing real-time analytics
- Provisioning dynamic dashboards



Big Data: A Flow Perspective



Dialog

Visualization
Verbalization
Summarization

Analytics

Statistics & Econometrics
Text Mining
Natural Language Processing
Machine Learning

Data Management

Extract-Transform-Load (ETL) + Curate

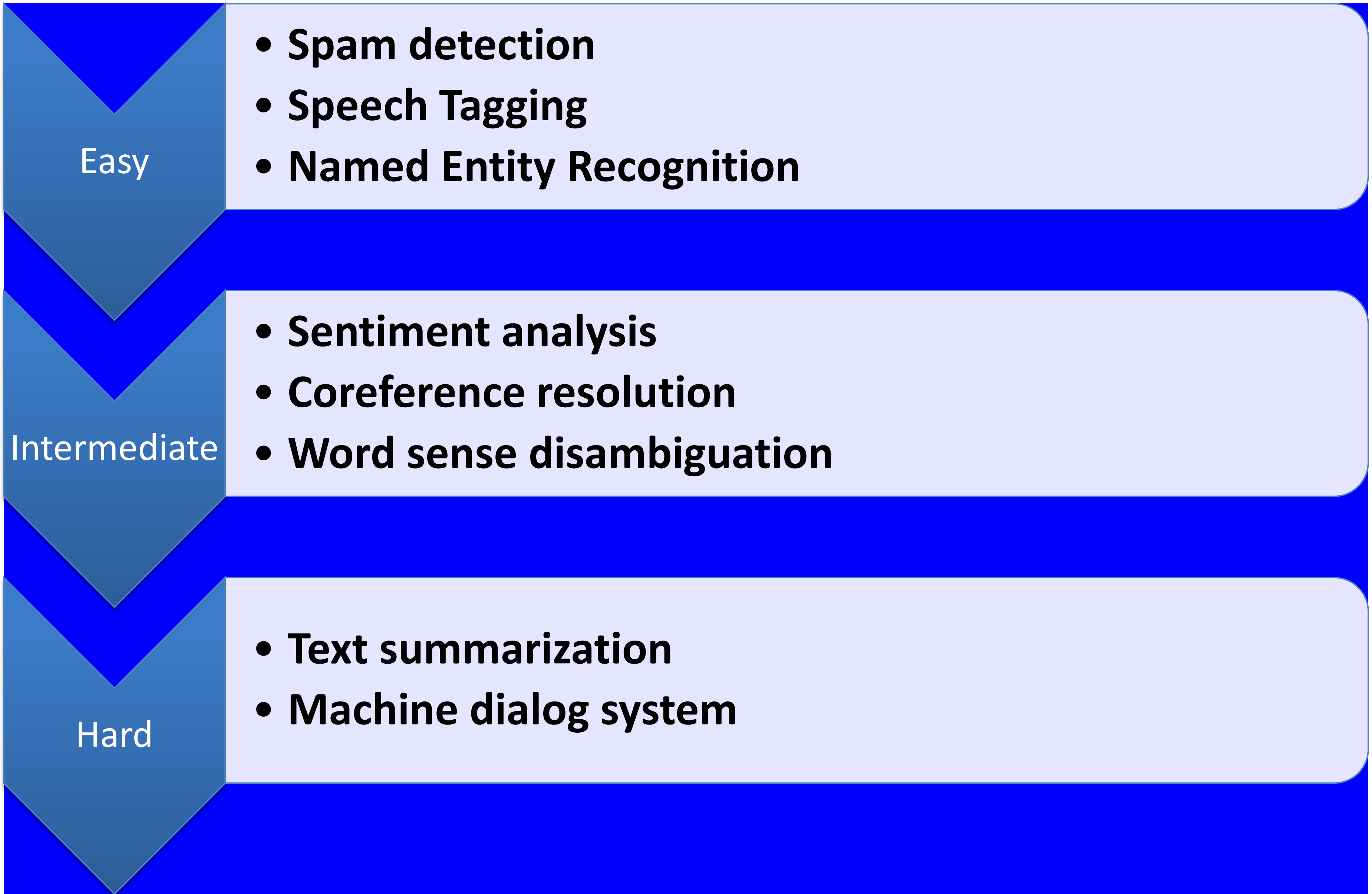
Variety Velocity Volume Veracity

Big Data Platform: Understanding the Co-Evolution of Capabilities and Governance

1. What are the capabilities of an enterprise-ready Big Data platform?
2. How should firms develop these capabilities?
3. How should the Big Data platform be governed?

Generating Insight from Text Analytics and Natural Language Processing

Progress in National Language Processing



Easy

- Spam detection
- Speech Tagging
- Named Entity Recognition

Intermediate

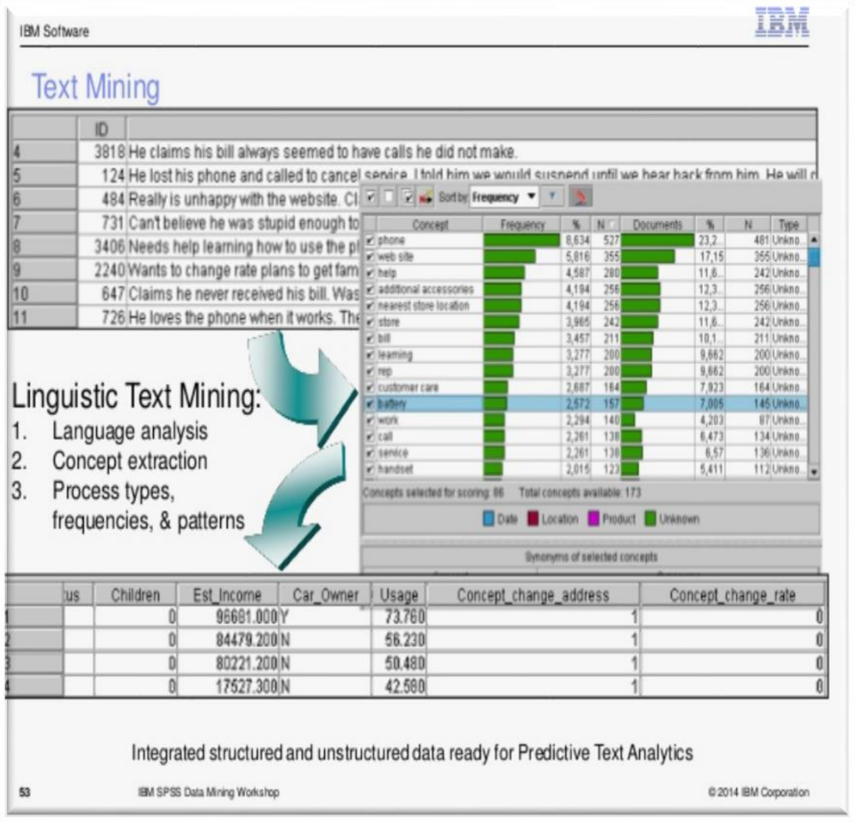
- Sentiment analysis
- Coreference resolution
- Word sense disambiguation

Hard

- Text summarization
- Machine dialog system

Using Text Analytics to Discover Themes and Derive (Valid) Measures

- Examine whether **revenue recognition complexity** increases the probability of restating reported revenue
- Measure **revenue recognition complexity** using (1) the number of words and recognition methods from the revenue recognition disclosure in the 10K reports and (2) a factor score based on the number of words and methods



IBM Software

IBM

Text Mining

	ID	
4	3818	He claims his bill always seemed to have calls he did not make.
5	124	He lost his phone and called to cancel service. I told him we would suspend until we hear back from him. He will d
6	484	Really is unhappy with the website. Cl
7	731	Can't believe he was stupid enough to
8	3406	Needs help learning how to use the pl
9	2240	Wants to change rate plans to get fam
10	647	Claims he never received his bill. Was
11	726	He loves the phone when it works. The

☑ □ ☑ Sort by: Frequency ▼

	Concept	Frequency	%	N	Documents	%	N	Type
☑	phone		8,634	527		23,2...	481	Unkno...
☑	web site		5,816	355		17,15	355	Unkno...
☑	help		4,587	280		11,6...	242	Unkno...
☑	additional accessories		4,194	256		12,3...	256	Unkno...
☑	nearest store location		4,194	256		12,3...	256	Unkno...
☑	store		3,965	242		11,6...	242	Unkno...
☑	bill		3,457	211		10,1...	211	Unkno...
☑	learning		3,277	200		9,662	200	Unkno...
☑	rep		3,277	200		9,662	200	Unkno...
☑	customer care		2,687	164		7,923	164	Unkno...
☑	battery		2,572	157		7,005	145	Unkno...
☑	work		2,294	140		4,203	87	Unkno...
☑	call		2,261	138		6,473	134	Unkno...
☑	service		2,261	138		6,57	136	Unkno...
☑	handset		2,015	123		5,411	112	Unkno...

Concepts selected for scoring: 86 Total concepts available: 173

🟦 Date 🟥 Location 🟪 Product 🟩 Unknown

Synonyms of selected concepts

Linguistic Text Mining:

1. Language analysis
2. Concept extraction
3. Process types, frequencies, & patterns

	us	Children	Est_Income	Car_Owner	Usage	Concept_change_address	Concept_change_rate
1		0	96681.000	Y	73.760	1	0
2		0	84479.200	N	56.230	1	0
3		0	80221.200	N	50.480	1	0
4		0	17527.300	N	42.580	1	0

Integrated structured and unstructured data ready for Predictive Text Analytics

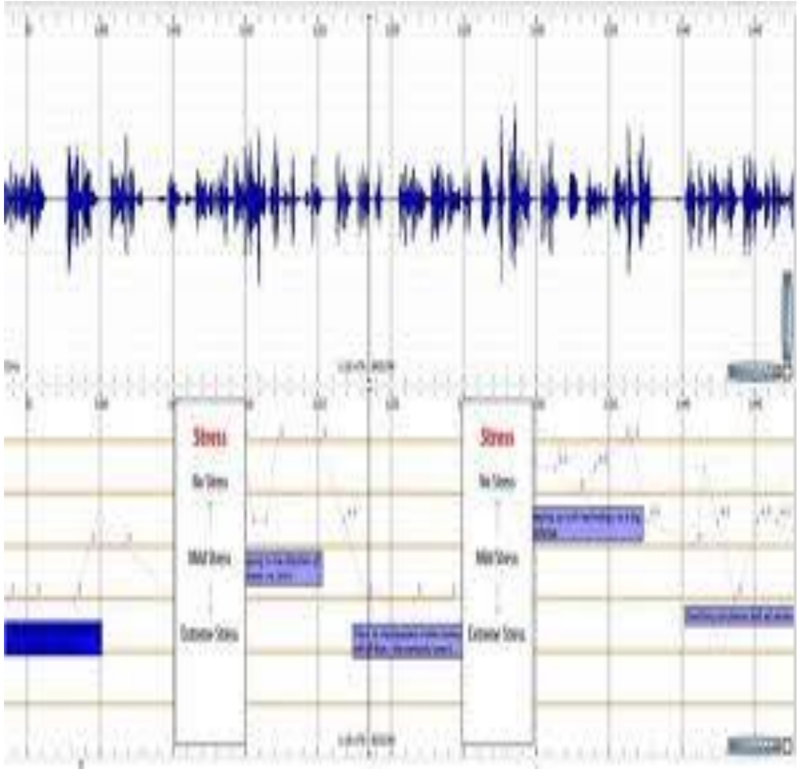
53 IBM SPSS Data Mining Workshop ©2014 IBM Corporation

Peterson, *Review of Accounting Studies*, 2012

Using Speech Analytics to Generate Consonance/Dissonance Markers

- Examine whether **vocal markers of cognitive dissonance** are useful for detecting financial misreporting

- Use **speech samples of CEOs during earnings conference calls**, and generate vocal dissonance markers using automated vocal emotion analysis software



Hobson, Mayew, & Venkatachalam, *Journal of Accounting Research, 2012*

Research Opportunities

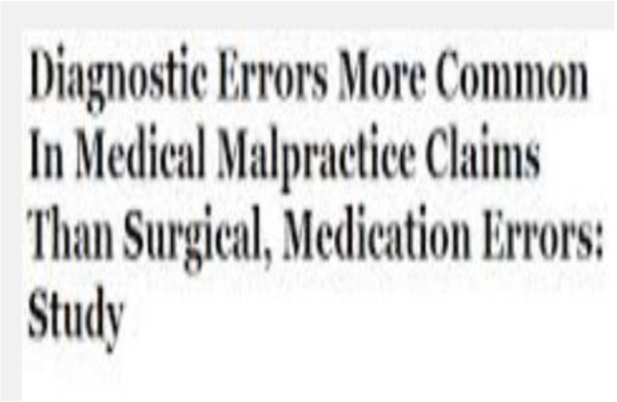
1. How can statistical and econometric methods that are used across a variety of IS problem domains be combined with a) text mining and b) speech analysis?

2. What additional insights are realized from this combination?

How Machine Learning Is Changing the Human-Machine Relationship

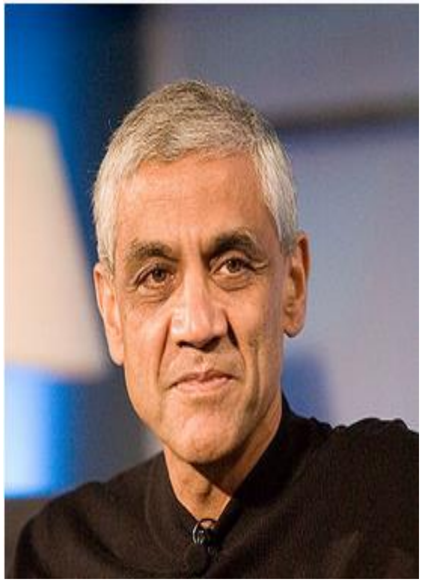
Medical Practice (Not Science): State of Affairs

- Misdiagnosis, conflicting diagnoses, general diagnostic error
- ICU misdiagnoses cause as many deaths as breast cancer
- Adverse drug interactions cause as many deaths as automobile accidents
- Preventable medical errors, often with clinical findings already in the medical record, are common.


Diagnostic Errors More Common In Medical Malpractice Claims Than Surgical, Medication Errors: Study

Medical Practice (Not Science): State of Affairs

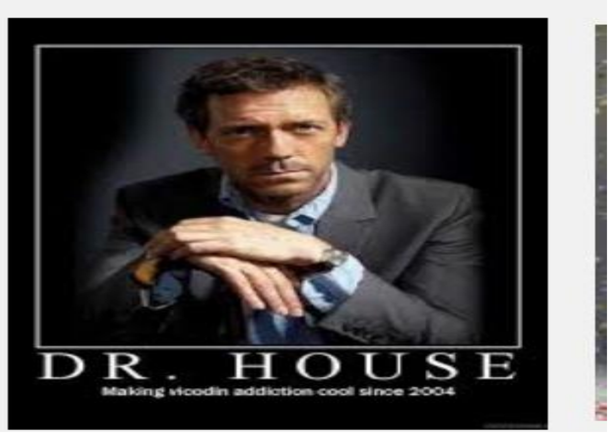
“Today’s diagnostic error rate is the equivalent of Google’s driverless car having one accident per week; while this would be unacceptable for self-driving cars; this failure rate is permissible in healthcare.”



Vinod Khosla
Khosla Ventures, Founder
Sun Microsystems, Co-Founder
General Partner, Kleiner-Perkins Caufield & Byers

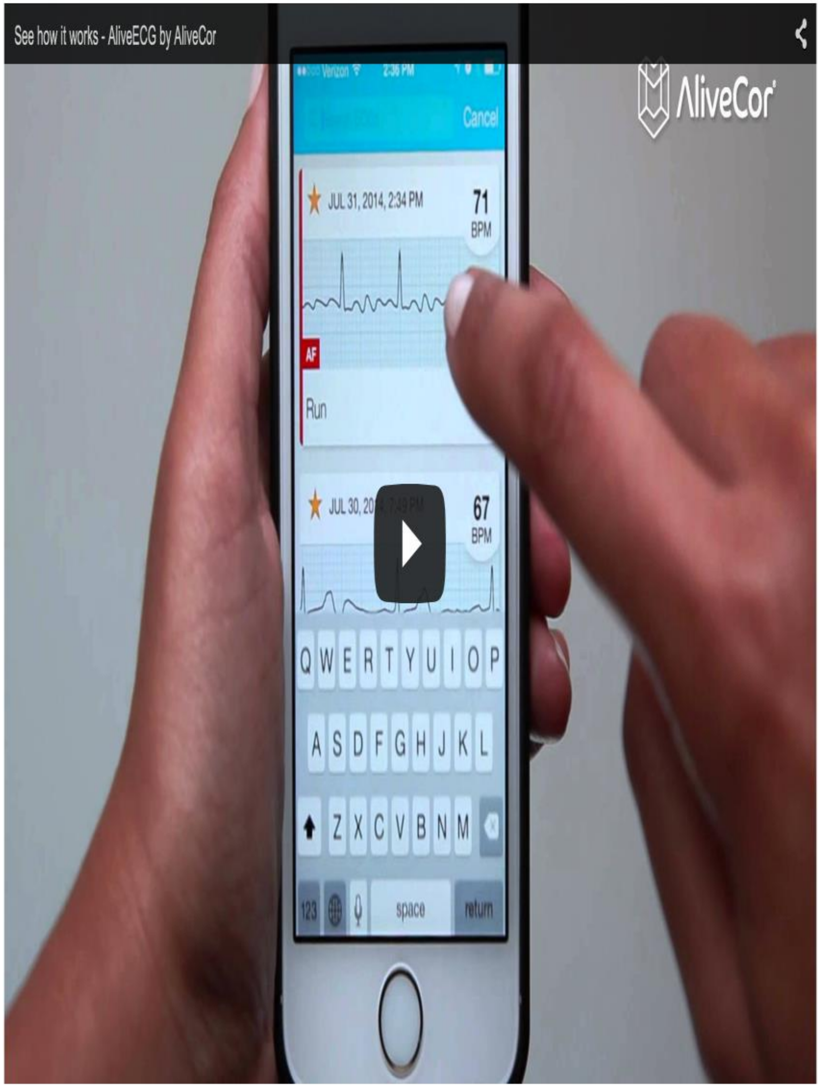
Source: http://techcrunch.com/2014/09/22/the-reinvention-of-medicine-dr-algorithm-version-0-7-and-beyond/

Will Machine Learning Lead to a Much Better Dr. House than Dr. House



Diagnostic errors due to premature closure, recency bias

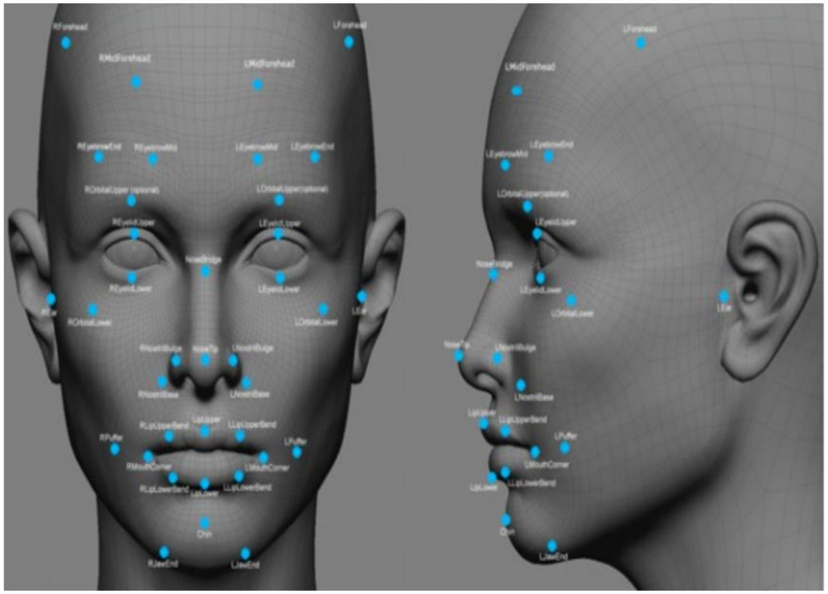
- Becoming more accurate at diagnosing
- Will create broad-scale access to wellness *and* sickness care
- Inexpensive data-gathering; continual monitoring and ubiquitous information leading to personalized, precise and consistent insights


See how it works - AliveECG by AliveCor
AliveCor
Verizon
2:36 PM
Cancel
JUL 31, 2014, 2:34 PM
71
BPM
AF
Run
JUL 30, 20
67
BPM
Q W E R T Y U I O P
A S D F G H J K L
Z X C V B N M
123
space
return

Facebook’s DeepFace

Verification

- Are two photos the same?
- DeepFace: 97.25% accuracy
- Humans: 97.53% accuracy



Progress in recognition: looking at a new photo and connecting it to the name of an existing user

IBM’s Watson

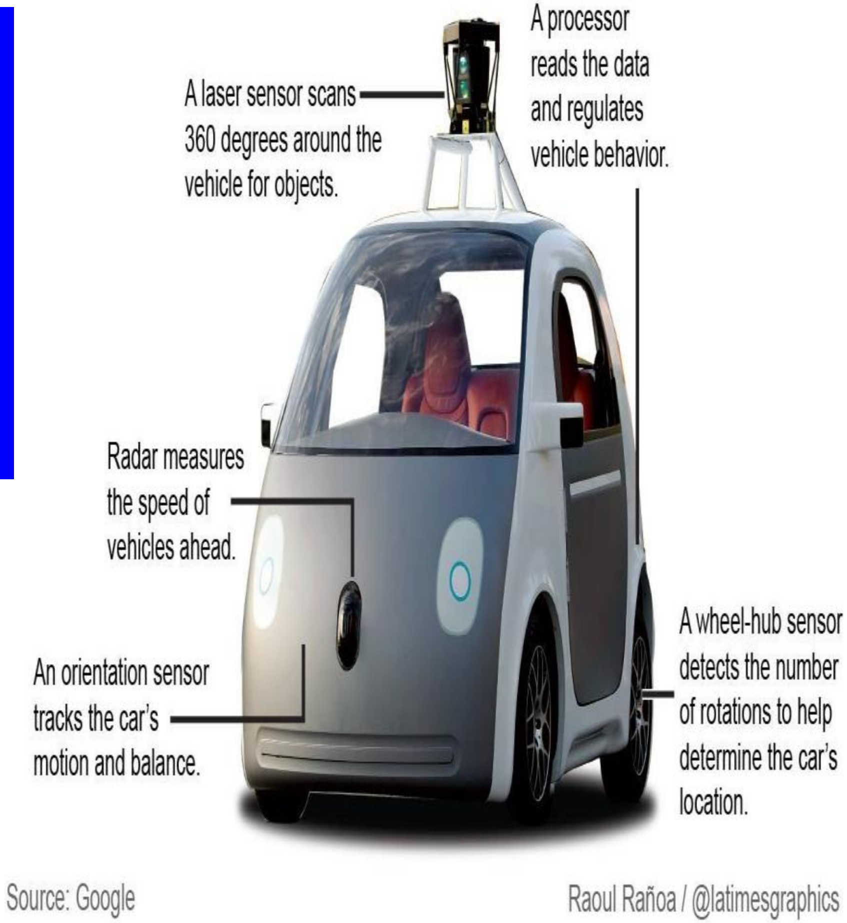
JEOPARDY
THE RIGHT WAY TO SOLVE THE LOOMING DEBT CRISIS.
UM... DUH, GUYS.
...AND WATSON WINS AGAIN.
-$3,792
What are TAX HIKES?
$40,961
What is Entitlement Reform?
-$3,538
WHAT ARE SPENDING CUTS?

"I'm Googling to the Grocery Store"

"Tests of Google's autonomous vehicles in California and Nevada suggests they already outperform human drivers."

Breakdowns:

- Heavy rain and snow-covered roads
- Encountering stalled vehicle over the crest of a hill
- Identifying debris in the middle of the road


A laser sensor scans 360 degrees around the vehicle for objects.
A processor reads the data and regulates vehicle behavior.
Radar measures the speed of vehicles ahead.
An orientation sensor tracks the car's motion and balance.
A wheel-hub sensor detects the number of rotations to help determine the car's location.
Source: Google
Raoul Rañoa / @latimesgraphics

Chris Urmson, Director of the Google car team:
http://www.forbes.com/sites/joannmuller/2013/03/21/no-hands-no-feet-my-unnerving-ride-in-googles-driverless-car/

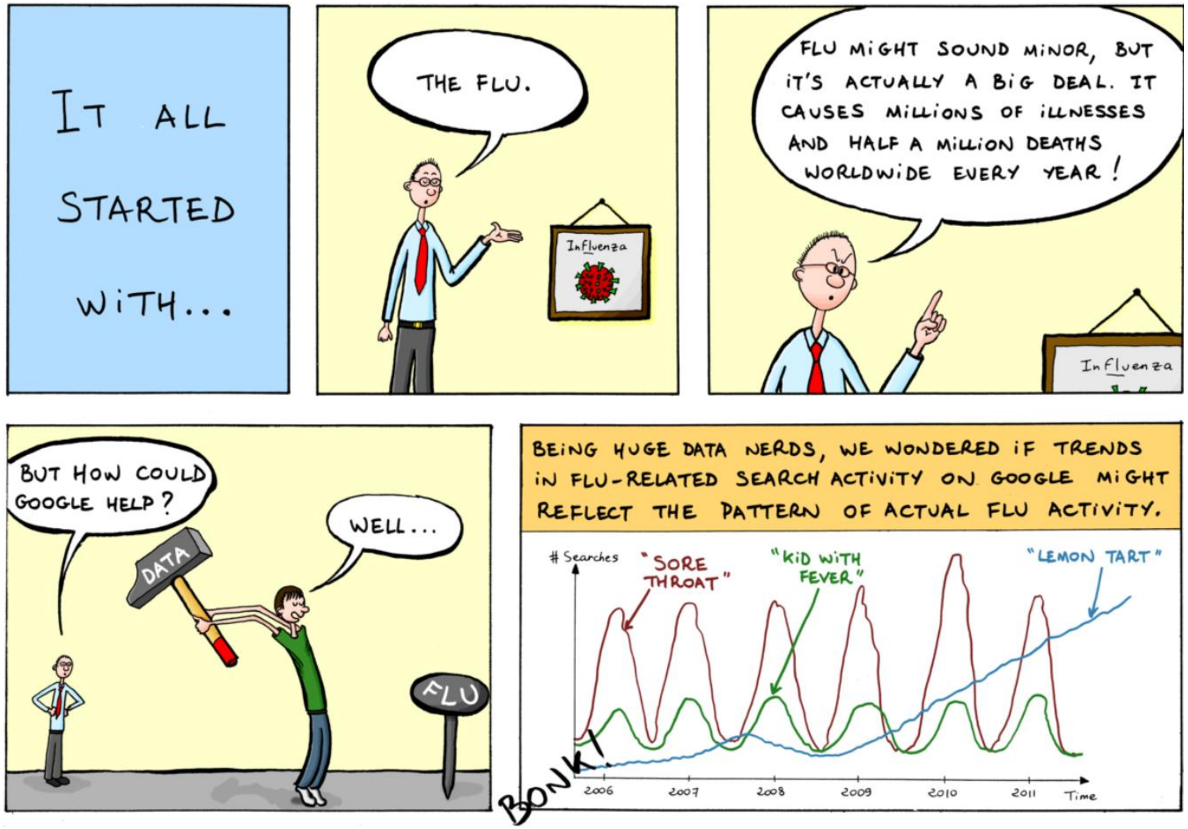
Human Expertise: ‘Miracle on the Hudson’

How Sullenberger Really Saved US Airways Flight 1549

IS Research Issues

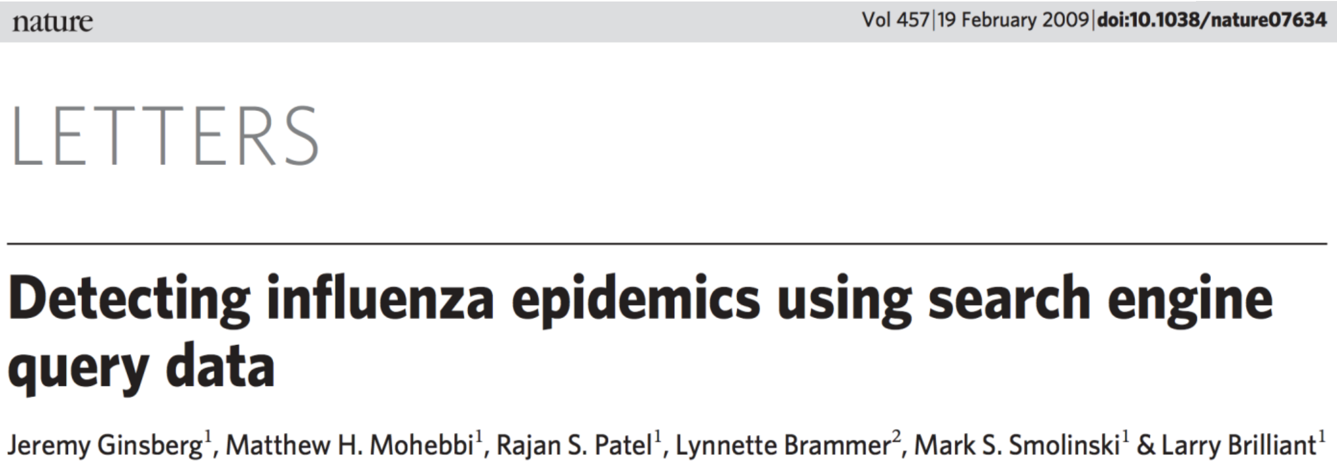
1. How is the human-machine relationship changing? How is the change to be managed?
2. How can causality-oriented IS research be combined with prediction-oriented machine learning?
3. What added insights result from this combination in different IS problem domains?
4. What are the templates for manuscripts that employ this combination?

Pitfalls with Big Data Analytics:
Insights from the Google Flu Tracker Parable


IT ALL STARTED WITH...
THE FLU.
Influenza
FLU MIGHT SOUND MINOR, BUT IT'S ACTUALLY A BIG DEAL. IT CAUSES MILLIONS OF ILLNESSES AND HALF A MILLION DEATHS WORLDWIDE EVERY YEAR!
Influenza
BUT HOW COULD GOOGLE HELP?
WELL...
DATA
FLU
BONK!
BEING HUGE DATA NERDS, WE WONDERED IF TRENDS IN FLU-RELATED SEARCH ACTIVITY ON GOOGLE MIGHT REFLECT THE PATTERN OF ACTUAL FLU ACTIVITY.
# Searches
"SORE THROAT"
"KID WITH FEVER"
"LEMON TART"
2006
2007
2008
2009
2010
2011
Time

Created by the Google Correlate team and Manu Cornet.
Inspired by the Google Chrome comic book.
© Google 2011

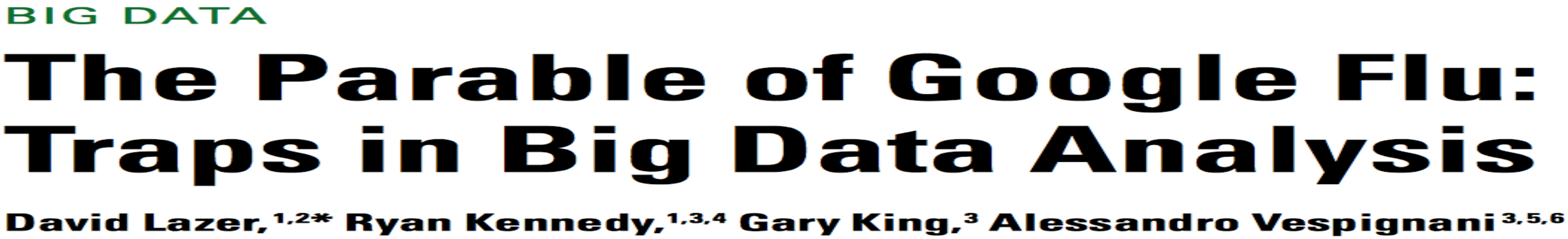
- "Tracking 45 flu-related search terms over billions of searches, monitoring trends and making correlations would win out. Google could tap the "collective intelligence" of society in real time, free of the human bias and hypotheses of traditional methods." ***New York Times***


nature Vol 457|19 February 2009|doi:10.1038/nature07634

LETTERS

Detecting influenza epidemics using search engine query data

Jeremy Ginsberg[1], Matthew H. Mohebbi[1], Rajan S. Patel[1], Lynnette Brammer[2], Mark S. Smolinski[1] & Larry Brilliant[1]



BIG DATA

The Parable of Google Flu: Traps in Big Data Analysis

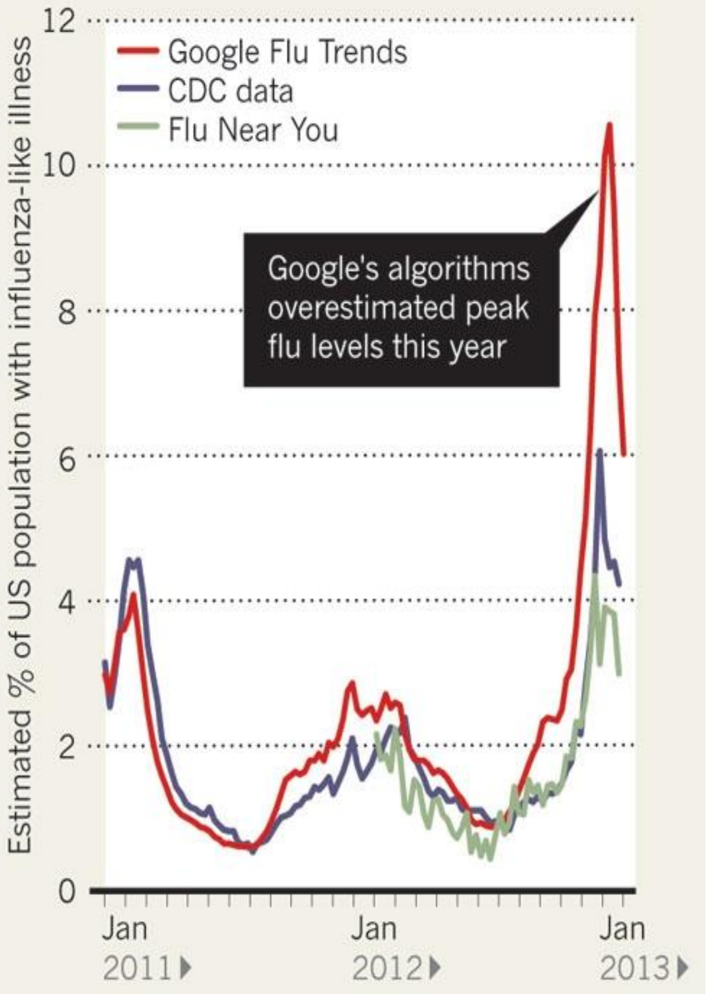
David Lazer,[1,2*] Ryan Kennedy,[1,3,4] Gary King,[3] Alessandro Vespignani[3,5,6]

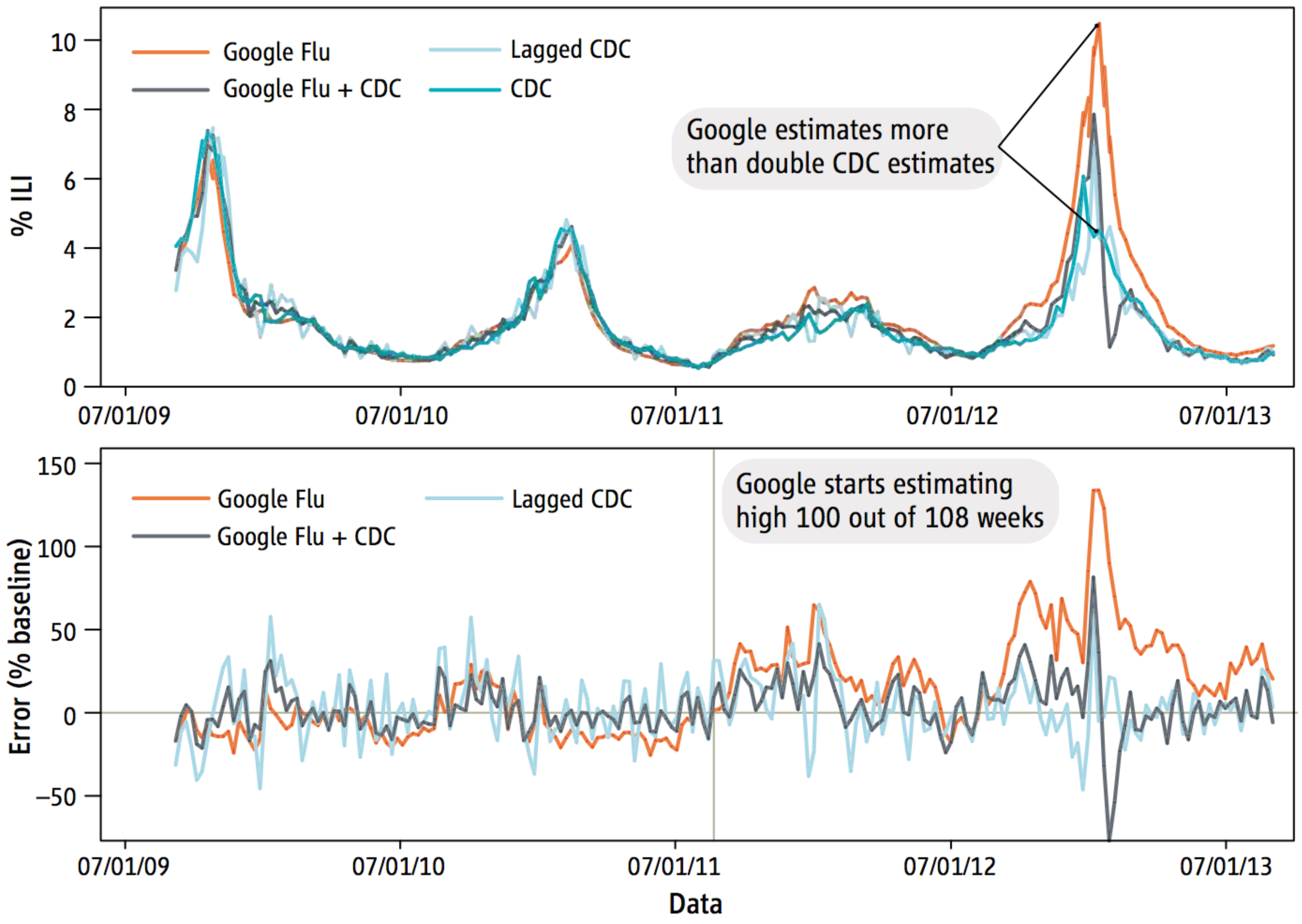
“**The problems we identify are not limited to GFT.** Research on whether search or social media can predict x has become common-place and is often put in sharp with traditional methods and hypotheses. **Although these studies have shown the value of these data, we are far from a place where they can supplant more traditional methods or theories.**” (emphasis added)

- Initial version part flu detector, part winter detector—high odds of findings search terms that match flu propensity but structurally unrelated
- Completely missed 2009 H1N1 nonseasonal pandaemic
- Algorithm updated in 2009, largely unchanged with few changes announced in October 2013
- Missed high for 100 out of 108 weeks since Aug 2011
- Errors are not random – seasonality and temporal autocorrelation

FEVER PEAKS

A comparison of three different methods of measuring the proportion of the US population with an influenza-like illness.

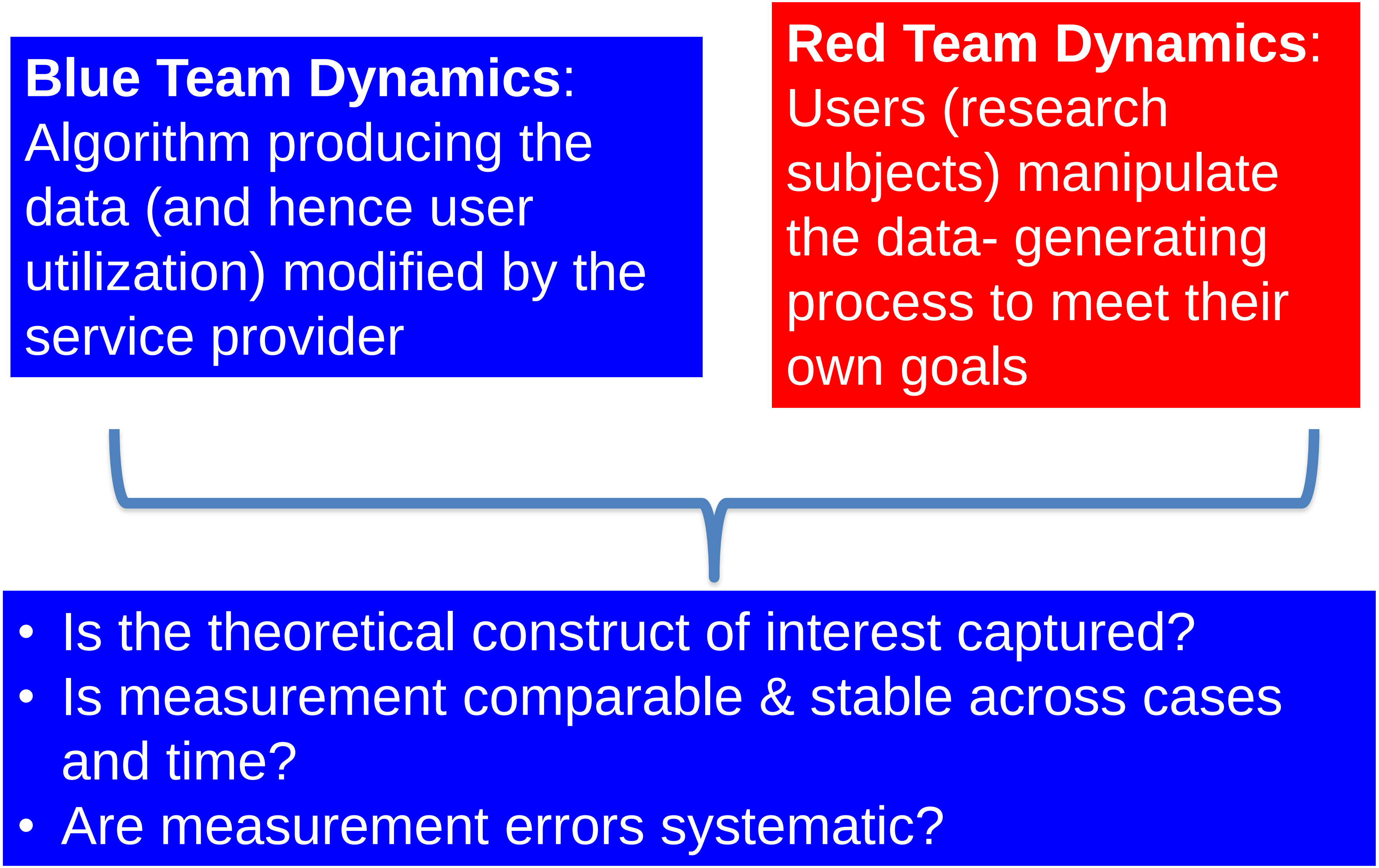




Google Flu
Google Flu + CDC
Lagged CDC
CDC
Google estimates more than double CDC estimates
% ILI
10
8
6
4
2
0
07/01/09
07/01/10
07/01/11
07/01/12
07/01/13
Google Flu
Google Flu + CDC
Lagged CDC
Google starts estimating high 100 out of 108 weeks
Error (% baseline)
150
100
50
0
−50
07/01/09
07/01/10
07/01/11
07/01/12
07/01/13
Data

Big Data Hubris

Algorithm Dynamics & Research Implications



Blue Team Dynamics: Algorithm producing the data (and hence user utilization) modified by the service provider

Red Team Dynamics: Users (research subjects) manipulate the data- generating process to meet their own goals

- Is the theoretical construct of interest captured?
- Is measurement comparable & stable across cases and time?
- Are measurement errors systematic?

“Blue Team” Dynamics

Google Flu Tracker Assumption: Relative search volume for certain terms statically related to external events (exogenously determined)

But: Google *changes* the data-generating process by providing users useful information quickly and, in part, to promote advertising revenue

Platforms such as Twitter and Facebook are always being re-engineered

Can studies conducted even a year ago on data collected from these platforms be replicated?

“Red Team” Dynamics

Economic or political gain motivators to manipulate the data-generation process

Campaigns and companies, aware that news media are monitoring Twitter, use tactics to ensure their candidate or product is trending

“Ironically, the more successful we become at monitoring the behavior of people using these open sources of information, the more tempting it will be to manipulate those signals.”

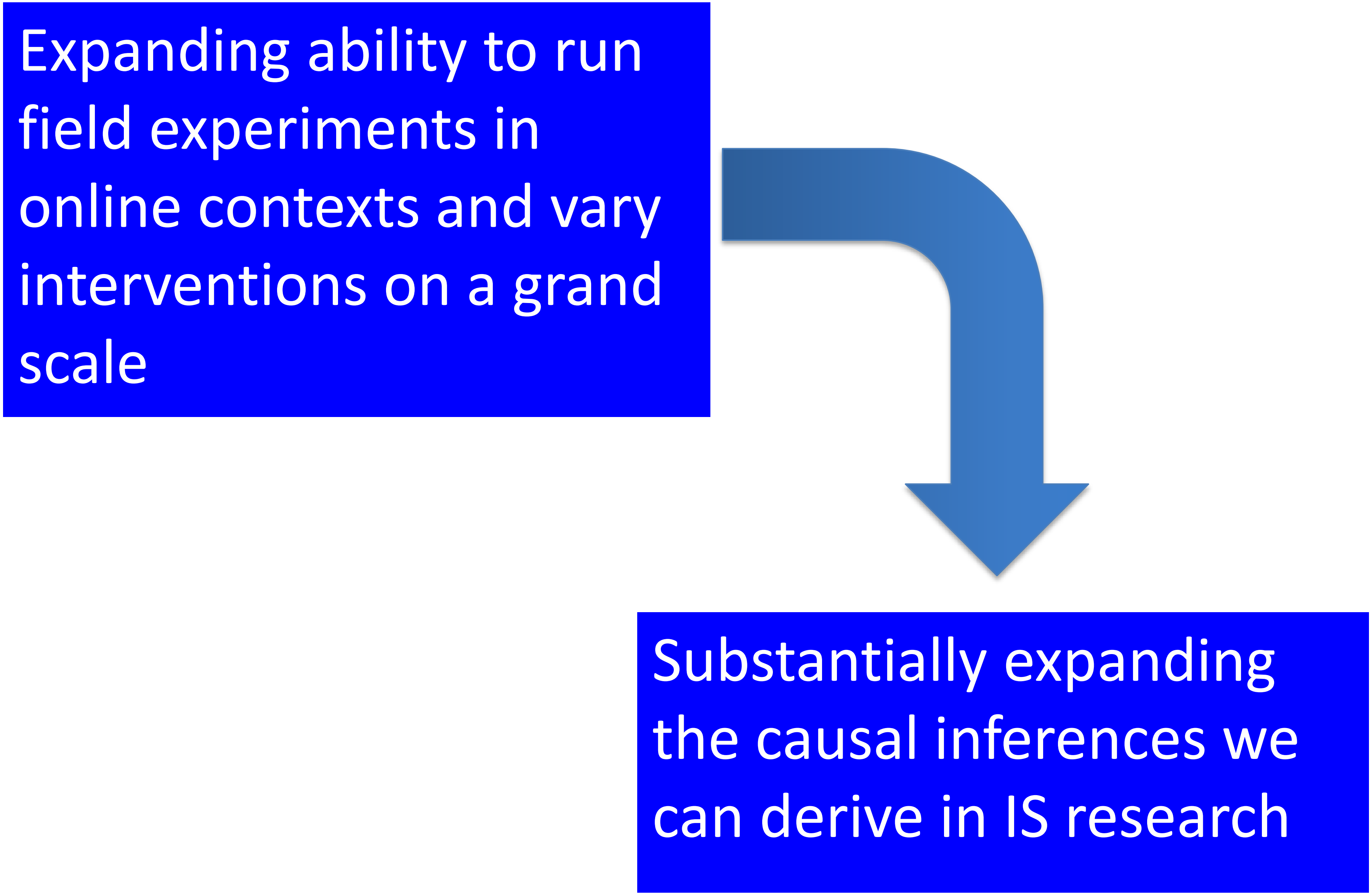
IS Research Issues

1. Evaluate user-generated online content for “blue team” and “red team” dynamics

2. Complement Big Data (especially when measurement quality is unclear) with traditional data collection/analysis that are based on reliable and valid instruments

3. Report on the measurement quality of Big Data

Randomized Online Experiments: Lessons from Facebook



Expanding ability to run field experiments in online contexts and vary interventions on a grand scale

Substantially expanding the causal inferences we can derive in IS research

The Experiment By Facebook

Determine whether the mood of users could be manipulated

Manipulated 689,000 users' home feeds: Some people shown content with a preponderance of happy/positive content; some shown content sadder than average.

Manipulated users were more likely to post either especially positive or negative words themselves.

Secret study in which people's Facebook posts were moved to influence moods has angered users. Have you lost trust in the network?

Visualizing in Context

- Representing tidal wave of data in a relevant and understandable fashion
- **Visualization decisions:**
  - *Web*: once a week, graph
  - *Augmented Reality*: real-time, real-place
- How should visualization decisions be made given characteristics of the user and the usage context?

Center for CEPRIN Process Innovation

J. Mack Robinson College of Business
GEORGIA STATE UNIVERSITY

Georgia State University

Comments welcome!

Arun Rai
arunrai@gsu.edu
Web site: arunrai.us

Presented at WeB
December 12, 2014