Limitations of Big Data for Solving Business Problems

Alan Montgomery
Associate Professor, Tepper School of Business
Carnegie Mellon University

HKU MSc (Ecom&Icomp) Experts Address
2 July 2014 (7-8pm)
ADC315, 3/F, HKU SPACE, Admiralty Centre, 18 Harcourt Road, Hong Kong

Outline

- The Promise of Big Data
- Case Studies to Illustrate the Limitations of Big Data
  - Assessing Model Uncertainty in Financial Risk Management
  - Retail Price Optimization using Business Rules
  - Predicting Flu Trends
- Conclusions about Limitations of Big Data

The Promise of Big Data

Press about Big Data

What is Big Data?

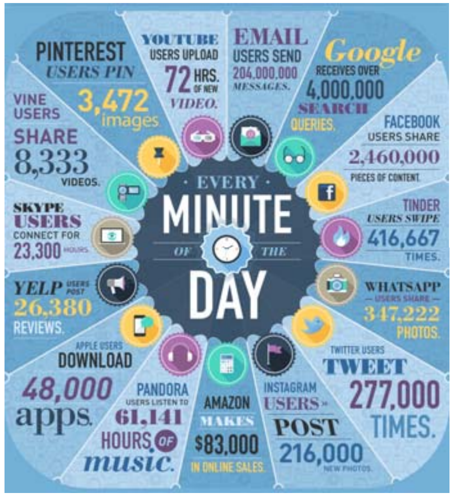
- Four basic components:
  - Massive datasets
  - Unstructured data
  - Collected as a by-product from transactions (not for decision making)
  - Populations not samples

Related to Business Analytics, Data Analysis, Data Mining, Data Science, Machine Learning, Statistics

Consumers generate Big Data

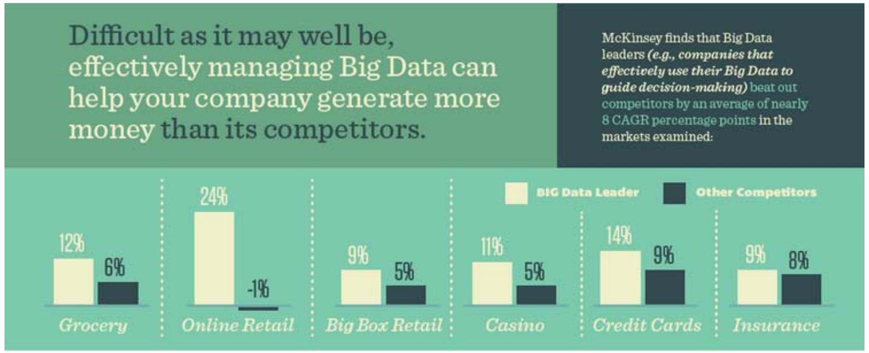
WAL★MART
ALWAYS OBEY ALWAYS WAL-MART
Always
# ITEMS SOLD 4
facebook
facebook
ebY
Active Selling (4)
Sold - Marked as Paid (1)
Google UK
twitter
OMG church was freakin' awesome this morning!!!
angrymulletbear

How much data is generated every minute?


PINTEREST USERS PIN 3,472 images
YOUTUBE USERS UPLOAD 72 HRS. OF NEW VIDEO.
EMAIL USERS SEND 204,000,000 MESSAGES.
Google RECEIVES OVER 4,000,000 SEARCH QUERIES.
VINE USERS SHARE 8,333 VIDEOS.
FACEBOOK USERS SHARE 2,460,000 PIECES OF CONTENT.
SKYPE USERS CONNECT FOR 23,300 HOURS.
TINDER USERS SWIPE 416,667 TIMES.
· EVERY · MINUTE OF THE DAY
YELP USERS POST 26,380 REVIEWS.
WHATSAPP — USERS SHARE — 347,222 PHOTOS.
APPLE USERS DOWNLOAD 48,000 apps.
PANDORA USERS LISTEN TO 61,141 HOURS OF music.
AMAZON MAKES $83,000 IN ONLINE SALES.
INSTAGRAM USERS POST 216,000 NEW PHOTOS.
TWITTER USERS TWEET 277,000 TIMES.

- Global Internet population is 2013 is 2.4 billion people
- Source domo.com

Competitive Advantage from Big Data



Difficult as it may well be, effectively managing Big Data can help your company generate more money than its competitors.

McKinsey finds that Big Data leaders *(e.g., companies that effectively use their Big Data to guide decision-making)* beat out competitors by an average of nearly 8 CAGR percentage points in the markets examined:

Source: McKinsey, http://www.domo.com/learn/infographic-sensory-overload

Quotes on Big Data

"Information is the oil of the 21st century, and analytics is the combustion engine"

Peter Sondergaard of the Gartner Group

"Data is the new science. Big Data holds the answers"

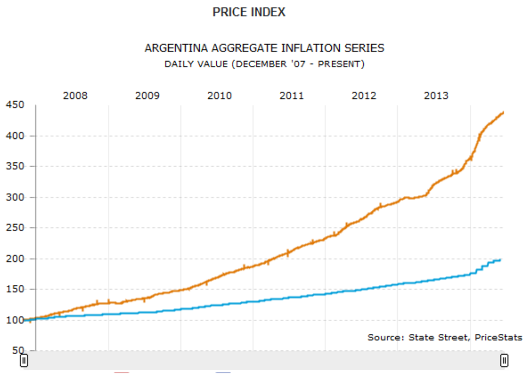
Pat Gelsinger, COO of EMC

"I keep saying that the sexy job in the next 10 years will be statisticians. And I'm not kidding"

Hal Varian, Chief Economist, Google

Big Data Example: Billion Prices Project

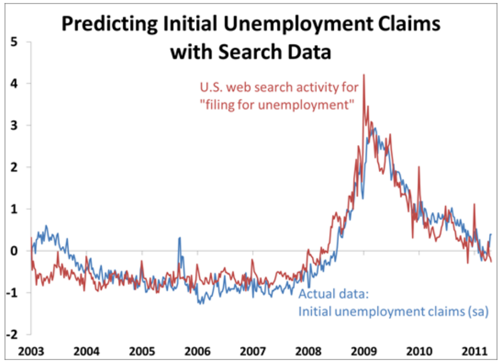
- MIT project which aggregates price information from dai price fluctuations of million items sold by ~300 online retailers more than 70 countr
- Contrast with CPI th focuses on a basket items that are monitored periodically



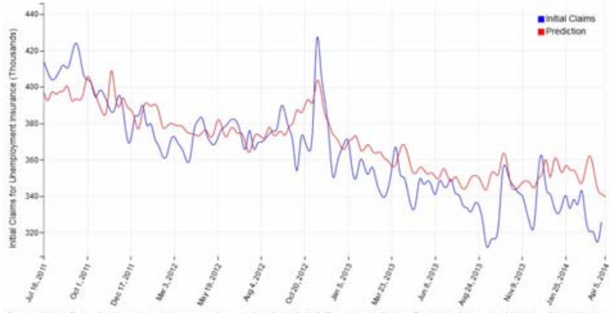
Source: State Street, PriceStats

Big Data Example: Google Correlates

- Provide Google with your weekly time series and it will tell you which search terms are most closely correlated with your data
- Question: What predicts Initial Unemployment Claims? Answer: "filing for unemployment"
- Provides a leading indicator based upon search


Predicting Initial Unemployment Claims with Search Data

Big Data Example: Tweeting about Unemployment

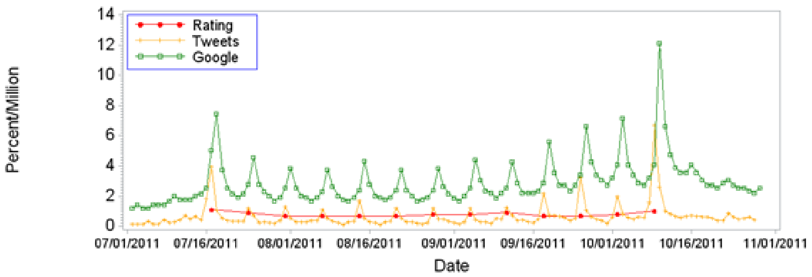


Sources: Initial Claims for Unemployment Insurance (seasonally adjusted), U.S. Department of Labor; Prediction, University of Michigan Social Media Job Loss Index

- Tweets like "I just lost my job. Who's buying my drinks tonight?" can be used to predict unemployment.
- Predicts 15-20% of the variance of the prediction error of the consensus forecast for initial claims.

Source: Antenucci et al, NBER Working Paper 20010

Big Data Example: Predicting TV Ratings



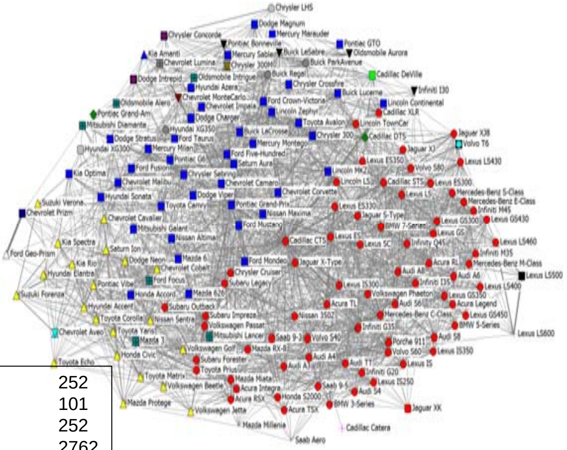
- Data for “Breaking Bad” in 2011 shows that using Tweets can accurately predict TV ratings
- Online search is a very weak predictor of TV show demand, but using sentiment of Tweets can explain up to 90% of variation

Source: Liu, Singh and Srinivasan (2014)

Big Data Example: Understanding Market Structure

Message #1199 **Civic vs. Corolla** by mcmanus *Jul 21, 2007 (4:05 pm)*
Yes DrFill, the Honda car model is sporty, reliable, and economical vs the **Corolla** that is just reliable and economical. Ironically its Toyota that is supplying 1.8L turbo ... Neon to his 16 year old brother. I drove it about 130 miles today. Boy does that put all this **Civic** vs. **Corolla** back in perspective! The Neon is very crudely designed and built, with no low ...

Audi A6	Honda Civic	252
Audi A6	Toyota Corolla	101
Honda Civic	Audi 6	252
Honda Civic	Toyota Corolla	2762
Toyota Corolla	Audi A6	101
Toyota Corolla	Honda Civic	2762



Source: Netzer (2011)
“Mine Your Own Business”

Research Examples

- Predicting...
  - Book Sales (Gruhl et al 2005)
  - Movie Box-Office (Mishne and Lance 2006)
  - Opinion Polls (O’Connor et al 2010)
  - Elections (Tumasjan et al 2010)
  - Stock Market Performance (Bollen, Mao and Zeng 2011)
- Analyzing blogs and online reviews
  - TV shows (Godes and Mayzlin 2004)
  - Movies & Phone Subscriptions (Onishi and Manchanda 2012)
  - Stock Prices (Tirunillai and Tellis 2012)

Industry Examples

- **GE, ‘Industrial Internet’**
  - Help airlines predict mechanical malfunctions and reduce flight cancellations
- **Kaggle, DIY data scientists cash-prize challenges**
  - Farms out complex ‘data challenges’ that come with cash prizes
- **Ayasdi, Visualizing Big Data**
  - Generate 3D maps that unearth new trends in genetic traits of cancer survivors, track E-coli outbreaks
- **IBM, Smarter Cities**
  - Improve traffic flow by predicting points of congestion
- **Weather Company**
  - How does unique climate data in each locate effect purchases (e.g., target antifrizz shampoo in humid climates)
- **Gnip, Monitoring Social Media Streams**
  - Lets customers monitor and parse through social media streams by attributes like keywords, trends and locations

Advantages of Big Data

Can detect patterns by leveraging these qualities:

- Massive
- Immediate and timely
- Predictive
- Free (or inexpensive)

Summary

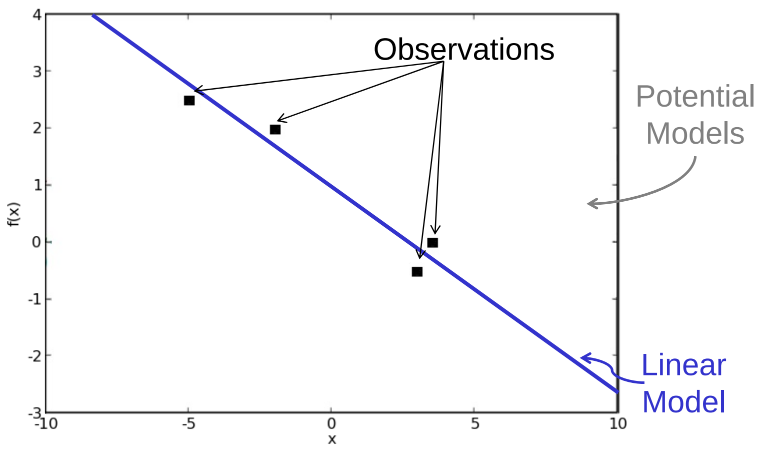
- Big Data has a huge potential to shape our lives through changes in business, government, and science, or society in general
- It is a by-product of our electronic lives and generally the reason it is collected has nothing to do with analysis or learning
- There are many questions about how data should be collected and analyzed, and how to protect privacy

Assessing Model Uncertainty in Financial Risk Management

Making Decisions with Big Data

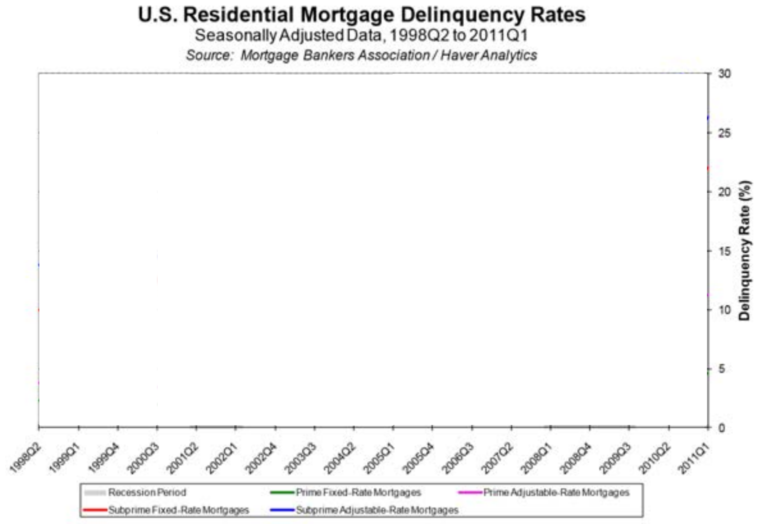
- Data mining models look for patterns in data that can be used to make better predictions and decisions
- The problem is “that all models are wrong; the practical question is how wrong do they have to be to not be useful”? (Box and Draper 1987)
- The Federal Reserve Bank has mandated banks to evaluate their risk exposure to quantitative models. But industry surveys (PWC 2013) show no commonly accepted standard for evaluating risk.

Illustration: Uncertain Form



Models have a huge impact on interpolations and extrapolations

Illustration: Insufficient Data


U.S. Residential Mortgage Delinquency Rates

Seasonally Adjusted Data, 1998Q2 to 2011Q1

Source: Mortgage Bankers Association / Haver Analytics

Problem

- We want to make inferences about an unknown property (θ). Typically we assume a model is known, but its parameters are not.
- Ignoring model uncertainty yields biased inferences:

  $$E[\theta \mid M] \neq E[\theta]$$

  and results in overconfidence:

  $$\mathrm{Var}[\theta \mid M] \leq \mathrm{Var}[\theta] = \mathrm{Var}_M\left[\mathrm{E}[\theta \mid M]\right] + \mathrm{E}_M\left[\mathrm{Var}[\theta \mid M]\right]$$

How important is this bias and overconfidence?

Overcoming Overconfidence

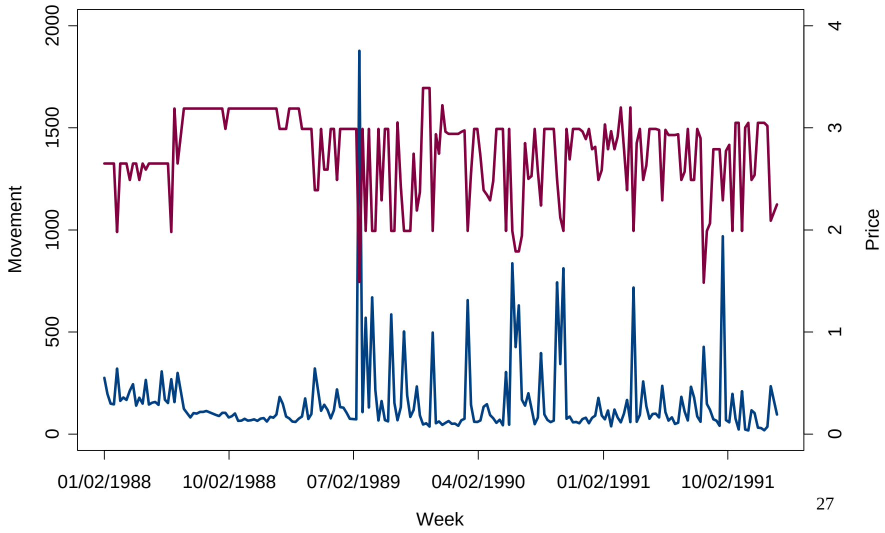
- If we identify and estimate the model using the same data as we are for making predictions then we are prone to be *overconfident* in our models.
- To compensate for this overconfidence we need to consider…
  - How much are we learning about functional form observed data
  - That many errors are correlated (unemployment, hurricanes, …) but most models assume independence, especially if the model is trained in a good economic cycle and we want to forecast in a bad economic cycle
  - Relationships may not be stable over time, economic relationships may be impacted by business cycles (which tend to be slow, infrequent)

Retail Price Optimization using Business Rules

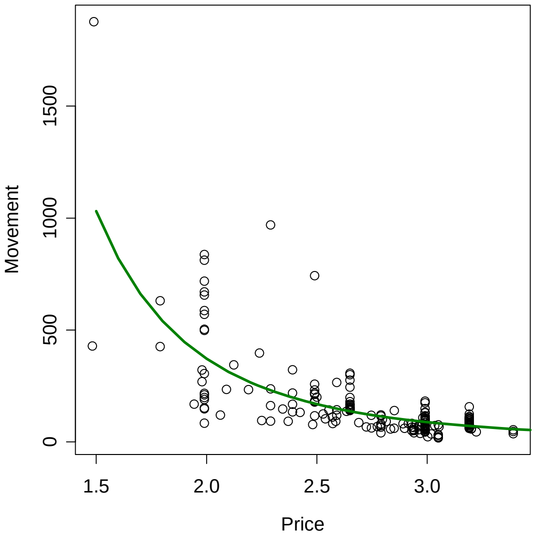
Price Optimization in Practice

- Huge growth of price optimization in practice for both retailers and manufacturers
- Gartner Marketscope states that “through 2010, price optimization technology will have a more direct impact on increasing revenue or margins than any other CRM technology”
- The Yankee Group estimated that more than one billion dollars would be spent on these systems in 2007
- Anecdotal reports suggest increases in gross margins in the range of 2-8%, retailers typically have gross margins of about 25% and have annual revenues of $2.5 trillion. Suggests benefit for retailers alone would be between $12.5b and $50b annually.

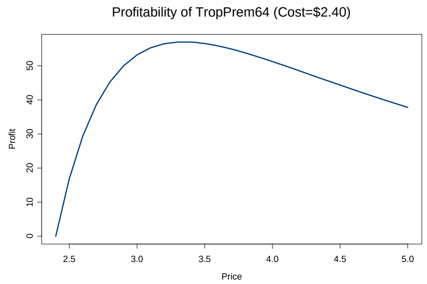
Weekly Movement and Price of TropPrem64



Movement vs Price of TropPrem64



Optimal Product Pricing



- Profits:

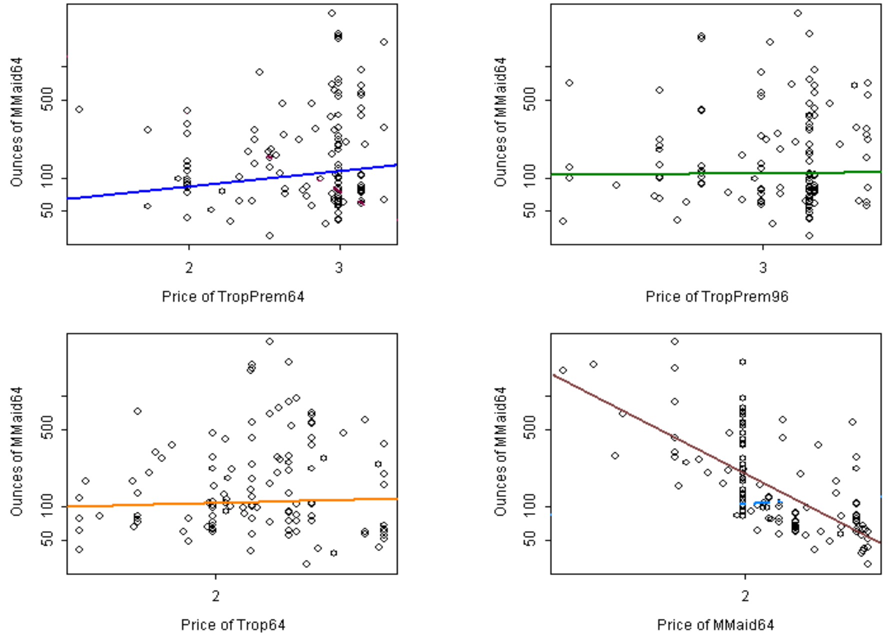
$$\Pi = (p-c)q$$

- Optimal pricing rule:

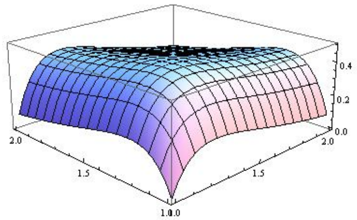
$$p^* = \frac{\beta}{\beta+1}c$$

- Where price elasticity measures demand responsiveness to price changes:

$$\beta = \frac{\partial q}{\partial p}\frac{p}{q} \approx \frac{\%\Delta q}{\%\Delta p}$$



Optimal Product Line Pricing



- Total Profits:

$$\Pi = \sum_{i=1}^{M} (p_i - c_i) q_i$$

- Optimal pricing rule:

$$p_i^* = \frac{\beta_{ii}}{\beta_{ii} + 1 + \sum_{j \neq i} \mu_j \beta_{ji} s_j / s_i} c_i$$

- Where cross price elasticities measure competitive effects:

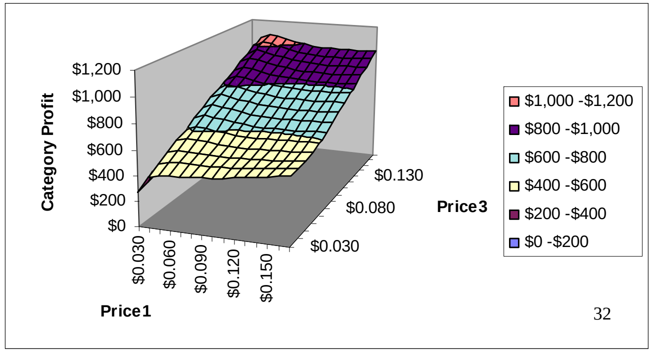
$$\beta_{ij} = \frac{\partial q_i}{\partial p_i}\frac{p_i}{q_i} \approx \frac{\%\Delta q_i}{\%\Delta p_j},$$

$$\mu_i = \frac{p_i - c_i}{p_i}, \; s_i = \frac{p_i q_i}{\sum_j p_j q_j}$$

Joint Price Optimization

Problem: Category Profits (the sum of products from each of the products) rises if all prices go up

Intuition: Model predicts substitution is constant, but will substitution really be the same for a $10 carton juice versus one at $1,000?



Business Rules

Current pricing solutions frequently implement constraints that reflect "business rules", which codify manager knowledge:

1. Allowed number and frequency of markdowns (e.g., at least a week between two consecutive markdowns)
2. Min-max discount levels or maximum lifetime discount
3. Minimum number of weeks before an initial markdown can occur
4. Types of markdowns allowed (e.g., 10%, 25%, ...) or the permissible set of prices
5. The "family" of items that must be marked down together

Source: Elmaghraby and Keskinocak (2003; Management Science)

Why use business rules?

- Answer: to "improve" the pricing solution and find a better one than would be afforded without these constraints
- Examples:
  - Strategic decision (Elmaghraby and Keskinocak 2003)
  - Ensure desired positioning of the product (Hawtin 2002)

DemandTec's Rule Relaxation Approach

1. *Group price advance or decline rules.* The user sets a maximum weighted group price advance or decline to 10%.
2. *Size pricing rules.* The user goes with the default that larger items cost less per equivalent unit than smaller identical items.
3. *Brand pricing rules.* For soft drinks, the user designates the price of brand A is never less than the price of brand B. For juices the user designates that brand C is always greater than Brand D.
4. *Unit pricing rules.* The user goes with the default that the overall price of larger items is greater than the overall price of smaller identical items.
5. *Competition rules.* The user designates that all prices must be at least 10% less than the prices of the same items sold by competitor X and are within 2% of the prices of the same items sold by competitor Y.
6. *Line price rules.* The user designates that different flavors of the same item are priced the same.

Source: Neal et al. (2010), Patent #7617119

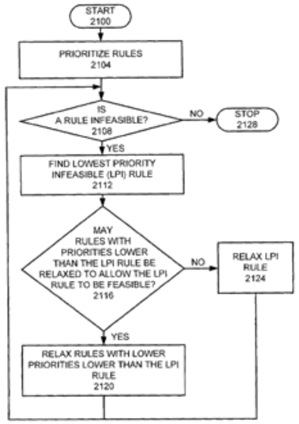
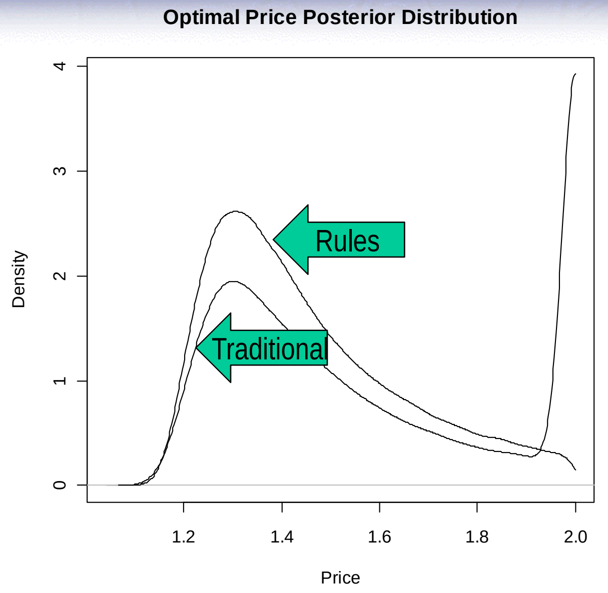


FIG. 21



Conclusion

- Current optimal pricing practitioners and researchers are introducing information in an ad hoc manner by relying upon "business rules" or constraints.
- An appropriate Bayesian data mining methods avoids ad hoc "corrections" to the predictions (or posterior) and says that the knowledge should be brought in a priori
- Leads to better decision support systems that reflect "expert" knowledge efficiently.

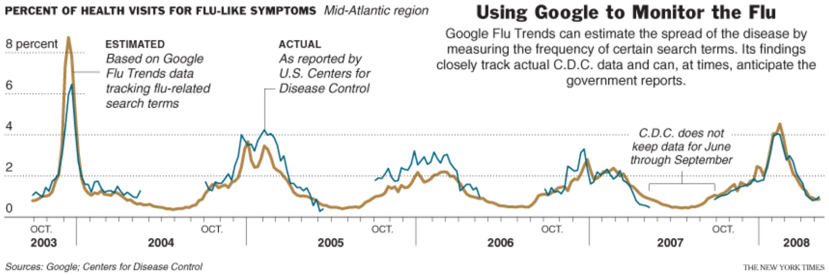
Predicting Flu Trends

Source: Lazer, Kennedy, King and Vespignani (2014), "The Parable of Google Flu: Traps in Big Data Analysis", *Science*

Google Flu Trends

In 2008 Google released an experiment called Flu Trends to predict the number of flu cases (as reported by the CDC) using searches from about 40 flu-related queries

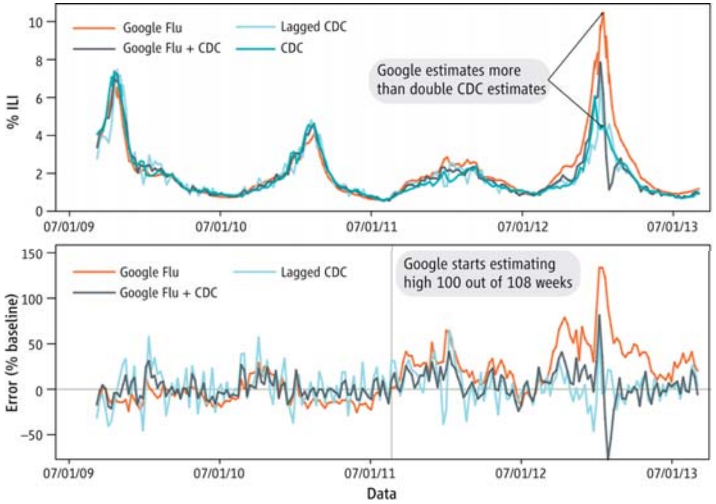
"The earlier the warning, the earlier prevention and control measures can be put in place, and this could prevent cases of influenza," said Dr. Lyn Finelli, lead for surveillance at the influenza division of the C.D.C. Source: NYT 12-11-2008



Using Google to Monitor the Flu

Google Flu Trends can estimate the spread of the disease by measuring the frequency of certain search terms. Its findings closely track actual C.D.C. data and can, at times, anticipate the government reports.

GFT overestimation.GFT overestimated the prevalence of flu in the 2012–2013 season and overshot the actual level in 2011–2012 by more than 50%.

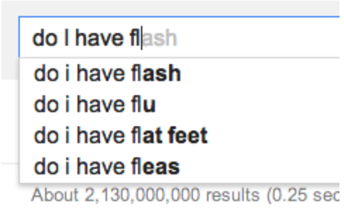


D Lazer et al. Science 2014;343:1203-1205

Published by AAAS

Big Data Headache

- Google's own autosuggest feature may have driven more people to make flu-related searches during the 2013 flu season
- Which misled its forecasting system and overstate the number of cases



do I have flash

do i have flash

do i have flu

do i have flat feet

do i have fleas

About 2,130,000,000 results (0.25 sec

Conclusion

- "GFT was like the bathroom scale where the spring slowly loosens up and no one ever recalibrated", David Lazer (Northeastern University)
- Google constantly makes tweaks to its general search algorithm averaging more than one a day, and the introduction of its "autosuggest" feature may led to more searches on influenza
- Lesson: the underlying patterns within social media and online behavior change, need to recalibrate their accuracy.

Conclusions

Limitations

- The promise of Big Data is that we can solve problems faster, cheaper and better.
- The problem is that Big Data is still just data, and we need to know its biases
  - Historical data may not be representative of future data
  - Participants in social media data may not be representative of society
  - The collection and use of "Big Data" changes through time
- Knowledge (managerial or theoretical) is still useful
  - The data we observe is influenced by our past decisions, which is a function of our "models", need to consider this feedback relationship

Disadvantages of Big Data

- It may not be representative
  - Who writes reviews? Really excited customers and really disappointed ones
- Data quality may be poor
  - Consumers generate Big Data for themselves not for data miners
- Privacy and confidentiality issues
  - How can we protect consumers?
- Difficult to assess accuracy and uncertainty
- The past may not be representative of the future

Future of Big Data

- Big Data is a rich source of information but to find the best solutions to business problems we need
  - To integrate economic and managerial knowledge
  - Be aware of biases in the data
  - Understand the differences between correlation and causation