Curses and Blessings of an (almost) Data-Complete Science: Big Data and the Social Sciences
Abstract:

Big Data has turned the social sciences from a traditionally data-poor science into arguably the most data-complete science to date – and this basically “overnight”. With over 99% of all of human kinds’ technologically mediated information in digital format, and a mobile penetration of 98% worldwide, the digitalization of human interaction produces an impressively detailed digital footprint of everything that’s relevant for the social sciences. Each and every digital communication inevitably leaves a trace that can be analyzed to better understand and influence social conduct. This renders many traditional survey and data collection and production processes obsolete. While creating unprecedented opportunities for private actors and lots of low hanging fruits for academic research, it also creates challenges that call for a profound paradigm shift in our relation to data.
"The biggest stumbling block… is obtaining the data to parameterize and validate… using automated cameras and image recognition… motion-activated cameras… continuous plankton recorders towed beneath ships…"
Storage

in optimally compressed MB

www.martinhilbert.net/WorldInfoCapacity.html
Digital Footprint (real interactions recorded anyways)

n = N (no sampling, but potential bias)

Data-fusion (unstructured and incomplete)

Real-time (dynamic)

Spooky accuracy through ML (no need for theory)
TWITTER: 2nd language

Blue - Spanish,
Light Green - Korean,
Fuchsia - Russian,
Red - Portuguese,
Yellow - Japanese,
Pink - Dutch,
Grey - Danish,
Coral - Indonesian.


N = n

Using data records like call duration and call frequency, one can predict socio-economic, demographic, and other behavioral trades with 80-85% accuracy.

“...human mobility traces are highly unique. ...four spatio-temporal points are enough to uniquely identify 95% of the individuals.”

(a) Rwanda 2005/09: mobile phone penetration of 2-20%

(b) LatAm economy 2009/10: mobile phone penetration of 60-80%

Source: (a) Blumenstock and Eagle (2012); (b) Frias-Martinez and Virseda (2013).
Real-time

U.S. Bureau of Labor Statistics
- > 100s staff visiting > 90 cities
- 80,000 prices
- Cost: US$ 250 million/year

PriceStats
- 17 staff
- 300 online retailers; > 70 countries
- daily 5,000,000 prices

Sources:
http://www.pricestats.com/about-us/meet-the-team;
www.economist.com/node/21548242;
http://www.inflacionverdadera.com
Choose a representation that can use unsupervised learning on unlabeled data, which is so much more plentiful than labeled data.

“Big” doesn’t need to know Why

Thomson Reuters MarketPsych Indices (TRMI)
18,864 separate indices, 119 countries, updated each minute (!)
Without any information about a Facebook user beyond a list of his friends, one can accurately predict his sexual orientation.
"Facebook Likes, can be used to automatically and accurately predict...: sexual orientation, ethnicity, religious and political views, personality traits, intelligence, happiness, use of addictive substances, parental separation, age, and gender..."

"(i) computer predictions based on ... Facebook Likes are more accurate (r = 0.56) than those made by the participants' Facebook friends (r = 0.49); (iii) computer personality judgments have higher external validity when predicting life outcomes such as substance use, political attitudes, and physical health; for some outcomes, they even outperform the self-rated personality scores..."

Meaningful business impact

Mattersight, a leader in enterprise analytics focused on customer-employee interactions, offers analysis showing that a behavioral mismatch between a customer and employee, as opposed to a favorable behavioral connection, has a significant impact on business outcomes across industries, even in highly specialized contact center functions:

<table>
<thead>
<tr>
<th>Business Outcome</th>
<th>Impact of Behavioral Connection</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sales Conversion Rate</td>
<td>85% to 230%</td>
</tr>
<tr>
<td>Customer Retention/Attrition Rate</td>
<td>25% to 50%</td>
</tr>
<tr>
<td>Customer Service Cost/Efficiency</td>
<td>35% to 45%</td>
</tr>
<tr>
<td>Student Enrollment Rate (Private Sector Education)</td>
<td>200% to 700%</td>
</tr>
<tr>
<td>Debt Cure Rate (Collections Organization/Function)</td>
<td>Over 1400 Basis Points</td>
</tr>
</tbody>
</table>

Matching Personality Types:
- Call average from 10 min to 5 min
- Customer Satisfaction from 47% to 92%

EMOTIONS-DRIVEN (30% of the population)
THOUGHTS-DRIVEN (25%)
REACTIONS-DRIVEN (20%)
OPINIONS-DRIVEN (10%)
REFLECTIONS-DRIVEN (10%)
ACTIONS-DRIVEN (5%)

Obama 2012 campaign

Data

- **US$1 billion investment**; core group of **40 engineers**
  (from Twitter, Google, Facebook, Craigslist, stem cell, professional poker players...)
- Project Narwhal: **16 million unique voter profiles**:
  - email sign-ups, zip codes, profession, voter registrations, volunteering & donation record, Tweets, Facebook postings and network ties, TV Watching behavior through 20 million set-top boxes, etc.
- **62,000 computer simulations** of likely voter behavior

Outcome

- Identified the 20% of Obama’s 2008 vote that shifted into the undecided column, ranking them on a 0-10 persuasion score
- Obama paid 35% less per broadcast commercial than Romney
  (40,000 more spots on the air, spending $90 million less!)
- Present tailor made campaign promises (agreeable adds; etc)
- Guide volunteers in phone and door-to-door campaigns
- Email donation requests, raising $181 million/month
- Predict States voting outcome at an accuracy of 0.5 percent
- **Change voting behavior of 78 % of targeted undecided voters through Facebook**

Sources:
Consumers’ financial vulnerability:
- “Social Influencer”
- “Rural and Barely Making It”
- “Ethnic Second-City Strugglers”
- “Retiring on Empty: Singles”
- “Tough Start: Young Single Parents”
- “Credit Crunched: City Families”
- “Transitory lifestyles: military personnel”
- “Elderly Opportunity Seekers: elderly looking for ways to make money”
- “Oldies but Goodies: gullible, want to believe their luck can change”


Source: http://www.youtube.com/watch?v=wqjKTW3wjZ8
The long-begun end of “free will”?

Amazon to ship things before you've even thought of buying them?

Amazon files a patent for “anticipatory package shipping.” The idea is that the company will know from your buying patterns what you're likely to want next.

Homicide Parole candidates

- dataset > 60,000 crimes
- with some 300 predictors (nature of crime, age, repetition)
  - 60 – 70 % correct who commit homicide

Pre-punishment vs. Free Will?

- already insurance premiums per age or gender (punishment)

Predictive Policing LADP & SantaCruz

- Data on crimes, weather, buses, parks...
- Models from Earthquake aftershock
- Predictions to 500^2 feet / 50^2 m
  - Crimes down 13 %; burglaries 11 %; car theft 8 % (while other districts went up during same period)


"Your recent Amazon purchases, Tweet score and location history makes you 23.5% welcome here."
Proxies are just proxies

Fruit prices to detect violence in Jalalabad Afghanistan

...a “big breakthrough” (DARPA’s Director Regina Dugan) that impressed a group of four-star generals...

JSOC drone operator: “It’s of course assumed that the phone belongs to a human being who is nefarious and considered an ‘unlawful enemy combatant.’ This is where it gets very shady…”

Sources:
Computational Social Science

I. **Big Data Blessings**
   - Produced anyways
   - \( n = N \) (Volume)
   - Data-fusion (Variety)
   - In real-time (Velocity)
   - Spooky accuracy through ML

II. **Big Data Curses:**
   - Social Transparency & Polarization
   - Data as commodity
   - Confusing statistical variables (proxies) with reality