Big Data is here to stay, now what?

Hasan Hboubati
hhbo64@gmail.com

Hasan Hboubati

15 years IT/Software/Leadership

SDLC

Passion for Big Data

Twitter: @hbo64

BigDataIsHere.blogspot.com

AGENDA

- Big Data: Introduction
- Datafication, Data Exhaust
- Value Proposition
- Big Data and Big Brother
- People's Data
- Getting to Know Ourselves
- SDLC Implications

Big Data: An Introduction

Where did it come from?

Data has always been there

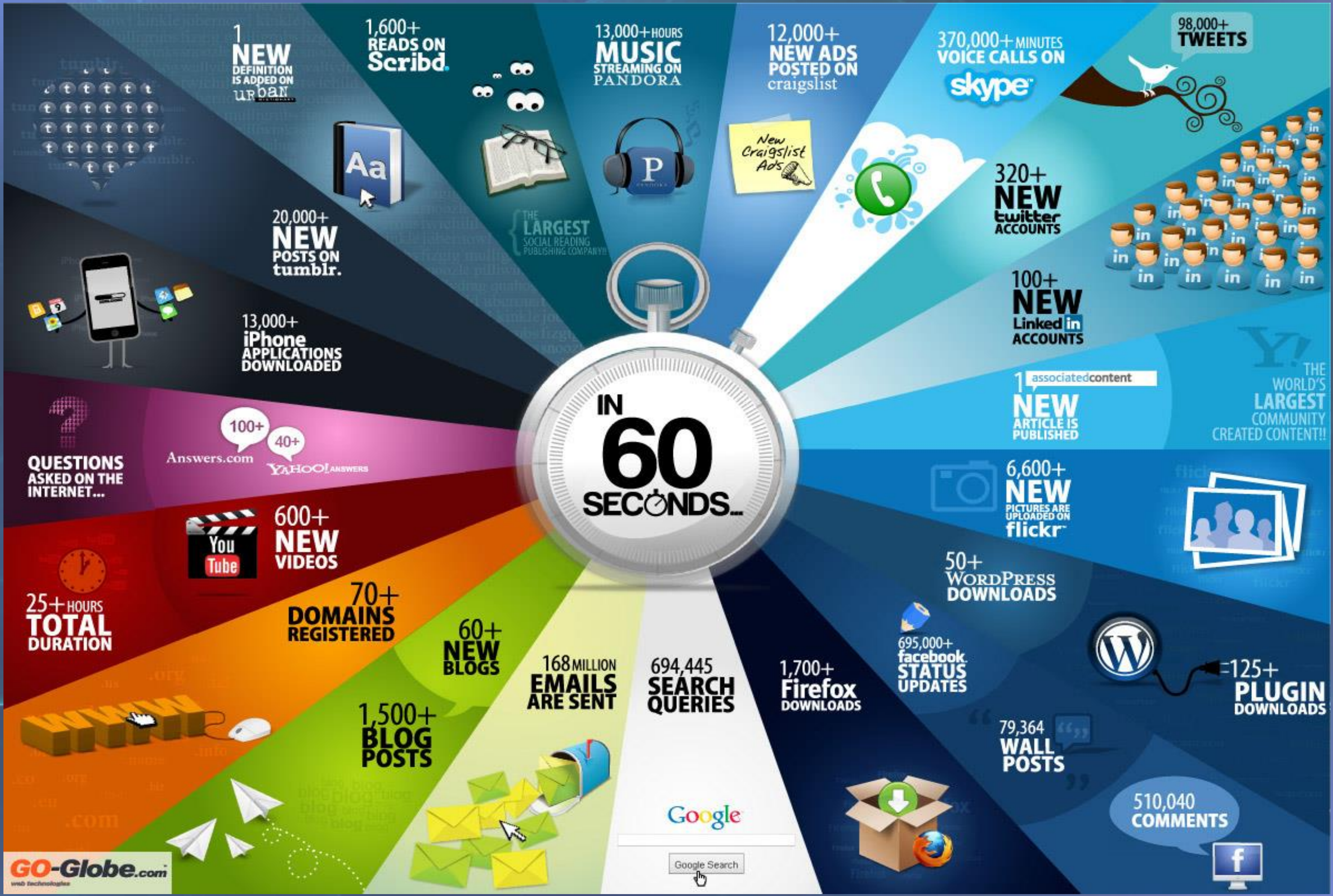
Data collection is easier than ever

Data storage is cheaper than ever

Data analysis is easier with new tools

More data is added every minute

Added in 60 seconds



IN 60 SECONDS...
1 NEW DEFINITION IS ADDED ON urban DICTIONARY
1,600+ READS ON Scribd.
THE LARGEST SOCIAL READING PUBLISHING COMPANY!!
13,000+ HOURS MUSIC STREAMING ON PANDORA
12,000+ NEW ADS POSTED ON craigslist
New Craigslist Ads
370,000+ MINUTES VOICE CALLS ON skype
98,000+ TWEETS
320+ NEW twitter ACCOUNTS
100+ NEW LinkedIn ACCOUNTS
1 associatedcontent NEW ARTICLE IS PUBLISHED
THE WORLD'S LARGEST COMMUNITY CREATED CONTENT!!
6,600+ NEW PICTURES ARE UPLOADED ON flickr
50+ WORDPRESS DOWNLOADS
125+ PLUGIN DOWNLOADS
695,000+ facebook STATUS UPDATES
79,364 WALL POSTS
510,040 COMMENTS
1,700+ Firefox DOWNLOADS
694,445 SEARCH QUERIES
Google
Google Search
168 MILLION EMAILS ARE SENT
1,500+ BLOG POSTS
60+ NEW BLOGS
70+ DOMAINS REGISTERED
25+ HOURS TOTAL DURATION
600+ NEW VIDEOS
You Tube
QUESTIONS ASKED ON THE INTERNET...
100+ Answers.com
40+ YAHOO! ANSWERS
13,000+ iPhone APPLICATIONS DOWNLOADED
20,000+ NEW POSTS ON tumblr.
GO-Globe.com web technologies

Big Data: An Introduction

- All human knowledge until 2002
- Football field filled with DVDs

Big Data: An Introduction

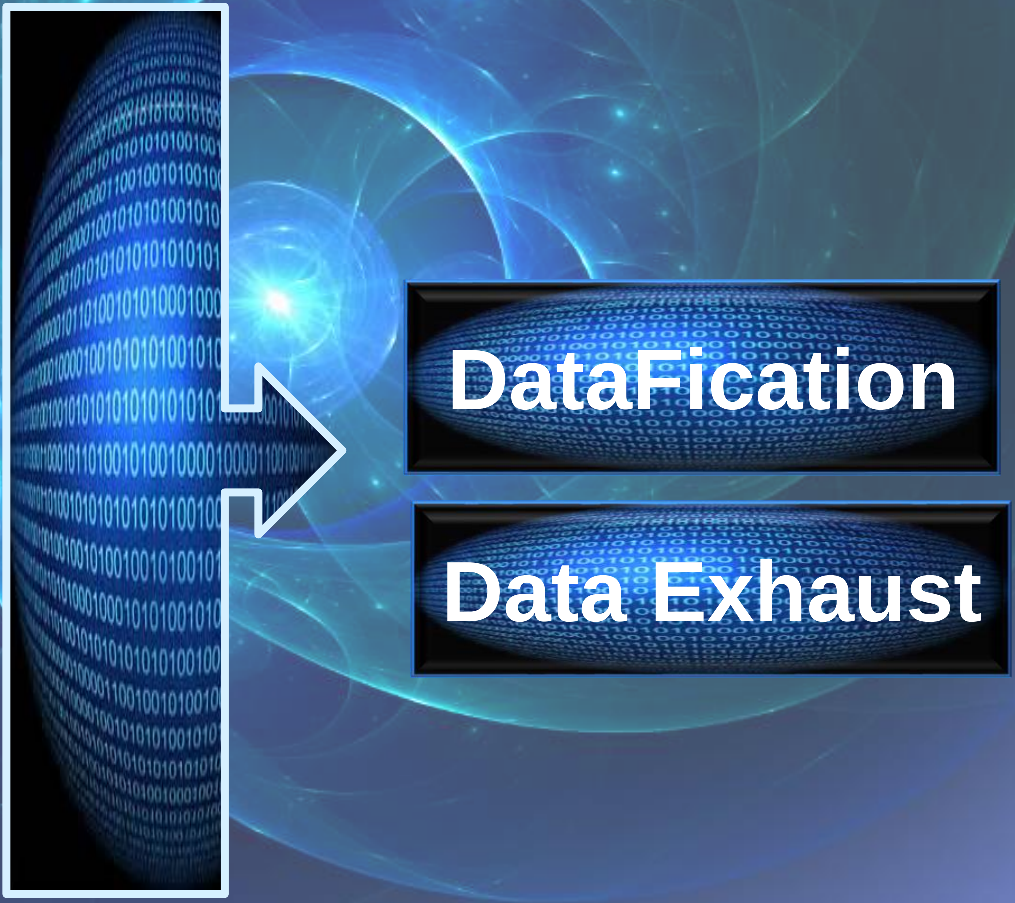
- Every 10 minutes

We create the same amount of data that has been created since recorded history until 2002

One Day’s worth of Data
144 Metlife Stadiums

Big Data: An Introduction

- Where is it coming from
  - Social media
  - Governments
  - Science
  - Shopping
  - UN
  - Universities
  - Businesses
  - Mobile
  - Location
  - Health
  - Manufacturing
  - Transportation
  - …



DataFication

Data Exhaust

AGENDA

- Big Data: Introduction
- **Datafication, Data Exhaust**
- Value Proposition
- Big Data and Big Brother
- People's Data
- Getting to Know Ourselves
- SDLC Implications

Datafication

- Turn a process or activity that was previously invisible into data
- That data can then be tracked, monitored, and optimized, leading to new opportunities — and new challenges

Datafication

- Facebook has “datafied” our friends network
- Google has “datafied” our search
- LinkedIn has “datafied” our professional connections
- Twitter has “datafied” news and real time information
- Waze has “datafied” our driving
- Amazon has “datafied” how we read books, what we read, and how fast we read
- Car manufacturers are “datafying” cars, tires, maintenance indicators, sitting positions
- Companies are “datafying” HR data, looking for trends, (who is leaving, why are they leaving, etc.)

Data Exhaust

- Unstructured **information** or **data** that is a by-product of the **online activities** of Internet users
- Collecting and analyzing data exhaust can provide **valuable insight** into the purchasing **habits** of **consumers**
- Also called digital exhaust

Data Exhaust

- We leave behind a trail (or a digital smell, if you will)
- Marketers are very interested in tracking this information so that they can display appropriate ads based on so-called "behavioral targeting"
- Huge competitive advantage for companies
- Big barrier to entry against rivals
- Amazon, Google, Facebook, etc.
  - Performance is a function of collected data exhaust
  - Incorporate back into the business

AGENDA

- Big Data: Introduction
- Datafication, Data Exhaust
- **Value Proposition**
- Big Data and Big Brother
- People's Data
- Getting to Know Ourselves
- SDLC Implications

Value proposition

- Forrester Research estimates that organizations effectively utilize less than 5% of their data
- 95% of what your organization “knows” isn’t being used to its full potential

Value proposition

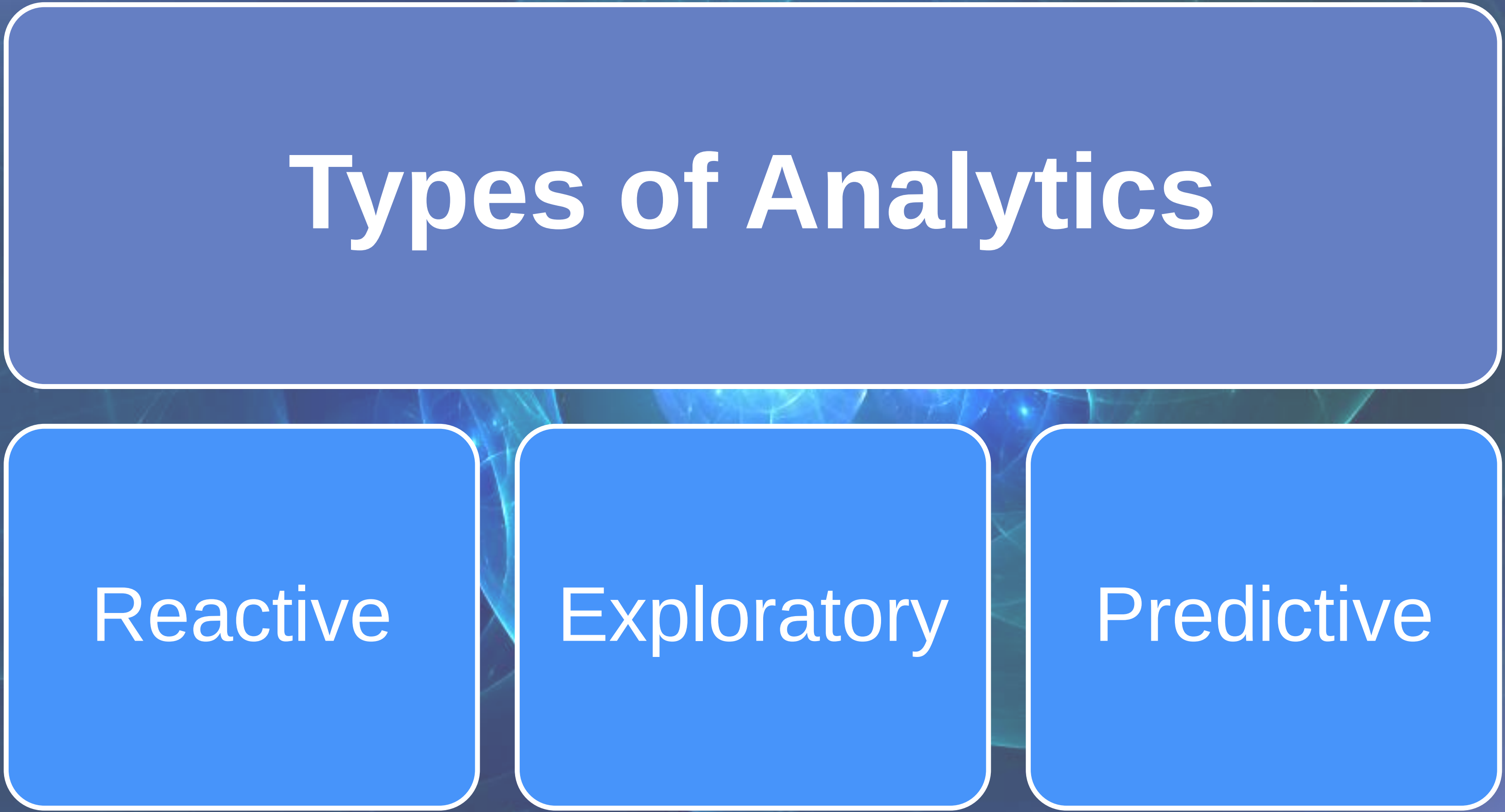
- **20 years ago**, Lew Platt, former CEO of HP, said, “If only HP knew what HP knows, we’d be three times more productive.”
- No lack of data
  - Your company is collecting data today (customer, vendor, process, performance, HR, etc.)
- Use data to implement
  - Business Intelligence
  - Predictive Analytics

Value Proposition

Imagine if

- You had the **power** to anticipate problems
- Your could do something to **fix** issues **before** they happen
- You didn't have to **explain** to shareholders why targets were missed
- You had the **power** to **preemptively** address problem areas
- You could **predict** the future
- You could use
  - social media
  - server logs
  - login stats
  - images search
  - shopping habits
  - equipment sensors
  - GPS information

Value proposition



Types of Analytics

Reactive

Exploratory

Predictive

Value proposition

- Google Analytics: examine the “compare to” click
- Google’s: “did you mean?”
- Prove you are human
- Value of data resides in its potential use (Latent value of data)
  - Use cell phone records to see protesters coming
  - Location-based advertising
  - Car seat censors to prevent theft

AGENDA

- Big Data: Introduction
- Datafication, Data Exhaust
- Value Proposition
- **Big Data and Big Brother**
- People's Data
- Getting to Know Ourselves
- SDLC Implications

Big Data and Big Brother

- Commercial enterprises are not the only ones collecting data and sniffing data exhaust
- Governments all over the world have amassed huge amounts of data about everything that we do
- Recent NSA revelations only scratch the surface of the amount/types of data that the US government is collecting around the world (even the German Chancellor, and US public officials are not immune)

Big Data and Big Brother

Using Predictive Analytics:

- Use Big Data predictions about people to punish them *before* they *may* have acted
- Parole boards base decisions to release inmates from prison
- Police use “Predictive Policing”
  - Streets, groups, individuals get extras scrutiny
- Richmond, VA police correlate crime data with pay day, sports events, gun shows…

Big Data and Big Brother

- When it comes to privacy, Big Data is a **game changer**
- **Dormant** value unknown at collection time
- Big Data is often used for purposes other that for which it was collected
- Three core strategies have lost effectiveness

  Notice of consent, Opt out, Anonymization
- Laws and regulations have not kept up
- Privacy through consent - Privacy through accountability
- Collector’s duty to use data responsibly
- Time limits on data storage
- Guaranteeing human agency

AGENDA

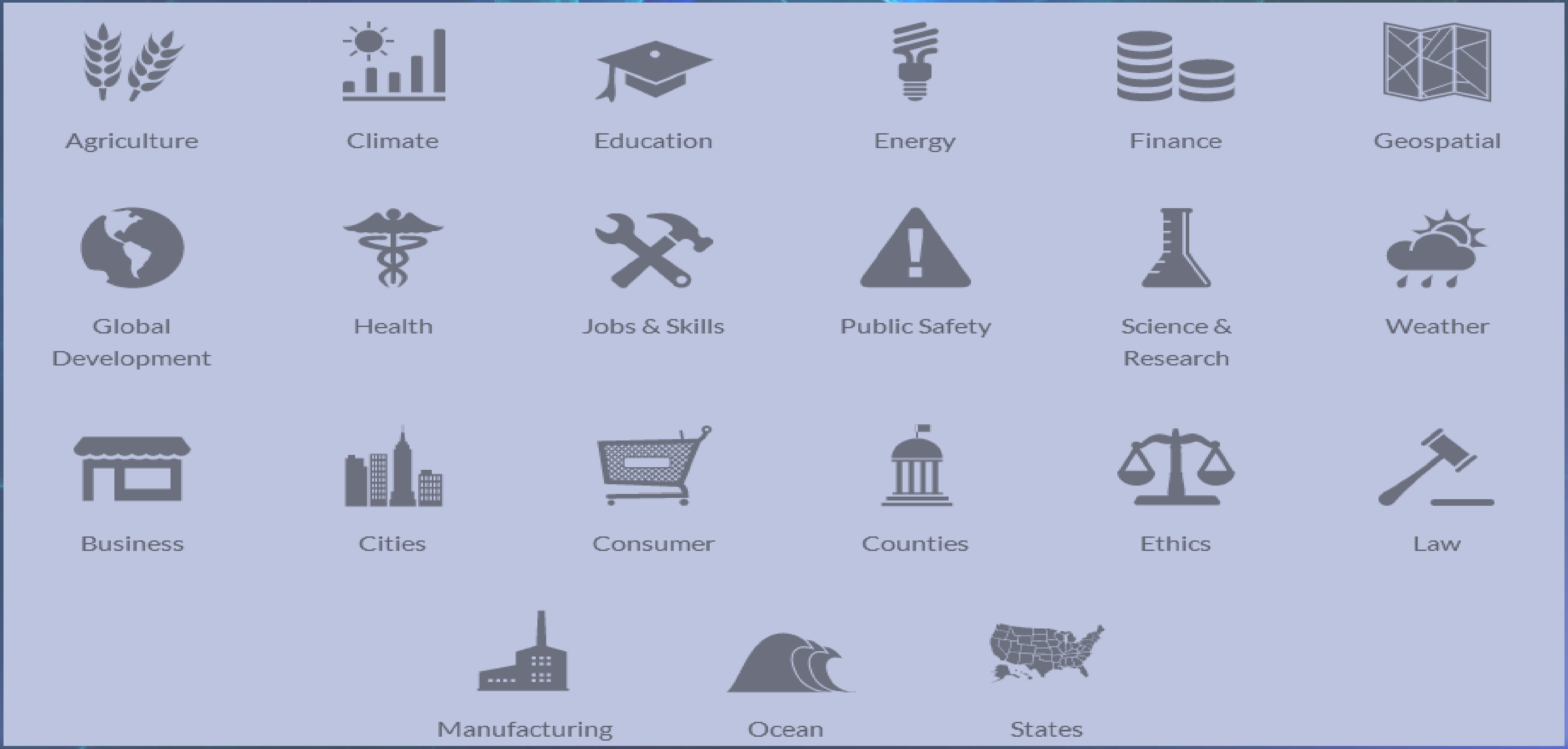
- Big Data: Introduction
- Datafication, Data Exhaust
- Value Proposition
- Big Data and Big Brother
- **People's Data**
- Getting to Know Ourselves
- SDLC Implications

Big Data: People's Data

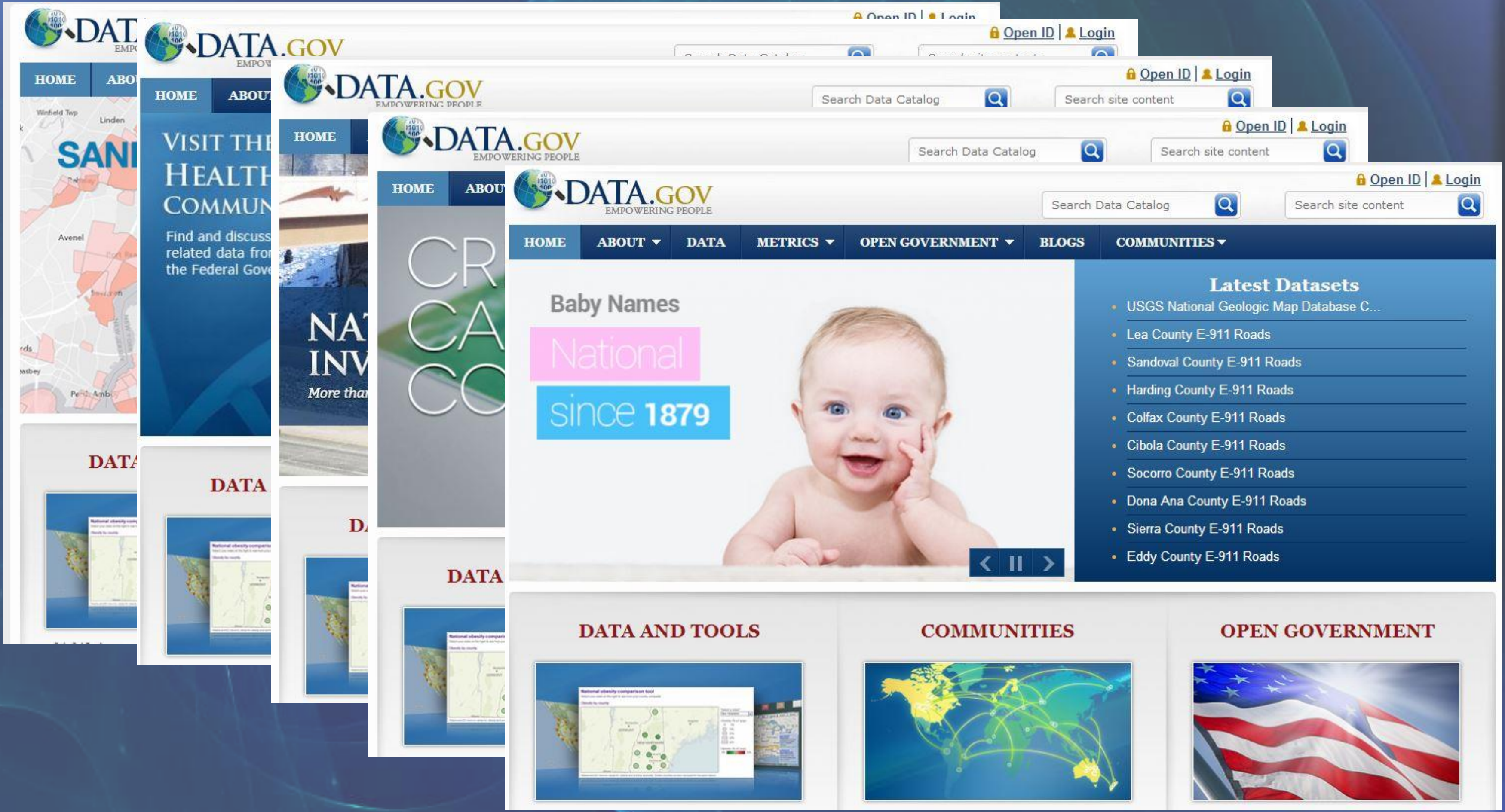
- Privacy concerns are real
- But not all doom and gloom
- People are demanding to get their data back (or at least access to it)
- Recent movement afoot that government data is really owned by the people
- Sites like Data.gov is one of many examples of open government movement to give access to that data (when possible)

Big Data: People's Data

- 91,641 datasets
- 349 citizen-developed apps
- 137 mobile apps
- 175 agencies
- 87 galleries
- 295 Government APIs



Big Data: People’s Data



Latest Datasets

- USGS National Geologic Map Database C...
- Lea County E-911 Roads
- Sandoval County E-911 Roads
- Harding County E-911 Roads
- Colfax County E-911 Roads
- Cibola County E-911 Roads
- Socorro County E-911 Roads
- Dona Ana County E-911 Roads
- Sierra County E-911 Roads
- Eddy County E-911 Roads

Big Data: People’s Data

Code for America	• improves the **relationships** between **citizens** and **government**
Sunlight Foundation	• uses the power of the Internet to catalyze **greater** government **openness** and transparency
Open Knowledge Foundation (UK)	• promotes **open data** and open content including government data, publicly funded research and public domain cultural content
Data Market (Iceland)	• example of a big data middleman company

AGENDA

- Big Data: Introduction
- Datafication, Data Exhaust
- Value Proposition
- Big Data and Big Brother
- People's Data
- **Getting to Know Ourselves**
- SDLC Implications

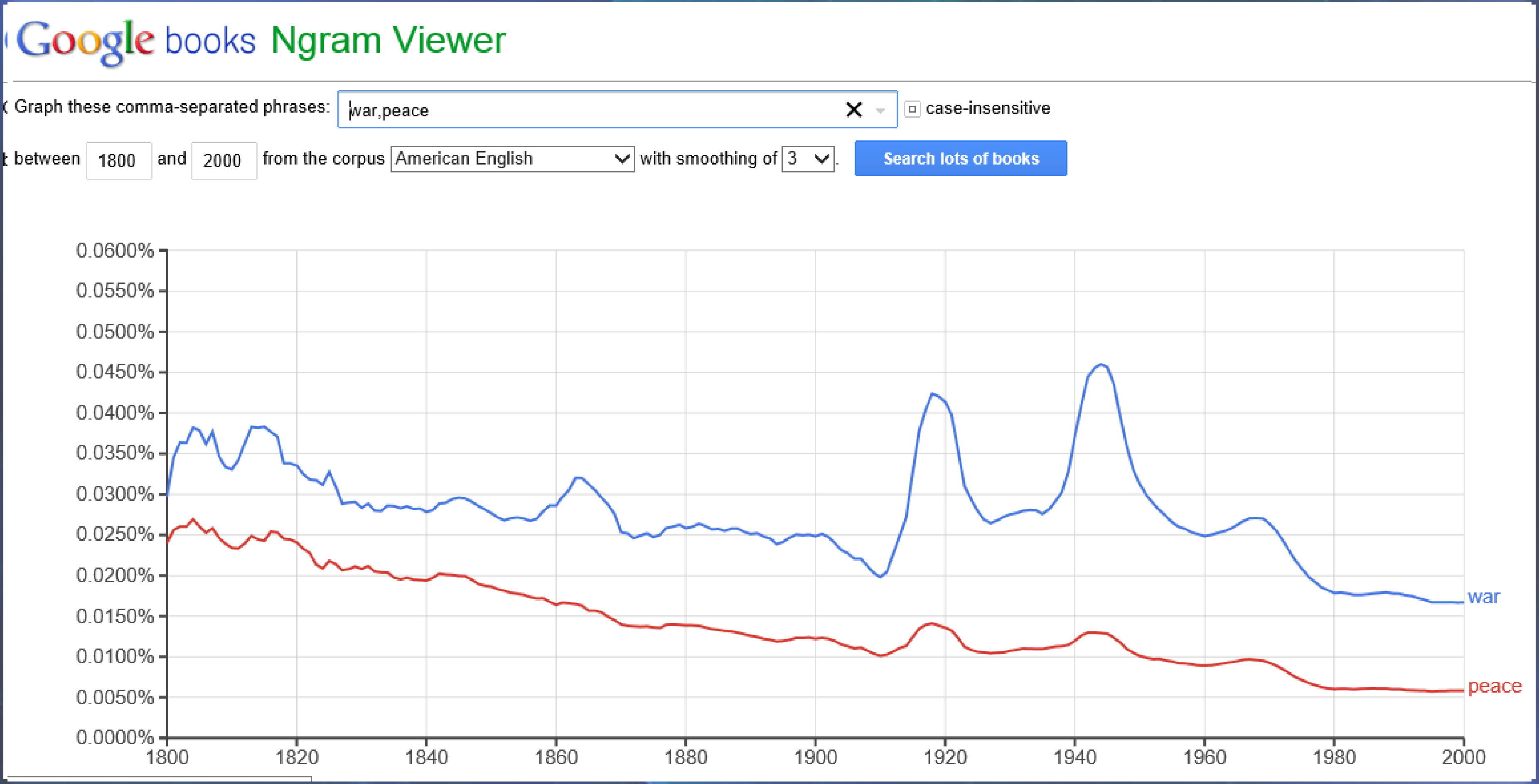
Getting to know ourselves

- Google nGram Project: Google digitized millions of books since the 1800s as part of its Google Books search engine.

  http://books.google.com/ngrams

- A great way to find out more about ourselves and make human connections

Getting to know ourselves



Graph these comma-separated phrases: war,peace □ case-insensitive

between 1800 and 2000 from the corpus American English with smoothing of 3.

Getting to know ourselves

- http://www.gapminder.org/
- Fact based world wide view
- Fighting **ignorance** with fact-based **worldviews** everyone can **understand**
- World **solutions** (example HIV in Africa) need to examine trends and offer **strategies** based on each **country** and segment within that country
- Big data will allow us to successfully do just that

AGENDA

- Big Data: Introduction
- Datafication, Data Exhaust
- Value Proposition
- Big Data and Big Brother
- People's Data
- Getting to Know Ourselves
- **SDLC Implications**

SDLC Implications
SW Development

- Software development:
  - Making software widgets (Manufacturing)
- Analytics projects:
  - Finding gold (Mining)
- Long live the prototype
- Business gaining IT expertise
- Leaving IT departments out of the loop
- IT – Business partnership

SDLC Implications - QA

- No requirements in the traditional sense
- How do we test big data analytics projects?
- What then is a pass/fail criterion?
- If the user clicks on a data point in a dashboard, what should they see?
- What do you expect them to see?
- Side-by-Side Testing

SDLC Implications - QA

- Using Analytics of your customer data (social media), you can focus your QA activities on functionality that is most used by the end user
- This is easy today as a one-off activity
- Continuous feedback loop - customer analytics to constantly inform your test activities
- Increase QA performance
- Reduce your testing timeline and cost
- Automate high risk test cases
  - Higher rate of use
  - Higher rate of failure

HR Implications

- Hiring and keeping good data scientists
- Promotions: Chief Analytics Officer
- Challenging environment
- People skills

THANK YOU

Q & A

References

- http://www.ted.com/playlists/56/making_sense_of_too_much_data.html
- http://blogs.forrester.com/brian_hopkins/11-09-30-big_data_will_help_shape_your_markets_next_big_winners
- http://info.sdlcpartners.com/blog/?Tag=Predictive%20Analytics
- Forrester:http://www.slideshare.net/hortonworks/forrester-mb-bigdatahortonworksv1
- TED Videos

Power of Human connections

- Khan Acadmy: https://www.khanacademy.org/
- Flipping Education: http://www.forbes.com/sites/pascalemmanuelgobry/2012/12/11/what-is-the-flipped-classroom-model-and-why-is-it-amazing-with-infographic/
- MIT: https://immersion.media.mit.edu/viz#
- Minority report: