The Intersection of Big Data, Data Science, and The Internet of Things

Bebo White
SLAC National Accelerator Laboratory/Stanford University
bebo@slac.stanford.edu
SLAC is a US national laboratory operated by Stanford University.

SLAC lives for data and high performance computational analysis!

SLAC collects data from a wide variety of different devices/sensors.
Data from the Linear Coherent Light Source (LCLS)

~1.3 million DVDs/month

Is this Big Data?

CITA’15, August 2015
About My Talk Title

• Have *Big Data, Data Science, and The Internet Of Things* somehow lost their relevant meaning or are they just *buzzwords* or *hype*?

• How can I possibly put them together in a single talk? - hype³

• I believe they are all a part of a *future convergence*

• “Singularity/convergence” was the CITA’11 theme and we are closer to it now!
Big Data

Internet of Things

Data Analytics

Pretty well understood; New technologies

Predicted; But not well understood

The exciting new platform!
• Big Data - not new, only matter of scale

• Data Analytics - not new, but grown in complexity to accommodate the “Big Data 4Vs”

• IOT - nothing like it before
• One of the big promises of “Big Data” and “The Internet of Things” is supposed to be *insight* and *knowledge*

• The key component to bridge the gap between data, insight, and knowledge is *analytics*

• The key concepts are *platform* and *innovation*
Why Platform?
4th Platform Critical Elements
• “Big Data,” “The Internet of Things,” and “Data Science” are “innovation platforms”

• Infrastructures upon which to build advanced applications

• “Big Data” is maybe too vague (IMHO)
  • Concept has spun off new tools and solutions
  • But usage is perhaps too focused with too many expectations

• “The Internet of Things” may dwarf “Big Data”
4th Platform Innovation Possibilities

- Dynamic contextual services
- Shared perception and learning
- Ecosystem domination
- Instantly composable apps
- Fully digital business models
- Time augmentation
- Digital communities as institutions
Interesting (to me) that IOT lies midway between Big Data and Data Science
It's Official: The Internet Of Things Takes Over Big Data As The Most Hyped Technology

What does this mean?
But, instead of hype

- Let’s think *evolution* instead of *revolution*
- What do each of these elements individually contribute to a greater whole?
The Data Deluge

- From the beginning of recorded time until 2003, mankind generated 5 exabytes of data.
- In 2011, the same amount of data was generated every two days.
- In 2013, the same amount of data was generated every 10 minutes.
- In 2015?
- Such numbers become almost meaningless…
<table>
<thead>
<tr>
<th>Memory unit</th>
<th>Size</th>
<th>Binary size</th>
</tr>
</thead>
<tbody>
<tr>
<td>kilobyte (kB/KB)</td>
<td>$10^3$</td>
<td>$2^{10}$</td>
</tr>
<tr>
<td>megabyte (MB)</td>
<td>$10^6$</td>
<td>$2^{20}$</td>
</tr>
<tr>
<td>gigabyte (GB)</td>
<td>$10^9$</td>
<td>$2^{30}$</td>
</tr>
<tr>
<td>terabyte (TB)</td>
<td>$10^{12}$</td>
<td>$2^{40}$</td>
</tr>
<tr>
<td>petabyte (PB)</td>
<td>$10^{15}$</td>
<td>$2^{50}$</td>
</tr>
<tr>
<td>exabyte (EB)</td>
<td>$10^{18}$</td>
<td>$2^{60}$</td>
</tr>
<tr>
<td>zettabyte (ZB)</td>
<td>$10^{21}$</td>
<td>$2^{70}$</td>
</tr>
<tr>
<td>yottabyte (YB)</td>
<td>$10^{24}$</td>
<td>$2^{80}$</td>
</tr>
</tbody>
</table>

dekabytes?   hectobytes?
Information from the Internet of Things:
We have gone beyond the decimal system

Today data scientist uses **Yottabyte** to describe how much government data the NSA or FBI have on people altogether.

In the near future, **Brontobyte** will be the measurement to describe the type of sensor data that will be generated from the IoT (Internet of Things).

**Yottabyte**
This is our digital universe today = 250 trillion of DVDs

1 EB of data is created on the internet each day = 250 million DVDs worth of information.
The proposed **Square Kilometer Array telescope** will generated an EB of data per day

**Exabyte**
500TB of new data per day are ingested in Facebook databases

**Petabyte**
The CERN Large Hadron Collider generates 1PB per second

**Zettabyte**
1.3 ZB of network traffic by 2016

**Terabyte**

**Gigabyte**

A device generating 1 MB/sec starting at the Big Bang
Where is This Data Coming From?

- EVERYWHERE - known and unknown; visible and invisible; with or with permission
- Any communication over a network involves transfer of data that is meaningful to someone
- Every e-mail, every tweet, every transaction, every social media interaction, etc., etc.
- Sensors - “The Internet of Things”
How is This Data being used (consumed)?

• The “poster children”/“large data generators” for datasets were:
  • Science
  • Finance
  • Government
  • etc.

• Now, we are the experiments creating the datasets
  • Facebook knows what food and music we like
  • Advertisers use cookies and intelligent algorithms to create personalization
  • Amazon even claims to know what we want to (or will) buy next
Big Data - a Possible Definition

• Refers to datasets whose size is beyond the ability of
  • Single storage devices
  • Typical database software tools to capture, store, manage, and analyze (McKinsey Global Institute)
  • This definition is not defined in terms of data size (which will increase)
  • It can vary by sector/usage
  • This is not a new issue
Beyond Capability

- 1956 technology
- 5Mb storage
- LCLS would require over 1 trillion units per month
Google processes 20 PB a day (2008) and crawls 20B web pages a day (2012).

eBay has >10 PB data, 75B DB calls per day (6/2012).

Facebook has >100 PB of user data + 500 TB/day (8/2012).

Amazon Web Services has S3: 449B objects, peak 290k request/second (7/2011) and 1T objects (6/2012).

JPMorgan Chase processes 150 PB on 50k+ servers running 15k apps (6/2011).


LHC processes ~15 PB a year.

CERN's LSST processes 6-10 PB a year (~2015).

SKA processes 0.3 – 1.5 EB per year (~2020).

640K ought to be enough for anybody.

How much data?
The Cloud/IOT is/will be a very “noisy” place

- An unbelievable of objects (theoretically more than $10^{38}$) will be able to talk to us and to each other (orders of magnitude more than now)

- We will be interested in hearing what some of them have to say

- How can we manage these conversations?

- Traditional interfaces break down
IOT is the Result of an Amazing Convergence

“We have the capability to fully connect and integrate a wide diversity of devices and objects into the online environment and interact with them”
“The Internet of Things (IOT) includes machine-to-machine (M2M) technology enabled by secure network connectivity and cloud infrastructure, to reliably transform data into useful information for people, businesses, and institutions”

“The Internet of Things is founded on familiar technologies - like sensors, networking and cloud computing - but its potential for transformation is incredible”

(Verizon)
What is IOT? (to me)

- A communications revolution
- *Communicating* not just copying
- An *interactive* universe of objects, things, data, ideas/concepts and processes both real & virtual
- A *mechanism* to interconnect devices, assets, processes, and systems to improve business models and profits, increase efficiency, and optimize use of resources
- All elements can be integrated & communicate - *co-dependency*
- A development *platform***
A platform is an application development environment - beyond the poster children - identifying opportunities.

The “Internet platform” resulted in e-mail, WWW, Twitter, etc., etc.

The strength of WWW is not just the number of servers or number of pages it has, but what we can do with them (e.g., E-Commerce, E-Learning, E-Government, etc.)
IOT as a *Platform* (2/2)

- The strength of IOT is not just what new (and perhaps bizarre) devices we can add to it
- A “Web of Things” can be built on the IOT
- What will IOT apps look like?; what will they do?
- Will there be a “killer app” for IOT?
Data is the language and currency of the 4th Platform
Challenges of Harnessing Big Data

• Not falling for the Big Data hype - why is Big Data better?
• Mining huge datasets; optimizing the signal-to-noise-ratio
• Identifying new analytic techniques and engines
• Shortages of Big Data experts
• Privacy, legal, and social issues
• Addition of the IOT is only going to increase these challenges by orders of magnitude (or not)
• How to avoid “information overload”
Data Analytics and Data Science

• “…[data] analytics is the process of obtaining optimal or realistic decision(s) based on existing data” (Wikipedia)

• “[data analytics is]…the extensive use of data, statistical and quantitative analysis, explanatory and predictive models, and fact-based management to drive decisions and actions” (Competing on Analytics: the New Science of Winning)

• There is nothing here about dataset size
“The theory is that you pump Big Data into the ‘black box’ of an analytics engine - most likely hidden on some unknown server in the cloud - and you get back a continuous stream of insights”
“When you have large amounts of data your appetite for hypotheses tends to get larger. And if it’s growing faster than the statistical strength of the data, then many of your inferences are likely to be false. They are likely to be ‘white noise.’ We have to have error bars around our predictions.”

-Michael Jordan, UC Berkeley
Understand when Big Data is better

- Outliers or small clusters
- Rare discrete values or classes
- Missing values
- Rare events or objects
• With the IOT we will have to learn to filter as well as analyze

• There will be some communications/data that we do not need to know, collect, or store or could be done at the “thing” level
The Fourth Quadrant

Big Data
Big Compute

Conventional
Computation

Social
Networking

More machines

More people

(From David de Roure)
Maybe sharing the Big Data Analytics
The Intersection of “Big Data,” IOT, and Data Analytics

A computationally-enabled sense-making network of expertise, data, software, models and narratives
Today’s Takeaways (1/2)

- New forms of data from new data sources will allow us to answer old questions in new ways and to ask and answer entirely new questions
- The future is being driven by a convergence of disruptive technologies
  - Data volume
  - Creative & realtime analytics
  - Computational infrastructure
  - Dataflows vs. datasets
  - Correlation vs. causation
  - Increasing automation
  - Machine2Machine data in the Internet of Things
  - Change in use - not just commercial data mining
Today’s Takeaways (2/2)

• The intersection/convergence of Big Data, Data Analytics, and the Internet of Things can lead to “The Fourth Platform”
  • We need to think of the sum rather than just the parts
  • Avoid the hype
  • The potential has only begun to be realized - embrace the platform
  • With the convergence comes big responsibility and unforeseen challenges

• We have an unprecedented opportunity to create knowledge - Let’s Do It!
The Distinguished Speakers Program is made possible by

For additional information, please visit http://dsp.acm.org/
About ACM

ACM, the Association for Computing Machinery is the world’s largest educational and scientific computing society, uniting educators, researchers and professionals to inspire dialogue, share resources and address the field’s challenges.

ACM strengthens the computing profession’s collective voice through strong leadership, promotion of the highest standards, and recognition of technical excellence.

ACM supports the professional growth of its members by providing opportunities for life-long learning, career development, and professional networking.

With over 100,000 members from over 100 countries, ACM works to advance computing as a science and a profession. www.acm.org
Thank You!
Questions? Comments?

bebo@slac.stanford.edu