What's Next for Big Data?

Lessons from a Decade+ Experiment in Big Data

A Discussion with IEEE BDIW

David Belanger PhD
Senior Research Fellow – Stevens Institute of Technology
dbelange@stevens.edu

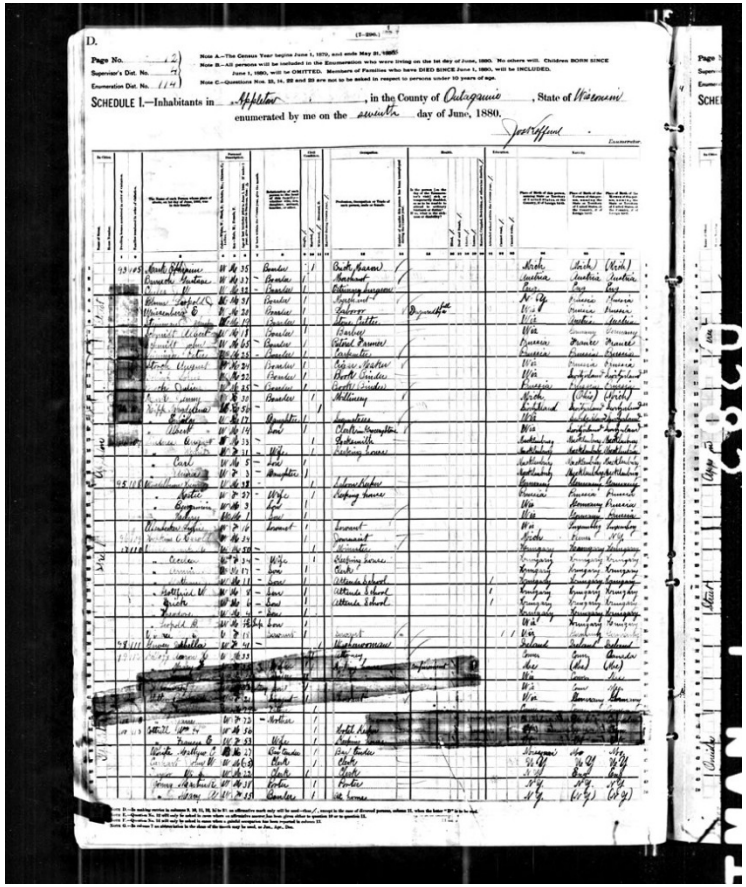
Introduction
What, How, Why

Big Data 1880

Census

Population: 50,189,209
Size: Low Gigabytes

Hollerith Tabulating Machine

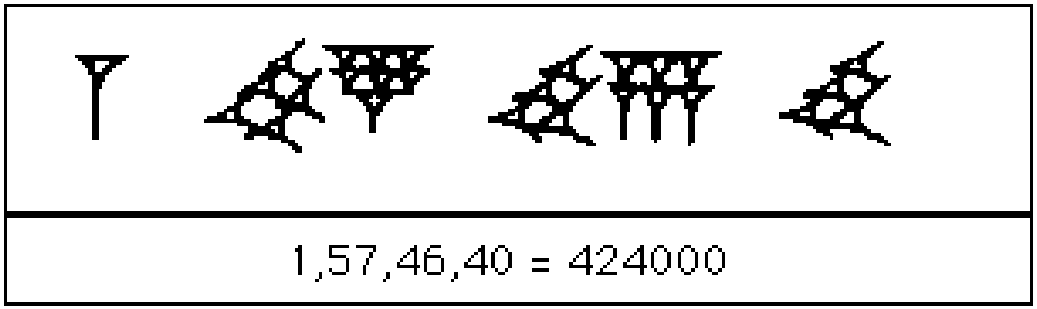


Source: http://www.winshuttle.com/big-data-timeline/

Big Data 2000 BC

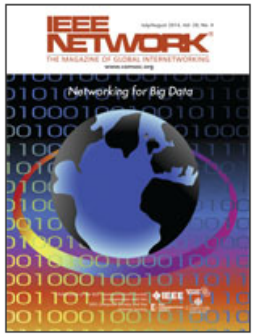
Base 60 Positional Arithmetic

1,57,46,40 in Babylonian numerals



Source: http://www-history.mcs.st-and.ac.uk/HistTopics/Babylonian_numerals.html

ComSoc – Big Data Contributions



This month's complimentary access to a select article is from current issue of IEEE Network

"Big Data: Transforming the Design Philosophy of Future Internet"

- ✓ Communications
- ✓ Conferences
- ✓ Publications
- ✓ Training

IEEE COMSOC TRAINING™ CONTINUING EDUCATION FOR COMMUNICATION PROFESSIONALS

IEEE COMMUNICATIONS SOCIETY

CONNECT WITH IEEE COMSOC TRAINING:

HOME CALENDAR INSTRUCTORS CUSTOM TRAINING IEEE CEUS / PDHS

THE BIG PICTURE FOR SMALL CELLS

Create a Small Cell Deployment Strategy

Instructor: Jonathan Levine

Wednesday, October 15, 2014 - 9:00am to 4:30pm EDT

Online via WebEx

EXPLORE THE WORLD OF BIG DATA DURING AT IEEE/IFIP IM'15 CONFERENCE IN OTTAWA, CANADA

Like 5 Tweet 2 +1 1 Share 4

IEEE Communications MAGAZINE

CURRENT ISSUE CALL FOR PAPERS GUEST EDITOR KIT REVIEWER GUIDELINES

AUTHOR GUIDELINES EDITORIAL BOARD THE PRESIDENT'S PAGE ADVERTISING

HOME » THE PRESIDENT'S PAGE - AUGUST 2014

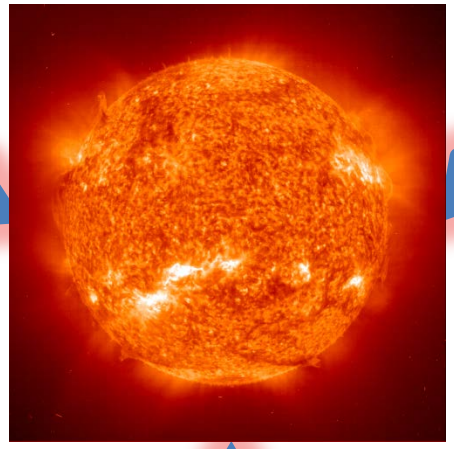
THE PRESIDENT'S PAGE - AUGUST 2014

Every year a full day in June is spent with Society Presidents or their delegates, discussing promising areas that may be worth exploring. The June meeting is open to all the operating units of IEEE. We typically have Educational Activities, Member and Geographic Areas, Marketing, Corporate, the Standards Association, as well as other unique areas, attending this open meeting. As an example, this June the area of Leveraging Big Data was identified. This is an area where a number of Societies are already involved, and what FDC will be doing, together with interested Societies, is considering how the theme of Leveraging Big Data can be addressed in a transversal way, including how to leverage data being created and harvested by IEEE and its Societies. The Future Directions Committee was charged by the NIC in 2013 and 2014 to develop a stand-alone Big Data initiative, as the Cloud Computing Initiative moves out of the FDC and into a sustainable environment within the IEEE operational units. Typically the FDC first begins incubating a new initiative within Technical Activities, but in the case of Big Data, the incubation has taken place within the Cloud Computing Initiative.

Big Data: What?, How?, Why?

DATA

- Volume
- Velocity
- Variety
- ++



TECHNOLOGY, GOVERNANCE, ORGANIZATION

"…datasets whose size is beyond the ability of typical database software tools to capture, store, manage, and analyze."
Technologies:

- Networking
- Computing
- Analysis
- Visualization
- Application

"Valuable capabilities that were formerly too difficult, costly, or simply not possible."

APPLICATIONS

http://teotw.net/energy_nuclear.html

Big Data Directions

1

Traditional Services TP, DW, Analytic Reports

2

Info in Flight
Real Time Stream Mining
Mining Unstructured Data
Mobility , Analysis @Scale,
BD in Cloud

3

Next-Generation Value from Data

Networked & Programmable World

Pervasive Monitoring/Control

Internet of Things

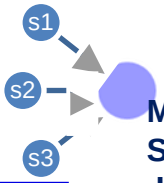
COSM
Xively

OPEN DATA

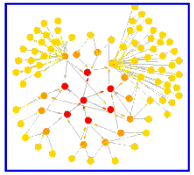
Immersive, Augmented Reality Interfaces

Diversification of Applications & Users

Next Gen Analytics, Prediction



Multiple RT Streams / Integration



Data Stream Mining

Speech/Text Mining

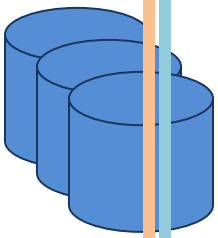
Anywhere, AnyDevice

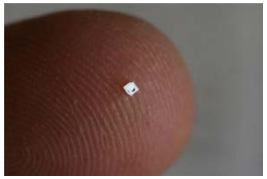
DATA

Networked & Programmable World: IoT + Big Data

Monitor

Analyze

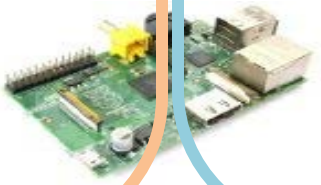




Instrument

Decide

Control



Internet of Things

Big Data

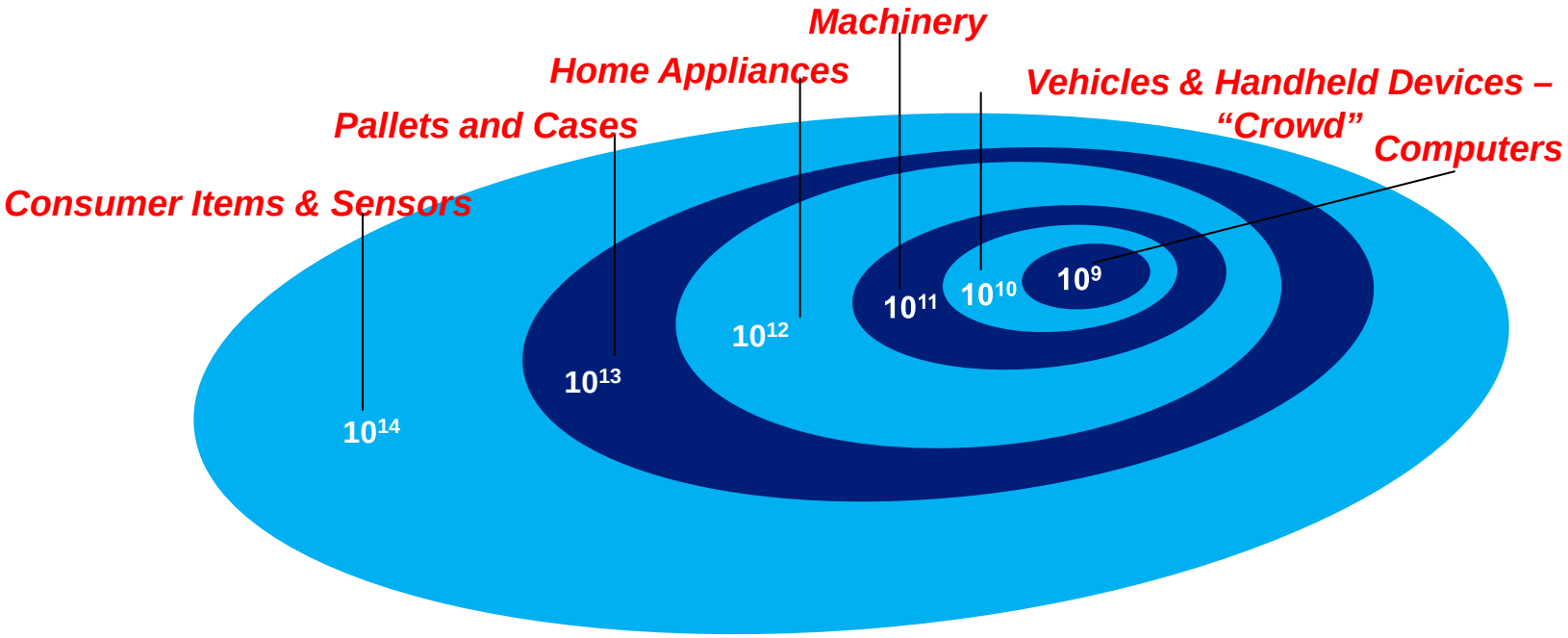
Some Things That Make a Difference DATA

Change	Classical	Big Data	Example
Granularity	Transactional or Aggregate	Elementary, Personalized	Web Transaction vs Click Stream
Signal Strength	Strong	Σ Weak	Google Flu
Latency	Transactional or Long	Streams and Real time	Location
Location	ZIP Code, Area Code, Nxx,	GPS, Lat/Long	Location Based Systems
Structure	Relational	Structured, SemiStructured, Unstructured, Graph	Text/Speech Analysis, (e.g. Twitter), Social Networks
Sources of Data	**Operational**	**Operational + Crowd + Sensors + IoT + Open**	**Dallas Museum of Art, data.gov - 156,584 datasets**
Data Integration @ Scale	**Hard, Join, Structured**	**Structured & Unstructured, Large Scale, Still not easy**	

The next “crowd” - Objects

Devices That Can Be Networked & IP Addressable

How can we best exploit the billions of devices, many mobile, intelligent, and video enabled as computing, sensing, and communications platforms?



- ***Invisible Computing - Will Far outnumber current IT Devices and People***
- ***Home Appliances***
- ***Consumer Items***
- ***Pallets and Cases***
- ***Machinery***
- ***Vehicles and Handheld Devices***
- ***Sensors – Machine to Machine Internet***

Integration

- Communications
- Entertainment

Technology

Some Things That Make a Difference Technology

Change	Classical	Big Data	Example
Computing Platforms	Large Symmetric Multiprocessors, expensive	Parallelism using commodity hardware	Cost reduced by x10, Cloud
Software Platforms	RDBMS, Analytic Software, Viz Software	Oriented to massive parallelism, Often Open Source	Map/Reduce, Hadoop, Storm,..
Data Base Systems	RDBMS, Transactional	Often Column Oriented, Availability Oriented, Document	Hbase, MongoDB, Cassandra, ….
Visualization	Static, Aggregate, Dashboards	Interactive, Drill Down	Swift
Streams vs Warehouses	Warehouse: Size: N View: Query Latency: High	Streams: Size: ∞ View: Window Latency: Low	Fraud, High Frequency Trading, Targeted Marketing
Networks	Access speed limited,	Access: Variety and Bandwidth; Core: Bandwidth & Low Latency	Ubiquitous.
Standards	Common	Evolving, Needed	Data Query Lang.

Dimensions of Technology

DATA

Structured – SemiStructured - Unstructured

Networking: Efficient, Reliable, Secure Data Transport.

Computing: Storage and Processing Architectures that operate at scale, and in Real Time

Analysis: Industry leading information mining and analysis technology.

Visualization: The most effective way to deliver information & alerts to decision makers

APPLICATIONS

Fraud/Security – Marketing – CFO – Hosting – Mobility - …

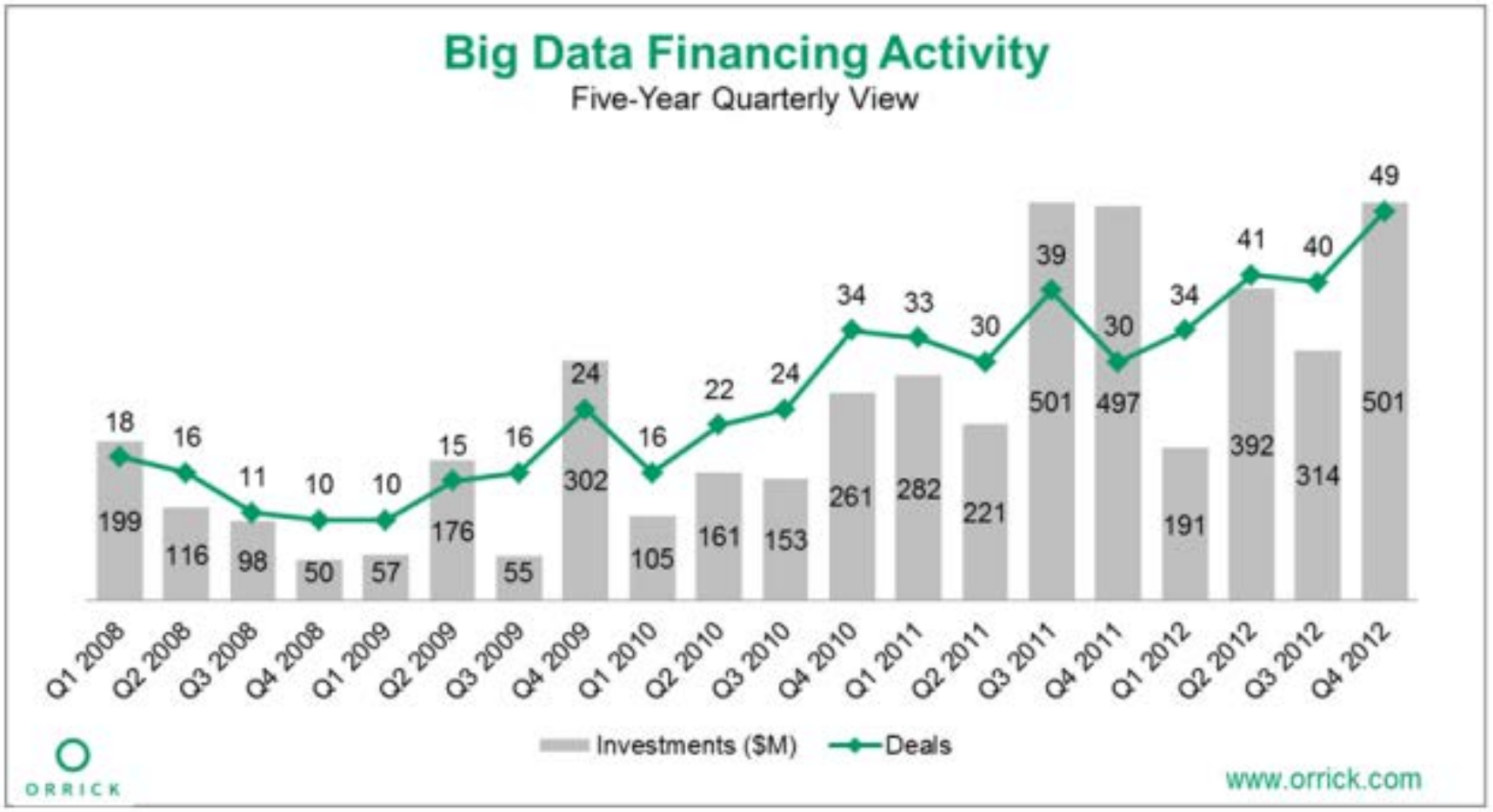
Source: AT&T Labs

What Got Us Here: Technology Startups/Open Source Landscape

Source: http://www.bigdata-startups.com/open-source-tools/

Investment in Big Data

Investment in Big Data Companies in the Big Data industry have received $4.9 billion in funding since 2008, with the bulk of financing in 2011 and 2012.



Source: http://www.kdnuggets.com/2013/03/top-big-data-vc-investors.html

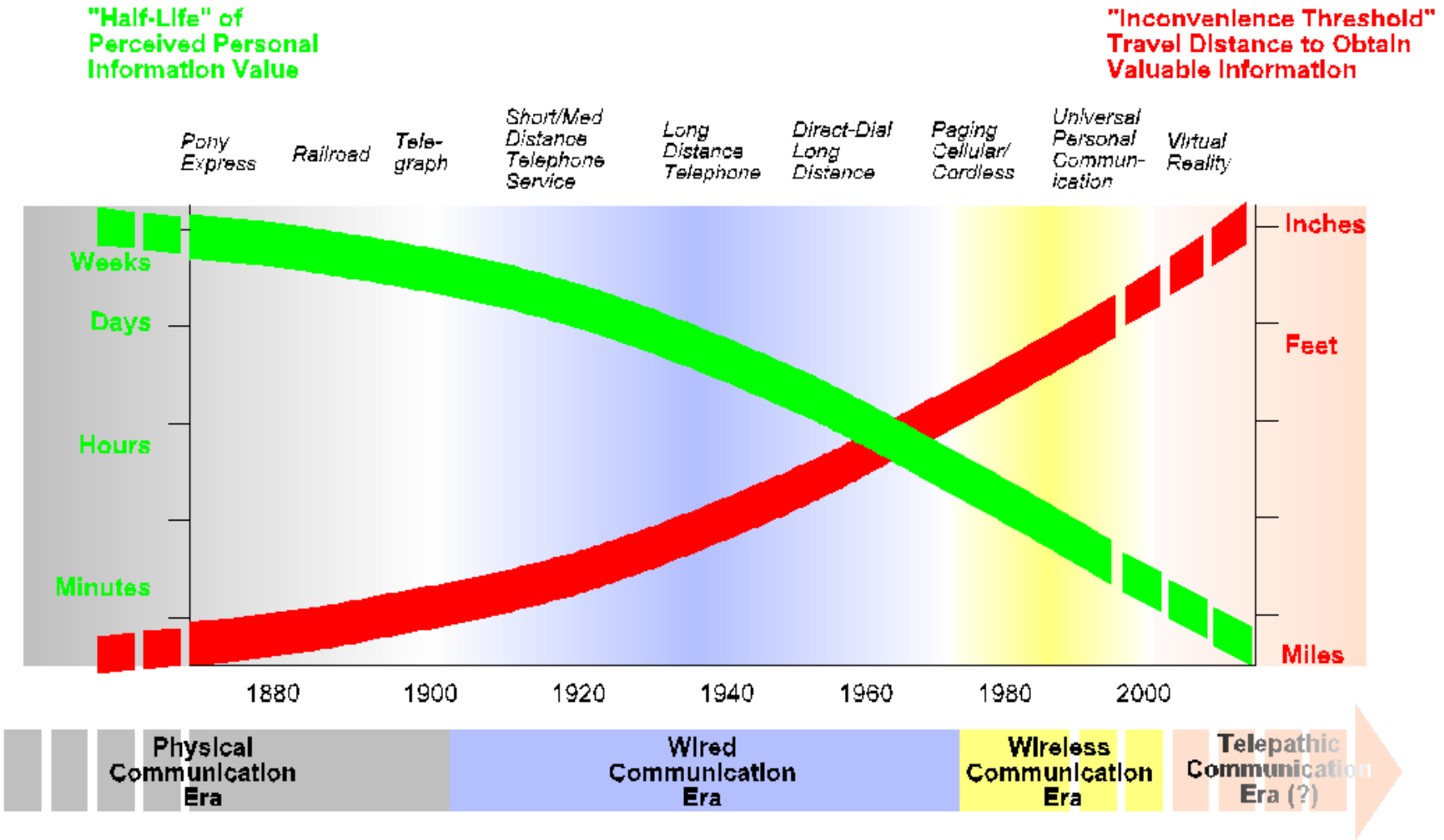
At the end of the day, Big Data is about taking advantage of scale rather than letting scale take advantage of you.

http://www.google.com/trends/explore#q=big%20data&cmpt=q

http://www.google.com/trends/explore#q=data%20scientist&cmpt=q

Applications
Making Scale Your Friend

Mobility Trends:
"Half Life" and "Inconvenience Threshold"

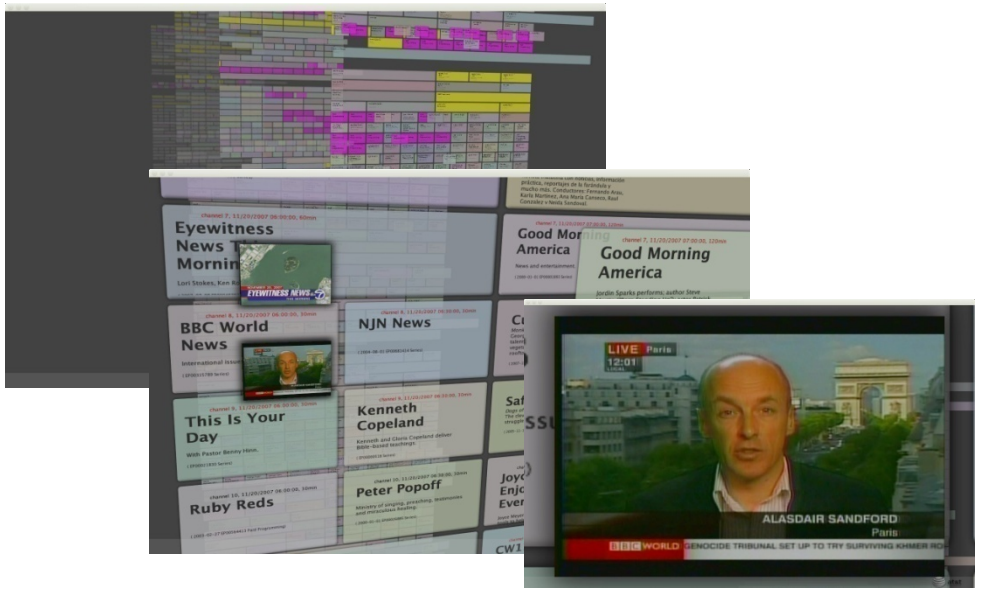


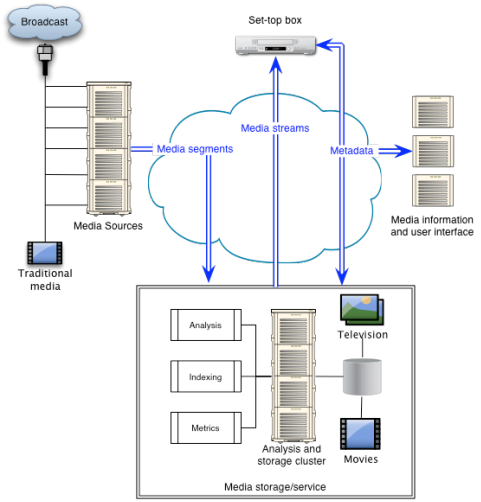
Each communication technology advance has shortened the useful life of information and increased the need to obtain new information more rapidly regardless of the situation or location...

Convergence of Communication and Entertainment

Can I find what I want and watch, when I want?

- **Base Issue**
  - How do we differentiate our media services in a world of commodity media services?
- **Vision**
  - Access to all media, from all time, from anywhere
- **Challenges**
  - User interface, especially search
  - Indexing
  - Storage
  - Distribution





Source: AT&T Labs

Some Things That Make a Difference Applications

Change	Classical	Big Data	Example
Granularity	Transactional or Aggregate	Atomic, Personalized	Individual's location at a specific time
Rare Events	Sample Based	Population, Signatures, Classification	Fraud, Security, Social Location
Broad Search	Indexed Retrieval, Flat Files	Structured and Unstructured data	Search Engines
Relationships	Customized	Graphs and (social) Networks	Recommender systems, CoI
Analytics at Scale	Sample	Population	Call Detail, Call Center
Personalized Traits	Survey based	Behavioral	Signatures, Targeted Marketing
Process Control	Silo'ed	End to End, Feedback Control	Virtual Integration
Recommendations, Learning	Rules	ML, Collaborative, Relationship based	Translation, Recommenders
Transparency	Within Silo	Across Processes	Provisioning
Prediction	Limited, Aggregate	Real time, Specific	Slipper
Text/Speech/Video/Image	Limited, often special purpose	Text, Speech Now; Others soon	Many uses of Twitter, Contracts.
Control	Customized, Rule Based	Learning Based	Driverless Cars

Canonical Examples of Big Data (1) Search

Volume

Variety

```
<!DOCTYPE HTML
PUBLIC "-//W3C//DTD
HTML 4.01//EN"
"http://www.w3.org/TR/html
4/strict.dtd"> <HTML>
<HEAD> <TITLE>My first
HTML document</TITLE>
</HEAD> <BODY>
<P>Hello world! </BODY>
</HTML>
```

http://www.w3.org/TR/REC-html40/struct/global.html

http://www.worldwidewebsize.com/

Canonical Examples of Big Data (2) Fraud

Volume

Velocity

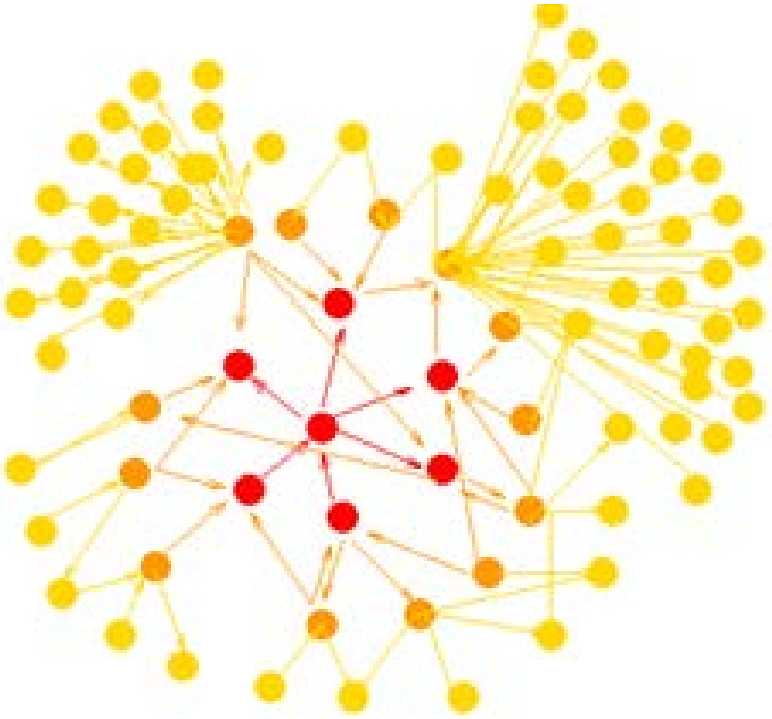
Initiating Phone Number	Terminating Phone Number	Time of Day	Day of Week	Duration	Bus/ Res	Lat	Long		

Canonical Examples of Big Data (3)
Marketing (Targeted/Viral/Recommenders)

Velocity

Volume

Variety



Canonical Examples of Big Data (4) Call Center

Velocity

Variety

http://www.featurepics.com/online/Call-Center-Photos214626.aspx

Canonical Examples of Big Data (5)
Machine Translation

Volume

Variety

Watson, come here, I want you

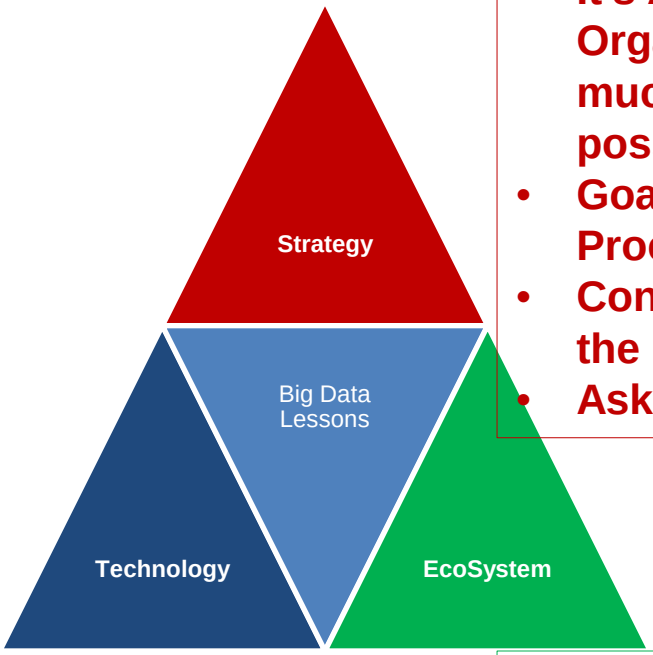
translate.google.com	**沃森，來到這裡，我要你**
Native	**沃森**，过来，我需要你。

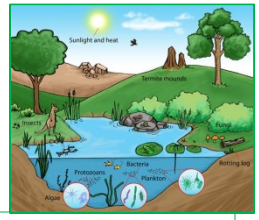
Lessons from Experience

Lessons from Experience

- It's All About the Data: Organic & Inorganic, As much & as fresh as possible
- Goal: Fundamental Process & Product Change
- Concentrate on Framing the Right Questions.
- Ask Big Questions

- Multidimensional Expertise is Essential -> Organizational Change
- The Nature of Analytics has Changed
- There are tradeoffs: e.g. Speed vs. Accuracy
- Think Steams of Data

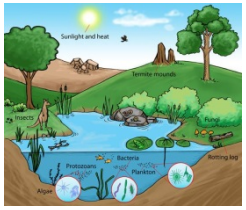




- Do Not Ignore the Broader World: Data, Technology, Practices
- We are entering a second generation, which will include many more players
- Customer Focus: Make Heroes of Your Partners
- RESEARCH – INCUBATION - PRODUCTION

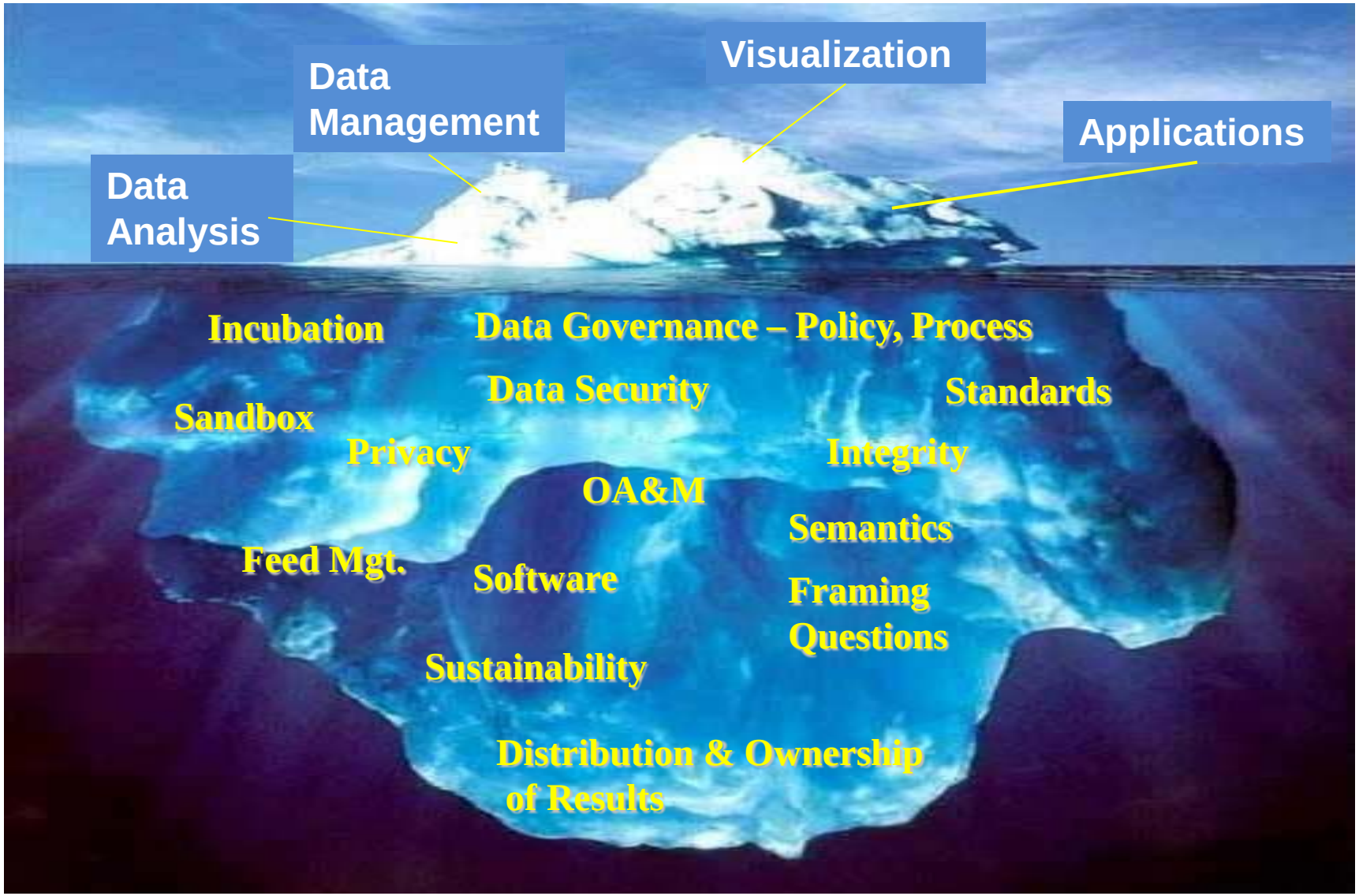
A Few More Lessons Learned





- **Technology (Leverage Aggressively):**
  - **Multidimensional technical expertise is essential:**
    - **Network – Computing – Data Analysis – Visualization - Domains**
  - **Data Feed management can scale linearly. That is really bad.**
  - **Interactive visualization can be the most effective tool for both the Understanding and Integration of data.**
  - **Tradeoffs: Speed vs. Accuracy : Rules vs. Differences : Weak & Strong Signals**
- **Strategy:**
  - **It's all about the Data**
    - **Collect all the data you can.**
    - **Collect, and use, the freshest data available.**
  - **Do not ignore unstructured data.**
  - **Occam's Razor will win over time, though it may be a while.**
- **Ecosystem (Leverage Expertise Wherever You Can):**
  - **In many areas, Standards have not become firm. Look for them where possible.**
  - **Open Data and Self Generated Data are competitive necessities**

A More Complete Picture



Meta Challenges

Then	Now
Significant systems containing sensitive data are not easily accessed. Complex semantics and poor integrity often exist, but impact is hidden because data is relatively closed. Integration, outside of joins, uncommon at scale.	Protection of SPI data a constant problem. Transparency of use, integrity of data a concern. Open data provides much more opportunity for interesting new apps from integration, and semantic confusion. Integration complex.

Security
Privacy
Integrity
Semantics
Integration

Data Quality and Integrity

Then	Now
Much of the burden of quality and integrity lies in the fact that the ACID properties and input rules of transactional systems are strictly enforced. There is a very rich technical ecosystem that has been built around integrity and is made available in most mature data management systems.	In many systems, at the volume, velocity, latency, and complexity expected, the levels of correctness required of transactional systems are neither possible nor necessary. Analytic techniques must take these changes into account. Given the very diverse nature of potential data sources, and consequent reduction in control over the data, this becomes a very challenging problem.

Organizing for Innovation - Maturity

Then	Now
Classical research or exploratory development teams create new products, often in large teams with significant timelines. Careful attention is paid to decision gates to prevent runaway costs. Due to costs, decisions often top down.	Small, elite teams create prototypes of potential products quickly, trial the prototypes, and, when successful, present for funding to go to market. Go to trial, very quick. Classical research provides technology base to prototyping teams, and partners with them.

Incubation and Production
Data OA&M
Data Governance

Data Governance

Then	Now
Data governance is typically the responsibility of the IT department, who, in cooperation with various stakeholders, develops policy, along with the processes to implement the policy.	Driven by the broader use models of big data, and the significance of non-IT stakeholders, policy should be done at a higher level. Protection of SPI data a constant problem. Transparency of use, and integrity of data are concerns. In general, the level of risk associated with the creation, curation, management, and use of data is considerably higher, hence governance is an even more critical task. The state of the art in this area is not yet mature.

Organization

Production

Incubation

R&D

Source: http://crocodilian.com/cnhc/potm-apr02.html
http://www.frauleindi.com/HiltonHeadNature/Wildlife/wildlife.htm
http://www.animaltrial.com/alligator.html

Some Thoughts

- Use of existing database, and expansion of data sources.
- Standards informed by breadth of expertise
- Leadership in second generation breadth and depth
- Integration with other emerging technologies such as IoT, Cloud, etc.

Thank You!