DO YOU KNOW BIG DATA? v.1*

DR. DAN LAW, CHARLIE GREENBACKER, JOHN EBERHARDT (ALTAMIRA)

WHAT IS BIG DATA?

Many definitions…

- **The Multiple V's:** *Data that brings challenges in Volume (size), Velocity (speed), Variety (formats), Veracity (accuracy), as well as Visualization, Value, Vendors, etc.*
- **McKinsey:** *"Datasets whose size is beyond the ability of typical database software tools to capture, store, manage, and analyze."*
- **Economist**: *"Society has more information than ever before and we can do things when we have a large body of information that simply we could not do when we only have smaller amounts"*
- **Wikipedia**: *"Big data is the term for a collection of data sets so large and complex that it becomes difficult to process using on-hand database management tools or traditional data processing applications"*
- **Adam Jacobs, 1010data**: *"Data whose size forces us to look beyond the tried-and-true methods that are prevalent"*
- **Dan Law, Altamira**: *"Data of any type with potential value that exceeds the analytic capabilities of traditional stand-alone solutions"*
- **John Eberhardt, Altamira**: *"Any data collection that cannot be managed as a single instance."*

WHAT TYPES OF DATA ARE IN BIG DATA?

- **Structured Data**
  - ✓ *Tables, Relational Data, etc. with semantics*
- **Unstructured Data**
  - ✓ *Raw Text, Images, Video, Audio*
- **Semi-structured**
  - ✓ *Hybrid data, such as documents with tables*
- **Metadata**
  - ✓ *Structured data about data, e.g. from/to*
- **Streaming Data**
  - ✓ *Data that moves across networks at high speed*
- **Temporal Data**
  - ✓ *Data including trends / activities in time*
- **Geospatial Data**
  - ✓ *Data that includes information on positions in space (regions, points, tracks, shapes)*
- **And many others…**

HOW DO WE EXTRACT KNOWLEDGE FROM BIG DATA?

By teaching computers to extract knowledge:

- ✓ *By agreeing upon and defining semantic concepts of knowledge in one or more "knowledge representations" (for example, a fixed ontology, auto-generated ontology, user-defined tags)*
- ✓ *By building **transforms** to map semantic content from **structured data** into knowledge representations*
- ✓ *By building **classifiers** to extract semantic content from **unstructured data** and to map that extracted content into knowledge representations*
- ✓ *By building **analytics** to correlate / fuse / perform reasoning on extracted semantic content to generate even more semantic content, and to map that information into a knowledge representations*

And by doing this on a large scale…

- ✓ *By dividing a problem into pieces and executing in parallel (e.g. MapReduce)*
- ✓ *By building clever indexes of knowledge, so that you can search it quickly…*
- ✓ *By using high performance computers (HPCs) or other fast electronics (e.g. FPGAs, ASICs, Optics)*
- ✓ *A mixture of the above… (e.g. Netezza, YarcData, Next Generation Oracle)*

WHAT DO WE DO WITH KNOWLEDGE WE EXTRACT?

We can estimate and visualize parameters in data using statistics

- ✓ *We can describe data, explore correlations, discover patterns, predict outcomes, etc. through "observational studies"*
- ✓ *We need to account/correct for Bias and Confounding, for example by introducing elements of chance!*
- *We need to consider selection bias, measurement bias, analysis bias, error, confounding variables*

We can implement rules to trigger actions in response to discovered knowledge

WHAT TYPES OF VISUAL TECHNIQUES ARE THERE?

TYPE	UTILITY	PROS	CONS
Spreadsheets	Viewing tabular data	Simple/Common	Can't see patterns
Common Charts	Viewing numeric data	See Patterns/Trends	Hard to pivot/explore
Graphs	Exploring networks	Powerful analysis	Complex / Intensive
Geospatial views	Viewing data in space	Intuitive maps	Graphics intensive
Temporal views	Viewing data in time	Find patterns/trends	Not all data temporal
Spatiotemporal	Both space & time	Powerful analysis	Uncommon, Intensive
3D Views	Viewing complex data	More immersive	Graphics Intensive
Spatiotemporal	Immersive visualization	Intuitive / Powerful	Specialized Hardware

WHAT TYPES OF STATISTICAL ALGORITHMS ARE THERE?

ALGORITHM	UTILITY	PROS	CONS
Linear	Providing point estimates	High precision, easy	Not qualitative, high curation burden
Non-Linear	Processing complex systems	Supports more complex data, complex decisions	Limited inference, high supervision needed
Fuzzy Logic /Neural	Representing highly complex, qualitative systems	Complex inference, messy data	Lower precision, seed value bias
Probabilistic	Distribution, probability oriented	Complex dependencies, fuzzy decisions	Lower precision, no point estimates, see value bias
Graph	Representation of data	Represent large sets, easy interaction	Limited inference, computationally challenging

HOW BIG IS BIG DATA?

It can be REALLY BIG!:

- *Internet traffic is now ~5 Zettabyte per year (IBM)*
- *1 Zettabyte = 1 billion terabytes*
- *Visa processes 150 Million transactions per day (VISA)*
- *Library of Congress holds 3.2 Petabytes of data*
- *207 Terabytes of video loaded daily on YouTube (2012)*
- *50 billion devices connected to the Internet by 2020 (IDC)*
- *50 Billion photos on Facebook in 2010*
- *400 Million Tweets per day (Washington Post)*
- *Seagate sold 330 Exabytes of hard drives in 2011*
- *LHC produces 500 Exabytes of particle collision data per day CERN*
- *Current iPhone 5s: 76 Gigaflops*
- *Fastest supercomputer: 50 Petaflops*
- *Interesting Comparison: Human Brain has 100 Billion Neurons (100 Giga-Neurons), 100 Trillion Synapses (100 Tera-Synapses), neurons "fire" 1-1000 times/second (100 Giga-fires to 100 Tera-fires per second)*

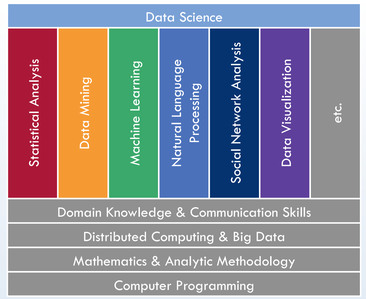
WHAT IS A DATA SCIENTIST?

People or teams that possess a special mix of skills:

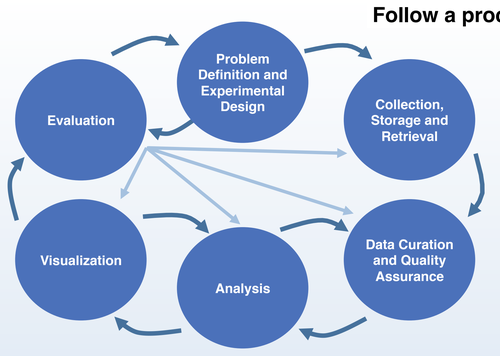
- *They are "T-shaped" (see graphic at right)*
- *They have mastered all foundational areas of data science (horizontals in the graphic)*
  - *Computer Programming*
  - *Mathematics & Analytic Methodology (Stats)*
  - *Big Data Technologies*
  - *Communications Skills*
- *They possess deep expertise in at least one specialty area (verticals in the graphic)*

Popular data scientist tools:

- *R, Python, Mahout, Pandas, Many Others…*
- *See http://oss4ds.com*



HOW DO WE IMPLEMENT BIG DATA SOLUTIONS?



Follow a process, that considers:

- *Team Skills*
- *Problem Definition*
- *Experimental Design*
- *Success/Evaluation Criteria*
- *Data, Curation & QA*
- *Solution Design*
  - *Infrastructure*
  - *Ingest & Storage*
  - *Analytics*
  - *Visualization*
  - *Security*
  - *Privacy & Ethics*
  - *Budget & Schedule*

Not a linear process

- *Try an agile approach…*

HOW DO WE ADDRESS PRIVACY & ETHICS IN BIG DATA?

Privacy. Be sure to comply with:

- *The 4th Amendment to the Constitution*
- *Electronic Communications Privacy Act*
- *Foreign Intelligence Surveillance Act*
- *The Privacy Act*
- *Executive Order 12333*
- *USA PATRIOT Act*

Ethics. Consider:

1. *Respect for Persons / informed consent*
2. *Beneficence*
3. *Justice*
4. *Respect for Law and Public Interest*

HOW DO WE SECURE BIG DATA?

By addressing the notorious nine:

1. *Data Breaches*
2. *Data Loss*
3. *Account Hijacking*
4. *Insecure APIs*
5. *Denial of Service*
6. *Malicious Insiders*
7. *Abuse and Nefarious Use*
8. *Insufficient Due Diligence*
9. *Shared Technology Issue*

By using big data to secure big data

- ✓ *Collect & analyze activity data, network data, audits, provenance, pedigree, lineage*

And by using:

- ✓ *Risk Management: ICD 503*
- ✓ *Access controls, IDAM, biometrics, PKI, physical security, cell-level security, smart data, encryption*
- ✓ *CND, Anti-Malware, anti-virus*

WHAT ARE LEADING BIG DATA TOOLS?

Big data tools fall along a "stack" spanning infrastructure to visualization

STACK ELEMENT	USED FOR	OPEN SOURCE EXAMPLES	COTS EXAMPLES
Visualization	• User Interface • Web-based tools	• D3js, 3js, Gephi, Ozone	• Tableau, Centrifuge, Visual Analytics
Analytics	• Machine learning • Statistical tools	• R, Mahout, Titan, OpenCV, Lumify, Hive, Pig, Spark	• SAS, SPSS, MapR, Palantir
Data Store	• Data & Metadata • Source Data • Indexes	• HDFS, Accumulo, MongoDB, Cassandra, Titan, Neo4j, MySQL	• Oracle, Marklogic, YarcData, Teradata
Ingest	• Transformation / Normalization • Ingest / Streams Processing	• Storm, Hadoop/MapReduce	• Splunk, SAS, Oracle, IBM
Infrastructure (IaaS, PaaS)	• CM, Scheduling, Monitoring • Application Operating Systems • Computers, Networks	• Linux, OpenShift, OpenStack, Puppet, Zookeeper, Oozie, HDFS, Kafka, JBoss, Xymon	• AWS, Azure, Cloudera, Red Hat, Rackspace, vendor specific

Select key components of Hadoop Ecosystem:

- *HDFS (Storage), MapReduce (Distributed Processing), Accumulo (Secure data store, Indexing)*

WHAT QUESTIONS SHOULD WE ASK ABOUT DATABASES?

- **What type of data do can we store?**
  - ✓ *Structured, unstructured, relational, graphs, entities…*
  - ✓ *Big Files (e.g. imagery)? Small files (e.g. text)?*
- **How is data ingested into the database?**
  - ✓ *Streaming? Batch?*
- **What does the database cost?**
  - ✓ *License costs? O&M costs? License restrictions?*
- **What hardware is required?**
  - ✓ *Commodity? Proprietary?*
- **How scalable is the database?**
  - ✓ *Gigabytes? Terabytes? Petabytes? Exabytes? Yottabytes?*
- **Is the database fault tolerant?**
  - ✓ *Does it need to be? What about COOP?*
- **Can we perform analytics using the database?**
  - ✓ *e.g. MapReduce?*
- **What is the latency for queries? Or for analytics?**
  - ✓ *e.g. milliseconds? days?*
- **Is it optimized for particular features?**
  - ✓ *Fast writes? Fast reads? Ease of use?*
- **How many users can the system support?**
  - ✓ *Scaling for data does not necessarily imply scaling for a large number of users*
- **Is the database secure?**
  - ✓ *Does it provide access controls? Has it been accredited? To what level?*

WHAT QUESTIONS… ABOUT PREDICTIVE TOOLS?

- **How does this tool perform prediction?**
  - ✓ *answers should list algorithms used (refer to table at left for descriptions/pros/cons…)*
- **What types of data does the tool analyze?**
  - ✓ *e.g. structured, unstructured, hybrid (are these right for your mission?). Petabytes of data?*
- **Does the tool correct for, or enable one to correct for potential bias or confounding variables?**
  - ✓ *e.g. by introducing elements of chance or by counting everything*
  - ✓ *if not, you should be skeptical of the predictions a tool makes!*

WANT TO LEARN MORE?

- **George Mason University**
  - ✓ *GMU has both full semester graduate-level courses and two day certificate course in Big Data Practices*
- **Learn about Data Science and Big Data Tools at www.oss4ds.com**
  - ✓ *The instructors are part of a broader data science team that publishes open source resources to help you learn about Data Science and Big Data*
- **Explore integrated open source analytic platforms at www.lumify.io**
  - ✓ *Learn about open-source ingest, knowledge extraction, and link analysis from structured and unstructured big data*
- **Learn more about big data implementations at www.altamiracorp.com**

A FEW USEFUL CONTACTS

**Please email Dan Law if you have comments or suggest any improvements to this poster*

ALTAMIRA: DAN LAW, DAN.LAW@ALTAMIRACORP.COM
AMAZON: HTTP://AWS.AMAZON.COM/FEDERAL/
CENTRIFUGE: INFO@CENTRIFUGESYSTEMS.COM
CLOUDERA: 866-843-7207
DATASTAX: INFO@DATASTAX.COM
IBM: 800-333-6705
KOVERSE: INFO@KOVERSE.COM
MARKLOGIC: INFO@MARKLOGIC.COM
MONGODB: 866-237-8815
ORACLE: 800-633-0738
SPLUNK: 866-438-7758
YARCDATA: 925-264-4700