Agenda

- What is Big Data?
- What generates Big Data?
- Use Cases for Big Data
- Open Research Questions

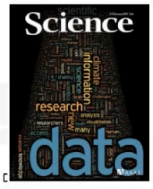
Big Data Definitions - NASA

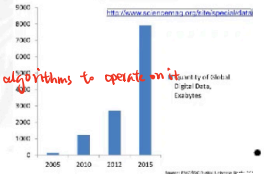
Big Data: Big Data describes the situation in which collection, storage and analysis of data exceeds the capability and capacity of conventional methods or software systems.

This state necessitates new architectural approaches in data management, artificial intelligence, statistics and visualization that change the paradigm by which data is collected, stored, processed, transmitted and analyzed, often through scalable cyberinfrastructures and algorithms.

U.S. National Science Foundation

- The Big Picture
  - Astronomy, Earth science, planetary science, life/physical science all *drowning in data*
  - Fundamental technologies and emerging techniques in archiving and data science
    - Largely center around *open source* communities and related systems
- Research challenges (adapted from NSF)
  - *More data is being collected than we can store*
  - *Many data sets are too large to download*
  - *Many data sets are too poorly organized to be useful*
  - *Many data sets are heterogeneous in type, structure*
  - *Data utility is limited by our ability to use it*
- Big Data Content Detection and Analysis
  - *Research methods for integrating intelligent algorithms for data triage, subsetting, summarization*
  - *Construct technologies for smart data movement*
  - *Evaluate cloud computing for storage/processing*
  - *Construct data/metadata translators "Babel Fish"*





A Summary of Some Web Facts

- **How many websites?** *1 billion+ (see Netcraft survey)*
- **How are they distributed across TLDs or across countries?**
  - *112 million out of 148 million belong to .com or about 72%*
- **How many web pages are there?** *1 trillion unique URLs from Google found in 2008,*
  see http://googleblog.blogspot.com/2008/07/we knew web was big.html
- **How many pages of different content types: HTML, PDF, Word, Excel, PPT, etc?**
  - *There are thousands of different content types*
- **How much storage is required to hold a single snapshot of the Web?**
  - *1 trillion web pages at 100K bytes per page requires 100 petabytes*
  - *Google processes 24 petabytes per day, http://en.wikipedia.org/wiki/Petabyte*
    The Internet Archive has 10 petabytes as of 10/2012, http://en.wikipedia.org/wiki/Internet Archive
  - *1 petabyte storage costs under $1,000*
- **What are the languages in which the documents are written?**
  - *As of 04/2013, about 55% is in English; other popular languages include: Russian, German, Chinese,*
    see http://en.wikipedia.org/wiki/Languages_used_on_the_Internet
- **How often are web pages changing**
  - *Studies done in 1999, 2002, 2004 are now old and out-of-date*
- **General properties of the Web graph**
  In-degree and out-degree distribution follows a power law
- **Categories of Content: pornography, spam, mirrors**
  Presumably there is a lot of the above, but little concrete data on how much

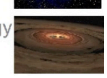
Credit: Ellis Horowitz – Chris Mattmann, 2012-2018

SKA: Key Science


Emerging from the Dark Ages & the Epoch of Reionization


Strong-field Tests of Gravity with Pulsars and Black Holes


Galaxy Evolution, Cosmology, & Dark Energy


The Cradle of Life & Astrobiology


Origin & Evolution of Cosmic Magnetism

Exploring the Universe with the world's largest radio telescope

National Institutes of Standards & Technology (NIST) – Big Data WG

- http://bigdatawg.nist.gov/home.php
- Big Data refers to the inability of traditional data architectures to efficiently handle the new datasets. Characteristics of Big Data that force new architectures are:
  - **Volume** (i.e., the size of the dataset);
  - **Variety** (i.e., data from multiple repositories, domains, or types);
  - **Velocity** (i.e., rate of flow); and
  - **Variability** (i.e., the change in other characteristics).

More data collected than we can store

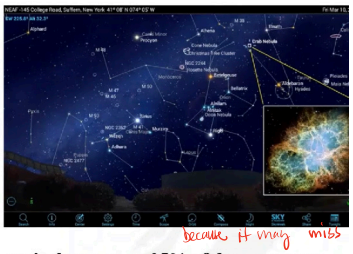
- Data is collected faster than we can store it
- Rationale
  - Instrument resolutions – cameras take bigger and better pictures nowadays (10+ megapixels on your phone, compared to in specialized instruments)
  - Data taken everywhere – "everyone is a sensor"
    - Cameras driving this
- Some examples

Square Kilometre Array

- What is it?
- Next generation radio astronomy instrument that will be built jointly by South Africa and Australia to image the sky like never before
- What's the status?
- Currently being built over the next decade
- Why do you care
  - *700 TB data/sec, 1000s of data types, 1,000,000s of users, geographically distributed, etc.*

Change in Astronomy - Time Domain

- Observation
  - Something missed, just *re-observe* on next pass
- Doesn't work anymore
  - What if on next pass, the object is no longer there
- Translation: we can't throw away 95% of data anymore



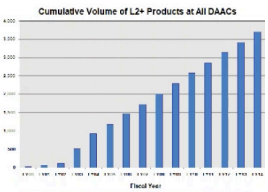
2 Many datasets too large to download

Volume + variety

- Chunking
- Organization of the data
  - Smaller or larger files?
  - By Space/Time?
  - By Type? Movies, Audio, Images, versus Text.
  - By metadata? By WHO is in the file?
- Implications on architectural style
  - Synchronous versus Asynchronous communication
  - PubSub versus *Streaming*
- Compression
  - BZip, LZO compression, etc.

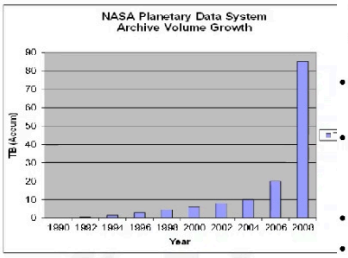
NASA Earth Science Datasets

- Increasing data volumes requiring new approaches for data production, validation, processing, archiving, discovery and data transfer/distribution (E.g., *scalability relative to available resources*)
  - Archives roughly doubling in size every two years* Shift from compute to data intensive *
- Increased emphasis on usability of the data (E.g., discovery, access and analysis)
- Data is DOUBLING every two years! Currently 10s-100s of Petabytes
- Increasing diversity of data sets and complexity for integrating across experiments
  - the benefits to science in bringing together and creating "fused" data products from multiple sources is critical in areas such as climatology where baseline data records are needed across measurements **
- Increasing distribution of coordinated processing and analysis (E.g., *federation*)
- Increasing desire for PIs to have integrated tool sets to work with data products with their own environments (E.g. *perform their own analysis locally*)



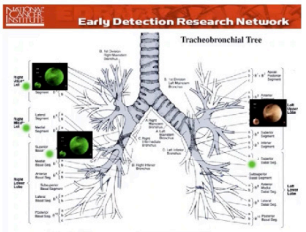
Example: Volume and Velocity

- Content repositories are growing rapidly in size
- At the same time, we expect more immediate dissemination of this data
- How do we distribute it…
- In a performant manner?
- Fulfilling system requirements?



- Content repositories are growing rapidly in size
- At the same time, we expect more immediate dissemination of this data
- How do we distribute it…
- In a performant manner?
- Fulfilling system requirements?

- Content repositories are growing rapidly in size
- At the same time, we expect more immediate dissemination of this data
- How do we distribute it…
- In a performant manner?
- Fulfilling system requirements?



3 Many datasets are too poorly organized to be useful

how we keep and store data

- Data *Forensics*
- Sometimes for scientific analysis
  - NASA Lunar Recovery Project
    - http://www.nasa.gov/topics/moonmars/features/LOIRP/
  - International Environmental Data Rescue Organization (IEDRO)http://www.theatlantic.com/technology/archive/2014/08/the-quest-to-scan-millions-of-weather-records/378962/
- Sometimes for defense and military
  - Seal Team 6 – cache of disk and data bricks/forensics

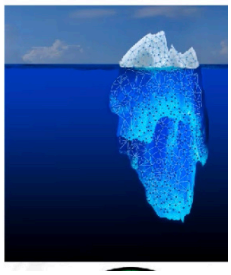
Binary Data

- How do you define "binary"??
  - Only things with a mimetype that starts text/ ?
    - Only "application/octet-stream"?
    - Only "application/x-<whatever>"?
  - Or do you want to include application/xml files?
  - Or things that extend from XML like DIF and FictionBook?
  - Only things that contain ascii-printable characters?
  - Other?
- Real question from Apache Tika list: 1/11/18

The World Wide Web

search web ⇒ search surface

- Data organized around "sites"
- Sites can come & go
- Surface web – 3-5% can see
- "Deep web" – 95-97%
  - Behind forms, Ajax/JS, and/or heterogeneous data
- Dark Web….



4 Many datasets are heterogeneous in size and structure

- Images
- Videos
- Audio
- PDF, documents
- Science data
- Code

Heterogeneous data sets

- File oriented
- Streaming data
- Databases
- NoSQL
- We will come back to this!

The Five V's

- **Volume** – more data collected than we can store
  - Total Volume at inception; at end of the project
  - Implications on Storage, on Services, on Architectures
- **Velocity** – too large to download and more than we can store
  - How fast is it coming
  - Matters – I/O – how fast the disk can write the data
  - Triaging the data
- **Variety** – too unorganized to be useful – heterogeneous in size/structure
  - How many data and metadata types – how are they described?
- **Veracity** – limited by our ability to use the data
  - Confidence in the data – how big is the error bar? Is this a guess?
- **Value** – limited by our ability to use the data
  - What decisions can be made on this data, and how easily is it to turn this data into knowledge and value?

Data utility is limited by our ability to use the data

- Ability to search and interact with information
- Self describing data
  - Linked open data, RDF, Microformats, etc.
- *Metadata* as a first class citizen
  - *Data about data*

Understanding the right ways to present information

2012 U.S. Presidential Initiative in Big Data

- "Big Data is a Big Deal"
  - https://www.whitehouse.gov/blog/2012/03/29/big-data-big-deal
- U.S. calls for 100M investment in Big Data Research
  - NSF, NIH, USGS, DARPA amongst agencies to respond
  - Really "took" from existing pot
  - Not new money
- DARPA XDATA one of the first responses to this

Big Data Use Cases

- *How do we handle 700 TB/sec of data coming off the wire when we actually have to keep it around?*
  - Required by the Square Kilometre Array
- *Joe scientist says I've got an IDL or Matlab algorithm that I will not change and I need to run it on 10 years of data from the Colorado River Basin and store and disseminate the output products*
  - Required by the Western Snow Hydrology project
- *How do we compare petabytes of climate model output data in a variety of formats (HDF, NetCDF, Grib, etc.) with petabytes of remote sensing data to improve climate models for the next IPCC assessment?*
  - Required by the 5th IPCC assessment and the Earth System Grid and NASA
- *How do we catalog all of NASA's current planetary science data?*
  - Required by the NASA Planetary Data System

Implications

- Need to be able to understand data
- Need to handle large amounts of data
- Need to handle data coming in fast
- Need to be able to identify data that is important versus data that is not
- Need to have confidence in our assessment of data and its value
- Need to be able to visualize, interact with, and find data

Lack of Data Legislation

lots of data collected by firms and govs

- US puts "few legal constraints on corporate exploitation and trafficking in personal data" – IEEE Computer 2013
- The following scenario
  - "Social media may know you are pregnant before your father does"
    - Combine user's
      - Web Search History – easy to get
      - Shopping history at Amazon – easy to get
      - Registry at Target – easy to get
      - Kaiser health record – harder
  - All enabled by ID

"The Unintended Consequences of Big Data"

- IEEE Computer 2013 – "Big Data's Big Unintended Consequences"
- Thinks of Big Data in a non-traditional fashion – one of the early topical papers that suggested
  - Vast amount of data collected in tons of domains, SETI, CERN, SKA, etc.
  - Corporations have seen Big Data as a tool for consumer marketing
  - Big Data has a political economy and is driven by corporations and governments not just looking inwards, but looking to agglomerate previously disconnected data

Another "unintended consequence"

- Data Silos
  - Bad, inefficient, no APIs, hard to interface with, costs more money for businesses and government
- YET
  - Beneficial if you value privacy considering that your footprint may be in MANY of those silos
  - Considered independently, it's fine
    - Your Yelp data for what food you like is fine by itself
    - Combine it with your MyFitnessPal data for your exercise routine, combined with your 2016 Taxes data on how much $$$ you make, and suddenly you have an AMAZING dataset to sell health insurers

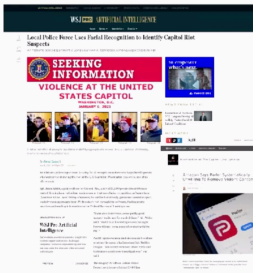
Data used against you

- Job applications
- Criminal records do not disappear and are not designed for corrective measure, but for permanent labeling
- Credit applications and credit worthiness

- The mantra is that there will be some "unintended consequences" that are worth having for the GREATER GOOD

US Capitol Insurrection – Jan 6, 2021

- Facial recognition from company's like ClearView AI



- Aided law enforcement along with social media posts from Parler to track down insurrectionists

Remainder of class

- Divide into groups of 5-6
- Pick three real world datasets
  - Characterize them in the sense of the five V's
    - How are they collected?
    - What is the temporal and spatial coverage
    - How are they structured?
  - State how the datasets are or are not Big Data
  - Consider the "unintended consequences" of these datasets
    - Can they be combined?
    - What can they be used for?
- Group presentations 5 slides (at least) and discussion

7min