About Data

Data Types
and Data Structures

Key Concepts & Definitions

- What is Data?
- What is Science?
- What is Data Science?
- What is Big Data Science/Analytics?

- Can't do **Big Data Science/Analytics** without knowing what these terms mean

Key Concepts Definitions

- What is **Data**?
  - "Measured" and "Collected" quantities and symbols intended to **represent** physical or intangible phenomena

Key Concepts Definitions

- What is **Science & Scientific Method**?
  - **Science**: systematic study of the structure and behaviour of the physical and natural world through observation and experiment (OED)
  - **Scientific Method**: systematic observation, measurement, and experiment, and the formulation, testing, and modification of hypotheses (OED)
  - **observations & measurements = data**
- **Assumption**: The fundamental premise of science is that the cosmos is understandable through use of the Scientific Method

Key Concepts Definitions

- What is **Data Science**?
  - the application of the Scientific Method to the definition, collection, representation, analysis, interpretation, and decision-making based on data
  - **data = evidence**

Key Concepts Definitions

- What is **Big Data Science/Analytics**?
  - The concepts, methods, and tools used to study datasets that are extreme in their size ("Volume"), rate of generation/acquisition ("Velocity"), or complexity ("Variety")
    - the traditional "Three Vs"…
    - …but there are more:
      - variability, veracity, validity, volatility, visualizability, value
  - "extreme" in the sense that current technologies are often *inadequate* for timely and efficient analysis

Data vs Information vs Knowledge

- **Data**: collection of symbols/measurements
  - {250, 300, 350, 400, …}
- **Information**: collection of symbols with <u>meaning</u>
  - {250ppm CO_2, 300ppm CO_2, 350ppm CO_2, 400ppm CO_2, …}
- **Interpretation -> <u>Knowledge</u>/Understanding:**
  - Atmospheric CO_2 is increasing…
    - …how?, why? action required? validation? replicatiion?

NOTE: “datum” (singular) is an element of a set of “data” (plural), but few authors refer to it that way in modern usage.

Basic Data Types

- **Nominal (Categorical)**
  - categories: qualitative, **no implied order or size**, *discrete*
    - color, gender, State, Country, words, …
- **Ordinal**
  - rank order: **no implied size metric**, *discrete*
    - 1: dislike < 2: neutral < 3: like
    - but like != 3 x dislike
- **Interval**
  - distance/difference measures have meaning, *continuous*
    - …but **no "zero" or origin**
    - degrees F or C, but 80° != 2 x 40°
- **Ratio**
  - **Size comparisons** have meaning, *continuous*
    - 80kg = 2 x 40kg
    - 0, differences, ratios have meaning
    - length, mass, time, distance, …
- **IMPORTANT**:
  - **Data type** determines what computations and statistical tests are appropriate or inappropriate!
  - e.g., can't calculate mean State of residence
- **Non-numeric data types**
  - media (images, video, audio)…often have above types for metadata

Data Structures

- Organized collections of various data types
- Examples
  - Arrays, Lists, records, tuples, DB records, matrices, …
  - Graphs, trees, hierarchies, …
- "Unstructured data"
  - no formal *metadata* model or schema
    - but may contain well-defined data types
  - freeform text, for example
    - e.g., email contents, social media postings, documents, …
  - IDC and EMC project that unstructured data will grow to 40 *zettabytes* by 2020, resulting in a 50-fold growth from the beginning of 2010. Computer World states that unstructured information might account for more than 70%–80% of all data in organizations.
  - See some cited references at:
    - https://en.wikipedia.org/wiki/Unstructured_data

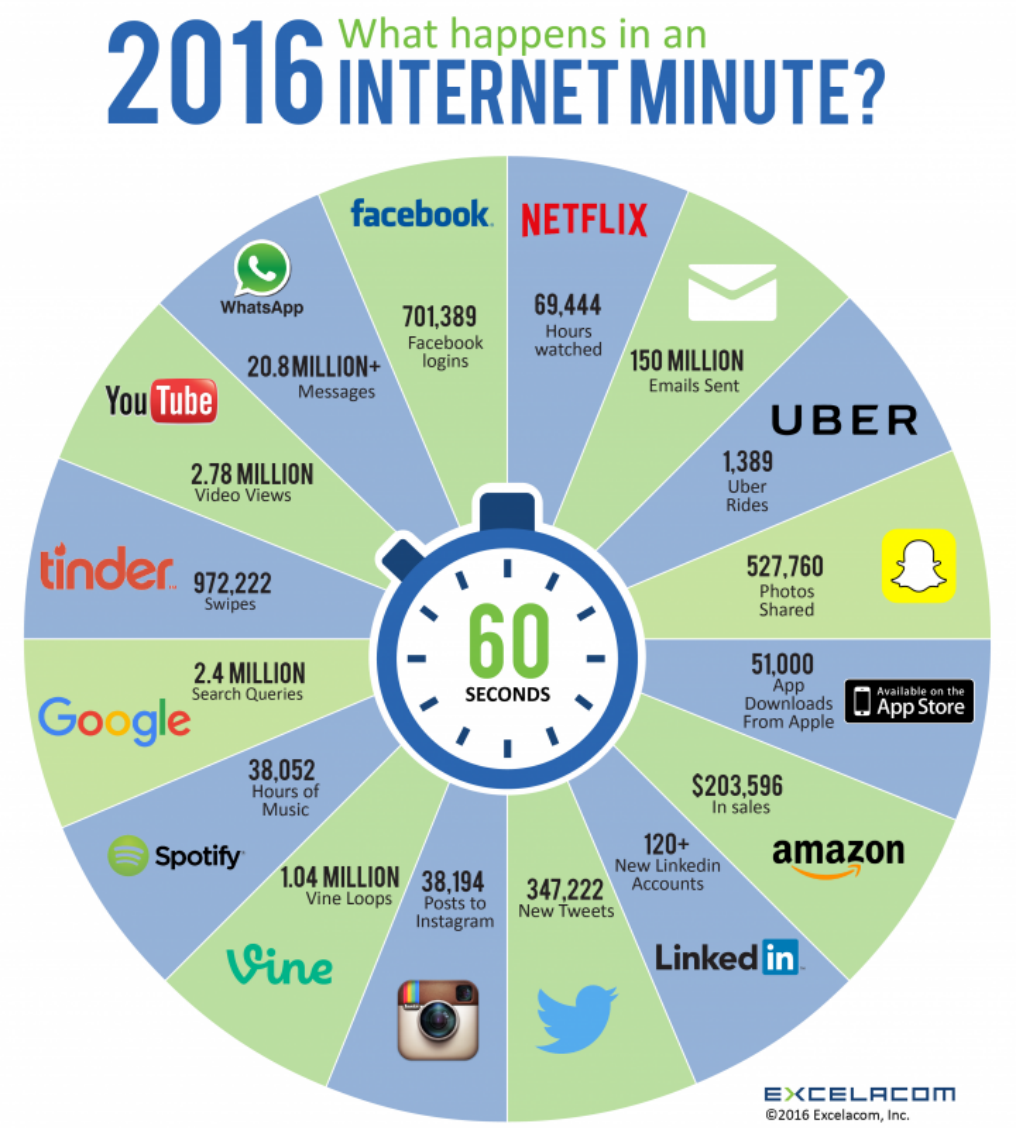
Some Data References

- Babbie, Earl, *The Practice of Social Research 14th Edition*, 2015, Wadsworth Publishing
  - https://www.amazon.com/Practice-Social-Research-Earl-Babbie/dp/1305104943/
- Elliott, A., and Woodward, W., *IBM SPSS by Example: A Practical Guide to Statistical Data Analysis 2nd Edition*, 2015 SAGE Publications
  - https://www.amazon.com/IBM-SPSS-Example-Practical-Statistical/dp/1483319032/
- Andrews, F., et al., *A Guide for Selecting Statistical Techniques for Analyzing Social Science Data 2nd Edition*, 1981, Survey Research Center, University of Michigan

"Big Data" *Human Perception* Issues

- Humans can't internalize ("get") size and time parameters
  - Size
  - Flow / Throughput
  - Reaction time / Latency…

Some Sources of Big Data



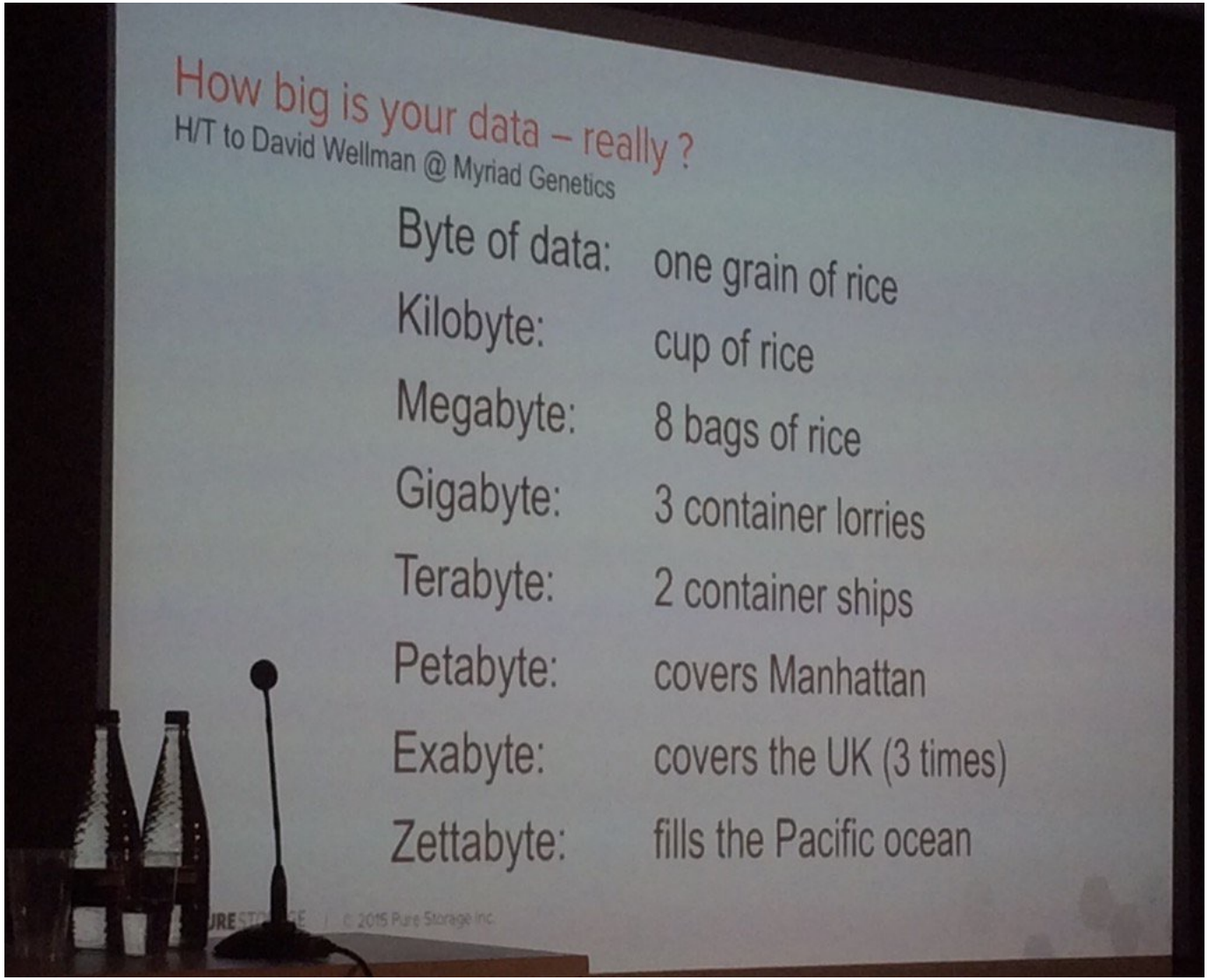
Understanding size

Multiples of bytes v·t·e

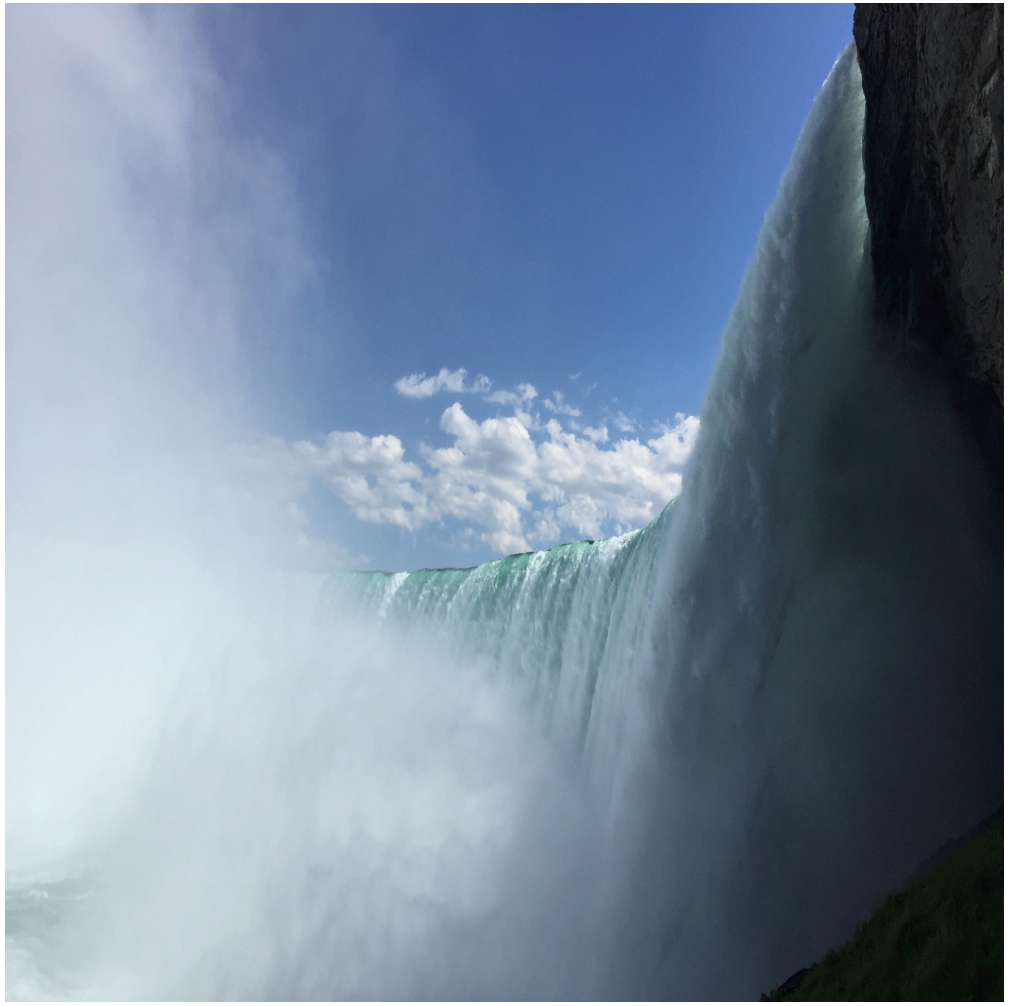
Decimal			Binary				
Value	Metric		Value	IEC		JEDEC	
1000	kB	kilobyte	1024	KiB	kibibyte	KB	kilobyte
1000^2	MB	megabyte	1024^2	MiB	mebibyte	MB	megabyte
1000^3	GB	gigabyte	1024^3	GiB	gibibyte	GB	gigabyte
1000^4	TB	terabyte	1024^4	TiB	tebibyte		–
1000^5	PB	petabyte	1024^5	PiB	pebibyte		–
1000^6	EB	**exabyte**	1024^6	EiB	exbibyte		–
1000^7	ZB	zettabyte	1024^7	ZiB	zebibyte		–
1000^8	YB	yottabyte	1024^8	YiB	yobibyte		–

Orders of magnitude of data

Understanding Size



Understanding flow (throughput): 1M gps



Understanding Latency

Level	Latency	Seconds	Time To Insight		
			Seconds	Hours	Days
NH L3 Cache	52 Clocks	1.62×10^{-8}	**1**		
Local Memory	150 ns	1.5×10^{-7}	9.23		
32 Core NUMA	3 Hops	5.19×10^{-7}	31		
SSD	125 us	1.25×10^{-4}	7,692	2.14	
Gb Ethernet	1 ms	1×10^{-3}	61,538	17	.71
Disk	3.6 ms	3.6×10^{-3}	**221,538**	**62**	**2.56**

- **In 30 seconds, you can block someone.**
- **In 2 hours, you can intercept someone**
- **In 2 days, actionable intelligence is forfeited**

Big Data Analytics Technical Issues

- **Ingest Rates** (**V**elocity)
  - Simultaneous ingest and query cause conflicts and delays in the system
- **Data Retention** (**V**olume)
  - Insufficient periods of time for identification of *trends* in data
  - Requirement for multi-year storage and *retention* = multi-PB+
- **Multiple Data Formats** (**V**ariety)
  - need to transform into *queryable* information
- **Poor Query Performance**
  - Hours-Days for results
    - reducing/limiting number of queries that can be executed
  - Challenge for the analyst to *visualize* the problem and refine query *dynamically*
  - Queries performed on separate computing platforms due to poor performance
    - Need networked *hierarchy* of systems to conduct queries
- We attempt to address these issues with *fundamental concepts of distributed computing*:
  - *Representation*
  - *Parallelism*
  - *Caching / Locality*
  - *Service component speed*