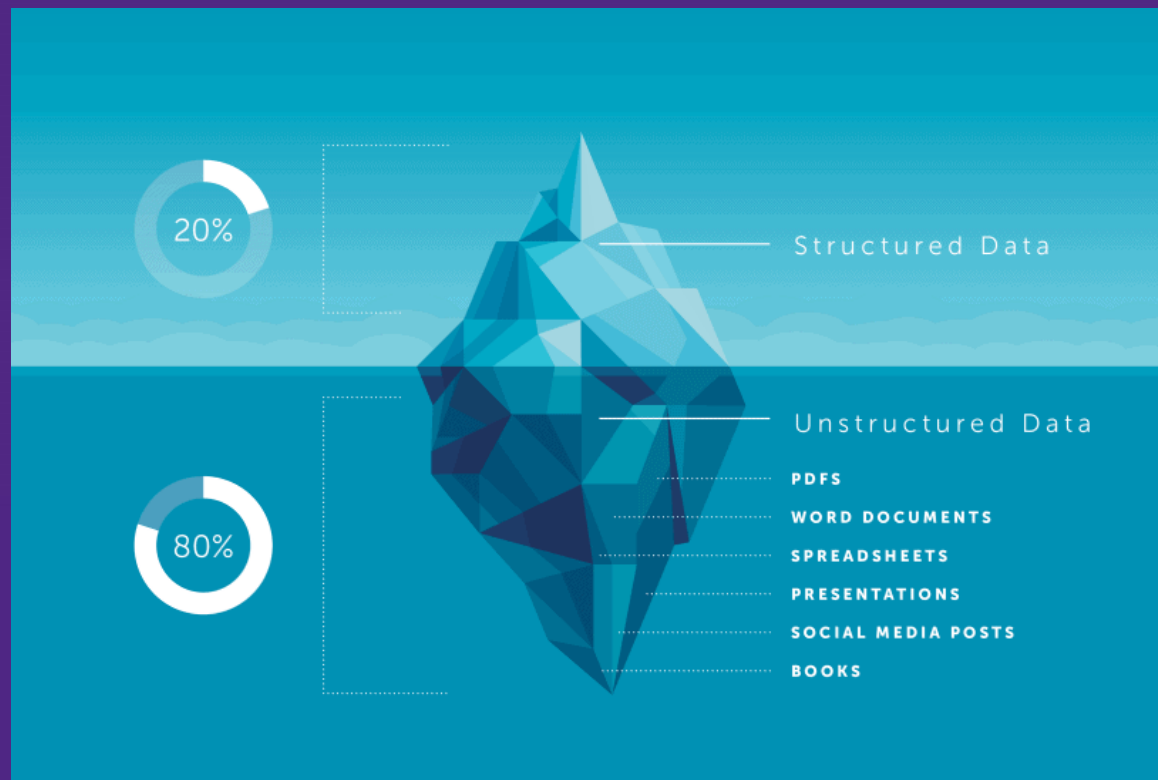


CS4417 / CS9647 / CS 9117

Unstructured Data



Course Content

- Analysis and Discovery (approx. ½)
 - Representing text and documents
 - Analyzing and discovering structure
- Systems and Applications (other ½)
 - Systems for large-scale unstructured data storage and processing
 - Contemporary software and applications

Tentative Topics

- Structure in Unstructured Data
- Text preprocessing and representation
- Document representation and retrieval
- Corpus structure
- Language structure
- Neural net language models, GPT
- Distributed programming models, MapReduce
- Systems and applications of distributed models
- Hadoop
- NoSQL, MongoDB

Logistics

- Lectures in NCB-113
Tuesday 3:30pm to 4:20pm
- Lectures in SEB-1200
Thursday 2:30pm to 4:20pm
- Instructor: Dr. Arshin Rezazadeh
- TAs: Chris Steward, Caro Strickland, Maxwell Yin
- Guest Lectures offered by Profs. from the CS Dept.

Communication

- *Using only OWL* for electronic communication
- OWL forums for questions and discussions
 - Can post anonymously if you want – instructor and TAs only can see poster's name
- OWL messaging to contact the instructional team directly
 - “Instructor” – Arshin
 - “Secondary Instructor” – Chris, Caro and Maxwell
- OWL will be monitored Monday – Friday, 9am to 5pm
- *Check the forums first*

Preparation

- I may post required readings and/or videos ahead of class.
- I will give at least one week's lead time; will send an OWL announcement.
- You are responsible for the assigned content, and it might be asked on the midterm/exam.

Evaluation

- 3 Assignments
 - 45% for undergrads, 35% for graduate students
- Midterm (Focus on first half; location TBA)
 - 20%
- Final (Focus on second half, but cumulative)
 - 25%
- Class participation (iClicker)
 - 10%
- Graduates only: Technical Topic Report
 - 10%

Technical Topic Report

- CS9647 and CS9117 Only
 - Individual, 2 Page report on a technology not covered in the course
 - *Check OWL assignment for full specification*
- Introduction
 - Overview of what the technology is and how it is applied
- Capabilities and Mechanisms
 - Describe what the technology can do and the techniques it uses
- Related Methods
 - Give a history of the development of the technology, and identify related technologies and competitors
- Opportunities
 - Speculate on how this technology could either evolve into new technologies or be applied in new ways in the future.

Questions?

- About topics?
- About logistics?
- About evaluation?

What is *Structured* Data?

- For this course, we are considering “structured” data to be things like
 - Relational databases
 - (Good) excel spreadsheets
 - (Closed-ended) survey data

What is Unstructured Data?

- Working definition: Unstructured data (or unstructured information) is information that does not have a pre-defined data model or is *not organized in a pre-defined manner*.
- Data may have *latent, implicit, or unknown* structure; natural language has *lots* of structure, but it does not have a complete formal description

Tasks Using Unstructured Data

- Information retrieval
 - Find documents relevant to a query
- Labeling
 - Spam/not spam
 - Sentiment analysis
 - Semantic association, annotating unstructured data with new information
- Structure-finding
 - Discover common topics
 - Discover networks of collaborators

Data Representations

- Often, a collection of data can be divided into ‘elements’
 - In statistics, sometimes called ‘units of analysis’
 - E.g., html page, photo, invoice, social media post
- Representations allow us to relate elements of unstructured data to each other.
- We will take a detailed look at how language is represented, and identify parallels to representing other kinds of data.

Example: Text Representations

- Bytes
 - Characters
 - Words
- Sentences
- Documents
 - We can structure these by tokenization, tagging (part of speech, emotional valence), others.
 - Vector representations (traditional and machine-learning derived)

Image Representations

— Bytes

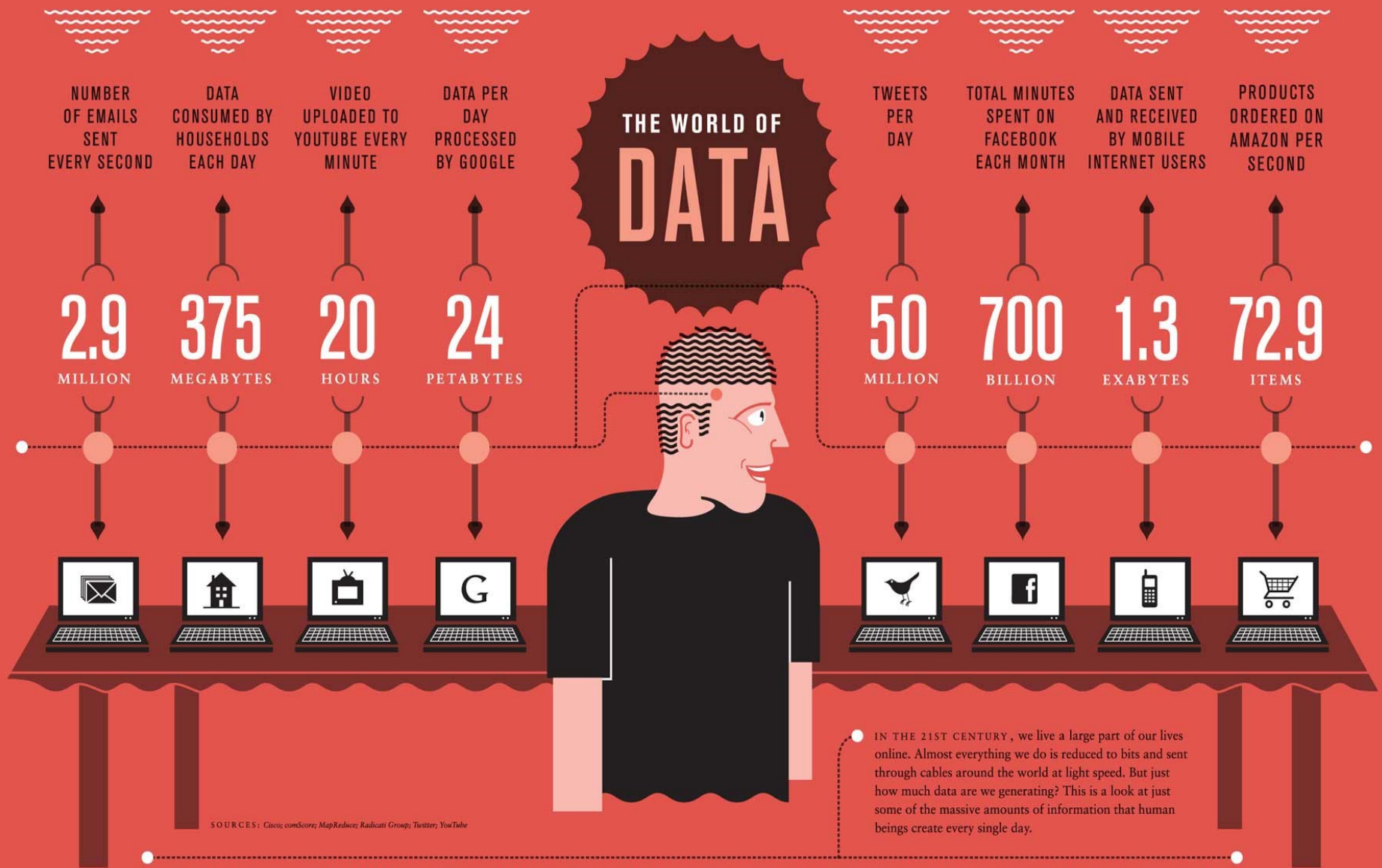
- Pixels

Images

- “Word analogs” like SIFT, SURFs. Invariant(ish) with respect to viewpoint.
- Compression/transformation-based representations
 - Unsupervised: PCA, manifold learning, autoencoders
 - Supervised: Classification, regression

Unstructured Data are often BIG

- Volume (*size*)
- Variety (*representation*)
- Velocity (*stored vs. streaming*)
- Veracity
 - Accuracy and precision
 - Errors, completeness, and integrity
- Validity
 - Data governance and management



A COLLABORATION BETWEEN GOOD AND OLIVER MUNDAY

IN PARTNERSHIP WITH **IBM**

As of 2011, the global size of data in healthcare was estimated to be

150 EXABYTES

[161 BILLION GIGABYTES]



By 2020, it's anticipated there will be

**420 MILLION
WEARABLE, WIRELESS
HEALTH MONITORS**



Variety

DIFFERENT FORMS OF DATA

**4 BILLION+
HOURS OF VIDEO**

are watched on
YouTube each month



**30 BILLION
PIECES OF CONTENT**

are shared on Facebook
every month



400 MILLION TWEETS

are sent per day by about 200
million monthly active users



Big Data Challenges

- Storage
- Computation
- Both of these ***must*** be distributed.

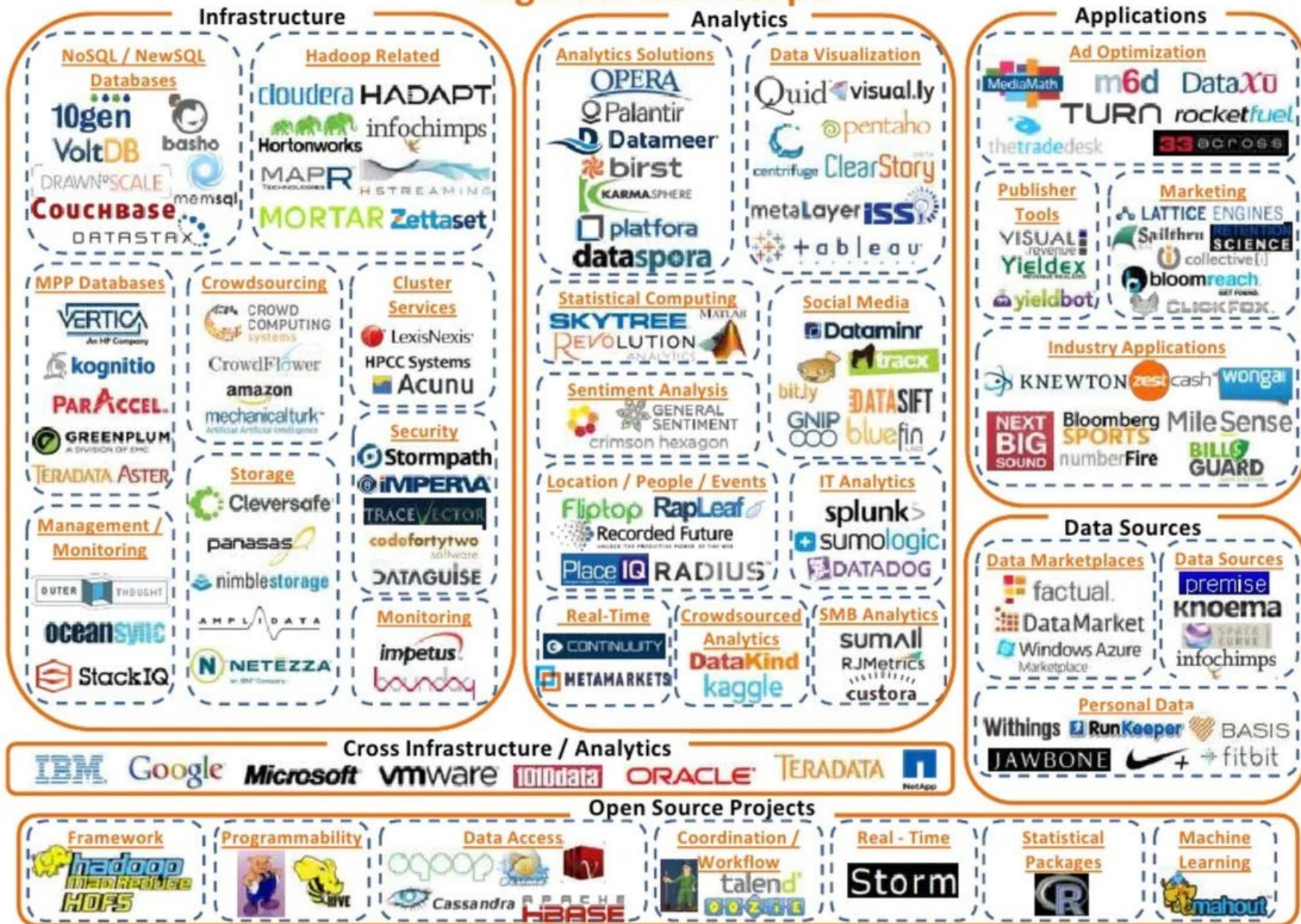
Big Data Solutions

- Storage
 - E.g., Hadoop Distributed File System, relies on replication, sharding
- Computation
 - E.g., MapReduce programming model, implemented by e.g., MongoDB, Hadoop, ...

Big Data Analytics

- Big data analytics platforms package up solutions to storage and computation challenges
- Can be hosted locally, or deployed to a cloud platform
- Cloud is increasingly common – We can watch the world grind to a halt when AWS goes down.
 - (Ofc the “cloud” is just somebody else’s computer(s).)

Big Data Landscape



© Matt Turck (@mattturck) and ShivonZilis (@shivonz)

Big Data Platforms - Features

- Databases for unstructured data e.g., MongoDB
- Services that provide machine learning
- Systems that provide storage for massive amounts of data, lots of processing power, and programming support for running multiple tasks
- Data integration from multiple sources
- Tools for specific types of analytics, e.g., customer profiles

Adapting to Big and Unstructured Data

- Enterprises are looking to a new generation of databases referred to as NoSQL
- Document-centric
- Flexible schema
- Often, no traditional relations (one-to-one, one-to-many, many-to-many; more like none-to-none)

Summary

- Unstructured data lacks *a priori* structure spec
- Key challenges
 - Representation
 - How do we represent unstructured data to accomplish the tasks we need to?
 - Storage and computation
 - How do we support the development and use of representations on big datasets?