



EECE5155: Wireless Sensor Networks and the Internet of Things

Josep Miquel Jornet

Associate Professor, Department of Electrical and Computer Engineering

Director, Ultrabroadband Nanonetworking Laboratory

Member, Institute for the Wireless Internet of Things

Northeastern University

jmjornet@northeastern.edu

www.unlab.tech

Module T7: Data Analytics

scitvlylA datA: T7 eluBOM

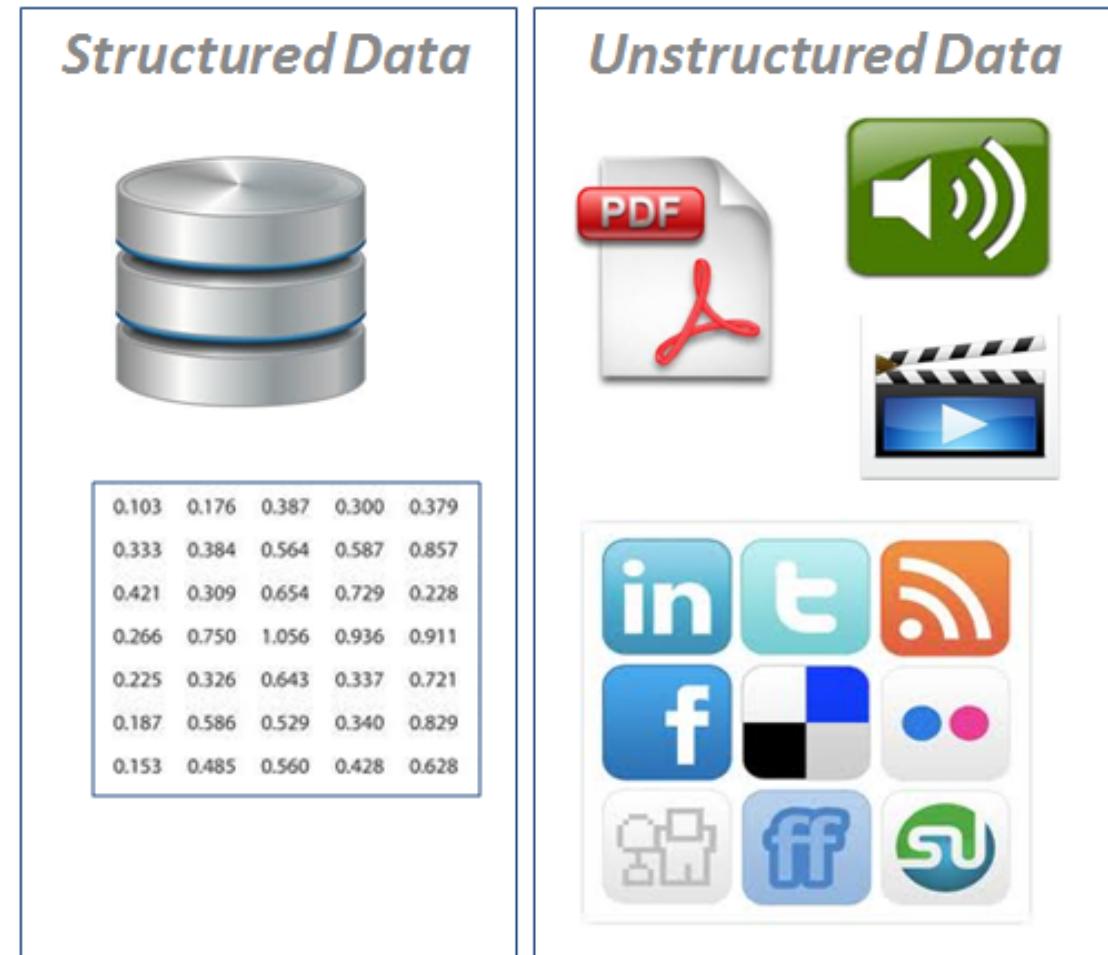
IoT = Data

Data = \$\$\$

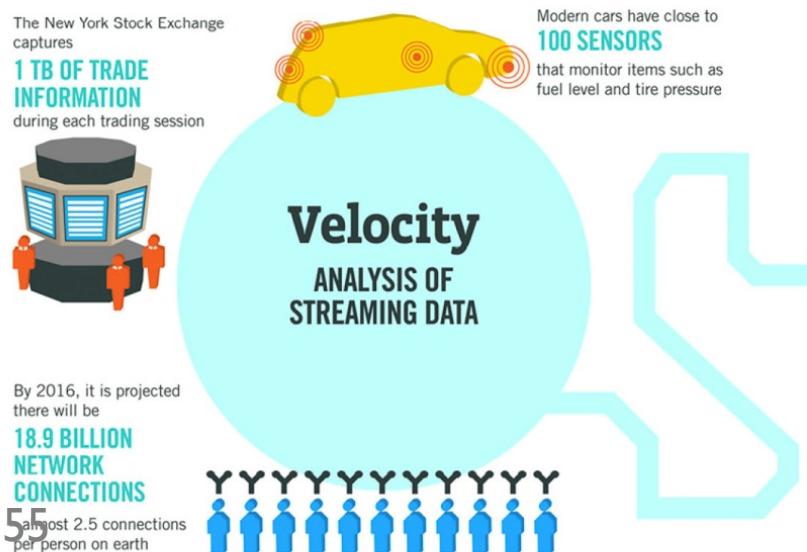
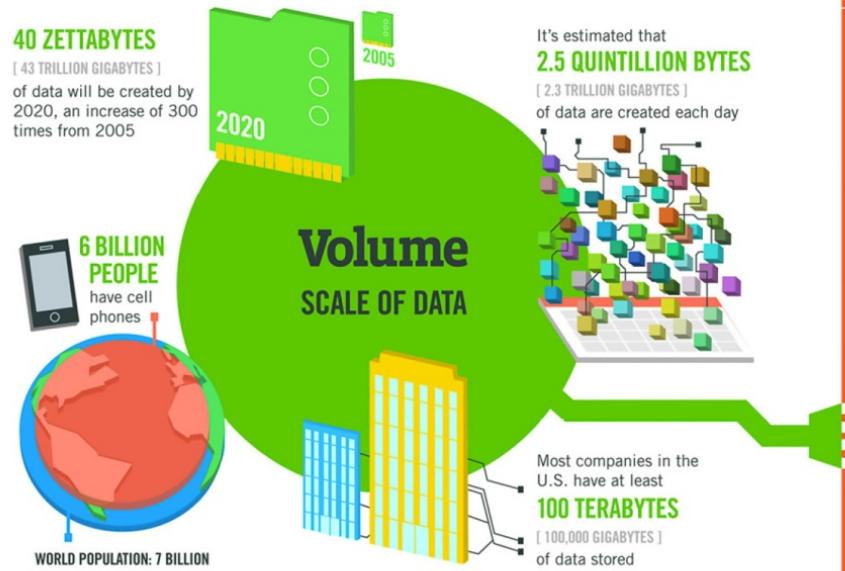
IoT = \$\$\$

Types of Data

- Based on data format:
 - **Structured:** has a pre-defined data model/schema/structure, relational in nature → Traditional databases
 - **Unstructured Data:** does not have a well-defined data model, the model is not relational → word documents, audio files, video files, tweets, ...



Big Data



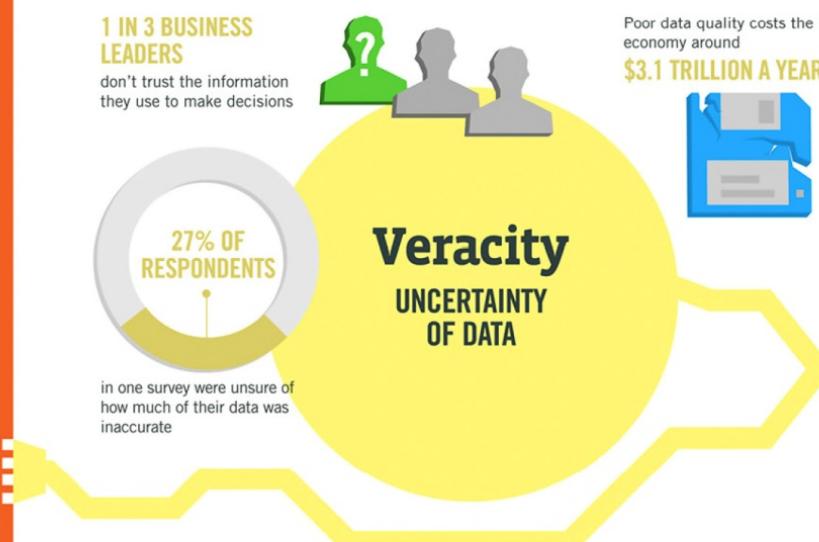
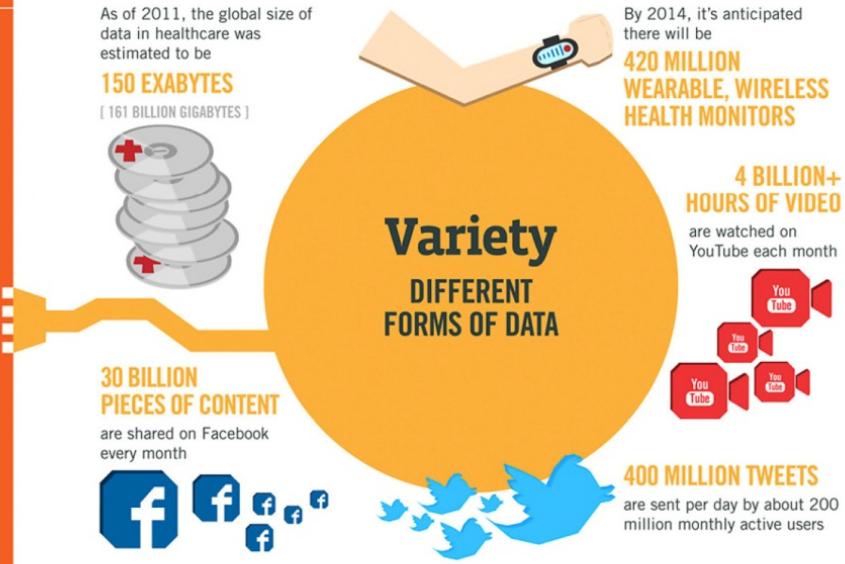
The FOUR V's of Big Data

From traffic patterns and music downloads to web history and medical records, data is recorded, stored, and analyzed to enable the technology and services that the world relies on every day. But what exactly is big data, and how can these massive amounts of data be used?

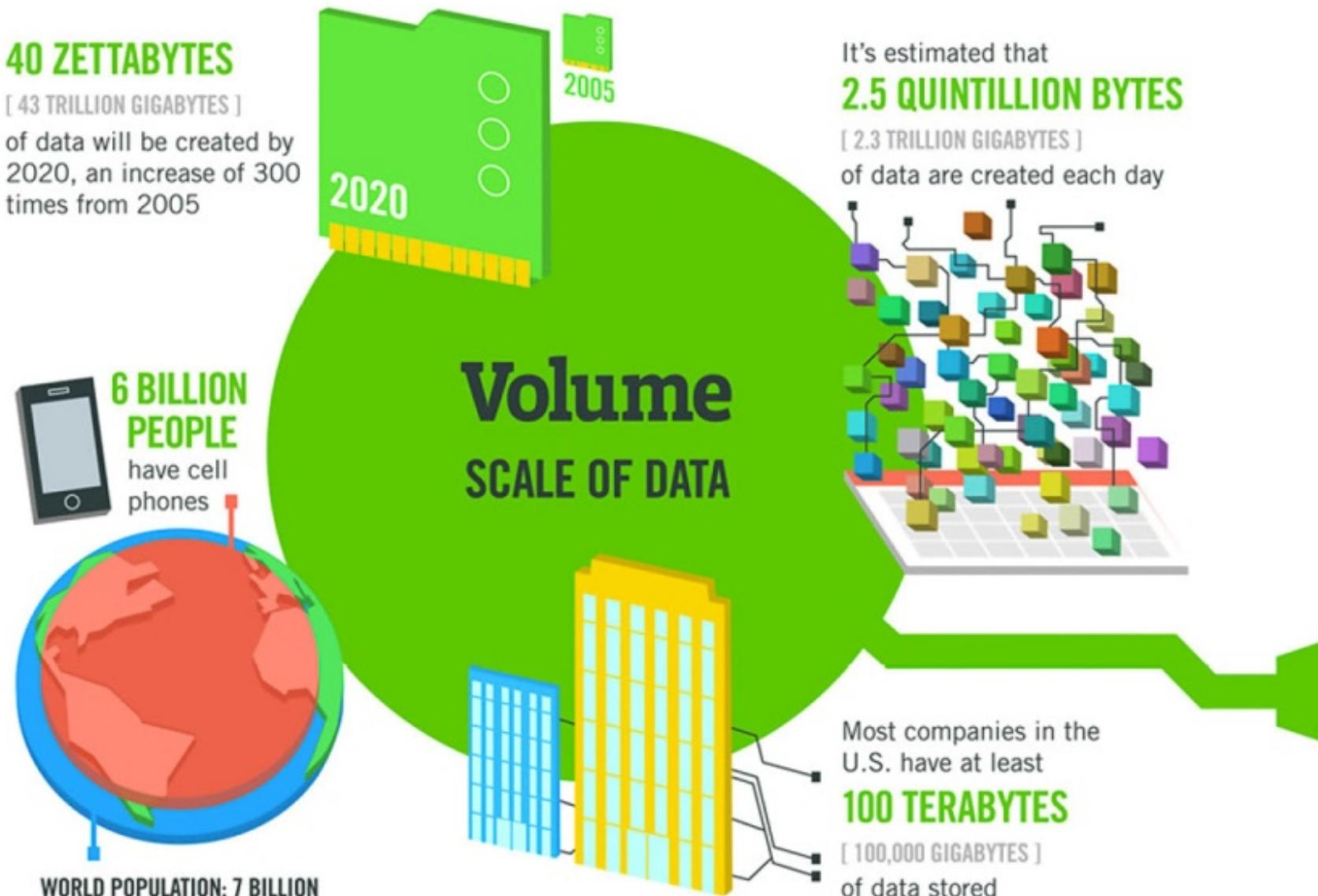
As a leader in the sector, IBM data scientists break big data into four dimensions: Volume, Velocity, Variety and Veracity.

Depending on the industry and organization, big data encompasses information from multiple internal and external sources such as transactions, social media, enterprise content, sensors and mobile devices. Companies can leverage data to adapt their products and services to better meet customer needs, optimize operations and infrastructure, and find new sources of revenue.

By 2015 4.4 MILLION IT JOBS will be created globally to support big data, with 1.9 million in the United States.



The Four V's of Big Data

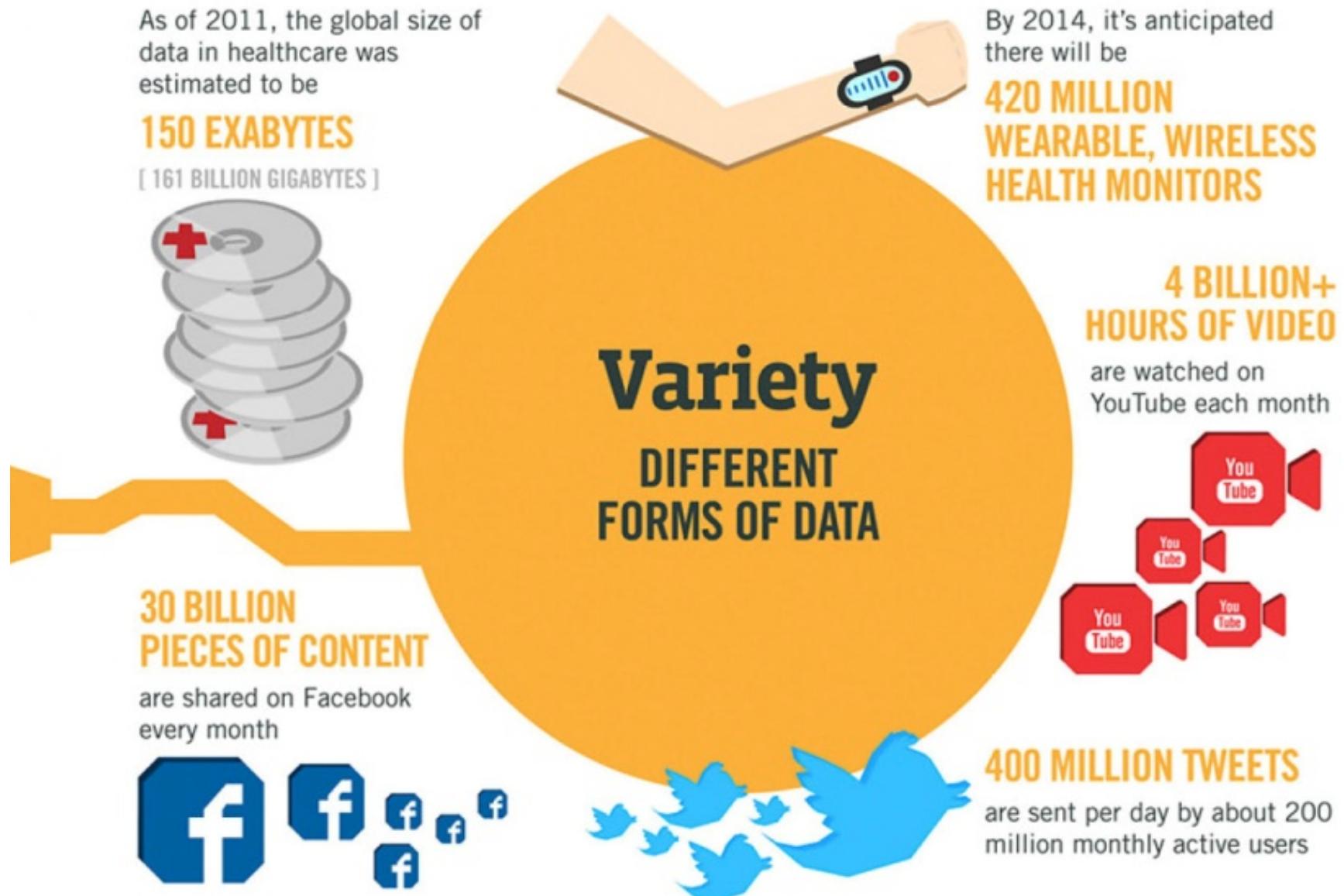


Think for a second...

- How much data (files, images, music, videos, ... → Gb) do you currently store:
 - Locally in your computer?
 - Locally in your phone?
 - Locally in your tablet?
 - Remotely in *a* cloud?
 - OneDrive
 - Google Drive
 - Dropbox
 - ...

And you are just one individual...

The Four V's of Big Data



Think for a second...

- How many times do you share your status on:
 - Instant messaging platforms?
 - Social networks?
 - Instagram
 - Twitter
 - Facebook
 - Pinterest
 - ...

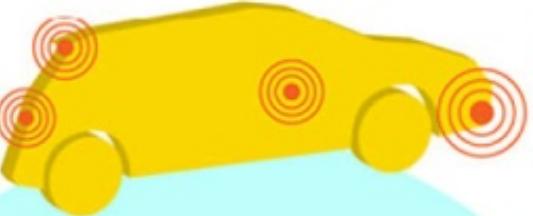
And you are just one individual...

The Four V's of Big Data

The New York Stock Exchange
captures

**1 TB OF TRADE
INFORMATION**

during each trading session



Modern cars have close to
100 SENSORS
that monitor items such as
fuel level and tire pressure

Velocity
ANALYSIS OF
STREAMING DATA

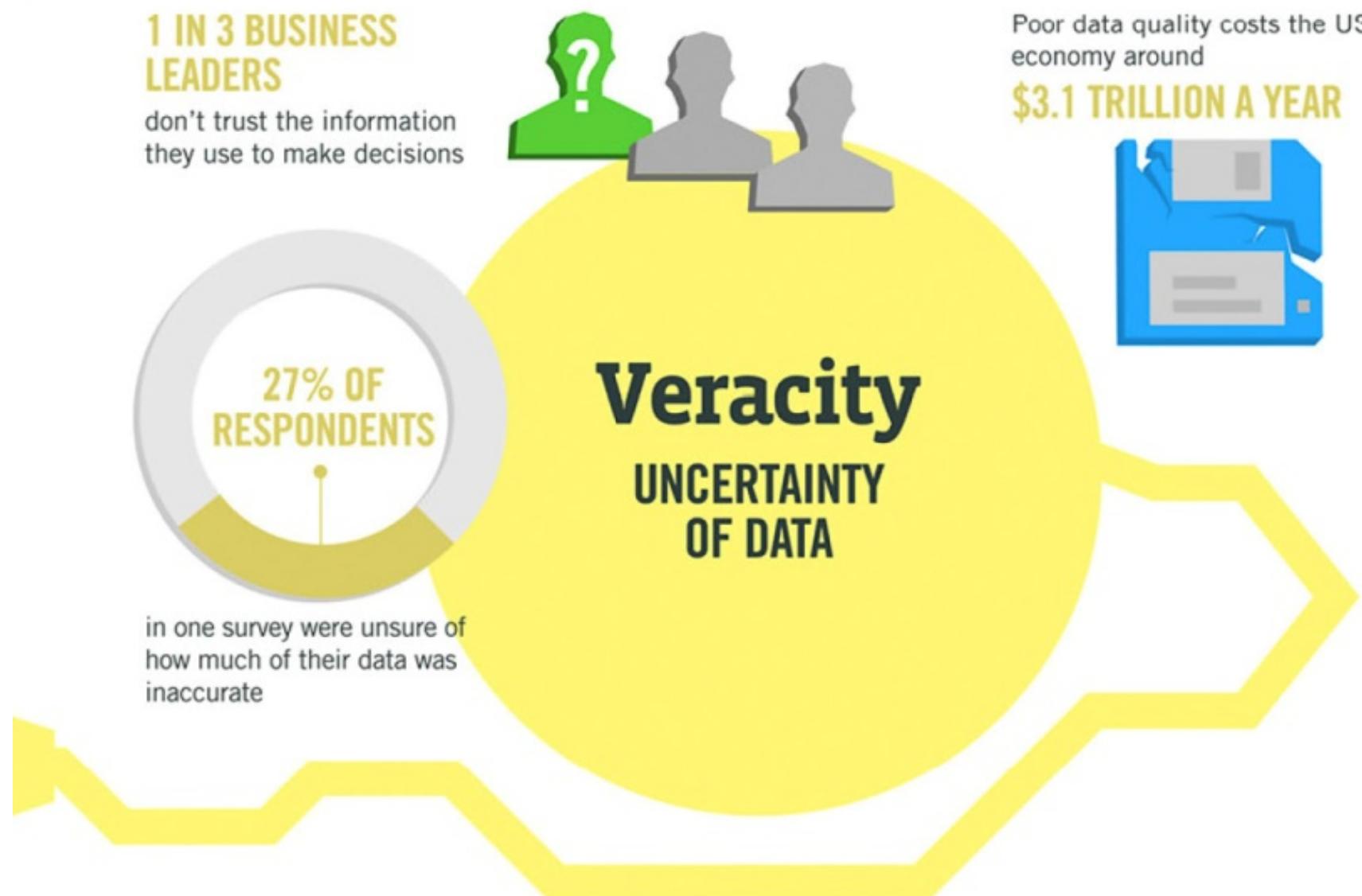
By 2016, it is projected
there will be

**18.9 BILLION
NETWORK
CONNECTIONS**

– almost 2.5 connections
per person on earth



The Four V's of Big Data



The Promise of Big Data

- Data contains information of great business value
 - If you can extract those insights, you can make better decisions!

nature International weekly journal of science

Home | News & Comment | Research | Careers & Jobs | Current Issue | Archive | Audio & Video | For Authors

News & Comment > News > 2017 > November > Article

NATURE | NEWS

Drug data reveal sneaky side effects

Mining of surveillance data highlights thousands of previously unknown consequences when drugs are taken together.

Heidi Ledford

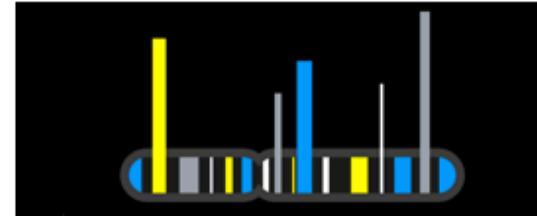
14 March 2012

Rights & Permissions

An algorithm designed by US scientists to trawl through a plethora of drug interactions has yielded thousands of previously unknown side effects caused by taking drugs in combination.

The work, published today in *Science Translational Medicine*¹, provides a way to sort through the hundreds of thousands of 'adverse events' reported to the US Food and Drug Administration (FDA) each year. "It's a step in the direction of a complete catalogue of drug–drug interactions," says the study's lead author, Russ Altman, a bioengineer at Stanford University in California.

Gene count



The most popular genes in the human genome

A tour through the most studied genes in biology reveals some surprises.

nature briefing

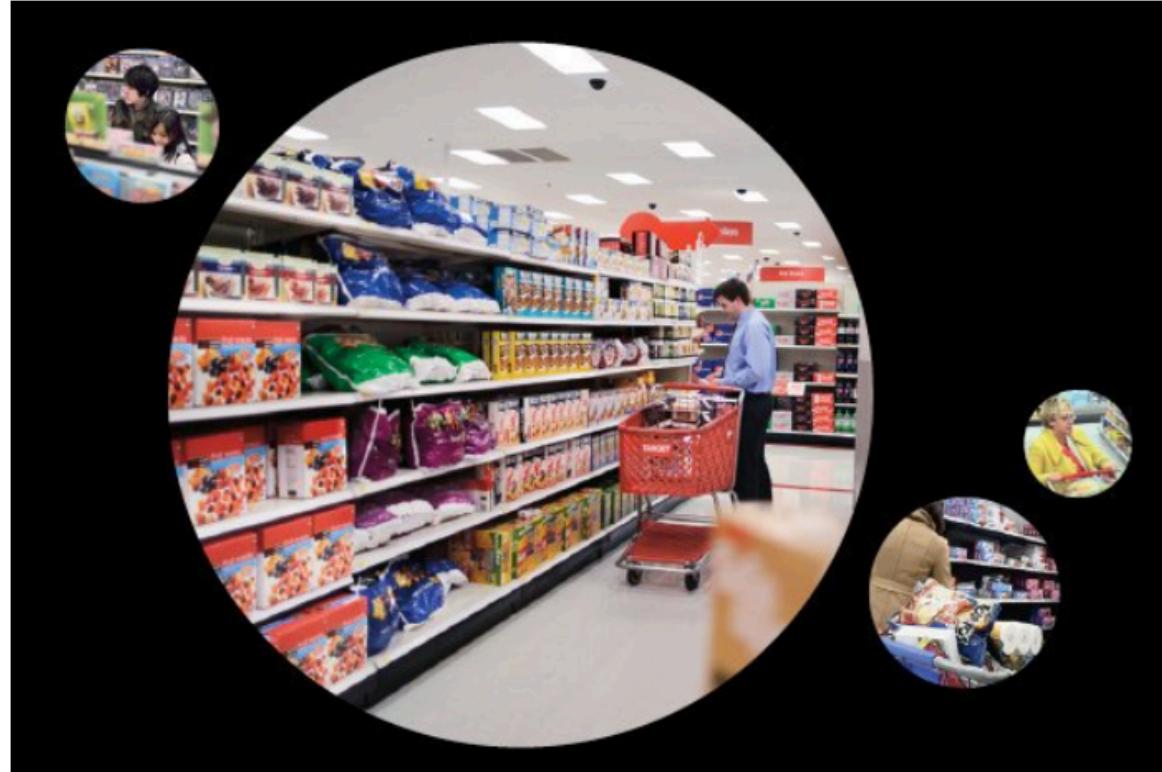


The best science news from Nature and beyond

Magazine

How Companies Learn Your Secrets

By CHARLES DUHIGG FEB. 16, 2012



Antonio Bolfo/Reportage for The New York Times

RELATED COVERAGE



Opinion | David Brooks
The Machiavellian Temptation MARCH 1, 2012

RECENT COMMENTS

Joanne February 22, 2012
Target, just because you can, doesn't mean you should.

Additional Examples

- **Sports**
 - Basketball increasingly driven by data analytics
 - Soccer beginning to follow
- **Entertainment**
 - Some series designed based on data analysis (e.g., House of Cards)
- ...

Data Analytics

- **Definition:** Process of
 - Cleansing,
 - Transforming, and
 - Modeling datawith the goal of
 - discovering useful information,
 - suggesting conclusions, and
 - supporting decision-making
- **Not only for IoT:** Data analysis has multiple facets and approaches, encompassing diverse techniques under a variety of names, in different business, science, and social science domains

Steps in Data Analytics

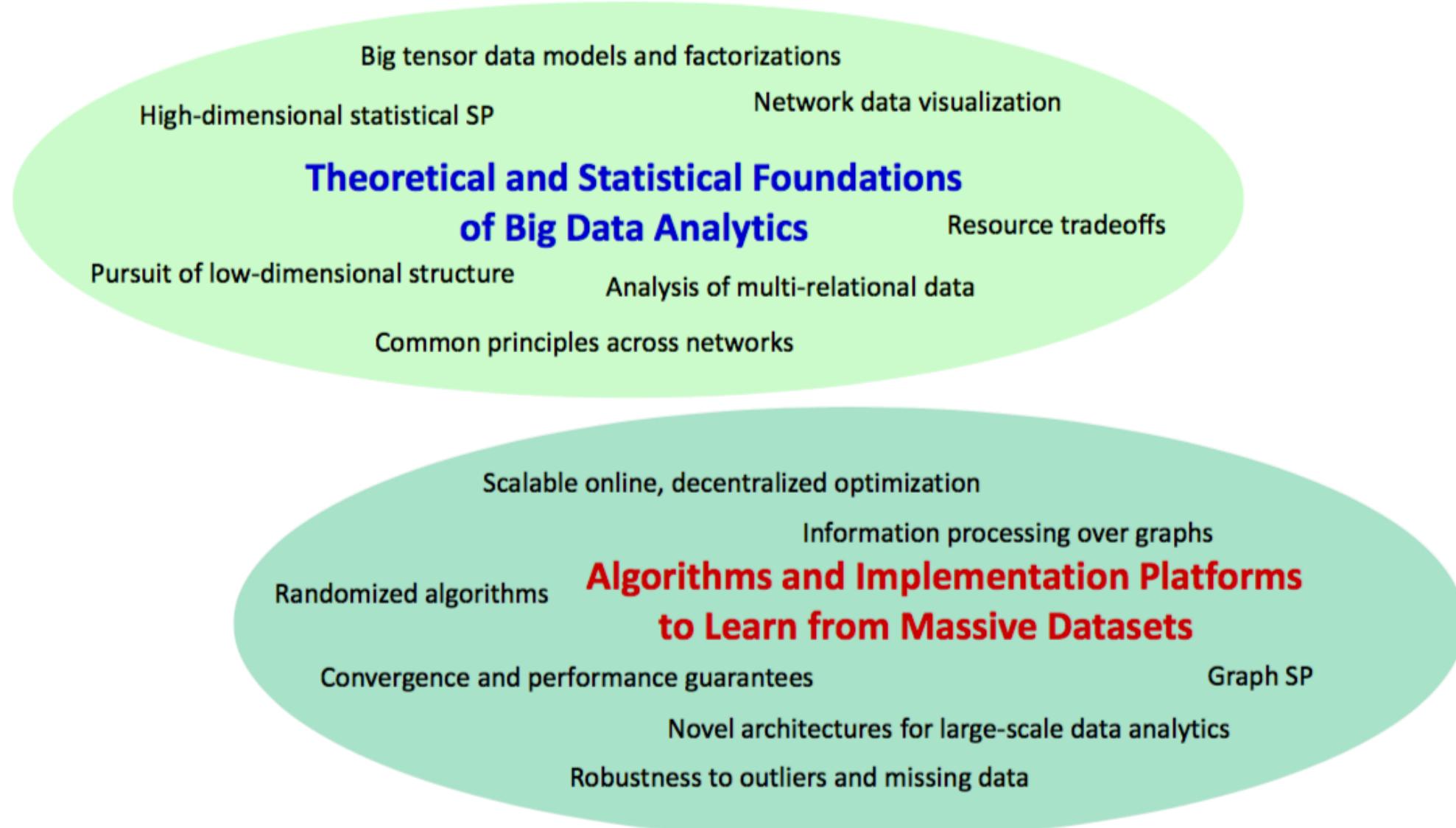


- 1. Collection:** Structured, unstructured and semi-structured data from multiple sources
- 2. Ingestion:** Loading vast amounts of data onto a single data store
- 3. Discovery & Cleansing:** Understanding format and content; clean up and formatting
- 4. Integration:** Linking, entity extraction, entity resolution, indexing and data fusion
- 5. Analysis:** Intelligence, statistics, predictive and text analytics, machine learning
- 6. Delivery:** Querying, visualization, real time delivery on enterprise-class availability

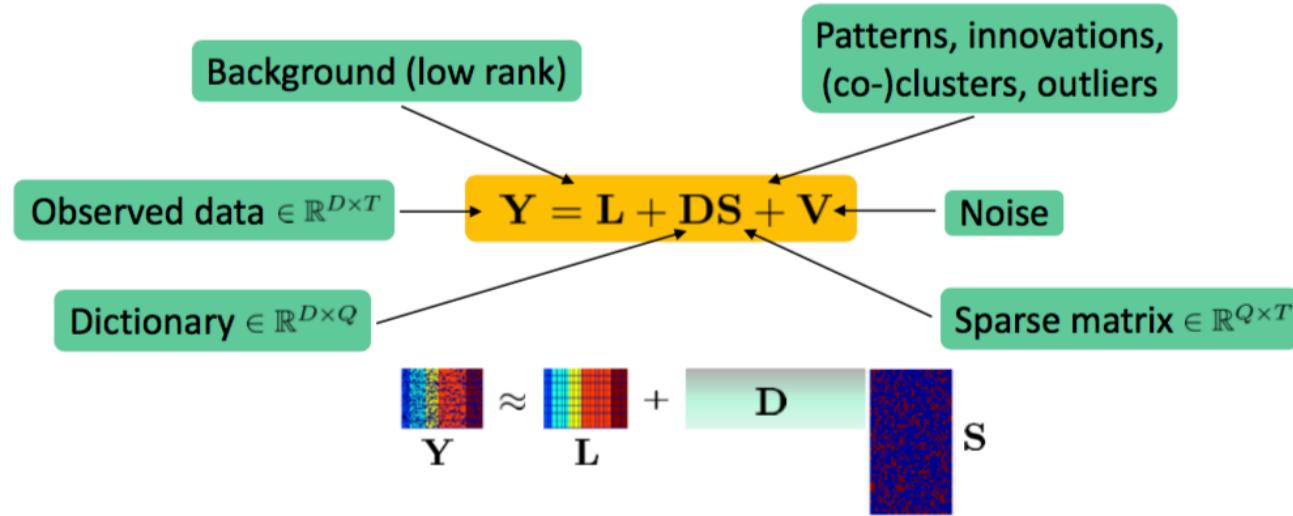
Challenges in Data Analytics

- Data collection from
 - **Multitude** of sensors, with different firmware versions
 - **Bad data** from damaged sensors
- Huge volume of data requires
 - **Decentralized** and **parallel** processing
 - **Security** and **privacy** measures
- Modern massive datasets involve many attributes
 - **Parsimonious** models to ease interpretability
 - Enhanced **predictive** performance
- Real-time streaming data requires
 - **Online** processing
 - Quick-rough answer VS slow-accurate answer?

Research Opportunities



Mathematical Models, Algorithms



Types of algorithms

Clustering
Association learning
Parameter estimation
Recommendation engines

Classification

Similarity matching
Neural networks
Bayesian networks
Genetic algorithms

Big Data Tools

ORACLE®



mongoDB®



+ a b l e a u®



cloudera®
Ask Bigger Questions



Oracle SQL

1. What type of data does it support?

SQL: Structured Query Language

2. What functions/steps of data analytics does it support?

1. Ingestion
2. Cleansing
3. Integration
4. Analysis
5. (Delivery)

3. What algorithms does it implement?

- Oracle Machine Learning
- Oracle Data Miner

4. Is it compatible with common cloud platforms?

- AWS, Azure,...

5. Open-source or private?

- Private



Hadoop

- 1. What type of data does it support?**
 - Unstructured data (but works with structured too)
- 2. What functions/steps of data analytics does it support?**
 - Distributed data storage and processing, i.e.,
 1. Ingestion
 2. Cleansing
 3. Integration
 4. Analysis
 5. (Delivery)
- 3. What algorithms does it implement?**
 - MapReduce, can handle other user-defined algorithms
- 4. Is it compatible with common cloud platforms?**
 - AWS, Azure,..
- 5. Open-source or private?**
 - Open source





1. What type of data does it support?

- Unstructured and structured data

2. What functions/steps of data analytics does it support?

1. Ingestion
2. Cleansing
3. Integration
4. Analysis
5. (Delivery)

3. What algorithms does it implement?

- Dimensionality reduction, machine learning

4. Is it compatible with common cloud platforms?

- AWS, Azure, (on top Hadoop)

5. Open-source or private?

- Open source

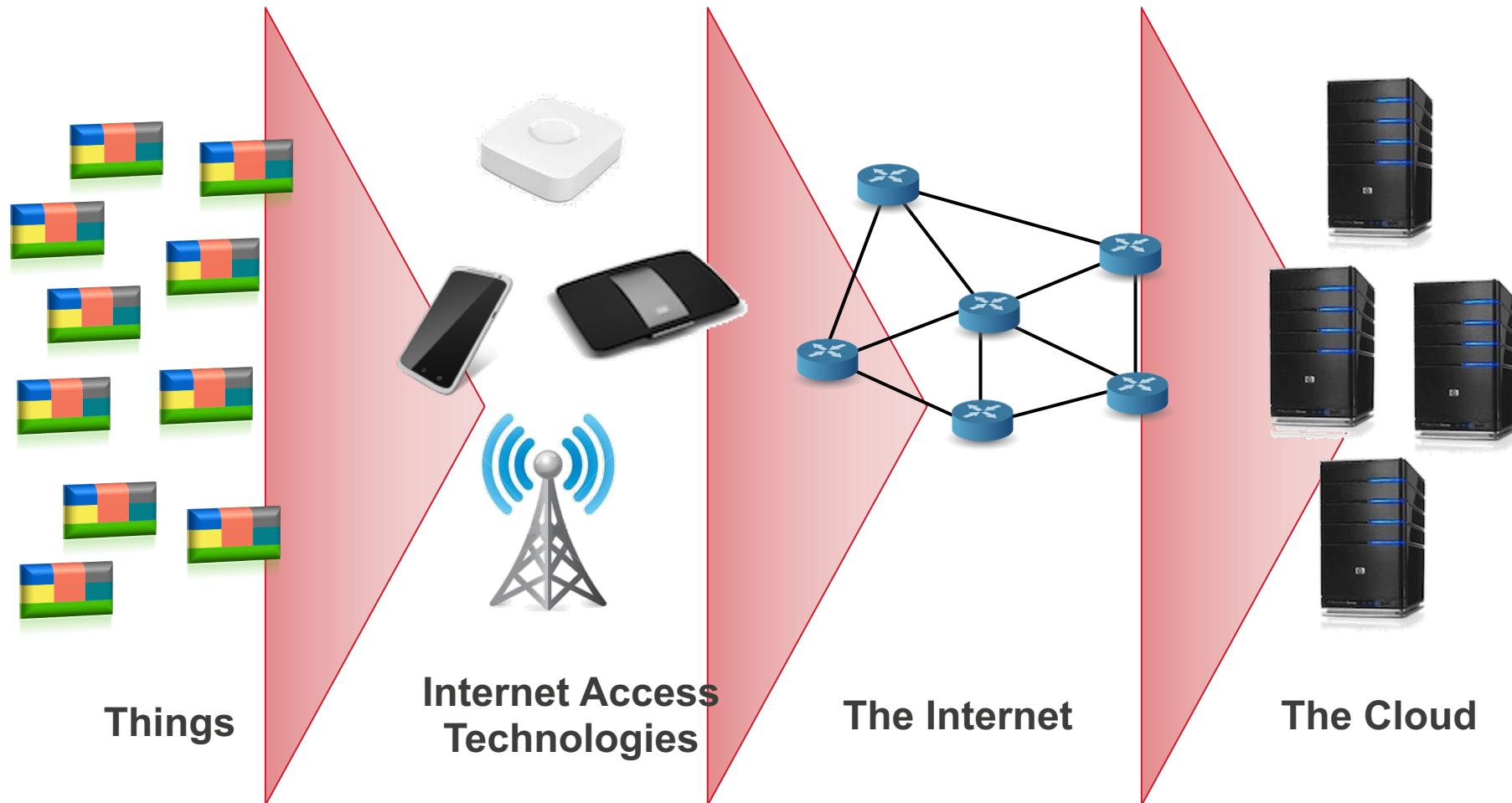
Need for Standardization

- Big Data use cases, definitions, vocabulary and reference architectures (e.g., system, data, platforms, online/offline)
- Specifications and standardization of metadata including data provenance
- Query languages including non-relational queries to support diverse data types and Big Data operations
- General and domain specific ontologies and taxonomies for describing data semantics including interoperation between ontologies
- Big Data security and privacy access controls
- Remote, distributed, and federated analytics (taking the analytics to the data) including data and processing resource discovery and data mining
- Human consumption of the results of big data analysis (e.g., visualization)
- Interface between relational (SQL) and non-relational (NoSQL)
- Big Data Quality and Veracity description and management
- ...

Course Contents

- **Module T1:** Introduction to the Internet of Things ✓
- **Module T2:** Data Acquisition ✓
- **Module T3:** Local Data Processing ✓
- **Module T4:** Data Communication ✓
- **Module T5:** Data Streaming ✓
- **Module T6:** Data Storage & Cloud ✓
- **Module T7:** Data Analytics ✓

The Internet of Things



Next Steps

- **Research paper video presentation:**
 - Due on December 13
- **Homework assignment:**
 - Due on December 13
- **Third laboratory assignment:**
 - Optional, due on December 15