Indian Institute of Science
Bangalore, India
भारतीय विज्ञान संस्थान
बंगलौर, भारत

Department of Computational and Data Sciences

SE256:Jan16 (2:1)

L1:Introduction
Scalable Systems for Data Science

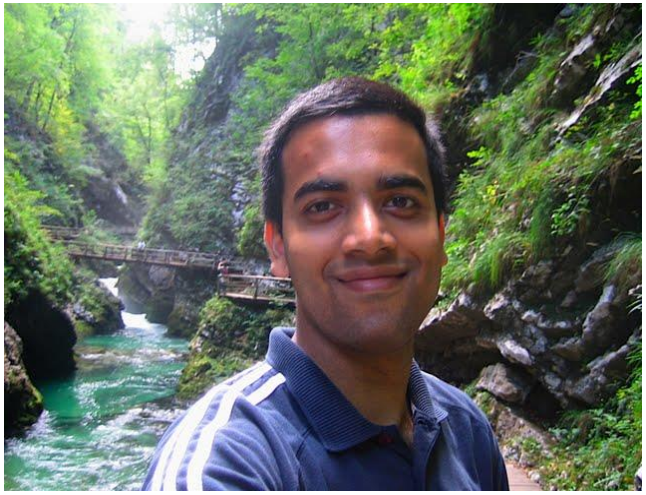
Yogesh Simmhan
Partha Talukdar

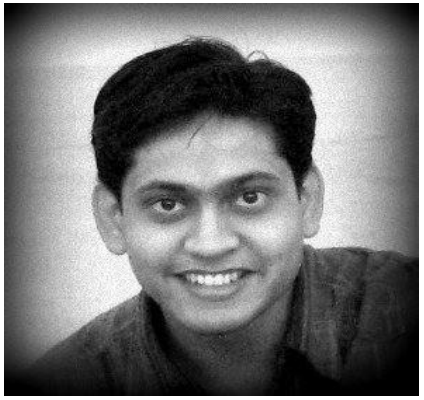
About This Course

www.serc.iisc.in/courses/se256/

Who are we?

- Instructors: Yogesh Simmhan, Partha Talukdar
- Teaching Assistant: Ravikant Dindokar





What is this course about?

- Fundamental systems aspects of big data platforms,
- How these platforms can be used to build large-scale data intensive applications.
- *Why Big Data platforms are necessary?*
- *How they are designed? What are the programming abstractions (e.g. MapReduce) that are used to compose data science applications?*
- *How the programming models are translated to scalable runtime execution on clusters and Clouds (e.g. Hadoop)?*
- *How do you design algorithms for analyzing large datasets?*
- *How do you map them to Big Data platforms?*

What is this course about?

- Platforms used for developing applications over large columnar/tuple based data,
  - Map Reduce/Apache Hadoop,
  - Streaming data like Apache Storm and Spark Streaming,
  - Graph/linked data like Apache Giraph
- Declarative tools
  - NoSQL databases and TensorFlow.
- Scale up of data analysis techniques
  - clustering, collaborative filtering, frequent itemset mining, classification, graph analytics,
  - Apply them over large datasets
- *Why do platforms scale, and how can you use them to scale data science applications?*

Pre-requisites

- Data Structures, Programming and Algorithm concepts. Programming experience required, preferably in Java.

Other Data Science Courses

- Foundations of Data Science
- Data Analytics
- Data Analysis and Visualization

- Introduction to Cloud Computing, Parallel Programming
- Topics in Web-scale Knowledge Harvesting, Data Mining, Deep Learning, Machine Learning, Artificial Intelligence, Probability & Statistics
- Bioinformatics, Chemoinformatics, Video Analytics

Lecture Schedule

- Wed 2-4PM, SE202
- Big Data & Platform Design Goals (YS)
  - Big Data & other computing platforms
  - Runtime Platforms: Hadoop, HDFS.
- Data-intensive algorithms and applications (PPT)
  - Analytics: clustering, collaborative filtering, frequent itemset mining, Classification, graph analytics, etc.
- Programming models data streams and graphs (YS)
  - Dataflow, BSP, Pregel.
    Runtime Platforms: Apache Storm, Apache Giraph, Apache Spark
- Big Tools: NoSQL databases, SparkQL, TensorFlow, etc (YS/PPT)

Tutorials

- Time TBD with Ravikant, Students
  - Check room availability
  - 330-430PM Mondays
- Programming aspects
- Assignment guides
- Installation issues
- Bring laptops if possible

Text Book

- Select chapters from **Mining of Massive Datasets**, *Jure Leskovec, Anand Rajaraman and Jeff Ullman*, 2nd Edition (v2.1), 2014.

Assignments & Exams (tentative)

- Assignment 0: Jan 12
- Assignment 1: Jan 26
- Assignment 2: Feb 23
- Midterm: Mar 2
- Assignment 3: Mar 15
- Project Topic Finalized: Mar 16
- Project Midterm Report: Apr 5
- Project Final Report: Apr 26
- Final Exam: *TBD*

Interaction

- http://www.serc.iisc.in/courses/se256/
- se256.jan16@mailman.serc.iisc.in
- Class/lab discussions

Academic Integrity

- Students must uphold IISc's Academic Integrity guidelines. While these are common sense, it is helpful to review them since failure to follow them will lead to sanctions and penalties.
- **This includes a reduced or failing grade in the course. Severe cases of academic violations will be reported to the Institute and may lead to an expulsion.**

http://www.iisc.ernet.in/students-corner/existingstudents-academicintegrity.php

Academic Integrity

- Learning takes place both within and outside the class. Hence, **discussions** between students and reference to **online material** is **encouraged** as part of the course to achieve the intended learning objectives.
- However, while you may learn from any valid source, you **must form your own ideas** and **complete problems and assignments by yourself**.
- *All works submitted by the student as part of their academic assessment must be their own!*

Academic Integrity

- **Plagiarism**: *Verbatim* reproduction of material from external sources (web pages, books, papers, etc.) is **not acceptable**.
- If you are *paraphrasing* external content (or even your own prior work) or were otherwise *influenced* by them while completing your assignments, projects or exams, **you must clearly acknowledge them**.
- *When in doubt, add a citation!*

Academic Integrity

- **Cheating**: While *you may discuss* lecture topics and broad outlines of homework problems and projects with others, **you cannot collaborate** in completing the assignments, **copy** someone else's solution or **falsify results**.
- You **cannot use** notes or unauthorized resources during exams, or copy from others.
- The narrow exception to collaboration is between team-mates when competing the project, and even there, *the contribution of each team member* for each project assignment should be **clearly documented**.

Academic Integrity

- **Classroom Behaviour**: Ensure that the course atmosphere, both in the class, outside and on the online forum, is conducive for learning.
- *Participate* in discussions but *do not dominate* or be abusive. **There are no "stupid" questions** ☺ Be considerate of your fellow students and avoid disruptive behaviour.

Big Data & Why is it Important

What is Big Data?

Data whose characteristics exceeds the capabilities of conventional *algorithms, systems and techniques* to derive useful value.

https://www.oreilly.com/ideas/what-is-big-data

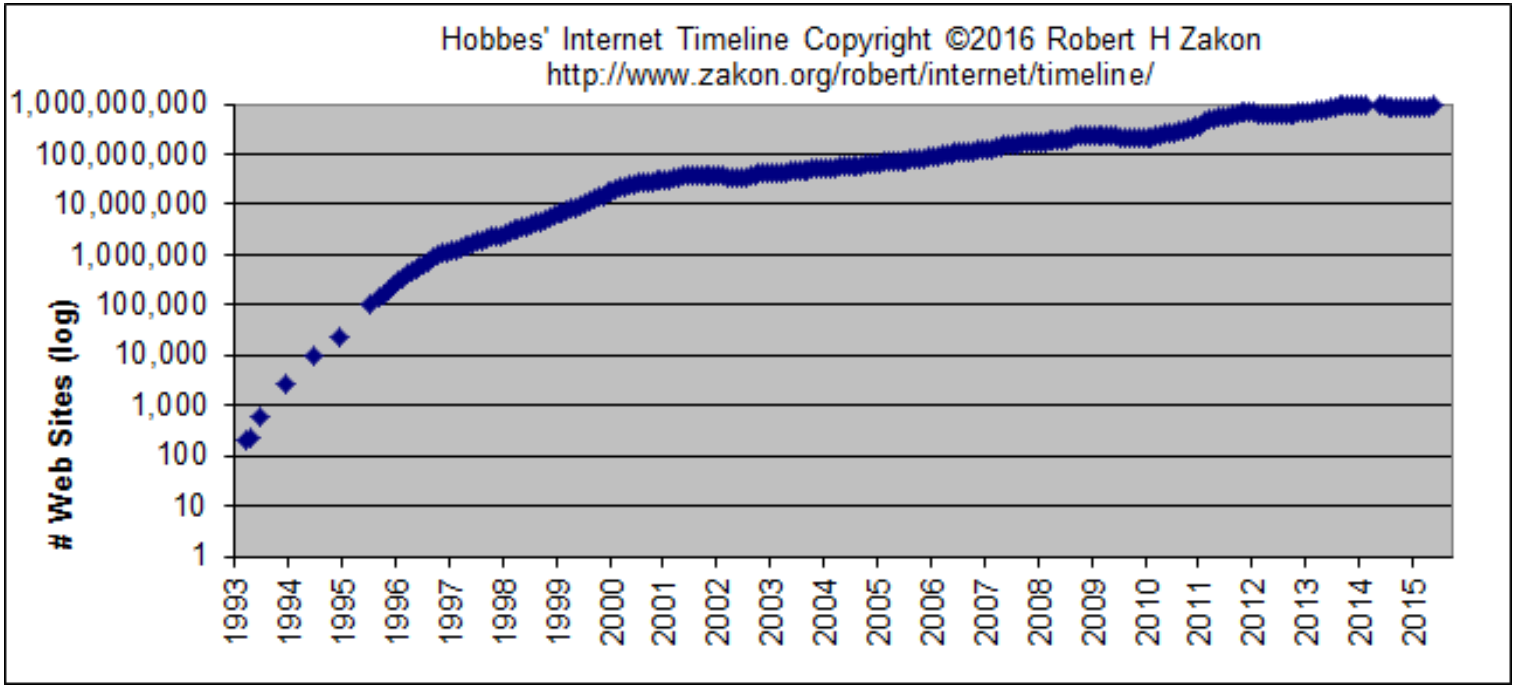
The term *is* fuzzy … *Handle with care!*

Wordle of "Thought Leaders'" definition of Big Data, © Jennifer Dutcher, 2014
https://datascience.berkeley.edu/what-is-big-data/

So, where does Big Data come from?

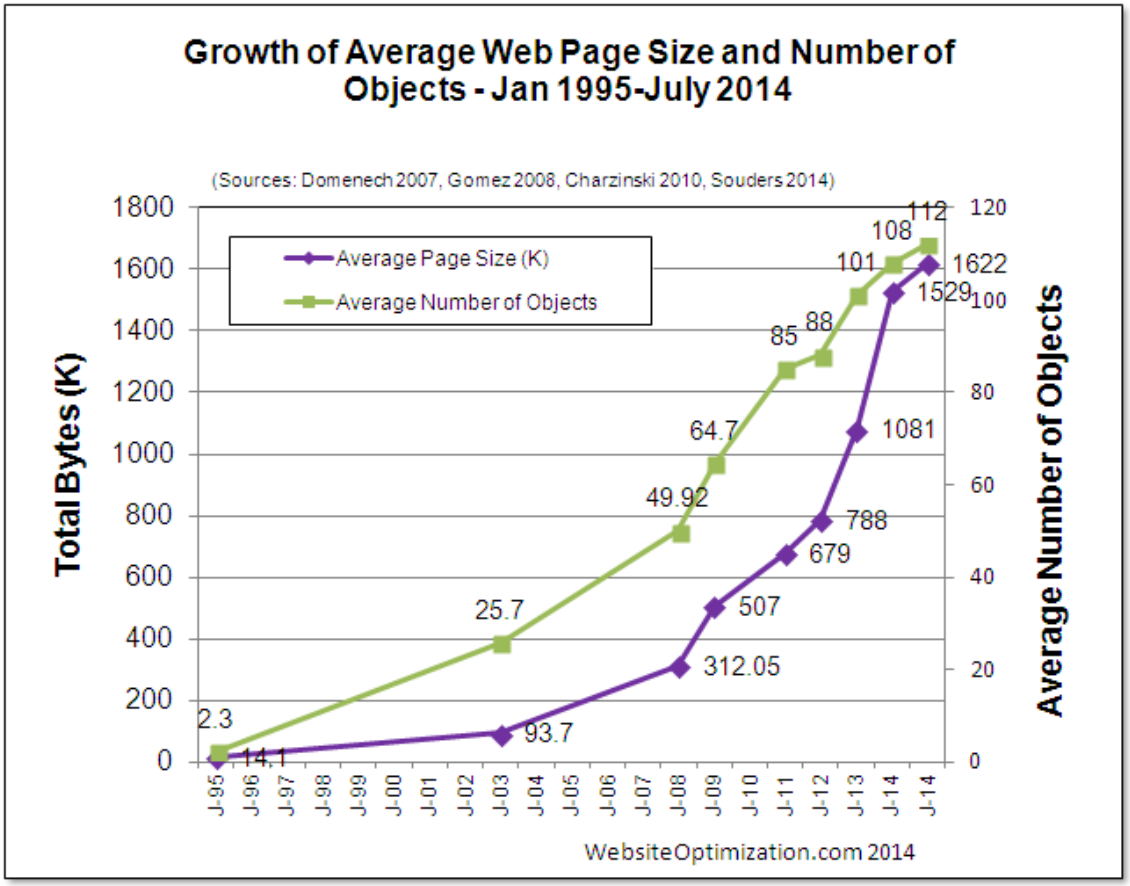
Web & Social Media

- World Wide Web & Search engines



Web & Social Media

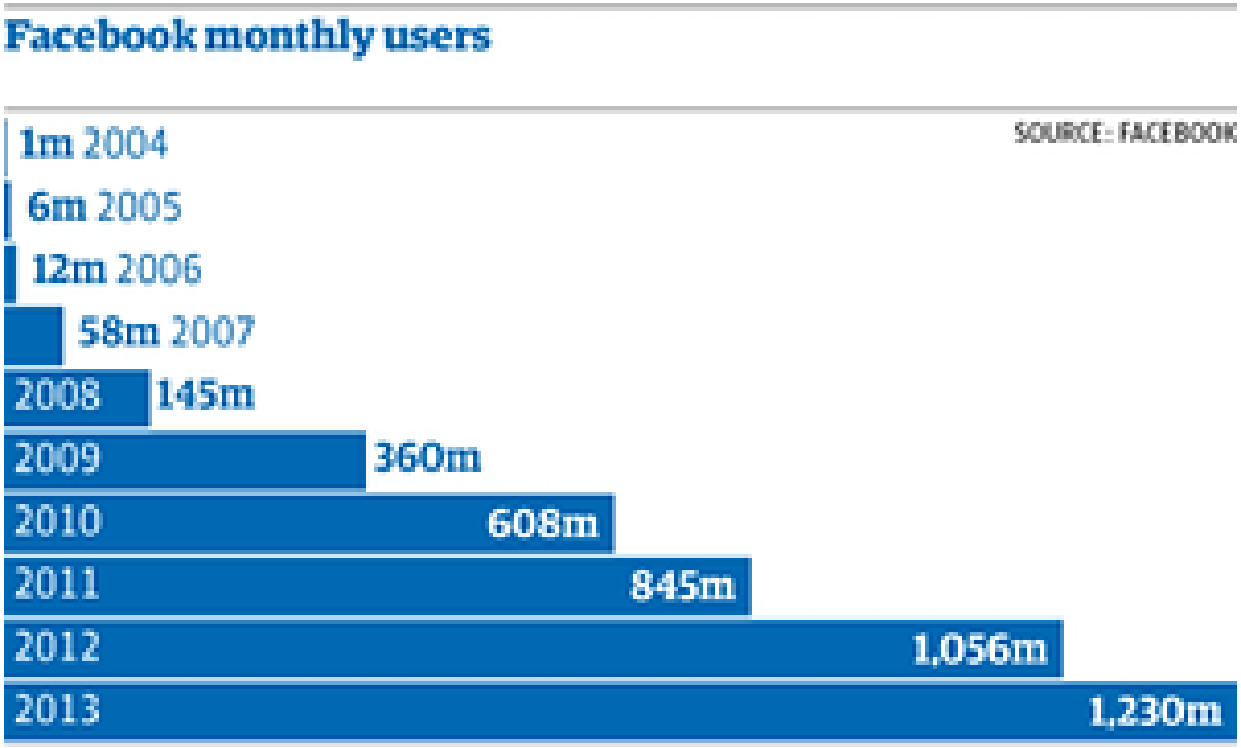
- World Wide Web & Search engines



http://www.yottaa.com/company/blog/application-optimization/a-brief-history-of-web-page-size/

Web & Social Media

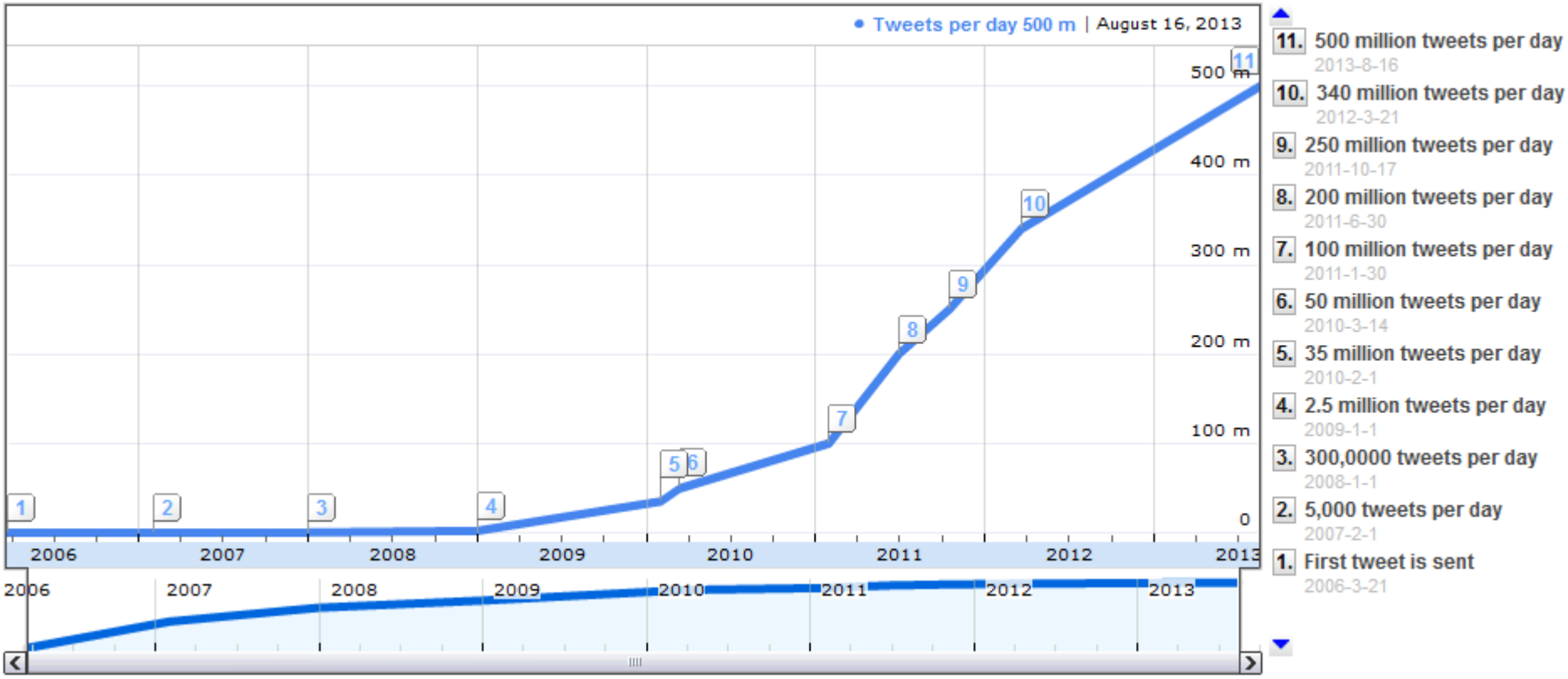
- Social Networks & Micro-blogs



http://www.theguardian.com/news/datablog/2014/feb/04/facebook-in-numbers-statistics

Web & Social Media

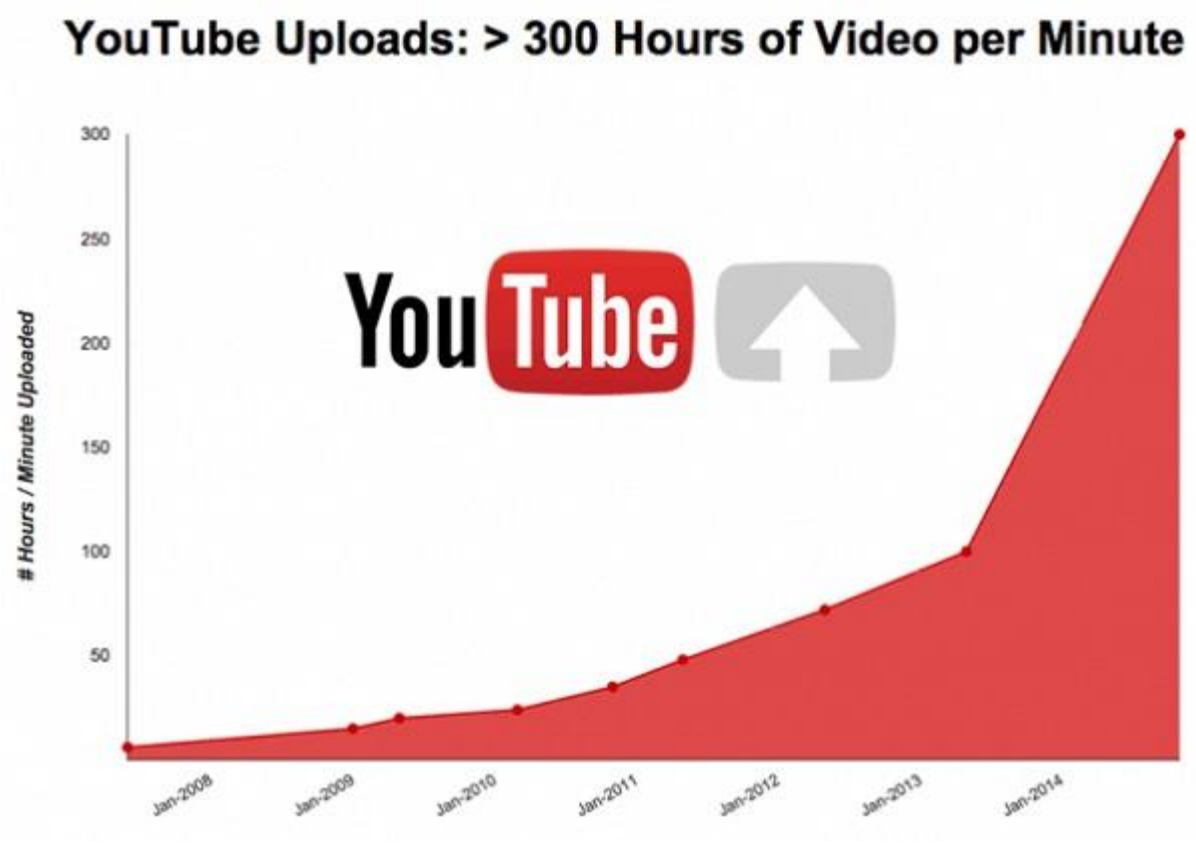
- Social Networks & Micro-blogs



http://www.internetlivestats.com/twitter-statistics/

Web & Social Media

- Youtube



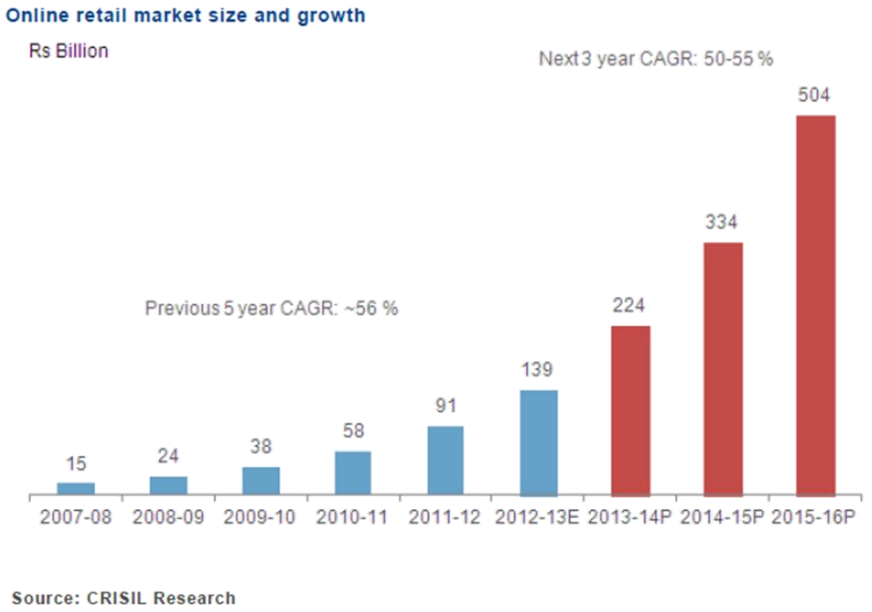
http://www.tubefilter.com/2014/12/01/youtube-300-hours-video-per-minute/

Web & Social Media

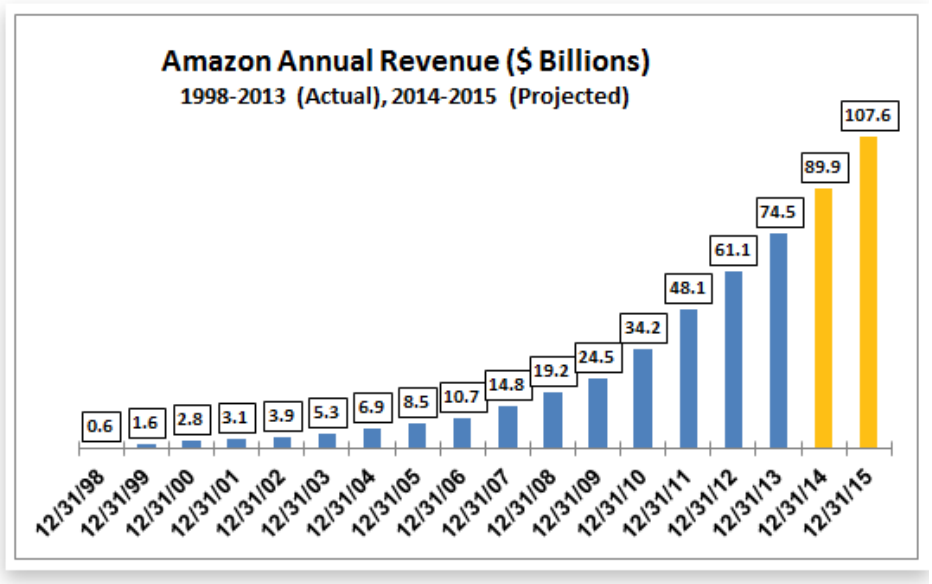
- The Netflix Prize (2006-2009)
  - "An open competition for the best collaborative filtering algorithm to predict user ratings for films, based on previous ratings"
  - Training set: 99,072,112 ratings
    - `<user, movie, date of grade, grade>`
  - Qualifying set: 2,817,131 ratings
    - `<user, movie, date of grade>`
- Siri, Cortana, Now, Echo

Enterprises & Government

- Online retail & eCommerce



Source: CRISIL Research

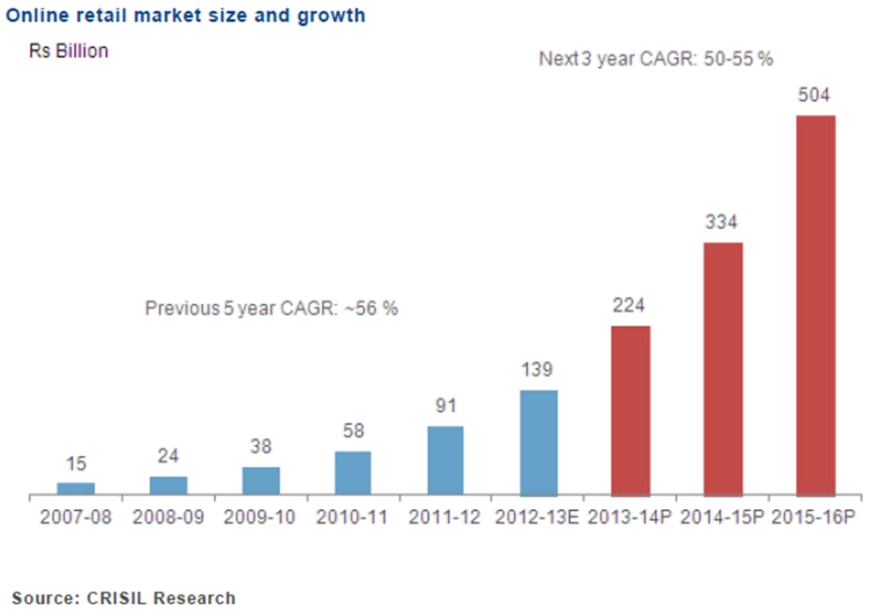


http://blogs.ft.com/beyond-brics/2014/02/28/online-retail-in-india-learning-to-evolve/

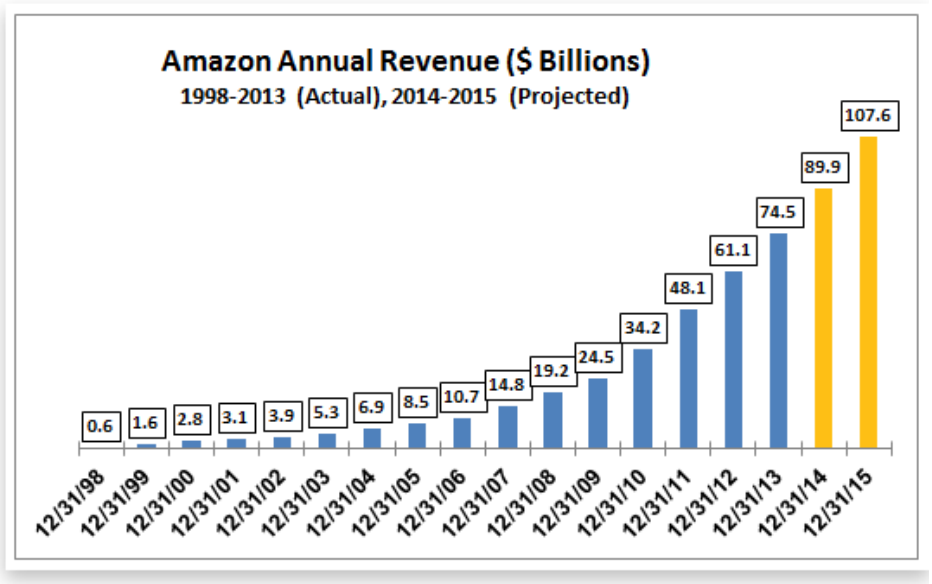
http://www.peridotcapital.com/2014/04/amazon-sales-growth-projections-for-next-two-years-appear-overly-optimistic.html

Enterprises & Government

- Online retail & eCommerce



Source: CRISIL Research



http://blogs.ft.com/beyond-brics/2014/02/28/online-retail-in-india-learning-to-evolve/

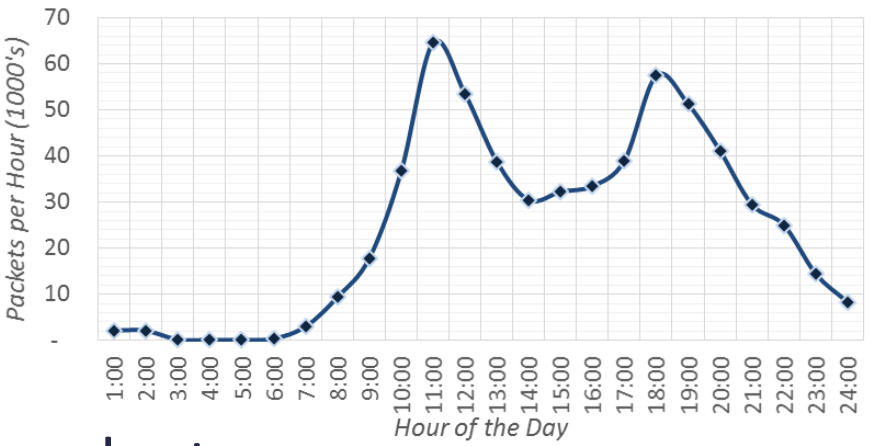
http://www.peridotcapital.com/2014/04/amazon-sales-growth-projections-for-next-two-years-appear-overly-optimistic.html

Enterprises & Government

- Logistics
  - FedEx handles ~2.2 billion transactions a day.
- Indian Railways
  - 150,000 Tickets Per Hour

http://pivotal.io/big-data/case-study/distributed-in-memory-data-management-solution
http://www.digitalistmag.com/future-of-work/2013/11/01/big-data-delivered-fedex-25-years-0887095

Aadhaar Data Streams



- Aadhaar packet sizes and rates
  - Hourly rates for a 24 hour period, Sep, 2015
- Mean packet size is **~3MB**
- Bi-modal rate distribution
  - After morning & evening sessions
  - **~600K/day** now. Peak was **1.3M/day** in 2013
- Soon >20 Millions of authentications per day

Science: Pan-STARRS Sky Survey

www.ps1sc.org

- *Discover & characterize Earth-approaching objects that might pose a danger to our planet.*
- One of the largest telescopes
  - 1.4 Gigapix camera world's largest!
- Scan $^{2}/_{3}^{rds}$ of sky, 3 times/month
  - **1 PB** of images, **30 TB** of processed data/year
  - **150 M** detections / night
  - **5.5 Billion** objects, **350 Billion** detections

The Andromeda Galaxy, M31

Dome of PS1 telescope at Haleakala

SDSS, PS, LSST, SKA, …

- Sloan Digital Sky Survey…first all digital survey…precursor to PS
- LSST: 10M events/night. 37B objects, 7T obs, 30T measurements per year*
- Square Km Array
  - Radio Telescope, 300 dishes
  - 68Tb/s to be sampled
  - 10PB images/day. 1PB catalog.

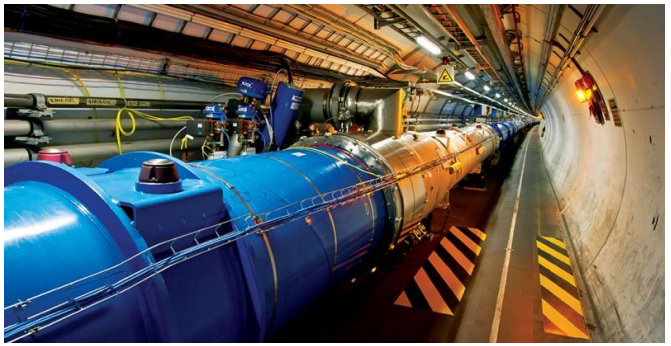
Graphs

Next Gen Sequencers

- TB of data per run

Large Hadron Collider

- Petabytes of data per run



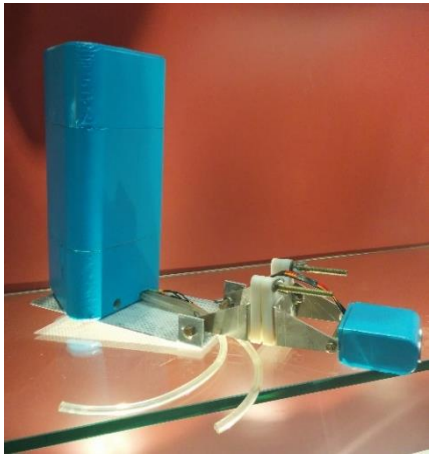
*Introduction to LSST Data Management, Jeffrey Kantor

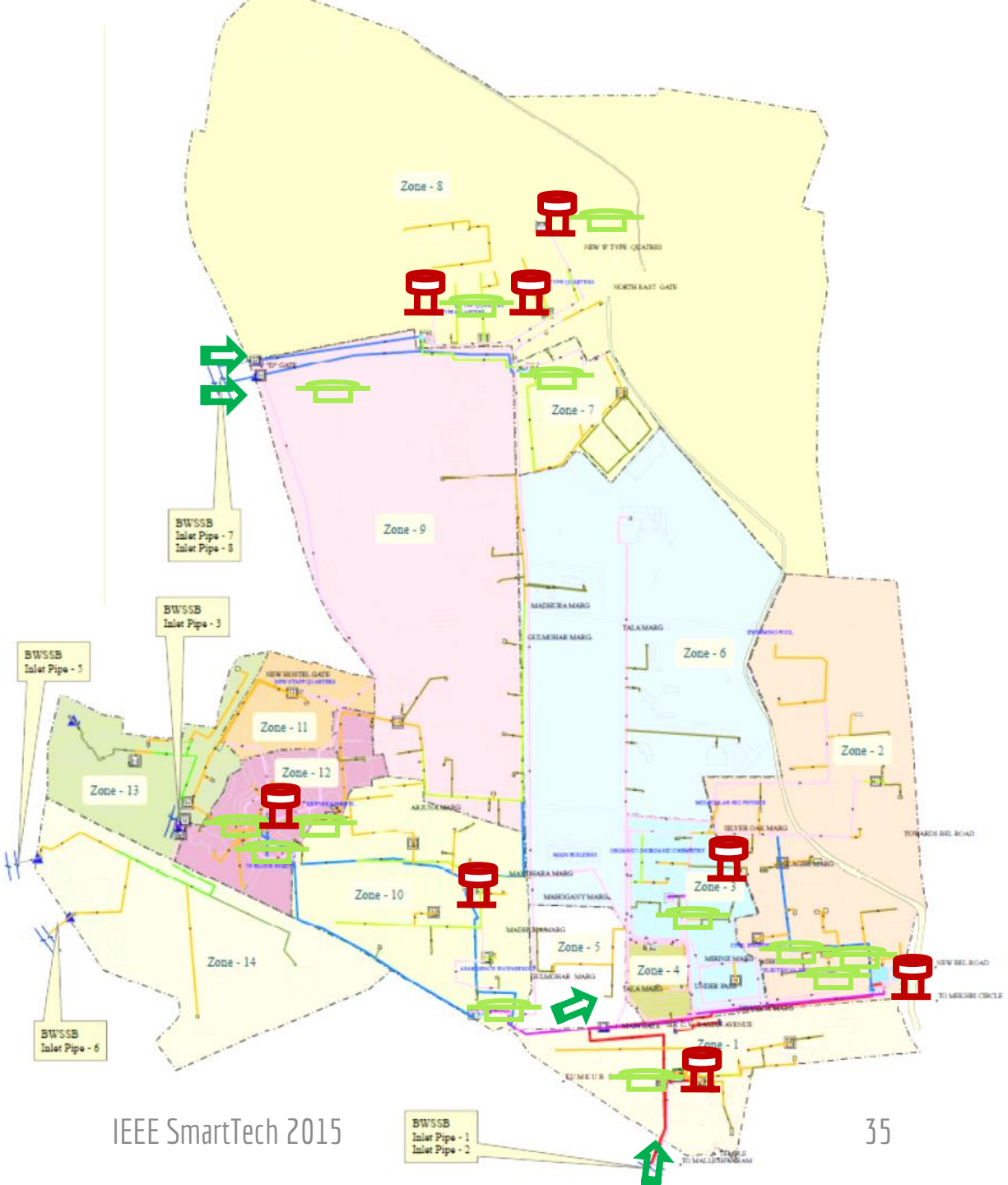
Square Kilometre Array Computational Challenges, Paul Alexander

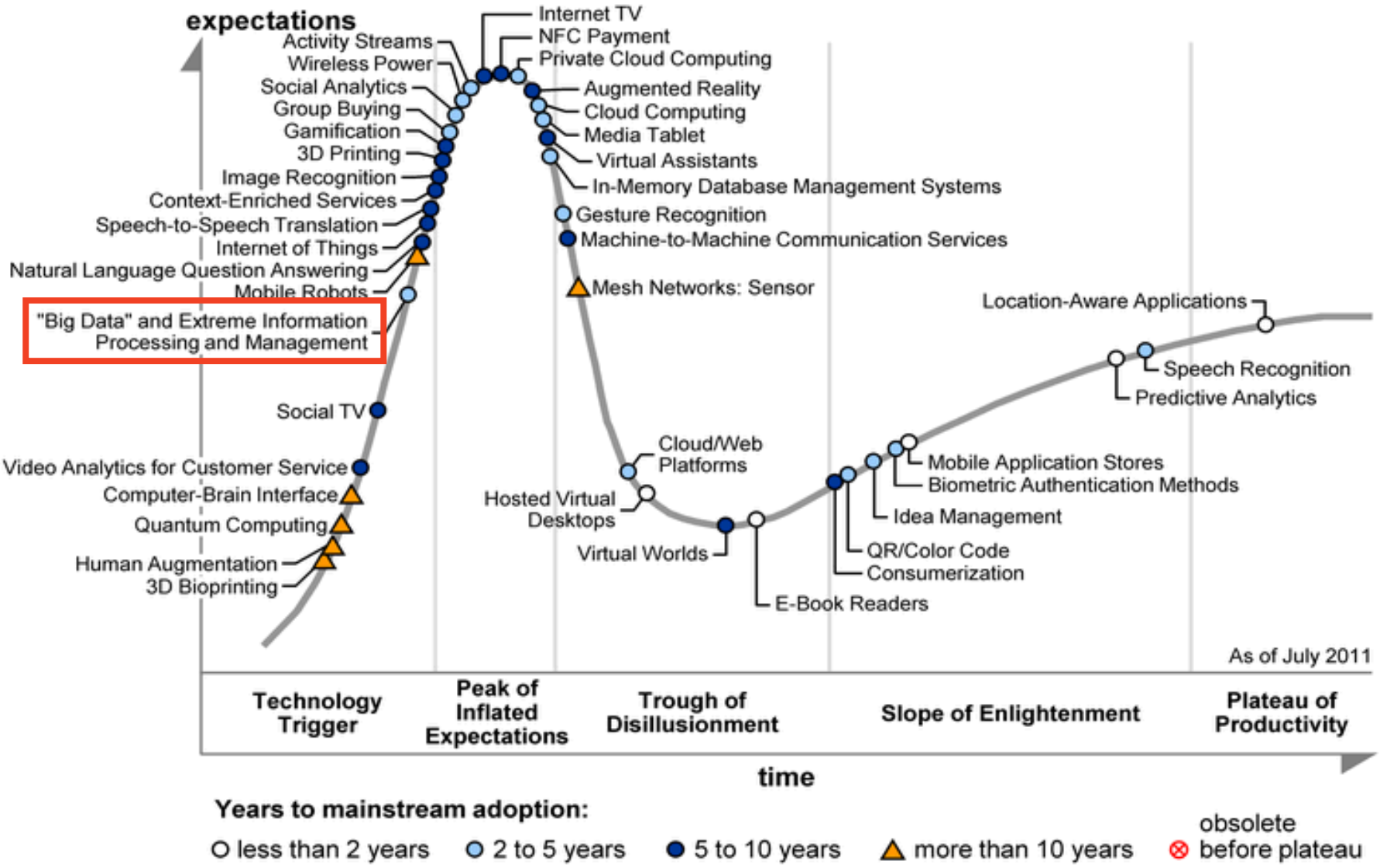
home.cern

Internet of Everything

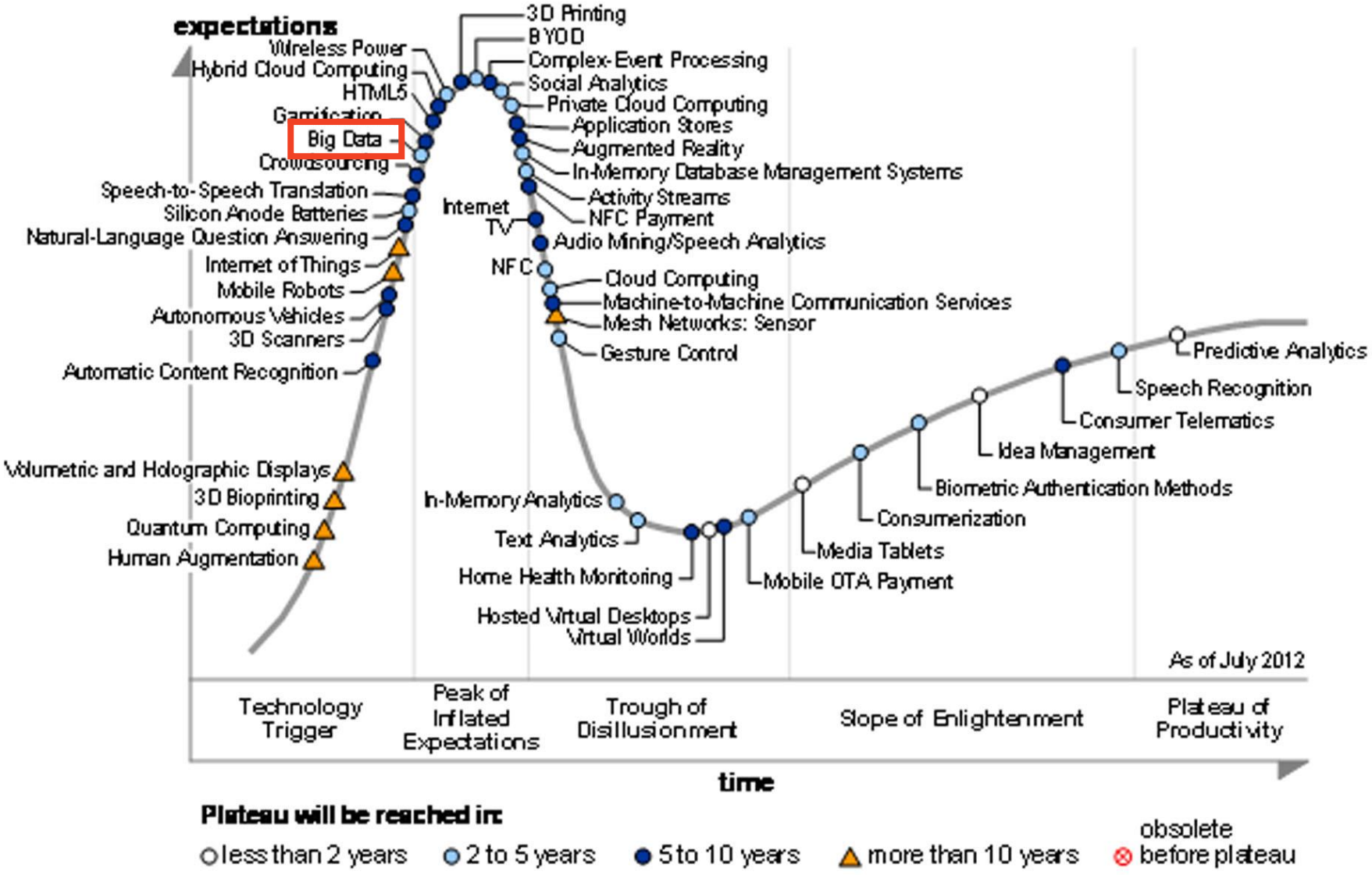
- Personal Devices: Smart Phones, Fitbit
- Smart Appliances
- Smart Cities: Power, Water, Transportation, Environment
- Smart Retail



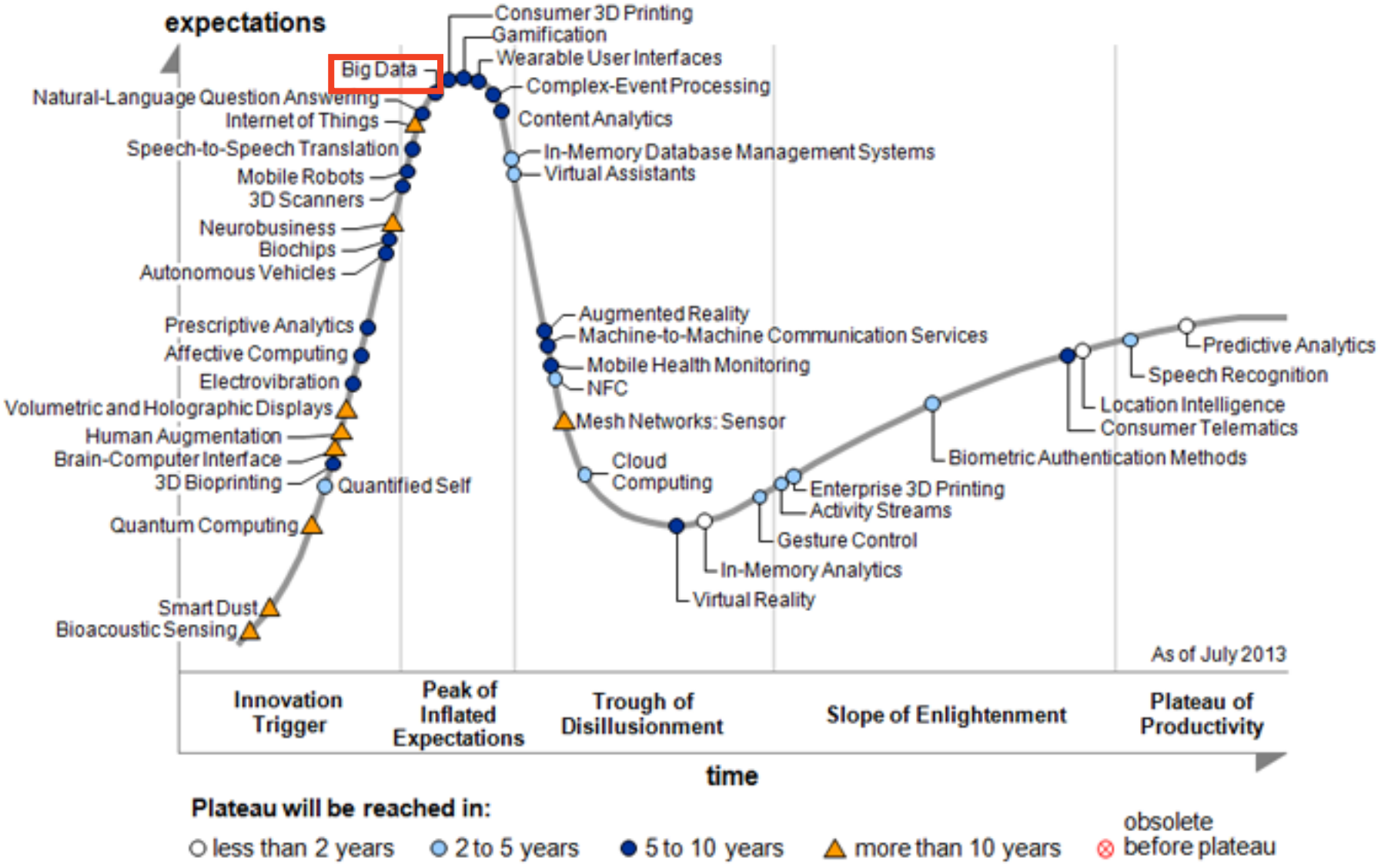

Zone - 8
Zone - 7
Zone - 9
Zone - 6
Zone - 2
Zone - 11
Zone - 12
Zone - 13
Zone - 10
Zone - 3
Zone - 5
Zone - 4
Zone - 14
Zone - 1
BWSSB
Inlet Pipe - 7
Inlet Pipe - 8
BWSSB
Inlet Pipe - 3
BWSSB
Inlet Pipe - 5
BWSSB
Inlet Pipe - 6
BWSSB
Inlet Pipe - 1
Inlet Pipe - 2



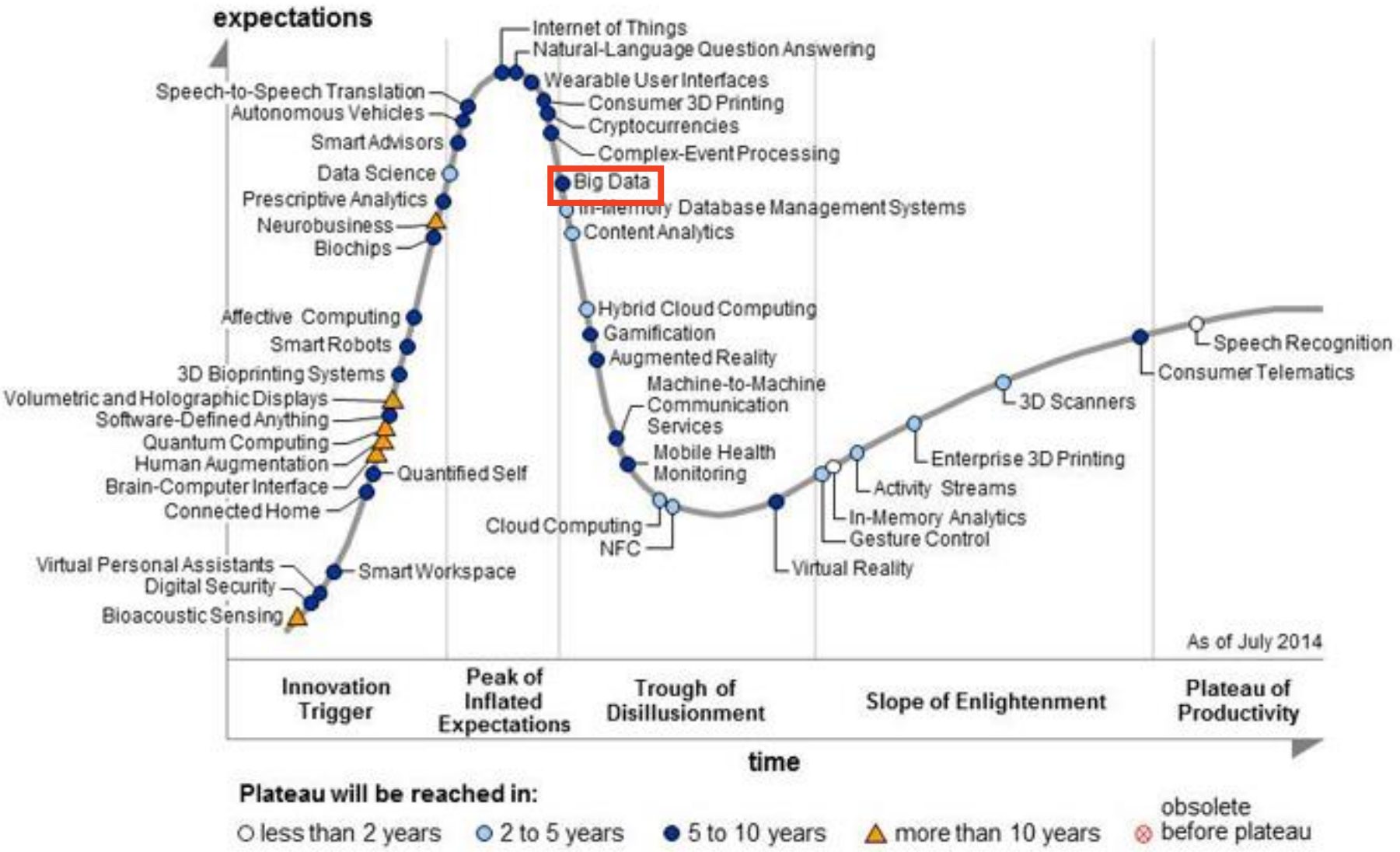
http://www.kdnuggets.com/2015/08/gartner-2015-hype-cycle-big-data-is-out-machine-learning-is-in.html



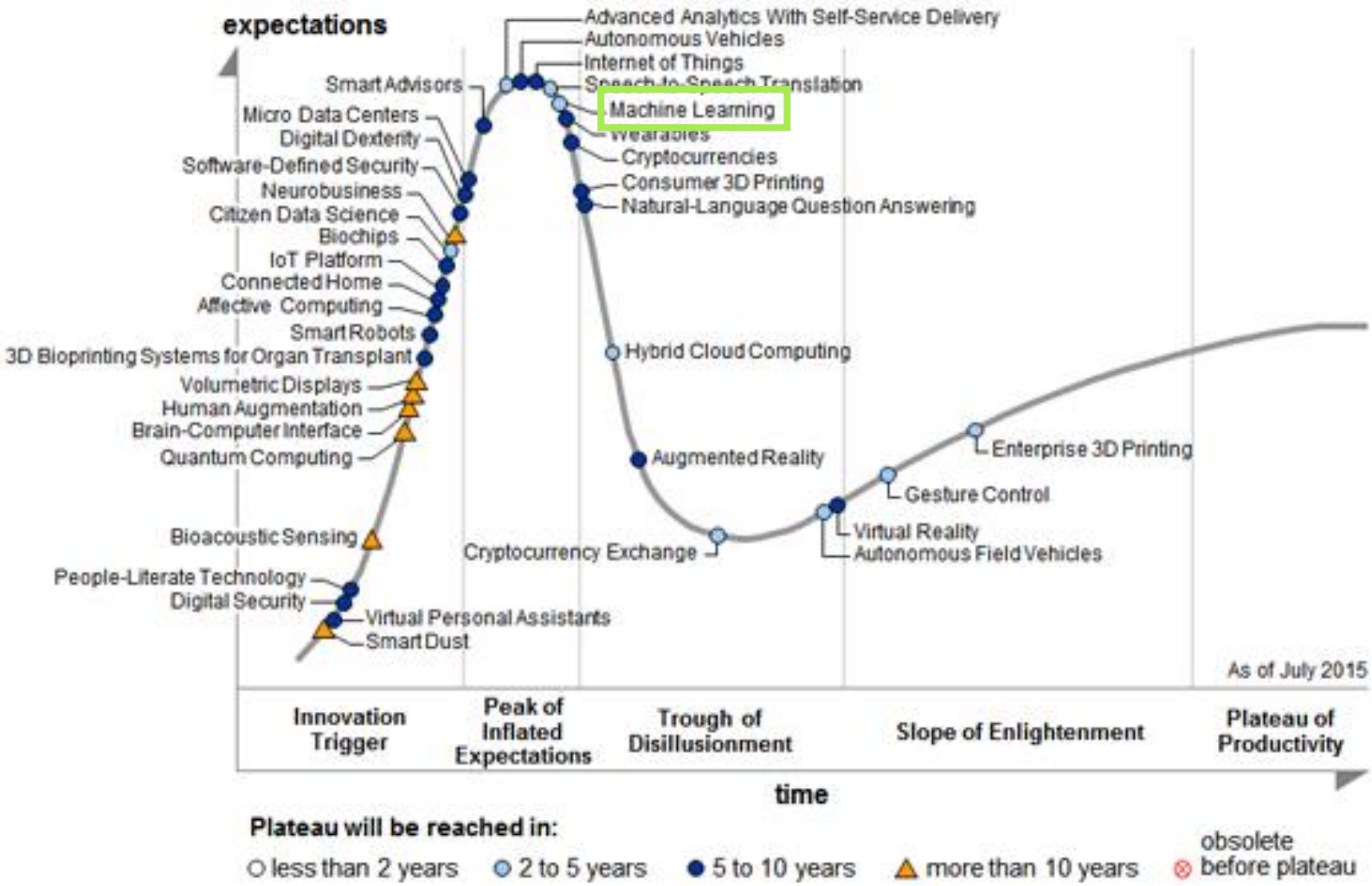
http://www.kdnuggets.com/2015/08/gartner-2015-hype-cycle-big-data-is-out-machine-learning-is-in.html



http://www.kdnuggets.com/2015/08/gartner-2015-hype-cycle-big-data-is-out-machine-learning-is-in.html



http://www.kdnuggets.com/2015/08/gartner-2015-hype-cycle-big-data-is-out-machine-learning-is-in.html



http://www.kdnuggets.com/2015/08/gartner-2015-hype-cycle-big-data-is-out-machine-learning-is-in.html

Why is Big Data Difficult?

Evolving Nature of Science Data

Large Hadron Collider

PS1 telescope

Illumina NGS @ IISc

Bluetooth Mote @ IISc

Smart Meter @ LADWP

Few Instruments, Large Data **Volume**

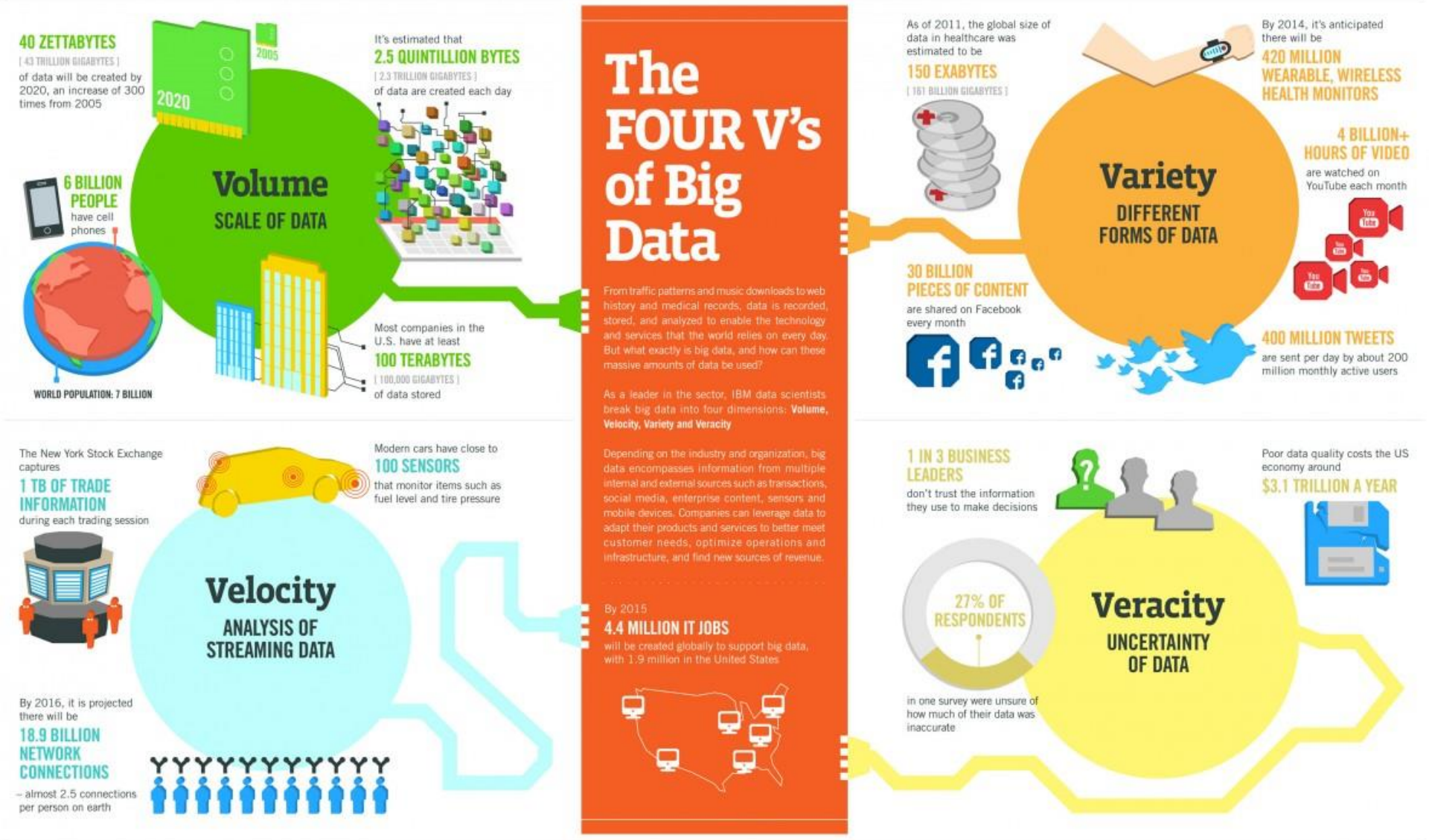
10^2 Sources
TB's Data
Days to Proc.

Many Devices, Volume & Velocity

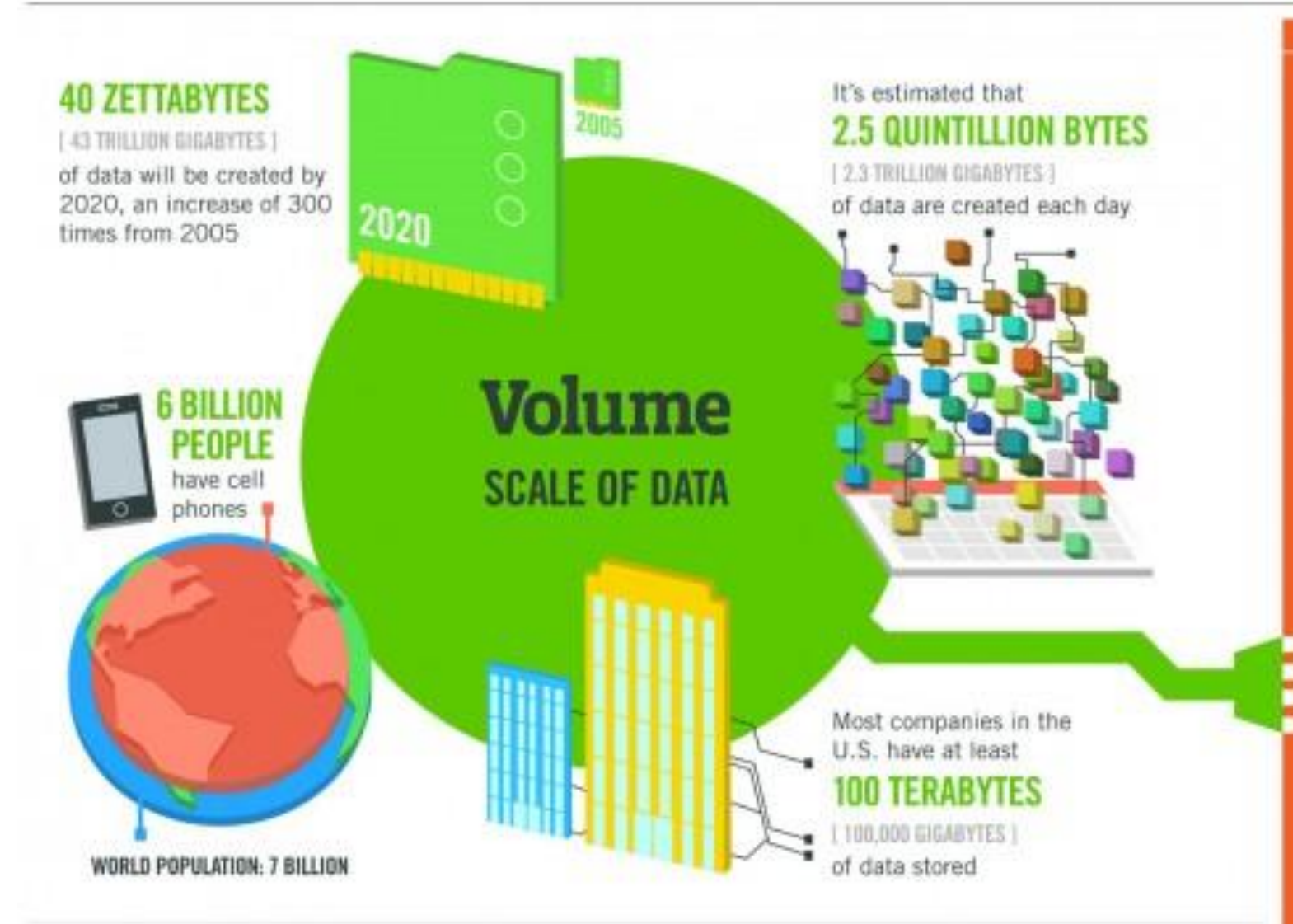
10^5 Sources
GB's Data
Hours to Proc.

Numerous Sensors, High data **Velocity**

10^8 Sources
MB's Data
<Mins to Proc.



http://www.ibmbigdatahub.com/infographic/four-vs-big-data



The FOUR of Big Data

From traffic patterns and r
history and medical rec
stored, and analyzed to
and services that the wo
But what exactly is big d
massive amounts of data

As a leader in the secto
break big data into fou
Velocity, Variety and Verac

Depending on the industr
data encompasses info
internal and external sourc

http://www.ibmbigdatahub.com/infographic/four-vs-big-data

WORLD POPULATION: 7 BILLION

U.S. have at least

100 TERABYTES

[100,000 GIGABYTES]

of data stored

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

But what exactly is big d
massive amounts of data

As a leader in the sect
break big data into foc
Velocity, Variety and Vera

Depending on the indus
data encompasses info
internal and external sour
social media, enterpris
mobile devices. Compan
adapt their products and
customer needs, opti
infrastructure, and find

By 2015

4.4 MILLION IT JO

will be created globally
with 1.9 million in the

Sources: McKinsey Global Institute, Twitter, Cisco, Gartner, EMC, SAS, IBM, MEPTEC, QAS

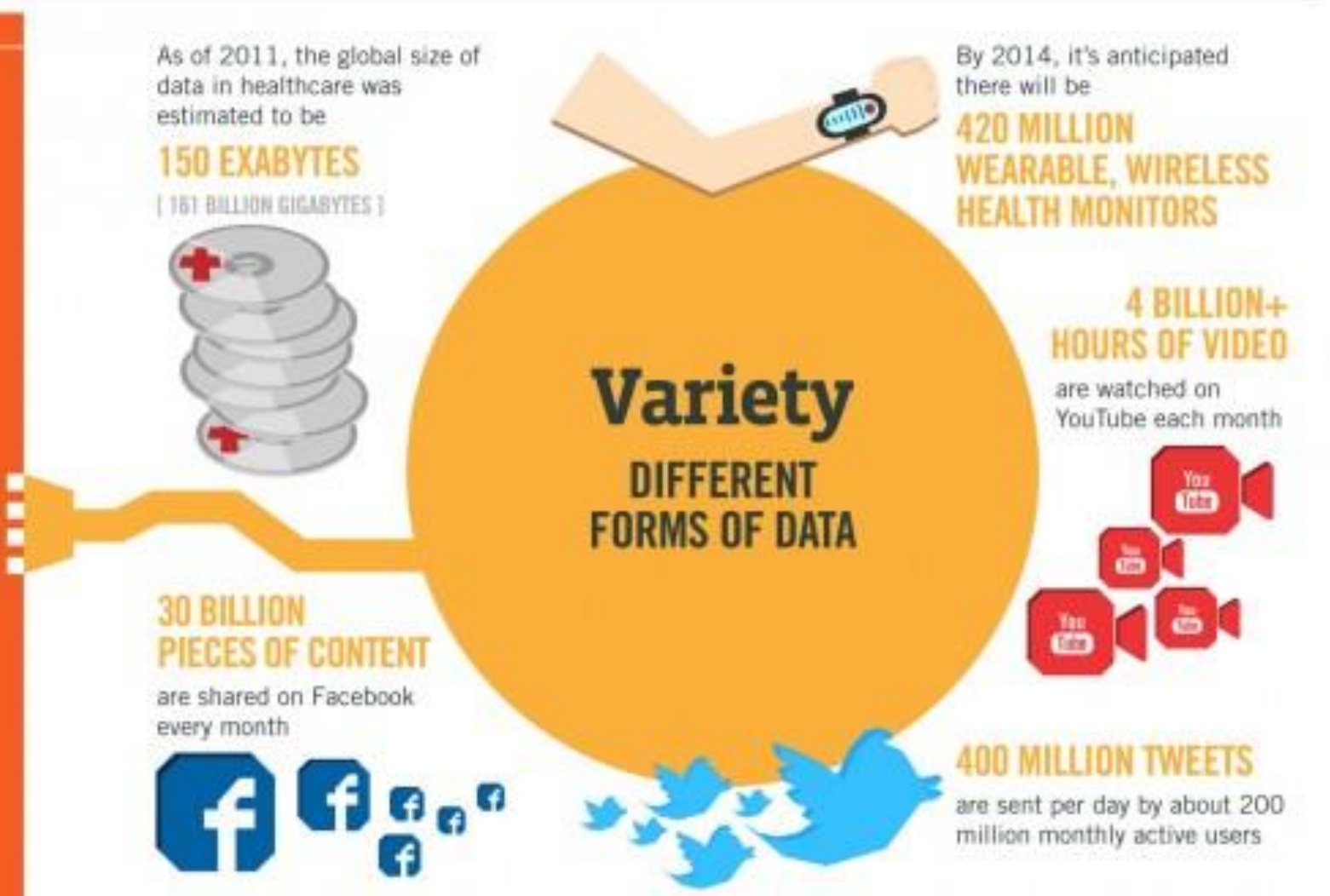
http://www.ibmbigdatahub.com/infographic/four-vs-big-data

R V's ig a

nd music downloads to web records, data is recorded, to enable the technology world relies on every day. g data, and how can these ata be used?

ector, IBM data scientists four dimensions: **Volume**, **eracity**

ustry and organization, big nformation from multiple ources such as transactions, ise content, sensors and



http://www.ibmbigdatahub.com/infographic/four-vs-big-data

ata, and how can these
be used?

or, IBM data scientists
ur dimensions: **Volume,**
city

try and organization, big
rmation from multiple
ces such as transactions,
e content, sensors and
ies can leverage data to
services to better meet
mize operations and
new sources of revenue.

BS

to support big data,
United States

are sent per day by about 200 million monthly active users

1 IN 3 BUSINESS LEADERS

don't trust the information they use to make decisions

Poor data quality costs the US economy around

$3.1 TRILLION A YEAR

in one survey were unsure of how much of their data was inaccurate

Veracity

UNCERTAINTY OF DATA

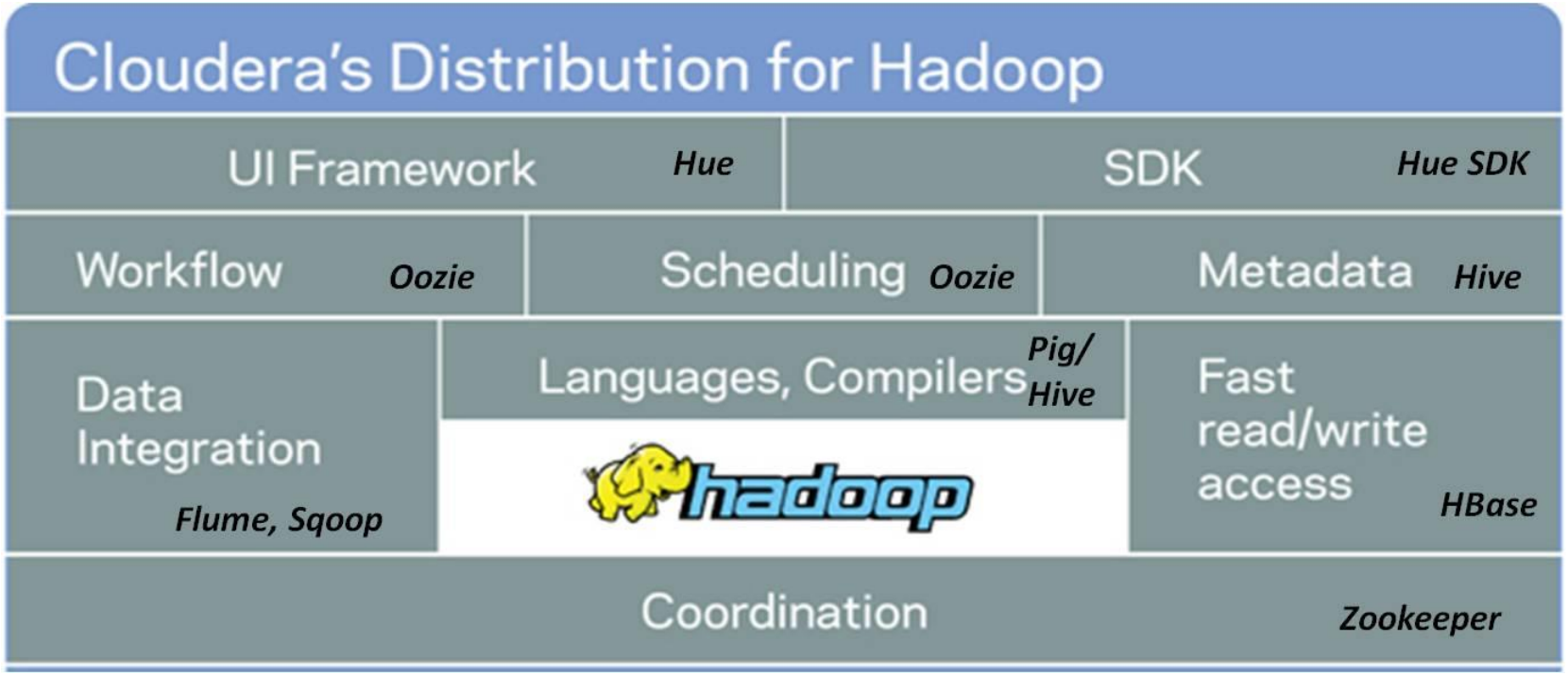
http://www.ibmbigdatahub.com/infographic/four-vs-big-data

IBM

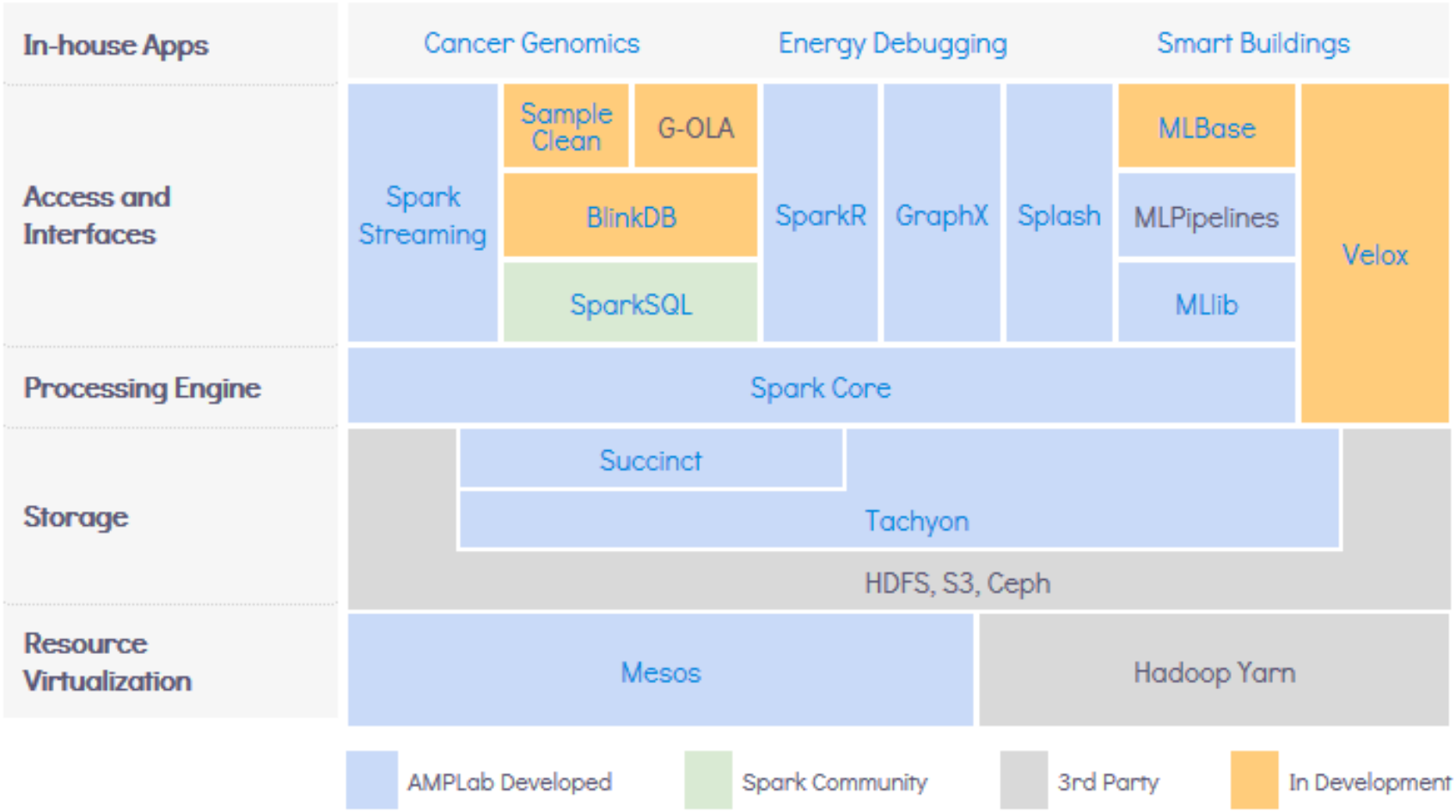
Data Platforms

- Acquire, manage, process Big Data
- At large scales
- To meet application needs

The Platform Stack: Cloudera



The Platform Stack: BDAS



https://amplab.cs.berkeley.edu/software/

Reading & Assignment

- Textbook chapter 2.1, 2.2
- Assignment 0
  - Install Hadoop 2.6.x+ in pseudo distributed mode on your laptop