BİL595
Distributed Data Processing and Analysis
«Big Data»

Spring 2017

Erdogan Dogdu

Çankaya University

Department of Computer Engineering

Outline

- Motivation for the course
- Course logistics
- Schedule
- Evaluation

Big data

- **V**olume
- **V**ariety
- **V**elocity
- Other V's
  - Veracity
  - Validity
  - Volatility

Why study «big data»?

- It is a trendy topic

Who is doing «big data»?

- All major IT companies
- Large corporations, banks, retail, all ...
- Academia
- Some examples are on the next few slides

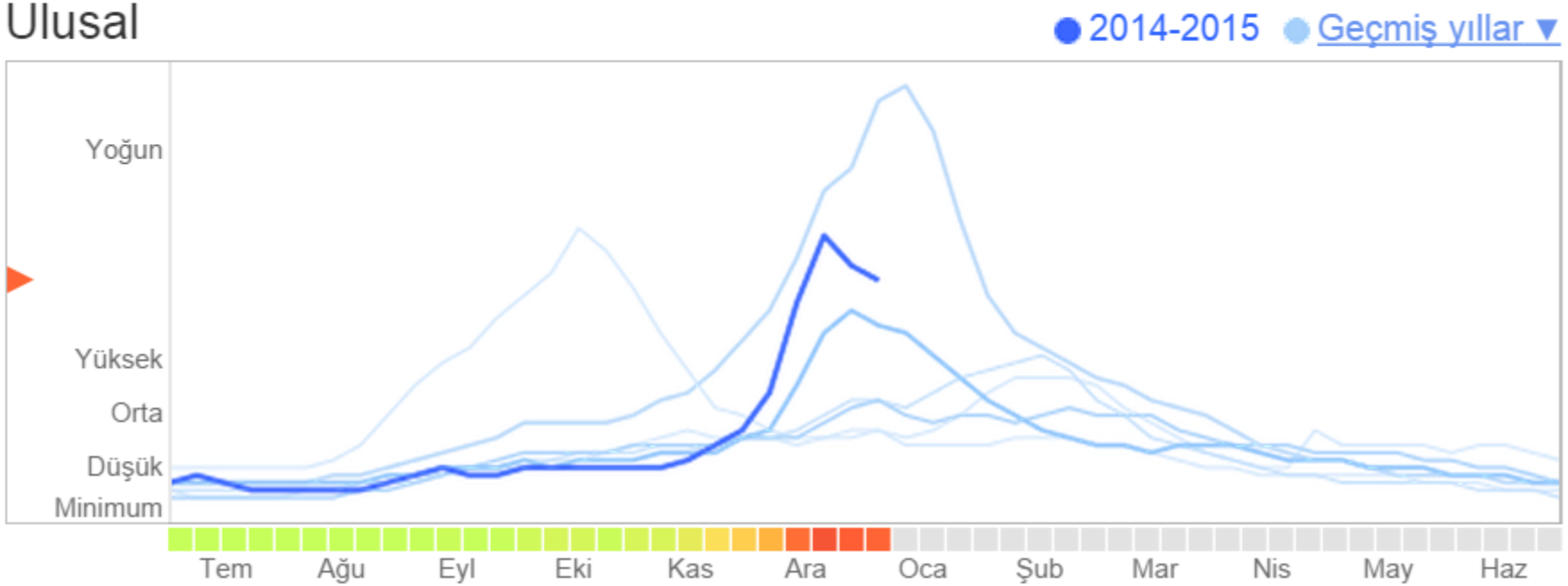
Google trends

www.google.org/flutrends/us/#US

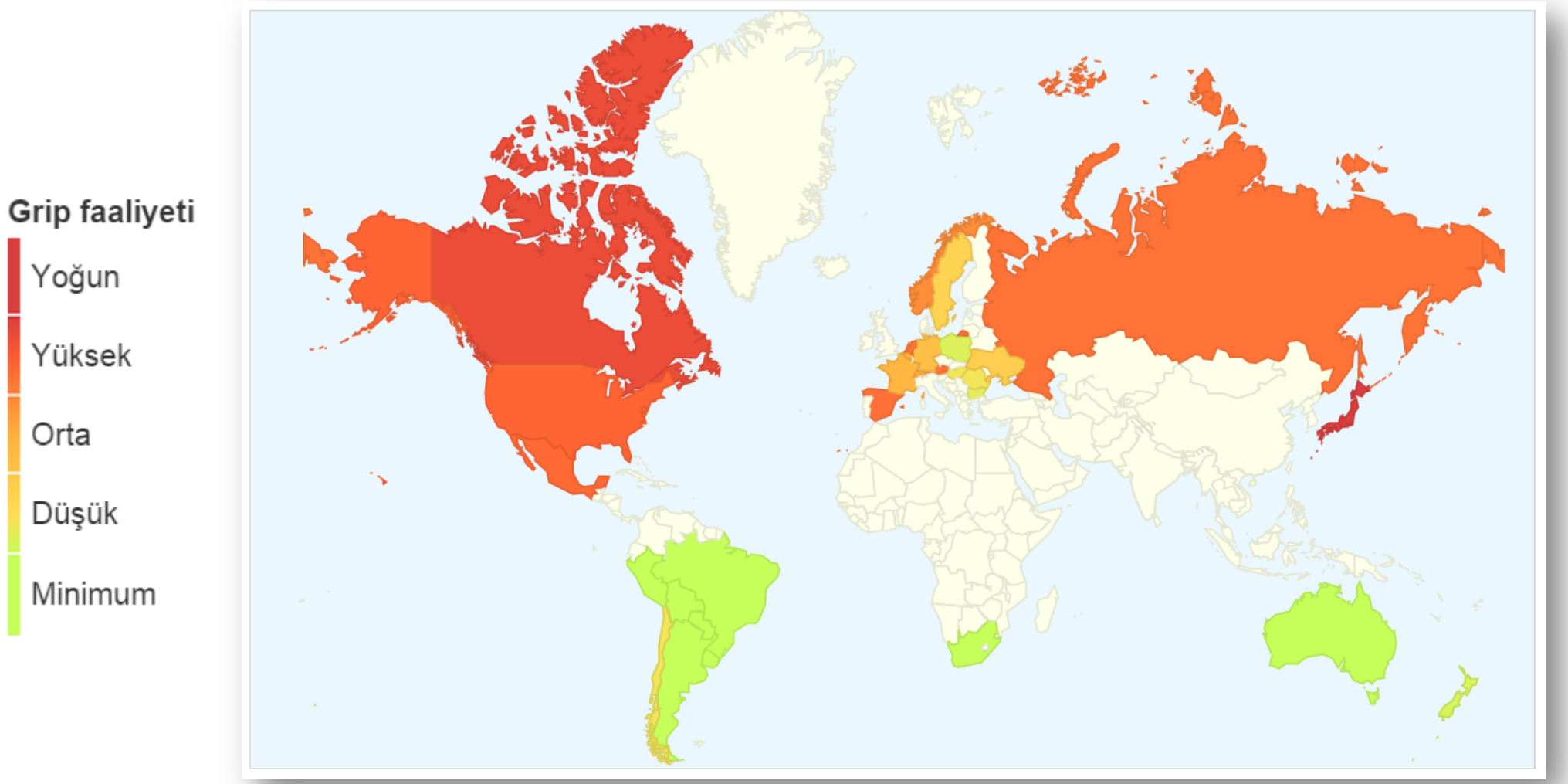
Grip trendlerini araştırın - ABD

Belirli arama terimlerinin, grip faaliyetlerini izlemek için iyi göstergeler olduğunu fark ettik. Google Grip Trendleri, grip faaliyetlerini öngörebilmek için toplu Google arama verilerini kullanmaktadır.
Daha fazla bilgi edinin »

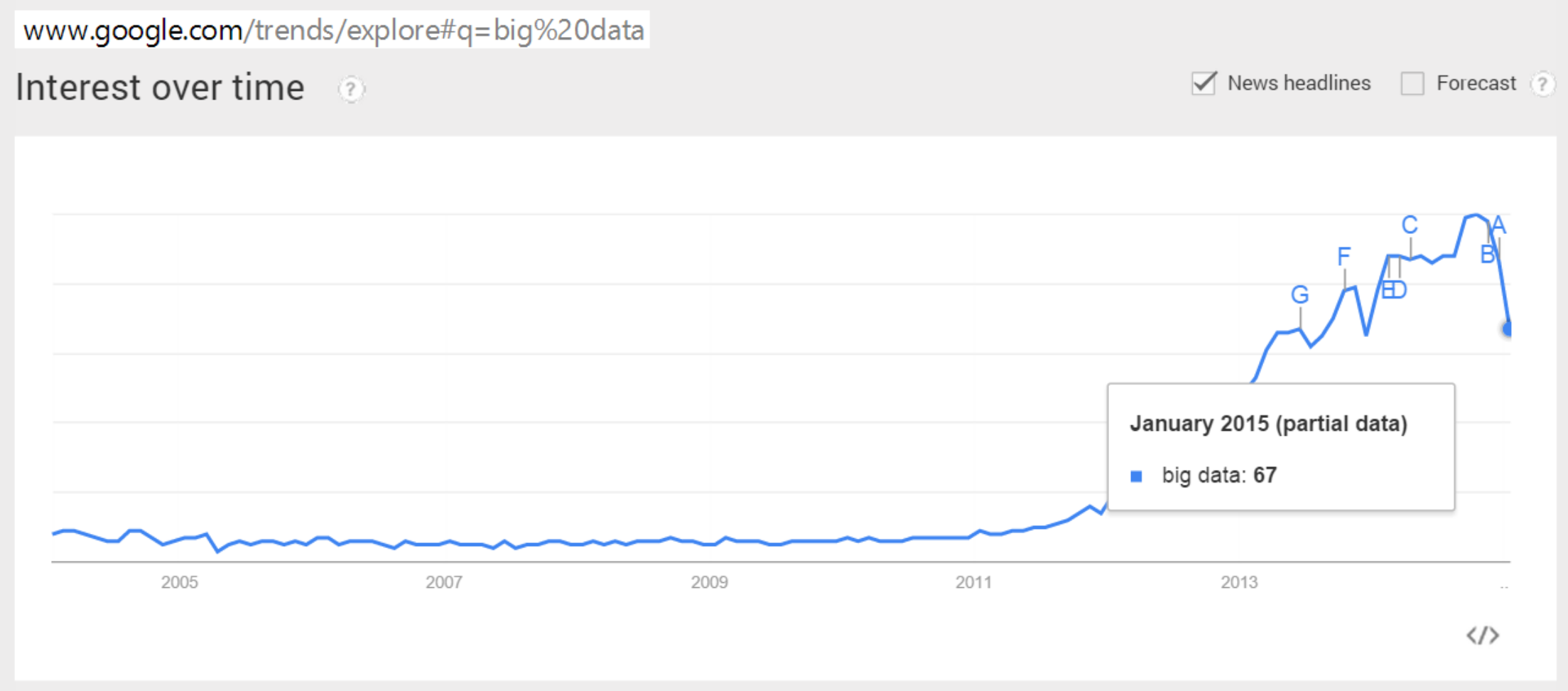
Ulusal



Google trends – flu trends worldwide



Google trends about «Big Data»



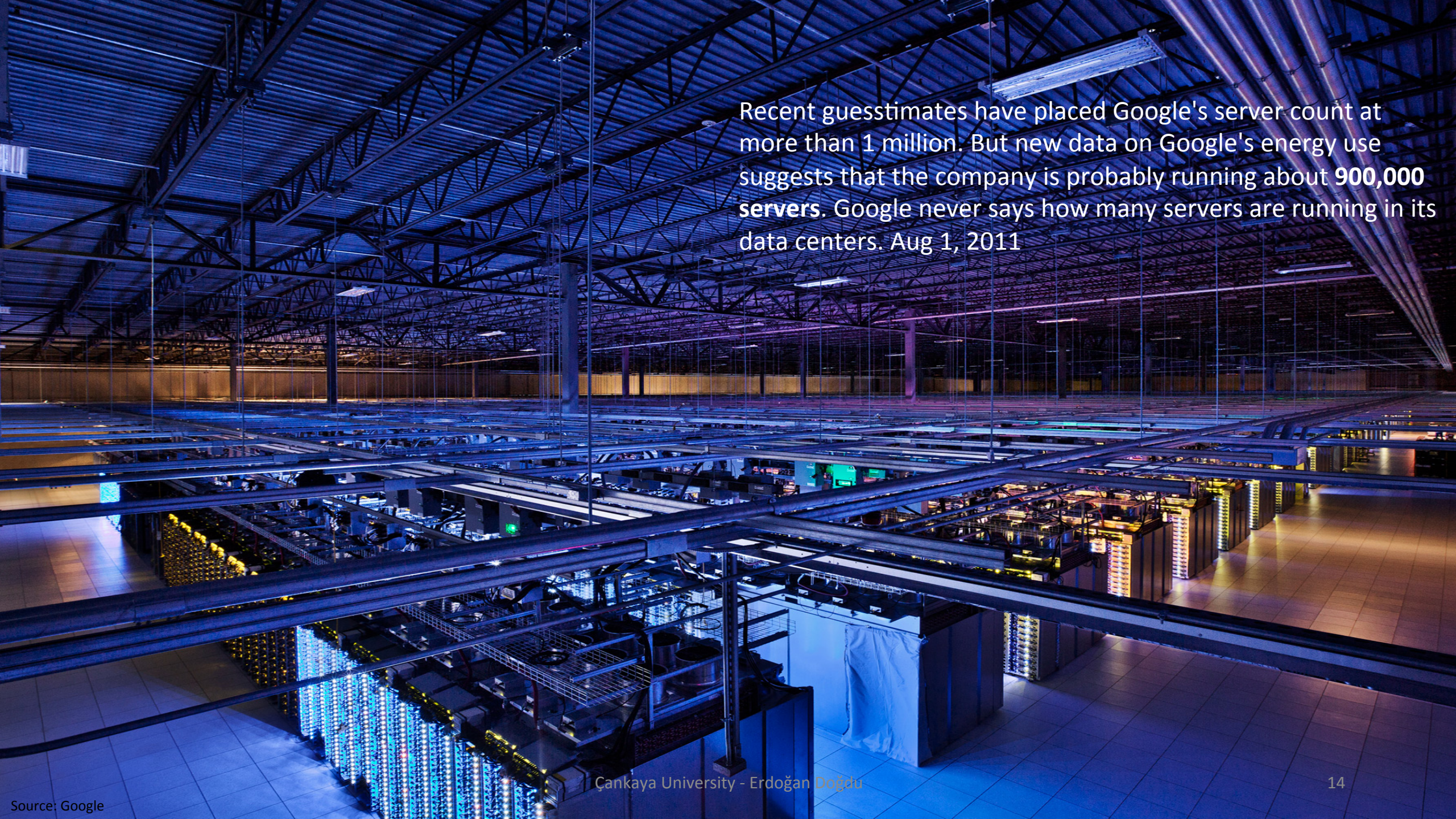
Source: Google

Source: Wikipedia (The Dalles, Oregon)

Source: Google

Source: Google

Source: Bonneville Power Administration



Recent guesstimates have placed Google's server count at more than 1 million. But new data on Google's energy use suggests that the company is probably running about **900,000 servers**. Google never says how many servers are running in its data centers. Aug 1, 2011

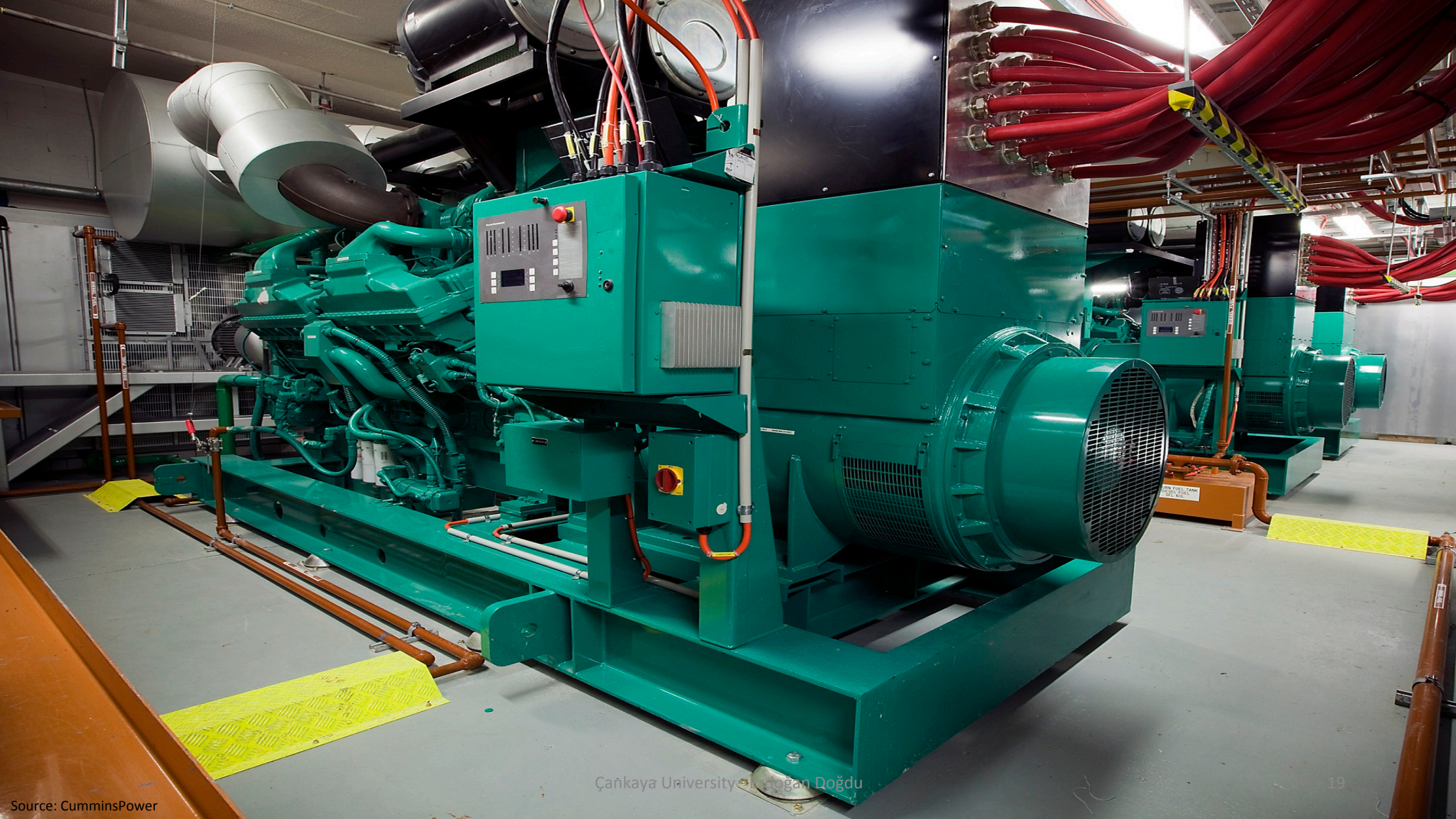
Source: Google

Source: Google

Source: Facebook

Source: Google

Source: Google



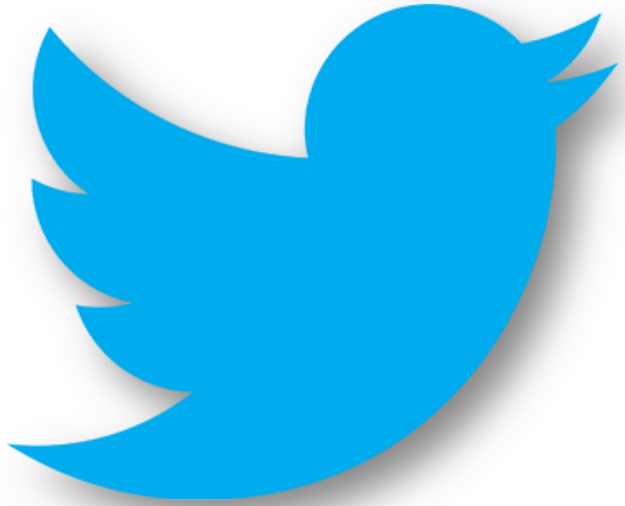
Source: CumminsPower

Source: Google

Twitter – topic trends

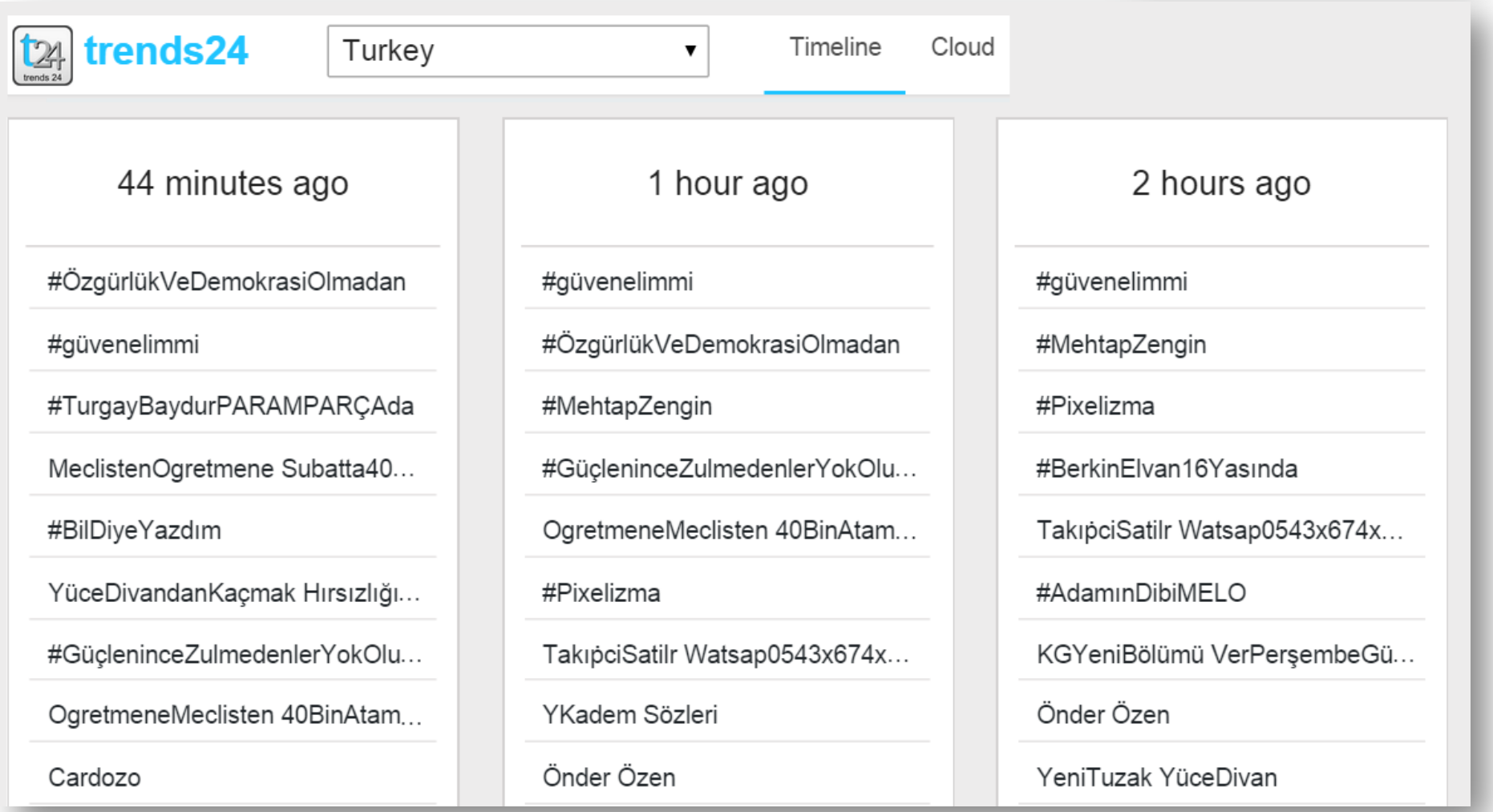
United States Trends · Change

#alreadyfailedresolutions
#ReplaceMovieLinesWithCoffee
Touch My Body
#DayOfPositivity
Ghost in the Shell
#RIPBelieberTaiane
Akinfenwa
#NationalBirdDay
Messi to Chelsea
Lol Blah

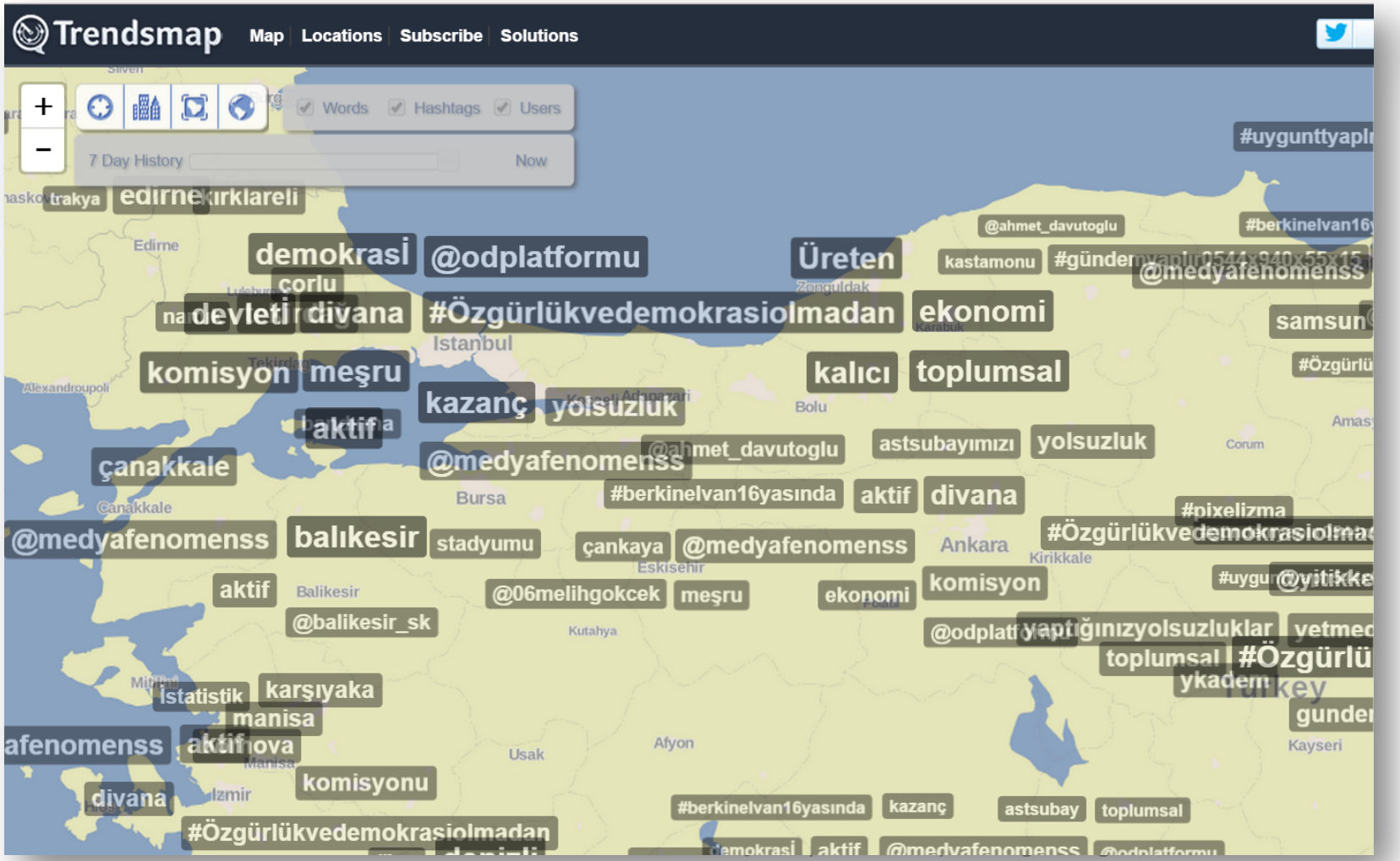


Twitter topic trends

trends24.in/turkey/



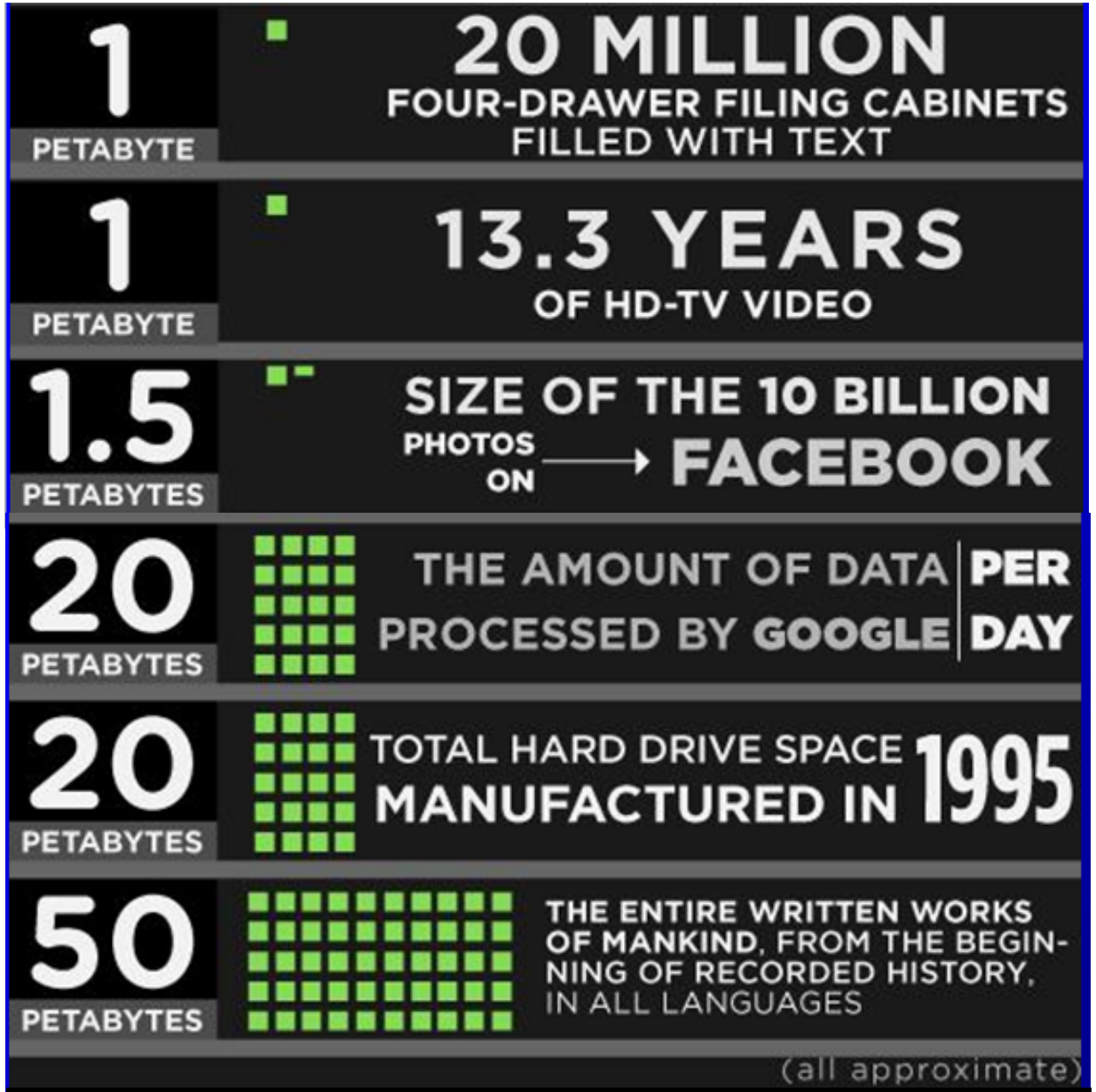
TrendsMap.com (on Twitter data)



trendsmap.com

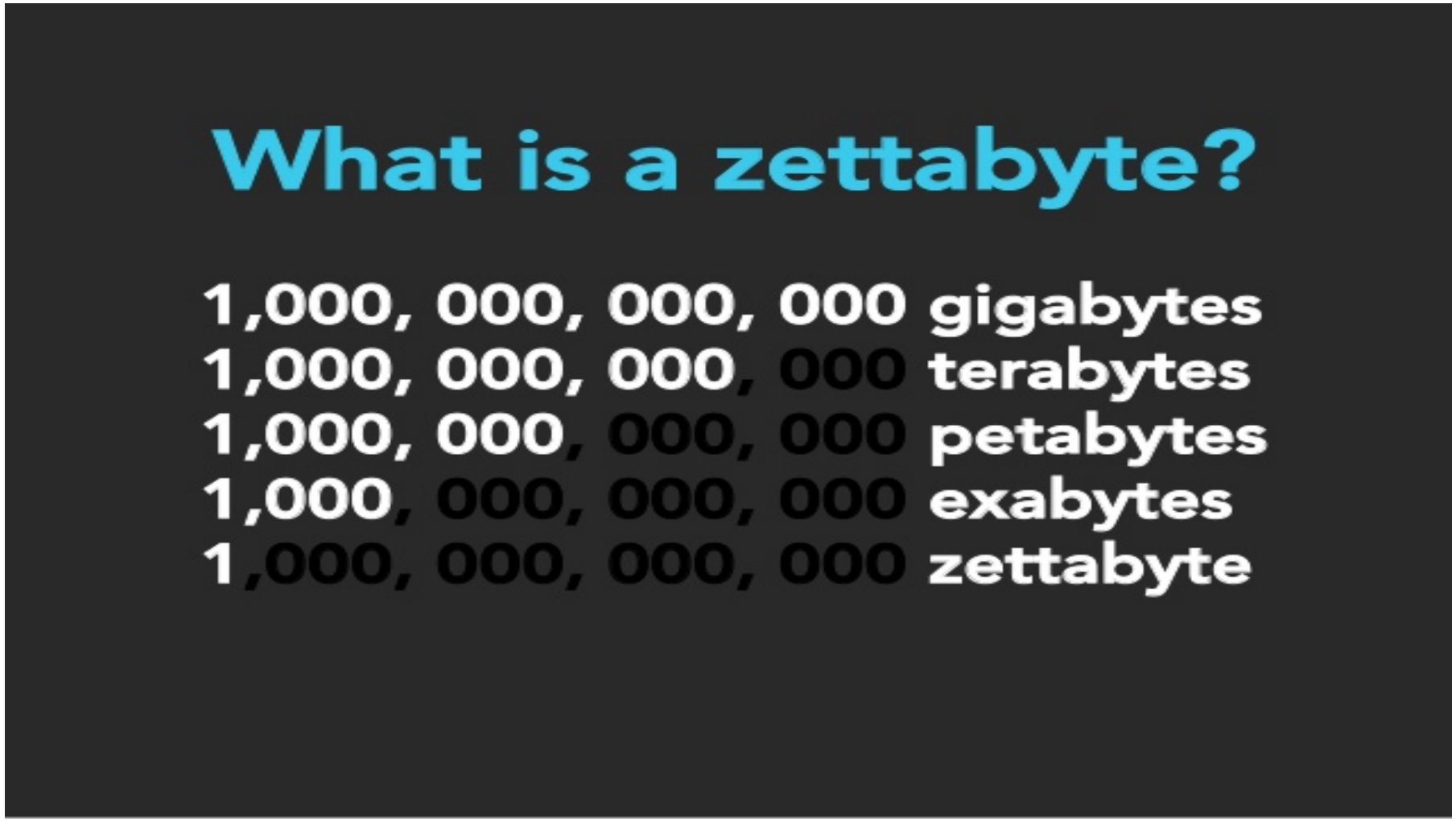
How big is Big Data?

- *1 byte = 8 bit*
- *1 MB = 10^6 B = 1 million byte*
- *1 GB = 10^9 B = 1 billion byte*
- *1 TB = 10^{12} B*
- *1 PB = 10^{15} B = 250.000 DVD*
- *1 EB = 10^{18} B*
- *1 ZB = 10^{21} B*
- *1 YB = 10^{24} B*



Facebook currently stores more than 100 petabytes of data.

Zettabyte


What is a zettabyte?
1,000, 000, 000, 000 gigabytes
1,000, 000, 000, 000 terabytes
1,000, 000, 000, 000 petabytes
1,000, 000, 000, 000 exabytes
1,000, 000, 000, 000 zettabyte

Google™ Processes 20 PB a day (2008)
Crawls 20B web pages a day (2012)
Search index is 100+ PB (5/2014)
Bigtable serves 2+ EB, 600M QPS (5/2014)

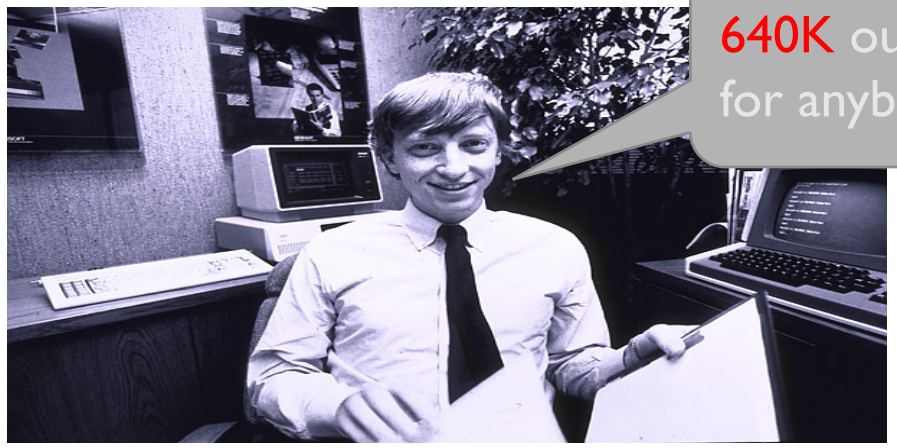
YAHOO!® 19 Hadoop clusters: 600 PB, 40k servers (9/2015)

ebay® Hadoop: 10K nodes, 150K cores, 150 PB (4/2014)

300 PB data in Hive + 600 TB/day (4/2014)

facebook

amazon web services™ S3: 2T objects, 1.1M request/ second (4/2013)

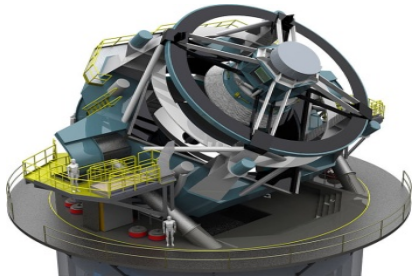


INTERNET ARCHIVE

JPMorganChase 150 PB on 50k+ servers running 15k apps (6/2011)

LHC: ~15 PB a year

CERN



LSST: 6-10 PB a year (~2020)

SKA: 0.3 – 1.5 EB per year (~2020)

How much data?

Source: http://lintool.github.io/UMD-courses/bigdata-2015-Spring

Source: Wikipedia (Everest)

Science

Data-oriented (e-science)

Maximilien Brice, © CERN

Engineering

Efficient use of data

Search, recommendation, prediction, …

Source: Wikipedia (Three Gorges Dam)

Knowing customers

Data → Knowledge → Competitive advantage

Ecommerce

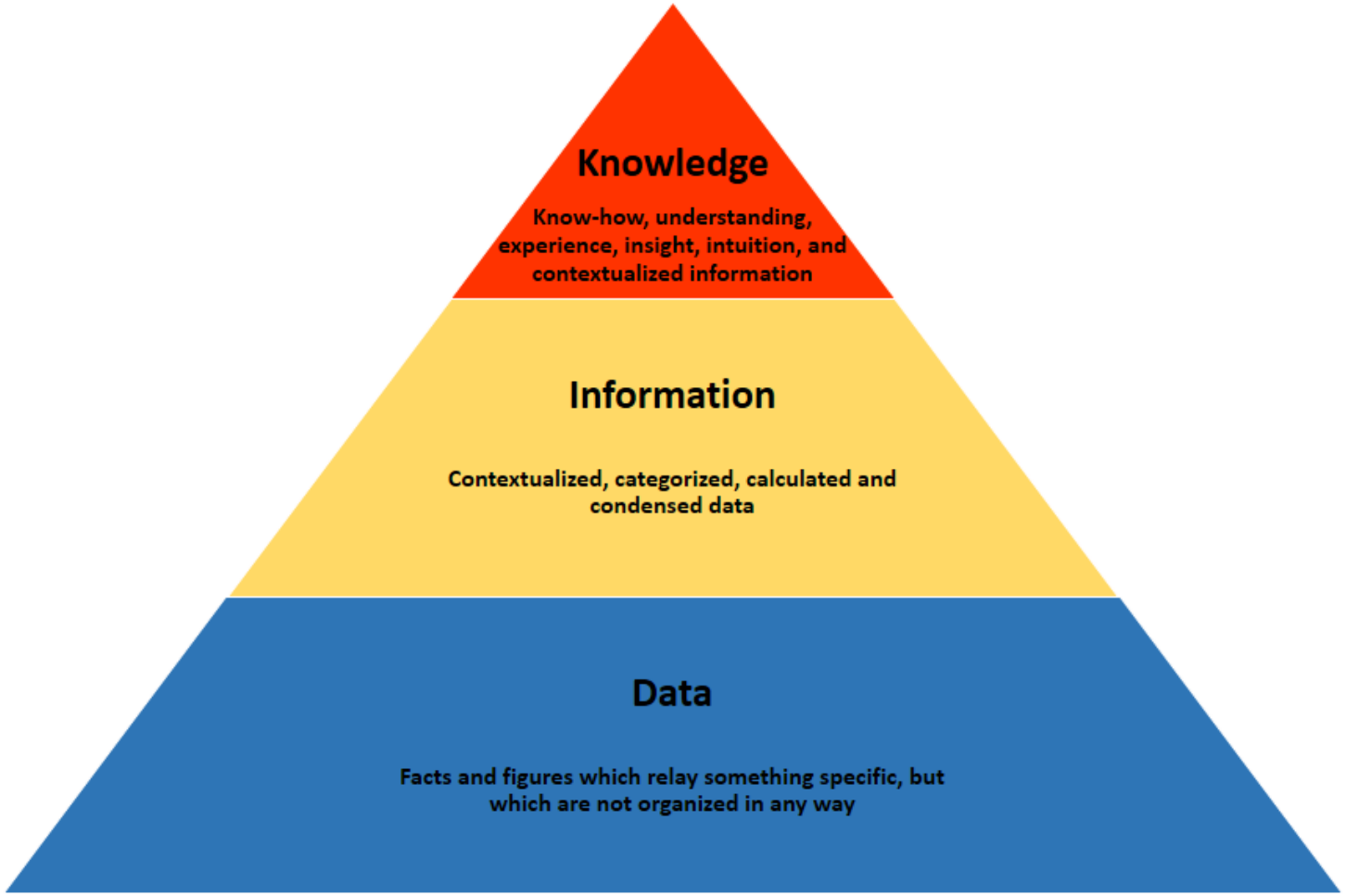
Source: wikiedia (Shinjuku, Tokyo)

Society

To know individuals/society

Computational social sciences

Data → Information → Knowledge



Data Analytics / Data Science

- Data → Information → Knowledge
- Finding patterns
- Classification
- Prediction
- Data mining
- Business intelligence
- Big data analysis
  - Applying known methods on big data in parallel

Web intelligence using big data

- Online advertisement – predicting interest
- Consumer sentiment – predicting behavior
- Detecting events – predicting impact
- Intelligent question answering – Watson, Google knowledge graph
- Categorizing, recognizing people, faces, people
- Intelligent public services – smart grids, water distribution, etc.
- Analysing **all** email and watching Web activity – predicting terrorists
- …

Big data resources

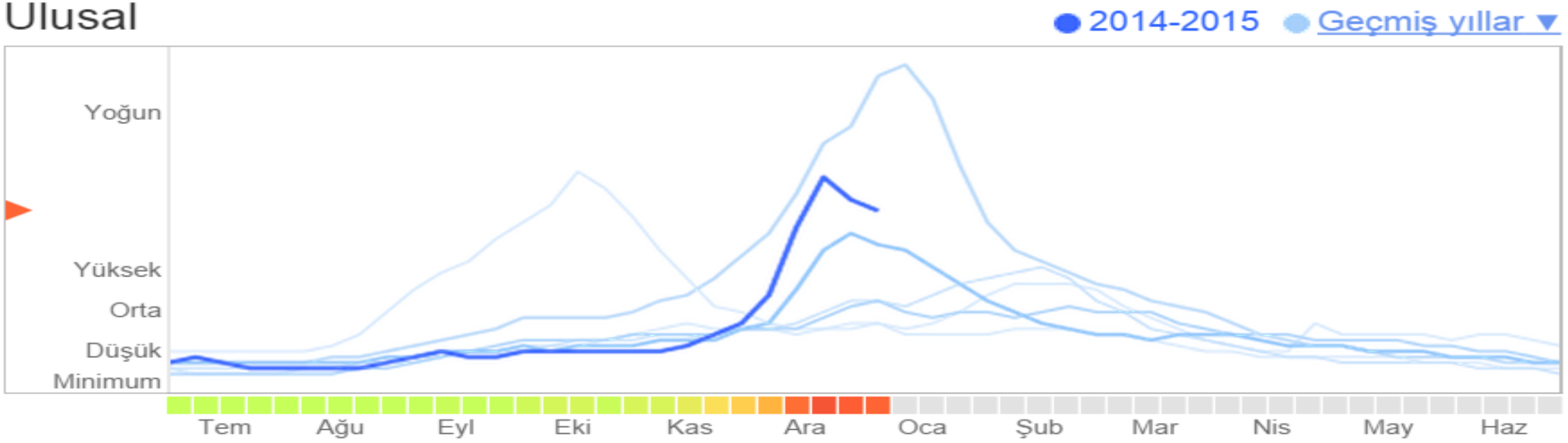
- People
  - Social media, web, blogs, forums, …
- Sensors / devices
  - Smart phones (GPS, …), industrial tools/robots, smart meters/grids (utilities), cars (200+ sensors), …
- Internet of Things (IoT) (30+ billion by 2020)
- Web of Things

Google Trends

Grip trendlerini araştırın - ABD

Belirli arama terimlerinin, grip faaliyetlerini izlemek için iyi göstergeler olduğunu fark ettik. Google Grip Trendleri, grip faaliyetlerini öngörebilmek için toplu Google arama verilerini kullanmaktadır.
Daha fazla bilgi edinin »

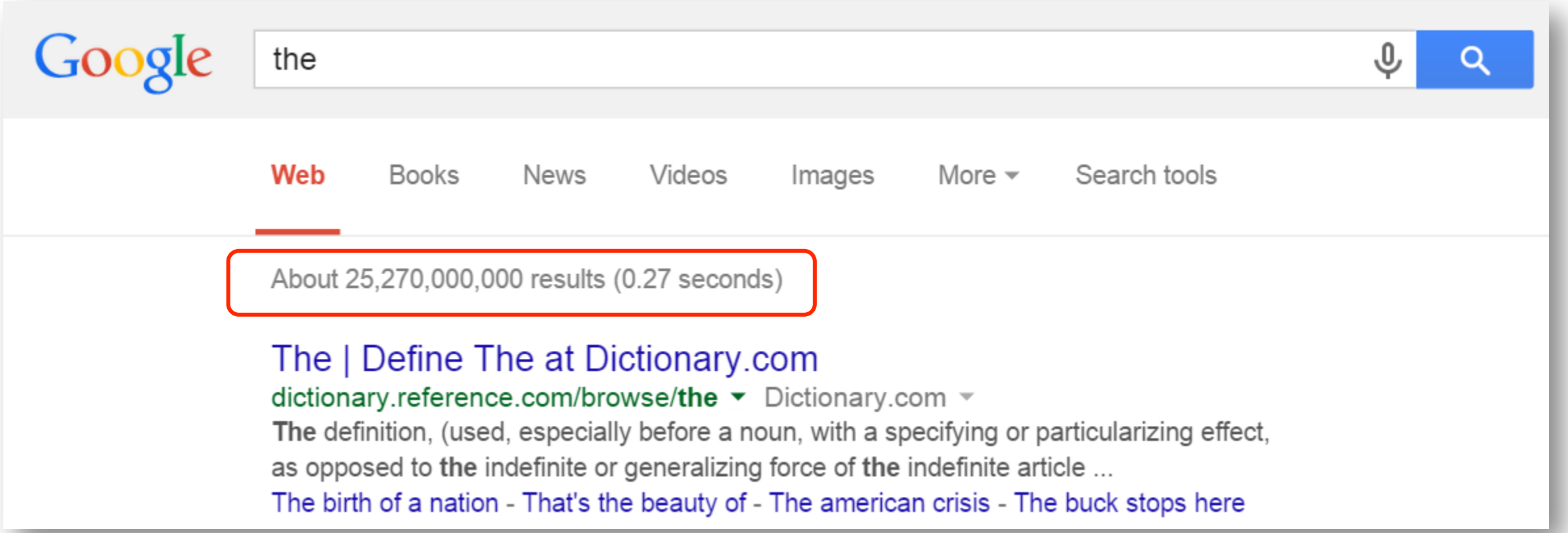
Ulusal



www.google.org/flutrends/us/#US

Where is Big Data?

- Web pages. How many?



Web in numbers

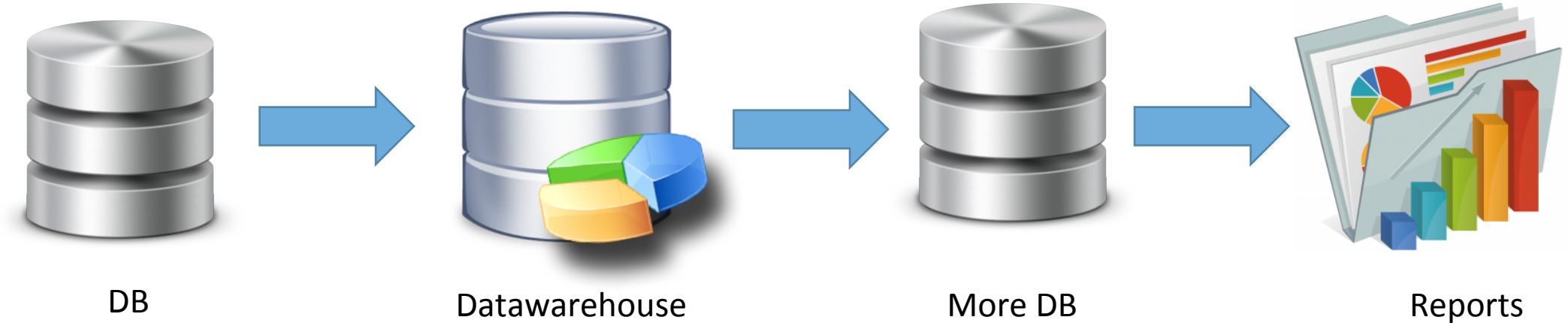
- Facebook, 1 billion users (Sep 2013)
- Twitter, 200 million users, 400 million tweets daily (60% from mobile devices) (Sep 2013)
- Google, 100 billion queries a month (May 2013)

In constrast, typical large enterprises:

- 5.000-50.000 servers
- Terabytes of data, millions of tx/day

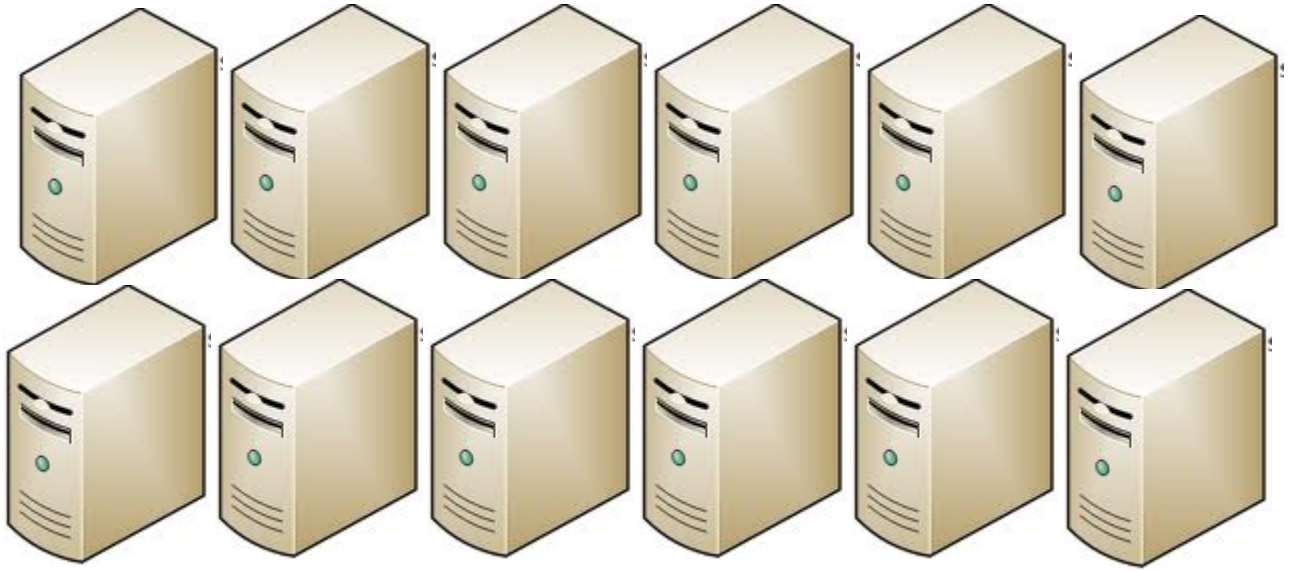
Big data technology

- Traditional «**business intelligence**» using databases



Big data technology

- Facebook, Twitter, LinkedIn, eBay, Amazon did not use «traditional databases» for big data
  - Massive **parallelism**
  - **Map-Reduce** paradigm



Big Data Analytics

- Data collection
- Data storage
- Data management
- Data analysis

Big Data Jobs

- **10 hot job titles** that did not exist 5 years ago
- ***LinkedIn*** *study on* ***259 million members*** *(November 2013)*

1. iOS Developer
2. Android Developer
3. Zumba Instructor
4. Social Media Intern
5. **Data Scientist**
6. UI/UX Designer
7. **Big Data Architect**
8. Beachbody Coach
9. Cloud Services Specialist
10. Digital Marketing Specialist

- http://talent.linkedin.com/blog/index.php/2014/01/top-10-job-titles-that-didnt-exist-5-years-ago-infographic

Top Jobs (at the end of 2014)

- LinkedIn, 330 million members, released the **top 25 skills of 2014** members got hired for and recruiters searched for:
  1. Statistics Analysis and Data Mining
  2. Middleware and Integration Software
  3. Storage Systems and Management
  4. Network and Information Security
  5. SEO/SEM Marketing
  6. Business Intelligence
  7. Mobile Development
  8. Web Architecture and Development Framework
  9. Algorithm Design
  10. Perl/Python/Ruby
- http://blog.linkedin.com/2014/12/17/the-25-hottest-skills-that-got-people-hired-in-2014/

LinkedIn Top Skills Summary

- **Data and cloud reign supreme:** I smell a dynasty in the making! **Cloud and distributed computing** has remained in the **#1 spot** for the past two years and is the **Top Skill** on almost every — including France, Germany, India, Ireland, Singapore, the U.S., and Spain. Following closely on its heels is **statistical analysis and data mining**, which came in **#2 last year**, and #1 in 2014. These skills are in such high demand because they're at the cutting edge of technology. Employers need employees with cloud and distributed computing, statistical analysis and data mining skills to stay competitive.
- https://blog.linkedin.com/2016/10/20/top-skills-2016-week-of-learning-linkedin

Kariyer.net top trends

- Jan 5th, 2015
  - SQL: 934 jobs
  - Java: 713
  - C#: 432
  - C++: 142
  - Hadoop: 13
- Feb 14, 2017
  - SQL: 833 jobs
  - Java: 442
  - C#: 434
  - C++: 192
  - Hadoop: 21
  - Spark: 18

Data scientists on demand

UC Berkeley Breeds Data Scientists Online: $60K, 18 Months

Want a data science Master's degree? UC Berkeley's $60,000 online program will make a data scientist out of you in 18 months.

The bad news about the University of California at Berkeley's new Master of Information and Data Science (MIDS) program is that it's expensive. Really expensive. The good news is that since data scientists are in such high demand, graduates are pretty much guaranteed a job with a generous income.

http://www.informationweek.com/big-data/big-data-analytics/uc-berkeley-breeds-data-scientists-online-$60k-18-months/d/d-id/1278764

MIT Big Data Class

- 545 USD
- 3500 students

MIT's 6-Week Big Data Online Class Wins Fans

Six-week data science class, just $545, is a hit with thousands of professionals who can't quit their day jobs, school officials say.

What's the best way to learn about big data? An online program may be the preferred option for students who can't attend on-campus classes during the day or evening. Schools increasingly are catering to the needs of the online student who, unlike most undergrads, tends to have a full-time job.

Massachusetts Institute of Technology (MIT) is adding two new sessions to its online big data course taught by its Computer Science and Artificial Laboratory (CSAIL) faculty. Last year's inaugural session proved a big success, drawing more than 3,500 participants from 88 countries, the school said.

http://www.informationweek.com/big-data/big-data-analytics/mits-6-week-big-data-online-class-wins-fans/d/d-id/1316148

Course objectives

- Understand big data **concepts**
- Learn **distributed data processing algorithms**, techniques and methods on big data.
- Learn (some) **data analysis methods** on big data.

Learning outcomes

- Write **map/reduce** methods to process big data
- Use advanced **distributed data processing techniques** and tools on big data
- **Develop map/reduce based applications** for processing big data
- Understand big data **analysis methods** and techniques
- Choose appropriate big data analysis methods for specific big data problems and apply

Books

- **Data-Intensive Text Processing with MapReduce**
  Jimmy Lin and Chris Dyer
  Morgan & Claypool Publishers, 2010
  http://lintool.github.io/MapReduceAlgorithms
- **Mining Massive Data Sets**, 2nd Ed.
  Jure Leskovec, Anand Rajaraman, Jeff Ullman
  http://mmds.org

Other books

- **Hadoop: The Definitive Guide, 3rd Edition**, by Tom White
  - O'Reilly Media/Yahoo Press
  - Print ISBN: 978-1-4493-1152-0 | ISBN 10: 1-4493-1152-0
  - Ebook ISBN: 978-1-4493-1151-3 | ISBN 10: 1-4493-1151-2
- MapReduce Design Patterns, by Donald Miner and Adam Shook
  - O'Reilly Media
  - Print ISBN: 978-1-4493-2717-0 | ISBN 10: 1-4493-2717-6
  - Ebook ISBN: 978-1-4493-4197-8 | ISBN 10: 1-4493-4197-7

Resources

- *Harness the Power of BigData*, McGraw-Hill, 2013 http://public.dhe.ibm.com/common/ssi/ecm/en/imm14100usen/IMM14100USEN.PDF
- *Understanding the BigData*, McGraw-Hill, 2012 http://public.dhe.ibm.com/common/ssi/ecm/en/iml14296usen/IML14296USEN.PDF
- *Hadoop for Dummies*, Robert D. Schneider, Wiley, 2012 http://public.dhe.ibm.com/common/ssi/ecm/en/dcm03002usen/DCM03002USEN.PDF
- *Hadoop Documentation,* http://hadoop.apache.org/docs/current
- *Big Data University*, http://bigdatauniversity.com

Lecture topics

- Introduction
- MapReduce Algorithm Design
- Technologies: HDFS, Hadoop, Pig, Hive, Hbase
- Spark
- Analyzing Text
- Analyzing Graphs
- Analyzing Relational Data
- Data Mining
- Applications

Grading

Work	%
Assignments	20%
Exam (quiz,midterm)	40%
Project/Research	40%

The End