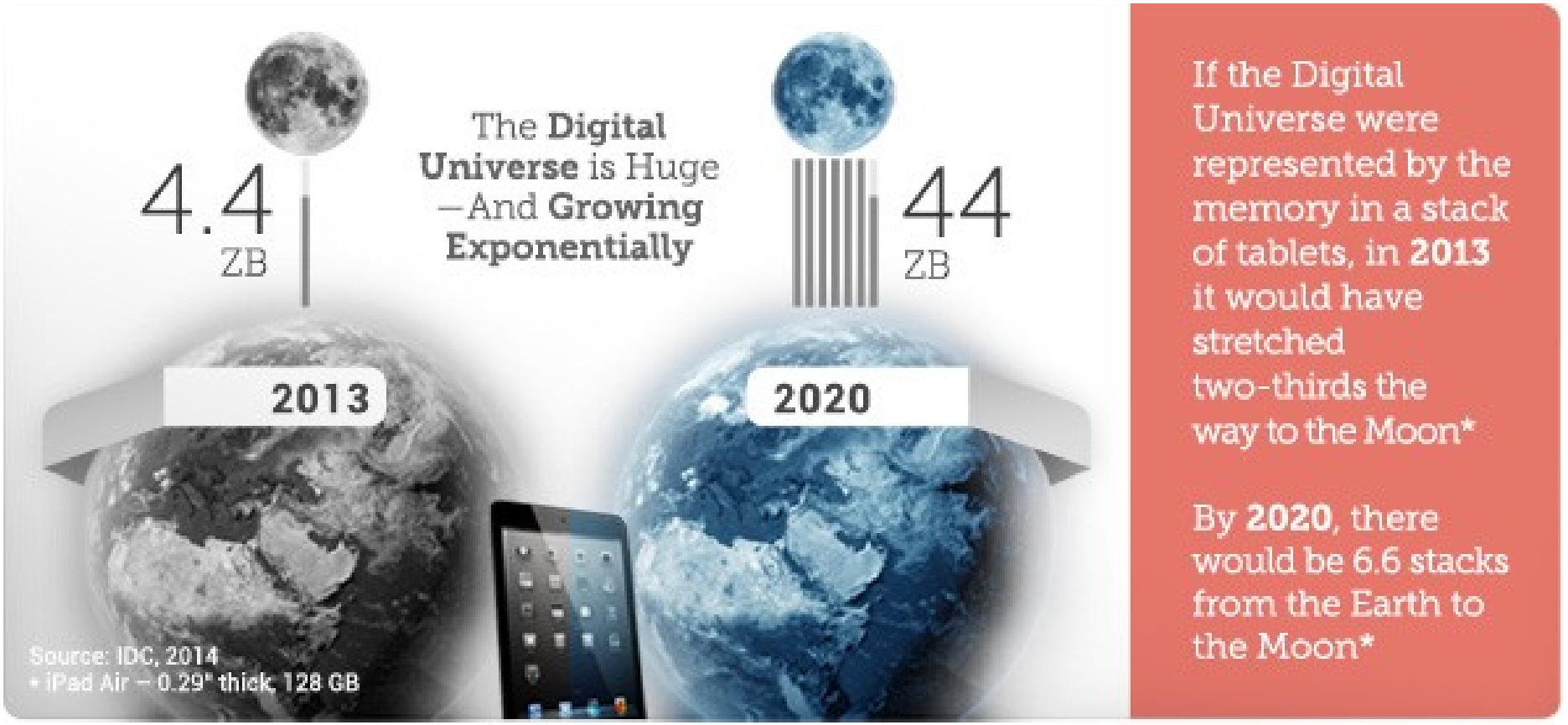
What is ”Big Data”?

- A fuzzy concept: no generally accepted formal definition of what ”big data” means or is.

- Excerpts from answers to a Berkeley University survey of 40 experts:
  - ”a quantity of data that’s so large that traditional approaches to data analysis are doomed to failure”
  - ”a lot of different data coming fast and in different structures”
  - ”a buzzword to mean anything related to data analytics or visualization”
  - ”analysis for data that’s really messy or where you don’t know the right questions or queries to make”
  - ”you capture and store data on a very large volume ... in order to make sense of it later”
  - ”big data is data that breaks Excel”

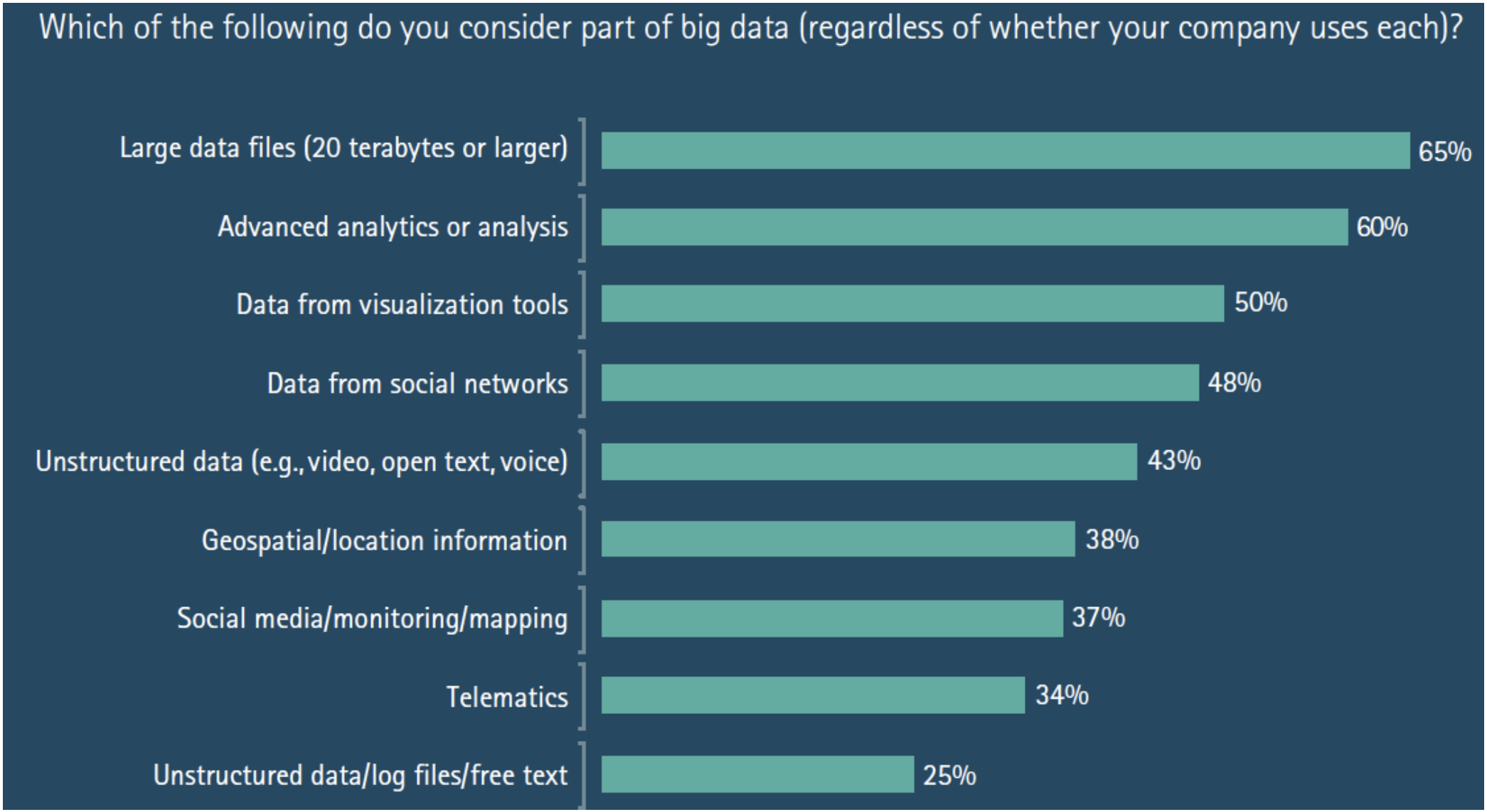
Units for measuring (literally) big data

- 1 terabyte (TB) = 10^{12} bytes = 1000 GB.
- 1 petabyte (PB) = 10^{15} bytes = 1 million GB.
- 1 exabyte (EB) = 10^{18} bytes = 1 billion GB.
- 1 zettabyte (ZB) = 10^{21} bytes = 1000 billion GB.
- 1 yottabyte (YB) = 10^{24} bytes = million billion GB.



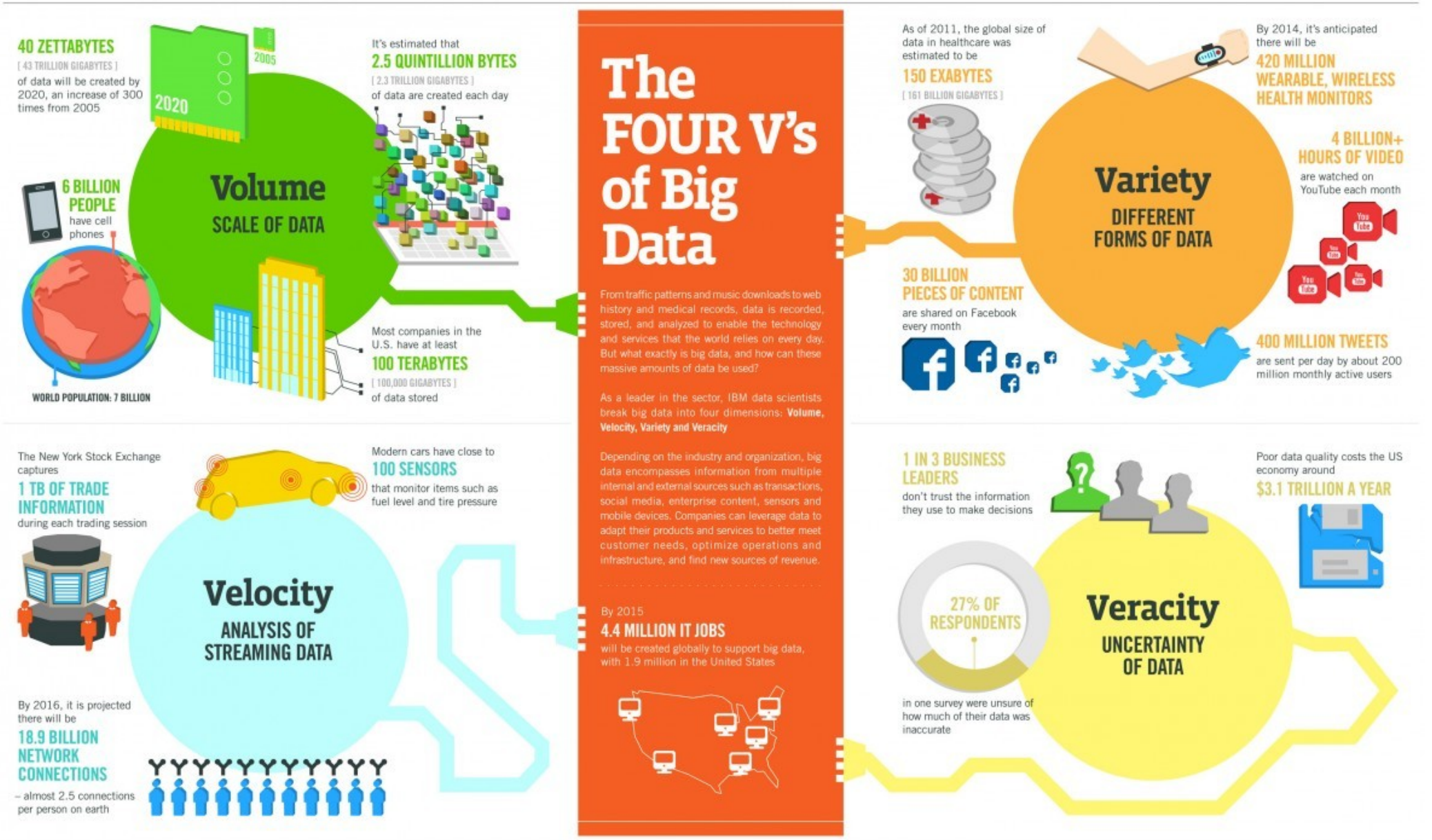
What is ”Big Data”?...

- A result from a Big Data survey by Accenture in 2014:



Characterizing Big Data by various V's

- A famous characterization of big data by IBM: **the four V's of big data**.



Characterizing Big Data by various V's...

- Later several more V's of big data have been proposed, such as:
  - **Variability**.
    - Variance in the meaning. Seemingly the same data can be interpreted differently depending on context (e.g. in case of natural language).
      - E.g. detecting which comments are true and which sarcastic.
  - **Visualization**.
    - One important aspect of (big) data analysis is how to present the results in an understandable form.
      - Visualization may also be helpful during the analysis (e.g. give hints about what to look and with what kind of methods).
  - **Value**.
    - The hidden potential value of big data (that may be realized by proper analysis of the data).

Examples of (big) data

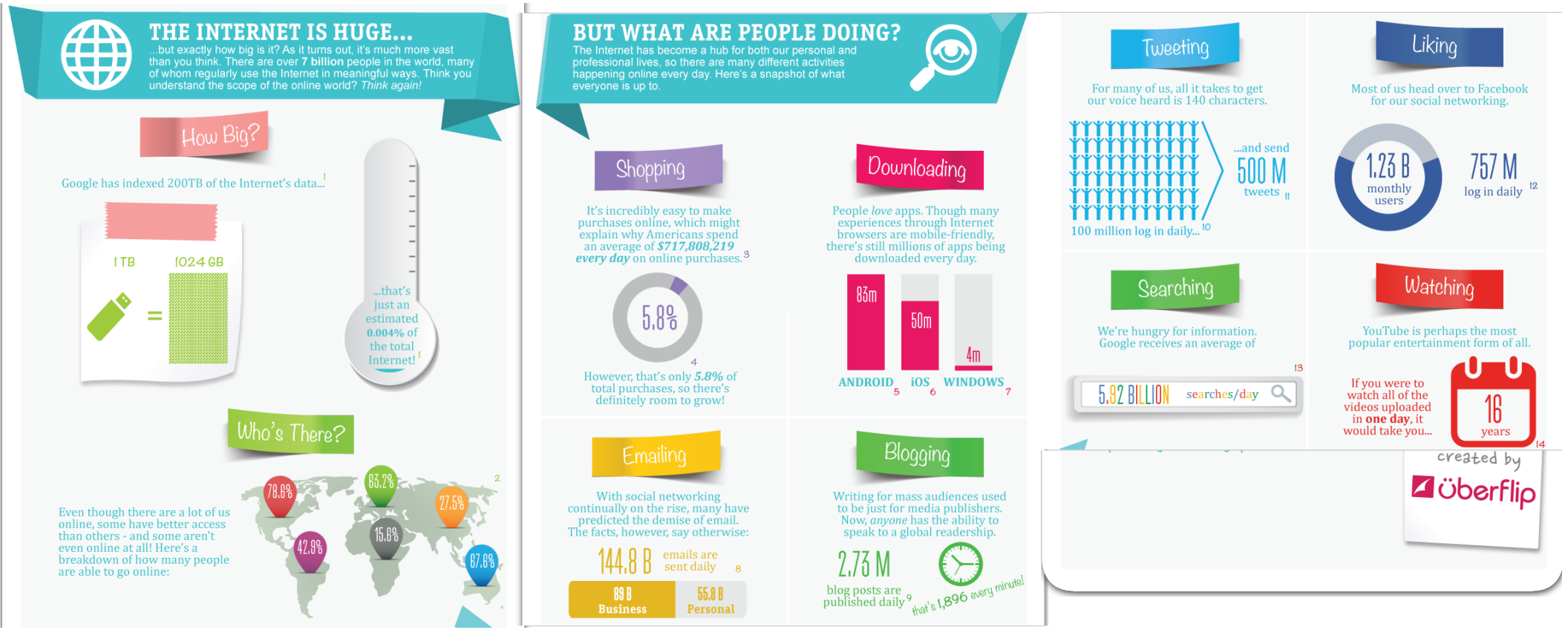
- A quote from 2013: ”90% of all data in the world has been generated in the last 2 years”.

- Sequencing and analyzing DNA genomes (e.g. for cancer research):
  - The genome of a single human: 3 billion DNA bases (nucleotides).
  - The first complete human genome sequence was sequenced in 2003.
    - Took 15 years and between 1-3 billion USD.
  - At present: a genome sequenced using ≈ one day and 10000 USD.
    - Genome sequencing seems to become routine → lots of data.

- Data from various physical sensors (e.g. in scientific experiments).
  - The CERN particle collider: ≈ 30 PB of data per year.
  - BRAIN research project (mapping the human brain’s function): could be **yottabytes** of data (1 yottabyte = 10^{24} bytes = million billion GB).
  - US national weather service: ≈ 7 PB of weather data per year.

Examples of (big) data...

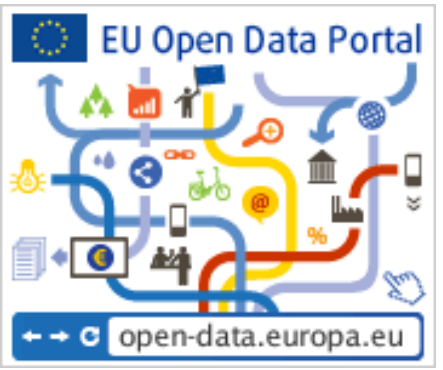
- The internet (as illustrated by Überflip in 2014):



Examples of (big) data: open data

- The global open data initiative: ”make Government data … available for anyone, anywhere to download … without charge”.

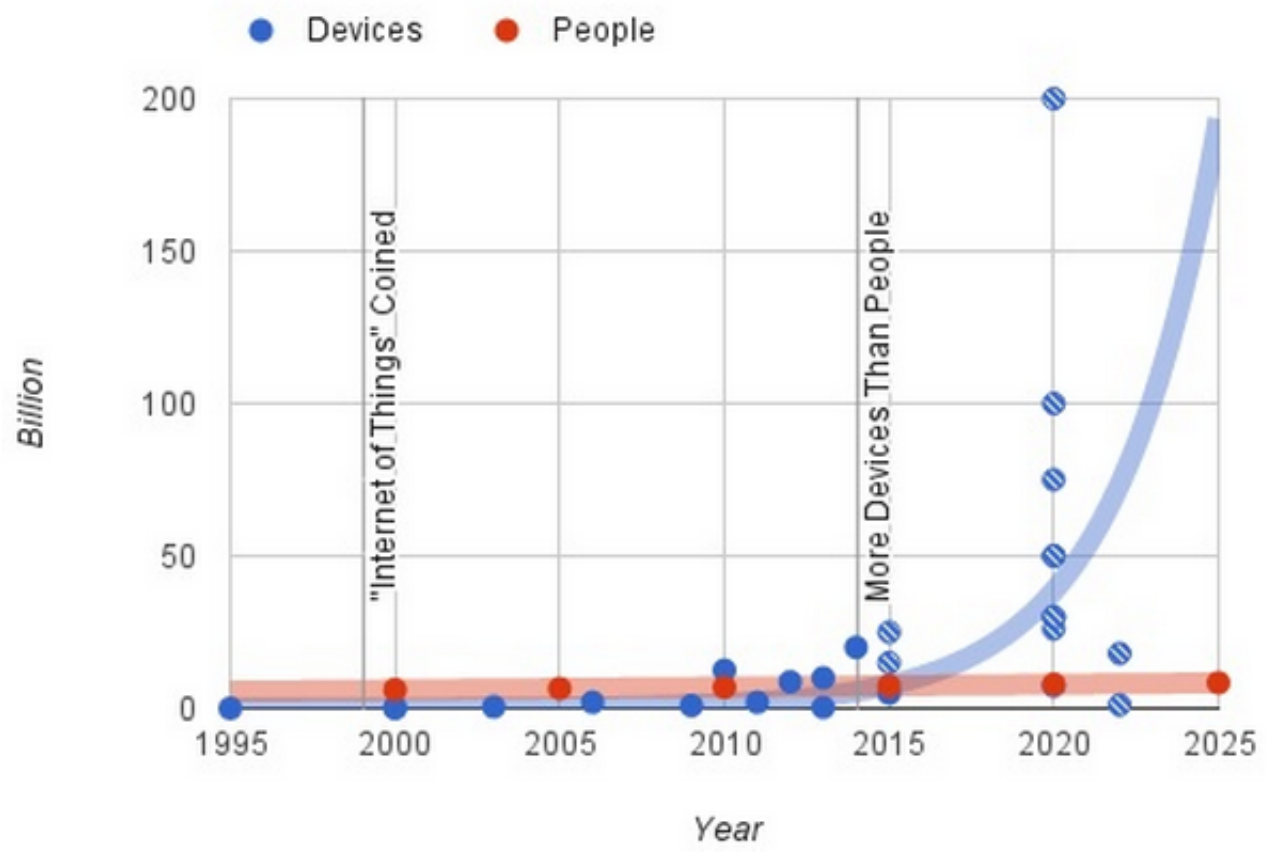
- Various governments etc. maintain listings of openly available data.
  - Finland: www.avoindata.fi
  - European Union: open-data.europa.eu
  - United States: www.data.gov
  - (and so on; too many to list...)



- For example the public bus transportation data in Tampere is open.
  - Includes timetables, real-time bus location information, etc.

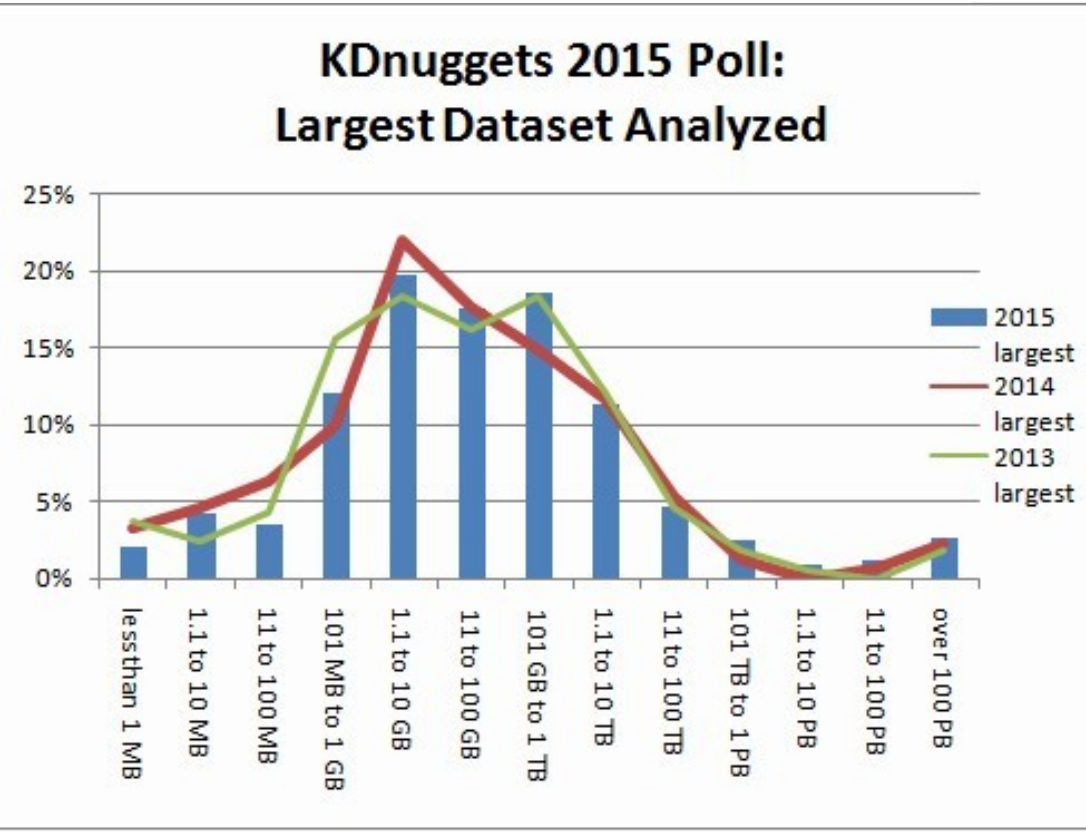
Examples of (big) data: Internet of Things (IoT)

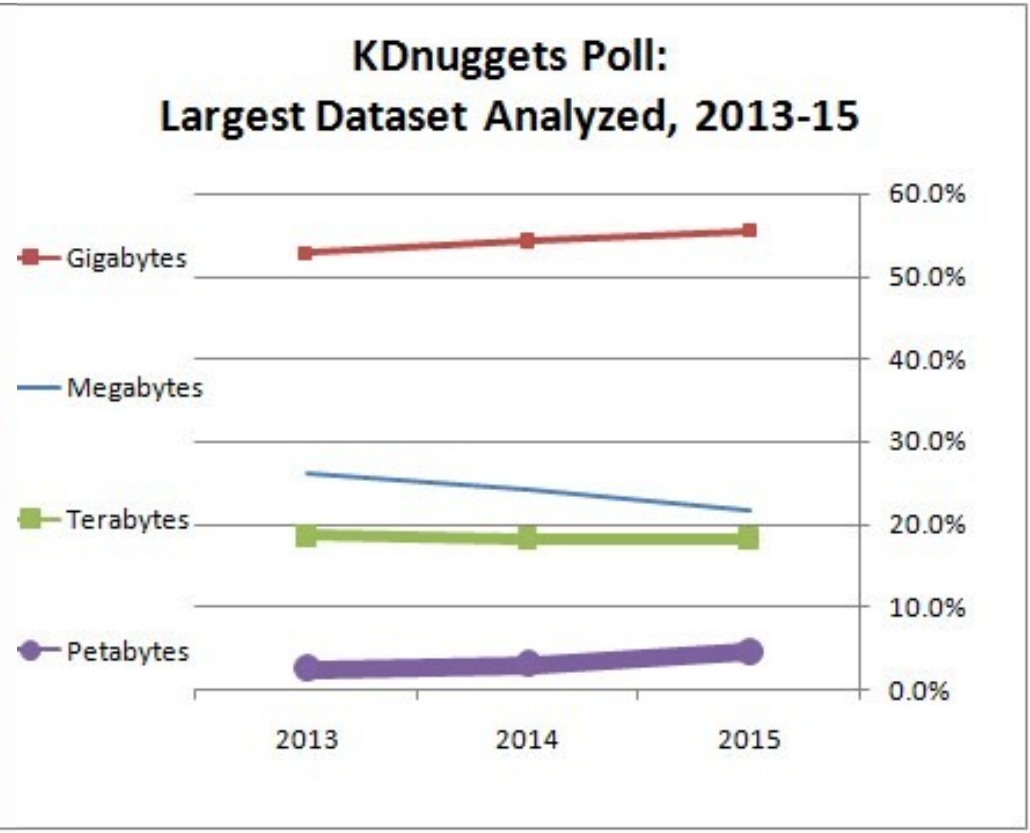
- Internet of Things: the emergence of more and more physical devices (many of which are not computers) that are connected to the internet.
  - In 2012: at least 1.3 billion devices were on the internet.
  - Brookings Institute: in 2020 there will be 50 billion connected devices!
    - The volume and usage of various types of sensor data will grow **fast**.



Big data trends

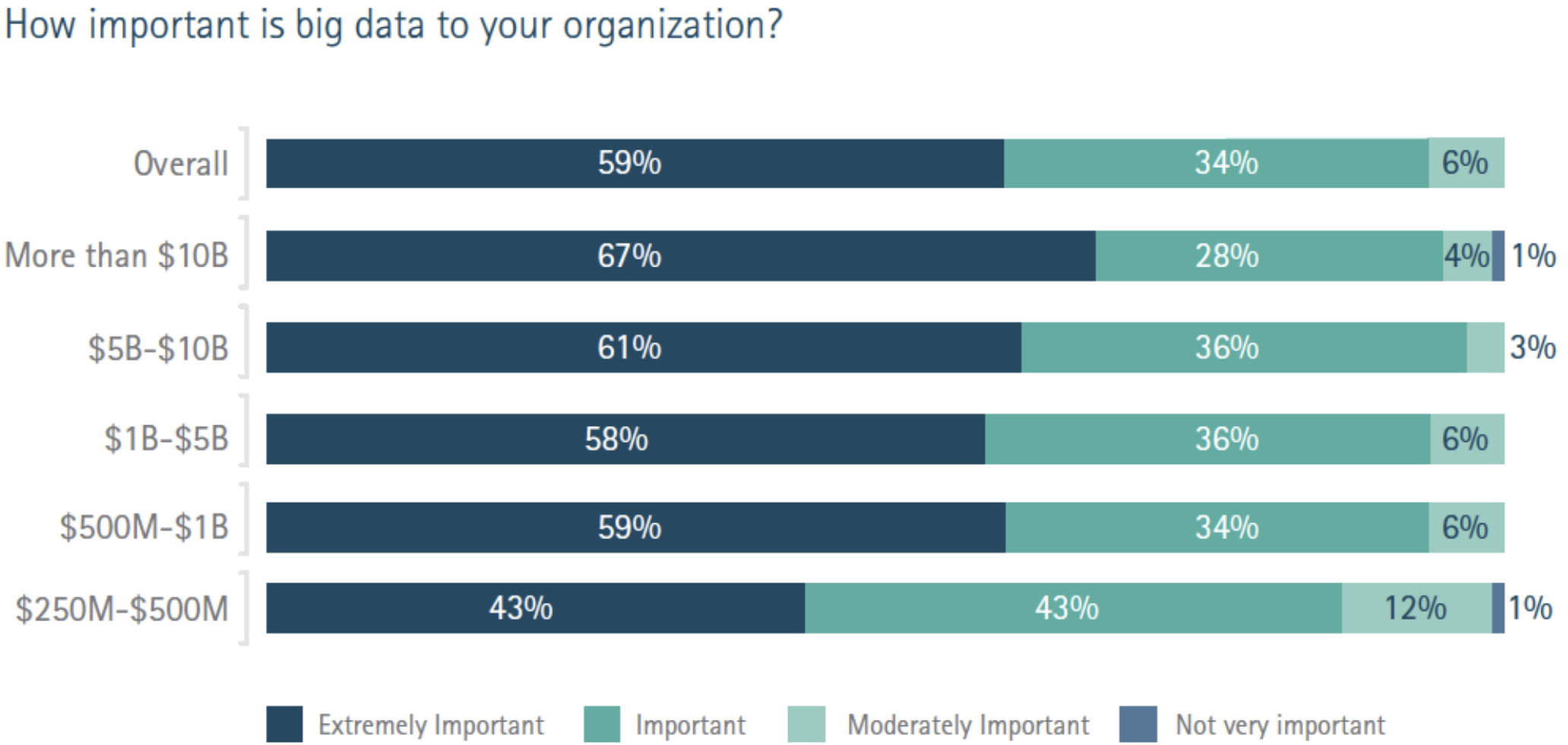
- A survey by KDnuggets regarding the sizes of analyzed data sets:
  - Roughly 23% at least 1 terabyte.
  - Almost 5% at least 1 petabyte.





Big data trends...

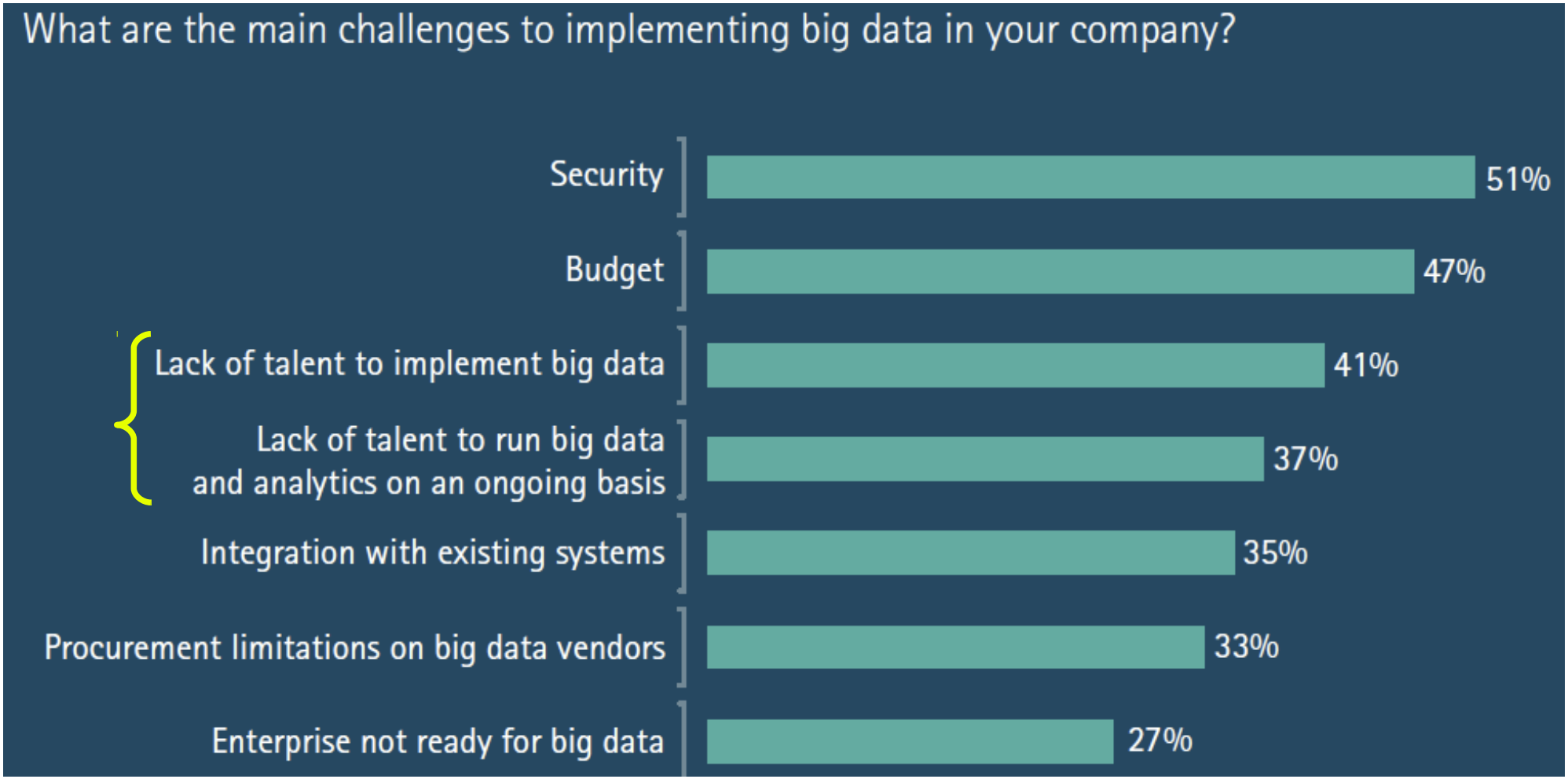
- Accenture: business interest in big data is very high.



- The answers in the table are grouped according to the market valuations of the surveyed companies.

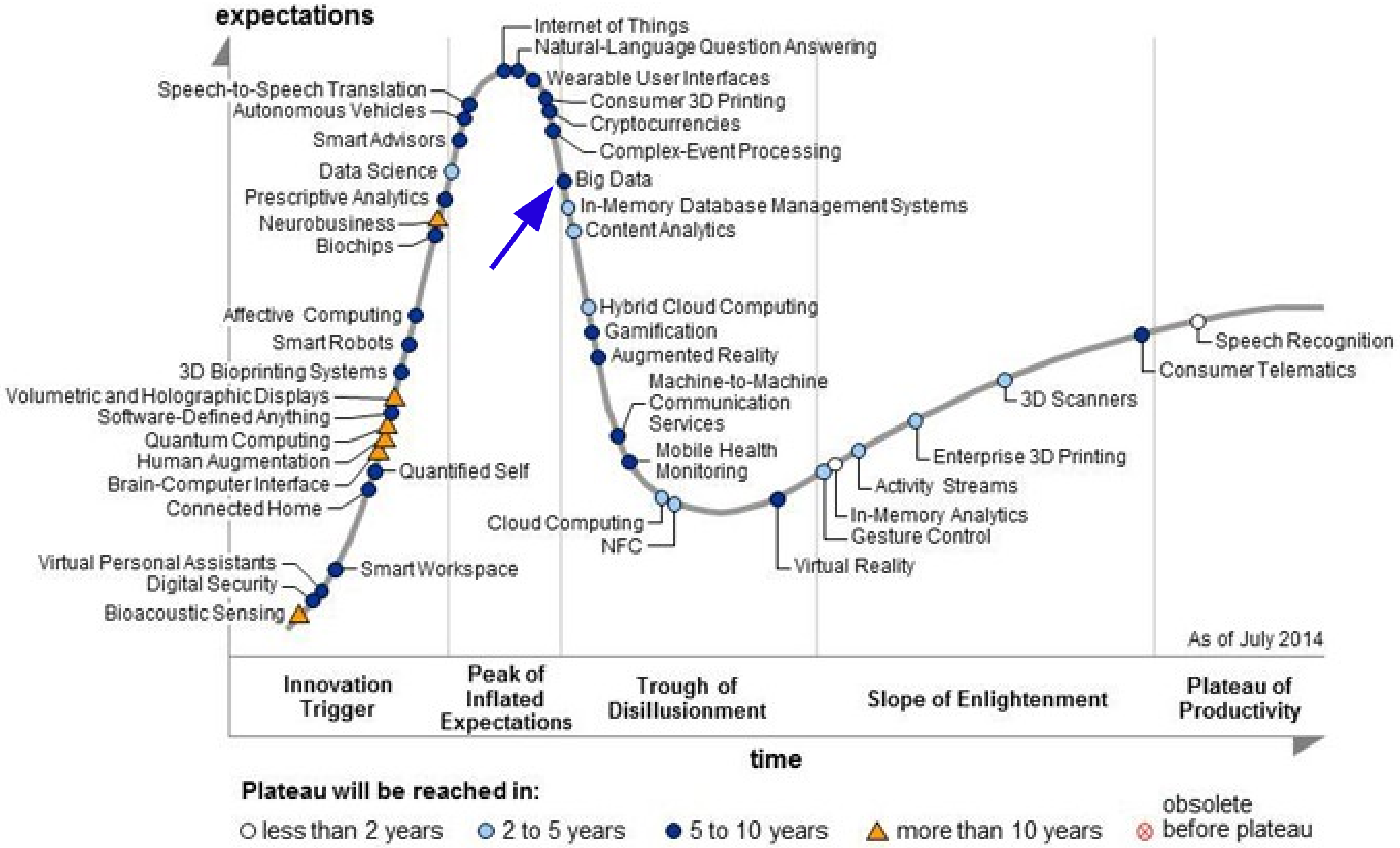
Big data trends...

- Demand for people with big data analyzing skills (aka ”data scientists”).
  - McKinsey: 2018 demand for data scientists 60% greater than supply?
  - Accenture survey in 2014: companies report lack of data scientists.



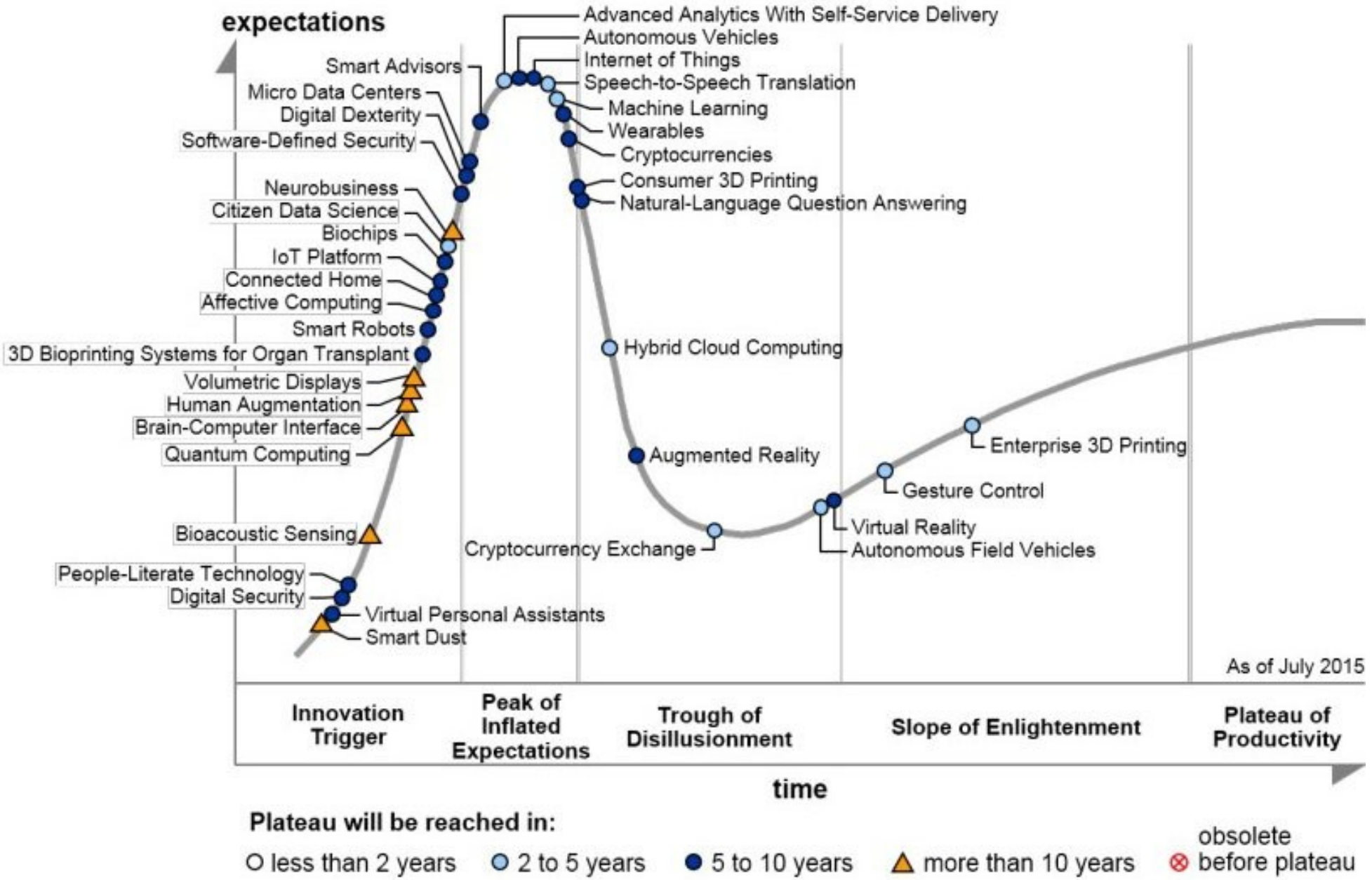
Big data trends...

- The Gartner hype cycle in 2014:



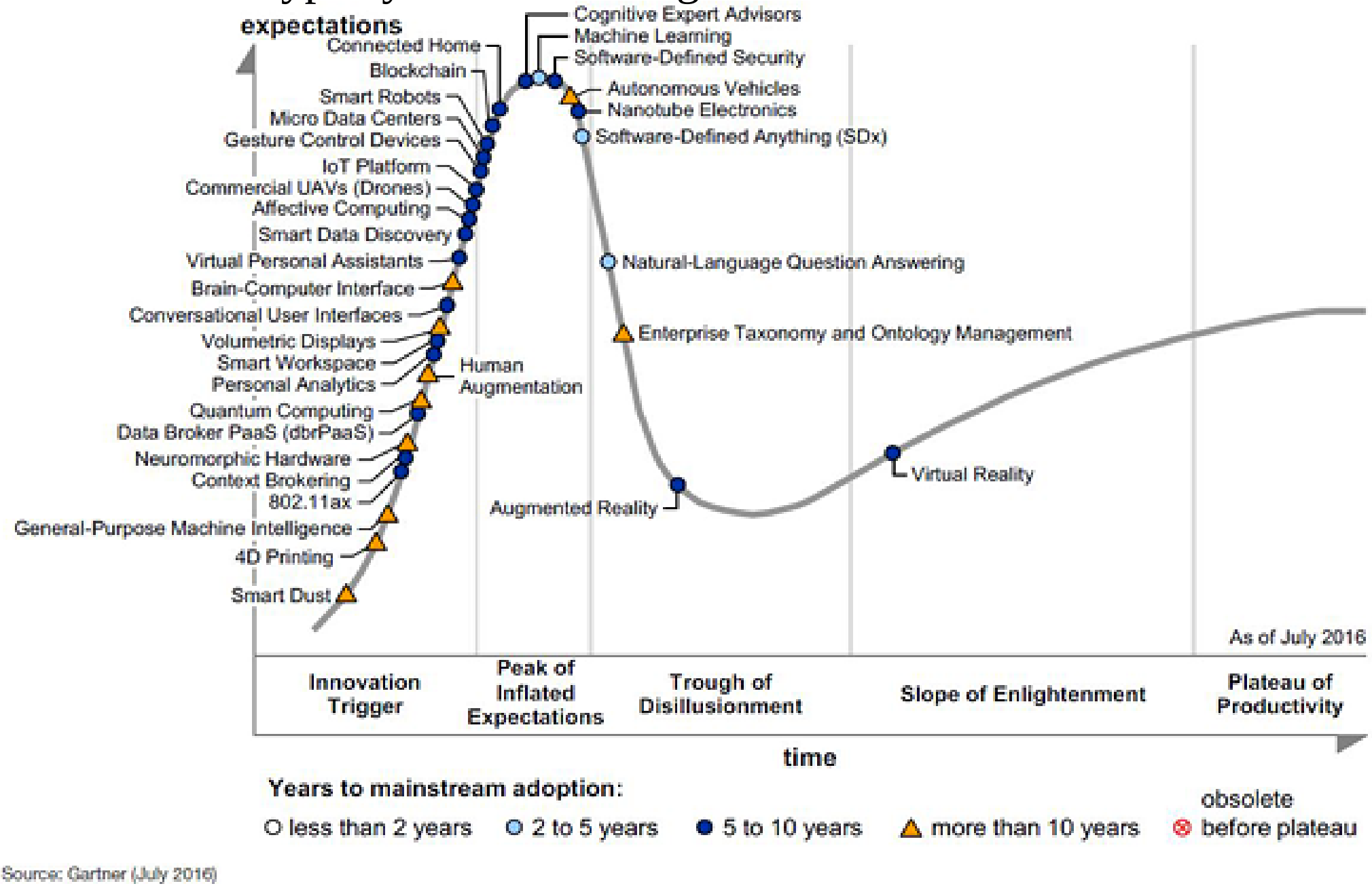
Big data trends...

- The Gartner hype cycle in 2015: big data...?



Big data trends...

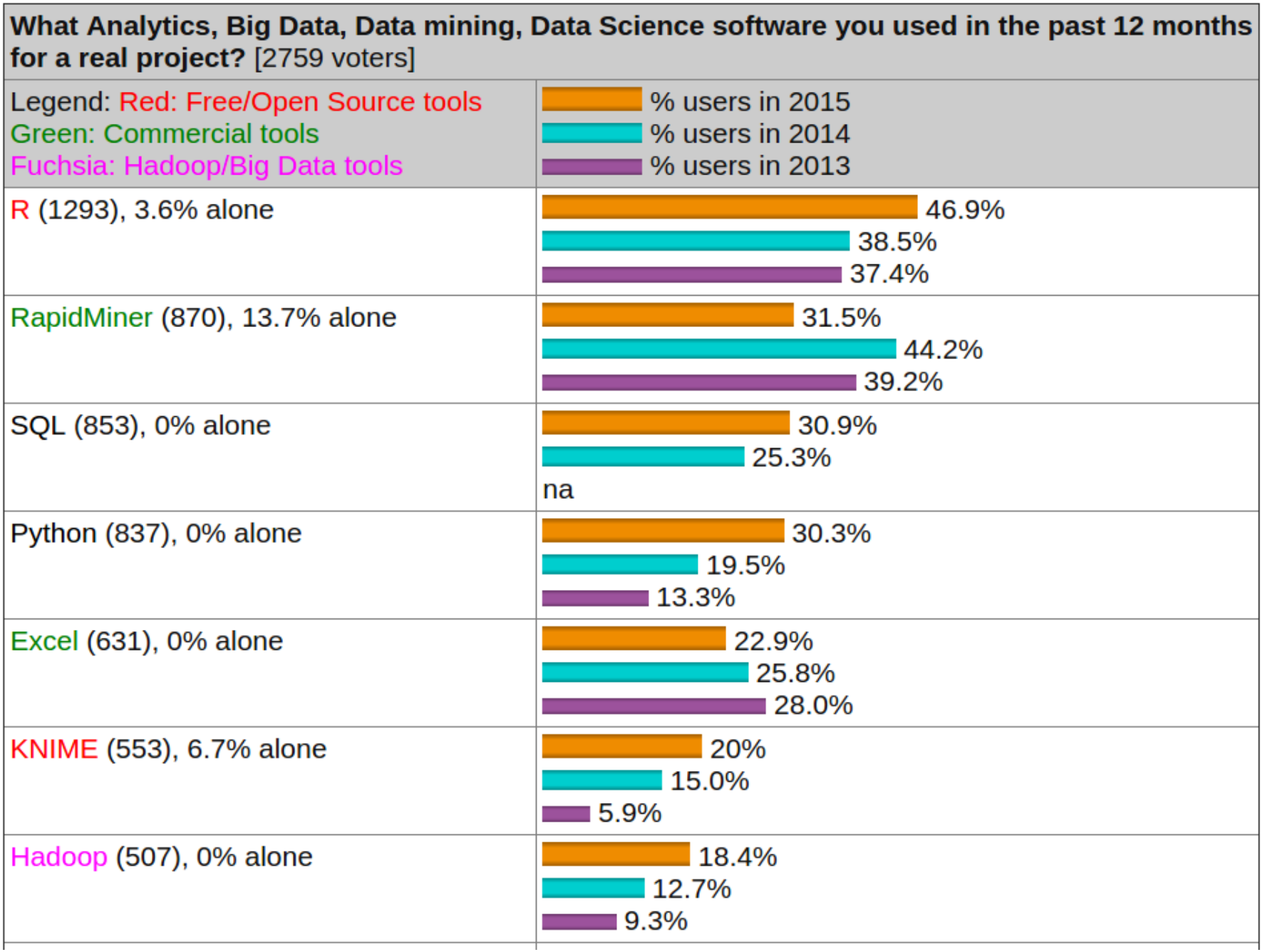
- The Gartner hype cycle in 2016: big data...?



Source: Gartner (July 2016)

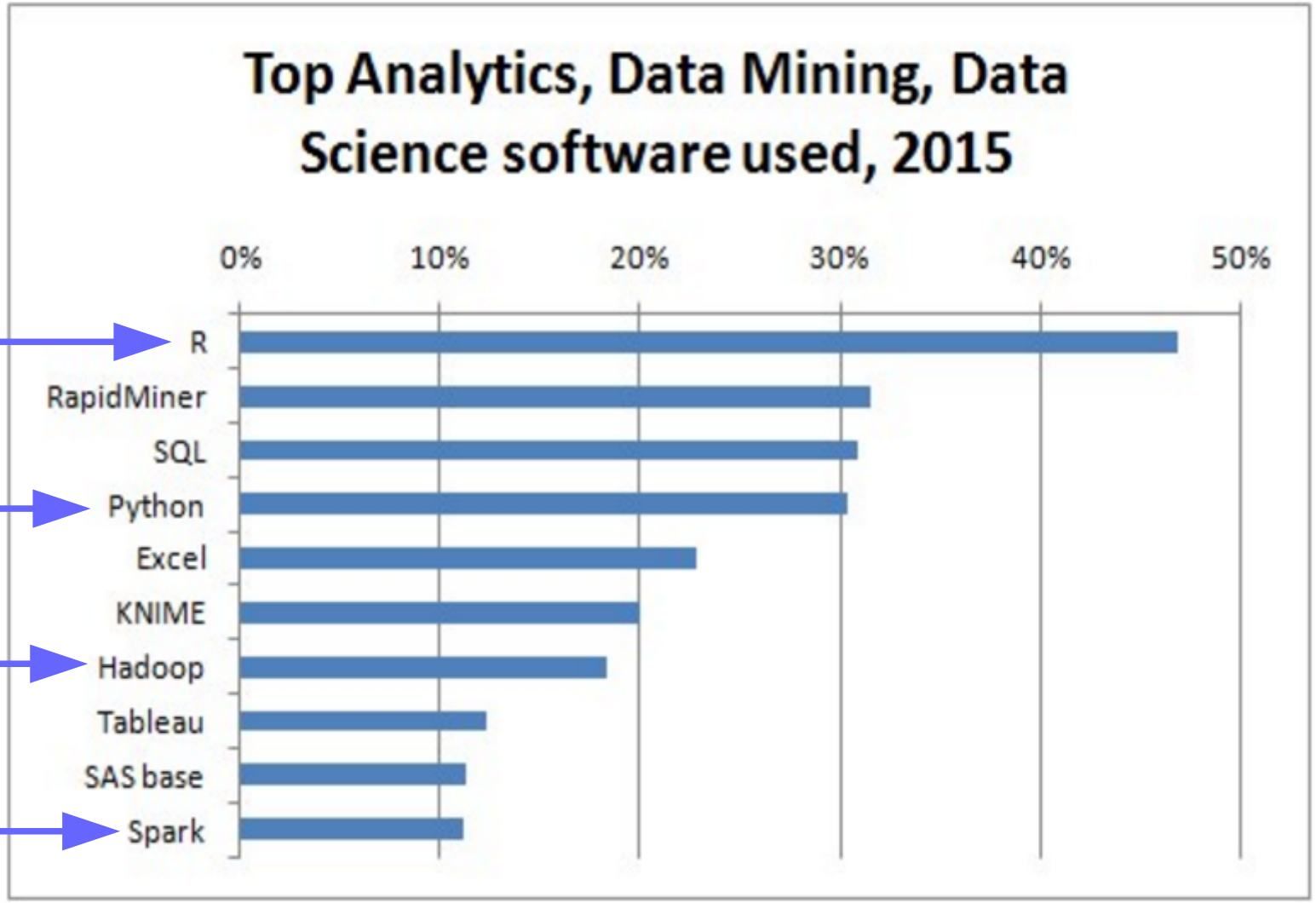
Data analysis tools / programming languages

- KDnuggets poll 2015:



Data analysis tools / programming languages...

- KDnuggets poll 2015:



- In this course we will mostly use Python and R programming languages.
  - And Hadoop and Spark distributed computation frameworks.
    - With these the main language is Python, but also a tiny bit of Java.

Python

- A high-level programming language credited with high productivity.
  - Relatively few lines of code.
    - Python needs only 20-35% of the lines of corresponding Java code?
  - The code is (typically) run directly with a code interpreter.
    - No need to separately compile the code.
  - The Python interpreter can also be used in an interactive manner.
    - The commands / statements are executed on-the-fly as you type them.
- Very extensive assortment of libraries is available.
  - Networking, text processing, machine learning, etc...
- Available for all major computing platforms. (www.python.org)
  - Note: Python is already preinstalled in Mac OS X and Ubuntu Linux!

Python...

- The popularity of Python as a general programming language, July 2016:

July 2016	July 2015	Programming Language	Ratings	Change
1	1	Java	19.804%	+2.08%
2	2	C	12.238%	-3.91%
3	3	C++	6.311%	-2.33%
4	5	Python	4.166%	-0.09%
5	4	C#	3.920%	-1.73%
6	7	PHP	3.272%	+0.38%
7	9	JavaScript	2.643%	+0.45%
8	8	Visual Basic .NET	2.517%	+0.09%
9	11	Perl	2.428%	+0.62%
10	12	Assembly language	2.281%	+0.75%
11	15	Ruby	2.122%	+0.74%
12	13	Delphi/Object Pascal	2.045%	+0.57%

- Ranking according to www.tiobe.com.

Python...

- The very first homework in this course: learn Python!

- In practice: all students need to complete the online Python course at www.codecademy.com/tracks/python.
  - The website estimates that the course takes ≈ 13 hours to complete.
  - Note that the Codecademy course teaches Python version 2.7.

- Complete the Python course by the deadline for exercise batch 1 (12.9).
  - One question in exercise batch 1 asks you to submit a screen capture of the list of completed skills in your Codecademy profile.

- **Note**: completing the Codecademy Python course is required regardless of whether you already have prior Python programming skills.