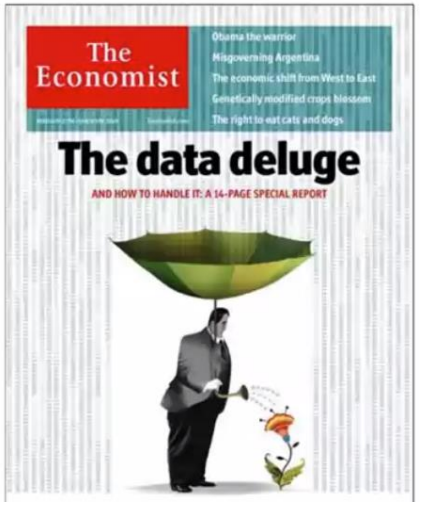
The Data Scientist's Toolbox

Eunbyeol Lee

2019.11.15

What is data science?



- Data science is using data to answer questions.

- Statistics, computer science, mathematics, data cleaning and formatting, and data visualization.

- A data scientist is broadly defined as someone who combines the skills of software programmer, statistician, and storyteller/artists to extract the nuggets of gold hidden under mountains of data.

http://www.economist.com/node/15557443

What is big data?

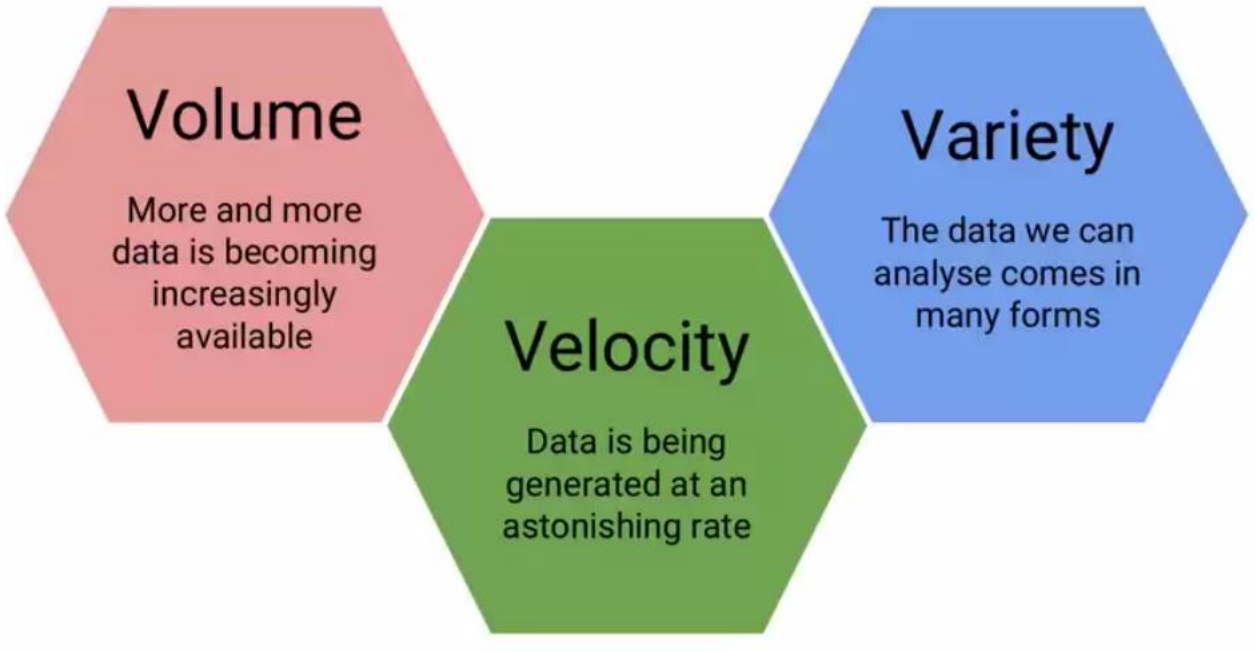
- This has created the perfect storm in which we enrich data and the tools to analyze it, rising computer memory capabilities, better processors, more software and now, more data scientists with the skills to put this to use and answer questions using this data.

Value	Prefix
10^{24}	Yotta
10^{21}	Zetta
10^{18}	Exa
10^{15}	Peta
10^{12}	Tera
10^{9}	Giga
10^{6}	Mega

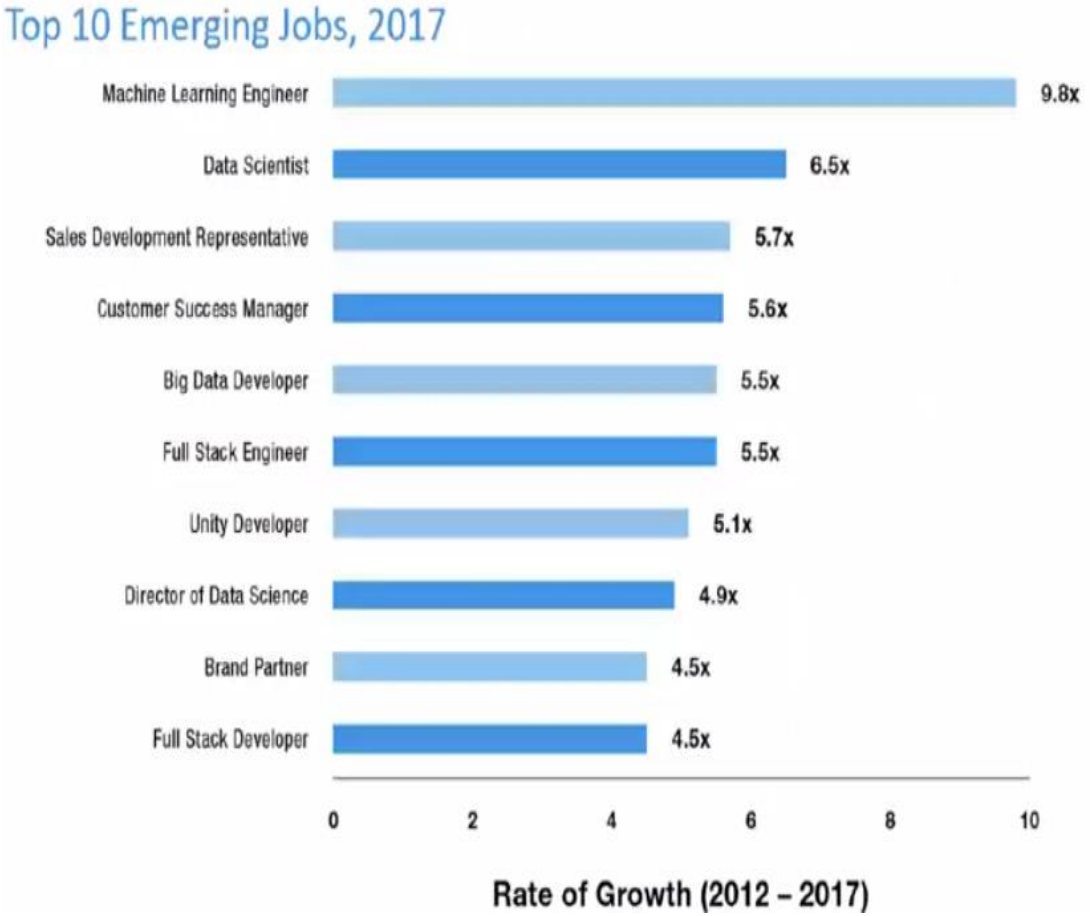
There is an estimated 1.2 zettabytes worth of information currently available - and this number is growing exponentially.

What is big data?

- So, we've talked about what data science is and what sorts of data it deals with, but something else we need to discuss is what exactly a data scientist is.



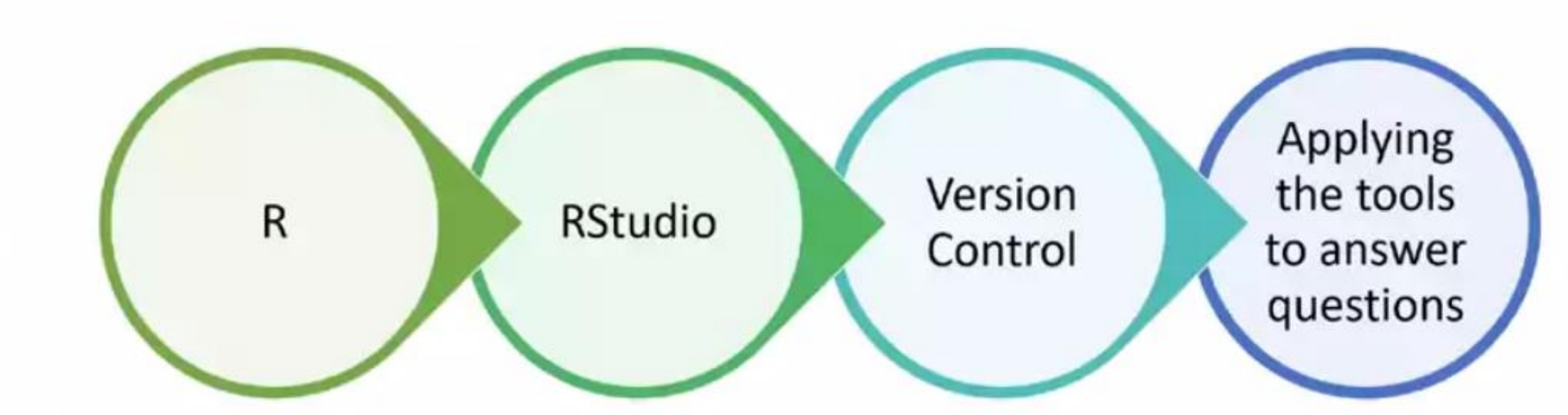
A huge need for individuals with data science skills



- Not only are machine-learning engineers, data scientists, and big data engineers among the top emerging jobs in 2017.
- Not only do we have more and more data, and more and more tools for collecting, storing, and analyzing it, but the demand for data scientists is becoming increasingly recognized as important in many diverse sectors, not just business and academia.

Adapted from: https://economicgraph.linkedin.com/research/Linkedins-2017-US-Emerging-Jobs-Report

In this course



What is data?

"A set of values of qualitative or quantitative variables"

Set: In statistics, the population you are trying to discover something about

Variable: Measurements or characteristics of an item

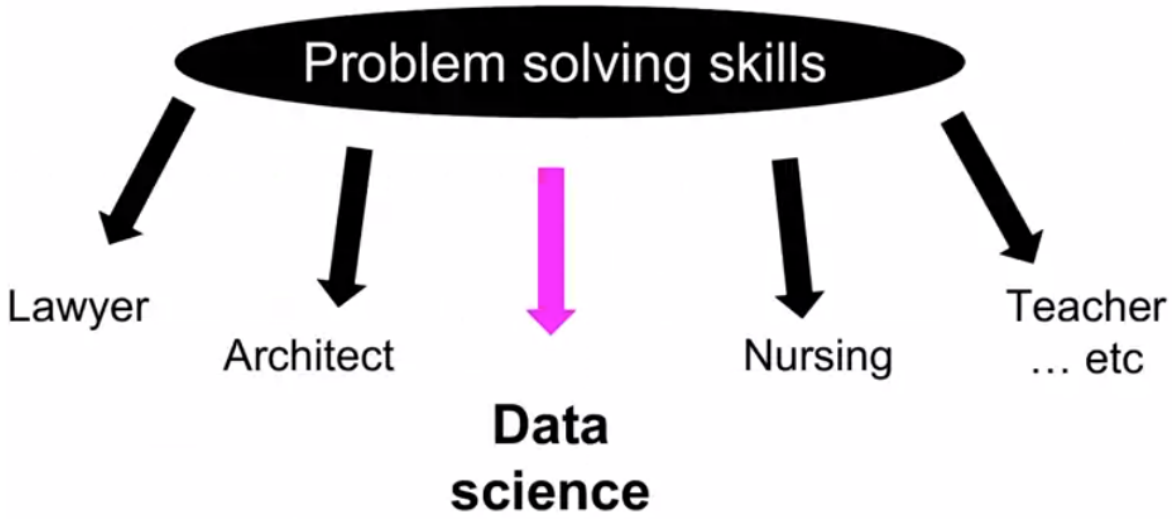
Qualitative variable: Measurements or information about qualities

Quantitative variable: Measurements or information about quantities or numerical items

What is data?

- Sequencing data
- Population census data
- Electronic medical records (EMR), other large databases
- Geographic information system (GIS) data (mapping)
- Image analysis and image extrapolation
- Language and translations
- Website traffic
- Personal/Ad data (eg: Facebook, Netflix predictions, etc)

Getting help



stackoverflow

CrossValidated

coursera

Asking questions on forums - details to include:

- The question you are trying to answer
- How you approached the problem, what steps you took to answer the question
- What steps will reproduce the problem (including sample data for troubleshooters to work from!)
- What was the expected output
- What you saw instead (including any error messages you received!)
- What troubleshooting steps you have already tried
- Details about your set-up, eg: what operating system you are using, what version of the product you have installed (eg: R, Rpackages)

Titling forum posts

Bad:

- HELP! Can't fit linear model!
- HELP! Don't understand PCA!

Better:

- R 3.4.3 lm() function produces seg fault with large data frame (Windows 10)
- Applied PCA to a matrix - what are U, D, and Vt?

Even better:

- R 3.4.3 lm() function on Windows 10 -- seg fault on large dataframe
- Using principal components to discover common variation in rows of a matrix, should I use, U, D or Vt?