CENG 499

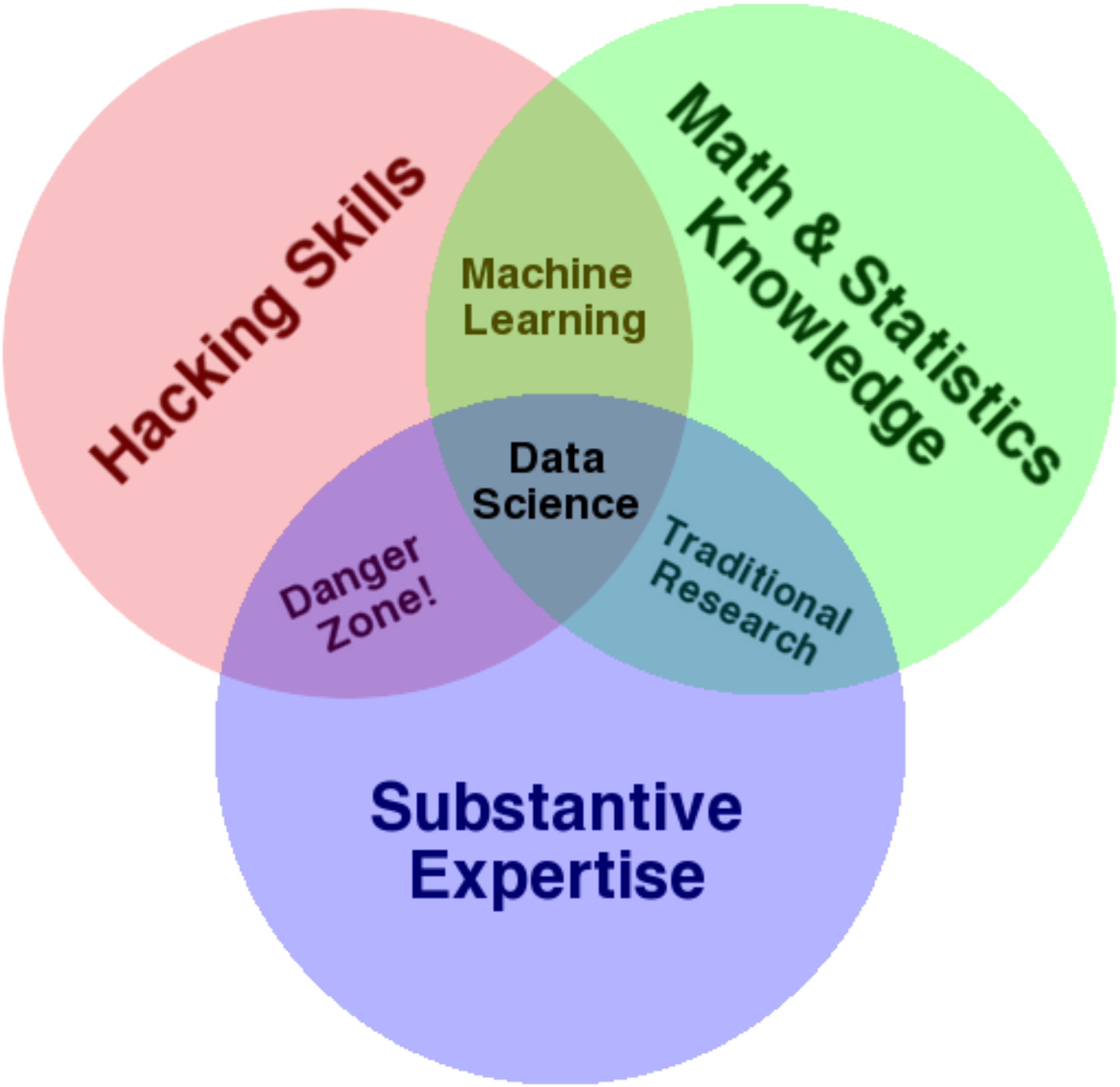
Introduction to Data Science

Erdoğan Doğdu

"Sweetheart, my neural net predicts that you and I are 98.9% compatible. Will you be my Valentine?"

What is Data Science?

- **Data science**, also known as **data**-driven **science**, is an interdisciplinary field of scientific methods, processes, and systems to extract knowledge or insights from **data** in various forms, either structured or unstructured, similar to **data** mining.
  - – Wikipedia https://en.wikipedia.org/wiki/Data_science



Source: http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram

Data Scientist

- Someone who extracts insights from messy data.
- Someone who knows more statistics than a computer scientist and more computer science than a statistician.

Data science process

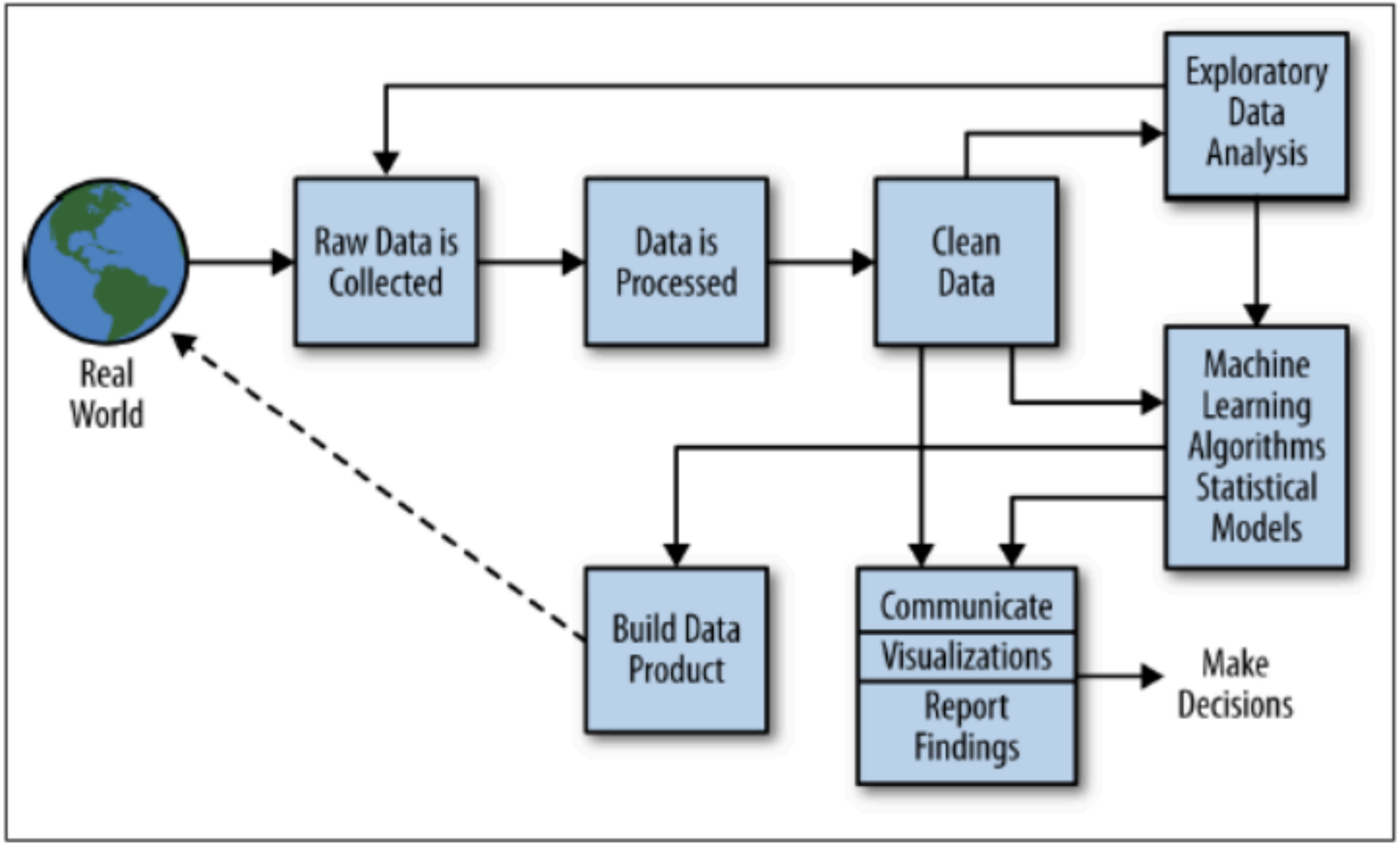


Figure 2-2. The data science process

DDS, p.41

Data Science

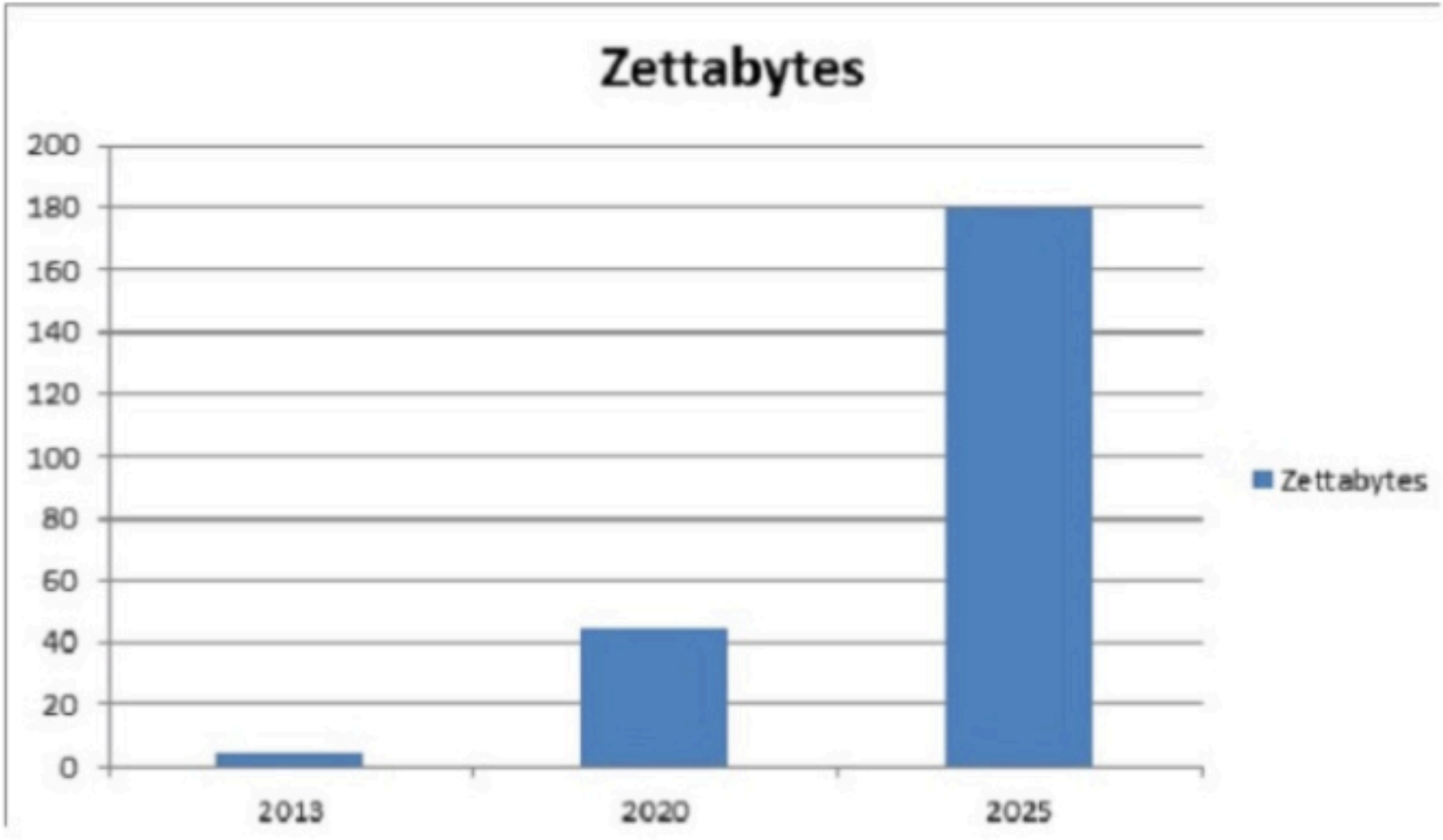
- "The ability to take data—to be able to **understand** it, to **process** it, to **extract value** from it, to **visualize** it, to **communicate** it—that's going to be a hugely important skill in the next decades, ... because now we really do have **essentially free and ubiquitous data**."
  - Hal Varian, Google's Chief Economist The McKinsey Quarterly, Jan 2009

Big Data

- Are you producing data right now?
- What kind?
- Where is it stored?
- How big is that data?
- Who is using it for what purposes?
- Can you avoid producing data?
- Can we live without big data?

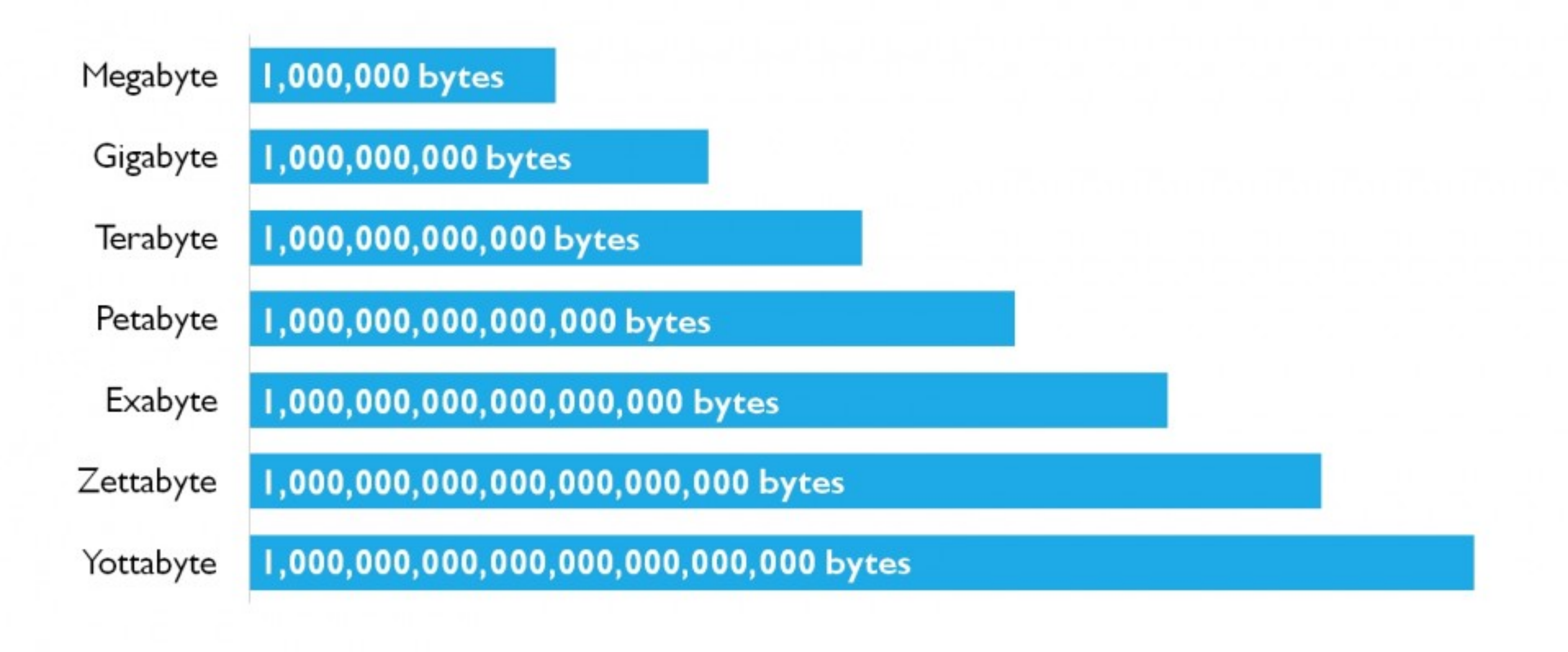
Amount of Data Created Annually to Reach 180 Zettabytes in 2025

Posted on March 7, 2016

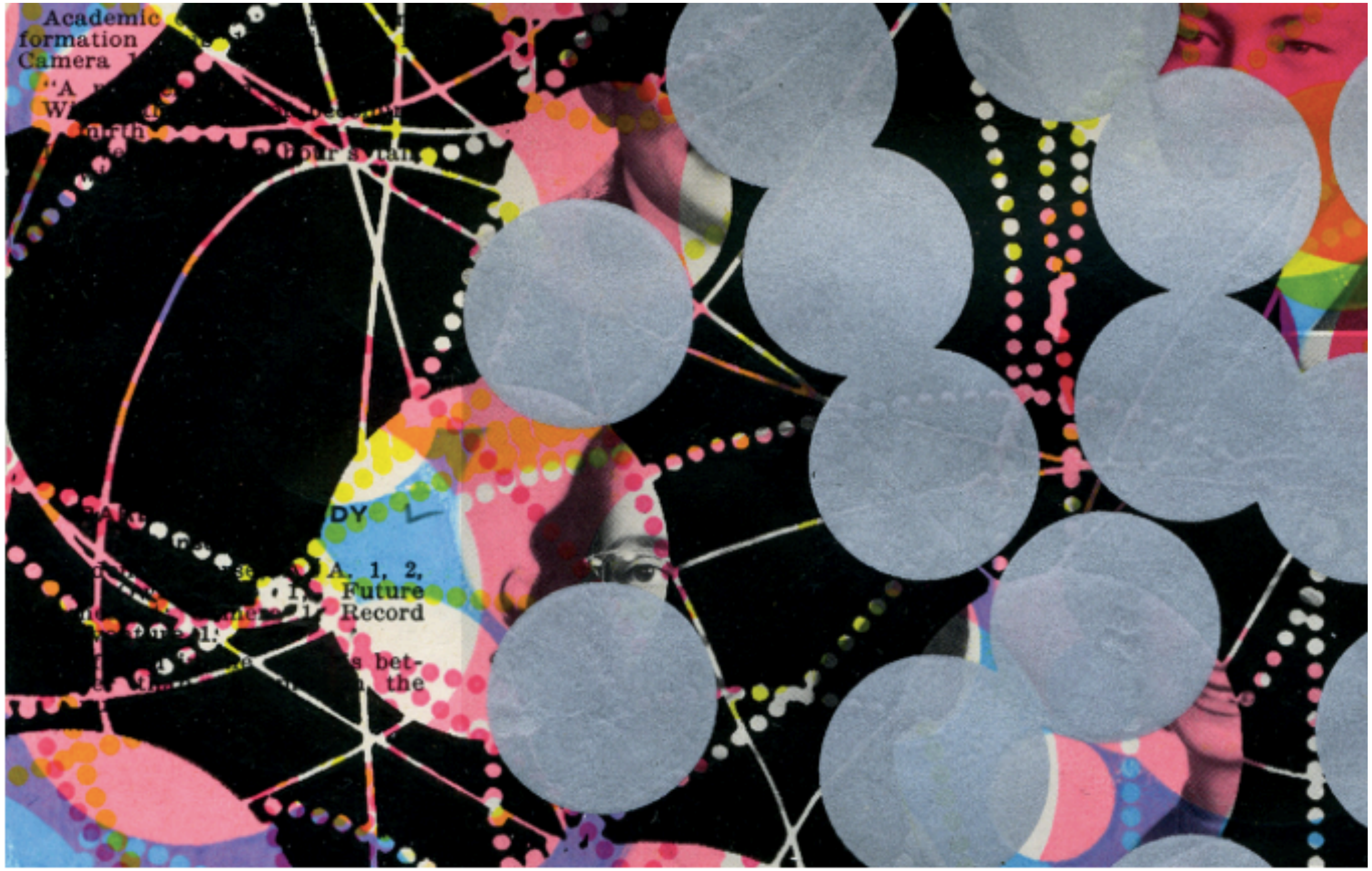


Source:IDC

1 zettabyte = 10^{21} bytes = 1 billion TB



Harvard Business Review



DATA

Data Scientist: The Sexiest Job of the 21st Century

by Thomas H. Davenport and D.J. Patil

FROM THE OCTOBER 2012 ISSUE

LinkedIn The Most Promising Jobs 2018

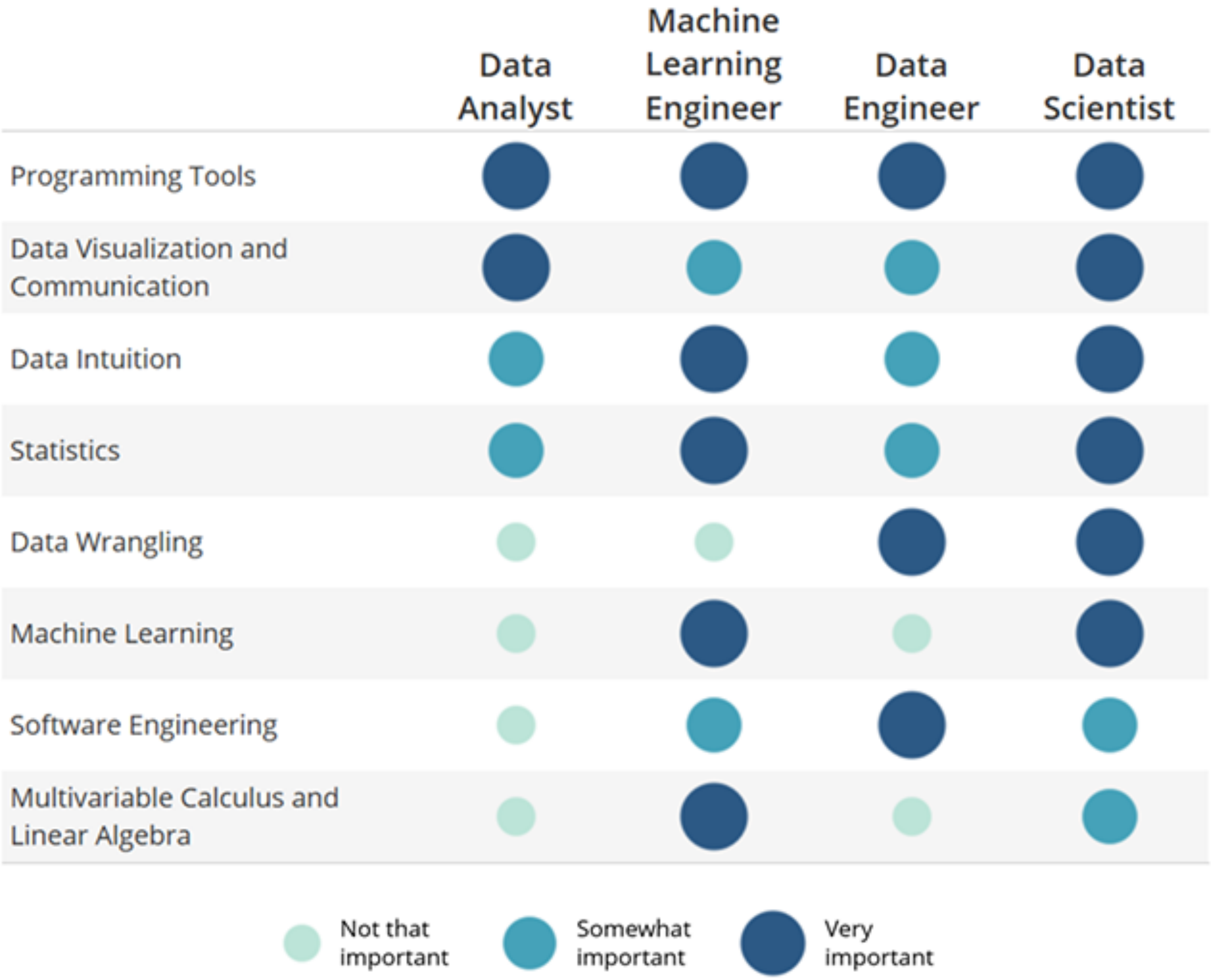
- **9. Data Scientist**
  - Median Base Salary: $113,000
  - Job Openings (YoY Growth): 2,100+ (45%)
  - Career Advancement Score (out of 10): 8
  - Top Skills: Python, Data Analysis, Machine Learning, SQL, Statistics

  - https://blog.linkedin.com/2018/january/11/linkedin-data-reveals-the-most-promising-jobs-and-in-demand-skills-2018

Top Skills (LinkedIn)

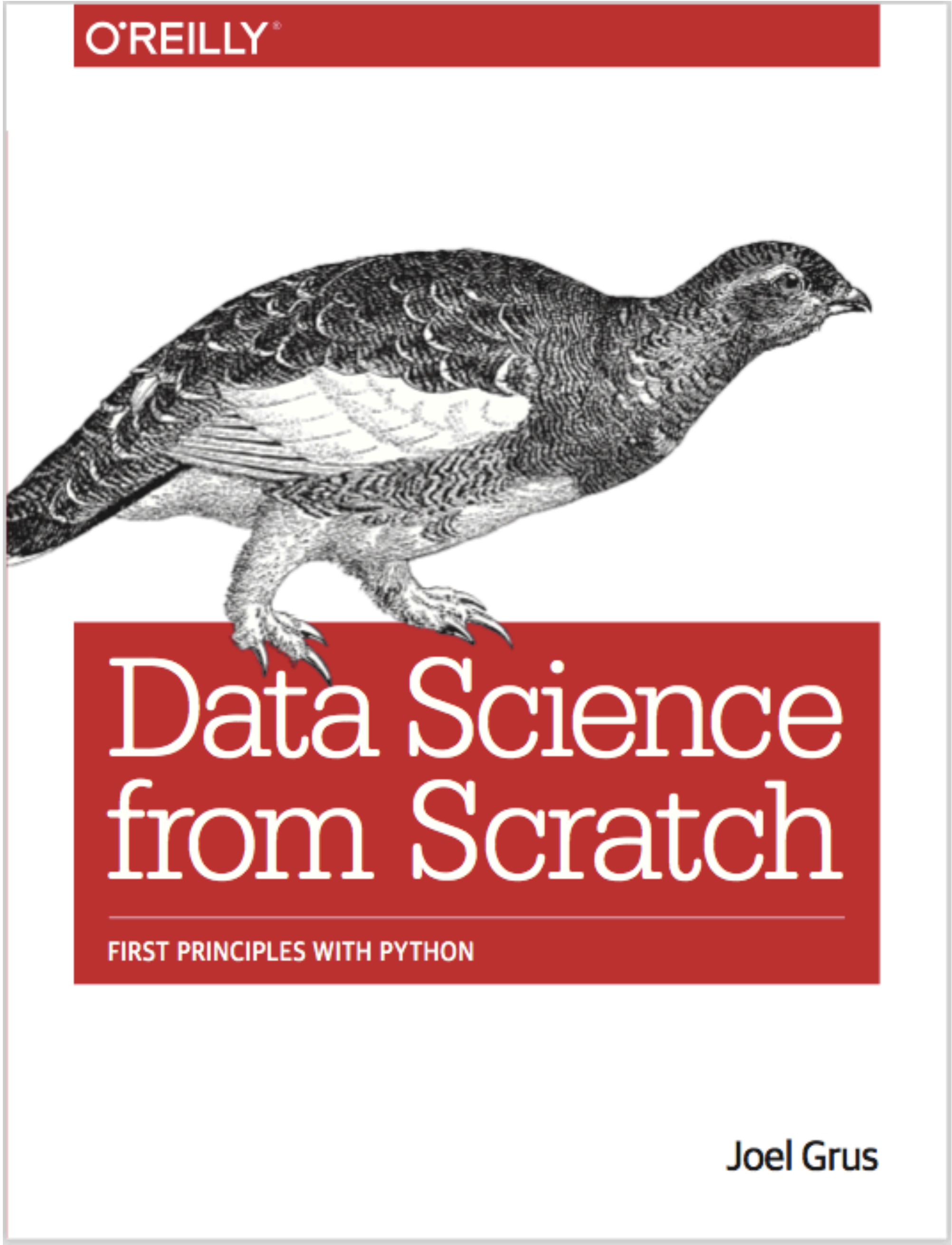
- **1. Cloud and Distributed Computing**
  - **Related Jobs:** Platform Engineer ($120,000), Cloud Architect ($135,000)
- **2. Statistical Analysis and Data Mining**
  - **Related Jobs:** Business Analyst ($72,000), Data Analyst ($62,000), Statistician ($90,200)
- **7. Data Presentation**
  - **Related Jobs:** Graphic Designer ($45,000), Data Scientist ($113,000), Business Consultant ($83,000)
- **12. Data Engineering and Data Warehousing**
  - **Related Jobs:** Software Engineer ($95,000), Database Developer ($85,000), Data Analyst ($62,000)
- https://www.linkedin.com/pulse/skills-companies-need-most-2018-courses-get-them-paul-petrone/?trk=li_corpblog_jobs_skills_2018

Skills to Jobs



	Data Analyst	Machine Learning Engineer	Data Engineer	Data Scientist
Programming Tools	●	●	●	●
Data Visualization and Communication	●	◐	◐	●
Data Intuition	◐	●	◐	●
Statistics	◐	●	◐	●
Data Wrangling	○	○	●	●
Machine Learning	○	●	○	●
Software Engineering	○	◐	●	◐
Multivariable Calculus and Linear Algebra	○	●	○	◐

○ Not that important ◐ Somewhat important ● Very important



O'REILLY®

Data Science from Scratch

FIRST PRINCIPLES WITH PYTHON

Joel Grus

The End