Overview: Data Science & Data Engineering Perspectives

Marco Morales
marco.morales@columbia.edu

Nana Yaw Essuman
nanayawce@gmail.com

GR5069: Applied Data Science for Social Scientists

Spring 2021
Columbia University

What is Data Science?

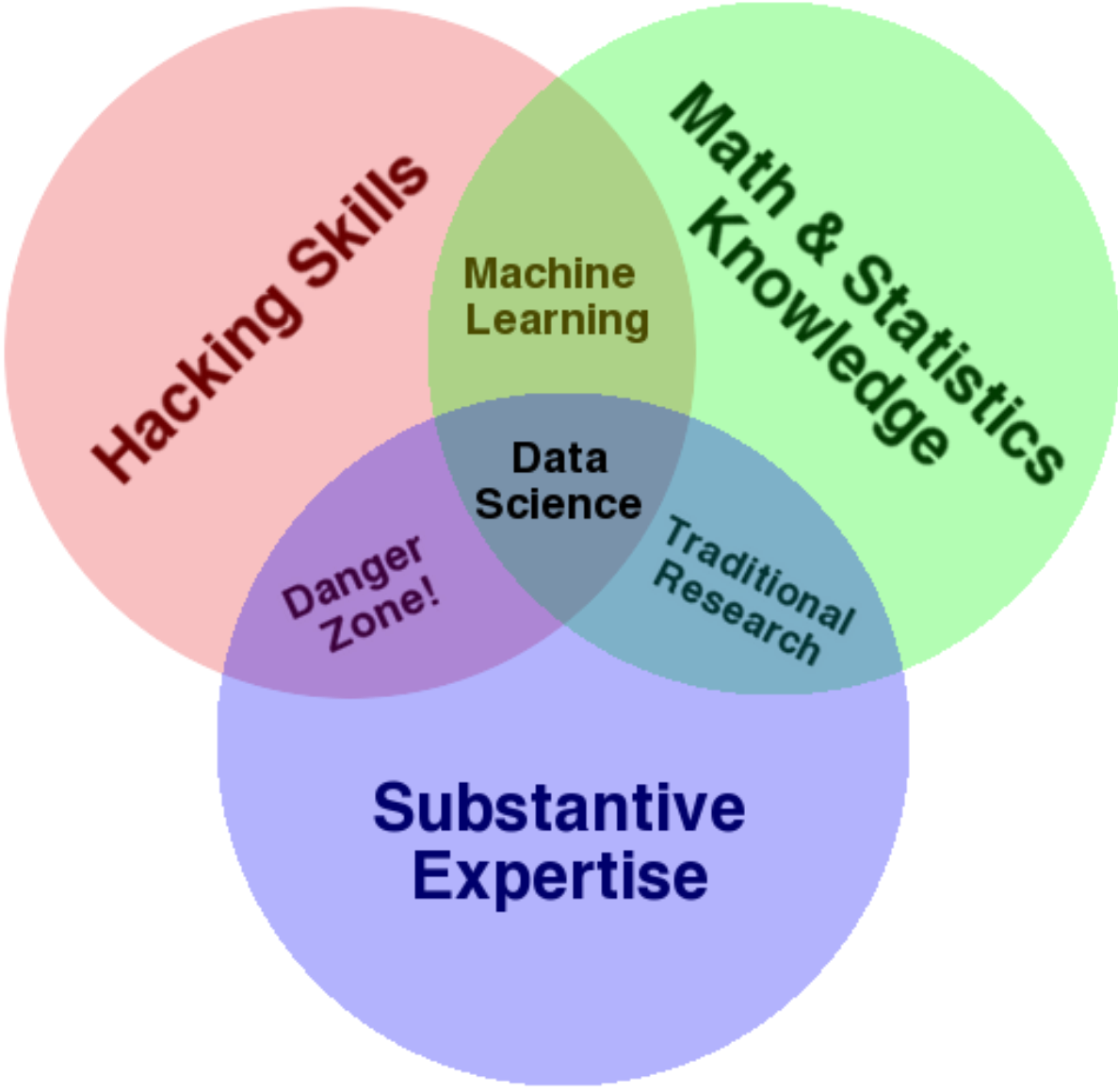
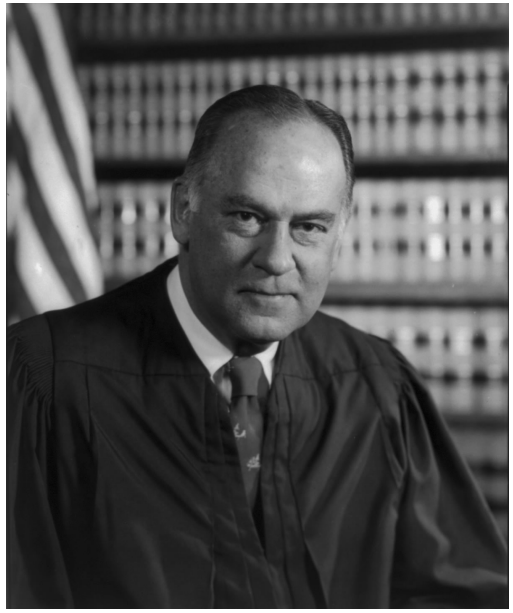


Figure: Drew Conway (2013)

Defining Data Science



*"I shall not today attempt further to define the kinds of material I understand to be embraced within that shorthand description; and perhaps I could never succeed in intelligibly doing so. But **I know it when I see it** [...]. "*

Justice Potter Stewart, *Jacobellis v Ohio*, 378 U.S. 184 (1964)

is it in the algorithms?

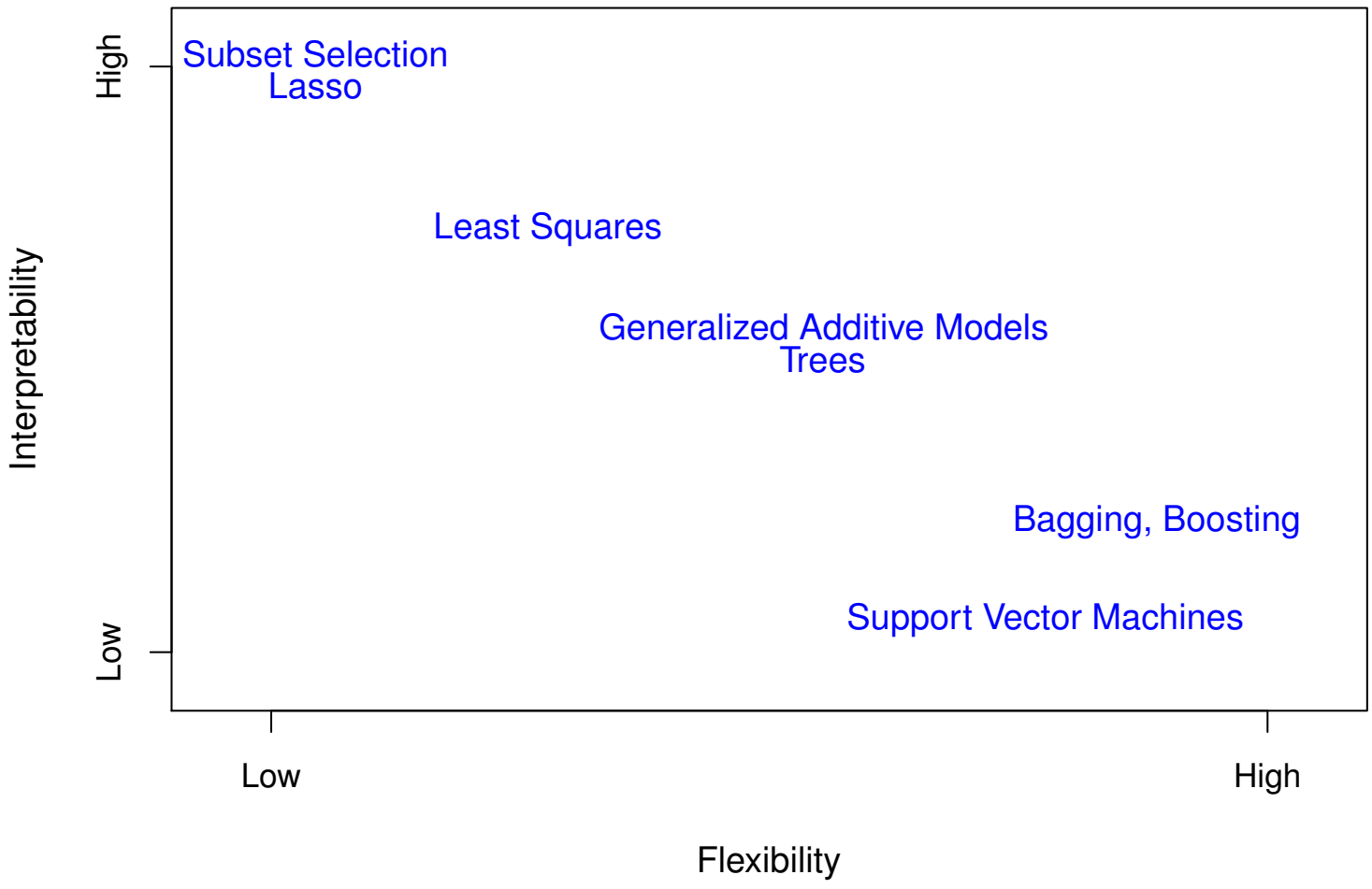


Figure: James et al. (2016)

is it in the algorithms?

- is it really that **different from applied statistics**?
- after all, ML is also **statistical learning**...
- and many algorithms were developed first or have equivalents in Statistics
- a growing movement in Data Science for **model interpretability** (and away from the black box)

is it in the tech stack?

is it in the tech stack?

- tech stack more relevant from the **engineering perspective**
  - what tools are more relevant for which purposes?
  - what tools are “scalable” in the context of this project?
  - tools are tools are tools
- most (new) technolgies are created (and deprecated) faster than we can adopt them

is it in the big data?

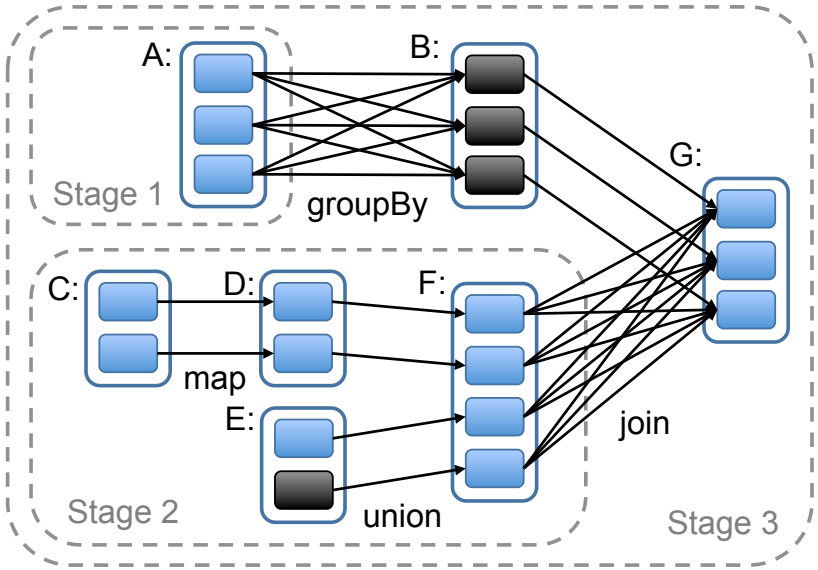


Figure 2.5. Example of how Spark computes job stages. Boxes with solid outlines are RDDs. Partitions are shaded rectangles, in black if they are already in memory. To run an action on RDD G, we build build stages at wide dependencies and pipeline narrow transformations inside each stage. In this case, stage 1's output RDD is already in RAM, so we run stage 2 and then 3.

Figure: Matei Zaharia (2014)

is it in the big data?

- the “big” in **big data** is relative to **computing capabilities**
  - until recently, driven by Moore’s “law”
- big data capabilities $\approx$ **efficient distributed computing**
- **reality check:** big data tools perform mostly **basic tasks** today
  - we’re only beginning to scratch the surface
  - promise in techniques that require **a lot** of data

is it in the predictive "focus"?

is it in the predictive "focus"?

- despite popular belief, **not all data science is predictive**
  - **inference** is a growing part of Data Science
  - **prediction** may be a large part of Data Science **education**
  - ...though not necessarily **practice**
- more important in some industries than others

is it the techniques to exploit data?

VB CHANNELS EVENTS NEWSLETTERS Search

AI

AI predictions for 2019 from Yann LeCun, Hilary Mason, Andrew Ng, and Rumman Chowdhury

KHARI JOHNSON @KHARIJOHNSON JANUARY 2, 2019 7:25 AM

Above: Left to right: Cloudera machine learning general manager Hilary Mason, Accenture global responsible AI lead Rumman Chowdhury, Facebook AI Research director Yann LeCun, and Google Brain cofounder Andrew Ng

is it the techniques to exploit data?

- although not always evident, there's **little consensus in the meaning of terms to designate techniques**
  - every few months a new fad term appears: ML, Reinforcement Learning, Deep Learning, AI (e.g. Artificial Intelligence, Augmented Intelligence), Cognitive Computing...
  - academics and practitioners usually **mean different things** when they use them...
  - meanings become even fuzzier when **consultants** come into the mix
- in reality, **very few problems require** (and have the necessary data needed by) **the most advanced techniques**

is it in the “unicorns”?

is it in the “unicorns”?

- Data Science is **collaborative** in nature
  - no single person possesses all
    - skills
    - substantive knowledge
    - expertise
- ~~most~~ many data scientists **are scholars** by training
  - ... but do **not exclusively** work in academia
- which means that **data scientists are** (have to be):
  - more **applied**
  - less theoretical
  - more focused on **results**

Different focus: who is a Data Scientist?

Data Scientist:

The Sexiest Job of the 21st Century

Meet the people who can coax treasure out of messy, unstructured data.

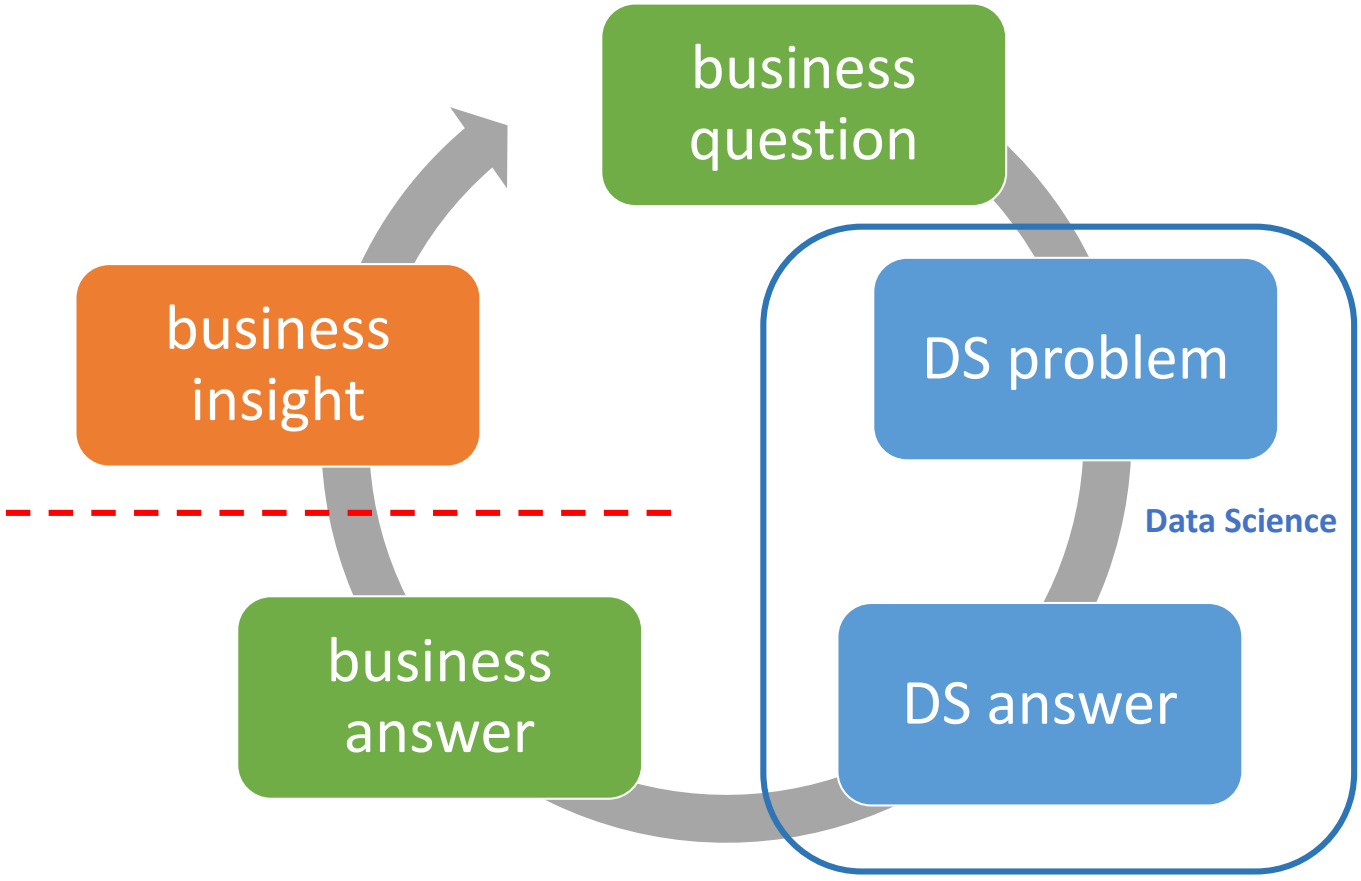
by Thomas H. Davenport and D.J. Patil

When Jonathan Goldman arrived for work in June 2006 at LinkedIn, the business networking site, the place still felt like a start-up. The company had just under 8 million accounts, and the number was growing quickly as existing members invited their friends and colleagues to join. But users weren't seeking out connections with the people who were already on the site at the rate executives had expected. Something was apparently missing in the social experience. As one LinkedIn manager put it, "It was like arriving at a conference reception and realizing you don't know anyone. So you just stand in the corner sipping your drink—and you probably leave early."

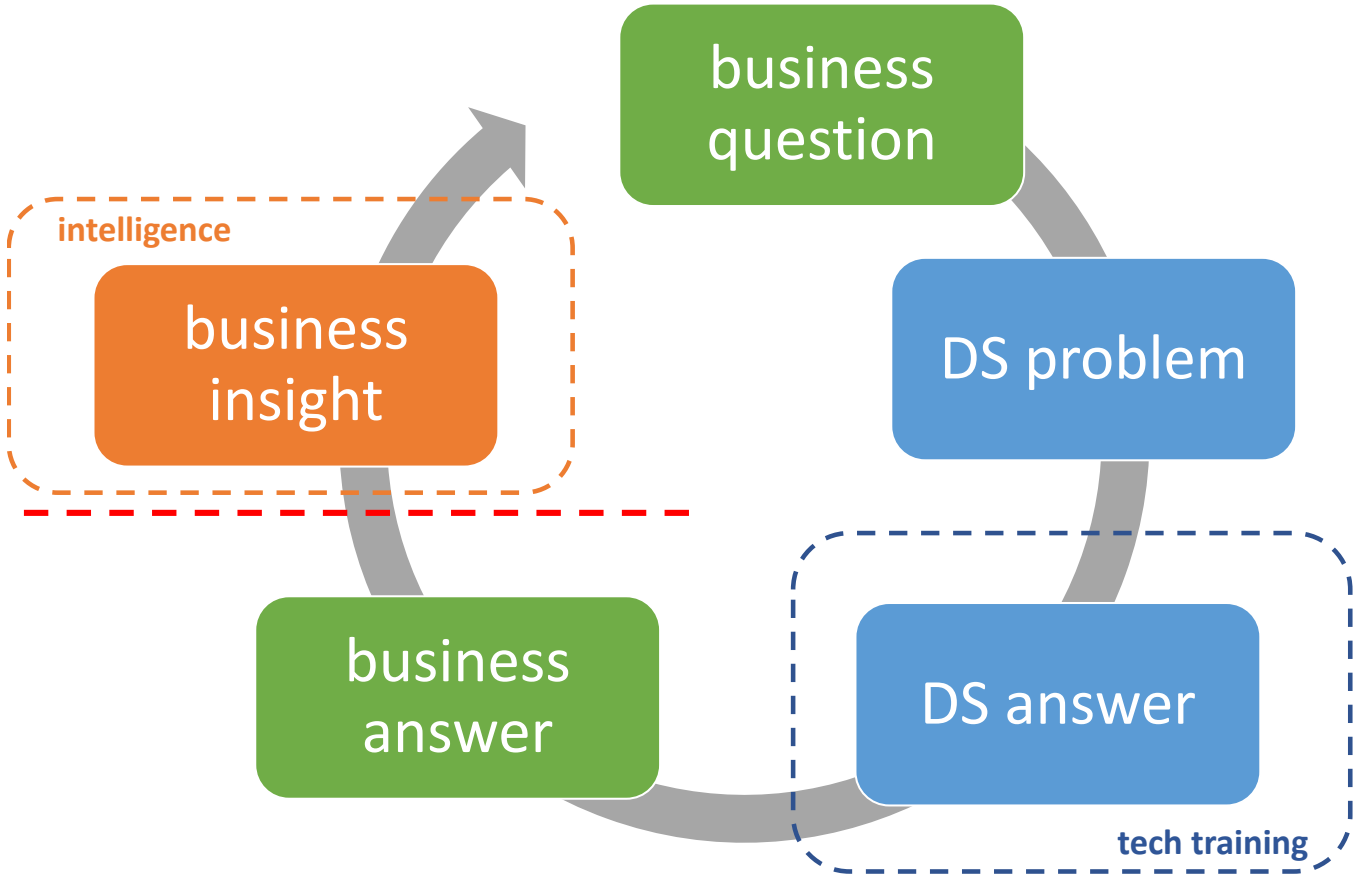
What does a Data Scientist do?

1. **learn** from data (evidence-based)
2. generate predictive or inferential **answers**
3. create reproducible and transferable **outputs**
4. (potentially) **scalable** products
5. (if lucky) **inform decision-making** with alternatives

What does a Data Scientist do?



What does a Data Scientist do?



What must a Data Scientist learn to do?

1. ask the **right questions**
   - turn **business questions into DS questions**
   - turn **DS answers into business answers**
2. **collaborate/coordinate** with data scientists with different skillsets
3. **learn** fast and constantly
   - pick up techniques quickly
   - leverage in-team knowledge to accelerate learning
4. **communicate effectively**
   - **explain** complicated techniques to technical and non-technical audiences
   - **translate** between business, expert stakeholders, enginering teams, DS teams

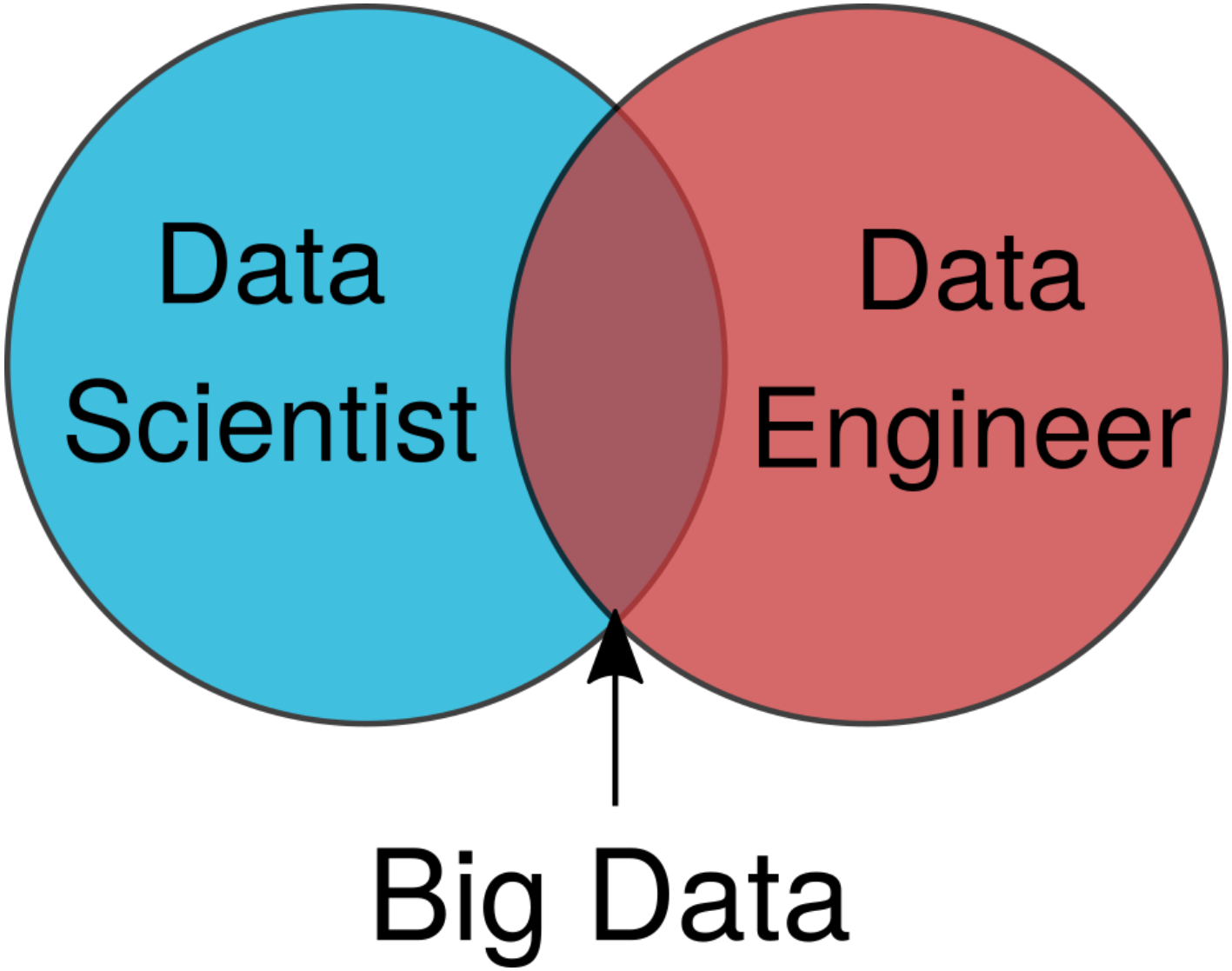
What skills must a Data Scientist have?

- **CRITICAL THINKING** (about data)
- **coding** (hacking)
- **data transformation** (ETL)
- **data exploration / visualization**
- **database usage**
- **modeling / analysis**
- **communication**
- **collaboration**

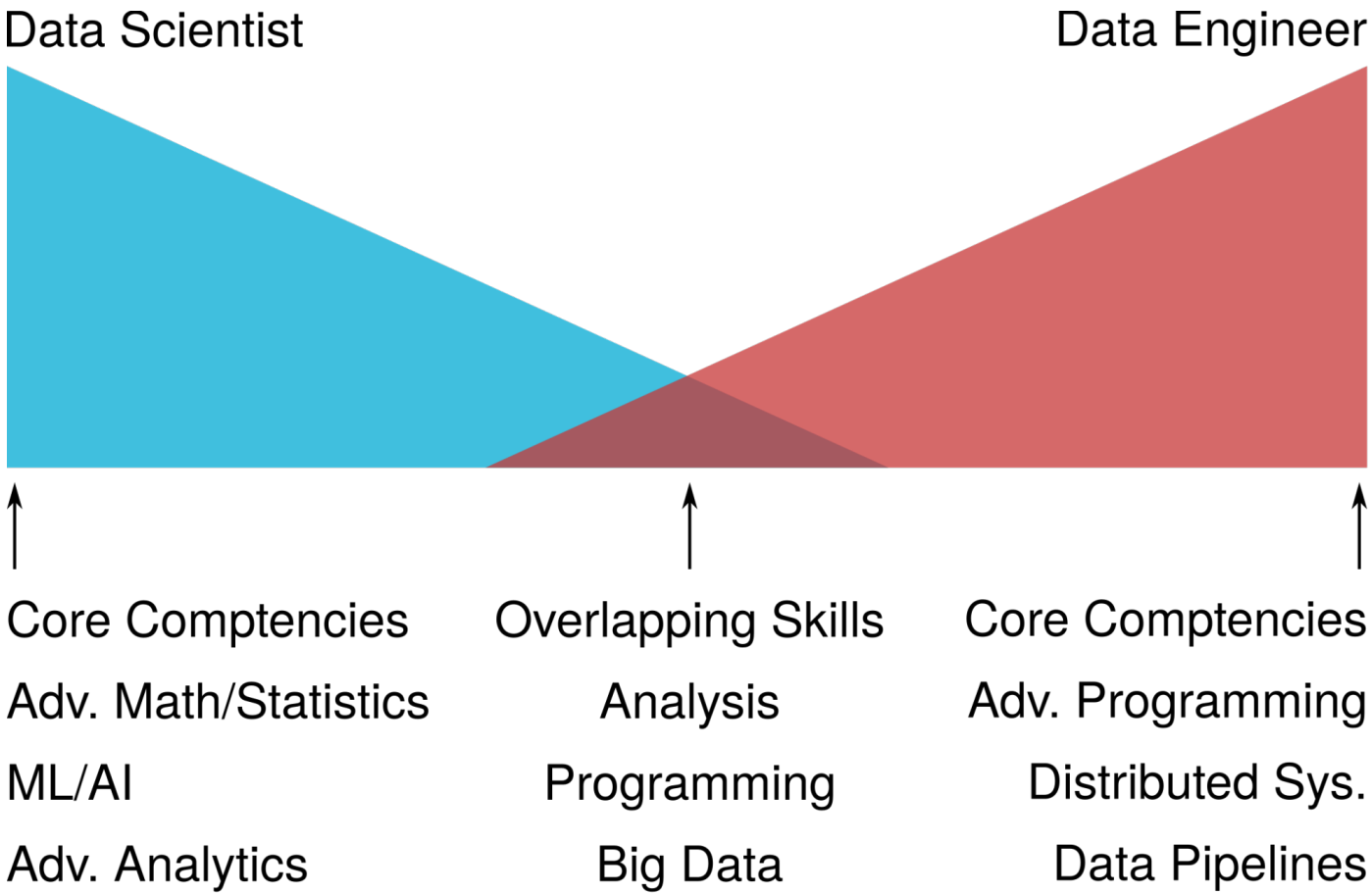
Defining Data Engineering

- Data engineering focuses on data, its curation, movement, storage, security, and processing. Data Engineers work on building batch & real-time pipelines, applications, APIs, and systems that produce, process, and consume data to meet business needs
- To compare; data engineering converts raw data into knowledge data
- Data Scientists deal with using this data produced by data engineers to generate insights & to predict the future using data from the past

The Overlapping Venn Diagram



Data Scientists v Data Engineers



What skills must a Data Engineer have?

- **programming skills** (Advanced)
- **data architecture**
- **distributed systems**
- **cloud computing**
- **database design**
- **data analysis**
- **communication**
- **collaboration**

Overview: Data Science & Data Engineering Perspectives

Marco Morales
marco.morales@columbia.edu

Nana Yaw Essuman
nanayawce@gmail.com

GR5069: Applied Data Science for Social Scientists

Spring 2021
Columbia University