

Overview: Data Science & Data Engineering Perspectives

Marco Morales Nana Yaw Essuman
marco.morales@columbia.edu nanayawce@gmail.com

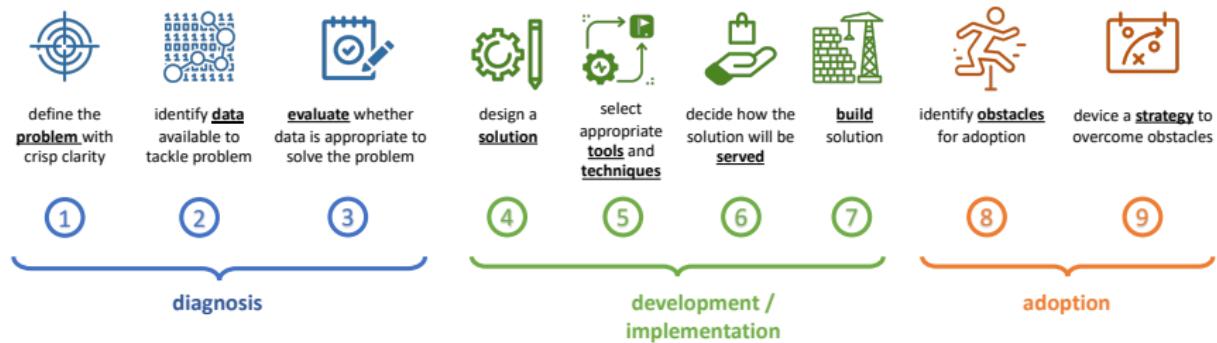
GR5069: Applied Data Science
for Social Scientists

Spring 2022
Columbia University

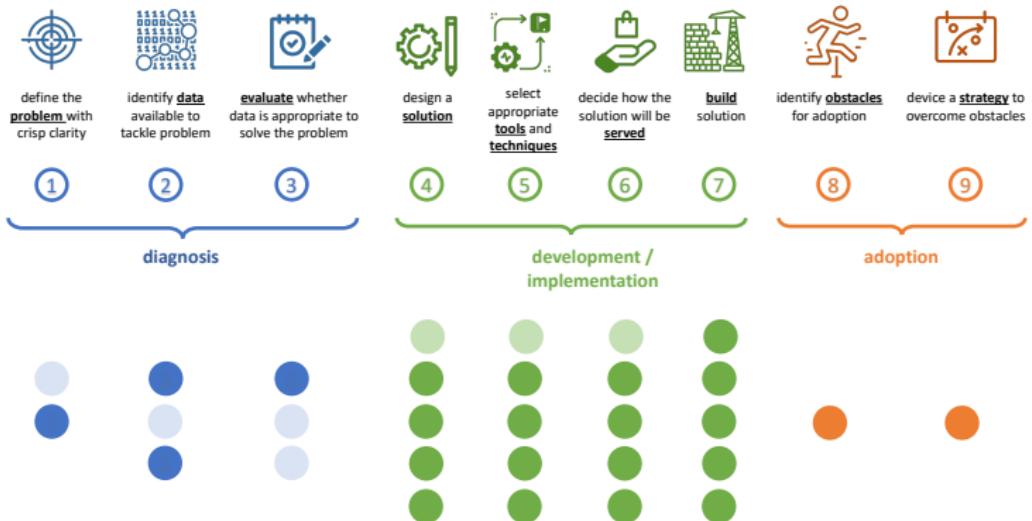
Why this class?



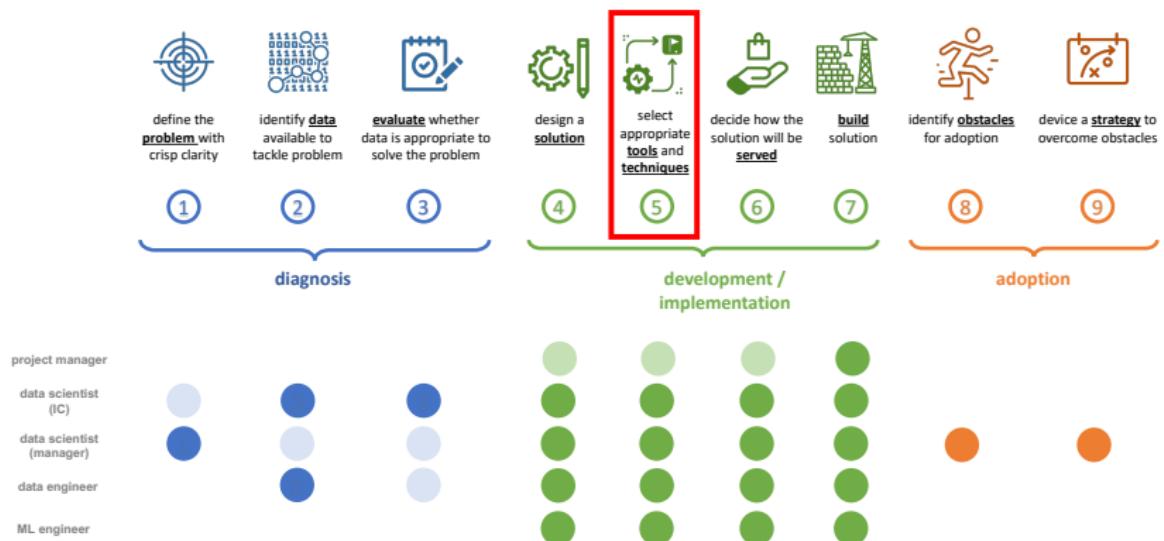
what does the Data Science Shop do?



how does the Data Science Shop operate?



how does the Data Science Shop operate?



Overview: Data Science & Data Engineering Perspectives

Marco Morales Nana Yaw Essuman
marco.morales@columbia.edu nanayawce@gmail.com

GR5069: Applied Data Science
for Social Scientists

Spring 2022
Columbia University

Course Objectives

1. teach processes & practices at the **intersection of Data Science & Data Engineering**
2. sharpen technical skills to **extract knowledge & generate insights**
3. working effectively in **highly-structured collaboration tools**
4. enhance **soft skills**

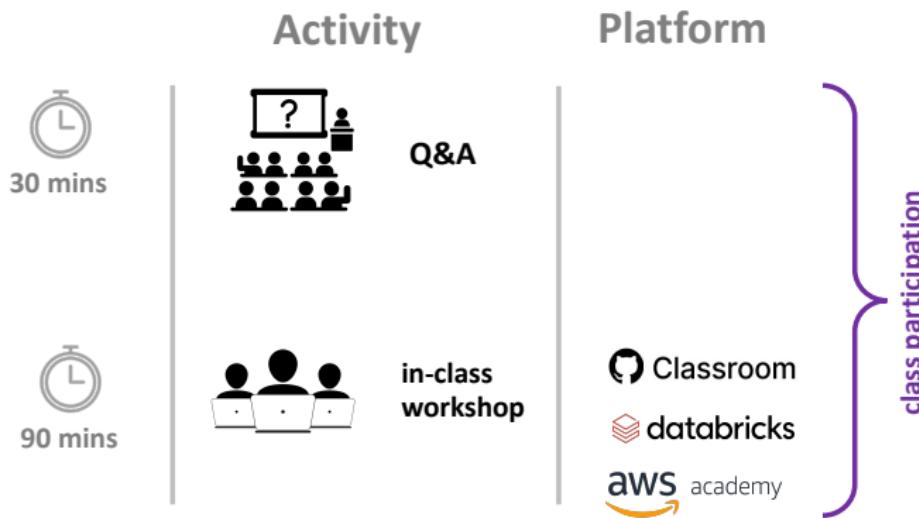
Course Dynamics: the **asynchronous** component

in your own time

Materials	Activity	Submission	
 GitHub  annotated materials	 read / watch / listen	 questions/ comments	
 canvas  recorded lecture	 watch	 questions/ comments	
 Classroom  take-home exercises	 code	 deliverable	 Classroom  databricks  aws academy

Course Dynamics: the **synchronous** component

all together



All class materials in the course's GitHub repo

clone and pull before each class

The screenshot shows the GitHub repository page for 'marco-morales / QMSS-GR5069_Spring2022'. The repository is public and contains 33 commits. The commits are listed in a timeline, mostly from 2 days ago, showing updates to syllabus files, annotated materials, notebooks, and code. The repository has 0 stars, 1 watching, and 0 forks. It uses Jupyter Notebook (97.2%) and Python (1.9%) as languages. The README.md file contains information about the course: 'QMSS GR5069 - APPLIED DATA SCIENCE FOR SOCIAL SCIENTISTS'. It lists the instructor as Marco Morales, Columbia University, and co-instructor as Nana Yaw Essuman, Columbia University. TAs listed are Ludovico Genovese, Columbia University, and Prerit Jain, Columbia University. A note at the bottom states: 'This repository is a companion to the course Applied Data Science for Social Scientists taught at the Quantitative Methods in the Social Sciences program over the Spring of 2022.'

Code

main · 1 branch · 0 tags

Go to file Add file · Code ·

About

Class repository for GR5069 - Applied Data Science for Social Scientists (Spring 2022)

data-science social-sciences educational-materials course-repository

README

0 stars 1 watching 0 forks

Languages

Jupyter Notebook 97.2% Python 1.9% R 0.4%

QMSS GR5069 - APPLIED DATA SCIENCE FOR SOCIAL SCIENTISTS

Instructor: Marco Morales, Columbia University
Co-Instructor: Nana Yaw Essuman, Columbia University

TAs: Ludovico Genovese, Columbia University
Prerit Jain, Columbia University

This repository is a companion to the course Applied Data Science for Social Scientists taught at the Quantitative Methods in the Social Sciences program over the Spring of 2022.

A GitHub Classroom link to each in-class & take-home exercise will create your personal repo submit by pushing to your assignment's repo

The screenshot shows a GitHub Classroom interface. At the top, there is a dark header bar with the GitHub Classroom logo on the left and 'GitHub Education' on the right. Below the header, the URL 'Classrooms / QMSS-GR5069-Spring-2022-classroom' is visible. The main title of the classroom is 'QMSS-GR5069-Spring-2022-classroom'. Underneath the title, it says 'QMSS-GR5069-Spring-2022'. Below the title, there are navigation links for 'Assignments' (with a count of 1), 'Students' (0), 'TAs and Admins' (2), and 'Settings'. A red box highlights the 'Assignments' link. The main content area is titled 'Assignments' and contains one item: 'class-exercise-1', which is described as an 'Individual assignment'. To the right of this item is a 'New assignment' button and an 'Invite link' dropdown menu. The bottom of the page features a standard presentation navigation footer with icons for back, forward, search, and other controls.

Course communications



- ▶ **Slack** is the **preferred method of communication** for this course
- ▶ post questions in appropriate **channels**
- ▶ DM classmates & instructors / TAs
- ▶ **aim: collaboration** – if you have a question, likely others have it also (and perhaps an answer to share)
- ▶ we will **reserve email for official communications**

Tech stack for this class



Instructor Dynamics: the Statler and Waldorf approach



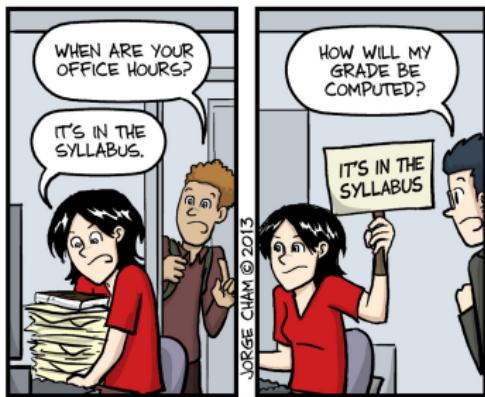
Course Outline

```
outline\
| -- week 1 : OVERVIEW: DS & DE PERSPECTIVES
| -- week 2 : WORKSHOP - VERSION CONTROL & GitHub
| -- week 3 : WORKSHOP - SETTING UP PROJECTS:
              DS & DE PERSPECTIVES
| -- week 4 : WORKSHOP - CODING ETIQUETTE
| -- week 5 : WORKSHOP - DATA PIPELINE IN PRACTICE
| -- week 6 : WORKSHOP - MISSING DATA & DATA QUALITY
| -- week 7 : WORKSHOP - MODEL DEPLOYMENT & VERSIONING,
              WORKING ENVIRONMENTS (DEV, STAGING, PROD)
| -- week 8 :          -- ACADEMIC HOLIDAY --
| -- week 9 : WORKSHOP - INTERACTIVE WORKING SESSION
| -- week 10 : WORKSHOP - EXPLANATION v PREDICTION
| -- week 11 : WORKSHOP - MODEL EVALUATION
| -- week 12 : WORKSHOP - DATA VISUALIZATION
| -- week 13 : WORKFLOW COLLABORATION
| -- week 14 : WORKSHOP - PRESENTING RESULTS
```

Course Requirements

- ▶ Take-home exercises (60%)
- ▶ Attendance & Class Participation (20%)
- ▶ Project Presentation (20%)

Word to the wise...



WWW.PHDCOMICS.COM

This message brought to you by every instructor that ever lived.



Overview: Data Science & Data Engineering Perspectives

Marco Morales Nana Yaw Essuman
marco.morales@columbia.edu nanayawce@gmail.com

GR5069: Applied Data Science
for Social Scientists

Spring 2022
Columbia University

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)

What is Data Science?

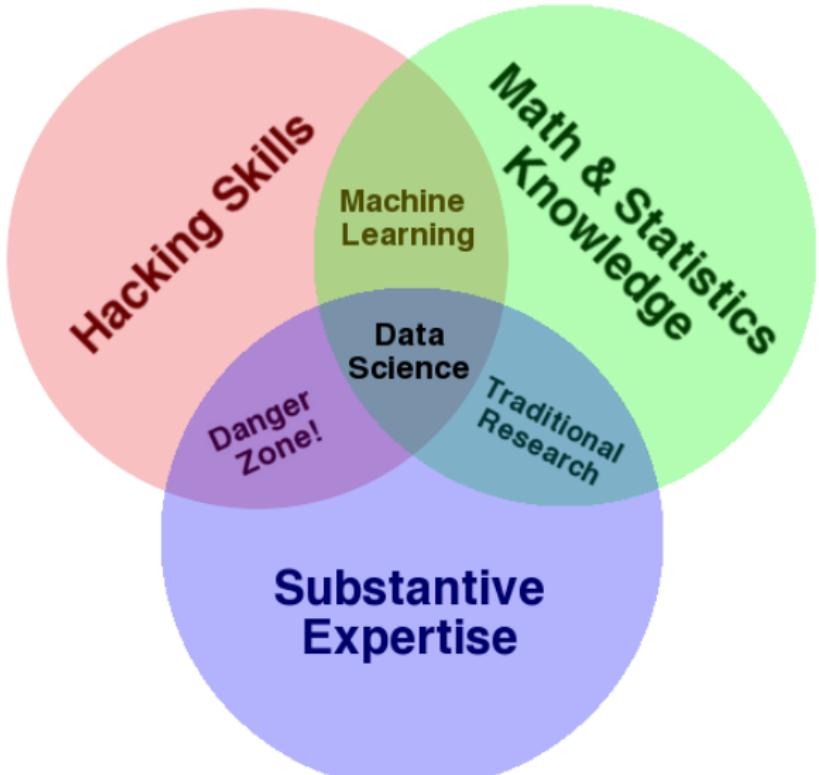


Figure: Drew Conway (2013)

is it in the algorithms?

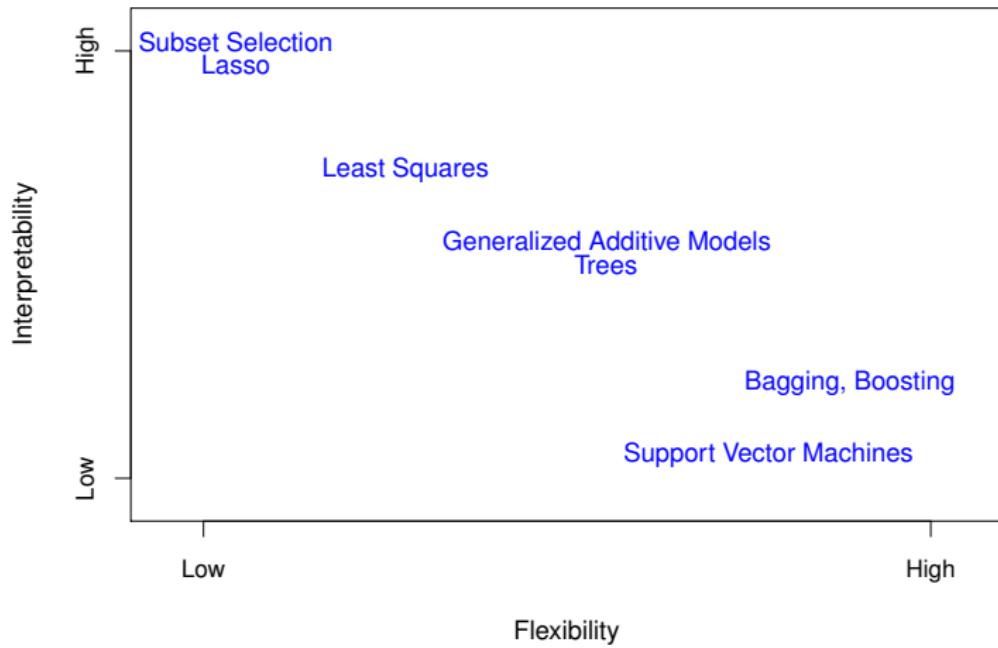


Figure: James et al. (2016)

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) technologies are created (and deprecated) faster than we can adopt them

is it in the big data?

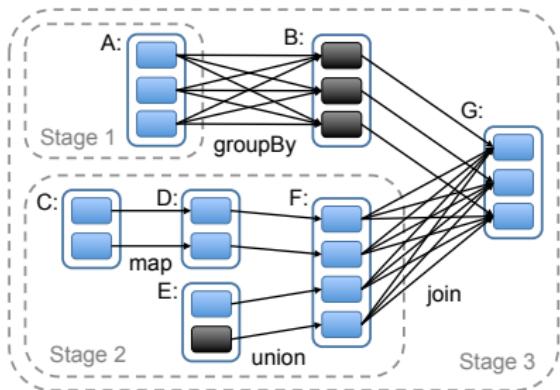


Figure 2.5. Example of how Spark computes job stages. Boxes with solid outlines are RDDs. Partitions are shaded rectangles, in black if they are already in memory. To run an action on RDD G, we build build stages at wide dependencies and pipeline narrow transformations inside each stage. In this case,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

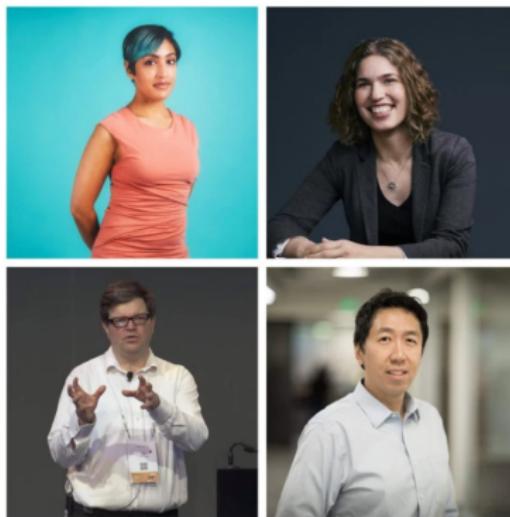
is it the techniques to exploit data?

VB CHANNELS ▾ EVENTS ▾ NEWSLETTERS f t in F RSS 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, **not all 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**

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

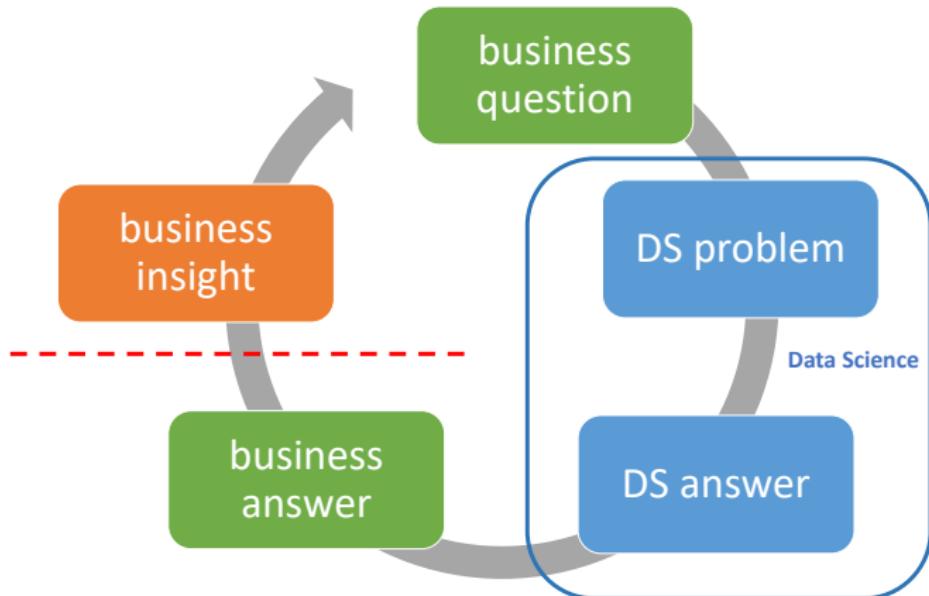
W

hen 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."

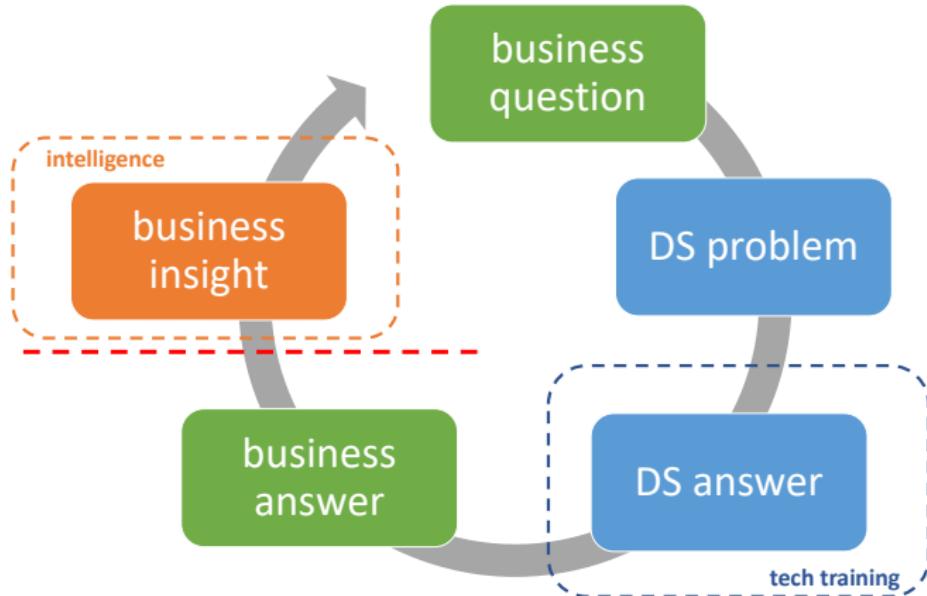
What does a Data Scientist do?

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** to produce **correct answers**
 - ▶ turn **business questions** into **DS questions**
 - ▶ turn **DS answers** into **business answers**
2. **collaborate/coordinate** with data scientists and engineers 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, engineering teams, DS teams

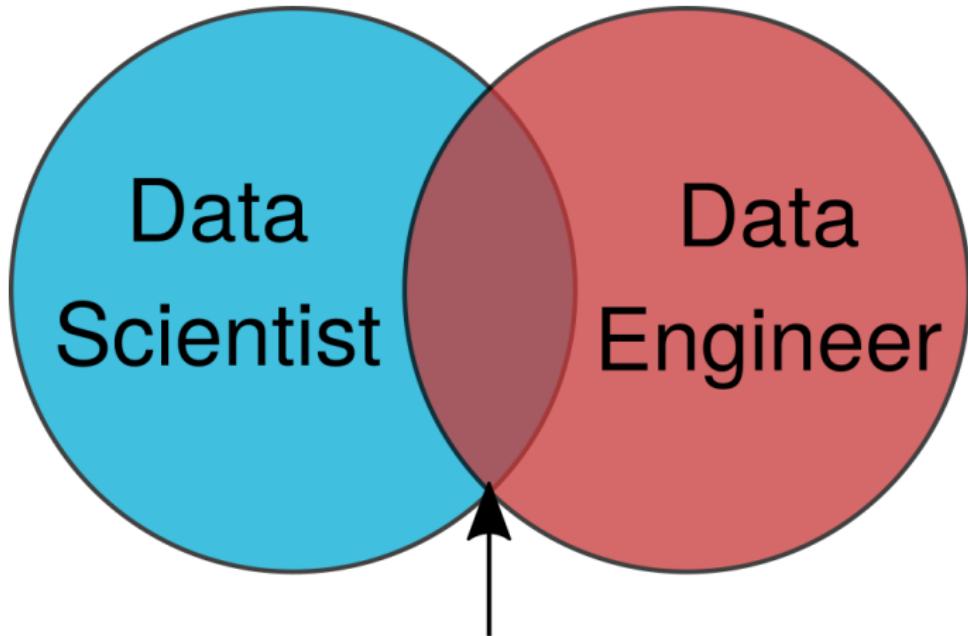
What skills must a Data Scientist have?

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

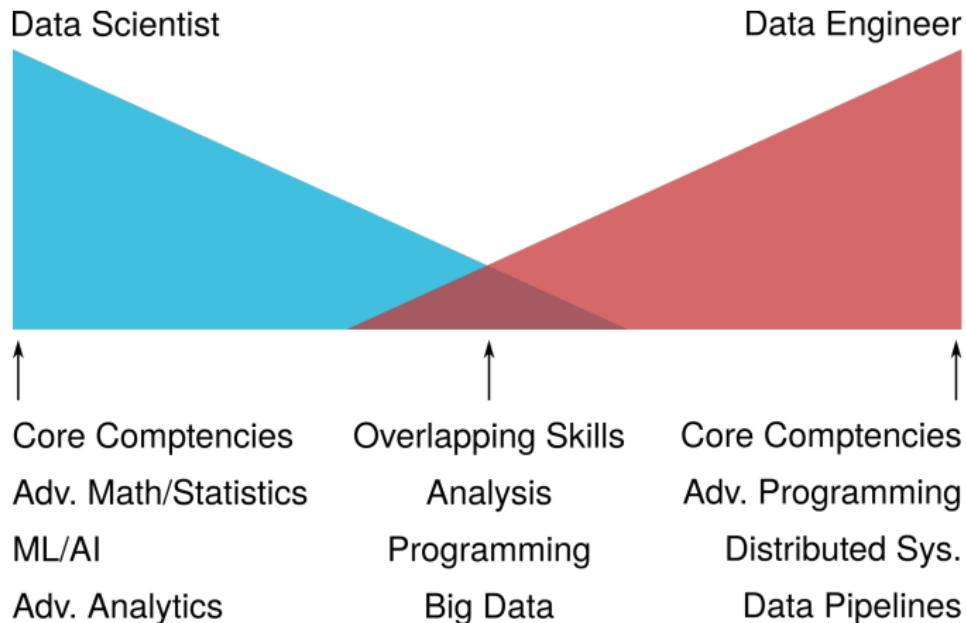
Defining Data Engineering

- ▶ 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 Nana Yaw Essuman
marco.morales@columbia.edu nanayawce@gmail.com

GR5069: Applied Data Science
for Social Scientists

Spring 2022
Columbia University