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 2020
Columbia University

About us

Instructor:	**Marco Morales**
email:	marco.morales@columbia.edu
Office:	509E International Affairs Building
Office Hours:	Weds 8-10PM, and by appointment
Co-Instructor:	**Nana Yaw Essuman**
email:	nanayawce@gmail.com
Office:	509E International Affairs Building
Office Hours:	Weds 8-10PM, and by appointment
TA:	**Ummugul Bezirhan**
email:	ub2126@tc.columbia.edu

Why this class?

Course Outline

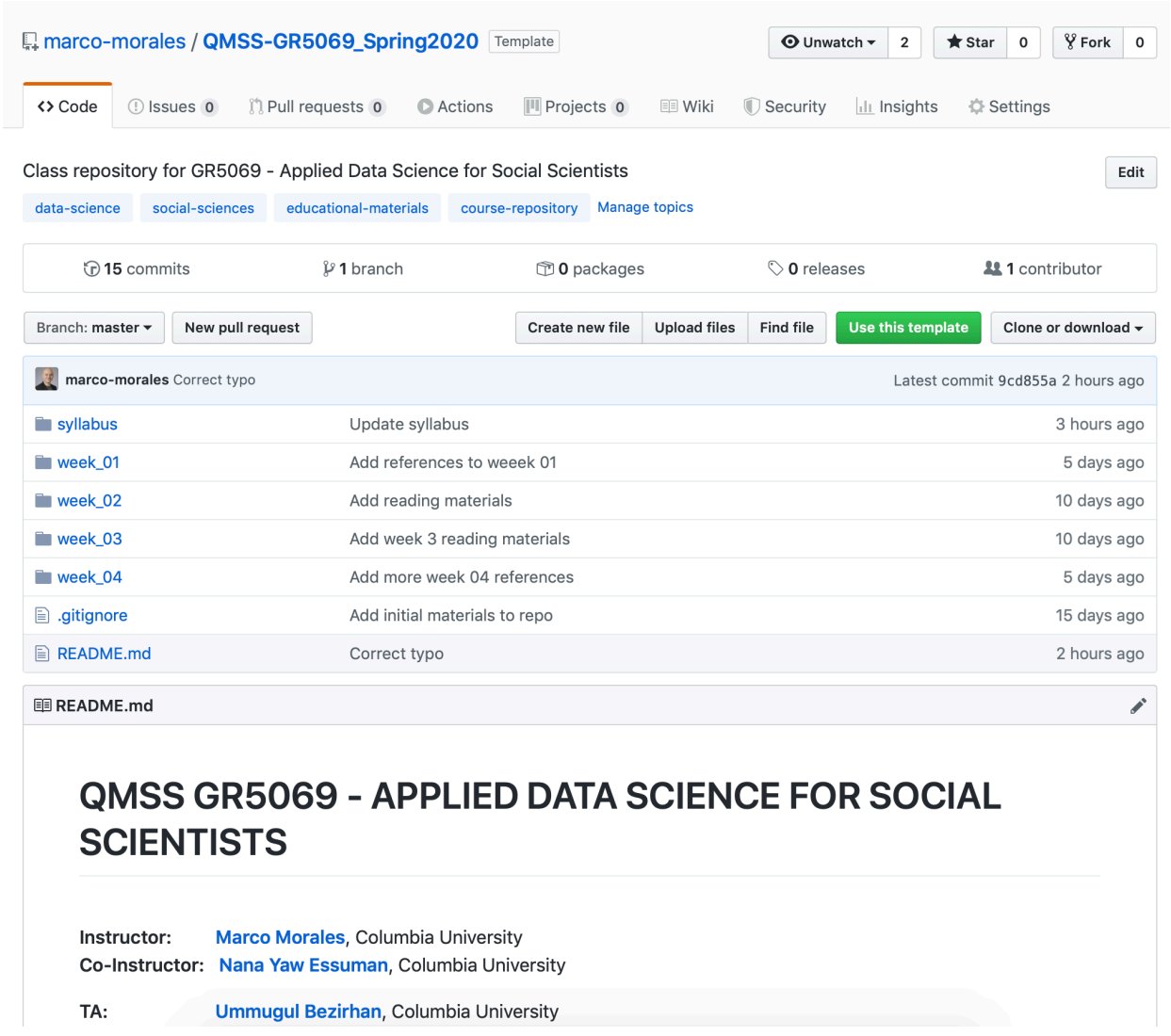
```
outline\
| -- week  1 : OVERVIEW: DS & DE PERSPECTIVES
| -- week  2 : WORKSHOP - SETTING UP PROJECTS:
                          DS & DE PERSPECTIVES
| -- week  3 : WORKSHOP - VERSION CONTROL & GitHub
| -- week  4 : WORKSHOP - CODING ETIQUETTE
| -- week  5 : WORKSHOP - DATA PIPELINE IN PRACTICE
| -- week  6 : MISSING DATA & DATA QUALITY
| -- week  7 : MODEL DEPLOYMENT & VERSIONING,
                WORKING ENVIRONMENTS (DEV, STAGING, PROD)
| -- week  8 : WORKSHOP - INTERACTIVE WORKING SESSION
| -- week  9 :       -- ACADEMIC HOLIDAY --
| -- week 10 : EXPLANATION v PREDICTION
| -- week 11 : CONDITIONAL RELATIONSHIPS IN THE DATA
| -- week 12 : MODEL EVALUATION
| -- week 13 : DATA VISUALIZATION
| -- week 14 : WORKFLOW COLLABORATION
| -- week 15 : PRESENTING RESULTS
```

Course Dynamics

- **Class structure:**
  - sessions will alternate between **short topical lectures** and **applied in-class workshops**
  - curated list of **suggested readings** provided; read as much as you can before class
- **In-class & take-home exercises** submitted through **GitHub Classroom**
- **Course Project**
  - **team assignment** to develop throughout the course
  - teams should be prepared to **share progress every week**
  - presentation will count as **final exam**
- there is a strict **Late Submission Policy**

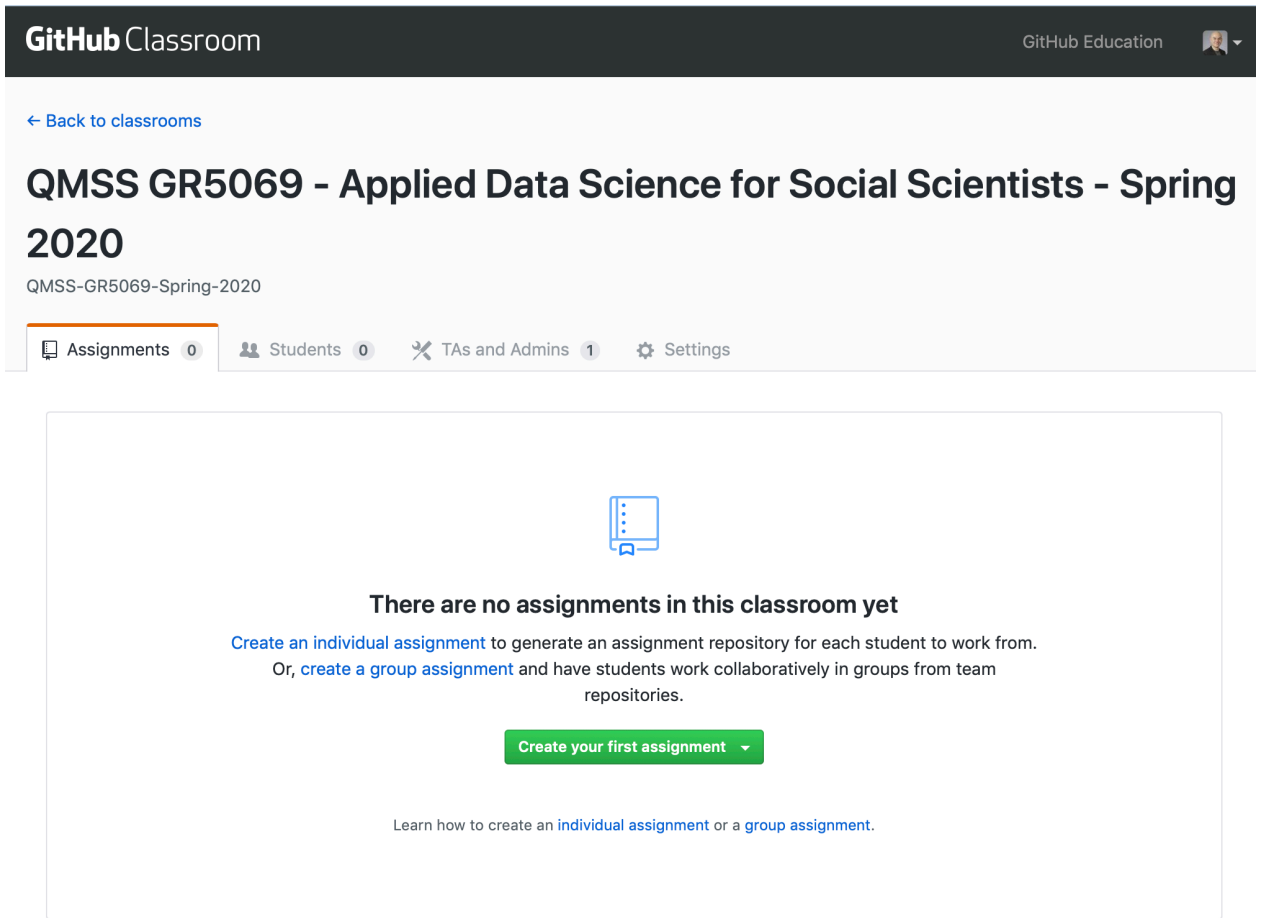
All class materials in the course's GitHub repo

clone and `pull` before each class



A **GitHub Classroom** link to each in-class & take-home exercise will create your personal repo

submit by `push`ing to your assignment's repo



Instructor Dynamics

Tech stack for this class

Course Requirements

- **Take-home exercises (30%)**
- **Class Participation (10%)**
- **Course Project (40%)**
- **Project Presentation (20%)**

Course Timeline

```
outline\
| -- week  1 : OVERVIEW: DS & DE PERSPECTIVES
| -- week  2 : WORKSHOP - SETTING UP PROJECTS:
                          DS & DE PERSPECTIVES
| -- week  3 : WORKSHOP - VERSION CONTROL & GitHub
| -- week  4 : WORKSHOP - CODING ETIQUETTE
| -- week  5 : WORKSHOP - DATA PIPELINE IN PRACTICE
|      |-- Homework #1
|
| -- week  6 : MISSING DATA & DATA QUALITY
|      |-- Homework #2
\
| -- week  7 : MODEL DEPLOYMENT & VERSIONING,
                WORKING ENVIRONMENTS (DEV, STAGING, PROD)
|      |-- Homework #3
|
| -- week  8 : WORKSHOP - INTERACTIVE WORKING SESSION
| -- week  9 :       -- ACADEMIC HOLIDAY --
| -- week 10 : EXPLANATION v PREDICTION
|      |-- Homework #4
|
| -- week 11 : CONDITIONAL RELATIONSHIPS IN THE DATA
| -- week 12 : MODEL EVALUATION
| -- week 13 : DATA VISUALIZATION
|      |-- Homework #5
|
| -- week 14 : WORKFLOW COLLABORATION
| -- week 15 : PRESENTING RESULTS
```

What is Data Science?

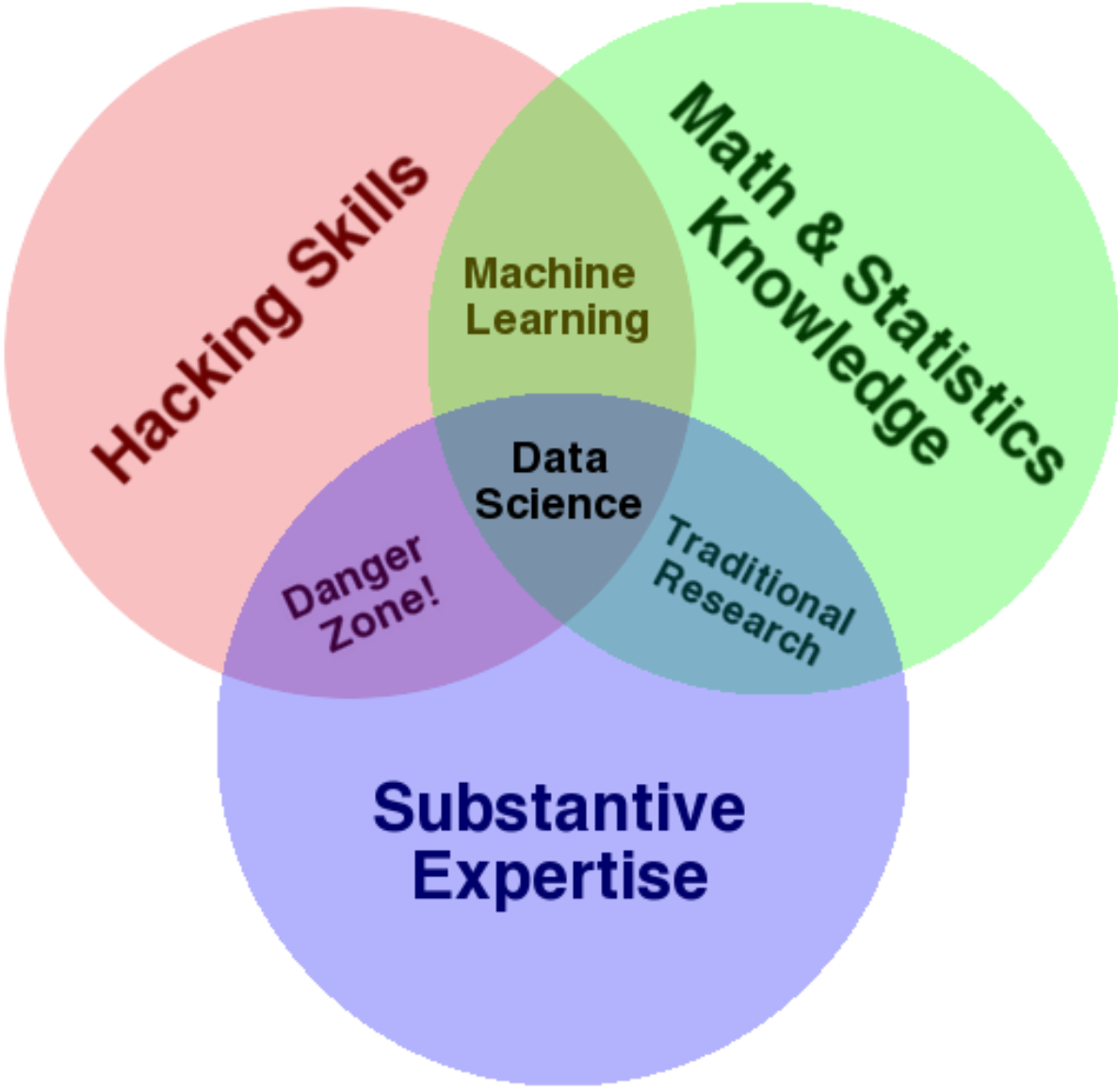
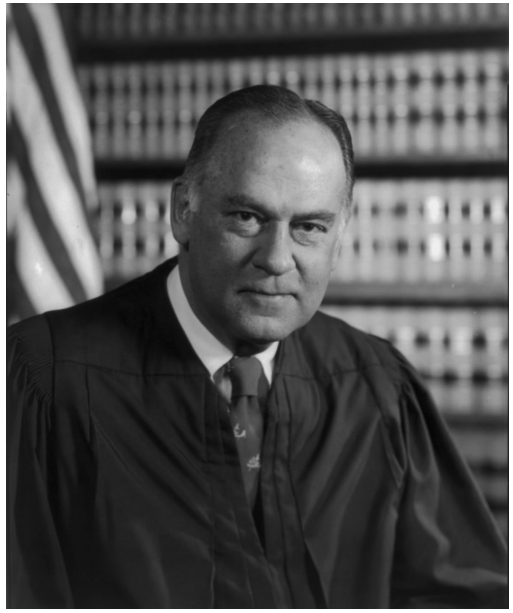


Figure: Drew Conway (2013)

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)

Defining Data Science

is it in the algorithms?

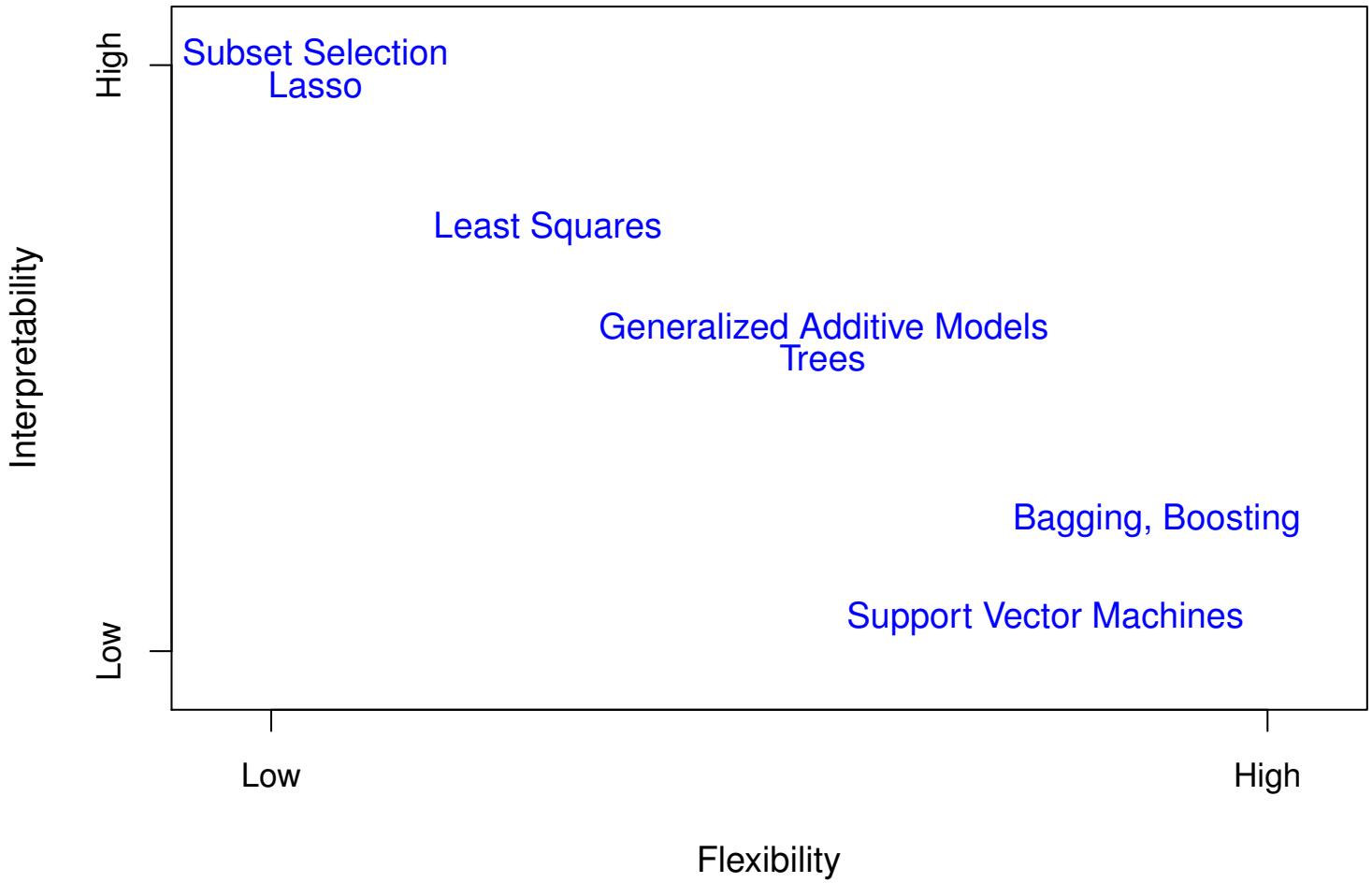


Figure: James et al. (2016)

Defining Data Science

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)

Defining Data Science

is it in the tech stack?

Defining Data Science

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

Defining Data Science

is it in the big data?

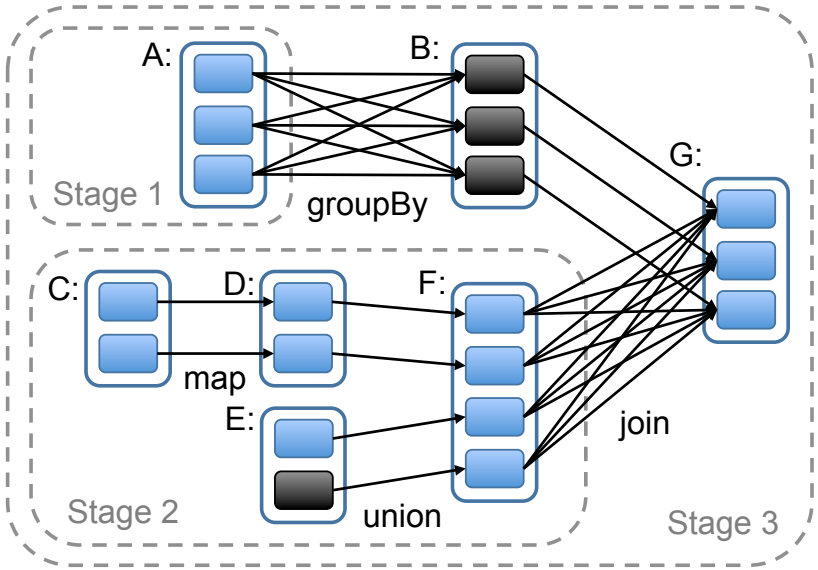


Figure 2.5. Example of how Spark computes job stages. Boxes with solid outlines are RDDs. Partitions are shaded rectangles, in black if they are already in memory. To run an action on RDD G, we build build stages at wide dependencies and pipeline narrow transformations inside each stage. In this case,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

Figure: Matei Zaharia (2014)

Defining Data Science

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

Defining Data Science

is it in the predictive "focus"?

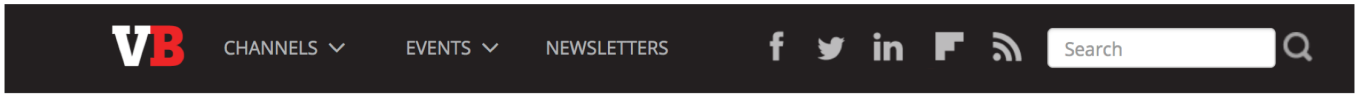
Defining Data Science

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

Defining Data Science

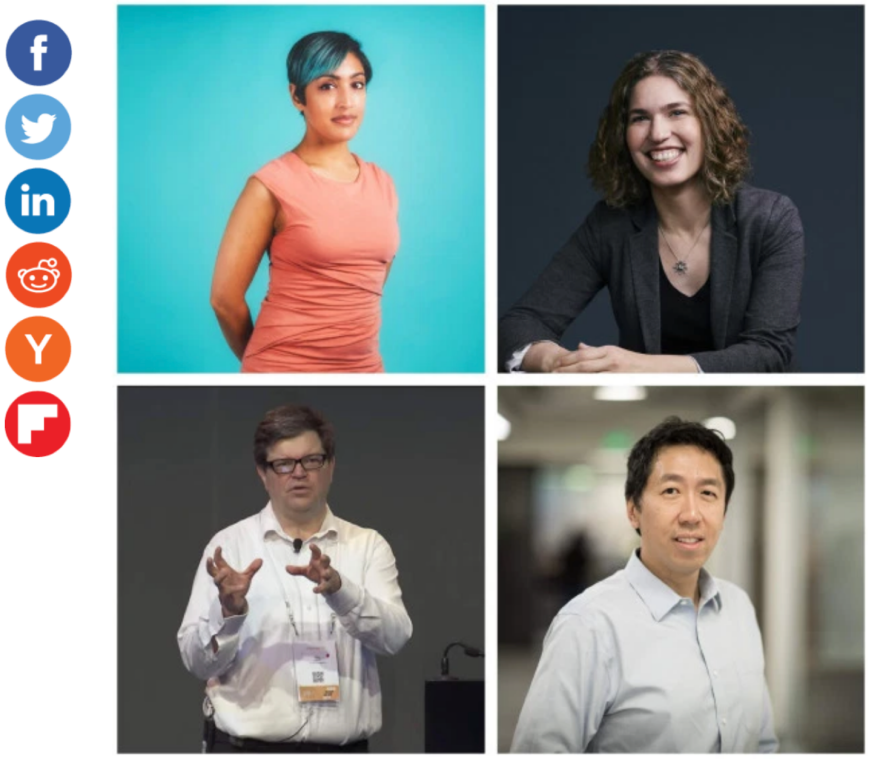
is it the techniques to exploit data?



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

Defining Data Science

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**

Defining Data Science

is it in the “unicorns”?

Defining Data Science

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.

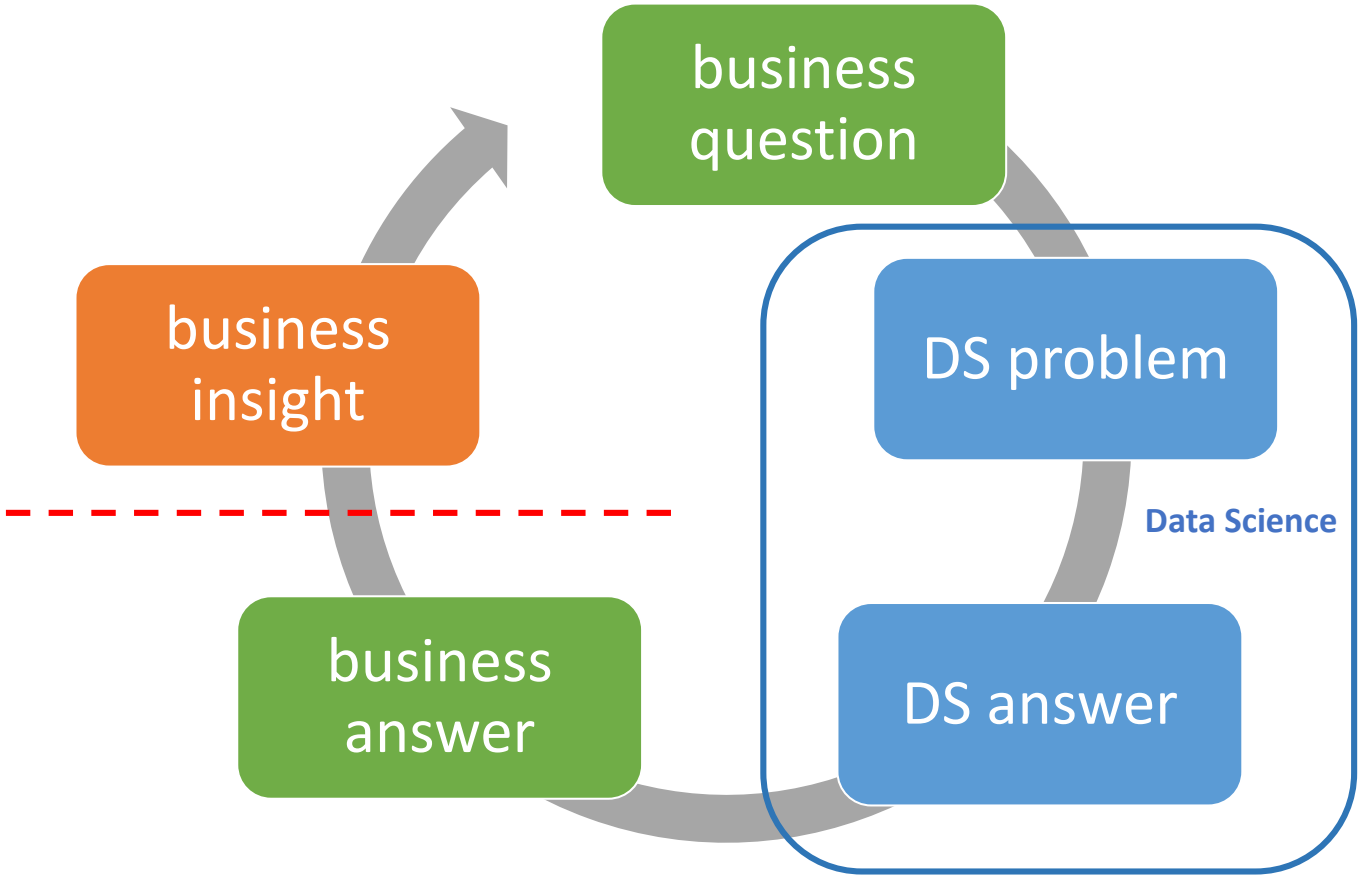
by Thomas H. Davenport and D.J. Patil

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."

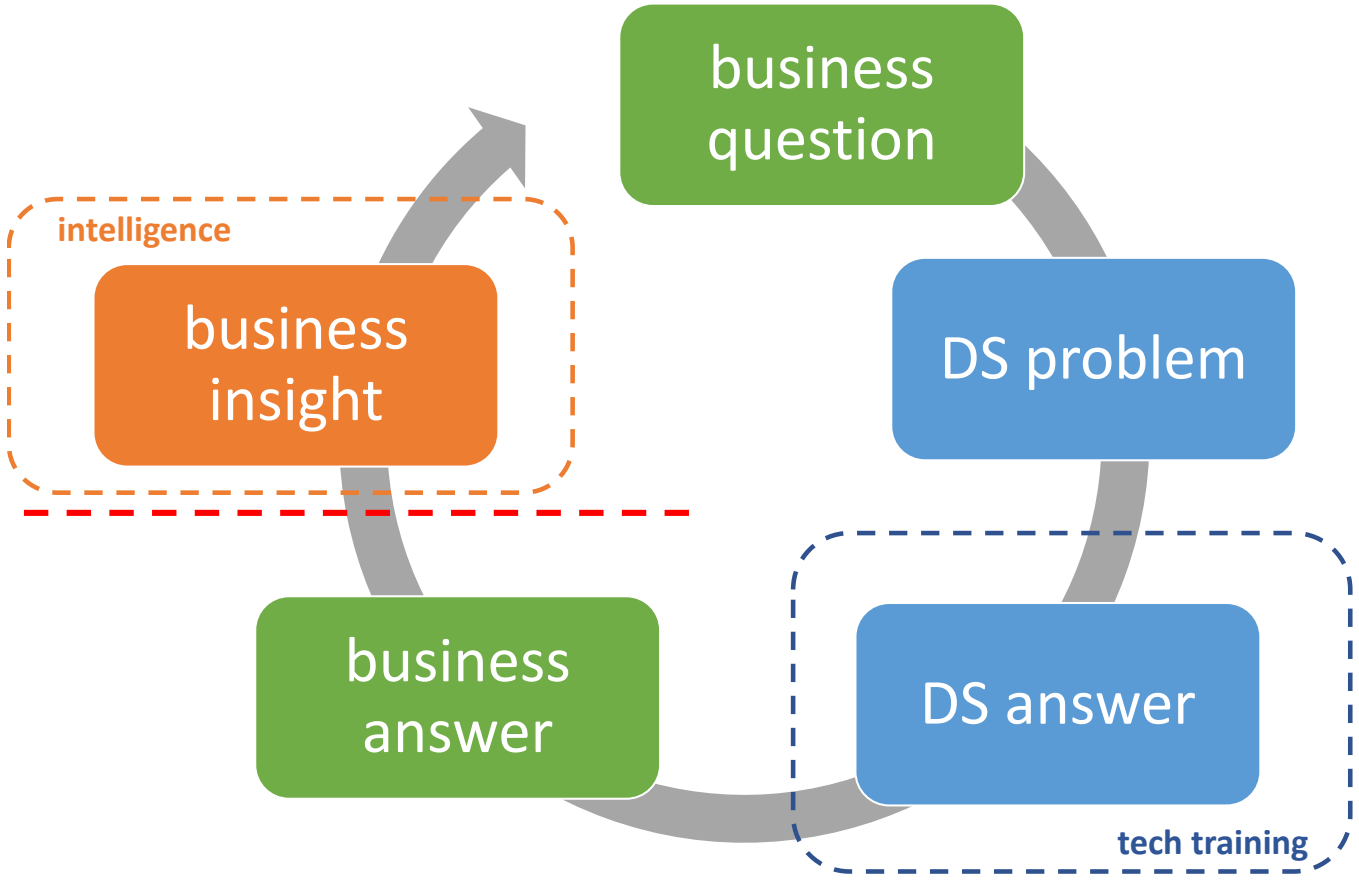
What does a Data Scientist do?

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

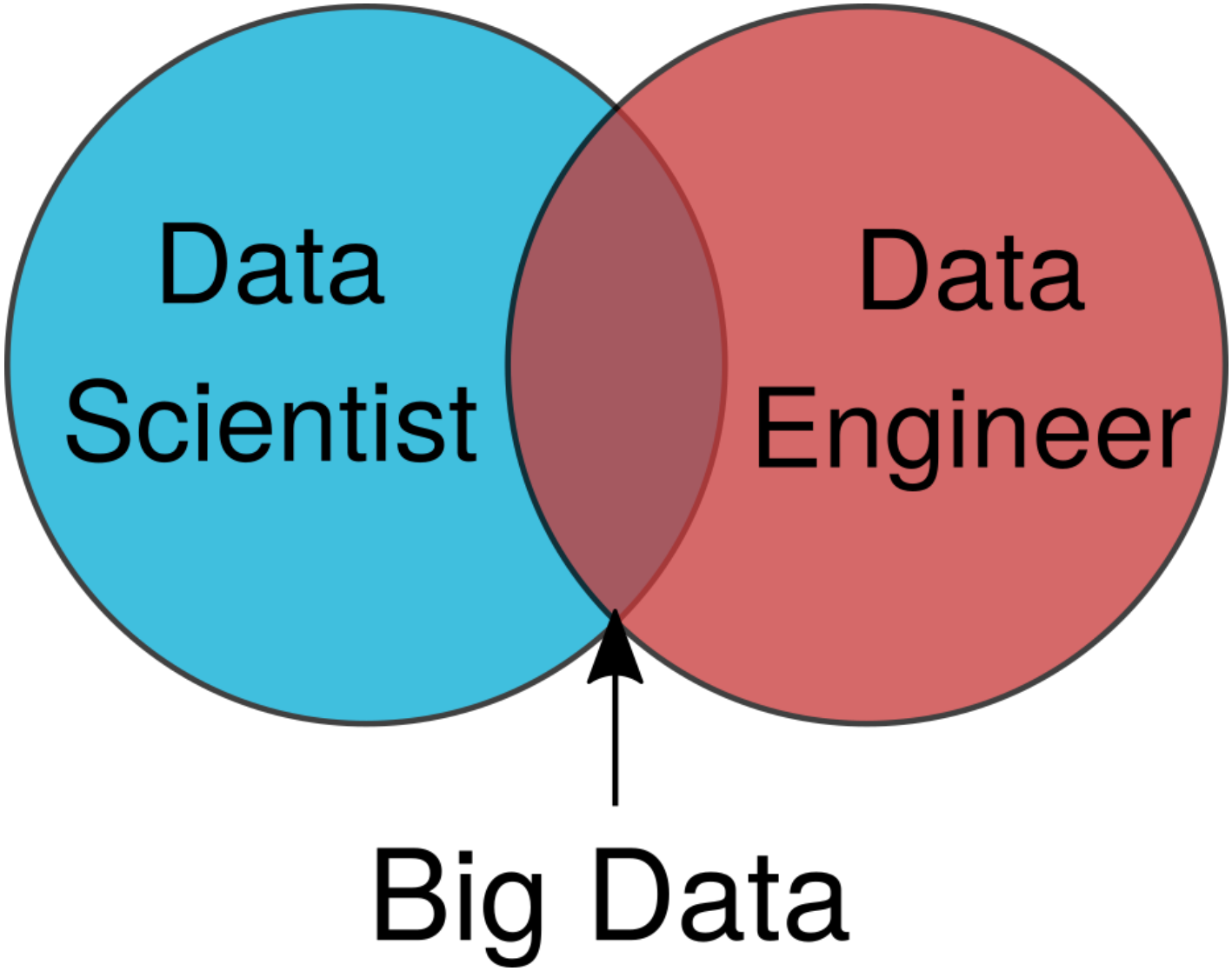
What skills must a Data Scientist have?

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

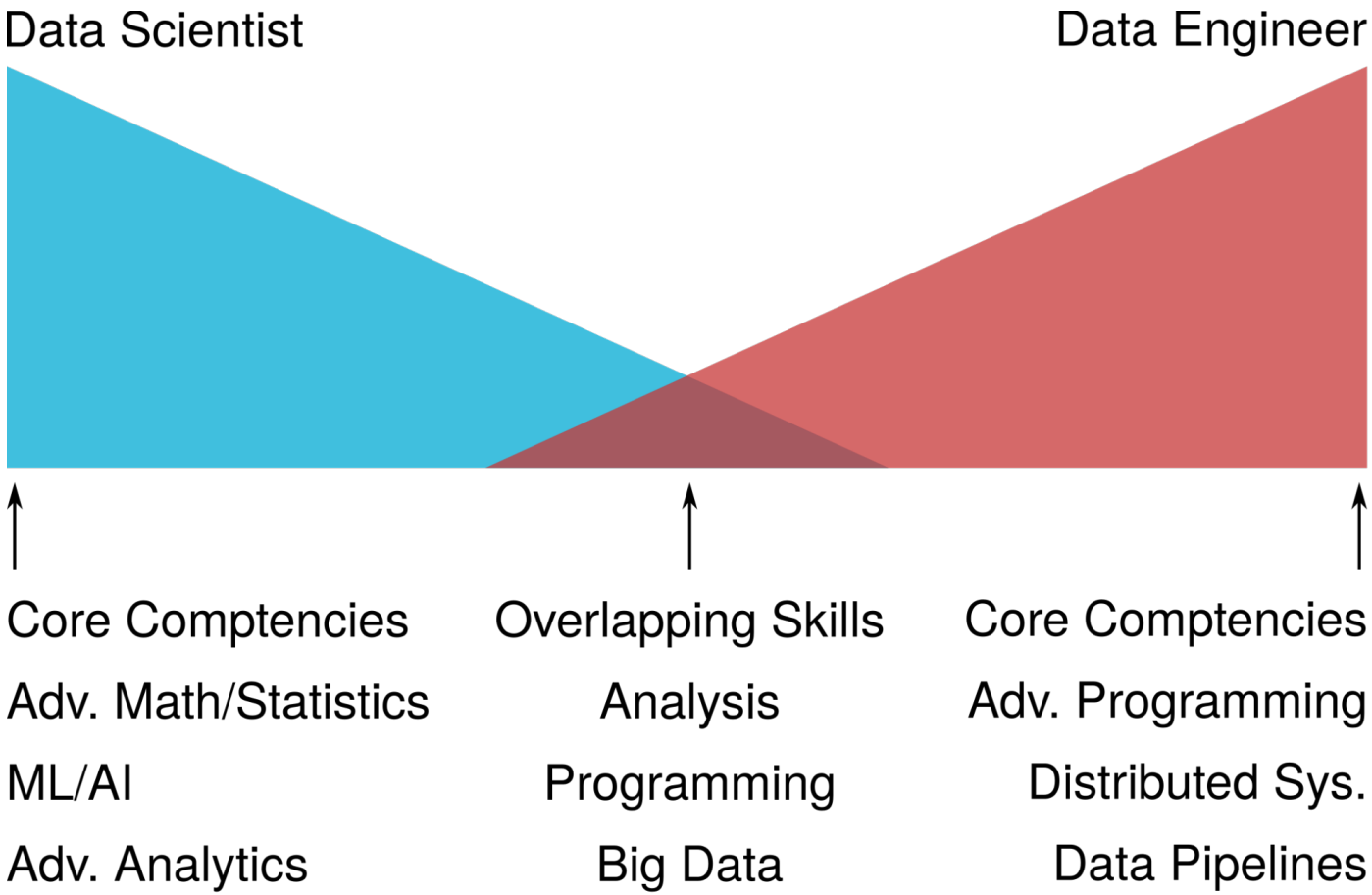
Defining Data Engineering

- 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**

Some Best Practices

Best Practice #1: the Data Science project

- two necessary characteristics of DS projects:
  - **reproducible**
    - a tenet of science (and of hacking too!)
  - **structured**
    - anyone can “understand” the project
- save time for you (and future you), as well as others collaborating in the project
- enabling scaling up of projects if/when needed

Structuring DS projects

a thin layer...

```
project\
|
| -- src                    <- Code
|
| -- data                   <- Inputs
|
| -- reports                <- Outputs
|
| -- references             <- Data dictionaries,
|                              explanatory materials.
|
| -- README.md
```

Best Practice #2: methods to carry out DS projects

the AGILE way...

- **AGILE** is one common method in DS environments
- main entities:
  i) Dev team
  ii) Product Owner
  iii) Scrum Master
- main principle: break project down into tasks and iterate

Carrying out DS projects

the AGILE way: product development

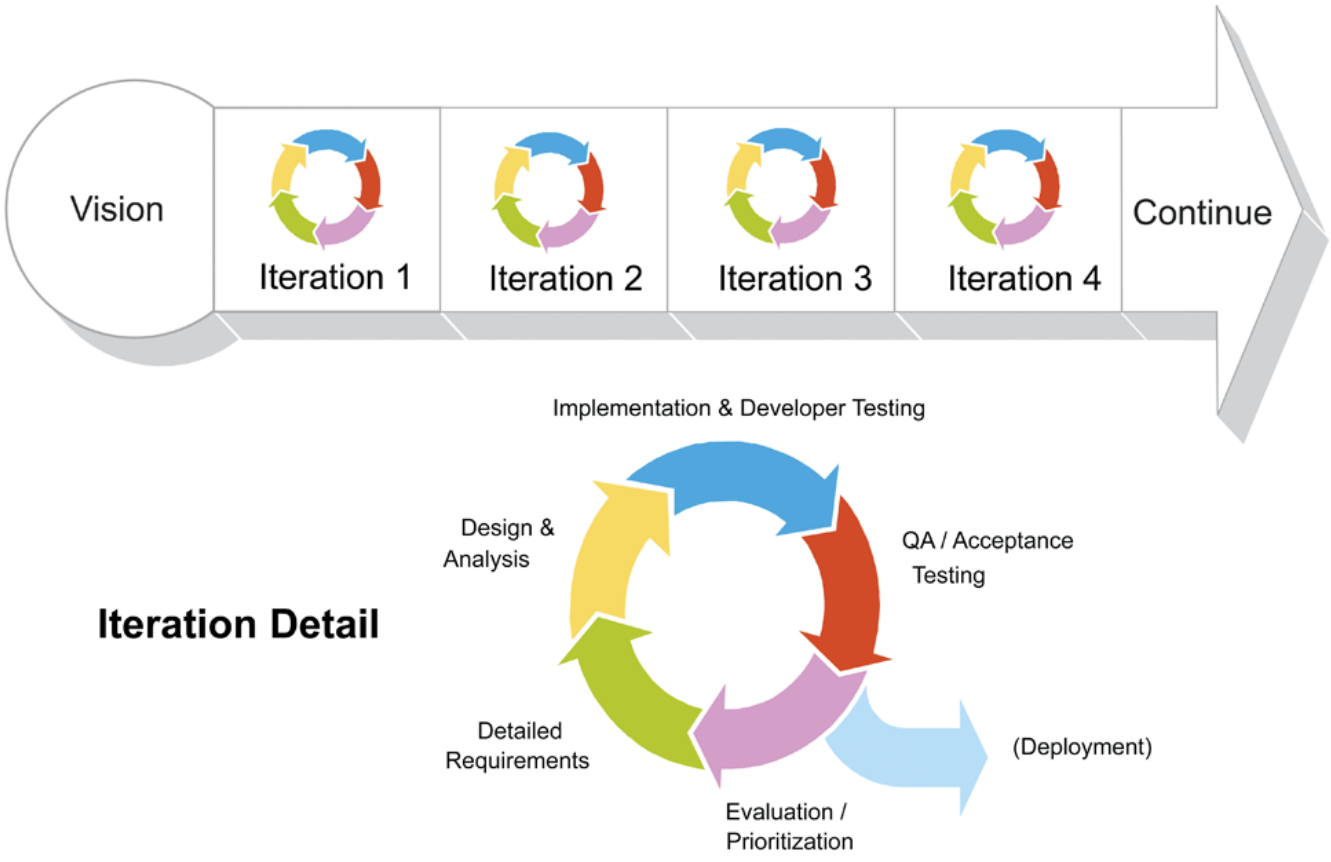


Figure: SCRUM Reference Card

Carrying out DS projects

the AGILE way: Backlog

ETL	Exploration	Analysis	Output
- input data	- descriptives	-modeling	- graphs
- clean data	- visualization		- report
- reshape data			- presentation

- each element to be broken down into **tasks**
- define taks to complete on each **sprint**
- **important concept:** definition of **done**

Carrying out DS projects

the AGILE way: Sprints

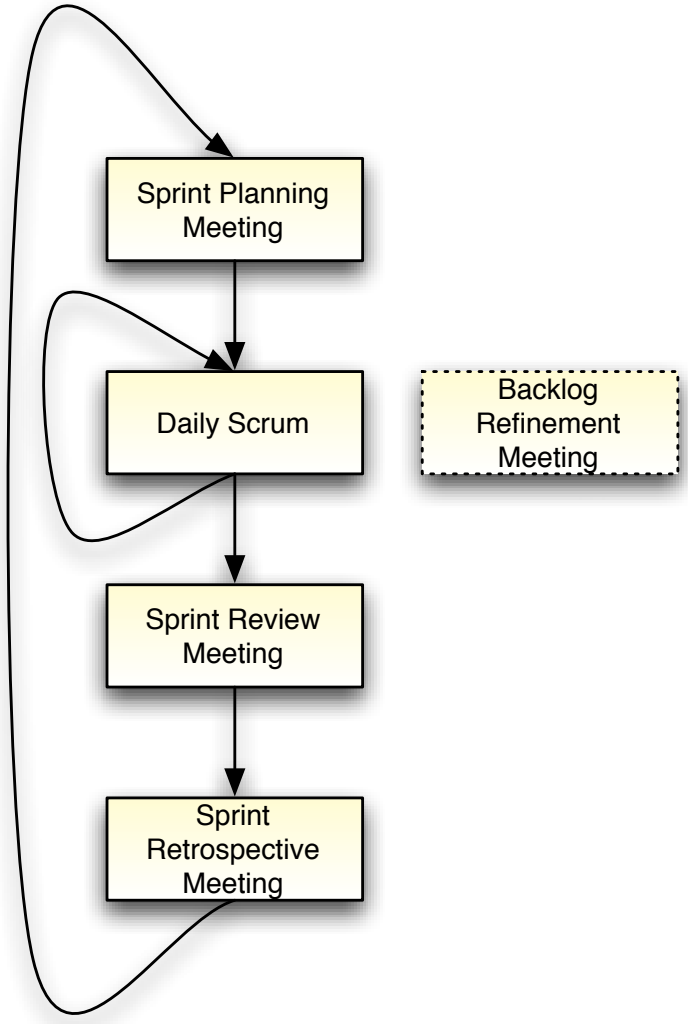


Figure: SCRUM Reference Card

Carrying out DS projects

the Kanban alternative...

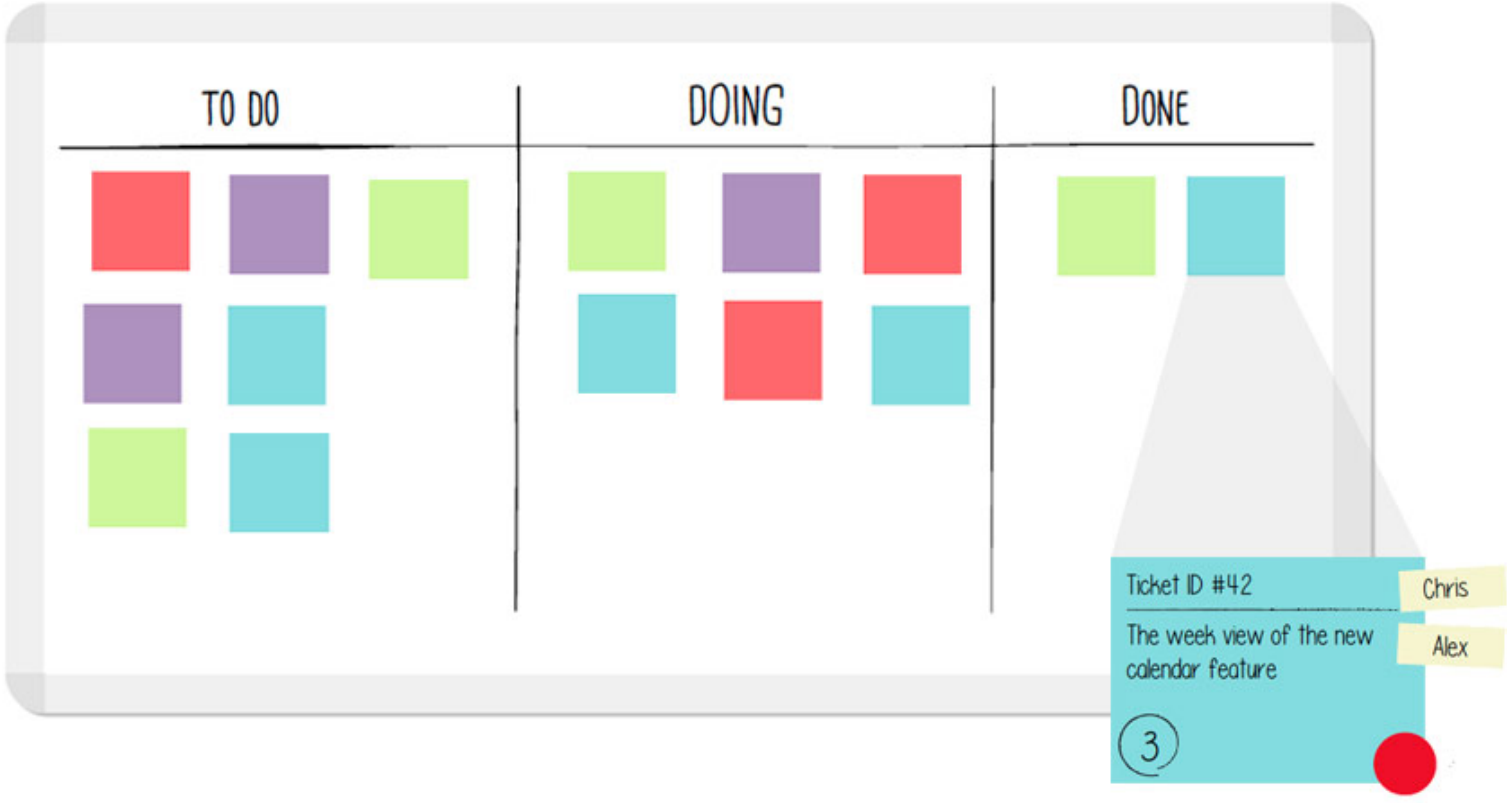
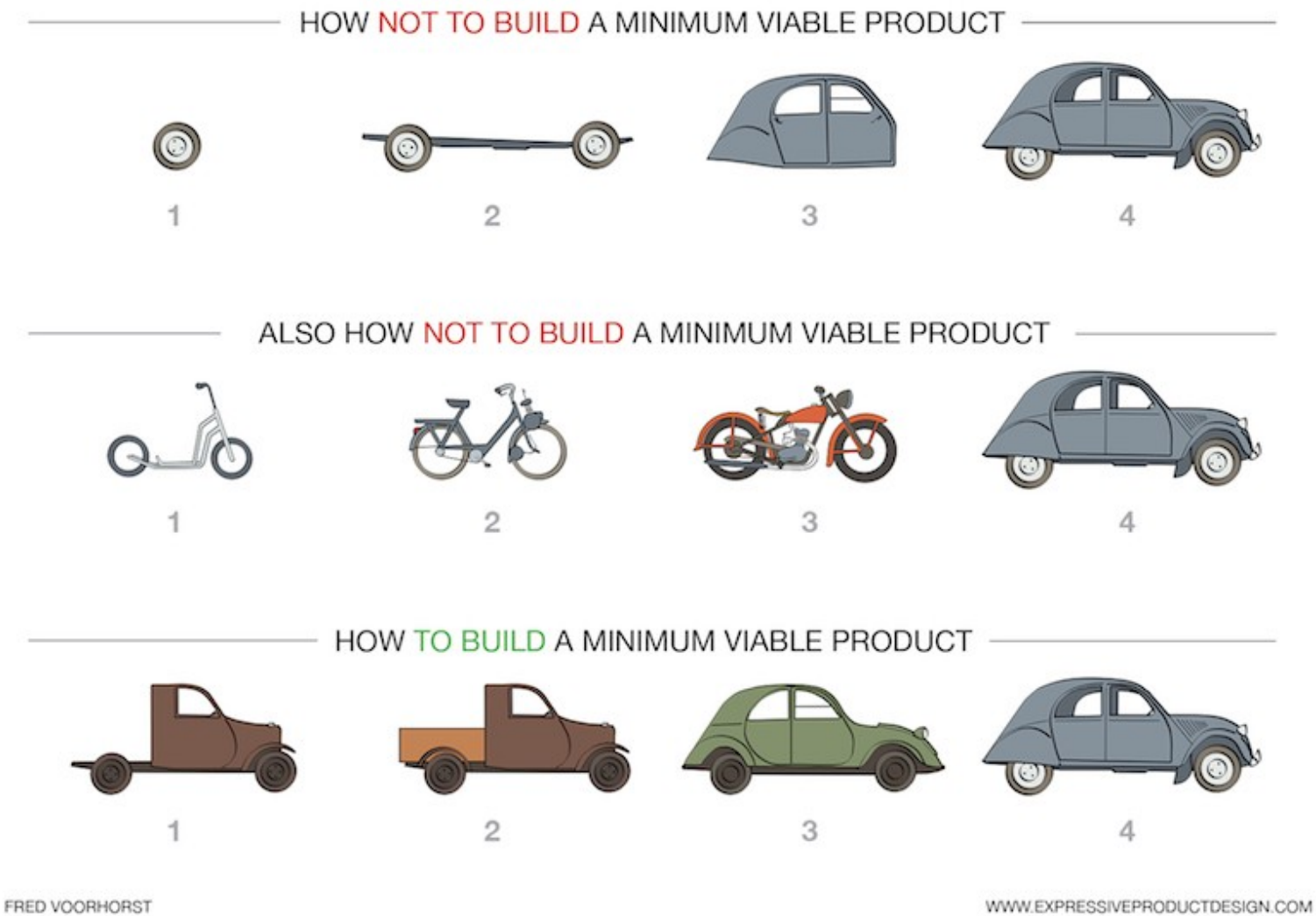


Figure: LeanKit.com

Best practice #3: a minimum viable product

to MVP or not to MVP?



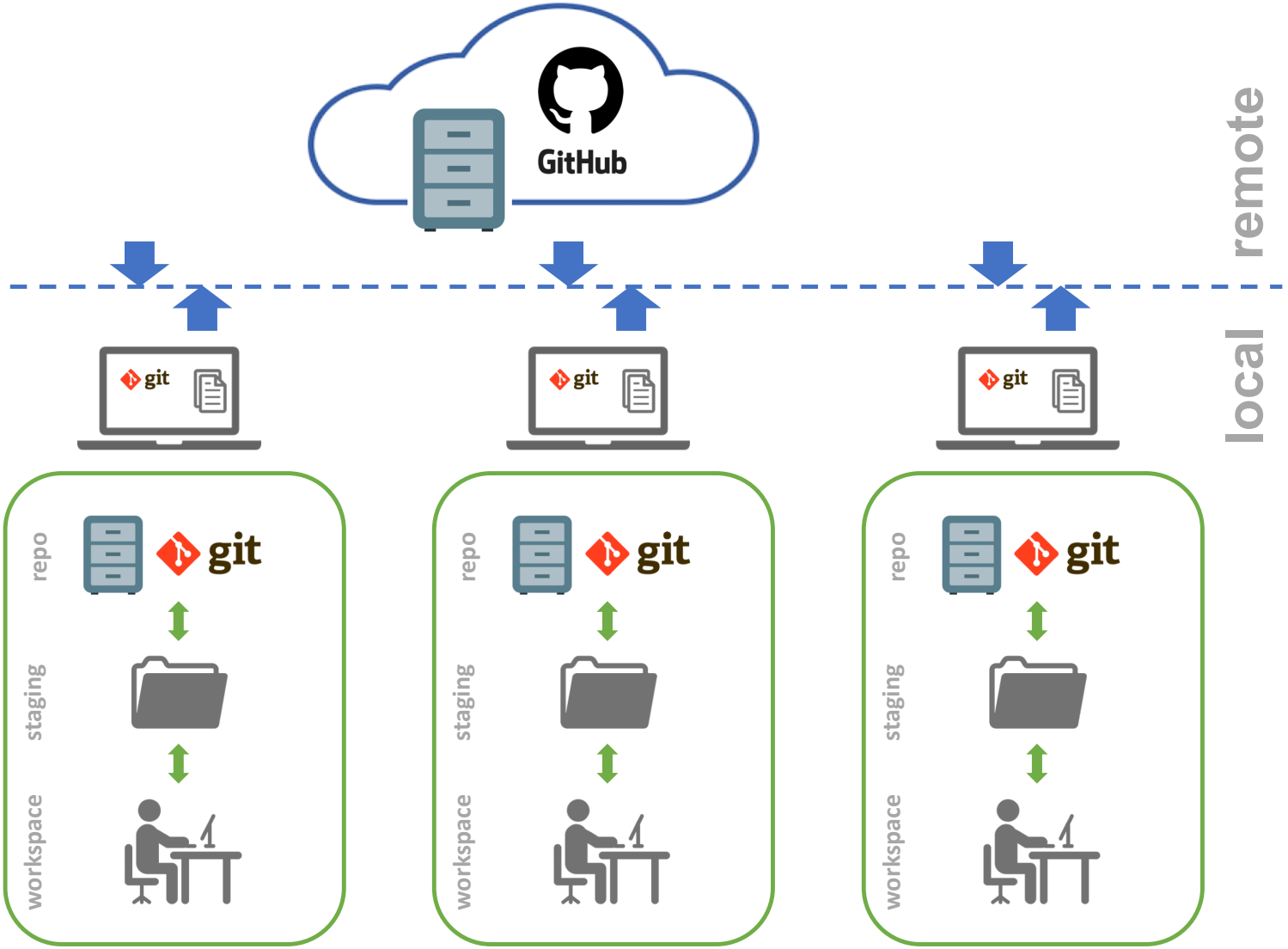
Best practice #4: collaborating using version control

Version control (and Git): though this be madness...

- **version control** allows you to keep track of changes/progress in your code
  - keeps “snapshots” of your code over time
  - helpful to debug, and to enhance reproducibility
  - also great for team collaboration (everyone can see who changed what!)
- **Git** is a version control software
- **GitHub** is an online Git repository (on steroids)
  - widely used by data scientists (and scholars lately)
  - not (strictly) a “software development” tool

Collaborating on DS projects

Version control (and Git): ...yet there is method in't!



Best Practice #5: real-time collaboration

Slack: some etiquette...

- mention people (i.e. **@marco-morales**) when speaking to them directly on a channel
  - people will not be notified unless you mention them
- use **@channel** and **@here** with care
  - **@here** notifies all people currently active in the channel
  - **@channel** notifies all members of the channel
  - **@everyone** notifies all members of the workspace
- be mindful of other people's time and schedules

Best Practice #5: real-time collaboration

Slack: some useful gimmicks...

- Slack works on Markdown, so it's simple to format the text of your messages
- easy to share snippets of code, text, data
- can edit messages after sending them (nice alternative to document)
- integrations with other apps

Housekeeping

IFF you're going to tale this course, make sure to have

a) installed **R** | **R Studio**, and/or **Anaconda** ... we're a language-agnostic course

b) installed **Git** and **Atom** (no other clients for now)

c) followed instructions to join **GitHub Classroom**, **Databricks Community**, **AWS Educate**, and **Slack**

d) cloned the course **GitHub** repo

https://github.com/marco-morales/QMSS-GR5069_Spring2020

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 2020
Columbia University