Best Practices in Data Science for Social Scientists

Marco Morales
mam2519@columbia.edu

GR5069
Topics in Applied Data Science
for Social Scientists

Spring 2018
Columbia University

RECAP: What is Data Science?

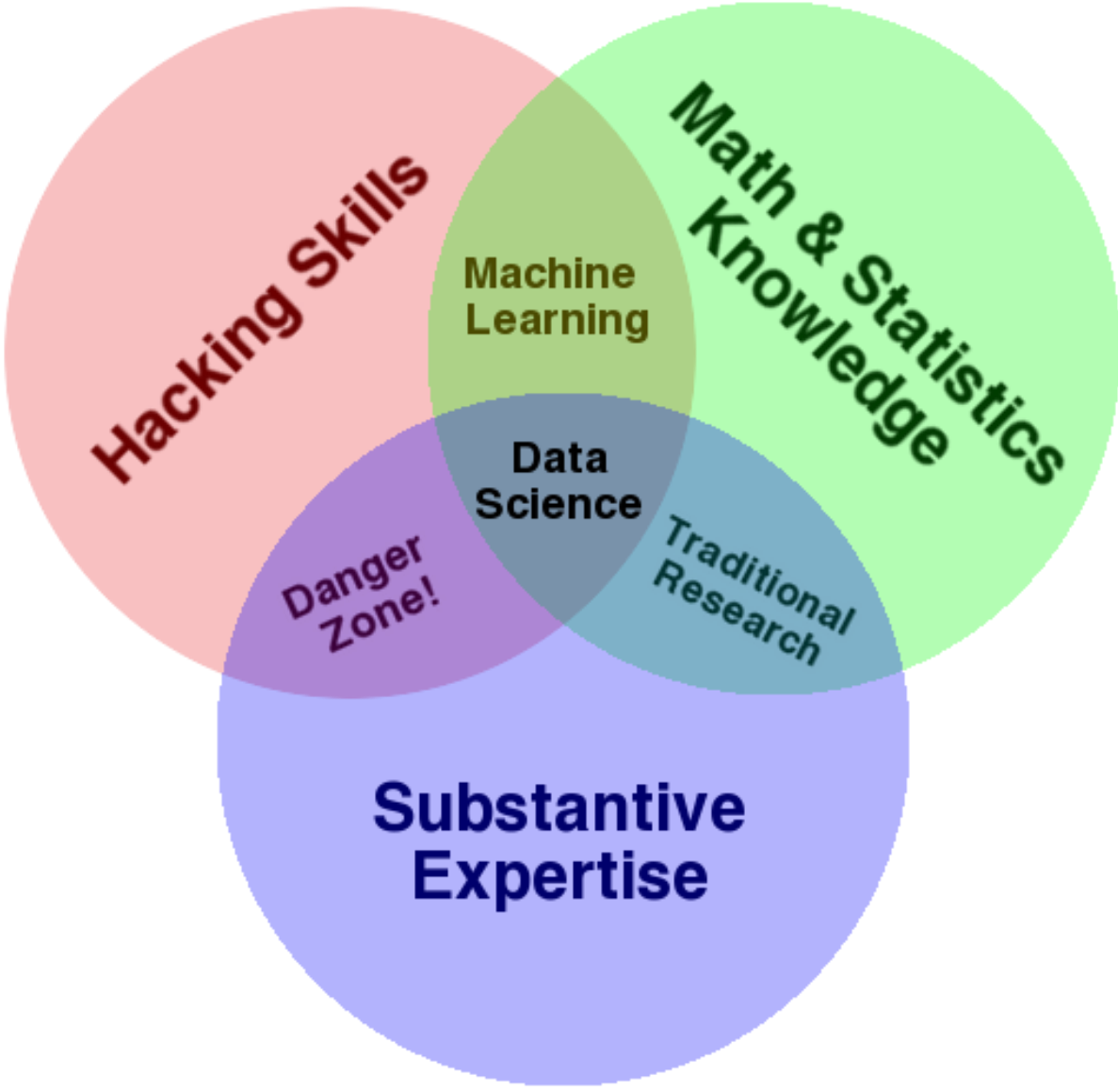

Figure: Drew Conway (2013)

What is Data Science?

is it the tools?

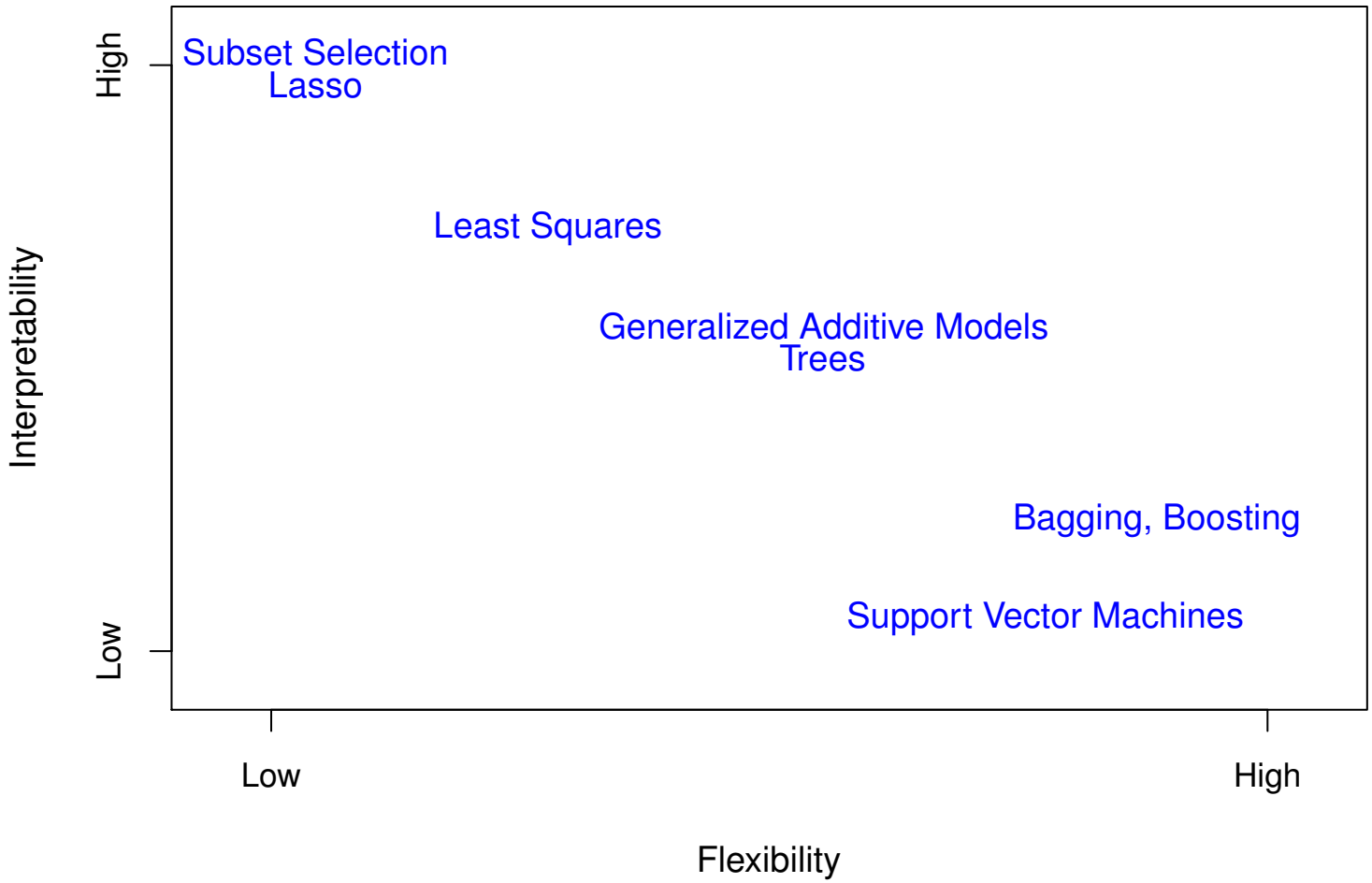

Figure: James et al. (2016)

What is Data Science?

is it the tools?

- is it really that **different from applied statistics**?
- after all, ML is also **statistical learning**...

What is Data Science?

is it big data?

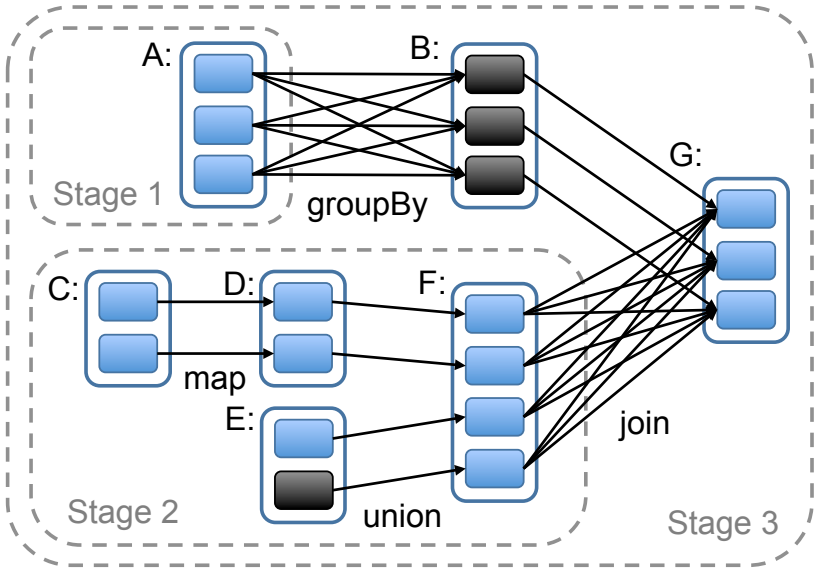

Figure 2.5. Example of how Spark computes job stages. Boxes with solid outlines are RDDs. Partitions are shaded rectangles, in black if they are already in memory. To run an action on RDD G, we build build stages at wide dependencies and pipeline narrow transformations inside each stage. In this case,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

Figure: Matei Zaharia (2014)

What is Data Science?

is it 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 VERY basic tasks
  - we're only beginning to scratch the surface
  - promise in techniques that require **a lot** of data

What is Data Science?

is it the predictive "focus"?

What is Data Science?

is it 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 DS **education**
  - ...though not necessarily **practice**
- more important in some industries than others

What is Data Science?

the unicorn myth

What is Data Science?

the unicorn myth

- Data Science is **collaborative** in nature
  - no single person possesses all
    - skills
    - substantive knowledge
    - expertise
- most 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**

What is Data Science?

in reality...

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 is Data Science?

now, seriously...

- what a data scientist does:
  - learn from data (evidence-based)
  - generate predictive or inferential insights
  - create reproducible and transferable products
  - (potentially) scalable products
- skill(set) of a data scientist:
  - coding (hacking)
  - data transformation (ETL)
  - data exploration / visualization
  - database usage
  - modeling / analysis
  - communication
  - collaboration

A 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
| -- TODO                   <- (opt)
| -- LabNotebook            <- (opt)
```

Carrying 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

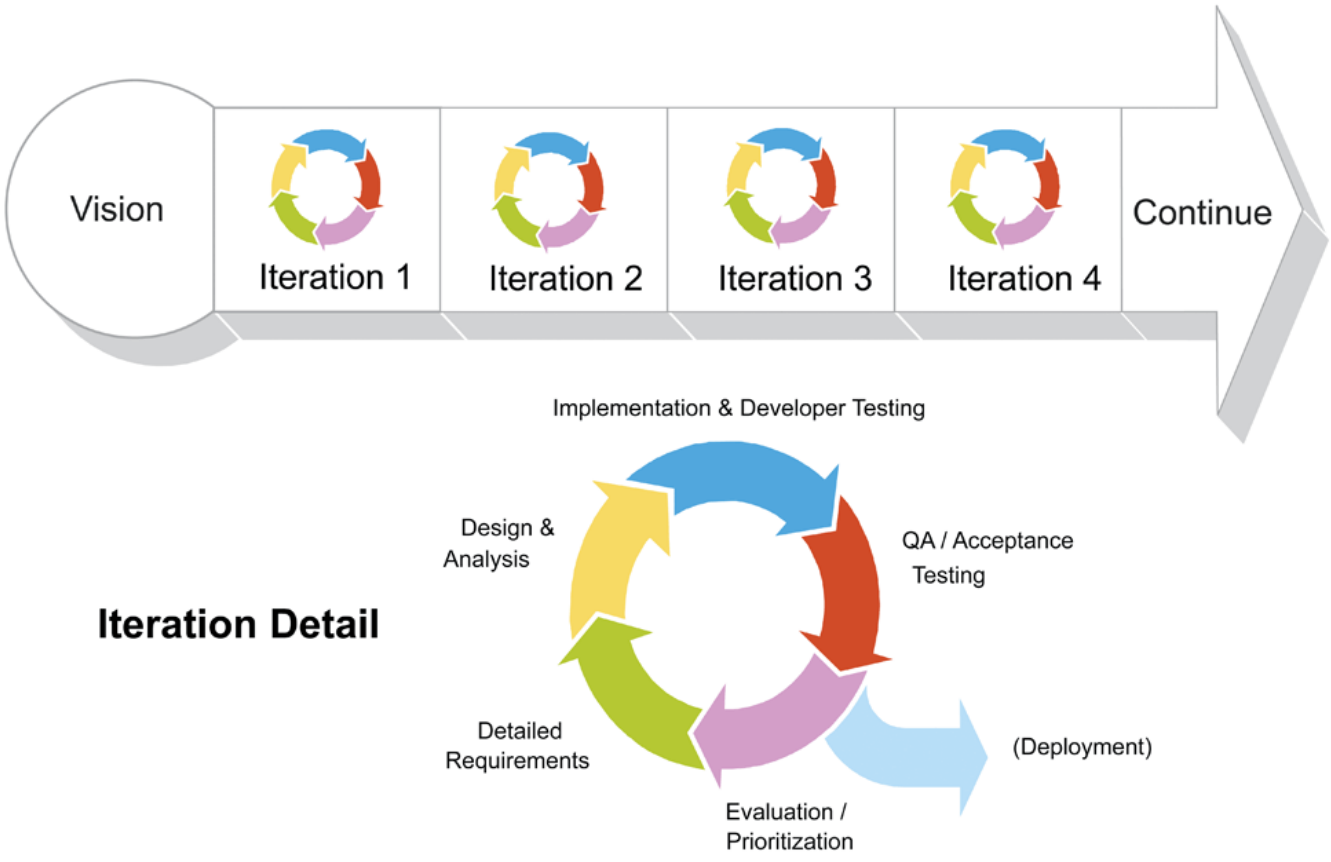

Figure: SCRUM Reference Card

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

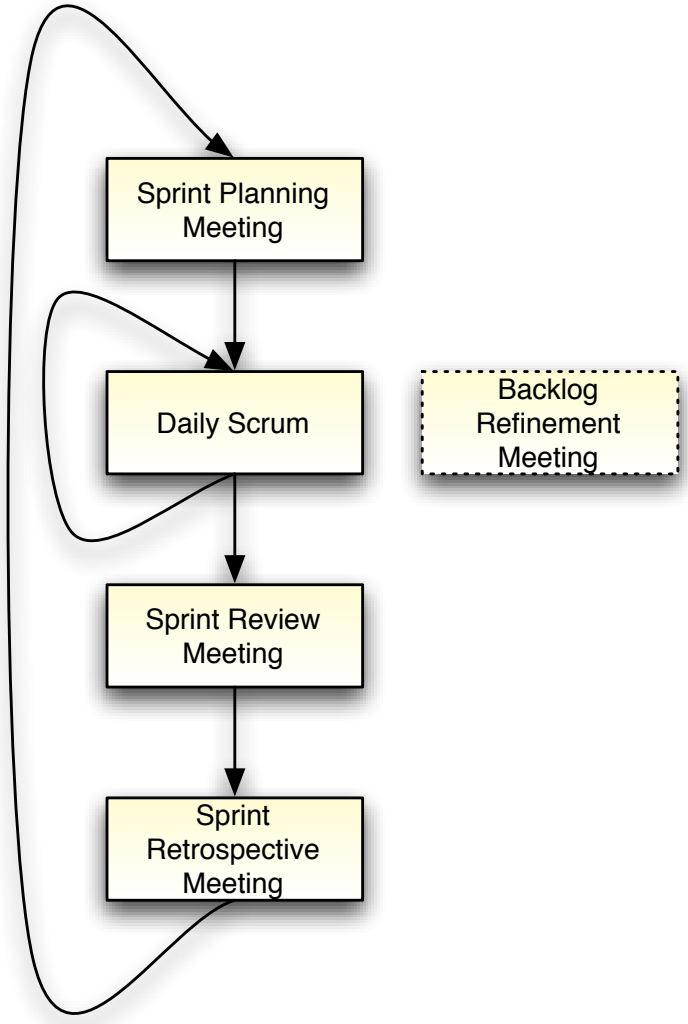

Figure: SCRUM Reference Card

Carrying out DS projects

the Kanban alternative...

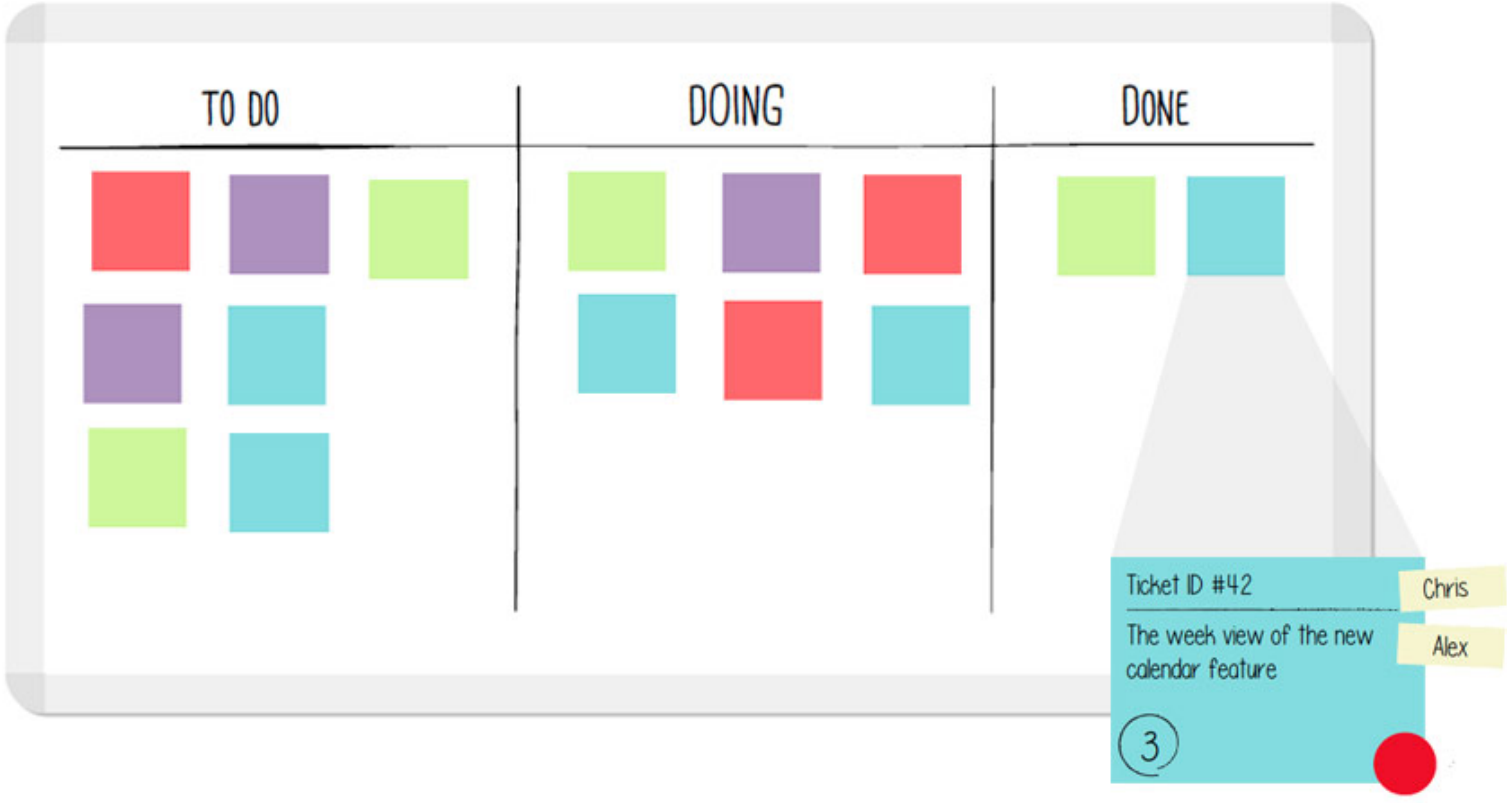

Figure: LeanKit.com

Collaborating on DS projects

Slack: getting started...

- after signing up for the course's Slack workspace
- add your name to your profile:

  @xyz2209 might not make it easy for people to find you

- join all class-related channels and stick to their purpose
  - channels serve to order conversations
  - you will not get notified of messages on channels you are not a member of
- create channels for your teams or other purposes

Collaborating on DS projects

Slack: class-related channels...

- **#anything-git**: solving Git/GitHub questions collaboratively
- **#anything-r**: solving R questions collaboratively
- **#anything-tidyverse**: solving tidyverse questions collaboratively
- **#anything-viz**: solving visualizations in R questions collaboratively
- **#datachallenge-n**: collaboration on solving each data challenge
- **#general**: all class-related communications, announcements and questions
- **#random**: everything else

Collaborating on DS projects

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 team
- be mindful of other people's time and schedules

Collaborating on DS projects

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

Collaborating on DS projects

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!

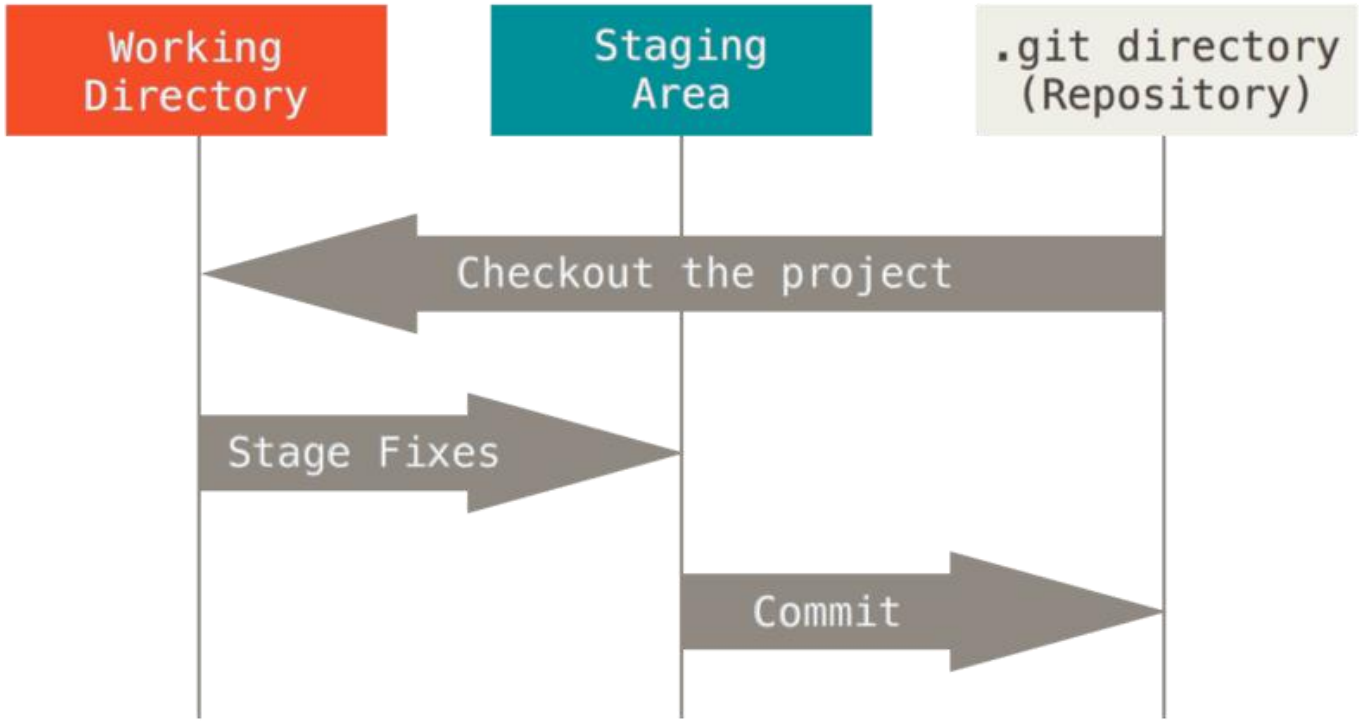

Figure: Pro Git,2nd Edition

Housekeeping

IFF you'll be taking the course, by next week make sure to

a) sign up for the **Slack** workspace

https://columbia-gr5069.slack.com/signup

b) clone the course **GitHub** repo

https://github.com/marco-morales/QMSS-GR5069_Spring2018

- You'll be randomly assigned to teams. Be prepared by week 4 to:
  - communicate your project
  - create a backlog
  - have planning session

Best Practices in Data Science for Social Scientists

Marco Morales
mam2519@columbia.edu

GR5069
Topics in Applied Data Science
for Social Scientists

Spring 2018
Columbia University