The Data Science Workflow

CSE 491H/691H – Intro to Data Science and Big Data

Last Time

- Worked on Python
- Started Learning about Generators

Today

- Administrative Stuff
- Go over Python solutions
- The Data Science Workflow
- Let’s talk Git…

Piazza Poll

- Finally up
- Please vote – your feedback will help me make the course better.

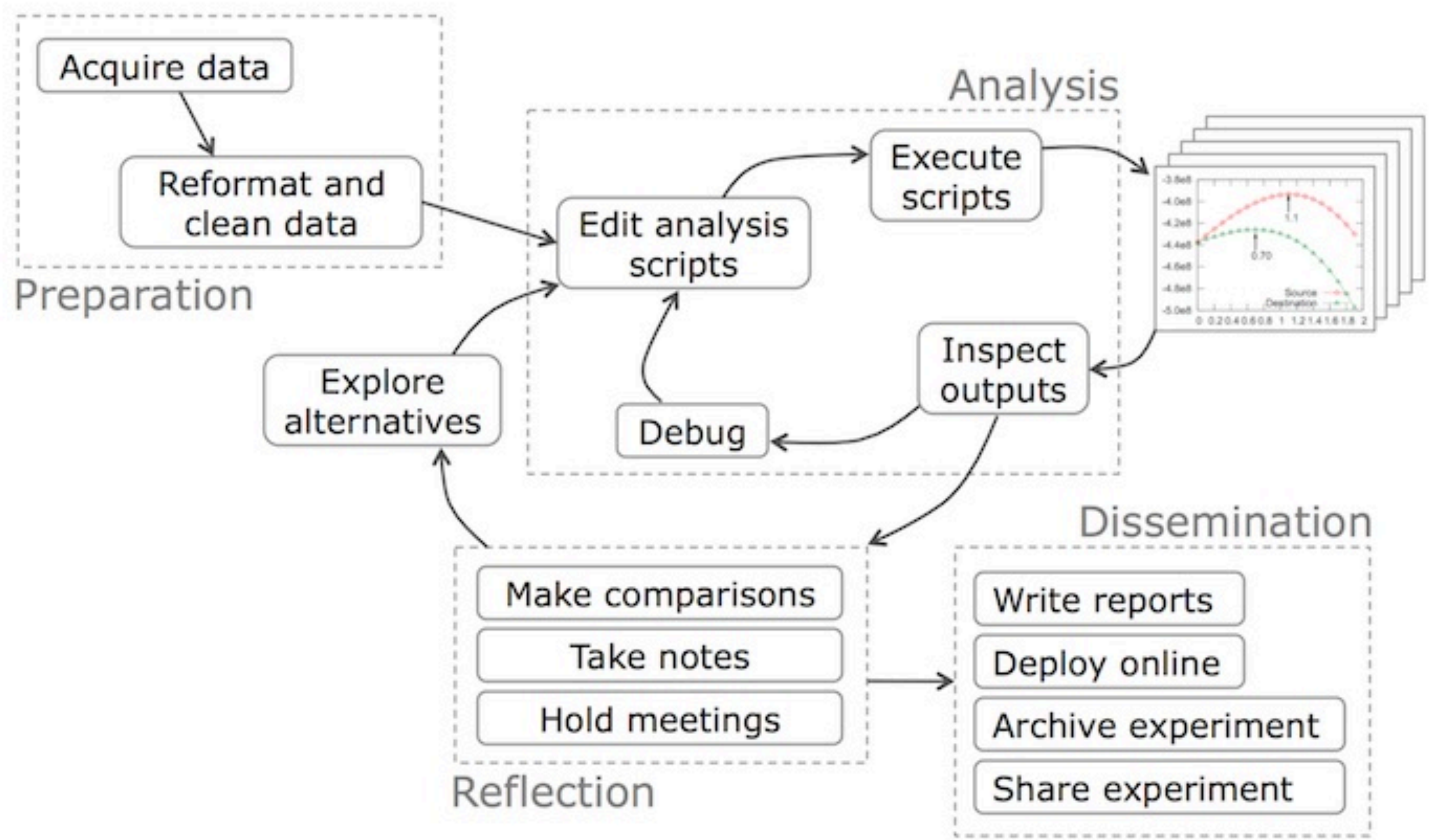
Python Generators

- Let’s walk through the code…

The Data Science Workflow

- We want to turn data analysis into a process.
- We'll look at a particular method, proposed by Philip Guo
  - You may have heard of him from *The Ph.D. Grind.*
    - http://www.pgbovine.net/PhD-memoir.htm
  - We're going to focus on his recent CACM article.
    - http://cacm.acm.org/blogs/blog-cacm/169199-data-science-workflow-overview-and-challenges/fulltext

The Data Science Workflow



Preparation

Acquire data

Reformat and clean data

Analysis

Edit analysis scripts

Execute scripts

Inspect outputs

Debug

Explore alternatives

Reflection

Make comparisons

Take notes

Hold meetings

Dissemination

Write reports

Deploy online

Archive experiment

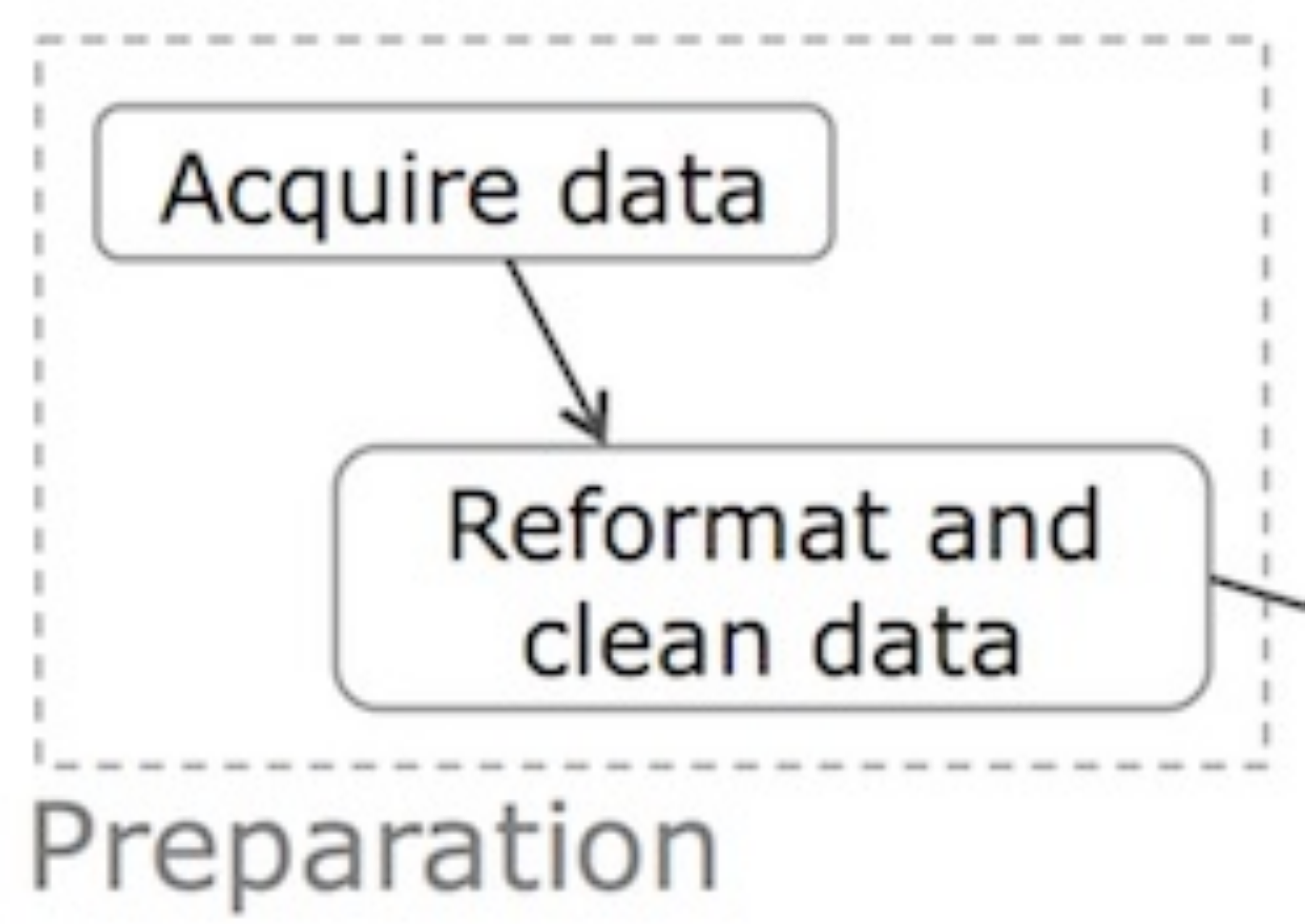
Share experiment

Philip Guo, *Data Science Workflow: Overview and Challenges*

The Main Phases

- Preparation
- Analysis
- Reflection
- Dissemination

Preparation



Preparation

- Where does data come from?
  - CSV files
  - APIs
  - Streaming APIs
  - Scientific Equipment
  - Logs
  - Excel Spreadsheets
  - Websites
  - Freeform text

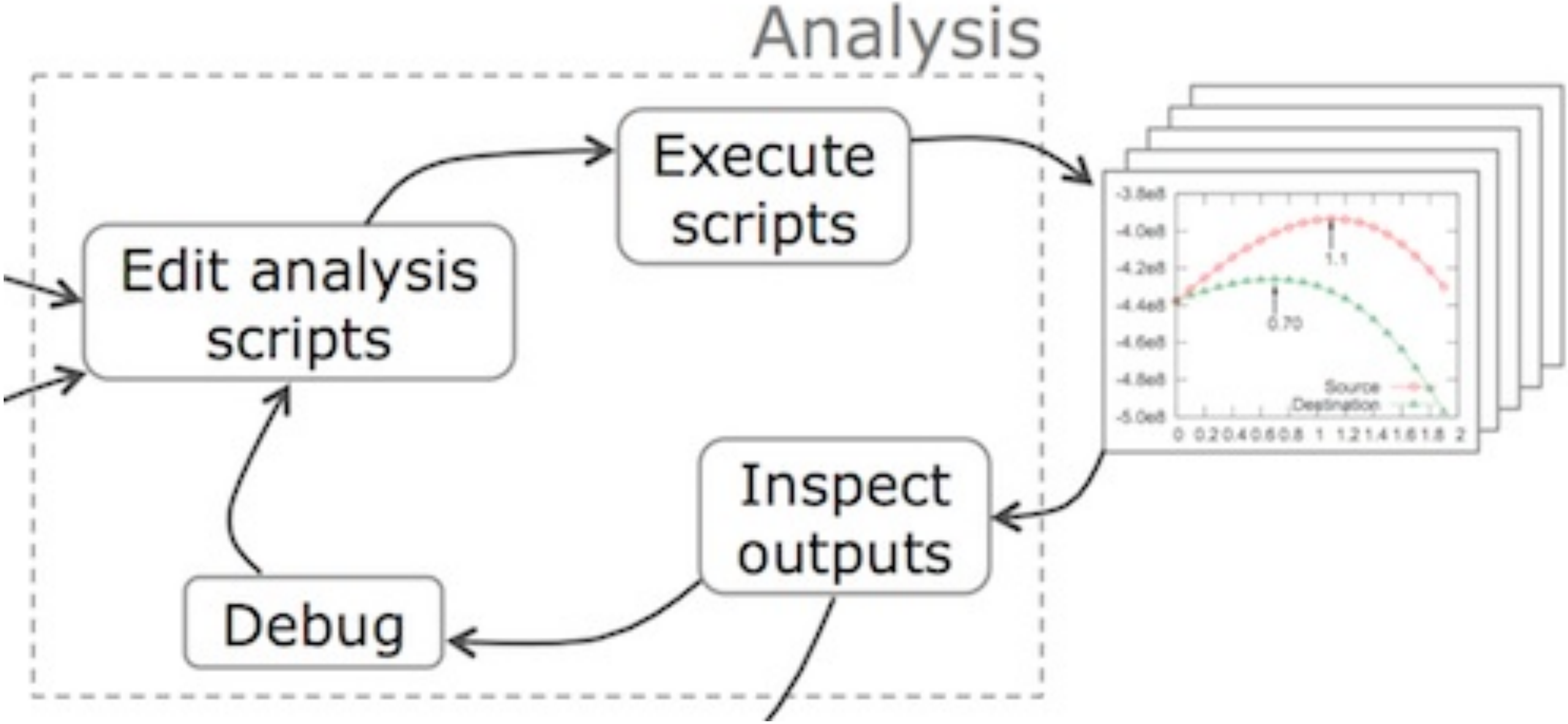
Preparation

- How do we get the data?
  - Need to parse CSV files
  - Need to pull from web APIs
  - Need to scrape webpages
  - Need to parse text
    - Sometimes for very deep definitions of “parse.”
- We’ll look at Python libraries for all of these operations in the coming weeks.

Preparation

- How do we clean and reformat data?
  - Data is rarely if ever in the nice clean format you might see in stats classes.
  - Missing values…
  - Lots of formats
    - Text
    - XML
    - CSV
  - This is often the most tedious part and valuable part of the process.

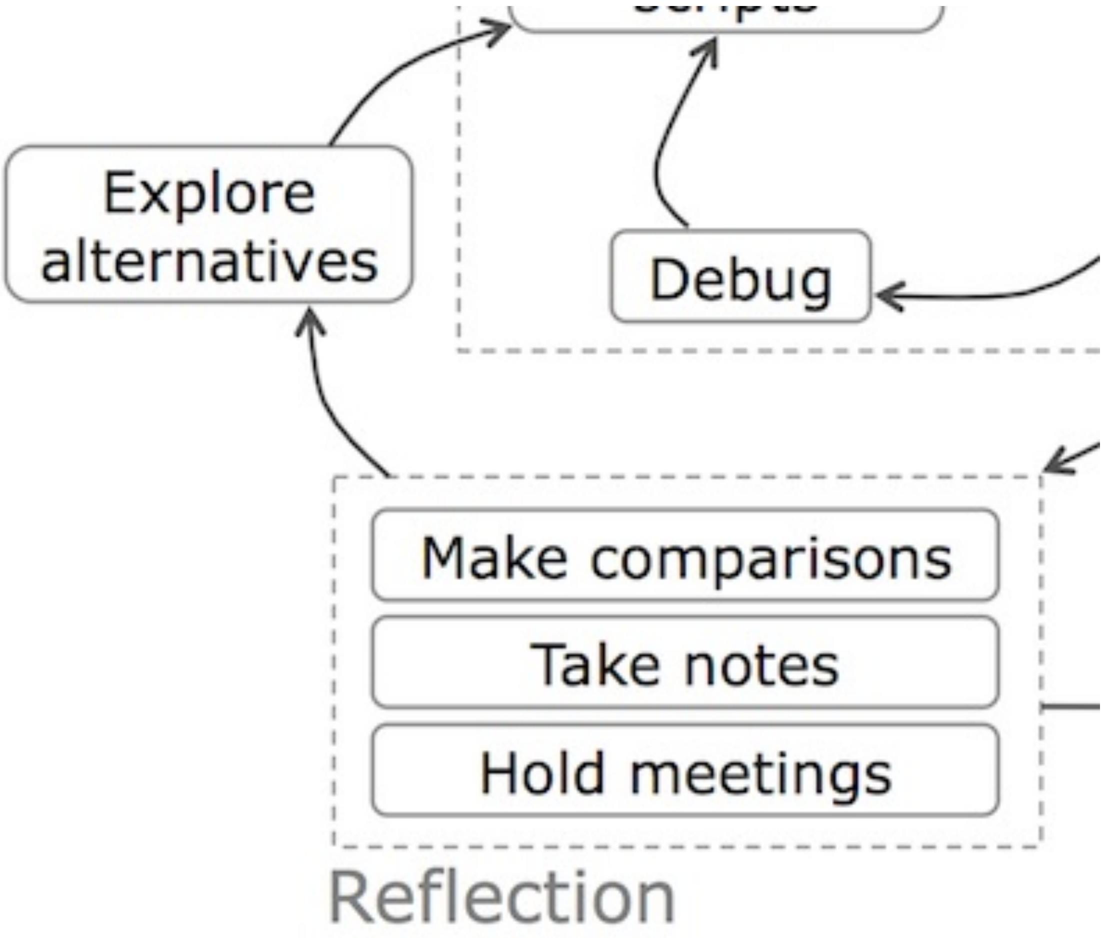
Analysis


Analysis
Edit analysis scripts
Execute scripts
Inspect outputs
Debug
-3.8e8
-4.0e8
-4.2e8
-4.4e8
-4.6e8
-4.8e8
-5.0e8
1.1
0.70
Source
Destination
0 0.2 0.4 0.6 0.8 1 1.2 1.4 1.6 1.8 2

Analysis

- What kinds of analysis can we do?
  - Inference
    - Describe the population we're looking at data for.
    - We won't spend a bunch of time on this – take a stats class.
  - Prediction
    - Regression
    - Classification

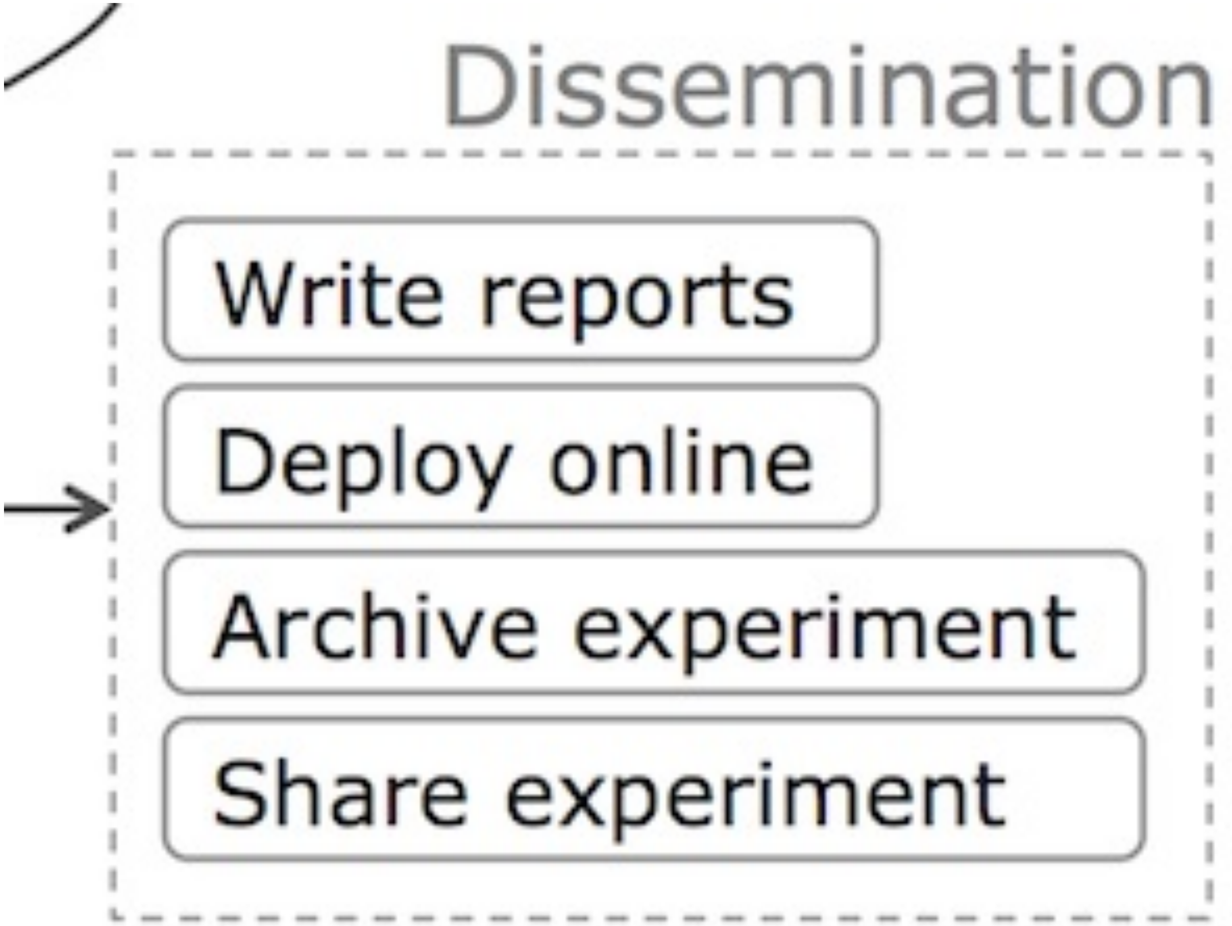
Reflection


Explore alternatives
Debug
Make comparisons
Take notes
Hold meetings
Reflection

Reflection

- This is the part we'll focus on the least.
  - We will talk about comparisons we can make between models and analyses.
  - Won't talk about meetings or taking notes.

Dissemination



Dissemination

Write reports

Deploy online

Archive experiment

Share experiment

Dissemination

- We're going to look at two aspects of dissemination in particular
  - Visualization
  - Presentations

Questions?

- Go read the paper. You'll be expected to know details on the exam.
  - What's *provenance*?

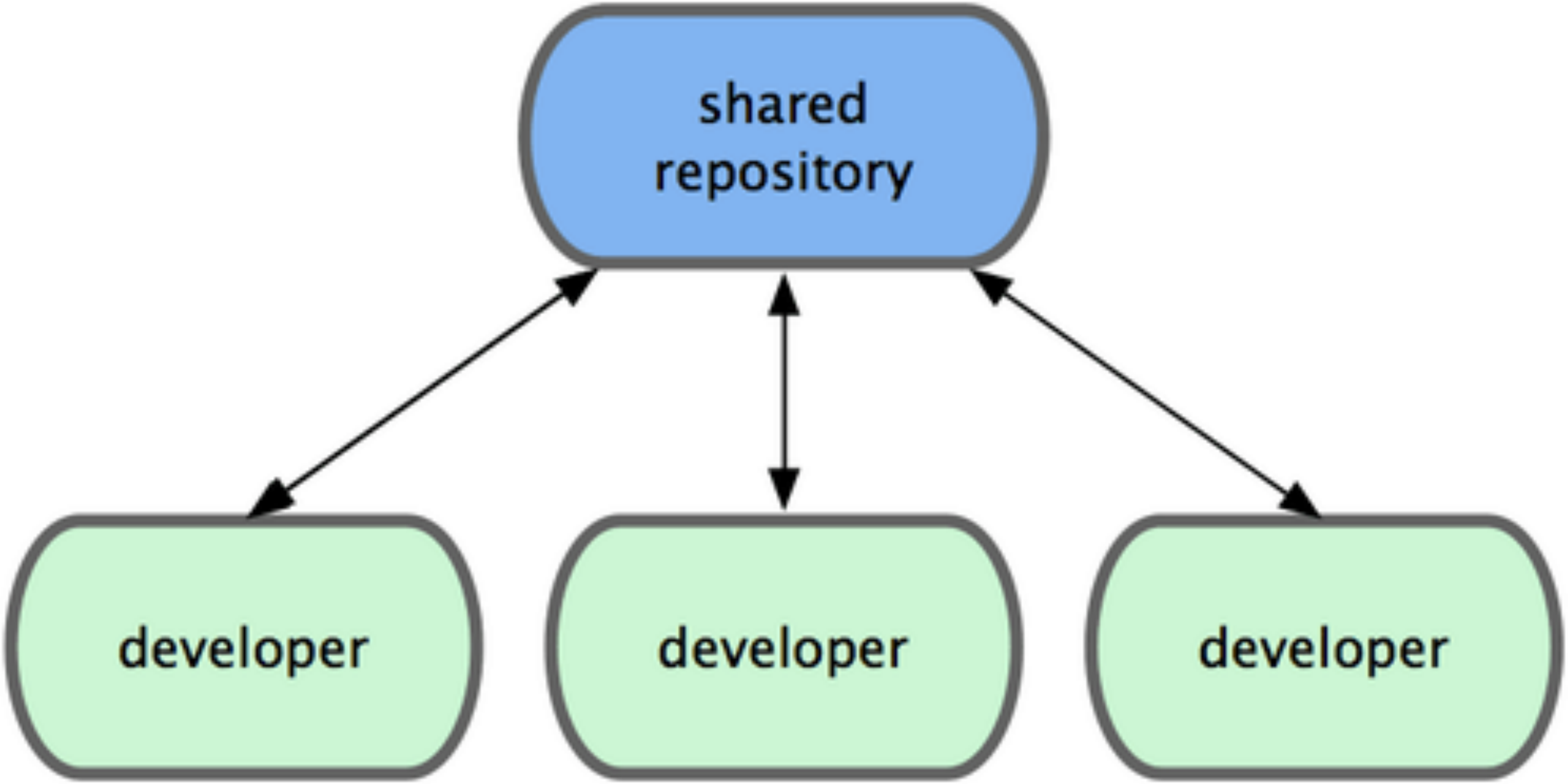
Git

- Let’s talk Git & Github…

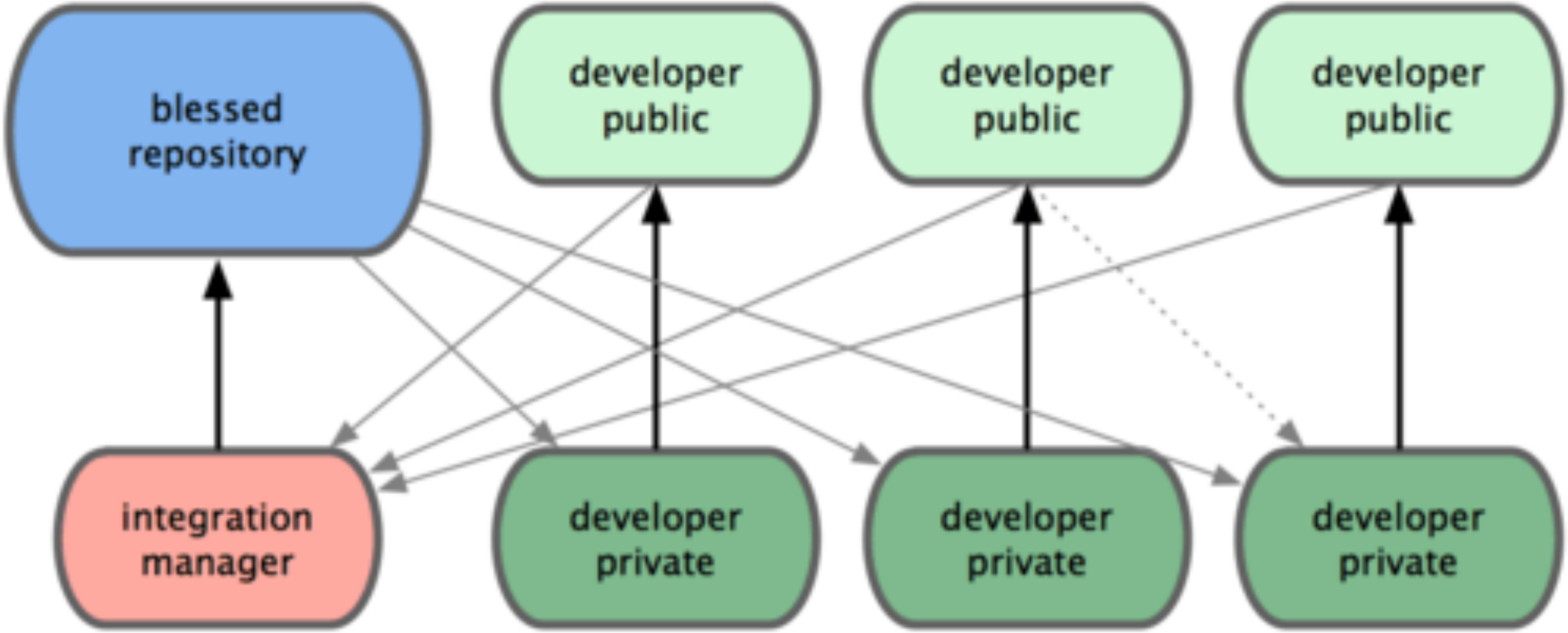
Look at the Git Book…

- http://git-scm.com/book
- That's where the diagrams come from…
- But first…
  - Do we all understand branching and merging?

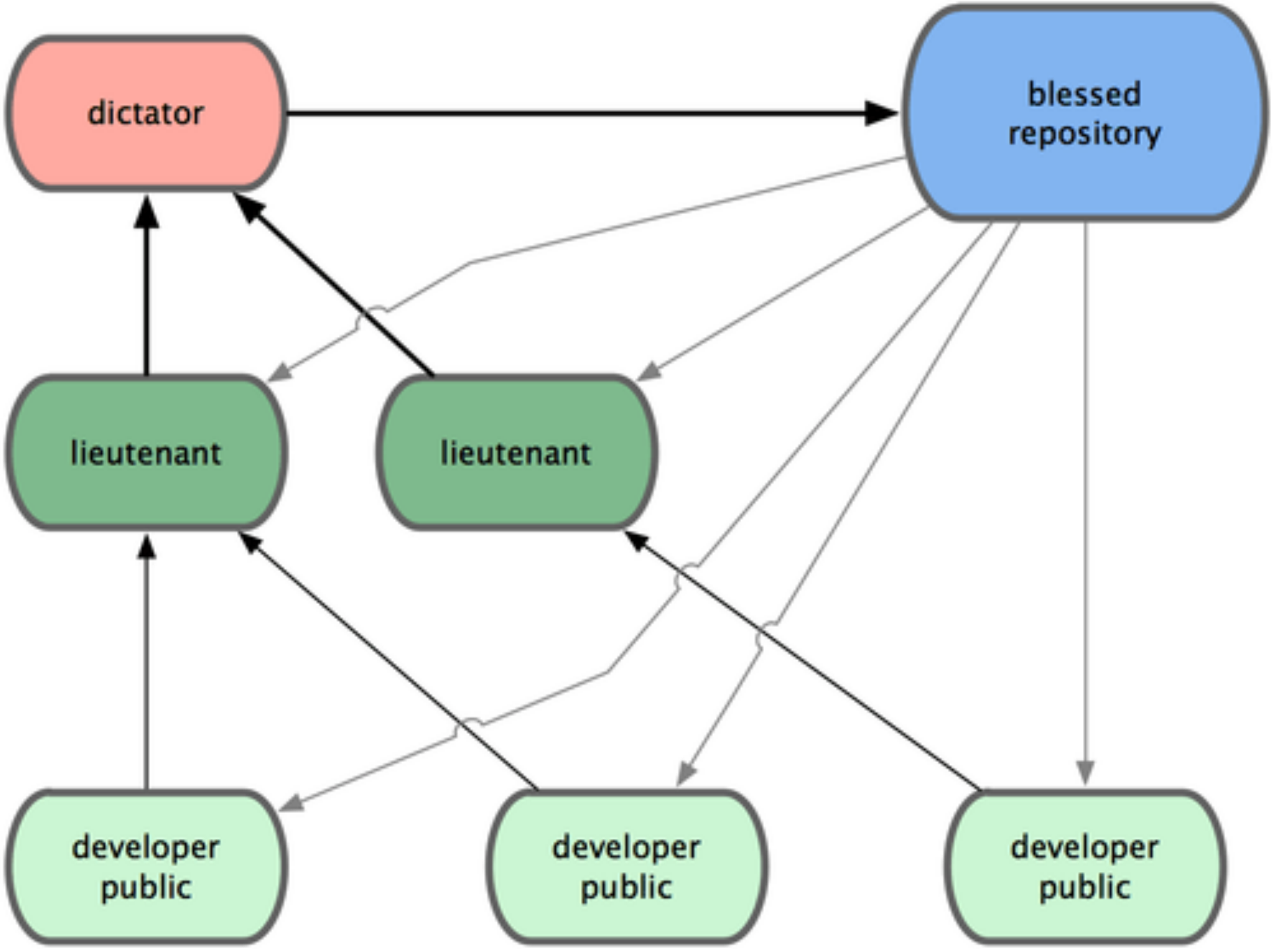
Centralized Workflow



Integration Manager Workflow



Dictator Workflow


dictator
blessed repository
lieutenant
lieutenant
developer public
developer public
developer public