

# So you want to become a Data Scientist?

Lecture 1

Sep 4, 2019



# What is data science?

- Data Science represents a new approach to
  - Acquire knowledge,
  - Collect evidence,
  - Form decisions,
  - Make predictions.
- The end points are:

**knowledge, evidence, decisions and predictions.**
- Driven by breakthroughs in technologies.
- Enabling faster solutions to traditional evidence-based practices.
- Creating solutions that would not be otherwise possible.

# A real example - AlphaGo

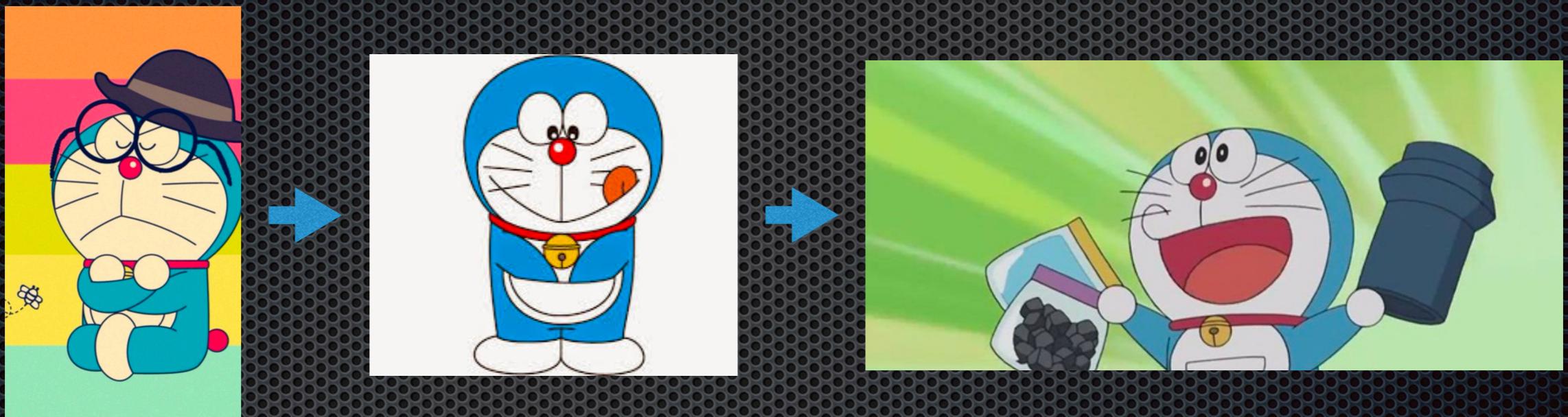
- Real world problem - playing Go
- What is the scientific problem?



“The most serious mistakes are not being made as a result of wrong answers. **The true dangerous thing is asking the wrong question.**”

*–Peter Drucker*

# A simplified data project cycle



Real world  
question/  
problem

What data/  
tools can  
help?

Problem  
solving

Value

data  
generating  
system

Extract

Understand

data

results  
data

Interpret

Engineer

data  
processin

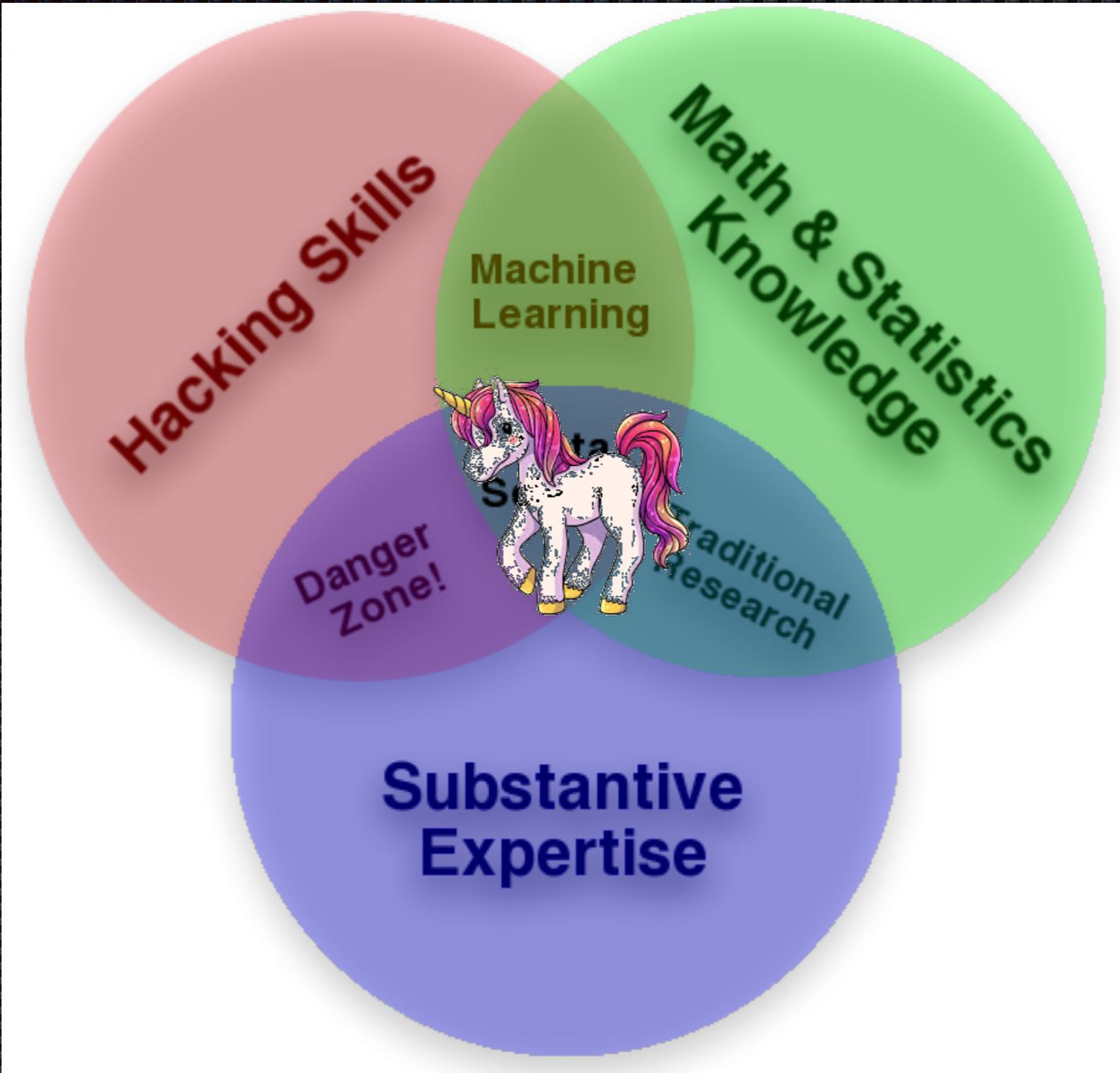
Design

data mining,  
exploratory data analysis,  
machine learning,  
statistical modeling

statistical modeling

# Foundations of data science

- Data engineering
- Software engineering
- Machine learning
- Statistics



# Data Science Skill Set

- How to **think** about data versus problem:
  - ▣ Mathematics/Statistics/Machine Learning
- How to **handle** data
  - ▣ **Technologies: Python, Java, Hadoop, Spark, etc**
- Teamwork and collaboration skills - how to **work** with others.
- How to turn data into business intelligence:  
find **value** in your data
  - ▣ Innovation, intellectual curiosity
  - ▣ Problem-solving skills
- How to convince others about your data science results
  - ▣ Visualization, story telling
- **Communication** skills

# How to fail as a data scientist\*

- Focusing only on the solution
- Forgetting the basics
- Ineffectively communicating
- \* From <https://www.techrepublic.com/article/how-to-fail-as-a-data-scientist-3-common-mistakes/>

# How this course can help

- No formal instruction on statistics/machine learning topics.
- Not intended to be a comprehensive data science bootcamp.
- Project-based course. Learning by doing.
- Project-based learning
  - Problem identification via teamwork and discussion.
  - Problem solving by using existing skills or new skills, learn new things “on the job”, and learn from your peers.
  - Present your codes, your results and your story (try to sell them).
  - There will be things I cannot answer but let’s learn together.

Stay Hungry. Stay Foolish.

*-Steve Jobs*

# Project-Based Learning

# **Project-Based Learning**

## **Integrating 21<sup>st</sup> Century Skills**



# Learning Objectives

- Become self-directed learners
- Develop problem-solving skills
- Teamwork skills: collaboration, reasoning and communication
- Self-assessment skills
- Presentation and critique skills
- “Initial stimulus” and experience for more fun in data science
- Try to become the master of your toolkit

# Student-centered Approach

- I am not to lecture here but to facilitate active learning.
- I will design open-ended challenges, each of which focuses on a slightly different area in data science.
- In each challenge,
  - Start with information/knowledge we already have (maybe not you but your teammate) about the problem.
  - Identify knowledge/skills we need to solve the problem.
  - Articulate the above thinking process in a team and implement an inquiry as a team
- I will provide case studies and tutorials to provide guidance on aspects of the above processes.

# Communicate!



Communication is everything

# Channels of Communication

- During class time
  - Brainstorm
  - Ask questions during tutorial
- Before and after classes
- On Piazza (*show piazza*)
- If you have questions
  - Online Q&A (live or not)
  - Email

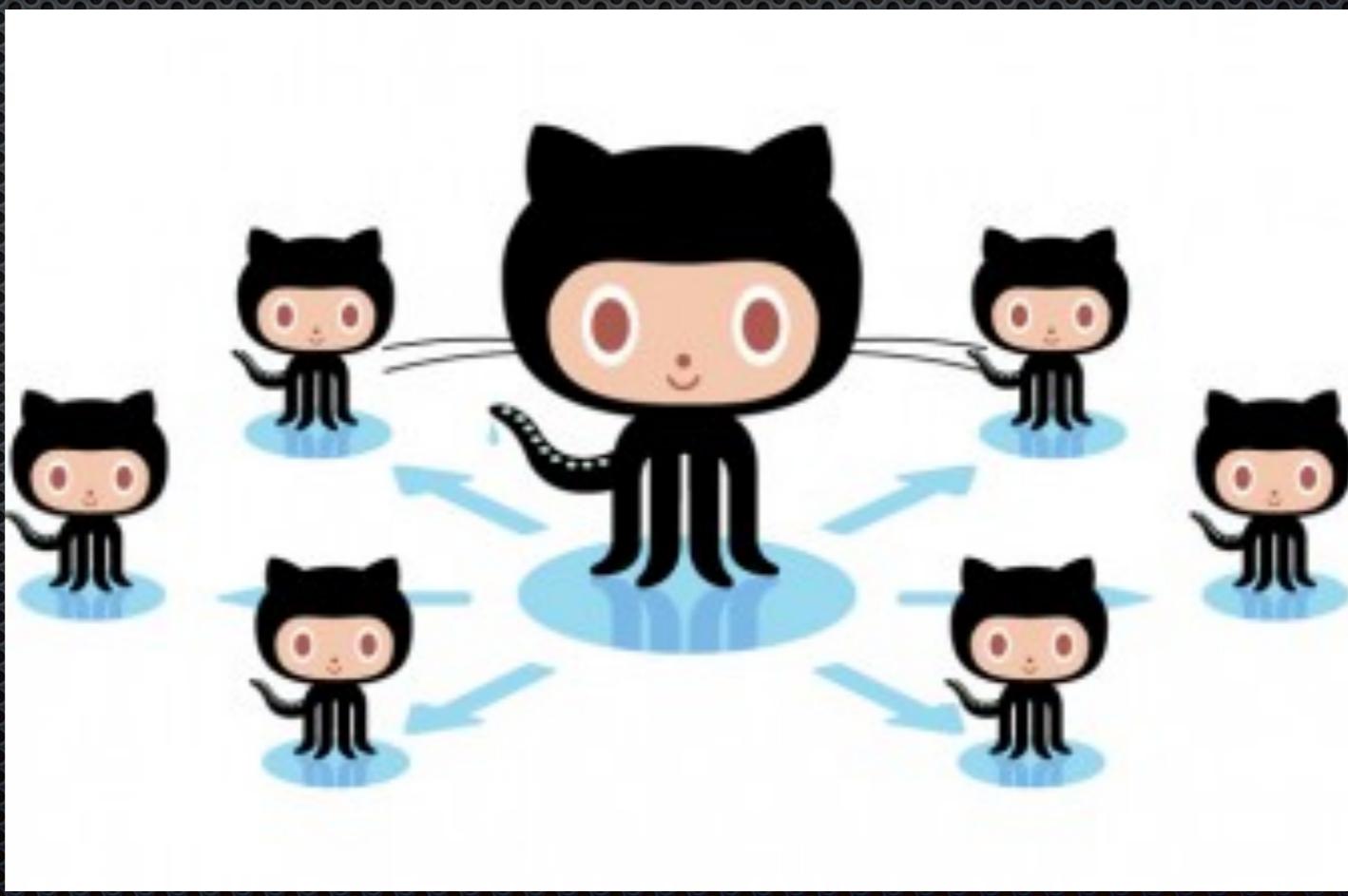
# Group Projects

# Working Together

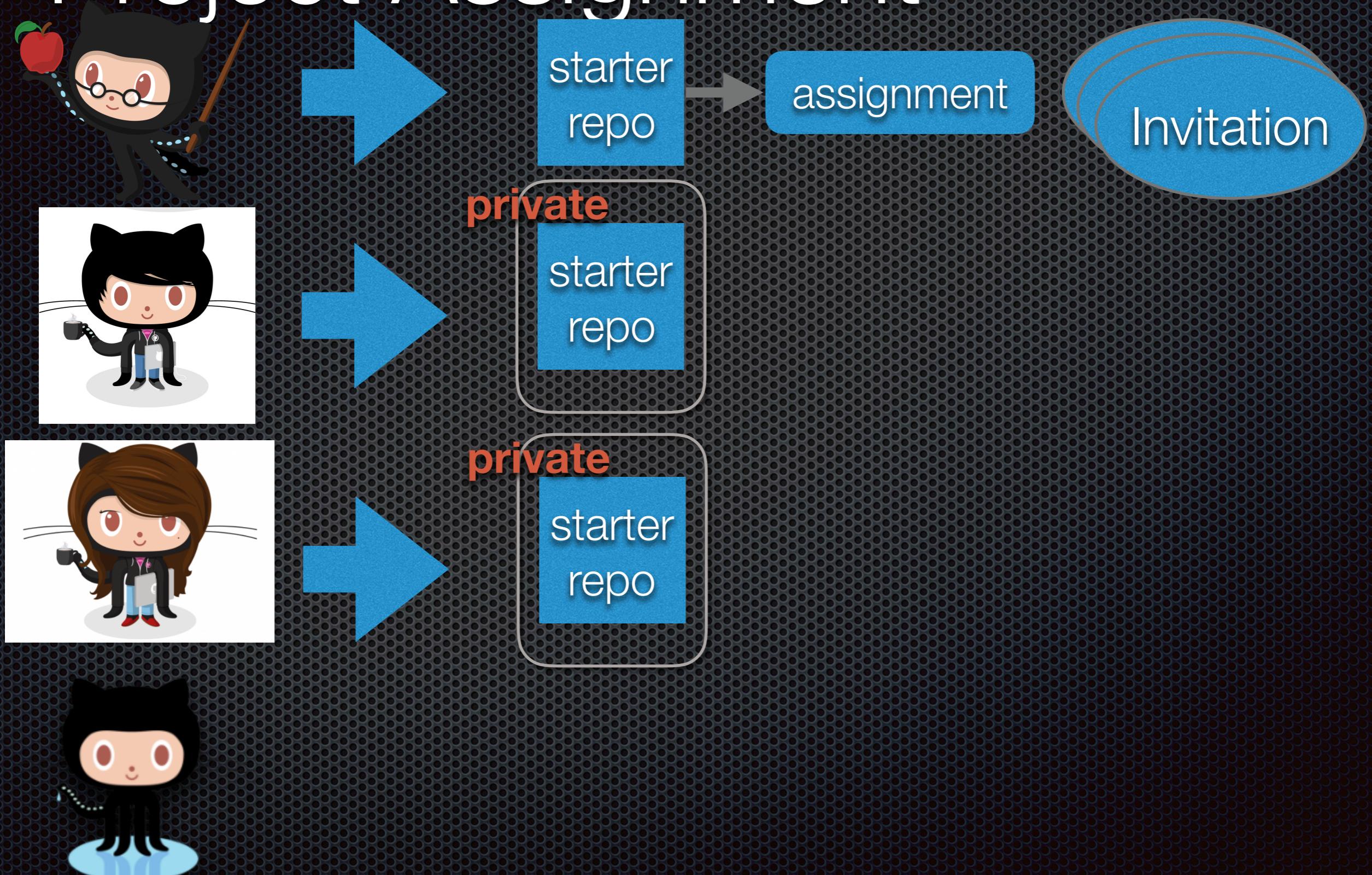
- You don't have to be in the same room at the same time to work together.
- Here are several ways you will work together in this course
  - Face-to-face brainstorm
  - Online discussion in group forum
  - Online video chat (say, via Google Hangout) with screen share.
  - **GitHub collaboration**
- Learning is not a zero-sum game.

# Learning on GitHub

- This semester we will use Classroom for GitHub
- It allows the instructor to create parallel private repositories for groups to collaborate.



# Project Assignment



# Project Assignment

- Teacher creates starter code folder
- Teacher creates groups with group numbers (off GitHub)
- Teacher shares the group info with students (especially group number) on Piazza
- Teacher create assignments (private) and set the option for “new set of groups”
- Send invitation link to students with instruction
  - First, check whether your teammate already created a team for your group from the “Join an existing group”.
  - If you cannot find your group’s name (as assigned in the Excel name), please create the team using precisely the name specified in the Excel file.
- The Project name and membership can be managed later but the most important part is we get all the teams/groups set up automatically.
- Everyone from your team should install Git, GitHub Desktop and use Git with Rstudio.

# Applied Data Science

## Tutorial 1: reproducible data analysis

# Improve Reproducibility

- Setup project folder
- Documentation
- Project history and source control

# Project Setup

- Rstudio really makes it easy to keep track of a project.
  - First, identify a working folder.
  - Inside the working folder, create the following subfolders.
    - data: data used in the analysis. Read only
    - doc: the report or presentation files
    - figs: contains the figures. only contains generated files. Images used for report should be put in a separate image folder under doc.
    - lib: various files with function definitions and code.
    - output: analysis output, processed datasets, logs, or other processed things. only contains generated files.

# Use knitr for reproducible data analysis

- knitr is an R package that processes R markdown files.
- An R markdown file follows the markdown syntax and contains R code blocks.
- An R markdown file can be “knitted” into either a html page or PDF document that reproduces a data analysis.
- It shows both the code *chunks* and the results produced.
- One can also include seamlessly project discussion, method section (with LaTeX support) and results discussion.
- It should be viewed as a data analysis documentation, rather than a report though, as the analysis needs to presented in a chronological order.

# DPLYR

- Data manipulation using five key verbs
  - filter
  - select
  - mutate
  - arrange
  - summarise
- along with "by group" adverb.

Now lets  
Look at Project 1