

So you want to become a Data Scientist?

Lecture 1

Sep 5, 2018

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 simplified data project cycle



Real world
question/
problem



What data/
tools can
help?



Problem
solving

A real example - search evaluation

Search results from “data science textbooks”:

The image shows two identical search result pages for "data science textbooks". The results are organized into three columns: "Popular on the web", "Shop for data science... on Google", and "Data Science Certificates - Or Master's Degree?".

Popular on the web:

- Deep learning
- Data Science for Business – Tom Fawcett
- Practical Data Science with R – Nina Zumel
- The Art of R Programming – Norman Matloff
- Storytelling With Data – Nathan Yau
- Predictive Analytics – Eric Siegel, Jeff Elton
- Bayesian Reasoning and Machine Learning – David Barber
- Pattern Recognition and Machine Learning – Christopher Bishop
- Advanced R – Hadley Wickham
- Building Data Science Teams – Eli Pelizzetti, Jim Hall

Shop for data science... on Google:

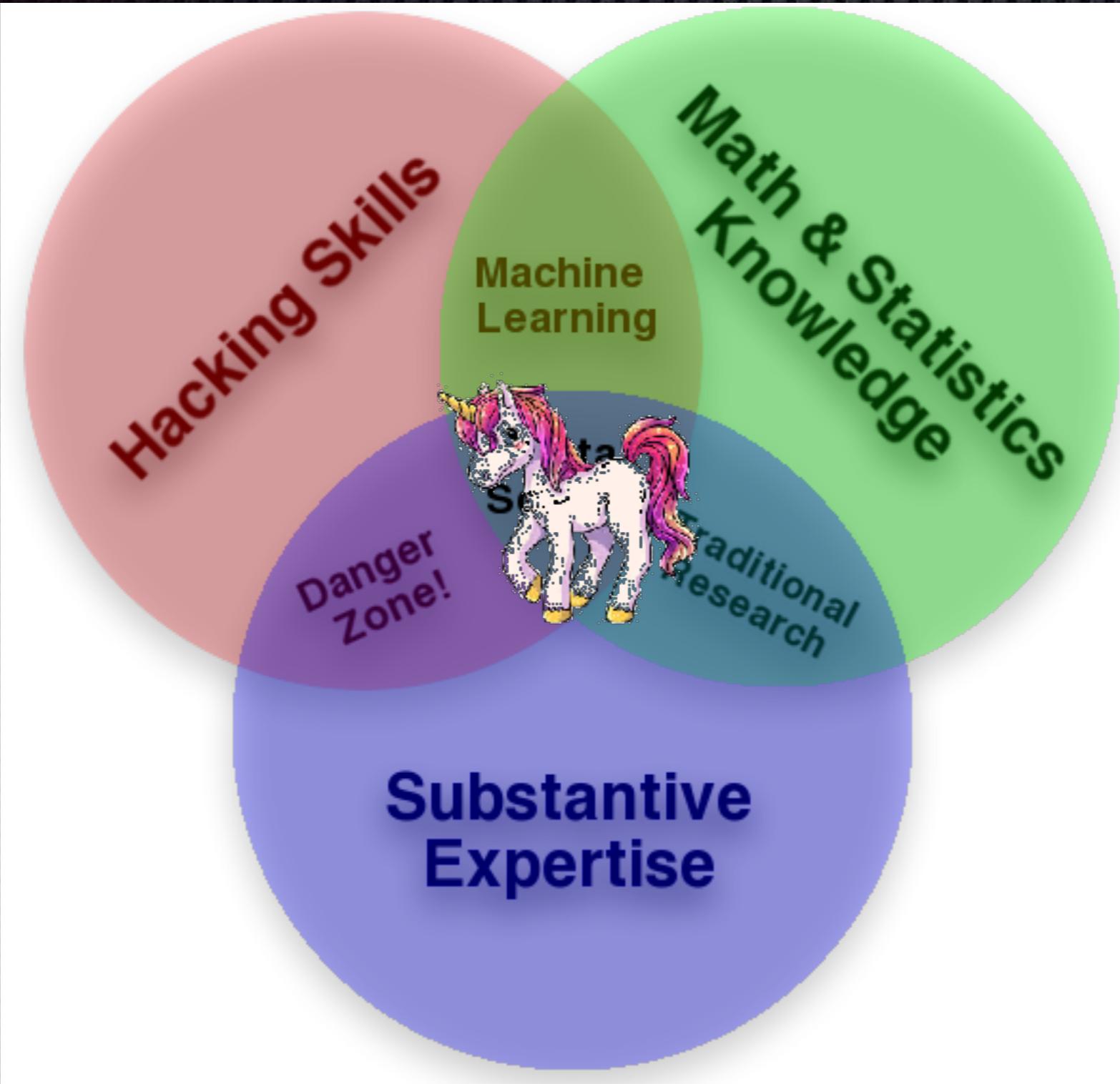
- Algorithms for Data Science – \$68.76 Barnes & Noble Free shipping
- Beginning Data Science with R – \$99.99 Google Express Free shipping
- Data Science and Big Data... – \$28.82 Barnes & Noble 21% price drop
- Data Science from Scratch – \$14.95 Textbooks.com
- Introduction to Data Science... – \$39.99 Prentice Hall
- Web and Network Data Science... – \$79.95 Barnes & Noble

Data Science Certificates - Or Master's Degree?

- Data Science Certificates - Or Master's Degree? - Saint Mary's College
- Graduate Programs
- More the right choice for your career. Apply to Saint Mary's College today!
- A Data Give In. Better Advance Your Data Career. Advanced Data Analysis - Accounting Audit/Review Program Overview • Request Information • From Our Experts • Admissions
- Become A Data Scientist - 12 Week Data Science Courses
- www.gutenberg.org
- 4.5 ★★★★☆ rating for generalists
- Learn Python 3.6. Unix & More. Join Our Tech Community Today.
- Bring Ideas To Life. Advance Your Career. Work At A Tech Startup. Learn Cutting-Edge Skills.
- Complete EGL, Spark, Machine Learning, Python, Big Data, AI/ML, Testing, Predictive Statistics, Data Mining, Data Science Career. Python Programming. Graduate Team Training. French Edition
- Data Scientist Masters Program - 124 industry-based projects
- www.apimil.com/Data_Scientist/certification
- Membership from industry experts, industry-recommended learning paths. Start Now!
- Learn without Fixed-Pass / High-Pass/Race. Instructor and Training.
- Big Data & Hadoop Training. Data Science Masters. Python. Data Science with R. Python
- Data science textbook - Amazon Official Site - amazon.com
- www.amazon.com/books/computers
- Business & Finance. Resources of Companies & Internet Book Titles, for Less.

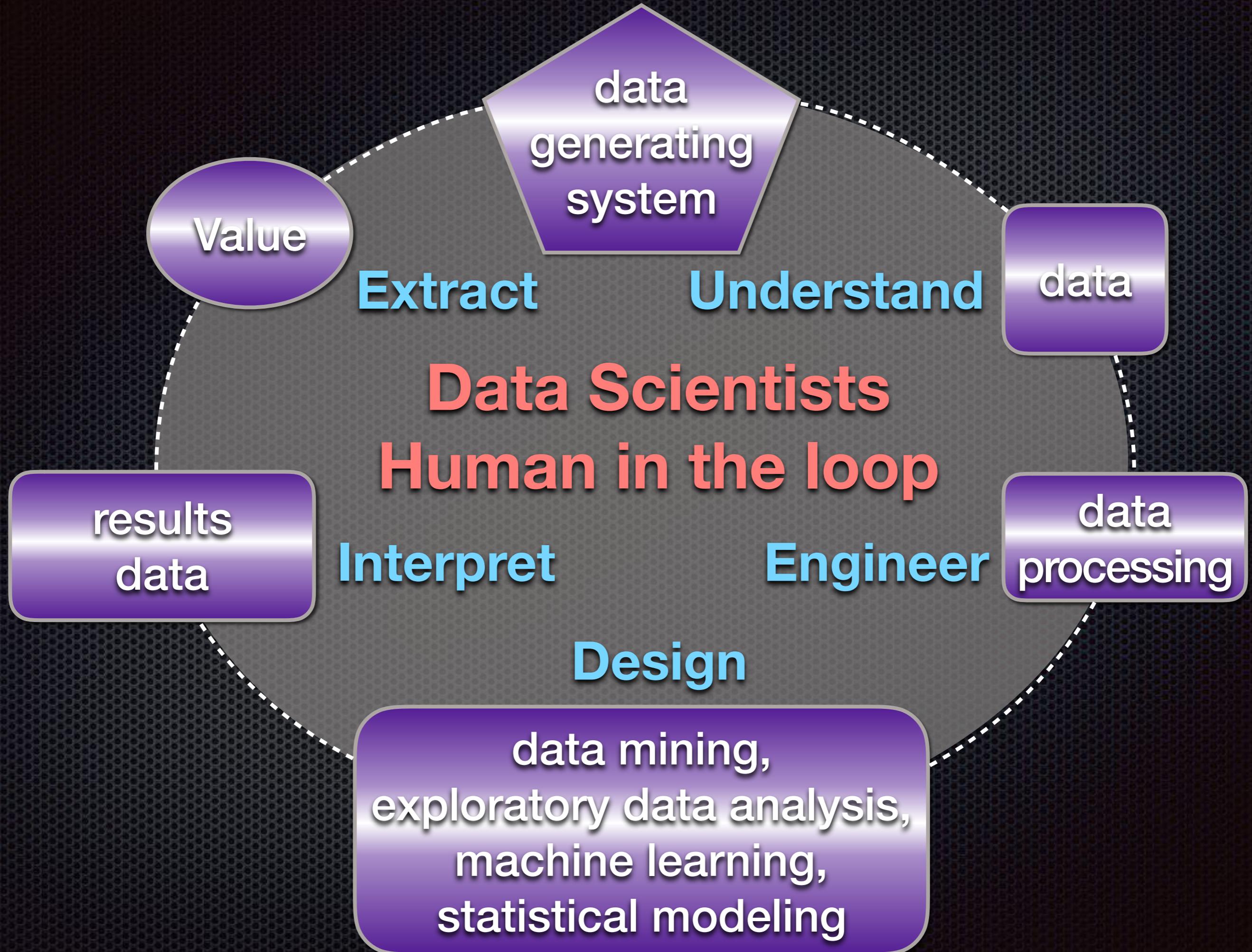
Foundations of data science

- 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 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 21st 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

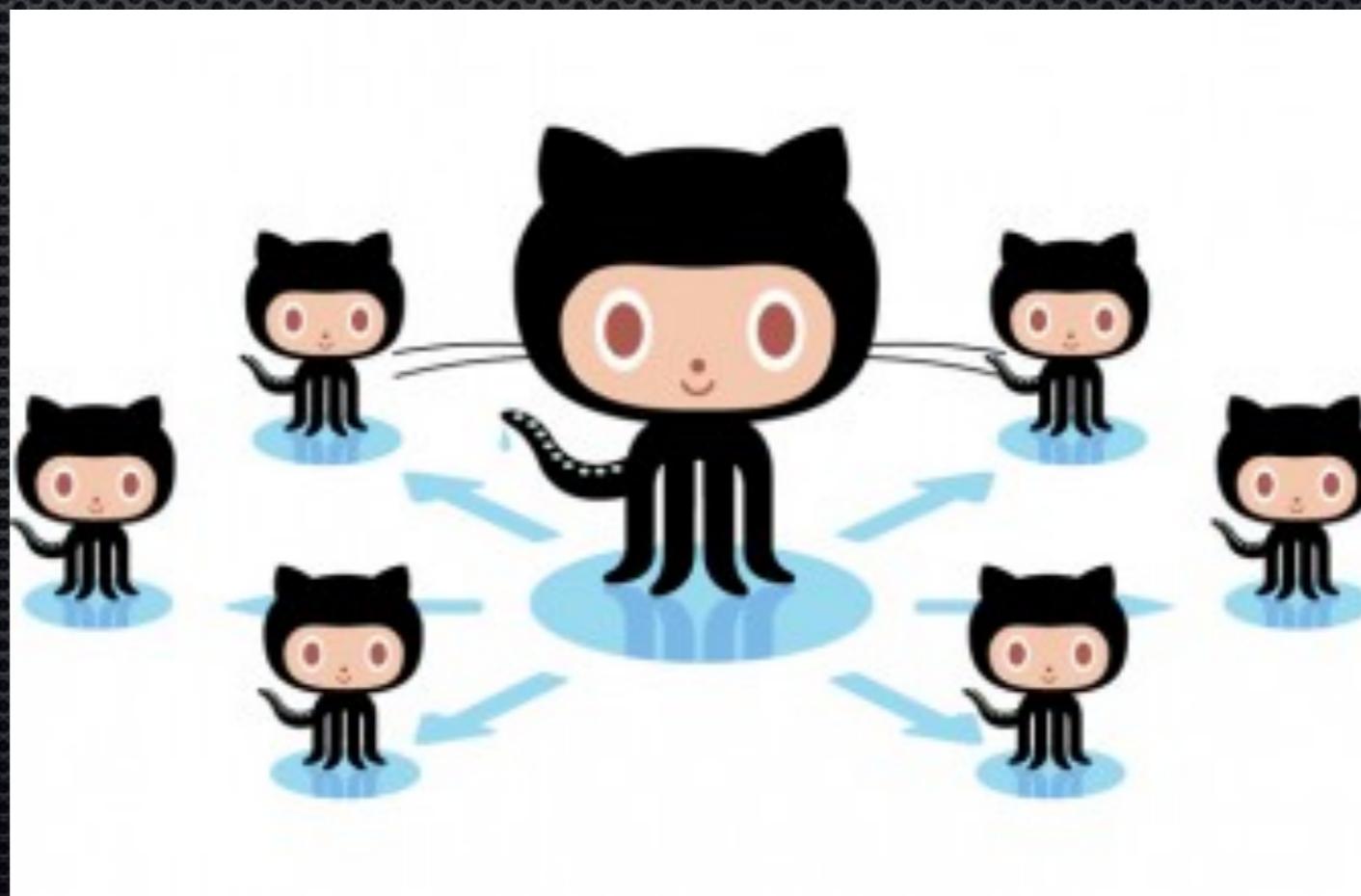
Group Projects

Working Together

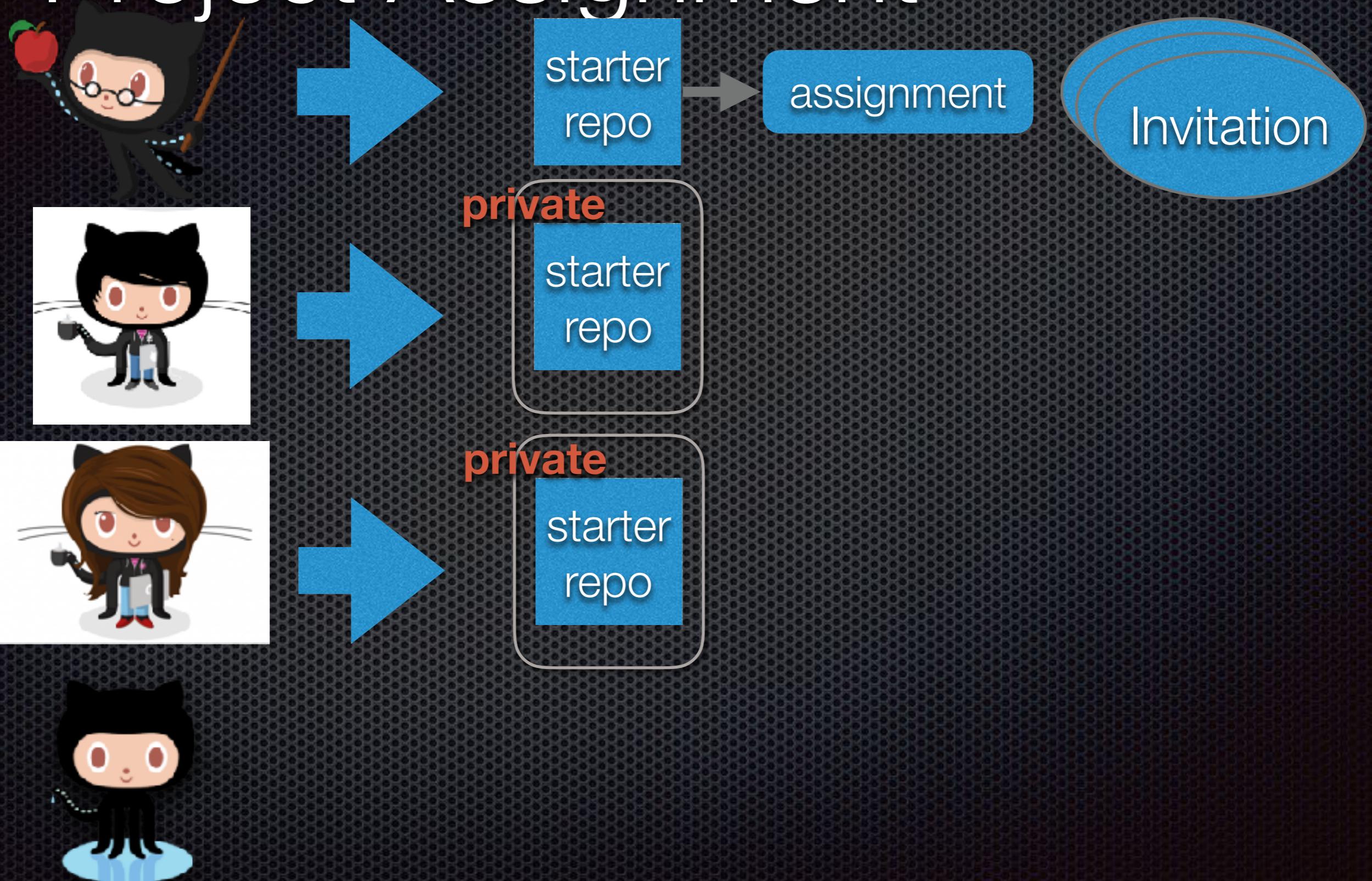
- You don't have to be in the same room at the same time to work together.
- Here are several ways you will work together in this course
 - Face-to-face brainstorm
 - Online discussion in group forum
 - Online video chat (say, via Google Hangout) with screen share.
 - **GitHub collaboration**
- Learning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

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