

Data 601 @ UMBC

Introduction

August 27, 2020

These notes are CC BY 4.0

If it were a real license, I'd use <http://matt.might.net/articles/crap/>

Outcomes for this evening

- Get to know your instructor and your classmates.
- Understand the course assignments and grading distributions.
- Understand the course rules/expectations and your rights.
- Understand course schedule for weeks.
- Learn the necessary tools for DATA 601.
- Understand what is data science.

Same and Different

Rules

1. Will be assigned randomly to groups of 3 people.
2. With your group choose 3 things you all have in common (or completely different). (5 mins)
3. Later on you will share these with everyone.
4. Be creative!

Let me know you more!

- Please go to Blackboard > Assignments
- Take the survey: “Data Science Toolkit”



Course Logistics

- Course Logistics
- What is Data Science?
- Software tools
- What are we not covering?
- Soft skills
- Homework

How can we make this environment optimal for learning?



Share

Ground Rules

- Be Professional
- Be Kind
- Be Constructive
- Be Active
- Be Inclusive
- Fair
- Last but the most important one:) Can you guess?



icon source: <http://www.chicagonow.com/quilting-sewing-creating/files/2014/09/comments-icon.jpg>

Logistics

- Schedule: 4:30 – 5:30, Break, 5:45 – 6:45
- I value being punctual (start of class, breaks, end of class)
- Raise your hand if you have a question/answer
- Don't apologize for asking a question or for not knowing something
- I find it acceptable for you to occasionally not participate
- Tell me if you cannot hear me or if you cannot understand me
- Slides will be provided after lecture

Learning Objectives

- What do I want you to leave the course with at the end of the semester?
 - Solid understanding of what data science is and what it isn't
 - Why data science is super cool.
 - Why it is not that cool.
 - Life cycle of a data science project.
 - Who are the partners of data scientist in a work environment?
 - Data visualization tools in Python
 - Data wrangling tools in Python
 - What machine learning is and what it isn't and it's place in data science.
 - Commonly used tools in data science outside of Python world.

Grading

Basics:

- Attendance/Engagement: 15%
- Homeworks: 30%
- Midterm Project 30%
- Final Project 30%



Grading Cont.

Attendance/Engagement:

- What counts as being present:
 - You showed up at most 15 min after class started. And you are late no more than 3 times throughout the semester.
 - You finished the readings, take the in class quizzes successfully and engaged with the material.

Homework

I am planning to give 3 mini Projects:

Due dates and tentative topics:

- Week-4: EDA with Python (matplotlib, seaborn)
- Week-7: Data acquisition and wrangling (pandas, api)
- Week-12: Unsupervised techniques

Tentative deliverables:

- Github repo: Code, ReadMe
- Blogpost

Projects

- Think of them as bigger homework.
- You can build your projects on previous homework

Tentative Schedule:

- Project-I: Week-9: Supervised Learning Project
- Project-II: Capstone

Deliverables:

- Github: Code, ReadMe, Presentation, Technical Notebook
- Blogpost

What do you want to learn in this class?



Activity: verbal popcorn; record answers on board



What is Data Science?

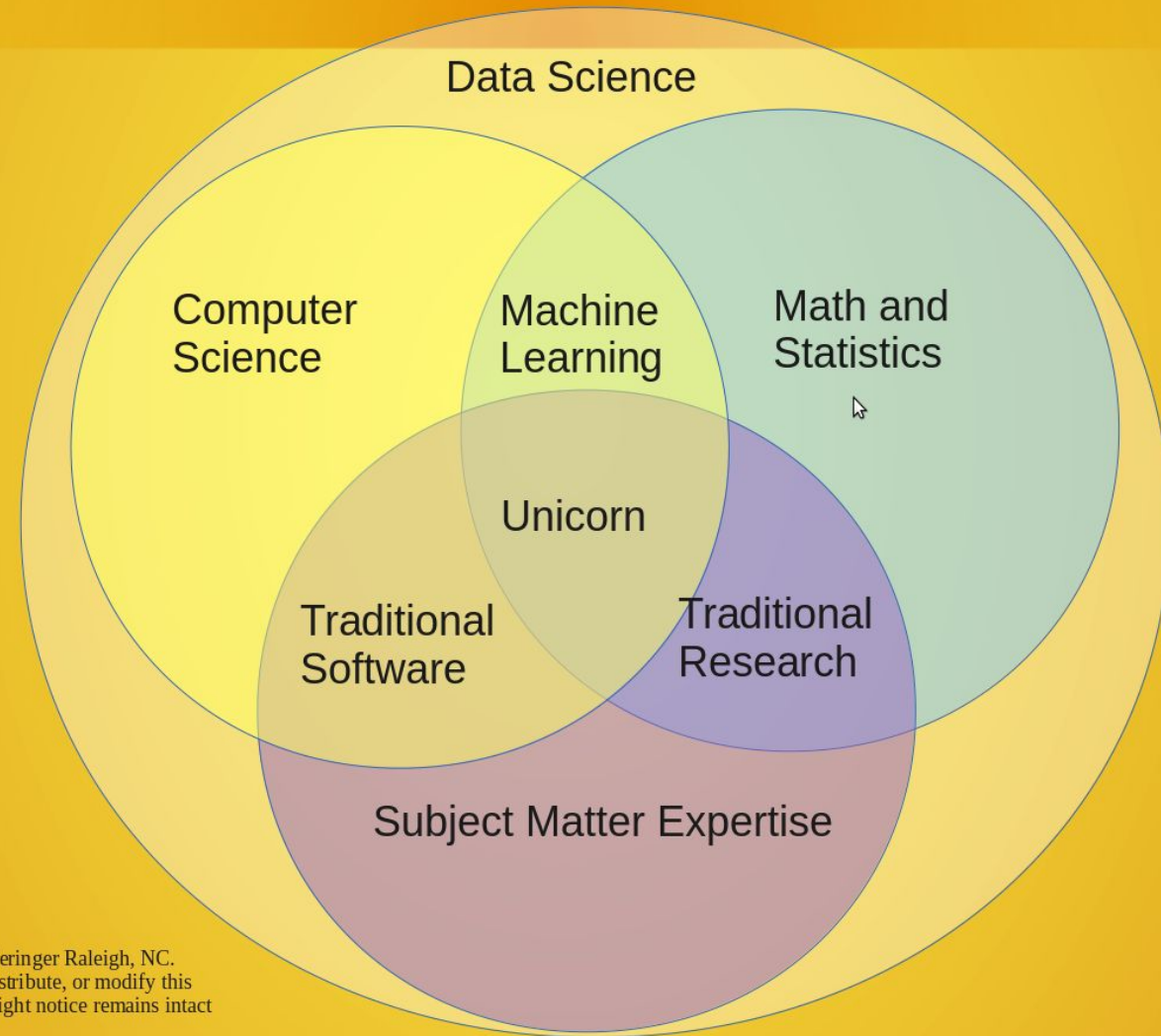
- ~~Course Logistics~~
- What is Data Science?
- Software tools
- What are we not covering?
- Soft skills
- Homework

What do you think data science is?

Please take the survey and share your thoughts!

Data Science Venn Diagram v2.0

There's a lot to cover



Copyright © 2014 by Steven Geringer Raleigh, NC.
Permission is granted to use, distribute, or modify this
image, provided that this copyright notice remains intact

Suggested reading:

<https://www.kdnuggets.com/2016/10/battle-data-science-venn-diagrams.htm>

!

What is Data Science?

Data science is an interdisciplinary field that uses

- scientific methods,
- processes,
- algorithms and
- systems

to **extract knowledge and insights** from many **structural and unstructured data**.

The Six Divisions of Data Science

- Data Exploration
- Representation and transformation of data
- Computing with data
- Data modeling
- Visualization and presentation
- Science about DS

Roles in Data Science Projects

DATA SCIENCE – A TEAM EFFORT				
	 Data Engineers	 Data Scientists	 Software Engineers	 Data Storyteller/Translators
What They Do	<ul style="list-style-type: none"> • Create Data pipelines. • Evaluate Databases • Design Schemas • Perform ETL 	<ul style="list-style-type: none"> • Apply statistical/Machine learning techniques to solve business problems • Perform R&D • Innovate new solutions • Develop Data science products 	<ul style="list-style-type: none"> • Help design UI (front end coding) • Do backend coding • Help deploy data science solution in production • Automate the entire process 	<ul style="list-style-type: none"> • Communicate Data Science solutions in Business friendly/ non technical terms • Understand business requirements and translate them to Data science problems • Design persuasive Data visualizations
Skill Set	<ul style="list-style-type: none"> • Knowledge of Databases • Scripting skills (Linux commands) • Knowledge of Cloud technologies • SQL commands 	<ul style="list-style-type: none"> • Knowledge of statistical and mathematical concepts • Knowledge of various statistical/ML algorithms • Scripting skills (R/Python) • SQL commands 	<ul style="list-style-type: none"> • Knowledge of Programming concepts • Programming languages • Knowledge of Databases • Knowledge of Restful APIs • Scripting skills (Linux commands) 	<ul style="list-style-type: none"> • High level understanding of statistics and ML concepts • Business acumen • Good soft skills • Creativity • Persuasion and articulation
Tools Used	   	   	   	   

Assumption in this class

- In class we will assume you are a lone data scientist on an island with an internet connection.
- This is not the typical case -- you'll have coworkers, customers, bosses, competitors, collaborators, peers.

Example of how class \neq real world

- This class will not use competitive grading. (Imagine if it were.)
 - As an employee at a company, you may be competing for a bonus or promotion
- > consequence: personal and organizational politics factor into the work environment

Data science is an active field with lots of jargon

There will always be something you haven't heard of before.

- Know enough to be conversant with peers
- Be curious about new topics
- Research concepts and labels before using them

Reference: <http://www.datascienceglossary.org/>

Why learn data science?

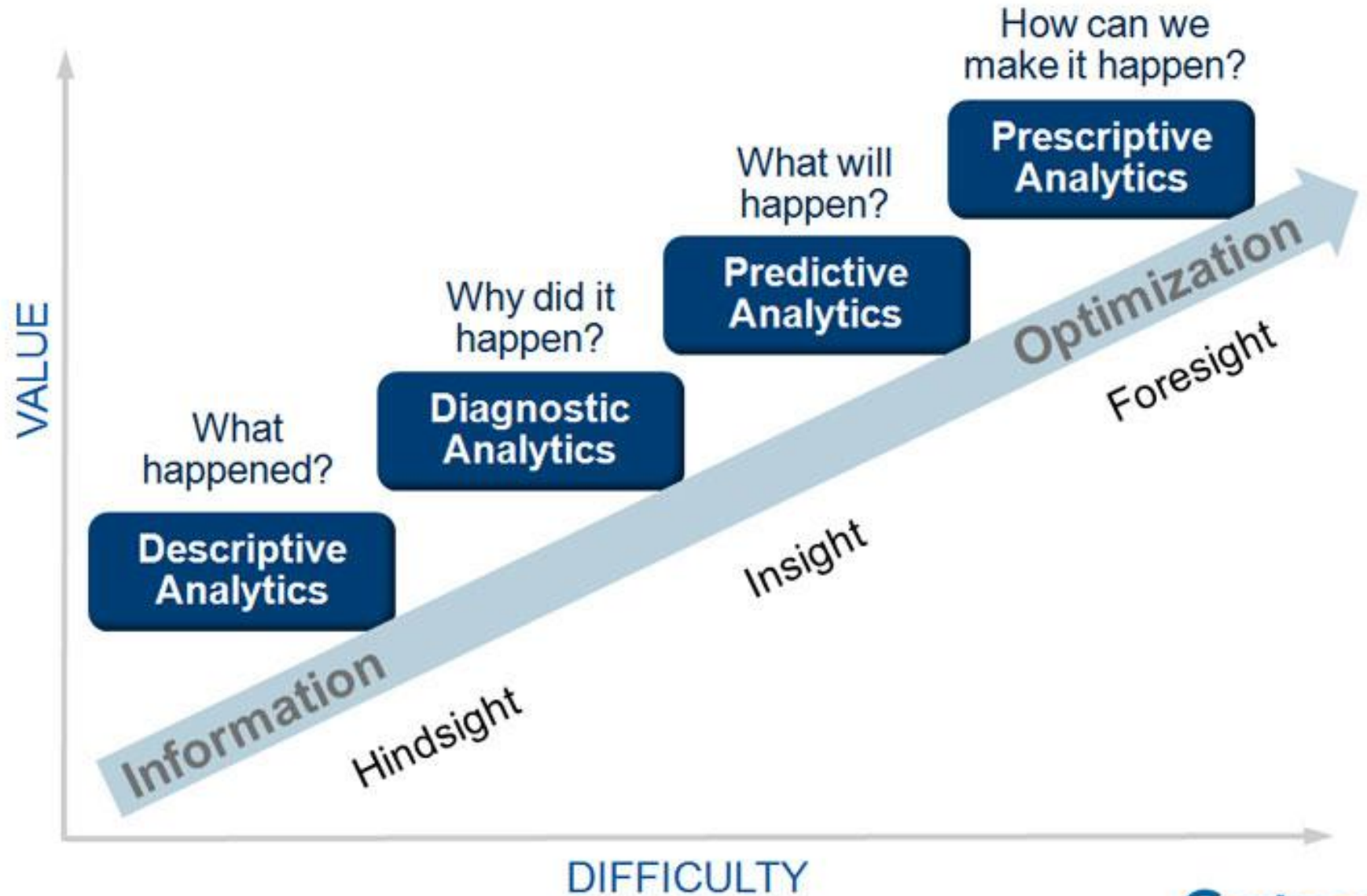
Explore: **identify patterns**

Predict: **make informed guesses**

Infer: **quantify what you know**

Motives:

- Make money
 - Employment
 - Promotion
- Help people
- Gain new knowledge



Large scale use cases with lots of data

- Google's search engine
- Recommendations from Amazon and Netflix
- Bank and Credit Card fraud detection
- Logistics (DHL, UPS) of fleet management
- Healthcare records from patients

Each depends on availability of compute and data

Small scale use cases with not much data

As a business employee or bureaucrat or politician

- How do I improve decision making process?
- How do I evaluate the outcome of decisions?
- How do I decrease the risk when faced with an opportunity?
- How do I convince other stakeholders of the best course of action?

While not taking too much time, spending too much money, using the resources I already have access to, and in a way that is convincing?

Software Tools

- ~~Course Logistics~~
- ~~What is Data Science?~~
- Software tools
- What are we not covering?
- Soft skills
- Homework

Tools of the Trade

- Terminal (Command Line Interface)
- Python & Jupyter Notebook
 - Anaconda
- Git & GitHub

But before we go into details, we have a very quick poll about the operating systems you are using.

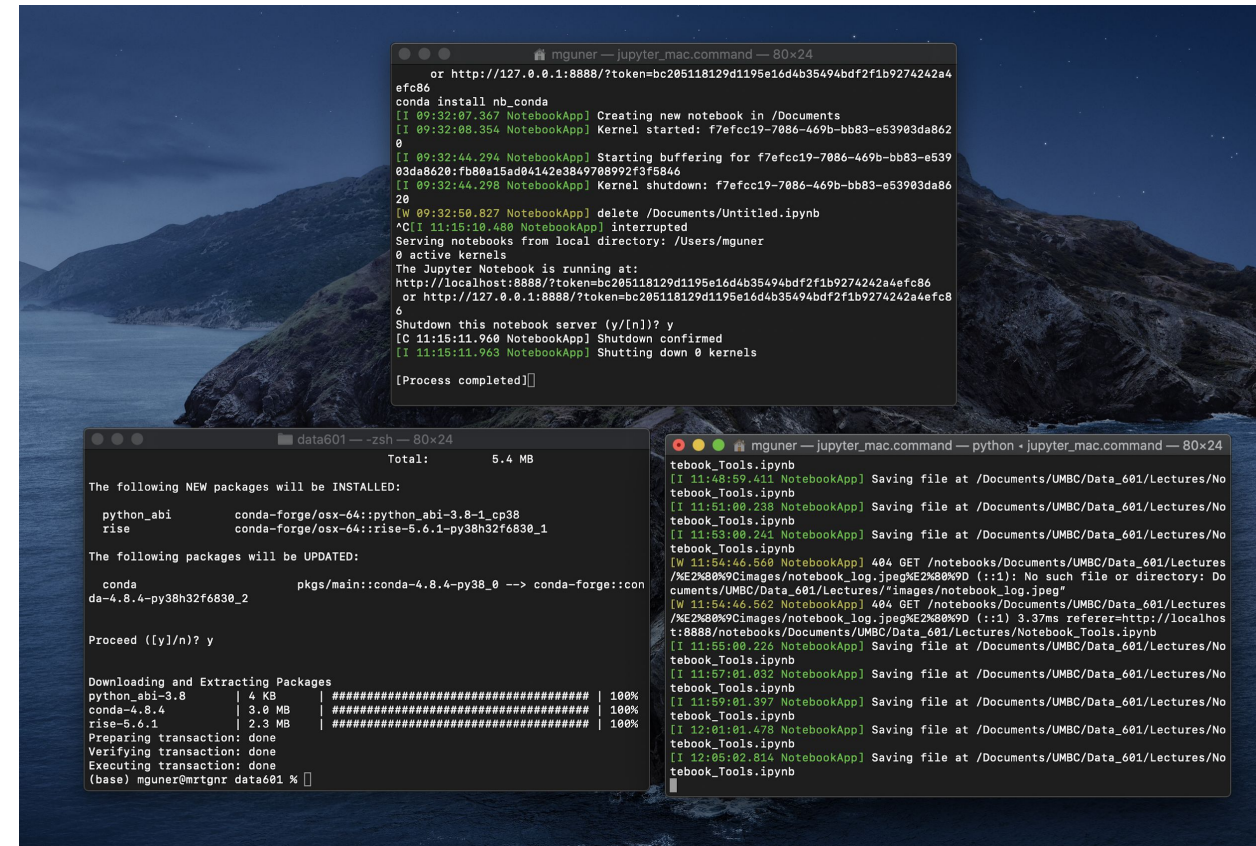
Anaconda installation

- What is Anaconda?
- Why do we need it?
- Now let's install it!
 - <https://www.anaconda.com/products/individual>
- This will take a while. In the meantime we can discuss terminal.

Terminal (Command Line Interface) Basics

- What is Command Line Interface (CLI)?
- Why do we need it?
- Basic commands

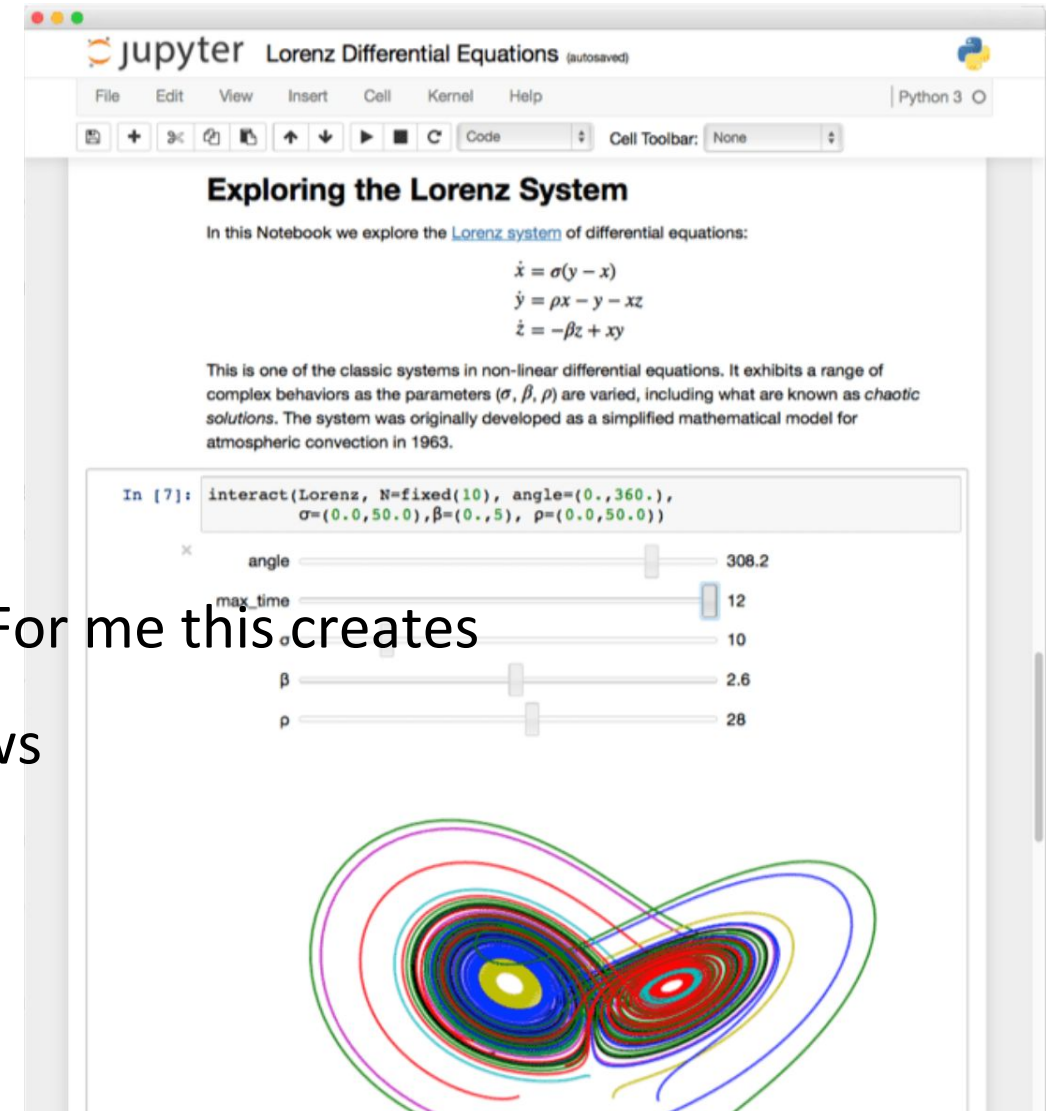
- pwd: print working directory
- cd : current directory
- ls: list files
- mkdir: make directory
- touch: creates files
- cat: show the content of a file
- echo: prints text



The image displays three terminal windows on a macOS desktop background. The top window, titled 'mguner — jupyter_mac.command — 80x24', shows the installation of Jupyter Notebook using 'conda install nb_conda'. It includes logs for creating a new notebook, starting the kernel, and shutting down the server. The bottom-left window, titled 'data601 — zsh — 80x24', shows the installation of 'python_abi' and 'rise' packages using 'conda-forge/osx-64:python_abi-3.8-1_cp38' and 'conda-forge/osx-64:rise-5.6.1-py38h32f6830_1'. It also shows the download and extraction of these packages. The bottom-right window, titled 'mguner — jupyter_mac.command — python · jupyter_mac.command — 80x24', shows the saving of a file named 'tebook_Tools.ipynb' to the directory '/Documents/UMBC/Data_601/Lectures/No'.

Jupyter Notebooks

- Code mode
- Markdown mode
- Extensions:
 - TOC, spellchecker, autopep8, hinterland? (For me this creates problems sometimes) and RISE if time allows
- Shortcuts



Git & Github

- What is Git?
- Why do we need it?
- Installation: Git is already installed for MacOS computers but you might want to update it.
 - check from terminal: `git --version`
- Create and account at Github and create first repo.

What are we not covering?

- ~~Course Logistics~~
- ~~What is Data Science?~~
- ~~Software tools~~
- What are we not covering?
- Soft skills
- Homework

Processes external to data science

Exploration may be enough and the effort terminates



Processes external to data science

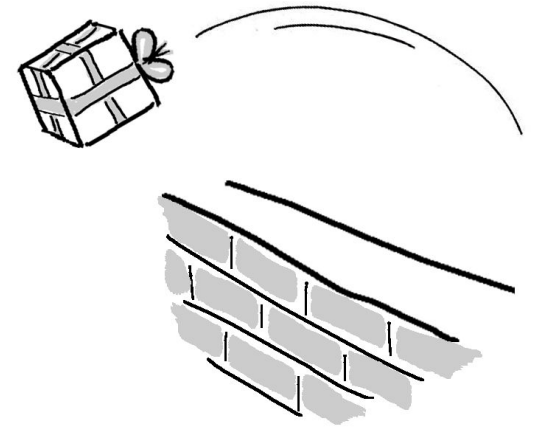
~~Exploration may be enough and the effort terminates~~



Additional refinement is often needed;
data science is often merely the start of
an investment

Not covered: product integration

- There's a complex network of dependencies (i.e. software engineers, managers) of which data science is one component.
- Downstream consumers of your output are likely to be software developers who use containers and support users.
- This class is focused on the data science; not with integration.



See <http://dev2ops.org/2010/02/what-is-devops/>



Not covered: security

We focus on data science techniques; these do not emphasize secure design of software.



Soft skills

- ~~Course Logistics~~
- ~~What is Data Science?~~
- ~~Software tools~~
- ~~What are we not covering?~~
- ~~Data Formats~~
- **Soft skills**
- Homework

Data Science is more than Math and Software

Human interaction in data science

- Discovering stakeholders
- Negotiating with data owners
- Customer engagement

<https://hbr.org/2017/01/the-best-data-scientists-get-out-and-talk-to-people>

Iterating with customers

- As a data scientist, you'll often be working for someone other than yourself.
- Expect under-specified requirements from customers. Iterate.
- Provide incomplete solutions rather than waiting until the product is perfect.

https://en.wikipedia.org/wiki/Minimum_viable_product

When to persist,
When to change course,
When to seek help



Try attacking the challenge for 30 minutes
Then seek help or do something else for a while

https://en.wikipedia.org/wiki/Pomodoro_Technique

Pro-tip when seeking help

How to ask well-formed questions:

<https://stackoverflow.com/help/how-to-ask>

[Intentional sidetrack to StackOverflow.]

Ask technical questions:

- *Poor*: "I don't understand Python dictionaries" (--> online tutorials)
- *Better*: "When is it appropriate to use a key-value pair?"
- *Poor*: If I submitted this assignment as is, what score would I get?
- *Better*: I am planning to submit the attached assignment, but currently there's an error in the third cell. I've searched online but don't find any references to the error message. Can you provide guidance?



Emotions in Data Science

- As a data scientist, most of your time will be spent in a desert of uncertainty, frustration, and doubt.
- There will be rare short-lived interspersed spikes of excitement and happiness due to events like getting a new dataset, creating a new analytic, getting a new result, or being thanked by a stakeholder.

This experience is normal and does not go away.

See also the psychology of slot machines

Online resources (in addition to books)

- Slack channel for Data 601: <https://umbcdatasci.slack.com/messages/>
- Meetups
 - <https://www.meetup.com/topics/data-science/>
 - <https://www.meetup.com/BigDataBaltimore/>
 - <https://www.meetup.com/Statistical-Seminars-DC/events/254200651/>
- News and blogs
 - <https://www.kdnuggets.com/>
 - <https://news.ycombinator.com/>
 - <https://hackernoon.com/>
 - <https://www.reddit.com/r/datascience/>
 - <https://dataelixir.com/newsletters/>
- Online courses
 - Coursera
 - <https://www.coursera.org/learn/machine-learning/home/welcome>