GENERAL ASSEMBLY

DATA SCIENCE CLASS 1: INTRODUCTION AND TOOLS

WELCOME!

- Introduction
- Data Science
- Course Overview
- Technical Review
  - Unix Commands
  - Git
  - Ipython
- Data Exploration
- Visualizations

INTROS

▸INTROS

Someone else in the room has the same number as you do. Find them!
Only rule is each of you needs to ask each other one question before comparing numbers:

- If you had to give up a favorite food, which would be the hardest?
- Favorite color, and why
- What do you do if someone says “I could care less”
- Strangest thing you’ve eaten, and when
- When someone says “Good question”, what do you think it means?
- Longest you’ve gone without sleeping?
- Anything else!

Who are you?

- What's your name?
- Why are you taking this course?
- Which company, in your opinion, has the coolest data?

I. WHAT IS DATA SCIENCE?

Wikipedia: "*Data science is the interdisciplinary field about process and systems to extract knowledge or insights from data in various forms, either structured or unstructured, which is a continuation of some of the data analysis fields such as statistics, data mining, and predictive analysis*"

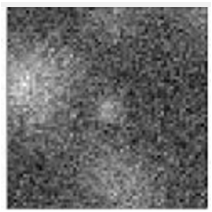
- A set of tools and techniques used to extract useful information from data.

- A set of tools and techniques used to extract useful information from data.
- An interdisciplinary, problem-solving oriented subject.

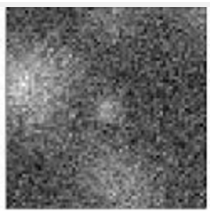
- A set of tools and techniques used to extract useful information from data.
- An interdisciplinary, problem-solving oriented subject.
- The application of scientific techniques to practical problems.

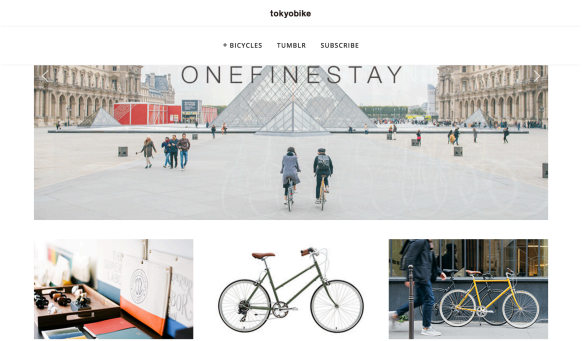
- A set of tools and techniques used to extract useful information from data.
- An interdisciplinary, problem-solving oriented subject.
- The application of scientific techniques to practical problems.
- A rapidly growing field.


NETFLIX
amazon
facebook
Google™
LinkedIn
FiveThirtyEight
Nate Silver's Political Calculus
foursquare
Etsy
2012
BARACKOBAMA.COM
Microsoft



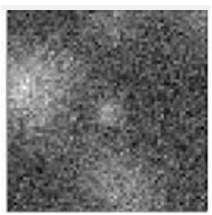
Cancer?

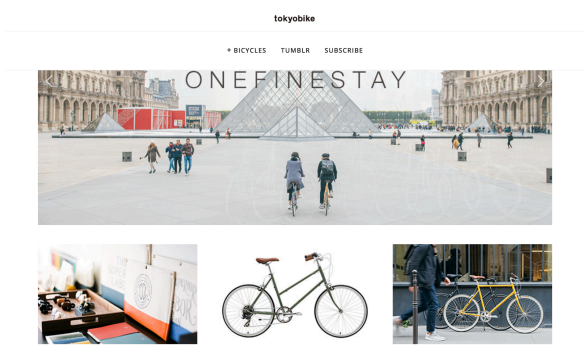




What kind of website?


FDA




tokyobike
+ BICYCLES TUMBLR SUBSCRIBE
ONEFINESTAY


venmo
Venmo 4:31 PM 81%
Nicole P paid Kristin W
Like Comment
Justice G paid Silas D
Breast implants
Like Comment
Richard B paid Kevin B
Like Comment
Corey S paid Edwin A
Thaiiiiii
Like Comment
Angela B paid Erin B
Monster sitting
Like Comment
Rikki M paid Jodi F
M
Like Comment
Cindy F paid Noa K

Fraudster?

WHAT ARE PEOPLE LOOKING FOR?

The Ideal Candidate Has a Deep-seated Passion For Problem Solving, Experience Working On Open-ended Projects, And a Proven Ability To Come Up With Creative, Elegant Solutions To Complex Issues. Experience With Specific Tools Is Less Important Than Aptitude And Drive, But At a Minimum We Would Expect

- proficiency with at least one general purpose programming language
- familiarity with relational databases and enough SQL skill to get by
- 1-2 years of academic or professional experience in a quantitative role

An Exceptional Candidate Might Have

- comfort working with Unix tools to streamline data retrieval, preparation, and distribution
- a track record of leading data analysis projects from start-to-finish
- knowledge in domains related to SeatGeek such as event ticketing, OTC markets, or online advertising
- experience using Hadoop / MapReduce or similar in production environment
- passion for data visualization as a tool for exploration and communication

The Tools We Use

We do research and development work in a custom environment optimized for repeatability and collaboration. You **absolutely do not** need experience with all of these, but we thought you might be curious.

- **Languages:** Python for web services and product devlopment, R for analysis and prototyping, Drake for repeatable analyses
- **Datastores:** MySQL, Redshift, Elasticsearch, Redis
- **Monitoring:** Graphite/StatsD
- **Version control:** Git

WHAT ARE PEOPLE LOOKING FOR? 18

Main Job Tasks and Responsibilities:

- Develop a deep familiarity with myriad data sources including operational databases, big data platforms, messaging frameworks, internal tools, and external integrations
- Identify and document the business logic required to clean, normalize, and model disparate source data into a conformed and dimensional representation of our business
- Understand the complex lineage of our data end-to-end; from source systems, to the data warehouse and data marts, all the way down to analytic products and constituent business metrics
- Take ownership of our data and constantly improve its quality and integrity; proactively identify data quality issues at every step and coordinate the implementation of required fixes and QA validations
- Partner with business stakeholders to understand their unique analytic requirements and assist in the development of metrics, reports, dashboards, visualizations and the design of schemas and data flows to enable them

Required Skills:

- 5+ years experience analyzing and manipulating complex, large-volume data from a wide-range of production systems and data platforms:
    - RDBMS (MySQL, Oracle, etc)
    - Columnar MPP Databases (AWS Redshift preferred)
    - NoSQL (Cassandra preferred)
    - REST APIs (GA, Salesforce, etc)
    - Bonus: Experience working with big data e.g. Hadoop/Hive/Spark
- Very strong SQL skills!
- Experience with a variety of data modeling techniques - emphasis on dimensional star schemas

WHAT ARE PEOPLE LOOKING FOR?

Required-skills

- Perform large-scale data analysis and develop effective statistical models for segmentation, classification, optimization, time series, etc.
- Design and implement reporting dashboards that track key business metrics and provide actionable insights
- Identify actionable insights, suggest recommendations and influence the direction of the business by effectively communicating results to cross functional groups
- Work closely with Product or Engineering & Operations teams to proactively create rule and manage decisions
- Prioritize leads so that the teams work on the most valuable cases
- Suggest improvements in the tools and techniques to help scale the team

Experience

- BS, MS or PhD degree in a quantitative discipline (applied mathematics, statistics, computer science, operations research, or related field) from a leading academic institution
- 3+ years of experience in solving analytical problems using quantitative approaches (or equivalent)
- Strong Programmer - Python, Perl, Java, and/or C++. Experience with relational database (SQL, PL*SQL) is a plus
- A passion for problem-solving, comfort with ambiguity, and creativity
- Demonstrated success presenting complex research data (qualitative and quantitative) in a clear and compelling manner that inspires action
- Experience utilizing both qualitative analysis (e.g., content analysis, phenomenology, hypothesis testing) and quantitative analysis techniques (e.g., clustering, regression, pattern recognition, descriptive and inferential statistics)

WHAT MAKES A GOOD DATA SCIENTIST?

MODERN DATA SCIENTIST

Data Scientist, the sexiest job of the 21th century, requires a mixture of multidisciplinary skills ranging from an intersection of mathematics, statistics, computer science, communication and business. Finding a data scientist is hard. Finding people who understand who a data scientist is, is equally hard. So here is a little cheat sheet on who the modern data scientist really is.

MATH & STATISTICS

- Machine learning
- Statistical modeling
- Experiment design
- Bayesian inference
- Supervised learning: decision trees, random forests, logistic regression
- Unsupervised learning: clustering, dimensionality reduction
- Optimization: gradient descent and variants

PROGRAMMING & DATABASE

- Computer science fundamentals
- Scripting language e.g. Python
- Statistical computing packages, e.g., R
- Databases: SQL and NoSQL
- Relational algebra
- Parallel databases and parallel query processing
- MapReduce concepts
- Hadoop and Hive/Pig
- Custom reducers
- Experience with xaaS like AWS

DOMAIN KNOWLEDGE & SOFT SKILLS

- Passionate about the business
- Curious about data
- Influence without authority
- Hacker mindset
- Problem solver
- Strategic, proactive, creative, innovative and collaborative

COMMUNICATION & VISUALIZATION

- Able to engage with senior management
- Story telling skills
- Translate data-driven insights into decisions and actions
- Visual art design
- R packages like ggplot or lattice
- Knowledge of any of visualization tools e.g. Flare, D3.js, Tableau

http://www.marketingdistillery.com/2014/11/29/is-data-science-a-buzzword-modern-data-scientist-defined/

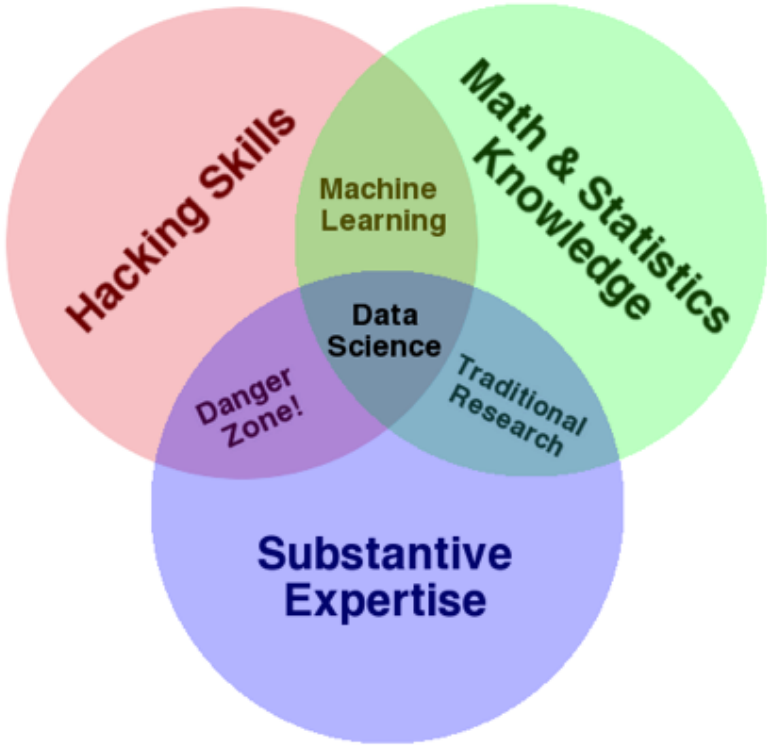
Michael E. Driscoll
@medriscoll

Following

Data scientists: better statisticians than most programmers & better programmers than most statisticians bit.ly/NHmRqu @peteskomoroch

Reply Retweet Favorite More Pocket

THE QUALITIES OF A DATA SCIENTIST



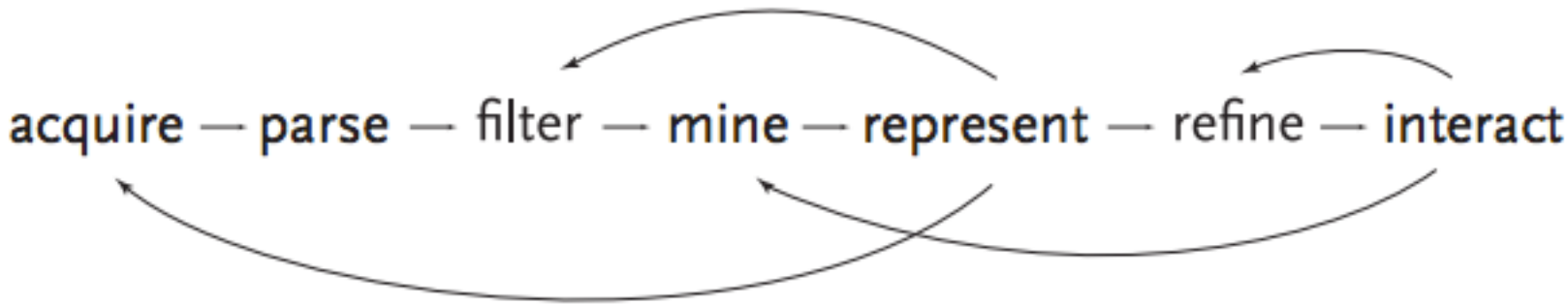
source: http://www.dataists.com/2010/09/the-data-science-venn-diagram/

II. THE DATA SCIENCE WORKFLOW

from Jeff Hammerbacher:

- 1. Identify problem
- 2. Instrument data sources
- 3. Collect data
- 4. Prepare data (integrate, transform, clean, impute, filter, aggregate)
- 5. Build model
- 6. Evaluate model
- 7. Communicate results

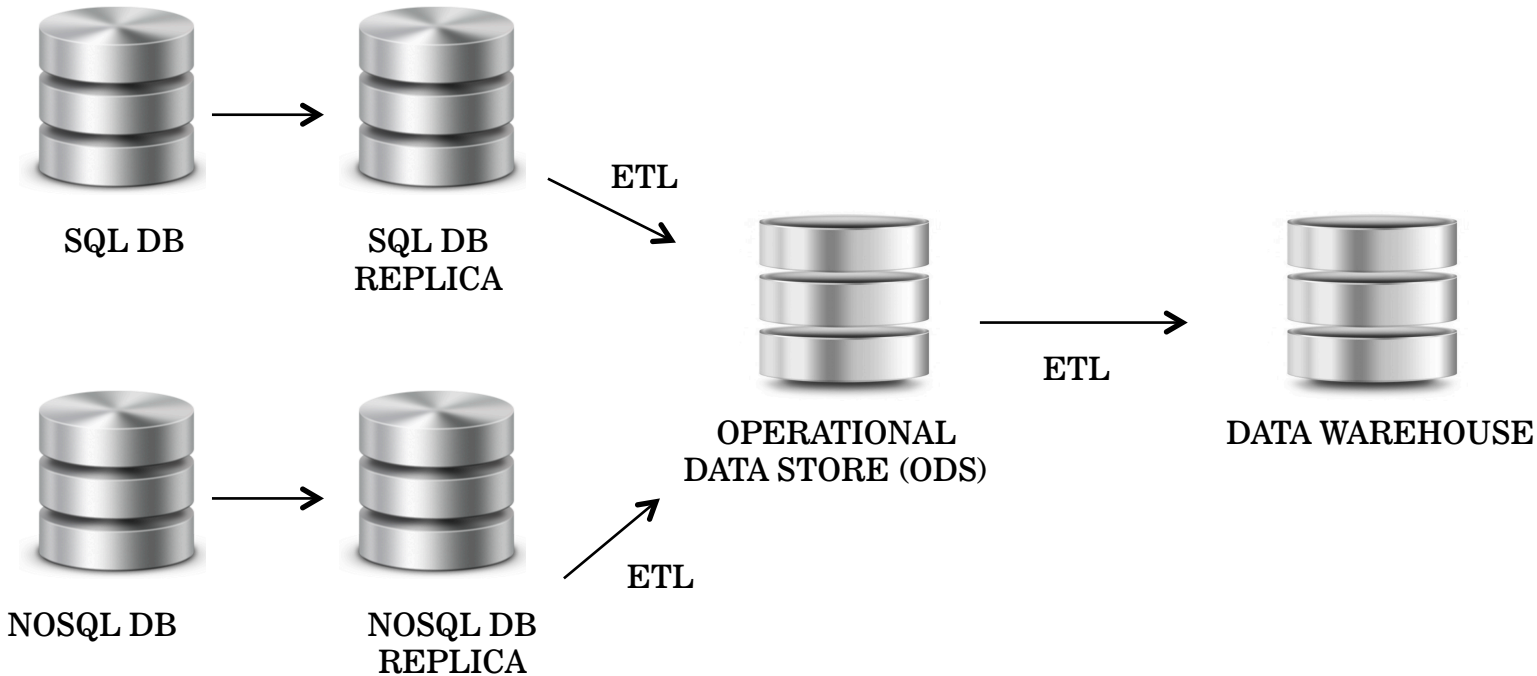
source: http://berkeleydatascience.files.wordpress.com/2012/01/20120117berkeley1.pdf



ZIPDECODE

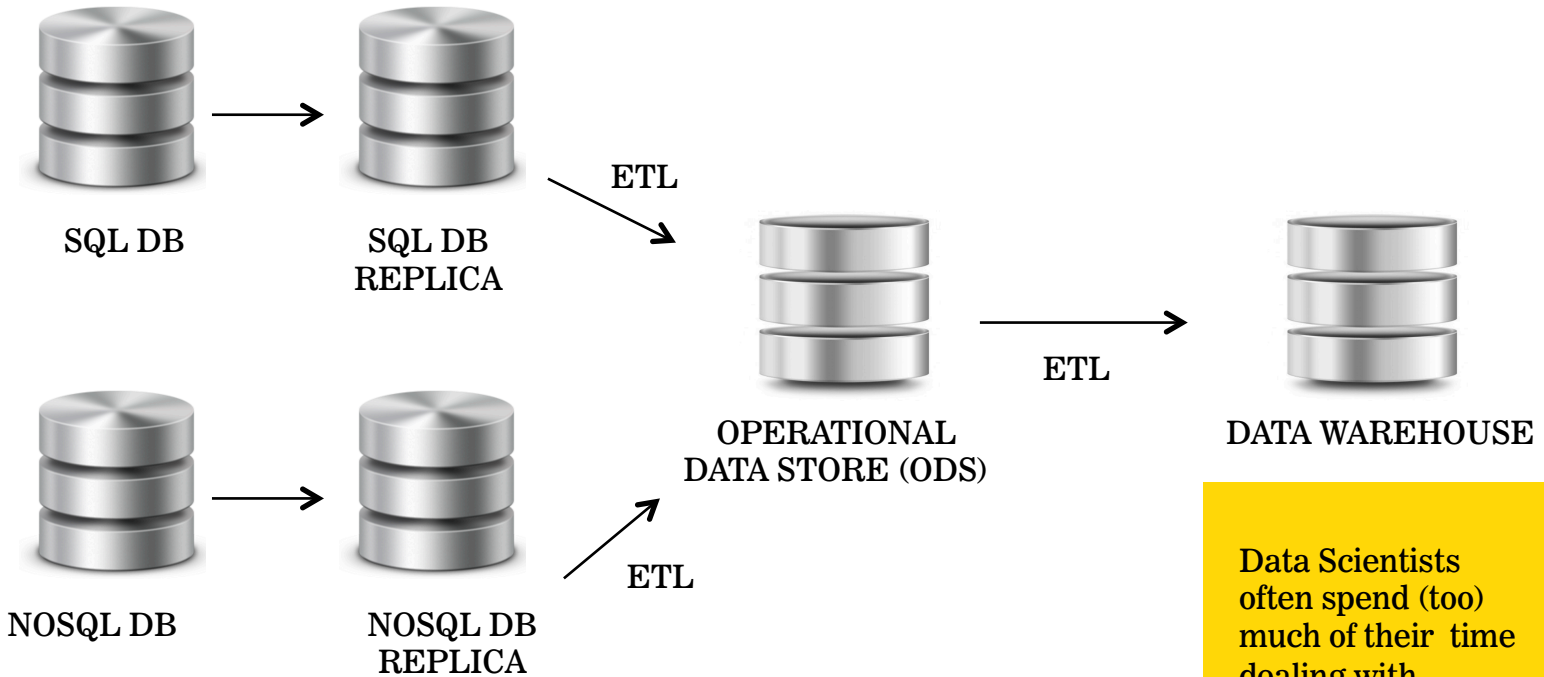
source: http://benfry.com/phd/dissertation-110323c.pdf

GETTING THE DATA EXAMPLE



*ETL: EXTRACT, TRANSFORM, LOAD

GETTING THE DATA EXAMPLE



Data Scientists often spend (too) much of their time dealing with complicated ETL processes

*ETL: EXTRACT, TRANSFORM, LOAD

Visualization & Communication are critical components to effective data storytelling

Examples:
- "11,419 people killing in 2013" *vs* Periscope

- "10% of users visit a the home page of a certain site and then stop" *vs* sunburst viz

10 minutes: You work for a mobile payments company and have a problem: Many of your users type in a user's name but sometimes pay the wrong person if the name is common (i.e. pay the wrong John Smith)! Describe the workflow of a data scientist who is attempting to solve this issue.

COURSE OVERVIEW

UNIT 3: FURTHERING MODELING TECHNIQUES

K-Means Clustering II	Lesson 8
Ensemble Techniques	Lesson 9
Decision Trees and Random Forests	Lesson 10
Support Vector Machines	Lesson 11
Dimensionality Reduction	Lesson 12
Recommendation Systems	Lesson 13

UNIT 4: OTHER TOOLS

▸ Database Technologies	Lesson 14
▸ Network Analysis	Lesson 15
▸ Map-Reduce	Lesson 16
▸ Final Project Working Session	Lesson 17
▸ Final Project Working Session	Lesson 18
▸ Where To Go Next	Lesson 19
▸ Final Project Working Session	Lesson 20
▸ Final Project Presentations	Lesson 21
▸ Final Project Presentations	Lesson 22

COMPUTER SETUP

Have you:

- Installed python (anaconda)
- Create a Git account, downloaded Git, and set global config (https://git-scm.com/book/en/v2/Getting-Started-First-Time-Git-Setup)
- Installed cygwin (windows)
- Launched ipython notebook
- Able to import:
    - Scikit-learn
    - Numpy
    - Pandas
    - nltk
    - Statsmodels
    - matplotlib

III. WORKING AT THE UNIX COMMAND LINE

EXERCISE – WORKING AT THE UNIX COMMAND LINE

KEY OBJECTIVES

- Navigate the filesystem
- Create, move, copy, and delete files & directories
- View & search files
- Edit & interact with files
- Combine steps
- Learn more

TOOLS

- ls, cd
- cat, touch, mv, cp, mkdir, rm, rmdir
- head, tail, less, cat, grep
- vim, tr, sort, uniq, wc
- pipe (|)
- man, apropos

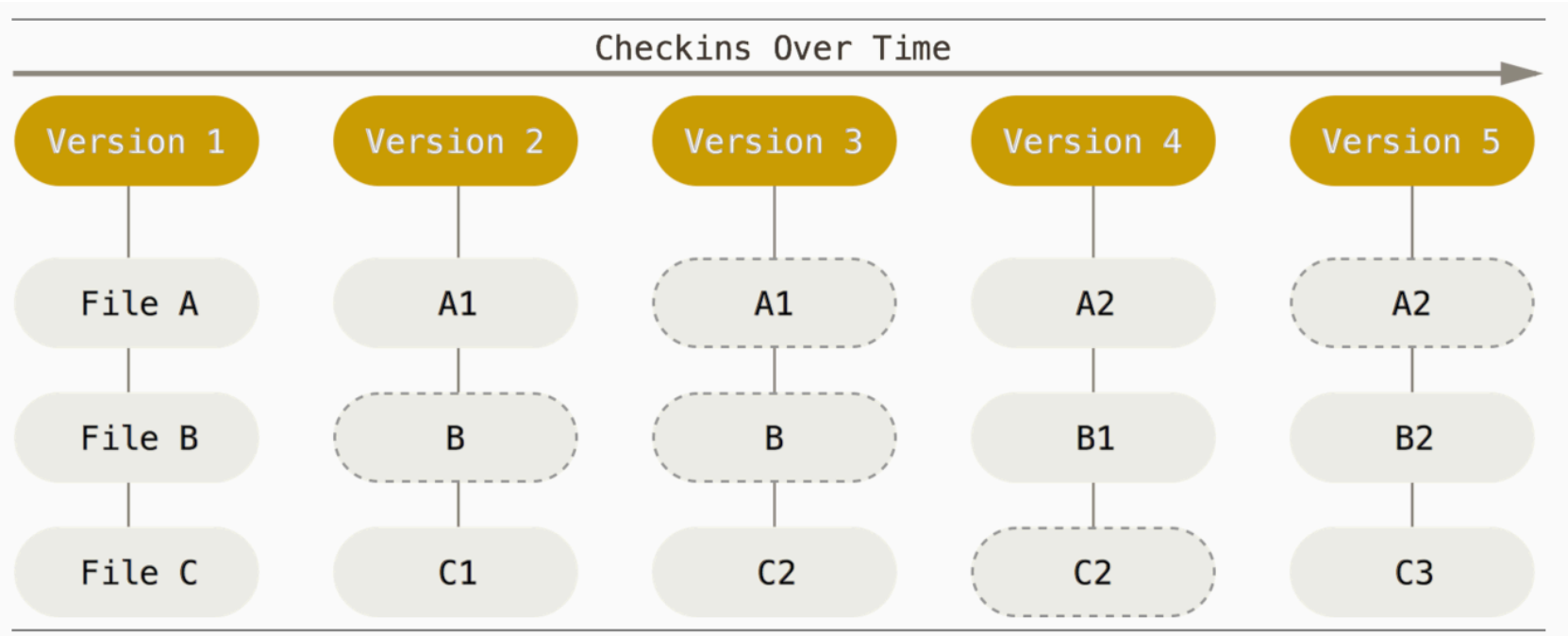
NOTE

Being comfortable at the command line makes your life much easier!

OTHER TOOLS

▸GIT

Git is a version control system. It stores a copy of all changes you make to files over time



▸GIT

Main commands:

- git clone <repo name>
  - This command copies all the code from a given repo

- git pull
  - This command updates all changes made to a repo since you last cloned

▸IN CLASS WORK

Lab time: pull the class directory and work on the lab found in class #1