INTRODUCTION TO DATA SCIENCE

MOHAMMAD NAYEEM TELI

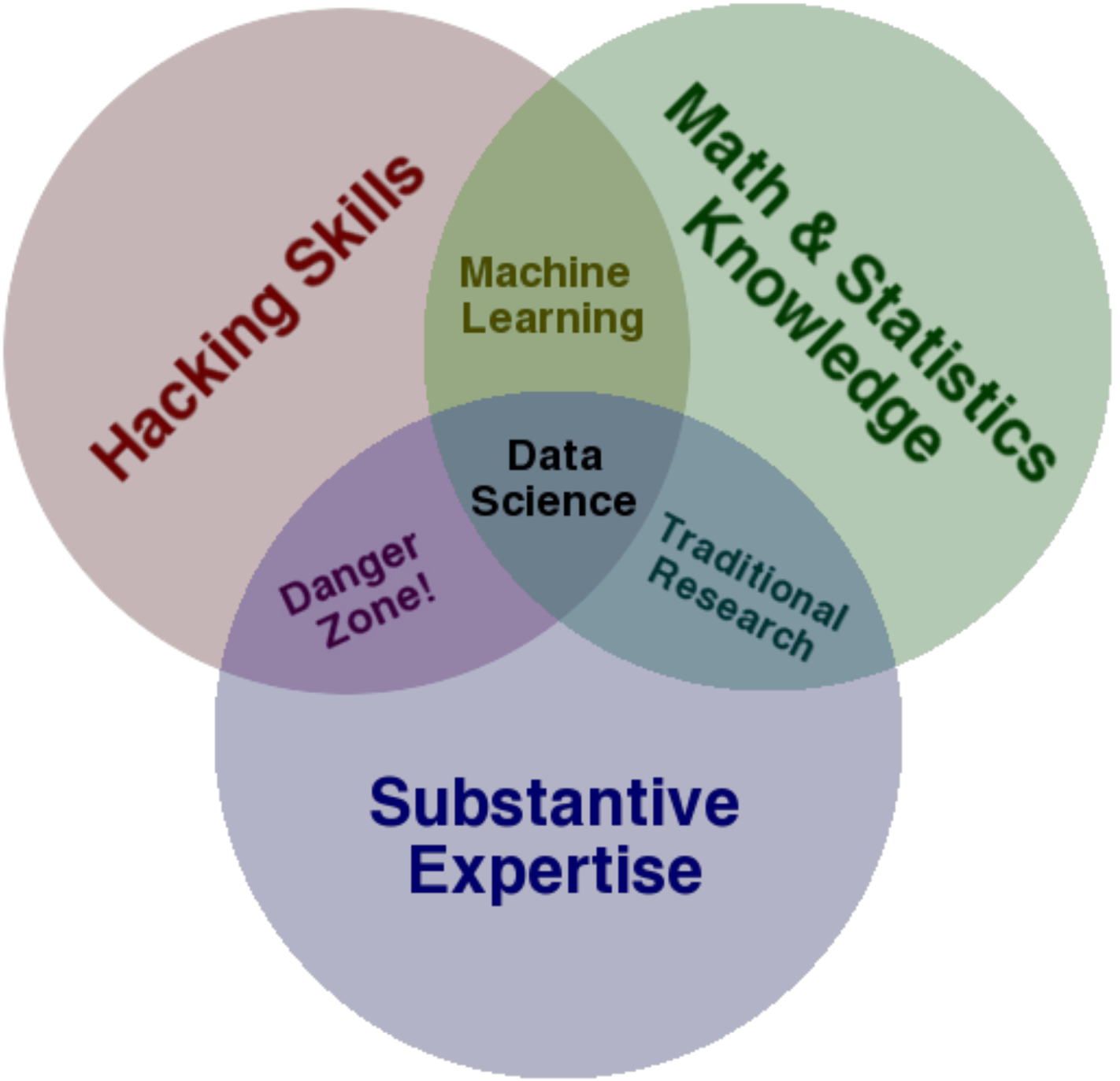
Lecture #1 – 05/2/2019

CMSC320
MTuWThF
2:00pm – 3:25pm

COMPUTER SCIENCE
UNIVERSITY OF MARYLAND

Data science is the application of computational and statistical techniques to address or gain [managerial or scientific] insight into some problem in the real world.

Zico Kolter
Machine Learning Prof, CMU



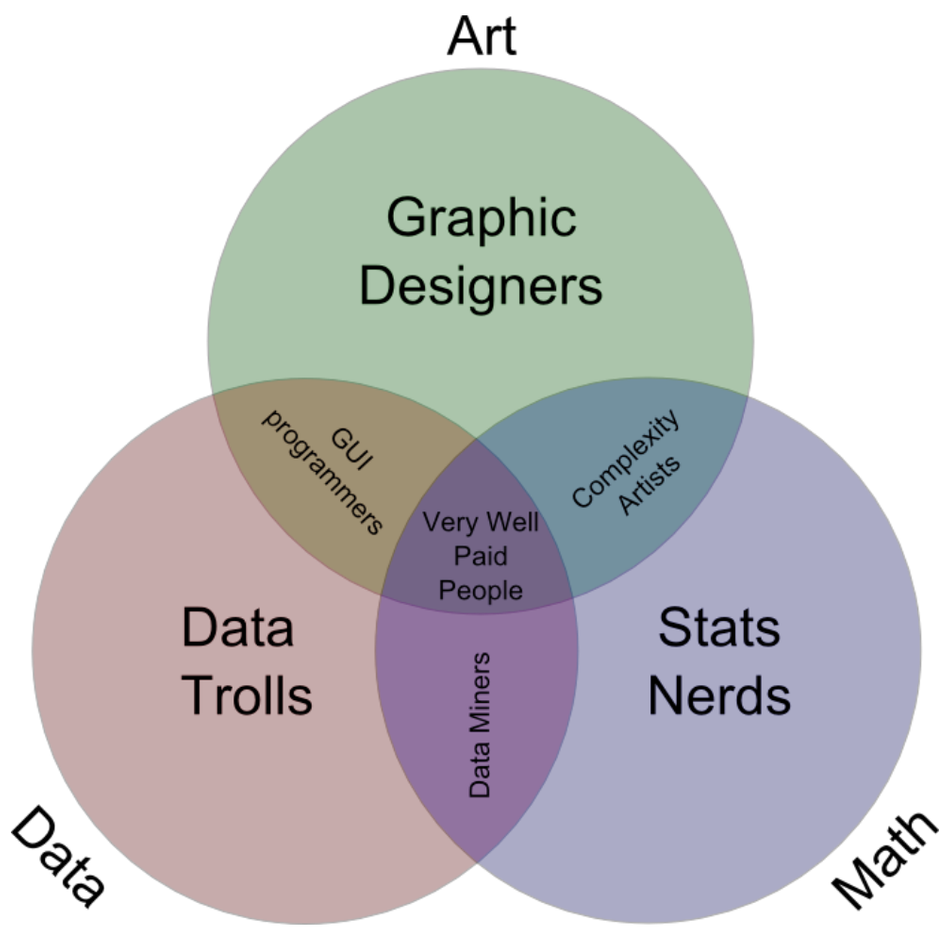
Drew Conway
CEO, Alluvium (analytics company)

MANY DEFINITIONS

Broad: necessarily larger than a single discipline

Interdisciplinary: statistics, computer science, operations research, statistical and machine learning, data warehousing, visualization, mathematics, information science, …

Insight-focused: grounded in the desire to find insights in data and leverage them to inform decision making



Tuomas Carsey, UNC

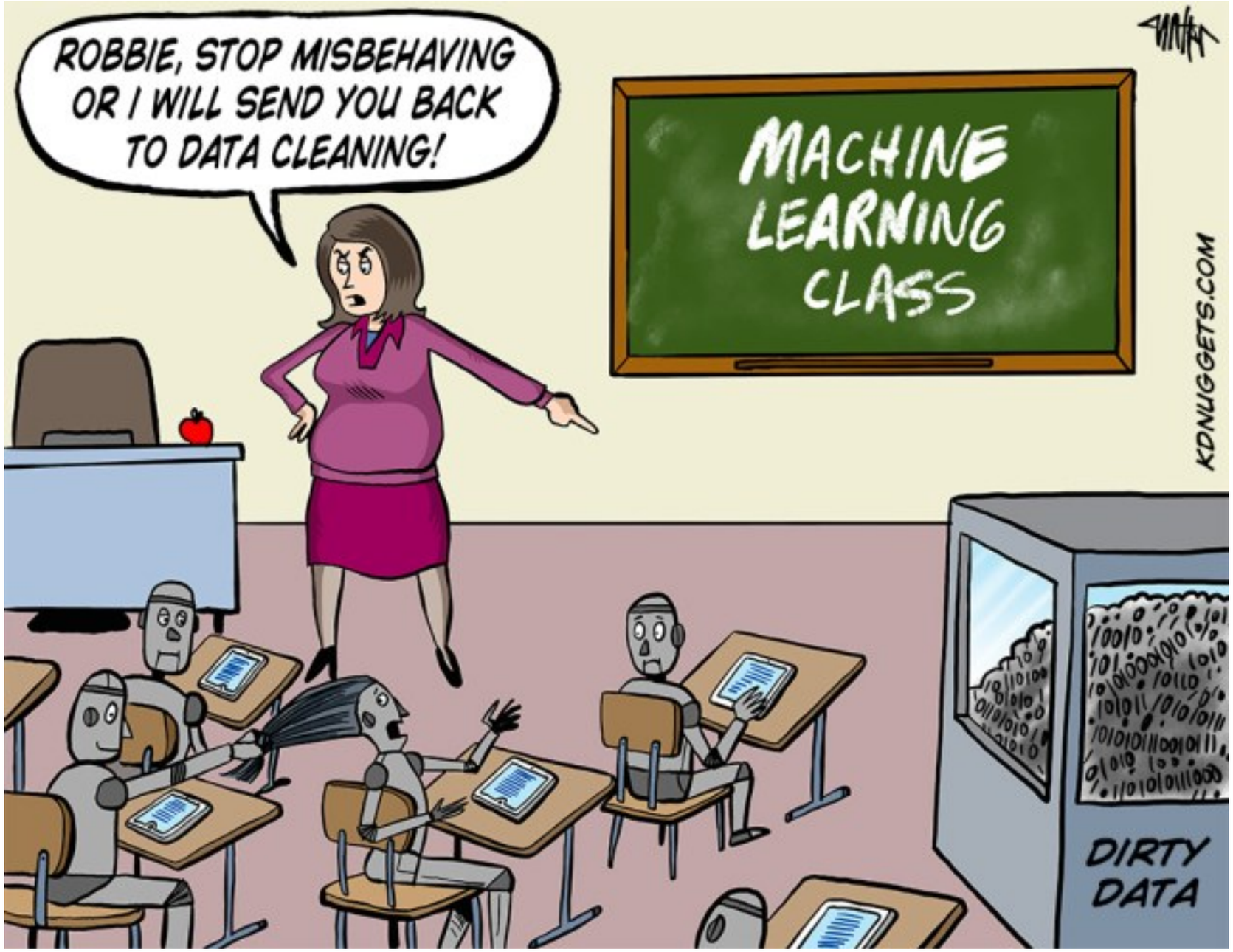
DATA SCIENCE LIFECYCLE

1. Collect Data

Remember Mr Pooter is not just a 'patient', he's an important source of valuable and readily marketable data!

DATA SCIENCE LIFECYCLE

2. Scrub Data

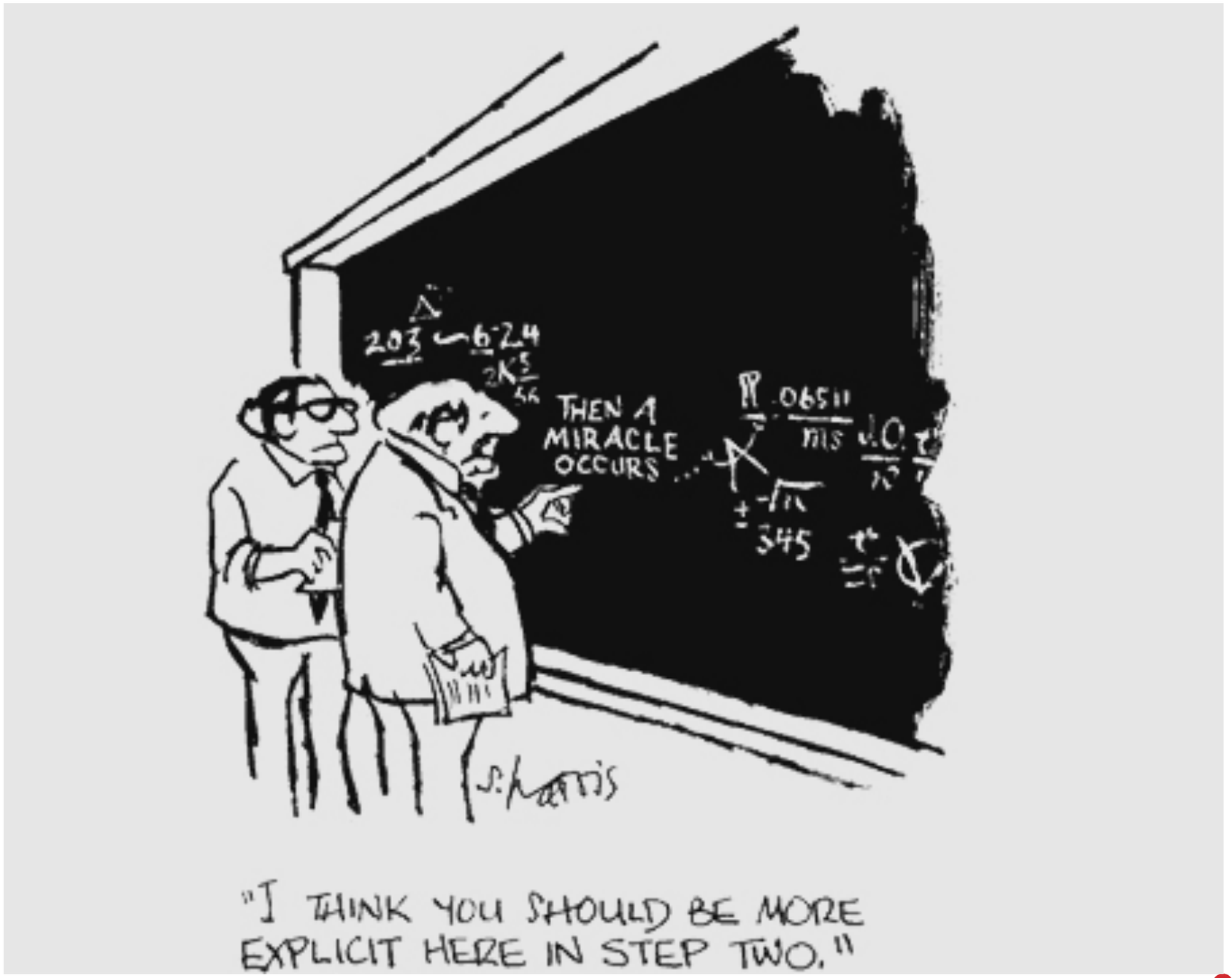


DATA SCIENCE LIFECYCLE

3. Explore Data

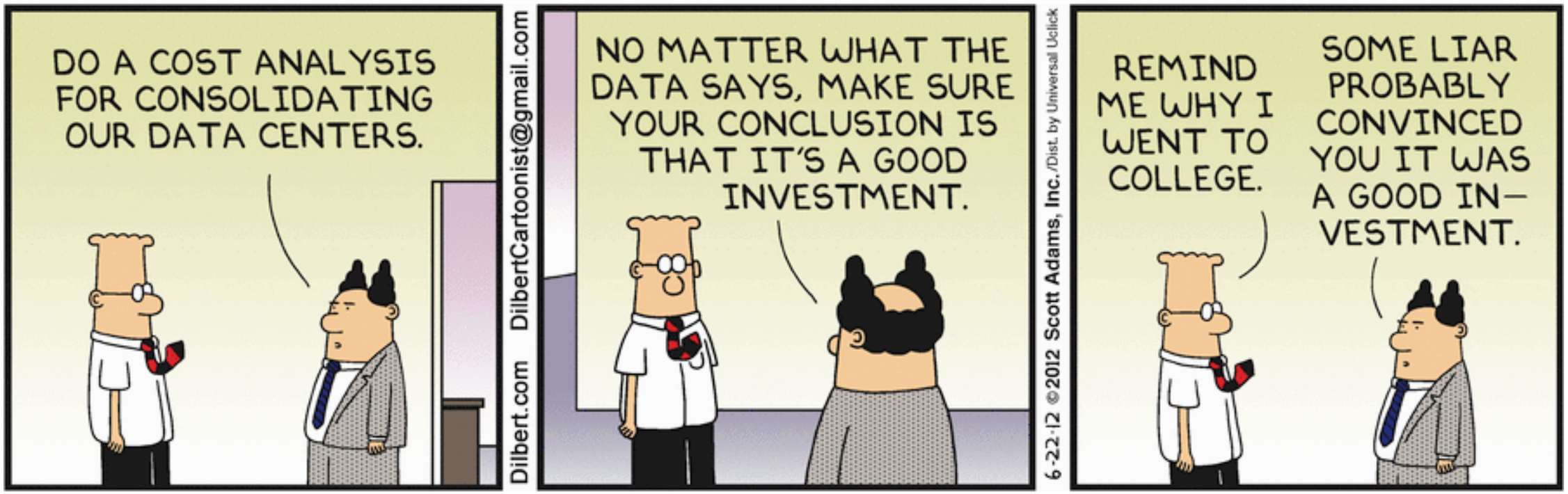
DATA SCIENCE LIFECYCLE

4. Build a model

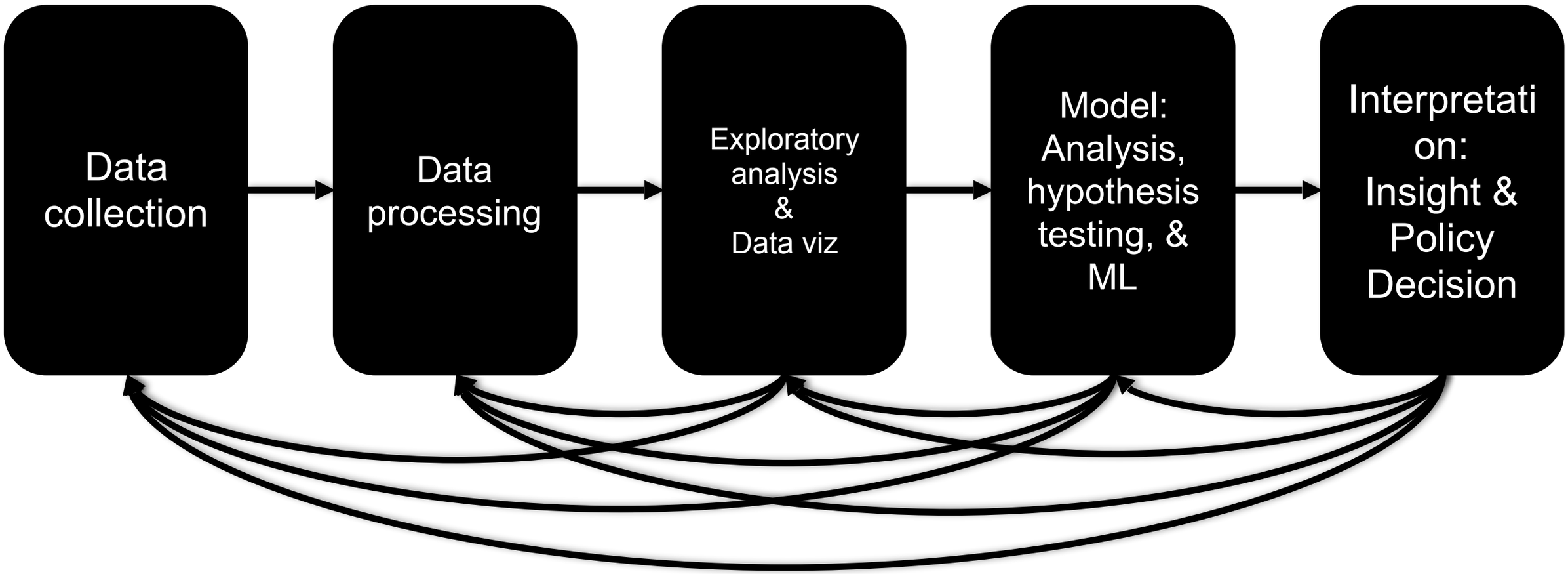


DATA SCIENCE LIFECYCLE

5. Interpretation



THE DATA LIFECYCLE



“The ability to take data—to be able to understand it, to process it, to extract value from it, to visualize it, to communicate it—that’s going to be a hugely important skill in the next decades, not only at the professional level but even at the educational level for elementary school kids, for high school kids, for college kids.”

Hal Varian
Chief Economist at Google

MOTIVATION

Explosion of data, in pretty much every domain

- Sensing devices and sensor networks that can monitor everything 24/7 from temperature to pollution to vital signs
- Increasingly sophisticated smart phones
- Internet, social networks makes it easy to publish data
- Scientific experiments and simulations → astronomical data volumes
- Internet of Things
- Dataification: taking all aspects of life and turning them into data (e.g., what you like/enjoy has been turned into a stream of your "likes")

How to handle that data? How to extract interesting actionable insights and scientific knowledge?

Data volumes expected to get much worse

FOUR V'S OF BIG DATA

Increasing data Volumes

- Scientific data: 1.5GB per genome -- can be sequenced in .5 hrs
- 500M tweets per day
- 2.5 Quintillion bytes of data created every day

Variety:

- Structured data, spreadsheets, photos, videos, natural text, ...

Velocity

- Sensors everywhere -- can generate high-rate "data streams"
- Real-time analytics requires data to be consumed as fast as it is generated

Veracity

- How do you decide what to trust? How to remove noise? How to fill in missing values?

THIS COURSE

End-to-end data science lifecycle

Acquiring, wrangling, cleaning, and integrating data; Setting up pipelines for ETL

Data modeling

Information Visualization

Ethics, Privacy, and Reproducibility

Feel free to tell me if there are topics that you think we should cover…

Info: **http://www.cs.umd.edu/class/summer2019/cmsc320/**
Piazza: **https://piazza.com/umd/summer2019/cmsc320**
Gradescope: https://www.gradescope.com/courses/50663
ELMS: (everyone should be registered automatically)

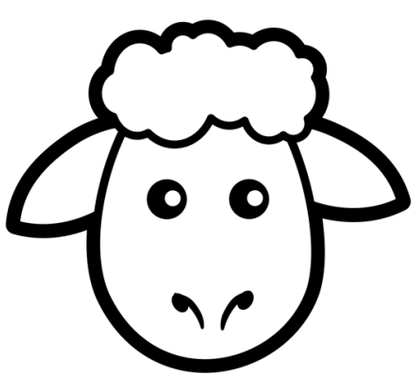
PREREQUISITE KNOWLEDGE

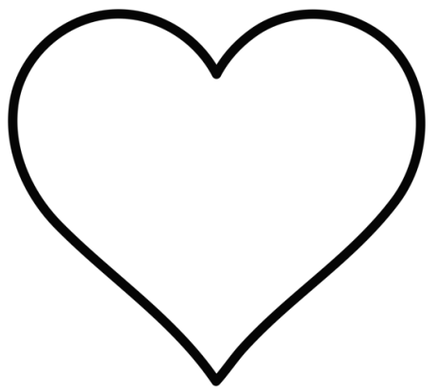
Aimed at folks with some CS knowledge – but likely accessible to others with programming experience and mathematical maturity.

We do not assume:

- Experience with Python, pandas, scikit-learn, matplotlib, etc …
- Deep statistics or any ML knowledge
- Database or distributed systems knowledge

We do assume: You want to be here!





(TENTATIVE) COURSE STRUCTURE

First few lectures: intro & primers in the Python data science stack

Next : data collection & management

- Best practices, data wrangling, exploratory analysis, ethics, debugging, visualization, etc …

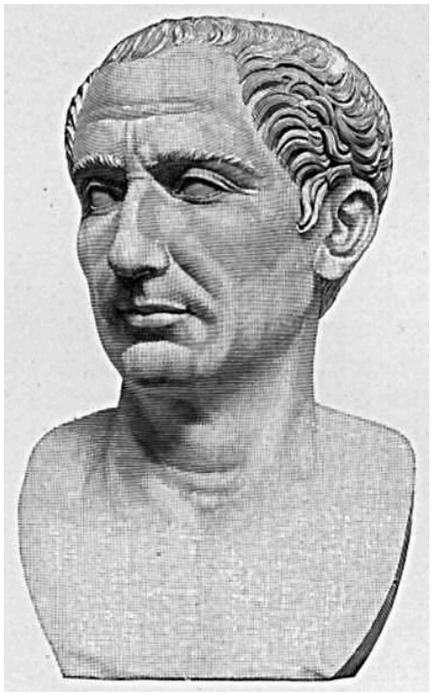
Next lectures: statistical modeling & ML

- Statistical learning, regression, classification, cross-validation, model evaluation, hypothesis testing, etc …

Midterm

Final lectures: advanced topics

- Dimensionality reduction, distributed learning, big data, distributed computation
- *Either* group presentations or more lectures



Ambitious …

GRADE #1: MINI-PROJECTS

Students will complete four mini-project assignments:

- **Case studies** meant to mimic what you, a future data scientist, will see in industry. They should be fun ☺️.

The rules:

- Allowed: small group **discussions**
- Required: individual **programming & writing**
- Never allowed: public posting of solutions

Deliverable:

- Turn in an .ipynb of a Jupyter notebook on ELMS and a pdf on Gradescope.

GRADE #2: READING HOMEWORKS

We will post (bi)weekly reading assignments. Mix of:

- Blog posts
- Academic articles
- News articles

Weekly quiz to be taken in class, every Monday.

Individual quiz grades are pass/fail:

- At least 60% correct → Pass
- Less than 60% correct → Fail

Must take at least ten of these quizzes over the semester

GRADE #3: MINI-TUTORIAL

In lieu of a final exam, you'll create a mini-tutorial that:

- Identifies a raw data source
- Processes and stores that data
- Performs exploratory data analysis & visualization
- Derives insight(s) using statistics and ML
- Communicates those insights as actionable text

Individual or group project

Will be hosted publicly online (GitHub Pages) and will strengthen your portfolio.

READY-MADE DATASET REPOSITORIES

https://www.data.gov/

- US-centric agriculture, climate, education, energy, finance, health, manufacturing data, …

https://cloud.google.com/bigquery/public-data/

- BigQuery (Google Cloud) public datasets (bikeshare, GitHub, Hacker News, Form 990 non-profits, NOAA, …)

https://www.kaggle.com/datasets

- Microsoft-owned, various (Billboard Top 100 lyrics, credit card fraud, crime in Chicago, global terrorism, world happiness, …)

https://aws.amazon.com/public-datasets/

- AWS-hosted, various (NASA, a bunch of genome stuff, Google Books n-grams, Multimedia Commons, …)

NEW DATASET IDEAS

Fraternal Order of Police vs Black Lives Matter

Linking finance data to ${anything_else}

Something having to do with Pokémon statistics?

Look through http://www.alexa.com/topsites and scrape something interesting!

University of Maryland-related, or College Park-related, stuff

- Check out http://umd.io/ – open source project; maybe your data collection and cleaning scripts can be added to this!

Honestly, pretty much anything! Just document everything.

Reproducibility!