SPRING 2024
DSE I2450 - 3GG
BIG Data &
Scalable Computation

Professor
Madeline Blount
she/her

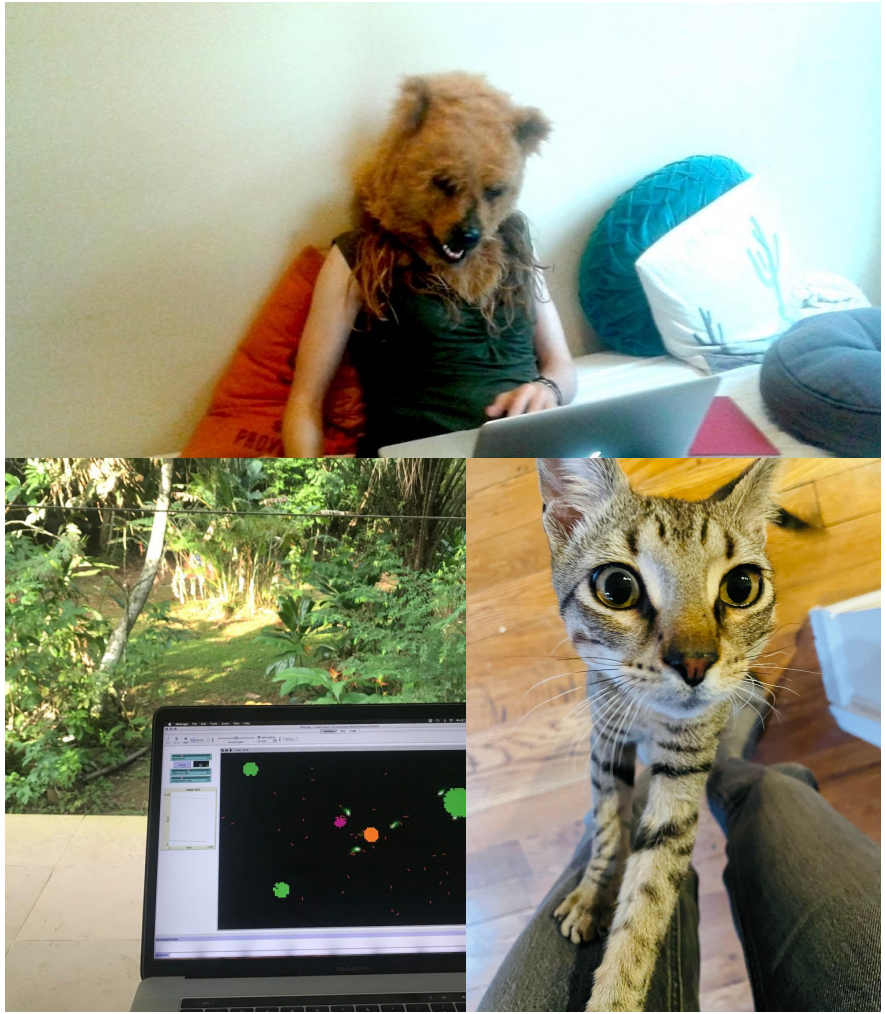
Week 0

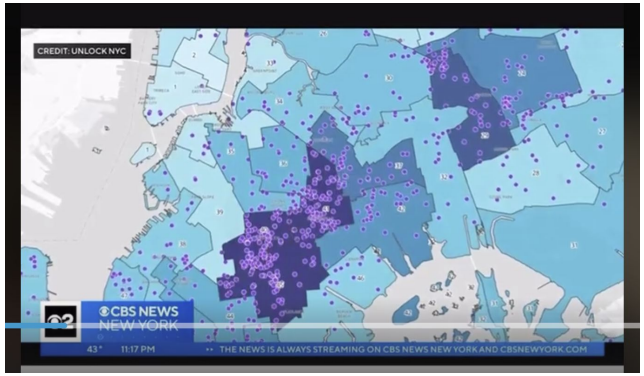
Dall-E 3: “inside a big data system, use your imagination”

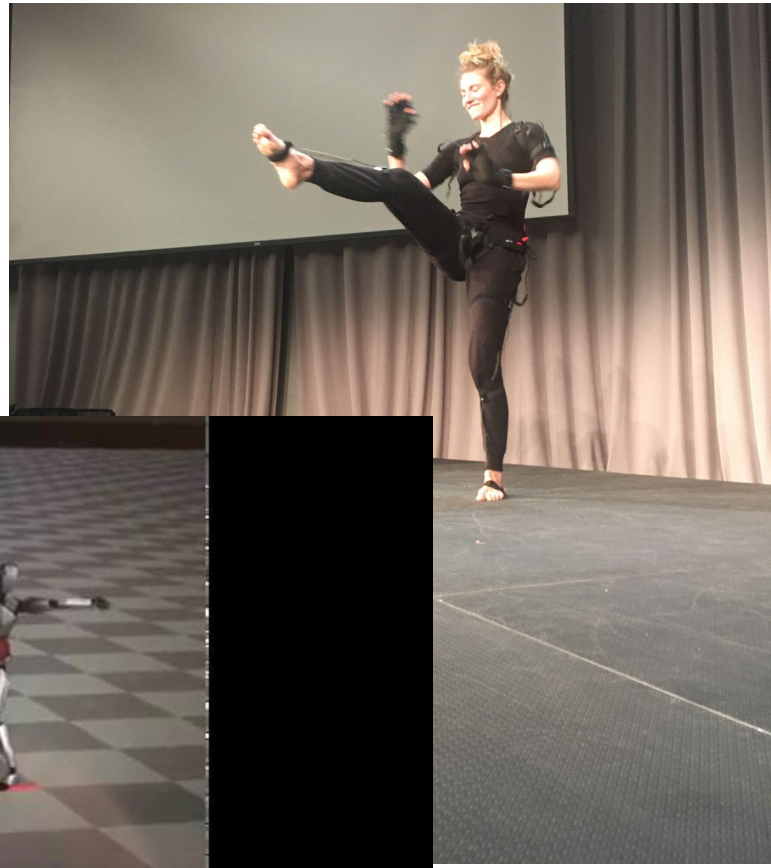
Professor
Madeline Blount
she/her

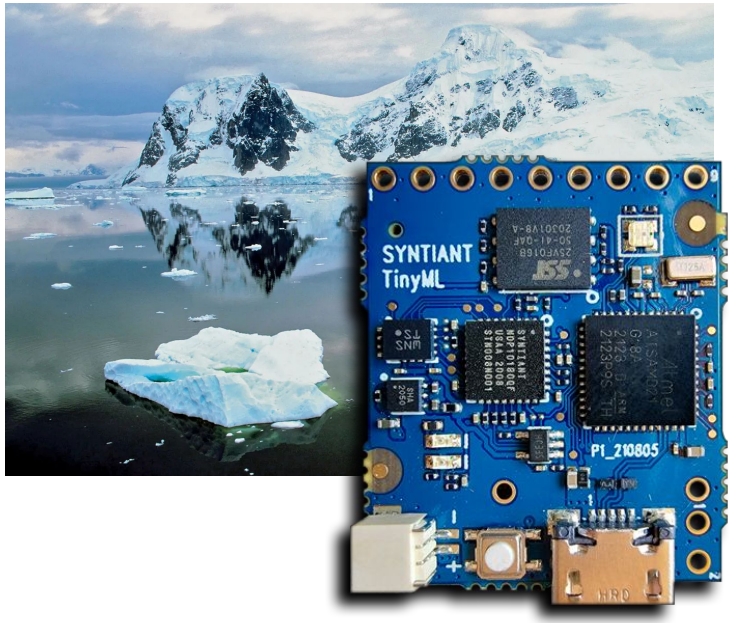
transdisciplinary
technologist

data engineering




CREDIT: UNLOCK NYC
CBS NEWS NEW YORK
43°
11:17 PM
THE NEWS IS ALWAYS STREAMING ON CBS NEWS NEW YORK AND CBSNEWYORK.COM




SYNTIANT
TinyML
Pi_210805

"BIG DATA"

The New York Times

Reading by the Numbers: When Big Data Meets Literature

The New York Times

UNBOXED

Big Data, Trying to Build Better Workers

The New York Times

NEWS ANALYSIS

The Age of Big Data

Share free access 82

By Steve Lohr
Feb. 11, 2012

The New York Times

Will You Graduate? Ask Big Data

By Joseph B. Treaster
Feb. 2, 2017

The New York Times

Opinion

OP-ED CONTRIBUTOR

A.I. and Big Data Could Power a New War on Poverty

By Elisabeth A. Mason
Jan. 1, 2018

Gov. Hochul Announces CUNY to Receive $75 Million from the Simons Foundation, Largest Donation in University History

January 17, 2024

$50 Million Will Create 25 More Faculty Positions and a New Master's Degree in Computational Science

$25 Million Contribution Designated for University AI Research

Jan. 17, 2024 CUNY

BIG DATA IS … RELATIVE

- "Big Data usually refers to a dataset that is **too big to fit into your available memory**, or too big to store on your own hard drive, or too big to fit into an Excel spreadsheet." - Hilary Mason (2012)

- **"available memory"** = RAM on 1 machine, processing

- Dataset too large to be dealt with via "traditional" data-processing techniques (wikipedia)

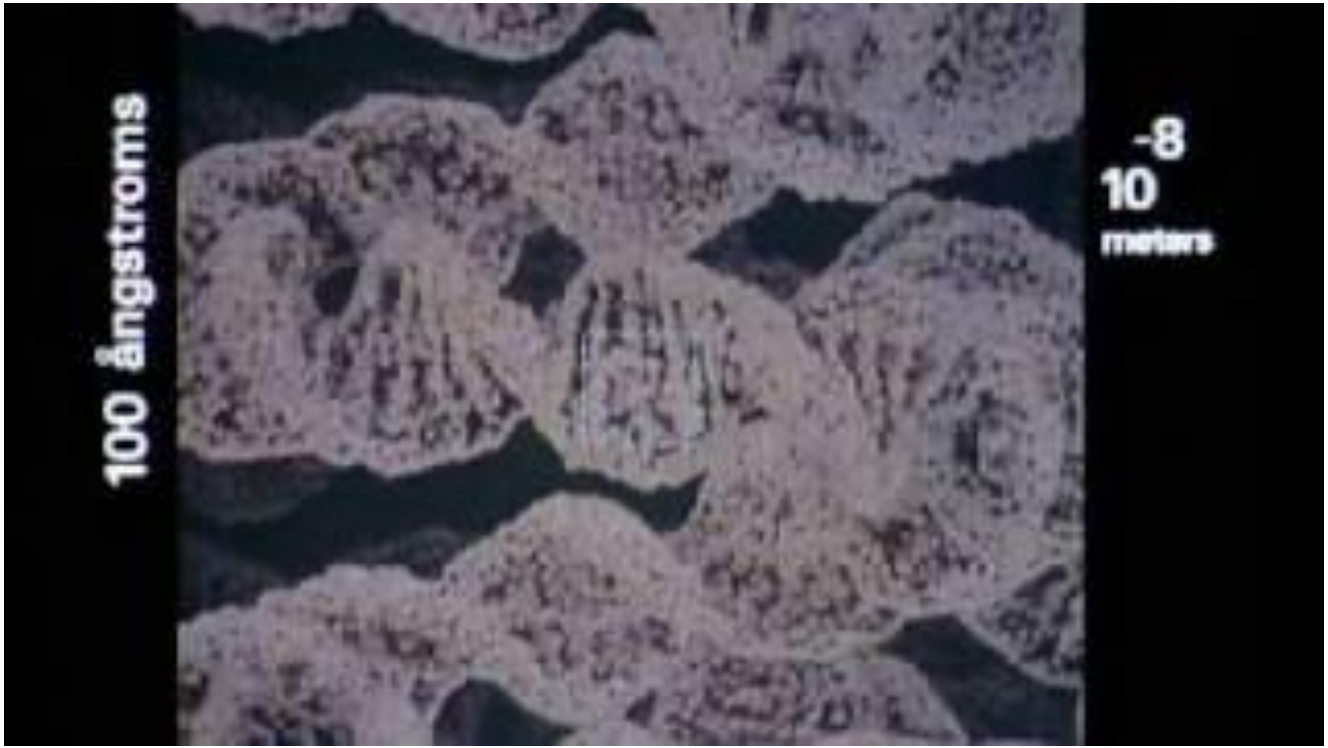
- Loose definition

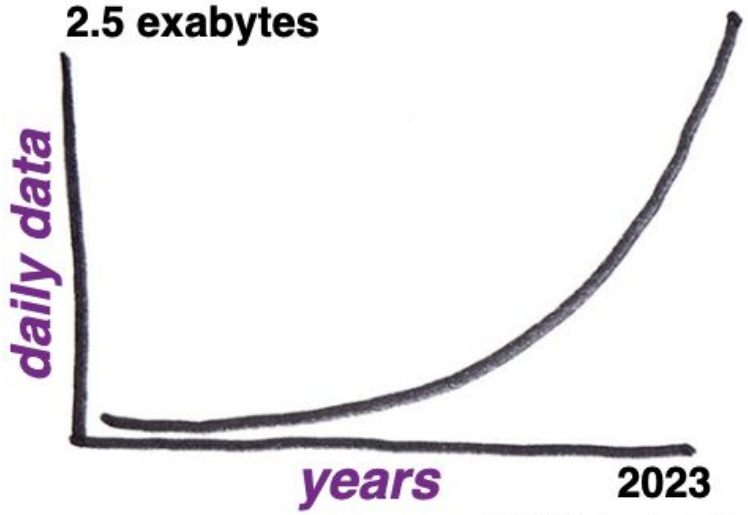

CARDINALPATH
(V) @SUM(B3...B5)
A B C D E F
145
231
198
574
It doesn't fit in Excel!
@SHamelCP
VisiCalc, 1979
Apple II, 8k of memory

HOW BIG IS THE INTERNET?

- **Very** difficult question!
- "Zettabyte" era since 2010's, estimates
  - Zettabyte = 1 million petabytes
  - Exabyte = 1 thousand petabytes
  - Petabyte = 1 million gigabytes
  - Gigabyte = 1 thousand megabytes
  - Megabyte = 1 million bytes
  - Byte = 8 bits, 8 1/0

Thinking about scales:
POWERS OF 10, Eames 1977





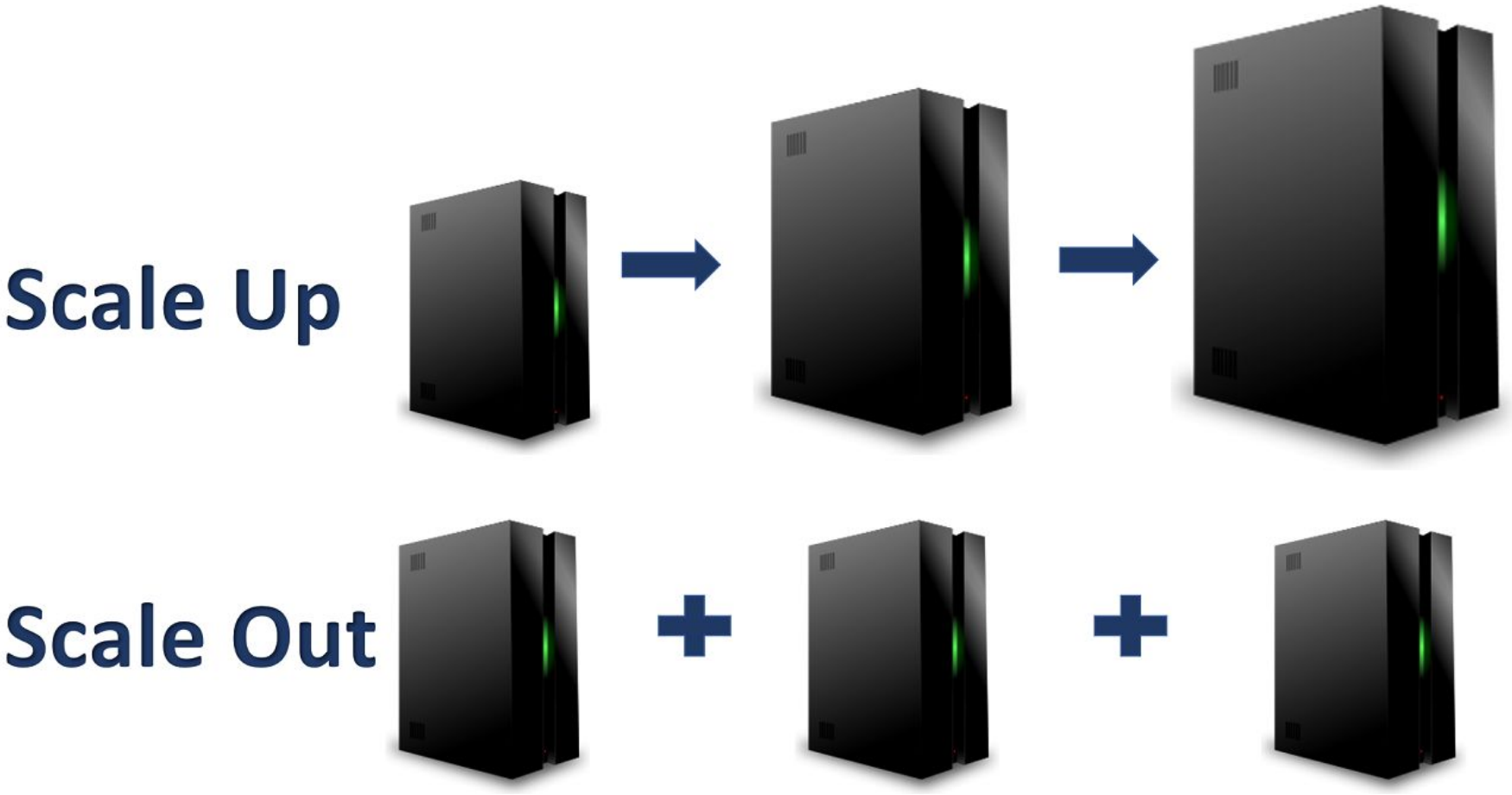
[IBMbigdata]

Every two days we create as much data as much we did from the dawn of humanity to 2003.

[Eric Schmidt, Google]

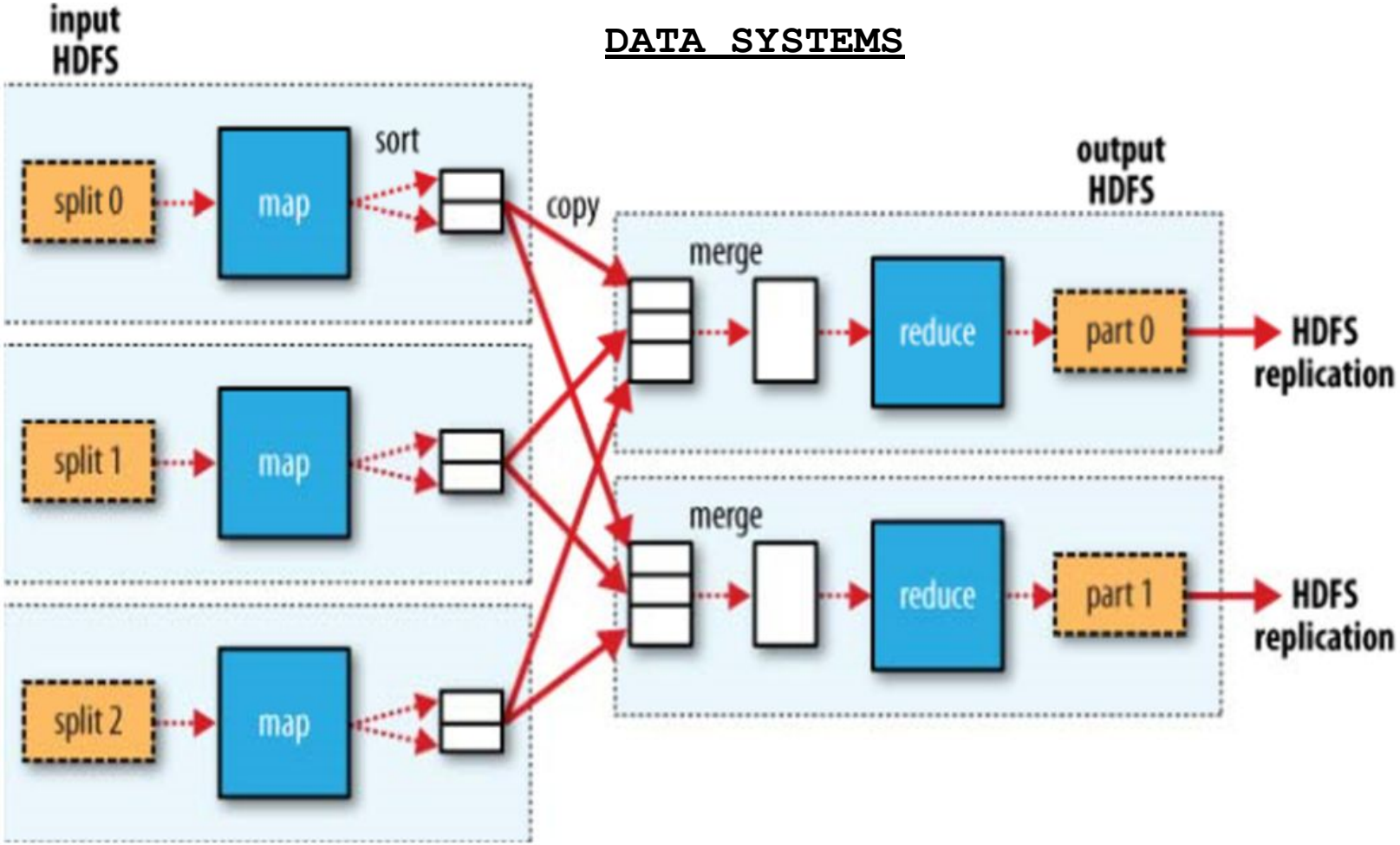
Source: Harvard CS165, Stratos

“SCALABLE COMPUTATION”



Source: Avantra.com

DATA SYSTEMS



Google data center

The Dalles, Oregon

Getty Images, OPB, Google

Hardware inside a (small) data center: University of Buffalo

Trying to get us a **tour of a data center** in NYC

(375 Pearl St. - TBD)

data systems are everywhere…

facebook

Google™

Source: Harvard CS165, Stratos

what will we do in this class?

- explore foundational concepts and architectures of big data systems (incl. Hadoop, MapReduce, Spark)
- work hands-on with big data programming paradigms via libraries (Python, SQL)
- explore current cloud platforms leveraging big data clusters (Databricks, Hugging Face, MongoDB, etc.)
- prepare for an evolving ecosystem by *learning to learn and teach* new technologies

what will we do in this class?

- compare, evaluate, and critique new research and new tools in the field
- interrogate the concept of big data as a form of knowledge production
- delve into the real limitations and ethical urgencies surrounding big data's current, growing role in sociotechnical systems

how will we do this class?

- **HYBRID!** 🏙️ some in-person, 🏠 some on zoom, 🦋 some asynchronous
- more **research seminar** than lecture course
- read original papers + new research, original documentation + new tools
- self-paced programming assignment (mid-term)
- self-directed research on a topic

Dall-E 3

Dall-E 3

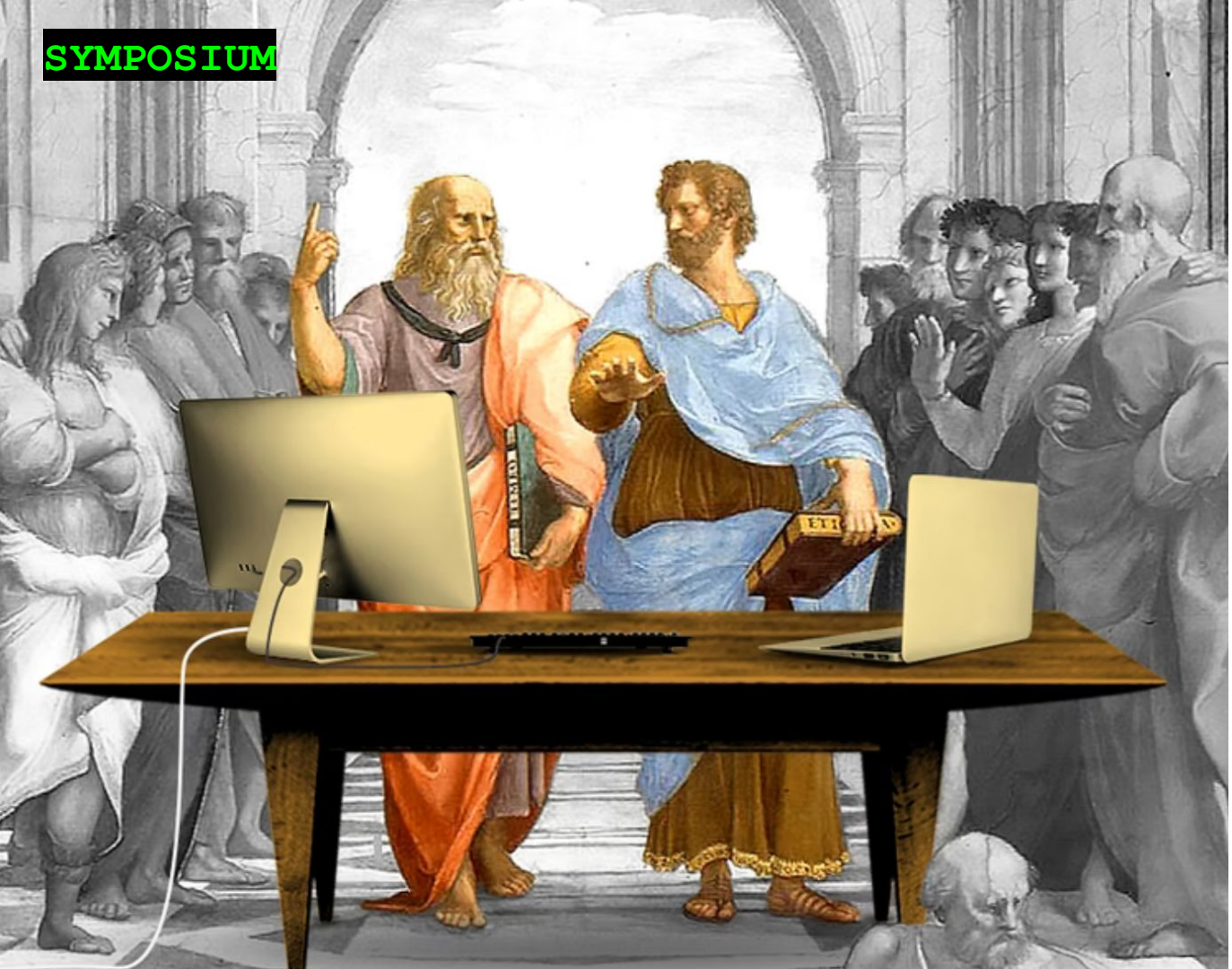
how will we do this class?

- Phase I: Weeks 0-5
  - foundations of infrastructure/systems
  - research papers
  - programming assignment

- Phase II: Weeks 6-9
  - cloud demos
  - light class reading
  - self-directed reading: bibliography

- Phase III: Weeks 10-end
  - class **symposium**
  - final paper/project work

SYMPOSIUM



- Choose topic + partner (start early!)
- Turn in abstract + bibliography (requirements on syllabus)
- READ! READ and READ. And tinker.
- Symposium Weeks 10+11, 30 min.

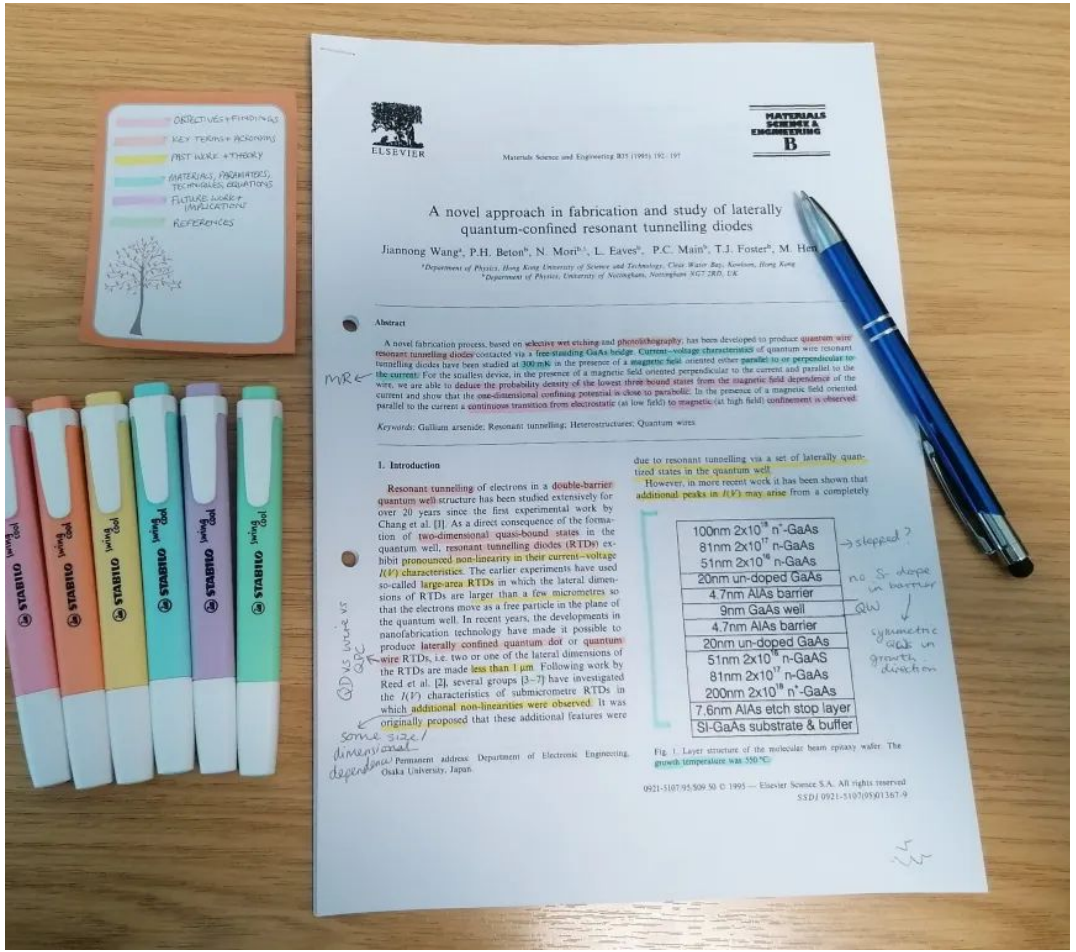
Reading original research papers = challenging!

Draw notes all over them

Expect to have lots of questions

Cross reference with other listed resources

Goal: to be able to discuss mechanics of system as a class, "how does this work"?



1. Read posted chapter/article

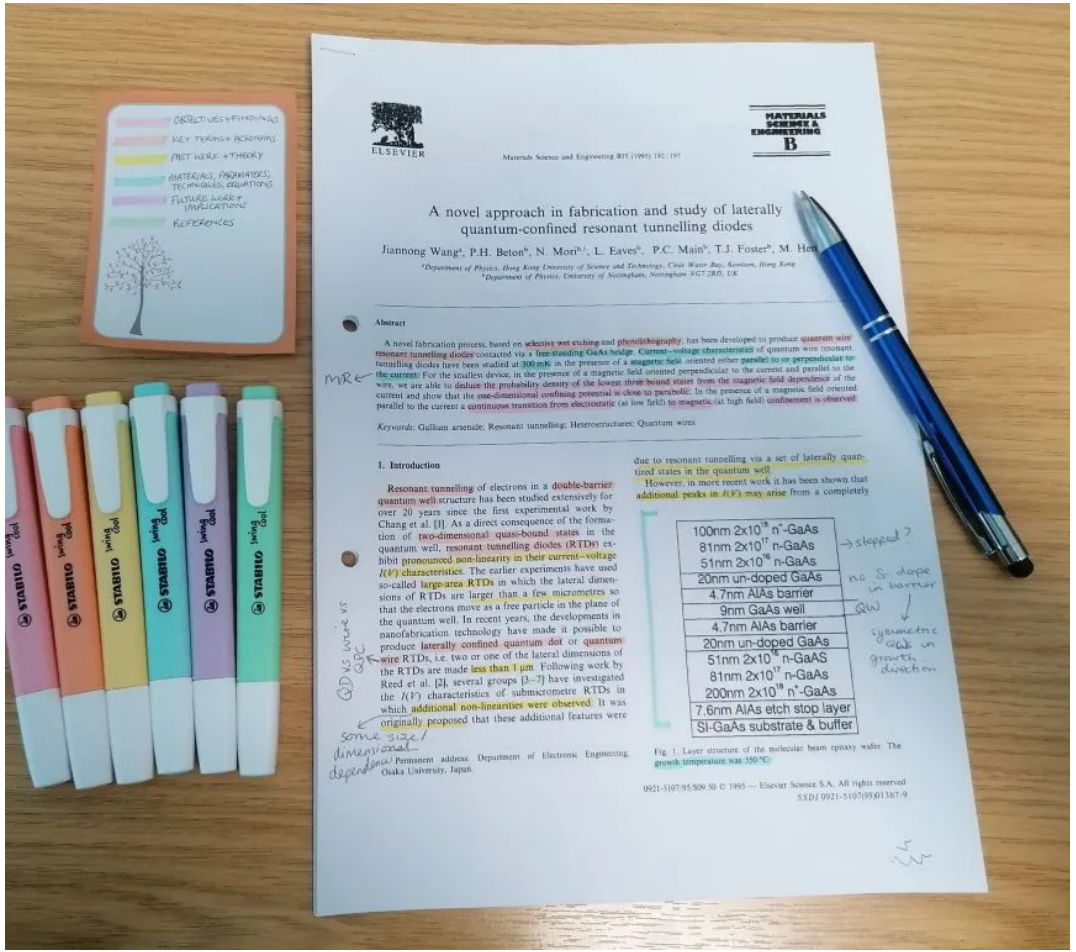
2. Read original paper

3. Make notes of terms/concepts you don't understand

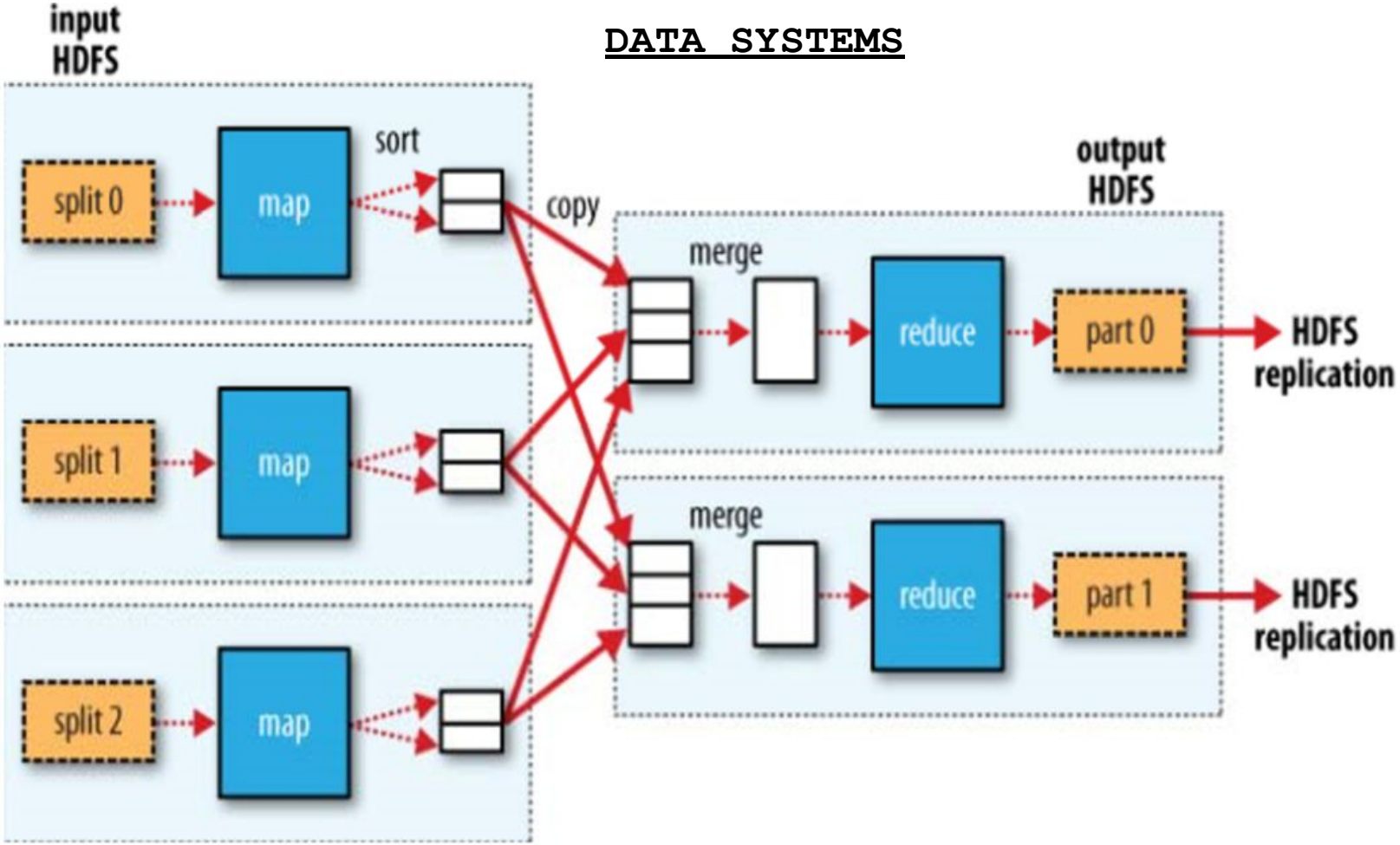
4. Look up what you need

5. Recommended videos, articles - SIMPLE!

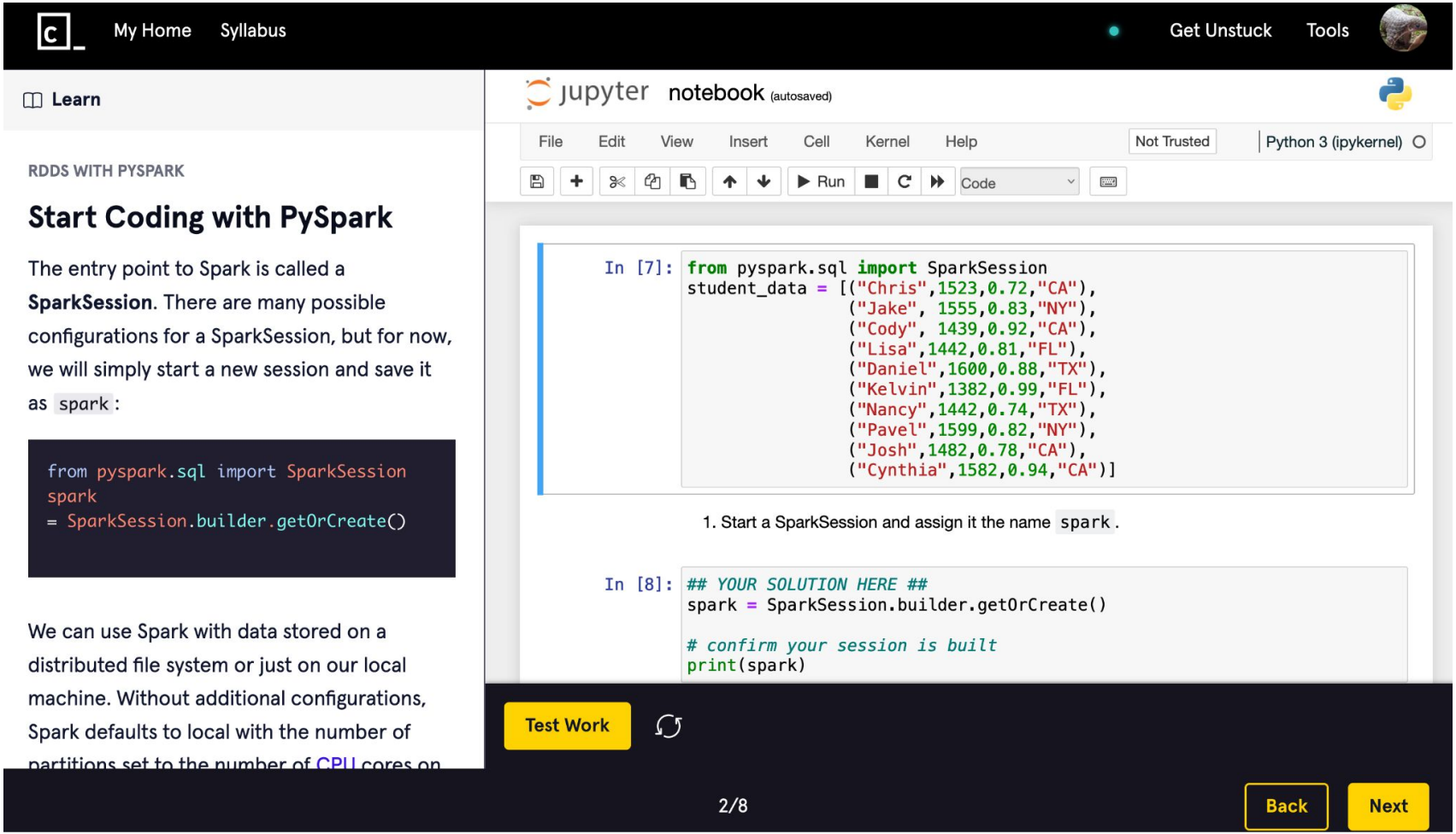
Idea to be able to <u>explain the system to someone else</u>



DATA SYSTEMS



programming assignment: codecademy, Python + SQL, self-paced



🤖 code recommendations!

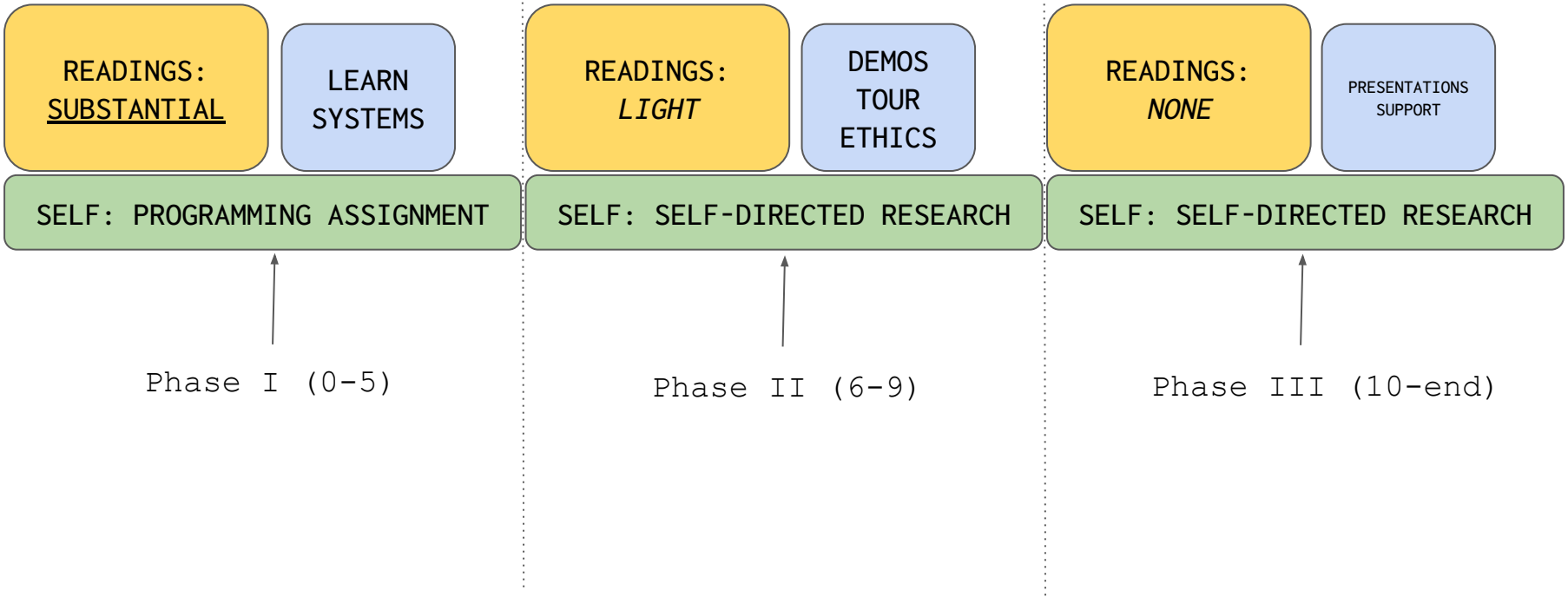
- brush up your Python 🐍
- particularly: lambda functions
  - `x = lambda a : a + 10`
  - `print(x(5))`
- review SQL
  - `SELECT select_list`
  - `FROM schema_name.table_name`
  - `WHERE column_name`
  - `IN (values);`

how will we do this class?

= assigned readings

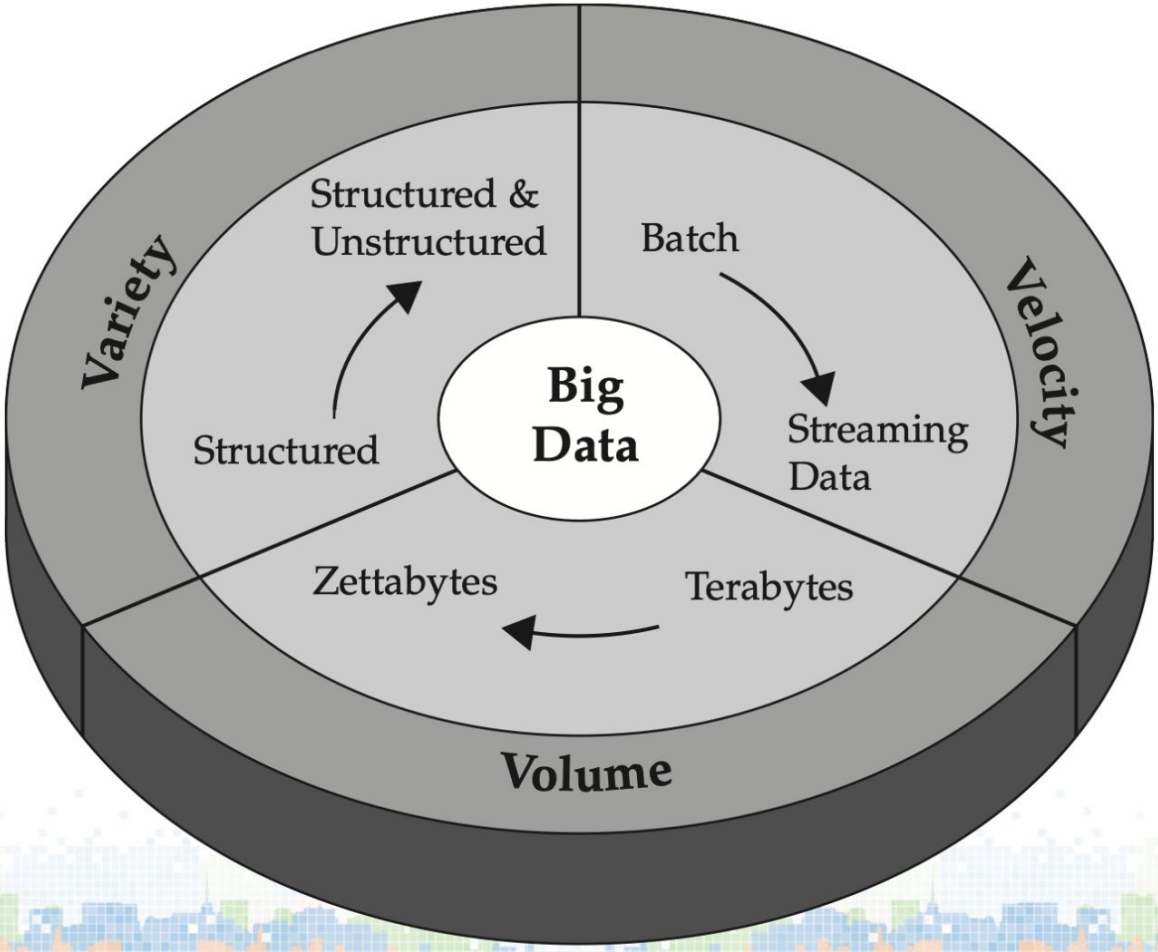
= in-class work

= self-directed work



📅 what you are responsible for:

- Weekly: active participation
- Weekly: 1 Discord post, update (see syllabus)
- 1 programming mid-term assignment (March 8th)
- 1 abstract + bibliography proposal (March 13th)
- 1 symposium presentation (30 min., Week 10 or 11)
- 1 final paper/project + reflection (May 20th)



[Source: IBM, 2012]
[Source: MapR 2014]