W205
Storage & Retrieval

Live Session - Week 1

Uri Schonfeld
shuri@berkeley.edu

datascience@berkeley

Agenda

- Who are we (you and I)
- First Lab
- What is Data Science?
- The course:
    - What you'll learn.
    - Structure: Lectures, live sessions, labs, exercises.
    - what are the live sessions for?
- Async session short review:
    - Data Dimensions
    - Processing Dimensions
- Data Science Case Studies
- Weekly reading material: discussion
- Data Sciencing a practice run

Uri Schonfeld
shuri@berkeley.edu

Checklist

- 1. Start audio
- 2. Enable webcam for participants and turn my webcam on.
- 3. Clear notes and chats, one by one, and clear them.
- 4. Check breakout rooms, make sure all cleared up.
- 5. Start recording before students come in, or have a reminder.
- 6. Have student be a recording reminder.
- 7. Merge students audio.
- 8. While people come in, drag them to their breakout room.

Uri Schonfeld
shuri@berkeley.edu

Who are we?

Uri Schonfeld
shuri@berkeley.edu

datascience@berkeley

Who are You

- Background.
- Computer science / programming background.
- Work.
- Education.
- Fun fact about you?

Uri Schonfeld
shuri@berkeley.edu

Takeaways

- Diverse starting points.
- Diverse experience.
- Whatever you're missing (linux, ec2, …) you'll need to work harder to catch up.
- Those who know will benefit from helping others.
- You'll get a lot by getting to know each other.

Uri Schonfeld
shuri@berkeley.edu

Who am I?

- Uri Schonfeld
- shuri@berkeley.edu

Course Overview & Intro

Jari Koister

Uri Schonfeld
shuri@berkeley.edu

datascience@berkeley

Admin

- Syllabus: https://github.com/UC-Berkeley-I-School/w205-fall-16-labs-exercises/tree/master/Course-Information
- Linux basics:
  - https://drive.google.com/file/d/0B-ddPrjoCXsFSEVQY3dIcEhmMkk/view?usp=sharing

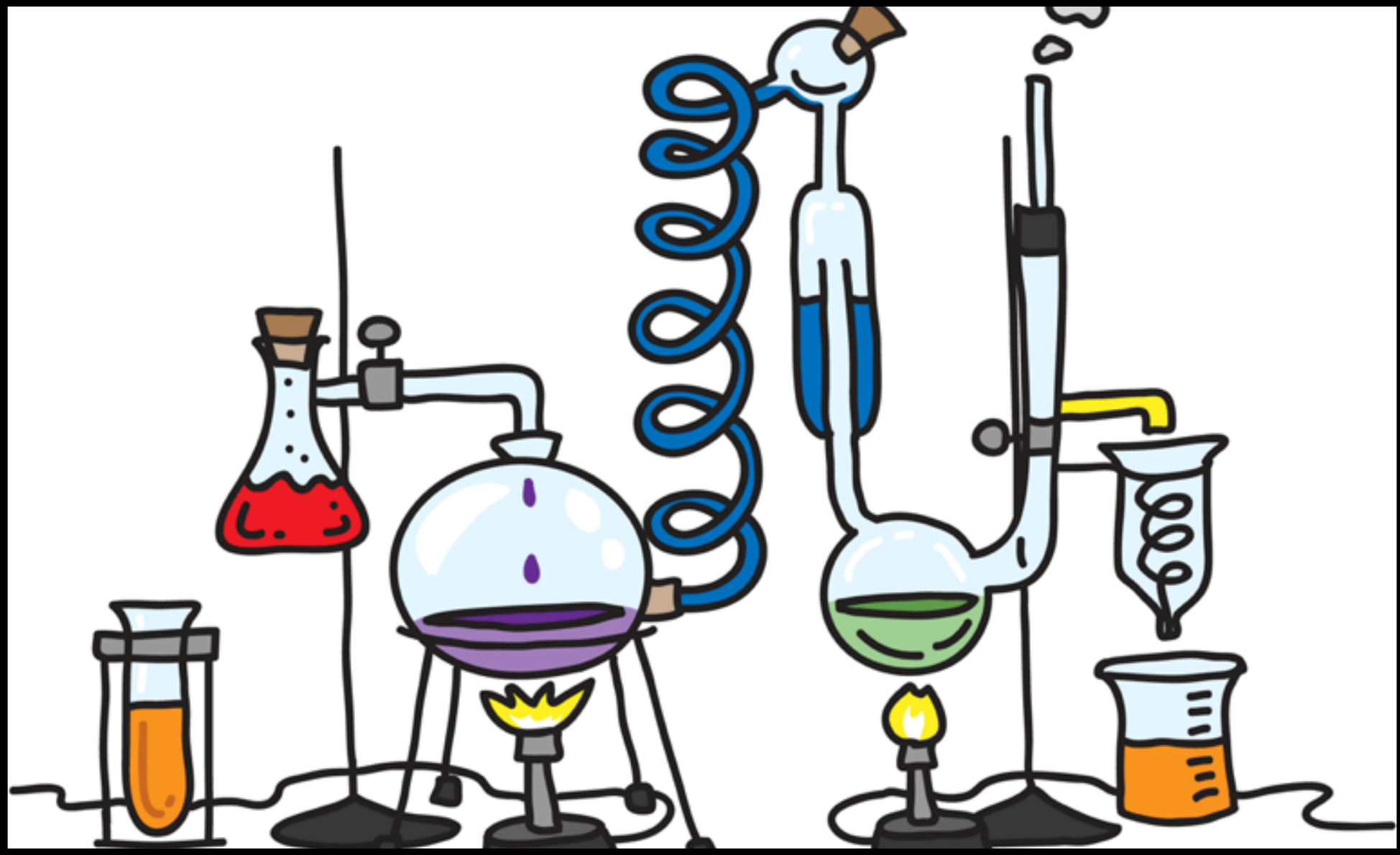
- Apply early for the educational credit, it may take a day or two (maybe more if the whole US is applying at once.
- Slack: Can be helpful, check it out, I'm on email or the section wall.

Clarification

- Network latency is the term used to indicate any kind of delay that happens in data communication over a network. Network connections in which small delays occur are called low-latency networks whereas network connections which suffers from long delays are called high-latency networks.

- What is Network Latency? - Definition from Techopedia

- https://www.techopedia.com/definition/8553/network-latency



First Lab

Uri Schonfeld
shuri@berkeley.edu

Lab 1

- Deadlines: Lab 1 due Tuesday September 6th Midnight.
- Note that you should use U.S East (Virginia) or you won't find the right AMI and things won't work.
- Excellent AWS Walk through by Arash (no need to copy linked to in Lab 1)
  - https://drive.google.com/file/d/0B6706xGNaPPyNEZHTkR5R19xcjA/view
- Note you have to use US East (Virginia)

Lab 1: Extra Recommendation

- This is a very simple lab.
- If you are unfamiliar with AWS, ssh, Linux, Bash use this week to take close the gap.
- Linux commands I recommend to know:
    - ls, find, mkdir, cd, rmdir, cp, rm, mv, cat, grep, less, more, sort, chmod, pwd, ps, (kill), date
    - screen: will allow you to keep your bash running when disconnected.
        - A command line editor: vim, vi, emacs, pico,… (I love vim; type ‘:q’ to quit :)
        - How to pipe commands (ls | sort | grep “^[a-z]”)
        - redirect output (ls > list_of_files.txt ; cat list_of_files.txt)
        - using simple variables(e.g. ENV=“staging”; DATESTAMP=`date`)
- So for example:
    - Create a bash script that accepts two arguments: a directory and a suffix.
    - Finds all the files in that directory that end with suffix
    - dumps them (redirect) to a temp file in ‘/tmp/‘ named according to the date (YYYY.mm.dd.HH.MM.SS.tmp)

Uri Schonfeld
shuri@berkeley.edu

Lab 1: Extra Recommendation

- Make sure you understand:
  - What is AWS?
  - What is an EC2 Instance? Does it always have memory? CPU? Does it have to have a “hard disk”? Why would it have to have more?
  - Spot instance vs reserved instance vs dedicated instance vs regular instance vs EBS backed instance?
  - If it crashes, will it still hold your files? What if you reboot? Does it depend on anything?
  - What’s an EBS? How is it different from S3? Could you use either to store files? Do you need a a file system on top of EBS? on top of S3?

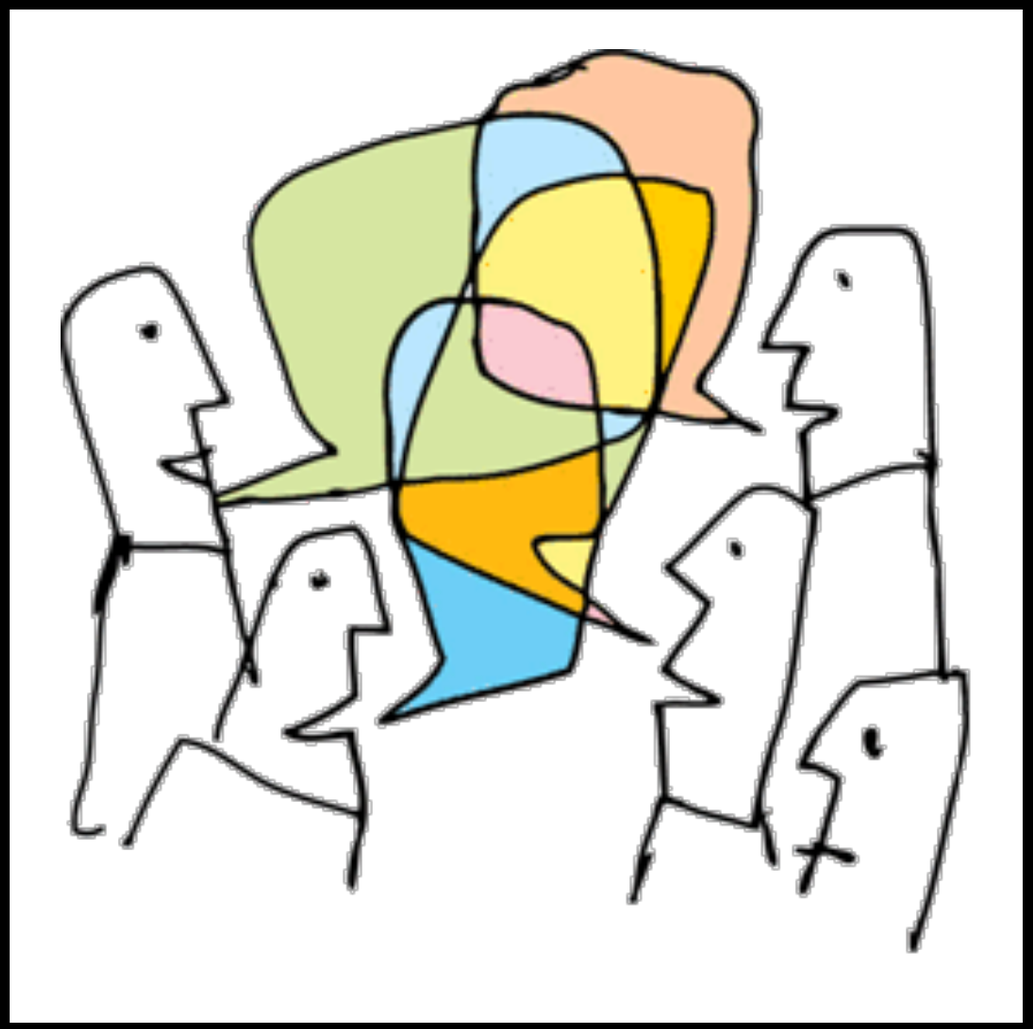
Uri Schonfeld
shuri@berkeley.edu

Lab 1: Extra Recommendation

- Cool thing about AWS?
  - you can always terminate and start over
- Get stuck? try again.
- Get stuck? Google your error message.
- Get stuck? reach out for help.
- All else fails? office hours
- Can’t wait? email me and I’ll try to help.

Uri Schonfeld
shuri@berkeley.edu

What is Data Science? (discussion)



Uri Schonfeld
shuri@berkeley.edu

Or Machine Learning

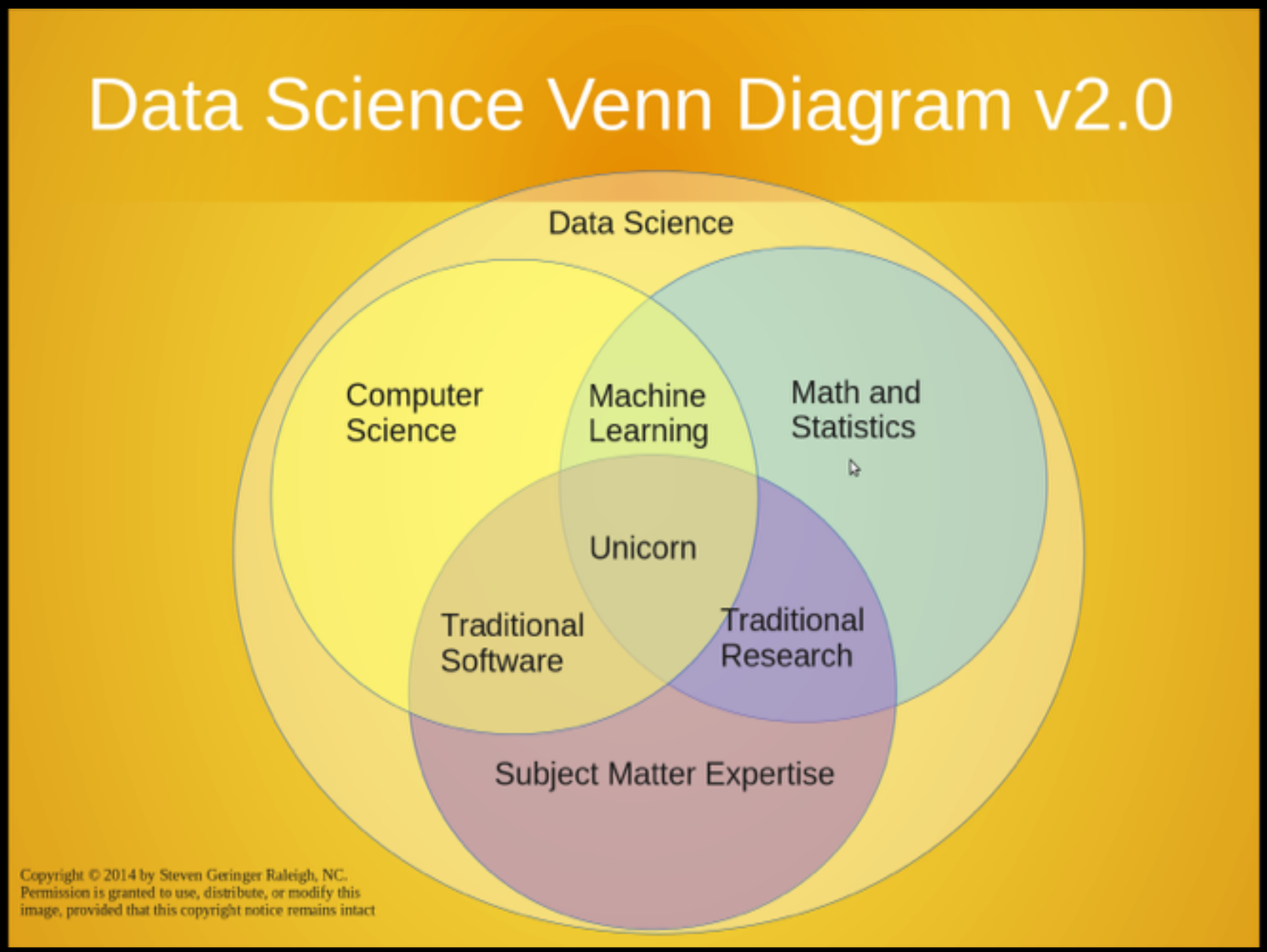
- Pattern Recognition?
- Statistical learning?

Uri Schonfeld
shuri@berkeley.edu

How about Data Mining…?

Uri Schonfeld
shuri@berkeley.edu

What is Data Science?



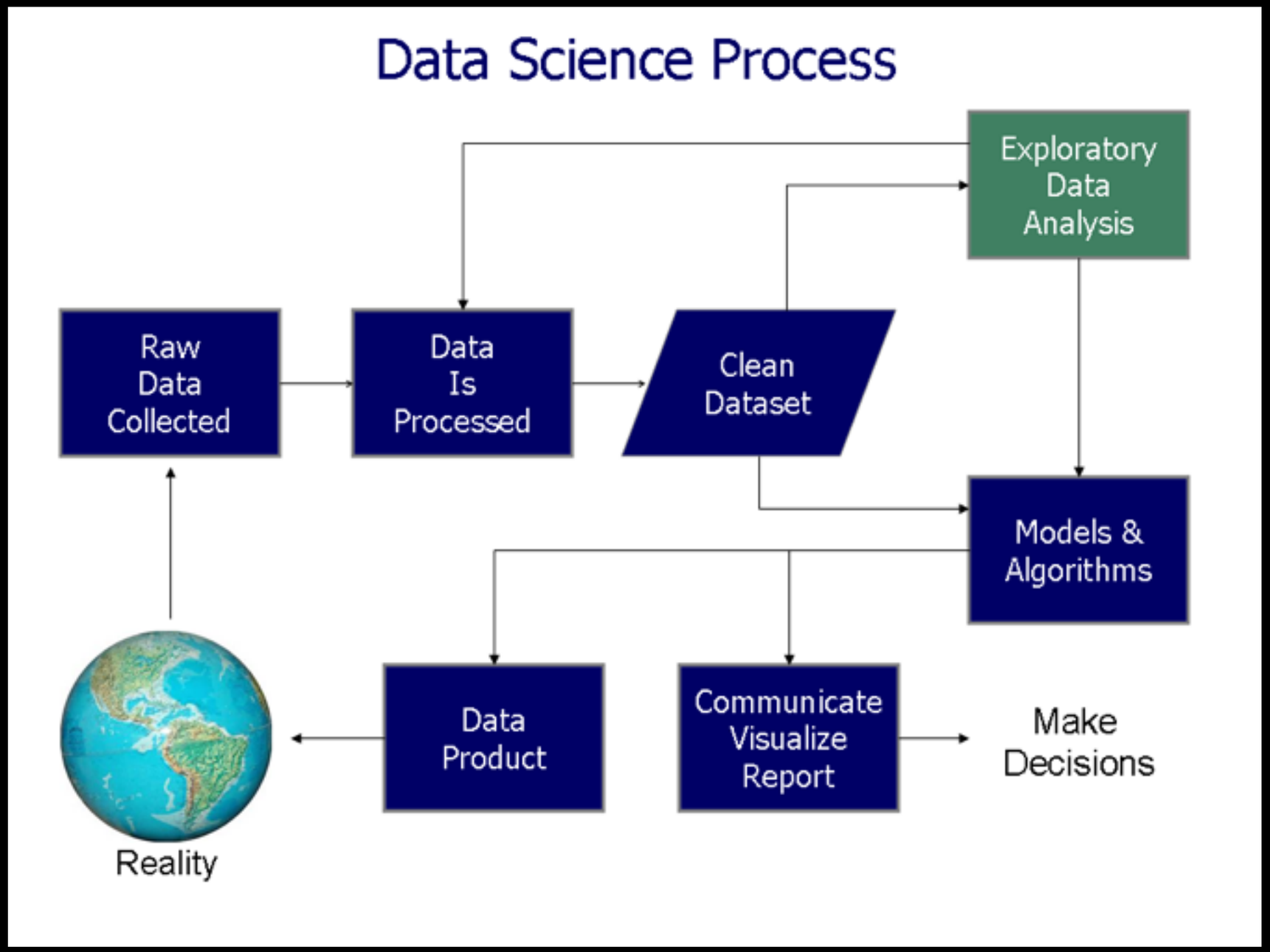
Uri Schonfeld
shuri@berkeley.edu

What is Data Science

- https://hail-data.quora.com/What-is-Data-Science
- Data Science is the practice of:
- **Asking questions** (formulating hypothesis), answers to which solve known problems or unearth unknown solutions that in turn drive business value,
- **Defining the data needed** or working with an existing data set and employing tools (computer science based) to collect, store and explore such data generally in huge volume & variety (probably more than 1 TB and 1000s of dimensions) ,
- **Identifying the type of analysis to be done** to get to the answers and performing such analysis by implementing various algorithms/tools (statistics based) in a distributed and parallel architecture,
- Communicating the insights gathered from the analysis in the form of simple stories/visualizations/ dashboards (the Data Product) that a non-data scientist can understand and build conversation out of it. (It should be kept in mind that a product can also be an piece of code that is internal to a company and is used by various departments. The presentation, maintenance, scalability, etc of the code are then the product features, which is often not practiced in many organizations)
- Building a higher level abstraction that does steps 2-4 in an autonomous way, predicting & taking actions on new data as they are fed to the system.

Uri Schonfeld
shuri@berkeley.edu

What is Data Science



Uri Schonfeld
shuri@berkeley.edu

What is Data Science?

- (Nate) Silver (statistician, writer, entrepreneur, Time's 100 most influential 2009,…) replied, "I think data-scientist is a **sexed up term for a statistician**", the reaction from the audience was for most, one of instantaneous laughter and applause. "Statistics is a branch of science. **Data scientist is slightly redundant** in some way and people shouldn't berate the term statistician."

Uri Schonfeld
shuri@berkeley.edu

So why storage and Retrieval?

Uri Schonfeld
shuri@berkeley.edu

So why storage and Retrieval?

- Complicated.
- Requires thinking ahead.
- Cool… Let someone else store and retrieve.
- Possible but cumbersome.

Uri Schonfeld
shuri@berkeley.edu

Course Overview

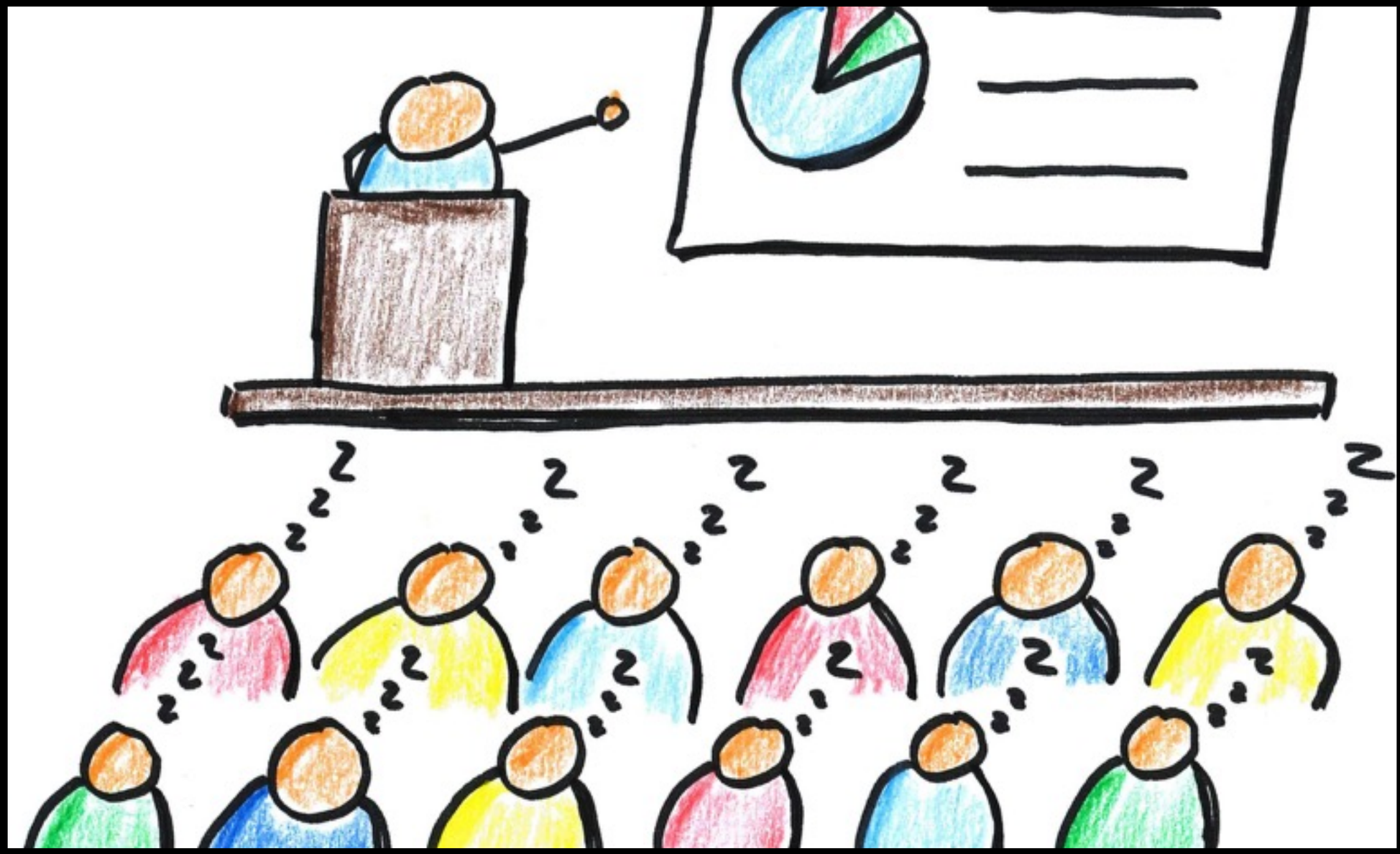
- Syllabus.
- Structure.
- Evaluation, requirements.

Uri Schonfeld
shuri@berkeley.edu

This week's Topics

- Data Driven Organization.
- Reference architecture.
- Dimensions Data.
- Dimensions Processing.

Uri Schonfeld
shuri@berkeley.edu



Review of Lectures

Evaluating Data and Processing Needs

Uri Schonfeld
shuri@berkeley.edu

datascience@berkeley

Data Driven Organization

- How Do DDOs Do It?
  - They collect data.
  - They develop intuition of the data.
  - They pose questions to answer or search the data for new insights.
  - They run experiments.
  - They make decisions and draw insights.

Uri Schonfeld
shuri@berkeley.edu

Reference Model: Processing Dimensions

- Query selectivity: (example numbers) how much of the data you expect to process.
    - High (<20% selected)
    - Low(>80% data)
- Query execution time: expected query response time
    - Short (< hour)
    - Long (< 24 hours)
- Aggregation:
    - Simple (counters)
    - Advanced: Roll-ups, Drill downs
- Processing time:
    - Short (< hour)
    - Long (< 24 hours)
- Joins: None, basic, advanced
- Precision: Exact, Approximate, Lossy

Uri Schonfeld
shuri@berkeley.edu

Reference Model: Data Dimensions

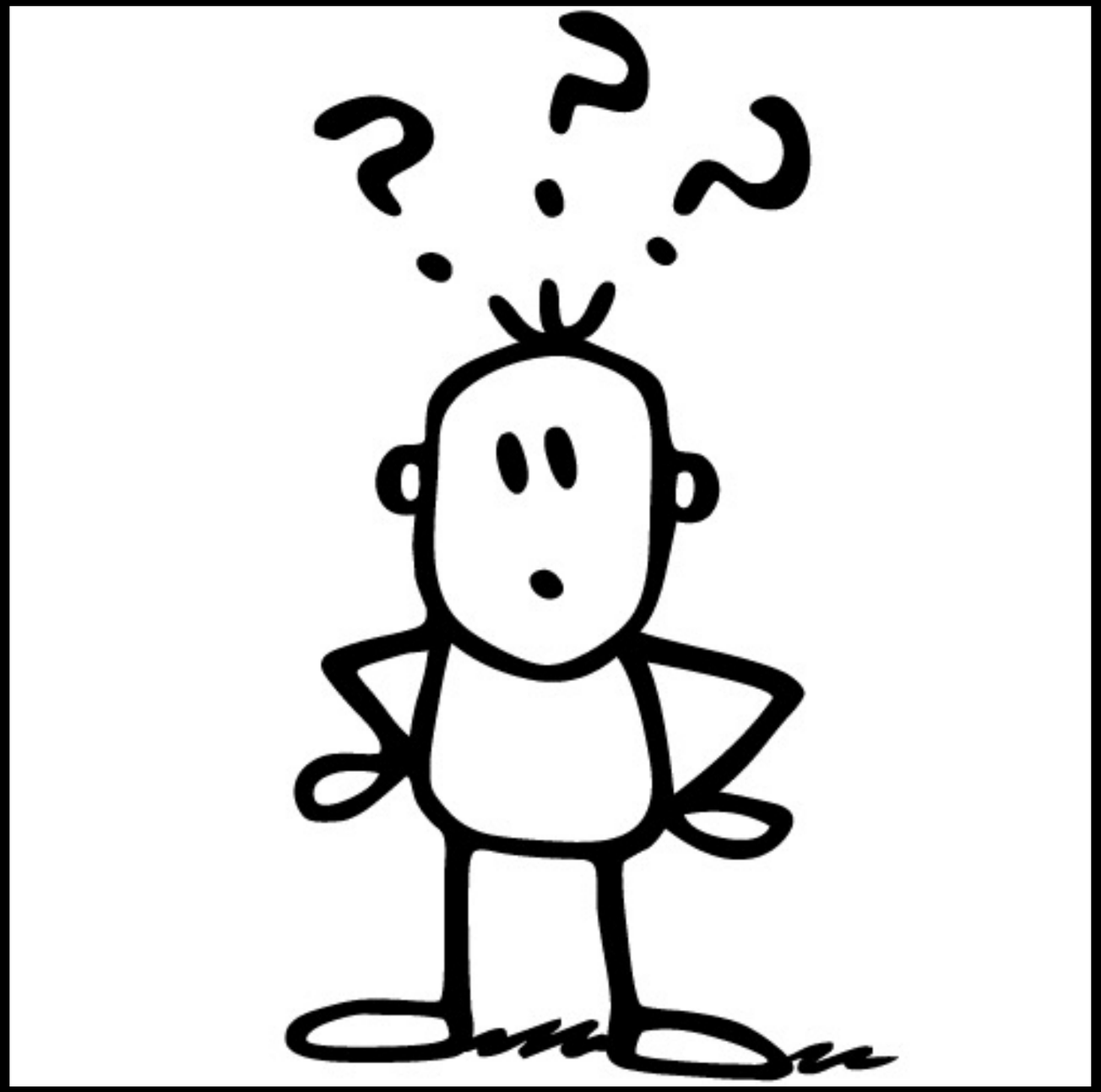
- Structure, Unstructured, Semi-Structures
- Size: Megabytes, Giga, Tera, 100s if Tera, Peta
- Sink latency: velocity of the data as they arrive
    - Very high: > 100s updates/second
    - Low: daily
- Source latency: how quickly data are reflected in the indexing layer
    - High: real time:
    - Low: daily
- Quality: ability to deal with bad or low quality data
    - High: can compensate and handle in an automated fashion
    - Low: can not handle bad or low quality data
- Completeness Requirement: how well the system deals with incomplete data
    - Incomplete: Does not require data to be complete
    - Complete: Does.

Uri Schonfeld
shuri@berkeley.edu

Reference Model: Processing Dimensions

- This is a starting point, not an end-all solution.
- Some specifics may be confusing or unfamiliar to others.
- Think and develop your own on a case-by-case basis.

Uri Schonfeld
shuri@berkeley.edu



Questions

Uri Schonfeld
shuri@berkeley.edu