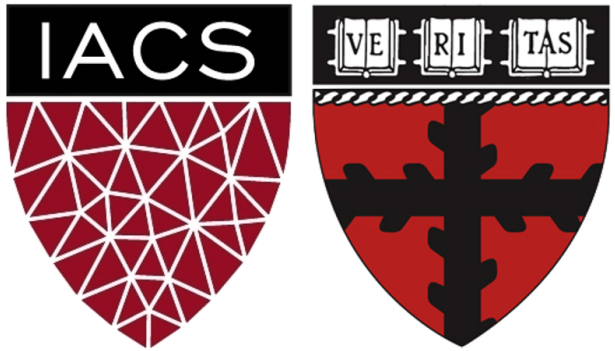
# Lecture 1: Introduction

**AC295**

Advanced Practical Data Science

Pavlos Protopapas



# Outline

1 : Why you should take this class and why not

2: Who are we

3: Course structure and activities

4: Expectations

5: Workload

6: Logistics

7: Grades

# Why you should take this class

**Because you want to learn how to**:

- Put your model in production
- Integrate and orchestrate applications
- Deploy increasing amount of data
- Take advantage of available models
- Evaluate and debug model using visualization

If you have attended **ComputeFest** and found the topics interesting this class will also be interesting

# Why you shouldn’t take this class

You are **not** familiar with most of the concepts covered in CS109A/B

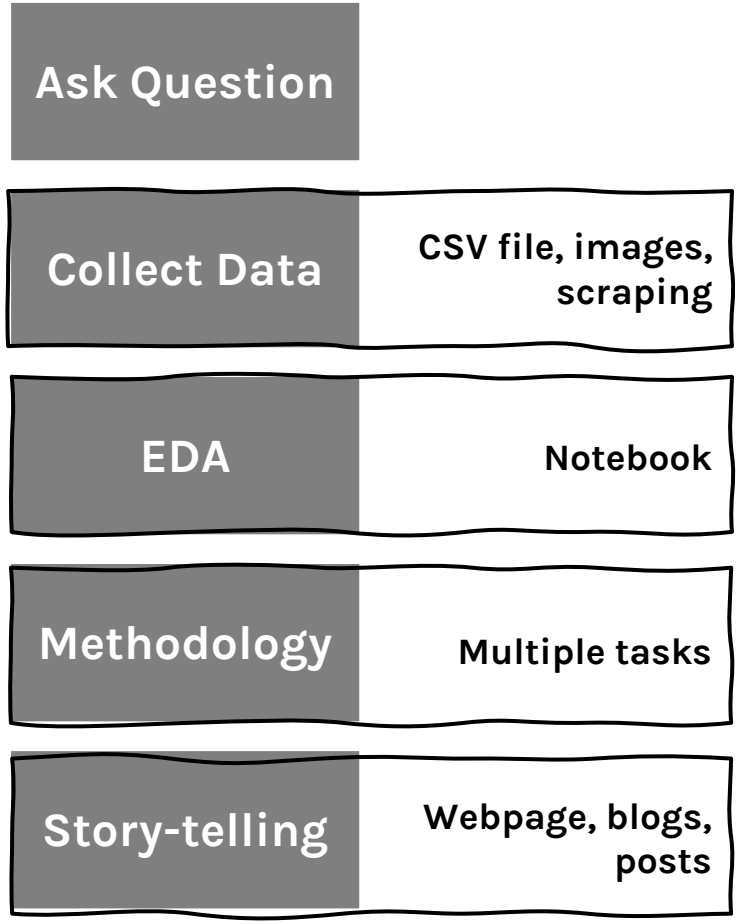
**For example:**

- Basic Machine Learning
- CNNs, RNNs, Autoencoders, GANs, etc
- Basic linux commands

**Remember**, this course will be offered again in the fall!

# Data Science Series to Real World

## Data Science Series 109A/B




## Real World

Manage larger database

Learn packages to process larger amount of data

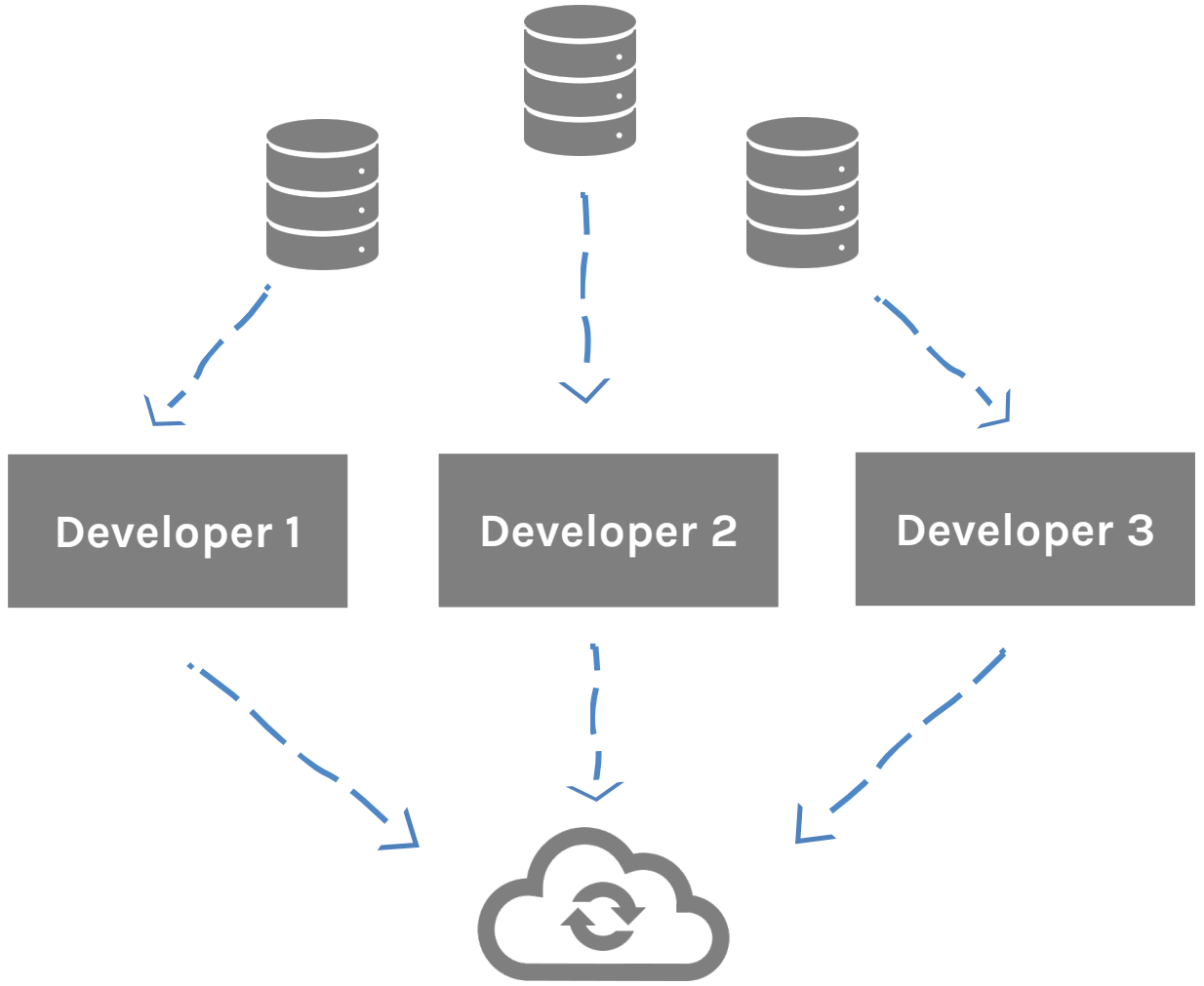
Handle complex team dynamics and orchestrate applications

# Data Science Series to Real World (cont)

Fragmented database

Multitude requirements and applications

Recombine and deploy

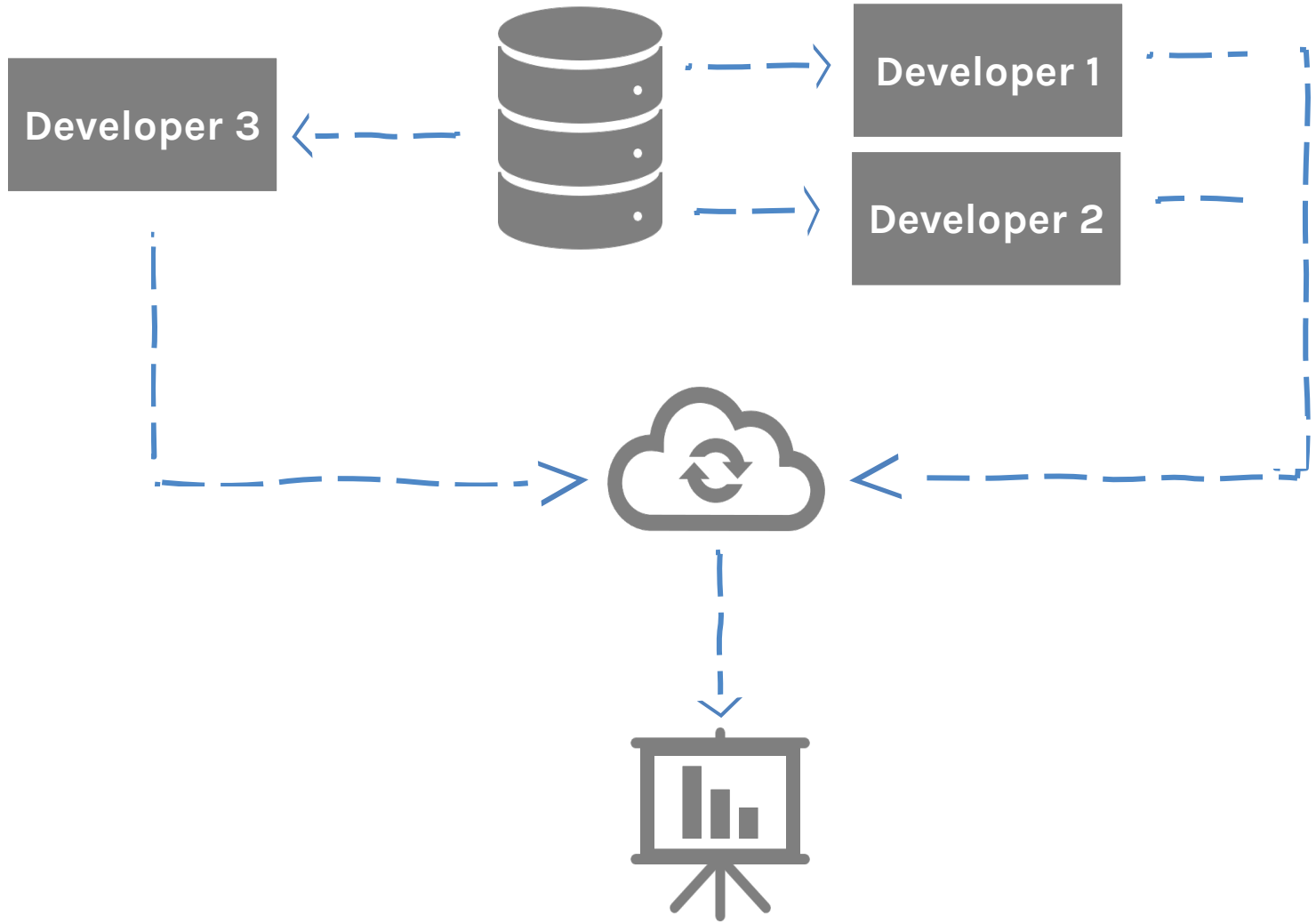


# Data Science Series to Real World (cont)

Multiple tasks or models (i.e. Ensemble)

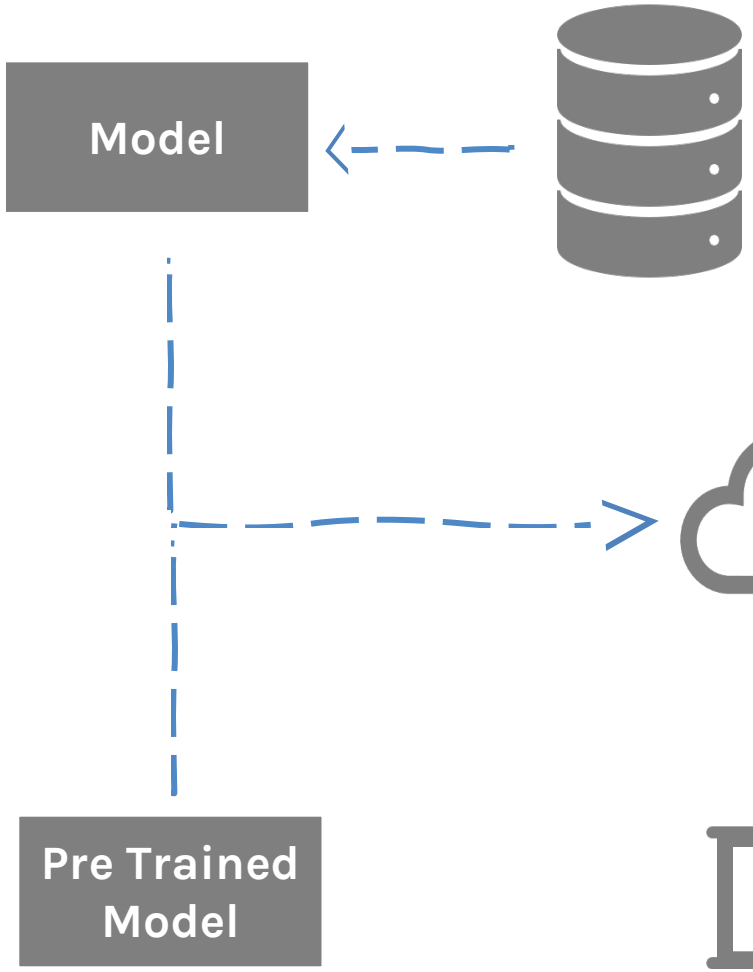
Recombine results

Present results



# Data Science Series to Real World (cont)

Model too expensive to train
Or not enough training data




Use pre-trained model

Final Results

Present results

# Who?

## Pavlos Protopapas

Teaches CS109(a/b), the data science capstone course, and AC295 (advanced practical data science). Research in astrostatistics: machine learning, statistical learning, big data for astronomical problems.

He has picked some new hobbies besides 109s and **eating**:

Going to BSO (see you there), cross country ski (completed Engadin skimarathon), cheese making and being a TikToker (check me out @pavlosprotopapas)

# Who? (cont)

**Michael S. Emanuel**

After 17 years in finance, mainly fixed income portfolio management, Michael started a second career and is completing the Masters of Data Science program at Harvard. He is a father of two small children who occasionally crash IACS events and enjoys distance running and classical music.

# Who? (cont)

## Andrea Porelli

Urban planner turned into data hacker. He likes to break things just for the sake of putting them back together (most of the time). Committed to apply Data Science to change something. So far, he managed to change himself the most –thanks IACS- and look forward to pass it over.

# Who? (cont)

## Giulia Zerbini

Data Designer. Creative technologist at *The Visual Agency* in Milan, MA Graduate at Politecnico di Milano. Designing and developing visualizations and interfaces based on data. Passionate about using visualizations for discovering patterns in data and communicating information in intuitive terms to a broad audience.

# Course Structure and Activities

**Modules:**

1. Deploy data science (integration + scalability)
2. Transfer learning and distillation
3. Visualization as investigative tool

**Activities:**

lectures, reading discussions, exercises, quizzes, practicums, projects

**Lectures:** Tuesday and Thursday 4:30-5:45 pm in Cruft 309

**Office Hours:** TBD

# Topics

Deploy data science (integration + scalability)

A. Virtual Environments, Virtual Boxes, and Containers

B. Kubernetes

C. Dask

# Topics (cont)

Transfer learning and distillation

A. Basic Transfer Learning and SOTA Models

B. Transfer Learning across Tasks

C. Distillation and Compression

# Topics (cont)

Visualization as investigative tool

A. Introduction and Overview of Viz for Deep Models

B. Convolutional Neural Networks for Image Data

C. Recurrent Neural Networks for Text Data

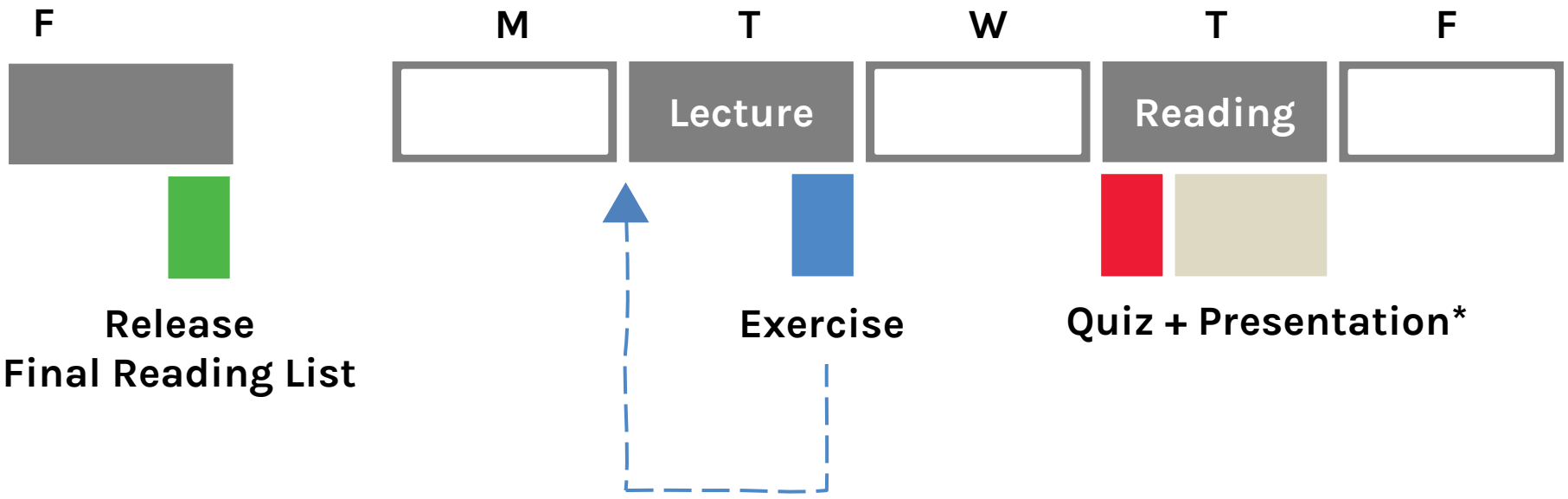
# Calendar

> *Link to Calendar* <

| Week | Date | Lecture # | Topics |
|---|---|---|---|
| 1 | 1/28 | 1 | Introduction |
| | 1/30 | 2 | Set up |
| 2 | 2/4 | 3 | Virtual Environments, Virtual Boxes, and Containers |
| | 2/6 | 4 | Journal Discussion |
| 3 | 2/11 | 5 | Kubernetes |
| | 2/13 | 6 | Journal Discussion |
| 4 | 2/18 | 7 | Dask |
| | 2/20 | 8 | Journal Discussion |
| 5 | 2/25 | 9 | Practicum |
| | 2/27 | 10 | Practicum |
| 6 | 3/3 | 11 | Intro to Transfer Learning: Basic Transfer Learning and SOTA Models |
| | 3/5 | 12 | Journal Discussion |
| 7 | 3/10 | 13 | Transfer Learning across Tasks |
| | 3/12 | 14 | Journal Discussion |
| | 3/17 | | Spring Recess |
| | 3/19 | | |
| 8 | 3/24 | 15 | Distillation and Compression |
| | 3/26 | 16 | Journal Discussion |
| 9 | 3/31 | 17 | Practicum |
| | 4/2 | 18 | Practicum |
| 10 | 4/7 | 19 | Introduction and Overview of Viz for Deep Models |
| | 4/9 | 20 | Journal Discussion |
| 11 | 4/14 | 21 | Convolutional Neural Networks for Image Data |
| | 4/16 | 22 | Journal Discussion |
| 12 | 4/21 | 23 | Recurrent Neural Networks for Text Data |
| | 4/23 | 24 | Journal Discussion |
| 13 | 4/28 | 25 | Practicum |
| | 4/30 | 26 | Practicum |
| 14 | 5/5 | 27 | Project |
| | 5/7 | 28 | Project |
| 15 | 5/12 | 29 | Project |
| | 5/14 | 30 | Project |

# Course Structure and Activities

## Regular week schedule



# Workload

## Regular Week

3 hours in class
3 hours reading
2 hours exercise
4 hours presentation*

~ 12 hours/week

* 1 presentation per module per group (3 total)

## Practicum and Project Week

~ 15 hours/week**

** 3 practicums and 1 final project (2 weeks long)

We will be asking for your feedback on the workload

# Expectations

## How to read and present class material

> *Link to Reading Guidelines* <

> *Link to Presentation Guidelines* <

# Logistics

## Fill up forms

Make group *

Sign-up presentation**

* Fill group components in each row

** Each group should pick one slot (white background) in each module

# Grades

| Assignment | Final Grade Weight |
|---|---|
| Quizzes | 9% |
| Exercises | 9% |
| Presentations | 15% |
| Practicums | 45% |
| Projects | 20% |
| Participation | 2% |
| **Total** | **100%** |

# Final Details

- We will be using ED for discussions, announcements and quizzes.
- Submissions for exercises, reports, presentations etc we will be using github  (details soon).

This is the first time we are offering the course, so your feedback will be vital in tuning it this year and improving it for future years.

However, we are making every effort to have a well organized course and we promise you an exciting semester full of learning!

# THANK YOU

**AC295** **Advanced Practical Data Science**

**Pavlos Protopapas**