

CS 8803-MDS

# Human-in-the-loop Data Analytics

---

Lecture 1

08/21/23

Kexin Rong

# Today's Class

The essentials

What is this class

Overview of course topics

Getting to know you

# The essentials

Instructor: Kexin Rong

Office: Klaus 3322

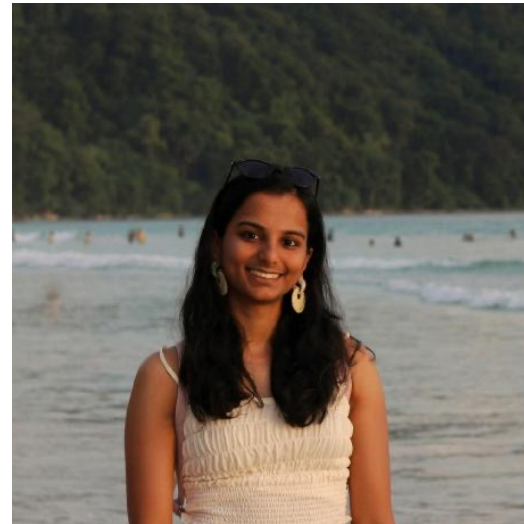
Email: [krong@gatech.edu](mailto:krong@gatech.edu)

TA: Ashmita Raju

Email: [ashmita.raju@gatech.edu](mailto:ashmita.raju@gatech.edu)

TA: Saumia Singhal

Email: [ssinghal79@gatech.edu](mailto:ssinghal79@gatech.edu)



# The essentials

Course website: <https://kexinrong.github.io/fa23-cs8803/>  
schedule, assignment, and course material

Canvas: submitting assignments

Piazza: discussing course contents

Email: meeting/extension requests  
mention CS8803 in the email title

OH: Thursday/Friday, time TBD  
<https://forms.gle/RrPBVLJbMQfYG7Jn6>  
Also available by appointment

# Course Learning Objectives



Learn about a research area: Data Management

Subarea: human-in-the-loop data analytics

Important in modern data-driven world/data-science

Primarily through student-led paper discussions

Some overview lectures and tutorials

Get hands-on research experience

Critically read and evaluate papers

Technical presentation

Conduct novel research

# Grading

Participation:	45%
Paper reviews	15%
Presentation	20%
Participation	10%
Assignment:	10%
Project:	45%
Project Proposal	4%
Progress Report	6%
Draft Paper	5%
Final Presentation	10%
Final Paper	25%



## DISCLAIMER

Second offering  
Beware of hiccups

# Paper Reviews

15% of grade (lightly graded)

Submit **midnight before class**. Not late submissions accepted

Need to submit at least **10 reviews** over the semester

At least **one review from each topic** (4 topics in total)

The review should cover the following key questions:

- What problem is the paper trying to solve?

- Why is the problem important?

- What sets it apart from prior work?

- What are the key technical ideas?

- What are the main areas of improvements and open questions?

# Participation

10% of grade

Goals: assess understanding, get feedback, make the class more fun

Any participation is good participation! Not necessary that

- you ask “good” questions
- answer questions “correctly”
- you attend and are super engaged every class

We expect you to attend most classes ( $\geq 80\%$ )

- We'll take attendance for  $N$  randomly-selected classes. You get penalized if you miss more than  $M$  attendance-taking classes. Tentative:  $N = 10, M = 2$



# Class Format: Role-Playing Paper-Reading Seminars



Adapted from Alec Jacobson and Colin Raffel: <https://colinraffel.com/blog/role-playing-seminar.html>

# Presentation

20% of grade

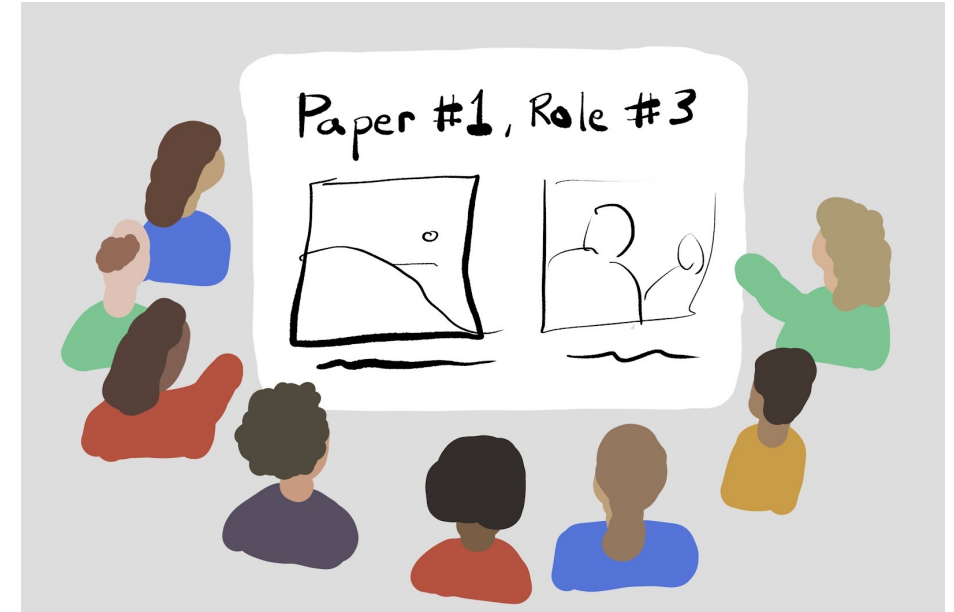
Before class

read paper and submit reviews

During class

(1-2 **paper author roles**) 15-20 min presentation

(4-5 **accessory roles**) 5 min presentation each



# Role: Paper Author



15-20 minutes (~1 slide/minute)

~15 min if one presenter

~20 min if two presenters (work together)

Imagine you are the author of the paper who is presenting your work at a conference. In your talk, you should probably address the following:

Why should people care about your work?

What are the key technical challenges and solutions?

How did you evaluate your hypothesis?

What are the main takeaways?

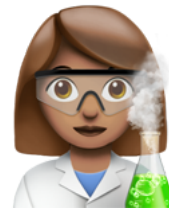
Can reference authors' slides, but **don't use without modification!**

# Accessory roles \* 4

5min: always start with a *one-slide summary* of the paper

Roles available

- academic researcher
- archaeologist
- industry practitioner
- hacker (takes 10 min)



# Role: Researcher



You're a researcher who is working on a new project in this area.

Propose an **imaginary follow-up project** *not just* based on the current paper but only possible due to the existence and success of the current paper.

Could be the start of your own project =)



# Role: Archaeologist

This paper was found buried under ground in the desert. You're an archeologist who must determine where this paper sits in the context of previous and subsequent work.

**Prior work archaeologist:** survey 2-3 important papers that substantially influences the current paper.

**Follow-up work archeologist:** survey 2-3 *newer* papers that are heavily influenced by this current paper.

Helpful tools: <https://www.connectedpapers.com/>

# Role: Industry Practitioner



You work at a company or organization developing an application or product of your choice.

Bring a convincing pitch for [how the method/system in the paper could fit into products and workflows at your company](#), and discuss at least one positive and negative impact of this application.

# Role: Hacker



You're a hacker who needs a demo of this paper ASAP.

Implement a [small part or simplified version](#) of the paper on a small dataset or toy problem. Prepare to share the core code (e.g., a Colab notebook) of the algorithm to the class and demo your implementation.

Do not simply download and run an existing implementation – though you are welcome to use (and give credit to) an existing implementation for “backbone” code.

Could be the starting point of your own project / baseline.



# Sign up for presentation <https://tinyurl.com/2s3amhrv>

## Credit system

every 5min presentation = 1 credit = 4% of grade

paper author role: 2~3 credits

accessory roles: 1 credit (hacker gets 2 credits)

## Rules

need  $\geq 5$  credits over the semester

need to be in the paper author role at least once

presenters for the [first three papers](#) get 1 extra credit  
(maximum 1 extra credit per person)

## Examples

1 solo paper author role + hacker role + archaeologist role

1 shared paper author role in the first 3 paper + 2 accessory roles

# Sign up for presentation <https://tinyurl.com/2s3amhrv>

Don't modify other people's slots without asking

Slots are frozen **one week** prior to the actual presentation

1<sup>st</sup> presentation is 8/30, signup open till 8/28

If you have a scheduling change after the freeze, please find a classmate to sub in your slot

No show to presentation – penalty of 2 credits (e.g., you'd need  $5 + 2 = 7$  credits for the semester).

# Grading

Participation: 45%

Paper reviews 15%

Presentation 20%

Participation 10%

Assignment: 10%

Project: 45%

Project Proposal 4%

Progress Report 6%

Draft Paper 5%

Final Presentation 10%

Final Paper 25%

Any questions?

# Research Project

45% of grade

Main criteria

something “new” + relevant to course topics

Project milestones

Week 5: project proposal

Week 10: project update

Week 15: draft paper + peer review

Week 16: project presentation

Week 16: final report

# What we expect for research projects

Teams of 1-3 (subject to change depending on final class size).  
Expected work proportional to team size

Projects are evaluated based on "completeness", not on "interestingness" of ideas

- Is the problem well-defined and motivated?

- Is related work thorough?

- Does the evaluation test the proposed hypothesis?

- Is the writing overall clear and easy to follow for a technical expert in the field?

# Different flavors of project

Benchmarking/new datasets/user study

Show: new insights and understanding

Tool/system/interface

Show: easy of use, scalability, design novelty

Algorithm

Show: novelty, correctness, scalability

Reproduce and extend

Show: assumptions/contexts that have changed

# Example projects from last year

Workload-Aware Adaptive Sample Update for AQP (reproduce and extend)

- Update precomputed samples when the target query workload changes

Evaluation of Scatterplot Sampling Techniques for Exploratory Trend Analysis of Massive 2D Datasets (benchmark)

- How does sampling rate and sample methods affect user's perception of trends in scatter plots?

Interactive Compositional Querying of Video Data (tools)

- Interface for exploratory video analytics using domain hints (e.g., OCR)

SQL Q-Suggest: Context-Aware SQL Prediction and Auto-Completion with Q-Learning (algorithm)

- Prediction on future sessions of SQL queries

More on last year's course website: <https://kexinrong.github.io/fa22-cs8803/schedule/>

# A note on ChatGPT/LLMs

Yes we know you are going to use them.

Yes it is ok to use them.

... with some conditions.

- DO use them to polish & condense your writing.
- MAYBE use them to come up with the title of your project (after writing the report).
- DON'T use them to write about scientific ideas. They are not so great at generating new research ideas / criticizing existing ones.
- Please note on your writing if & how you used ChatGPT



# Grading

Participation: 45%

Paper reviews 15%

Presentation 20%

Participation 10%

Assignment: 10%

Project: 45%

Project Proposal 4%

Progress Report 6%

Draft Paper 5%

Final Presentation 10%

Final Paper 25%

Any questions?

# Grading

Participation: 45%

Paper reviews 15%

Presentation 20%

Participation 10%

Assignment: 10%

Project: 45%

Project Proposal 4%

Progress Report 6%

Draft Paper 5%

Final Presentation 10%

Final Paper 25%

WIP

Related to topic 2: data producer

Tentative timeline: 09/25-10/06

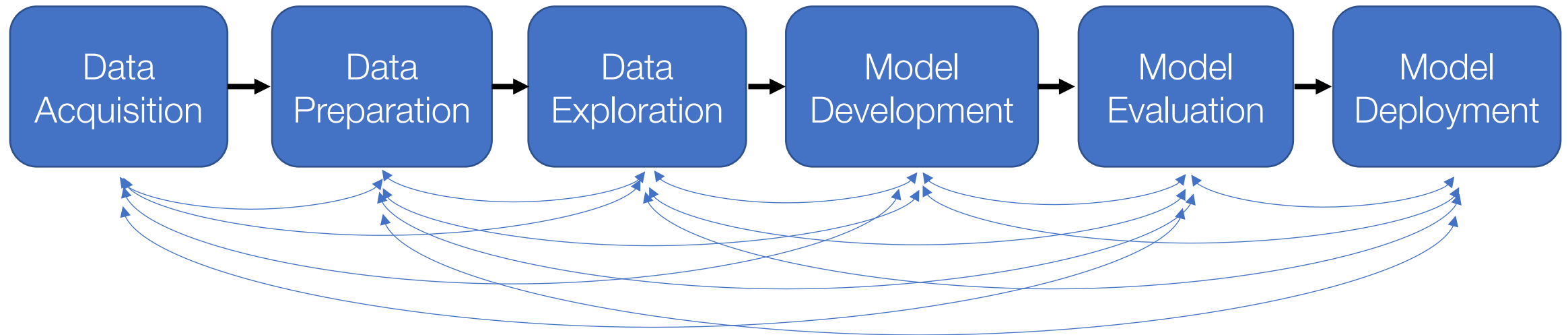
# Overview of Course Topics

Human-in-the-loop What is [human-in-the-loop](#)?

Data Analytics What is [data analytics](#)?

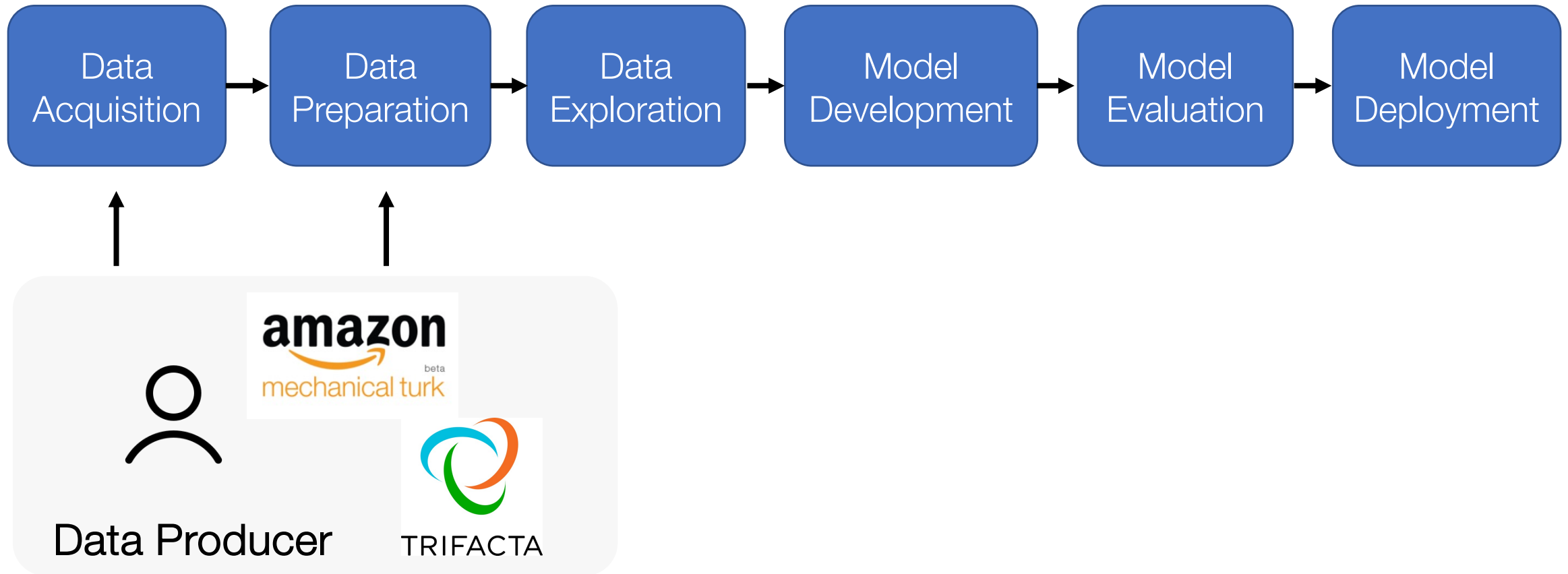
# The Data Analytics lifecycle in a bird's eye view

“Only a fraction of real-world ML systems is composed of ML code” [1]



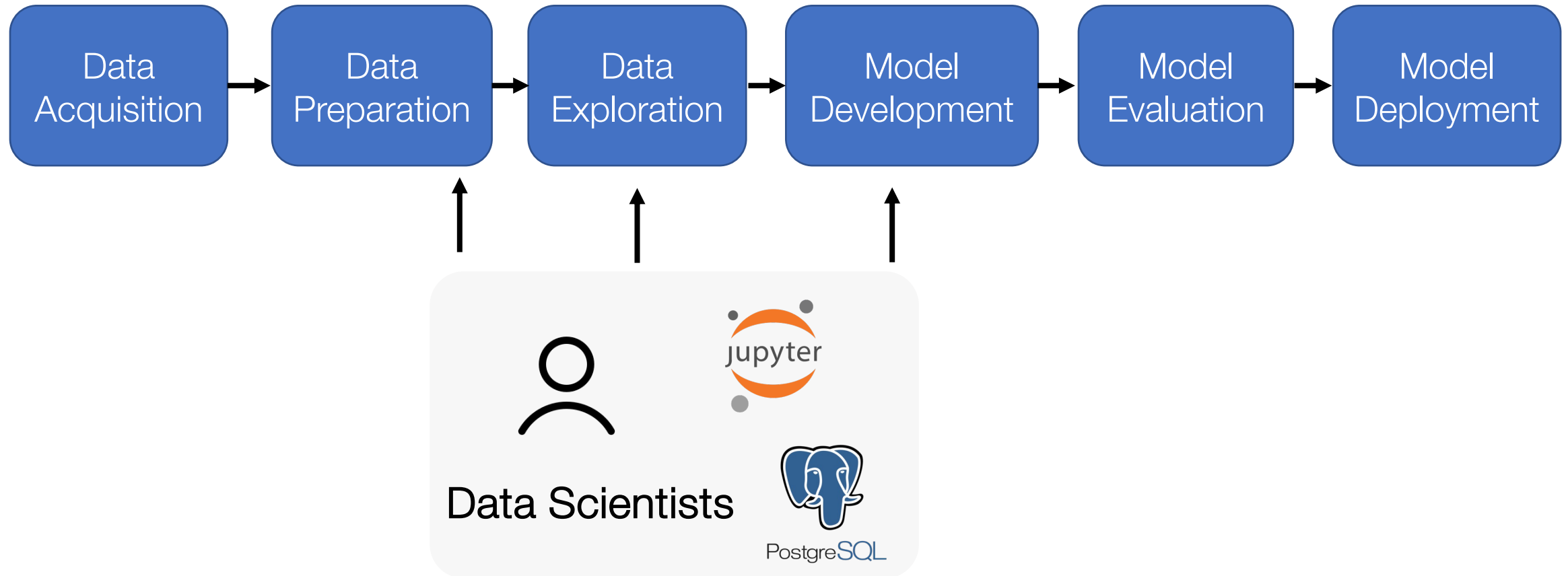
The machine learning lifecycle is complex and iterative process  
Humans play an important role in almost all steps of the lifecycle

# Human roles in data analytics



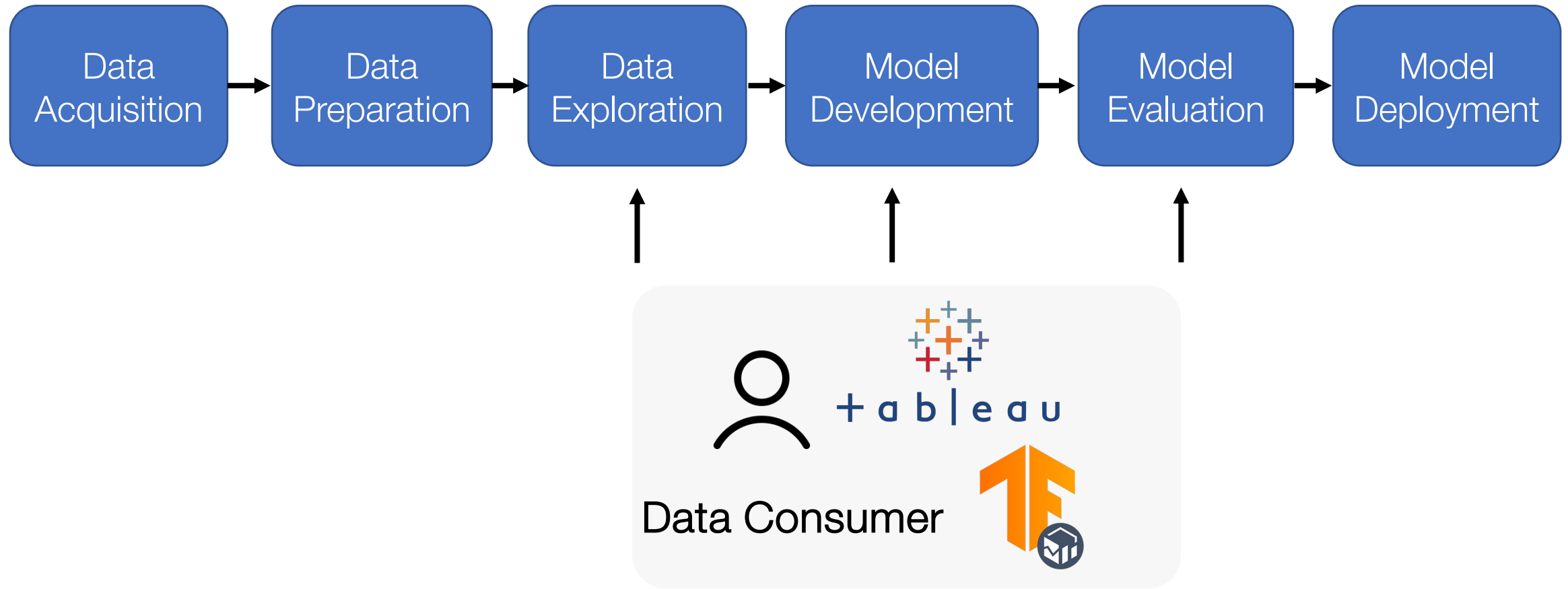
Humans play an important role in almost all steps of the lifecycle

# Human roles in data analytics



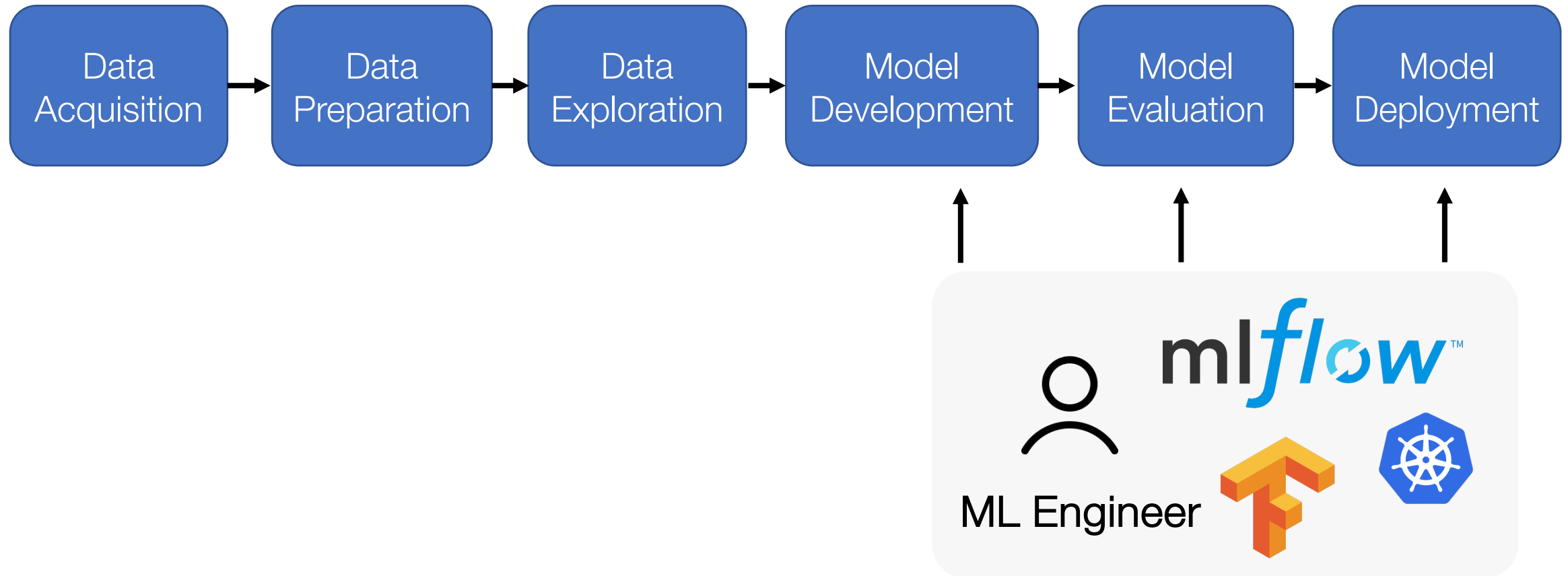
Humans play an important role in almost all steps of the lifecycle

# Human roles in data analytics



Humans play an important role in almost all steps of the lifecycle

# Human roles in data analytics



Humans play an important role in almost all steps of the lifecycle



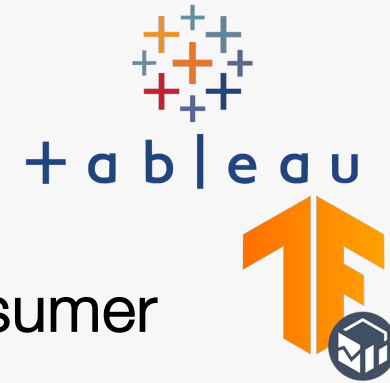
# Human roles in data analytics



Data Producer



Data Consumer



Data Scientists



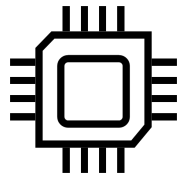
ML Engineer



# Humans are the main bottleneck in data analysis

## #1 They are expensive

With \$100, you can get



Compute: Up to 100 CPUs/day



People: 0.66 person/day



Yet the success of ML depends on them

**OpenAI has hired an army of contractors**  
to do what's called "data labeling"



# Humans are the main bottleneck in data analysis

## #2 They have limited attention span and working memory

How long will users wait for the computer to respond before they become annoyed?

**Table 1. Acceptance of Delay as a Function of Response Time<sup>a</sup>**

Standard response time (seconds)	Total number of trials	Average delay (seconds)	Standard deviation (seconds)	Trials attention key pressed (%)
2	11,634	1.98	0.53	1.42
4	9,754	3.50	2.08	17.44
8	10,103	2.27	5.63	82.92

← 8-sec was generally not acceptable

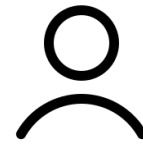
<sup>a</sup> Data from Williams [1973].

## The Magical Number Seven, Plus or Minus Two

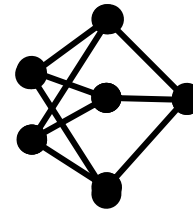
People can rapidly recognize approximately seven "chunks" of information at a time, and hold time in short-term memory for 15-30 seconds

# Humans are the main bottleneck in data analysis

#3 They speak a different language and it's not always easy to communicate that to machines



I see a player passes a soccer ball to another player on his team

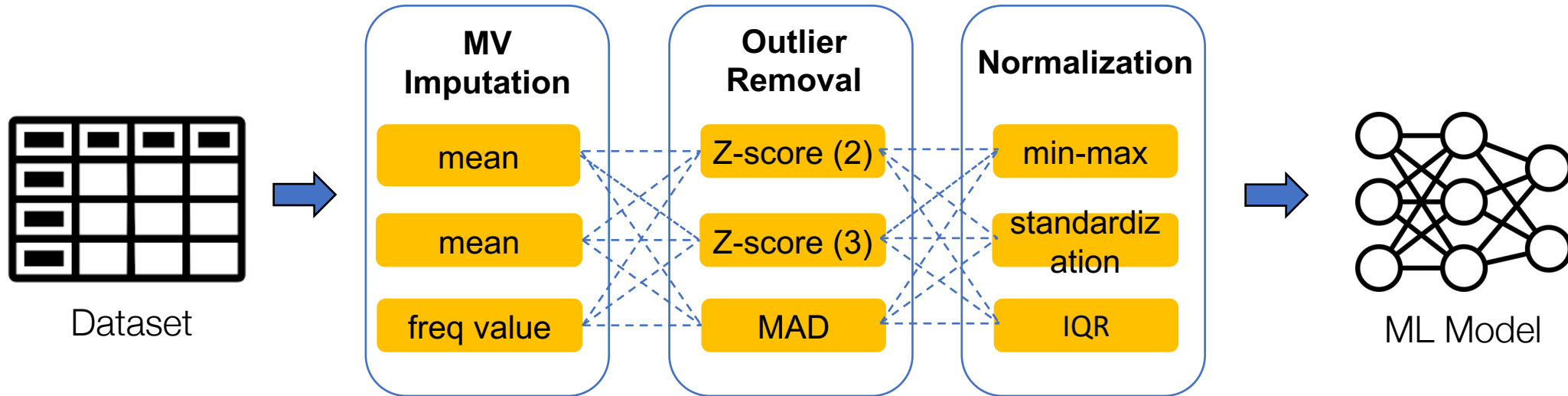



I see large boxes and small boxes moving around


There's often a mismatch between users' intentions to specify high-level objectives and the low-level primitives exposed by the tools

# Humans are the main bottleneck in data analysis

#4 They are bad at solving complex problems and often go for ad-hoc heuristics



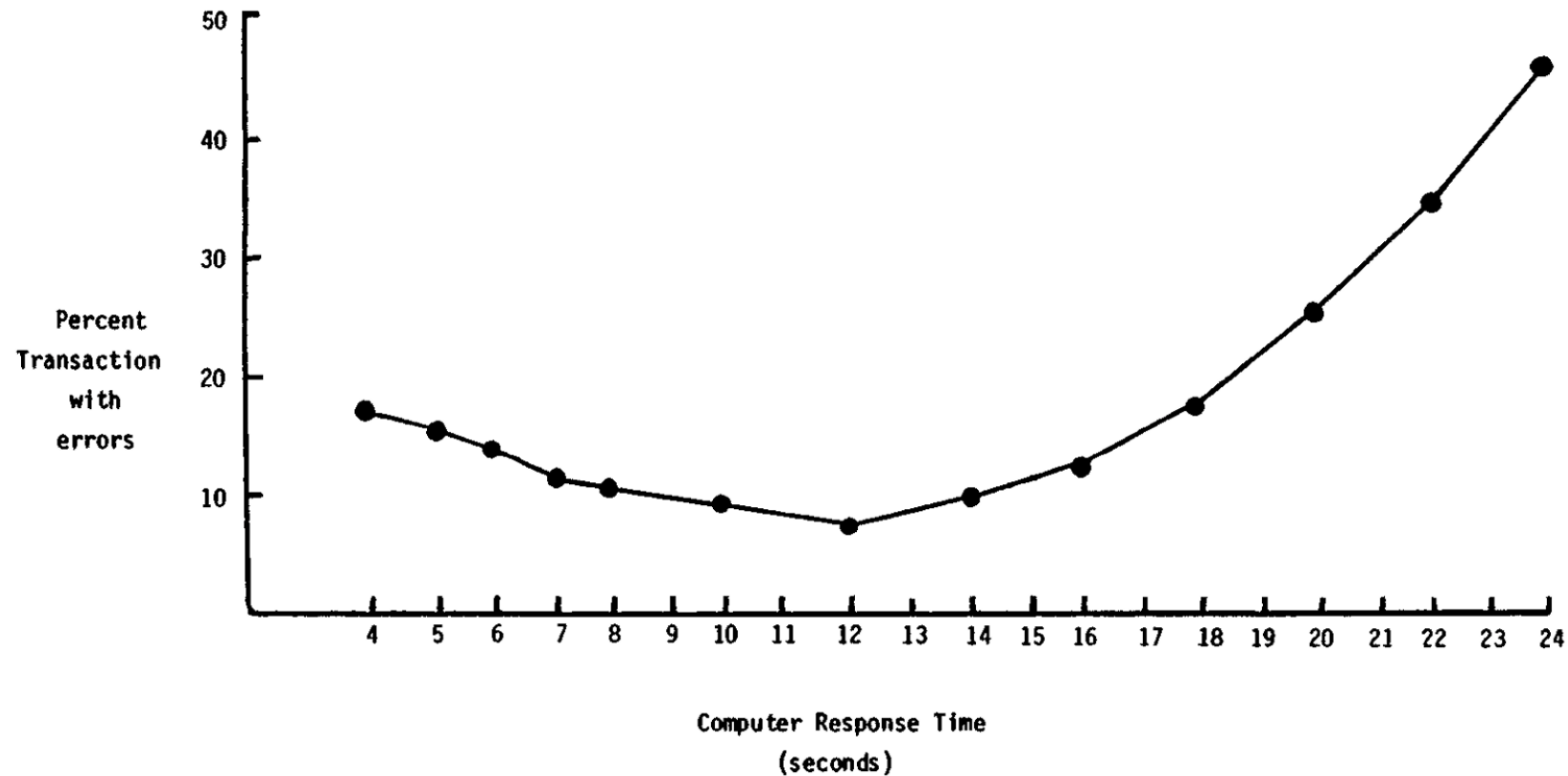
 Should I switch their order?  
Which method do I use?  
What threshold do I use?  
...

 How about I just go with  
what worked last time!

The sub-optimal decisions could compromise system performances and model accuracy

# Data analytics can be made more effective when considering the human aspect

#1 Interactivity: response time can directly impact user productive

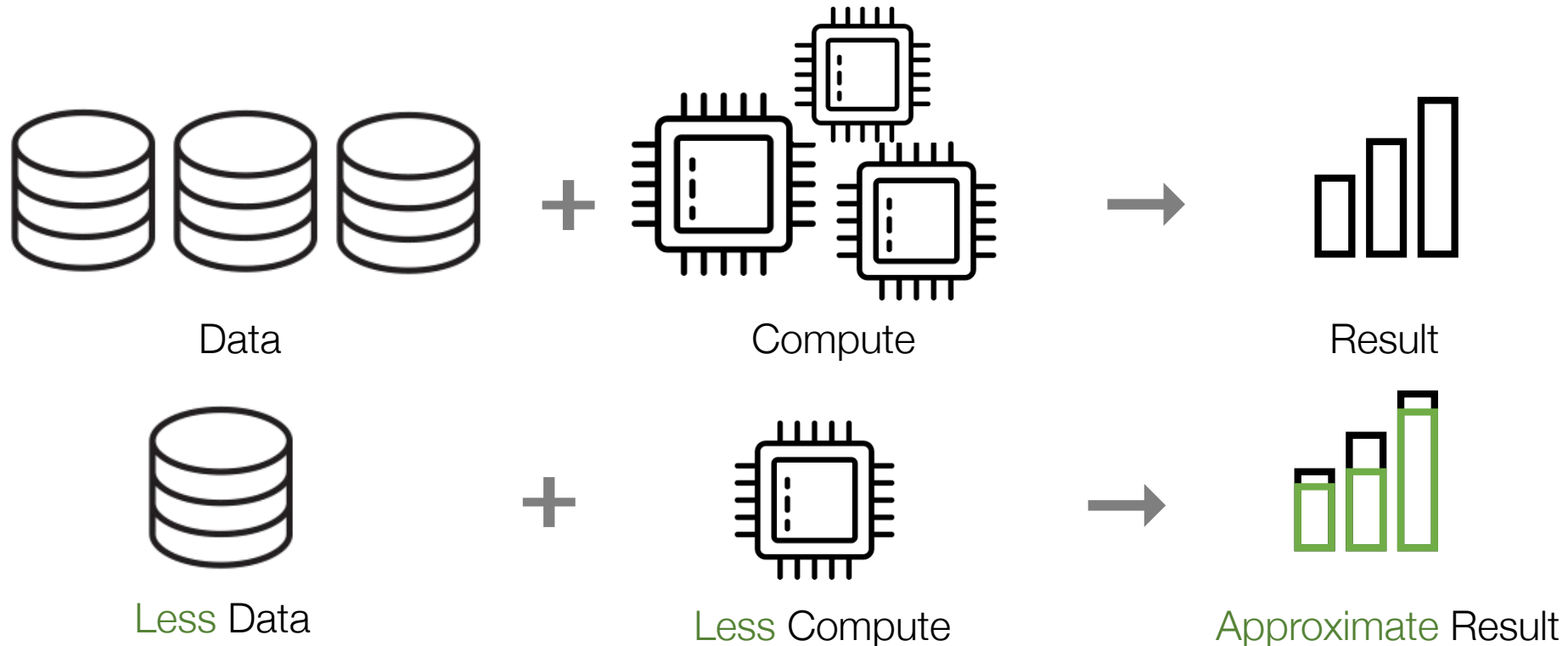


**Figure 4.** Error rates as a function of response time for a complex telephone circuit layout task by Barber and Lucas [1983].

# Data analytics can be made more effective when considering the human aspect

#1 Interactivity: response time can directly impact user productive

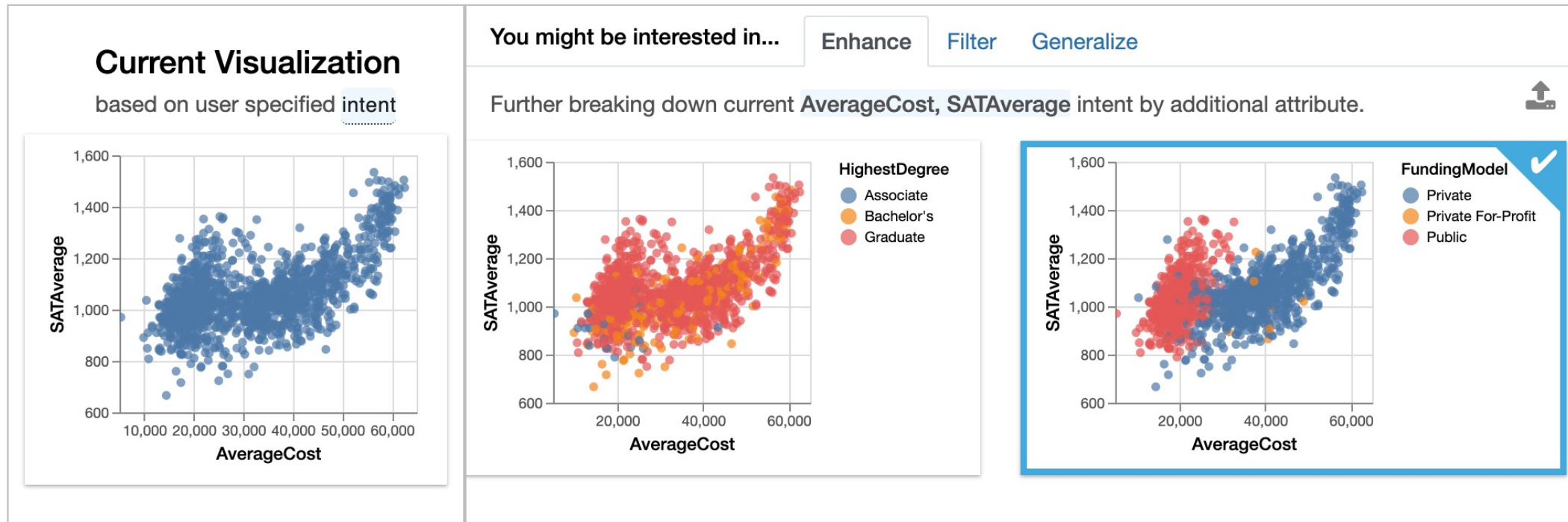
Example from this class: approximation techniques



# Data analytics can be made more effective when considering the human aspect

## #2 Intelligence: help users better navigate the large search space

Example from this class: visualization recommendation



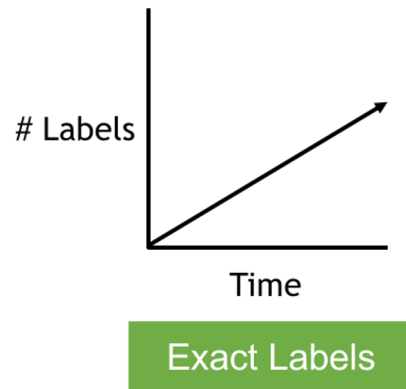


# Data analytics can be made more effective when considering the human aspect

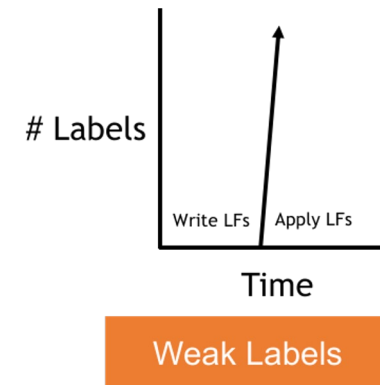
**#3 Interface:** better capture user's feedback and domain knowledge

Examples from this class: scalable label generation

Labeling individual data points



Writing Labeling Functions (LFs) where each LF abstracts a supervision source (e.g. heuristics, existing models, external KBs, ...)



# Overview of course topics

~4 papers per topic:

1. Approximation I: approximate query processing  
Sampling and sketching
2. Approximation II: approximate nearest neighbor search  
Application in data management and machine learning
3. Data Producers  
Data preprocessing and labeling
4. Data Consumers  
Visualization and recommendation

# Schedule

Tentative Schedule (subject to changes)

Date	Topic	Content	Presenter
W1: Aug 21	Lecture	Course Introduction and Logistics	Instructor
W1: Aug 23	Lecture	Research Skills (Part 1)	Instructor

W2: Aug 28	Lecture	Approximate Query Processing	Instructor
W2: Aug 30	Approximation	<a href="#">BlinkDB: Queries with Bounded Errors and Bounded Response Times on Very Large Data</a>	
W3: Sep 04	No Class (Labor Day)		

# Getting to know you

Your name?

Which department/program are you in?

What are you hoping to get from the class?

# Next Class

What is research?

How to develop a novel idea?

How to read a paper?

The peer review process

## Your task:

Office hour poll: <https://forms.gle/RrPBVLJbMQfYG7Jn6>

Sign up for presentation: <https://tinyurl.com/2s3amhrv>

Start looking for project teammates