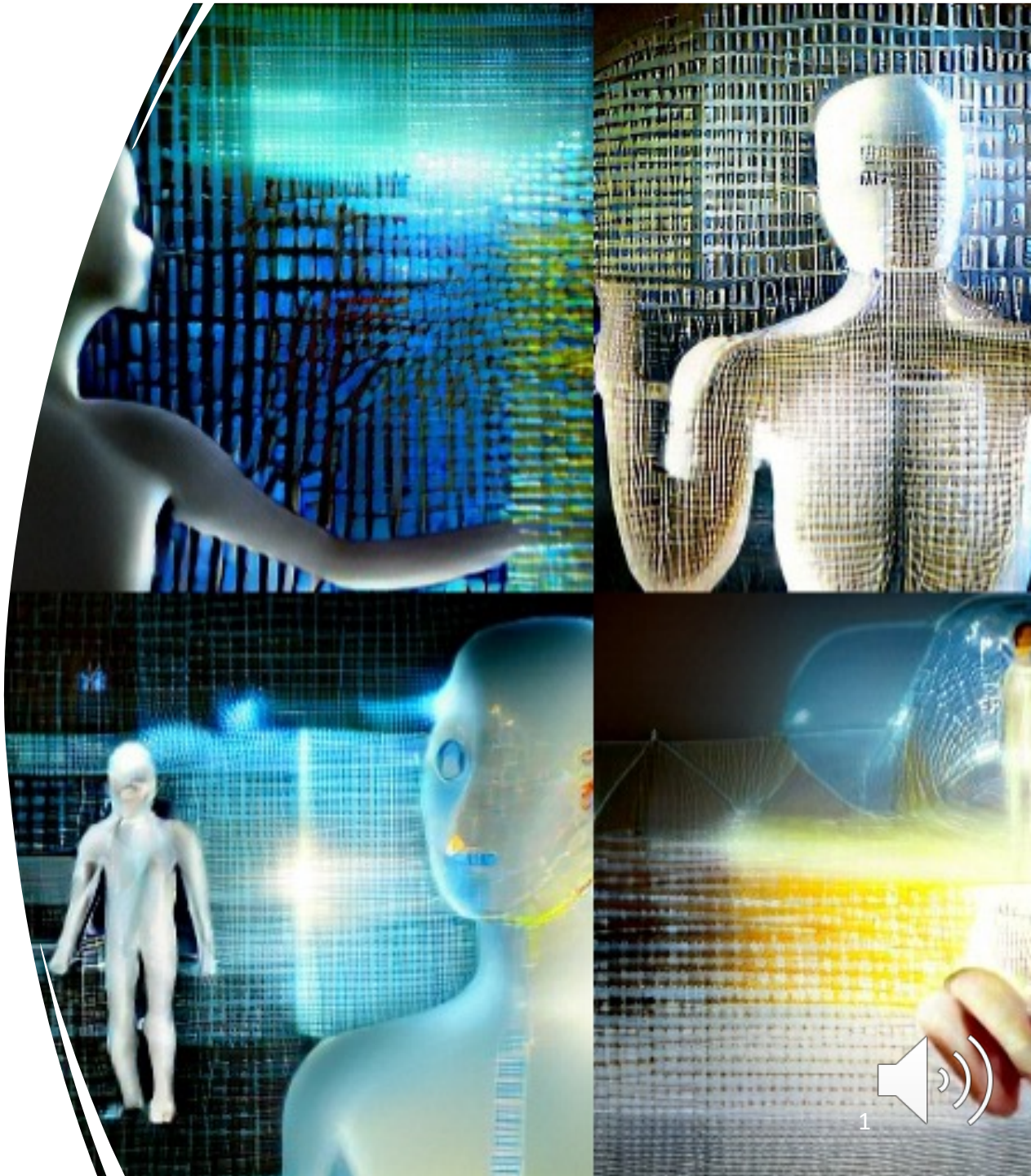
CS 8803-MDS

Human-in-the-loop Data Analytics

Lecture 1

08/22/22



Today's Class

The essentials

What is this class

Getting to know you

Overview of course topics

The essentials

Instructor: Kexin Rong
Office: Klaus 3410
Email: krong@gatech.edu

TA: Kaushik Ravichandran
Email: kravicha3@gatech.edu

OH: Thursday/Friday, time TBD
https://forms.gle/Faqb4N8Q9Uq7gyQk8
Also available by appointment

The essentials

Course website: https://kexinrong.github.io/fa22-cs8803/

gradescope: submitting assignments

Entry Code: NXPRRE

Please add yourself in if you are currently in the waiting list!

Email: meeting/extension requests

mention CS8803 in the email title

Teams: connecting with classmates

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

Get hands-on research experience

- Critically read and evaluate papers
- Technical presentation
- Conduct novel research

Grading

Participation:	50%
Paper reviews	15%
Presentation	25%
Participation	10%

Project:	50%
Project Proposal	5%
Intermediate Report	10%
Project Presentation	10%
Final Report	25%

No curve, everyone can get A!

New course
First time I'm teaching at Tech
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

Shared with paper presenters

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

Class Format:
Role-Playing Paper-Reading Seminars

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

Presentation

25% 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

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 x 4

5min

always start with a *one-slide summary*

Roles available

peer reviewer

archaeologist

academic researcher

industry practitioner

Role: Peer Reviewer

The paper has not been published yet and is currently submitted to a top conference where you've been assigned as a peer reviewer.

Complete a full review of the paper based on prompts of the official review form of the top venue in this research area (e.g., *VLDB, SIGMOD and CHI)*:

- Overall evaluation: {Accept, Weak Accept, Weak Reject, Reject}
- Summary of contribution
- Describe in detail all strong points, labeled S1, S2, S3, etc.
- Describe in detail all opportunities for improvement, labeled O1, O2, O3, etc.

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.

Find and report on one *older* paper cited within the current paper that substantially influenced the current paper and one *newer* paper that cites this current paper.

Role: Academic 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: Industry Practitioner

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

Bring a convincing pitch for why you should be paid to implement the method in the paper, and discuss at least one positive and negative impact of this application.

Sign up for presentation https://bit.ly/3CnBOSa

Credit system

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

paper author role: 2~3 credits

accessory roles: 1 credit

Rules

need >=5 credits over the semester

need to be in the paper author role at least once

presenters for the first 3 papers get 1 extra credit

(maximum 1 extra credit per person)

Examples

1 solo paper author role + 2 accessory roles

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

Sign up for presentation https://bit.ly/3CnBOSa

Don’t modify other people’s slots without asking

Slots are frozen one week prior to the actual presentation

1st presentation is 8/31, signup open till 8/29

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

No show to presentation

-1 credit

Grading

Participation:	50%
Paper reviews	15%
Presentation	25%
Participation	10%

Project:	50%
Project Proposal	5%
Intermediate Report	10%
Project Presentation	10%
Final Report	25%

Any questions?

Research Project

50% of grade

Main criteria

something “new” + relevant to course topics

Project milestones

Week 5: project proposal	5%
Week 10: intermediate report	10%
Week 15: peer review	
Week 16: project presentation	10%
Week 16: final report	25%

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

Many possible projects

An interface for querying relations between multiple time series

A scatter-plot tool that scales to 10M datapoints

A data exploration system for large JSON dumps

A jupyter notebook extension that helps track dependencies between cells

A tool to visualize differences between two versions of the dataset

An algorithm to find related datasets from heterogenous data sources

Starting from a new application

Extending an existing algorithm to a new setting or domain

Many public data and workflows

Kaggle

UCI Machine Learning Repository

Github

StackOverflow

OpenAIRE Research Graph

Data.gov

Come to office hours and we are here to help!

Grading

Participation:	50%
Presentation	25%
Paper reviews	15%
Participation	10%

Project:	50%
Project Proposal	5%
Intermediate Report	10%
Project Presentation	10%
Final Report	25%

Any questions?

Getting to know you

Your name?

Which department/program are you in?

Fun fact?

What are you hoping to get from the class?

Overview of Course Topics

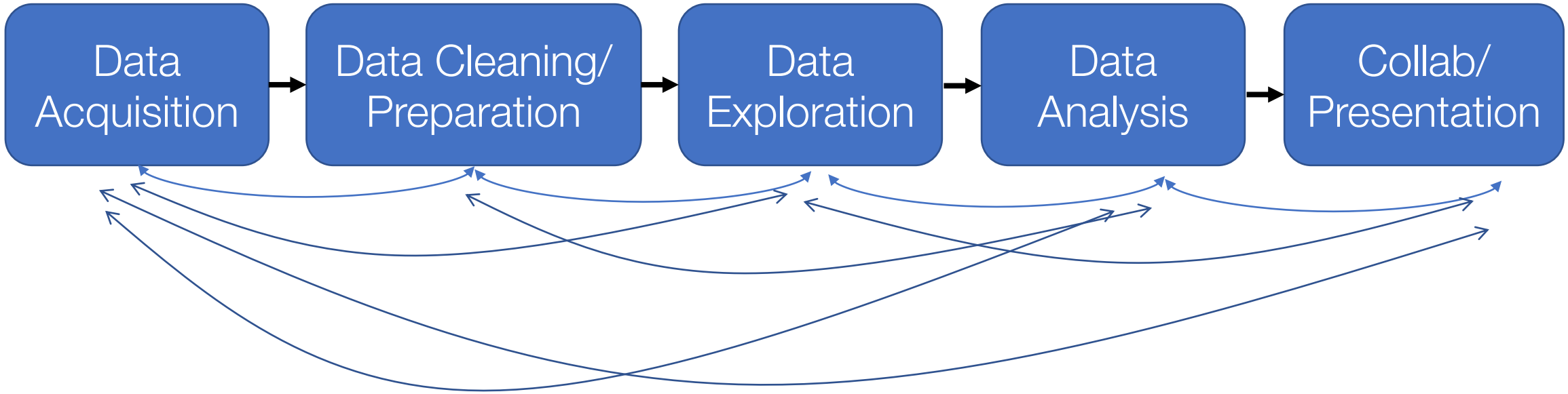
Human-in-the-loop

What is human-in-the-loop?

Data Analytics

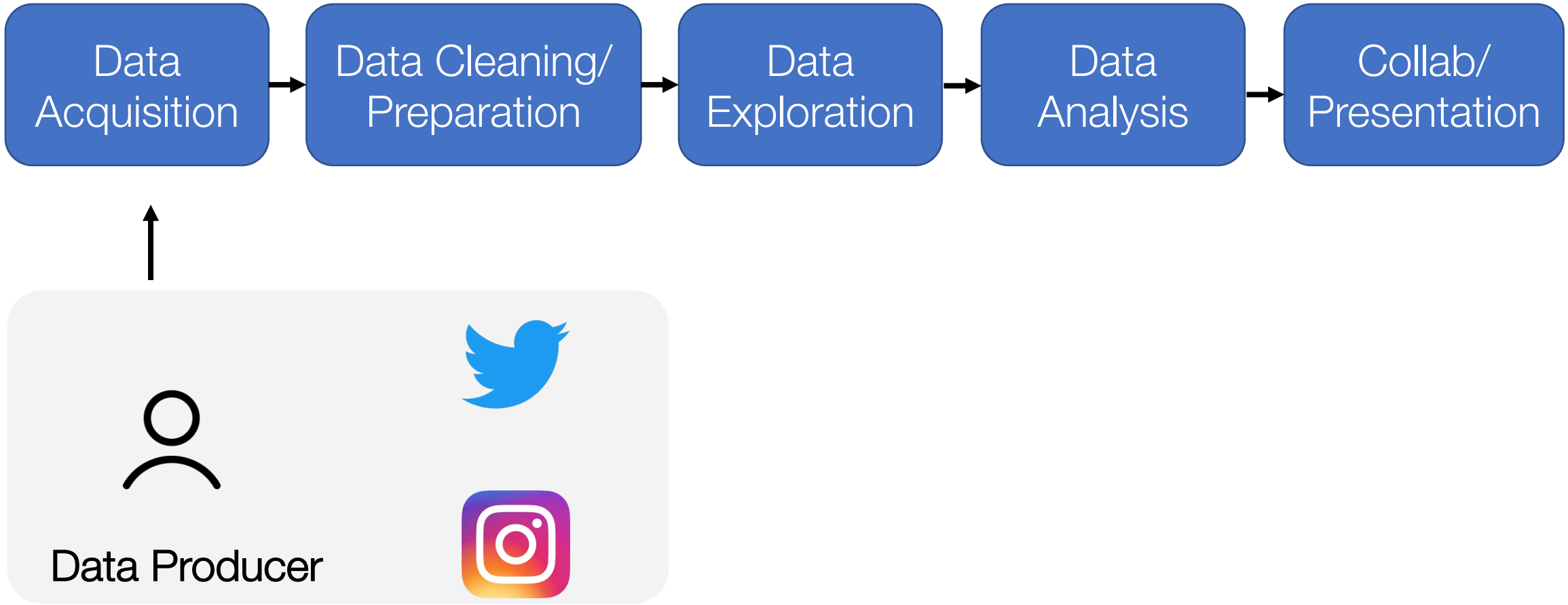
What is data analytics?

The Data Analytics Lifecycle

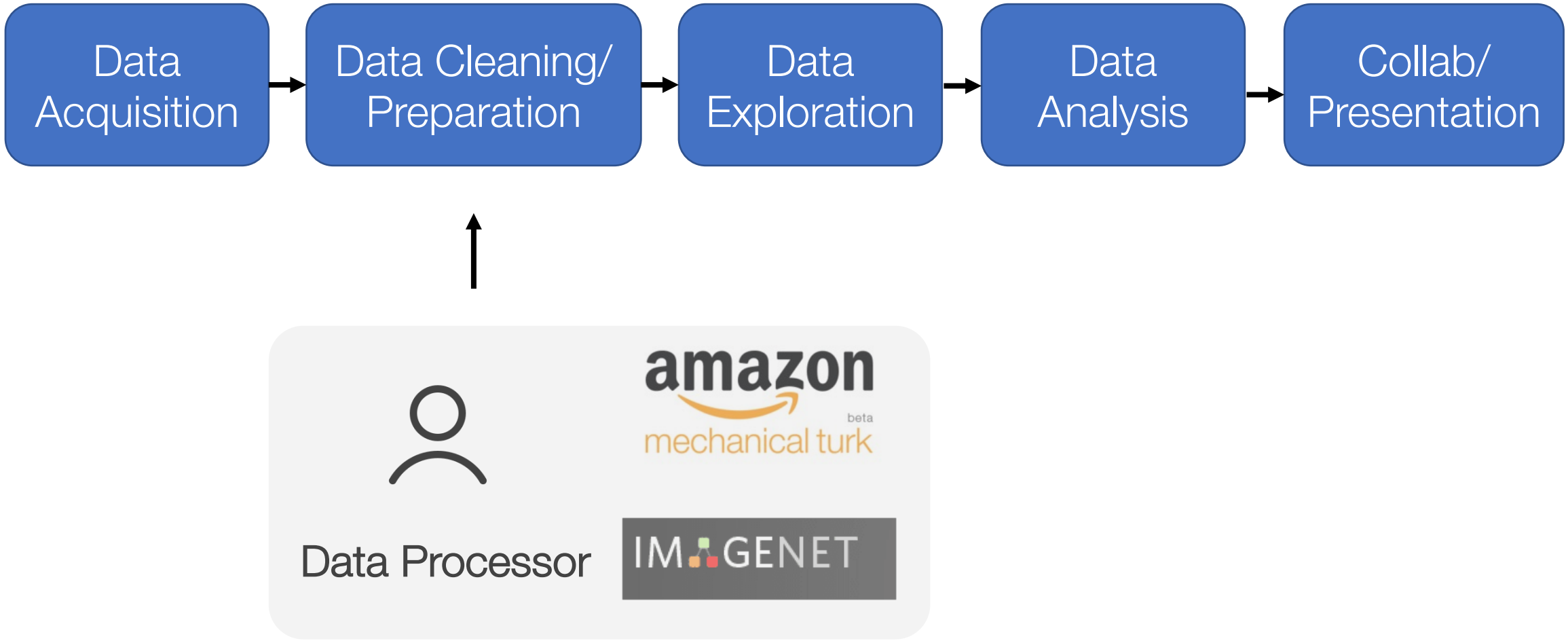


We have a scalability challenge and a need for systems/tools at every single phase of the lifecycle

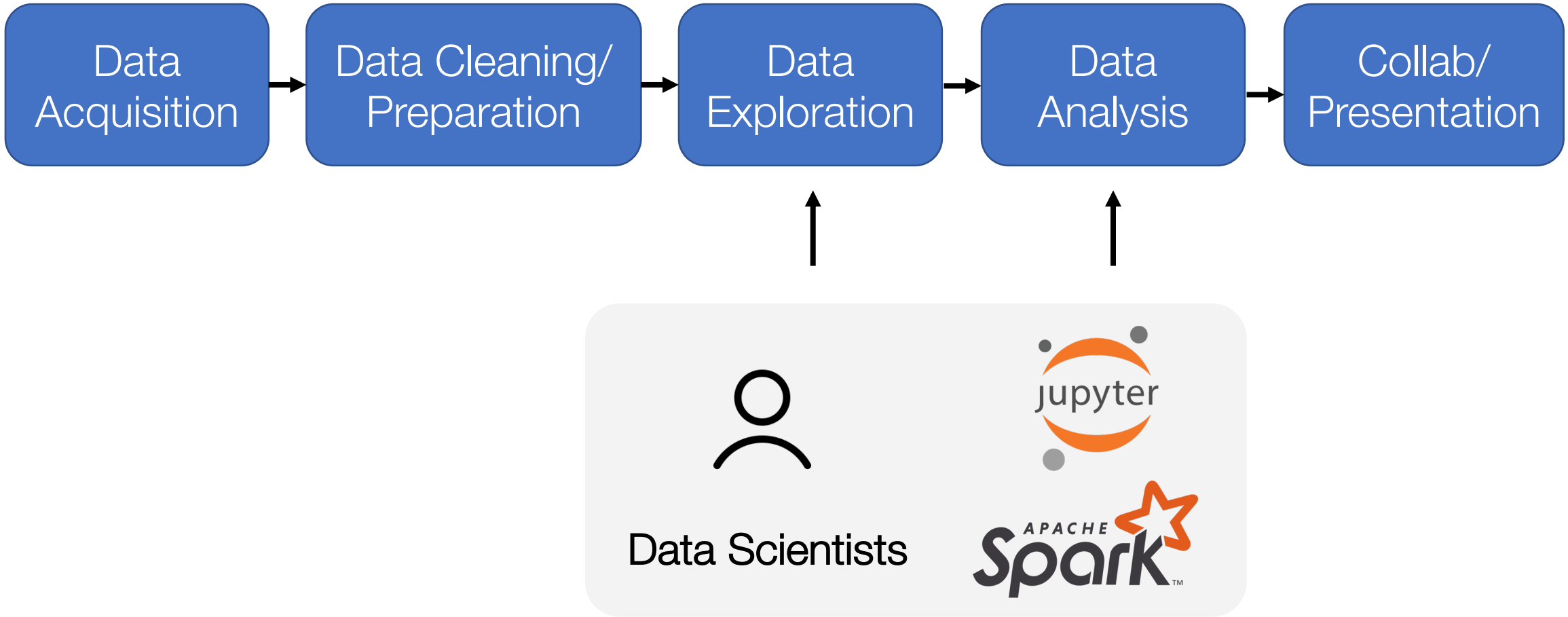
Human roles in data analytics



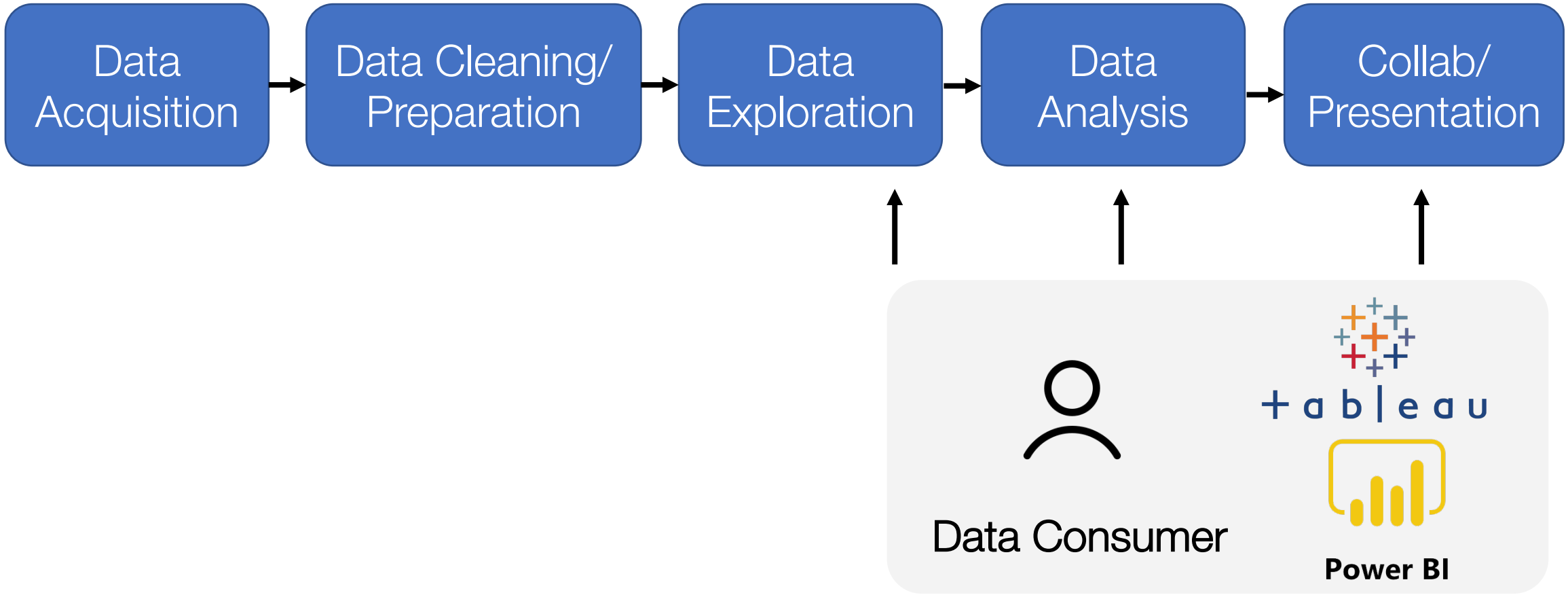
Human roles in data analytics



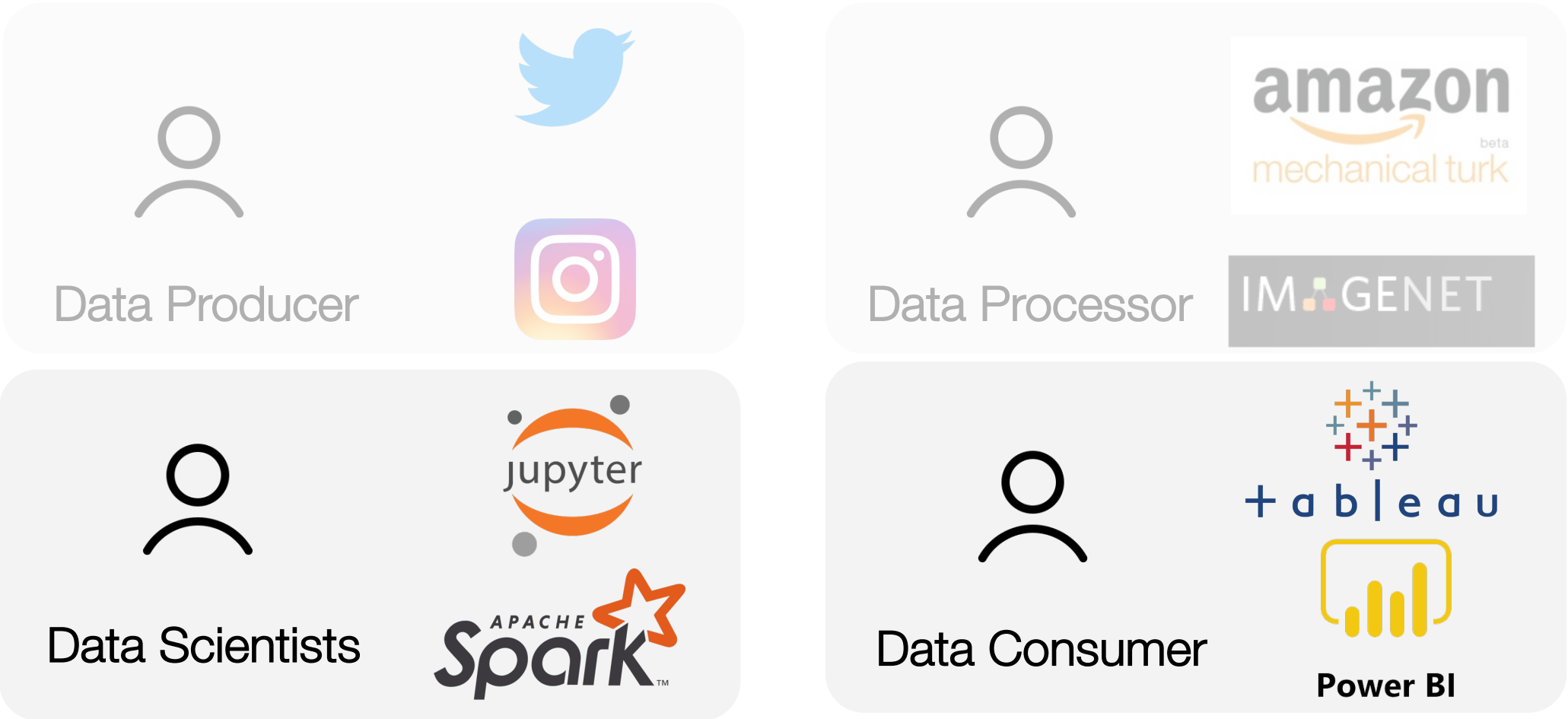
Human roles in data analytics



Human roles in data analytics



Human roles in data analytics



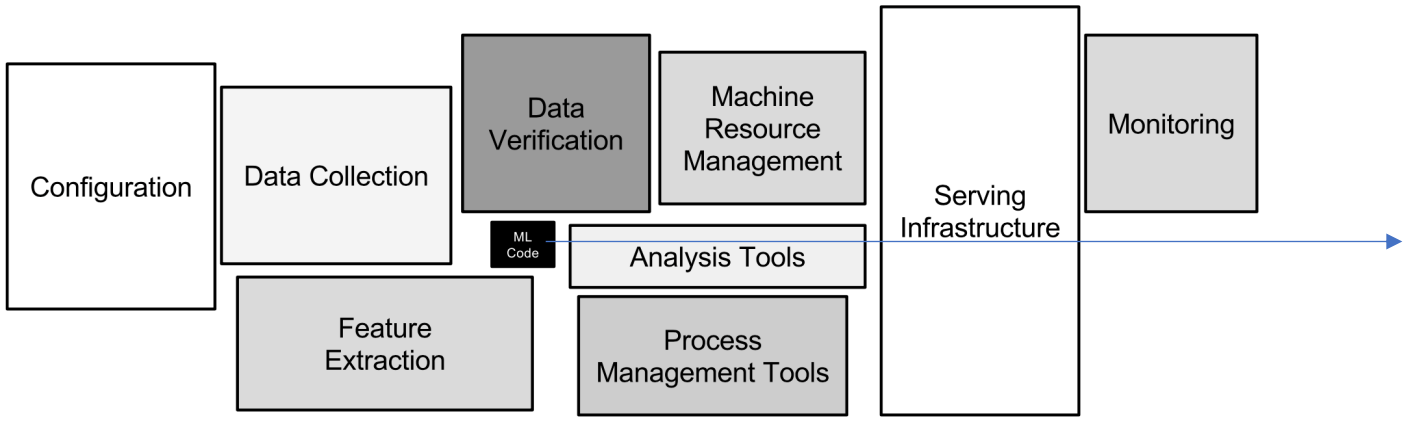
Why do we care about humans?

It's the golden area for data*

*for the best-funded, best-trained engineering teams

Hidden Technical Debt in Machine Learning Systems

D. Sculley, Gary Holt, Daniel Golovin, Eugene Davydov, Todd Phillips
{dsculley,gholt,dgg,edavydov,toddphillips}@google.com
Google, Inc.



"Only a fraction of real-world ML systems is composed of ML code"

Why do we care about humans?

It's the golden area for data*

*for the best-funded, best-trained engineering teams

Powerful
Require expertise

?

Usable
Limited scale

Why do we care about humans?

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

Some examples from this class:

- How to build scalable spreadsheet systems
- How to provide context/explanation for anomaly detection
- How to specify visualization through examples

Next Class

What is research?

How to read a paper?

The peer review process

Your task:

Enroll in gradescope (NXPRRE)

Office hour poll: https://forms.gle/Faqb4N8Q9Uq7gyQk8

Sign up for presentation: https://bit.ly/3CnBOSa