# Unit 1: What is data science?

# Data Collection & Annotation

Lesson 5

Derek Ruths

# Overview of unit

- What is data science?
- What is data?
- **Data collection and curation**
- What is data analytics?
  - How does data analytics fit into data science?
- What is machine learning?
  - How does machine learning fit into data science?

# Lesson overview

## Objectives

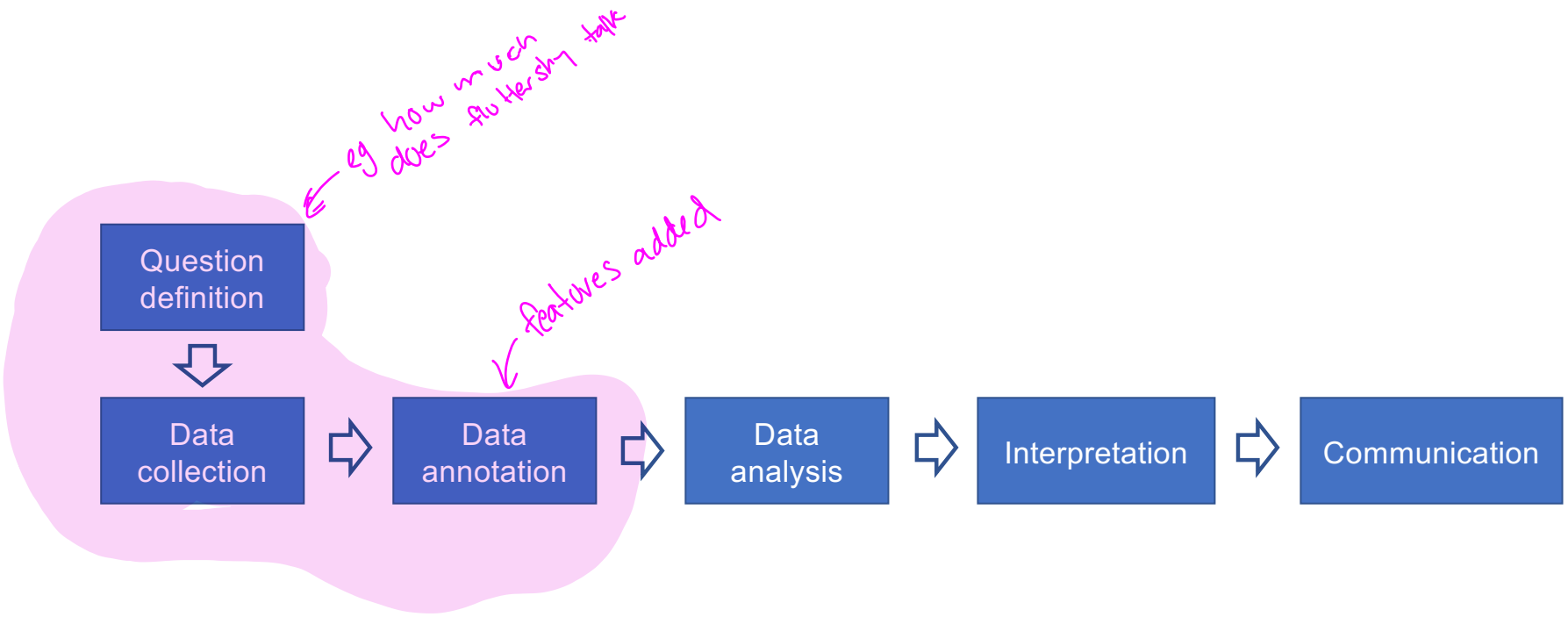
- Understanding of objectives and outputs of
  - Data collection
  - Data annotation

## Outline

- Data science phases
- Data collection
- Data annotation


# In My Little Pony, how often does Fluttershy talk to other ponies?

# Data science project phases



# Data collection

Obtaining and arranging data so it can be worked with.

Activities:

- Decide what to collect
- Obtaining the raw data
- Identifying anomalies
- Correcting issues in the data
- Standardizing the structure of the data

eg inconsisten naming in MLP scripts

# Data annotation

- Applying/inferring features that will be used for analysis.

Three primary ways of generating features:

- Human annotation
  - Expert annotation
  - Crowd sourced annotation
- Machine learning

. or both

# Lesson wrap-up

## Takeaways

- Fluttershy is a shy, nuanced pony
- Data collection can be very involved
- Data annotation can involve humans or machines (or both)

## Up next

- What is data analysis?