Data Science: Tools & Process

Lesson 2

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON

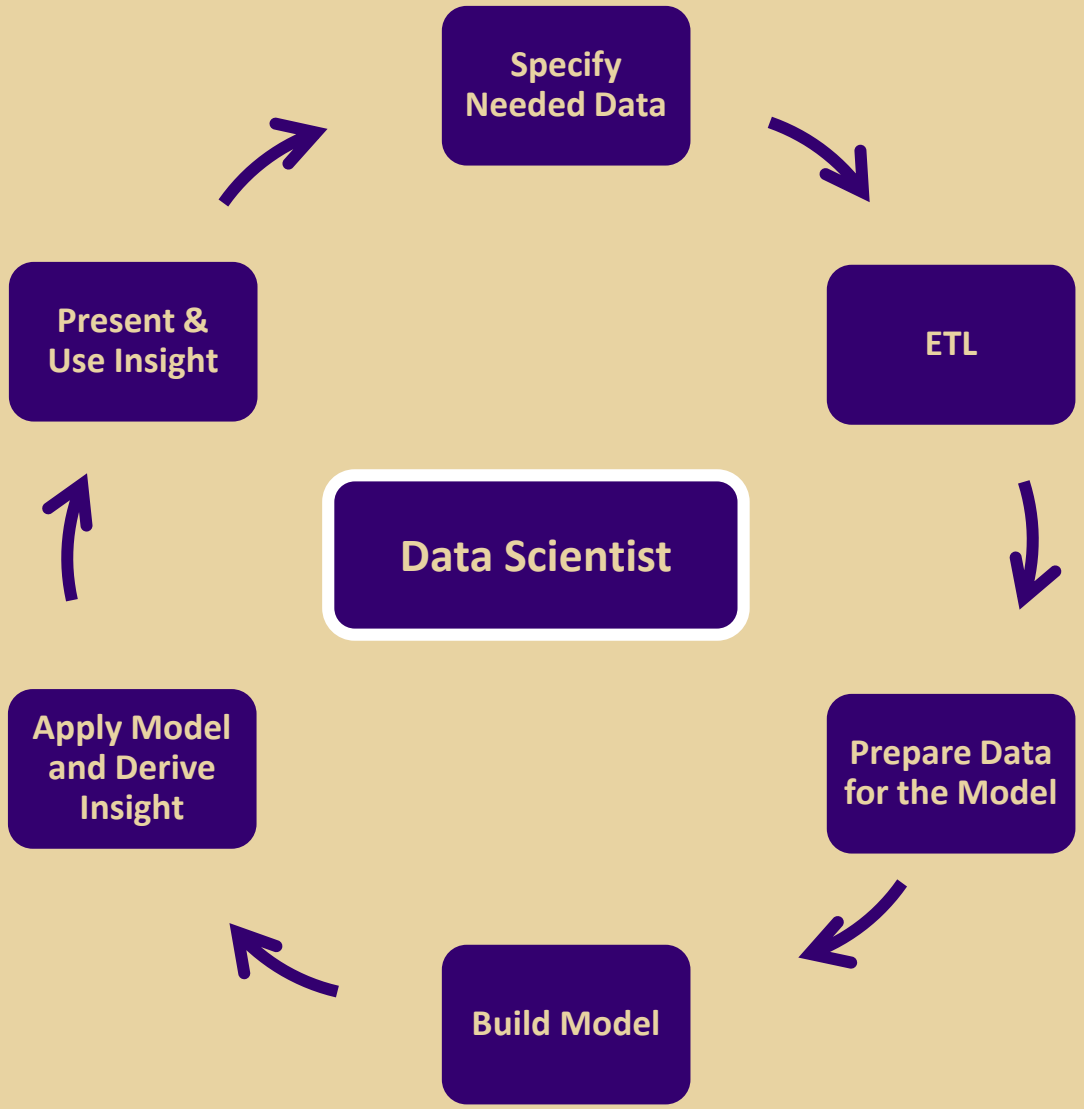
PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON

Where do Data Scientists spend most of their time in tackling problems?

Data Science Cycle

Which part of the cycle is the time consuming?

Why?



Data Flow Diagrams

A How-to for Milestone Project 1

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON

Data Flow

- Required for Data Processing
- SSADM specifies Data Flow Diagrams (DFD)

Four components of a DFD:

- Terminator
- Store
- Process
- Data Flow

What is a Data Flow Diagram?

A defined language in the structured systems analysis and design method (SSADM).

- for describing processes that involve movement and transformation of data.

Dataflow diagrams (DFD,) define processes and do not necessarily represent components. DFDs processes are easily related to development tasks.

DFD Symbols

Complete Rectangles = start or terminate process

– either generate or consume data.

Rectangles without sides = stores, like databases.

Ellipse = a process that transforms data.

Arrow = data.

Data Flow Diagram Example

What are popular tourist locations to photograph?

DFD Example: Image Aggregation Story

Describe, in a few sentences, a data science task that interests you.

1. Data are extracted and processed from images on cell phones
2. The processed data are combined
3. The combined data are used to derive meaning, like: Which are the popular tourist locations?

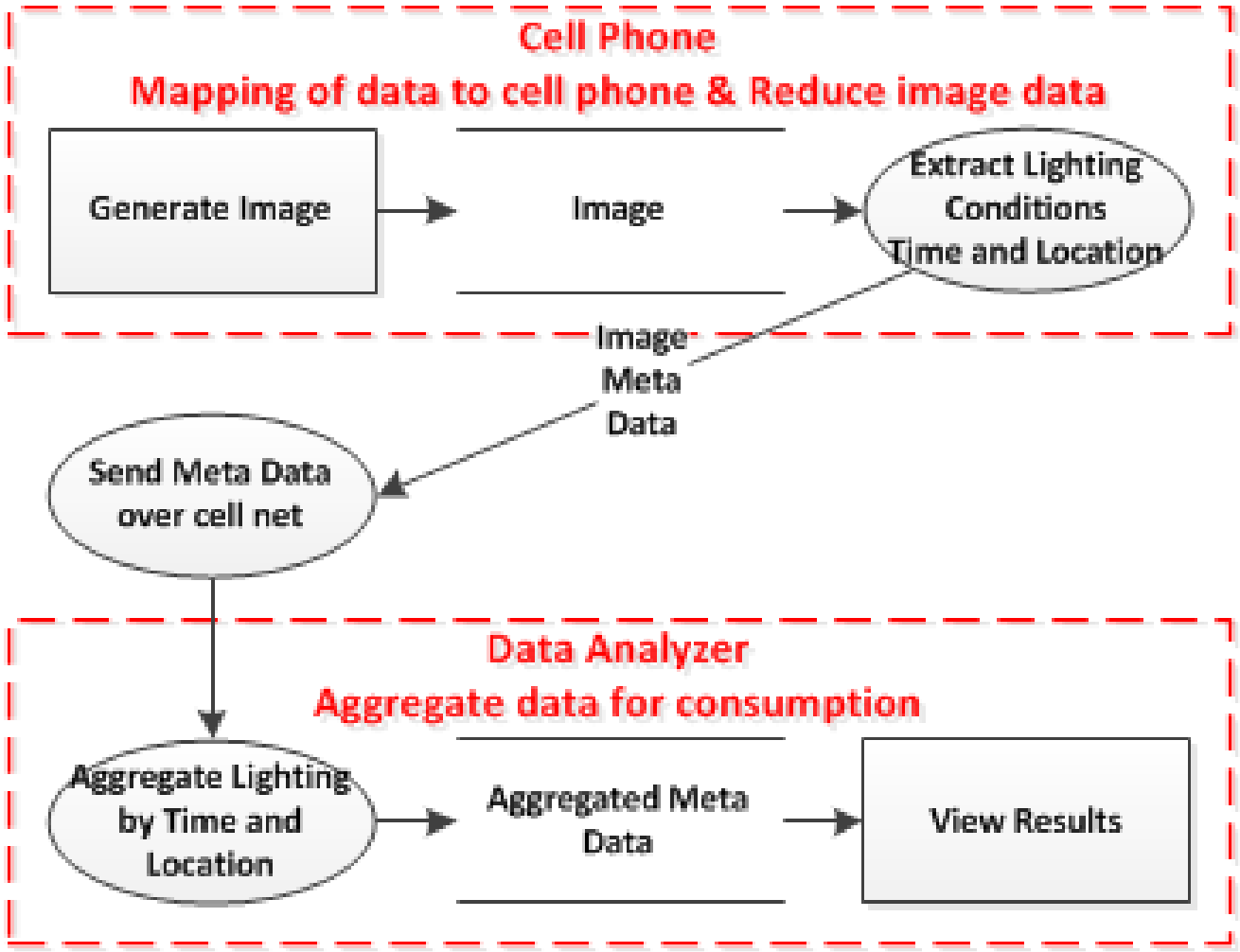
Construct a data flow diagram that depicts the data processing that is required to complete the task in item 1

DFD: Image Aggregation Steps

Collect and aggregate cell phone camera images

1. The image is taken (Image is mapped to cell phone)
2. Image is associated with cell location and time
3. The image data is extracted (Data Reduction)
4. The data (Image characteristics, time, and location) are sent
5. The data are collected and aggregated by location and time
6. The data are viewed

Image Aggregation DFD



A closer look at DFD

Understanding the components

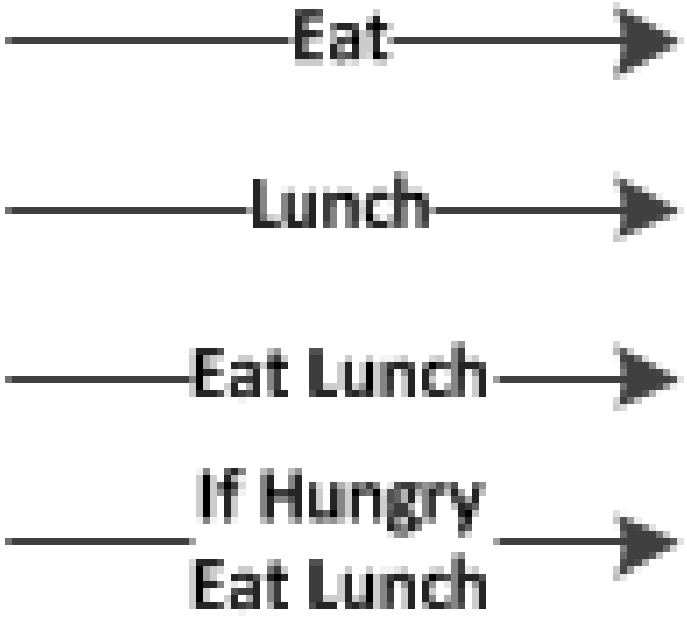
Data Flow: DFD Arrow

An arrow represents data or data flow. The arrow is labeled by the name of the data. Example:

An arrow is necessary to connect the other data flow components. Every data flow component must have at least one arrow.

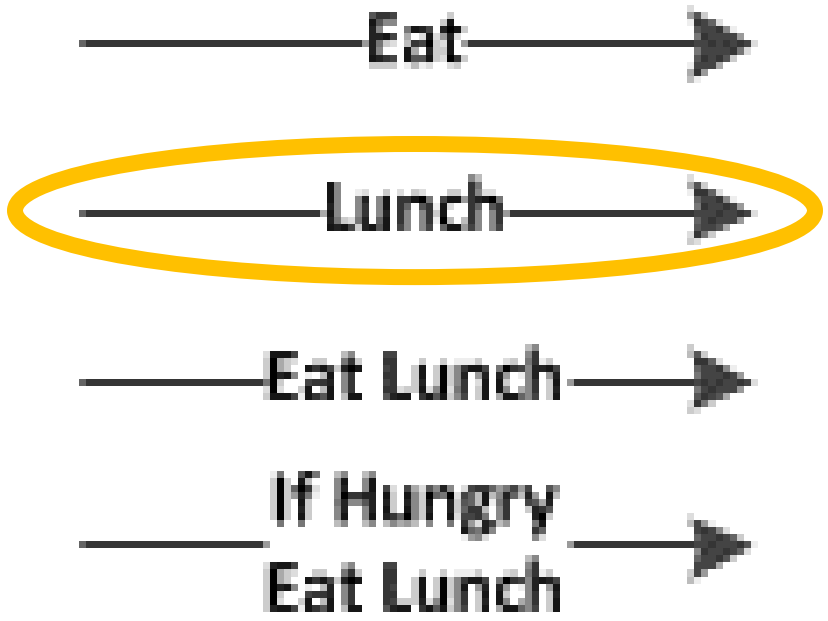
Data Flow Practice: DFD Arrow

Which example is correct?



Data Flow Practice: DFD Arrow

Which example is correct?



Data Flow: DFD Process

Represented by an ellipse

Takes in data from one or more data sources, transforms the data, and then outputs the data.

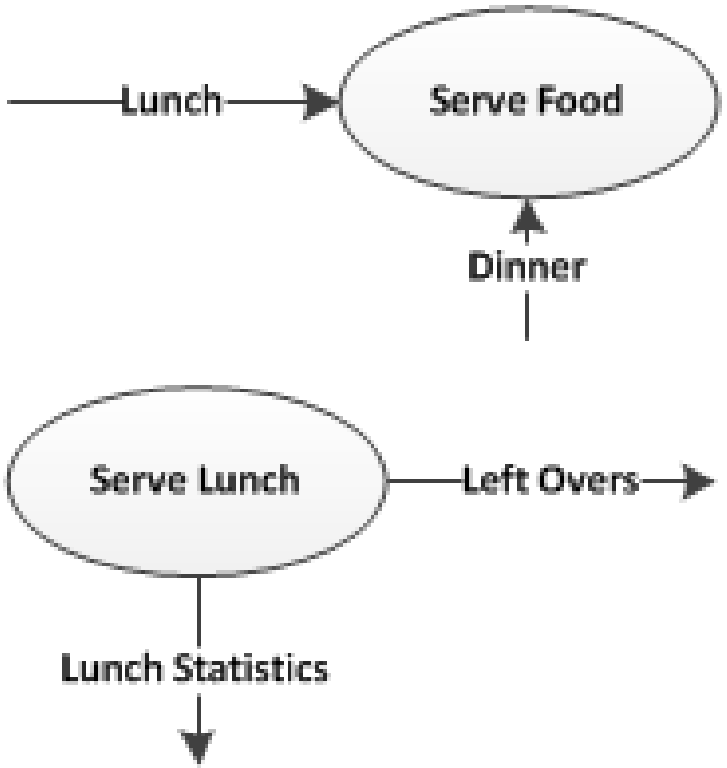
- A process must have at least one input arrow
- A process must have at least one output arrow.

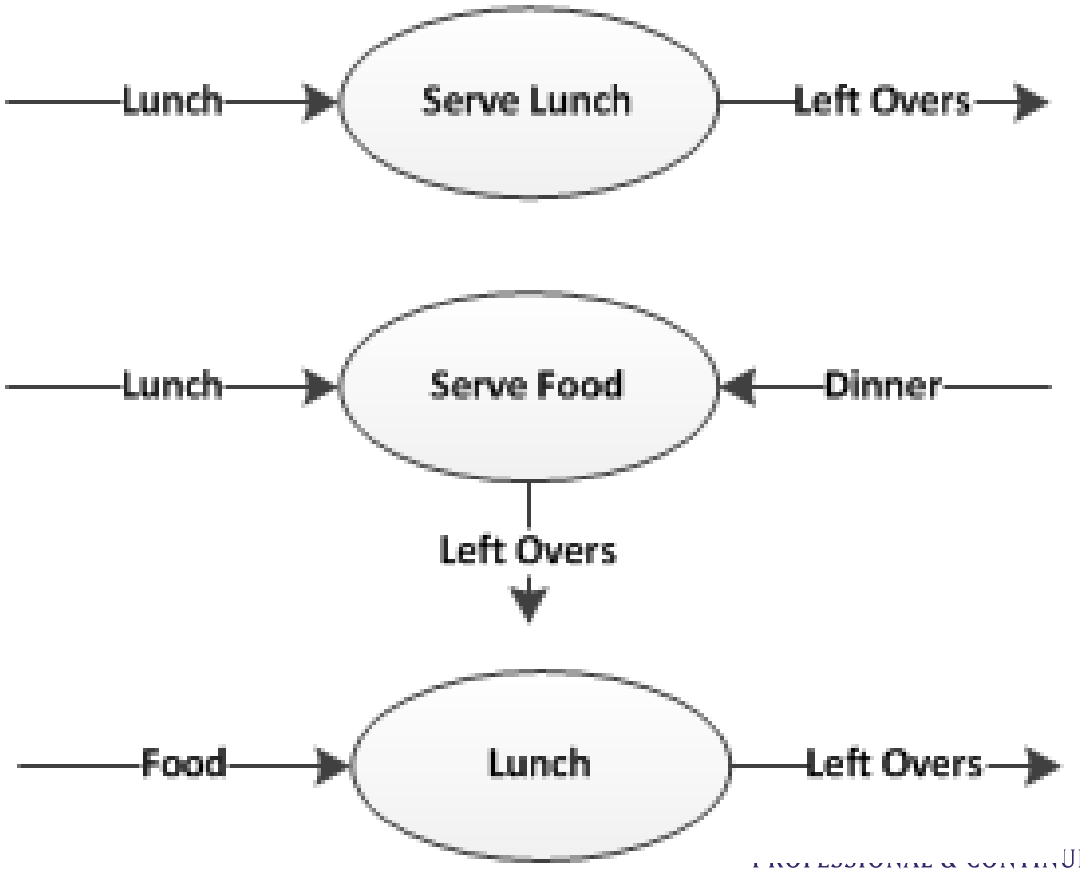
A process is labeled with a verb, like “Brighten”

Example:

Data Flow Practice: DFD Process

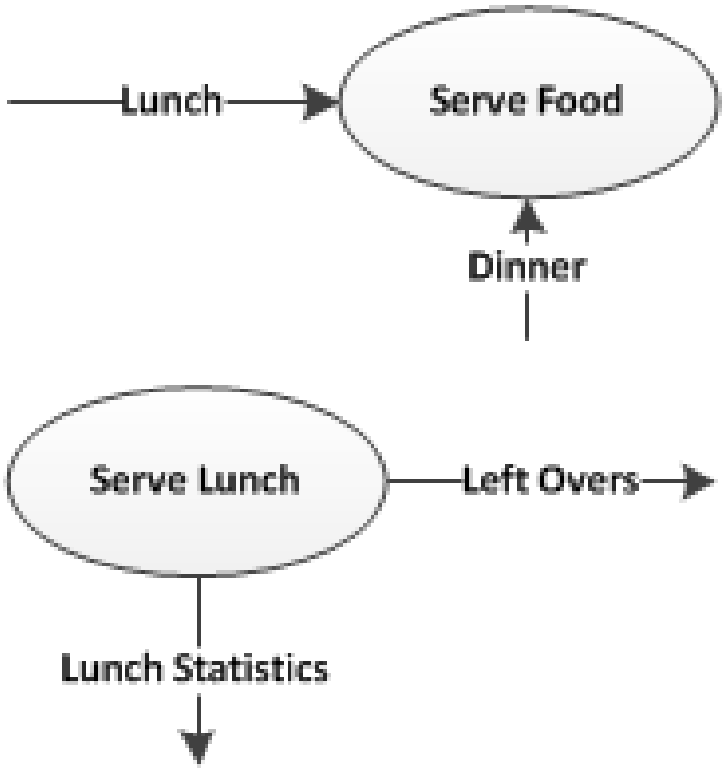
Which of these are correct?

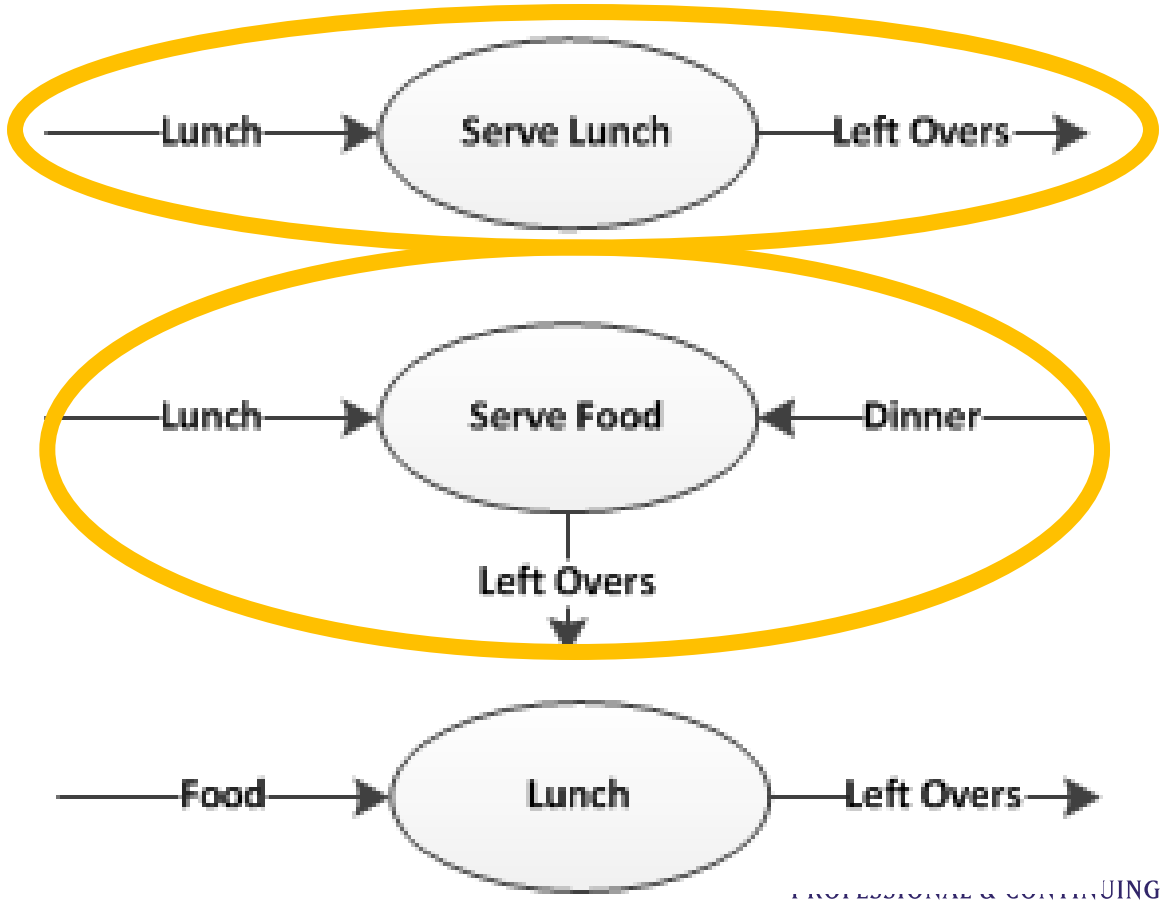




Data Flow Practice: DFD Process

Which of these are correct?





Data Flow: DFD Terminator

Represented by a rectangle with all four sides drawn.

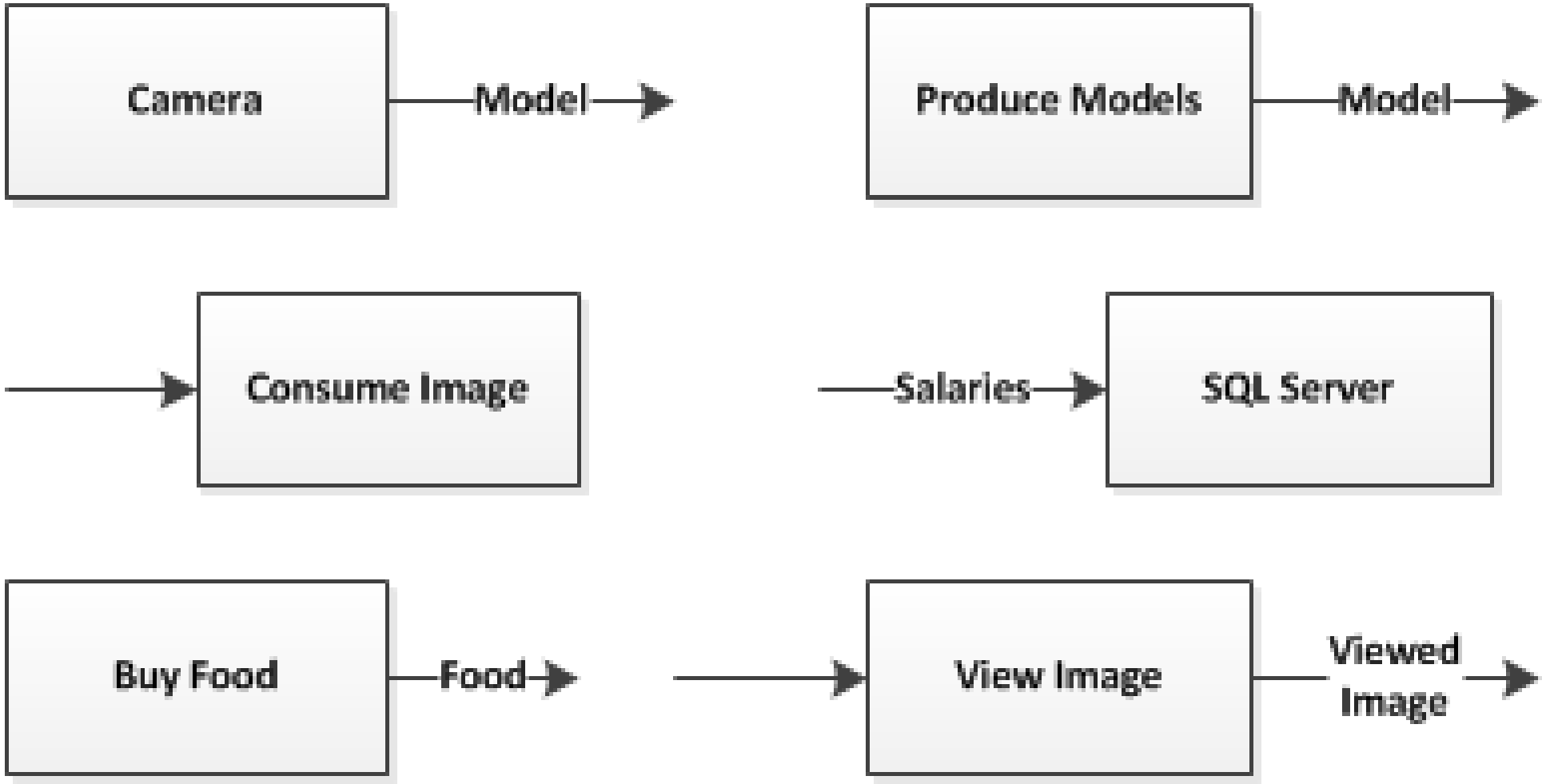
A process that either generates or consumes data.

This process may reference a component like: Get data from Internet or View data in Monitor

Example:

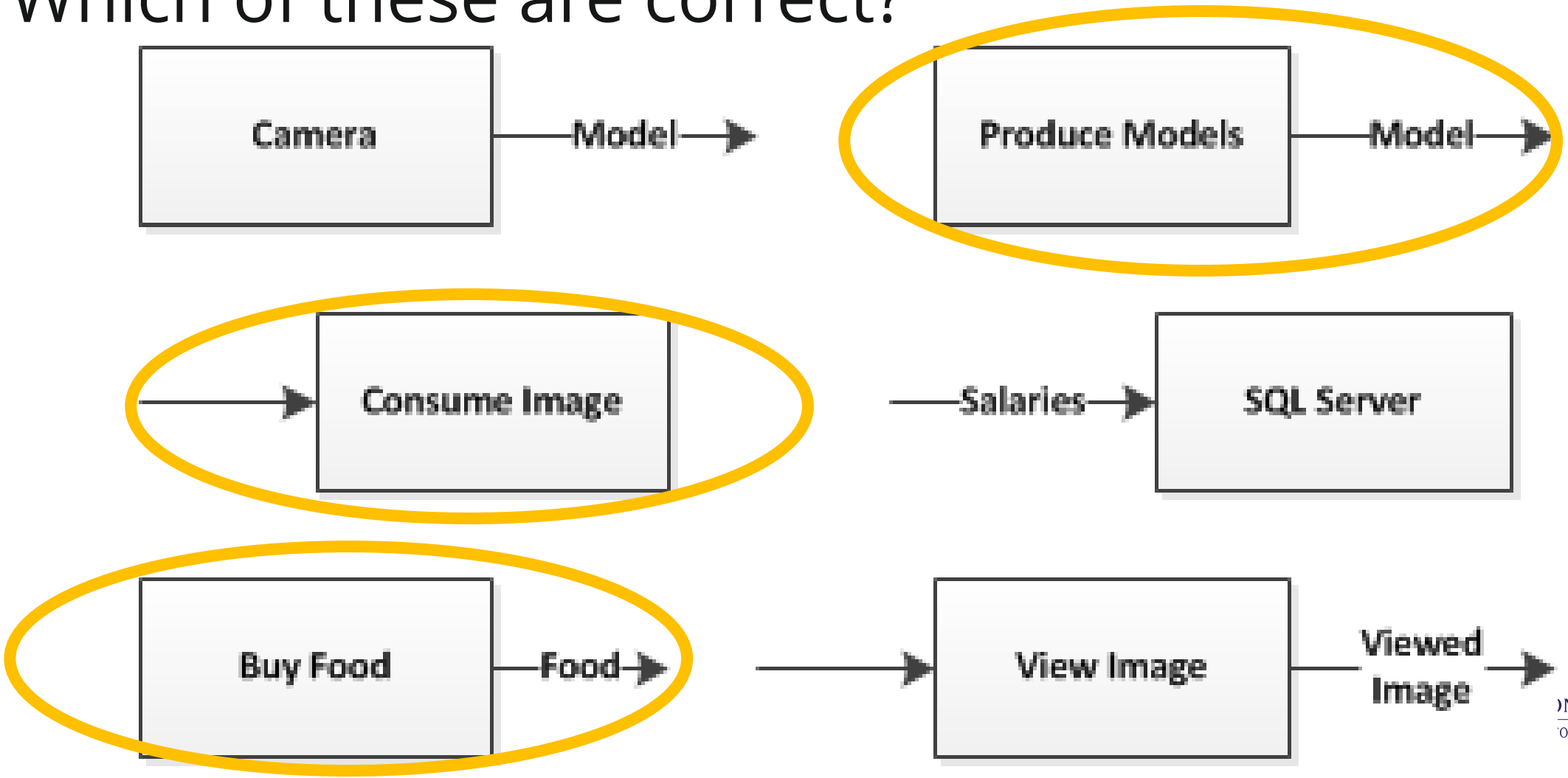
Data Flow Practice: DFD Terminator

Which of these are correct?



Data Flow Practice: DFD Terminator

Which of these are correct?



Data Flow: DFD Store

Represented by a rectangle that is missing the right-hand side or both the right- and left-hand sides.

A place where the data is persisted. Typical stores are text files, websites, and relational data bases.

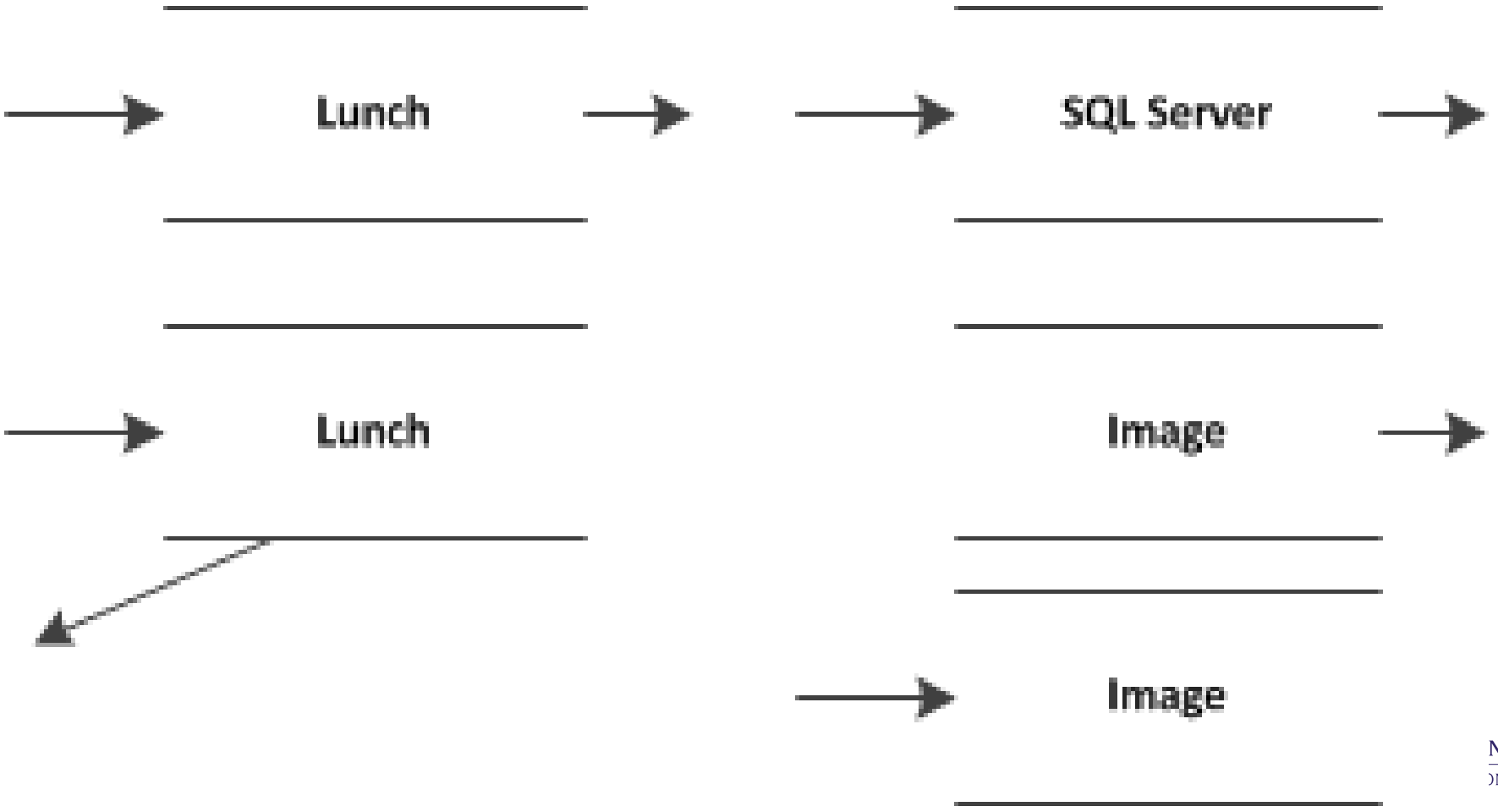
- A store has at least one input arrow
- A store has at least one output arrow
- Typically, the input and output arrows are not labeled.

The name of the store describes the nature of the data (not the nature of the data base)

Example:

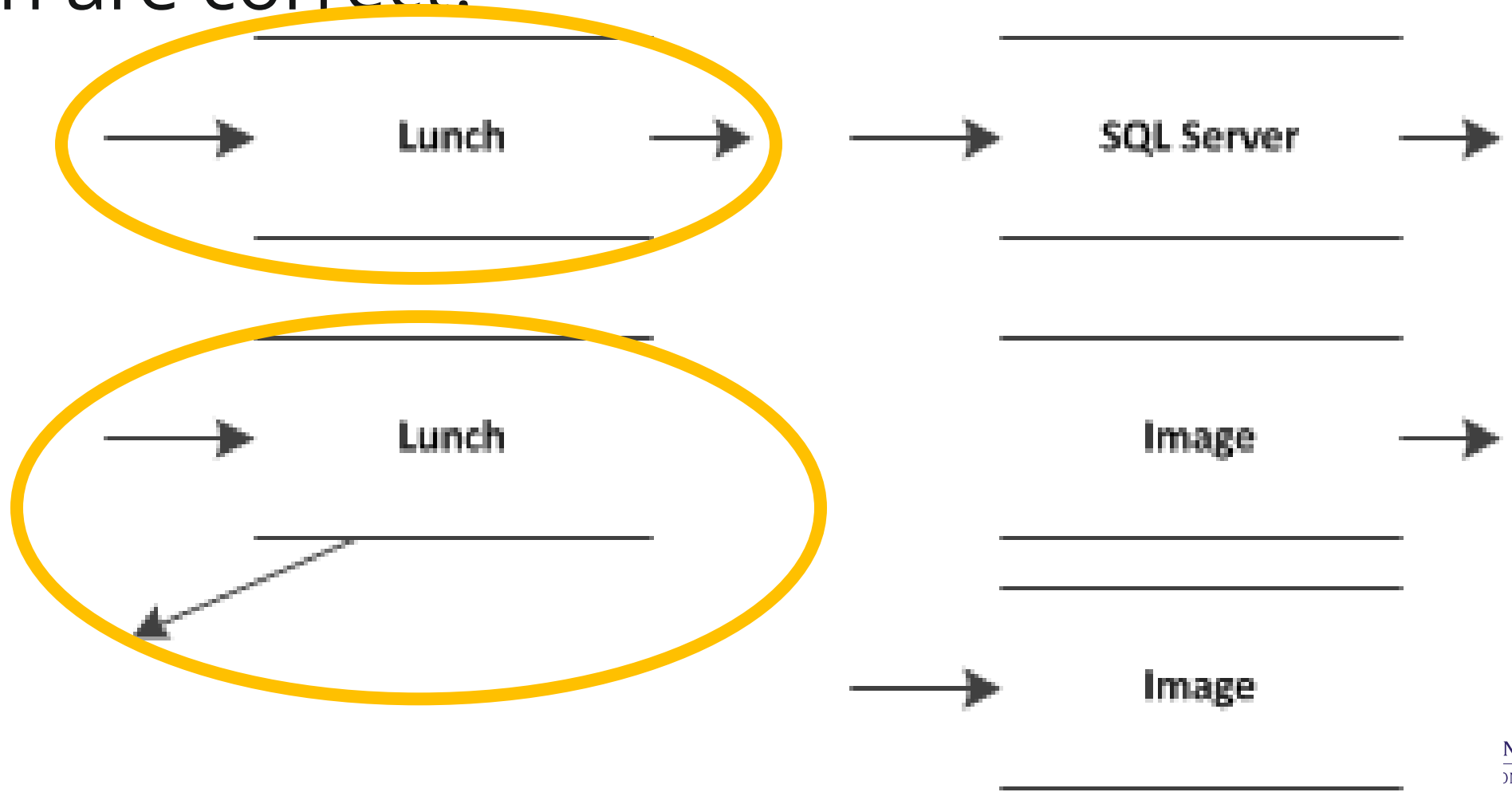
Data Flow Practice: DFD Store

Which are correct?



Data Flow Practice: DFD Store

Which are correct?



DFD: Digital Pathology

An Example

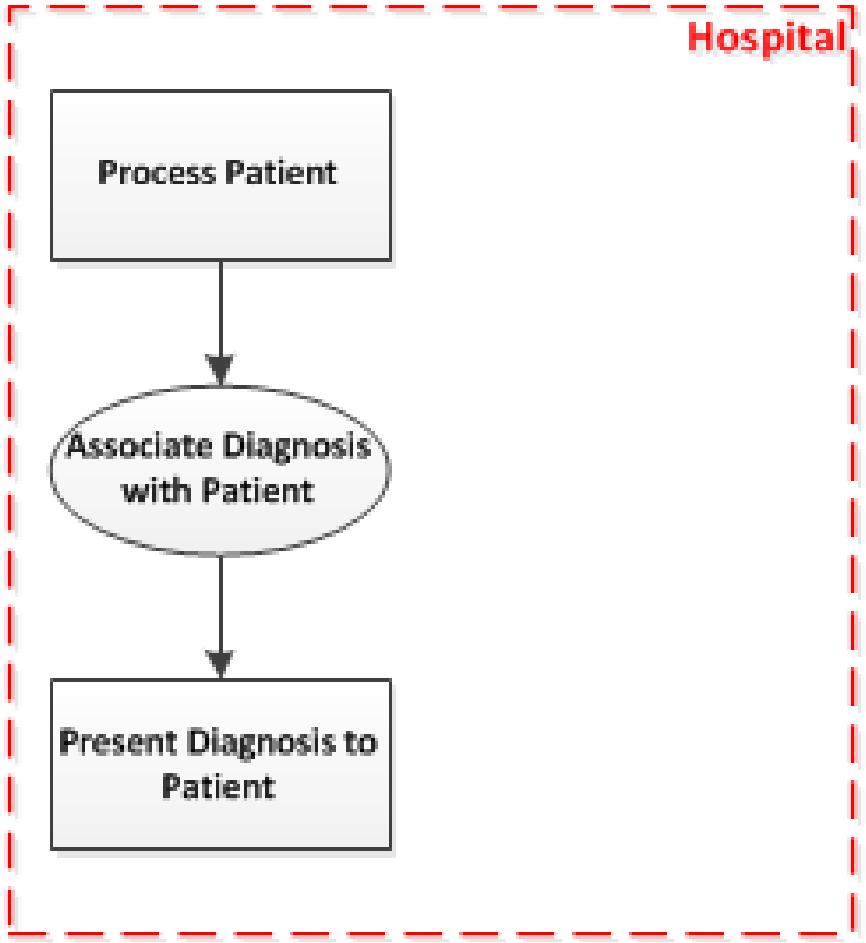
DFD Example: Digital Pathology

Many blood disorders manifest themselves through easily recognizable morphological changes, but the affected cells may be as few as one in a hundred thousand.

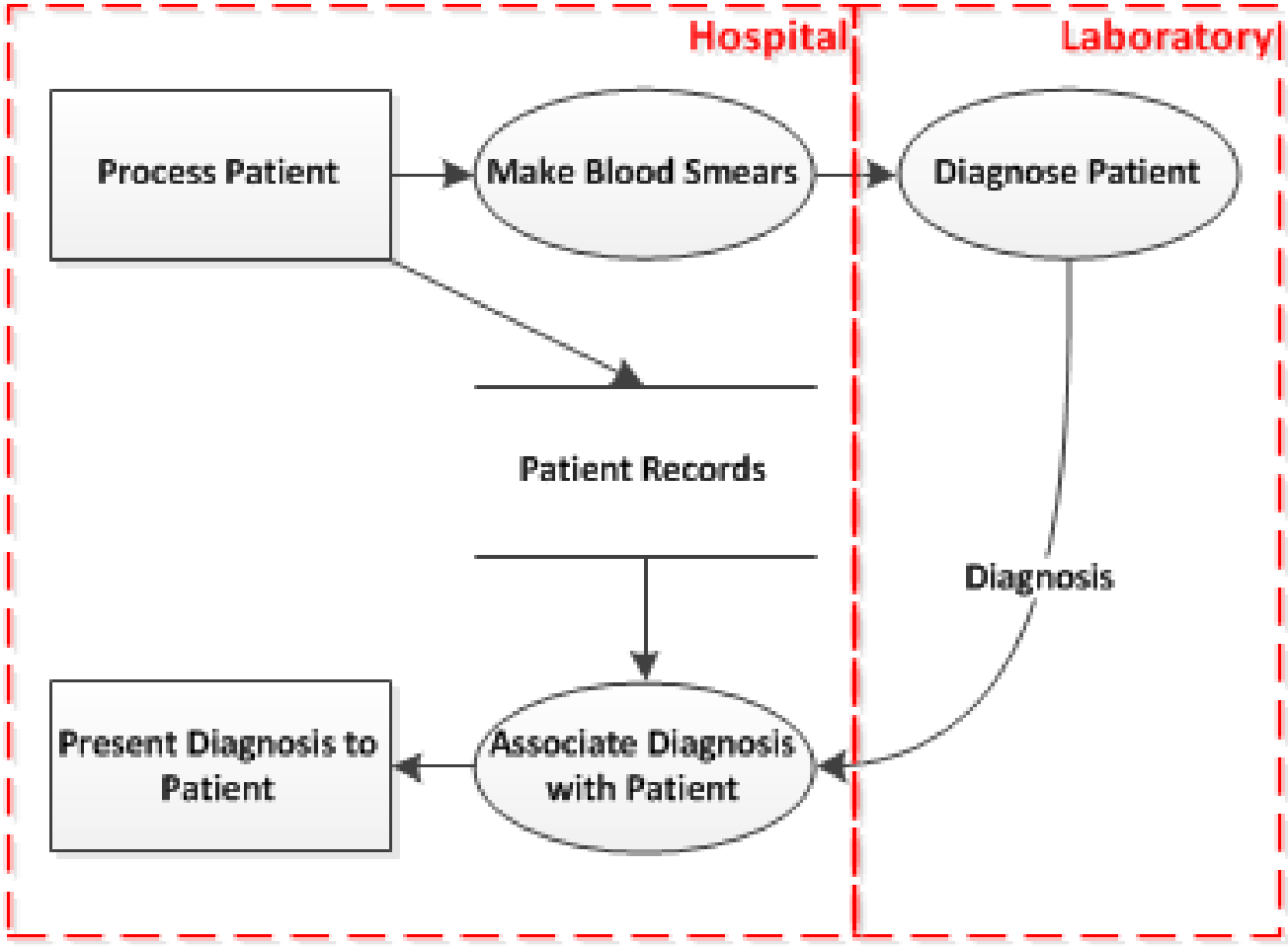
Given the scarcity and cost of pathologists, it is not possible to routinely screen for these blood disorders. We would like to find an automated way of diagnosing such disorders.

We use a pathologist to score aberrant cells and correlate these findings with shape characteristics determined by image segmentation.

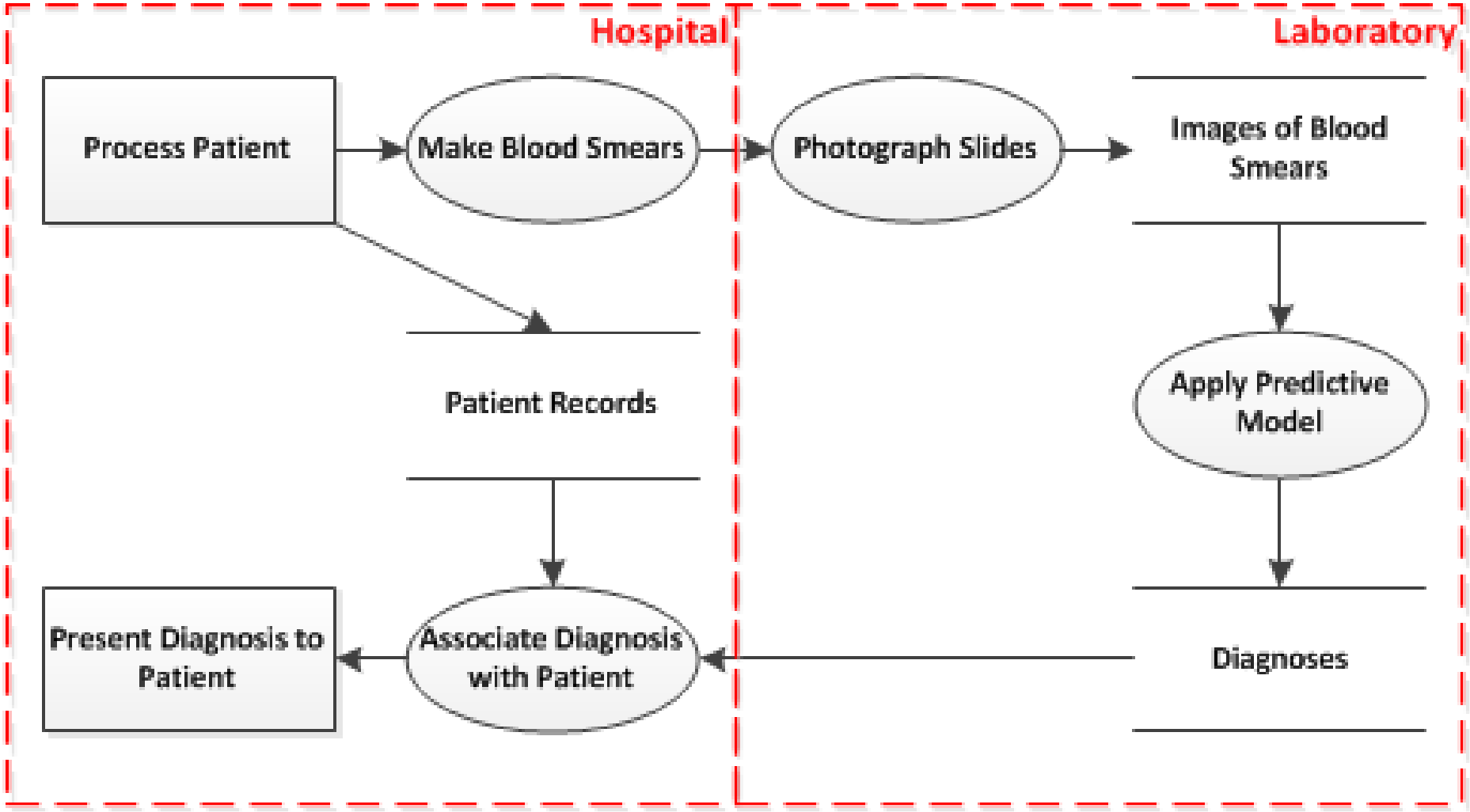
DFD Example: Digital Pathology



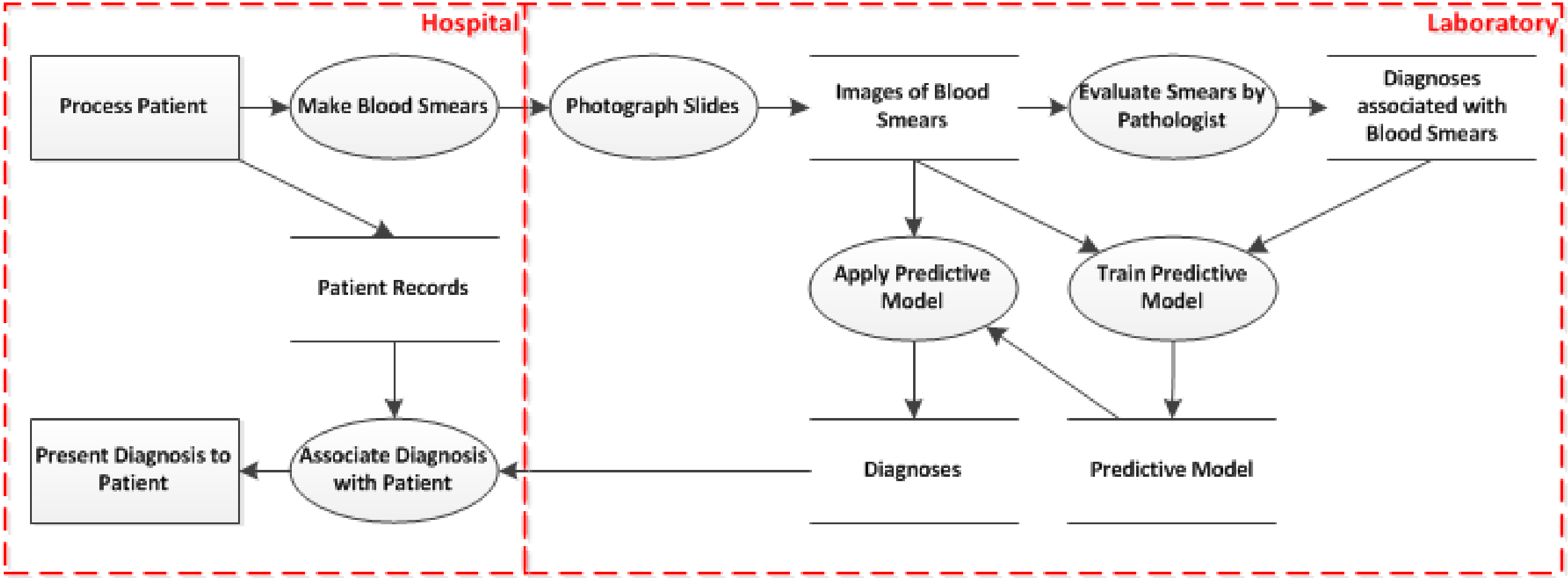
DFD Example: Digital Pathology



DFD Example: Digital Pathology



DFD Example: Digital Pathology



Milestone Project 1: Data Flow Diagram

1. Describe, in a few sentences, a data science task that interests you.
2. Construct a data flow diagram that depicts the data processing that is required to complete the task in item 1

Submit a PDF of the diagram with description of the task.

Data Flow Diagrams

A How-to for Milestone Project 1

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON