Data Science: Tools & Process

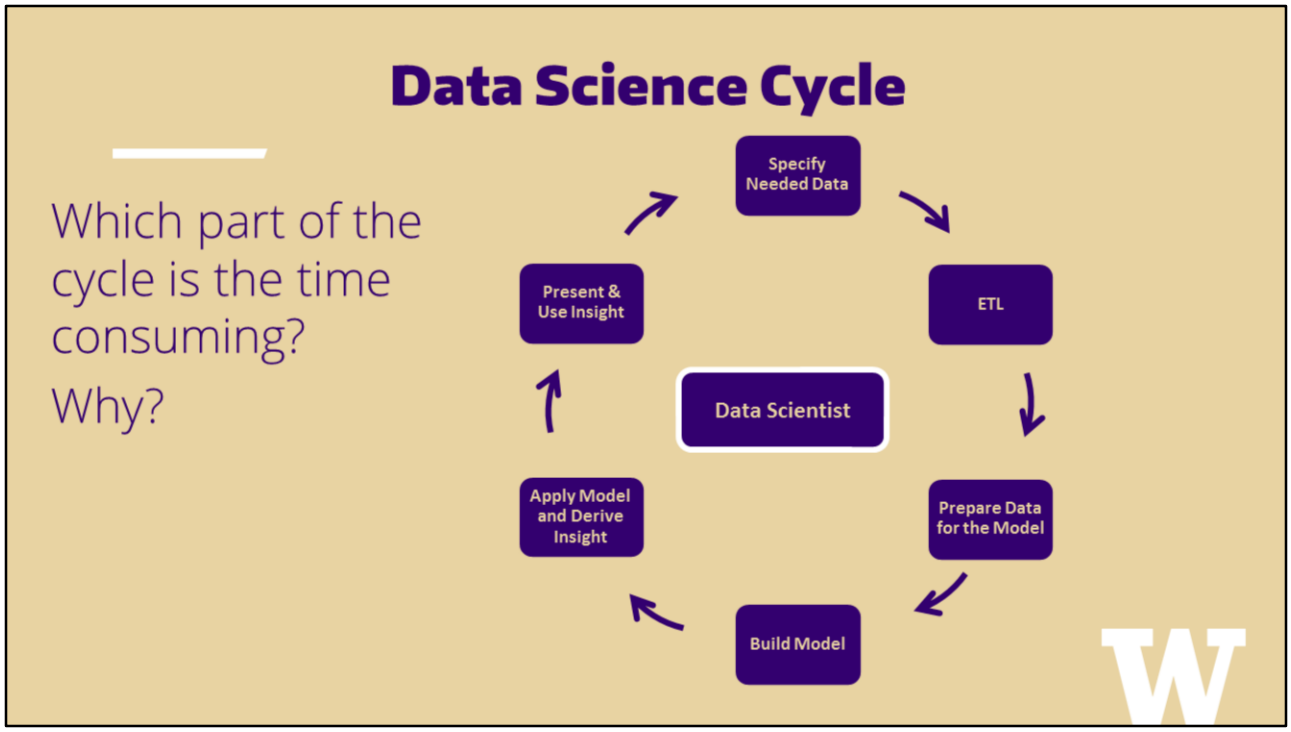
Week 2

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON

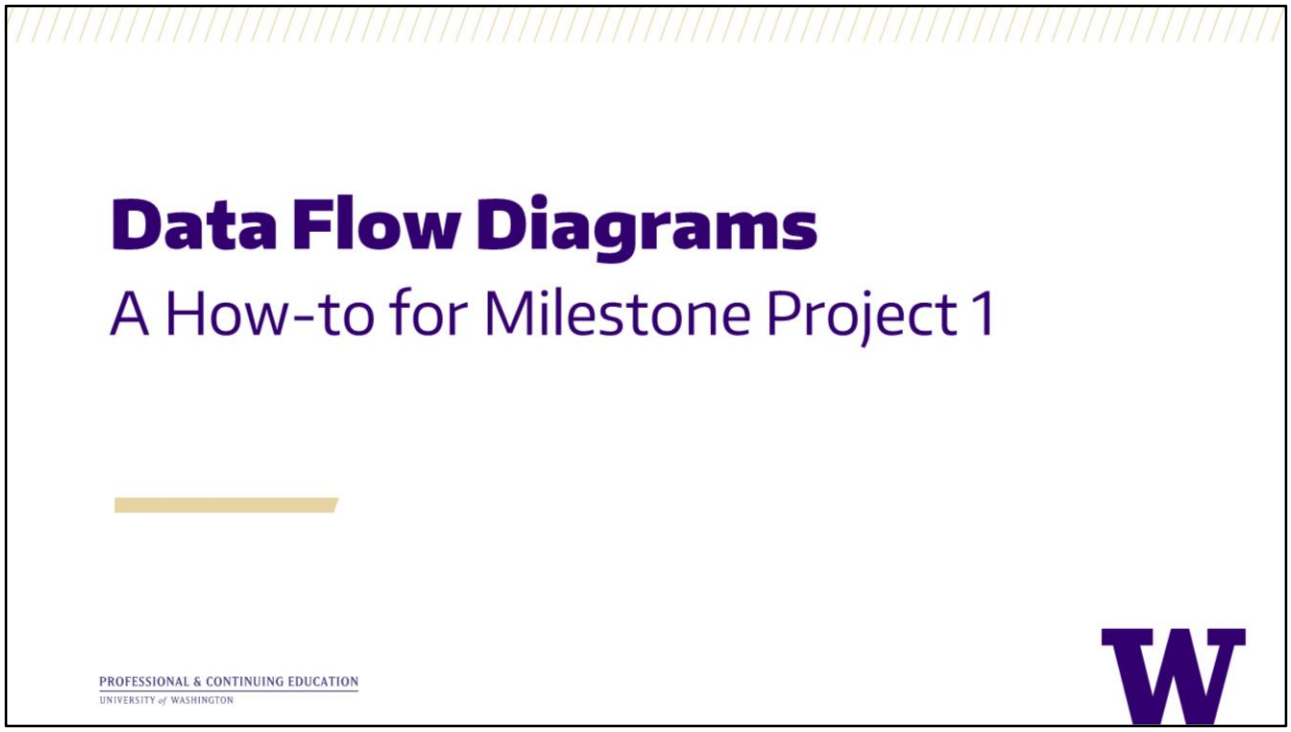
Data Flow Diagrams

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON

Recap from last time.



Data scientists, and everyone else, waits and spends the most amount of time on the ETL (Extract, Transform, Load) process. Data scientist spend the most amount of work in the "Prepare Data for the Model" step. Generally, data scientists and their business managers claim that two thirds of the cycle time is spent on these two processes. In my experience, these two processes take 85% of the data science cycle time and about two thirds of the data scientist's work.



So knowing that preparing the data takes some time, it is important to know how the data will be used. The project for Milestone 1 is a Data Flow Diagram of an interesting Data science problem. This week's homework you'll practice making a Data Flow Diagram, and next week, you will turn in your final version for the Milestone.

Data Flow

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

Four components of a DFD:

- –Terminator
- –Store
- –Process
- –Data Flow

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON

The glue that links these items together is data flow. UML describes a means of diagramming data flows. The diagrams are called data flow diagrams and they have three major components: Processes for data transformation, Stores for data persistence, and terminators that are poorly defined processes and are primarily used to start and stop data flows.

http://en.wikipedia.org/wiki/Data_flow_diagram

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.

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON

These diagrams are called Data Flow Diagrams (DFD). A benefit of DFDs are that they are a defined language in the structured systems analysis and design method (SSADM).
DFD is a particularly good language for describing processes that involve movement and transformation of data.

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.

The complete rectangles are starting or terminating processes that either generate or consume data.
The rectangles without sides are stores, like databases.
An Ellipse represents a process that transforms data.
An arrow represents data.

Data Flow Diagram Example

What are popular tourist locations to photograph?

Let’s take a look at an interesting Data Science problem.

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

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON

Here's a sample of how the Data Flow Diagram works.
First think of an interesting data science task. What are the steps involved?

DFD: Image Aggregation Steps

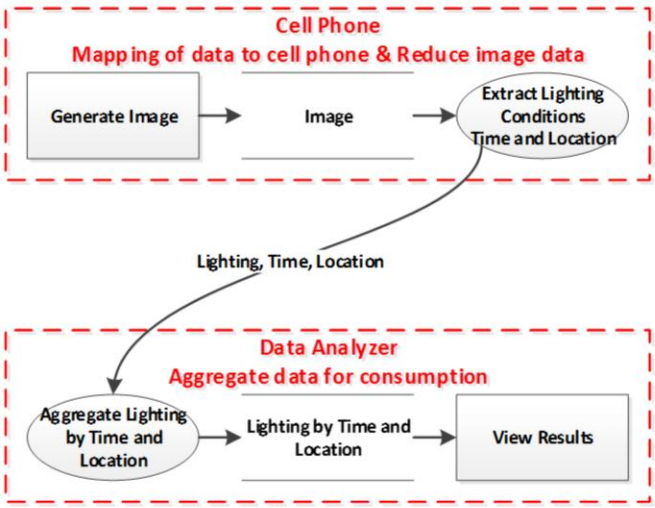
Collect and aggregate cell phone camera images

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

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON

Here are the steps involved in aggregating the image data

Image Aggregation DFD



This is an example of a Data Flow Diagram that you will submit for Milestone 1 project.
Here's what those steps look like.
Remember closed rectangle are stop and starts. Generate and consume data. These are verbs.
The open rectangle is where the data is stored, in this case in the image itself or the combined data of all images. These are nouns.
The oval is a process that happens to the data sore. These are verbs.
The arrows represent the data itself, notice we transform the image data into image meta data.

A closer look at DFD

Understanding the components

Ok, let’s take an even closer look at 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.

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON

See info on slide.
Always working with data, because we're showing how the data is flowing.

Data Flow Practice: DFD Arrow

Which example is correct?

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON

Which one of these arrows represents data?

Data Flow Practice: DFD Arrow

Which example is correct?

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON

Lunch, is a noun, is the data.

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:

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON

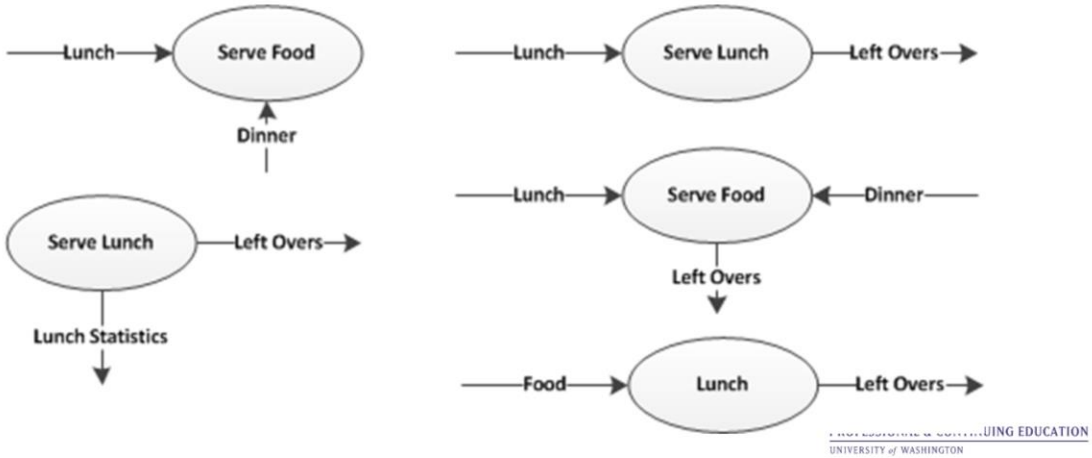
See info on slide

A process is represented by an Ellipse or oval.

A process takes in data from one or more data sources, transforms the data, and then outputs the data.

Data Flow Practice: DFD Process

Which of these are correct?

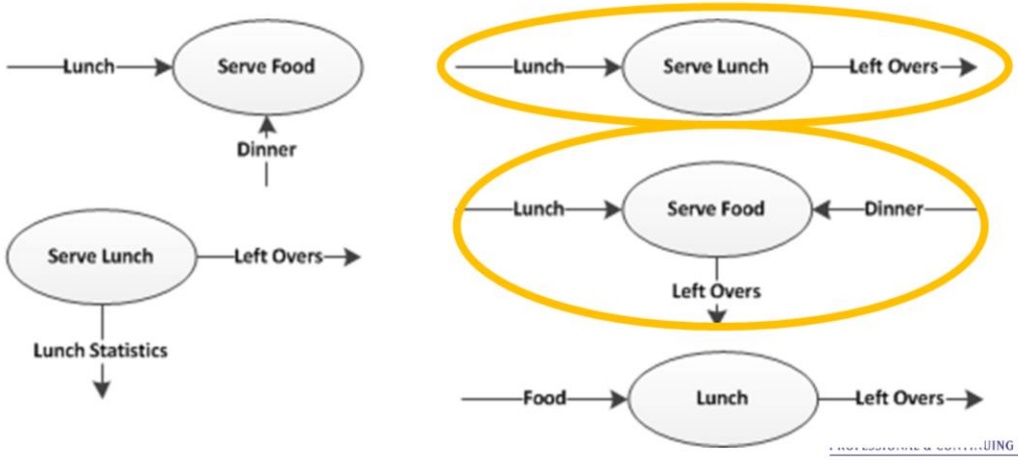


UIING EDUCATION
UNIVERSITY of WASHINGTON

More than 1 of these is correct.

Data Flow Practice: DFD Process

Which of these are correct?



UNIVERSITY of WASHINGTON

Verb in the ellipse.
Remember at least one input and at least one output.

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:

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON

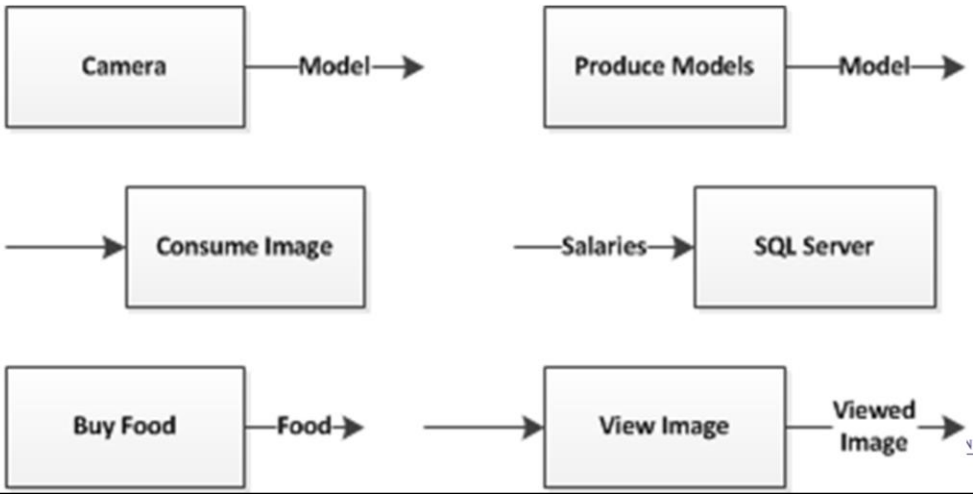
See info on slide

A terminator is represented by a rectangle with all four sides drawn. Also called **sources** (supplies data) and **sinks** (receives data)

A terminator is a process that either generates or consumes data. This process may reference a component like: Get data from Internet or View data in Monitor

Data Flow Practice: DFD Terminator

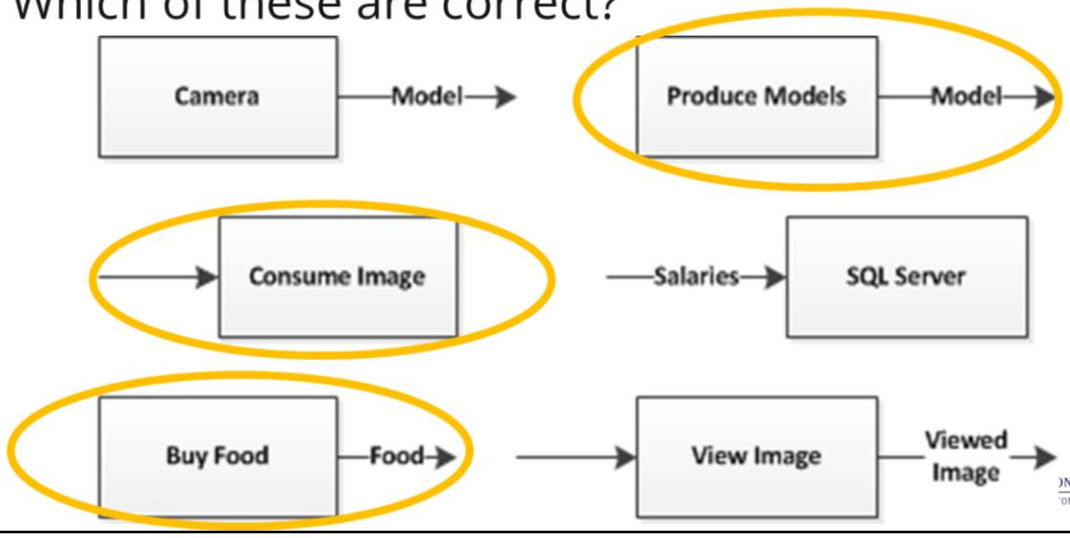
Which of these are correct?



More than 1 is correct

Data Flow Practice: DFD Terminator

Which of these are correct?



Generate or consume data.
Only 1 arrow because it begins or ends the process.

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:

→ Pathology Slides →

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON

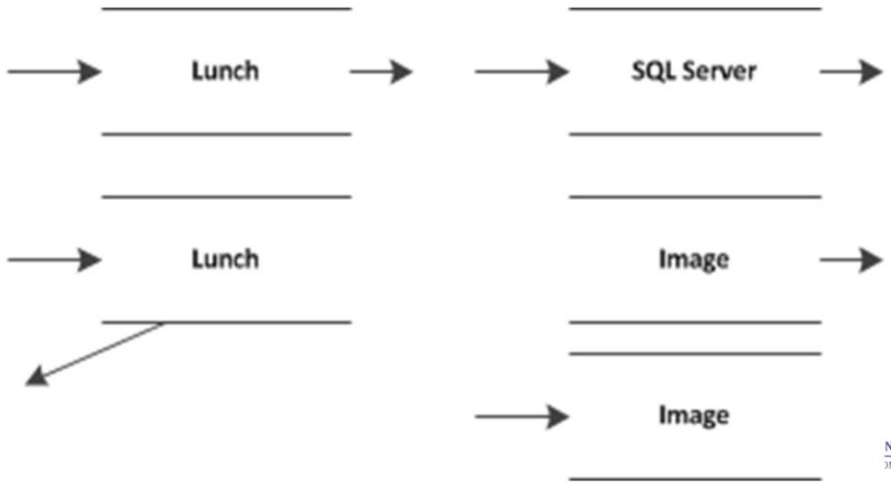
Stores are represented by a rectangle that is missing the right-hand side or both the right- and left-hand sides.
A store is a place where the data is persisted. Typical stores are text files, websites, and relational data bases.

- 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)

Data Flow Practice: DFD Store

Which are correct?

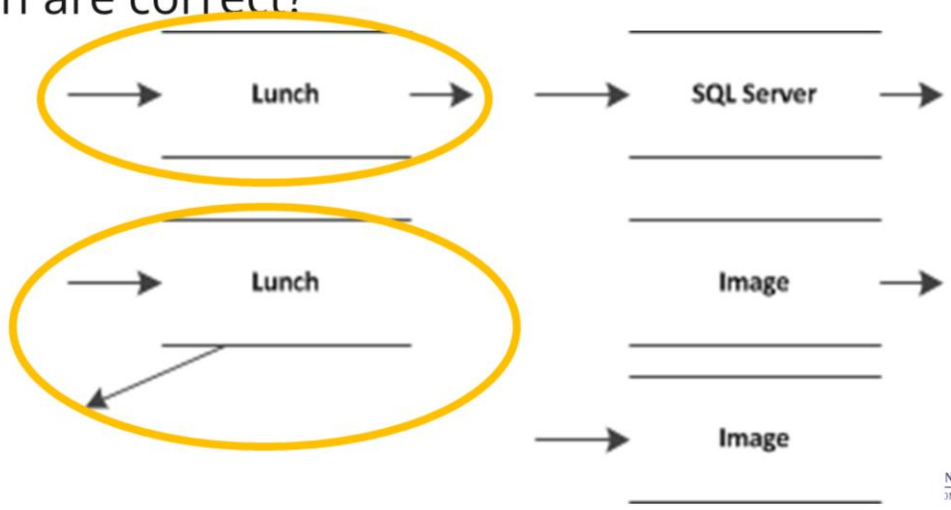


NTINUING EDUCATION
ON

More than 1 is correct.

Data Flow Practice: DFD Store

Which are correct?



Lunch is a type of data. SQL server is the nature of the DATABASE
The language demands that stores have both input(s) and output(s). Often the DFD designer wants to start or end with a data store, like it is shown for the two examples labeled “Image”. This is not allowed! Instead prepend or append a terminator. The prepended terminator might be labeled: “Generate Images”. The appended terminator might be labeled: “Present Images”.

DFD: Digital Pathology

An Example

Let’s try another 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.

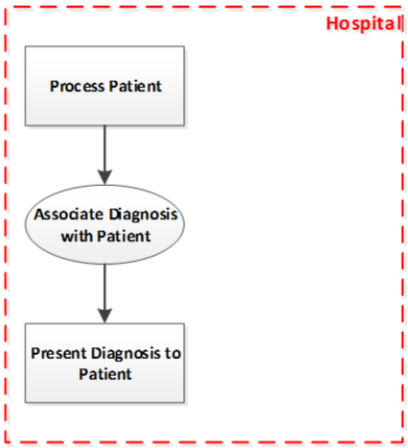
Given the scarcity and cost of pathologists, it is not possible to routinely screen for these blood disorders. We would like to find an automated way of diagnosing such disorders.

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

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON

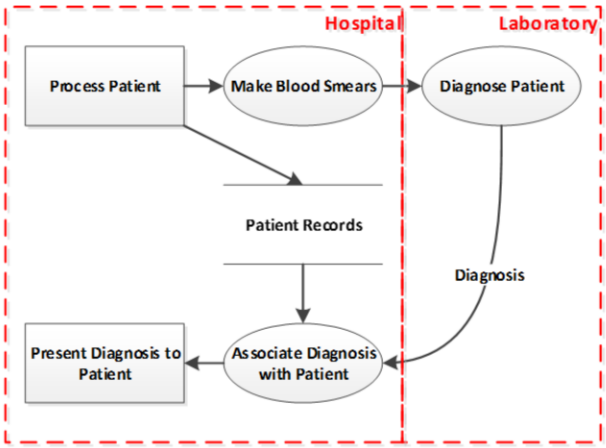
Digital Pathology is a way to screen for blood disorders based on the image of the cells.

DFD Example: Digital Pathology



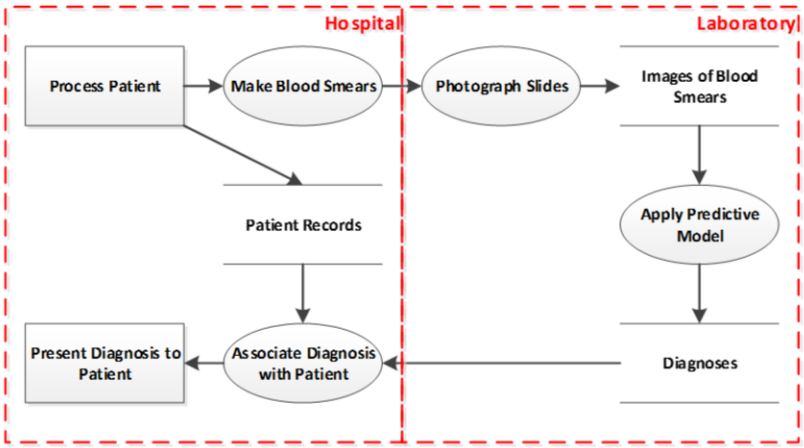
On the surface, at the hospital it seems fairly simple:
Generate data by processing the patient
Associate the diagnosis with the patient is the action that is performed, transforming raw data into diagnosis data
And then we consume that data by presenting it to the patient.

DFD Example: Digital Pathology



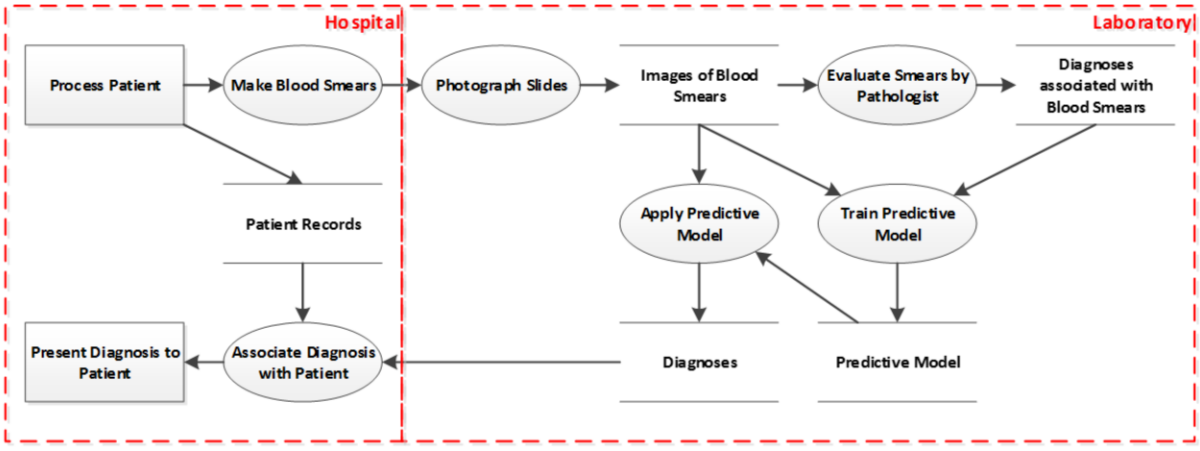
But there's more involved. The Laboratory is responsible for transforming some of that data. Constructing a diagram may take a few iterations. When complete it is easy to read and understand. As you build the diagram, think about what kind of data is involved, where does that data come from, how does it change?
See slide

DFD Example: Digital Pathology



The lab process is even more involved when we want to digitize this process because they use images of blood.
See slide image

DFD Example: Digital Pathology



Here's where data science comes into play with training those images with a live human and then modeling that with the data.
Part 2 of this course deals with transforming that data into useable forms to create models in Part 3.
See slide image

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.

Summary Slide.
HW this week is to submit a draft of this diagram. Most students do not get it right the first time and need additional feedback to think about what they are trying to describe with the flow of data.