

# **Data Science: Tools & Process**

## **Week 3**



# Agenda

---

- **Announcements**

- Some of today's resources (\*.pdf, \*.py, etc.) will have the numeral 4 instead of 3
- Discussions: Please post at least once a week in forum. Should we create a LinkedIn group with the other class?
- Milestones: Project 1. Redo the DFD; Project 2. Select a dataset from Milestone 2 Overview

- **Review DFD (Last Week's Assignment)**

- **Lesson 03 Quiz 0 (Quiz on Sparse Matrices and DFD)**

- **Aberrant Data Lab I (L04-A-1-RemoveOutliers.py)**

- **Break**

- **Aberrant Data Lab II (L04-A-2-ReplaceOutliers.py & L04-A-Outliers.pdf)**

- **Lesson 03 Quiz 1 (Quiz on Outliers)**

- **Data Type Lab (L04-B-1-DataTypes.py & L04-B-DataTypes.pdf)**

- **Break**

- **Remove Missing Values Lab (L04-B-2-RemoveMissing.py)**

- **Lesson 04 Quiz 2 (Quiz on Missing values and Data Types)**

- **Boolean Indexing (Time Permitting)**



# **Data Flow Diagrams**

## **Review for Milestone Project 1**



# **DFD: Digital Pathology**

**An Example**

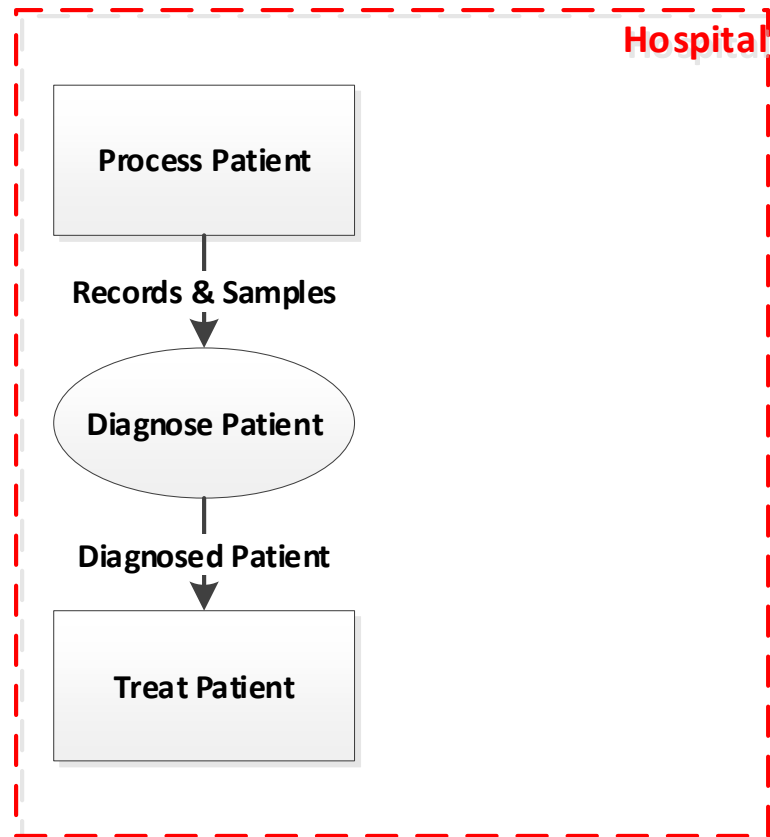
**W**

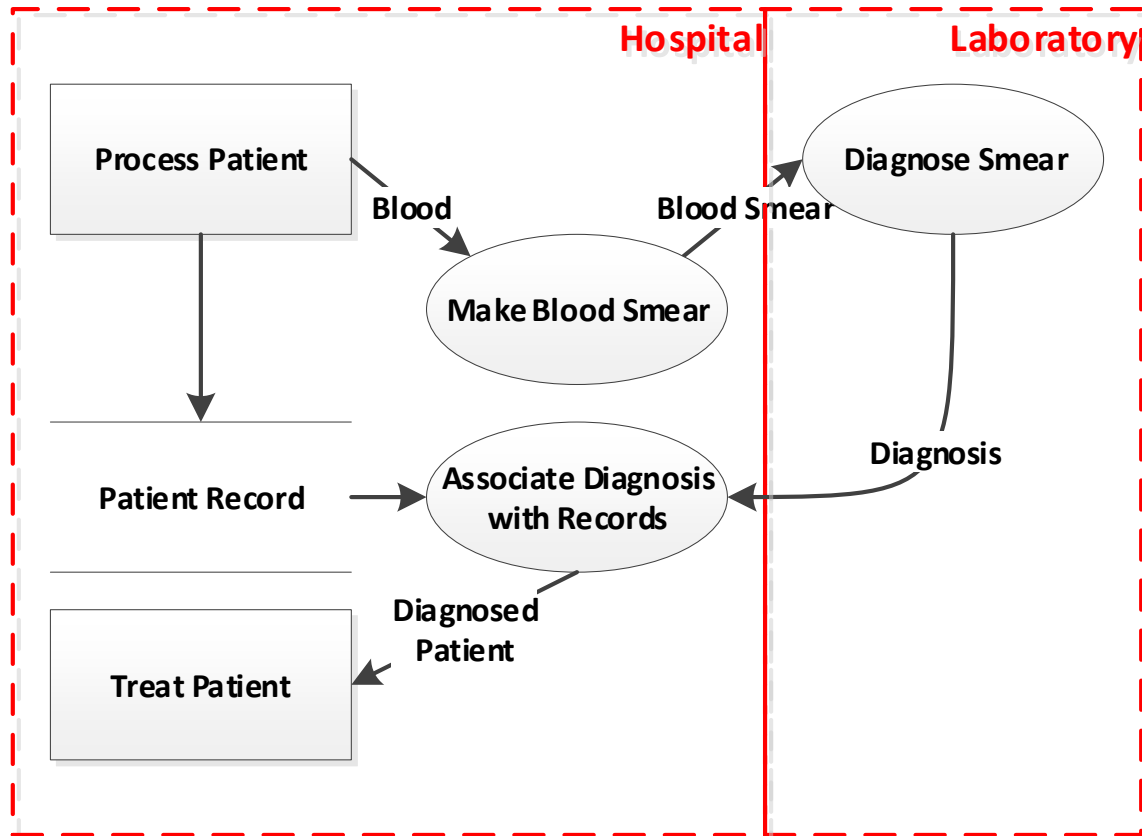
# DFD Example: Digital Pathology

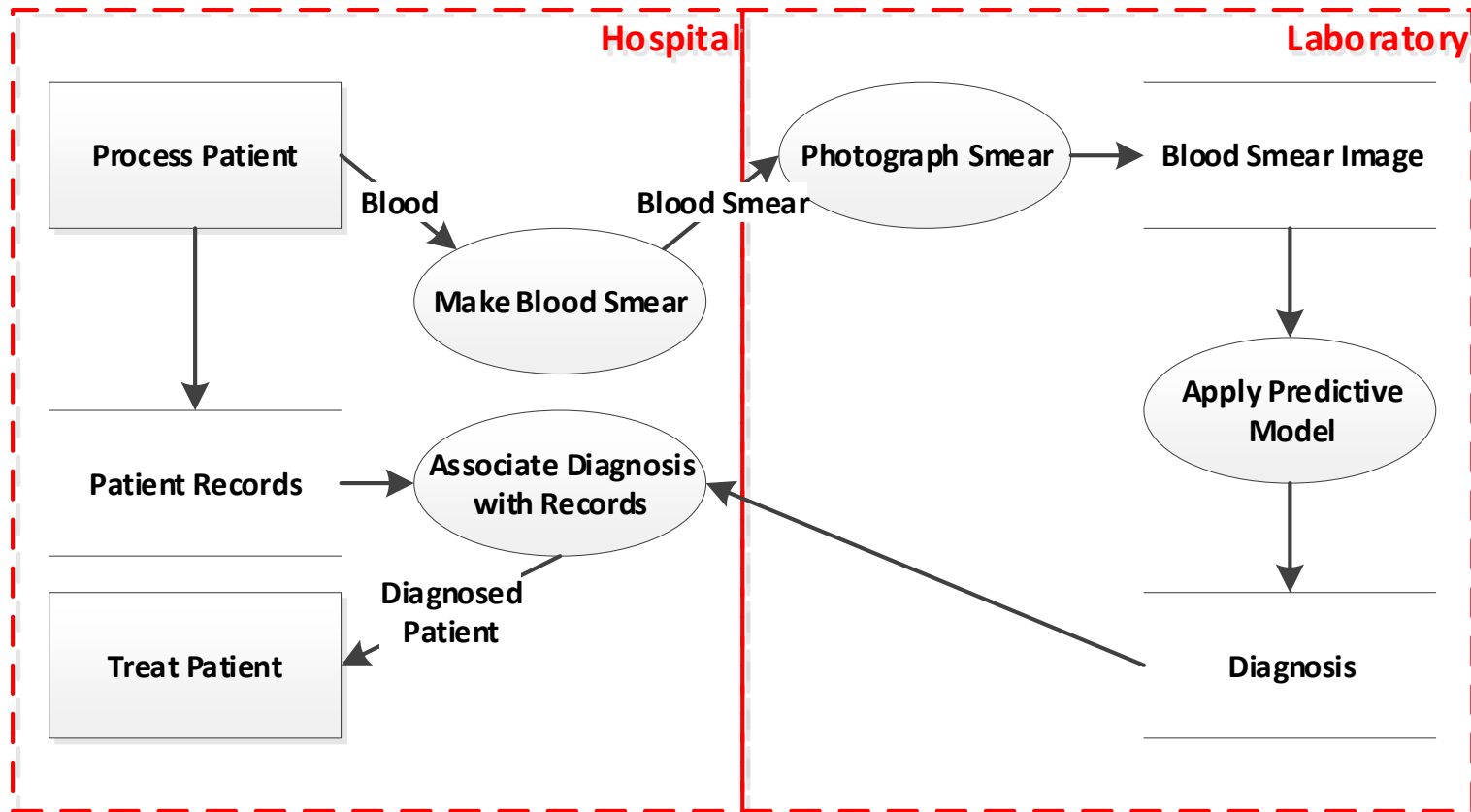
Many blood disorders manifest themselves through easily recognizable morphological changes, but the affected cells may be as few as one in a hundred thousand.

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.

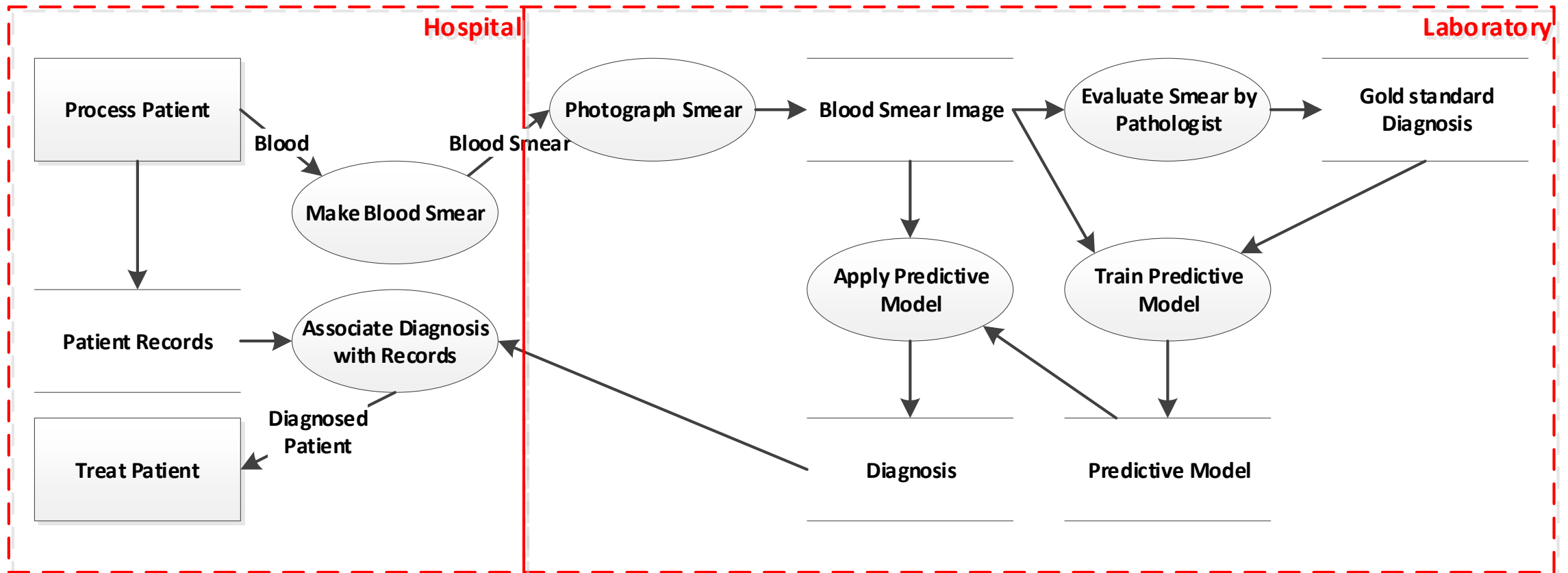
We use a pathologist to score aberrant cells and correlate these findings with shape characteristics determined by image segmentation.











# 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.

# Lesson 03 Quiz 0

---

## Quiz on Sparse Matrices and DFD

Remember last week's punch line:  
Datasets are multi-dimensional sparse matrices

# Aberrant Data Lab I

- Files for Lab:
  - L04-A-1-RemoveOutliers.py

# Break

Break

# Aberrant Data Lab II

- Files for Lab:
  - L04-A-2-ReplaceOutliers.py
  - L04-A-Outliers.pdf

# Lesson 03 Quiz 1



## Quiz on Outliers and Missing Values

# Data Type Lab

- Files for Lab:
  - L04-B-1-DataTypes.py (We need to understand data types in the context of missing values)
  - L04-B-DataTypes.pdf



# Break



## Break

# Remove Missing Values Lab



- File for Lab:
  - L04-B-2-RemoveMissing.py

# Lesson 03 Quiz 2



## Quiz on Data Type

# Boolean Indexing?