DATA 3464: Fundamentals of Data Processing

Intro to the course

Charlotte Curtis

January 6, 2026

Meet your instructor

Name: Charlotte Curtis

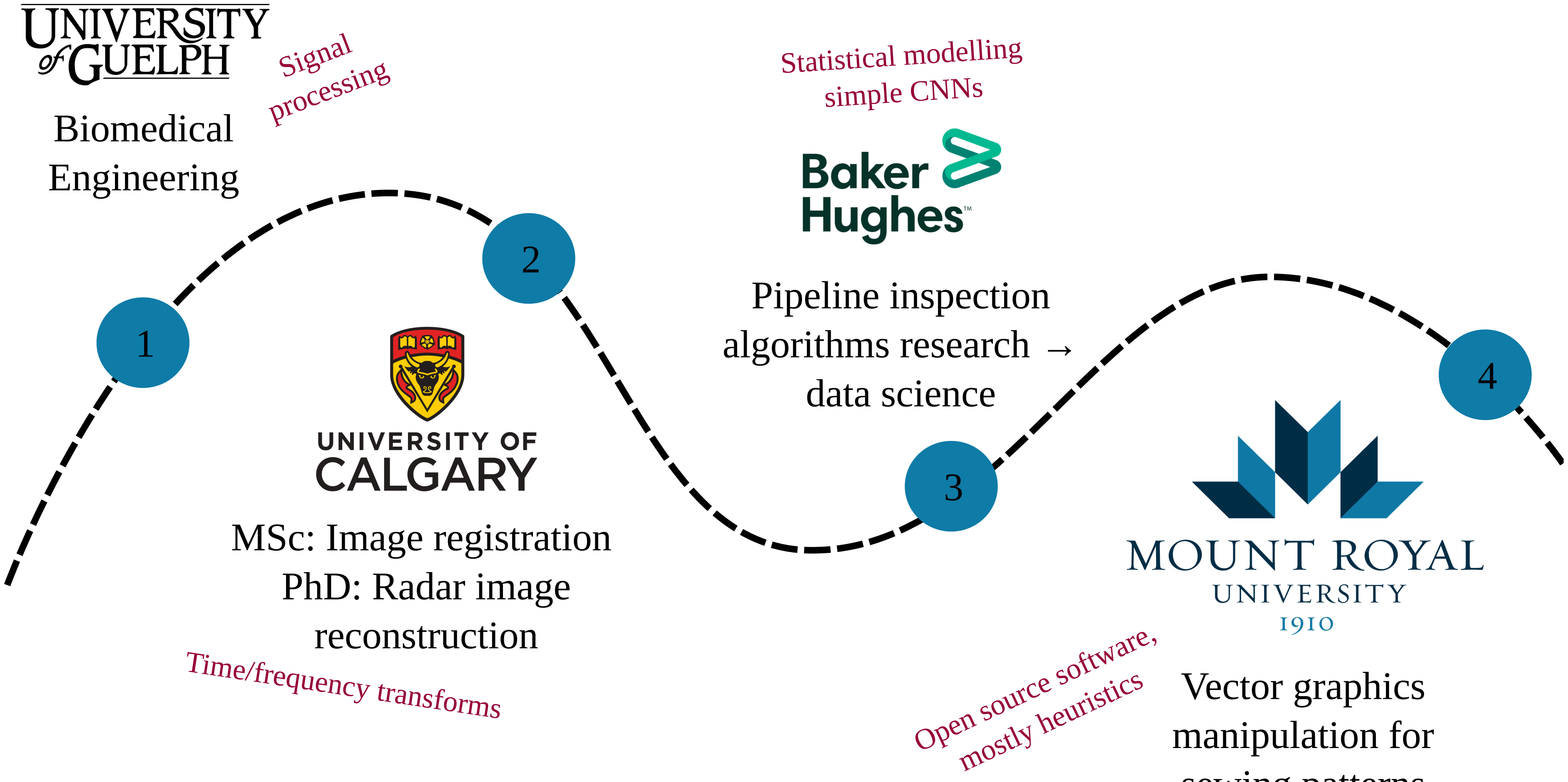
Pronouns: She/her

Office: B102-4

Email: ccurtis@mtroyal.ca

Office hours: Book here

My Background



1

UNIVERSITY of GUELPH

Biomedical Engineering

Signal processing

2

UNIVERSITY OF CALGARY

MSc: Image registration
PhD: Radar image reconstruction

Time/frequency transforms

3

Statistical modelling
simple CNNs

Baker Hughes™

Pipeline inspection algorithms research → data science

4

MOUNT ROYAL UNIVERSITY 1910

Vector graphics manipulation for sewing patterns

Open source software,
mostly heuristics

Another new class!

> *This course introduces techniques for ethically and responsibly* ***wrangling*** *and manipulating datasets to make them appropriate for addressing the question at hand. Topics may include cleaning and transforming data, integrity and quality measures, common file formats, feature selection and engineering, and generating features from unstructured sources such as text and images.*

Grade Assessment

Component	Weight
Tutorial exercises	10%
Assignments	30%
Midterm exam	25%
Final exam	35%

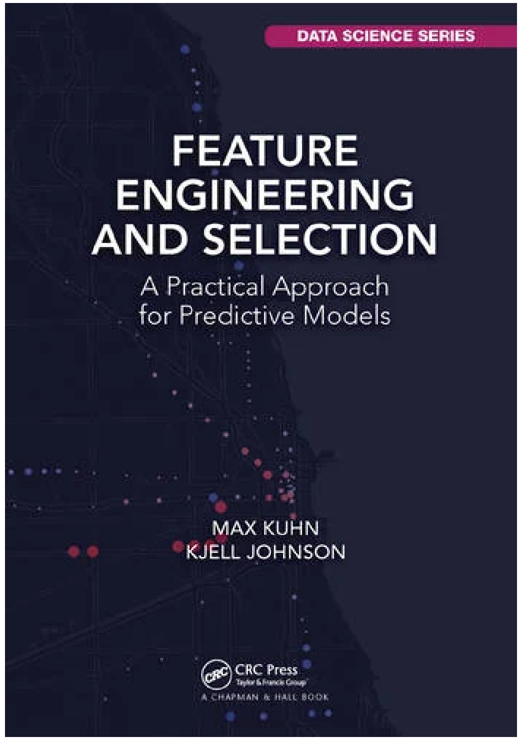
Bonus marks may be awarded for *substantial* corrections to materials, submitted as pull requests

Source repo: https://github.com/mru-data3464/w26

Rendered at: https://mru-data3464.github.io/w26

Textbook(s)

- http://www.feat.engineering/
- Additional texts/websites as needed
- All the documentation!
  - Pandas
  - Numpy
  - SciPy
  - Scikit Learn
  - Matplotlib
- ... or the R tidyverse



DATA SCIENCE SERIES

FEATURE ENGINEERING AND SELECTION

A Practical Approach for Predictive Models

MAX KUHN
KJELL JOHNSON

CRC Press
Taylor & Francis Group
A CHAPMAN & HALL BOOK

> *Don't just rely on AI summaries!*

Speaking of AI...

In this course (and others, and your career), you will need to know:

- **What** to do, and **why**
- **How** to do it

(also when and who)

> *Which of these things seem appropriate for AI assistance?*

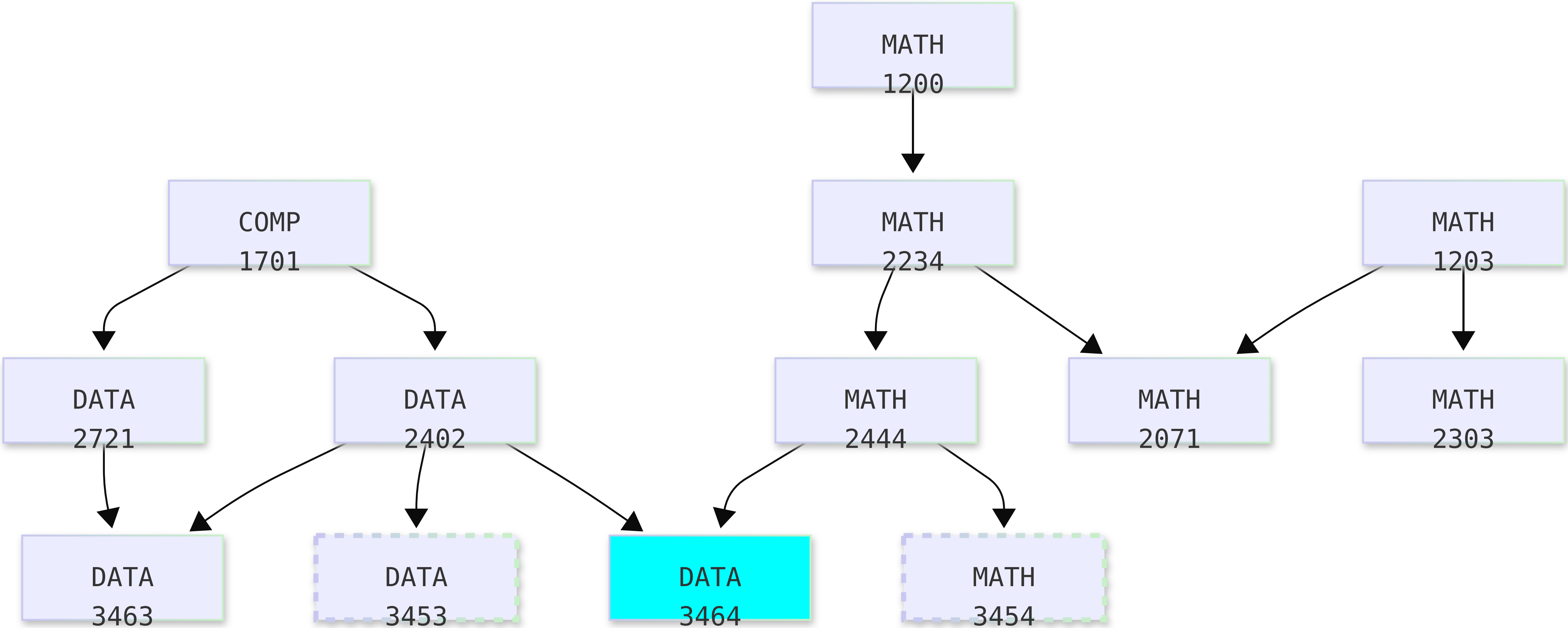
The plan - before Reading Week

Week	Topic	Chapter (ish)
1	Review and overview	1-2
2	Exploring data, sampling, splitting	3-4
3	Representing categorical data	5
4	Numeric transformations, dimensionality reduction	6
5	Dealing with missing values	7-8
6	Feature selection	10

The plan - after Reading Week

Week	Topic
7	Midterm
8	Extracting data from text
9	Image representation and processing
10	Data labelling and augmentation
11	Processing pipelines
12	Supervised and unsupervised learning
13	Project presentations, buffer time

Core courses so far



What do you know about...

- Various probability distributions
- Linear and logistic regression
- Data quality measures
- Data stewardship best practices
- Document parsing, web scraping, audio/video feature detection
- Linear algebra and array programming
- Prediction tasks: classification and regression
- Clustering and anomaly detection
- Evaluation metrics
- Basic data visualization (scatter plots, histograms, etc)

What do you want to know about?

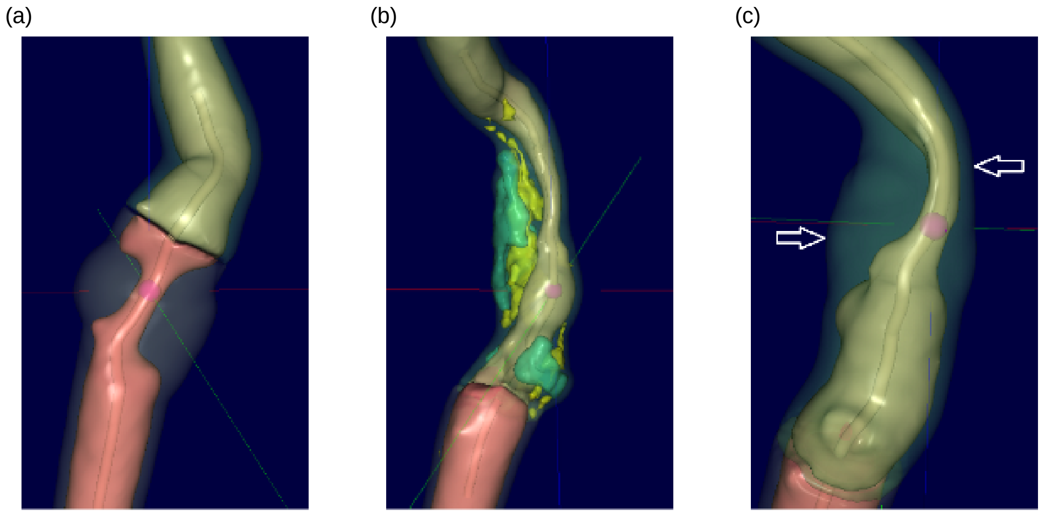
Examples of Subject Matter

- Finance
- Real estate
- Transportation
- Climate
- Politics
- Biology
- Chemistry
- Malware

Examples of Data types

- Unstructured text
- Structured text (e.g. csv, HTML)
- PDF
- Word documents
- Images
- Audio
- Video

Case study: risk of ischemic stroke



- Arterial stenosis can predict risk
- Plaque composition plays a role
- Features extracted from CT images
- Other risk factors (demographics, lifestyle) added to dataset

Chapter 2:

http://www.feat.engineering/stroke-tour

> *Many decisions in the data analysis process are subjective - I will often make different decisions than the textbook*

From data to prediction

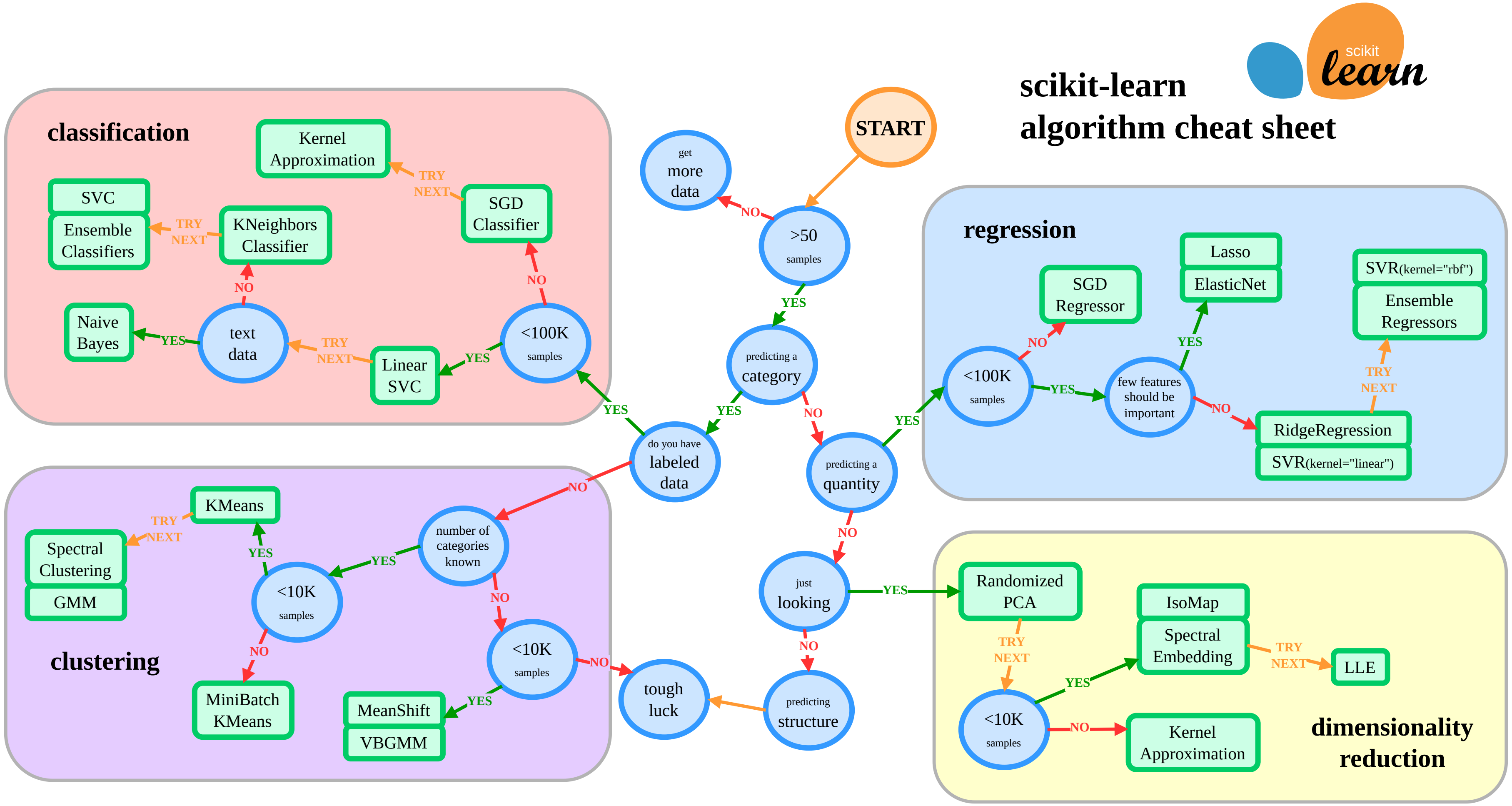
1. Understand the problem and define the task
2. Collect, anonymize and organize the data
3. Extract features
4. Explore the dataset
5. Select a model and preprocess
6. Train the model
7. Evaluate, fine-tune, iterate
8. Deploy and maintain your system

Adapted from Aurélien Géron's ML Project Checklist

Applied to the stroke example

1. What is the problem? What do we need to do?
2. (Collect, anonymize and organize the data) - Done for us
3. (Extract features) - Done for us
4. **Explore the dataset**
    - A critically important component, DO NOT OFFLOAD TO AI
    - This can even be where the data sciencing stops and we jump straight to visualizations and communicating insights!
    - Check out Data for Good case studies

5. Select a model and preprocess



scikit-learn
algorithm cheat sheet

Source: scikit-learn documentation

7. Evaluate, fine-tune, and iterate

- In my example, I jumped straight to testing on the held-back test set
- This is a terrible idea! We have no confidence that the model actually worked. We could be:
    - overfitting to the training data
    - making incorrect assumptions about the data
    - applying inappropriate transformations, or missing some
    - using the wrong model altogether

> *Validation needs to happen before the final testing*

Terminology for evaluation (classification)

- **True positive**: predicted positive, label was positive (TP) ✓
- **True negative**: predicted negative, label was negative (TN) ✓
- **False positive**: predicted positive, label was negative (FP) ✗ (type I)
- **False negative**: predicted negative, label was positive (FN) ✗ (type II)
- **Accuracy** is the fraction of correct predictions, given as:

$$\text{accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Precision and recall

- **Precision**: Out of all the positive **predictions**, how many were correct?

$$\text{precision} = \frac{TP}{TP + FP}$$

- **Recall**: Out of all the positive **labels**, how many were correct?

$$\text{recall} = \frac{TP}{TP + FN}$$

- **Specificity**: Out of all the negative **labels**, how many were correct?

$$\text{specificity} = \frac{TN}{TN + FP}$$

Confusion matrix

	Predicted Positive	Predicted Negative
True Positive	TP	FN
True Negative	FP	TN

- The axes might be reversed, but a good predictor will have strong diagonals
- There's also the **F1 score**, or harmonic mean of precision and recall:

$$F1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

Check out the Wikipedia page for more ways of describing the same information

ROC Curves

- The receiver operating characteristic curve is a plot of the **true positive rate** (recall or sensitivity) vs. **false positive rate** (1 - specificity) as the detection threshold changes
- The diagonal is the same as random guessing
- A perfect classifier would hug the top left corner

> *Fun fact: the name comes from WWII radar operators, where true positives were airplanes and false positives were noise*

Coming up next

- Lab: basic regression, show me where you're at
- Lectures: exploratory data analysis
  - Summary statistics
  - Basic visualizations
  - When and how to split your dataset