Introduction to Data Science

Welcome!

Today we are going to get all of the administrative details. Here is a quick outline:

- – go through the syllabus
- – give a brief overview of the course material
- – tell you a bit about myself
- – install software
- – answer any additional questions

Syllabus

See syllabus posted on the course website.

Class Structure

Most course meetings are organized as follows:

— **homework I:** carefully read any posted notes on the website and formulate questions for the next class (30-90 minutes)
— **course form**: fill out at the start of class (< 1 minute)
— **discussion / slides:** review readings or previous notebooks, discuss any questions, perhaps start classwork together (15-30 minutes)
— **classwork**: work individually or in small groups to answer questions in the form of programming notebooks (45-60 minutes)
— **homework II**: finish classwork or review posted solutions (0-60 minutes)

All materials can be found on the course website.

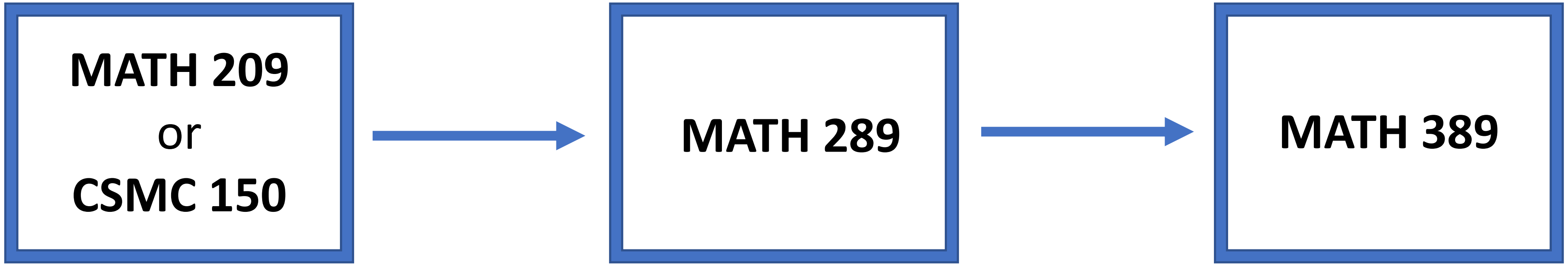
Data Science

What exactly is data science?

Data science is an interdisciplinary field concerned with drawing knowledge from data and communicating those results to various audiences. Data science needs to be learned *by doing* data science.

By the end of the semester, you will feel confident collecting and analyzing datasets from a variety of fields. You will be able to use these skills to address data-driven problems in a wide range of application domains.

Course Sequence


MATH 209
or
CSMC 150

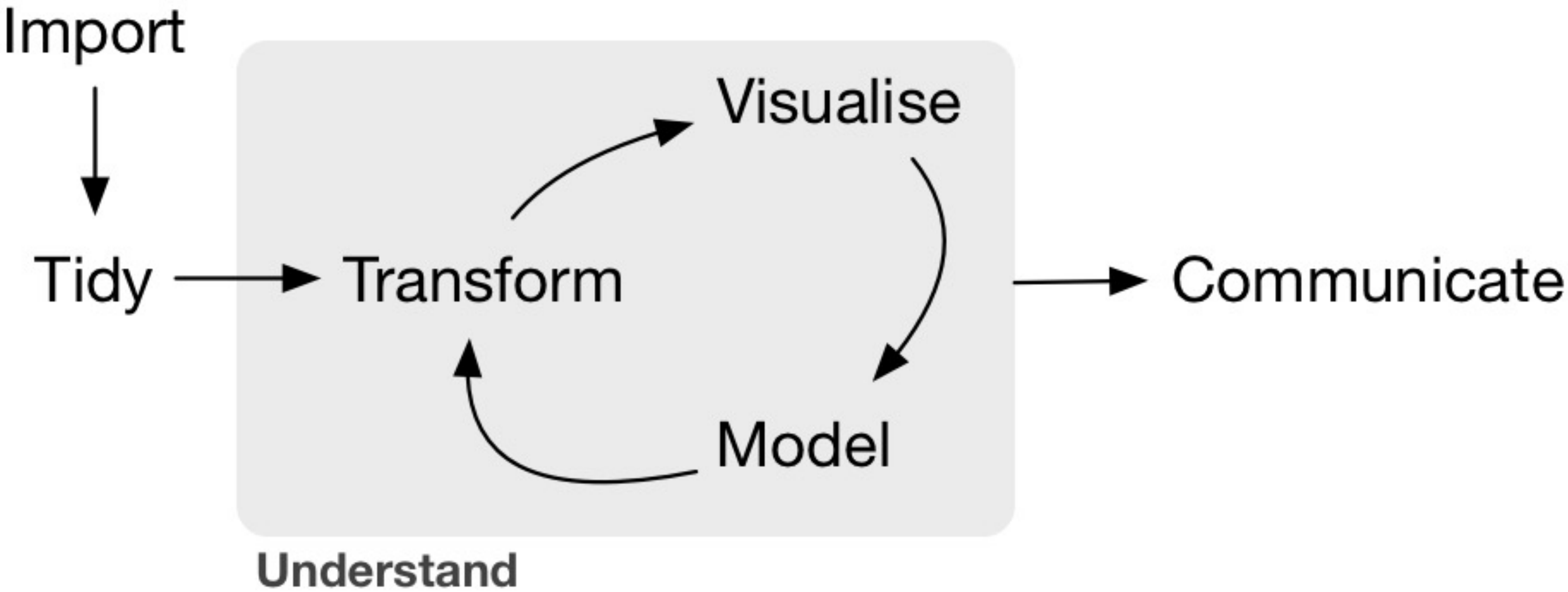
MATH 289

MATH 389

This course assumes you have some experience using code to manipulate data. It makes no assumptions about your knowledge of any specific programing language or knowledge of statistical inference.

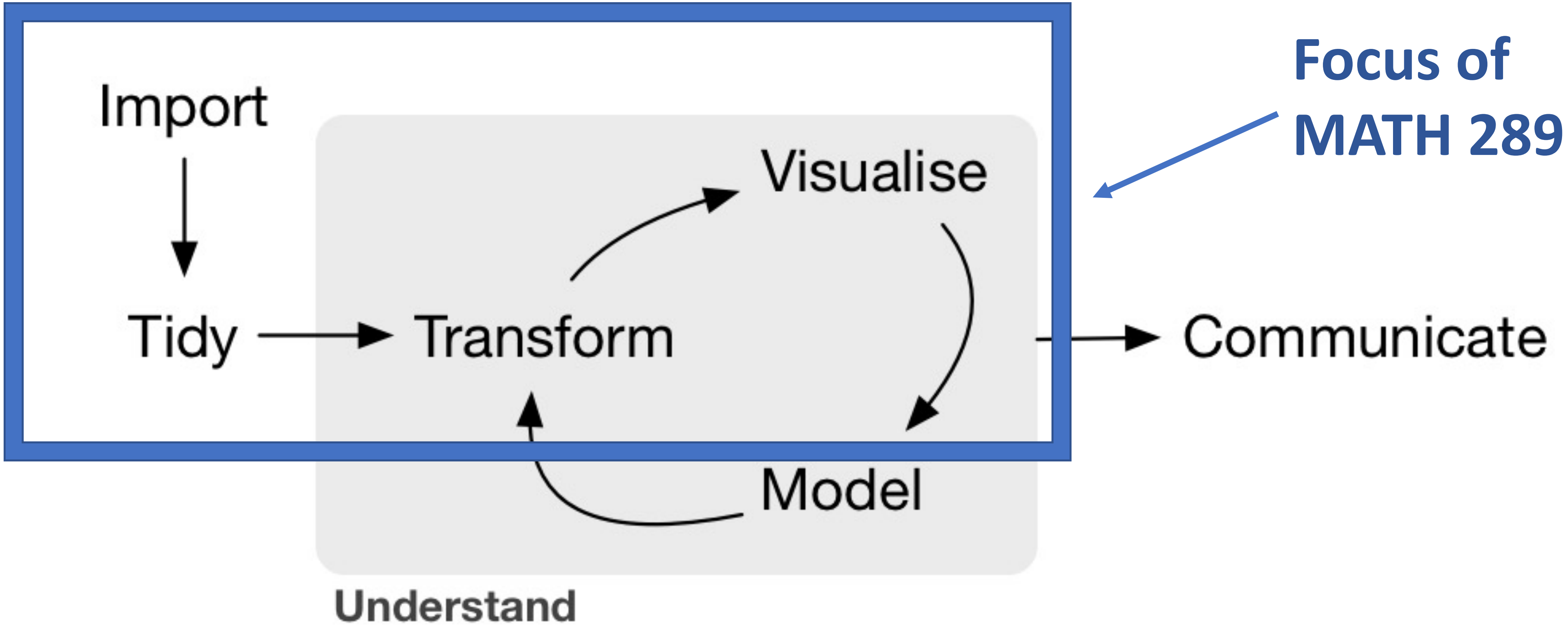
The class is designed as a year-long sequence paired with MATH 389; I strongly suggest taking both during the same academic year if possible.

Data Science Pipeline



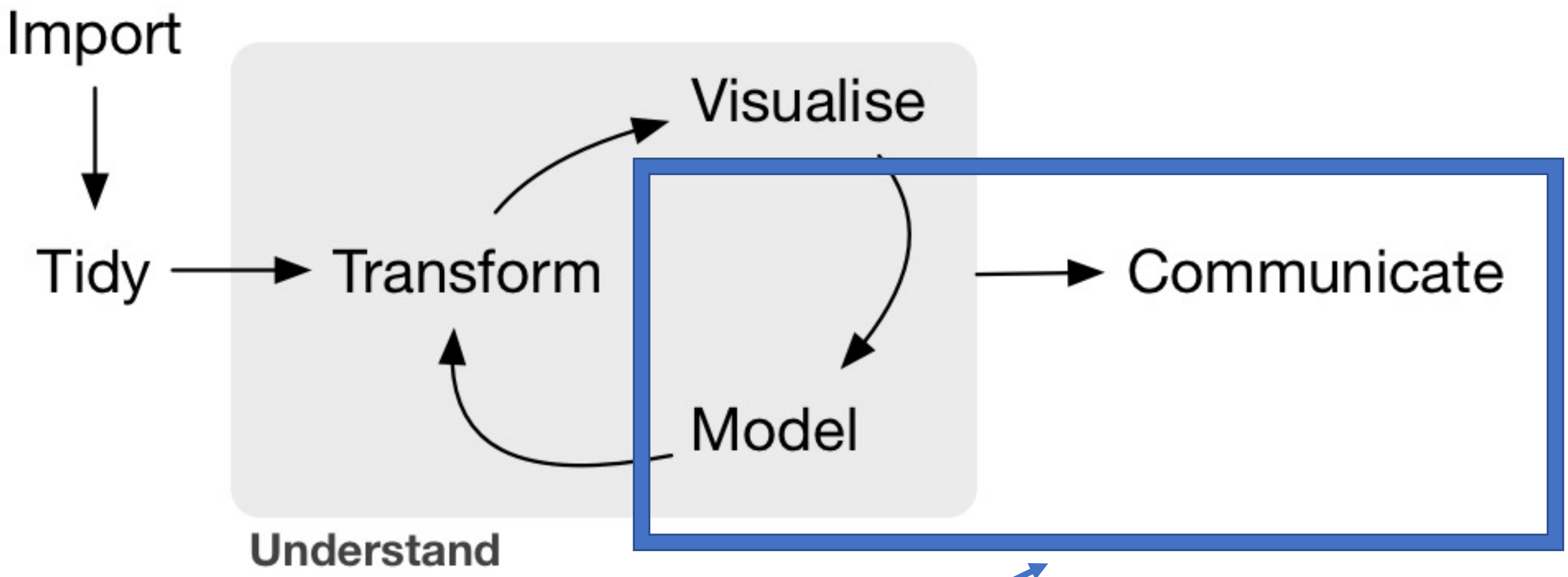
Grolemund & Wickham R4DS Illustration

Data Science Pipeline



Grolemund & Wickham R4DS Illustration

Data Science Pipeline



Grolemund & Wickham R4DS Illustration

Focus of MATH 389

Programming

There are several different programming languages for data science. By far the two most popular are R and Python.

We will be using R this semester but will learn a version that is easily adapted to other languages such as Python.

In the last week I will demo the use of Python and JavaScript based on the the class material.

Course Topics

See tentative topics posted on the course website.

About Me

- From New England: born in Maine, school in MA, ME, CT
- Moved to Richmond in 2016
- Research on large text and image datasets in linguistics and cultural studies

Photogrammar

R Journal

ACLS

Distant Viewing Lab

THE ANDREW W. MELLON FOUNDATION

NATIONAL ENDOWMENT FOR THE HUMANITIES

COLLEGIUM DE LYON
Université de Lyon

About Me

- Lots of industry experience in DS:
  - IBM (Healthcare)
  - Travelers (Insurance)
  - DARPA (social media)
  - AT&T (location analytics)
  - Telperian (pharmaceuticals)

About Me

- I have two Shih-Tzus: Roux and Sargent
- Roux is often in my office; please come say hello