Introduction

Data Management an

Kylie A. Bem

Northeastern Uni
Khoury College of Comput

Northeastern U

Basics

- DS 5110 Introduction to Data M
  - Sec 3: Tue 11:45 am - 1:25 pm, Th
  - Sec 4: Thu 6:00 pm - 9:20 pm
  - Dates: Jan 8, 2023 - Apr 27, 2023
  - Course website:
    - https://kuwisdelu.github.io/ds5110-spring24.
    - Link to all relevant resources available on C

Intros: Kylie Arie

- PhD Statistics, Purdue University
- Research includes:
  - Bioinformatics (mass spectrometry imaging)
  - Statistical computing languages (R development)
- Software (R packages):
  - "Cardinal" - mass spectrometry imaging tools
  - "matter" - out-of-memory computing
- 2015 John M. Chambers Statistical Software award for development of "Cardinal" package
- Published author of short science fiction and fantasy

Intros: TA

"Keep an open and curious mindset"

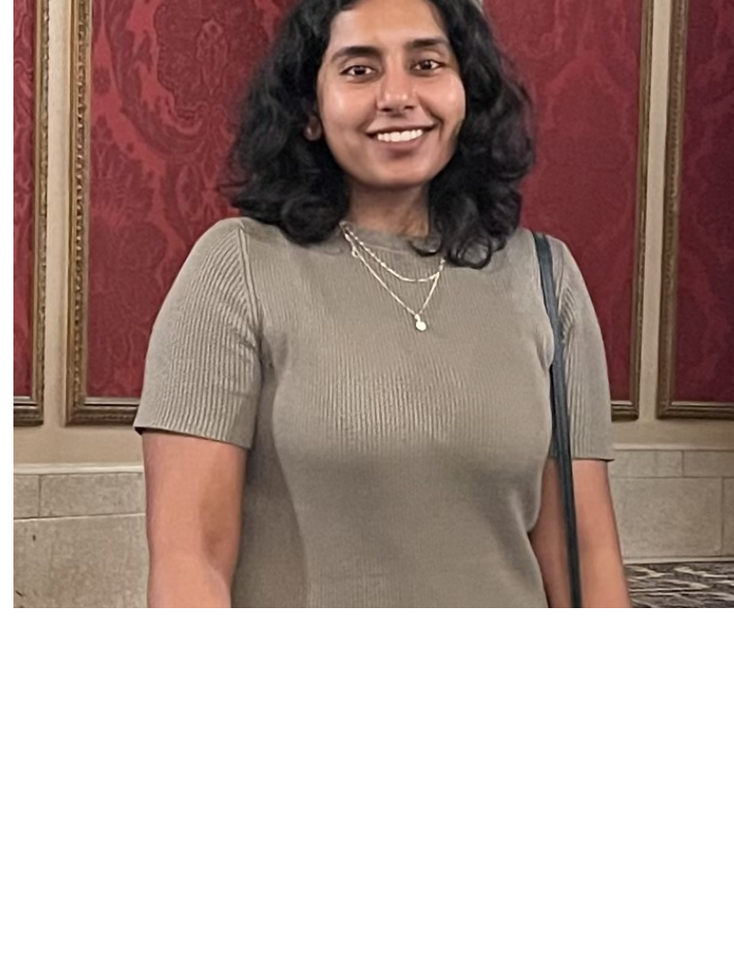
Jayatha

"Data science is about more than co

Sakshi

"Don't hesitate to ask questions"

"Engage in class and



Shubham

"Embrace challenges with curiosity and perseverance"

Goals for to

- Course expectations ar
- Tools and resources
- Introduction to data sci
- Introduction to R

COURSE EXPEC
AND POLI

Course mate

- Syllabus:
  - https://kuwisdelu.github.io/ds5110
- Piazza:
  - https://piazza.com/northeastern/sp
- Canvas:
  - https://northeastern.instructure.co

Zoom meet

- Classes will be hosted synchron
  - Meeting links for each class can be found
  - Meetings will be recorded and made avai
- Synchronous attendance is enc
  - Projects must be presented during the sc
  - Quizzes will have a limited time frame to

Announcements, questio

- Outside office hours, all course-r correspondence should use **Piaz**
  - https://piazza.com/northeastern/spring2024
  - Course material and office hours schedules
  - Use "Q&A" to ask asynchronous questions
  - Send a private note for course-related quer
- Do not email for course-related
- How to ask a good question:
  - https://stackoverflow.com/help/how-to-ask

Office hou

- Virtual office hours will be hel
  - Download at https://teams.northeastern
  - Live online during our scheduled office h
  - Chat or video call
- Or schedule a video call by ap
- In-person appointments availab

Assignments and

- Assignments and grading will us
  - https://northeastern.instructure.com
  - Semester schedule (subject to change)
  - Assignments must be submitted via Canv
  - Grades will be distributed via Canvas
- Do not email assignments — th

Grade breako

- Homework: 40%
- Quizzes: 30%
- Project: 30%

Homewor

- 5-6 individual homewor
- Practice on real-world d
- Submit on Canvas
  - R Markdown source code
  - PDF report (generated from R
- Due every ~1-2 weeks

Quizzes

- 2-3 cumulative quizzes
- Complete online during
- Test your conceptual ur
- Will replace a class mee

Project

- Teams of 2-5 students
- Complete a term project u
- Choose a team and submi
- Present in-class during fina
- Submit a final written repo
- Detailed guidelines will be discussed in class

Peer review

- Provide feedback for your
- Receive feedback on your
- Due after projects are sub
- Rubrics will be shared on discussed in class

Grading poli

- Late assignments will not rec
  - Assignments late by more than 1 day
  - Submit early to avoid technical issues
- Extensions may be given on a
  - Request at least 48 hours in advance
  - Require a reasonable justification
- Petitions for re-grades must b
  - Must be submitted within 1 week and
  - *The new grade may be lower than the o*

General poli

- Academic integrity policy
  - http://www.northeastern.edu/osccr/academ
  - Please remember you are here to learn, no
  - Use work with permission; cite it properly;
- Title IX policy
  - https://www.northeastern.edu/ouec/
  - Northeastern is committed to preventing c
  - You may talk to me, but I am a mandatory
  - Confidential resources are available if you n

General policies

- Please be kind and respectful to
  - Use other people's requested names and p
  - Be considerate of other people's identities a
  - Keep the classroom a safe and healthy envi
- Getting help
  - Please reach out as early as possible if you a
  - I am less able to make accommodations clo
  - Northeastern's WeCare office is a resource
    - https://studentlife.northeastern.edu/we-care/

TOOLS AND RE

Software

- Install R from CRAN
  - https://cran.r-project.org
- Install RStudio
  - https://rstudio.com/products/rs
- Install LaTeX
  - https://www.latex-project.org/g

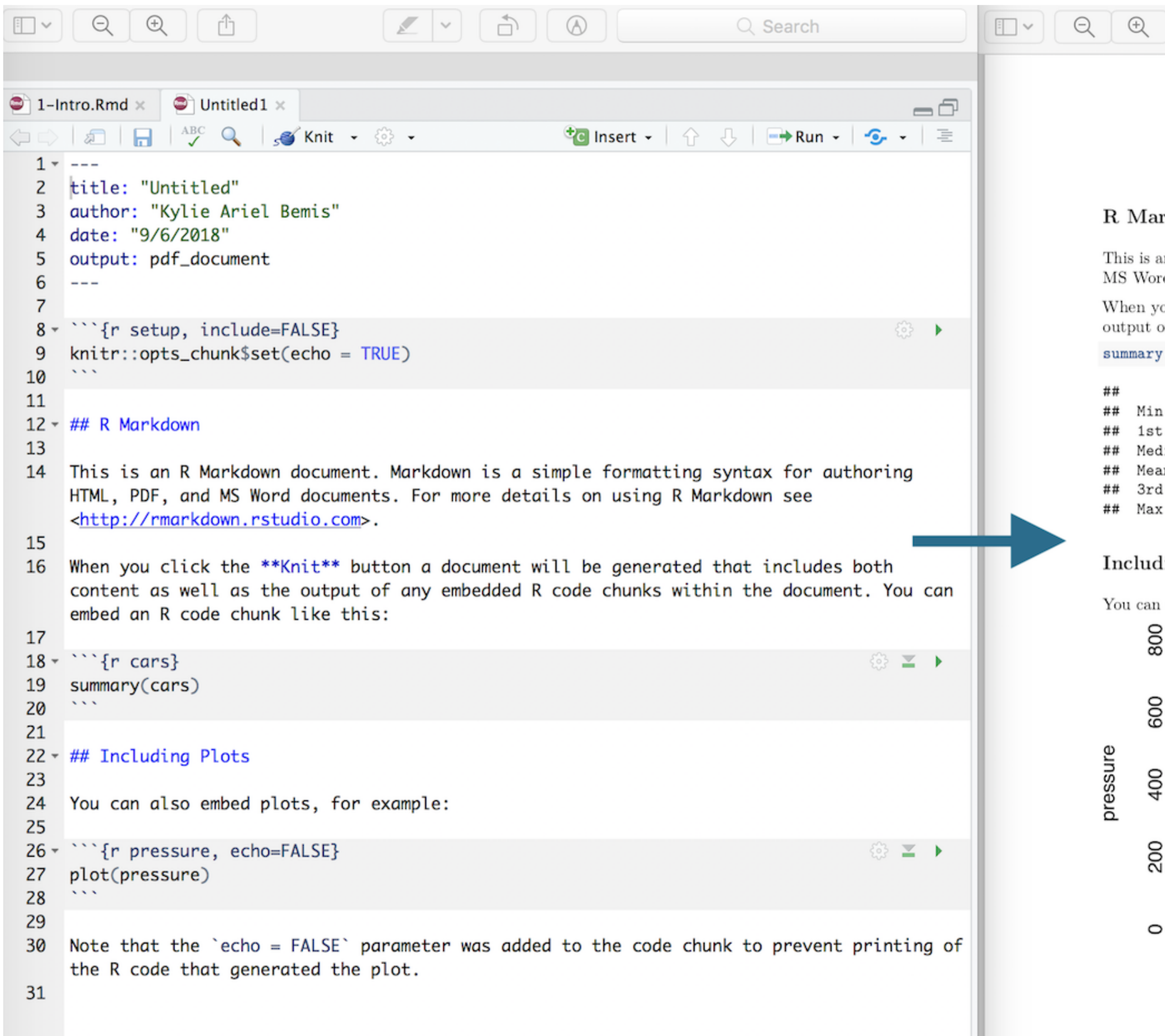
R packages

- Tidyverse
  - ggplot2, dplyr, etc.
  - `install.packages("ti`
- Datasets
  - `install.packages("ny`
  - `install.packages("ga`
- R Markdown
  - `install.packages("rm`

R Markdov

- Authoring framework for da
- Combines R programming w
- Code for analysis embedded
- Documents are fully reprodu
- Required for homework sub
  - Rmd — R + Markdown source code
  - PDF — polished, presentable report

R Markdov


Search
1-Intro.Rmd
Untitled1
Knit
Insert
Run
1 ---
2 title: "Untitled"
3 author: "Kylie Ariel Bemis"
4 date: "9/6/2018"
5 output: pdf_document
6 ---
7
8 ```{r setup, include=FALSE}
9 knitr::opts_chunk$set(echo = TRUE)
10 ```
11
12 ## R Markdown
13
14 This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see <http://rmarkdown.rstudio.com>.
15
16 When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:
17
18 ```{r cars}
19 summary(cars)
20 ```
21
22 ## Including Plots
23
24 You can also embed plots, for example:
25
26 ```{r pressure, echo=FALSE}
27 plot(pressure)
28 ```
29
30 Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.
31
2:1
Untitled
R Markdown
R Mar
This is a
MS Wor
When yo
output o
summary
##
Min
1st
Med
Mea
3rd
Max
Includ
You can
pressure
800
600
400
200
0
Note tha
generate

Reading and res

- *R for Data Science* by Wickha
  - http://r4ds.had.co.nz/
- *Advanced R* by Wickham
  - http://adv-r.had.co.nz/
- *Introduction to Statistical Learn*
  - https://hastie.su.domains/ISLR2/ISLRv2
- *Text Mining with R* by Silge an
  - https://www.tidytextmining.com/

How to ask fo

- Write titles that summar
  - Bad: "Doubt on homework"
  - Good: "Why does my function re
- Explain the problem
  - What is the problem you're tryin
  - What have you tried so far to fix
- https://stackoverflow.com

How to ask for h

- Provide a reproducible
  - Provide the minimum amount
  - Make sure your example repro
  - Ideally, the example should run
  - Provide the actual code, and *no*
- https://stackoverflow.co reproducible-example

INTRODUC
TO **DATA SC**

Learning goals in th

- What is data science?
- What does a data science
- What you will learn in this

Where did **data scienc**

- 1935 — R. A. Fisher's *Design of Ex*
  - "Correlation does not imply causation"
- 1974 — Peter Naur coins term "d
  - Also proposed "datalogy" to replace "comput
- 1977 — John Tukey's *Exploratory D*
  - Introduced the box-and-whisker plot and pop

- 1977 — The International Associatio
  Statistical Computing is established
- 1989 — The first Knowledge Discov
  Databases (KDD) workshop is held
- 2001 — William S. Cleveland publish
  "Data Science: An Action Plan for Ex
  the Technical Areas of the Field of St
  - First proposes "data science" as a separate discip
    distinct from statistics or computer science

- 2002 — *Data Science Journal* launch
- 2003 — *Journal of Data Science* laun
- 2005 - 2014
  - Many people write articles that struggle to pre
    "data science" but everyone seems to agree it's
  - "Sexy" is a frequent descriptor of the new disci
- 2015 - present
  - Northeastern launches Masters in Data Science

Defining data s

- Drew Conway writes in "The Data Science Venn Diagram" (2010):
  - "…one needs to learn a lot as they aspire to become a fully competent data scientist. Unfortunately, simply enumerating texts and tutorials does not untangle knots. Therefore… I present the Data Science Venn diagram… hacking skills, math and stats knowledge, and substantive expertise."

What is data sc

- A **data scientist** blends se
  of expertise to draw insig
  - Statistics and machine learning
  - Computer science and programm
  - Domain knowledge
- A **data scientist** leverages
  inferences that drive decis
  particular field of applicati

The data science pro

- Understand the problem
  - What is the purpose of this an
  - May require domain knowledg
- What data and tools ar
  - May need to pull from a databa
  - May need to identify appropria
  - Most data needs to be cleaned
  - What additional resources do y

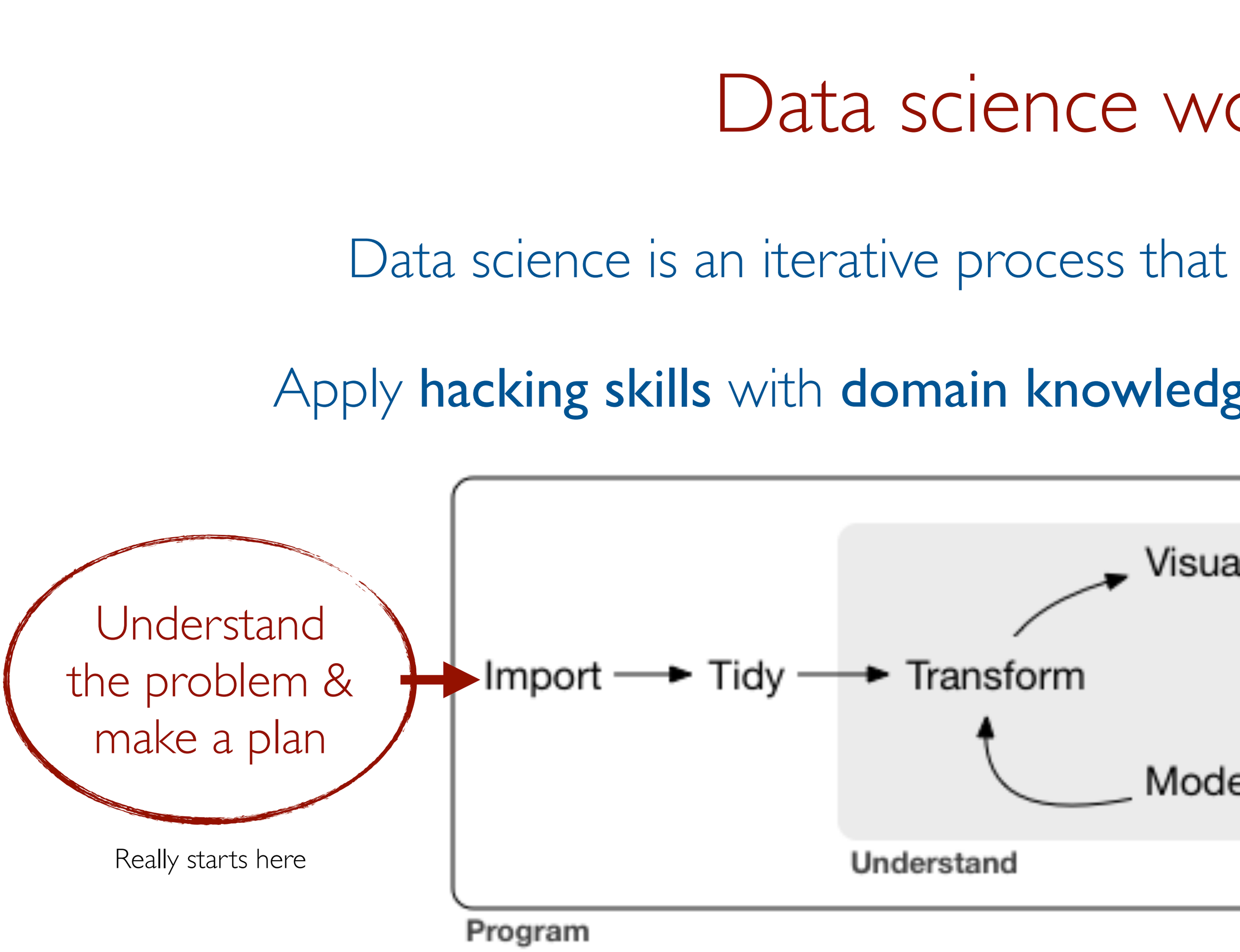
The data science pro

- Perform the analysis
  - What kind of analysis is approp
    - Regression or classification (supervised)
    - Clustering or dimension reduction (uns
    - Hypothesis testing? Statistical inference?
  - Visualize the cleaned data and
- Communicate the resul
  - What insights or decision-maki
  - Communicate the results in an
  - Do you need to deliver a prod

Data science wo

Data science is an iterative process that

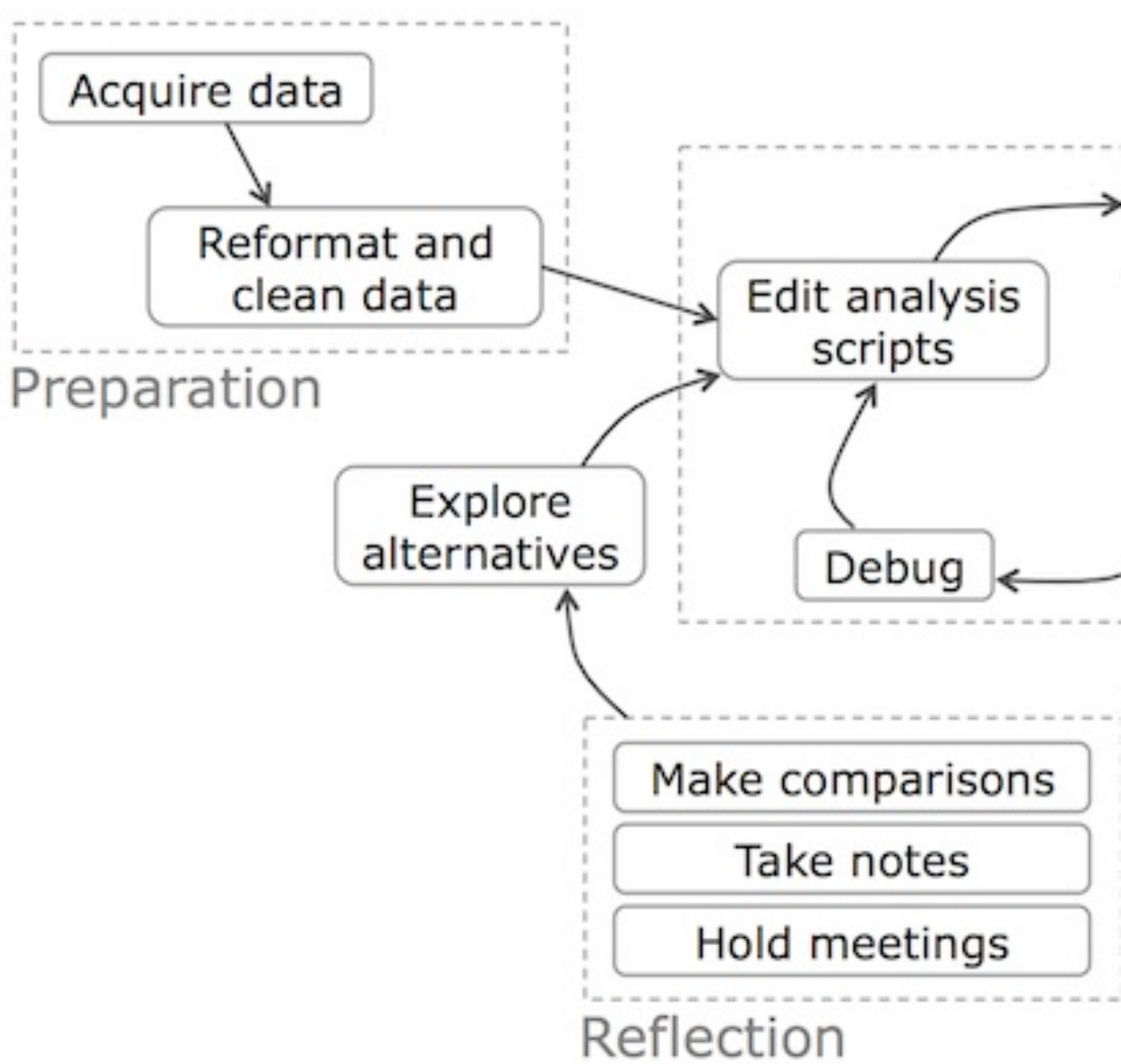
Apply **hacking skills** with **domain knowledg**



An example of the data sc
from *R for Data Science* by Wickh

Data science wo

Another perspec



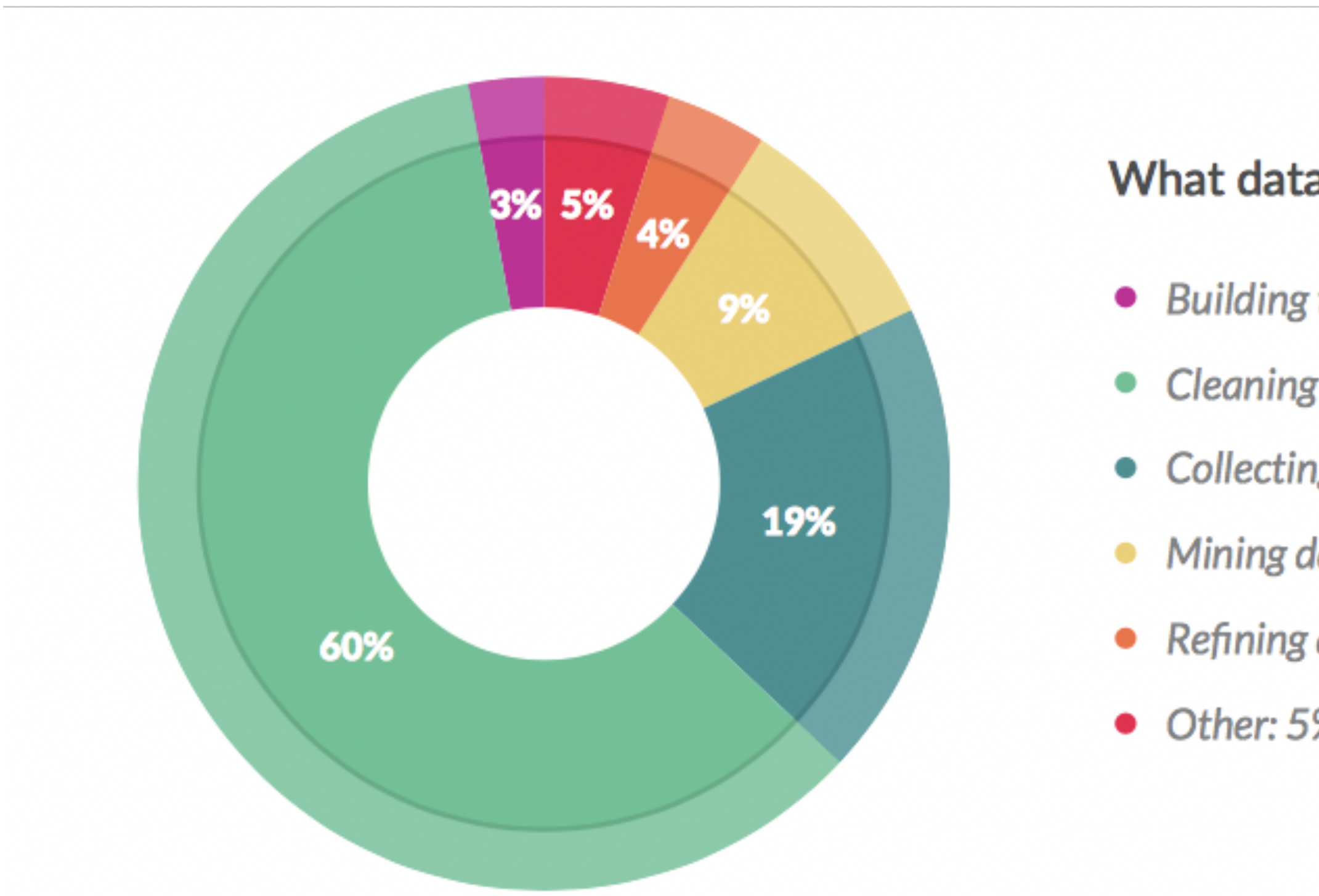
https://cacm.acm.org/blogs/blog-cacm/169199-data-science-v

What does a data s

- visualize data
- pull from a database
- scrape from a website
- clean and wrangle data
- run a regression
- build and train a model
- design an experiment

You should be able to do ~80% of this

What does a data s



https://visit.figure-eight.com/rs/416-ZBE-142/images/Crow

What you will (and won't)

- This course will discuss the fu
  management, processing, and e
  - How to clean and organize data for ana
  - How to visualize and perform basic data
  - How to interpret results and communic
- This course will not delve into
  theory, database design, or clo
  - But we will use machine learning tools a

Course outl

- Data visualization
  - Looking at data and understand
- Data wrangling
  - Cleaning and organizing data fo
- Relational data and databases
  - How data is stored, accessed, a
- Building models for data analysi
  - Basics of statistical modeling an
- Special topics (TBD)

INTRODUCTIO