



INST447 -0101

Fall 2020

Lecture 2

Virtual

Instructor: Bill Farmer

TA: Jonathan Chen

Grader: Jeffrey Chen

September 8, 2020

01

Admin

02

Projects

03

Data Science Pipelines

04

**Python and Jupyter
Notebooks**

05

Review Papers

06

Lab

07

Next Class

This Week

Time: Tuesday virtual

- Admin
 - Re-introduction
 - Class Structure
 - Office Hours
 - Syllabus Updates
- Projects
- Data science Pipelines
- Python and Jupyter Notebooks
- Simple Numpy and Pandas

**Time: Thursday Virtual
w/ optional live session**

- Live session
 - Review of paper?
 - Qs about software install
- Lab
- Next Week
- Questions

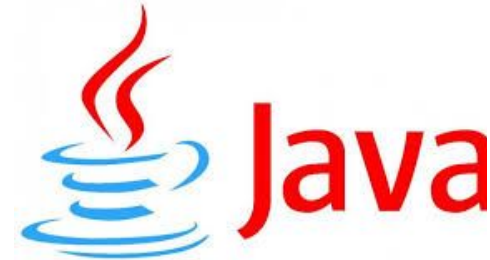
If you are tired, stand up in the back of class.

Use of phone during class for non-class purposes is rude.

Admin

Re-Introductions

- Vice President of Engineering -
 - Clarity Business Solutions, Inc.
- MS Information Technology in Software,
Carnegie Mellon University
- BS Computer Science, University of Pittsburgh
- 22 years Software Engineering experience
- Maryland Data Works Meetup
- Professional Internet & Data Person since 1998
- Parallel tinkerer - RaspberryPi, Arduino, Prusa 3D printer, SDR (beginner)
- Dad to 4 growing children and 3 dogs and 1 cat
- Books currently reading and/or audiobooks
 - Weapons of Math Destruction (UMD suggestion)
 - Infinite Powers - The Story of Calculus, the Language of the Universe



My expectations of you

- Watch the lectures.
 - If it's important enough for me to put in the lecture. You should pay attention to it.
- Do the assigned readings.
- Self Directed Learning
 - A **learning** strategy which allows learners take charge of their own **learning** process (diagnosis **learning** needs, identify **learning** goals, select **learning** strategies, and evaluate **learning** performances and outcomes).
 - I will provide reading assignments and additional videos. You have to let me know if you are not understanding the material.
 - It's ok to not understand at first. It can be frustrating but don't let that get you down. The instructional staff is here to help.
- Keep up with the labs and assignments.
- Communicate with me and the TA (please give me 24-48 hours to respond).
- If you are having difficulty, don't wait till the last moment, let me, Jonathan, or Jeffrey know as early as possible.

Your expectations of me

- Provide an inclusive and equitable classroom climate. (virtual - hybrid of recorded lectures and 'in-person' time)
- We have a shared goal of academic progress.
- Approachable
 - If I'm not reaching out to you as much as you would like, it is a two-way street!
 - REACH OUT TO ME or TA (but please give me 24-48 hours to respond)
- Provide clear weekly direction
 - I am still learn fairly new to ELMS/Canvas and am just finding all of the capabilities included with it
 - You may see some usability changes within the first two weeks.
- Provide fair assessment (grades)
 - If you do the work and you do it on time then you WILL pass this course.
 - It's not just about 'passing' the course, it's about learning the material.

General Class Structure



- Readings
- Lecture videos
- Labs
- Assignments
- Tests
- Group project
- Tuesdays
 - Slides and Video
 - Readings
 - Practice
- Thursdays
 - Live ~30 minutes session
 - Will be recorded
 - Labs

Admin

- Office Hours (need to schedule a time slot):
 - Tuesday 6-7:40 pm
 - Friday 8-10 am
 - Saturday 8-10 am
 - Sunday 4-6pm
 - By Appointment *
- Tentative Live class meetings - (This is not a lecture, more of a review to see if anyone needs anything, classroom time, view lectures prior, software issues)
 - Thursdays 12:30-1:15pm (Class originally scheduled to start @ 12:30)
 - We can add a couple of these at different times if needed
 - I am trying to be flexible

UTA - Jonathan Chen



- Currently is a senior in Information Science
- Email: jonnyapple985@gmail.com
- Office hours:
 - Thursday's 4-5
 - <https://umd.webex.com/umd/j.php?MTID=maa654cf56a69519872456151b7d2c073>

Grader - Jeffrey Chen



- iSchool Alumni - 2019
- Currently in second semester for the Master's in Information Systems program at the R.H. Smith School of Business
- jeffrey.chen@rhsmith.umd.edu








Syllabus Updates



- NTR

INST447 General Schedule

• Tentative as of 9/8

Sunday	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday
					 8-10am	 8-10am
		 Noon		 12:30-1:00pm		
 4-6pm						
		 6-7:40pm		 11:59pm		



Office Hours



Video Ready



Live Sessions



Lab Due

Data Science Skills



<https://towardsdatascience.com/data-science-minimum-10-essential-skills-you-need-to-know-to-start-doing-data-science-e5a5a9be5991>

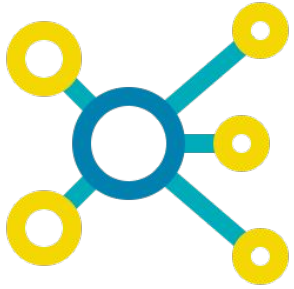
- Coding
- Math (stats)
- Ethics (notice I didn't put this last)
- Team player
- Lifelong learning
- Communication
- Real world project skills (PM)
- ML
- Data visualization
- Data wrangling and preprocessing

Projects

Projects



- Teams of 2, 3, or 4
 - I prefer sizes of 3 or 4
 - Project proposals due 10/1, so you have time
- API Keys
 - Twitter (see my submission process)
- Scraping
 - Reddit example (json)



Data Science Projects

Data science - the ability to take large amounts of data in many different formats and be able to understand it, to process it, to extract value from it, to summarize it, to visualize it, and to communicate it to others.

Data science is the field of study that combines domain expertise, programming skills, and knowledge of math and statistics to extract meaningful insights from data. Data science practitioners apply machine learning algorithms to numbers, text, images, video, audio, and more to produce artificial intelligence (AI) systems that perform tasks which ordinarily require human intelligence. In turn, these systems generate insights that analysts and business users translate into tangible business value. -datarobot.com



- Sentiment Analysis
- Customer Segmentation
- Recommending products
- Public Health Issues
- Manufacturing - predicting faults
- Financial Risk Analysis

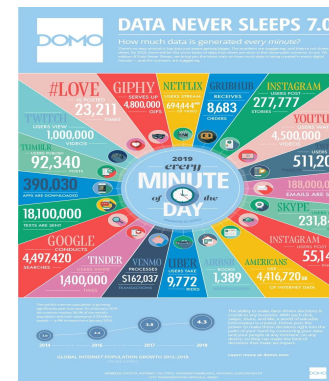
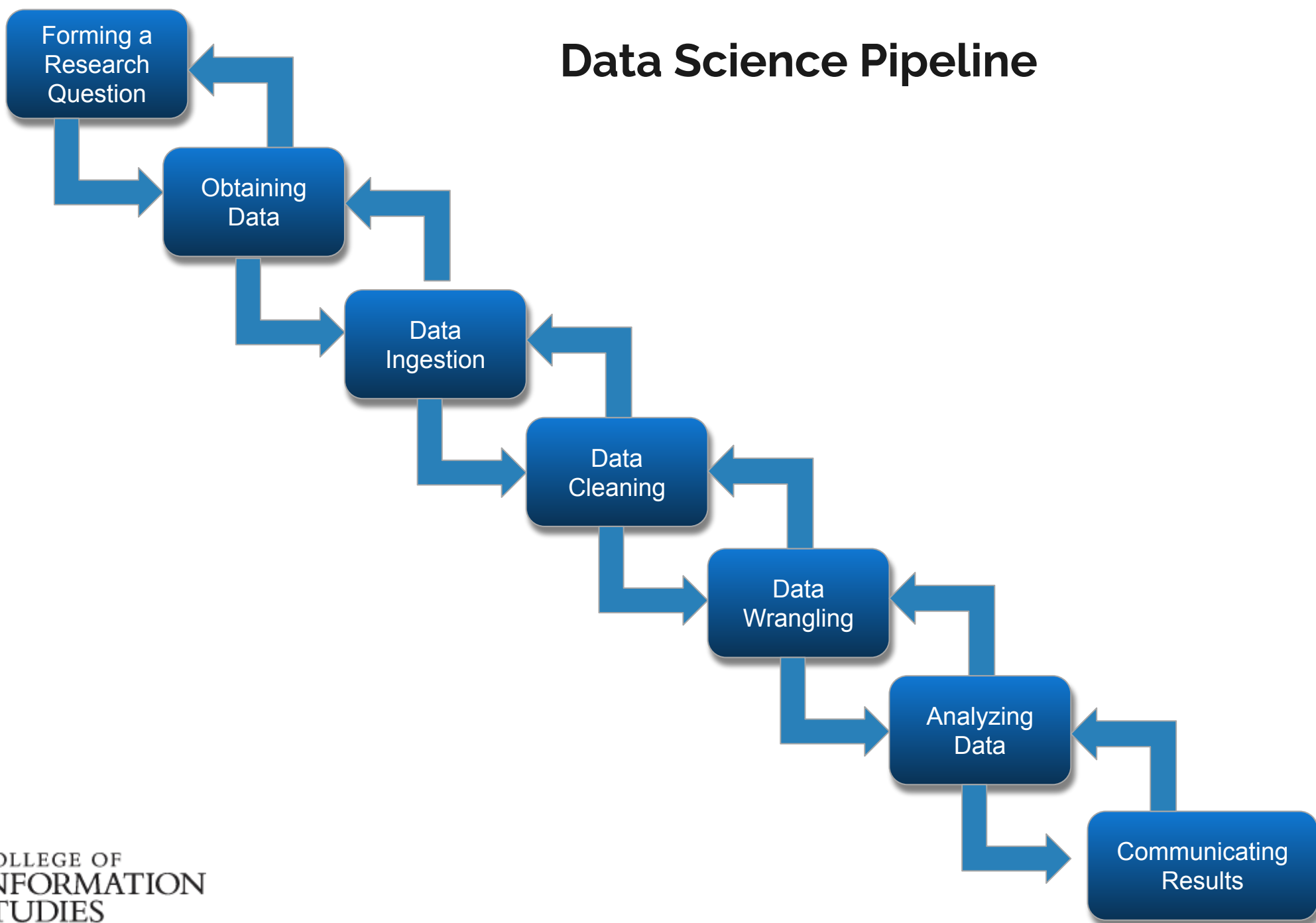


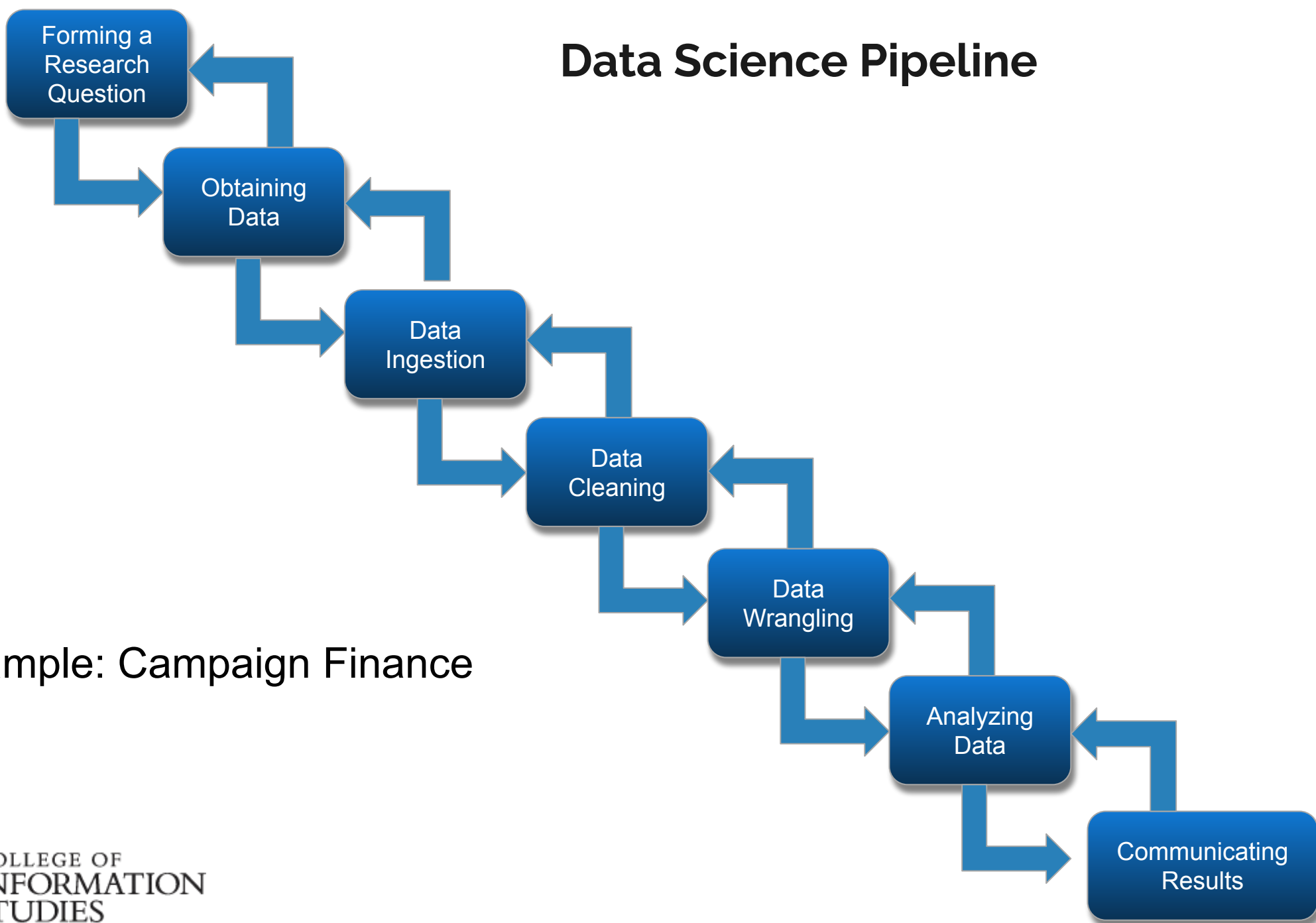
Illustration from:
<https://www.socialmediatoday.com/news/what-happens-on-the-internet-every-minute-2019-version-infographic/558793/>

Data Science Pipelines

Data Science Pipeline



Data Science Pipeline

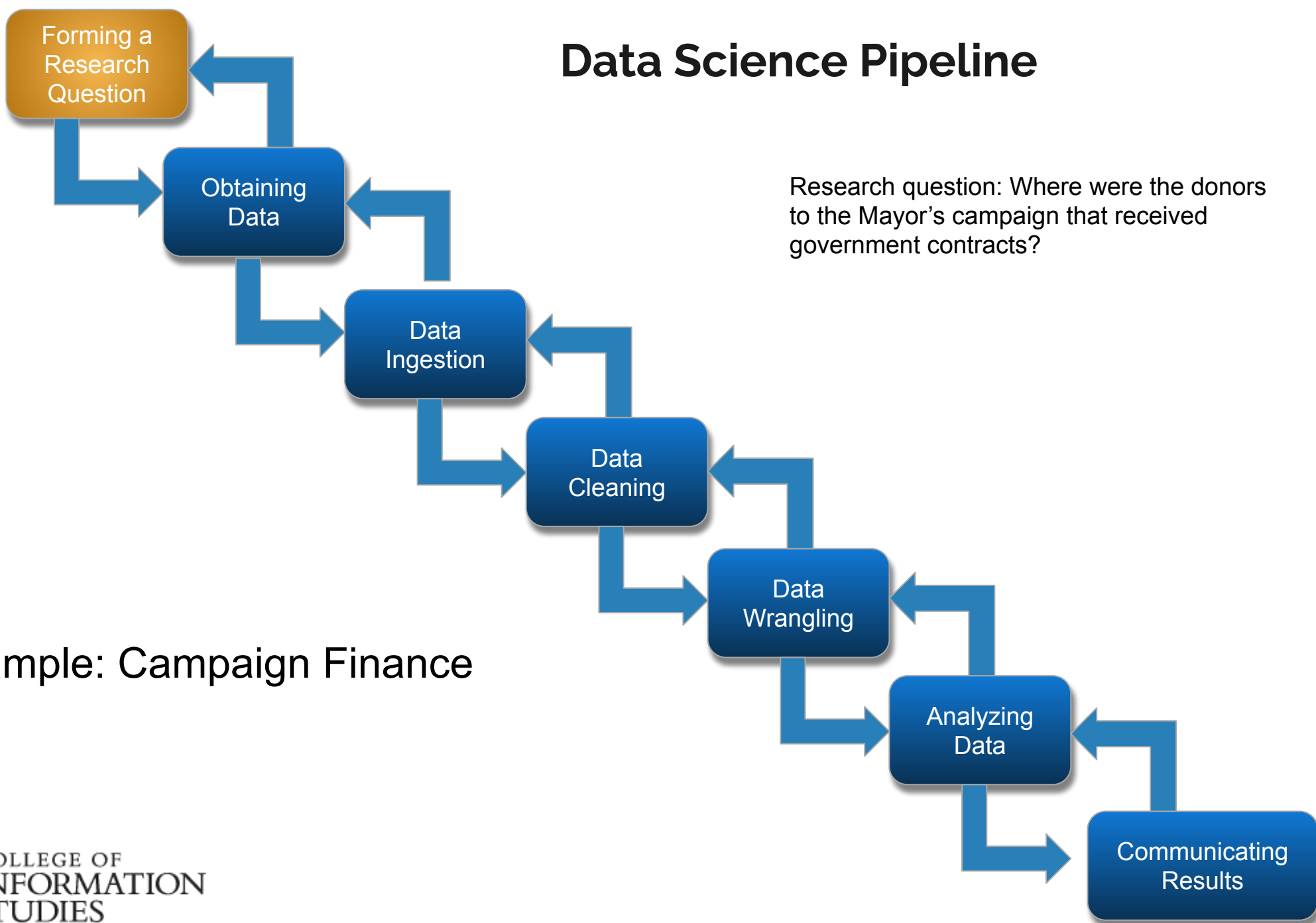


Example: Campaign Finance

Data Science Pipeline

Research question: Where were the donors to the Mayor's campaign that received government contracts?

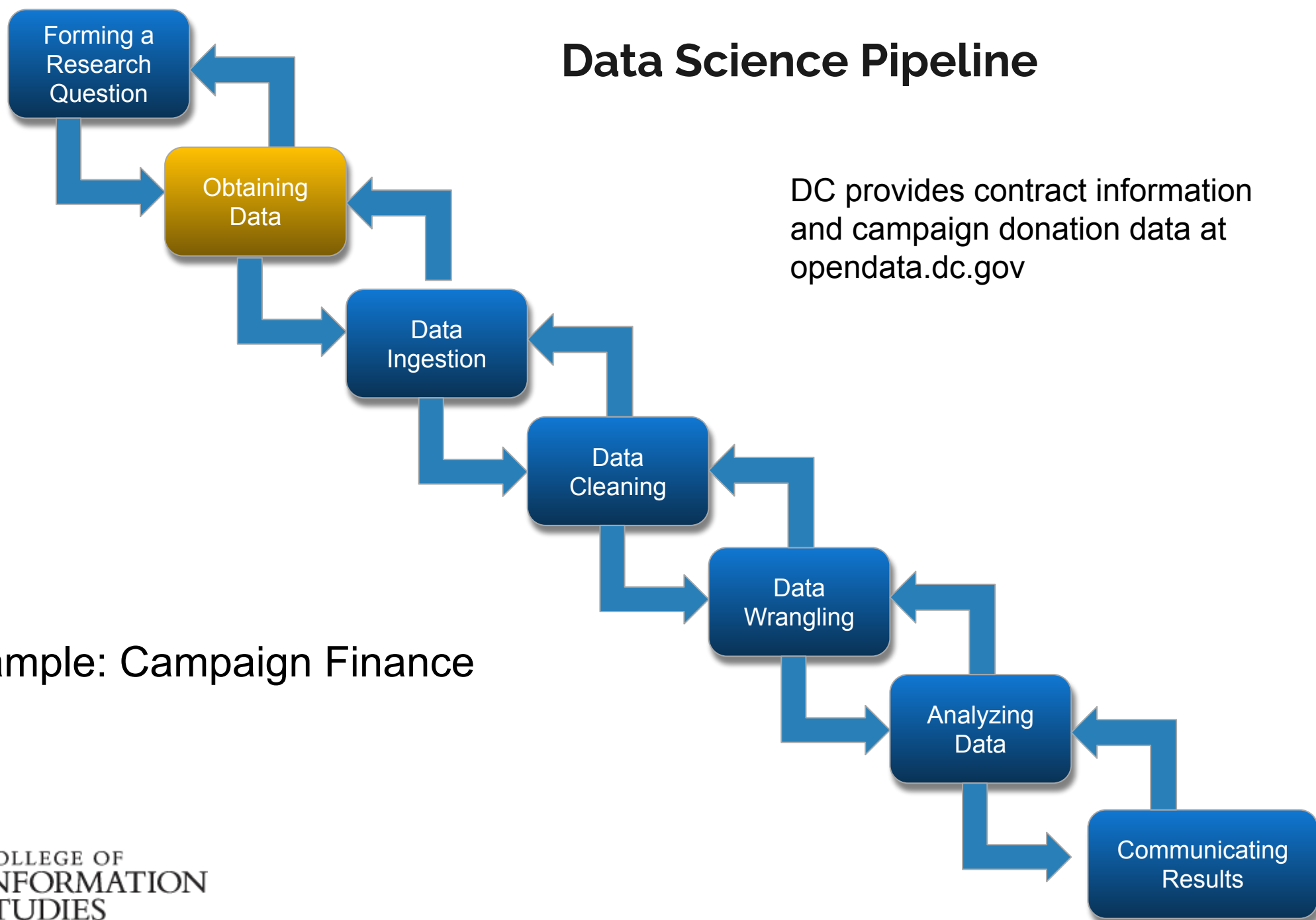
Example: Campaign Finance



Data Science Pipeline

DC provides contract information
and campaign donation data at
opendata.dc.gov

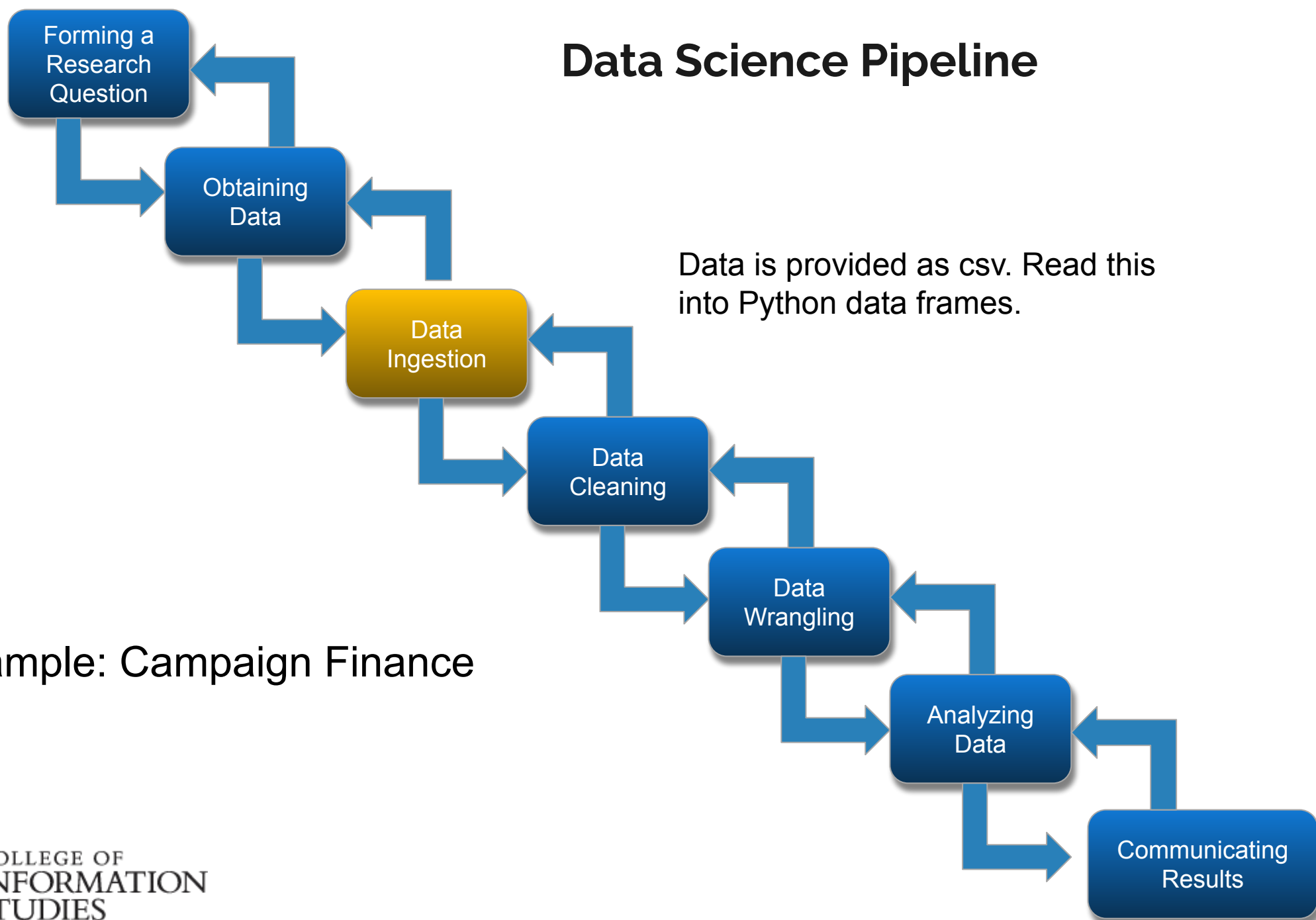
Example: Campaign Finance



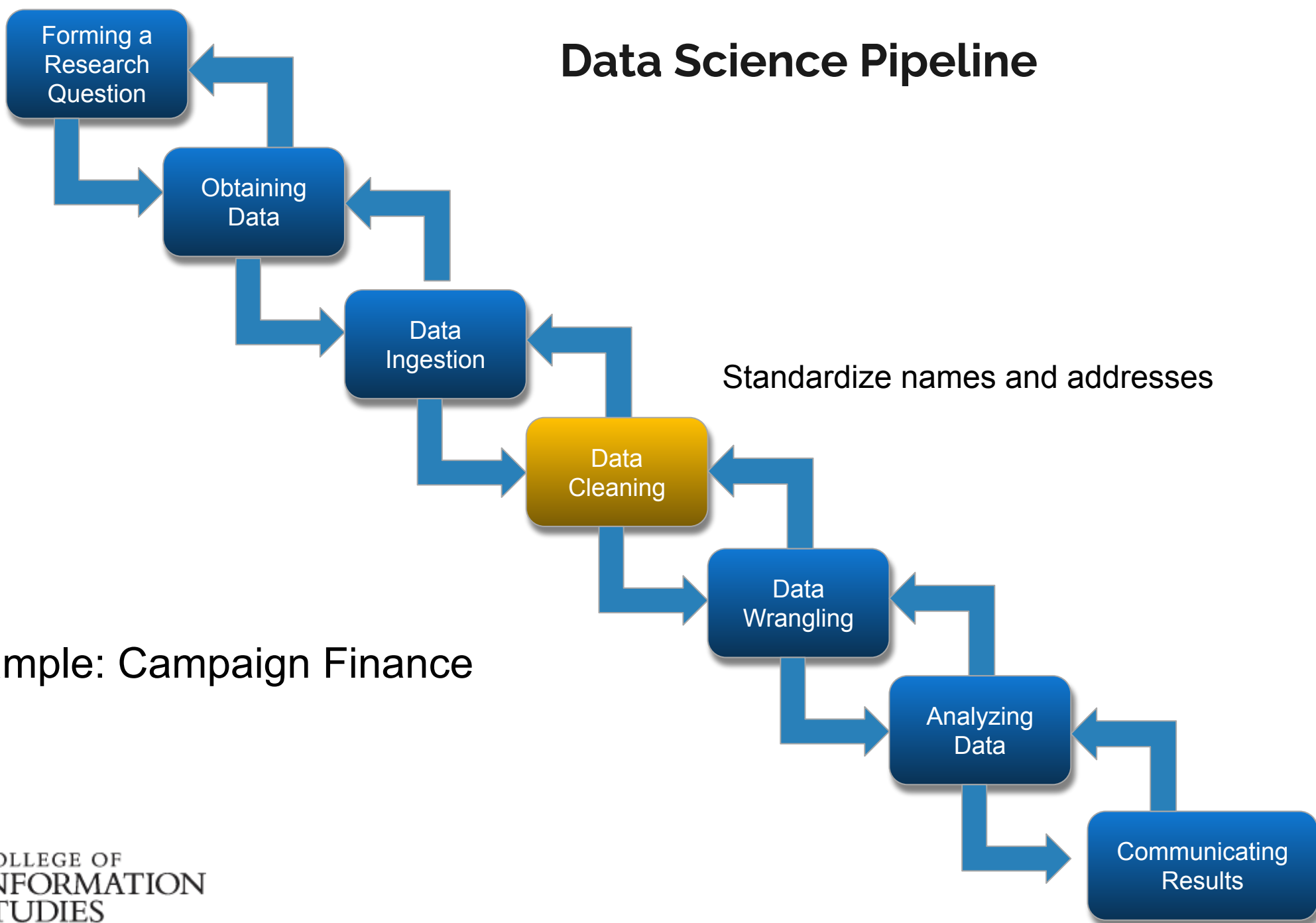
Data Science Pipeline

Data is provided as csv. Read this into Python data frames.

Example: Campaign Finance

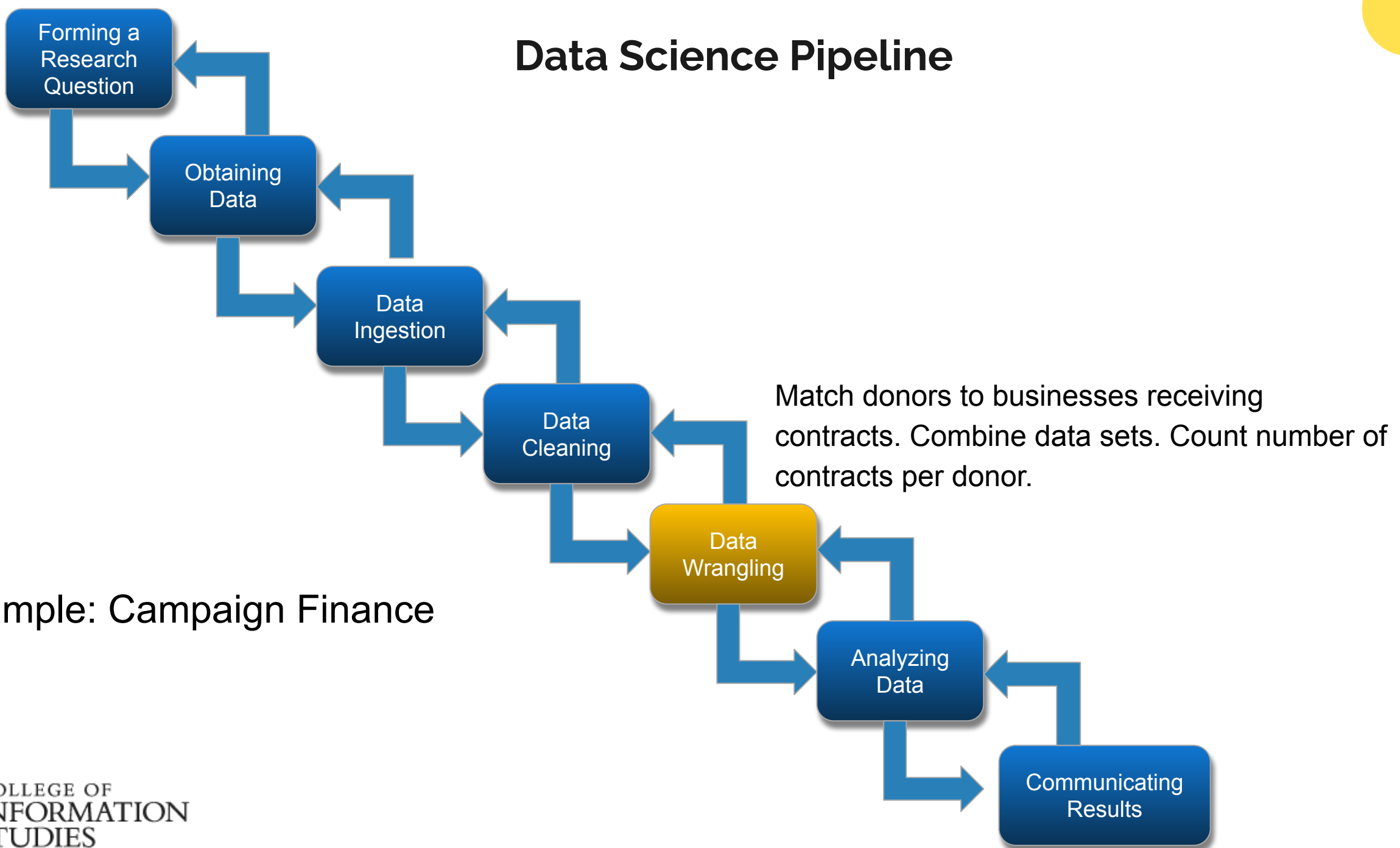


Data Science Pipeline

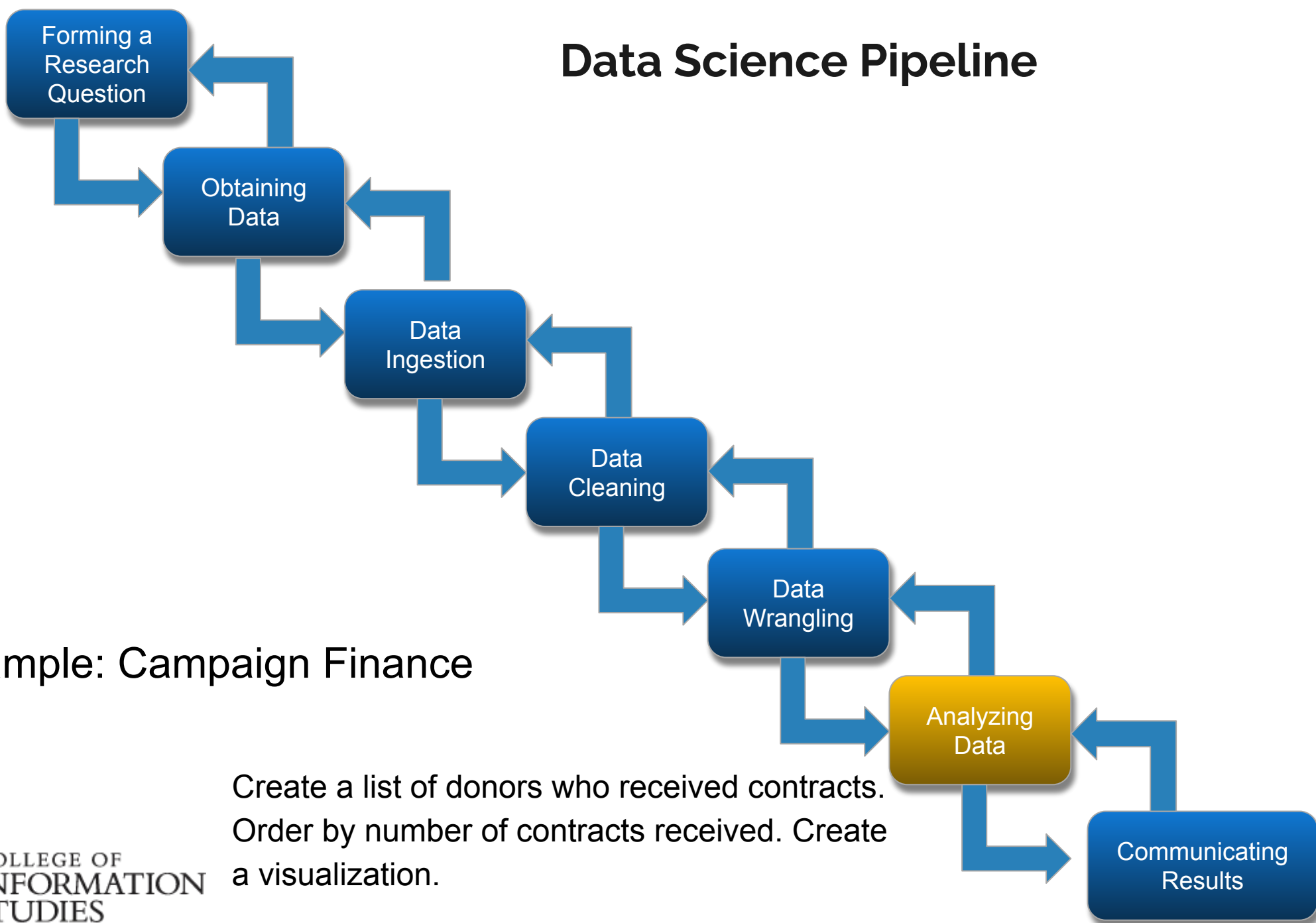


Example: Campaign Finance

Data Science Pipeline



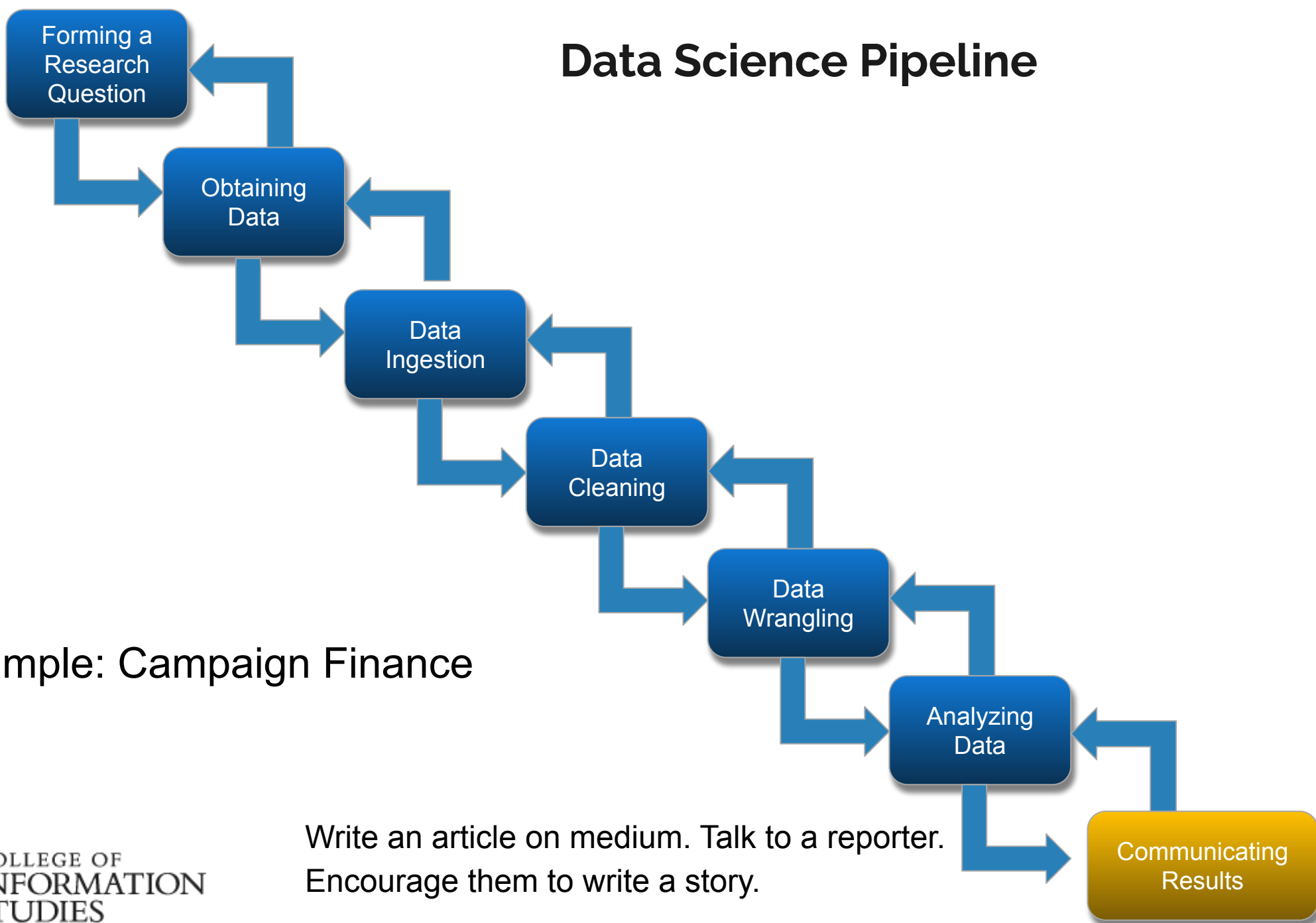
Data Science Pipeline



Example: Campaign Finance

Create a list of donors who received contracts.
Order by number of contracts received. Create
a visualization.

Data Science Pipeline



Software Install

Software Tools



- Python 3 & Jupyter Notebook
 - Method 1
 - Python 3 (<https://www.python.org/downloads>)
 - Pandas Data Analysis Library (pandas)
 - Other modules (e.g. numpy, plotnine)
 - Jupyter Notebooks (aka ipython) (<https://jupyter.org/install>)
 - blend narrative text
 - code
 - output
 - visualizations
 - Method 2
 - Install Anaconda (includes both) (<https://www.anaconda.com/distribution>)
- Open Refine *Week 3
 - <http://openrefine.org/download.html>

Python and Jupyter Notebooks

Python and Jupyter Notebooks review



- Python3 Installed
- Jupyter Notebooks Installed
 - Markdown
 - Save as
- Examples, Numpy and Pandas

Lab

Labs



- Focus on skills related to topic of week
 - e.g. regular expressions
- New data set(s) and research questions
- Work together in pairs in class
- Perform manipulations on data set(s) to answer research questions.
- Enter your answers in ELMS. Turn in *own* Jupyter notebook.

Next Week

Next Week



- Data Cleaning
- Proposal Brainstorm
- More Pandas
- Assignment 1 will be available
- * Much more reading for next week

**I appreciate your
attention
Hope to see you on
Thursday!**



Reference Material Install Software

4 Programming Assignments



- Work independently
- Deeper investigation into a data set and research question
- Turn in a well-structured and written report using Jupyter notebooks

For Thursday

How Twitter can predict an election



DiGrazia, McKelvey, Bollen et al. (2013) More tweets, more votes: Social media as a quantitative indicator of political behavior. PLoS One, 8, 1-5.

- What is this paper about?
- What do they argue and why?

How Twitter can predict an election



- What are the strengths of the paper?
- What are the weaknesses and limitations?

How Twitter can predict an election



How would you go about reproducing this study? What are the steps be detailed. Write down at least one step at each stage in the pipeline

Additional questions to consider to be detailed:

- How can you download data from Twitter?
- How big is a tweet?
- Will there be a problem storing 500 million?
- How will you know what congressional district the person tweeted from?