INST 728Q – Visual Analytics | Modified for BEA

Visual Data Science

Andrea Julca
College of Information Studies
University of Maryland, College Park

MOTIVATING QUESTIONS:

What is data science, really?

~~How do I get answers from data?~~

How does visual analytics fit in?

OUTLINE

- Defining data science
  - Extract, transform, load (ETL)
- Exploratory analysis and modeling
  - NLP – Natural Language Processing
- Streaming visualization

What is a "data scientist?"

"Data Scientist (n.): Person who is better at statistics than any software engineer and better at software engineering than any statistician." - Josh Wills

- Something of a marketing term, but careers and formal data science programs have sprung up around the concept

Data science competencies

Anderson et al. (2014):

- **Information retrieval**
- **Large or streaming data sets**
- **Databases**
- AI and statistical techniques
- Software development and algorithms
- Mathematics
- Communication
- Social, ethical, and legal awareness

Data science competencies

Anderson et al. (2014):

- Information retrieval
- Large or streaming data sets
- Databases
- **AI and statistical techniques**
- **Software development and algorithms**
- **Mathematics**
- Communication
- Social, ethical, and legal awareness

Data science competencies

Anderson et al. (2014):

- Information retrieval
- Large or streaming data sets
- Databases
- AI and statistical techniques
- Software development and algorithms
- Mathematics
- **Communication**
- **Social, ethical, and legal awareness**

Data science workflow

1. Scope out the problem or question
2. Knowledge search: Research and sensemaking
3. Data retrieval; extract, transform, load (ETL)
4. Exploratory analysis
5. Modeling

- System-building [sometimes]
- Versioning/archival
- Communication

Data science workflow

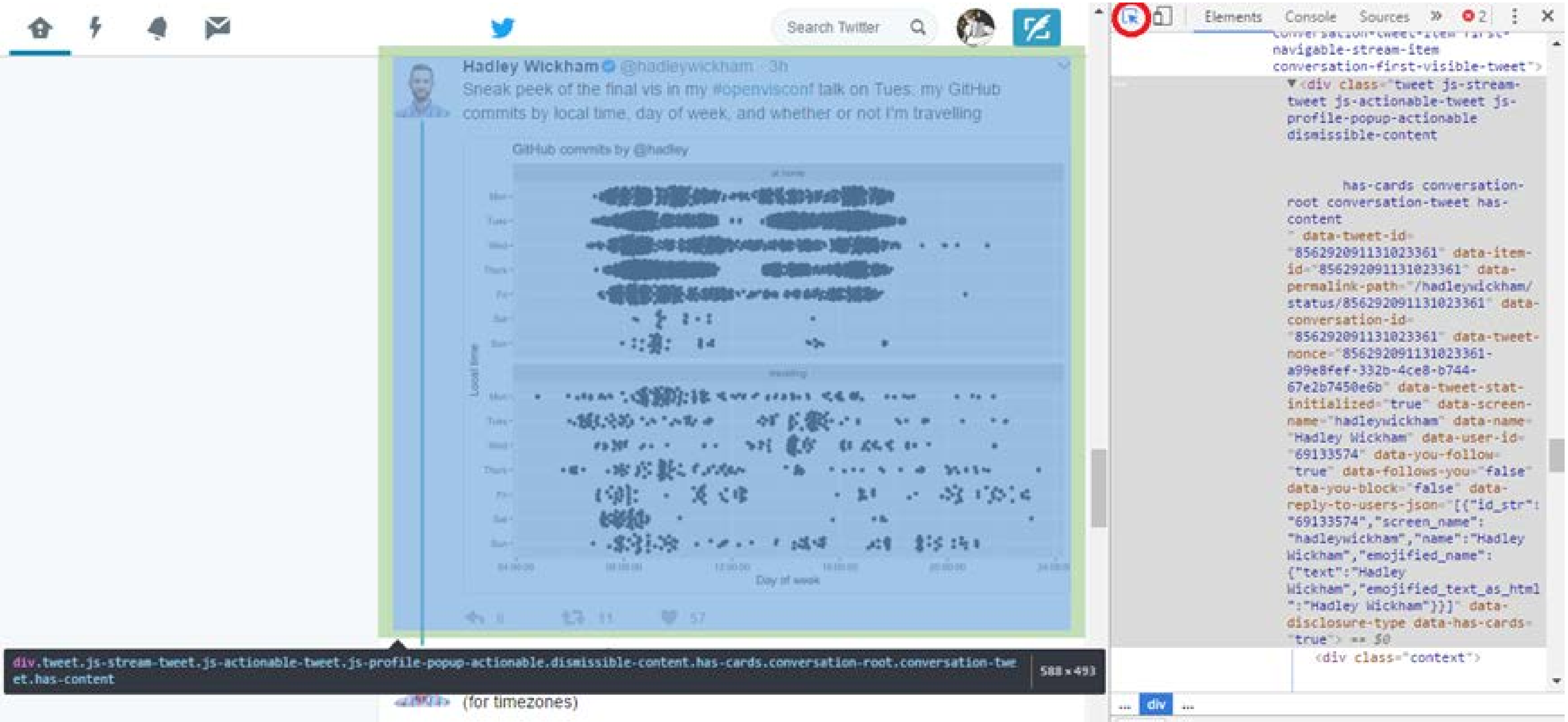
1. Scope out the problem or question
2. Knowledge search: Research and sensemaking
3. **Data retrieval; extract, transform, load (ETL)**
4. Exploratory analysis
5. Modeling

- System-building [sometimes]
- Versioning/archival
- Communication

Extract: Information retrieval

- Information systems: Get data from a database
- Information studies: "Everything is data"
  - Tables
  - Text
  - Images
  - Media files (video, audio)
  - Interviews?
  - Artifacts??
- Other examples?

Extract: "Webscraping"



Hadley Wickham @hadleywickham · 3h
Sneak peek of the final vis in my #openvisconf talk on Tues: my GitHub commits by local time, day of week, and whether or not I'm travelling

(for timezones)

```
navigable-stream-item
conversation-first-visible-tweet">
▼<div class="tweet js-stream-
tweet js-actionable-tweet js-
profile-popup-actionable
dismissible-content


       has-cards conversation-
root conversation-tweet has-
content
" data-tweet-id=
"856292091131023361" data-item-
id="856292091131023361" data-
permalink-path="/hadleywickham/
status/856292091131023361" data-
conversation-id=
"856292091131023361" data-tweet-
nonce="856292091131023361-
a99e8fef-332b-4ce8-b744-
67e2b7450e6b" data-tweet-stat-
initialized="true" data-screen-
name="hadleywickham" data-name=
"Hadley Wickham" data-user-id=
"69133574" data-you-follow=
"true" data-follows-you="false"
data-you-block="false" data-
reply-to-users-json="[{"id_str":
"69133574","screen_name":
"hadleywickham","name":"Hadley
Wickham","emojified_name":
{"text":"Hadley
Wickham","emojified_text_as_html
":"Hadley Wickham"}}]" data-
disclosure-type data-has-cards=
"true"> == $0
  <div class="context">
```

```
div.tweet.js-stream-tweet.js-actionable-tweet.js-profile-popup-actionable.dismissible-content.has-cards.conversation-root.conversation-tweet.has-content    588 × 493
```

... div ...

Windows: F12, Ctrl + Shift + I

Mac: Cmd + Opt + I

Transform [Info Systems]

Reshaping and restructuring data for the target database

- Clean
- Filter
- Apply models
- Business rules
- Aggregate
- Et cetera

Transform [Mathematics]

- Geometry:
  - Reflect
  - Rotate
  - Scale (resize)
  - Translate (shift position)

- Generally:
  - An invertible function mapping one domain to another

Transform [Comp & Data Sci]

Why not both?

[Scrape.R Demo]

Data science workflow

1. Scope out the problem or question
2. Knowledge search: Research and sensemaking
3. Data retrieval; extract, transform, load (ETL)
4. **Exploratory analysis**
5. **Modeling**

- System-building [sometimes]
- Versioning/archival
- Communication

“Homework” Exercise 1

1. Use R with "rvest" package to extract data from website of your choice
2. Use "tidytext" package to transform into tibble

See "Scrape.R" @https://goo.gl/z6OqUS

Load (Stage/Publish) & Archive

- We've "loaded" the data from our chosen website into the *R* environment
  - Not a reliable way to warehouse. Why? *Low permanence*
  - Also not a great publication / communication platform
- In a more complete information or business system, we would:
  - Perform further transformations
  - Load into database with well-defined schema (*higher permanence*)
  - We're skipping that today

Data science competencies

Anderson et al. (2014):

- Information retrieval
- **Large or streaming data sets**
- Databases
- AI and statistical techniques
- Software development and algorithms
- Mathematics
- Communication
- Social, ethical, and legal awareness

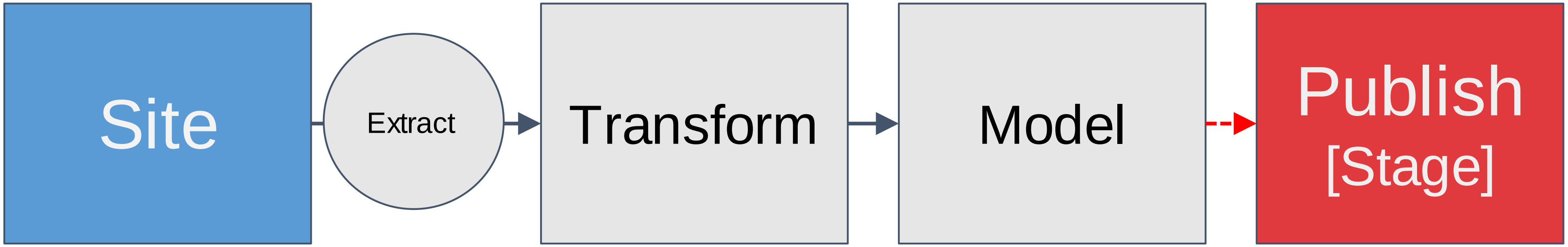
Streaming Visualization

What is "streaming?"

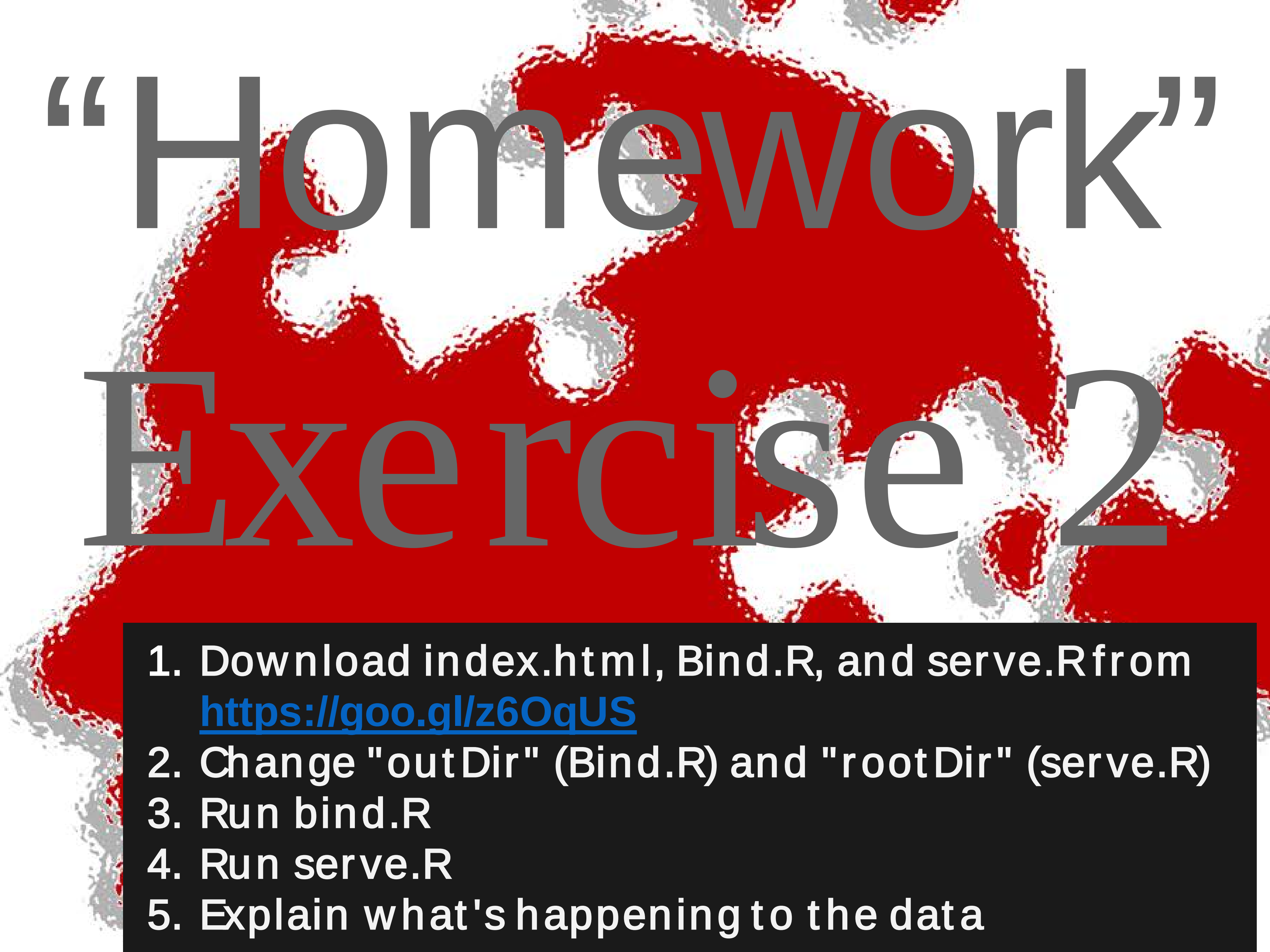
- Transfer of continuously-generated data in real time
- "Real-time" somewhat subjective, contextual

"Streaming visualization," then, is any vis that is continuously updated based on newly-generated, high frequency data

Our process so far & next step



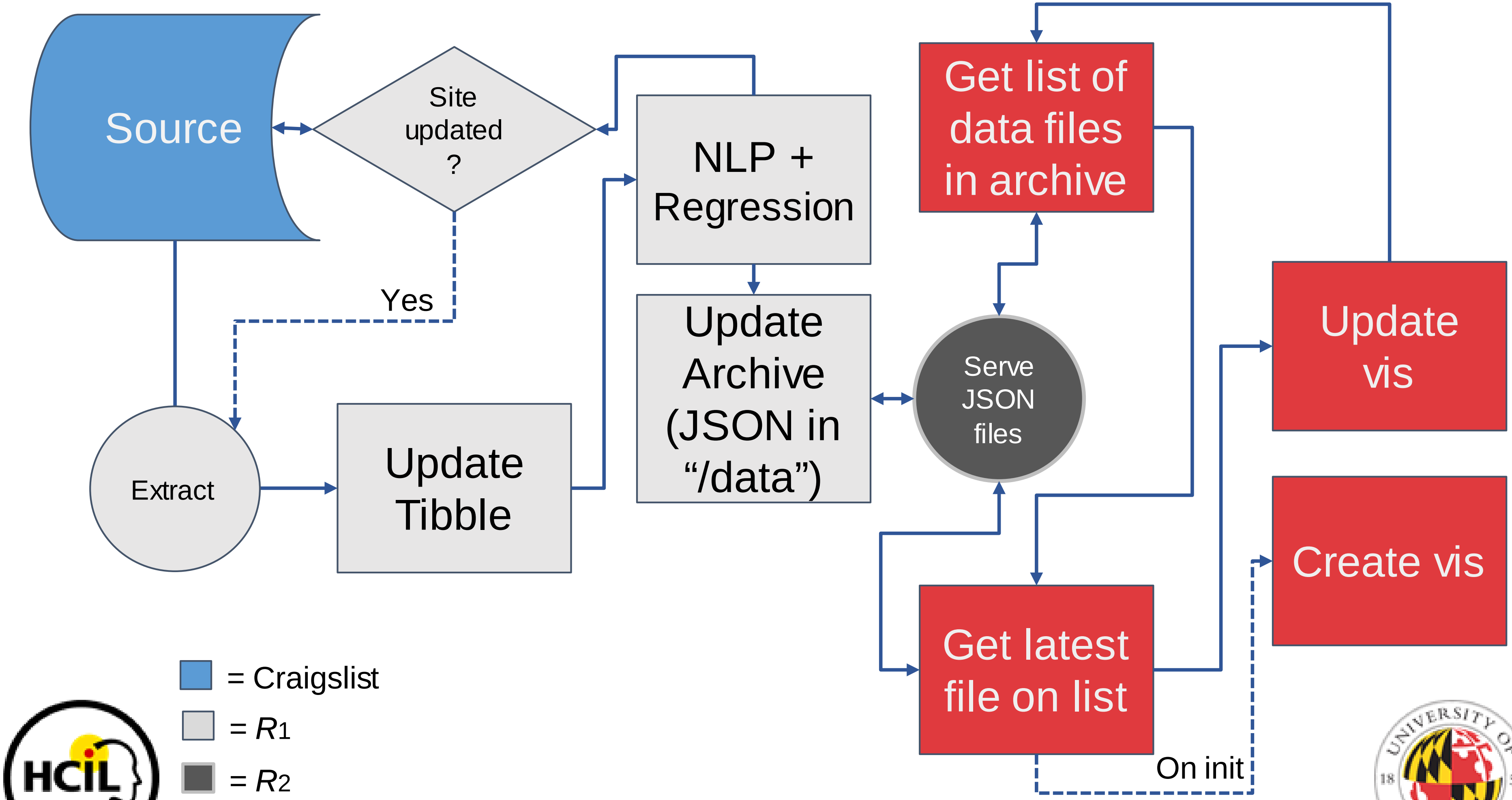
[Streaming Vis Demo]



“Homework” Exercise 2

1. Download index.html, Bind.R, and serve.R from https://goo.gl/z6OqUS
2. Change "outDir" (Bind.R) and "rootDir" (serve.R)
3. Run bind.R
4. Run serve.R
5. Explain what's happening to the data

Our finished network of continuous processes



🟦 = Craigslist

⬜ = *R*1

⬛ = *R*2

🟥 = D3 JS/CSS/HTML Vis

Data science workflow

1. Scope out the problem or question
2. Knowledge search: Research and sensemaking
3. Data retrieval; extract, transform, load (ETL)
4. Exploratory analysis
5. Modeling

- System-building [sometimes]
- Versioning/archival
- Communication

CLOSING REMARKS

QUESTIONS?