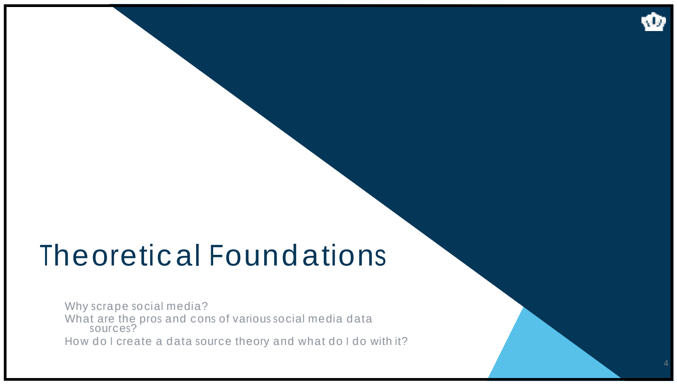
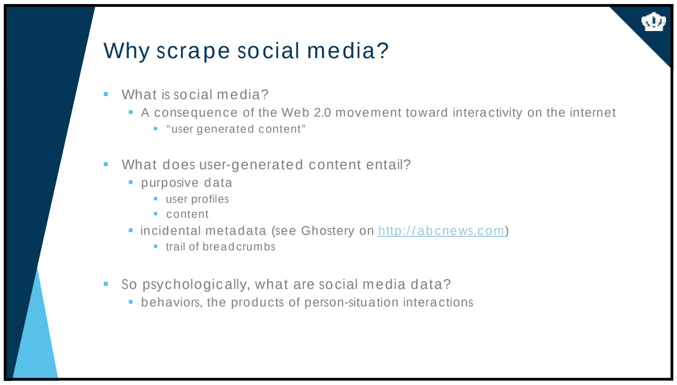


Theoretical Foundations

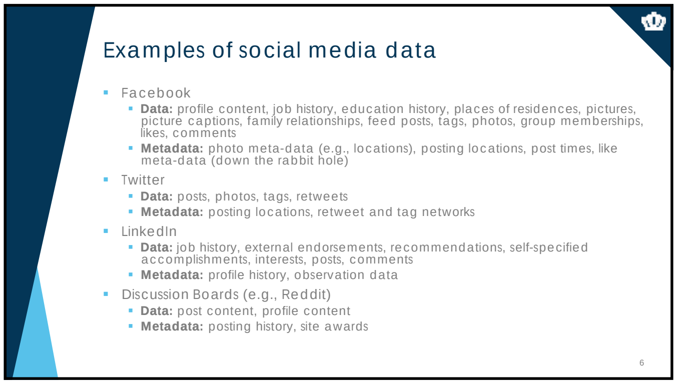
Why scrape social media?
What are the pros and cons of various social media data sources?
How do I create a data source theory and what do I do with it?

4



Why scrape social media?

- What is social media?
  - A consequence of the Web 2.0 movement toward interactivity on the internet
    - "user generated content"
- What does user-generated content entail?
  - purposive data
    - user profiles
    - content
  - incidental metadata (see Ghostery on http://abcnews.com)
    - trail of breadcrumbs
- So psychologically, what are social media data?
  - behaviors, the products of person-situation interactions



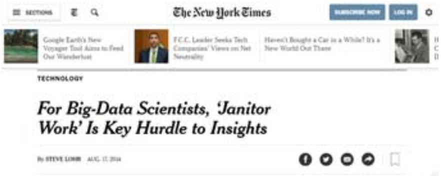
Examples of social media data

- Facebook
  - **Data:** profile content, job history, education history, places of residences, pictures, picture captions, family relationships, feed posts, tags, photos, group memberships, likes, comments
  - **Metadata:** photo meta-data (e.g., locations), posting locations, post times, like meta-data (down the rabbit hole)
- Twitter
  - **Data:** posts, photos, tags, retweets
  - **Metadata:** posting locations, retweet and tag networks
- LinkedIn
  - **Data:** job history, external endorsements, recommendations, self-specified accomplishments, interests, posts, comments
  - **Metadata:** profile history, observation data
- Discussion Boards (e.g., Reddit)
  - **Data:** post content, profile content
  - **Metadata:** posting history, site awards

6

So what can I do with web scraping?

- The first step of “big data science,” data wrangling/munging



SECTIONS
The New York Times
SUBSCRIBE NOW
LOG IN
Google Earth's New Voyager Tool Aims to Feed Our Wanderlust
F.C.C. Leader Seeks Tech Companies' Views on Net Neutrality
Haven't Bought a Car in a While? It's a New World Out There

TECHNOLOGY

For Big-Data Scientists, ‘Janitor Work’ Is Key Hurdle to Insights

By STEVE LOHR AUG. 17, 2014

https://www.nytimes.com/2014/08/18/technology/for-big-data-scientists-hurdle-to-insights-is-janitor-work.html

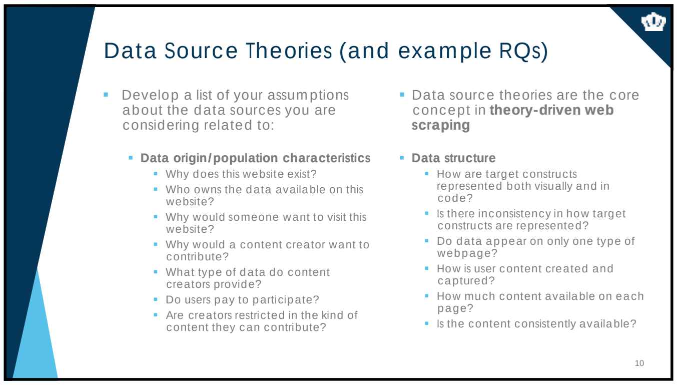
- Can be followed up with any sort of machine learning you want (e.g., OLS regression and Pearson’s correlation, naïve Bayes classifiers)

Who does this generalize to?

- That depends.

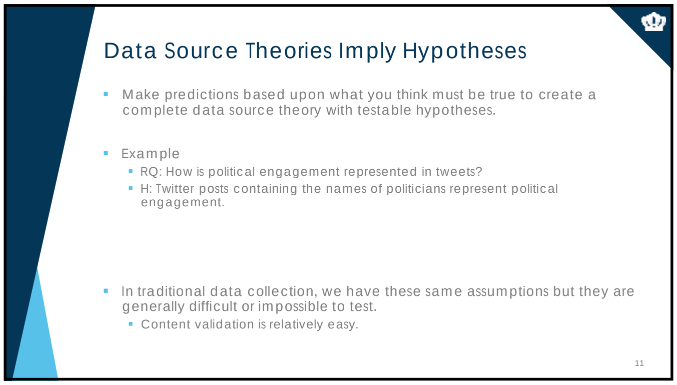
- **Landers, R. N.** & Behrend, T. S. (2015). An inconvenient truth: Arbitrary distinctions between organizations, Mechanical Turk, and other convenience samples. *Industrial and Organizational Psychology, 8,* 142-164.
  - Essentially all samples in I/O psychology are convenience samples, whether academic or practitioner research.
  - The primary questions we need to ask of any convenience sample in relation to generalizability are:
    - Omitted variables bias (endogeneity)
      - Causes of relationships/effects that come from outside our data source
    - Range restriction
      - Constraints on representativeness that comes from outside our data source

Endogeneity



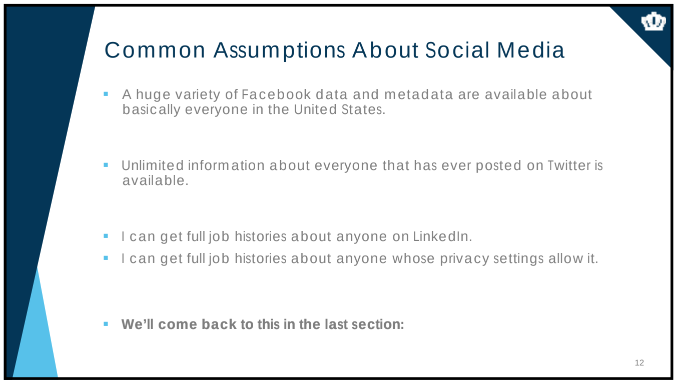
Data Source Theories (and example RQs)

- Develop a list of your assumptions about the data sources you are considering related to:
  - **Data origin/population characteristics**
    - Why does this website exist?
    - Who owns the data available on this website?
    - Why would someone want to visit this website?
    - Why would a content creator want to contribute?
    - What type of data do content creators provide?
    - Do users pay to participate?
    - Are creators restricted in the kind of content they can contribute?
- Data source theories are the core concept in **theory-driven web scraping**
  - **Data structure**
    - How are target constructs represented both visually and in code?
    - Is there inconsistency in how target constructs are represented?
    - Do data appear on only one type of webpage?
    - How is user content created and captured?
    - How much content available on each page?
    - Is the content consistently available?



Data Source Theories Imply Hypotheses

- Make predictions based upon what you think must be true to create a complete data source theory with testable hypotheses.
- Example
  - RQ: How is political engagement represented in tweets?
  - H: Twitter posts containing the names of politicians represent political engagement.
- In traditional data collection, we have these same assumptions but they are generally difficult or impossible to test.
  - Content validation is relatively easy.



Common Assumptions About Social Media

- A huge variety of Facebook data and metadata are available about basically everyone in the United States.
- Unlimited information about everyone that has ever posted on Twitter is available.
- I can get full job histories about anyone on LinkedIn.
- I can get full job histories about anyone whose privacy settings allow it.
- **We'll come back to this in the last section:**