# EXAMINING THE AUTHENTICITY AND RELIABILITY OF SOCIAL MEDIA DATASETS

Shawn Walker @walkeroh

Information School
UNIVERSITY of WASHINGTON

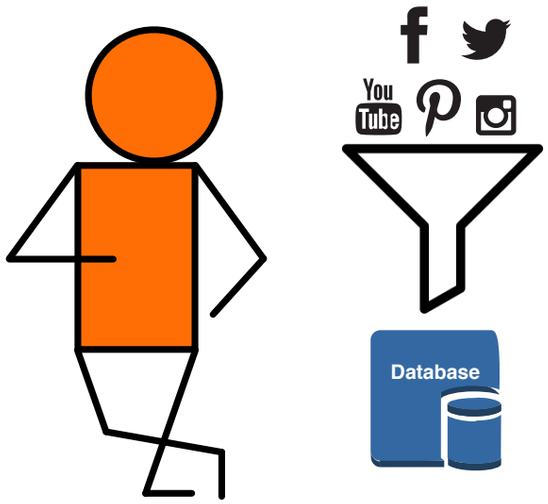
# Social Media Studies

- Elections
- Social Movements
- Disasters
- Epidemiology
- Death and Memorialization
- Depression and Suicidality
- Civil War Reenactment
- Patient Support and Information Seeking Behavior
- Viral Information Flows
- Community Policing
- Marketing
- Meme Propagation and Change
- Journalism
- Social Media Platforms

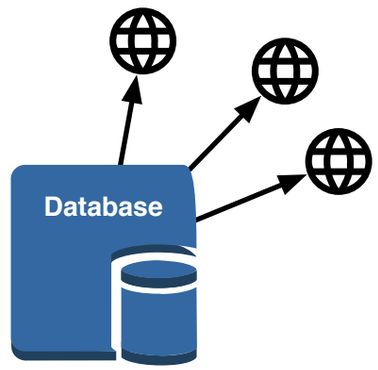
# So You Want to Conduct Social Media Research

- Seemingly Simple on the Surface
  - “Just Collect Data, Right?”

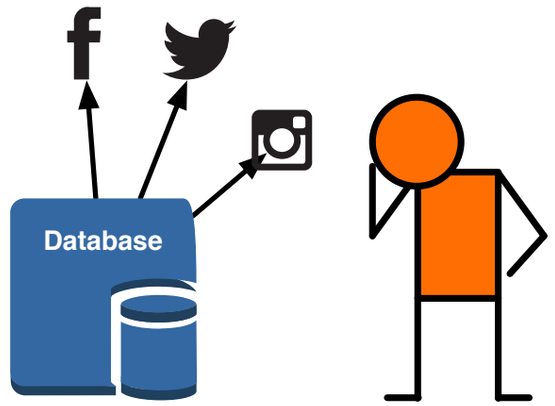
A researcher starts collecting social media data…



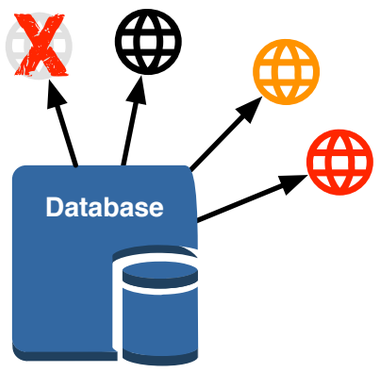

Some of the posts collected contain URLs…



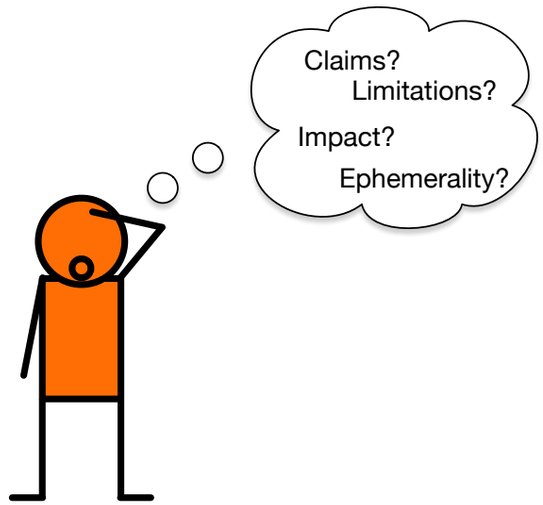

Data analysis begins weeks or months later…



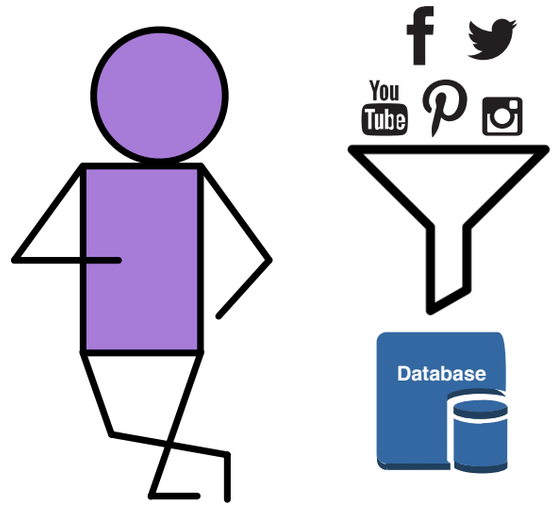

Some URLs have changed after data collection…

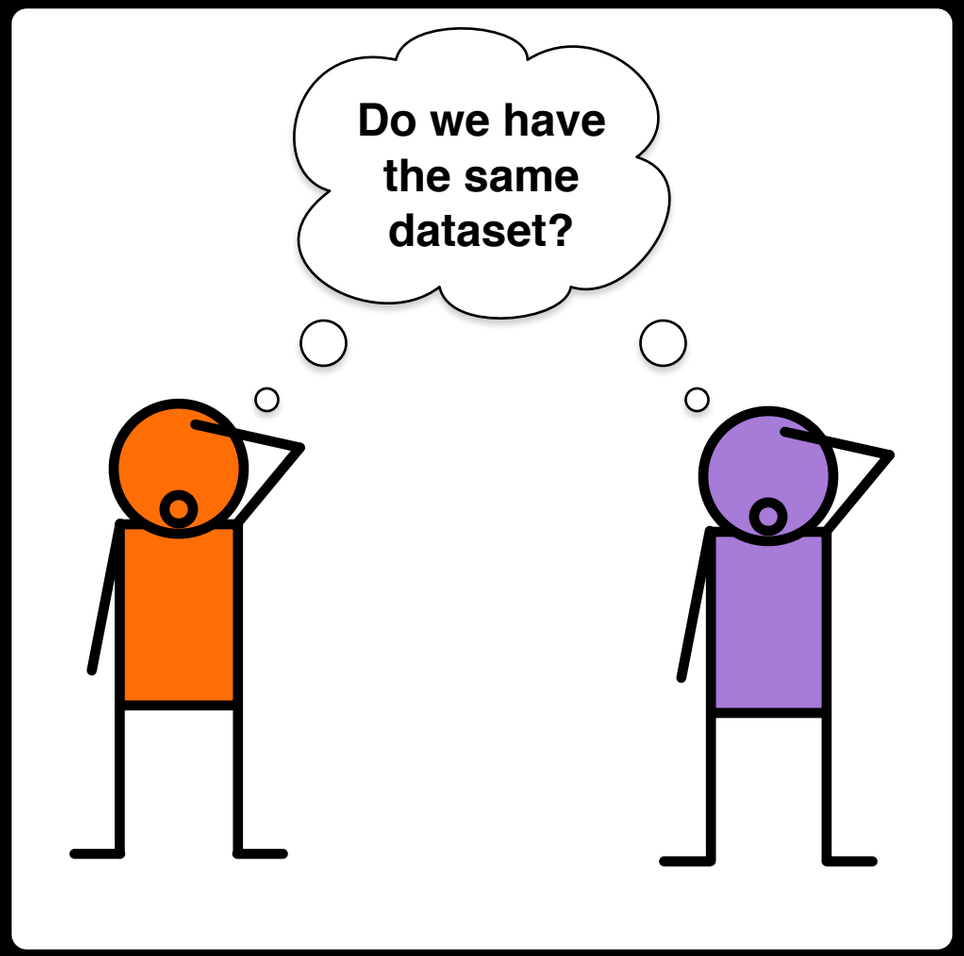



The researcher wonders about the impact…




Another researcher collects the same dataset…




Do we have the same dataset?

# Long List of Competencies

- Theory
- Subject Matter Expertise
- Data
- Computational Processes
- Technical Skills
- Research Design
- Preservation of Data and Metadata

# Social Media Studies Literature

- Methodology Literature is Tool Driven
  - Methodological and Epistemological Choices “Under the Hood”
- Critical Data Studies
  - Important Critiques
  - Lacking Tangible Solutions

- Limitations and appropriateness of existing methods
  - Assumptions met?
    - Content analysis assumes stability in dataset
    - Parametric statistics assume normalized distributions
    - Stratified sample requires knowledge of population and strata

# Where Does That Leave (New) Researchers?

# Ephemerality

- Often used, but rarely defined
- Not a New Concept
    - Film Studies — 1950s BBC Archives
    - Histories — Web Archives
    - Archives and Special Collections — Preservation
    - Data Curation
        - Data that cannot be reproduced, or reconstructed

RQ1: How does the ephemeral nature of social media data effect social media data (SMD) sets?

RQ1A: How does the ephemerality of SMD interact with the process of data collection to impact the reliability of social media data sets?

RQ1B: How does the ephemerality of SMD interact with the process of data collection to impact the authenticity of social media data sets?

RQ1: How does the ephemeral nature of social media data effect social media data (SMD) sets?

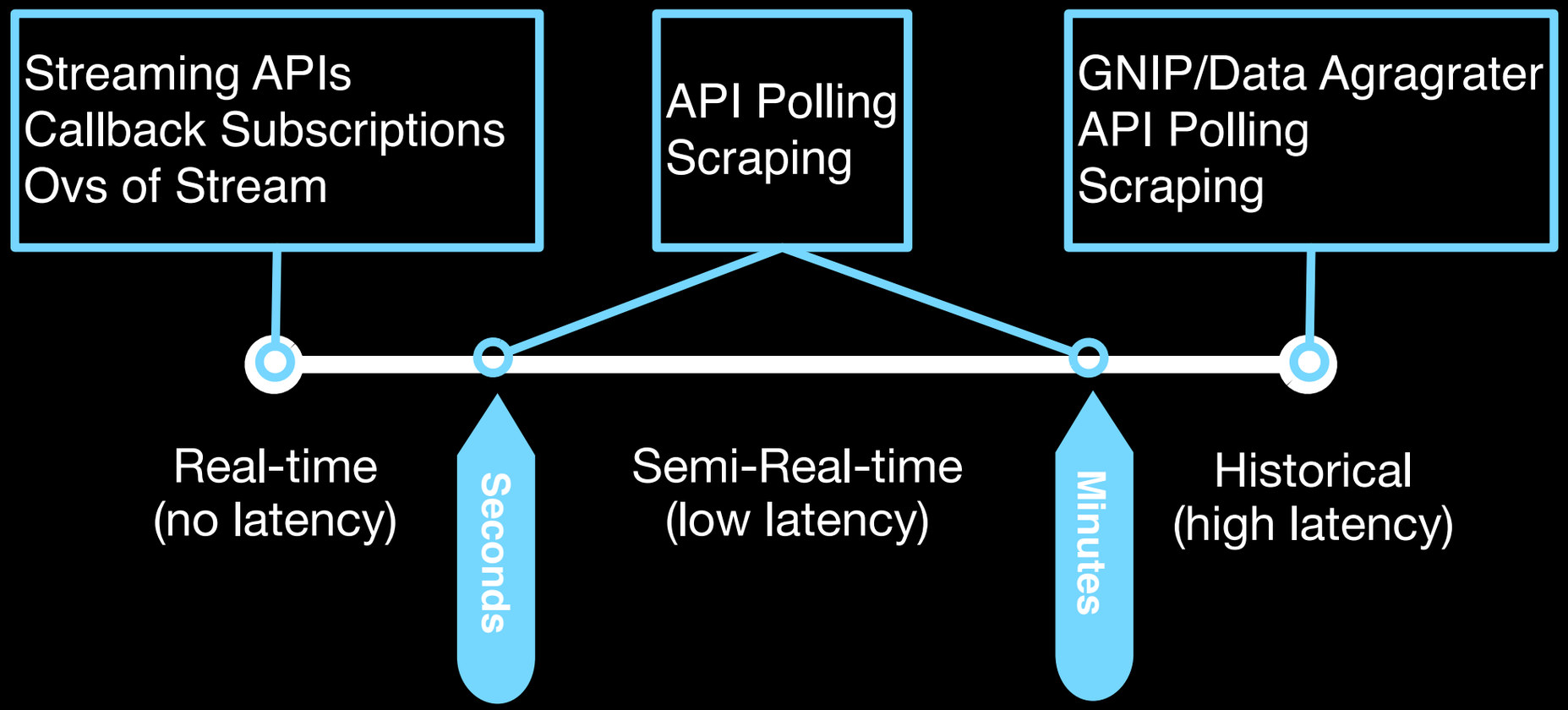
RQ1A: How does the ephemerality of SMD interact with the process of data collection to impact the **reliability** of social media data sets?

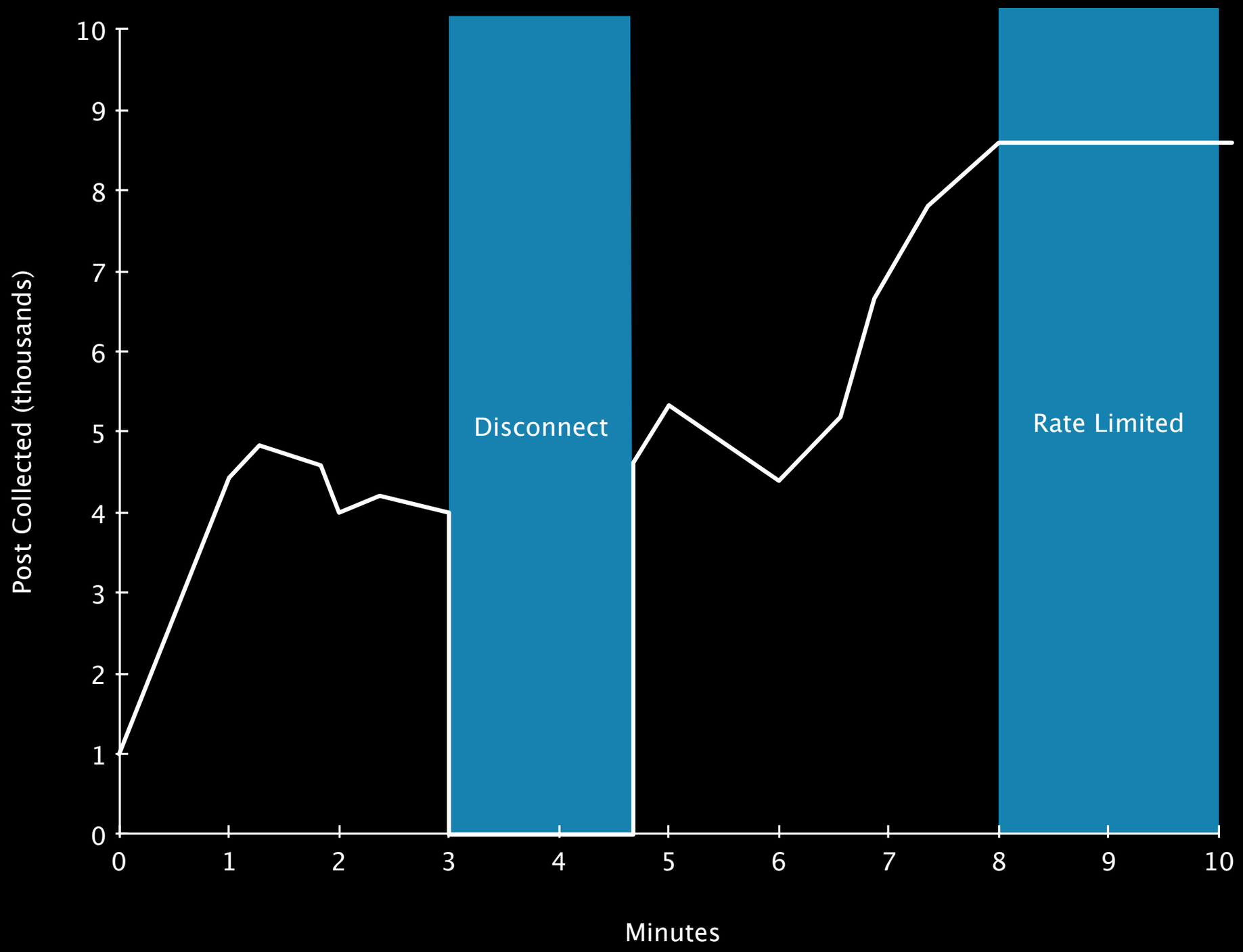
RQ1B: How does the ephemerality of SMD interact with the process of data collection to impact the **authenticity** of social media data sets?

# Theoretical Lenses

- Process Theory

- Archival Theory — Preservation of Electronic Records


Streaming APIs
Callback Subscriptions
Ovs of Stream
API Polling
Scraping
GNIP/Data Agragrater
API Polling
Scraping
Real-time
(no latency)
Seconds
Semi-Real-time
(low latency)
Minutes
Historical
(high latency)

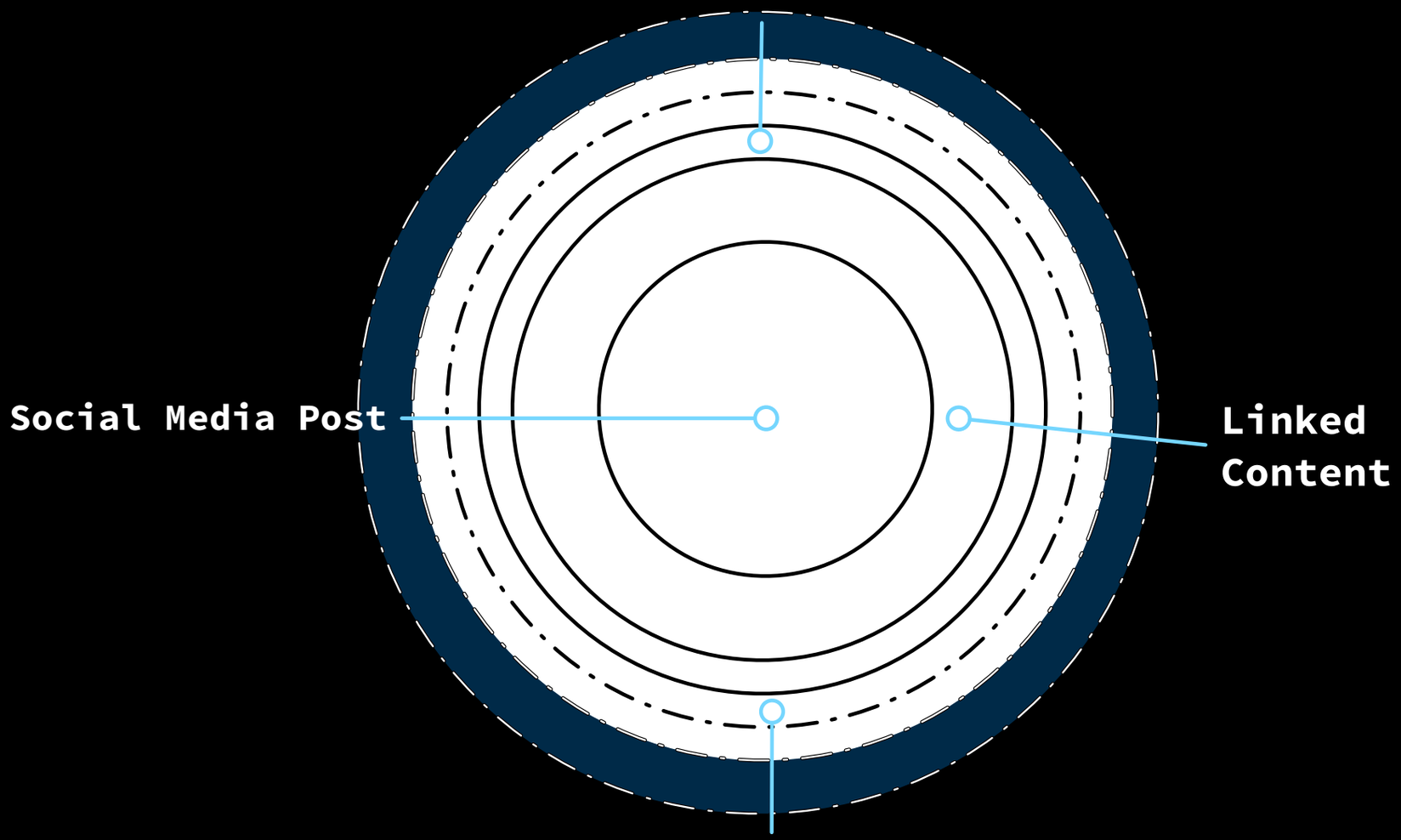

Post Collected (thousands)
10
9
8
7
6
5
4
3
2
1
0
Disconnect
Rate Limited
0
1
2
3
4
5
6
7
8
9
10
Minutes

# Archival Theory

**Record** → **Electronic Record**

↓

**Social Media as a Record**


Metadata
Social Media Post
Linked Content
Related Social Posts

**The White House** ✓
@WhiteHouse

Follow

You can now pick a student loan repayment option in 5 steps or less:
StudentLoans.gov/Repay #CollegeOpportunity


RETWEETS 232 LIKES 349

3:31 PM - 28 Apr 2016

```
{
  "created_at": "Thu Apr 28 22:31:32 +0000 2016",
  "id": 725814735917060097,
  "id_str": "725814735917060097",
  "text": "You can now pick a student loan repayment option in 5 steps or less: https://t.co/
n6Tmuk5Nn8 #CollegeOpportunity https://t.co/gXNLUIOaTQ",
  "truncated": false,
  "entities": {
    "hashtags": [
      {
        "text": "CollegeOpportunity",
        "indices": [
          93,
          112
        ]
      }
],
...
"urls": [
  {
    "url": "https://t.co/n6Tmuk5Nn8",
    "expanded_url": "http://StudentLoans.gov/Repay",
    "display_url": "StudentLoans.gov/Repay",
    "indices": [
      69,
      92
    ]
  }
],
"media": [
  {
    "id": 725814708343738368,
```

You can now pick a student loan repayment option in 5 steps or less: https://t.co/n6Tmuk5Nn8 #CollegeOpportunity https://t.co/gXNLUIOaTQ

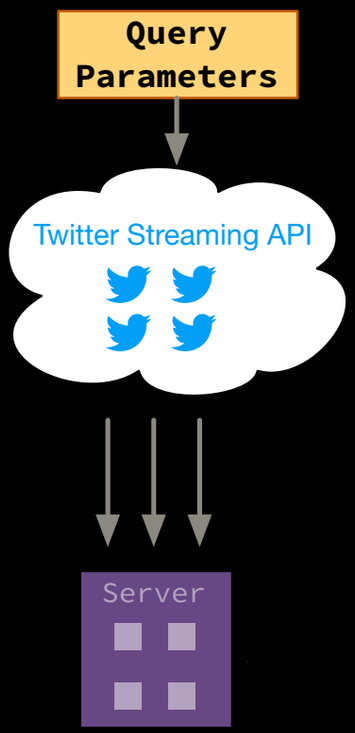
- Content of Embedded Links
- Final Destination of Shortened Links
- Embedded Images and Video
- Conversations Surrounding Hashtags and Keywords

# Reliability

- Track unique tweet IDs across three data collection points

# Authenticity

- Compare the metadata of tweets across three data collection points.
  - Twitter handle
  - Profile Description
  - Location
  - Profile Picture
  - Homepage URL
  - Tweet and User Statistics




Two Weeks

# Case Studies


# Occupy Wall Street

# Prototypical Features

- Multi-Year Time Scale
- Contentious Political Context
- Keyword Based Query
- High Account and Metadata Instability
- Unbounded Population

Oregon Department of Transportation


# Departments of Transportation

# Prototypical Features

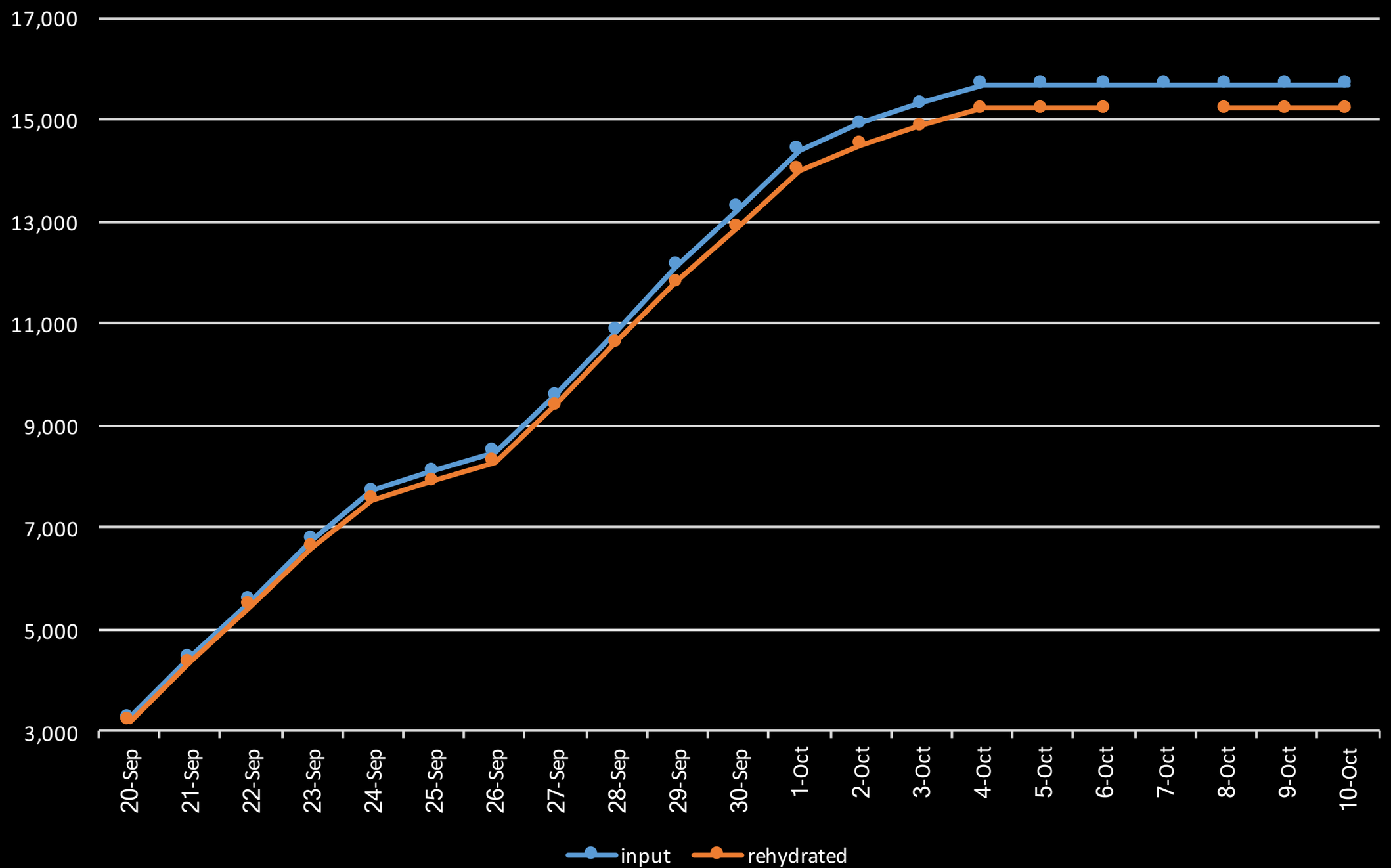
- Bounded Population
- Account Based Query
- High Metadata and Account Stability
- High URL Stability
- Every-Day Political Context

# RuPaul's Drag Race

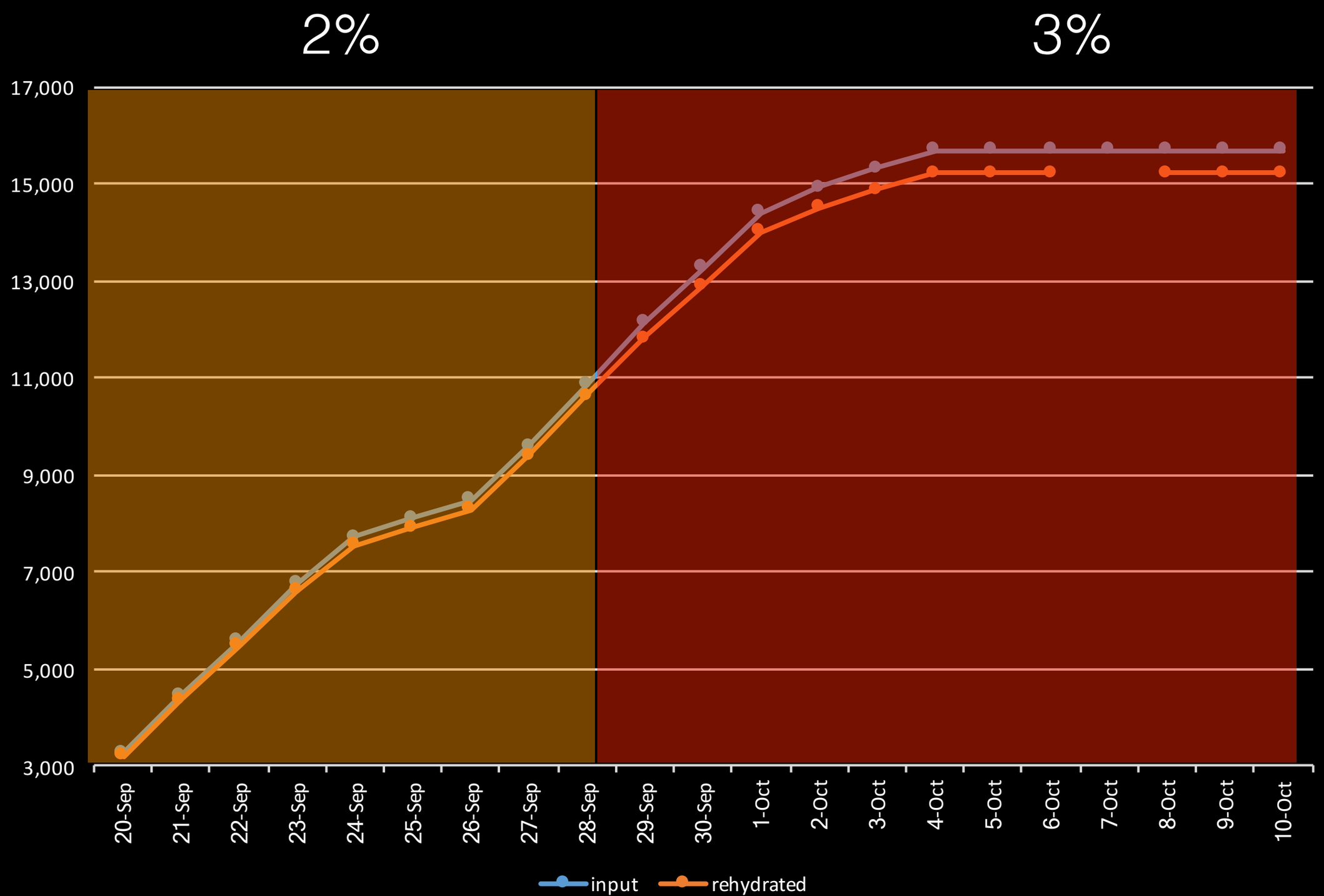
# Prototypical Features

- Mixed Keyword and Account Query
- Entertainment Context
- Media Intensive
- Semi-Bounded Population

# Reliability - Depts. Of Transportation

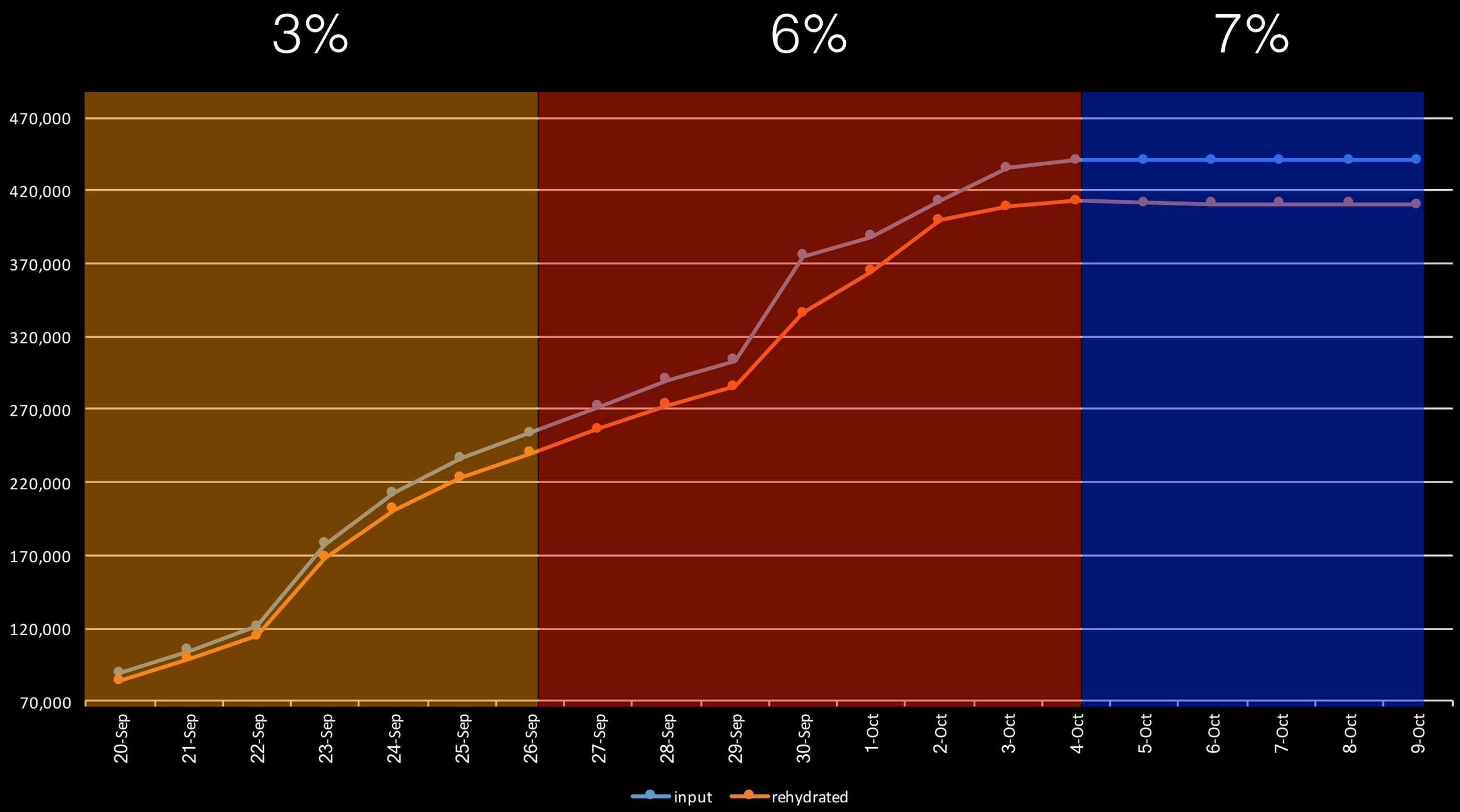


# Reliability - Depts. Of Transportation



# Reliability - Drag Race

# Reliability - Drag Race



# THANK YOU...

SHAWNW.IO

@WALKEROH

STW3@UW.EDU