

Lab 1

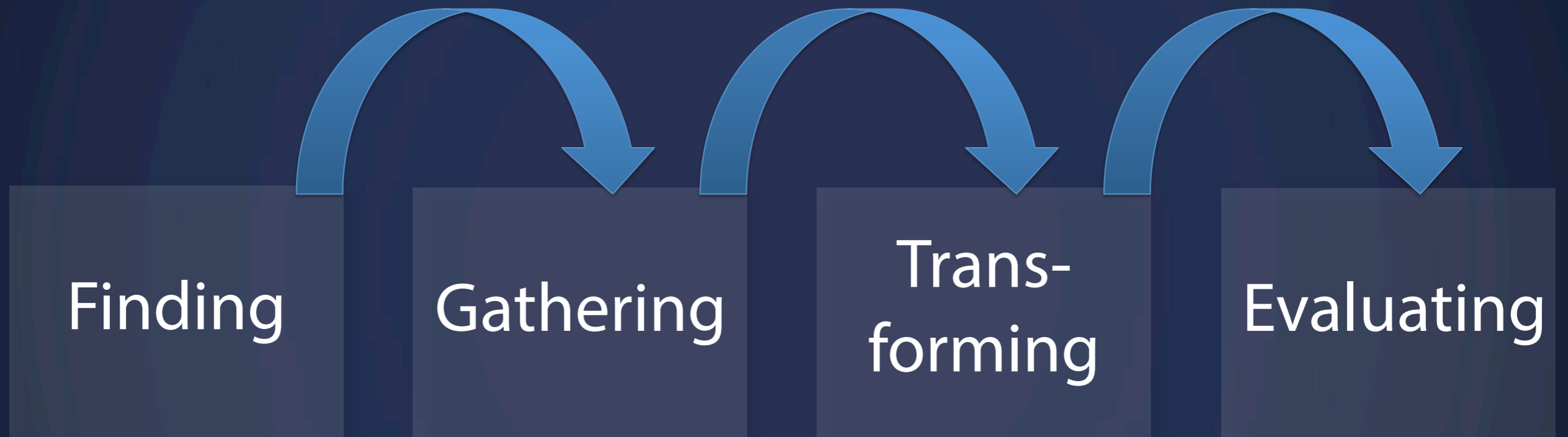
Data

Jan 24, 2013 – Michael Porath (@poezn)

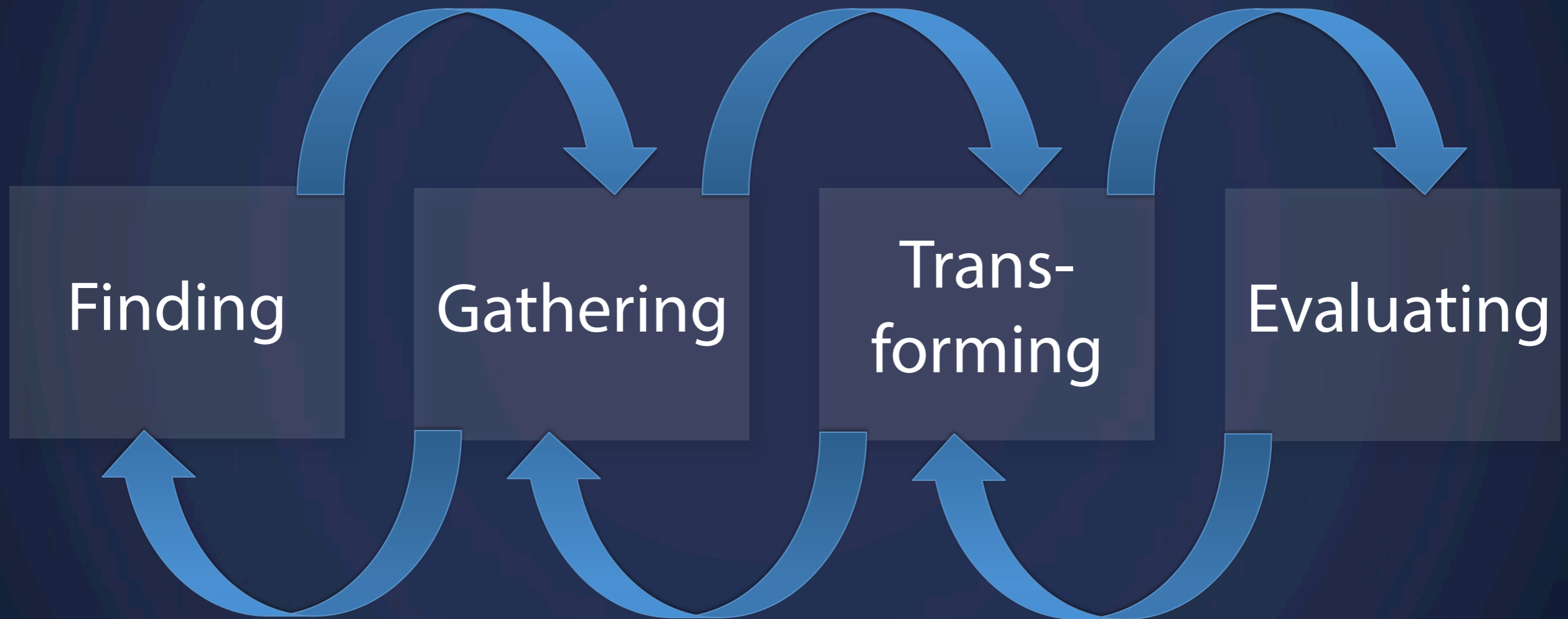
Process

Finding. Gathering. Transforming.

Process



Process





From (Idea To) Raw Data To Data



Finding data

The needle in the haystack

Open Government Data



Finding data

The needle in the haystack

Open Government Data



Product APIs



Finding data

The needle in the haystack

Open Government Data



Product APIs



Third Party Data Providers



Finding data

The needle in the haystack

Open Government Data



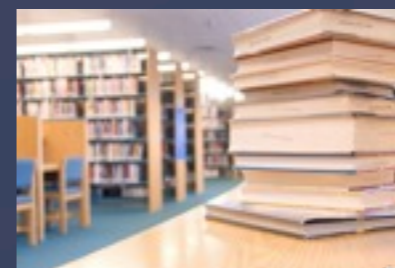
Product APIs



Third Party Data Providers



Anywhere



flickr / ccacnorthlib

| | A | B | C | D |
|----|---------|------|-------|---|
| 1 | Weekday | Hour | Count | |
| 2 | 0 | 0 | 1 | |
| 3 | 0 | 1 | 0 | |
| 4 | 0 | 2 | 1 | |
| 5 | 0 | 3 | 3 | |
| 6 | 0 | 4 | 0 | |
| 7 | 0 | 5 | 0 | |
| 8 | 0 | 6 | 0 | |
| 9 | 0 | 7 | 0 | |
| 10 | 0 | 8 | 0 | |
| 11 | 0 | 9 | 0 | |



Gathering data

The tedious part

Scraping

```
<td>Rank</td>
<td><a href="/index.php?sort=name">Name</a><td>Address</td>
<td>Neighborhood</td>
<td>Espresso <a href="/html/tasting-criteria.html">Espresso</td>
<td><a href="/index.php?sort=cafe">Cafe</a> <a href="/html/cate-criteria.html">Cafe</td>
<td><a href="/index.php?sort=overall">Overall</a> <a href="/html/overall.html">Overall</td>
</tr>
|
|  |

```

Collecting manually

Crowdsourcing



| | B | C | D | E | F | G | H |
|--------------|---------------|---------------|--------------|--------------|------|-----------------|---|
| link | name | address | neighborhood | rating_overs | id | rating_espresso | |
| /review-view | Blue Bottle C | 66 Mint St. | SOMA | 8.5 | 1064 | 8.5 | |
| /review-view | Blue Bottle C | 1 Ferry Build | Embarcadero | 8.1 | 1128 | 8.4 | |
| /review-view | Blue Bottle C | 315 Linden S | Hayes Valley | 8.3 | 820 | 8.4 | |
| /review-view | Coffee Bar | 1890 Bryant | Potrero Hill | 8.45 | 1059 | 8.4 | |
| /review-view | Epicenter Ca | 764 Harrison | SOMA | 8.3 | 1121 | 8.4 | |
| /review-view | Four Barrel C | 375 Valencia | Mission | 8.45 | 1070 | 8.4 | |
| /review-view | Mercury Caf | 201 Octavia | Hayes Valley | 8.2 | 1171 | 8.4 | |
| /review-view | Ritual Coffee | 1634 Jerrold | Bayview | 8.2 | 1016 | 8.4 | |
| /review-view | Ritual Coffee | 1026 Valenci | Mission | 8.45 | 843 | 8.4 | |
| /review-view | Cafe Capricci | 2200 Mason | North Beach | 8.05 | 1127 | 8.3 | |
| /review-view | Special Ytra | 46 Minna St | SOMA | 7.55 | 1177 | 8.2 | |

Transforming and evaluating data

Clean your data

- Missing data points
- inconsistent formats

Jan 24, 2013

2013/01/24

24/01/13

24th January 2013

Transforming and evaluating data

Clean your data

- Missing data points
- inconsistent formats

Jan 24, 2013

2013/01/24

24/01/13

24th January 2013

Identify your target format

SQL Database?

CSV?

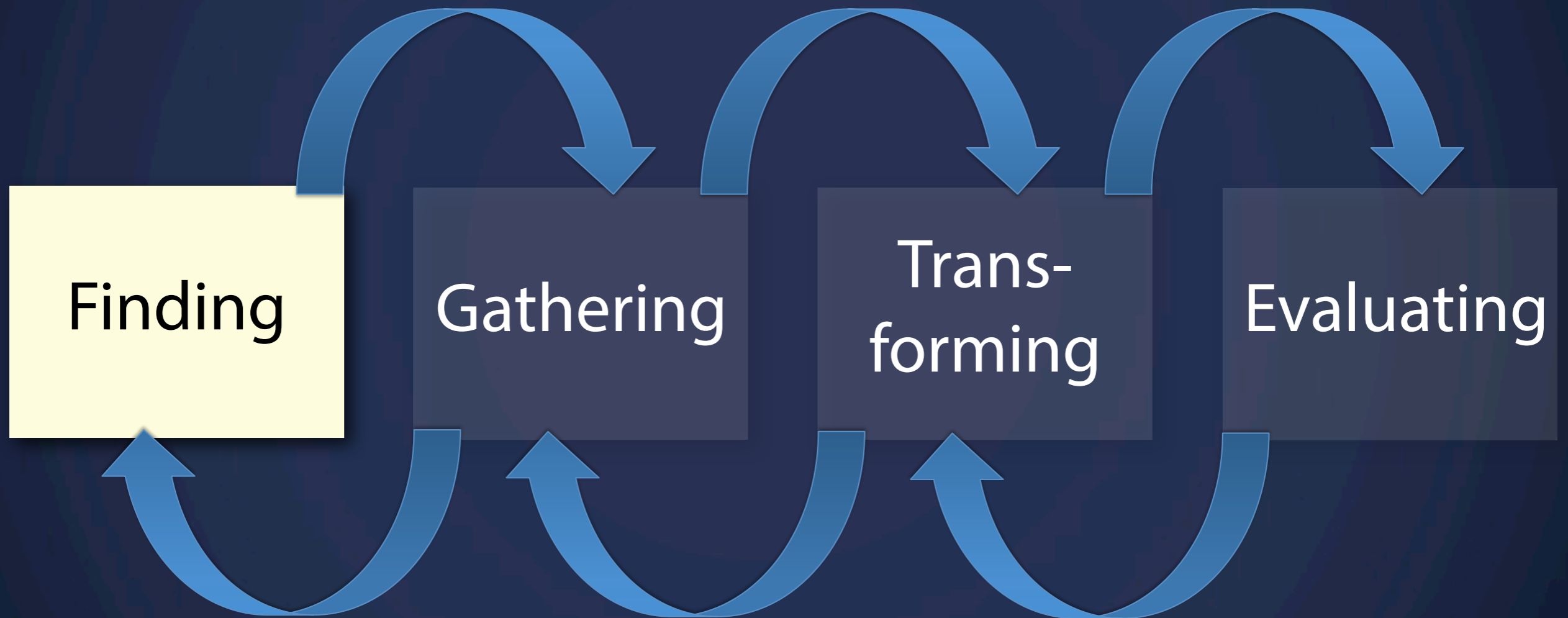
Excel?

JSON?

API?

Example

Process



THE NEW FILM BY
QUENTIN TARANTINO

DJANGO

UNCHAINED



The image is a movie poster for 'Django Unchained'. It features a bright red background with a large, black silhouette of a chain link running vertically down the center. At the bottom, two silhouettes of men in western attire are walking towards the viewer. The title 'Django' is written in large, white, stylized letters, and 'Unchained' is written in smaller, white, spaced-out letters below it. Above the title, the text 'THE NEW FILM BY QUENTIN TARANTINO' is written in white, bold, sans-serif font.

THE NEW FILM BY
QUENTIN TARANTINO
DJANGO
UNCHAINED

"I bet that's the most successful western so far"

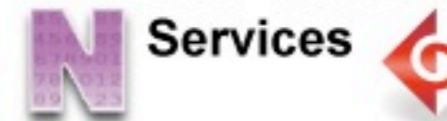
Question

Which is the most successful Western so far?

THE NUMBERS

BOX OFFICE DATA, MOVIE STARS, IDLE SPECULATION

Learn About Our
Research and Data
Services



Site Search

Google™ Cus Go

News

Latest News
Coming Soon
Trailers
The Crunch
Oscars

The Movies

Daily Chart
Weekend Chart
Theater Counts
Chart Archive
Movie Archive
Records
Top Rated
Popular
Budgets
Franchises
Keywords

Home Market

DVD Sales Chart
Blu-ray Sales Chart
2012 DVD Chart
2011 DVD Chart
2010 DVD Chart
Coming Soon
Archive

Market Analysis

Overview
2012
2011
2010
Distributors
Genres
MPAA Ratings
Sources
Prod'n Methods
Creative Types

International

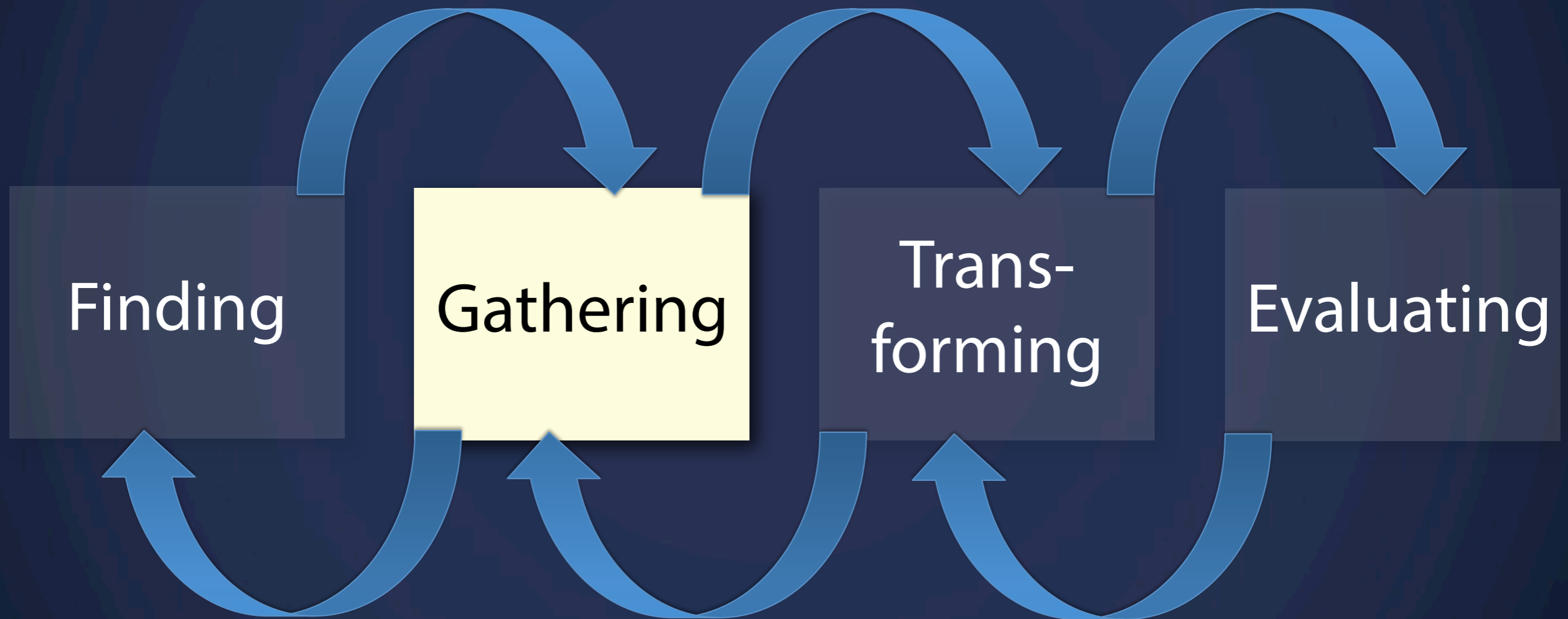
World Chart
News

Wednesday, January 23, 2013

Western Movies

| Movie | Release Date | Distributor | MPAA Rating | Domestic Gross | Inflation-Adjusted Gross |
|---|--------------|--------------------|-------------|----------------|--------------------------|
| Hondo | Jan 1, 1953 | | PG | \$8,200,000 | \$107,830,003 |
| The Alamo | Oct 24, 1960 | United Artists | PG-13 | \$7,900,000 | \$90,334,780 |
| Major Dundee | Apr 7, 1965 | Sony Pictures | PG-13 | \$14,873 | \$116,188 |
| Bandolero! | Jan 1, 1968 | 20th Century Fox | PG-13 | \$12,000,000 | \$72,274,806 |
| C'era una volta il West | May 28, 1969 | Paramount Pictures | PG-13 | \$5,321,508 | \$29,568,098 |
| Big Jake | Jan 1, 1971 | | PG-13 | \$7,500,000 | \$35,863,640 |
| Tombstone | Dec 25, 1993 | Walt Disney | R | \$56,505,065 | \$90,906,513 |
| Lightning Jack | Mar 11, 1994 | Savoy | PG-13 | \$16,821,273 | \$24,969,720 |
| Bad Girls | Apr 22, 1994 | 20th Century Fox | R | \$15,187,851 | \$25,815,733 |
| Maverick | May 20, 1994 | Warner Bros. | PG | \$101,631,272 | \$195,161,375 |
| Wyatt Earp | Jun 24, 1994 | Warner Bros. | PG-13 | \$25,052,000 | \$38,006,035 |
| Wagons East | Sep 9, 1994 | Sony Pictures | PG-13 | \$4,358,940 | \$8,429,424 |
| The Quick and the Dead | Feb 10, 1995 | Sony Pictures | R | \$18,552,460 | \$33,634,344 |
| Tall Tale | Mar 24, 1995 | Walt Disney | PG | \$8,247,627 | \$14,939,186 |
| Last of the Dogmen | Sep 8, 1995 | Savoy | PG | \$7,008,542 | \$12,666,709 |
| Wild Bill | Dec 1, 1995 | MGM | R | \$2,169,373 | \$3,934,790 |
| Dead Man | May 10, 1996 | Miramax | R | \$1,025,488 | \$1,830,567 |
| Ride With the Devil | Nov 24, 1999 | USA Films | R | \$630,779 | \$967,211 |
| Shanghai Noon | May 26, 2000 | Walt Disney | PG-13 | \$56,932,305 | \$83,338,756 |
| All the Pretty Horses | Dec 22, 2000 | Miramax | PG-13 | \$15,527,125 | \$22,275,742 |

Process



Tip

HTML pages to Excel/CSV/Spreadsheet

Google Docs

`=importHTML(URL, element, index)`

Tip

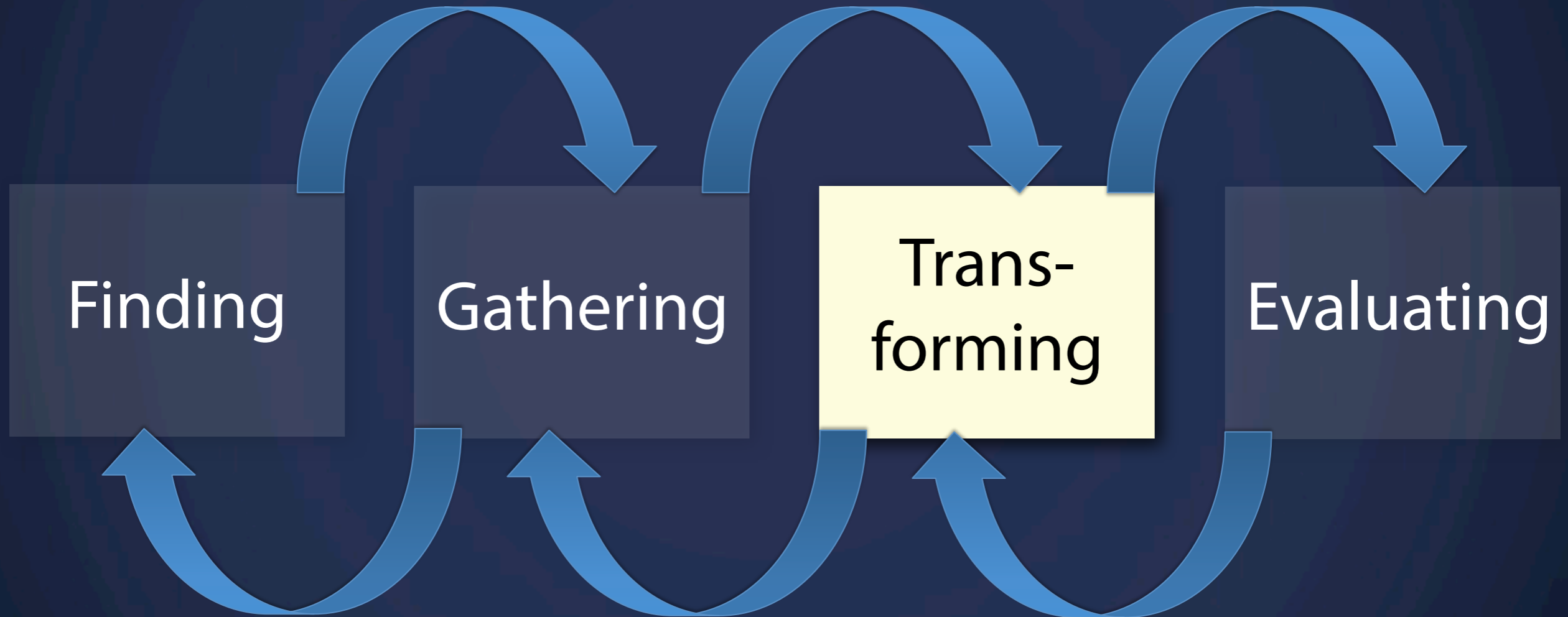
HTML pages to Excel/CSV/Spreadsheet

Google Docs

```
=importHtml(URL, element, index)
```

```
=importHtml(  
    "http://www.the-numbers.com/movies/genre/Western",  
    "table",  
    4  
)
```

Process



Data Cleaning

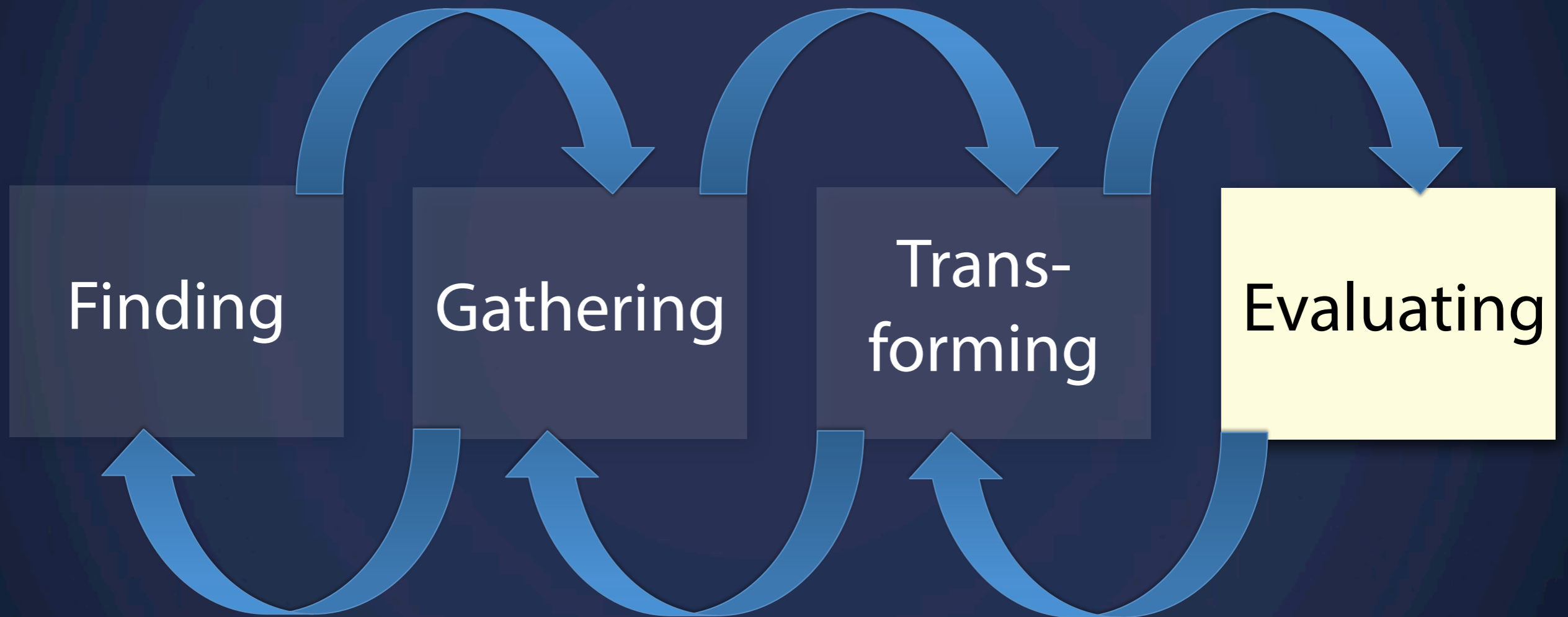
Stanford Text Wrangler

<http://vis.stanford.edu/wrangler/app/>

Open Refine

<http://openrefine.org/>

Process



Let's try again

<http://blogs.ischool.berkeley.edu/i247s13/lab-1-data/>

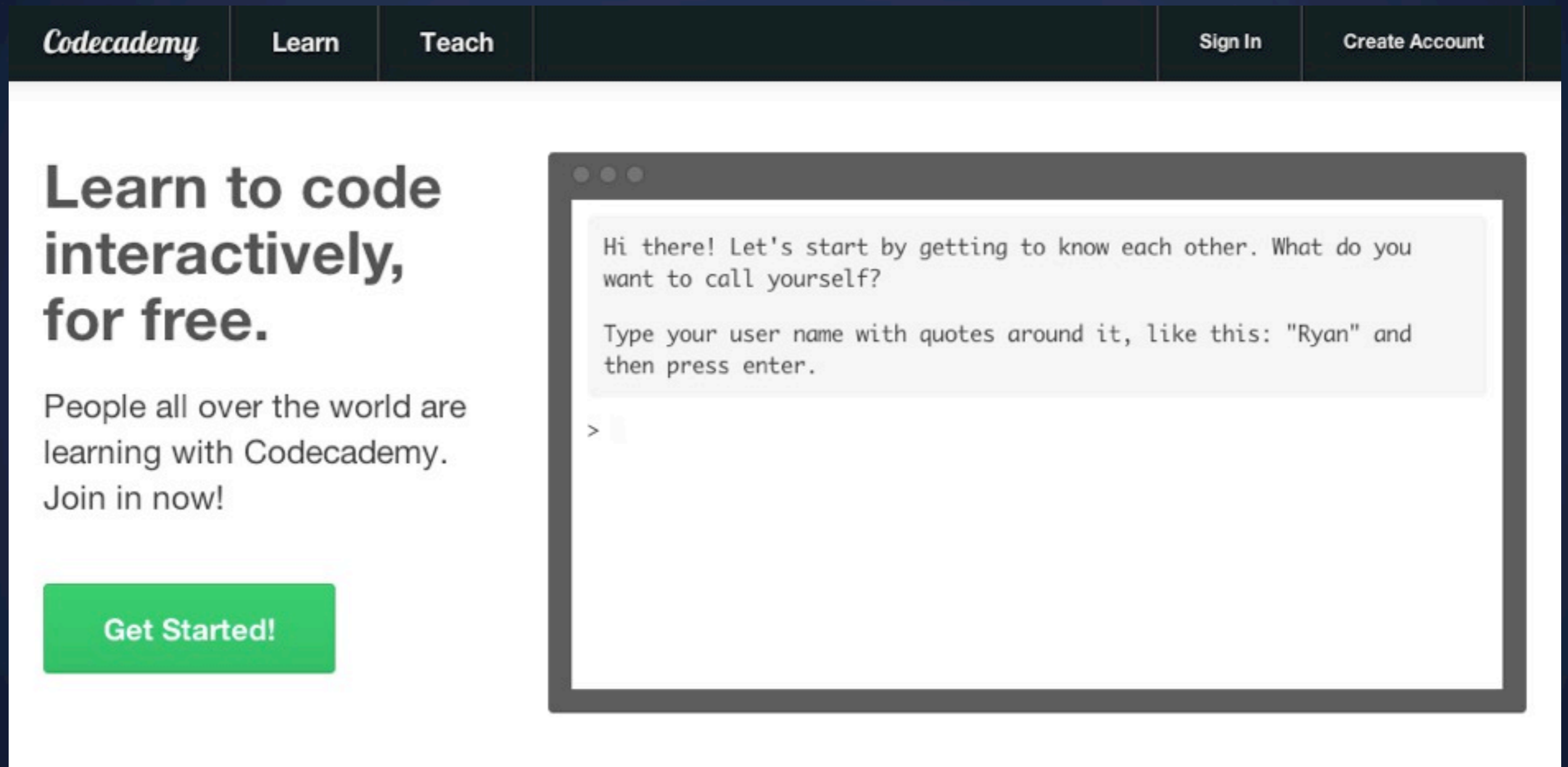
Questions?

Aaaaaargh!

I can't code!

Start somewhere

Just do it!



The screenshot shows the Codecademy website header with navigation links: Codecademy, Learn, Teach, Sign In, and Create Account. The main content area features a large heading, a subheading, a paragraph of text, and a green 'Get Started!' button. To the right is a terminal window with a message and instructions for a JavaScript exercise.

Codecademy Learn Teach Sign In Create Account

Learn to code interactively, for free.

People all over the world are learning with Codecademy. Join in now!

[Get Started!](#)

```
Hi there! Let's start by getting to know each other. What do you want to call yourself?  
  
Type your user name with quotes around it, like this: "Ryan" and then press enter.  
  
>
```

<http://www.codecademy.com/tracks/javascript>

Reminder

Assignment 1

Task Find and document 2 visualizations

- 1 online
- 1 “in the wild”

Deliverable 2 visualizations; 1 page writeup

Due Tuesday Jan 29, 3:00PM

More information on the class blog

Tuesday

From Data to Visualization