Art of the scrape!!!!

Show the internet
who's boss.
Scrape it!

Before we begin!

You might want….

A computer!

Server space!

Processing!

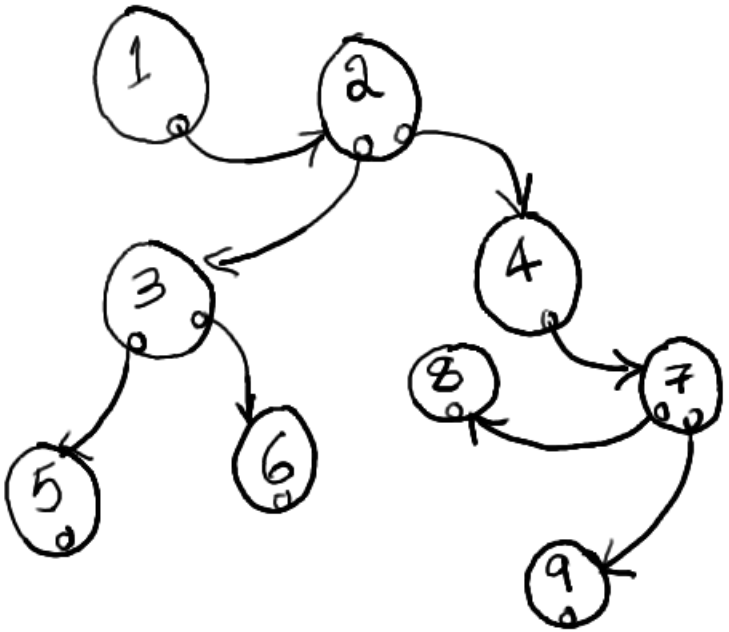
What's going on here?

Today we're going to….

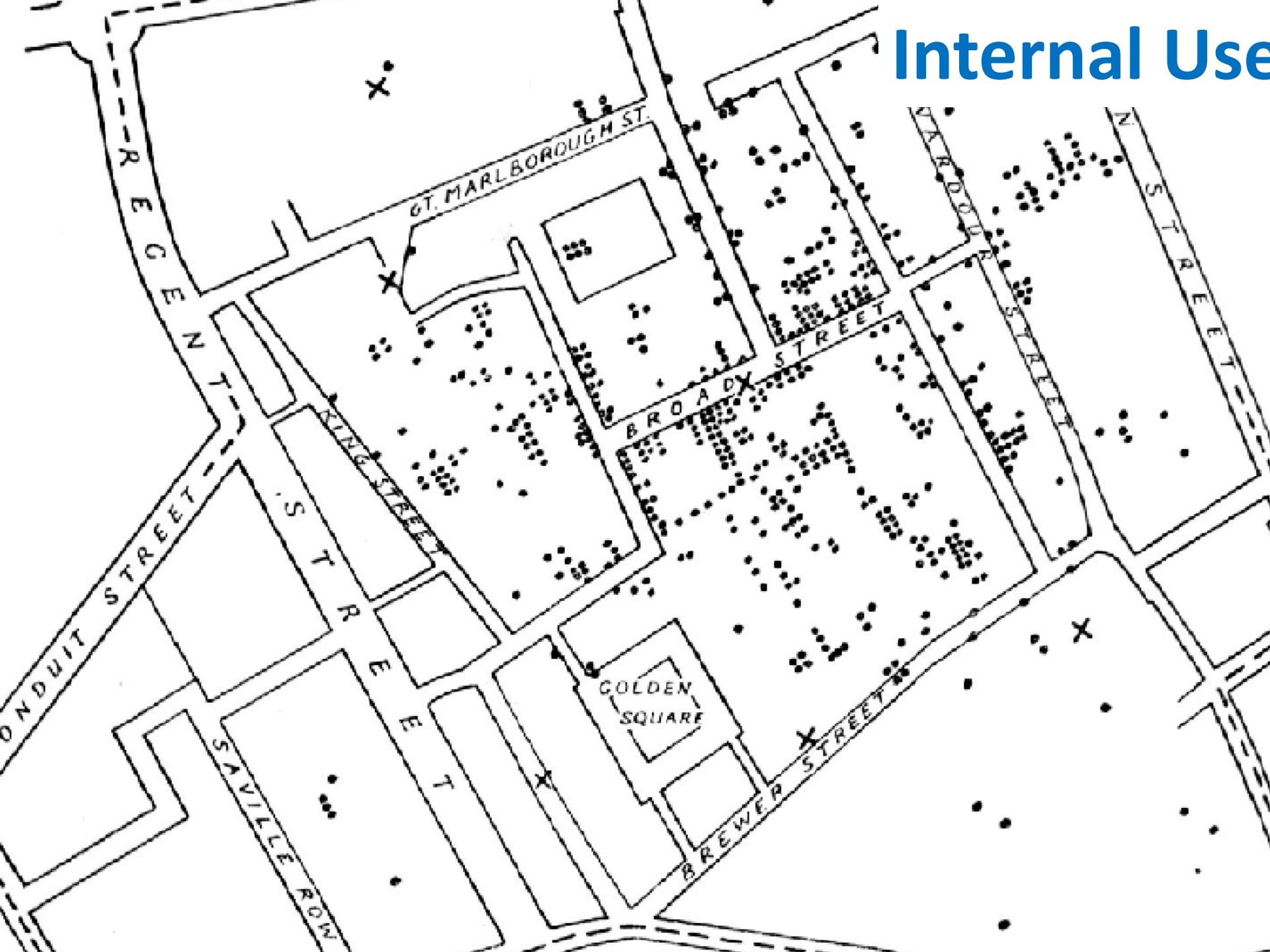
- Examine our data resources!
- Try some scraping!
- Try some pulling!
- Mess around with an API!
- Say hello to visualization!

Data? I hardly knew-a!

- Data: Any discreet unit and its meta information
- Useful data: More than one record of data...but that second record can be in your head!
- **Everything** is numbers!



Internal Use




External Use
50 St C·E
50 St 1·9
49 St N·R·W
southbound only
n-bound
5 Av/53 St E·V
51 St 6
47–50 Sts/ Rockefeller Ctr B·D·F·V
42 St Port Authority Bus Terminal A·C·E
Times Sq 42 St N·Q·R·S·W·1·2·3·7·9
42 St Bryant Pk B·D·F·V
5 Av 7
Grand Central 42 St S·4·5·6·7·Metro-North
Vernon Blvd–Jackson Av 7
Hunterspoint A
Hunters Point 7·LIRR
LINCOLN TUNNEL
Weehawken Ferry (Lincoln Harbor)
JAVITS CENTER
QUEENS MIDTOWN TUNNEL
NJTransit • Amtrak
34 St Penn Station A·C·E·LIRR
34 St Penn Station 1·2·3·9·LIRR
34 St Herald Sq B·D·F N·Q·R·V·W
33 St 6
28 St N·R·W
28 St 6
28 St 1·9
23 St C·E
23 St 1·9
23 St F·V
23 St N·R·W
23 St 6
14 St–Union Sq L·N·Q·R·W·4·5·6
L·N·Q·R·W
18 St 1·9
8 Av L
6 Av L
14 St A·C·E
14 St 1·2·3·9
14 St F·V
3 Av L
1 Av L
Hunters Point Ferry
Long Island City
GREENPOINT
Bedford Av L
8 St NYU N·R·W
Astor Pl 6
EAST VILLAGE
WASHINGTON SQUARE PARK
Christopher St PATH
W 4 St–Wash Sq Pk A·B·C·D·E·F·V
Bleecker St 6
B'way–Lafayette St B·D·F·V
Lower East Side 2 Av F·V
NOHO
EAST RIVER PARK
Ferries: E90 St • E34 St Hunters Point
Christopher St Sheridan Sq 1·9
GREENWICH VILLAGE
Houston St 1·9
Prince St N·R·W
Spring St C·E
Spring St 6
SOHO
LITTLE ITALY
Delancey St Essex St F·J·M·Z
LOWER EAST SIDE
WILLIAMSBURG BRIDGE
Canal St 1·9
Canal St A·C·E
Canal St J·M·N·Q R·W·Z·6
Grand St B·D
Bowery J·M·Z
TRIBECA
CHINATOWN
NAVY YARD
Franklin St 1·9
East Broadway F
MANHATTAN BRIDGE
Chambers St A·C
Chambers St 1·2·3·9
City Hall R·W
Chambers St J·M·Z
Brooklyn Bridge–City Hall 4·5·6
Park Place 2·3
World Trade Center E
BATTERY PARK CITY
FULTON LANDING
BROOKLYN BRIDGE
High St A·C
WTC Site
WTC E
Cortlandt St 1·9
Cortlandt St R·W
Fulton St–Broadway Nassau A·C·J·M·Z
BROOKLYN HEIGHTS
Court St
City Hall
Subway R W
Bklyn Bridge–City Hall
Subway J M Z 4 5 6
4,5,6 only
NYC Transit Bus
M1 Fifth/Madison Avs
M6 Broadway/Sixth Av
M15 First/Second Avs
M22 Madison St
M103 Third/Lexington Avs
B51 Downtown Brooklyn
Broadway–Nassau Fulton Street
Subway A C J M Z 2 3 4 5
NYC Transit Bus
Marcy Av
Subway J M Z
NYC Transit Bus
B24 Greenpoint Av
B39 Williamsburg Br
B44 Nostrand Av
B46 Utica Av
B60 Wilson Av
Q54 Metropolitan Av

Tell me more of this data of which you speak!

Real-time

- Blogs
- Twitter feed
- News feeds…
- Etc

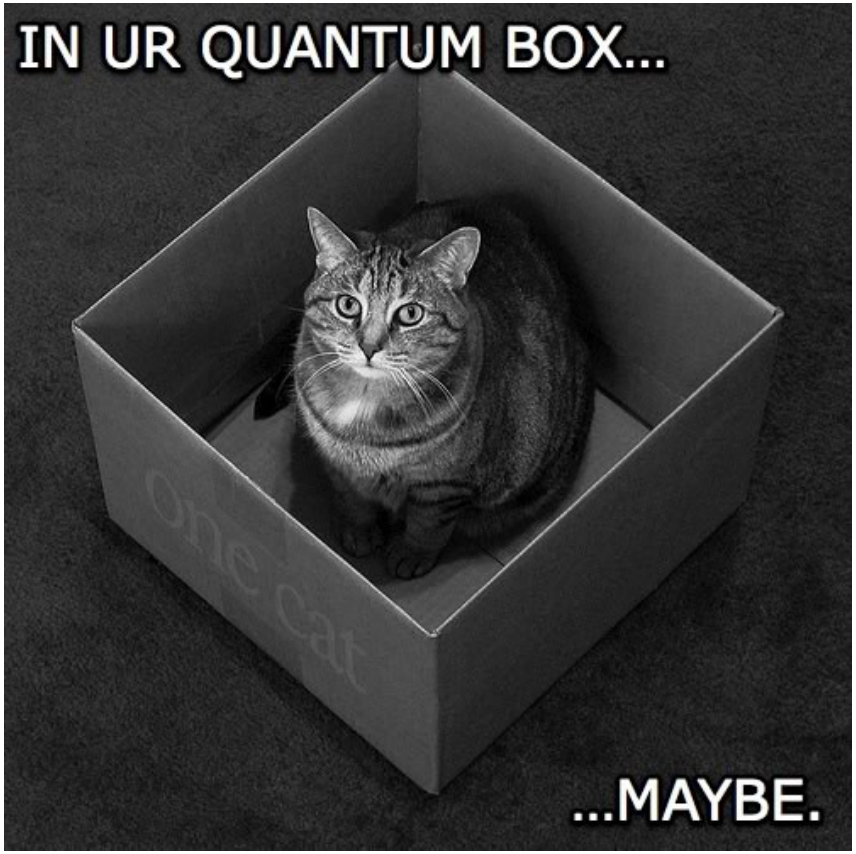
Static data sets

- Gov't census data,
- EPA data
- National League salaries
- etc

Data is Powerful!

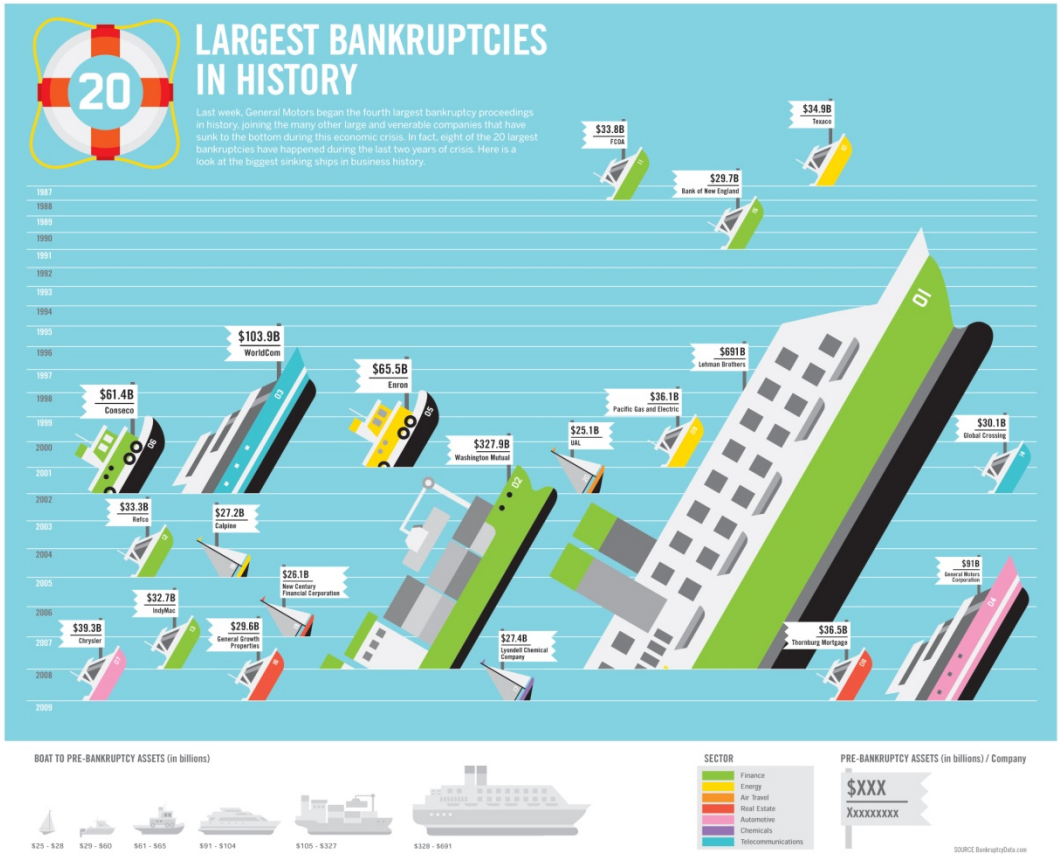
The act of measuring
something
solidifies its state.

Ahh, the power!!!



Data is misleading!

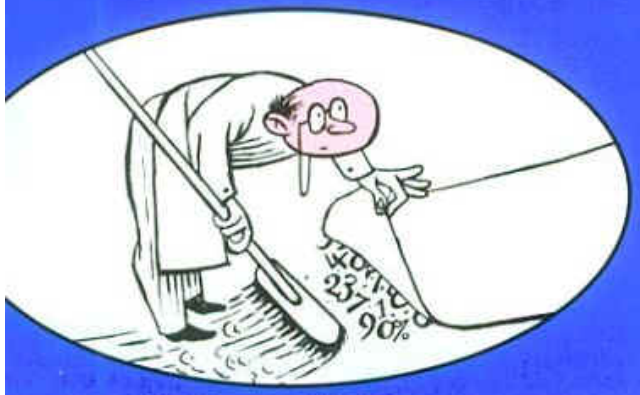
- Choosing one source over another
- Only portraying parts of the statistic
- Choosing a biased method of portrayal



HOW TO LIE WITH STATISTICS

Darrell Huff

Illustrated by Irving Geis



Over Half a Million Copies Sold—
An Honest-to-Goodness Bestseller

Information Overload: don’t believe the hype

Flavors of data

- **Indexed data** - documents, weblogs, images, videos, shopping articles, jobs ...

- **Cartographic and geographic data** - Geolocation software, Geovisualization

- **News Aggregators** - Feeds, podcasts:

DATATYPE!

- **Straight text**
- **CSV/ tab delimited**
- **XML/RSS/ATOM**
- **JSON**

Would it fit in here? Then its data!

VIA

- Text file
- Data feed
- Scraping html
- API
- Some combination

Could it potentially be transferred by this? Then it's grabable!

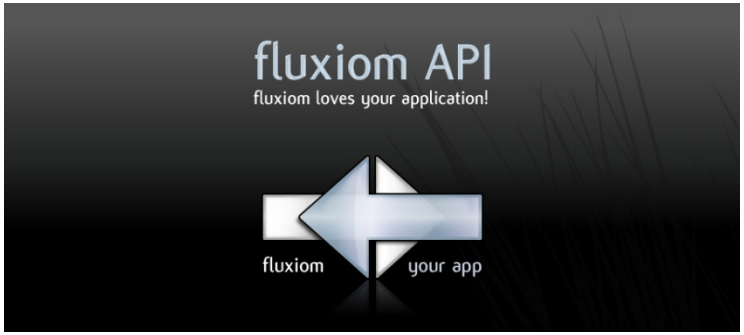
DESTINATION

- **Spreadsheet** *(by hand)*
- **Browser** *(direct, javascript, php, perl...)*
- **Database** *(via sql using php, perl, etc....)*
- **Application** *(Processing, java, python)*
- **A second API**

Mom, where does data come from?

HTML for scraping: Anywhere you can see text online

- Weather.com
- Yahoo trending topics



Preformatted data sets: Anywhere it's available

- Amazon data sets
- opendata.gov

Realtime rss feeds: Anywhere there's a data feed

- Any blog feed
- Any news feed

Personalized Awesome targeted data: Anywhere with an API.

- New York times API
- Twitter API

Choose wisely!

DATATYPE	VIA	DESTINATION
xml/rss	Browser	Excel
csv	text file	php: database
xml	api	php:browser
xml	api	javascript:browser
html	scraping	php browser
csv	text file	Processing
html	scraping	Processing
xml	browser	Processing *(through php)*

Example 1 and 2

Datatype	VIA	Destination
HTML	SCRAPING	BROWSER
(Weather info)	(PHP)	(Firefox, or whatever)

- Step one: Get to know your data:
  http://www.weather.com/weather/today/New+York+NY+10010?lswe=10010
- Step two: Set up the code

10010 Weather Forecast and Conditions - Mozilla Firefox

File Edit View History Bookmarks Tools Help

http://www.weather.com/weather/today/New+York+NY+10010?lswe=10010

Secure Search

Google simpleml Search Share Check Translate AutoFill

Google Calendar | http://ww...mple2.php | 10010 Weather F... | The Full Feed from H... | Google Hot Trends | Learning Processing:...

United States (English) | Local weather in 1-click | Put weather on my desktop | Customize weather.com | Sign In

The Weather Channel weather.com

Local Weather Site Web powered by Google

10010 Search

Maps | Video | Photos | World | Mobile | Alerts

Home | Weather News | Travel | Driving | Health | Home & Family | Sports | Outdoor Activities | Climate & Green | On TV

10010 Weather

Print Email Desktop Mobile RSS

See the Weather Your Way...

Traffic Reports | Life-saving Tornado Tips | Photos: Big Catches! | Raw: Huge Tornadoe

Choose Theme | Yesterday | Today | Hour-by-Hour | Tomorrow | Weekend | 10-Day | Month | Radar Map

New York, NY (10010) Weather

° F | ° C

Updated: Jun 19, 2010, 10:45am EDT

Right Now

76° F

Feels Like: **76° F**

Wind: **From SSW at 7mph**

Boat & Beach

Coastal Marine Summary

Sea Temp: **65° F**
Sea Height: **1.02 ft**

Hunters Point/ Newtown Creek Tide Times

Find: chxl Next Previous Highlight all Match case

Done

Secure Search

Example 1: Straight scrapin’

```
<?php

Get the data!

Do Something with it!

?>
```

Example 1

```
<?php

$url =
    'http://www.weather.com/weather/today/New+Y
    ork+NY+10010?lswe=10010';
$output = file_get_contents($url);

echo $output;

?>
```

Example 2: Scraping with a purpose

Get everything ready

Get the data!

Do Something with it!

Example 2

```
$currentTerm = NULL;  //we'll use this to hold the words!
$myUrl = "http://www.google.com/trends/hottrends/atom/hourly
$searchForStart = "sa=X\">";
$searchForEnd = "</a>";
```

```
$rawPage = file_get_contents($myUrl);
```

```
echo "<B>These are this hour's trending topics on Google!</b><BR><BR>";
while ($startPos = (strpos($rawPage, $searchForStart))) {  //as long as there's more stuff to find, find it!
                $endPos = strpos($rawPage, $searchForEnd); //And then find where it ends!
                $length = $endPos - $startPos;                    //How long is this string we've found, anyway?

                if ($startPos && $endPos) {                                        //Did we find something?  Then
      $currentTerm = substr($rawPage, ($startPos+strlen($searchForStart)), $length-6);
                              echo $currentTerm  . "<BR>";
                } //end if
 $rawPage = substr($rawPage, ($endPos + 4));
} //end while
```

Example 3

Datatype	VIA	Destination
XML	RSS FEED	BROWSER
(Huffington post)	(PHP)	(Firefox, or whatever)

- Step one: Get to know your data:
  http://feeds.huffingtonpost.com/huffingtonpost/raw_feed
- Step two: Set up the code

What's this xml stuff?

```
<introductory tags>
<entry>
        <title></title>
          <id></id>
          <published></published>
          <updated>2010-06-19T15:50:45Z</updated>
          <summary>summary>
          <author>
             <name></name>
             <uri>http://www.huffingtonpost.com/anne-naylor/</uri>
          </author>
          <content></ content>
</entry>
```

Example 3: XML makes things awesome

Get the data!

Get it in a form we can use

Do Something with it!

Example 3: XML makes things awesome

```
$url = "http://feeds.huffingtonpost.com/huffingtonpost/raw_feed";
$data = file_get_contents($url);
```

```
$xml = new SimpleXmlElement($data);
```

```
echo "<b>Here are the current popular posts from Huffington Post without the
     ads!</b><BR><BR><ul>";

foreach ($xml->entry as $item)  {  //navigate to the tag we want?

                                $myTitle = "unknown";  //initialize the variable so it's all set!
                                $myTitle = trim($item->title);
                                echo "<LI>" . $myTitle . "<br>";            //now print it!
            }//end foreach
echo "</ul>";
```

Example 4

Datatype	VIA	Destination
XML	data FEED	API and BROWSER
(US Exchange rates)	(PHP)	(Google Charts API Firefox, or whatever)

- Step one: Get to know your data:
  http://rss.timegenie.com/forex.xml
- Step two: Set up the code

API's? Eh?

- **Data** –all the types of data we discussed before

- **Functionality**

  Data converters: *language translators, speech processing, url shorteners)*

  Communication: *email, IM, notifications*

  Visual data rendering: *Information visualization, diagrams, maps*

  Security related : electronic payment systems, ID identification...

Example 4: Doing the two-step

Get the data!

Get it in a form we can use

Run it through a second Process

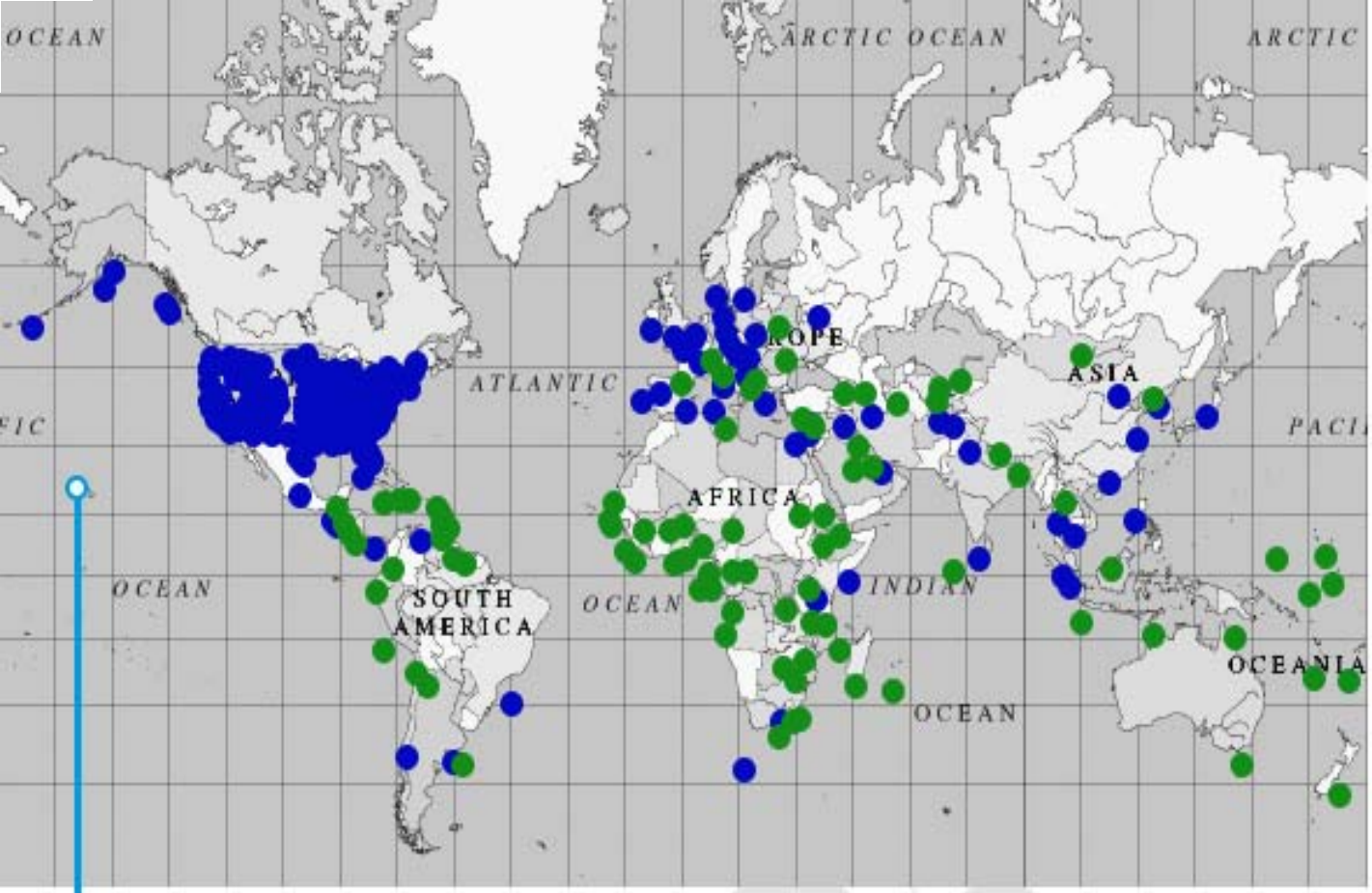
Do something with it (like displaying that baby!)

Bringing data into a higher-level Application…like processing!

- Install the simplml library:

  http://www.learningprocessing.com/tutorials/simpleml/

- Inspect your data for structure
- Write some code!
  - Declare your xml intent!
  - Make the request!
  - Process the request!
  - Do fun stuff with it!



Obama Is to Report Tuesday on Blagojevich Contacts

Honolulu, United States

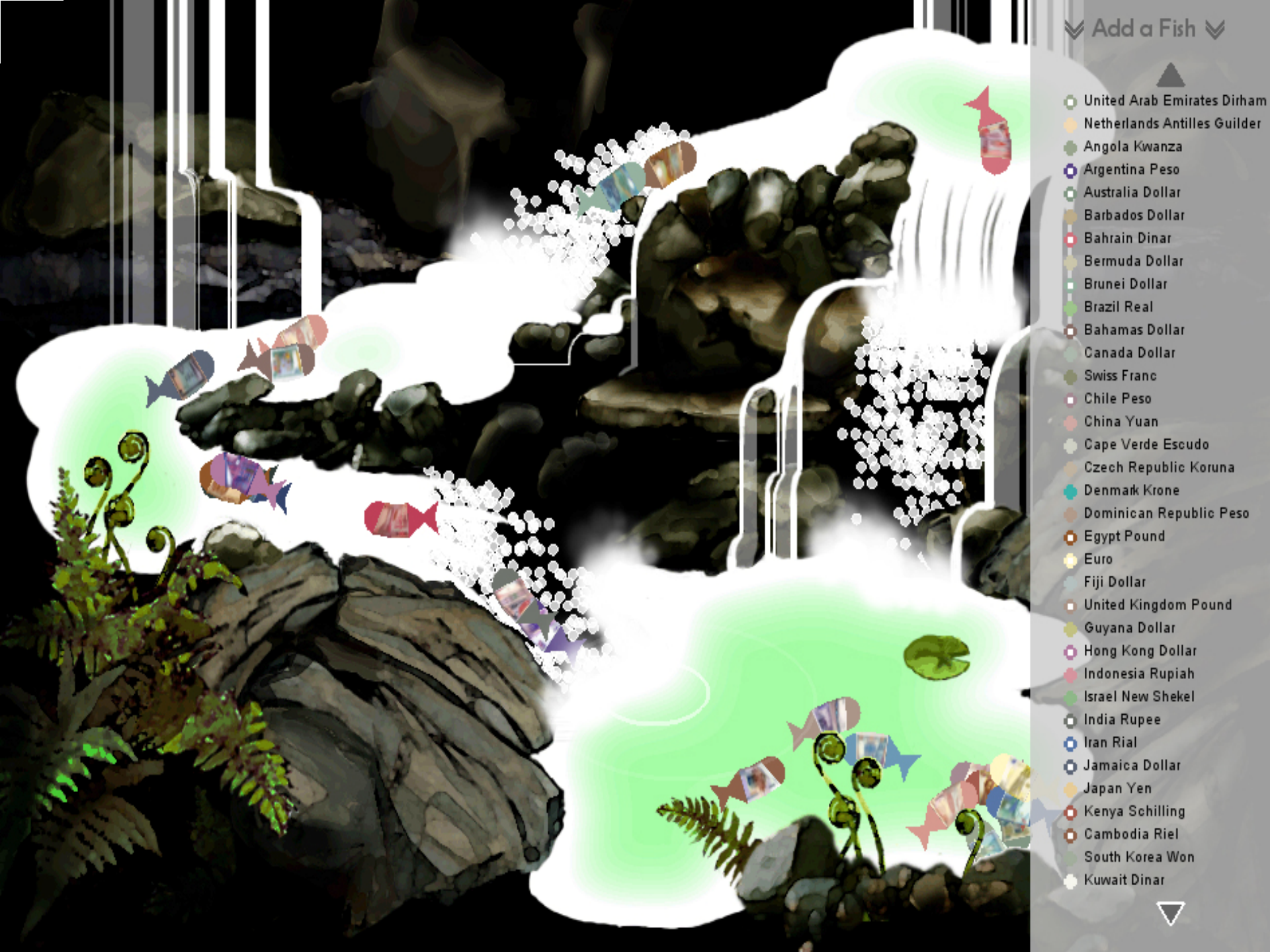
The memorandum will lay out a narrative in about a dozen paragraphs the president-elect's advisers had with the governor's office...

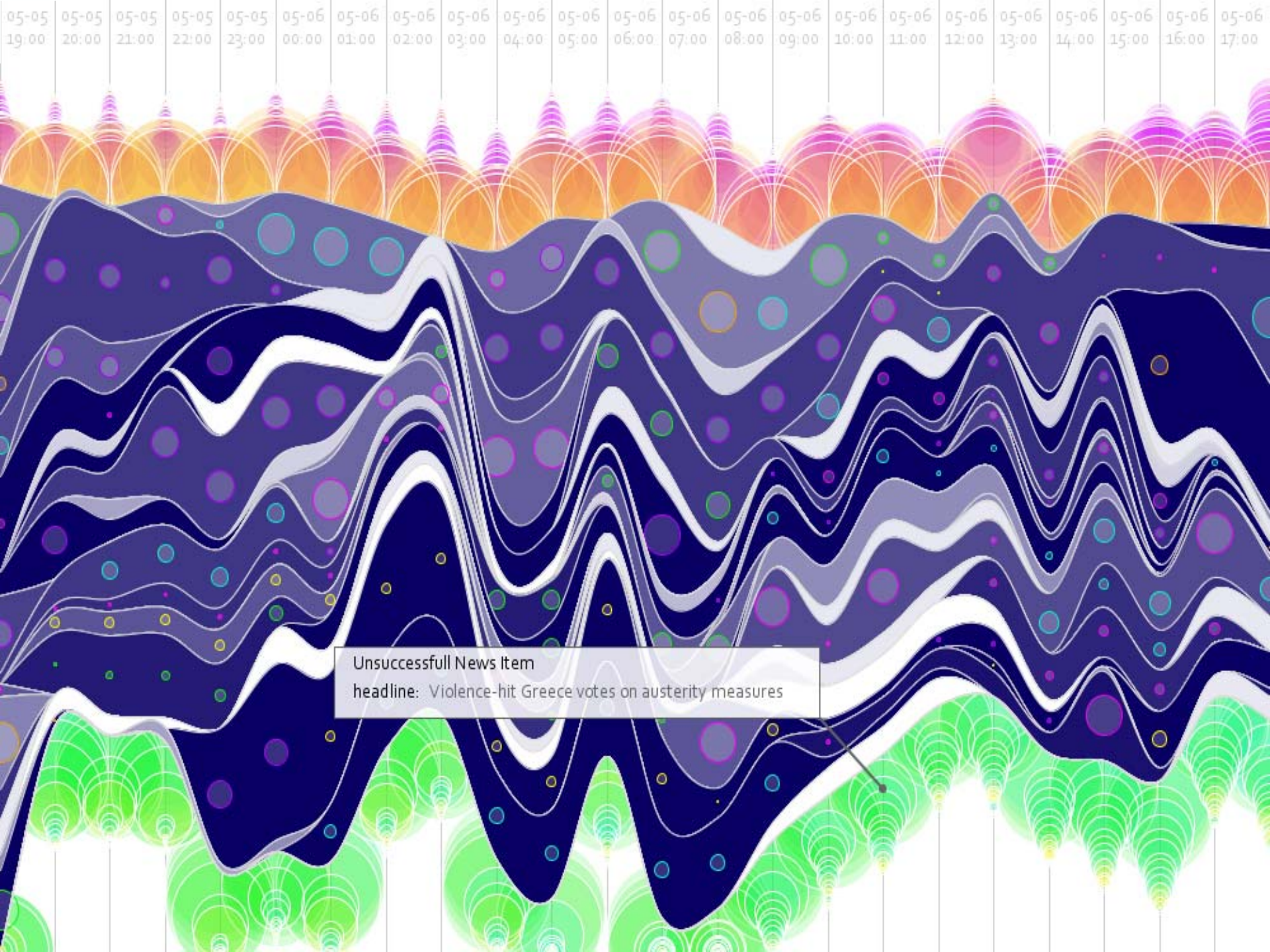
recently covered

- Washington DC, United States
- Kabul, Afghanistan
- Rome, Italy
- New Delhi, India
- Caracas, Venezuela
- Algiers, Algeria
- Buenos Aires, Argentina
- Vienna, Austria
- Brussels, Belgium
- Sarajevo, Bosnia and Herzegovina
- Rio De Janeiro, Brazil
- Phnom Penh, Cambodia
- Ottawa, Canada

recently missed

- Tirane, Albania
- Andorra la Vella, Andorra
- Luanda, Angola
- Yerevan, Armenia
- Canberra, Australia
- Baku, Azerbaijan
- Dhaka, Bangladesh
- Bridgetown, Barbados
- Minsk, Belarus
- Belmopan, Belize
- Porto-Novo, Benin
- La Paz, Bolivia
- Sucre, Bolivia


Add a Fish
United Arab Emirates Dirham
Netherlands Antilles Guilder
Angola Kwanza
Argentina Peso
Australia Dollar
Barbados Dollar
Bahrain Dinar
Bermuda Dollar
Brunei Dollar
Brazil Real
Bahamas Dollar
Canada Dollar
Swiss Franc
Chile Peso
China Yuan
Cape Verde Escudo
Czech Republic Koruna
Denmark Krone
Dominican Republic Peso
Egypt Pound
Euro
Fiji Dollar
United Kingdom Pound
Guyana Dollar
Hong Kong Dollar
Indonesia Rupiah
Israel New Shekel
India Rupee
Iran Rial
Jamaica Dollar
Japan Yen
Kenya Schilling
Cambodia Riel
South Korea Won
Kuwait Dinar


05-05 19:00
05-05 20:00
05-05 21:00
05-05 22:00
05-05 23:00
05-06 00:00
05-06 01:00
05-06 02:00
05-06 03:00
05-06 04:00
05-06 05:00
05-06 06:00
05-06 07:00
05-06 08:00
05-06 09:00
05-06 10:00
05-06 11:00
05-06 12:00
05-06 13:00
05-06 14:00
05-06 15:00
05-06 16:00
05-06 17:00
Unsuccessfull News Item
headline: Violence-hit Greece votes on austerity measures

Go out and do some scraping!

Zoe Fraade-Blanar

Fraade@gmail.com

www.binaryspark.com