GENERAL ASSEMBLY

DATA SCIENCE

DATA

LAST TIME:

I. PYTHON REVIEW

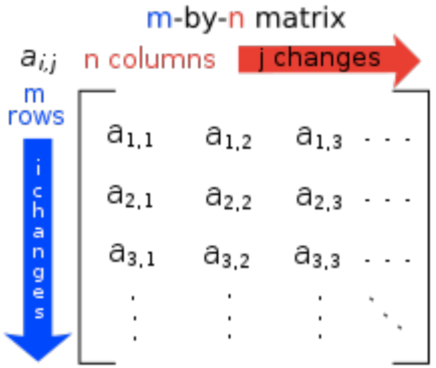
II. LINEAR ALGEBRA REVIEW

EXERCISES:

III. PYTHON

IV. NUMPY AND PANDAS

```
>>> a = [1,'b',True]
>>> a[2]
True
>>> a[1]='aa'
>>> a
[1,'aa',True]
```



QUESTIONS?

WHAT WAS THE MOST INTERESTING THING YOU LEARNT?

WHAT WAS THE HARDEST TO GRASP?

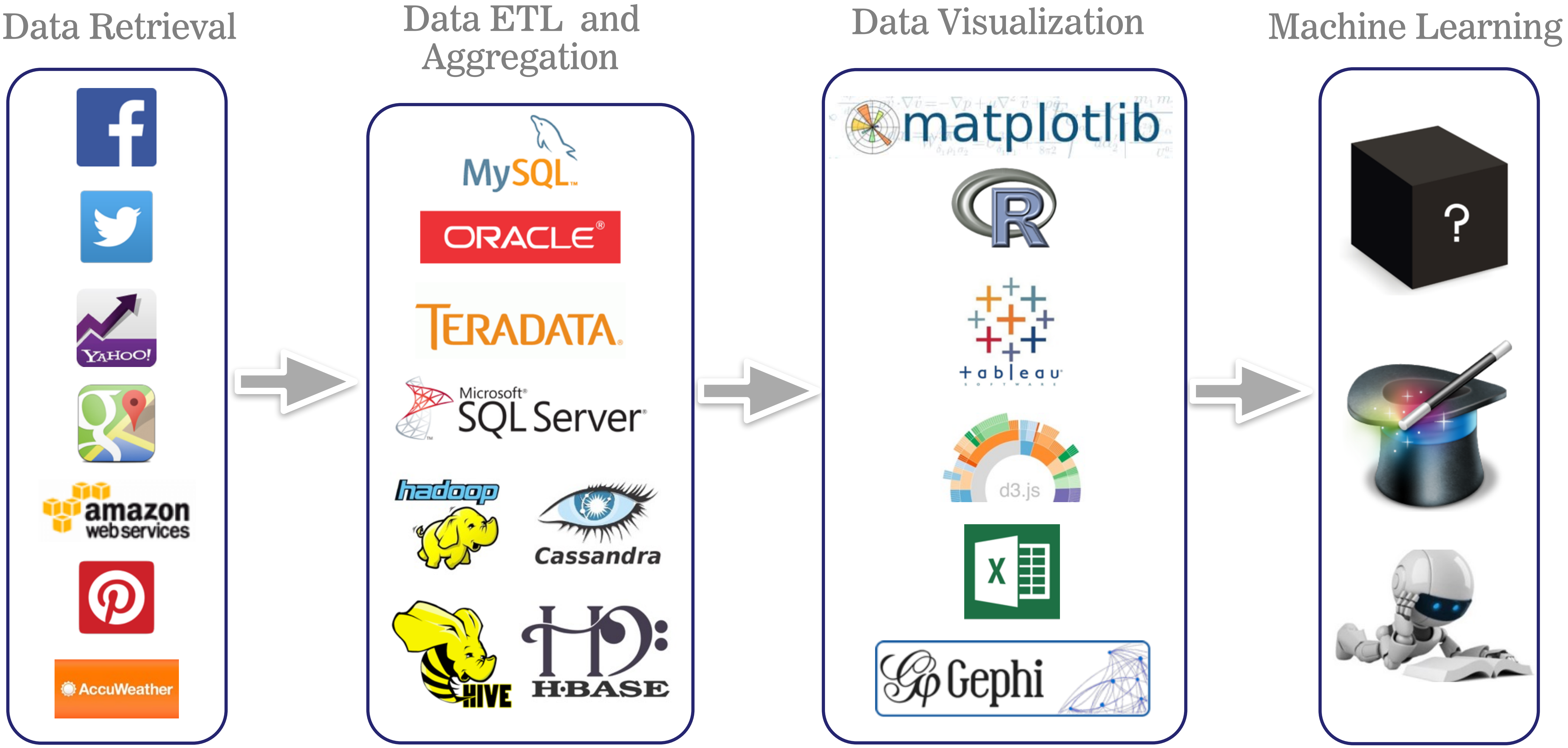
I. DATA SOURCES
II. DATA FORMATS
III. APIS
IV. CLEANING DATA
V. MISSING DATA

EXERCISES:
VI. KIMONO
VII. PANDAS
VIII. BOKEH

- LEARN ABOUT VARIOUS DATA SOURCES
- UNDERSTAND HOW TO EXTRACT DATA FROM APIS
- LEARN TO CLEAN AND IMPUTE DATA
- LEARN TO VISUALIZE THE DATA

WHERE DOES THE DATA COME FROM?

DATA FLOW


Data Retrieval
Data ETL and Aggregation
Data Visualization
Machine Learning
YAHOO!
amazon web services
AccuWeather
MySQL
ORACLE
TERADATA
Microsoft SQL Server
hadoop
Cassandra
HIVE
HBASE
matplotlib
R
tableau
d3.js
Gephi
?

DATA FLOW

Data Retrieval

YAHOO!
amazon web services
AccuWeather

Data ETL and Aggregation

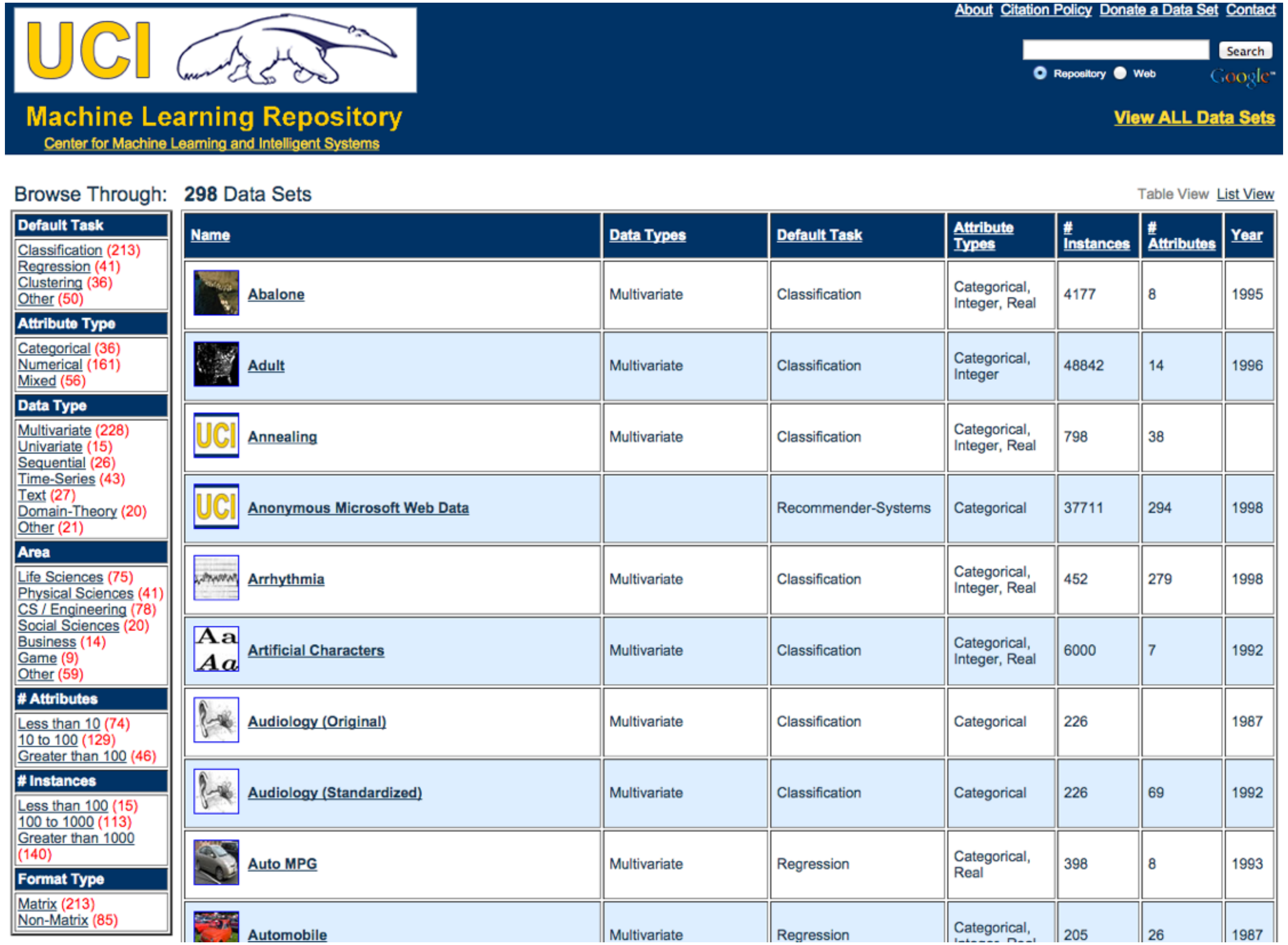
MySQL
ORACLE
TERADATA
Microsoft SQL Server
hadoop
Cassandra
HIVE
HBASE

Data Visualization

matplotlib
R
tableau
d3.js
Gephi

Machine Learning

DATA SOURCES



UCI Machine Learning Repository
Center for Machine Learning and Intelligent Systems

About Citation Policy Donate a Data Set Contact

Search
Repository Web Google

View ALL Data Sets

Browse Through: 298 Data Sets

Table View List View

Default Task
- Classification (213)
- Regression (41)
- Clustering (36)
- Other (50)

Attribute Type
- Categorical (36)
- Numerical (161)
- Mixed (56)

Data Type
- Multivariate (228)
- Univariate (15)
- Sequential (26)
- Time-Series (43)
- Text (27)
- Domain-Theory (20)
- Other (21)

Area
- Life Sciences (75)
- Physical Sciences (41)
- CS / Engineering (78)
- Social Sciences (20)
- Business (14)
- Game (9)
- Other (59)

 # Attributes
- Less than 10 (74)
- 10 to 100 (129)
- Greater than 100 (46)

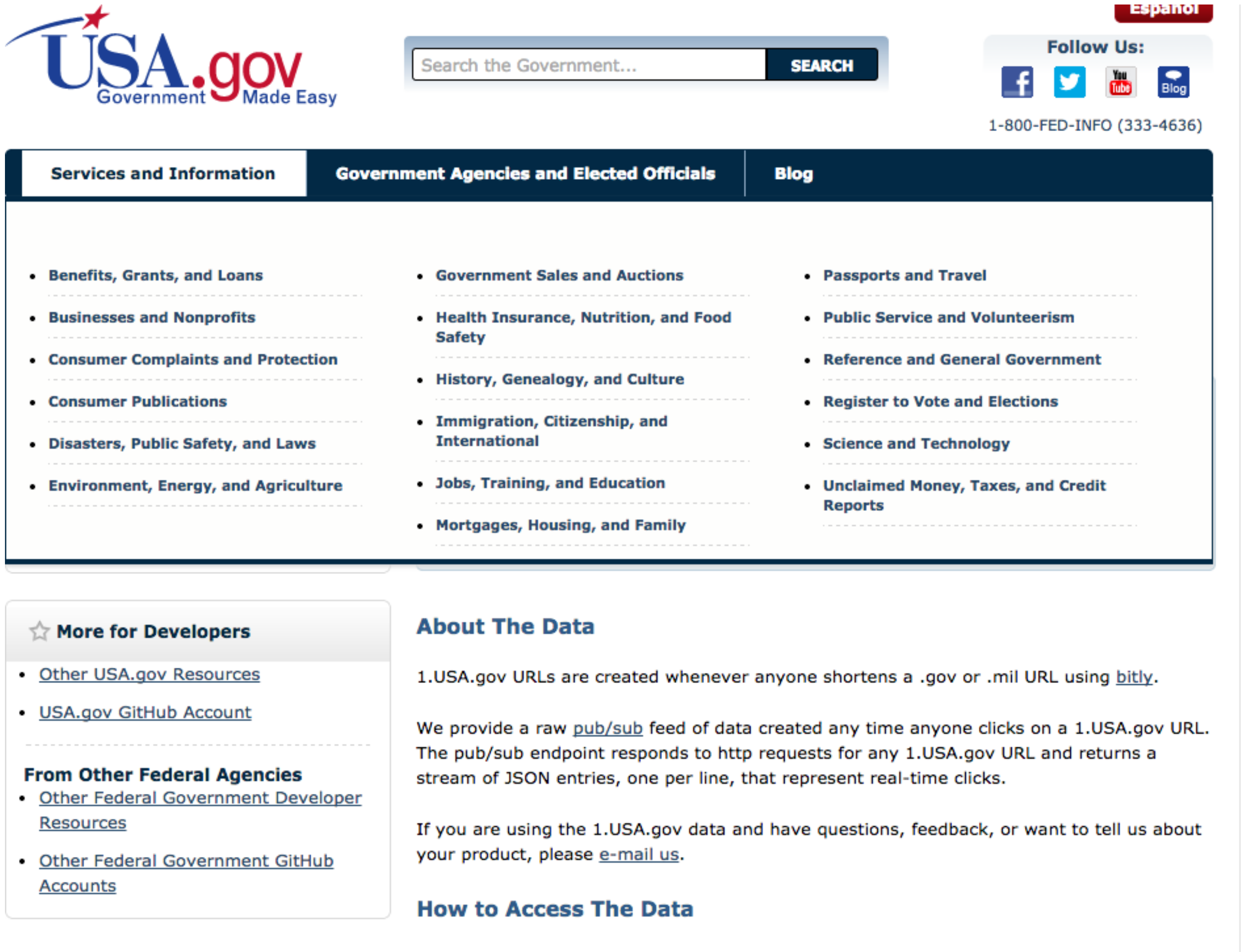
 # Instances
- Less than 100 (15)
- 100 to 1000 (113)
- Greater than 1000 (140)

Format Type
- Matrix (213)
- Non-Matrix (85)

Name	Data Types	Default Task	Attribute Types	# Instances	# Attributes	Year
Abalone	Multivariate	Classification	Categorical, Integer, Real	4177	8	1995
Adult	Multivariate	Classification	Categorical, Integer	48842	14	1996
Annealing	Multivariate	Classification	Categorical, Integer, Real	798	38	
Anonymous Microsoft Web Data		Recommender-Systems	Categorical	37711	294	1998
Arrhythmia	Multivariate	Classification	Categorical, Integer, Real	452	279	1998
Artificial Characters	Multivariate	Classification	Categorical, Integer, Real	6000	7	1992
Audiology (Original)	Multivariate	Classification	Categorical	226		1987
Audiology (Standardized)	Multivariate	Classification	Categorical	226	69	1992
Auto MPG	Multivariate	Regression	Categorical, Real	398	8	1993
Automobile	Multivariate	Regression	Categorical, Integer, Real	205	26	1987

Source: [*http://archive.ics.uci.edu/ml/datasets.html*](http://archive.ics.uci.edu/ml/datasets.html)

DATA SOURCES


Español

USA.gov Government Made Easy

Search the Government... SEARCH

Follow Us:

Blog

1-800-FED-INFO (333-4636)

Services and Information | Government Agencies and Elected Officials | Blog

- Benefits, Grants, and Loans
- Businesses and Nonprofits
- Consumer Complaints and Protection
- Consumer Publications
- Disasters, Public Safety, and Laws
- Environment, Energy, and Agriculture
- Government Sales and Auctions
- Health Insurance, Nutrition, and Food Safety
- History, Genealogy, and Culture
- Immigration, Citizenship, and International
- Jobs, Training, and Education
- Mortgages, Housing, and Family
- Passports and Travel
- Public Service and Volunteerism
- Reference and General Government
- Register to Vote and Elections
- Science and Technology
- Unclaimed Money, Taxes, and Credit Reports

More for Developers

- Other USA.gov Resources
- USA.gov GitHub Account

From Other Federal Agencies

- Other Federal Government Developer Resources
- Other Federal Government GitHub Accounts

About The Data

1.USA.gov URLs are created whenever anyone shortens a .gov or .mil URL using bitly.

We provide a raw pub/sub feed of data created any time anyone clicks on a 1.USA.gov URL. The pub/sub endpoint responds to http requests for any 1.USA.gov URL and returns a stream of JSON entries, one per line, that represent real-time clicks.

If you are using the 1.USA.gov data and have questions, feedback, or want to tell us about your product, please e-mail us.

How to Access The Data

Source: http://www.usa.gov/About/developer-resources/1usagov.shtml

DATA SOURCES

Source: http://www.kaggle.com/

1) PETE SKOMOROCH (LINKEDIN) HTTPS://DELICIOUS.COM/PSKOMOROCH/DATASET

2) HILARY MASON (ACCEL PARTNERS, BITLY) HTTPS://BITLY.COM/BUNDLES/HMASON/1

3) KEVIN CHAI (U. OF NEW SOUTH WALES, SYDNEY) HTTP://KEVINCHAI.NET/DATASETS

4) JEFF HAMMERBACHER (CLOUDERA) HTTP://WWW.QUORA.COM/JEFF-HAMMERBACHER/INTRODUCTION-TO-DATA-SCIENCE-DATA-SETS

5) JERRY SMITH (3I-MIND) HTTP://DATASCIENTISTINSIGHTS.COM/2013/10/07/DATA-REPOSITORIES-MOTHERS-MILK-FOR-DATA-SCIENTISTS/

6) GREGORY PIATETSKY-SHAPIRO (KDD) HTTP://WWW.KDNUGGETS.COM/DATASETS/INDEX.HTML

7) HTTP://WWW.QUORA.COM/DATA/WHERE-CAN-I-FIND-LARGE-DATASETS-OPEN-TO-THE-PUBLIC

8) HTTPS://GITHUB.COM/CAESAR0301/AWESOME-PUBLIC-DATASETS

Source: http://blog.mortardata.com/post/67652898761/6-dataset-lists-curated-by-data-scientists

PAIR EXERCISE:

CHOOSE A DATA SOURCE AND LOOK AT WHAT DATA YOU CAN GET

DISCUSS HOW YOU WOULD USE THE DATA

QUESTIONS?

JSON, CSV, ETC…

JSON (JavaScript Object Notation) is:

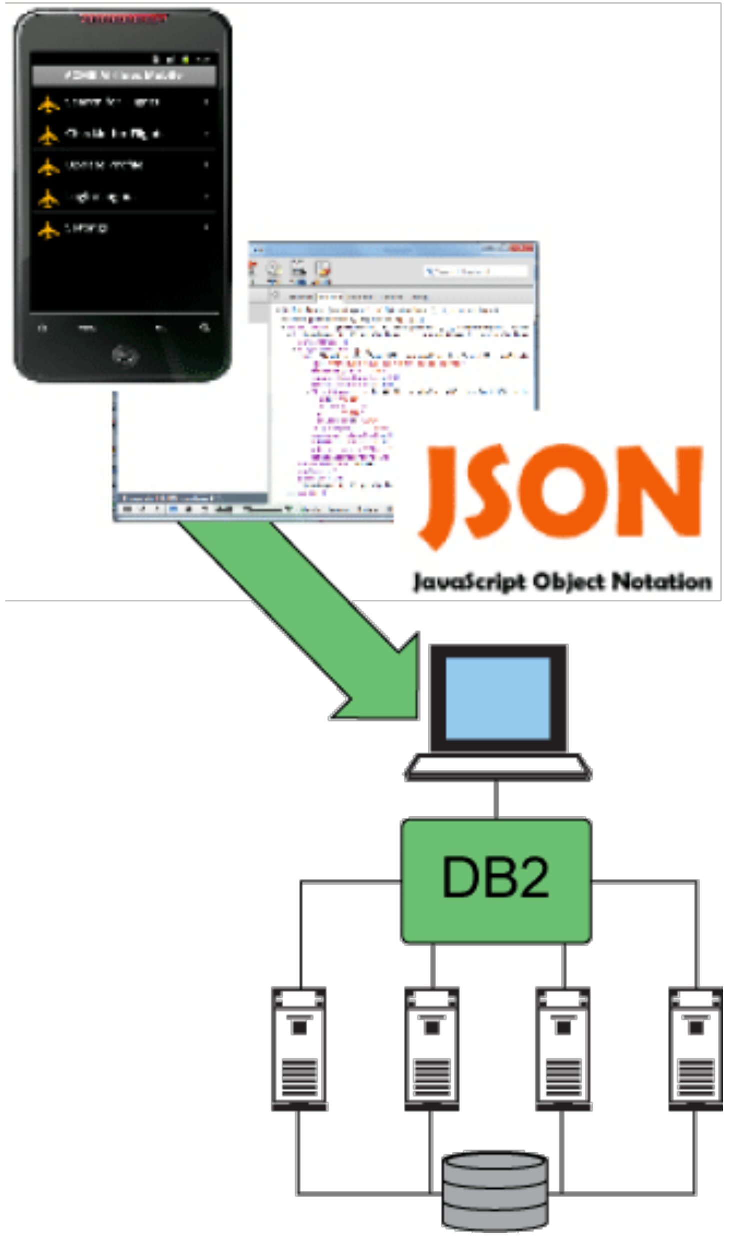
a lightweight data-interchange format

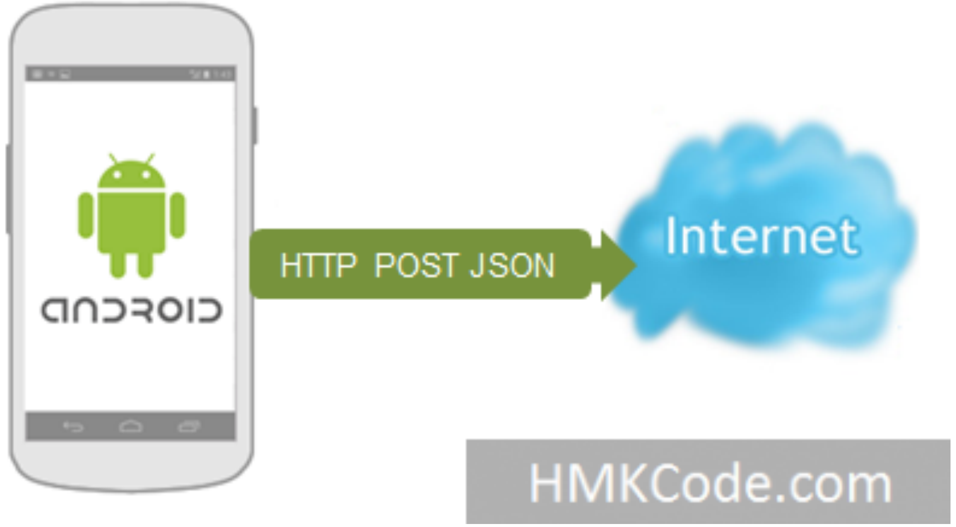
a string

JSON can be passed

between applications

easy for machines to parse and generate





JSON are passed through applications

as strings

and converted into native objects per language.

JSON are passed through applications

as strings

and converted into native objects per language.

```
{ "empinfo" :
    {
          "employees" : [
          {
            "name" : "Scott Philip",
            "salary" : £44k,
            "age" : 27,

          },

          {
            "name" : "Tim Henn",
            "salary" : £40k,
            "age" : 27,
          },

          {
           “name” : “Long Yong”,
           “salary” : £40k,
           “age” : 28,

          }
                    ]
      }
}
```

```
import json
py_object = [ { 'a':'A', 'b':(2, 4), 'c':3.0 } ]
json_string = json.dumps(py_object)
print ‘JSON:’, json_string
```

```
JSON: [{"a": "A", "c": 3.0, "b": [2, 4]}]
```

```
decoded = json.loads(json_string)
```

https://docs.python.org/2/library/json.html

CSV (Comma Separated Values):

```
name,game,points
John,basketball,3
Mary,volleyball,5
James,ping pong,2
…
```

CSV (Comma Separated Values):

- `easy to read and write`
- `structured like a table`
- `very common`
- `can export to/from MS Excel`

https://docs.python.org/2/library/csv.html

OTHER DATA FORMATS

txt

tsv

xml

dat

images

binary

etc…

APIs

APIs (Application Programming Interface) allow people to interact with the structures of an application

- get
- put
- delete
- update
- …

Best practices for APIs are to use RESTful principles.

Best practices for APIs are to use RESTful principles.

Representational State Transfer (REST)

RESTFUL EXAMPLE

RESTful API HTTP methods

Resource	GET	PUT	POST	DELETE
Collection URI, such as `http://example.com/resources/`	**List** the URIs and perhaps other details of the collection's members.	**Replace** the entire collection with another collection.	**Create** a new entry in the collection. The new entry's URI is assigned automatically and is usually returned by the operation.[9]	**Delete** the entire collection.
Element URI, such as `http://example.com/resources/item17`	**Retrieve** a representation of the addressed member of the collection, expressed in an appropriate Internet media type.	**Replace** the addressed member of the collection, or if it does not exist, **create** it.	Not generally used. Treat the addressed member as a collection in its own right and **create** a new entry in it.[9]	**Delete** the addressed member of the collection.

http://en.wikipedia.org/wiki/Representational_state_transfer

- The Base URL
- An interactive media type (usually JSON)
- Operations (GET, PUT, POST, DELETE)
- Driven by http requests

REST API EXAMPLE

Collection

GET https://api.instagram.com/v1/users/10

Operation

REST API EXAMPLE

GET https://api.instagram.com/v1/users/search/?q=andy

Querystring

https://dev.twitter.com/rest/public

https://developer.linkedin.com/docs/signin-with-linkedin

LIST OF PYTHON APIS

http://www.pythonapi.com/

PAIR EXERCISE: http://www.pythonapi.com/

1) CHOOSE 1 API: WHAT DATA YOU CAN GET?

2) INSTALL PYTHON MODULE, TRY TO EXTRACT DATA

3) DISCUSS: HOW COULD YOU LEVERAGE THAT API? HOW COULD YOU USE THE DATA?

KIMONO LABS

www.kimonolabs.com

kimono

Turn websites into structured APIs from your browser in seconds

Get started, click to install

QUESTIONS?

CLEANING DATA

DATAIST (HILARY MASON & FRIENDS)

1. Obtain - pointing and clicking does not scale (APIs, Python, shell scripting)
2. Scrub - “Scrubbing data is the least sexy part of the analysis process, but often one that yields the greatest benefits” (Python, sed, awk, grep)
3. Explore - look at the data (visualizing, clustering, dimensionality reduction)
4. Model - “All models are wrong, but some are useful” / models are built to predict and interpret!
5. Interpret - “The purpose of computing is insight, not numbers”

Source: http://www.dataists.com/2010/09/a-taxonomy-of-data-science/

DATAIST (HILARY MASON & FRIENDS)

1. Obtain - pointing and clicking does not scale (APIs, Python, shell scripting)
2. Scrub - “Scrubbing data is the least sexy part of the analysis process, but often one that yields the greatest benefits” (Python, sed, awk, grep)
3. Explore - look at the data (visualizing, clustering, dimensionality reduction)
4. Model - “All models are wrong, but some are useful” / models are built to predict and interpret!
5. Interpret - “The purpose of computing is insight, not numbers”

Source: http://www.dataists.com/2010/09/a-taxonomy-of-data-science/

FOR BIG-DATA SCIENTISTS, ‘JANITOR WORK’ IS KEY HURDLE TO INSIGHTS

From NYTimes on August 18, 2014:

“Data wrangling is a huge — and surprisingly so — part of the job,” said Monica Rogati, vice president for data science at Jawbone, whose sensor-filled wristband and software track activity, sleep and food consumption, and suggest dietary and health tips based on the numbers. “It’s something that is not appreciated by data civilians. At times, it feels like everything we do.”

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

DATA MUNGING IS AWESOME

Obtain Data
Scrub Data
Explore
Model Algorithms
iNterpret Results

80%

20%

Majority of time
is spent data munging

DATA CLEANSING

Data cleansing, data cleaning or data scrubbing is the process of detecting and correcting (or removing) corrupt or inaccurate records from a record set, table, or database.

Source: http://en.wikipedia.org/wiki/Data_cleansing

DATA CLEANSING

Data cleansing, data cleaning or data scrubbing is the process of detecting and correcting (or removing) corrupt or inaccurate records from a record set, table, or database.

Remove inconsistencies

Source: http://en.wikipedia.org/wiki/Data_cleansing

DATA CLEANSING

Data cleansing, data cleaning or data scrubbing is the process of detecting and correcting (or removing) corrupt or inaccurate records from a record set, table, or database.

Remove inconsistencies
Data type harmonization

Source: http://en.wikipedia.org/wiki/Data_cleansing

DATA CLEANSING

Data cleansing, data cleaning or data scrubbing is the process of detecting and correcting (or removing) corrupt or inaccurate records from a record set, table, or database.

Remove inconsistencies
Data type harmonization
Standardization, Normalization

Source: http://en.wikipedia.org/wiki/Data_cleansing

DATA CLEANSING

Data cleansing, data cleaning or data scrubbing is the process of detecting and correcting (or removing) corrupt or inaccurate records from a record set, table, or database.

Remove inconsistencies
Data type harmonization
Standardization, Normalization
Typos correction, Formatting (eg. timestamps)

Source: http://en.wikipedia.org/wiki/Data_cleansing

DATA CLEANSING

Data cleansing, data cleaning or data scrubbing is the process of detecting and correcting (or removing) corrupt or inaccurate records from a record set, table, or database.

Remove inconsistencies
Data type harmonization
Standardization, Normalization
Typos correction, Formatting (eg. timestamps)
Missing data

Source: http://en.wikipedia.org/wiki/Data_cleansing

DATA CLEANSING

Data cleansing, data cleaning or data scrubbing is the process of detecting and correcting (or removing) corrupt or inaccurate records from a record set, table, or database.

Remove inconsistencies
Data type harmonization
Standardization, Normalization
Typos correction, Formatting (eg. timestamps)
Missing data
Sorting

Source: http://en.wikipedia.org/wiki/Data_cleansing

MISSING DATA

MISSING DATA

- Understand the reasons why data are missing
- Random or not?
- If random, the data sample may still be representative of the population.
- If not random analysis may be harder

Source: http://en.wikipedia.org/wiki/Missing_data

MISSING DATA

- Understand the reasons why data are missing
- Random or not?
- If random, the data sample may still be representative of the population.
- If not random analysis may be harder

- Missing completely at random (MCAR)

Source: http://en.wikipedia.org/wiki/Missing_data

MISSING DATA

- Understand the reasons why data are missing
- Random or not?
- If random, the data sample may still be representative of the population.
- If not random analysis may be harder

- Missing completely at random (MCAR)
- Missing at random (MAR)

Source: http://en.wikipedia.org/wiki/Missing_data

MISSING DATA

- Understand the reasons why data are missing
- Random or not?
- If random, the data sample may still be representative of the population.
- If not random analysis may be harder

- Missing completely at random (MCAR)
- Missing at random (MAR)
- Missing not at random (MNAR)

Source: http://en.wikipedia.org/wiki/Missing_data

MISSING COMPLETELY AT RANDOM (MCAR)

- Missing value (y) neither depends on x nor y
- Example: some survey questions asked of a simple random sample of original sample

- When data are MCAR, the analyses performed on the data are unbiased; however, data are rarely MCAR.

Source: http://en.wikipedia.org/wiki/Missing_data

MISSING AT RANDOM (MAR)

- Missing value (y) depends on x, but not y
- Example: Respondents in service occupations less likely to report income

Source: http://en.wikipedia.org/wiki/Missing_data

MISSING NOT AT RANDOM (MNAR)

- The probability of a missing value depends on the variable that is missing

- Example: Respondents with high income less likely to report income

Source: http://en.wikipedia.org/wiki/Missing_data

TECHNIQUES TO DEAL WITH MISSING DATA

- Imputation, Partial imputation
- Deletion, Partial deletion
- Analysis
- Interpolation

Source: http://en.wikipedia.org/wiki/Missing_data

TECHNIQUES TO DEAL WITH MISSING DATA

- 1. Identify patterns/reasons for missing and recode
- correctly
- 2. Understand distribution of missing data
- 3. Decide on best method of analysis

Source: https://www.utexas.edu/cola/centers/prc/_files/cs/Missing-Data.pdf

LINKS

- https://www.utexas.edu/cola/centers/prc/_files/cs/Missing-Data.pdf
- http://www.uvm.edu/~dhowell/StatPages/More_Stuff/Missing_Data/Missing.html
- http://en.wikipedia.org/wiki/Missing_data
- https://www.coursera.org/course/getdata

VISUALIZATION

Data Retrieval

Data ETL and Aggregation

MySQL
ORACLE
TERADATA
Microsoft SQL Server
hadoop
Cassandra
HIVE
HBASE

Data Visualization

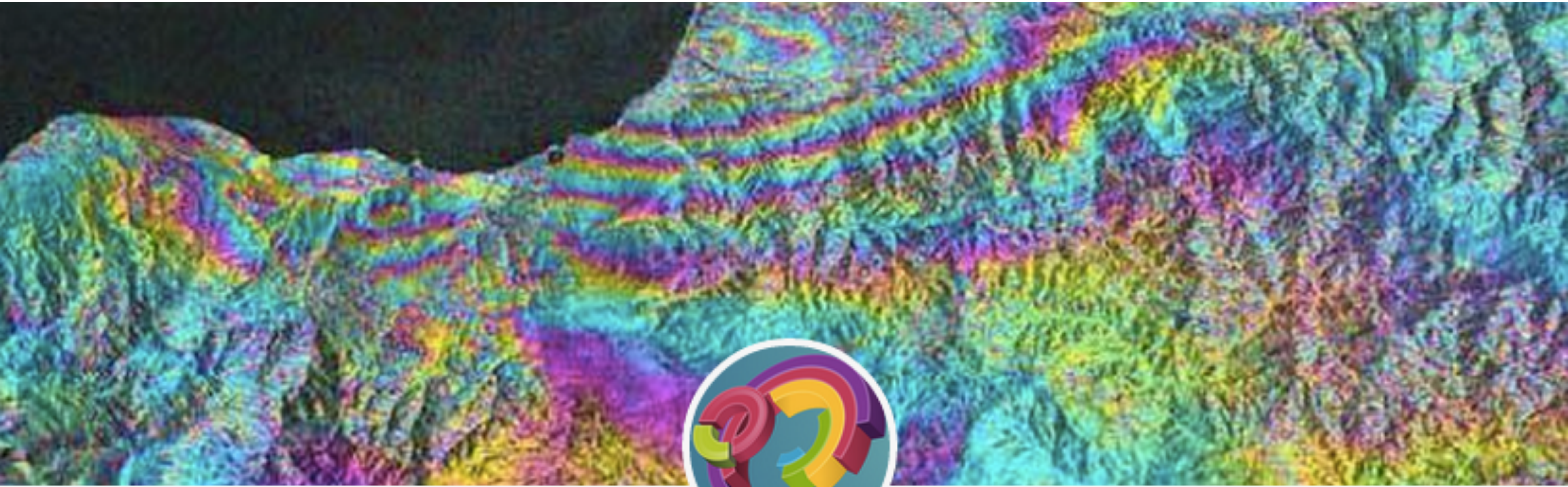
matplotlib
R
tableau
d3.js
Gephi

Machine Learning

Data visualization is the presentation of data in a pictorial or graphical format.

The same data can be represented in many forms and some can be more explanatory than others

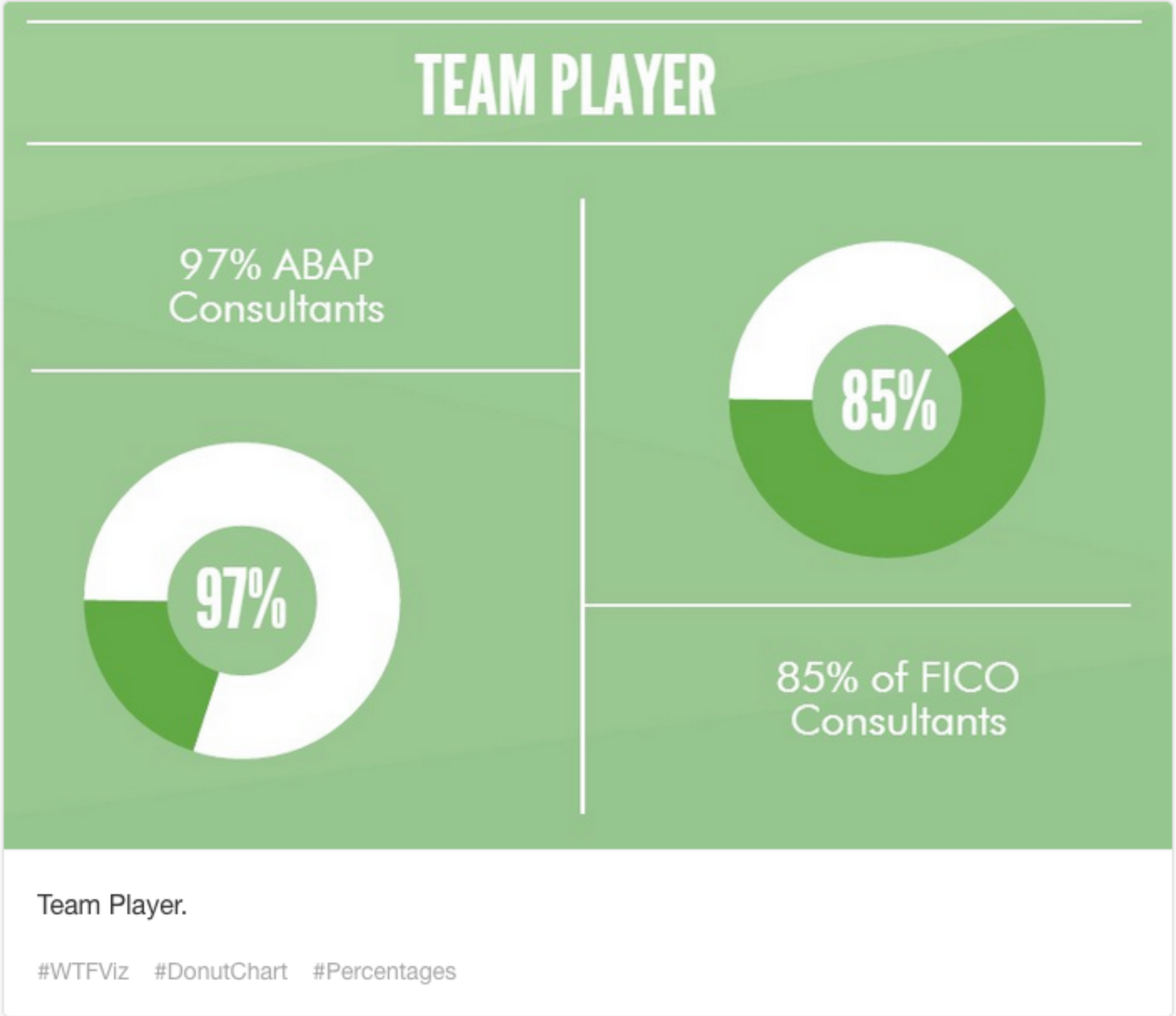
Clarity and accuracy are key



WTF Visualizations

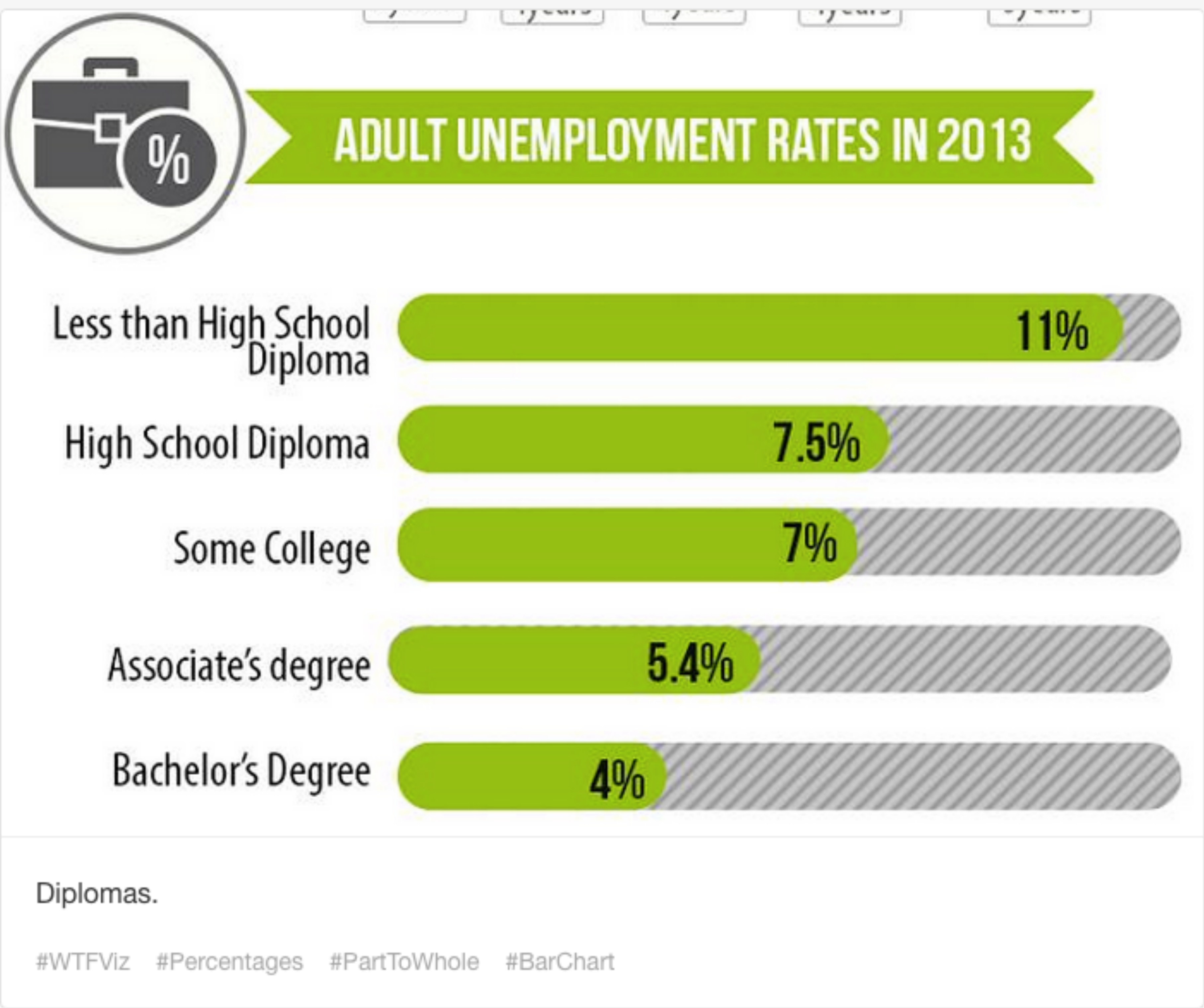
Visualizations that make no sense.
For a discussion of what is wrong with a particular visualization, tweet at us @WTFViz.
Check out our friends Thumbs Up Viz and accidental aRt, or submit.

VISUALIZATION



source: http://wtfviz.net/

VISUALIZATION



source: http://wtfviz.net/

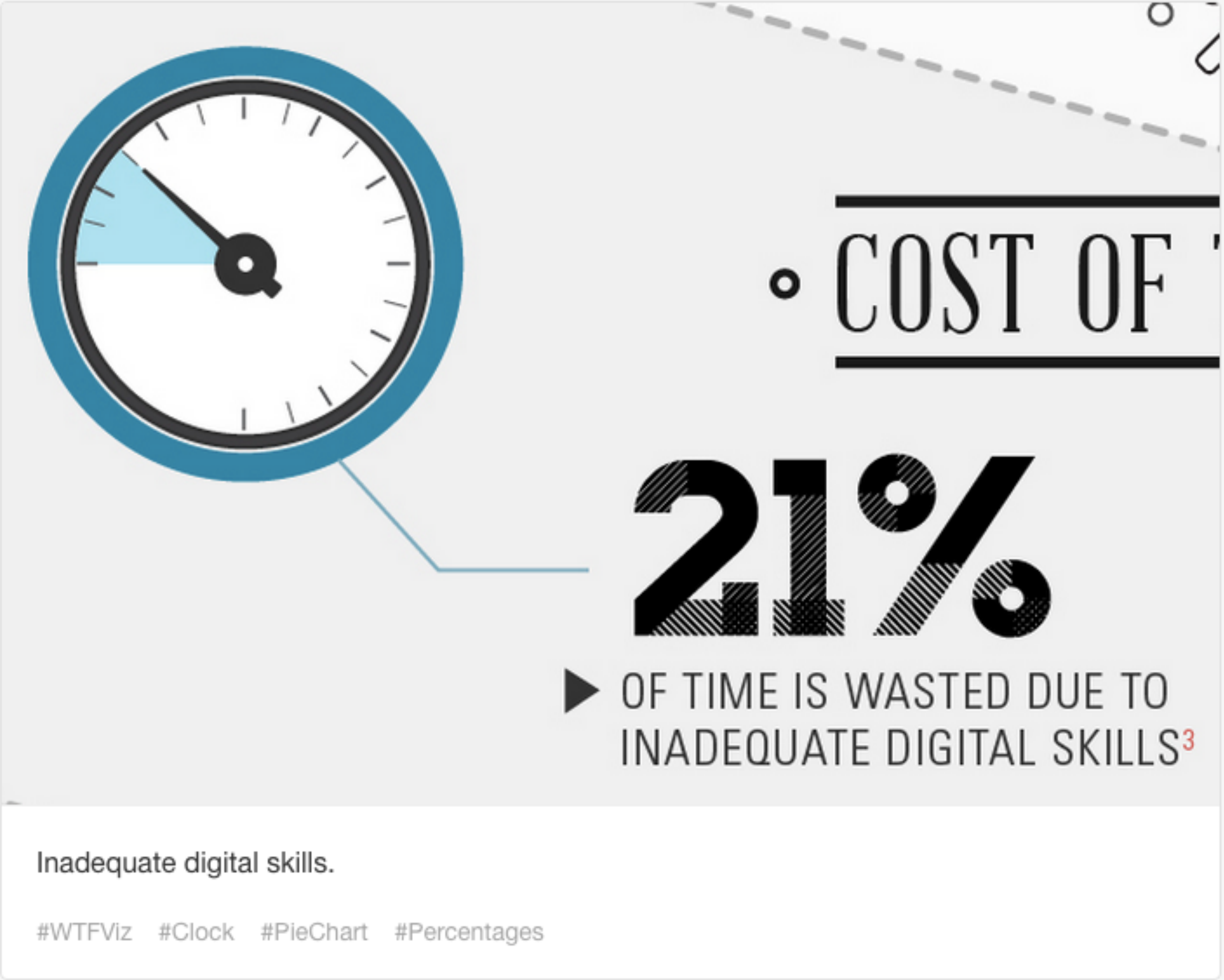
VISUALIZATION

Core skills.

#WTFViz #Radial #BarChart #LabelLines

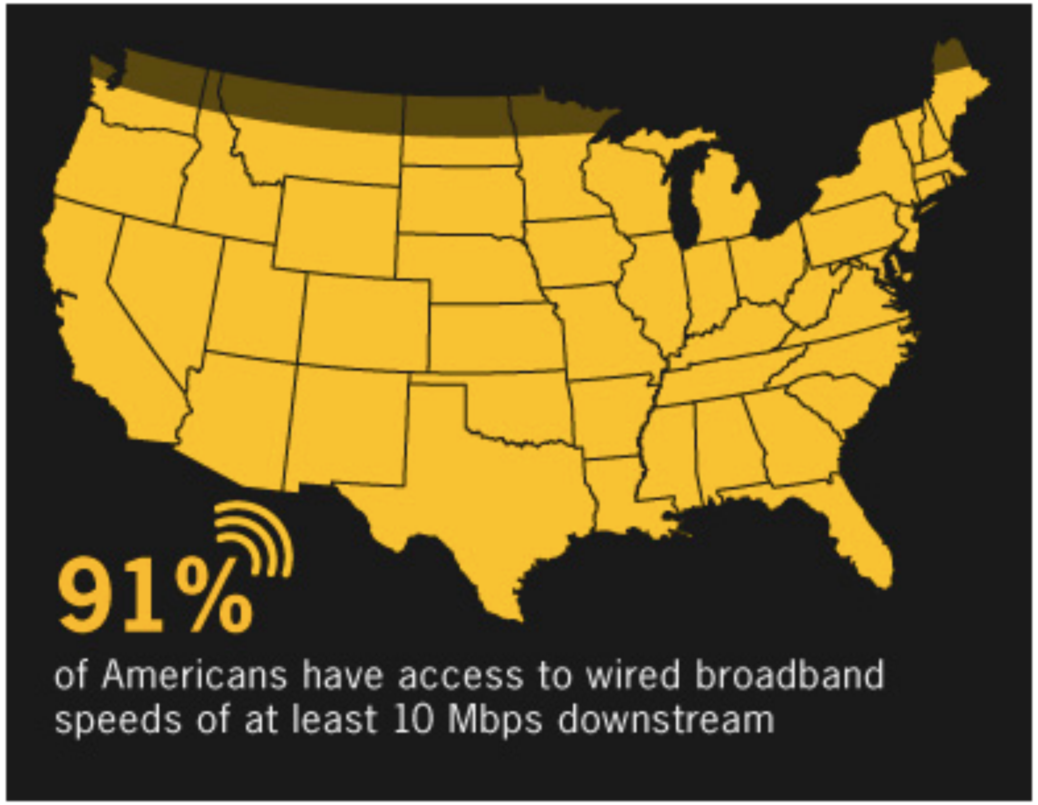
source: http://wtfviz.net/

VISUALIZATION



source: http://wtfviz.net/

VISUALIZATION



Northern regions.

#WTFViz #Map #Percentages

source: http://wtfviz.net/

Fundamental things:

1) choose the appropriate kind of graph
2) choose the right scale
3) label axes
4) use legends (when appropriate)

GALLERIES AND TOOLS

http://www.creativebloq.com/design-tools/data-visualization-712402
https://github.com/mikedewar/d3py
http://bokeh.pydata.org/en/latest/docs/gallery.html
https://github.com/mbostock/d3/wiki/Gallery

BOKEH LAB