GENERAL ASSEMBLY

DATA SCIENCE

CLEANING AND IMPUTING DATA

LAST TIME:

I. PYTHON REVIEW

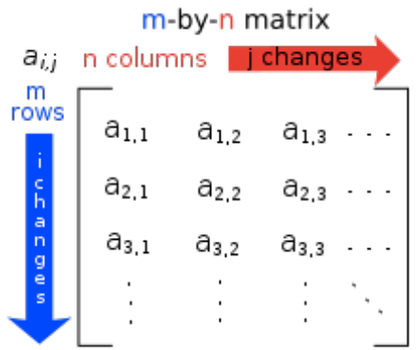
II. LINEAR ALGEBRA REVIEW

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. CLEANING DATA
II. MISSING DATA
III. VISUALIZATIONS

EXERCISES:
IV. NUMPY & PANDAS
V. BOKEH & MATPLOTLIB

- LEARN STRATEGIES TO CLEAN DATA
- LEARN STRATEGIES TO IMPUTE DATA
- LEARN TO VISUALIZE THE DATA

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
} 80%

Model Algorithms
iNterpret Results
} 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

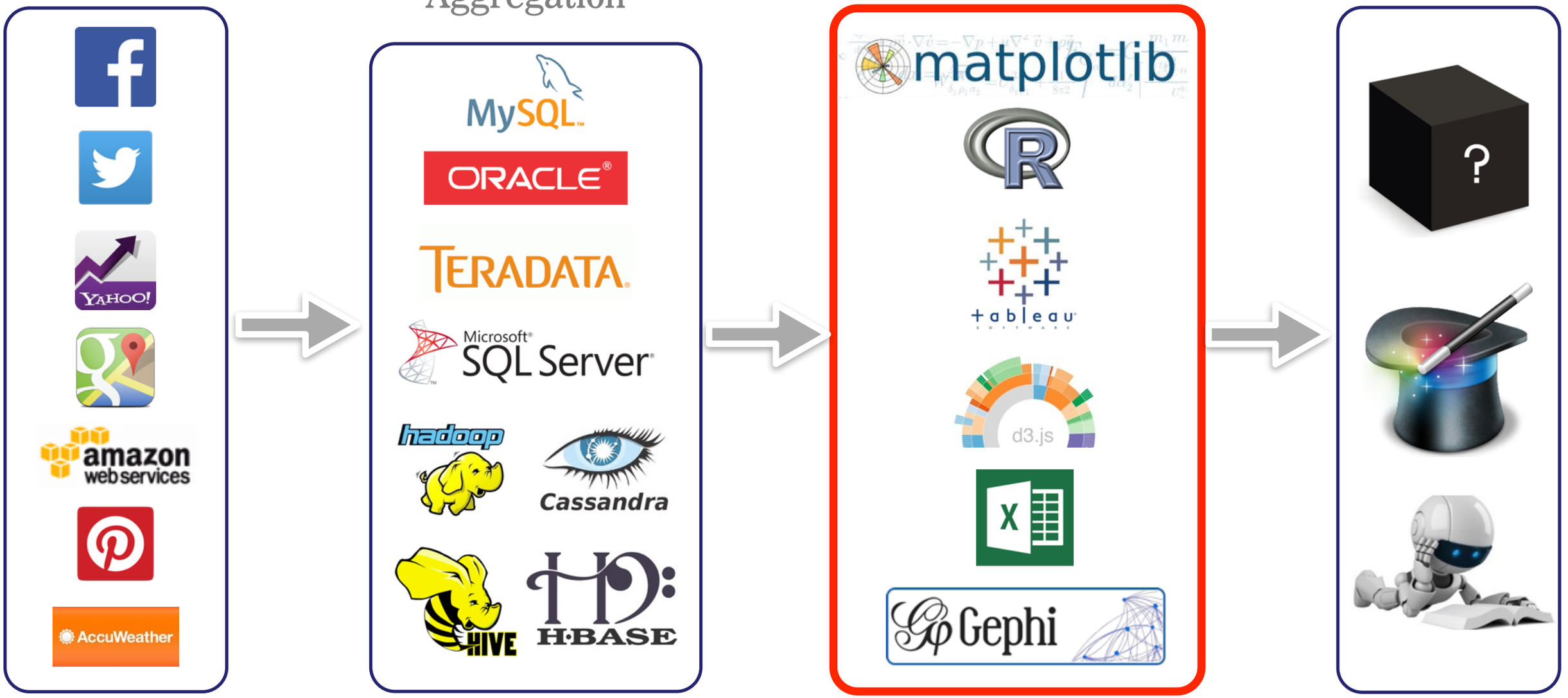
VISUALIZATION

Data Retrieval

Data ETL and Aggregation

Data Visualization

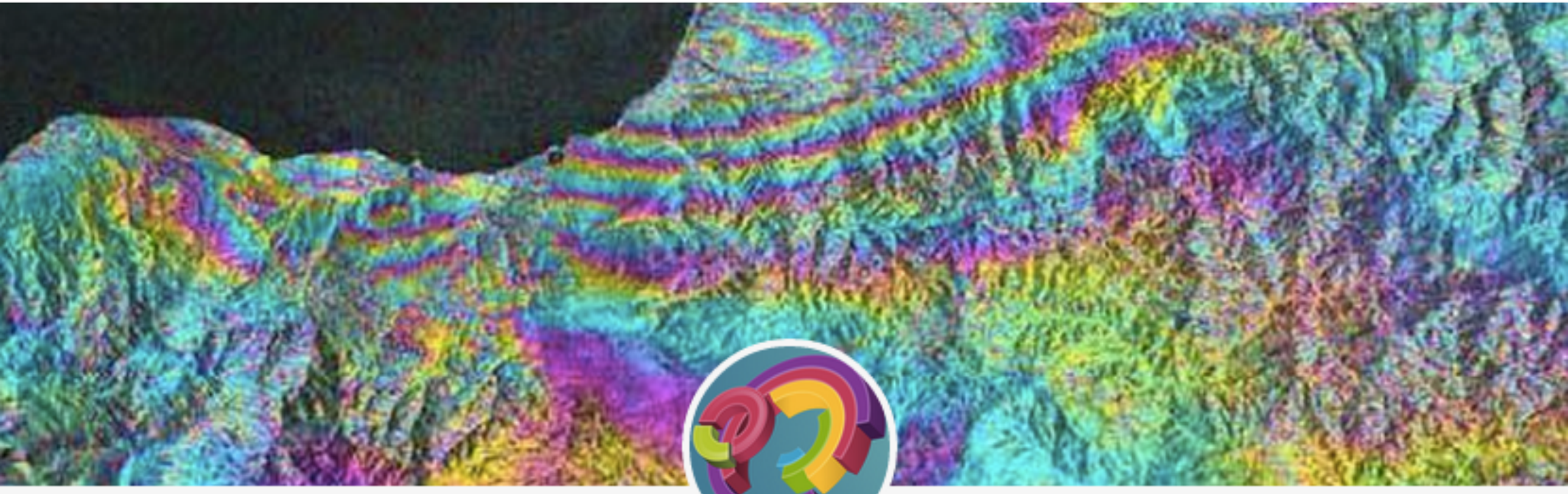
Machine Learning



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



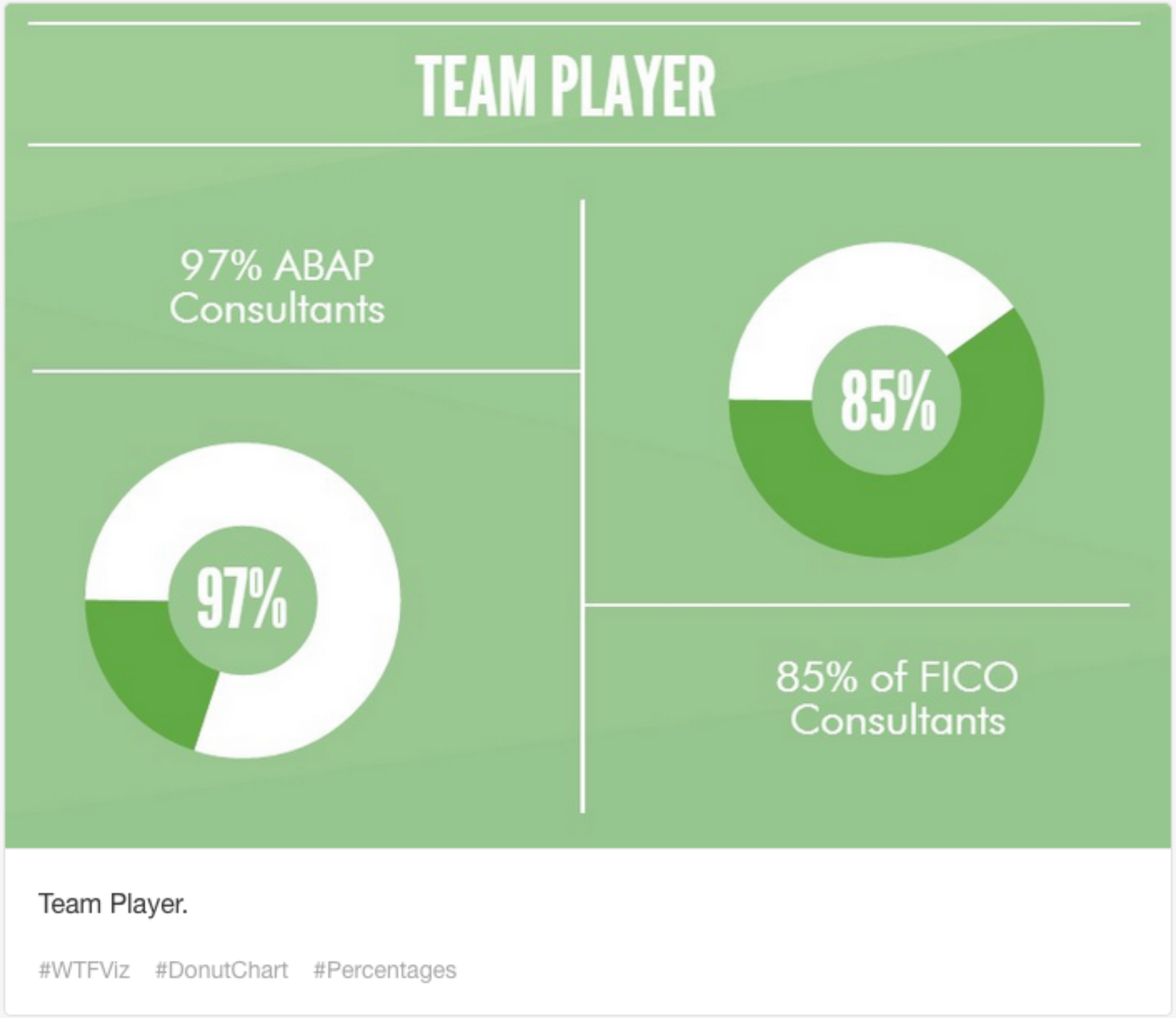
WTF Visualizations

Visualizations that make no sense.

For a discussion of what is wrong with a particular visualization, tweet at us @WTFViz.

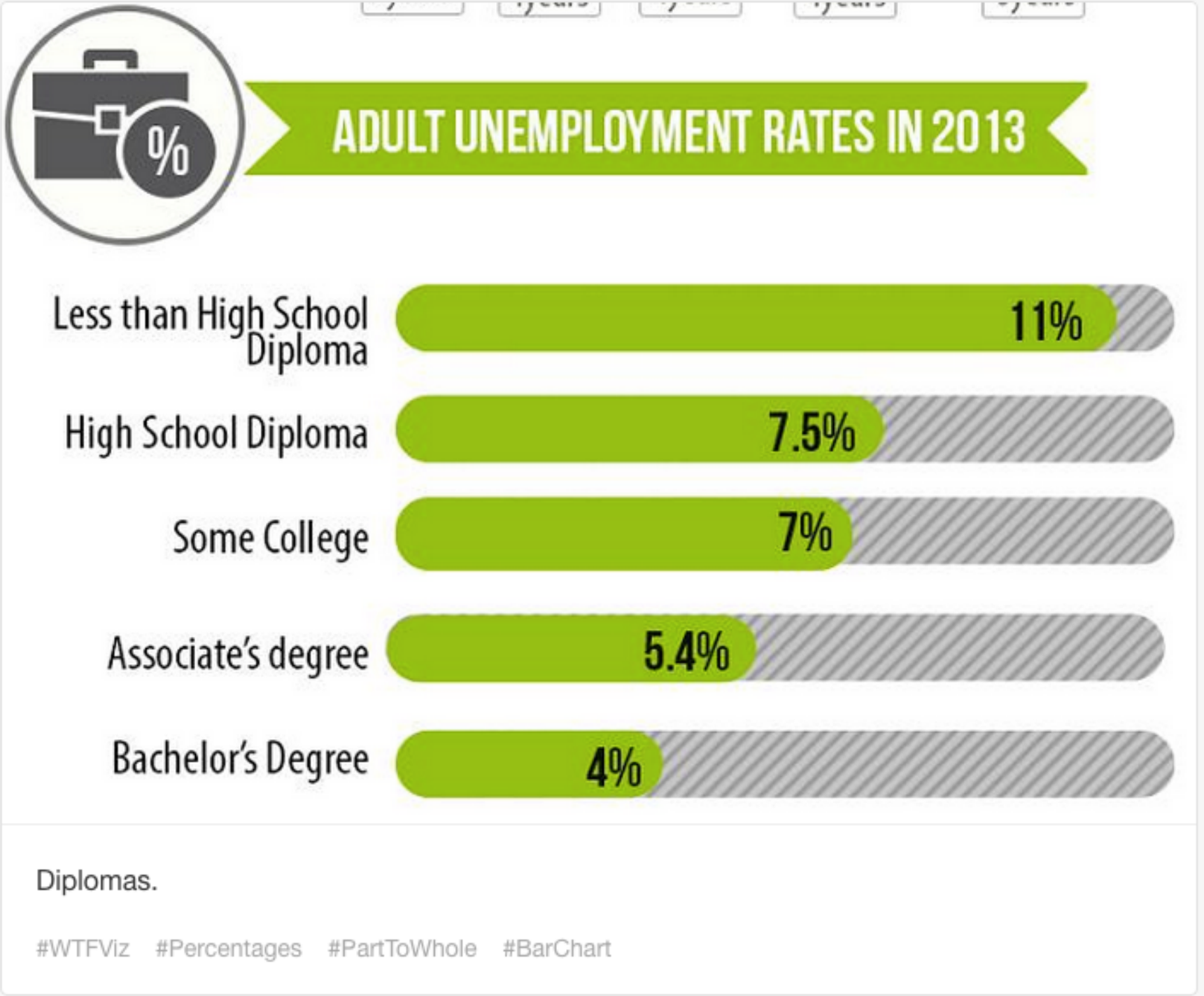
Check out our friends Thumbs Up Viz and accidental aRt, or submit.

VISUALIZATION



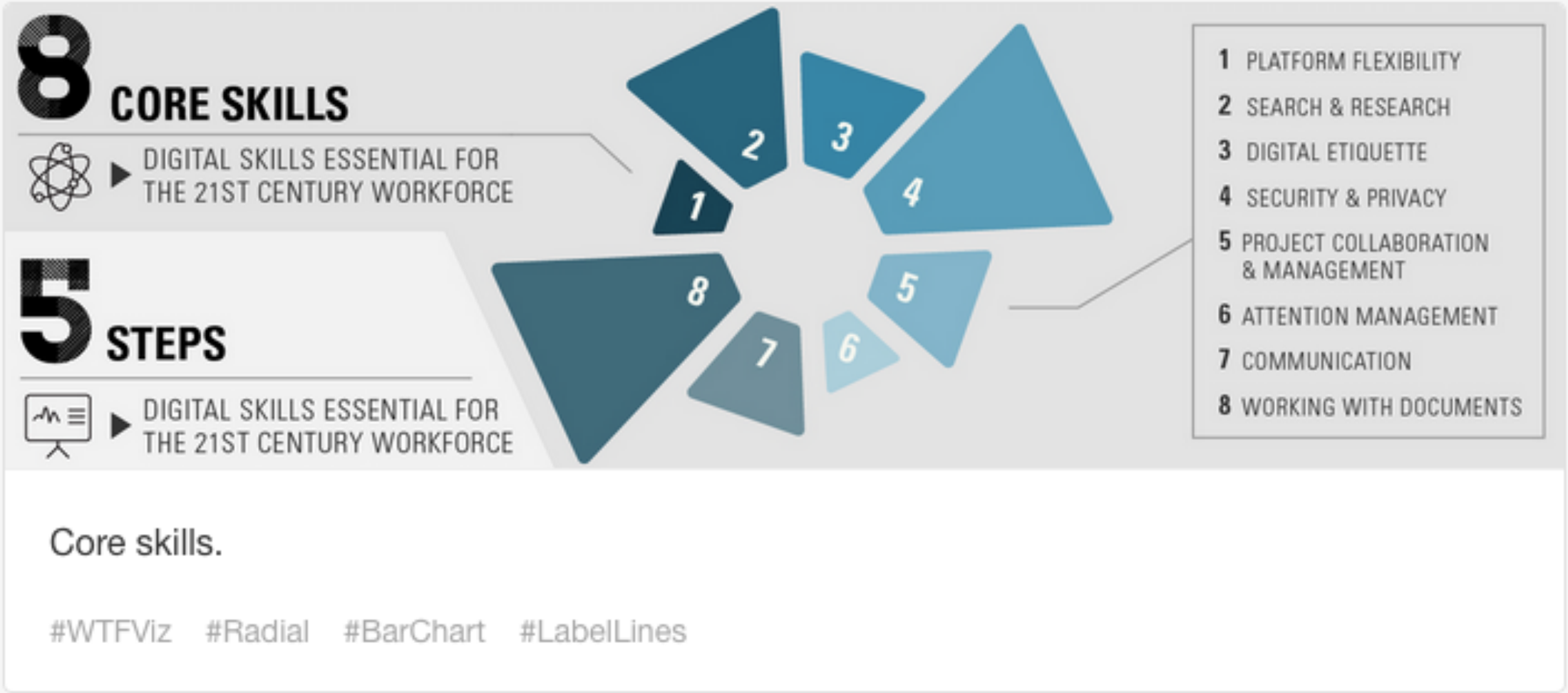
source: http://wtfviz.net/

VISUALIZATION



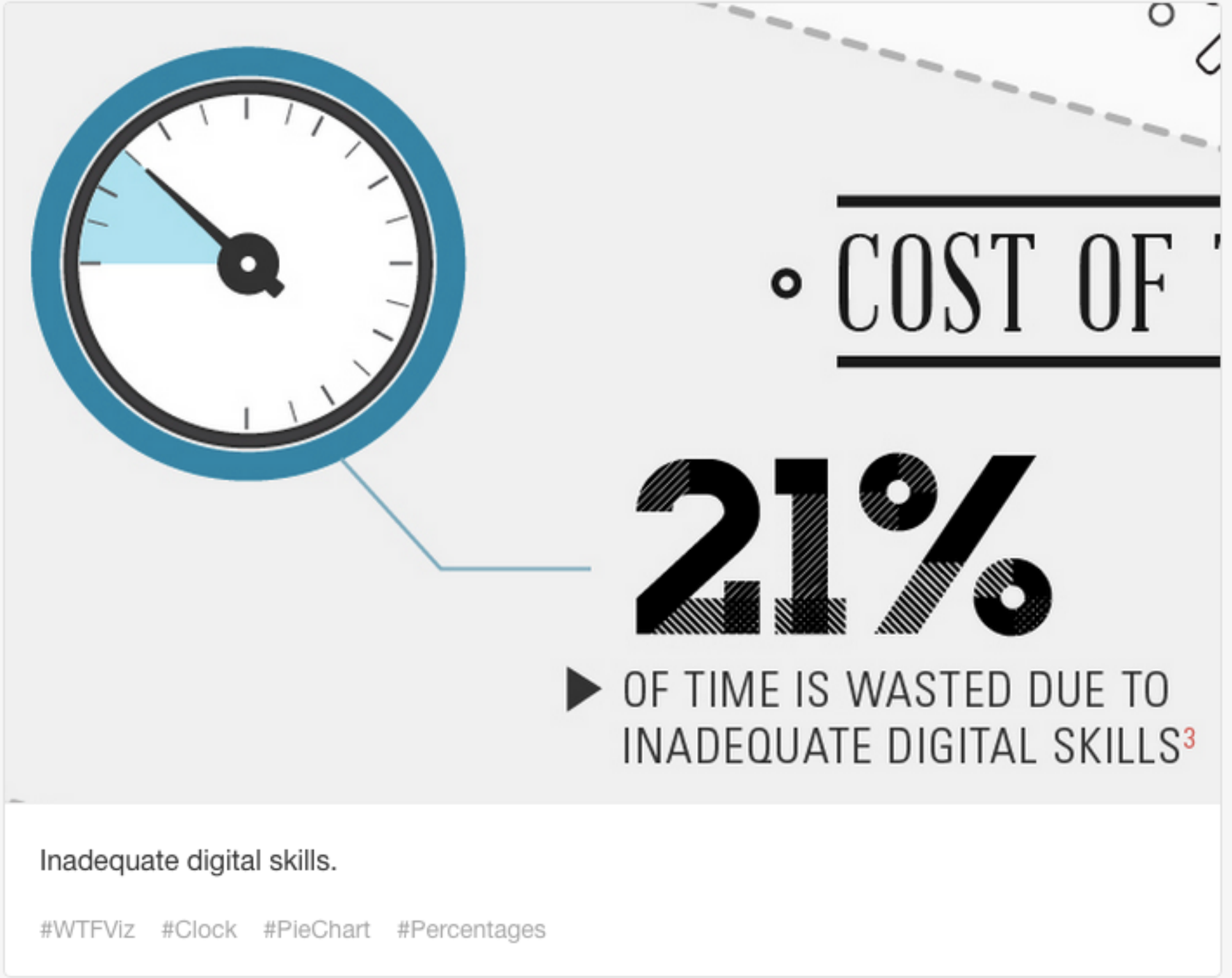
source: http://wtfviz.net/

VISUALIZATION



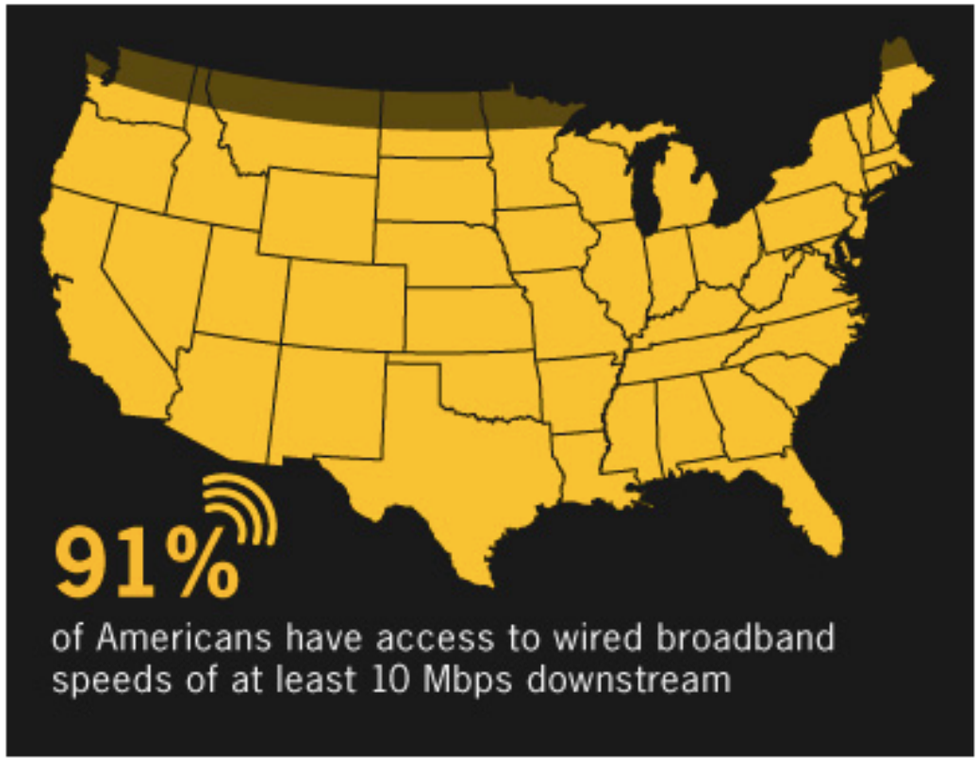
source: http://wtfviz.net/

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