

DATA SCIENCE

CLEANING AND IMPUTING DATA

LAST TIME:

I. PYTHON REVIEW

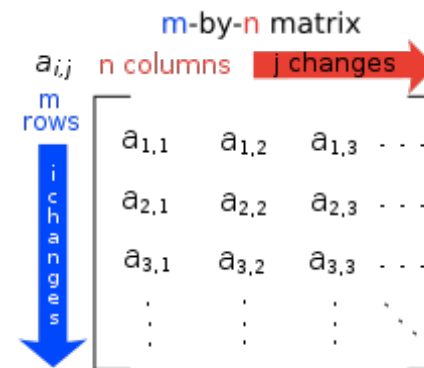
II. LINEAR ALGEBRA REVIEW

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

EXERCISES:

III. NUMPY

IV. PANDAS

V. BOKEH & MATPLOTLIB

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

INTRO TO DATA SCIENCE

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”

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”

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.”



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.

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 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

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

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)

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

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

INTRO TO DATA SCIENCE

MISSING DATA

CLEANING 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

CLEANING 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)

CLEANING 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)

CLEANING 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)

CLEANING 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.

CLEANING DATA

MISSING AT RANDOM (MAR)

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

CLEANING 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

CLEANING DATA

TECHNIQUES TO DEAL WITH MISSING DATA

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

CLEANING 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

CLEANING DATA

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>

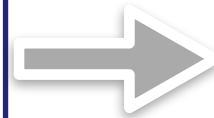
INTRO TO DATA SCIENCE

VISUALIZATION

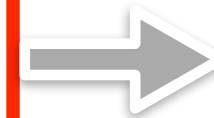
Data Retrieval



Data ETL and Aggregation



Data Visualization



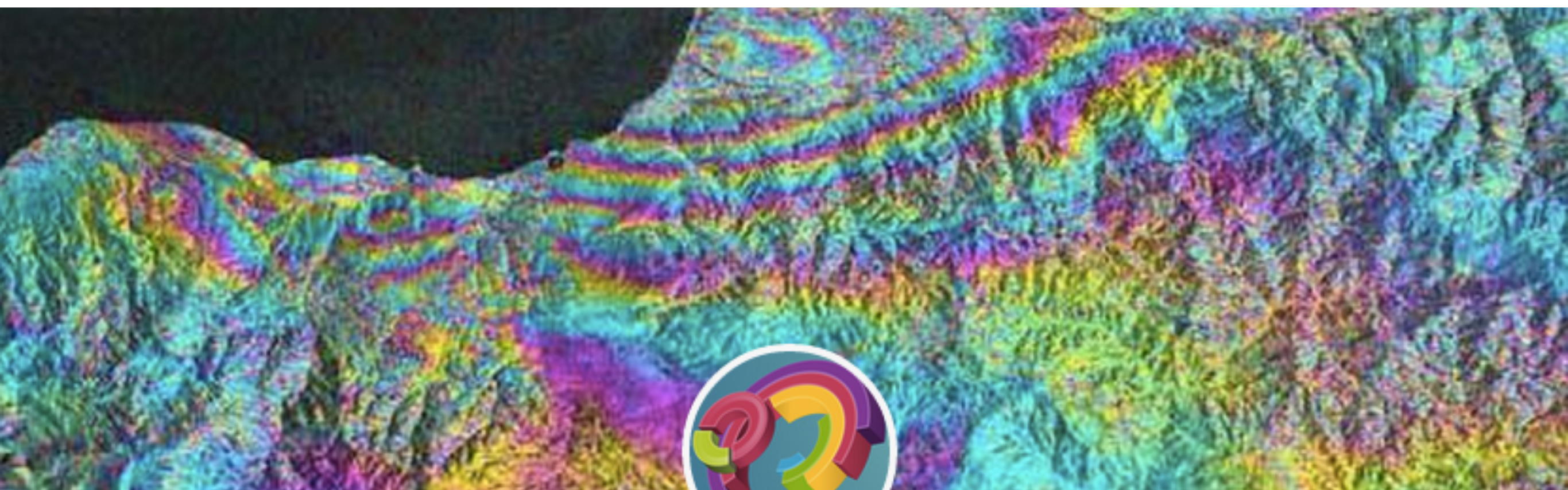
Machine Learning



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.

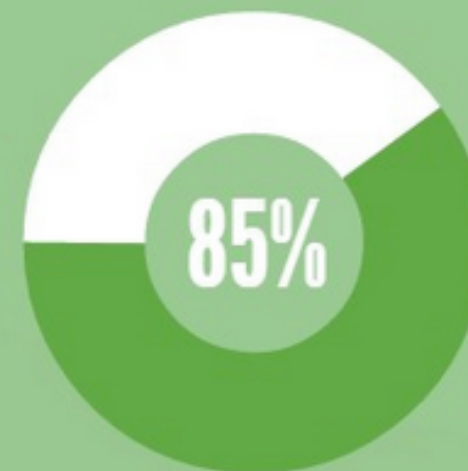
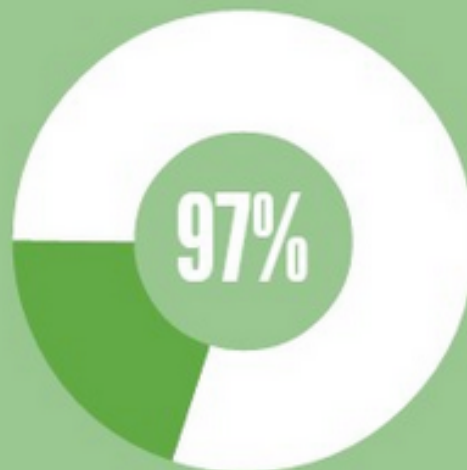
For a discussion of what is wrong with a particular visualization, tweet at us [@WTFViz](https://twitter.com/WTFViz).

Check out our friends [Thumbs Up Viz](#) and [accidental aRt](#), or [submit](#).

VISUALIZATION

TEAM PLAYER

97% ABAP
Consultants



85% of FICO
Consultants

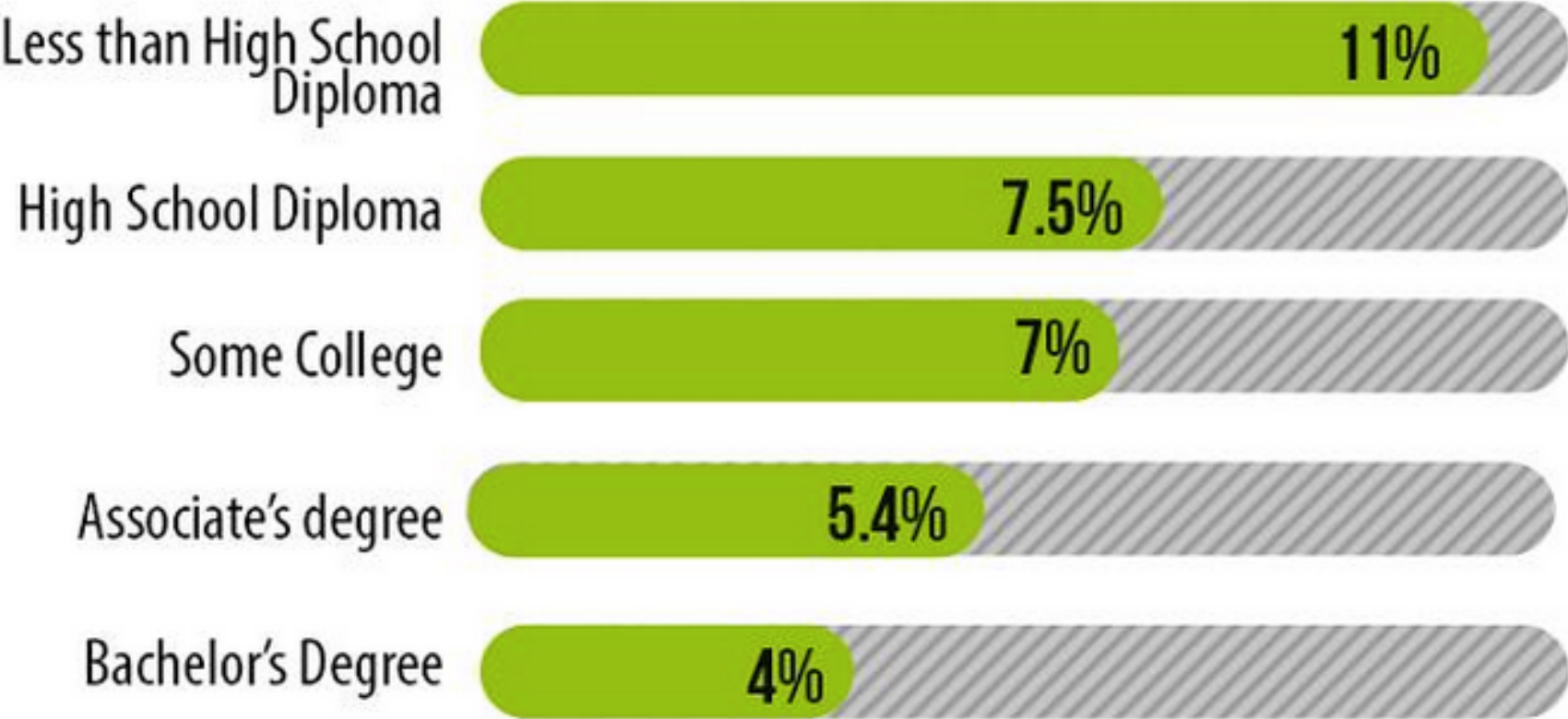
Team Player.

#WTFViz #DonutChart #Percentages

VISUALIZATION



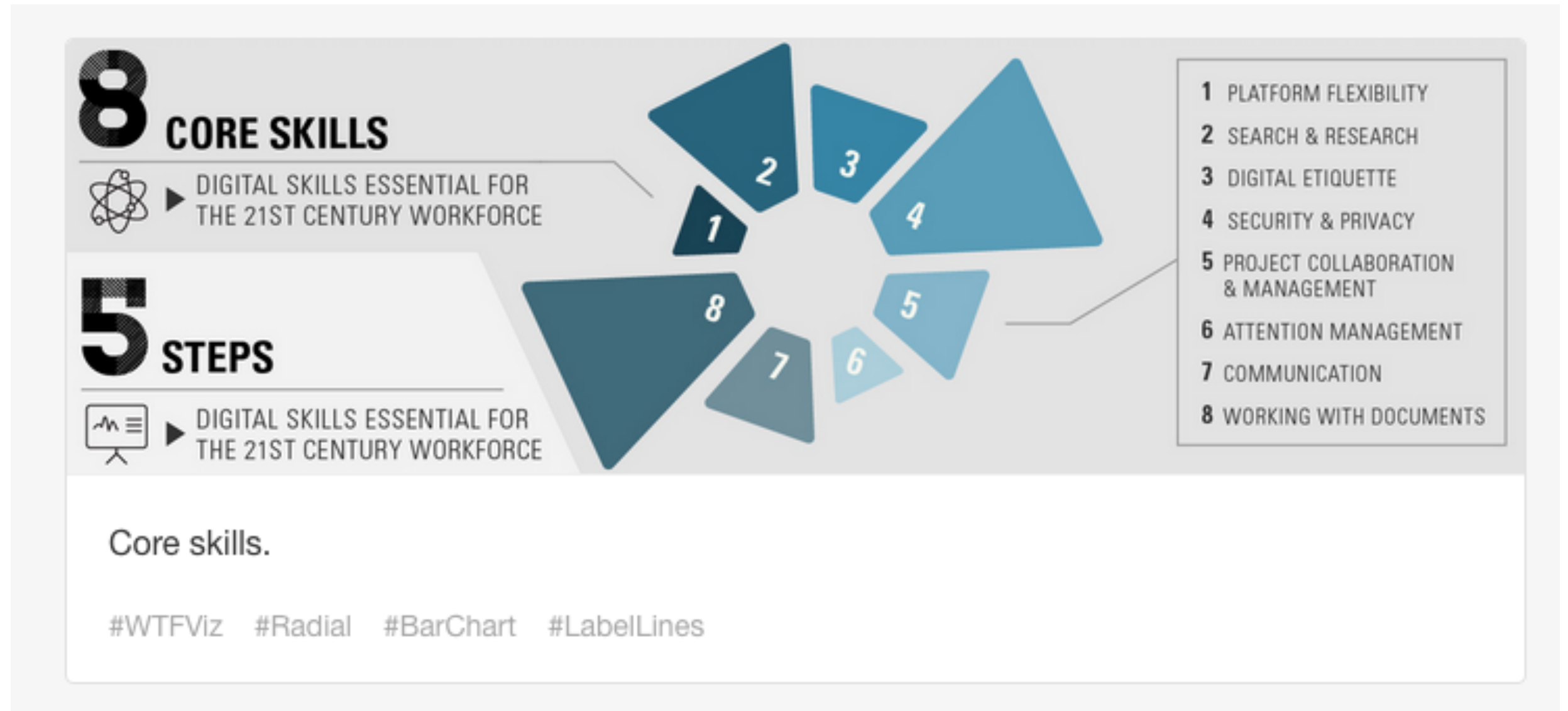
ADULT UNEMPLOYMENT RATES IN 2013



Diplomas.

#WTFViz #Percentages #PartToWhole #BarChart

VISUALIZATION



VISUALIZATION



• COST OF

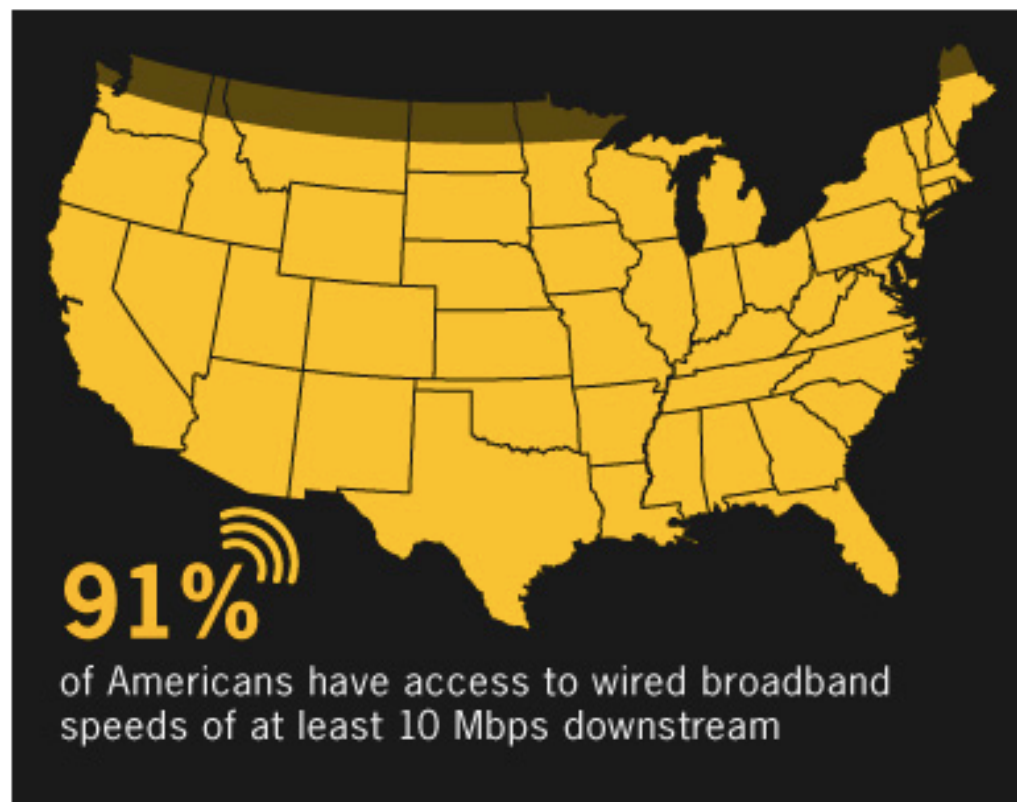
21%

► OF TIME IS WASTED DUE TO
INADEQUATE DIGITAL SKILLS³

Inadequate digital skills.

#WTFViz #Clock #PieChart #Percentages

VISUALIZATION



Northern regions.

#WTFViz #Map #Percentages

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>