

# **DATA SCIENCE**

## **LECTURE 3: CLEANING AND EXPLORING DATA (+ LINEAR ALGEBRA REVIEW)**

**FRANCESCO MOSCONI / ROB HALL / DAT-16**

## LAST TIME:

I. PYTHON QUICK REVIEW

II. DATA SOURCES

III. APIS

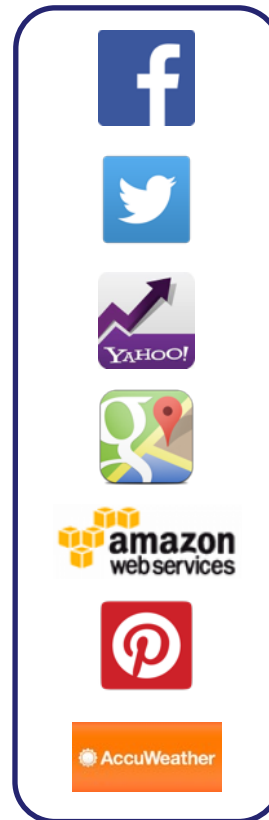
EXERCISES:

IV. PYTHON REVIEW

V. EXTRACTING DATA FROM API

## QUESTIONS?

Data Retrieval



Data ETL and Aggregation



kimono

Turn websites into structured APIs from your browser in seconds



Get started, click to install

# **QUESTIONS?**

**WHAT WAS THE MOST INTERESTING THING YOU LEARNT?**

**WHAT WAS THE HARDEST TO GRASP?**

**I. LINEAR ALGEBRA REVIEW**

**II. DATA CLEANING**

**III. DATA VISUALIZATION**

**EXERCISES:**

**IV. NUMPY**

**V. PANDAS**

**VI. BOKEH**

---

**INTRO TO DATA SCIENCE**

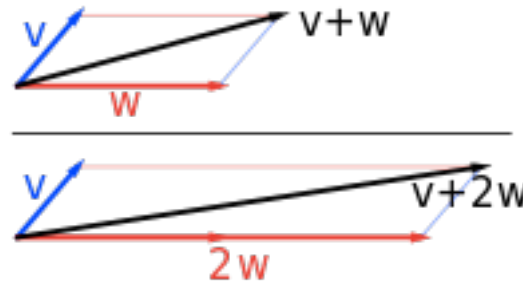
---

# **LINEAR ALGEBRA**

Linear algebra is the branch of mathematics concerning **vector spaces** and **linear mappings** between such spaces.

Linear algebra is the branch of mathematics concerning **vector spaces** and **linear mappings** between such spaces.

A **vector space** (also called a linear space) is a collection of objects called vectors, which may be added together and multiplied ("scaled") by numbers, called scalars in this context.



Linear algebra is the branch of mathematics concerning **vector spaces** and **linear mappings** between such spaces.

A **vector space** (also called a linear space) is a collection of objects called vectors, which may be added together and multiplied ("scaled") by numbers, called scalars in this context.

A **linear mapping** is a mapping  $V \rightarrow W$  between spaces that preserves the operations of addition and scalar multiplication.



Two operations are defined in a vector space

Two operations are defined in a vector space

**ADDITION:**

takes any two vectors  **$\mathbf{v}$**  and  **$\mathbf{w}$**  and outputs a third vector  **$\mathbf{v} + \mathbf{w}$**

$$\mathbf{z} = \mathbf{v} + \mathbf{w}$$

Two operations are defined in a vector space

**ADDITION:**

takes any two vectors  **$\mathbf{v}$**  and  **$\mathbf{w}$**  and outputs a third vector  **$\mathbf{v} + \mathbf{w}$**

$$\mathbf{z} = \mathbf{v} + \mathbf{w}$$

**SCALAR MULTIPLICATION:**

takes any scalar  $a$  and any vector  **$\mathbf{v}$**  and outputs a new vector  **$a\mathbf{v}$**

$$\mathbf{z} = a\mathbf{v}$$

---

**PROPERTY****MEANING**

---

**Associativity** of addition

$$u + (v + w) = (u + v) + w$$

*u, v and w are vectors in V, and a and b are scalars in F.*

PROPERTY

MEANING

**Associativity** of addition

$$u + (v + w) = (u + v) + w$$

**Commutativity** of addition

$$u + v = v + u$$

*$u, v$  and  $w$  are vectors in  $V$ , and  $a$  and  $b$  are scalars in  $F$ .*

PROPERTY	MEANING
Associativity of addition	$u + (v + w) = (u + v) + w$
Commutativity of addition	$u + v = v + u$
<b>Identity element of addition</b>	There exists an element $0 \in V$ , called the <b>zero vector</b> , such that $v + 0 = v$ for all $v \in V$ .

*$u, v$  and  $w$  are vectors in  $V$ , and  $a$  and  $b$  are scalars in  $F$ .*

PROPERTY	MEANING
Associativity of addition	$u + (v + w) = (u + v) + w$
Commutativity of addition	$u + v = v + u$
Identity element of addition	There exists an element $0 \in V$ , called the <b>zero vector</b> , such that $v + 0 = v$ for all $v \in V$ .
<b>Inverse elements of addition</b>	For every $v \in V$ , there exists an element $-v \in V$ , called the additive <b>inverse</b> of $v$ , such that $v + (-v) = 0$

*$u, v$  and  $w$  are vectors in  $V$ , and  $a$  and  $b$  are scalars in  $F$ .*

PROPERTY	MEANING
Associativity of addition	$u + (v + w) = (u + v) + w$
Commutativity of addition	$u + v = v + u$
Identity element of addition	There exists an element $0 \in V$ , called the <b>zero vector</b> , such that $v + 0 = v$ for all $v \in V$ .
Inverse elements of addition	For every $v \in V$ , there exists an element $-v \in V$ , called the additive <b>inverse</b> of $v$ , such that $v + (-v) = 0$
<b>Distributivity</b> (scalar and vector)	$a(u + v) = au + av$ $(a + b)v = av + bv$

*$u, v$  and  $w$  are vectors in  $V$ , and  $a$  and  $b$  are scalars in  $F$ .*



PROPERTY	MEANING
Associativity of addition	$u + (v + w) = (u + v) + w$
Commutativity of addition	$u + v = v + u$
Identity element of addition	There exists an element $0 \in V$ , called the <b>zero vector</b> , such that $v + 0 = v$ for all $v \in V$ .
Inverse elements of addition	For every $v \in V$ , there exists an element $-v \in V$ , called the additive <b>inverse</b> of $v$ , such that $v + (-v) = 0$
Distributivity (scalar and vector)	$a(u + v) = au + av$ $(a + b)v = av + bv$
Compatibility of multiplication	$a(bv) = (ab)v$

*$u, v$  and  $w$  are vectors in  $V$ , and  $a$  and  $b$  are scalars in  $F$ .*

PROPERTY	MEANING
Associativity of addition	$u + (v + w) = (u + v) + w$
Commutativity of addition	$u + v = v + u$
Identity element of addition	There exists an element $0 \in V$ , called the <b>zero vector</b> , such that $v + 0 = v$ for all $v \in V$ .
Inverse elements of addition	For every $v \in V$ , there exists an element $-v \in V$ , called the additive <b>inverse</b> of $v$ , such that $v + (-v) = 0$
Distributivity (scalar and vector)	$a(u + v) = au + av$ $(a + b)v = av + bv$
Compatibility of multiplication	$a(bv) = (ab)v$
Identity element of scalar multiplication	$1v = v$

*$u, v$  and  $w$  are vectors in  $V$ , and  $a$  and  $b$  are scalars in  $F$ .*

A linear transformation  $T$  between two vector spaces  $V$  and  $W$  is compatible with scalar multiplication and vector addition:

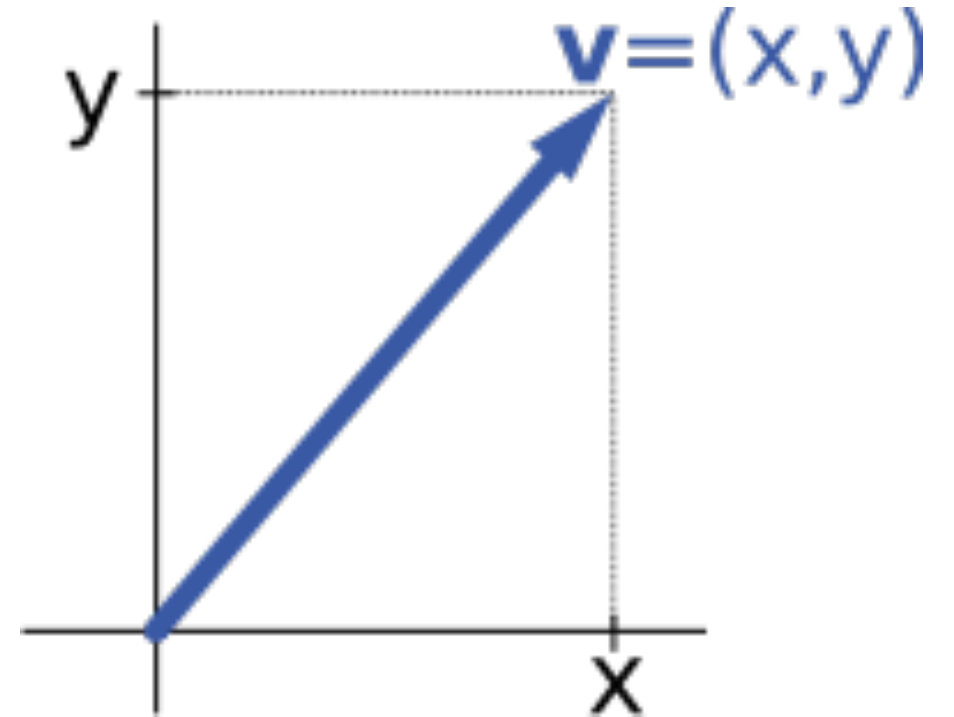
$$T: V \rightarrow W$$

satisfies:

$$T(a \mathbf{u} + b \mathbf{v}) = a T(\mathbf{u}) + b T(\mathbf{v})$$

Vectors can be represented by a list of numbers, their coordinates:

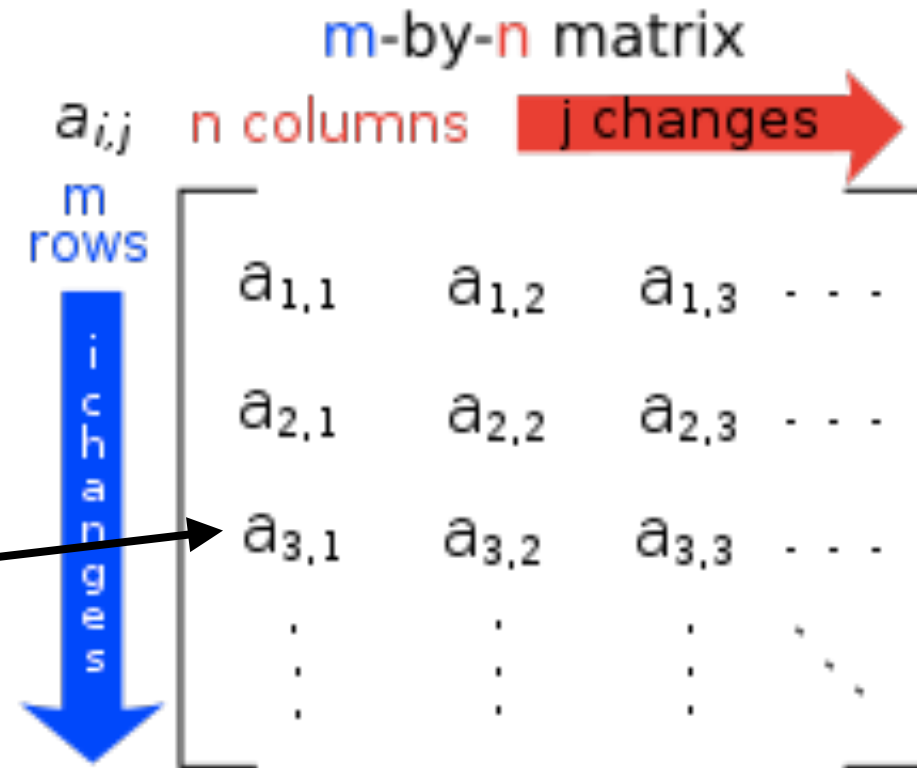
(give me some examples of vector quantities)



Linear mappings can be represented by **matrices**

Matrices are an array of real numbers with  $m$  rows and  $n$  columns

Each value in a matrix is called an entry.



The size of a matrix is defined by the number of rows and columns.

Examples:

Name	Size	Example
Row vector	$1 \times n$	$\begin{bmatrix} 3 & 7 & 2 \end{bmatrix}$
Column vector	$n \times 1$	$\begin{bmatrix} 4 \\ 1 \\ 8 \end{bmatrix}$
Square matrix	$n \times n$	$\begin{bmatrix} 9 & 13 & 5 \\ 1 & 11 & 7 \\ 2 & 6 & 3 \end{bmatrix}$

## Rule 1!

Matrices can be added together only when they are the same size. If they are not the same size, their sum is **undefined**.

$$[1 \ 3 \ 9 \ 2] + [2 \ 5 \ 9 \ 4] = [3 \ 8 \ 18 \ 6]$$

**Rule 1!**

Matrices can be added together only when they are the same size. If they are not the same size, their sum is **undefined**.

$$[1 \ 3 \ 9 \ 2] + [2 \ 5 \ 9 \ 4] = [3 \ 8 \ 18 \ 6]$$

$$[8 \ 72 \ 3 \ 1] + [17 \ 55 \ 3 \ 10] = ?$$



## Rule 2!

Matrices can be multiplied by a scalar (single entity) value.

Each value in the matrix is multiplied by the scalar value.

$$[1 \ 3 \ 9 \ 2] * 3 = [3 \ 9 \ 27 \ 6]$$

$$[8 \ 72 \ 3 \ 1] * 2 = ?$$

## Rule 3!

Matrices and vectors can be multiplied together given that the matrix columns are as wide as the vector is long.

What shape will the result take?

$$\begin{bmatrix} 1 & 3 & 9 & 2 \\ 2 & 4 & 6 & 8 \end{bmatrix} * \begin{bmatrix} 2 \\ 3 \\ 6 \\ 5 \end{bmatrix} = ?$$

$2 \times 4$                    $4 \times 1$

## Rule 3!

Matrices and vectors can be multiplied together given that the matrix columns are as wide as the vector is long.

**The result will always be a vector.**

$$\begin{array}{ccc} \begin{bmatrix} 1 & 3 & 9 & 2 \\ 2 & 4 & 6 & 8 \end{bmatrix} & * & \begin{bmatrix} 2 \\ 3 \\ 6 \\ 5 \end{bmatrix} = \begin{array}{l} (2 + 9 + 54 + 10) \\ (4 + 12 + 36 + 40) \end{array} = \begin{bmatrix} 75 \\ 92 \end{bmatrix} \\ 2 \times 4 & & 4 \times 1 \qquad \qquad \qquad 2 \times 1 \end{array}$$

## Rule 4!

Matrices can be multiplied together using the same rules that we have from matrix–vector multiplication.

**What shape will the result take?**

$$\begin{bmatrix} 1 & 3 & 9 & 2 \\ 2 & 4 & 6 & 8 \end{bmatrix} * \begin{bmatrix} 2 & 1 \\ 3 & 2 \\ 5 & 0 \\ 6 & 4 \end{bmatrix} = ?$$

## Rule 4!

Matrices can be multiplied together using the same rules that we have from matrix–vector multiplication.

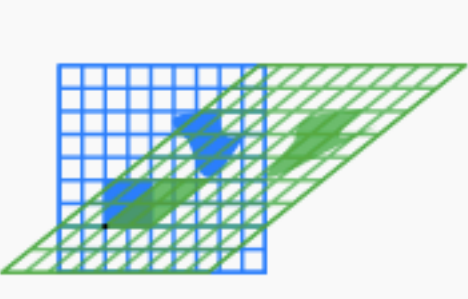
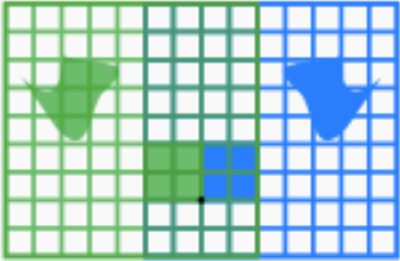
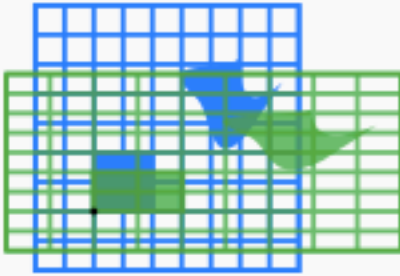

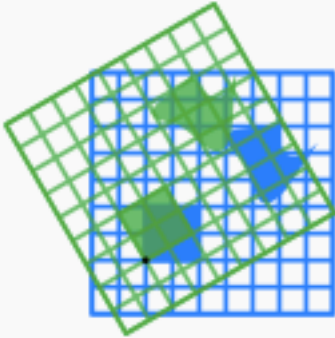
**The result will always be a matrix.**

$$\begin{bmatrix} 1 & 3 & 9 & 2 \\ 2 & 4 & 6 & 8 \end{bmatrix} * \begin{bmatrix} 2 & 1 \\ 3 & 2 \\ 5 & 0 \\ 6 & 4 \end{bmatrix} = \begin{bmatrix} (2+9+54+10) & (1+6+0+8) \\ (4+12+36+40) & (2+8+0+32) \end{bmatrix}$$

$\qquad\qquad\qquad = 75 \qquad\qquad\qquad = 15$   
 $\qquad\qquad\qquad = 92 \qquad\qquad\qquad = 42$

Here are some examples of operations in a 2D vector space with the corresponding matrix.

Each point in this space is represented by the vector of its coordinates  $P = (x, y)$

Horizontal shear with $m=1.25$ .	Horizontal flip	Squeeze mapping with $r=3/2$	Scaling by a factor of $3/2$	Rotation by $\pi/6^R = 30^\circ$
$\begin{bmatrix} 1 & 1.25 \\ 0 & 1 \end{bmatrix}$	$\begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix}$	$\begin{bmatrix} 3/2 & 0 \\ 0 & 2/3 \end{bmatrix}$	$\begin{bmatrix} 3/2 & 0 \\ 0 & 3/2 \end{bmatrix}$	$\begin{bmatrix} \cos(\pi/6^R) & -\sin(\pi/6^R) \\ \sin(\pi/6^R) & \cos(\pi/6^R) \end{bmatrix}$
				

---

## MATRICES

---

## LINKS

[https://en.wikipedia.org/wiki/Matrix\\_\(mathematics\)](https://en.wikipedia.org/wiki/Matrix_(mathematics))

<http://mathworld.wolfram.com/Matrix.html>

<http://ed.ted.com/lessons/how-to-organize-add-and-multiply-matrices-bill-shillito>

---

**INTRO TO DATA SCIENCE**

---

# **NUMPY LAB**



---

**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

---

**INTRO TO DATA SCIENCE**

---

# **WALK THE WALK OF CLEANING DATA**

---

## **CLEANING DATA**

---

# **DATA MUNGING TOOLS AND OPERATIONS**

- python, pandas
- sed, awk, bash, perl
- regular expressions
- text editor
- etc. etc. etc.



---

## **CLEANING DATA**

---

### **DATA MUNGING TOOLS AND OPERATIONS**

- python, pandas
- sed, awk, bash, perl
- regular expressions
- text editor
- etc. etc. etc.

### **IN SMALL GROUPS:**

- Choose one tool from the list
- investigate functionality
- find one example
- show use to class

---

## CLEANING DATA

---

### LINKS

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

---

**INTRO TO DATA SCIENCE**

---

# **PANDAS LAB**

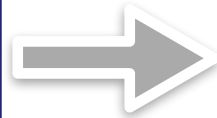
---

**INTRO TO DATA SCIENCE**

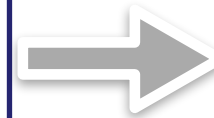
---

**VISUALIZATION**

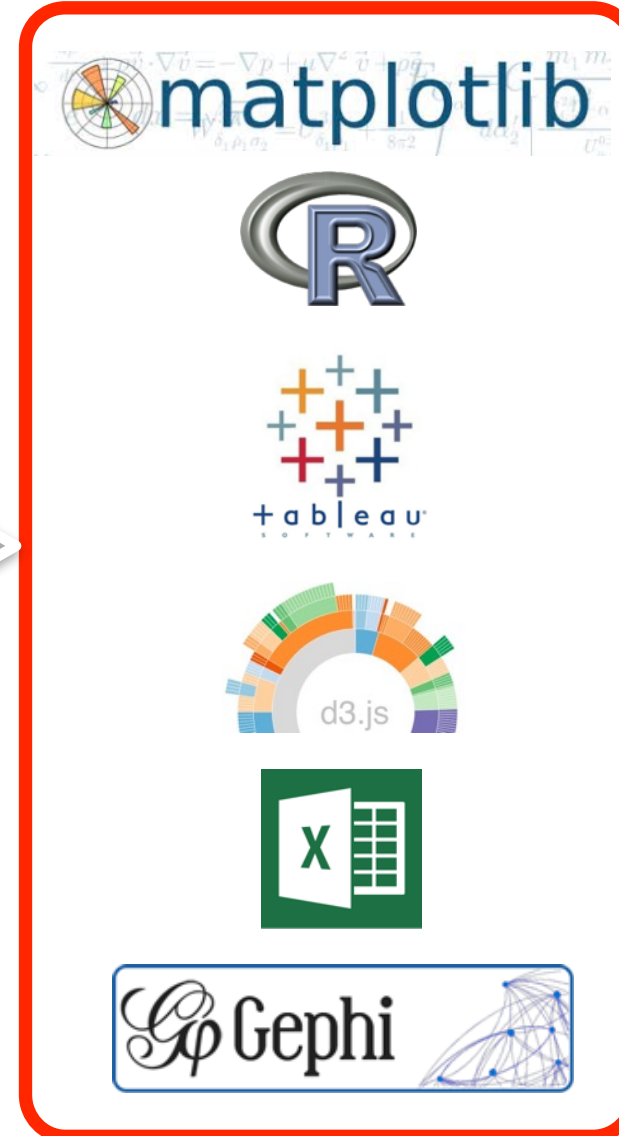
## Data Retrieval



## Data ETL and Aggregation



## Data Visualization



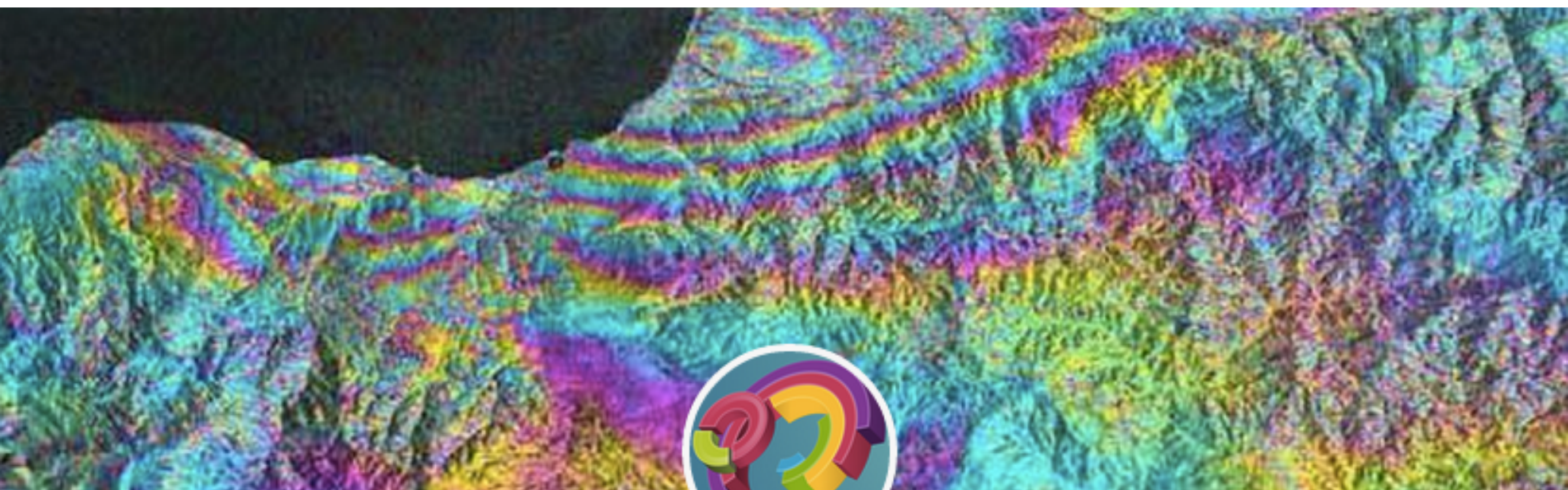
## Machine Learning



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

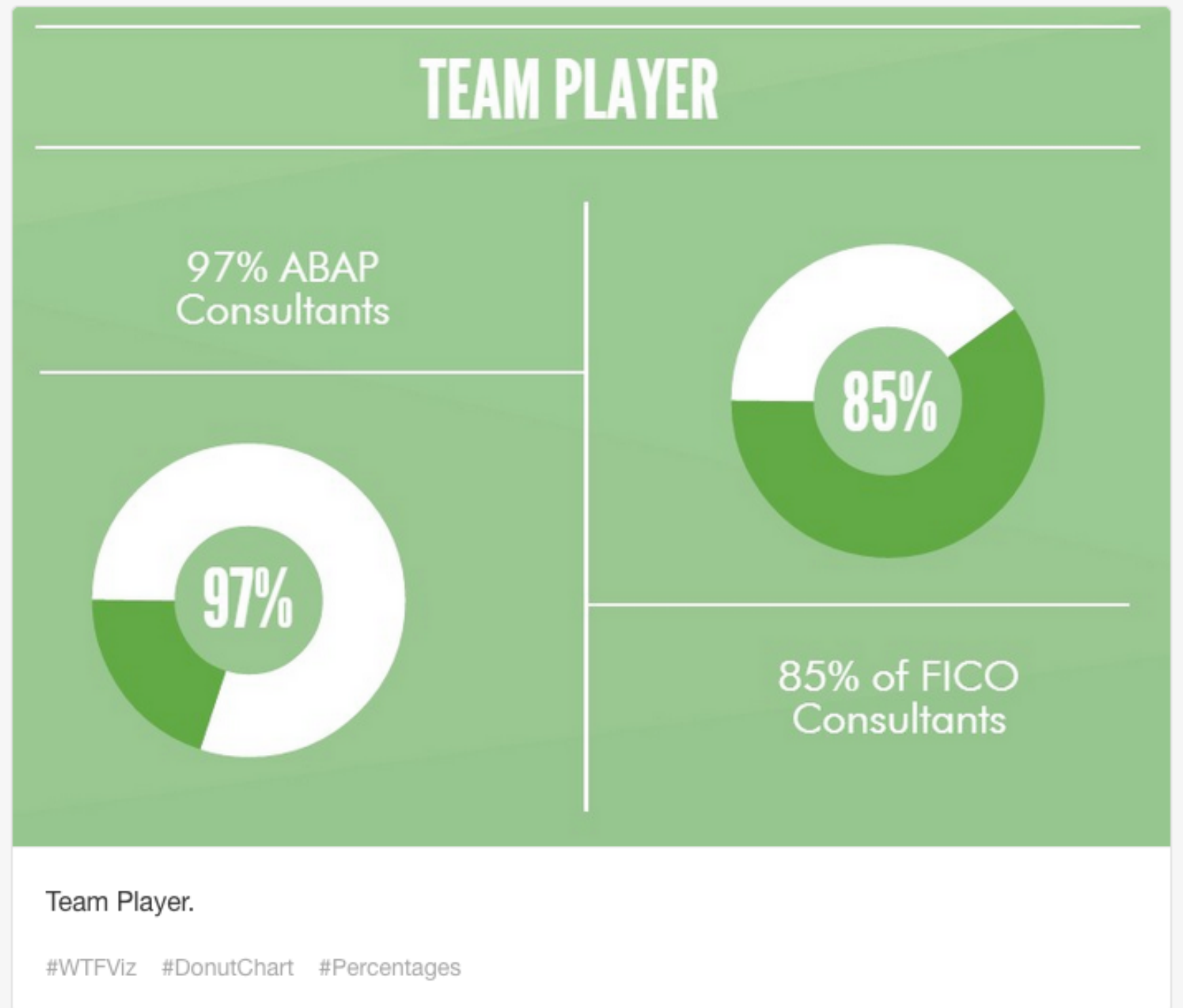
Visualizations that make no sense.

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

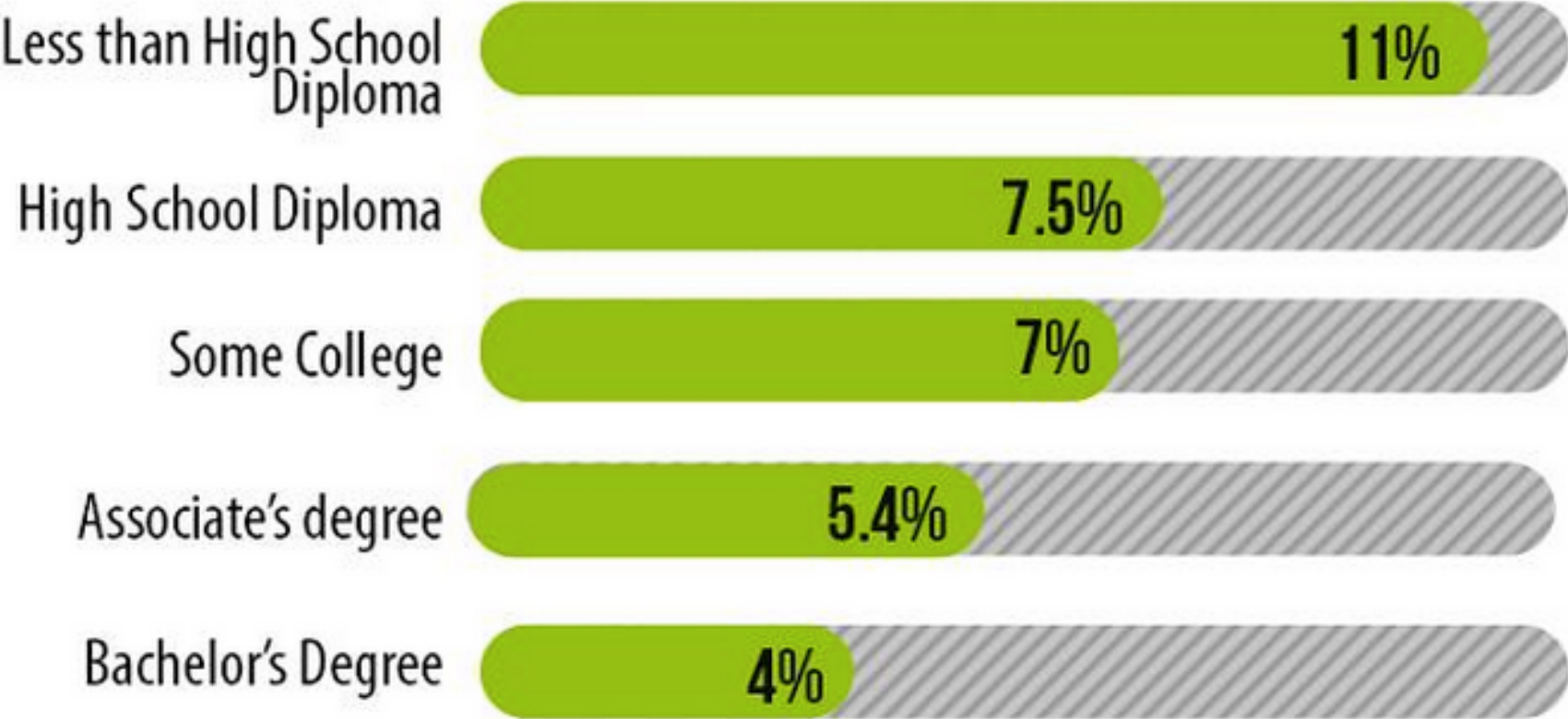




VISUALIZATION



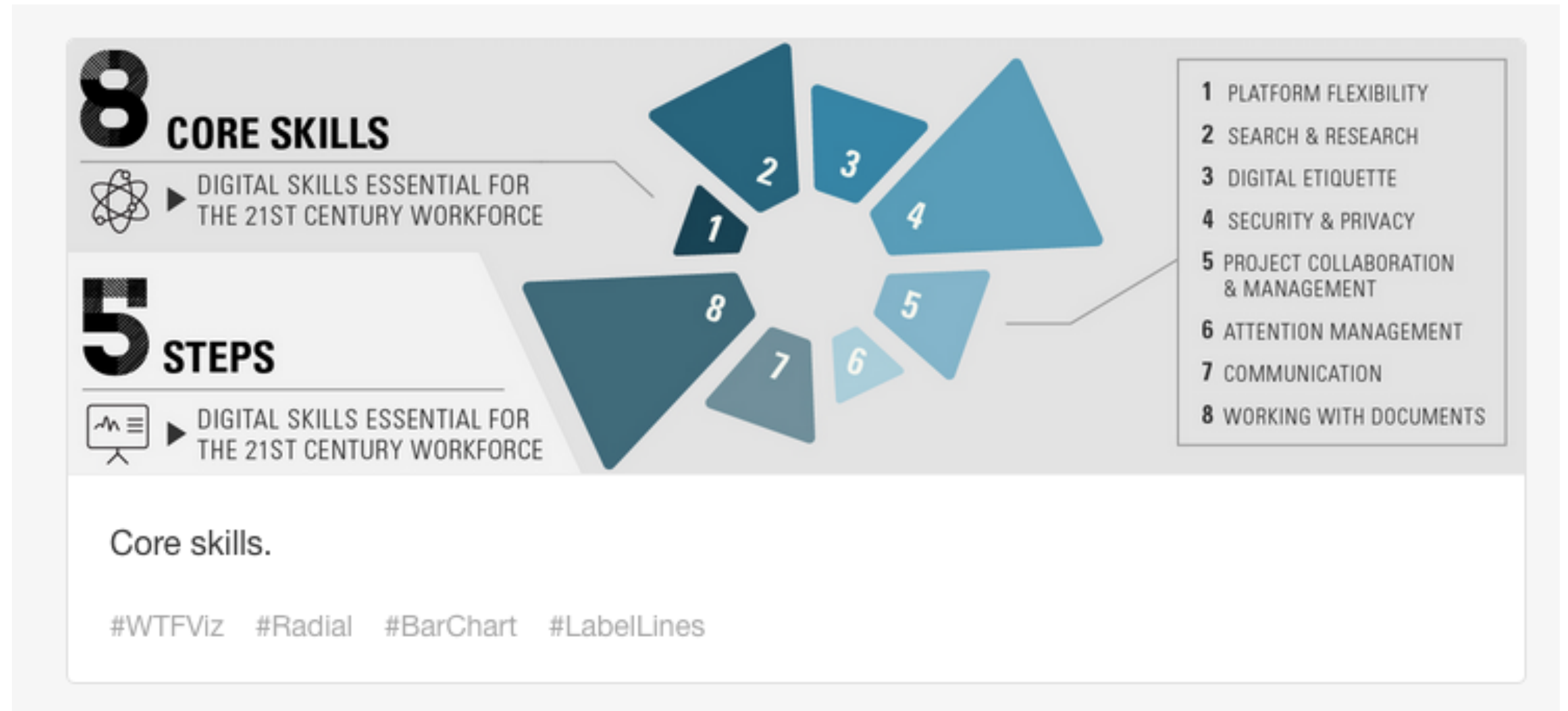
ADULT UNEMPLOYMENT RATES IN 2013



Diplomas.

#WTFViz #Percentages #PartToWhole #BarChart

# VISUALIZATION



## VISUALIZATION



• COST OF

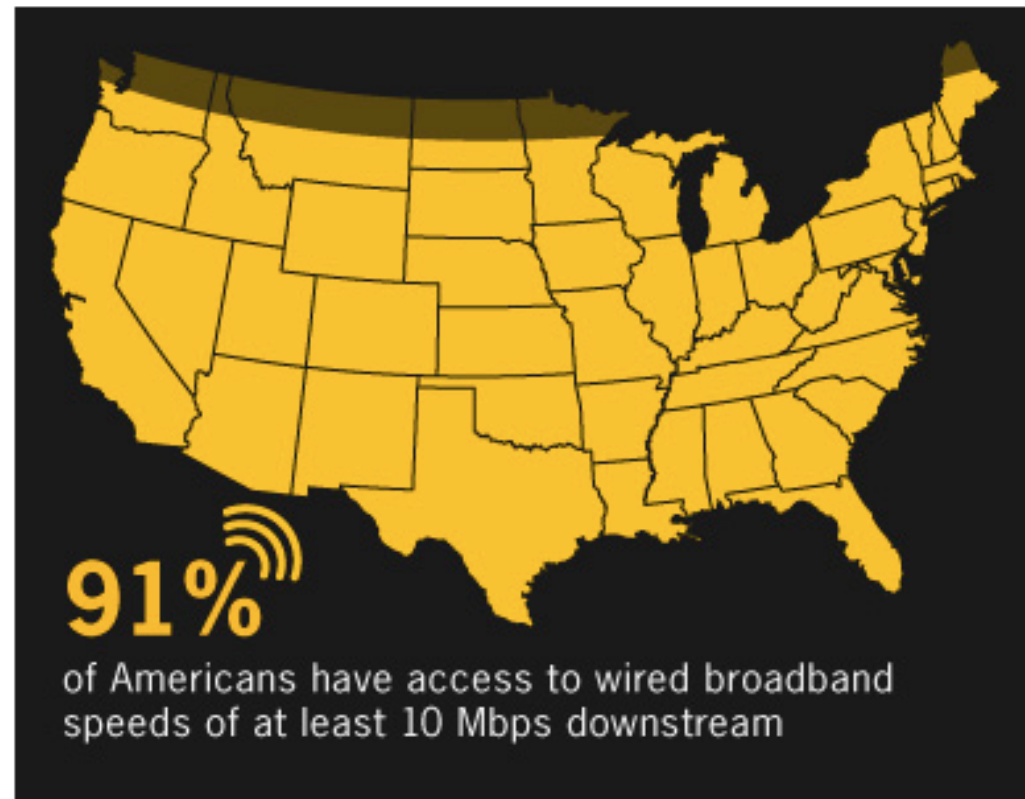
**21%**

► OF TIME IS WASTED DUE TO  
INADEQUATE DIGITAL SKILLS<sup>3</sup>

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>

---

**INTRO TO DATA SCIENCE**

---

# **BOKEH LAB**