CURATION AND DATA MANAGEMENT

FIRST WINTER SCHOOL 2025
Bankya
04-07 February 2025

Peter Stanchev
Kettering University, IMI-BAS

A

Funded by the European Union. Views and opinions expressed are however those of the auhor(s) only and do not necessarily reflect those of the European Union or the REA. Neither the European Union nor the granting authority can be held responsible for them.

Funded by the European Union

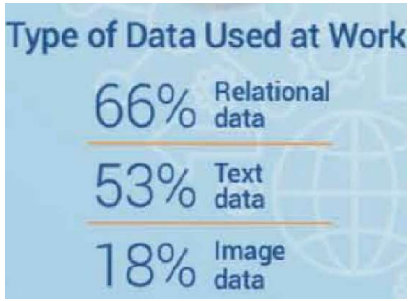
REINFORCING

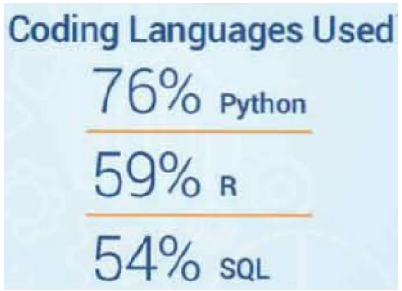
WHAT IS DATA SCIENCE?

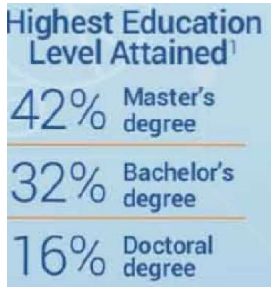
REINFORCING

WHAT IS DATA SCIENCE?

Data Science is the study of data. It is about extracting, analyzing, visualizing, managing and storing data to create insights. These insights help the companies to make powerful data-driven decisions. Data Science requires the usage of both unstructured and structured data. It is a multidisciplinary field that has its roots in statistics, math and computer science. It is one of the most highly sought after jobs due to the abundance of data science position and a lucrative pay-scale [towardsdatascience.com].







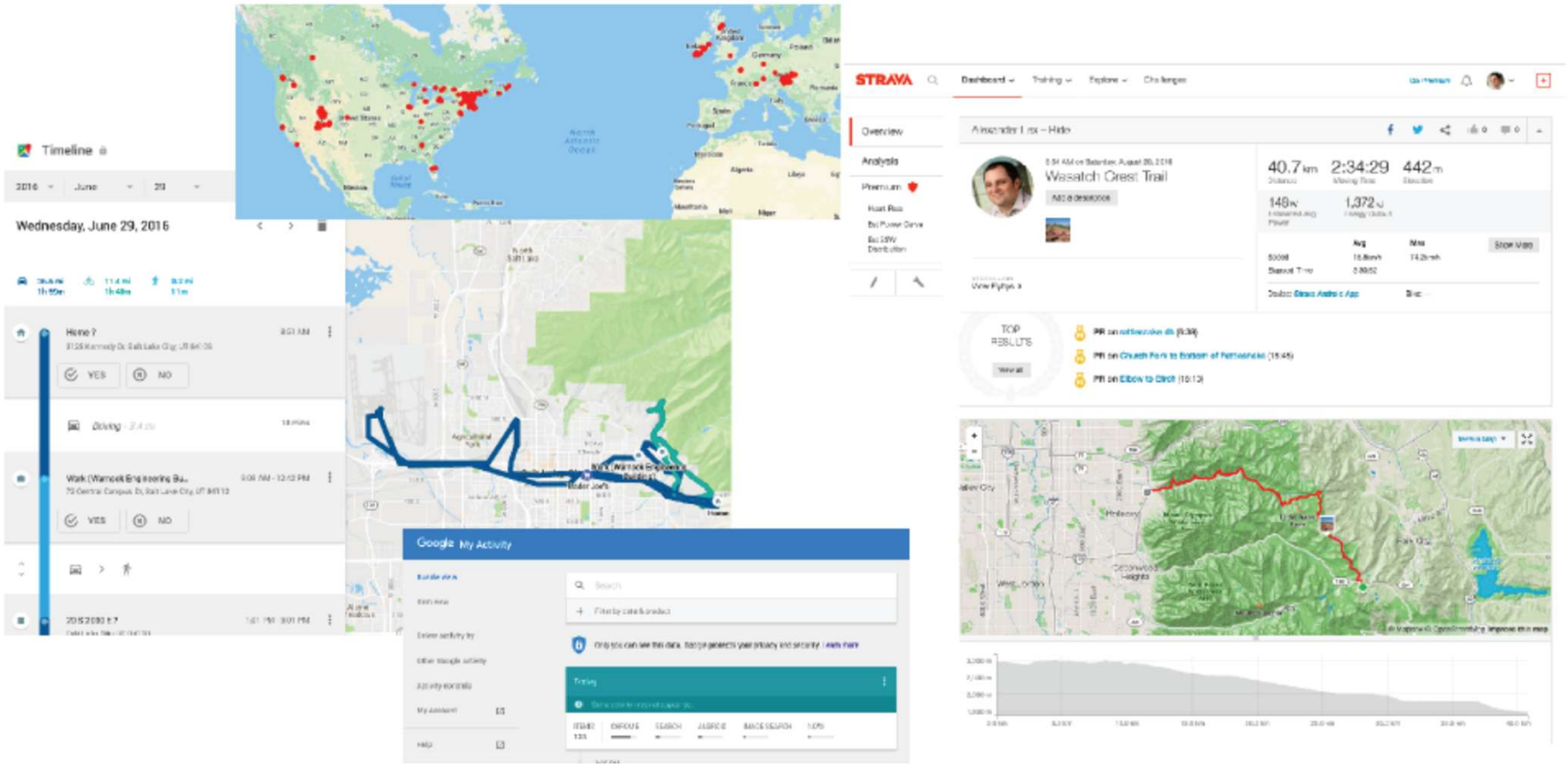
REINFORCING

WHY BECOME A DATA SCIENTIST?

- **It's in Demand**. Is predicted to create 11.5 million jobs by 2026.
- **Abundance of Positions**. Data Science is a vastly abundant field and has a lot of opportunities.
- **A Highly Paid**. Data Scientists make an average of $116,100 per year. The Senior Data Scientist make $138,068 in USA, $85,247 in Europe
- **Data Science Makes Data Better**. The researchers deal with enriching data and making it better for their company.
- **Data Scientists Are Highly Prestigious.** Data Scientists allow companies to make smarter business decisions.
- **Data Science Makes Products Smarter.** It has enabled computers to understand human-behavior and make data-driven decisions.
- **Data Science Can Save Lives**. Healthcare sector has been greatly improved because of Data Science.

REINFORCING

Example: Personal Data



REINFORCING

DATA CURATION

- Data curation is the process of creating, organizing and maintaining data sets so they can be accessed and used by people looking for information. It involves collecting, structuring, indexing and cataloging data for users in an organization, group or the general public.
- Collect the data sets.
- Cleanse the data
- Transform the data
- https://www.youtube.com/watch?v=s6s0BpxUlFo

REINFORCING

What data curation is...

...and why you need to learn it!

REINFORCING

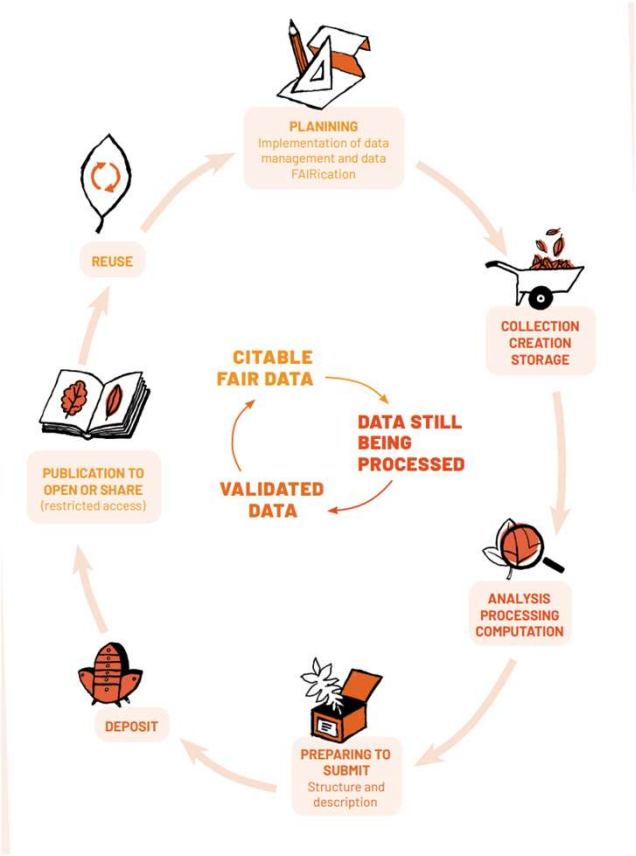
DATA ORGANIZATION

- Involves systematically arranging data in a structured format
- Here are some key points:
  - Categorize and classify data.
  - Use databases, spreadsheets, or other storage systems.
  - Ensure easy retrieval, analysis, and interpretation.

REINFORCING

DATA MANGEMENT

- Data management is the practice of collecting, organizing, managing, and accessing data to support productivity, efficiency, and decision-making.
- https://www.youtube.com/watch?v=ISSb39KlgoI



REINFORCING

REINFORCING

DIFFERENT DATA FORMATS

- Open file format
- JSON
- PDF
- Resource Description Fra...
- MPP
- Bib
- Simple Data Format
- Mdb
- Plain text
- Doc
- Comma-separated values
- Video
- Sas7Bdat
- Tab-separated values
- Lnk
- Xlk
- Metafile
- GeoTIFF
- Au file format
- EPUB
- Proprietary file format
- Data format management
- European Data Format
- Esri grid
- Device independent file f...
- Myd
- Hierarchical Data Format
- Audio file format
- Webarchive
- FITS
- NetCDF
- GeoPackage
- Document file format
- Nsf
- OPML
- Keyhole Markup Language
- PDF/E
- MP3
- Xls
- FASTA format
- OpenDocument
- Industry Foundation Clas...
- Shapefile
- Multimedia
- Spatial Data Transfer Stan...
- Optical disc image
- E00
- .dbf
- GeoJSON
- Office Open XML file for...
- Flat-file database

DIFFERENT DATA FORMATS

pclass	survived	name	sex	age	sibsp	parch	ticket	fare	cabin	embarked	boat	body	home.dest
1	1	Allen, Miss. Elisabeth Walton	female	29	0	0	24160	211.3375	B5	S	2		St Louis, MO
1	1	Allison, Master. Hudson Trevor	male	0.9167	1	2	113781	151.55	C22 C26	S	11		Montreal, PQ / Chesterville, ON

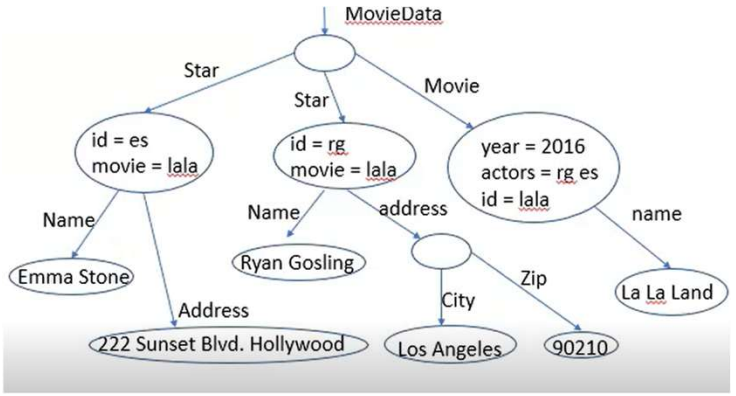
Excel files

REINFORCING

DIFFERENT DATA FORMATS

XML files

```
<MovieData>
<Star id="es" movie="lala">
<Name>Emma Stone</Name>
<Address>222 Sunset Blvd.
Hollywood</Address>
</Star>
<Star id="rg" movie="lala">
<Name>Ryan Gosling</Name>
<Address>
<City>Los Angeles</City>
<Zip>90210</Zip>
</Address>
</Star>
<Movie id="lala" year="2016" actors="rg
es">
<Name>La La Land</Name>
</Movie>
</MovieData>
```



DIFFERENT DATA FORMATS

Json files

```
{
   "firstName": "Jane",
   "lastName": "Doe",
   "hobbies": ["running", "sky diving", "singing"],
   "age": 35,
   "children": [
      {
         "firstName": "Alice",
         "age": 6
      },
      {
         "firstName": "Bob",
         "age": 8
      }
   ]
}
```

DIFFERENT DATA FORMATS

Web pages

ORBIT project
117 followers
2mo •

ORBIT is a new project funded by a cascading grant of the REINFORCING Project that aims to strengthen and expand the application of Open and Responsible Research and Innovation (#ORRI) practices in research within Sofia University St. Kliment Ohridski in collaboration with NGO #Links. The project started on 1 September 2024. Yesterday and today, some of our team members attended a vibrant and informative training in #Brussels. In the photo, Maria Zolotonosa from Stickydot presents best practices in citizen engagement.

REINFORCING

DIFFERENT DATA FORMATS

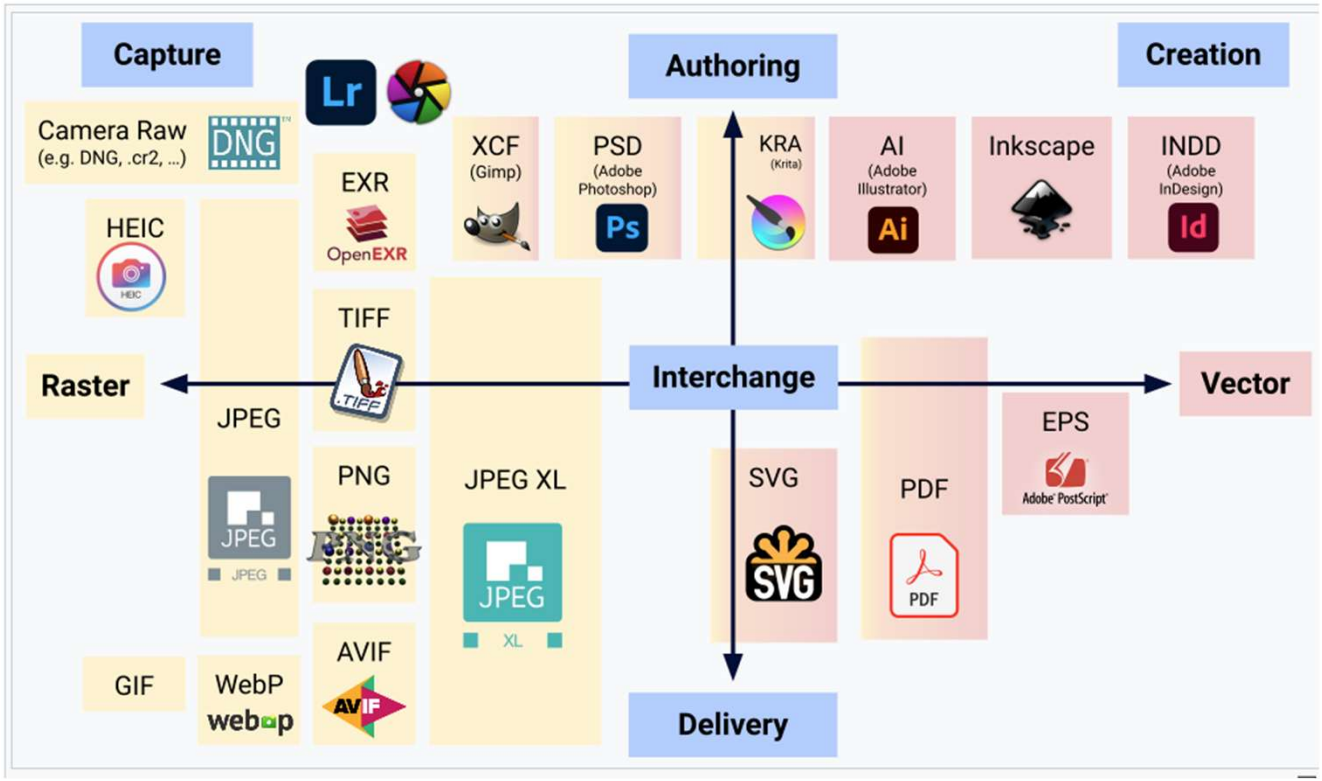
text files

experimental investigation of the aerodynamics of a wing in a slipstream .
an experimental study of a wing in a propeller slipstream was made in order to determine the spanwise distribution of the lift increase due to slipstream at different angles of attack of the wing and at different free stream to slipstream velocity ratios . the results were intended in part as an evaluation basis for different theoretical treatments of this problem .
the comparative span loading curves, together with supporting evidence, showed that a substantial part of the lift increment produced by the slipstream was due to a /destalling/ or boundary-layer-control effect . the integrated remaining lift increment, after subtracting this destalling lift, was found to agree well with a potential flow theory .
an empirical evaluation of the destalling effects was made for the specific configuration of the experiment .

REINFORCING

DIFFERENT DATA FORMATS

Image formats



REINFORCING

DIFFERENT DATA FORMATS

Data series

Price	Adj Close	Close	High	Low	Open	Volume
Ticker	TSLA	TSLA	TSLA	TSLA	TSLA	TSLA
Date						
2020-01-02 00:00:00+00:00	28.684000	28.684000	28.713333	28.114000	28.299999	142981500
2020-01-03 00:00:00+00:00	29.534000	29.534000	30.266666	29.128000	29.366667	266677500
2020-01-06 00:00:00+00:00	30.102667	30.102667	30.104000	29.333332	29.364668	151995000
2020-01-07 00:00:00+00:00	31.270666	31.270666	31.441999	30.224001	30.760000	268231500
2020-01-08 00:00:00+00:00	32.809334	32.809334	33.232666	31.215334	31.580000	467164500
...	...	...	...	...	...	...
2022-05-24 00:00:00+00:00	209.386673	209.386673	217.973328	206.856674	217.843338	89092500
2022-05-25 00:00:00+00:00	219.600006	219.600006	223.106674	207.669998	207.949997	92139300
2022-05-26 00:00:00+00:00	235.910004	235.910004	239.556671	217.886673	220.473328	106003200
2022-05-27 00:00:00+00:00	253.210007	253.210007	253.266663	240.176666	241.083328	89295000
2022-05-31 00:00:00+00:00	252.753326	252.753326	259.600006	244.743332	257.946655	101914500

REINFORCING

DATA CLEANING

REINFORCING

DATA CLEANING

- **Missing data**
  - **Removal**: When missing data severely compromises the quality of analysis, it may be necessary to remove those records.
  - **Imputation**: To avoid losing valuable data, you can fill in missing values using methods like mean, median, or mode imputation.
- **Removing Duplicates**
- **Correcting Inconsistencies**
  - **Standardization**: Consistent data formats across your dataset improves readability and usability.
  - **Validation**: Constraints and checks can prevent invalid data entries, maintaining data integrity.
- **Validating Accuracy**
  - **Cross-Verification**: Comparing your data against trusted external sources can help verify its accuracy and authenticity.
  - **Cross-Verification**: Comparing your data against trusted external sources can help verify its accuracy and authenticity.

REINFORCING

STANDARDIZATION

- **Common Standardization Targets**
  - **Date Formats**: Standardizing dates to a consistent format, such as ‘YYYY-MM-DD
  - **Text Case**: Converting text to a consistent case enhances searchability and uniformity across the dataset
  - **Consistent Units**: Standardizing measurement units prevents calculation errors and ensures clarity

REINFORCING

NORMALIZATION

```
3 data = [{'price':850000, 'rooms':4,'neighborhood':'airport'},
4   {'price':1000000,'rooms':3,'neighborhood':'downtown'},
5 {'price':2000000,'rooms':5,'neighborhood':'Foothill'}]
```

	neighborhood=Foothill	neighborhood=airport	neighborhood=downtown	price	rooms
0	0	1	0	850000	4
1	0	0	1	1000000	3
2	1	0	0	2000000	5

REINFORCING

NORMALIZATION

```
sample = ['problem of evil','evil queen','horizon problem']

from sklearn.feature_extraction.text import CountVectorizer

vec = CountVectorizer()
X = vec.fit_transform(sample)
pd.DataFrame(X.toarray(), columns =vec.get_feature_names
```

```
        evil  horizon  of  problem  queen
```

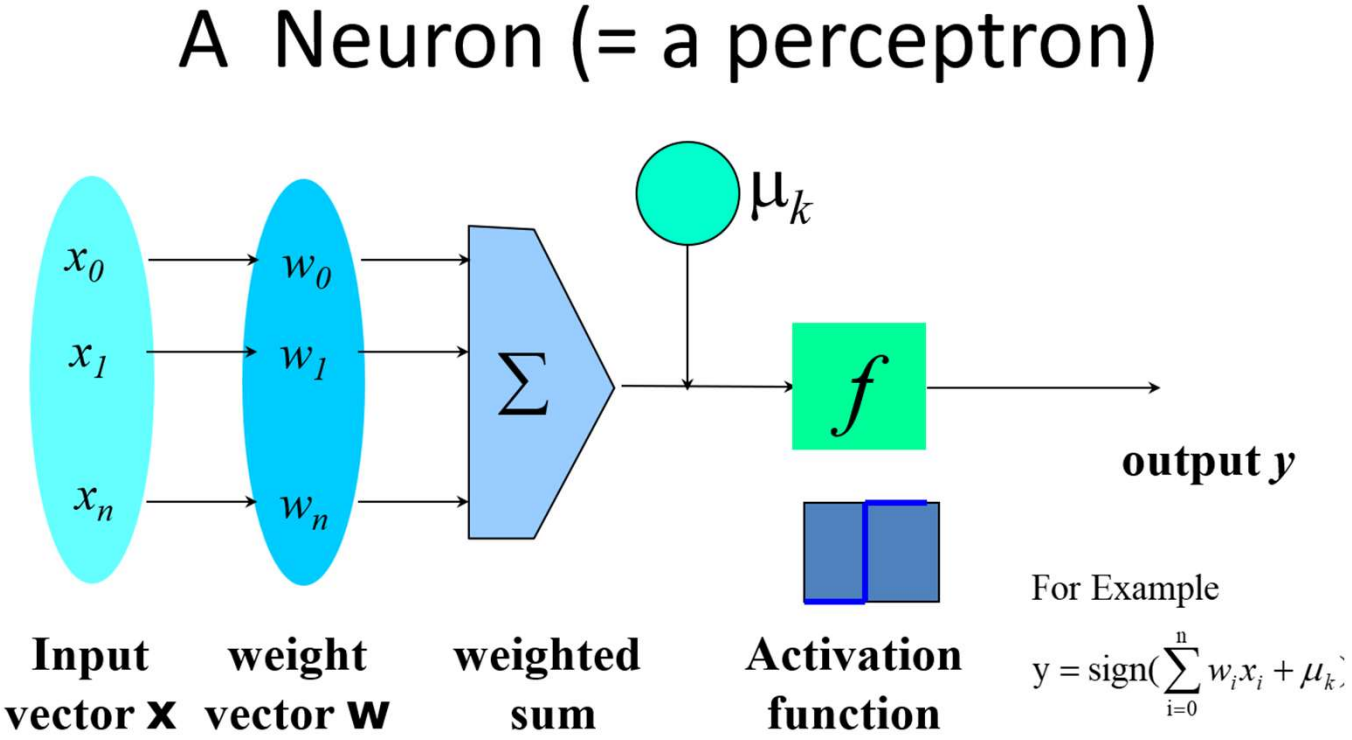
```
from sklearn.feature_extraction.text import TfidfVectorizer

vec = TfidfVectorizer()
X = vec.fit_transform(sample)
pd.DataFrame(X.toarray(), columns = vec.get_feature_names())
```

	evil	horizon	of	problem	queen
0	0.517856	0.000000	0.680919	0.517856	0.000000
1	0.605349	0.000000	0.000000	0.000000	0.795961
2	0.000000	0.795961	0.000000	0.605349	0.000000

REINFORCING

ANALYSIS



BINARY CLASSIFICATION: EXAMPLE

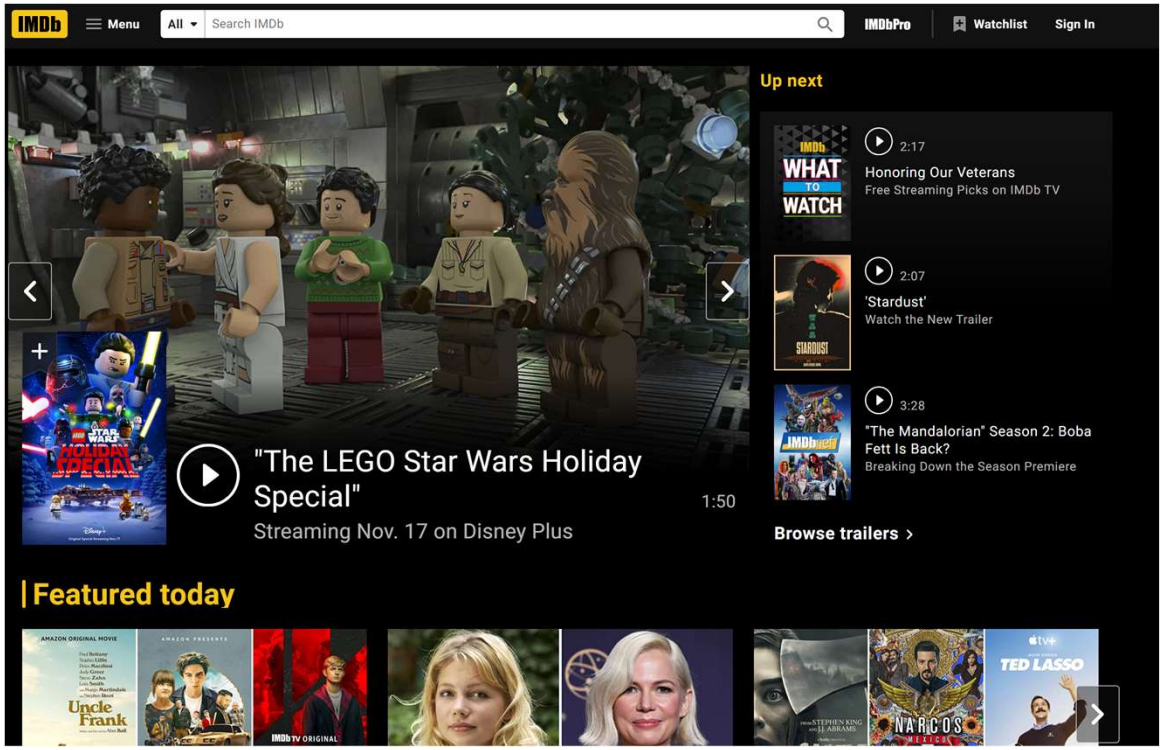
- 50, 000 reviews separated in training and testing sets
- In each set, 1/2 of the reviews are positive and the other half negative.

```
from keras.datasets import imdb

(training_data,training_labels), (testing_data,testing_labels) = imdb.load_data(num_words=10000)
```

- 10000 means that we only keep the top 10,000 most frequent words in the reviews.
- Each review will be represented as a list of words, e.g, (2,53,233,22) is a review of 4 words.
- A label of 0 means negative review and a label of 1 means positive review.

REINFORCING




```
1 results = model.evaluate(x_testing,y_testing)
2
```

```
782/782 [==============================] - 2s 2ms/step - loss: 0.2908 - accuracy: 0.8857
```

REINFORCING

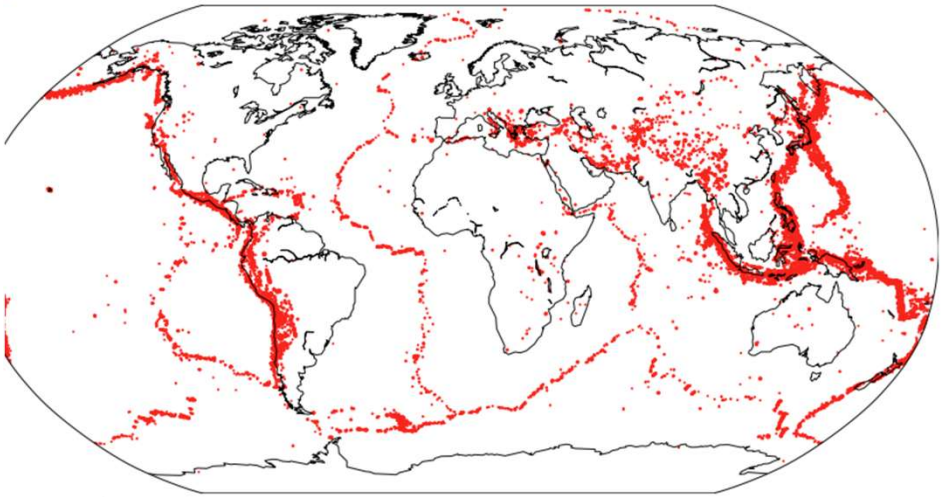
```
1 nba
```

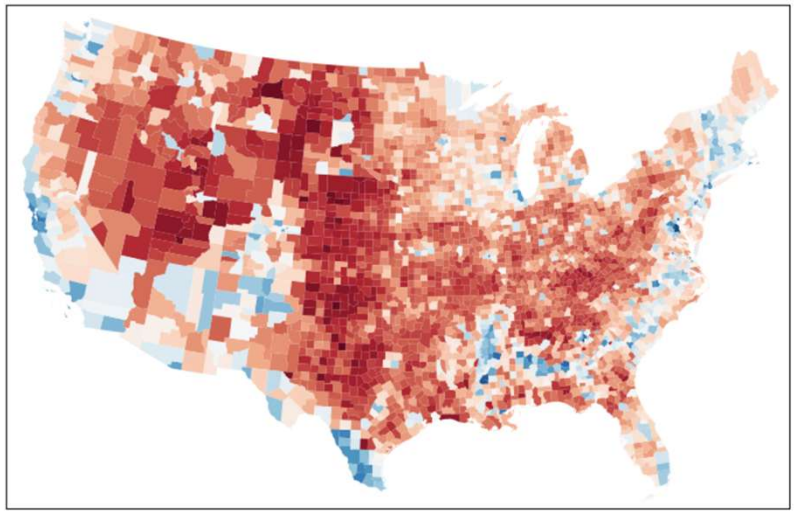
	player	pos	age	bref_team_id	g	gs	mp	fg	fga	fg.	...	drb	trb	ast	stl	blk	tov	pf	pts	season	season_end
0	Quincy Acy	SF	23	TOT	63	0	847	66	141	0.468	...	144	216	28	23	26	30	122	171	2013-2014	2013
1	Steven Adams	C	20	OKC	81	20	1197	93	185	0.503	...	190	332	43	40	57	71	203	265	2013-2014	2013
2	Jeff Adrien	PF	27	TOT	53	12	961	143	275	0.520	...	204	306	38	24	36	39	108	362	2013-2014	2013
3	Arron Afflalo	SG	28	ORL	73	73	2552	464	1011	0.459	...	230	262	248	35	3	146	136	1330	2013-2014	2013
4	Alexis Ajinca	C	25	NOP	56	30	951	136	249	0.546	...	183	277	40	23	46	63	187	328	2013-2014	2013

```
1 nba['distance'] = nba_normalized.apply(lambda row: euclidean_distance(row, player_normalized,distance_columns), axis=1)
2
3 nba.sort_values('distance',inplace=True)
4 print(nba)
5
```

```
            player pos  age bref_team_id   g  gs    mp   fg   fga    fg.  \
225     LeBron James  PF   29          MIA  77  77  2902  767  1353  0.567
17   Carmelo Anthony  PF   29          NYK  77  77  2982  743  1643  0.452
277       Kevin Love  PF   25          MIN  77  77  2797  650  1421  0.457
179    Blake Griffin  PF   24          LAC  80  80  2863  718  1359  0.528
110    Stephen Curry  PG   25          GSW  78  78  2846  652  1383  0.471
```

REINFORCING

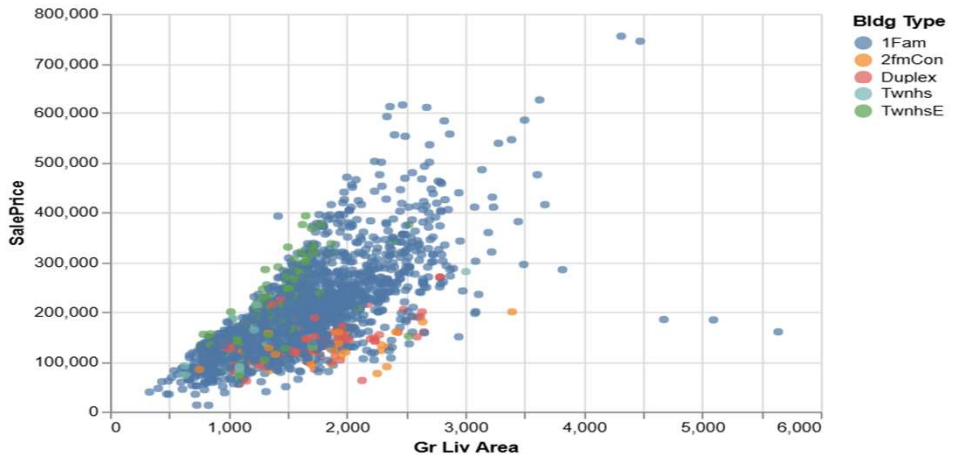




REINFORCING

PRESENTING THE DATA

```
1 import altair as alt
2
3
4 alt.Chart(housing).mark_circle().encode(
5     x="Gr Liv Area",
6     y="SalePrice",
7     color="Bldg Type"
8 )
9
10
11
```



REINFORCING

THANK YOU

Peter Stanchev

pstanche@kettering.edu

www.reinforcing.eu