PYTHON SEMINAR 2020

JENS HAHN

THEORETICAL BIOPHYSICS

TODAY

I Recap data analysis I

II Data analysis II

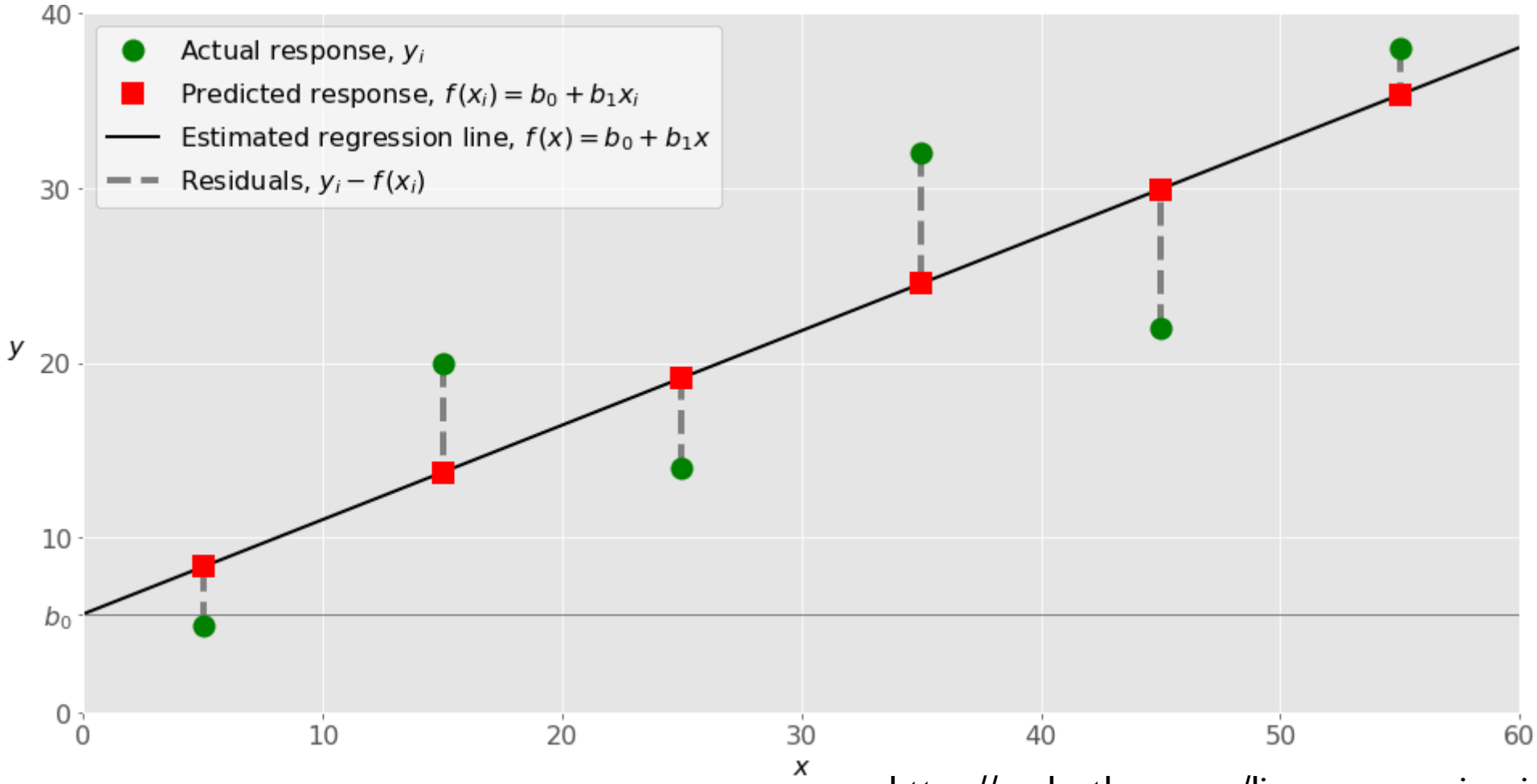
III Pandas

IV Assignment

I. RECAP CLASSES

- Numeric data set
- numpy arrays
- Masking of numpy arrays
- Mean, median, SD
- Interpolation/extrapolation

I. LINEAR REGRESSION



https://realpython.com/linear-regression-in-python/

I. ASSIGNMENT – TURN IT INTO A CLASS

- Write a class DataAnalysis
- Write methods for every task

- Import data
- Calculate mean & median
- Calculate SD
- Normalisation
- Interpolation
- Linear regression

II. DATA ANALYSIS

Large data sets (oil of the 21th century)

- Various data types (numbers, abbreviations, descriptions, dates…)
- Large amount of data (too much for Excel!)
- No clear questions before gathering data

1. Curate
2. Analyse
3. Visualise

II. DATA ANALYSIS

Data set: parking violation data of L.A.

- Ticket: Number, issuing time, fine
- Violation: Description and location
- Plate: State plate and expiry date
- Car: Make, style, colour
- Agency: Code and route

https://data.lacity.org/A-Well-Run-City/Parking-Citations/wjz9-h9np

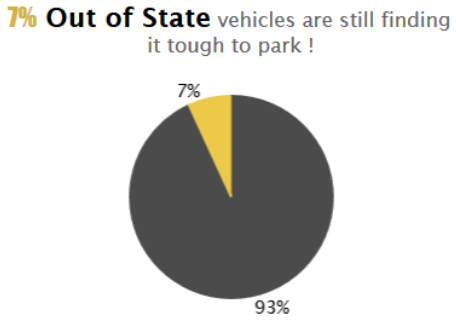
II. DATA ANALYSIS

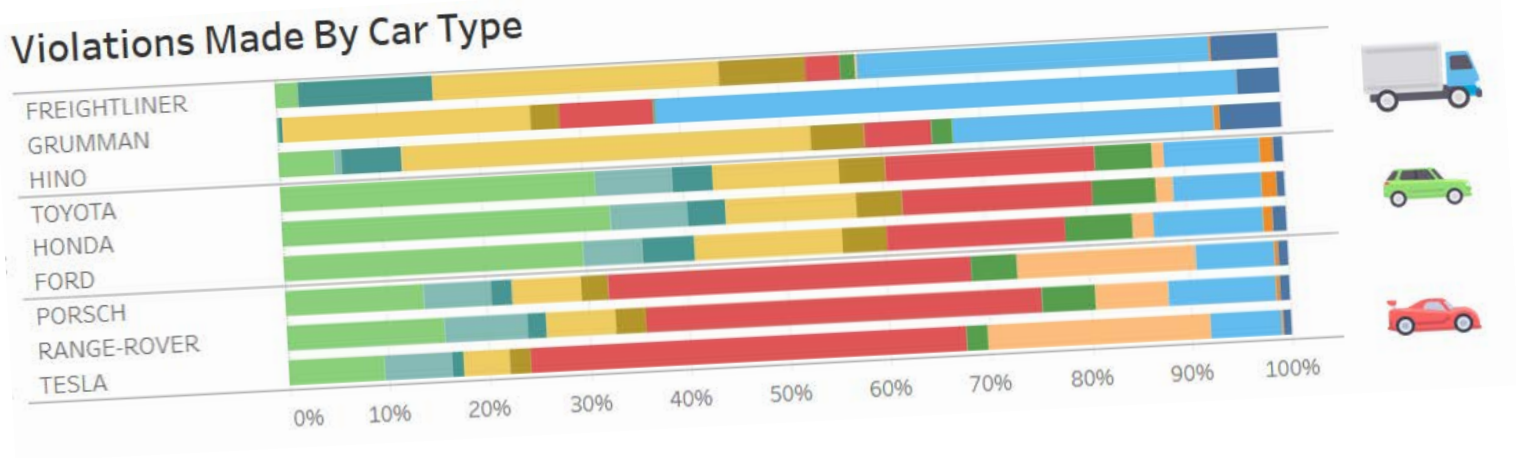
Data set: parking violation data of L.A.

1. Too big for Excel
2. What can we learn from the data? And how?

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	Ticket numl	Issue Date	Issue t	Meter	Marked	RP St	Plate Expir	VIN	Make	Body Style	Color	Location	Route	Agency	Violation code	Violation Description	Fine an	Latitude	Longitude
2	1103341116	21.12.2015 00:00	1251			CA	200304		HOND	PA	GY	13147 WELBY WAY	01521	1	4000A1	NO EVIDENCE OF REG	50	99999	99999
3	1103700150	21.12.2015 00:00	1435			CA	201512		GMC	VN	WH	525 S MAIN ST	1C51	1	4000A1	NO EVIDENCE OF REG	50	99999	99999
4	1104803000	21.12.2015 00:00	2055			CA	201503		NISS	PA	BK	200 WORLD WAY	2R2	2	8939	WHITE CURB	58	64399979	18026864
5	1104820732	26.12.2015 00:00	1515			CA			ACUR	PA	WH	100 WORLD WAY	2F11	2	000	17104h		64400411	18026862
6	1105461453	15.09.2015 00:00	115			CA	200316		CHEV	PA	BK	GEORGIA ST/OLYMPIC	1FB70	1	8069A	NO STOPPING/STANDING	93	99999	99999
7	1106226590	15.09.2015 00:00	19			CA	201507		CHEV	VN	GY	SAN PEDRO S/O BOYD	1A35W	1	4000A1	NO EVIDENCE OF REG	50	99999	99999
8	1106500452	17.12.2015 00:00	1710			CA	201605		MAZD	PA	BL	SUNSET/ALVARADO	00217	1	8070	PARK IN GRID LOCK ZN	163	99999	99999

II. DATA ANALYSIS





Fedex And **UPS** trucks recived enough tickets to account for **10 full-time traffic officers**.

https://public.tableau.com/profile/amit.n6941#!/

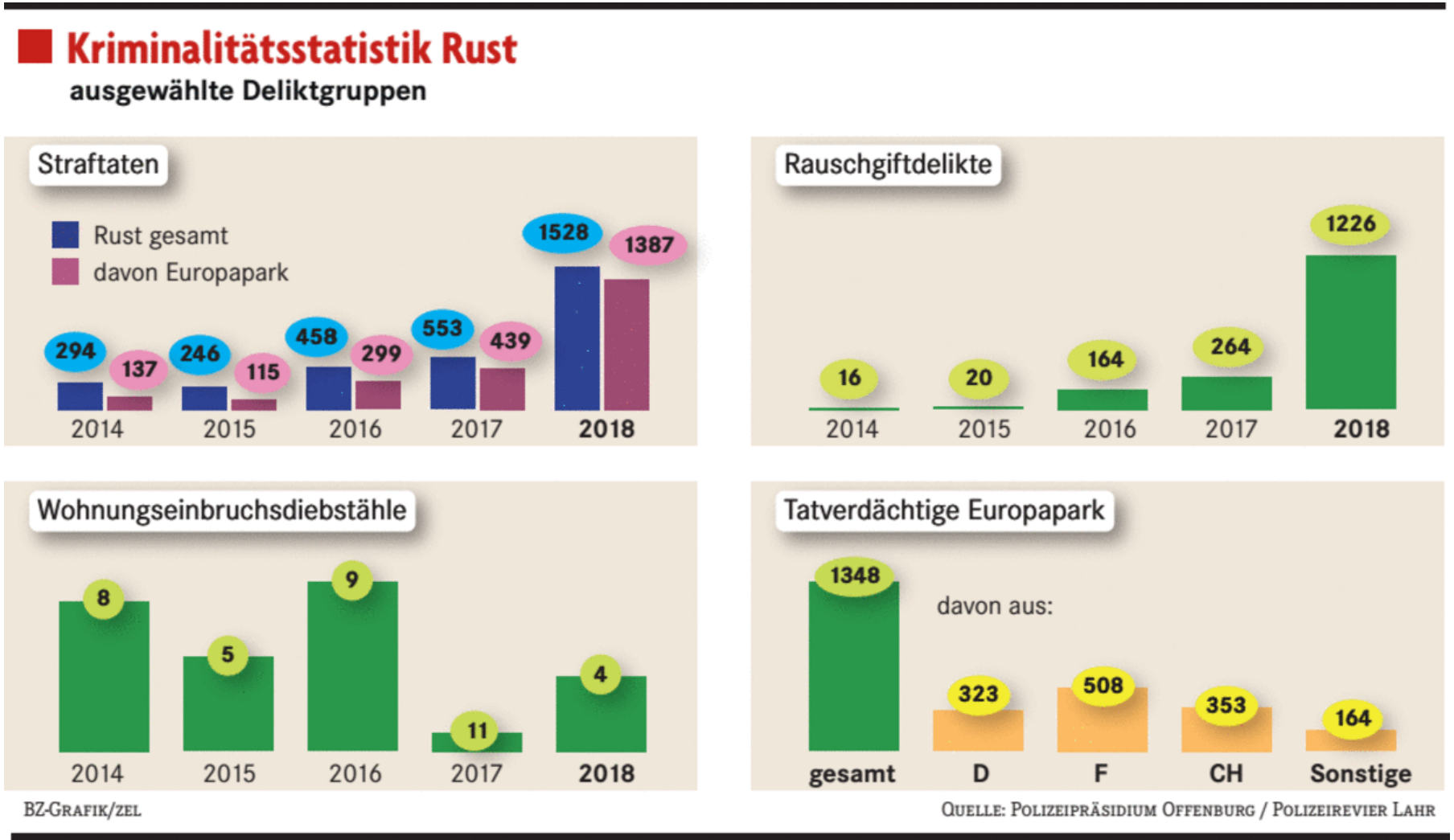
https://public.tableau.com/profile/jack.chen3893#!/

II. DATA ANALYSIS

Data set: parking violation data of L.A.

- How many out-of-state cars are parked wrong?
- Street with the most parking violations?
- Which car type and car colour is highest in parking violations?
- On which daytime the most cars are violating parking rules?

II. DATA ANALYSIS



III. PANDAS

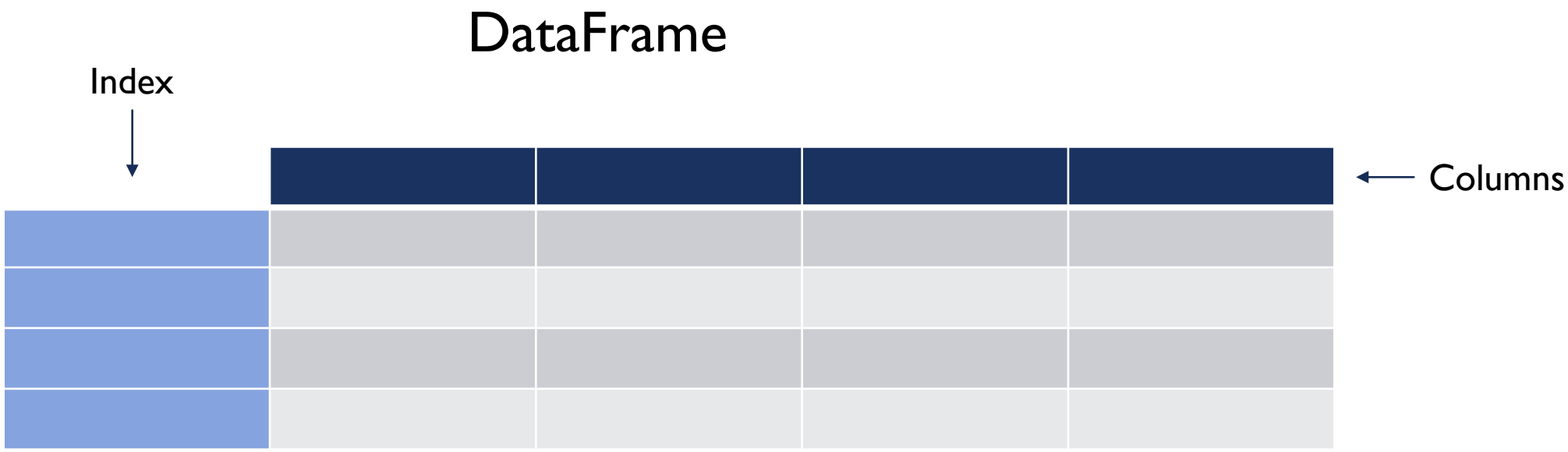
Python packages for data analysis (learned from R)

```
import pandas as pd
```

- Read data (csv, Excel, …)

```
df = pd.read_csv('./parking_data_small.csv')
```

III. PANDAS



III. PANDAS

Have a look on the data

- Data types: `df.dtypes`
- Dimensions: `df.shape`
- Top [Bottom] 5 rows: `df.head()` `[df.tail()]`
- Basic statistics: `df.describe()`

III. PANDAS

Address data

- Transpose data frame: `df.T`
- Indices `df.index`
- Columns `df.columns`
- Slicing: `df.loc[index_name, column_name]`

  `df.iloc[1, 4]`

III. PANDAS

More advanced

- Pick data:

```
df[df[column_name] == value]

df[df[column_name].isin([v1, v2])]
```

- Group data

```
df.groupby(column_name).mean()
```

- Count values

```
df[column_name].value_counts()
```

III. PANDAS

Pivot tables

```
df.pivot(values=,columns=,index=)
```

	A	B	C	D	E
0	0.250124	0.457986	0.146158	meep	1
1	0.333871	0.954610	0.911692	meep	2
2	0.432136	0.537708	0.001518	map	1

	A	B
D		
map	0.432136	0.537708
meep	0.291998	0.706298

	A		B	
E	1	2	1	2
D				
map	0.432136	NaN	0.537708	NaN
meep	0.250124	0.333871	0.457986	0.95461

IV. ASSIGNMENT

Data analysis of data

- Define questions (statistically correct)
- Think about how to find the answer in the data
- Extract the information from the data

V. FURTHER READING

Data analysis

- Data analysis in Python

  http://www.data-analysis-in-python.org/

- Coursera – data analysis

  https://www.coursera.org/learn/data-analysis-with-python

Python pandas

- Pandas tutorial

  https://www.python-kurs.eu/pandas.php

- Pandas documentation

  https://pandas.pydata.org/pandas-docs/stable/