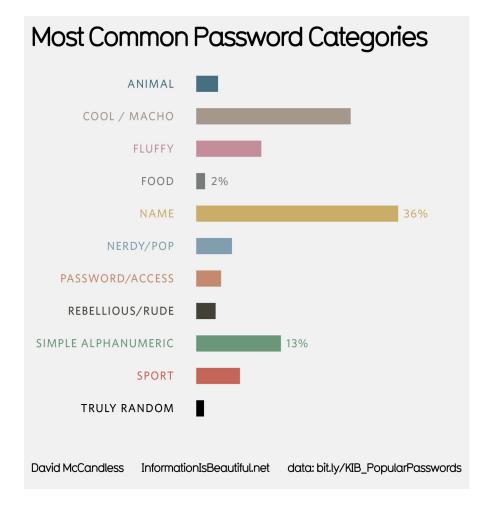# Case Study

Passwords

Prof. Dr. Jan Kirenz
HdM Stuttgart

Top ten rows of the passwords dataset

	rank	password	category	value	time_unit	offline_crack_sec	rank_alt	strength	font_size
0	1	password	password-related	6.91	years	2.170000e+00	1	8	11
1	2	123456	simple-alphanumeric	18.52	minutes	1.110000e-05	2	4	8
2	3	12345678	simple-alphanumeric	1.29	days	1.110000e-03	3	4	8
3	4	1234	simple-alphanumeric	11.11	seconds	1.110000e-07	4	4	8
4	5	qwerty	simple-alphanumeric	3.72	days	3.210000e-03	5	8	11
5	6	12345	simple-alphanumeric	1.85	minutes	1.110000e-06	6	4	8
6	7	dragon	animal	3.72	days	3.210000e-03	7	8	11
7	8	baseball	sport	6.91	years	2.170000e+00	8	4	8
8	9	football	sport	6.91	years	2.170000e+00	9	7	11
9	10	letmein	password-related	3.19	months	8.350000e-02	10	8	11

Bottom ten rows of the passwords dataset.

	rank	password	category	value	time_unit	offline_crack_sec	rank_alt	strength	font_size
490	491	natasha	name	3.19	months	0.08350	493	7	11
491	492	sniper	cool-macho	3.72	days	0.00321	494	8	11
492	493	chance	name	3.72	days	0.00321	495	7	11
493	494	genesis	nerdy-pop	3.19	months	0.08350	496	7	11
494	495	hotrod	cool-macho	3.72	days	0.00321	497	7	11
495	496	reddog	cool-macho	3.72	days	0.00321	498	6	10
496	497	alexande	name	6.91	years	2.17000	499	9	12
497	498	college	nerdy-pop	3.19	months	0.08350	500	7	11
498	499	jester	name	3.72	days	0.00321	501	7	11
499	500	passw0rd	password-related	92.27	years	29.02000	502	28	21

Data information

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 500 entries, 0 to 499
Data columns (total 9 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   rank              500 non-null    int64
 1   password          500 non-null    object
 2   category          500 non-null    object
 3   value             500 non-null    float64
 4   time_unit         500 non-null    object
 5   offline_crack_sec 500 non-null    float64
 6   rank_alt          500 non-null    int64
 7   strength          500 non-null    int64
 8   font_size         500 non-null    int64
dtypes: float64(2), int64(4), object(3)
memory usage: 35.3+ KB
```

Practice

Do these data come from an observational study or an experiment?

There are 500 rows and 7 columns in the dataset. What does each row and each column represent?

Variables and their descriptions for the passwords dataset.

Variable	Description
rank	Popularity in the database of released passwords.
password	Actual text of the password.
category	Category password falls into.
value	Time to crack by online guessing.
time_unit	Time unit to match with value.
offline_crack_sec	Time to crack offline in seconds.
strength	Strength of password, relative only to passwords in this dataset. Lower values indicate weaker passwords.

Practice

Determine whether each variable in the passwords dataset is numerical or categorical.

- For **numerical** variables, further classify them as **continuous** or **discrete**.
- For **categorical** variables, determine if the variable is **nominal** or **ordinal**.

Variables and their class for the passwords dataset.

variable	class	description
rank	double	popularity in their database of released passwords
password	character	Actual text of the password
category	character	What category does the password fall in to?
value	double	Time to crack by online guessing
time_unit	character	Time unit to match with value
offline_crack_sec	double	Time to crack offline in seconds
rank_alt	double	Rank 2
strength	double	Strength = quality of password where 10 is highest, 1 is lowest, please note that these are relative to these generally bad passwords
font_size	double	Used to create the graphic for KIB

Resources

The slides are based on the excellent book "Introduction to Modern Statistics" by Mine Çetinkaya-Rundel and Johanna Hardin.

The online version can be **accessed for free**:

https://openintro-ims.netlify.app/data-applications.html