BDA 2020/21
Datasets for projects

more details at this link:
http://bit.ly/bda2021_datasets

	set 15 MAR 10.00 12.00	set 15 MAR 14.00 16.00	set 15 MAR 16.00 18.00	set 16 MER 10.00 12.00	set 16 MER 14.00 16.00	set 16 MER 16.00 18.00	set 17 GIO 10.00 12.00	set 17 GIO 14.00 16.00	set 17 GIO 16.00 18.00	set 18 VEN 10.00 12.00	set 18 VEN 14.00 16.00	set 18 VEN 16.00 18.00
19 partecipanti +	✓2	✓11	✓12	✓3	✓4	✓4	✓3	✓3	✓6	✓3	✓3	✓2
Luca Pappal...												
Tommaso Cavalieri		✓										
Vitalba Macaluso		✓	✓									
Aldo Iannarelli		✓	✓									
Daniele Tribuzio		✓	✓									
Malick jobe			✓									
Davide Domenico C...			✓			✓			✓			✓
Alessandro Poggiali	✓	✓		✓	✓	✓	✓			✓	✓	
Marco Sorrenti					✓							
Cinzia Lestini		✓	✓		✓			✓	✓			
Lucrezia Orlandini		✓					✓			✓	✓	
BALDASSARE BONGI...		✓	✓	✓				✓				
Clara D'Apoli		✓	✓									
Luca Corbucci			✓			✓			✓			
andrei sauca			✓						✓	✓		✓
Simone Rossi			✓						✓			
Ivan Ogando		✓	✓						✓			
Marco Marino	✓			✓	✓	✓	✓				✓	
Jose Gonzalez		✓						✓				

List of datasets

- Heart Failure Prediction
- Google Play Store Apps
- Red Wine Quality
- Modelling earthquake damage
- Predict Flu Vaccines
- Pump it up: data mining the water table
- Predicting Disease Spread

Module 3: laboratory for interactive project development

- Create teams of “data analysts”
- Choose a dataset among those proposed

1. *October*: 1st Mid Term (Data Understanding and Project Formulation)
2. *November*: 2nd Mid Term (Model implementation and evaluation)
3. *December*: 3rd Mid Term (Model interpretation and explanation)
4. *January*: Exam (Final Project results)

Building Teams...

- 3 teams already registered
  I TeamIDI (4), BigProblem (4), MMG (3)

- Register your team by September 27th
- Write me an email if you cannot find any team

- Some people are in a team, but not in the pre-registration. Please, register!

Heart Failure Prediction

12 variables
(age, anaemia, creatinine, diabetes, sex, etc.)

1 target
(1 patient has died, 0 has not)

Cardiovascular diseases (CVDs) are the number 1 cause of death globally (31% of all deaths worldwide).

- **unbalanced data set**
- **small data set**
  (300 instances)

kaggle

https://www.kaggle.com/andrewmvd/heart-failure-clinical-data

Google Play Store Apps

14 variables
(name, category, user rating, reviews, dimension of the app, downloads, price, age group, genre, date, version, ...)

target: success of the app (e.g., the rating)

https://www.kaggle.com/lava18/google-play-store-apps

Red Wine Quality

11 variables based on physicochemical tests

1 target: *quality* score in [0, 10]

fixed acidity, acidity, citric acid, sugar, chlorides, free and total sulfur dioxide, density, pH, sulphates, alcohol

ordered classes
unbalanced dataset

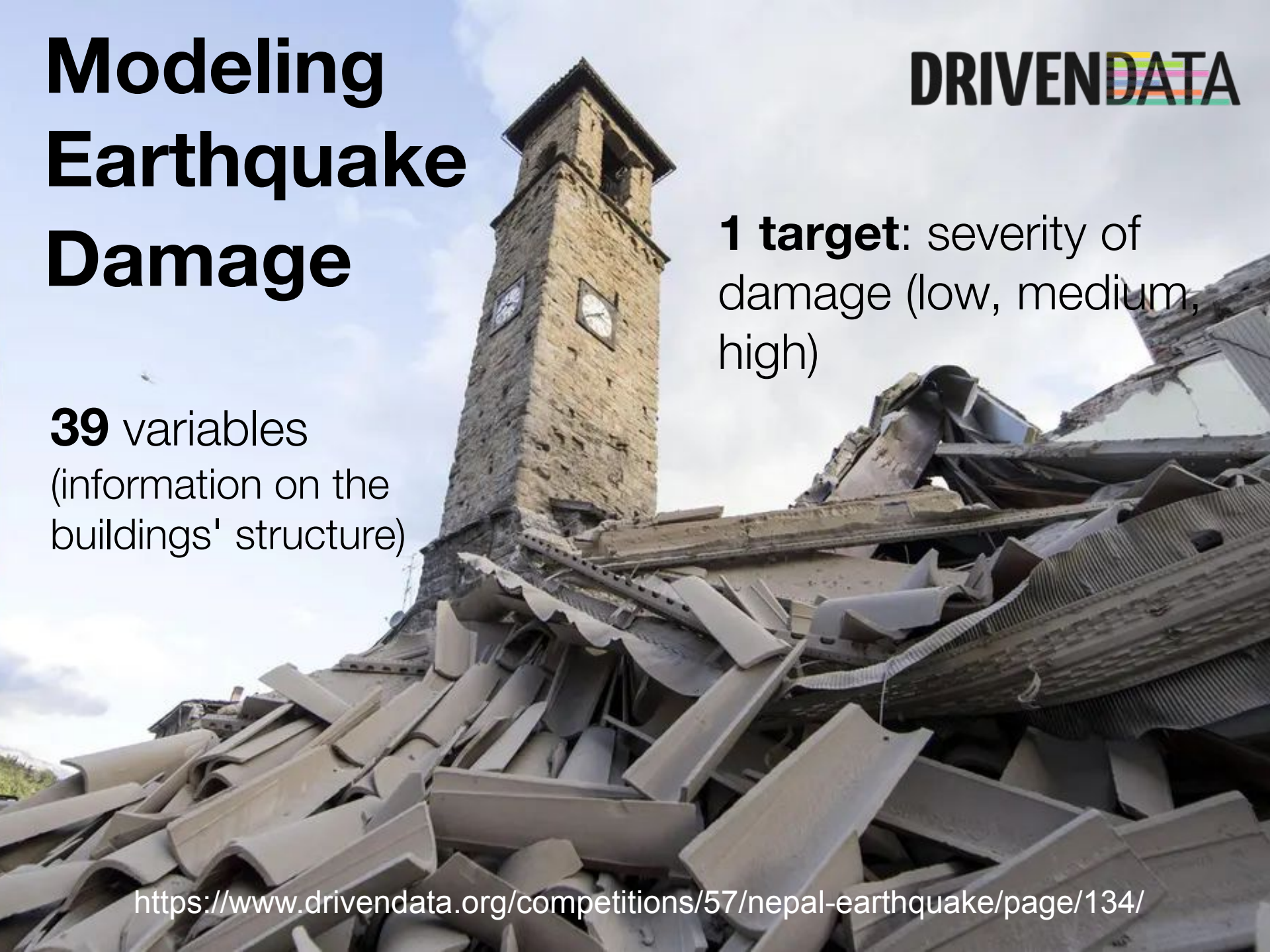
https://archive.ics.uci.edu/ml/datasets/wine+quality
https://www.kaggle.com/uciml/red-wine-quality-cortez-et-al-2009

Modeling Earthquake Damage

DRIVENDATA

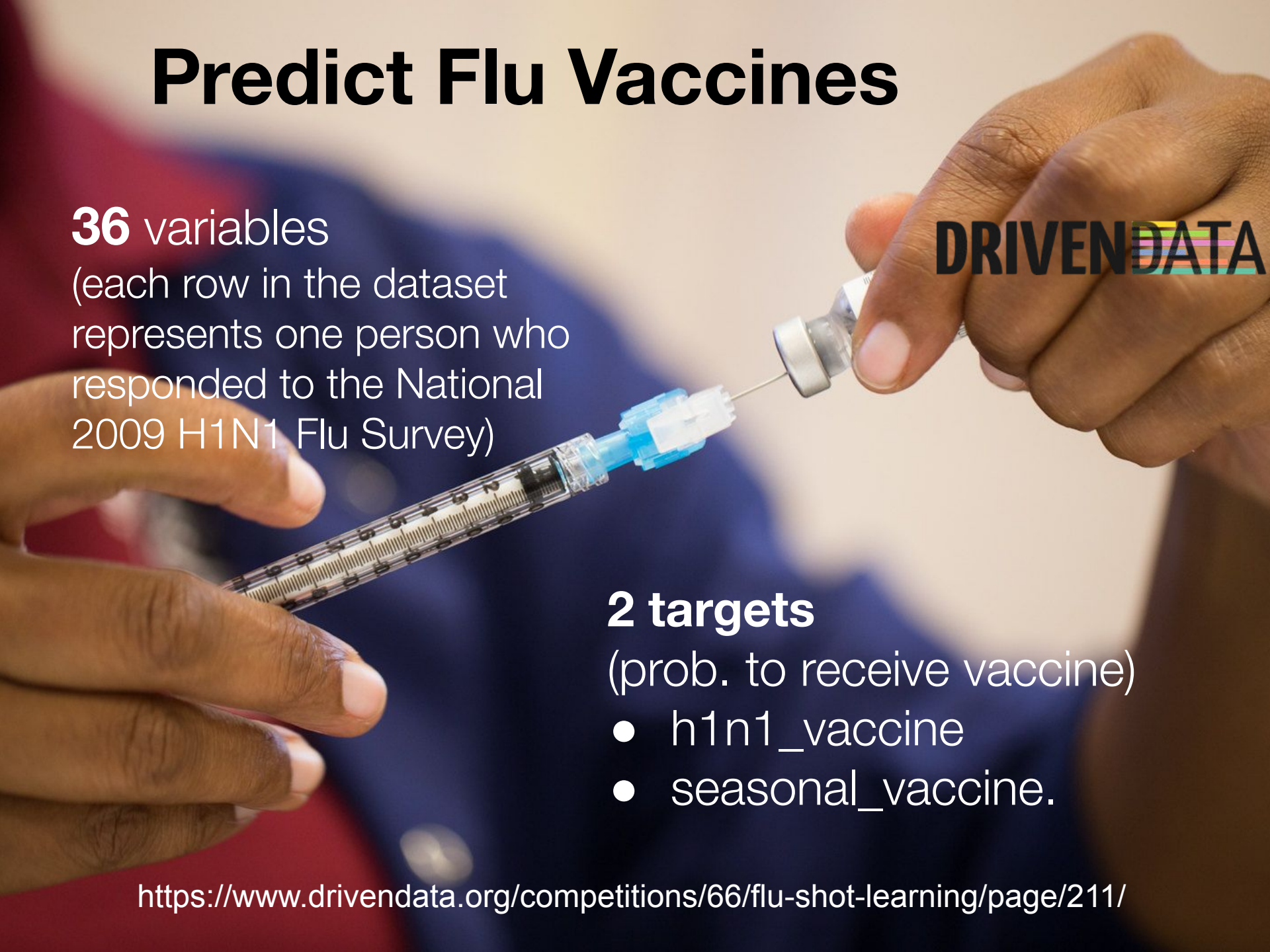
1 target: severity of damage (low, medium, high)

39 variables (information on the buildings' structure)



https://www.drivendata.org/competitions/57/nepal-earthquake/page/134/

Predict Flu Vaccines



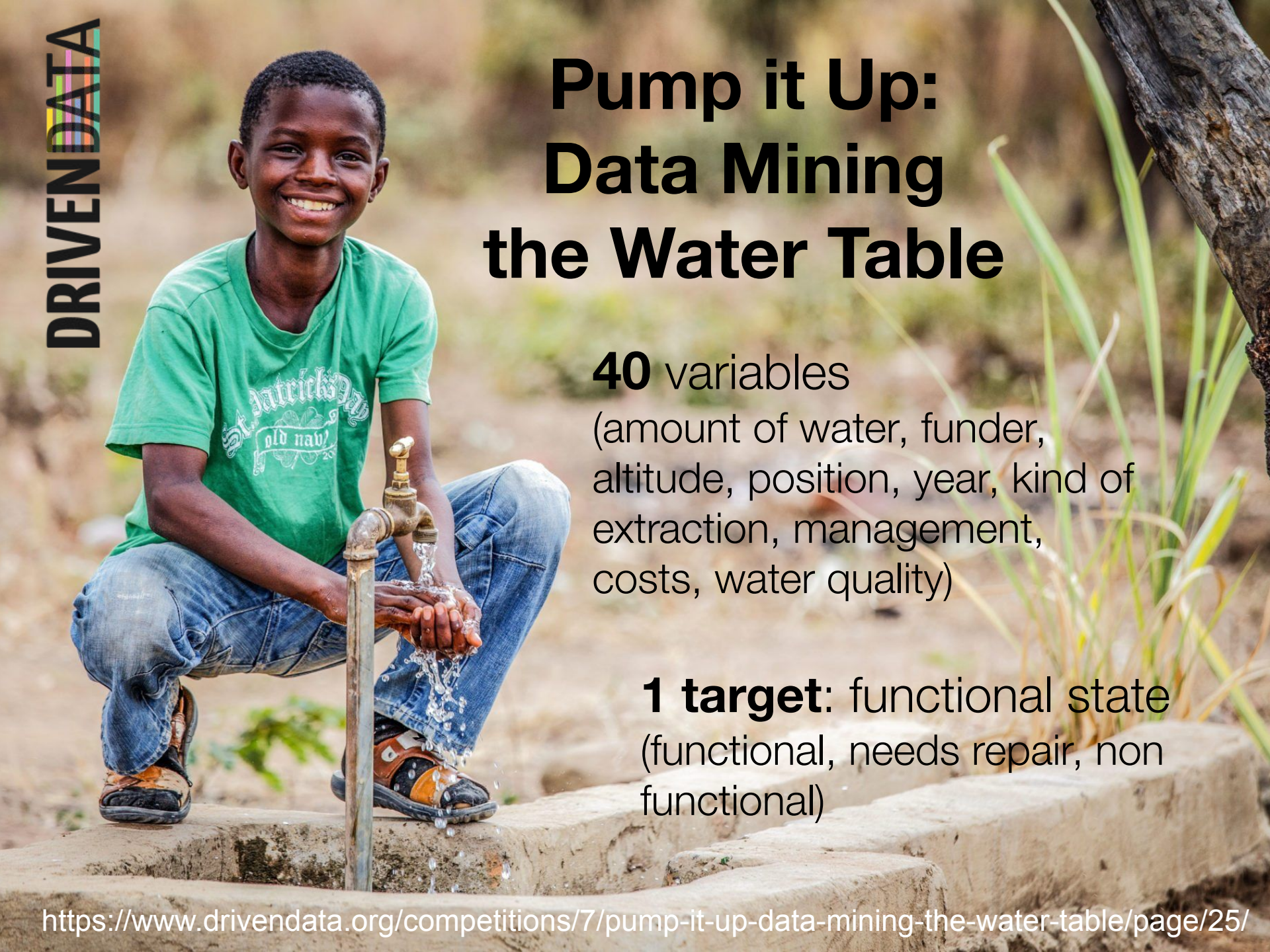
36 variables
(each row in the dataset represents one person who responded to the National 2009 H1N1 Flu Survey)

DRIVENDATA

2 targets
(prob. to receive vaccine)

- h1n1_vaccine
- seasonal_vaccine.

https://www.drivendata.org/competitions/66/flu-shot-learning/page/211/


DRIVENDATA
Pump it Up:
Data Mining
the Water Table
40 variables
(amount of water, funder, altitude, position, year, kind of extraction, management, costs, water quality)
1 target: functional state
(functional, needs repair, non functional)
https://www.drivendata.org/competitions/7/pump-it-up-data-mining-the-water-table/page/25/

Predicting Disease Spread

DRIVENDATA

20+ variables
(temperature, precipitation, humidity, vegetation, and more.)

1 target: total number of cases for each (city, year, weekofyear).

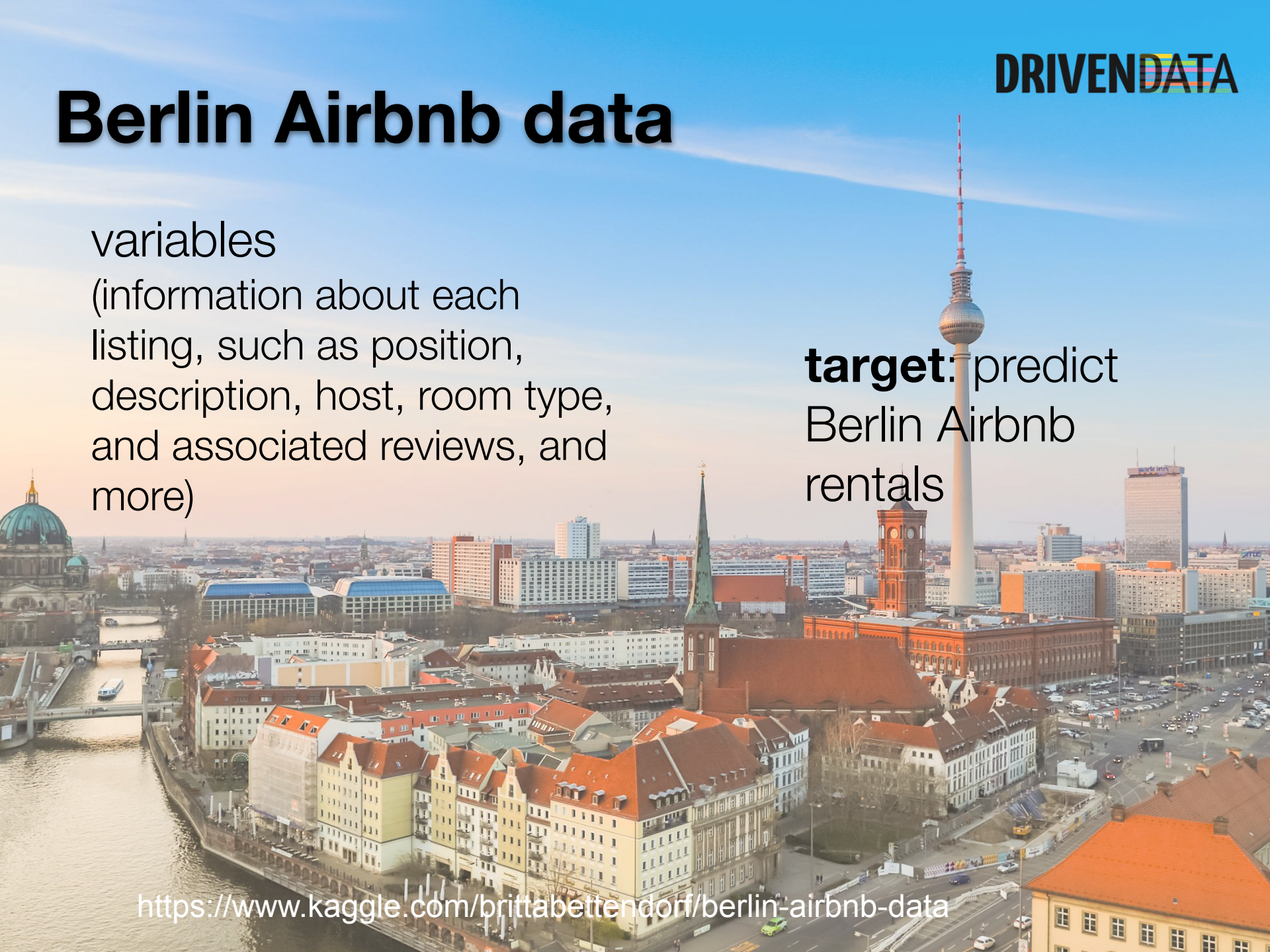
https://www.drivendata.org/competitions/44/dengai-predicting-disease-spread/page/82/

Berlin Airbnb data

DRIVENDATA

variables
(information about each listing, such as position, description, host, room type, and associated reviews, and more)

target: predict Berlin Airbnb rentals



https://www.kaggle.com/brittabettendorf/berlin-airbnb-data