

# Brust Krebs Diagnosis

Karla Araceli Luque Cobos, Maria Beatriz Walter Costa und  
Ulrike Rösler  
Betreuer: Anatol Reibold

Big Data Analytics (alphatraining)

Oktober/2019

# Hypothesis und Probleme

- Hypothesis: es ist möglich mit numerische Parametern Brustkrebs zu diagnostizieren (gut oder böseartig)
- Problem I: Machine Learning models erstellen (von gut und böseartige Stichproben)
- Problem II: neue Patientinnen klassifizieren (Diagnostics: g/b)
- Multivariate Dataset von 1995
  - 569 Patientinnen
  - 357 gutartige (63%), 212 böseartige (37%)

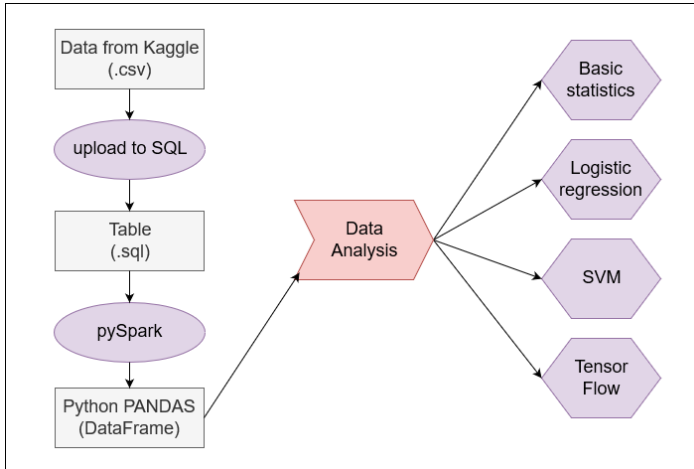
# Dataset

- Wie hat man dieser dataset gemacht?
- Krebs Brustzelle sammeln
- Bilder auf Brustzellen machen (nuclei)
- Umsetzung von Bilder zu Parametern mit Image Recognition Methoden
- Output: Tabelle mit der numerische Parametern
- Source von University of Wisconsin (Kaggle): <https://www.kaggle.com/uciml/breast-cancer-wisconsin-data>

# Parametern von der Tabelle

- 1 ID number
- 2 Diagnosis (M = malignant, B = benign)
- 3 Ten real-valued features are computed for each cell nucleus:
  - a) radius (distance means from center to perimeter points)
  - b) texture (std deviation of gray-scale values)
  - c) perimeter
  - d) area
  - e) smoothness (local variation in radius lengths)
  - f) compactness ( $perimeter^2 / area - 1.0$ )
  - g) concavity (severity of concave portions of the contour)
  - h) concave points (number of concave portions of the contour)
  - i) symmetry
  - j) fractal dimension ("coastline approximation" - 1)

# Data Analysis (Pipeline)



# Ergebnisse

Python Jupyter Notebook:

```
http://localhost:8888/notebooks/Documents/Costa_MB/  
breast_cancer_diagnostics/jupyter_projekt.ipynb
```