eu2020.de

DESTATIS
Statistisches Bundesamt

uRos2020, 2. December 2020

Jörg Feuerhake, Kerstin Lange,
Annelen Siegismund, Elsa Vigneau

Classifying the Country of Birth

Task: What is the Country of Birth?

Paddington image source: stencil street art "Migration is not a crime" often attributed to Banksy, attribution is doubted

Items available

country of origin

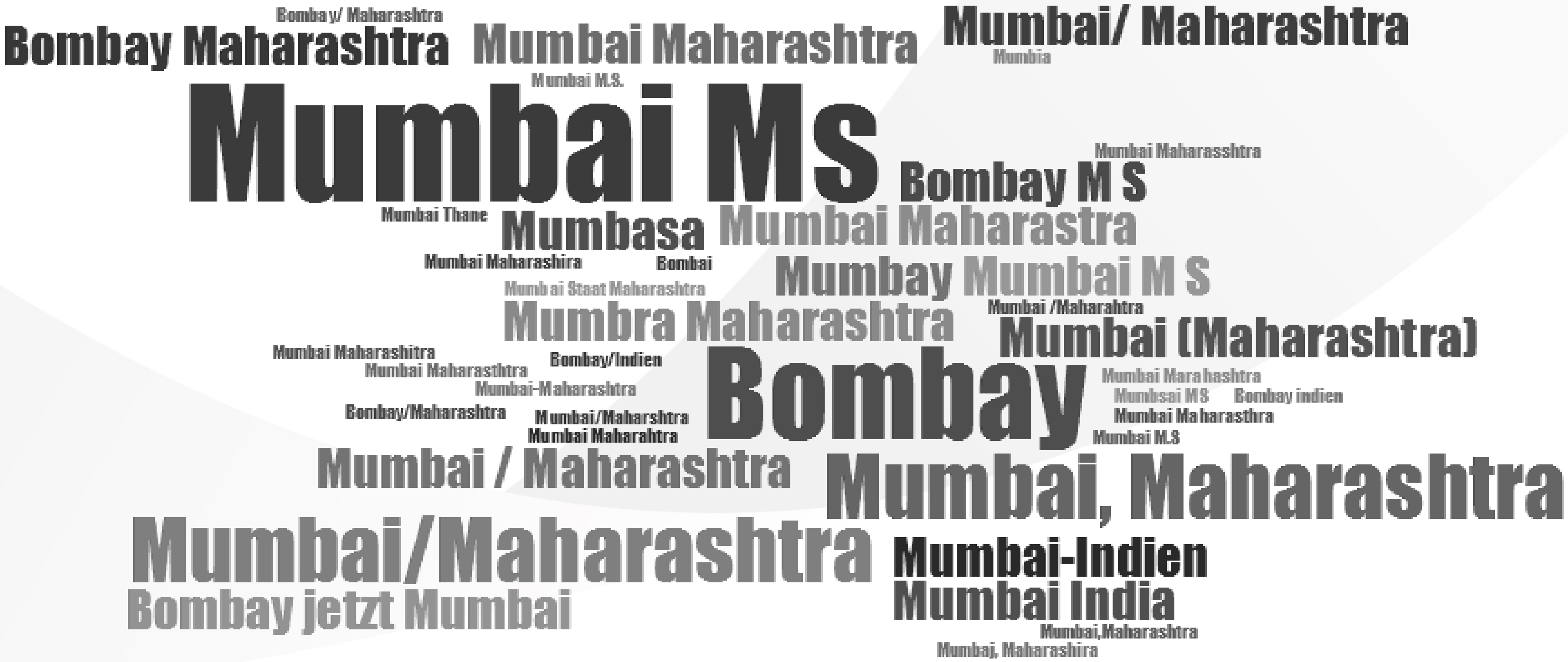
nationality

date of birth

religion

place of birth

Quality of „Place of Birth“



Federal Statistical Office of Germany (Destatis)

image source: Jason Davies Word Cloud Generator https://www.jasondavies.com/wordcloud/ cited 7.Feb.2020

Two Approaches

Lead File Based

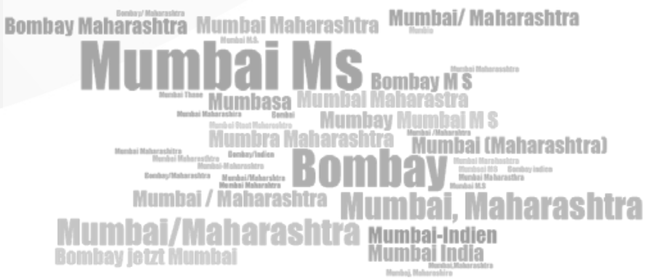
Supervised Learning

image source: © Capcom/Nintendo Street Fighter II 1993

Lead File Based Approach

👍 **Results easy to explain**

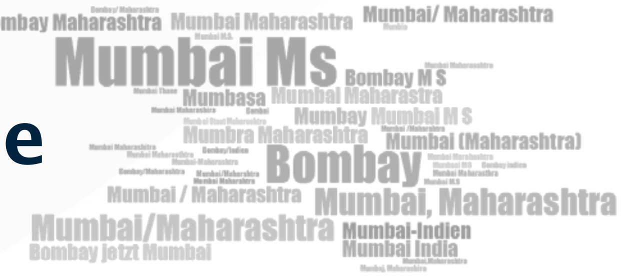
⚠️ **Lead file maintenance**



emoji source: Noto Color Emoji Pie, Apache License Version 2.0

Supervised Learning Approach

👍 **Lead file maintenance obsolete**



⚠️ **Quality of Training Data matters**

Supervised Learning Approach

N-Grams vectorise birth places:

DRESDEN → **[DR,RE,ES,SD,DE,EN]**

… tested methods:

Naïve Bayes & Random Forest

Results 2018

Country of Birth		Lead File		Supervised Learning	
Original	Imputation	Frequency	Percent	Frequency	Percent
missing	**X**	**3 513 541**	**52,3%**	**3 537 299**	**52,7%**
X	X	2 816 551	41,9%	2 769 265	41,2%
X	Y	80 074	1,2%	384 103	5,7%
X	unknown	306 253	4,6%	25 752	0,4%
		6 716 419	100,0%	6 716 419	100,0%

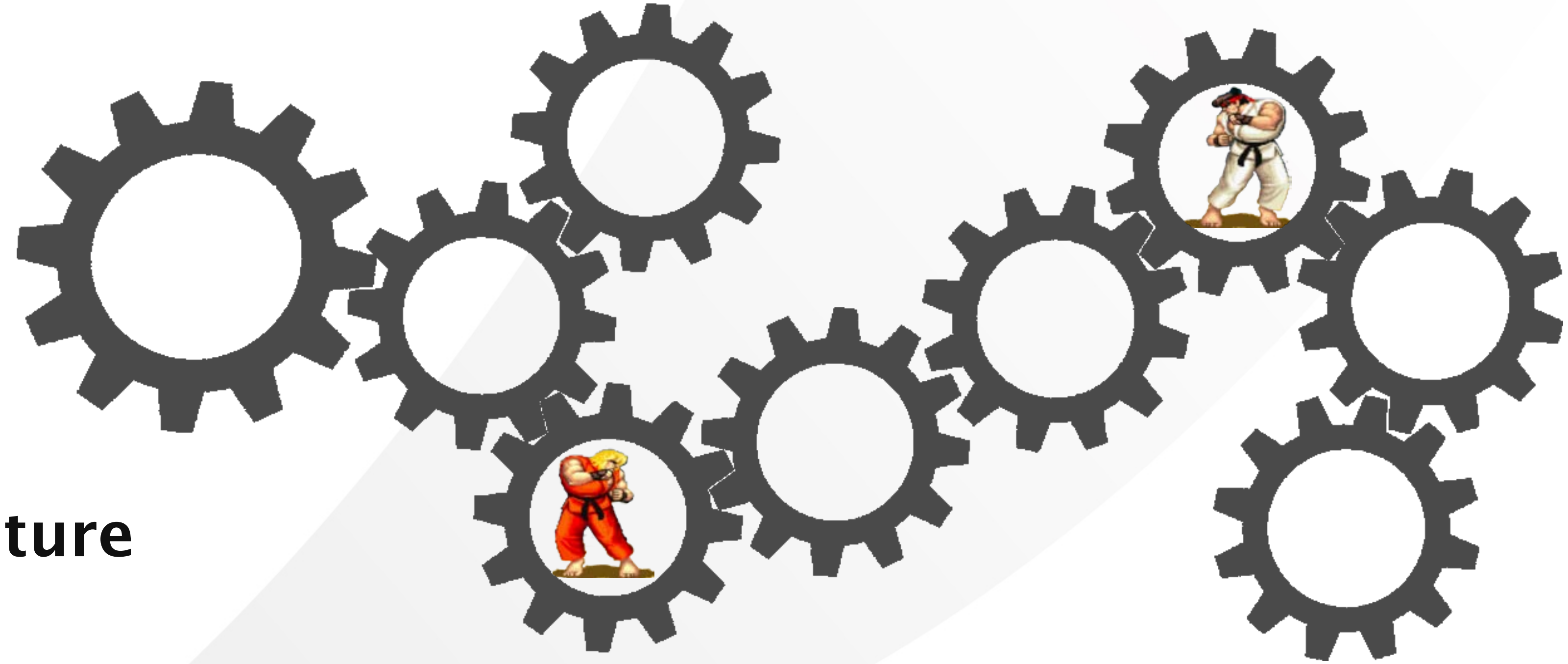
Results

Country of Birth		Lead File	Supervised Learning
Original	Imputation	Percent	Percent
missing	X	94,2%	93,9%
X	X		
X	Y	1,2%	5,7%
X	unknown	4,6%	0,4%
		100,0%	100,0%

Results

Country of Birth		Lead File	Supervised Learning
Original	Imputation	Percent	Percent
missing	X	52,3%	52,7%
X	X	41,9%	41,2%
X	Y	1,2%	5,7%
X	unknown	4,6%	0,4%
		100,0%	100,0%

System Integration



Rich client infrastructure

4-digit number of daily cases

Cases should classified automatically

Lessons Learned – Lead File Based

It works! 👍

System integration 🍰

Lead File maintenance 🧑‍🔧

Lessons Learned – Supervised Learning

It works! 👍

Lead File maintenance 🚮

System integration 🤯

Results

🥇 **Results & Ease of Integration**

🥈 **Role in Lead File Maintenance**