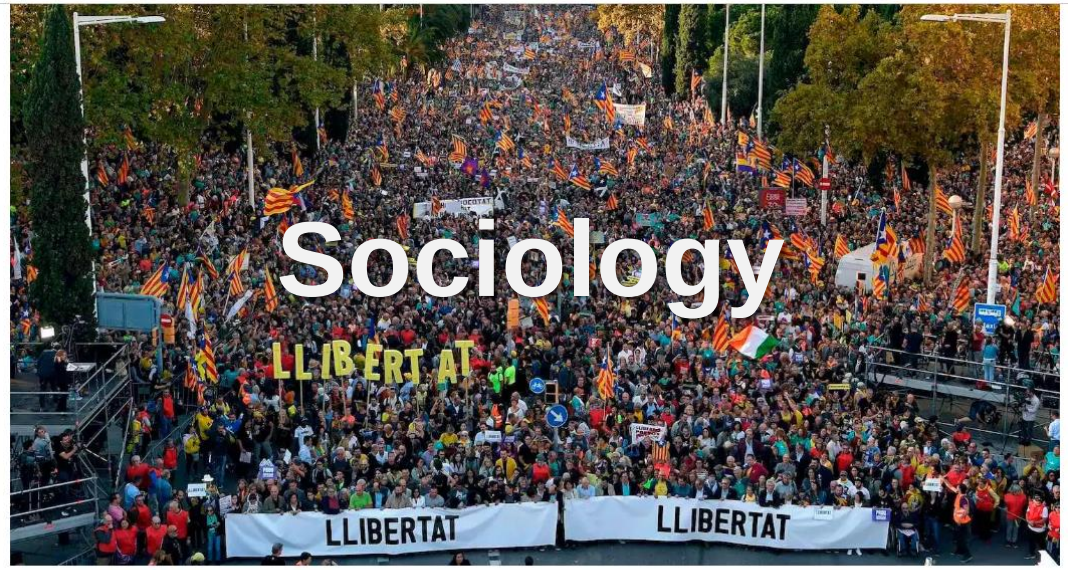
DETECT ETHNIC SPATIAL DISTRIBUTION USING SURNAME DATA

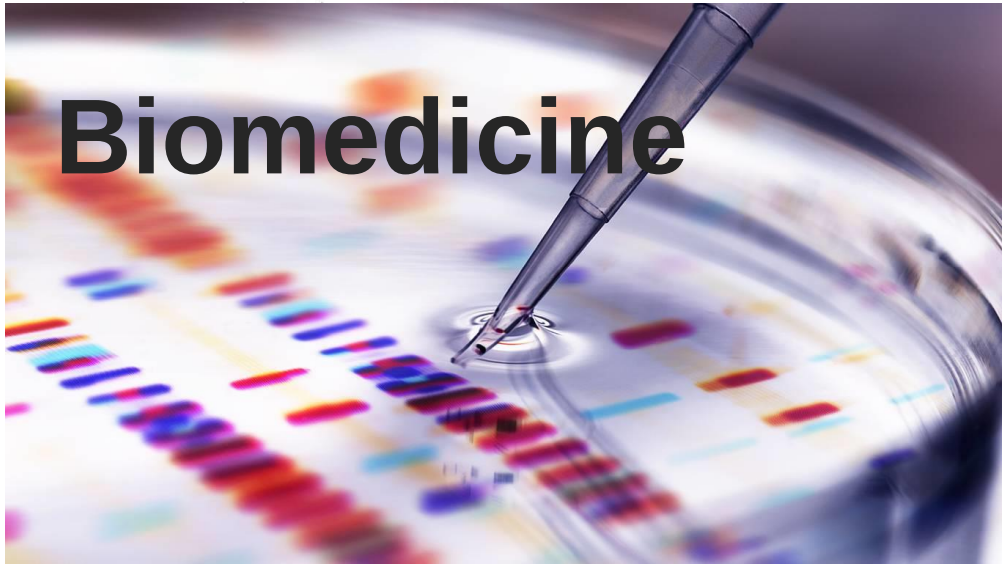
JOOMI JUN, TAKAYUKI MIZUNO

(SOKENDAI, NII)

ETHNICITY

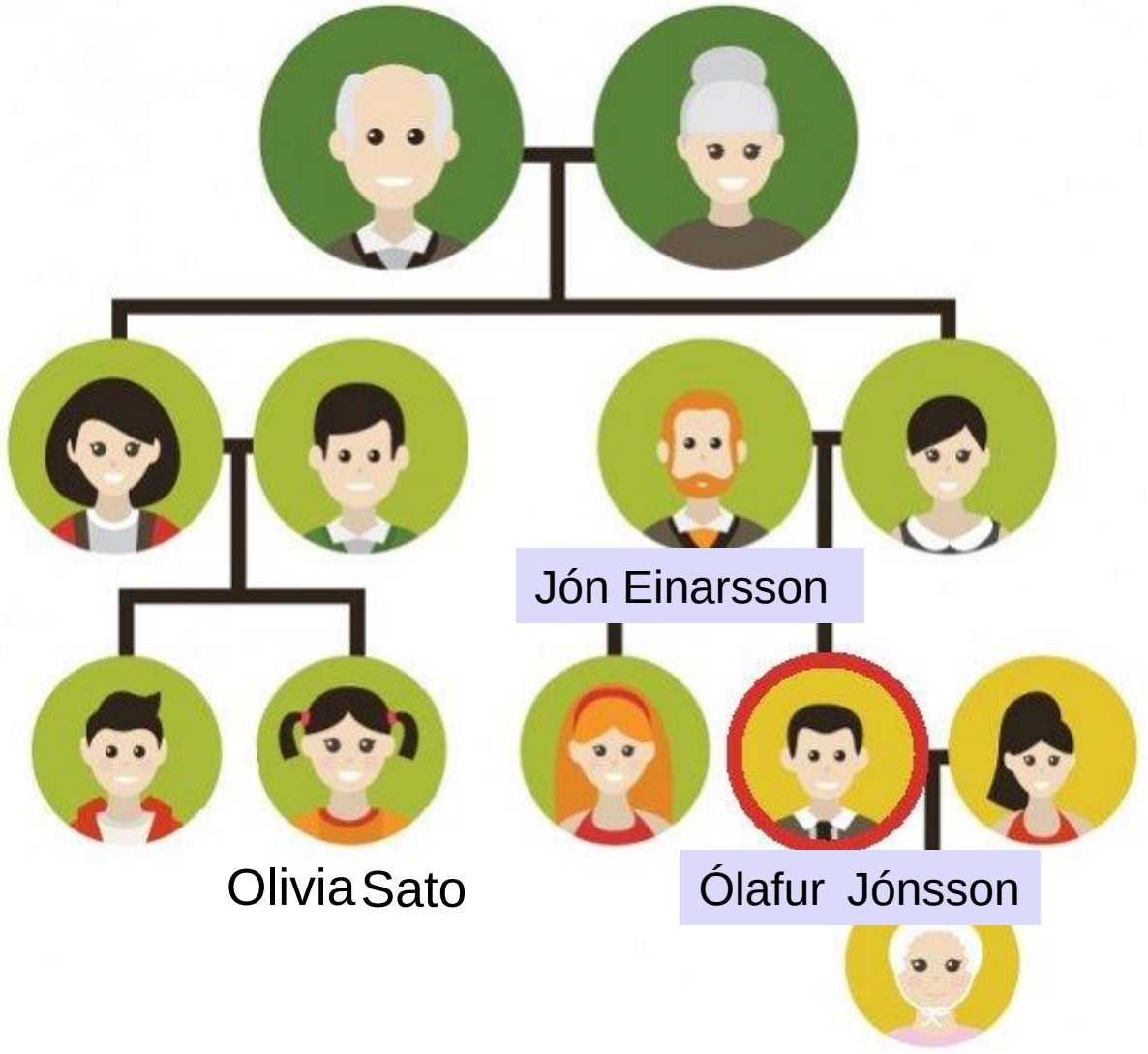


People march during a pro-independence demonstration against the conviction of Catalan separatist leaders in Barcelona on Oct. 26. LLUIS GENE/AFP VIA GETTY IMAGES



NAME DATA

- Names include information
  - First name : gender, trend of age, culture, religion
  - Surname : family roots, nationality, ethnicity



DATASET

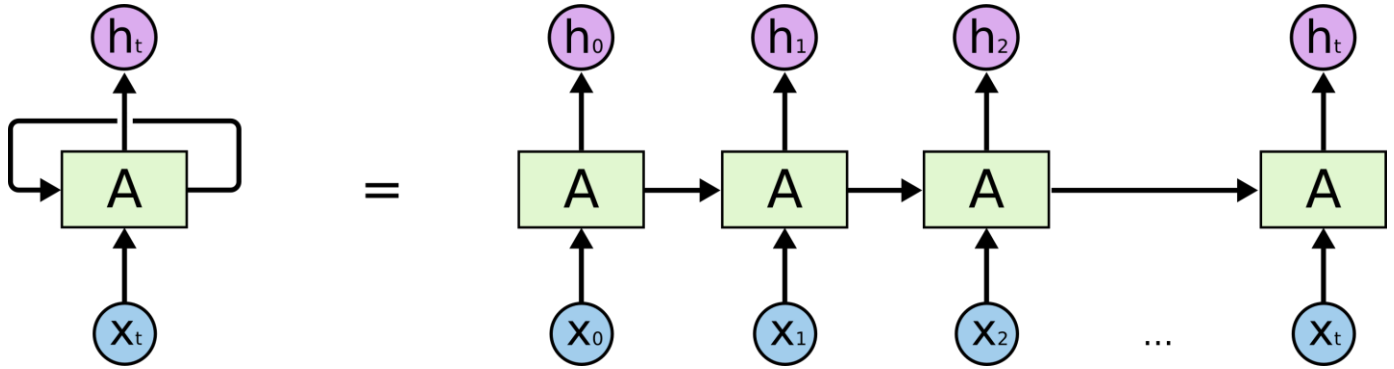
- Large scale surname dataset
- ORBIS 2016 dataset (by Bureau van Dijk)– it includes 35 million cases of the names and nationalities of company executives and individual shareholders in 203 countries

Data sample from ORBIS dataset

Title	Full Name	surname	Place of Birth(city)	Nationality	Date of Birth (year)
Mr	Alka***A	A	RIYADH	Saudi Arabia	1966
Mr	Dar*** Franceschetto	Franceschetto	BARBARANO VICENTINO	Italy	1962
Ms	Mar*** Ivascu	Ivascu	NA	Romania	1959
Ms	Kar*** Raloff	Raloff	NA	Germany	1969
Mr	Khu*** yan	yan	NA	China	1979

Surname [Raloff] - Nationality [Germany]

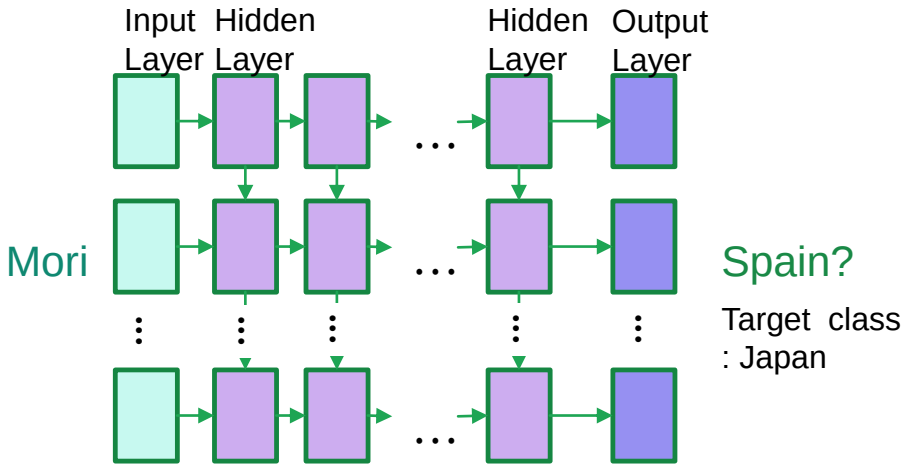
USING RNN(RECURRENT NEURAL NETWORK)



http://colah.github.io/posts/2015-08-Understanding-LSTMs/

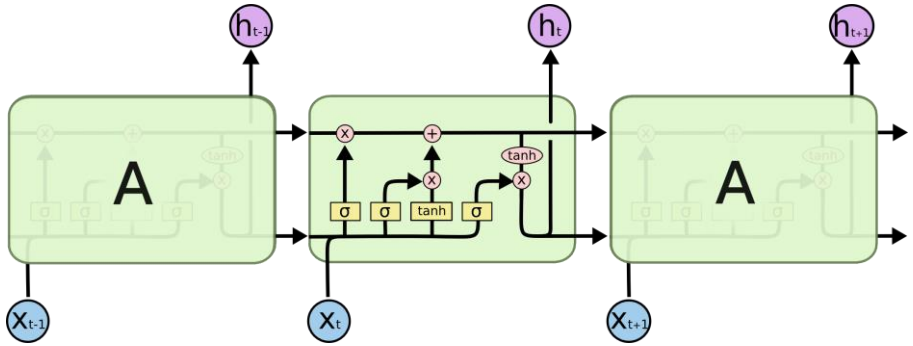
Learning
: **Mori / Japan**

Smith / UK
Zhang / China
Bellinazzi / Italy

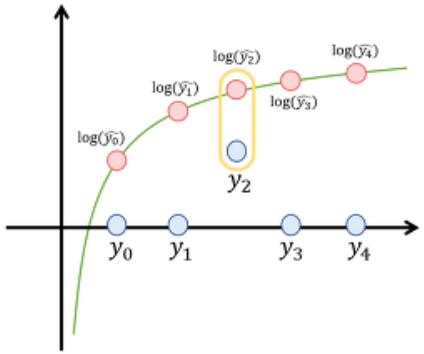


TRAINING

- 77 countries (classes)
- Using LSTMs (Long Short Term Memory) networks : to solve Vanishing Gradient problem
- Iteration 2,000,000 learning rate 0.0005, 3-layer of LSTMs



Structure of LSTMs



Class : One – hot vector

Classification problem:
Learn as decrease Cross Entropy and get parameters

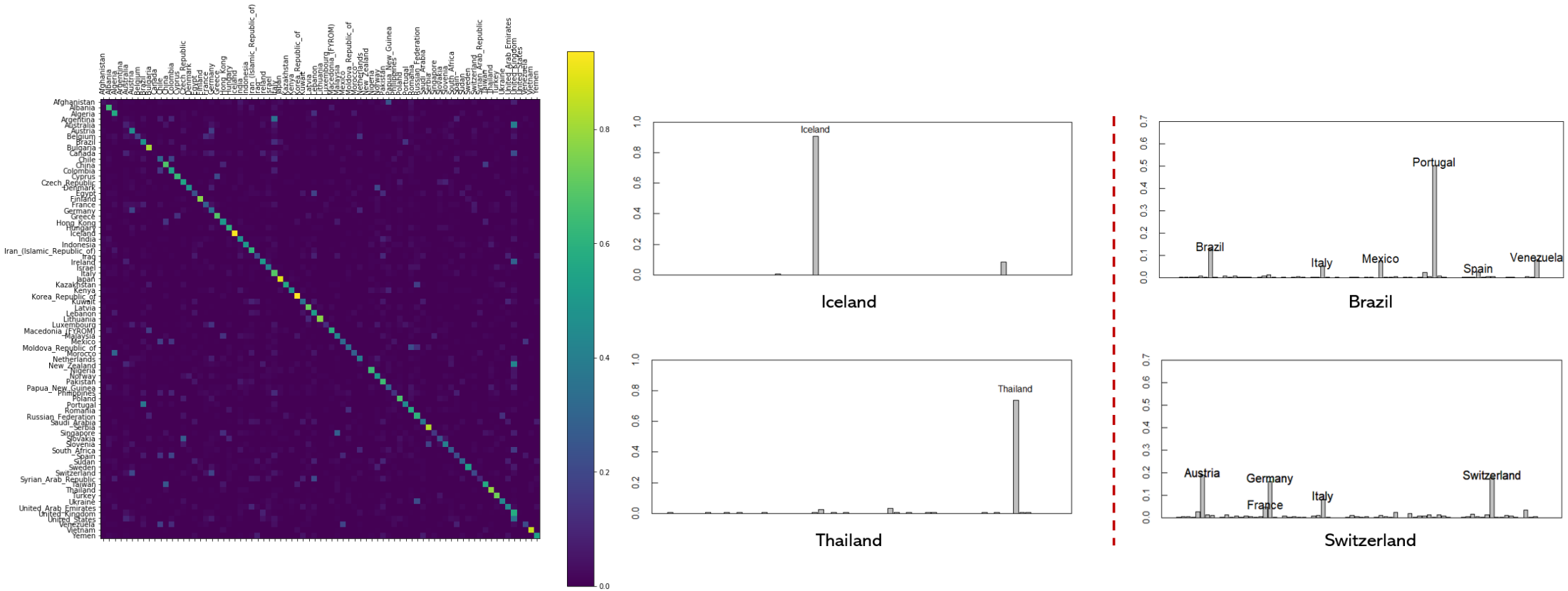
Cross Entropy

$$L(y, \hat{y}) = -\sum_{i=1}^{N} y_i log\, \hat{y}_i$$

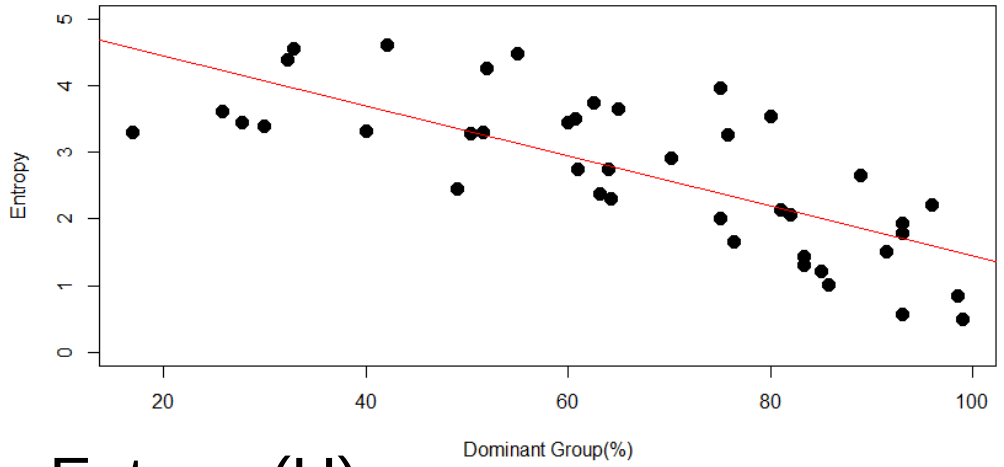
y : true distribution learned
$\hat{y}$: distribution result from classifier
N : class (country)

TEST RESULT

- Classification 77-country



ENTROPY



- Entropy(H)

$$H(X) = -\sum_{i=1}^{n} p(x_i) log_2 p(x_i)$$

- Dominant group

  percentage of the largest ethnic group

Rank	Country	Dominant Ethnic Group	%	Entropy
1	Rep. of Korea	Korean	99	0.48573
2	Iceland	Icelandic	93	0.561222
3	Japan	Japanese	98.5	0.853479
4	Vietnam	Vietnamese	85.7	1.0056
5	Bulgaria	Bulgarian	85	1.223591
		…		
38	Belgium	Flemish	52	4.254518
39	Canada	Canadian	32.3	4.38808
40	Luxembourg	Luxembourgers	55	4.486005
41	Philippines	Visayan	32.9	4.563974
42	Afghanistan	Pashtun	42.1	4.621831

(ref. Wikipedia / World Population review / the world fact book – CIA)

METHOD : SPATIAL DISTRIBUTION

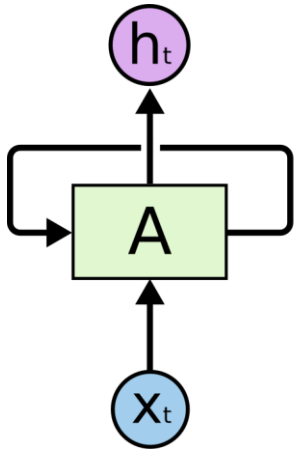
- The similarity between ethnic distributions -> visualize similarity of ethnic composition spatially
- Measure Similarity : JSD (Jensen-Shannon Divergence)
- Extract Cluster : MapEquation (based on random walk)

JSD : $D_{JS} = \frac{1}{2} D_{KL}(P||\frac{P+Q}{2}) + \frac{1}{2} D_{KL}(Q||\frac{P+Q}{2})$

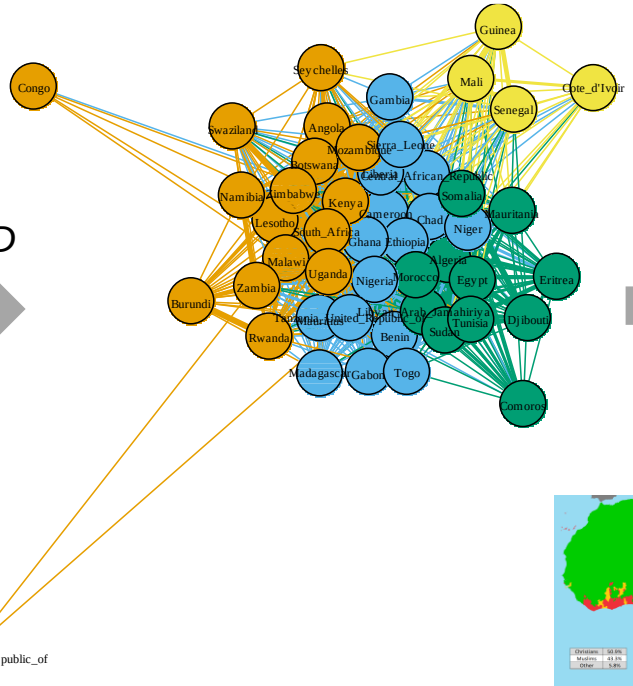
KLD : $D_{KL} = (\mathrm{P}||\mathrm{Q}) = \int_x P(x)(logP(x) - log\, Q(x))$

SPATIAL DISTRIBUTION OF AFRICA

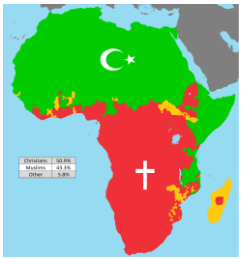
- 48-African nation



Clustered by JSD



Clustered map



Religions map

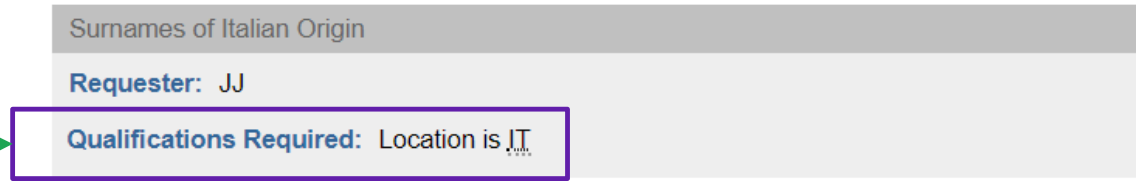
Language map

PROBLEMS

- Similar countries
  - Select unique countries as classes to identify
    - UK, USA, Australia, Ireland, New Zealand etc. → UK
  - UK, Germany, Spain, Israel, UAE, Saudi Arabia, South Africa, Zimbabwe, Italy, France, China, Russia, Japan, India (14)
- Ethnicity
  - Compared with human recognition

USING MTURK

- Set location option



If your work result is unreliable, we can reject your work.

- Question

Please Select ALL surnames of Italian Origin from the list.

- Set First 2, last 2 surnames for filtering

☐ Poli

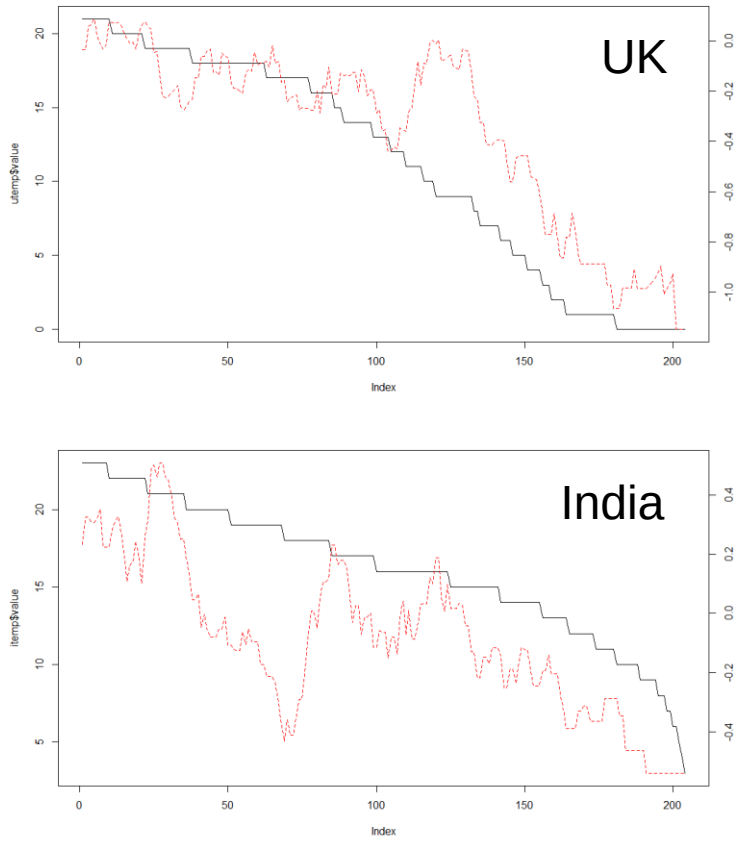
☐ Balbo

☐ Fazzari

☐ Fumante

☐ Stuppia

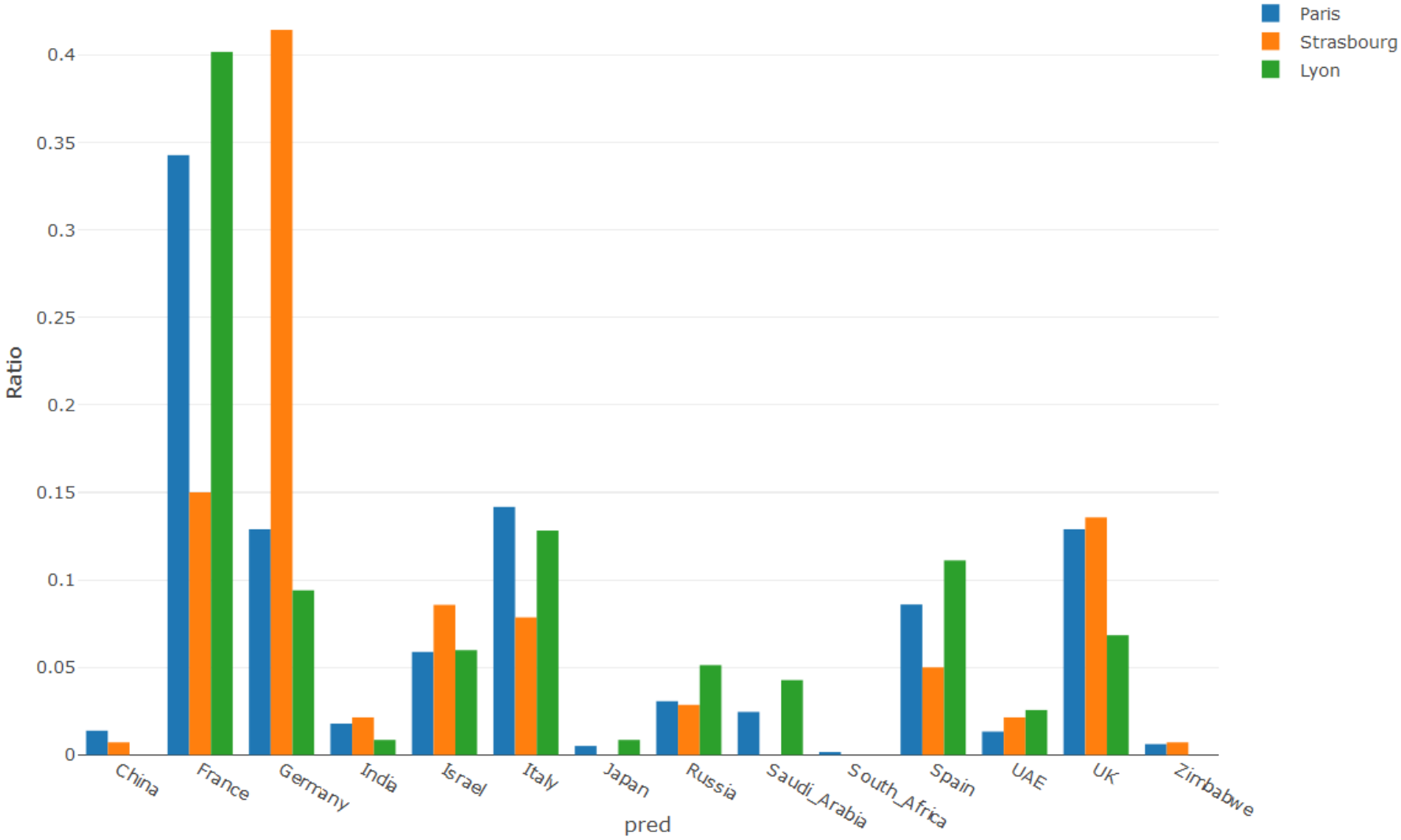
COMPARE THE RESULT



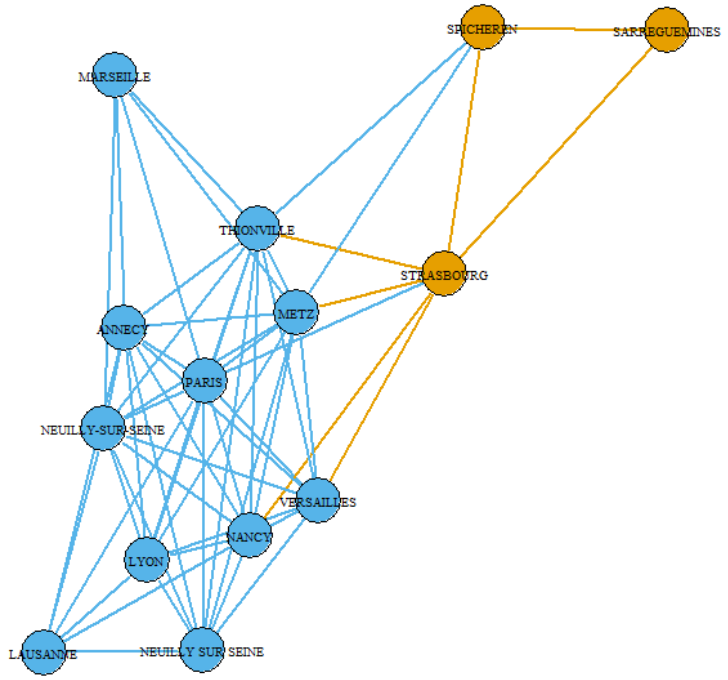
- Avg. of Top1 accuracy : 67%

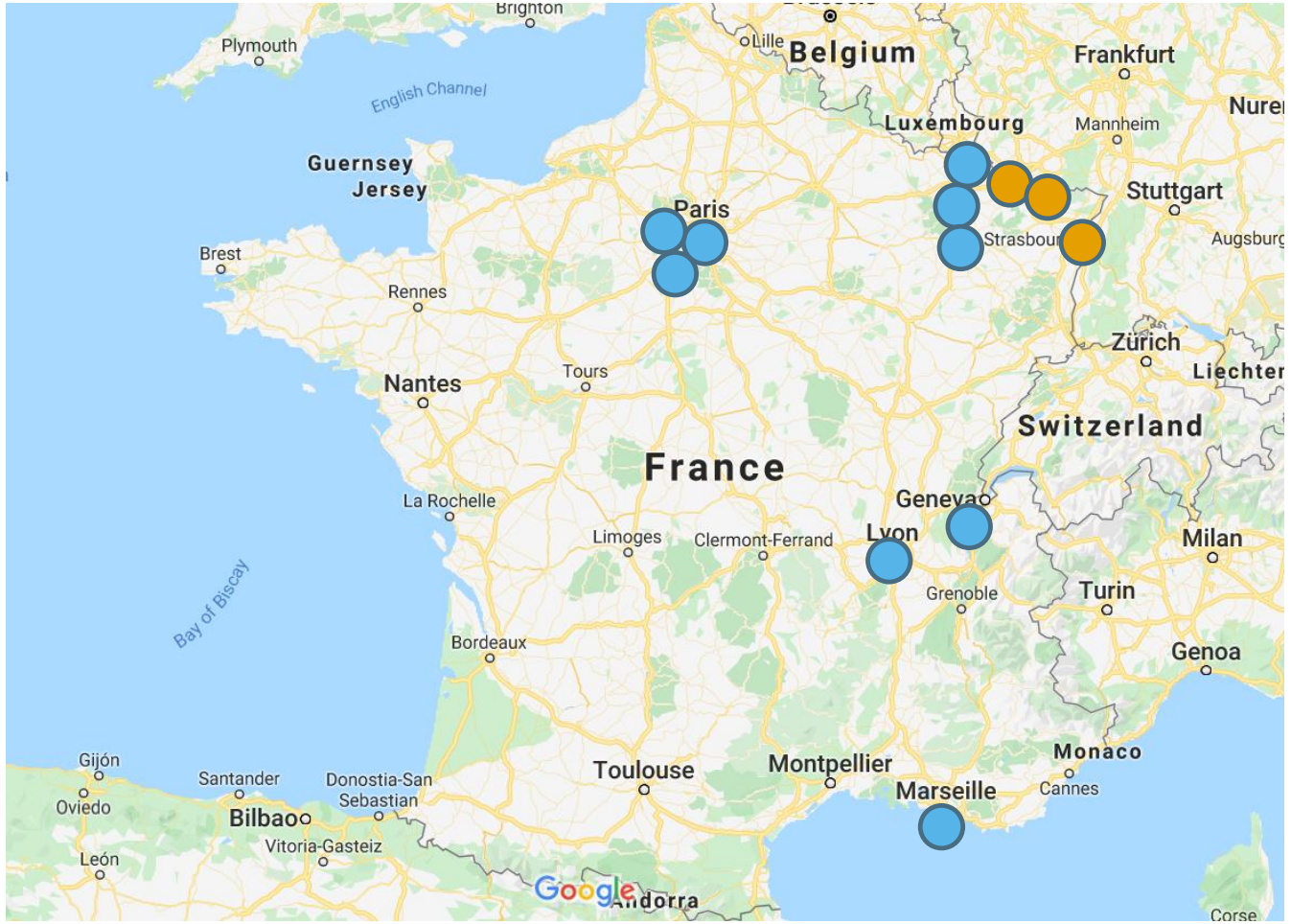
Nation	Balanced Accuracy
UK	0.60
Germany	0.68
Italy	0.70
India	0.60
France	0.76
Spain	0.72

SPATIAL DISTRIBUTION OF FRANCE

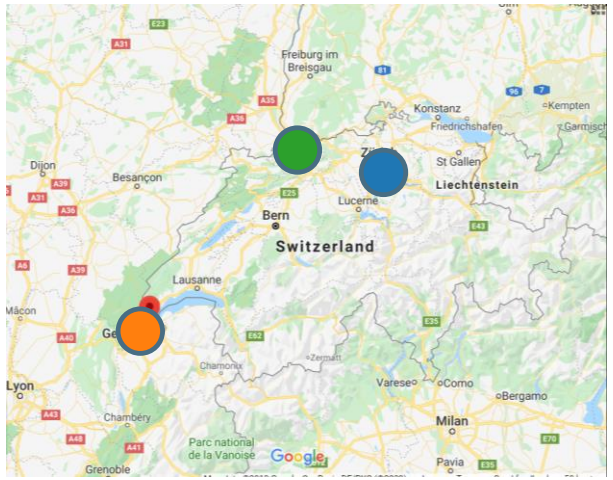


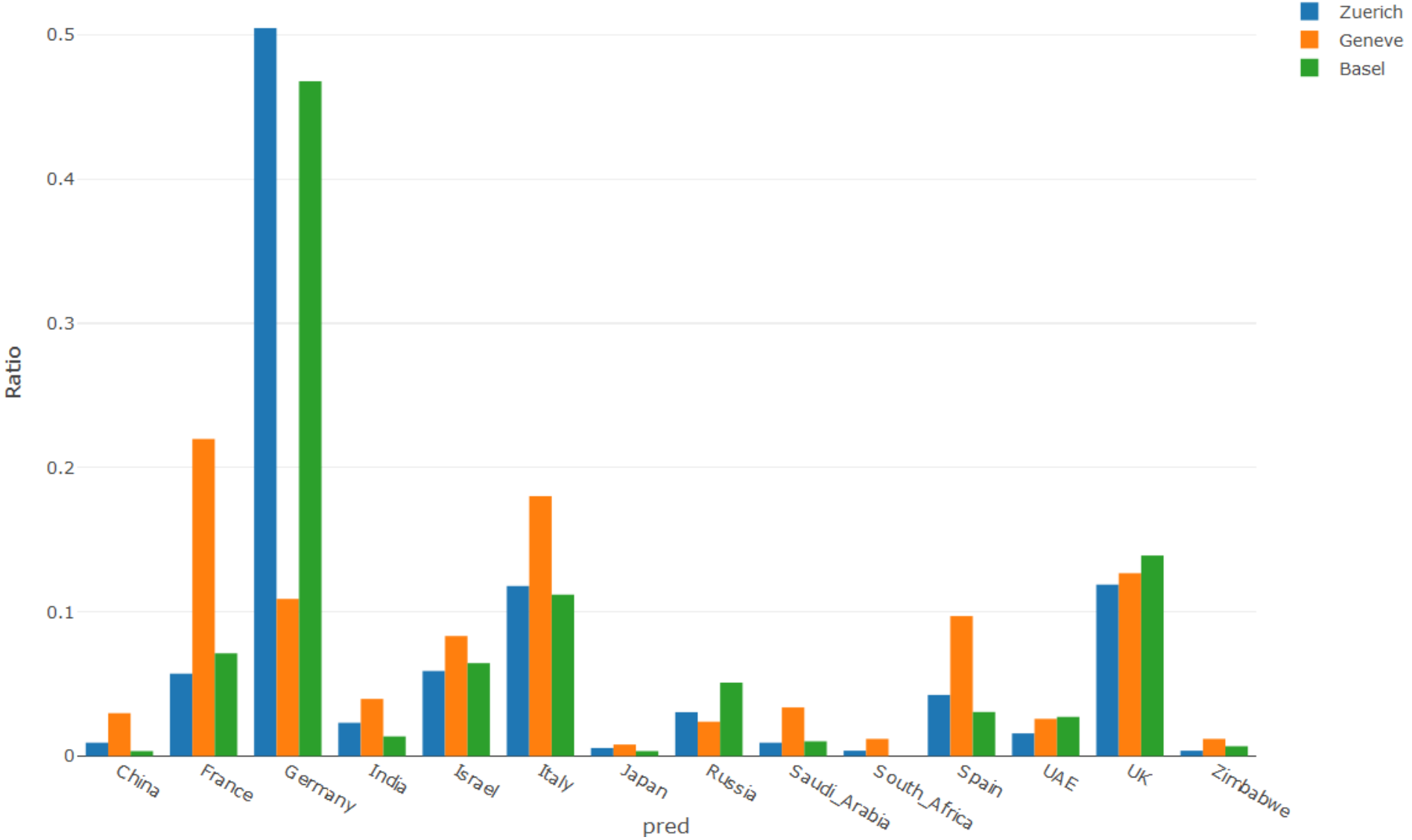
SPATIAL DISTRIBUTION OF FRANCE



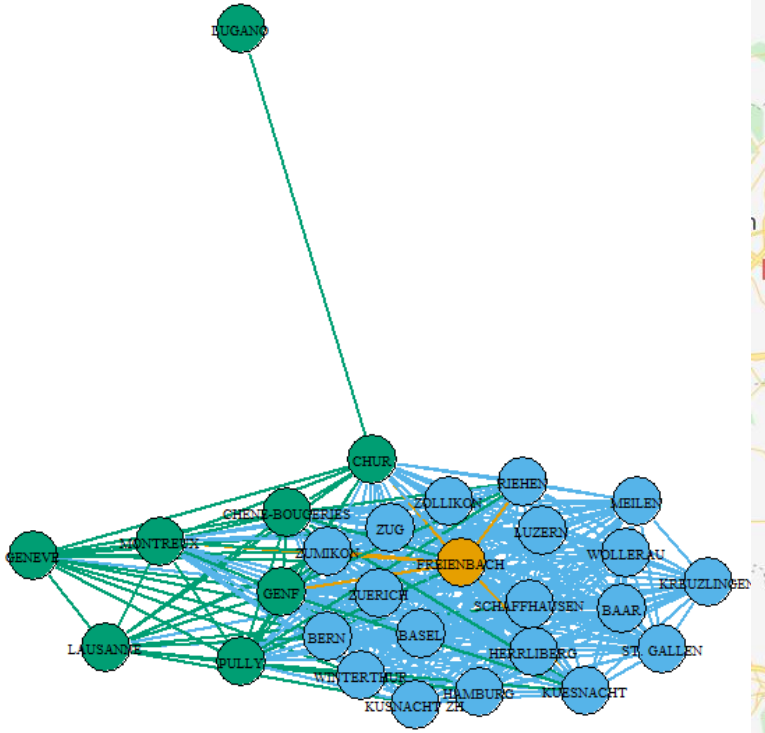


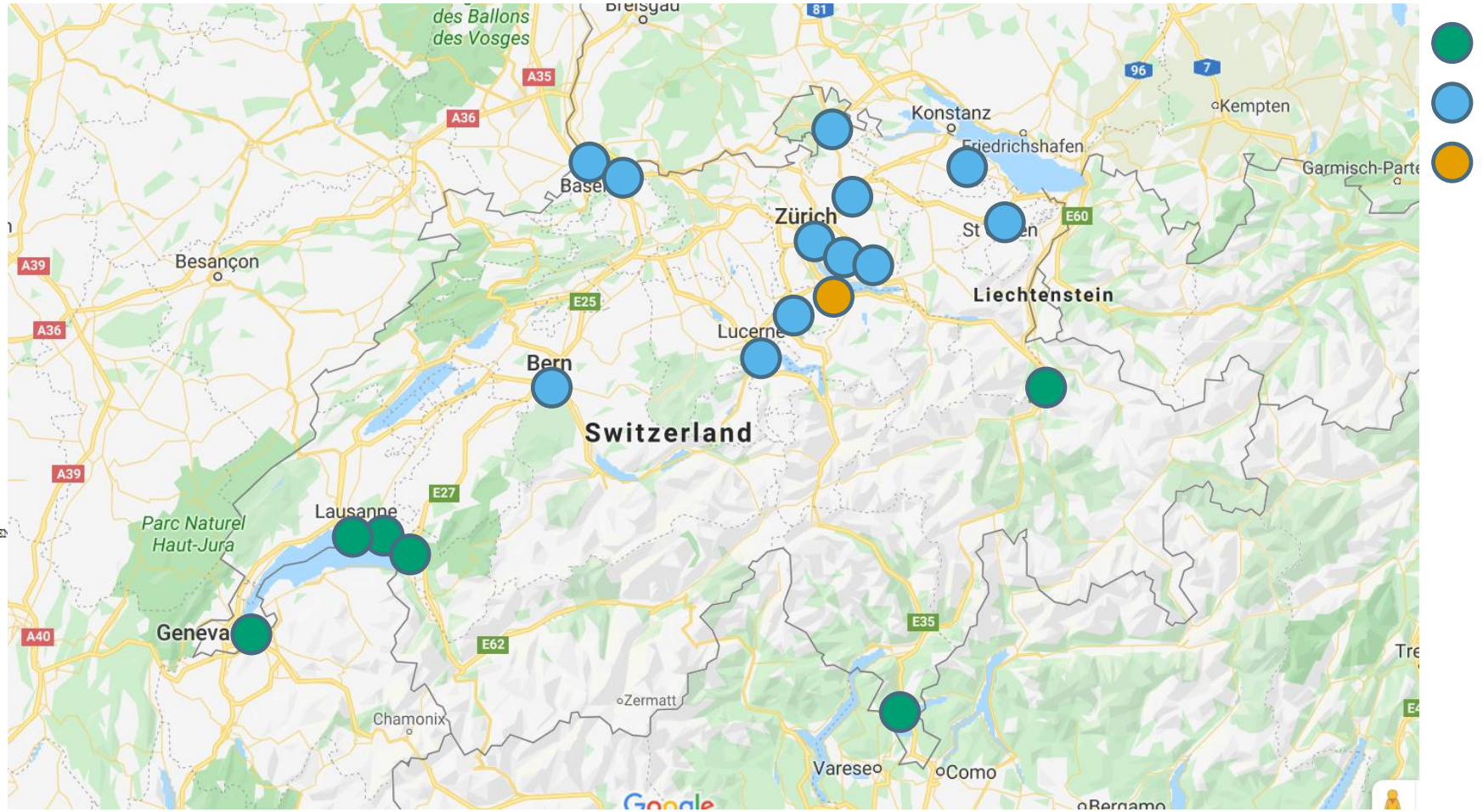
SPATIAL DISTRIBUTION OF SWITZERLAND





SPATIAL DISTRIBUTION OF SWITZERLAND



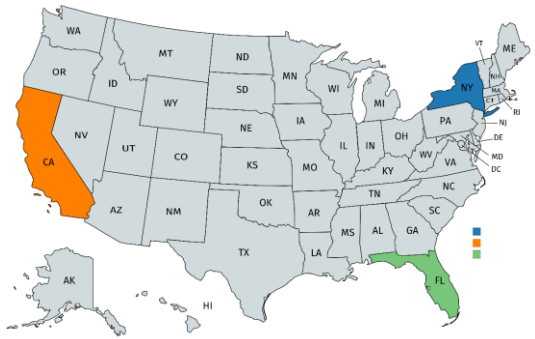


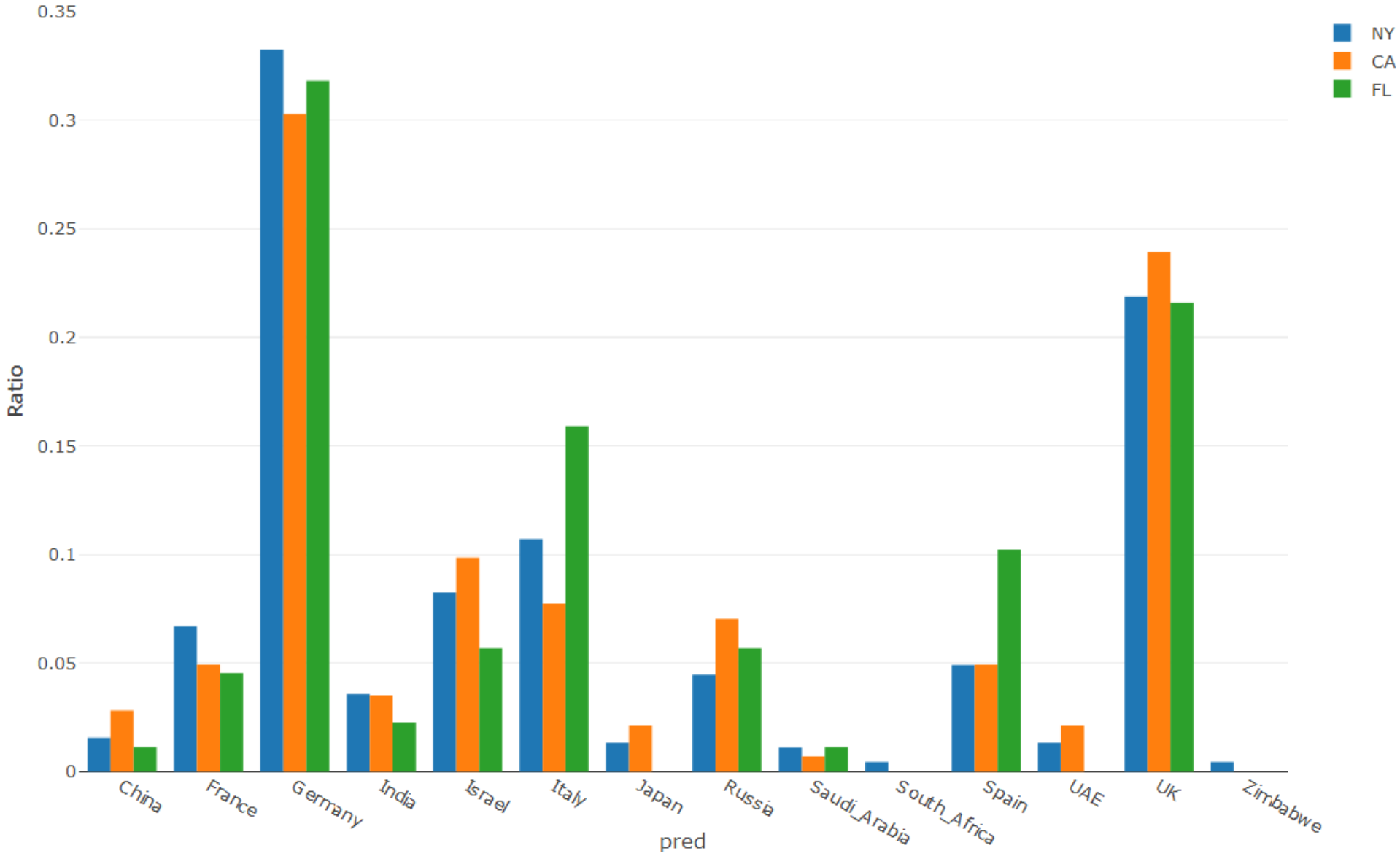
SUMMARY

- We used large-scale surname data of business people and RNN to build surname-origin of ethnicity classifier.
- We extract spatial distribution of ethnicity using an ethnic distribution that came from the classified result.

REF.

SPATIAL DISTRIBUTION OF USA





SPATIAL DISTRIBUTION OF USA