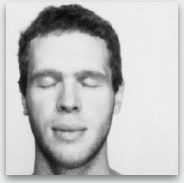
Namograph.antonomase.fr/

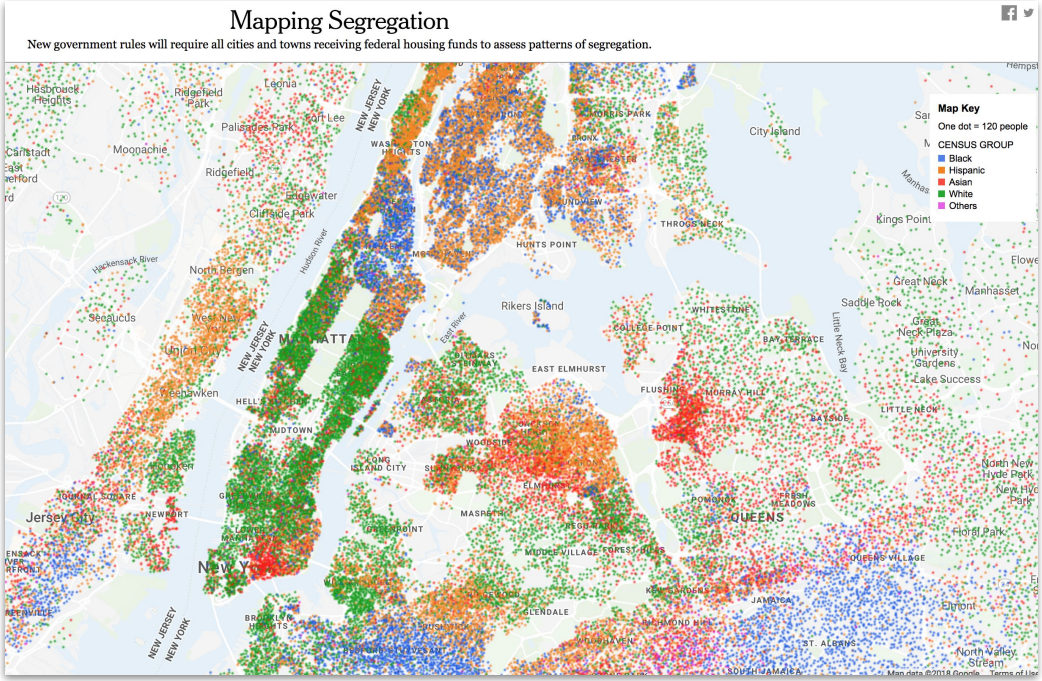
Antoine Mazieres
https://antonomase.fr/



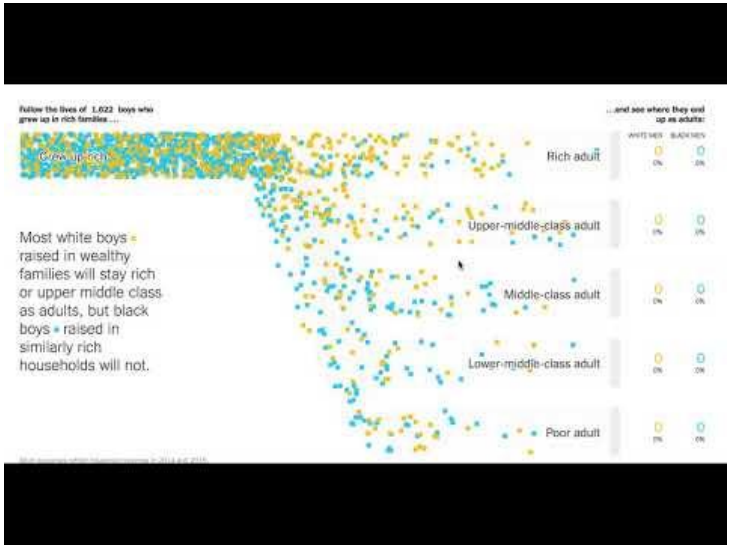
Camille Roth
https://camilleroth.eu/

Large-scale diversity estimation through surname origin inference

Ethnicity statistics and discrimination studies

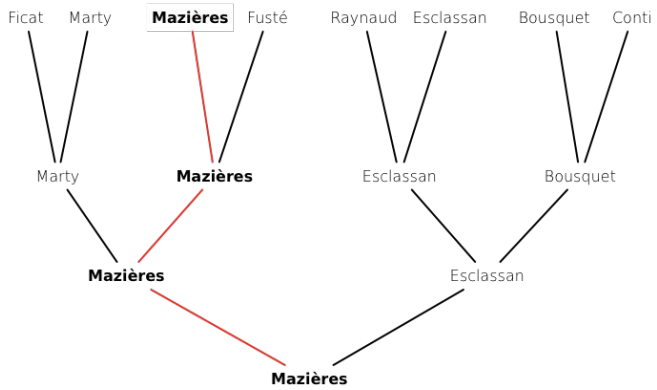


Source: https://www.nytimes.com/interactive/2015/07/08/us/census-race-map.html

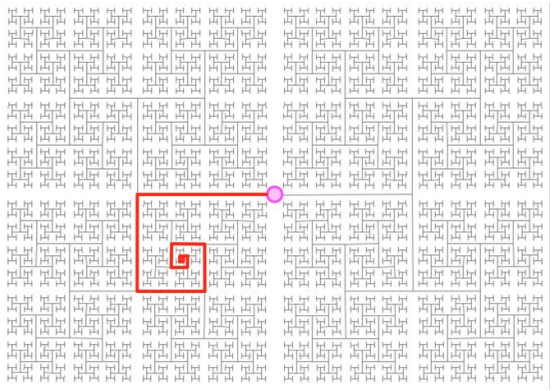


Source:
https://www.nytimes.com/interactive/2018/03/19/upshot/race-class-white-and-black-men.html

Why surnames (still) mean something ?



My surname over 3 generations



A surname over 12 generations

Endogamy makes the name

Ethnicity *versus* Origin for onomastics

Ethnicity

- Subjective
- *Plural*
- Changing
- Identified individually
- Sometimes polemical

Origin

- **Objective**
- *Plural*
- **Stable**
- **Aggregated estimations**
- **More explicit biases**

Crafting the learning data

- **Need** for **(Surname -> Origin)** tuples
- 25 millions (Author's surname - Affiliation's Country)
- Normalize and seeking "**Core Surnames**" with frequency and *Herfindahl-Hirschmann Index*: **650k** core names
- Features with ***n-grams***
- How to regroup countries ?

(Ambekar, 2009)

(Lee et al., 2017)

PubMed

(Torvik, 2016)

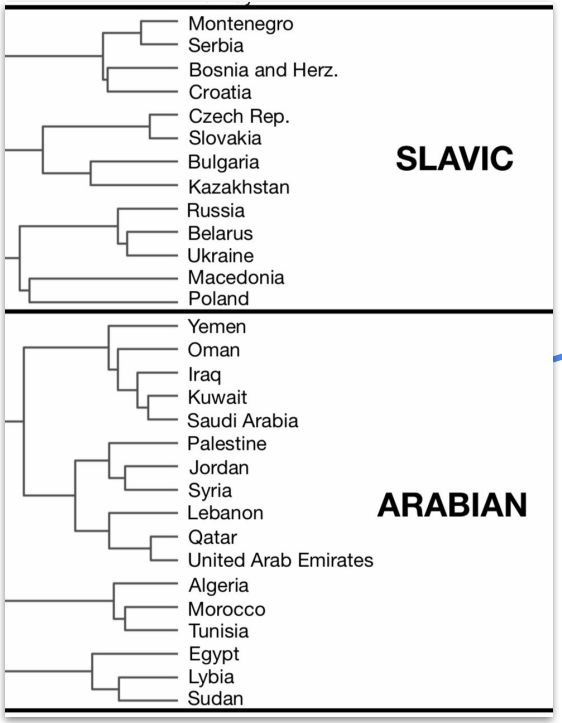
United States Census Bureau

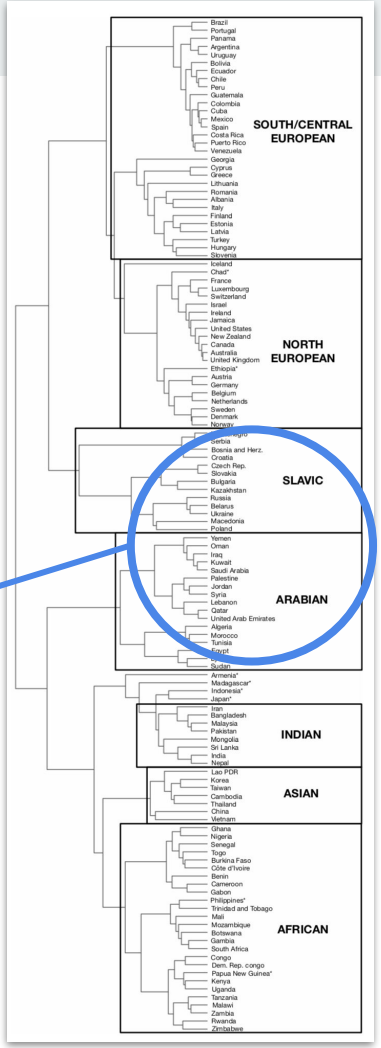
(Mislove et al., 2014)

(Mateos, 2014)

A data-driven typology of origins

- Regrouping ***n-grams*** **by countries**
- Hierarchical clustering (Ward) yields intuitive representation
- Hand-made tree cutting.
- A few irregularities





Learning a model to infer surname origin

- 650k (“core surnames” -> Origin)

  (15% for testing)
- Multinomial Naive-Bayesian

  Classifier w/ smoothing
- Heterogeneous results

Cluster	Core names		Class. Perf.	
	Total	*Evaluation*	*Precision*	*Recall*
African	30 748	4 529	0.43	0.61
Arabian	31 272	4 596	0.52	0.72
Asian	44 658	6 754	0.61	0.77
CS-European	189 624	28 668	0.81	0.71
Indian	68 145	10 067	0.63	0.72
N-European	216 465	32 469	0.78	0.62
Slavic	65 259	9 843	0.64	0.84
Total	*646 171*	*96 926*		

Model correction

$$P(\text{actual} = i | \text{guessed} = j) = C_{ij} / \sum_{j'} C_{ij'}$$

Guessed origin	Actual origin						
	Afr.	Arab.	Asian	CSE	Indian	NE	Slavic
African	**2763**	165	381	1081	460	1441	157
%	***61.0***	*3.59*	*5.64*	*3.77*	*4.60*	*4.44*	*1.60*
Arabian	159	**3292**	84	577	598	1549	77
%	*3.51*	***71.6***	*1.24*	*2.01*	*5.94*	*4.77*	*0.78*
Asian	319	113	**5200**	831	716	1147	174
%	*7.04*	*2.46*	***77.0***	*2.90*	*7.11*	*3.53*	*1.77*
CS-Eur.	258	128	274	**20364**	299	3535	324
%	*5.70*	*2.79*	*4.06*	***71.0***	*2.97*	*10.9*	*3.29*
Indian	273	487	420	991	**7226**	1862	191
%	*6.03*	*10.6*	*6.22*	*3.46*	***71.8***	*5.73*	*1.94*
N-Eur.	643	351	315	3254	609	**20183**	670
%	*14.2*	*7.64*	*4.66*	*11.4*	*6.05*	***62.2***	*6.81*
Slavic	114	60	80	1570	159	2752	**8250**
%	*2.52*	*1.31*	*1.18*	*5.48*	*1.58*	*8.48*	***83.8***
	4529	4596	6754	28668	10067	32469	9843
%	*100*	*100*	*100*	*100*	*100*	*100*	*100*

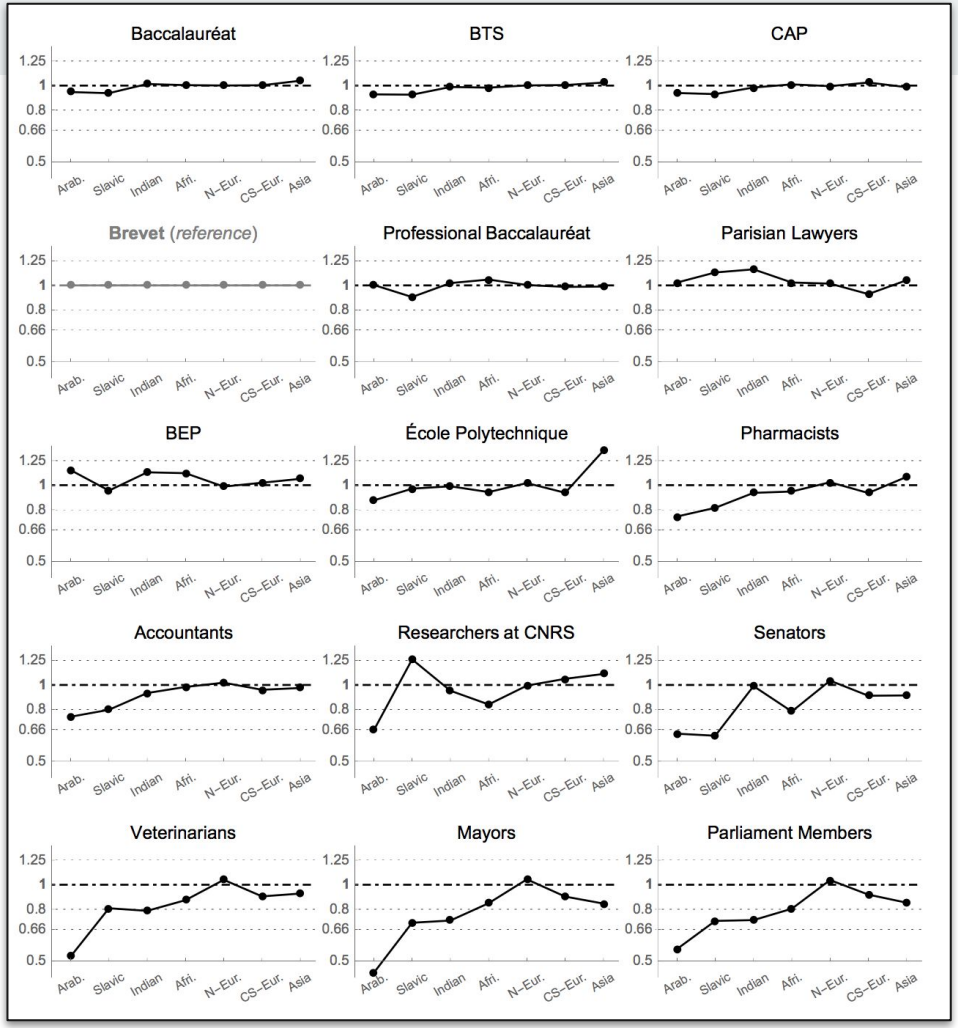
Measuring diversity

- Why **diversity** instead of **discrimination** ?
- Simple method: **ratio** between target and reference dataset
- ***Brevet des collèges*** as reference
- Targets:

Name	List of surnames of all …	nb. obs.
Brevet	Candidates to *Diplôme National du Brevet* in 2008 [5]	562,952
Baccalauréat	Candidates to the nationwide *Baccalauréat* (*Général* and *Technologique*) in 2008	435,645
BEP	Candidates to *Brevets d'Études Professionnelles* in 2008	116,814
CAP	Candidates to *Certificats d'Aptitude Professionnelle* in 2008	98,364
BTS	Candidates to *Brevets de Technicien Supérieur* in 2008	87,917
Professional Baccalauréat	Candidates to *Baccalauréats Professionnels* in 2008	80,672
Pharmacists	Pharmacists registered in their *Ordre Professionnel* in 2017 [6]	73,422
Mayors	Mayors of French cities ("communes") in 2014 [7]	36,628
Parisian Lawyers	Lawyers registered in the Parisian Bar Association in 2017 [8]	32,021
École Polytechnique	Students at *École Polytechnique* (1958-2016) [9]	23,058
Accountants	Accountants registered in their *Ordre Professionnel* in 2017 [10]	20,946
Veterinarians	Veterinary physicians registered in their *Ordre Professionnel* in 2017 [11]	15,710
Researchers	Researchers at *Centre National de la Recherche Scientifique* in 2017 [12]	12,657
Parliament Members	Parliament Members of *Assemblée Nationale* (1958-2016) [13]	8,326

Results

- Plots and origins ordered by proximity
- Clusters of types of occupational groups (political functions, state exams) and particular profiles
- Certain comparison with discrimination studies results



Questions

Advertisement

Project site: https://namograph.antonomase.fr/

Get in touch: {mazieres,roth}@cmb.hu-berlin.de