Matthias Niggli
Center for International Economics and Business CIEB
University of Basel

University of Basel
Center for International Economics and Business | CIEB

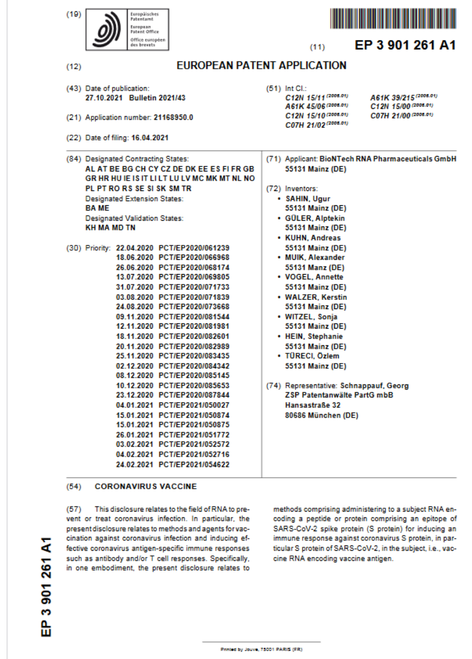
Moving On – Investigating Inventors' Ethnic Origins Using Supervised Learning

The Problem

- Patent data provides rich information about technical inventions and is frequently used in social sciences research.
- Unfortunately, it does **not disclose inventors' ethnic origins.**
- Current approaches use costly commercial tools to map inventors to ethnicities.
- Particular challenge for research at the intersection of immigration and innovation topics

Contributions

- Construct a new dataset of names that is representative for inventor origin analysis
- Train and validate an algorithm that predicts ethnic origins based on names
- Provide estimates for inventor ethnic origin compositions over time, across countries and technological areas.



(19) Europäisches Patentamt / European Patent Office / Office européen des brevets

(11) EP 3 901 261 A1

(12) EUROPEAN PATENT APPLICATION

(43) Date of publication: 27.10.2021 Bulletin 2021/43

(21) Application number: 21168950.0

(22) Date of filing: 16.04.2021

(51) Int Cl.: C12N 15/11 (2006.01) A61K 39/215 (2006.01) A61K 45/06 (2006.01) C12N 15/00 (2006.01) C12N 15/10 (2006.01) C07H 21/00 (2006.01) C07H 21/02 (2006.01)

(84) Designated Contracting States: AL AT BE BG CH CY CZ DE DK EE ES FI FR GB GR HR HU IE IS IT LI LT LU LV MC MK MT NL NO PL PT RO RS SE SI SK SM TR
Designated Extension States: BA ME
Designated Validation States: KH MA MD TN

(30) Priority: 22.04.2020 PCT/EP2020/061239; 18.06.2020 PCT/EP2020/066968; 26.06.2020 PCT/EP2020/068174; 13.07.2020 PCT/EP2020/069805; 31.07.2020 PCT/EP2020/071733; 03.08.2020 PCT/EP2020/071839; 24.08.2020 PCT/EP2020/073668; 09.11.2020 PCT/EP2020/081544; 12.11.2020 PCT/EP2020/081981; 18.11.2020 PCT/EP2020/082601; 20.11.2020 PCT/EP2020/082989; 25.11.2020 PCT/EP2020/083435; 02.12.2020 PCT/EP2020/084342; 08.12.2020 PCT/EP2020/085145; 10.12.2020 PCT/EP2020/085653; 23.12.2020 PCT/EP2020/087844; 04.01.2021 PCT/EP2021/050027; 15.01.2021 PCT/EP2021/050874; 15.01.2021 PCT/EP2021/050875; 26.01.2021 PCT/EP2021/051772; 03.02.2021 PCT/EP2021/052572; 04.02.2021 PCT/EP2021/052716; 24.02.2021 PCT/EP2021/054622

(71) Applicant: BioNTech RNA Pharmaceuticals GmbH
55131 Mainz (DE)

(72) Inventors:
- SAHIN, Ugur
  55131 Mainz (DE)
- GÜLER, Alptekin
  55131 Mainz (DE)
- KUHN, Andreas
  55131 Mainz (DE)
- MUIK, Alexander
  55131 Manz (DE)
- VOGEL, Annette
  55131 Mainz (DE)
- WALZER, Kerstin
  55131 Mainz (DE)
- WITZEL, Sonja
  55131 Mainz (DE)
- HEIN, Stephanie
  55131 Mainz (DE)
- TÜRECI, Özlem
  55131 Mainz (DE)

(74) Representative: Schnappauf, Georg
ZSP Patentanwälte PartG mbB
Hansastraße 32
80686 München (DE)

(54) CORONAVIRUS VACCINE

(57) This disclosure relates to the field of RNA to prevent or treat coronavirus infection. In particular, the present disclosure relates to methods and agents for vaccination against coronavirus infection and inducing effective coronavirus antigen-specific immune responses such as antibody and/or T cell responses. Specifically, in one embodiment, the present disclosure relates to methods comprising administering to a subject RNA encoding a peptide or protein comprising an epitope of SARS-CoV-2 spike protein (S protein) for inducing an immune response against coronavirus S protein, in particular S protein of SARS-CoV-2, in the subject, i.e., vaccine RNA encoding vaccine antigen.

EP 3 901 261 A1

Printed by Jouve, 75001 PARIS (FR)

Dataset: Olympic Athletes' Names

- 135'571 full names of athletes who have been starters for 191 different national teams in Olympic games from Athens 1896 to Rio 2016.
- Label ethnic origins of these names by mapping national teams to ethnic origins.
- Clean and enrich with inventor names to a final dataset of 95'202 labelled names.

➢ Use the letters a name consists of to predict the name's ethnic origin with an LSTM
➢ 91% F1-score across 17 ethnic origins

Table 1: Taxonomy of Ethnic Origins and National Teams

Ethnic Origin	National Teams / Countries
Anglo-Saxon	Great Britain, Ireland
Chinese	China
French	France
German	Germany
Hispanic-Iberian	Spain, Portugal, Mexico
India	India
Italian	Italy
Japanese	Japan
Korean	Korea
Arabic	Egypt, Syria, Saudi Arabia, Jordan, UAE, Tunisia, Algeria, Morocco
Persian	Iran
Slavic-Russian	Russia, Ukraine, Belarus
East-Europe	Poland, Czechoslovakia, Hungary
Balkans	Serbia, Croatia, Yugoslavia
Scandinavian	Sweden, Norway, Finland, Denmark, Iceland
South-East Asia	Vietnam, Thailand, Malaysia, Indonesia, Laos, Cambodia
Turkish	Turkey

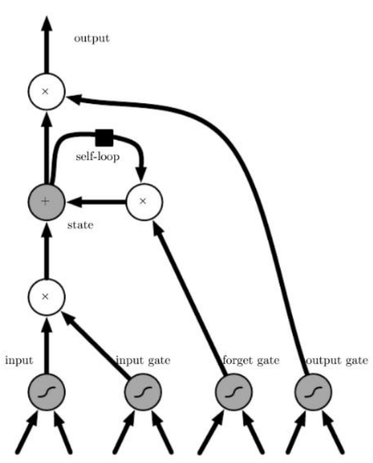
Name Transformation

Figure D.1: Example of the Encoded Name Mahatma Gandhi

	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x	y	z	END
0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
3	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
6	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0
8	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
9	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
10	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
11	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
12	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
13	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
14	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
15	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
16	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
17	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
18	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
19	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
20	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
21	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
22	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
23	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
24	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
25	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
26	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
27	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
28	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
29	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0

LSTM Network

Figure D.2: Visualization of an LSTM cell (taken from Goodfellow et al., 2016)



Performance

Table 2: Performance of the LSTM Classification Model

Ethnic Origin	Precision	Recall	F1 Score
Overall (weighted)	**0.910**	**0.910**	**0.910**
Anglo-Saxon	0.859	0.880	0.873
Arabic	0.911	0.927	0.919
Balkans	0.816	0.753	0.783
Chinese	0.932	0.938	0.938
East-Europe	0.908	0.913	0.910
French	0.911	0.893	0.902
German	0.805	0.860	0.832
Hispanic-Iberian	0.908	0.925	0.916
India	0.910	0.852	0.880
Italian	0.955	0.901	0.927
Japanese	0.972	0.993	0.982
Korean	0.925	0.955	0.940
Persian	0.897	0.902	0.900
Scandinavian	0.919	0.895	0.907
Slavic-Russian	0.961	0.956	0.958
South-East Asia	0.871	0.758	0.810
Turkish	0.923	0.966	0.944

Inventors' Ethnic Origins

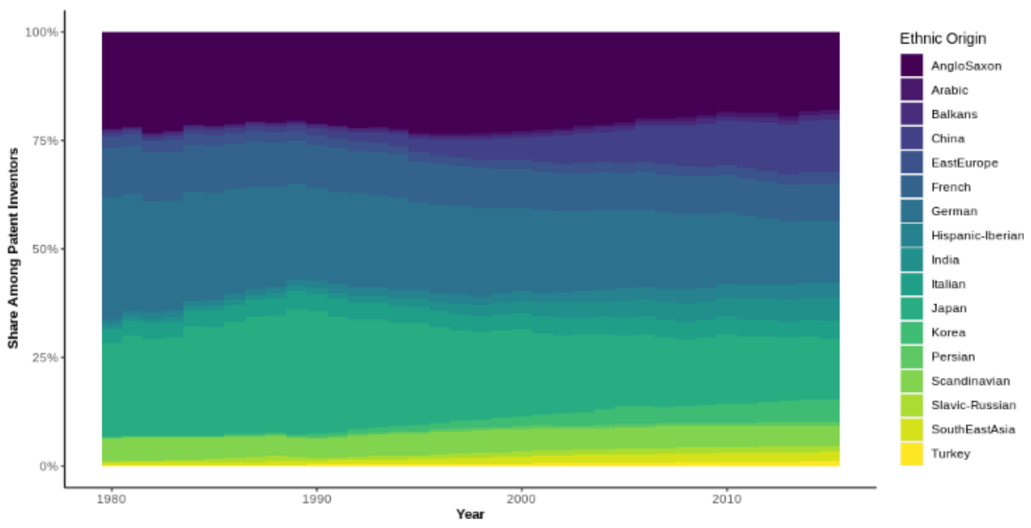
- Predict the ethnic origins of 2.68 million patent inventors from 1980 to 2015
- Global ethnic origin distribution has become more diverse. Mostly due to more Asian origin inventors (Chinese, Indian)
- Prevalence of foreign-origin inventors is especially high in the USA but has also increased in other high-income economies.
- Non-Western origin inventors are to be found mostly in the USA but not to European countries.

The USA has an extraordinary ability to attract global talents

vs.

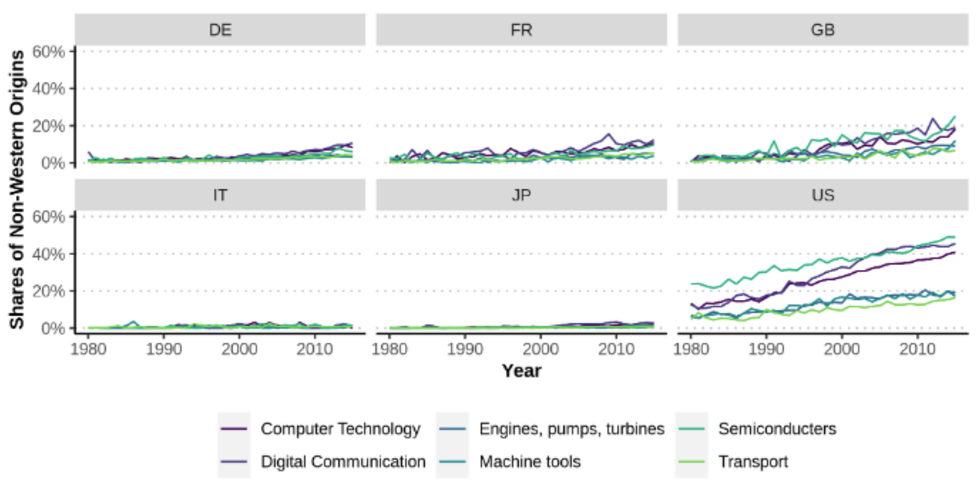
Continental European countries are lagging behind

Figure 1: The Global Ethnic Origin Composition of Inventors (1980-2015)



Notes: The graph shows the evolution of the prevalence of 17 ethnic origins among the annual stock of patent inventors between 1980 and 2015. Formally, it plots $\hat{\pi}^k_{j,t}$. The data for this plot is from the OECD and the USPTO.

Figure 3: Prevalence of Non-Western Ethnic Origin Across Technologies (1980-2015)



Notes: The graph shows the evolution of the aggregate prevalence of non-western ethnic origins among the annual stock of patent inventors in six technology fields for six high-income economies. Formally, it plots $\sum_{k=1}^{K} \hat{\pi}^k_{j,t}$, for each technology field j, with k corresponding to the following non-western ethnic origins: Arabic, Chinese, Indian, Persian, Slavic-Russian, Turkish and South-East Asian. The data for this plot is from the OECD and the USPTO.