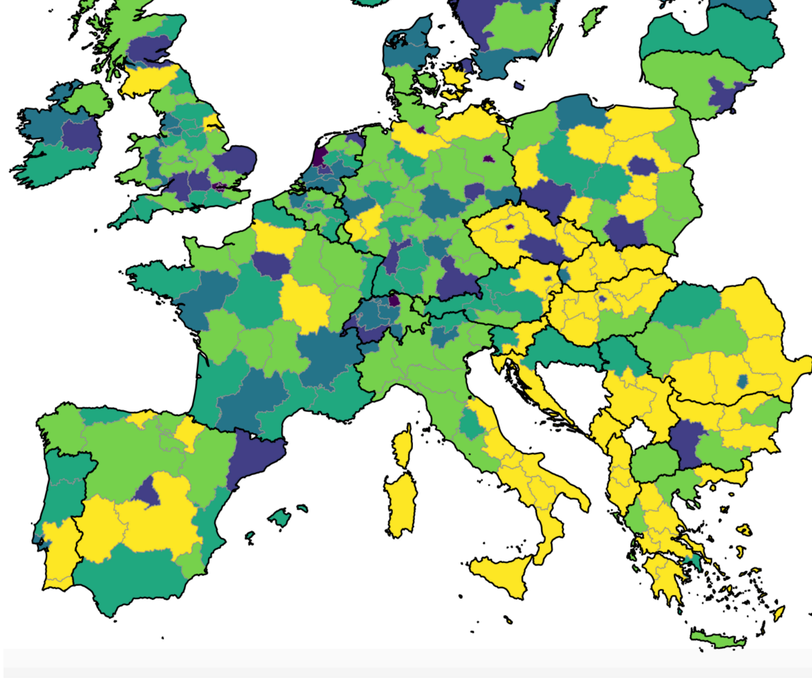
The Geography of OSS
From Countries to Regions



OFA Symposium 2021

Johannes Wachs
Vienna University of Economics and Business
Complexity Science Hub Vienna

WIRTSCHAFTS UNIVERSITÄT WIEN VIENNA UNIVERSITY OF ECONOMICS AND BUSINESS

COMPLEXITY SCIENCE HUB VIENNA

@johannes_wachs

OSS and the Economy

*An increase of **10%** in contributions would generate between **0.4% and 0.6% additional EU GDP** per year […and] more than **600 additional ICT start-ups** per year in the EU.*

- Blind et al.

OSS and the Economy

*An increase of **10%** in contributions would generate between **0.4% and 0.6% additional EU GDP** per year […and] more than **600 additional ICT start-ups** per year in the EU.*

- Blind et al.

Mechanisms:

- OSS as public good: reuse code, build on work of others.
- OSS use saves €.
- OSS is open/transparent -> high standards.

OSS and the Economy

*An increase of **10%** in contributions would generate between **0.4% and 0.6% additional EU GDP** per year […and] more than **600 additional ICT start-ups** per year in the EU.*

- Blind et al.

Mechanisms:

- OSS as public good: reuse code, build on work of others.
- OSS use saves €.
- OSS is open/transparent -> high standards.

-> But does it matter *where* OSS is created?

Geography and OSS

Distance is not dead!

- Likelihood of collaboration on GitHub decays exponentially with distance, a la gravity equation (Takhteyev & Hilts, 2010; Fackler & Laurentsyeva, 2020).

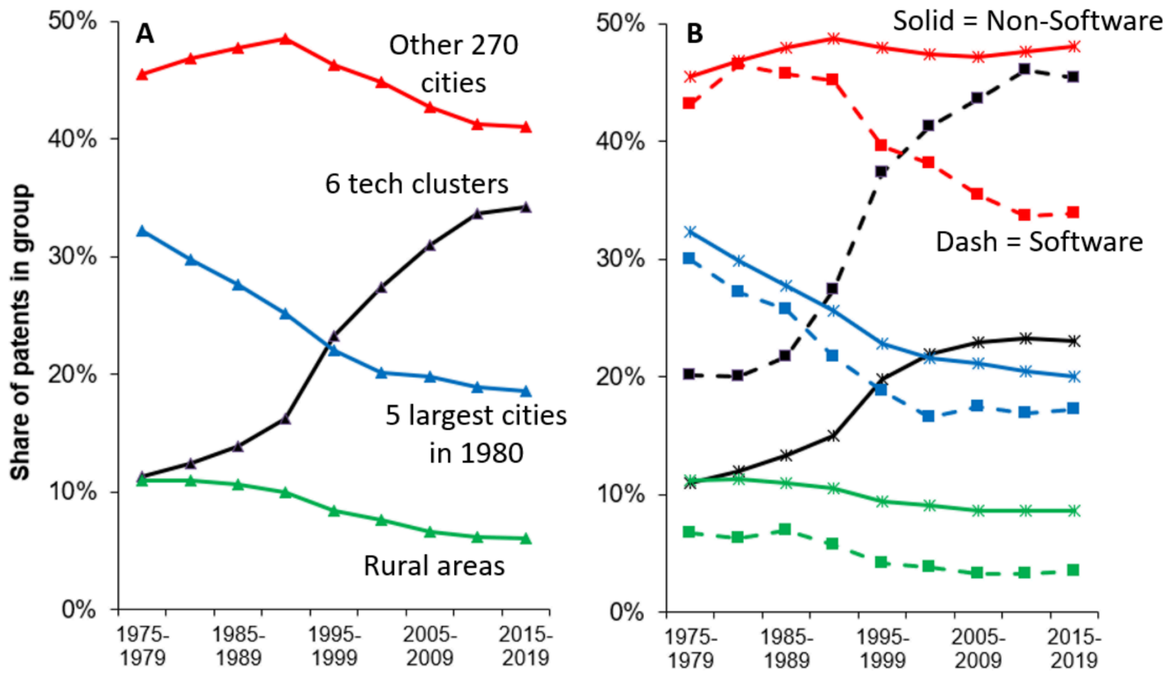
Many benefits of OSS accrue locally because:

1. Firms learn and gain feedback by contributing to OSS (Nagle 2018 - Org. Sci.)

2. Firms using OSS become more productive (Nagle 2018 - Man. Sci.)

3. Firms, workers, and investors use the information revealed by OSS contributions to make better choices (Wright et al. 2020).

Fig. 2: Spatial distribution of US patents



More broadly innovation in all fields is increasingly clustered in a few tech hubs. In the US this is largely driven by software patents.

Six Tech Clusters: SF, Boston, Seattle, San Diego, Denver, Austin

Chattergoon and Kerr, 2021

@johannes_wachs

With these ideas in mind, we would like to know:

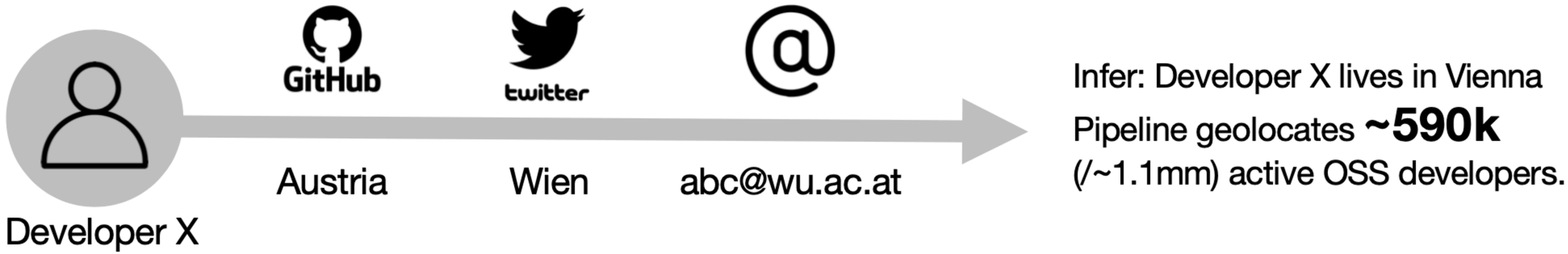
1. Does OSS activity cluster significantly in space?
2. If so, where are these hotspots?
3. Can we explain the ingredients needed for a place to promote and attract OSS development and developers?
4. Can we translate these into policy ideas?

Measuring the Geography of OSS in 2021

We* built a pipeline to generate geographic data on active OSS devs.

1. Look up “active” software developers on GitHub using GHArchive. Inclusion criteria: 100+ commits across 2019 & 2020.

2. Geolocate them using the Bing Maps API + Heuristics via user-provided locations on GitHub, Twitter, and commit email suffixes:

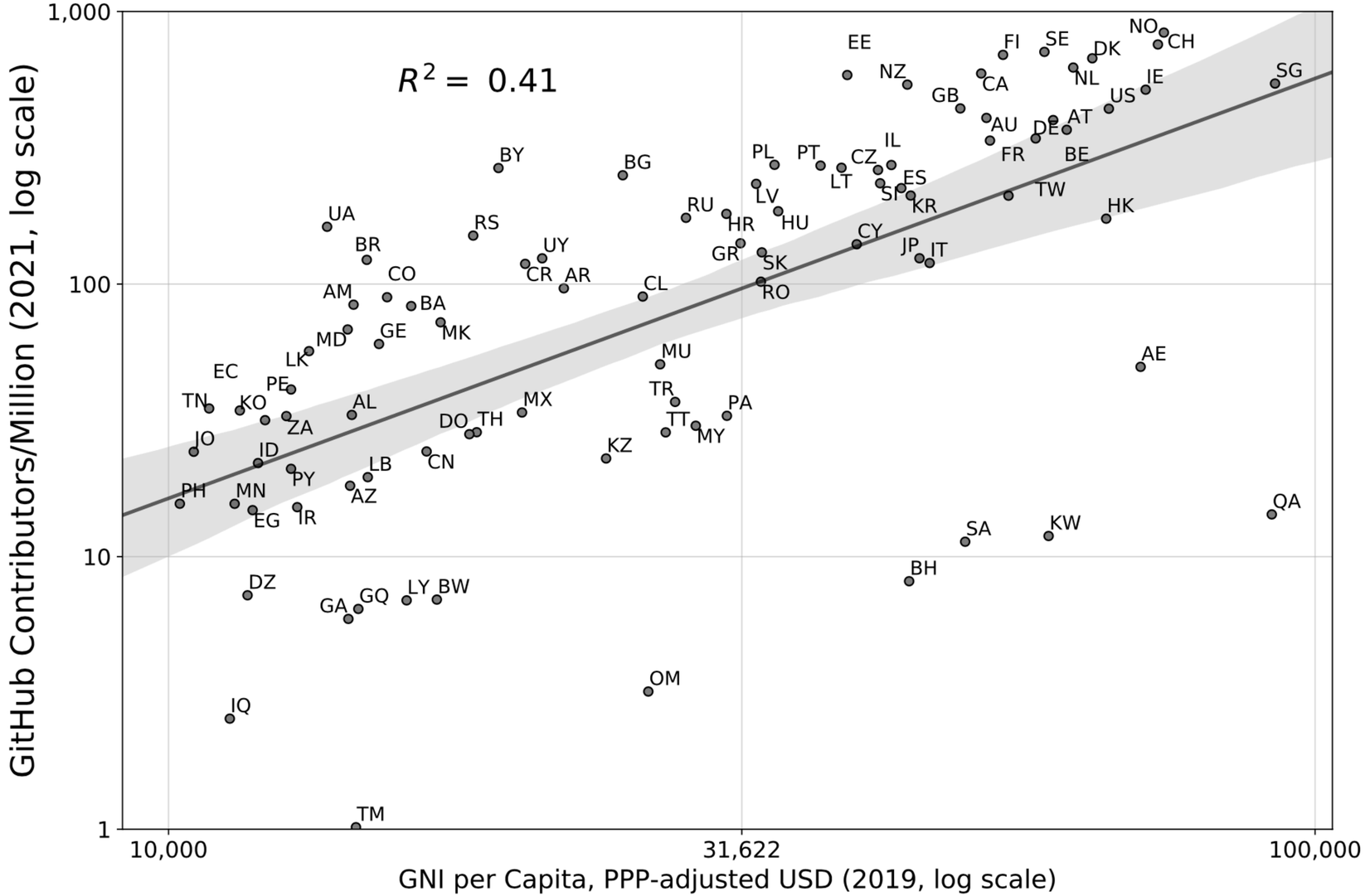


**joint work with Mariusz Nitecki, William Schueller, and Axel Polleres:*
The Geography of Open Source Software (2021) - preprint available at https://arxiv.org/abs/2107.03200.

	Sourceforge 2008		GitHub 2010		**GitHub 2021**		
Rank	Country	Share	Country	Share	Country	Share	Rank Chg. vs. 2008
1	United States	36.1	United States	38.7	United States	24.6	-
2	Germany	8.1	UK	7.7	China	5.8	↑ 4
3	UK	5.1	Germany	6.2	Germany	5.6	↓ 1
4	Canada	4.2	Canada	4.3	India	5.4	↑↑ 7
5	France	3.8	Japan	3.9	UK	5.0	↓ 2
6	China	3.1	Brazil	3.6	Brazil	4.4	↑↑ 6
7	Australia	2.7	France	3.2	Russia	4.3	↑↑ 6
8	Italy	2.6	Australia	3.1	France	3.8	↓ 3
9	Netherlands	2.5	Russia	2.3	Canada	3.8	↓↓ 5
10	Sweden	2.0	Sweden	2.2	Japan	2.7	↑↑ 5
11	India	1.9			South Korea	1.9	↑↑↑ 14
12	Brazil	1.8			Netherlands	1.8	↓ 3
13	Russia	1.6			Spain	1.8	↑ 1
14	Spain	1.6			Poland	1.8	↑ 2
15	Japan	1.3			Australia	1.8	↓↓ 8

Country shares of all active OSS contributors on GitHub in 2021
vs. previous snapshots from:
- Sourceforge *(Gonzalez-Barahona et al., 2008)*
- GitHub *(Takhteyev & Hilts, 2010)*.

Between countries: since 2010 more even distribution.



Not just a rich country vs poor country story!

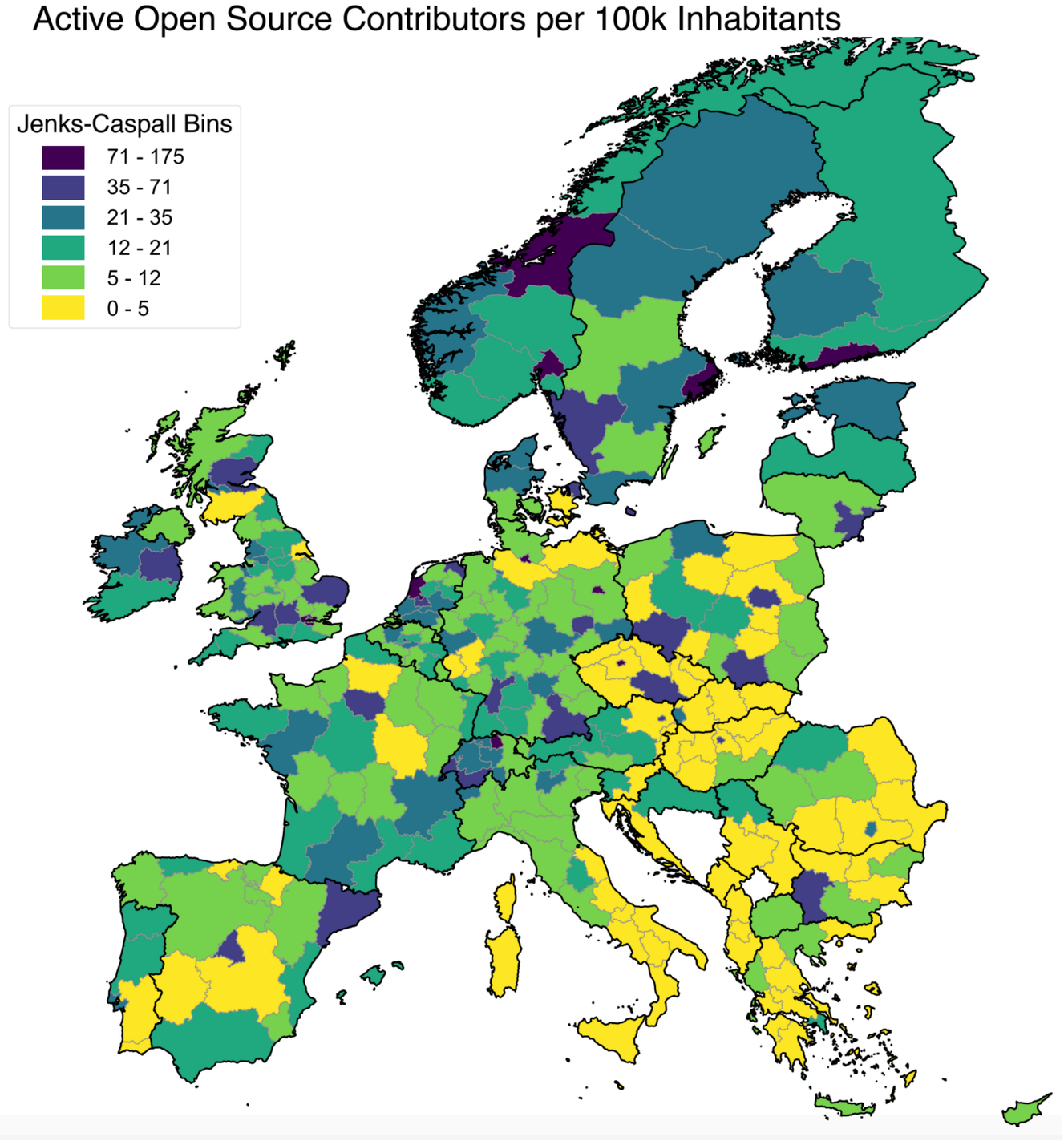
We *can* explain ~75% of variance between countries using GNI/pc and the UN's HDI as features in a regression model.

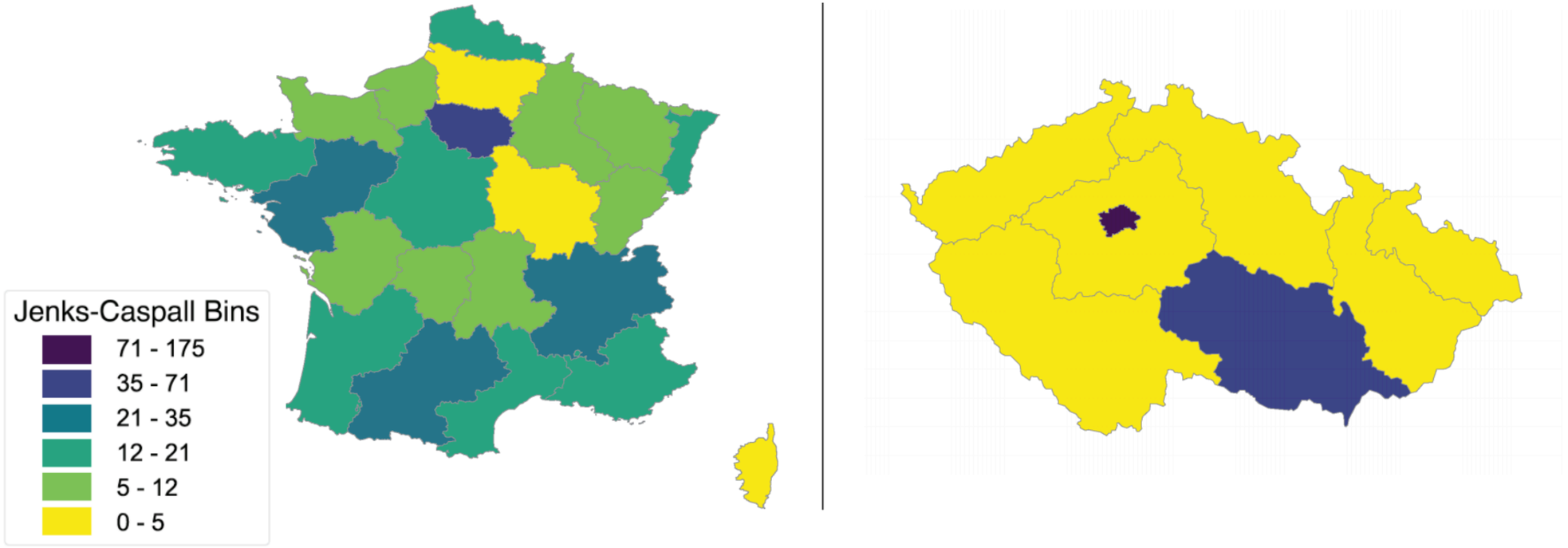
European Regions

We observed intense regional concentration at the NUTS 2 level.

Berlin: 175/100k

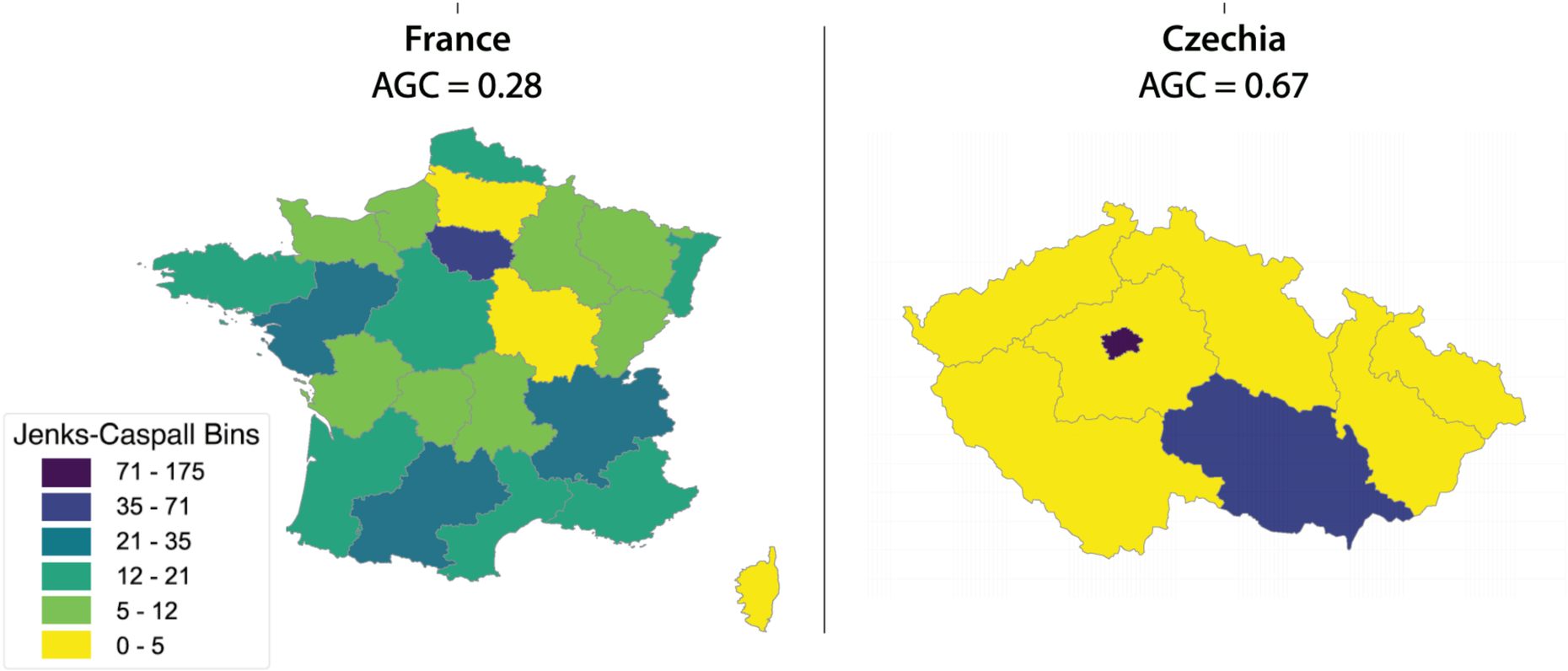
London, Zurich, Oslo,
Prague, Stockholm, Amsterdam
> 100/100k.



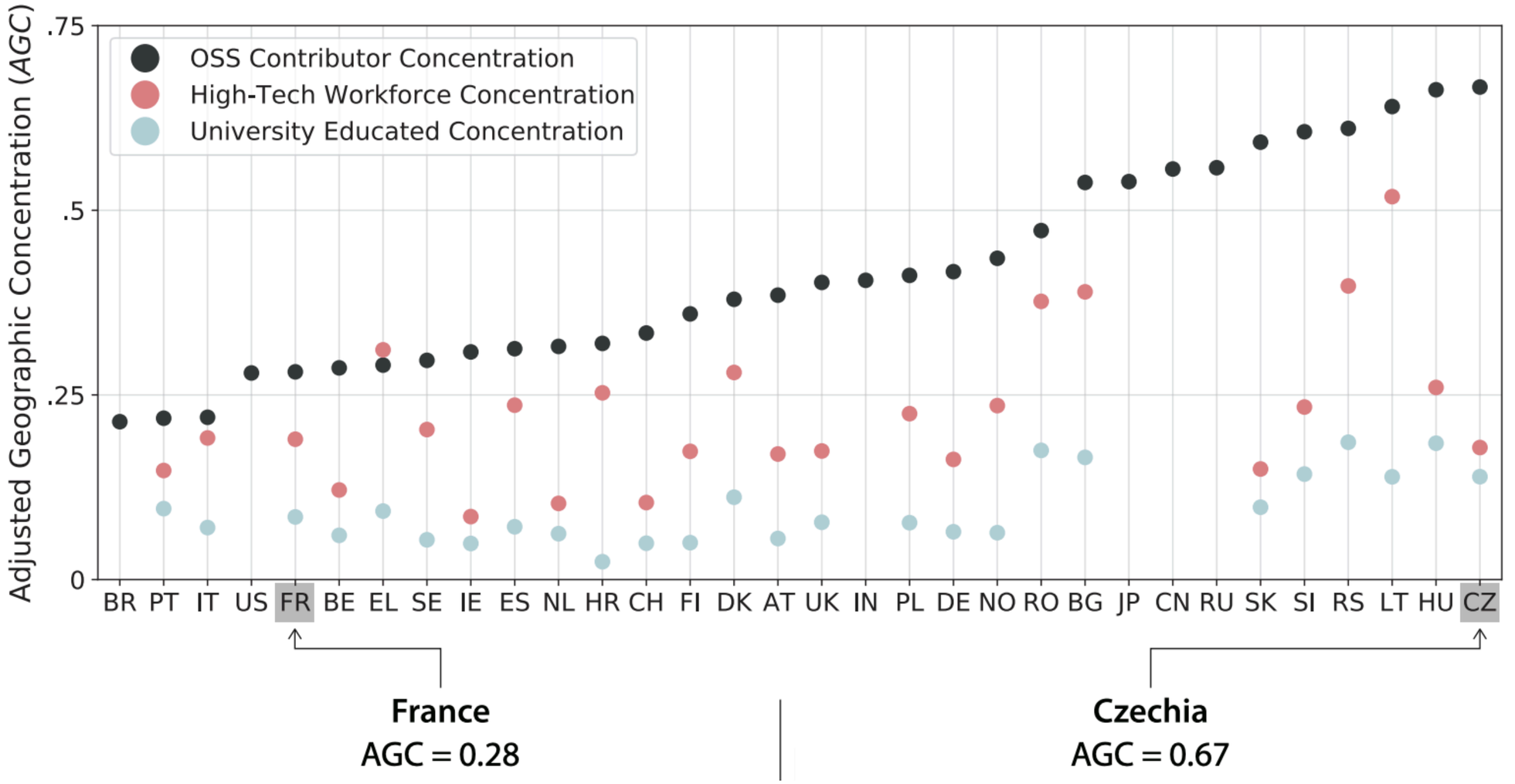


Compare Met. France with the Czech Republic. The differences between regions seem larger in the latter.

Can we measure the degree of concentration?



We adopt a measure of geographic concentration from the OECD called **Adjusted Geographic Concentration**. *AGC* ranges from 0 to 1. It is 0 if OSS developers are spread according to population, and is 1 if all OSS contributors are concentrated in the least populated region of the country.



One can also calculate the AGC for university educated individuals, or individuals working in High-tech fields. The concentration of OSS developers exceeds the concentration of these alternatives. OSS activity is **extremely** concentrated!

US Metropolitan Statistical Areas

MSA Name	Count Contributors	Population	Contributors/100k
San Jose-Sunnyvale-Santa Clara, CA	4,587	1,990,660	230
San Francisco-Oakland-Hayward, CA	10,702	4,731,803	226
Seattle-Tacoma-Bellevue, WA	8,830	3,979,845	221
Austin-Round Rock, TX	3370	2,227,083	151
Portland-Vancouver-Hillsboro, OR-WA	2,751	2,492,412	110
Boston-Cambridge-Newton, MA-NH	5,221	4,873,019	107
Denver-Aurora-Lakewood, CO	2,555	2,967,239	86
Raleigh, NC	1,159	1,390,785	83
Salt Lake City, UT	989	1,232,696	80
New York-Newark-Jersey City, NY-NJ-PA	11,579	19,216,182	60

Table 6: Top 10 US Metropolitan Statistical Areas with at least one million inhabitants ranked by active GitHub developers per capita.

Not just a European story!

We can estimate that 34% of all US-based OSS devs are in the 6 tech hubs of Chattergoon & Kerr (accounting for 33% of all US patents and 45% of software patents).

Explaining Variance

Recall that we can explain 75% of variance in OSS activity between countries with human and economic development indicators in a regression framework.

We can only explain **50%** of variance between NUTS 2 regions with a similar approach.

——> Local OSS activity is more idiosyncratic!

Strong correlates:
- Tertiary education
- Employment in high-tech industries
- Social trust measured via the European Values Survey

Revisiting our questions:

1. Does OSS activity cluster significantly in space? **YES**
2. Where are the hotspots? **In Europe: London, Berlin, Prague, etc.**
3. Can we explain the ingredients needed for a place to promote and attract OSS development and developers? **Not very well! Seems idiosyncratic…**
4. Can we translate these into policy ideas? **…**

Policy

Regional and city policy is quite different from national policy.

1. Often not as much power, but can be more flexible.
2. Rich literature on *cluster policy*: how to encourage agglomerations of specific activities.
    - Foster informal networks (key to Silicon Valley's flourishing - see Saxenian, 1990)
    - Give people opportunities to meet in person, encourage mentoring relationships
    - Advise firms on the benefits of OSS
    - Involve local universities ("Helix" models of innovation/development)

Data/code ——>

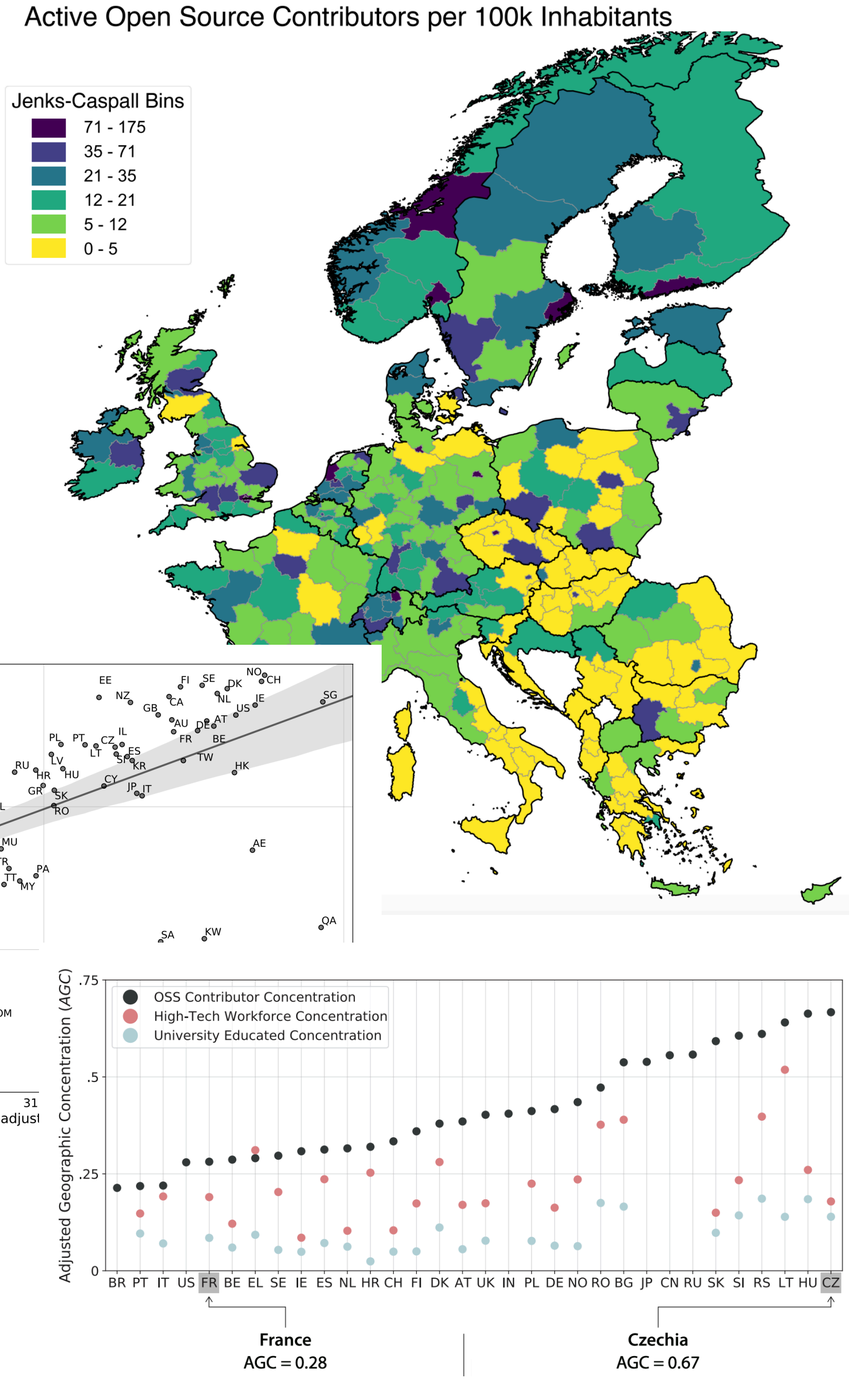
github.com/johanneswachs/OSS_Geography_Data

Contact:
johannes.wachs@wu.ac.at
@johannes_wachs



<—— Preprint/Report

https://arxiv.org/abs/2107.03200



Rank	Country	ISO2	Count Total Contributors	Pop. (mm)	Cont. / 100k
1	Iceland	IS	421	0.4	105
2	Switzerland	CH	7197	8.6	84
3	Norway	NO	4012	5.3	76
4	Sweden	SE	7323	10.3	71
5	Finland	FI	3813	5.5	69
6	Denmark	DK	3906	5.8	67
7	Netherlands	NL	10773	17.3	62
8	Canada	CA	22269	37.6	59
9	Estonia	EE	760	1.3	58
10	Luxembourg	LU	324	0.6	54
11	New Zealand	NZ	2642	4.9	54
12	Singapore	SG	3102	5.7	54
13	Ireland	IE	2531	4.9	52
14	United States	US	144371	328.2	44
15	United Kingdom	GB	29452	66.8	44
16	Australia	AU	10337	25.4	41
17	Germany	DE	33212	83.1	40
18	Austria	AT	3276	8.9	37
19	France	FR	22551	67.1	34
20	Belgium	BE	3935	11.5	34
21	Israel	IL	2488	9.1	27
22	Belarus	BY	2532	9.5	27
23	Portugal	PT	2802	10.3	27
24	Lithuania	LT	748	2.8	27
25	Poland	PL	10406	38.0	27
26	Czechia	CZ	2805	10.7	26
27	Bulgaria	BG	1755	7.0	25
28	Slovenia	SI	492	2.1	23
29	Latvia	LV	443	1.9	23
30	Spain	ES	10593	47.1	22
31	Malta	MT	112	0.5	22
32	Taiwan	TW	4979	23.6	21
33	South Korea	KR	10921	51.7	21
34	Hungary	HU	1813	9.8	18
35	Croatia	HR	742	4.1	18
36	Russia	RU	25271	144.4	18
37	Hong Kong	HK	1303	7.5	17
38	Ukraine	UA	7204	44.4	16
39	Serbia	RS	1039	6.9	15
40	Cyprus	CY	168	1.2	14
41	Greece	GR	1510	10.7	14
42	Slovakia	SK	719	5.5	13
43	Japan	JP	15706	126.3	12
44	Uruguay	UY	435	3.5	12
45	Brazil	BR	25891	211.0	12
46	Costa Rica	CR	593	5.0	12
47	Italy	IT	7204	60.3	12
48	Romania	RO	1979	19.4	10
49	Namibia	NA	260	2.5	10
50	Argentina	AR	4332	44.9	10

	Active GitHub Contributors per Million Inhab. (log, 2021)				
	(1)	(2)	(3)	(4)	(5)
PPP GNI per Cap. ('000 USD, 2019)	0.017*	-0.007	-0.002	0.004	-0.010*
	(0.009)	(0.007)	(0.009)	(0.008)	(0.006)
Internet Penetration (% of Pop., 2019)	0.043***	0.002	0.029***	0.026***	-0.003
	(0.005)	(0.008)	(0.005)	(0.006)	(0.009)
Population (log, 2019)	-0.145***	-0.071	-0.016	-0.143**	-0.038
	(0.052)	(0.044)	(0.046)	(0.056)	(0.057)
Human Development Index (2019)		11.709***			9.327***
		(1.528)			(1.553)
Index of Public Integrity (2019)			0.704***		
			(0.127)		
Economic Complexity Index (2019)				0.962***	0.683***
				(0.184)	(0.153)
Observations	174	173	115	150	149
Adjusted R^2	0.631	0.747	0.819	0.740	0.804
Residual Std. Error	1.191	0.986	0.788	1.016	0.882
F Statistic	81.3***	175.2***	195.8***	142.4***	155.6***

*p<0.1; **p<0.05; ***p<0.01

Table 4: Regression models (1-5) relating country-level counts of GitHub contributors per million inhabitants (log-transformed) and socio-economic indicators. While income and internet penetration alone account for nearly two-thirds of variance in OSS activity (1), human development (2), quality of political institutions (3), and economic complexity (4) significantly improve model fit above and beyond that baseline. A combined model (5) explains over 80% of variance. We report robust standard errors.

	Active GitHub Contributors/100k Inhab. European NUTS2 (log, 2021)						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Internet Penetration	0.010**	0.005	0.012***	0.012***	0.012**	0.011**	0.004
—(% of Pop. 2017)	(0.004)	(0.005)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)
GDP per Cap.	0.600***	0.609***	0.012	0.367**	0.248	0.269	0.352**
—(log Eur, 2017)	(0.140)	(0.154)	(0.256)	(0.147)	(0.176)	(0.202)	(0.151)
Population	0.143*	0.178	0.356***	0.166*	0.063	0.089	0.207**
—(log, 2020)	(0.084)	(0.113)	(0.131)	(0.092)	(0.086)	(0.085)	(0.081)
Population Dens.	0.043*	0.036	0.044	-0.018	0.055**	0.050*	0.030
—(log, 2017)	(0.023)	(0.024)	(0.029)	(0.022)	(0.027)	(0.028)	(0.022)
EVS Trust		0.315**					
—(2017)		(0.151)					
R&D Spend. per Cap.			0.113**				
—(log, 2017)			(0.051)				
% Empl. High-Tech				0.083***			
—(2019/20)				(0.011)			
Patents Elec-Eng./100k					0.058***		
—(log, 2017)					(0.022)		
Patents/100k						0.047*	
—(log, 2017)						(0.026)	
% with Tertiary Edu.							0.021***
—(2019/20)							(0.002)
Lambda	0.099***	0.114***	0.073***	0.083***	0.078***	0.084***	0.071***
—(est. spatial dep.)	(0.014)	(0.012)	(0.015)	(0.015)	(0.013)	(0.014)	(0.014)
Observations	276	198	258	262	258	258	276
Pseudo-R^2	0.392	0.429	0.417	0.509	0.411	0.399	0.536

*p<0.1; **p<0.05; ***p<0.01

Table 5: GMM spatial regression models (1-7) ([7]) relating EU NUTS2 counts of GitHub contributors per 100k inhabitants (log-transformed) and socio-economic indicators. Income, population, and internet penetration account for just over one third of variance in OSS activity (1). Social trust (2), R&D spending (3), employment in high tech sectors (4), innovation activity (5,6), and higher education (7) all explain additional variance above this baseline (from 1 to 14%). In each model the spatial autoregressive term lambda is positive and significant, indicating a positive adjacency relationship: neighboring regions tend to have similar levels of OSS activity even accounting for the features in each model.

MSA Name	Count Contributors	Population	Contributors/100k
Boulder, CO	995	326196	305
San Jose-Sunnyvale-Santa Clara, CA	4587	1990660	230
San Francisco-Oakland-Hayward, CA	10702	4731803	226
Seattle-Tacoma-Bellevue, WA	8830	3979845	221
Ann Arbor, MI	600	367601	163
Champaign-Urbana, IL	365	226033	161
Austin-Round Rock, TX	3370	2227083	151
Durham-Chapel Hill, NC	759	644367	117
Portland-Vancouver-Hillsboro, OR-WA	2751	2492412	110
Charlottesville, VA	238	218615	108
Boston-Cambridge-Newton, MA-NH	5221	4873019	107
Santa Cruz-Watsonville, CA	249	273213	91
Denver-Aurora-Lakewood, CO	2555	2967239	86
Madison, WI	574	664865	86
Raleigh, NC	1159	1390785	83
Salt Lake City, UT	989	1232696	80
Lafayette-West Lafayette, IN	166	233002	71
Trenton, NJ	251	367430	68
Gainesville, FL	227	329128	68
Santa Maria-Santa Barbara, CA	291	446499	65
New York-Newark-Jersey City, NY-NJ-PA	11579	19216182	60
Provo-Orem, UT	390	648252	60
San Diego-Carlsbad, CA	1903	3338330	57
College Station-Bryan, TX	148	264728	55
Pittsburgh, PA	1185	2317600	51
Fort Collins, CO	181	356899	50
Nashville-Davidson–Murfreesboro–Franklin, TN	952	1934317	49
Burlington-South Burlington, VT	104	220411	47
Athens-Clarke County, GA	101	213750	47
Atlanta-Sandy Springs-Roswell, GA	2668	6020364	44
San Luis Obispo-Paso Robles-Arroyo Grande, CA	124	283111	43
Washington-Arlington-Alexandria, DC-VA-MD-WV	2698	6280487	42
Eugene, OR	161	382067	42
Minneapolis-St. Paul-Bloomington, MN-WI	1554	3640043	42
Bellingham, WA	96	229247	41
Los Angeles-Long Beach-Anaheim, CA	5533	13214799	41
Chicago-Naperville-Elgin, IL-IN-WI	3876	9458539	40
Lincoln, NE	132	336374	39
Boise City, ID	272	749202	36
Rochester, NY	376	1069644	35
Tucson, AZ	347	1047279	33
Santa Rosa, CA	154	494336	31
Orlando-Kissimmee-Sanford, FL	804	2608147	30
Philadelphia-Camden-Wilmington, PA-NJ-DE-MD	1883	6102434	30
Vallejo-Fairfield, CA	138	447643	30
Charlotte-Concord-Gastonia, NC-SC	771	2636883	29
Kansas City, MO-KS	637	2157990	29
Rochester, MN	66	221921	29
Columbus, OH	603	2122271	28
Fargo, ND-MN	67	246145	27

Table 8: Top US MSAs with population of at least 250k, by developers per capita.