NATIONAL INSTITUTE OF SCIENCE AND TECHNOLOGY POLICY

プレプリントと研究論文の内容からみたCOVID-19研究の動向

2021年2月17日

第13回政策研究レビューセミナー

文部科学省　科学技術・学術政策研究所

第2調査グループ 小柴 等

概要

- **今回のコロナ禍において，研究動向の迅速な把握を企図し，プレプリントを用いた分析を試行**
  - プレプリントサーバを分野の代理変数とした分野動向把握
  - 自然言語処理を用いたトピックの抽出
  - 連絡先情報を用いた国・地域の推定
- **結果**
  - 投稿数は5月初旬をピークに，なだらかに減衰中
    - 週あたり450件程度の新規投稿はあるものの，ピークの800件程度からは半減
  - 具体のトピックは，社会・経済 や 公衆衛生 に関するものにシフト
    - 医療系の投稿数は安定・減衰傾向，人社系（SSRN）の数が相対的に伸び
  - 5月までの投稿数は 中国 がリード，その後は 米国 が首位に
    - 英国も米国と同様の傾向

COVID-19に関する既報（一部）

NISTEP Discussion Paper 181

- **COVID-19/SARS-CoV-2に関する研究の概況 2020.05.15**
  - 世界保健機関（WHO）で公開されているCOVID-19に関する文献リスト並びにプレプリントサーバbioRxivおよびmedRxivで公開されているCOVID-19/ SARS-CoV-2関連の論文リストを対象として，COVID-19/SARS-CoV-2に関する研究の概況把握を実施

NISTEP Discussion Paper 185

- **COVID-19研究に関する国際共著状況：2020年4月末時点のデータを用いた分析 2020.07.03**
  - COVID-19研究に関する国際共著状況の把握を目的として，国・地域別の文献産出状況や国際共著状況等の調査を実施

本日の内容はこの報告書がベース

NISTEP Discussion Paper 186

- **COVID-19/SARS-CoV-2 関連のプレプリントを用いた研究動向の試行的分析 2020.06.30 （2020.11.04 補遺 公表）**
  - COVID-19関連のプレプリント（arXiv, medRxiv, bioRxiv, chemRxiv, SSRN）を対象に，自然言語処理を用いたエマージング領域の把握も試行的に実施

背景

- **迅速性などの観点からプレプリントサーバが普及**
  - COVID-19の様な危急の要件にも合致すると期待される
- **プレプリントサーバ（PPS）は分野に違い**
  - 研究分野ごとに投稿するPPSが異なる傾向

▼

- **PPSによって，数の推移，トピック，分布などにはどんな違いがあるのか（ないのか）？**
- **PPSを用いた研究動向把握は可能か？**

分析対象としたPPS

- **arXiv**
  - プレプリントの老舗
  - 物理・情報系

arXiv.org

- **medRxiv・bioRxiv**
  - 医療，バイオ系

medRχiv THE PREPRINT SERVER FOR HEALTH SCIENCES
bioRχiv THE PREPRINT SERVER FOR BIOLOGY

- **SSRN**: Social Science Research Network
  - 人社系を中心に全般

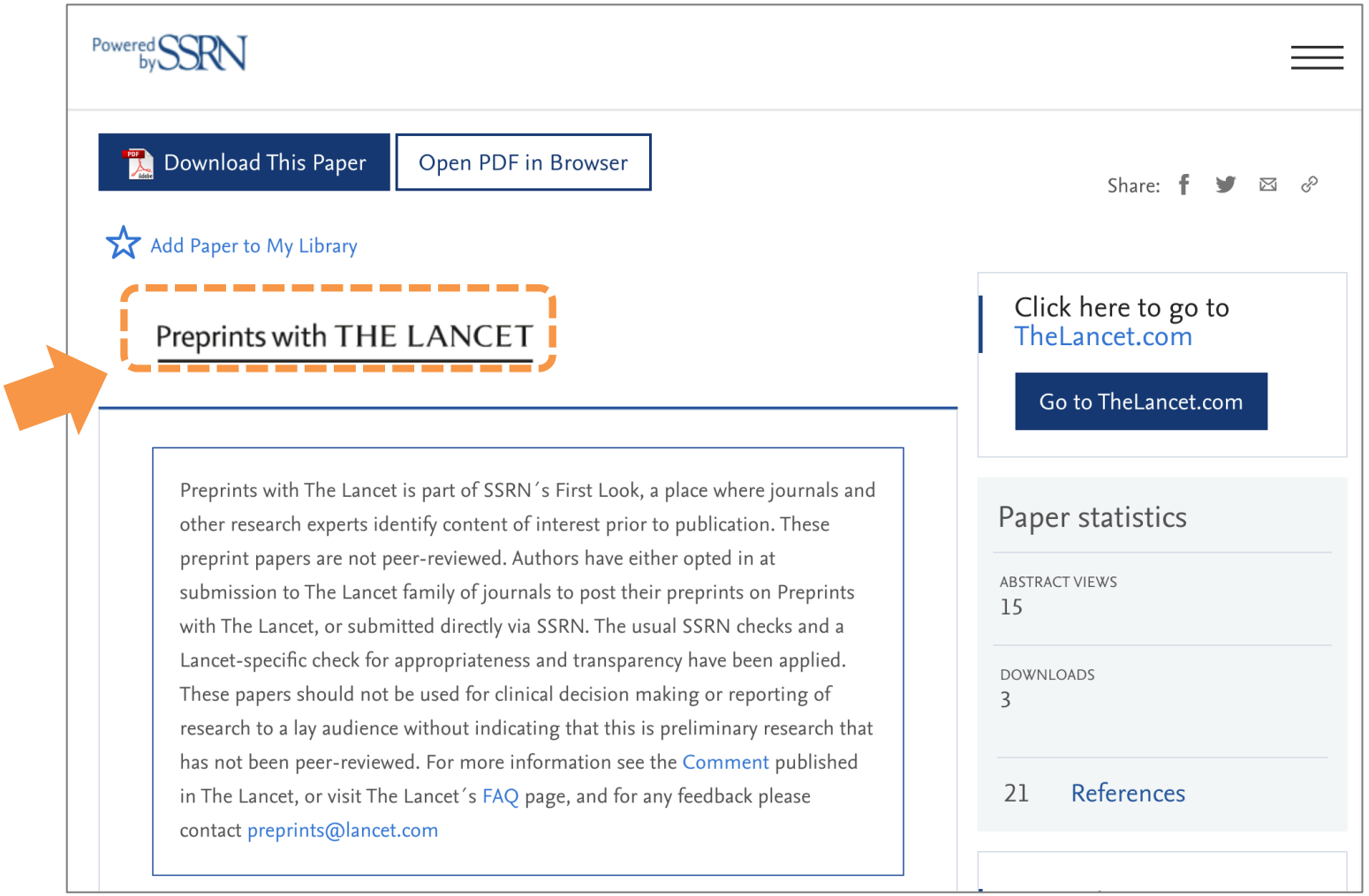
SSRN Tomorrow's Research Today

- **chemRxiv**
  - 化学系

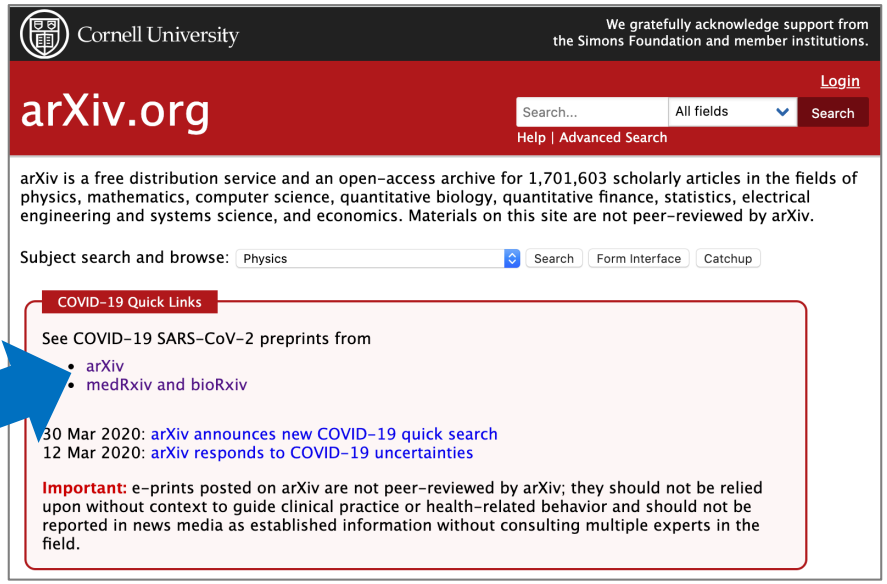
ChemRxiv™

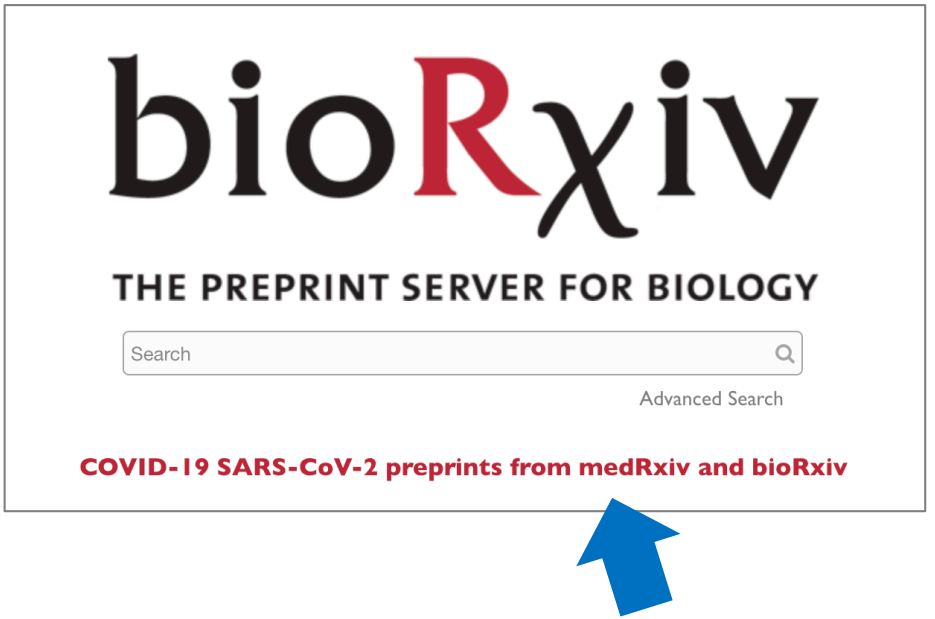
SSRN はさらに2分割

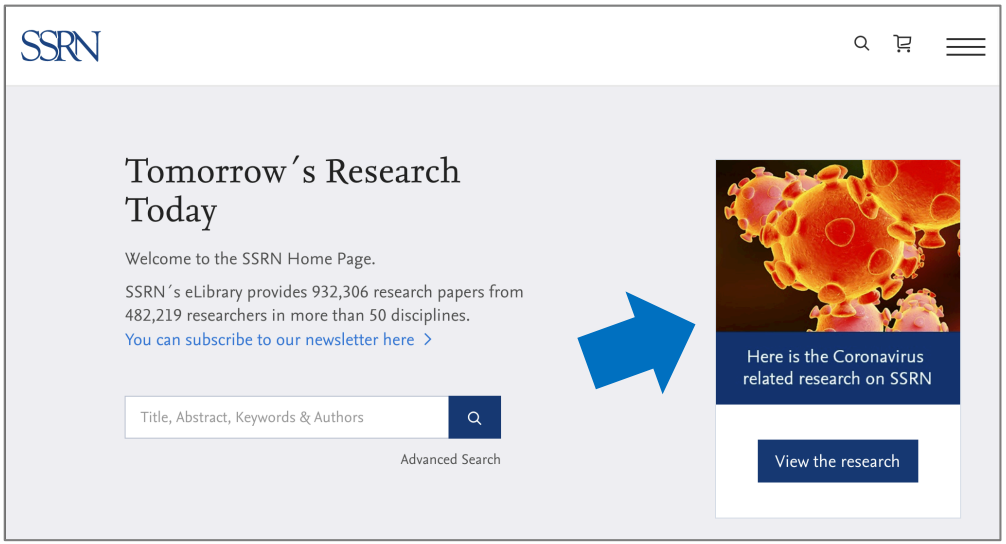
- 医療系の著名雑誌 Lancet 関連の投稿がある
  - これらを **SSRN Lancet として別に扱う**



対象となるプレプリント







基本的には各PPSの
"COVID-19 関連原稿リスト"
掲載のものを収集

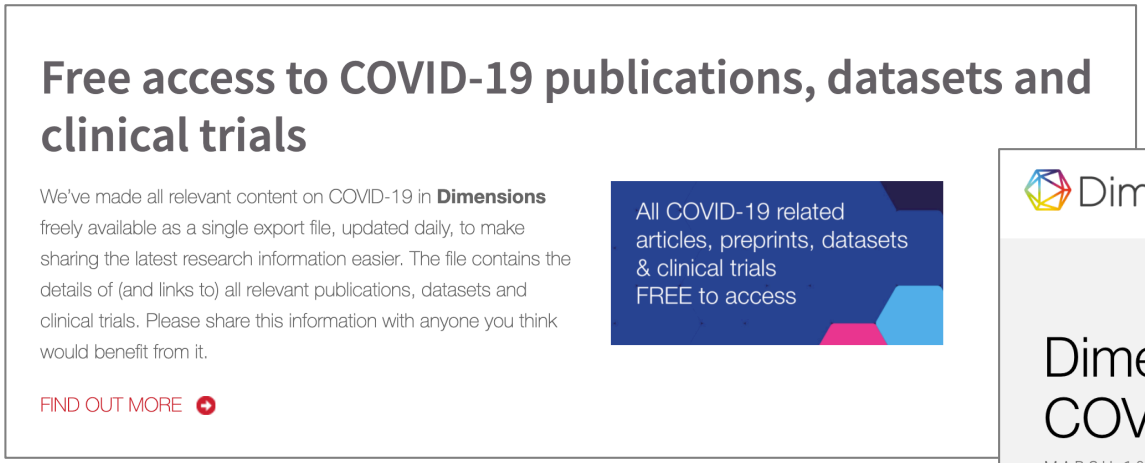
対象となるプレプリント

- **chemRxiv は COVID-19 の特別リスト無し**
  - chemRxiv の技術基盤である figshare を運営する Digital Science 社の文献DBサービス Dimensions で別途COVID-19関連の論文リストを提供

▼

  - Dimensions の COVID-19関連の論文リスト から，取り出した chemRxiv をベースに新たに整備
    - 新規追加分は SARS-CoV，COVID で検索した結果※

※ 上記の既存リストからキーワードを推定



データと分析内容

データソース

プレプリント
bioRxiv / medRxiv
arXiv, SSRN
chemRxiv

分析内容

時系列推移

トピック分析

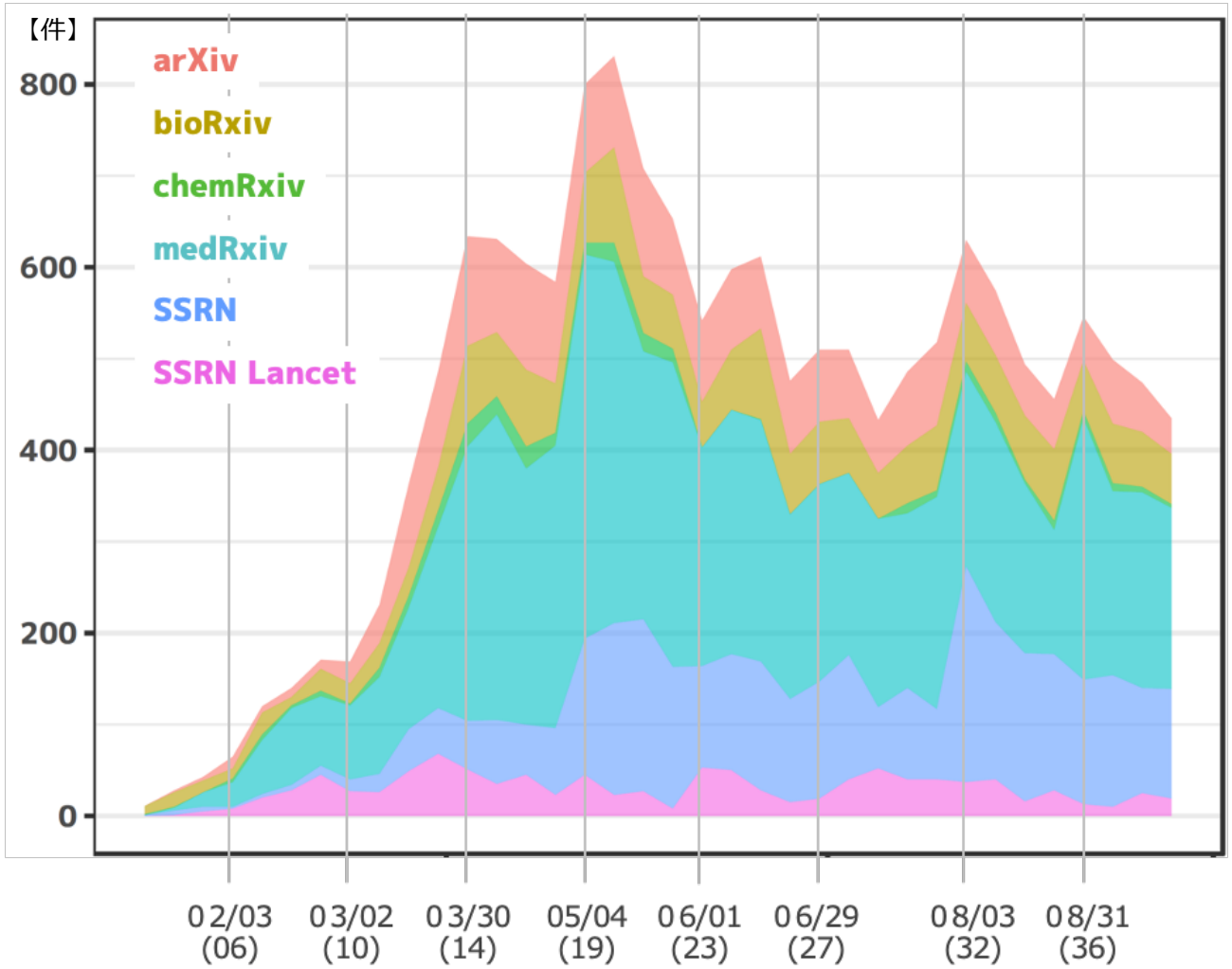
国際比較

- **収集日：** **2020年 9月末**
- **対象期間：** **2020年 第04週 (01/20から)**
  2020年 第39週 (09/27まで)
- **基準日：** **Posted Date**
  - PPS に 記事が投稿された日を基準として採用
- **件数：** **16,066 件**


NISTEP
NATIONAL
INSTITUTE OF
SCIENCE AND
TECHNOLOGY
POLICY

全体傾向

数の推移



括弧内は週番号

Week		arXiv	bioRxiv	ChemRxiv	medRxiv	SSRN	Lancet
2020-01-20	(04)	0	9	1	0	1	0
2020-01-27	(05)	2	16	1	3	5	1
2020-02-03	(06)	4	13	0	16	5	5
2020-02-10	(07)	13	11	4	27	2	8
2020-02-17	(08)	7	24	6	59	4	20
2020-02-24	(09)	10	9	3	84	6	28
2020-03-02	(10)	10	24	6	76	10	45
2020-03-09	(11)	24	22	2	81	13	27
2020-03-16	(12)	42	27	10	106	20	26
2020-03-23	(13)	91	31	13	133	46	49
2020-03-30	(14)	105	47	18	198	50	68
2020-04-06	(15)	120	85	25	300	53	51
2020-04-13	(16)	102	70	20	334	70	35
2020-04-20	(17)	116	84	24	280	55	45
2020-04-27	(18)	111	54	14	309	73	23
2020-05-04	(19)	97	76	13	420	149	45
2020-05-11	(20)	100	104	21	395	188	23
2020-05-18	(21)	118	62	20	293	188	27
2020-05-25	(22)	83	59	15	333	155	8
2020-06-01	(23)	89	50	0	239	111	53
2020-06-08	(24)	88	66	0	267	127	50
2020-06-15	(25)	79	99	1	264	141	28
2020-06-22	(26)	80	66	0	202	113	15
2020-06-29	(27)	79	68	0	216	128	19
2020-07-06	(28)	75	60	0	199	136	40
2020-07-13	(29)	58	50	0	206	67	52
2020-07-20	(30)	81	63	11	191	100	40
2020-07-27	(31)	91	71	7	232	77	40
2020-08-03	(32)	69	64	11	213	236	37
2020-08-10	(33)	71	63	10	219	172	40
2020-08-17	(34)	56	70	4	186	162	16
2020-08-24	(35)	55	78	10	136	149	28
2020-08-31	(36)	48	55	8	286	136	13
2020-09-07	(37)	70	65	9	201	144	10
2020-09-14	(38)	54	60	6	214	115	25
2020-09-21	(39)	39	55	4	198	120	19

※ 既報分も新たに収集・分析し直した結果，既報部分もわずかに数が異なる

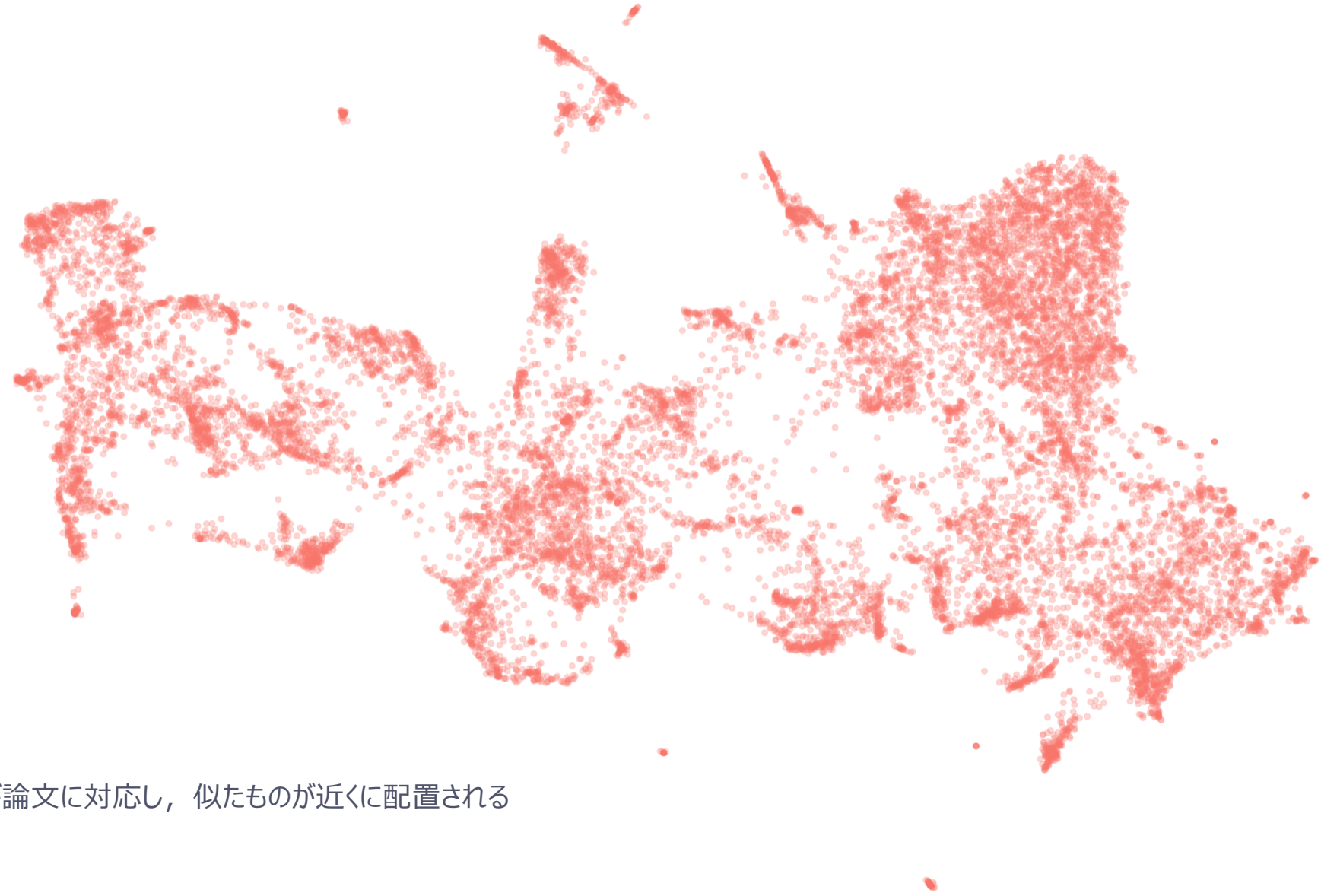

NISTEP
NATIONAL
INSTITUTE OF
SCIENCE AND
TECHNOLOGY
POLICY

トピックの分析

論文の可視化・トピックの抽出手法

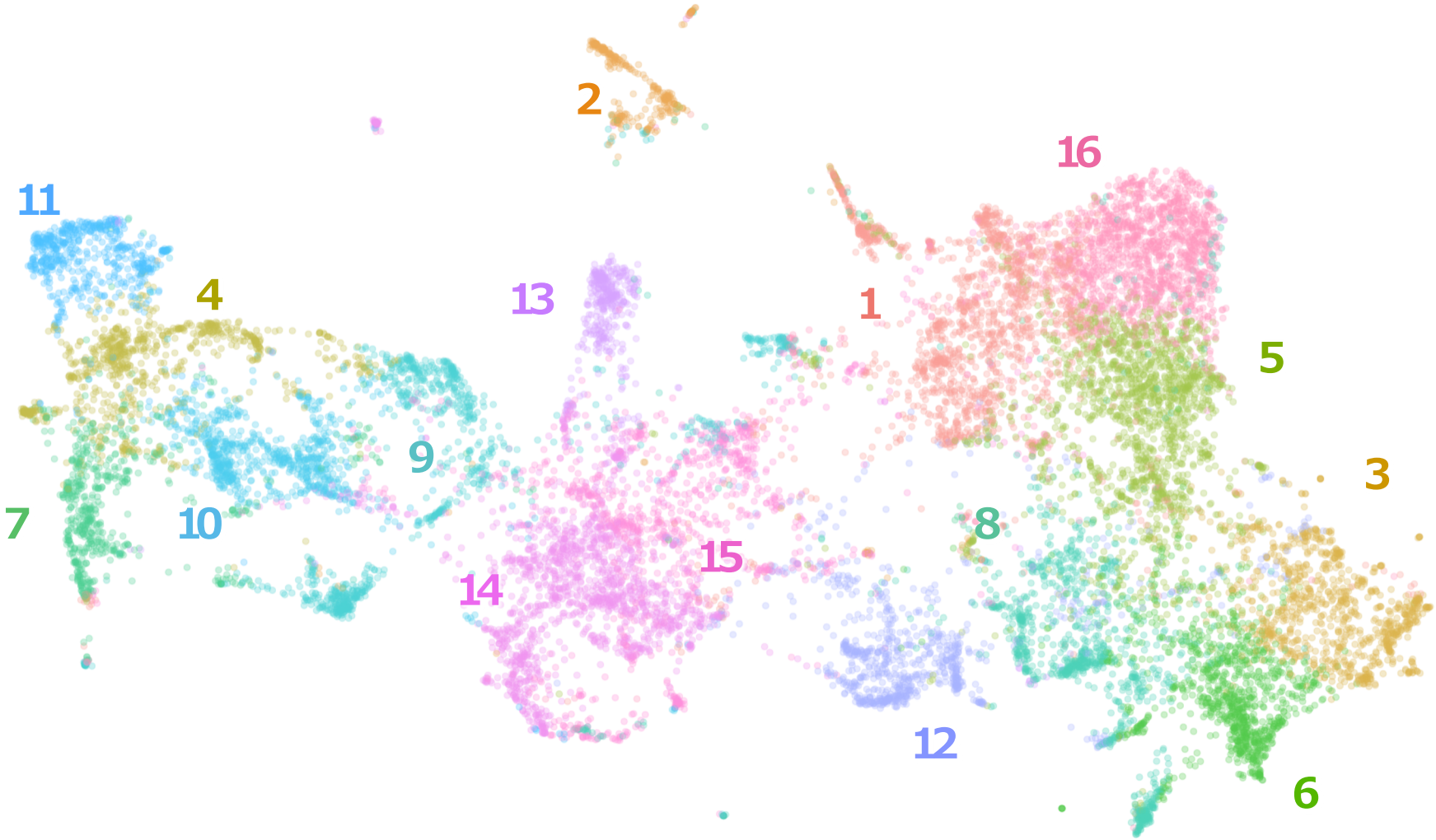
- **タイトル・概要を対象に分析**
- **分散表現化**
  - 上記対象を用い，fastTextで100次元の単語分散表現を算出
  - 単語分散表現の線形加算＆正規化を論文の分散表現に採用
- **トピック**
  - 論文分散表現に対し，K-means++ で16分割しトピックに採用
  - 各トピックの単語出現頻度からトピックをラベル付け
- **可視化**
  - UMAPにて100次元空間を2次元に圧縮

コンテンツ内容の分布



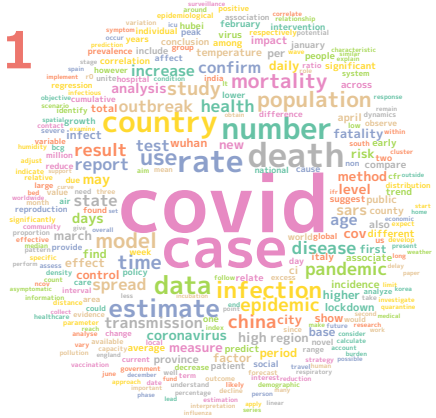
各点が論文に対応し，似たものが近くに配置される

コンテンツ内容の分布：トピック別



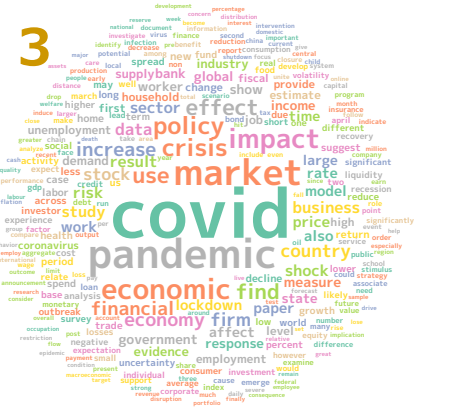
各点が論文に対応し，似たものが近くに配置される
この図での色は，高次元空間での距離に基づき16分類した結果

16のトピック

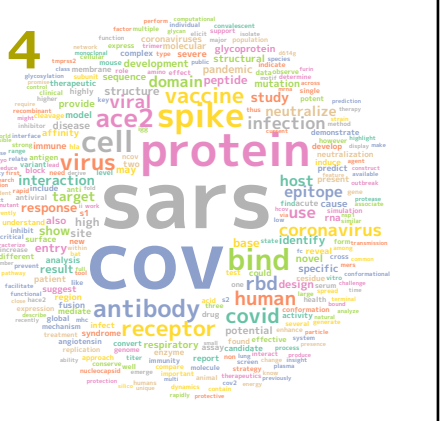


国別比較

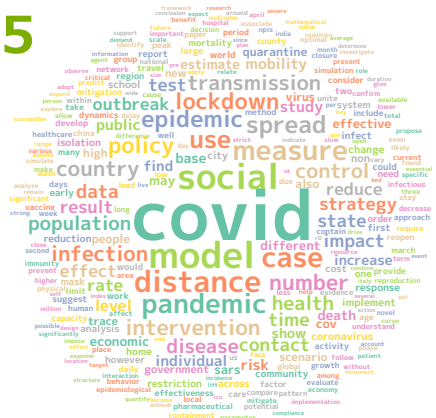
マスク・人工呼吸器



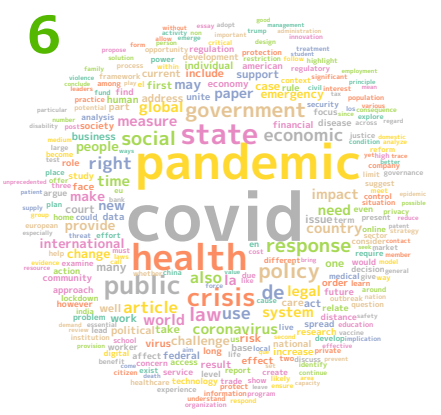
社会・経済・政策



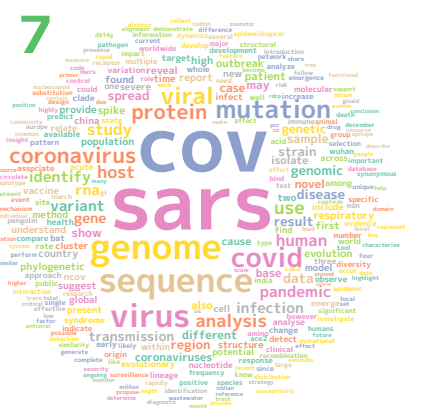
ワクチン開発



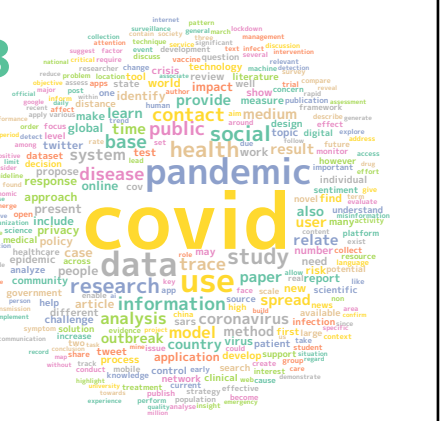
感染拡大



公衆衛生



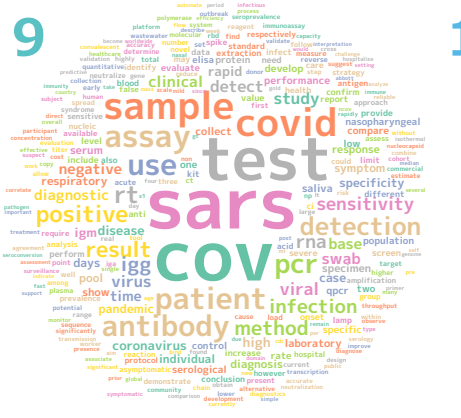
ゲノム解析



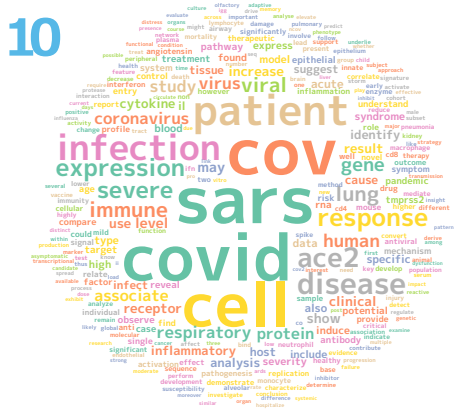
情報・データ分析

p.15 の分類ごとに頻出語と頻度を可視化したもの．紺色のラベルは専門家による解釈例

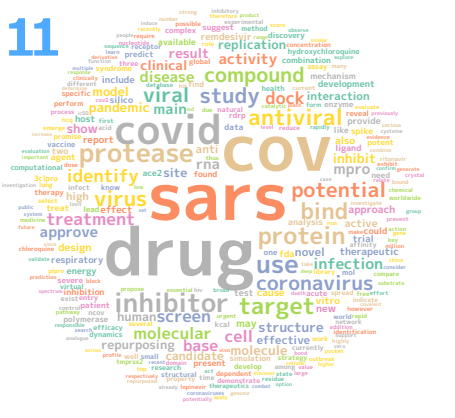
16のトピック



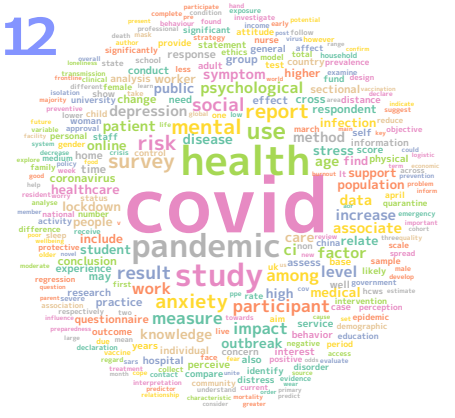
検出・検査



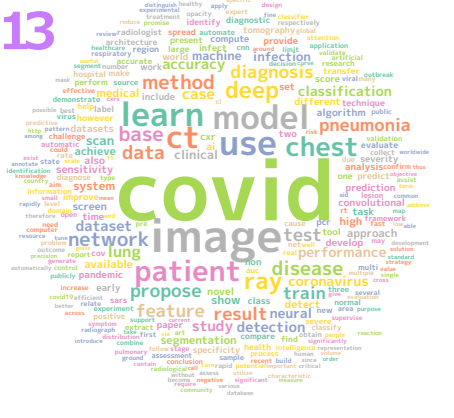
感染機構



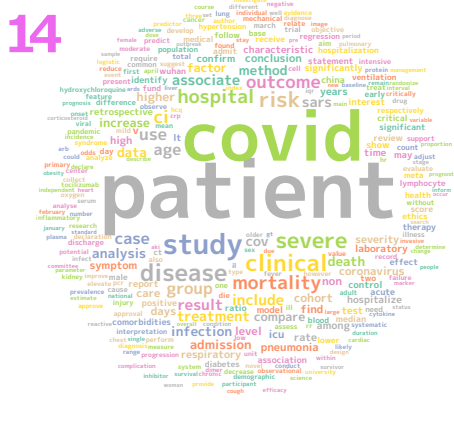
治療薬探索



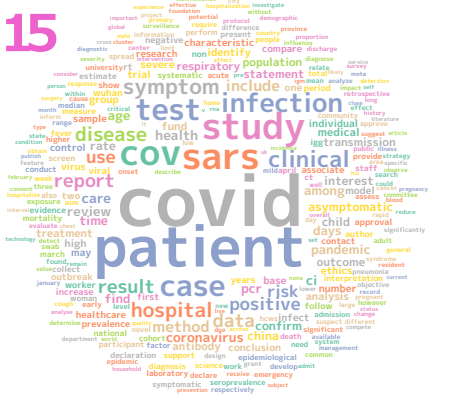
健康・不安



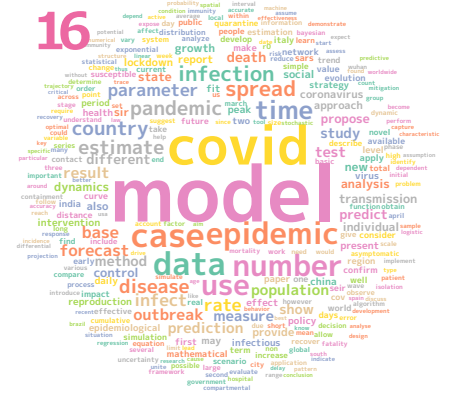
肺画像診断



患者病状



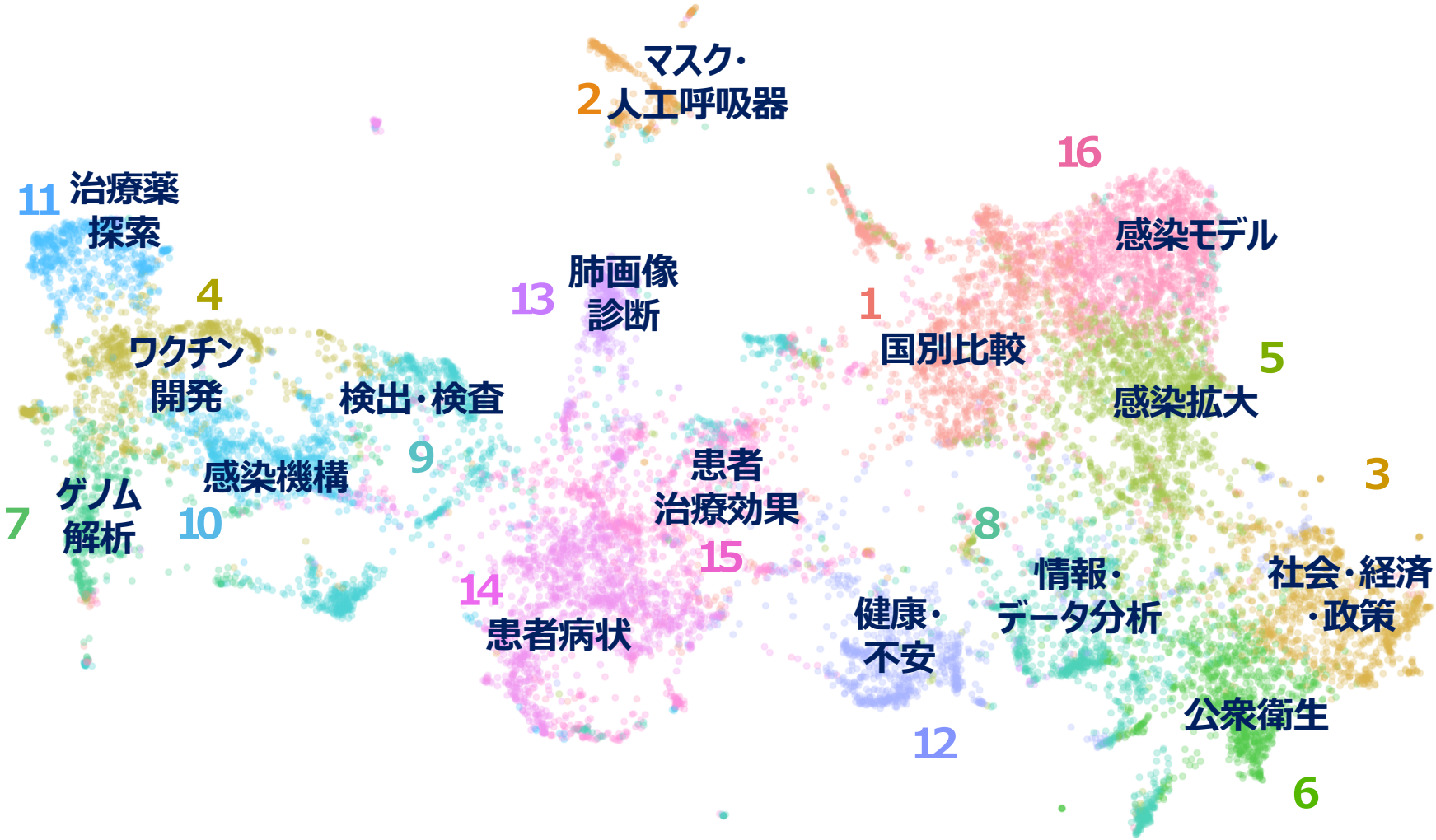
患者治療効果



感染モデル

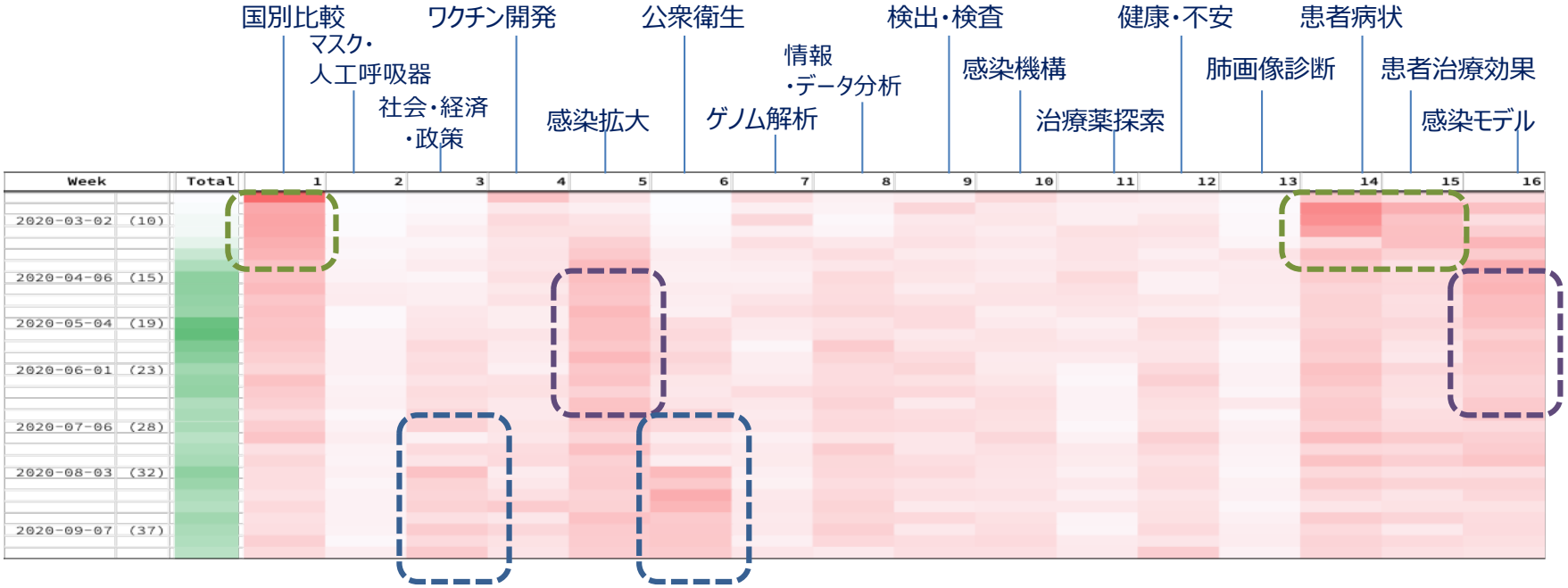
p.15 の分類ごとに頻出語と頻度を可視化したもの．紺色のラベルは専門家による解釈例

コンテンツ内容の分布：トピック別



各点が論文に対応し，似たものが近くに配置される
この図での色は，高次元空間での距離に基づき16分類した結果
文字は，p.16, 17 に示したワードクラウドに基づき専門家による解釈

トピックの時系列変化

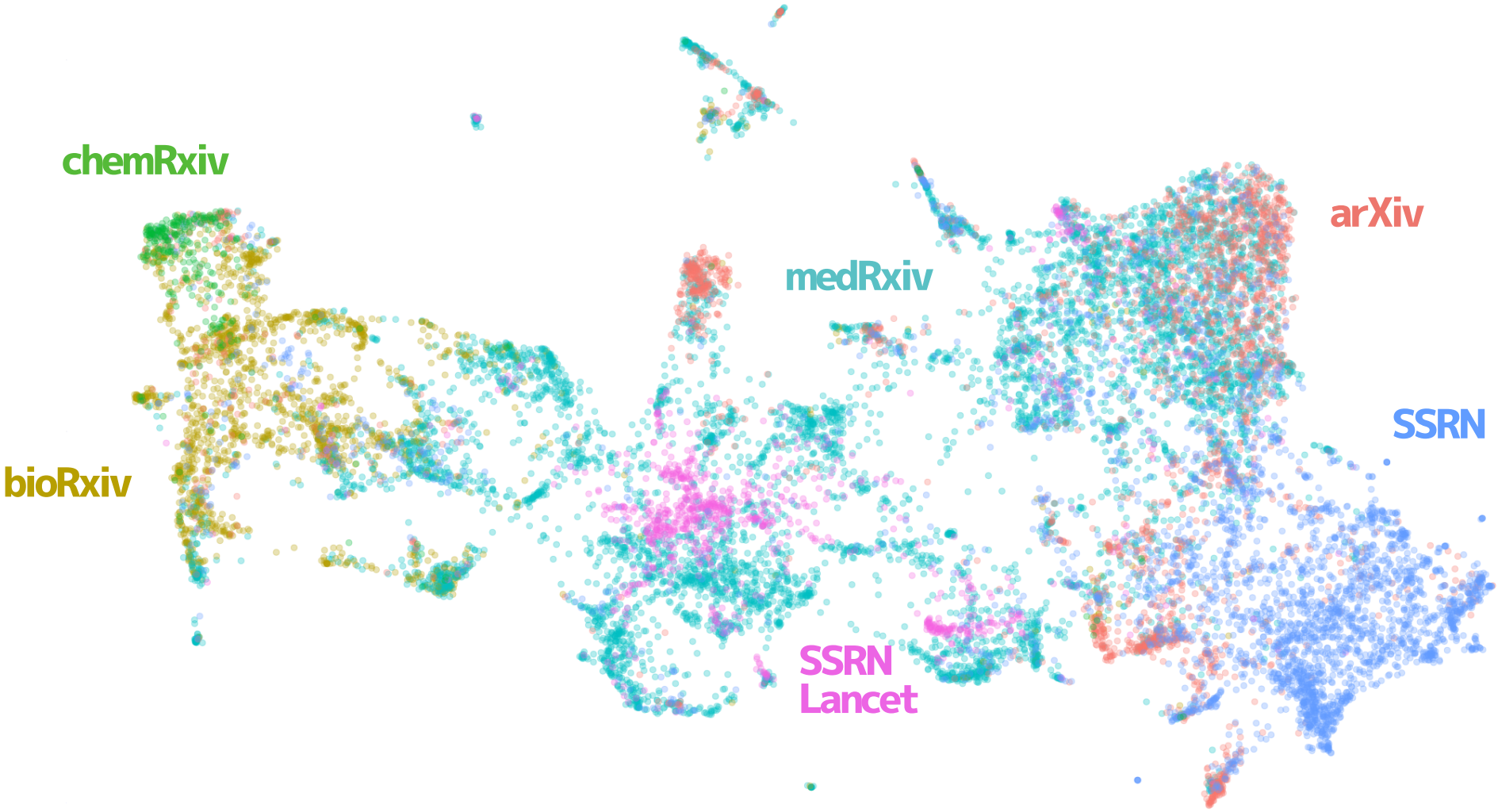


- 投稿数は5月の連休（05/04前後）をピークになだらかに減少中

- 2月～4月頃： 国別の比較や，患者の病状に関心
- 4月～6月頃： 感染拡大・感染モデル等に関心
- 7月～9月頃： 公衆衛生，社会経済系に関心
- その他 ： 検査や，感染機構，データ解析などは概ね一定の割合で推移

社会経済など，長期的課題・研究へとシフト

コンテンツ内容の分布：PPS別



各点が論文に対応し，似たものが近くに配置される
この図での色は，論文が投稿されたPPSに対応


NISTEP
NATIONAL
INSTITUTE OF
SCIENCE AND
TECHNOLOGY
POLICY

国・地域の分析

国・地域分析に関する留意点

※ 既報(*NISTEP Discussion Paper 181*）と同様

- 基本的に 筆頭著者・連絡著者 1名のみを対象に国・地域を判定※
  - 基本的にはメールアドレスを利用（SSRNのみ，所属名を利用）
  - 例えば，所属機関が米国の著者と，日本の著者の共著であっても片方のみ利用
  - 所属機関が米国であっても，Gmail 等のアドレスであると Unknown 判定，
    また，米国機関に所属でも例えば連絡先が出向元の .jp アドレスなら日本判定
    - 結果，全体の2割が国・地域判定不能（Unknown）

DP186 をはじめ，一般的な “論文分析” では
共著者全員を対象に，所属機関国籍で分析

▼

本報は条件が大きく異なり，単純比較が困難な点に注意

商用の論文DBでは著者の国のデータも整備・提供されていることが多いため，共著者を含めた分析も可能だが，
プレプリントはそれらの機械可読なデータが存在しないため，特に複数PPSを対象に同様の分析を行うことは困難

国・地域ごとのPPS別の投稿数

【件】

＊ 所属組織名に基づいて判定

	Region	Total	arXiv	bioRxiv	ChemRxiv	medRxiv	SSRN＊	SSRN Lancet＊
1	USA	4214	499 (11.8%)	661 (15.7%)	53 (1.3%)	1694 (40.2%)	1222 (29.0%)	85 (2.0%)
2	Unknown	3141	762 (24.3%)	184 (5.9%)	79 (2.5%)	1734 (55.2%)	318 (10.1%)	64 (2.0%)
3	China	1725	88 (5.1%)	246 (14.3%)	20 (1.2%)	752 (43.6%)	125 (7.2%)	494 (28.6%)
4	UK	1191	129 (10.8%)	105 (8.8%)	9 (0.8%)	656 (55.1%)	219 (18.4%)	73 (6.1%)
5	India	611	104 (17.0%)	70 (11.5%)	55 (9.0%)	124 (20.3%)	234 (38.3%)	24 (3.9%)
6	Germany	483	75 (15.5%)	87 (18.0%)	0 (0.0%)	206 (42.7%)	97 (20.1%)	18 (3.7%)
7	Italy	446	82 (18.4%)	46 (10.3%)	8 (1.8%)	184 (41.3%)	75 (16.8%)	51 (11.4%)
8	Canada	372	55 (14.8%)	67 (18.0%)	7 (1.9%)	142 (38.2%)	90 (24.2%)	11 (3.0%)
9	France	362	46 (12.7%)	55 (15.2%)	10 (2.8%)	190 (52.5%)	39 (10.8%)	22 (6.1%)
10	Australia	297	41 (13.8%)	31 (10.4%)	3 (1.0%)	107 (36.0%)	104 (35.0%)	11 (3.7%)
11	Brazil	253	48 (19.0%)	29 (11.5%)	5 (2.0%)	127 (50.2%)	34 (13.4%)	10 (4.0%)
12	Spain	248	33 (13.3%)	27 (10.9%)	5 (2.0%)	112 (45.2%)	36 (14.5%)	35 (14.1%)
13	Japan	201	24 (11.9%)	33 (16.4%)	8 (4.0%)	102 (50.7%)	20 (10.0%)	14 (7.0%)
14	Switzerland	164	29 (17.7%)	20 (12.2%)	2 (1.2%)	87 (53.0%)	19 (11.6%)	7 (4.3%)
15	Netherlands	157	15 (9.6%)	32 (20.4%)	1 (0.6%)	61 (38.9%)	39 (24.8%)	9 (5.7%)
16	Israel	87	9 (10.3%)	14 (16.1%)	0 (0.0%)	39 (44.8%)	24 (27.6%)	1 (1.1%)
17	Sweden	87	13 (14.9%)	11 (12.6%)	0 (0.0%)	49 (56.3%)	14 (16.1%)	0 (0.0%)
18	Bangladesh	85	16 (18.8%)	16 (18.8%)	2 (2.4%)	17 (20.0%)	32 (37.6%)	2 (2.4%)
19	Belgium	83	3 (3.6%)	9 (10.8%)	0 (0.0%)	52 (62.7%)	16 (19.3%)	3 (3.6%)
20	Korea	81	9 (11.1%)	18 (22.2%)	2 (2.5%)	23 (28.4%)	17 (21.0%)	12 (14.8%)

- **連絡著者メールアドレス／筆頭著者所属を元に国・地域を推定**
  - Gmail や Hotmail，記載なしについては ”Unknown” と判定（全数の約2割）
- **筆頭など1名のみが対象の点で，一般的な調査と異なる点に注意**

※：企業等(.com, .org, .edu)も可能な限りドメイン登録者国籍から国籍付与，SSRN系はメールアドレスがないため所属組織名から推定

参考：国・地域とトピック

【件】

Region	total	1 国別比較	2 マスク・人工呼吸器	3 社会・経済・政策	4 ワクチン開発	5 感染拡大	6 公衆衛生	7 ゲノム解析	8 情報・データ分析	9 検出・検査	10 感染機構	11 治療薬探索	12 健康・不安	13 肺画像診断	14 患者病状	15 患者治療効果	16 感染モデル
USA	4214	8.5%	3.1%	8.7%	6.8%	13.2%	10.4%	3.6%	6.8%	6.2%	6.3%	3.7%	4.0%	1.8%	6.6%	4.1%	6.2%
Unknown	3141	13.8%	1.8%	3.3%	2.6%	10.0%	3.6%	3.7%	7.8%	3.9%	2.7%	3.3%	6.2%	4.4%	10.7%	6.9%	15.2%
China	1725	10.7%	0.7%	1.7%	7.7%	3.9%	0.8%	3.3%	2.6%	5.4%	7.8%	2.8%	5.4%	3.0%	25.4%	14.7%	4.1%
UK	1191	9.8%	2.8%	5.7%	4.4%	11.7%	7.2%	2.4%	6.9%	5.5%	3.2%	1.8%	11.1%	1.2%	10.1%	9.7%	6.7%
India	611	5.7%	1.6%	5.6%	5.9%	10.1%	7.9%	7.4%	9.8%	2.1%	2.8%	12.1%	4.6%	2.1%	2.1%	2.9%	17.2%
Germany	483	7.2%	2.5%	7.7%	6.2%	11.8%	4.1%	3.1%	7.9%	9.7%	9.9%	3.1%	2.7%	0.8%	4.8%	5.4%	13.0%
Italy	446	11.2%	4.5%	4.0%	4.0%	9.4%	4.3%	6.7%	4.3%	6.3%	5.4%	3.8%	3.8%	1.1%	9.9%	7.2%	14.1%
Canada	372	8.1%	2.4%	6.7%	7.8%	14.0%	9.9%	5.9%	9.1%	8.9%	3.8%	1.6%	5.6%	3.0%	3.0%	5.1%	5.1%
France	362	9.9%	1.4%	1.4%	4.4%	10.8%	3.0%	3.9%	2.8%	9.1%	8.6%	4.7%	2.8%	1.4%	11.0%	7.7%	17.1%
Australia	297	6.1%	1.7%	12.8%	2.7%	14.1%	15.2%	5.7%	9.1%	3.0%	2.7%	1.0%	9.8%	1.3%	2.4%	7.7%	4.7%
Brazil	253	11.9%	1.6%	0.4%	3.2%	13.8%	4.3%	4.0%	8.3%	3.6%	6.7%	3.2%	7.5%	2.4%	4.3%	4.0%	20.9%
Spain	248	12.5%	0.4%	5.2%	3.6%	6.9%	2.8%	1.6%	4.4%	5.6%	5.6%	4.0%	4.8%	2.8%	16.1%	8.5%	14.9%
Japan	201	19.4%	2.0%	2.5%	7.0%	10.4%	1.0%	3.5%	4.5%	13.4%	4.0%	6.5%	6.0%	1.0%	3.5%	8.0%	7.5%
Switzerlan	164	4.3%	4.3%	3.7%	4.3%	12.8%	4.3%	3.0%	7.9%	12.2%	5.5%	1.8%	4.9%	3.7%	7.3%	6.1%	14.0%
Netherland	157	5.7%	3.8%	3.8%	8.9%	12.7%	12.7%	5.7%	8.9%	7.6%	10.2%	2.5%	1.3%	1.3%	5.7%	7.6%	1.3%
Israel	87	13.8%	3.4%	3.4%	6.9%	9.2%	12.6%	2.3%	5.7%	6.9%	3.4%	4.6%	3.4%	0.0%	5.7%	10.3%	8.0%
Sweden	87	6.9%	1.1%	2.3%	3.4%	23.0%	6.9%	1.1%	3.4%	8.0%	18.4%	0.0%	2.3%	0.0%	6.9%	3.4%	12.6%
Bangladesh	85	14.1%	0.0%	12.9%	4.7%	5.9%	8.2%	8.2%	17.6%	0.0%	5.9%	4.7%	5.9%	4.7%	2.4%	1.2%	3.5%
Belgium	83	10.8%	2.4%	9.6%	3.6%	15.7%	1.2%	4.8%	3.6%	7.2%	9.6%	1.2%	1.2%	1.2%	10.8%	8.4%	8.4%
Korea	81	9.9%	1.2%	7.4%	4.9%	14.8%	3.7%	3.7%	6.2%	3.7%	7.4%	8.6%	2.5%	4.9%	8.6%	7.4%	4.9%

- **大まかには分野・トピックとPPSが紐付くため，目立った知見はない**
  - 中国は患者の病状，日本では国別比較に関連する投稿が多い傾向がある

国・地域と投稿数の時系列推移

期間中の投稿数 Top10 に
日本 を 加えた11カ国のリスト

Week		Total	USA	China	UK	India	Germany	Italy	Canada	France	Australia	Brazil	Japan
2020-01-20	(04)	11	3	3	1	0	0	1	0	0	0	0	0
2020-01-27	(05)	28	4	14	1	0	0	0	0	0	0	0	0
2020-02-03	(06)	43	11	19	2	0	0	1	0	0	1	0	1
2020-02-10	(07)	65	12	23	2	0	3	0	0	0	0	0	0
2020-02-17	(08)	120	12	61	4	0	3	0	1	1	1	0	5
2020-02-24	(09)	140	12	89	3	0	0	2	0	0	1	0	2
2020-03-02	(10)	171	18	96	4	1	0	2	3	1	2	0	1
2020-03-09	(11)	169	16	86	8	0	1	4	1	2	0	1	4
2020-03-16	(12)	231	29	81	11	3	2	11	2	4	3	0	4
2020-03-23	(13)	363	73	99	15	6	4	9	10	4	4	4	4
2020-03-30	(14)	486	88	125	12	12	18	13	9	4	9	7	4
2020-04-06	(15)	634	157	101	26	16	20	29	14	12	7	7	2
2020-04-13	(16)	631	146	85	37	17	23	21	17	17	8	8	7
2020-04-20	(17)	604	124	76	31	21	23	14	15	17	11	7	9
2020-04-27	(18)	584	122	52	50	19	16	23	10	19	10	6	2
2020-05-04	(19)	800	181	60	58	27	20	25	21	20	12	9	8
2020-05-11	(20)	831	183	54	54	24	21	23	12	17	11	17	11
2020-05-18	(21)	708	122	23	42	7	11	12	5	20	10	10	7
2020-05-25	(22)	653	100	30	34	14	14	19	10	13	7	15	11
2020-06-01	(23)	542	89	23	30	13	15	14	4	7	7	11	8
2020-06-08	(24)	598	107	23	39	15	12	11	7	12	6	8	8
2020-06-15	(25)	612	110	14	46	11	14	9	10	14	9	14	10
2020-06-22	(26)	476	95	18	30	13	12	9	7	4	6	5	6
2020-06-29	(27)	510	103	18	33	5	14	5	12	8	2	10	3
2020-07-06	(28)	510	99	16	30	4	11	8	7	8	6	3	3
2020-07-13	(29)	433	74	16	33	10	12	4	8	18	6	5	3
2020-07-20	(30)	486	108	20	30	7	10	12	8	11	5	1	5
2020-07-27	(31)	518	122	17	31	8	10	15	7	8	6	6	9
2020-08-03	(32)	630	114	9	29	11	13	7	6	5	3	8	2
2020-08-10	(33)	575	110	19	40	12	11	7	14	9	7	9	1
2020-08-17	(34)	494	91	17	30	9	15	4	7	6	7	4	4
2020-08-24	(35)	456	72	11	20	12	12	6	10	5	3	12	6
2020-08-31	(36)	546	116	25	21	9	11	11	19	10	6	5	3
2020-09-07	(37)	499	89	22	32	11	16	4	12	8	6	10	3
2020-09-14	(38)	474	100	15	28	8	9	3	10	9	9	5	5
2020-09-21	(39)	435	78	15	31	6	10	2	4	7	5	6	4

【件】

- ピークは5月上旬
- 中国は5月までリード
  その後，米国が先導
- 5月からは英国も

まとめ

- 今回のコロナ禍において，研究動向の迅速な把握を企図し，プレプリントを用いた分析を試行
- 論文分析と比べて様々な留意点もあるものの，動向把握は可能で観測しておくべき対象となっていることも確認できた

NATIONAL INSTITUTE OF SCIENCE AND TECHNOLOGY POLICY

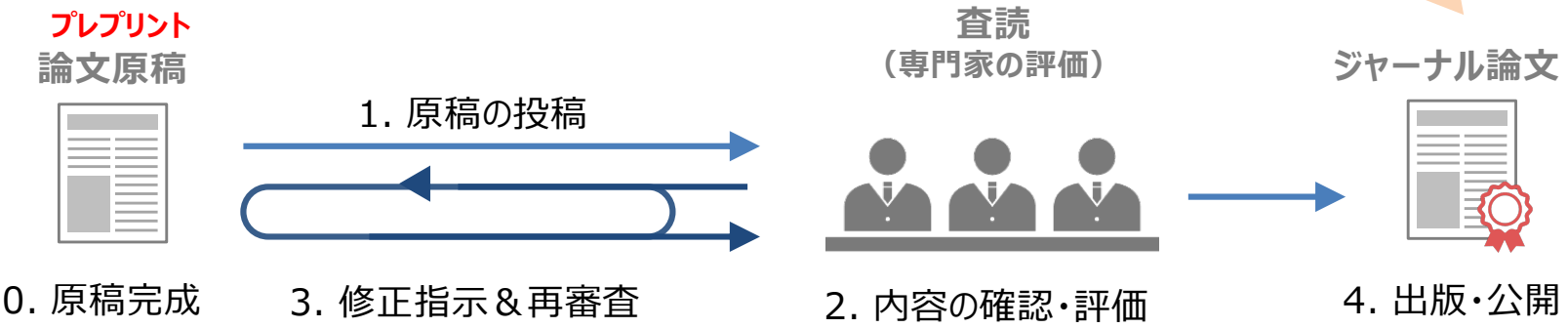
参考資料

プレプリントとは？

- **プレプリント = 査読（第三者による内容確認）前の論文原稿**
- **査読前段階で公開することで先取権を主張できることなどから近年広まりつつある**

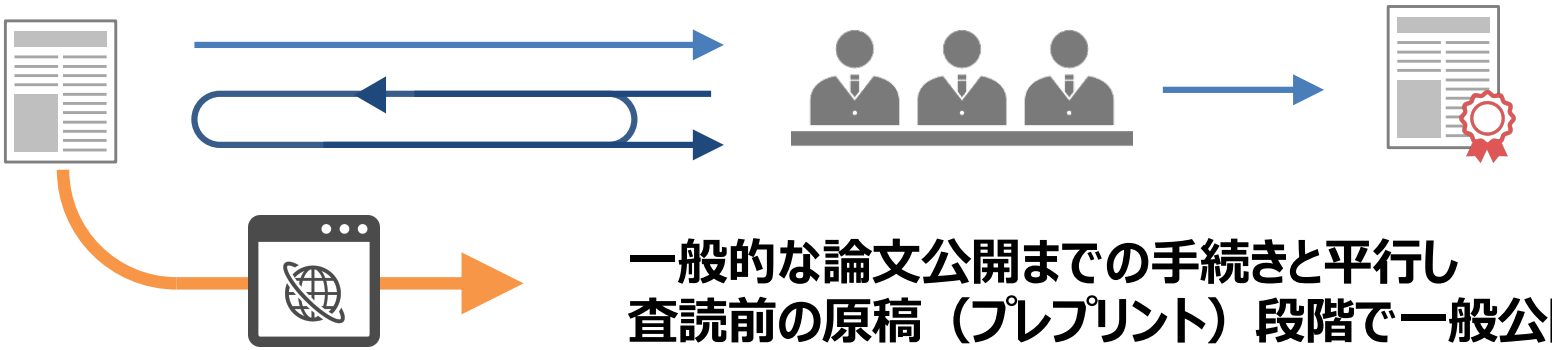
参考： 科学技術・学術審議会 ジャーナル問題検討部会 第7回（令和2年10月27日） https://www.mext.go.jp/content/20201026-mxt_jyohoka01-000010684_2.pdf

一般的な論文公開までの手続き　巧遅（正確性が高まるが、時間もかかる）



公開までのプロセスに概ね数ヶ月、場合により1年以上かかることもある
査読の結果不採録（公開見送り）になるケースも

プレプリントを用いる手続き　拙速＋巧遅（査読前にプレプリントとして公開し、ジャーナル論文の査読プロセスも平行）

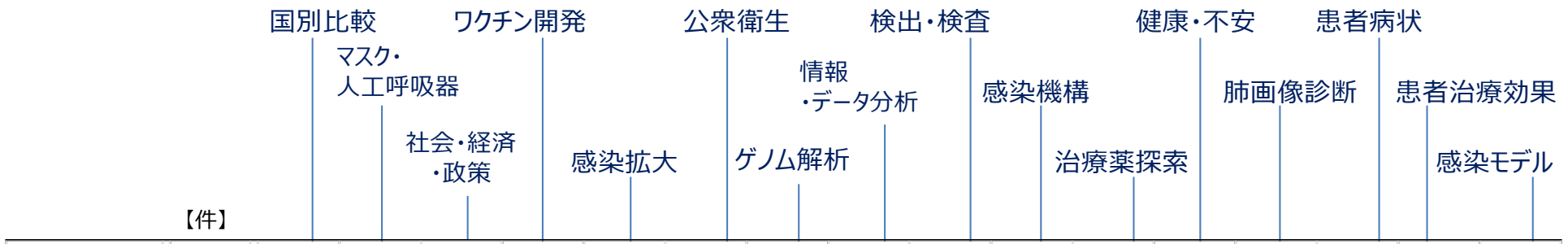


プレプリントサーバー
（プレプリントの公開場所）

一般的な論文公開までの手続きと平行し
査読前の原稿（プレプリント）段階で一般公開

原稿が完成した段階ですぐ公開できる
新型コロナウィルス感染症のような一刻を争うケースではこの速報性が役立つ面も

PPSの種類とトピック



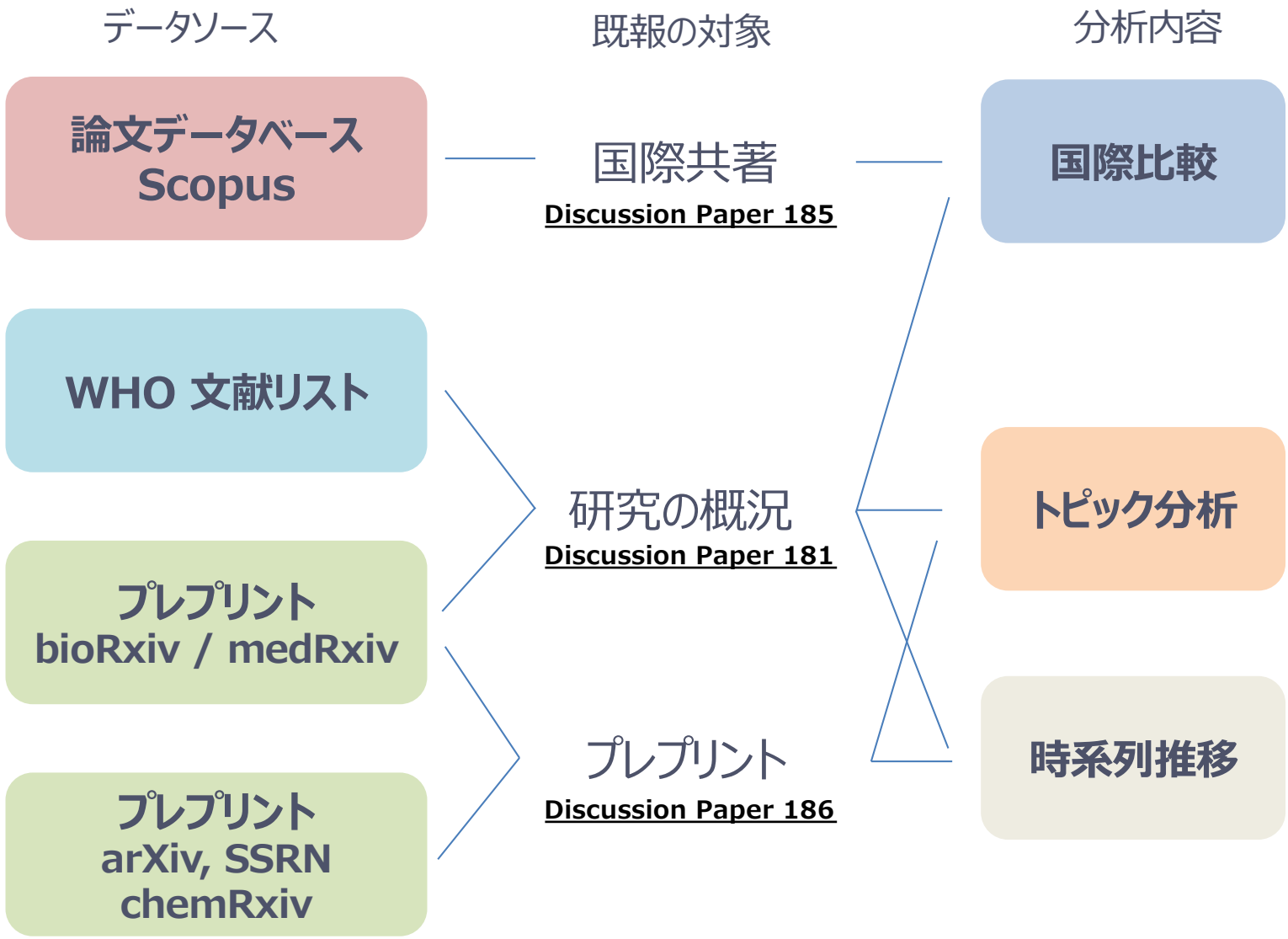
【件】

Kind	Total	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
arXiv	2337	5.6%	2.7%	2.7%	1.3%	15.6%	2.2%	1.5%	24.1%	0.9%	1.1%	2.8%	1.1%	10.3%	0.4%	0.8%	27.0%
bioRxiv	1932	0.3%	1.3%	0.0%	34.4%	0.3%	0.0%	21.4%	1.8%	8.4%	19.0%	11.9%	0.4%	0.3%	0.1%	0.2%	0.2%
ChemRxiv	298	1.3%	2.7%	0.0%	13.1%	0.0%	0.0%	1.3%	3.0%	1.7%	1.0%	74.8%	0.0%	1.0%	0.0%	0.0%	0.0%
medRxiv	7157	16.7%	3.1%	0.1%	0.7%	12.3%	0.2%	2.0%	3.0%	9.5%	4.5%	0.4%	8.0%	2.1%	16.2%	9.8%	11.5%
SSRN	3347	6.2%	0.7%	26.4%	0.9%	13.4%	30.2%	1.3%	7.6%	0.4%	2.3%	1.0%	4.0%	0.1%	1.0%	1.5%	3.2%
SSRN Lancet	1059	9.0%	0.2%	0.0%	0.0%	3.8%	0.5%	0.4%	1.8%	1.9%	2.1%	0.6%	14.8%	0.5%	31.5%	30.8%	2.3%

■ PPSの種別ごとの得意分野とトピックは概ね対応

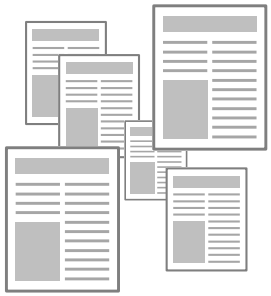
- arXiv　：数理モデル，データ分析に強み
- bioRxiv　：感染機構，ワクチン開発などに強み
- chemRxiv　：治療薬探索に強み
- medRxiv　：臨床を中心に，幅広くカバー
- SSRN　：人文社会系に強み
- SSRN L　：患者病状など臨床系に強み

既報におけるソースと分析内容の関係



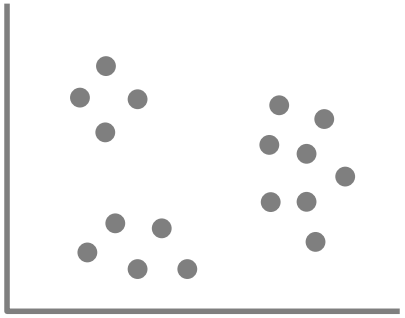
トピック抽出手法の概要① 分析の流れ

― 分散表現と類似度を用いたクラスタリング ―



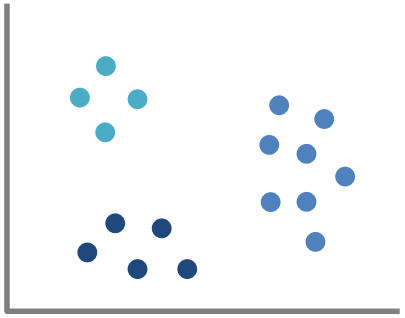
1. 対象文書の収集

fasttext
(word2vec)



2. 対象文書の数値化
（類似するものを近くに配置）

k-means



3. クラスタリング
（クラスタ数は人間が指定）

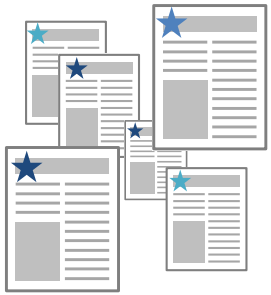
ポイントは，この数値化 部分

4. クラスタごとの頻出語抽出

動物

会議　計算機

5. クラスタの解釈（ラベリング）
（本作業は人間が実施）



6. 文書へのラベリング

トピック抽出手法の概要②
文書の数値化 (1:単語の数値化)

■ **基本戦略は，“同じような単語が出てくる文書は似ている”**

有効に機能するが，“ミカン” と “みかん”，“オレンジ” がすべて別の単語として扱われる
計算機は 記号として単語を見ていて，意味は見ていない（わからない）

■ **単語の類似度を算出する方法 ― 分散表現**

- “ミカン” と “みかん”，“オレンジ” は似たようなものだ… と，計算機に教えれば良い
- ある単語とセットで使われる単語同士は似ている
  - 朝ご飯 に **XX** を 食べた ← **XX** = パン，トースト，おにぎり，鮭 …など
    - パン，トースト，おにぎり，鮭… は，他の単語（例えば，机，本，猫 など）に比べて似ている

ニューラルネットを用いてこの単語の類似度を学習させ，100次元などの高次元空間上にマッピング

この高次元空間へマップした結果を分散表現 (Word Embedding) という

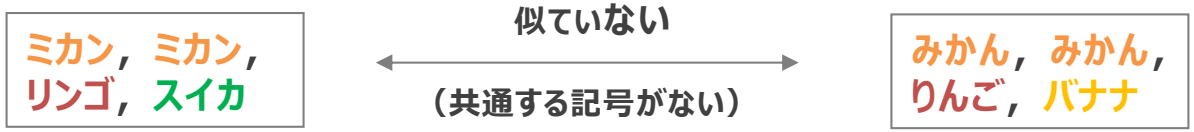
※ 具体的な手法としては word2vec，fasttext, BERT などが知られる

トピック抽出手法の概要③
文書の数値化 （2：文書の数値化）

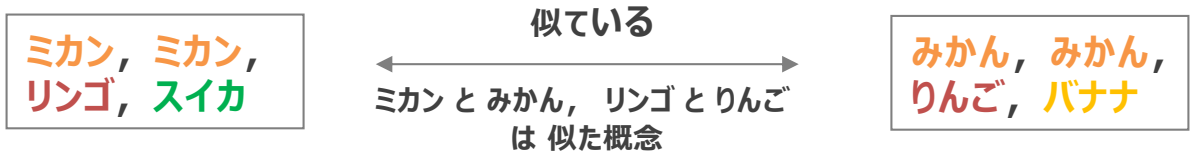
■ **単語の類似度から文書の類似度へ**

- ◆ 文書についても同様に数値化する手法がいくつか存在
- ◆ ここでは単純に，“単語の分散表現” を 平均化したものを “文書の分散表現” に
  - □ 文書中の単語数が同程度で，意味が似ていれば近くに配置されるはず

分散表現を用いない場合



分散表現を用いた場合



トピックの時系列変化（詳細版）

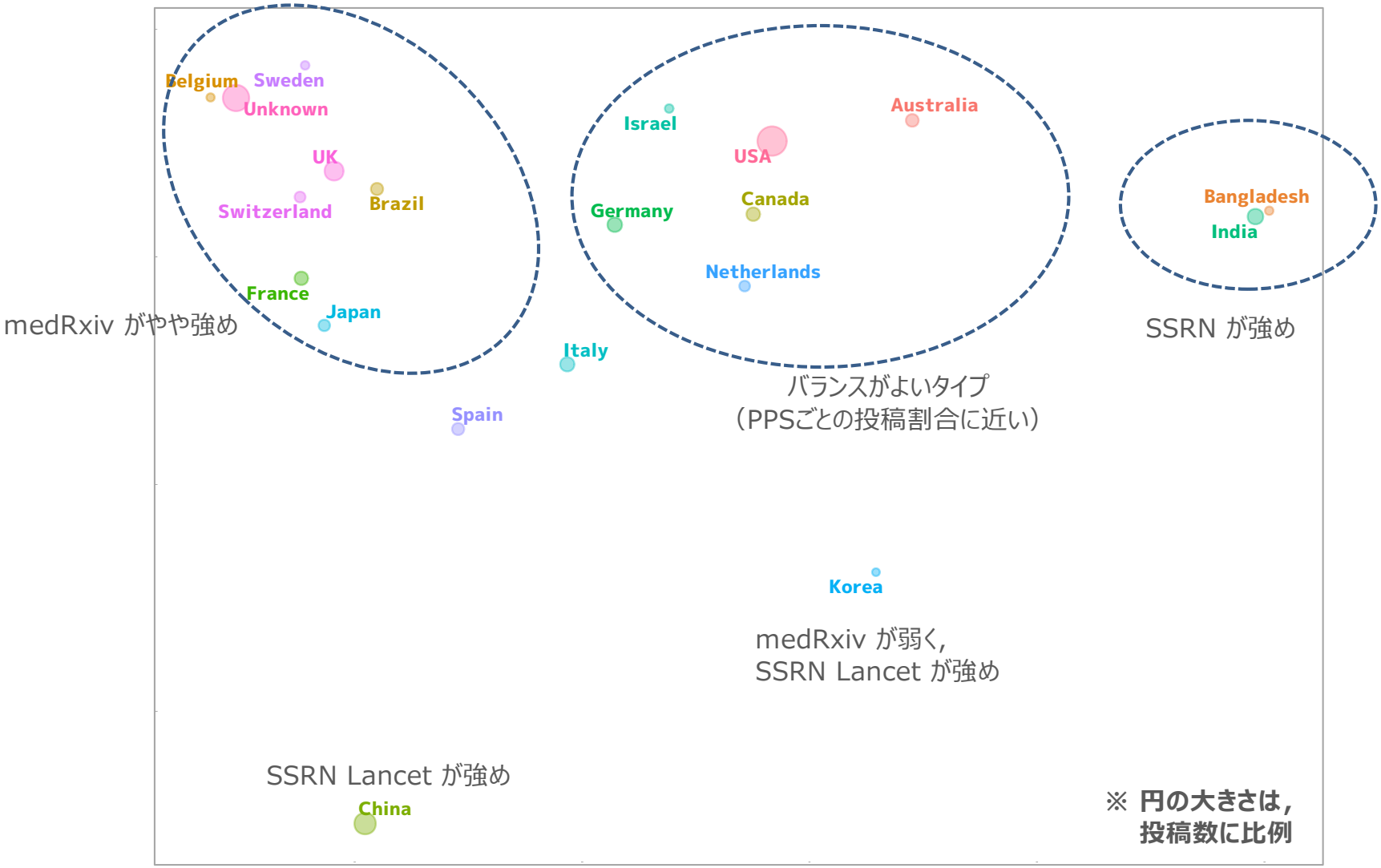
		【件】	1 国別比較	2 マスク・人工呼吸器	3 社会・経済・政策	4 ワクチン開発	5 感染拡大	6 公衆衛生	7 ゲノム解析	8 情報・データ分析	9 検出・検査	10 感染機構	11 治療薬探索	12 健康・不安	13 肺画像診断	14 患者病状	15 患者治療効果	16 感染モデル
Week		Total																
2020-02-17	(08)	120	30.0%	0.0%	0.8%	11.7%	5.0%	0.0%	6.7%	2.5%	3.3%	7.5%	4.2%	3.3%	0.8%	10.8%	6.7%	6.7%
2020-02-24	(09)	140	17.1%	0.0%	0.7%	4.3%	2.9%	0.0%	2.9%	2.1%	7.9%	4.3%	2.9%	2.9%	0.7%	23.6%	15.7%	12.1%
2020-03-02	(10)	171	18.1%	0.6%	1.8%	7.0%	5.3%	0.6%	7.6%	1.2%	4.1%	4.7%	2.9%	4.7%	1.2%	22.2%	11.7%	6.4%
2020-03-09	(11)	169	17.8%	0.6%	3.0%	5.9%	5.9%	1.2%	3.0%	1.8%	4.7%	3.6%	5.9%	5.3%	1.2%	18.3%	12.4%	9.5%
2020-03-16	(12)	231	16.0%	1.7%	1.7%	4.8%	9.1%	1.7%	6.1%	4.3%	5.6%	4.3%	5.2%	4.3%	1.7%	6.5%	12.6%	14.3%
2020-03-23	(13)	363	14.9%	1.4%	2.8%	5.2%	10.5%	3.3%	4.1%	6.3%	3.9%	3.3%	5.2%	3.0%	5.0%	12.4%	8.5%	10.2%
2020-03-30	(14)	486	11.9%	1.9%	3.5%	4.7%	13.4%	2.9%	3.3%	4.5%	5.1%	4.1%	4.7%	4.1%	3.5%	10.5%	5.6%	16.3%
2020-04-06	(15)	634	12.1%	2.1%	1.7%	5.8%	12.8%	2.7%	2.7%	7.1%	4.9%	3.8%	7.1%	2.5%	3.8%	9.9%	7.1%	13.9%
2020-04-13	(16)	631	13.5%	3.3%	3.0%	4.3%	11.7%	3.5%	3.3%	6.5%	3.5%	5.1%	5.9%	2.5%	3.6%	8.1%	7.3%	14.9%
2020-04-20	(17)	604	11.1%	3.5%	3.0%	5.5%	11.1%	3.1%	5.0%	6.6%	5.3%	5.1%	5.1%	4.5%	3.5%	8.8%	6.0%	12.9%
2020-04-27	(18)	584	12.3%	1.2%	4.5%	3.3%	13.9%	2.7%	6.2%	7.5%	6.8%	3.3%	4.6%	3.4%	2.4%	8.7%	6.2%	13.0%
2020-05-04	(19)	800	11.5%	1.1%	4.8%	3.3%	12.4%	6.4%	4.0%	5.1%	6.8%	4.0%	3.0%	6.5%	3.8%	8.1%	7.6%	11.8%
2020-05-11	(20)	831	11.8%	2.3%	5.2%	5.1%	12.4%	7.2%	3.5%	5.8%	4.1%	5.3%	3.9%	5.9%	2.5%	9.3%	6.5%	9.4%
2020-05-18	(21)	708	10.5%	2.3%	7.2%	4.0%	11.0%	6.4%	1.6%	10.0%	5.1%	5.1%	4.8%	4.8%	1.3%	10.0%	4.9%	11.2%
2020-05-25	(22)	653	9.3%	2.5%	6.1%	3.8%	13.9%	8.0%	3.2%	8.1%	6.7%	3.8%	3.8%	5.1%	1.5%	9.3%	4.4%	10.3%
2020-06-01	(23)	542	7.9%	2.6%	6.6%	2.8%	10.9%	6.8%	3.1%	6.6%	7.9%	3.9%	1.7%	7.2%	2.4%	12.0%	7.6%	10.0%
2020-06-08	(24)	598	12.0%	2.0%	5.2%	5.5%	11.7%	4.7%	3.7%	6.7%	5.7%	4.7%	1.3%	9.0%	2.3%	11.2%	6.0%	8.2%
2020-06-15	(25)	612	10.1%	2.6%	6.7%	5.9%	8.7%	4.4%	6.5%	7.4%	7.4%	6.0%	2.1%	6.9%	1.8%	9.2%	5.9%	8.5%
2020-06-22	(26)	476	9.0%	2.7%	6.1%	4.2%	11.8%	5.3%	4.4%	8.4%	4.0%	6.5%	2.1%	5.9%	4.2%	10.1%	5.0%	10.3%
2020-06-29	(27)	510	7.1%	1.4%	8.0%	4.1%	12.0%	8.2%	3.7%	7.1%	6.5%	5.9%	2.0%	5.3%	1.4%	10.4%	6.5%	10.6%
2020-07-06	(28)	510	9.6%	2.4%	8.6%	5.1%	9.8%	7.5%	4.5%	7.3%	6.9%	5.7%	1.4%	5.9%	2.5%	8.8%	6.9%	7.3%
2020-07-13	(29)	433	11.5%	3.2%	2.5%	4.6%	7.9%	3.7%	2.8%	5.1%	4.6%	7.6%	2.3%	9.0%	3.0%	12.9%	10.2%	9.0%
2020-07-20	(30)	486	5.8%	2.1%	6.8%	6.6%	12.1%	5.8%	3.7%	9.5%	4.5%	5.8%	3.9%	7.2%	2.1%	7.4%	7.0%	9.9%
2020-07-27	(31)	518	7.7%	1.9%	5.0%	6.9%	8.9%	3.1%	3.9%	7.5%	6.2%	5.2%	3.7%	5.6%	1.9%	11.8%	9.3%	11.4%
2020-08-03	(32)	630	6.3%	2.5%	11.9%	3.5%	9.7%	13.7%	3.0%	7.5%	7.0%	4.9%	2.7%	6.5%	1.6%	7.3%	5.9%	6.0%
2020-08-10	(33)	575	6.4%	1.7%	8.5%	4.3%	9.9%	10.6%	2.4%	6.4%	5.2%	4.7%	4.2%	6.6%	2.4%	9.7%	8.3%	8.3%
2020-08-17	(34)	494	5.5%	2.4%	8.1%	6.1%	8.5%	16.4%	4.3%	7.5%	5.9%	5.5%	2.8%	4.9%	2.6%	7.1%	5.1%	7.5%
2020-08-24	(35)	456	7.2%	2.4%	8.8%	10.1%	8.8%	14.5%	4.4%	8.3%	3.9%	4.4%	4.4%	5.5%	1.5%	4.8%	5.5%	5.5%
2020-08-31	(36)	546	5.9%	2.4%	6.0%	4.6%	11.9%	10.3%	4.0%	7.5%	7.1%	5.7%	2.0%	5.7%	2.6%	9.7%	8.1%	6.6%
2020-09-07	(37)	499	6.2%	2.2%	9.8%	7.6%	10.0%	10.6%	4.2%	9.0%	4.2%	6.2%	1.8%	7.0%	3.2%	7.0%	4.0%	6.8%
2020-09-14	(38)	474	9.1%	1.1%	7.0%	5.5%	9.9%	10.8%	2.7%	7.4%	6.3%	7.0%	3.8%	4.9%	3.4%	8.2%	7.0%	6.1%
2020-09-21	(39)	435	6.9%	3.2%	10.1%	5.1%	9.4%	10.1%	4.8%	3.9%	6.0%	6.0%	3.9%	9.4%	2.3%	7.8%	5.7%	5.3%

PPSの種類と時系列変化

Week		Total	arXiv	bioRxiv	ChemRxiv	medRxiv	SSRN	Lancet
2020-01-20	(04)	11	0	9	1	0	1	0
2020-01-27	(05)	28	2	16	1	3	5	1
2020-02-03	(06)	43	4	13	0	16	5	5
2020-02-10	(07)	65	13	11	4	27	2	8
2020-02-17	(08)	120	7	24	6	59	4	20
2020-02-24	(09)	140	10	9	3	84	6	28
2020-03-02	(10)	171	10	24	6	76	10	45
2020-03-09	(11)	169	24	22	2	81	13	27
2020-03-16	(12)	231	42	27	10	106	20	26
2020-03-23	(13)	363	91	31	13	133	46	49
2020-03-30	(14)	486	105	47	18	198	50	68
2020-04-06	(15)	634	120	85	25	300	53	51
2020-04-13	(16)	631	102	70	20	334	70	35
2020-04-20	(17)	604	116	84	24	280	55	45
2020-04-27	(18)	584	111	54	14	309	73	23
2020-05-04	(19)	800	97	76	13	420	149	45
2020-05-11	(20)	831	100	104	21	395	188	23
2020-05-18	(21)	708	118	62	20	293	188	27
2020-05-25	(22)	653	83	59	15	333	155	8
2020-06-01	(23)	542	89	50	0	239	111	53
2020-06-08	(24)	598	88	66	0	267	127	50
2020-06-15	(25)	612	79	99	1	264	141	28
2020-06-22	(26)	476	80	66	0	202	113	15
2020-06-29	(27)	510	79	68	0	216	128	19
2020-07-06	(28)	510	75	60	0	199	136	40
2020-07-13	(29)	433	58	50	0	206	67	52
2020-07-20	(30)	486	81	63	11	191	100	40
2020-07-27	(31)	518	91	71	7	232	77	40
2020-08-03	(32)	630	69	64	11	213	236	37
2020-08-10	(33)	575	71	63	10	219	172	40
2020-08-17	(34)	494	56	70	4	186	162	16
2020-08-24	(35)	456	55	78	10	136	149	28
2020-08-31	(36)	546	48	55	8	286	136	13
2020-09-07	(37)	499	70	65	9	201	144	10
2020-09-14	(38)	474	54	60	6	214	115	25
2020-09-21	(39)	435	39	55	4	198	120	19

- medRxiv の割合が大きいが，
  arXiv, medRxiv も含めて，
  4月末～5月上旬が投稿数のピーク
- SSRN はやや遅れて，
  5月中旬以降，と 8月以降に盛り上がり
- ChemRxiv は4月末～5月上旬と，
  8月上旬にわずかに盛り上がるが低調

国・地域とPPSの関係性



- **国・地域ごとに，各PPSにどのくらいの割合で投稿しているかについて多次元尺度法を用い，類似するものを近くに配置 （＝投稿割合の類似度マップ）**

メールアドレスからの国・地域名推定フロー