Topic Modelling for Vaccine Safety Signal Detection

Sedigheh Khademi, Pari Delir Haghighi

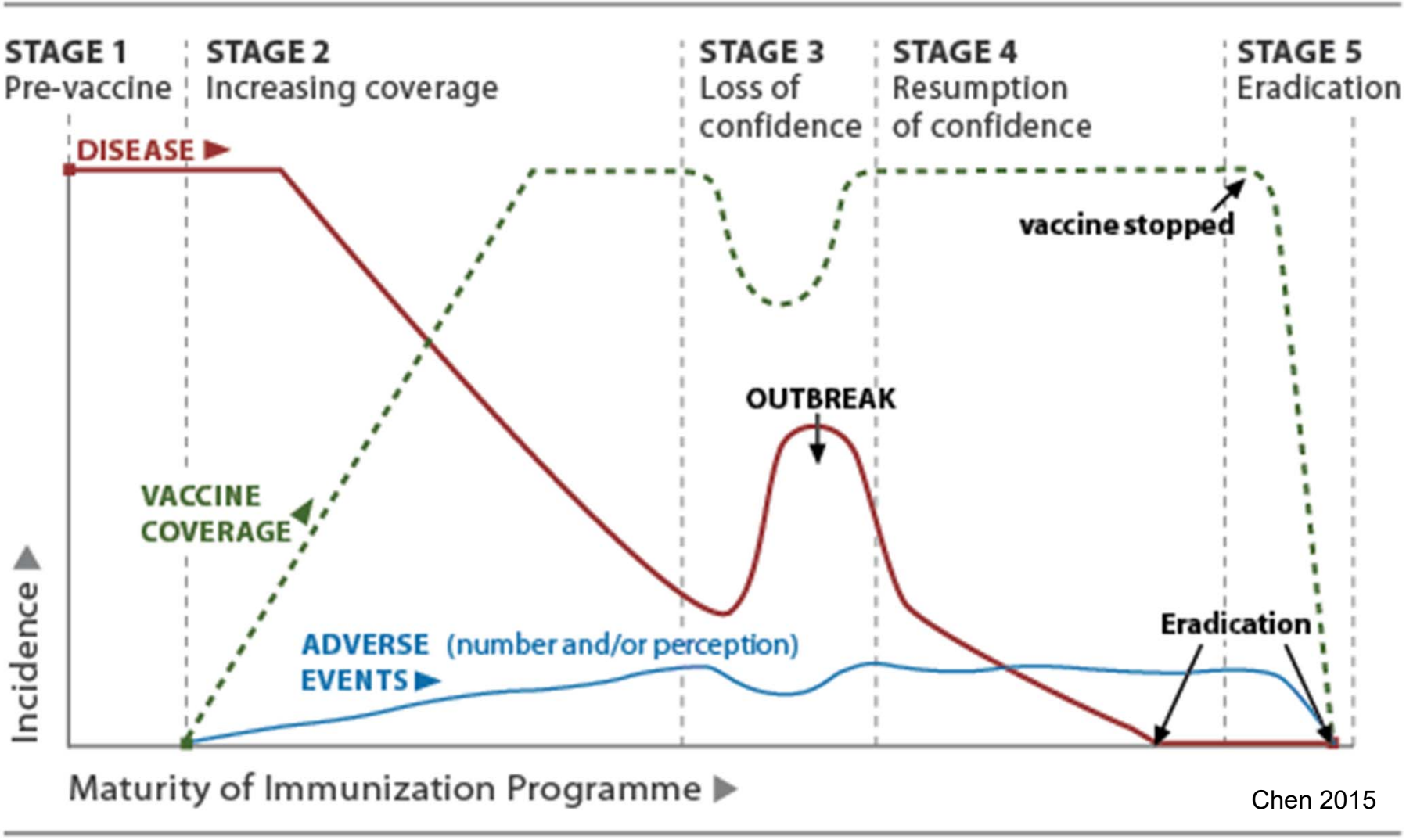
Monash University

Jan 2019

Vaccines Importance

alterdoctor.in/sentinel

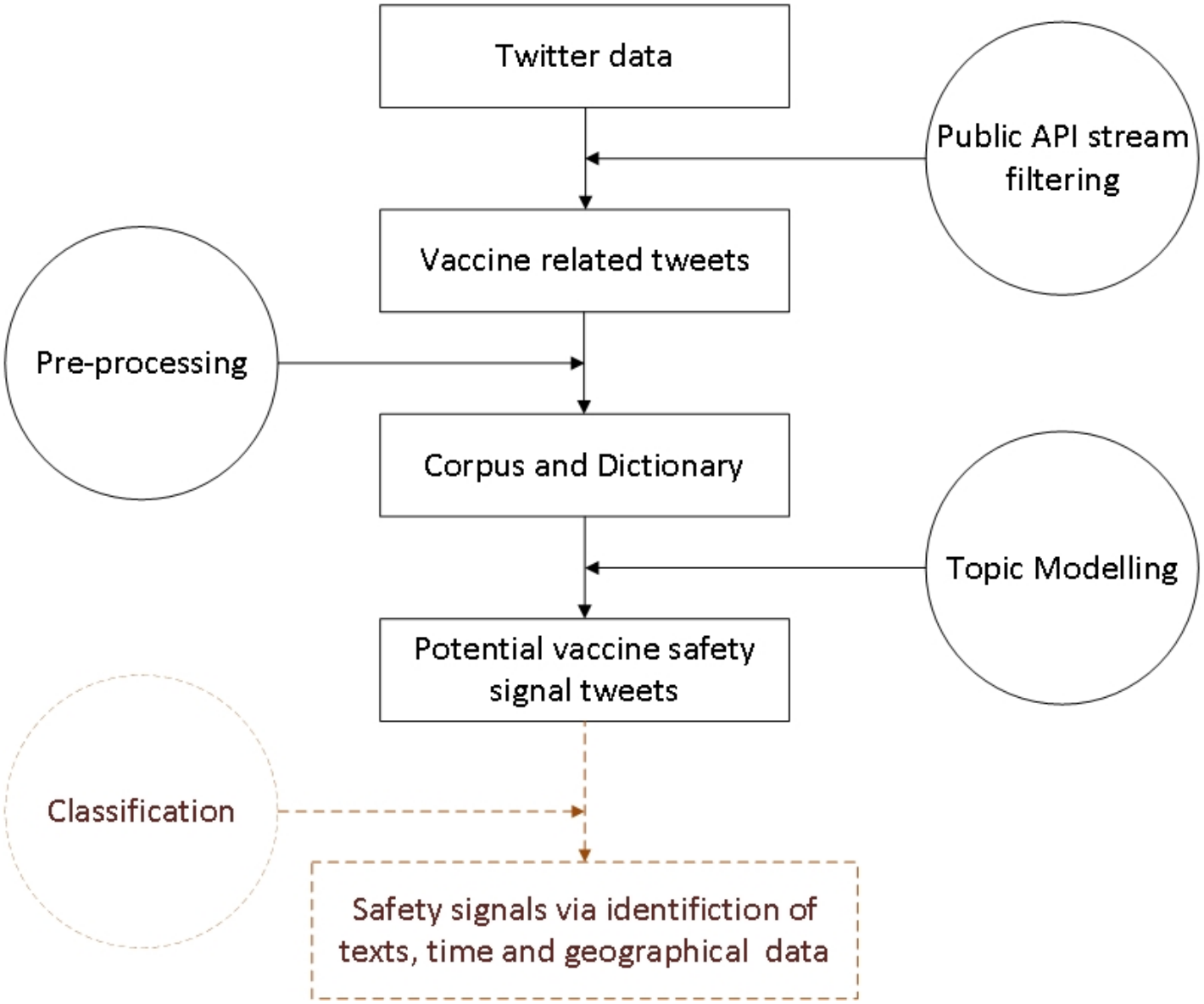
Potential evolution of an immunization program



Chen 2015

Social Media as a data source

Workflow


Twitter data
Public API stream filtering
Vaccine related tweets
Pre-processing
Corpus and Dictionary
Topic Modelling
Potential vaccine safety signal tweets
Classification
Safety signals via identifiction of texts, time and geographical data

Data

- Twitter data was collected from February to May 2018
- 341,507 tweets were collected
- Keyword search for "vaccination, vaccinations, vaccine, vaccines, vax, vaxx, vaxine, vaccinated, vacinated, flushot, flu shot"

Labelling data

Label	Topic
0	Vaccine Safety Signals
1	Enquiries / Discussions mentioning vaccines
2	Obvious sentiment against vaccines – anti-vax
3	Sentiment against anti-vax viewpoints, pro vaccines
4	Statements from vaccine related organizations
5	News articles and other factual or fake news
6	Nonsense / Spoof hijacking Vaccine meme
7	Everything else
11	Animal related
12	Advertising
99	Unlabelled data

Potential safety signal posts

Vaccinations suck when they make ur baby sick :(
I got a flu shot Tuesday and my arm seriously hurts so bad.
I got my flu shot on March 1 (03/01). By Friday (03/09), I had noticeably developed a runny nose, cough and sore throat.

Data Cleaning

- Removed duplicates
- Converted from Unicode to plain text
- Converted to lower case
- Eliminated URLs
- Removed the retweet tag and @user references and the hash symbol from hashtags
- Replaced text-based emoticons with plain English:

Vaccinations suck when they make ur baby sick :(

→

Vaccinations suck when they make ur baby sick <sad>

Data Cleaning

- Removed posts:
  - having less than five words:
    vaccines might cause autism
  - with a high number of non-unique (therefore repeated) words:
    Get your shots shots shots shots shots shots
  - with a low number of English words:
    Nice milestone for NVAXNVAX IBB XBIXBI FBT

Data was reduced from 341,507 to 329,842 documents

Data preparation for topic modelling

- Contractions were expanded prior to tokenisation:

Don’t →do not

they'd → they would

Data preparation for topic modelling

- Stop words were removed, except for “do” and “not”
  The source was nltk.corpus library, examples:

  me
  a
  if
  or
  as
  to
  from
  some

Data preparation for topic modelling

- Bigrams and trigrams were created

  Bigrams:

  arm_hurt

  side_effect

  sore_throat

  Trigrams:

  measles_mumps_rubella

  compromised_immune_system

  highly_contagious_respiratory

Data preparation for topic modelling

- Data was tokenized and lemmatized, retaining nouns, adjectives, verbs and adverbs

- A dictionary of lemmatized terms was constructed containing 100,148 tokens

- Dictionary was trimmed, removing words that occur in less than 20 or in more than 50% of the documents.

Data preparation for topic modelling

Token count: 9,986

Document count: 328,822

Topic modelling

- Aim to discover and annotate large archive of documents
- Set of techniques that analyse the words of the documents to discover the themes that run through them
- The number of topics to be discovered is predefined
- Do not require any prior annotation or labelling
- A document may have multiple topics

Topic modelling algorithms

Python library for:

- Scalable statistical semantics
- Analyze plain-text documents for semantic structure
- Retrieve semantically similar documents

MALLET is a Java-based package for:

- Statistical natural language processing
- Document classification
- Clustering, topic modeling, information extraction

JLDADMM

A Python library ideal for topic modelling on short text

What makes a useful model

- Differentiates topics in an understandable way
- Creates enough topics so that there is a clear difference between them, not so many that the differences are only slight
- Isolates topics that clearly identify documents containing safety signal

Topic model evaluation

- Manual examination of the topics and the most dominant words
- Coherence score
- Comparison with annotated documents

The best model's labelled document distribution

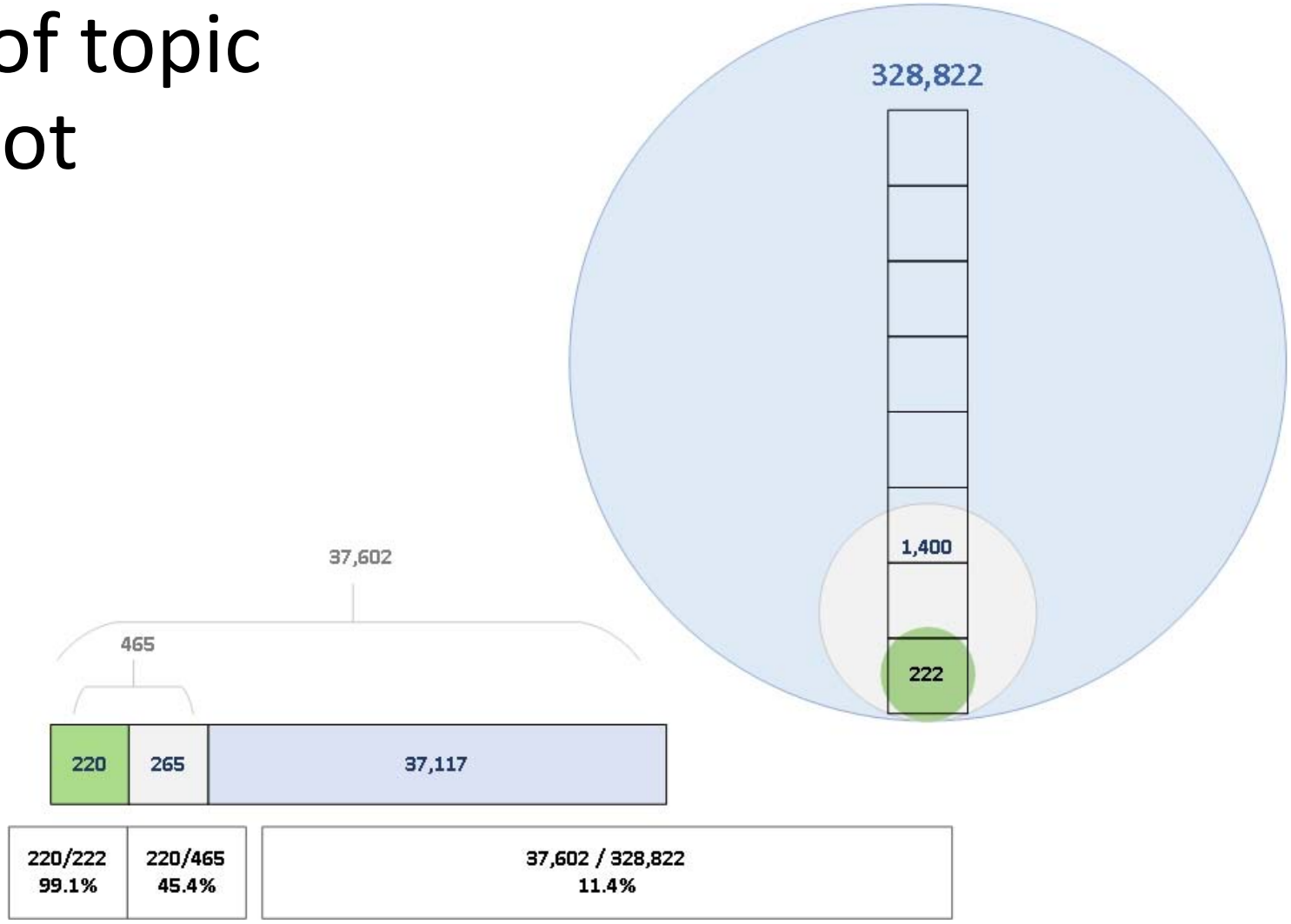
14 Topic model - the best topic for labelled records is get_flu_shot

	Labels										Labelled		Grand
Topics	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**11**	**12**	**Total**	**99**	**Total**
get_flu_shot	**220**	166	15	18	1	0	64	0	1	0	**485**	37,117	37,602
autism	0	44	36	141	0	0	220	0	1	0	442	46,407	46,849
not_vaccination	0	36	60	14	0	9	45	2	0	0	166	36,001	36,167
health_child	0	15	18	5	0	6	12	0	0	0	56	25,831	25,887
flu_shot	0	29	3	6	2	2	12	0	0	1	55	23,623	23,678
disease_child	1	22	3	8	2	1	9	0	0	0	46	20,372	20,418
pets_ads	0	10	0	1	4	1	13	0	2	1	32	19,136	19,168
vax_new	1	5	0	0	0	0	21	3	0	0	30	16,260	16,290
free_vaccination	0	17	0	0	5	2	0	0	0	2	26	18,195	18,221
cancer_research	0	6	0	0	10	4	2	0	0	0	22	19,528	19,550
research	0	5	0	0	4	3	2	1	0	0	15	22,647	22,662
vaccines_work	0	4	0	2	1	2	1	0	0	0	10	17,286	17,296
outbreak	0	3	0	0	0	2	3	0	0	0	8	13,098	13,106
hpv_cancer	0	5	0	1	0	0	1	0	0	0	7	11,921	11,928
Grand Total	**222**	**367**	**135**	**196**	**29**	**32**	**405**	**6**	**4**	**4**	**1,400**	**327,422**	**328,822**

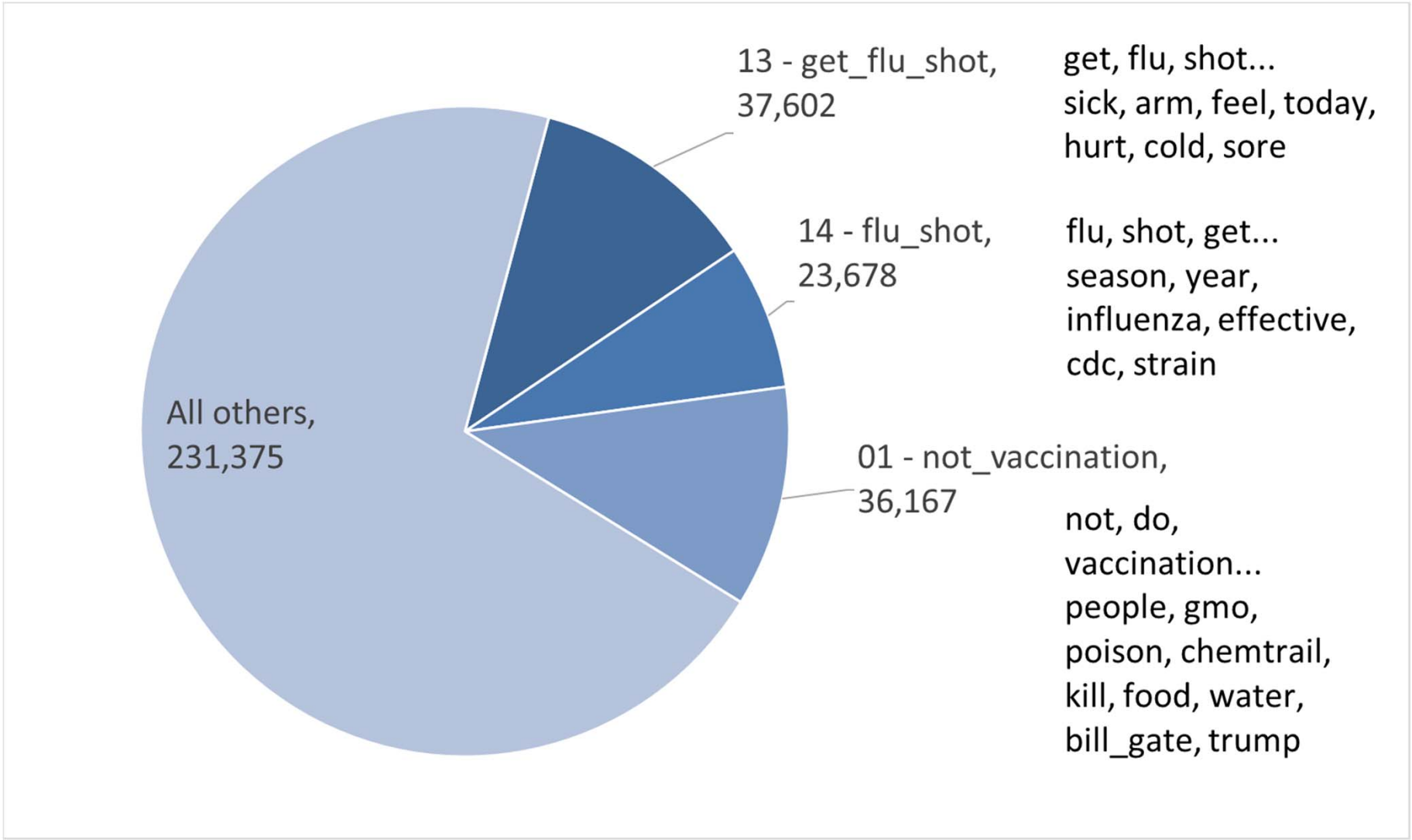
Labelling data

Label	Topic
0	Vaccine Safety Signals
1	Enquiries / Discussions mentioning vaccines
2	Obvious sentiment against vaccines – anti-vax
3	Sentiment against anti-vax viewpoints, pro vaccines
4	Statements from vaccine related organizations
5	News articles and other factual or fake news
6	Nonsense / Spoof hijacking Vaccine meme
7	Everything else
11	Animal related
12	Advertising
99	Unlabelled data

Summary of topic get_flu_shot

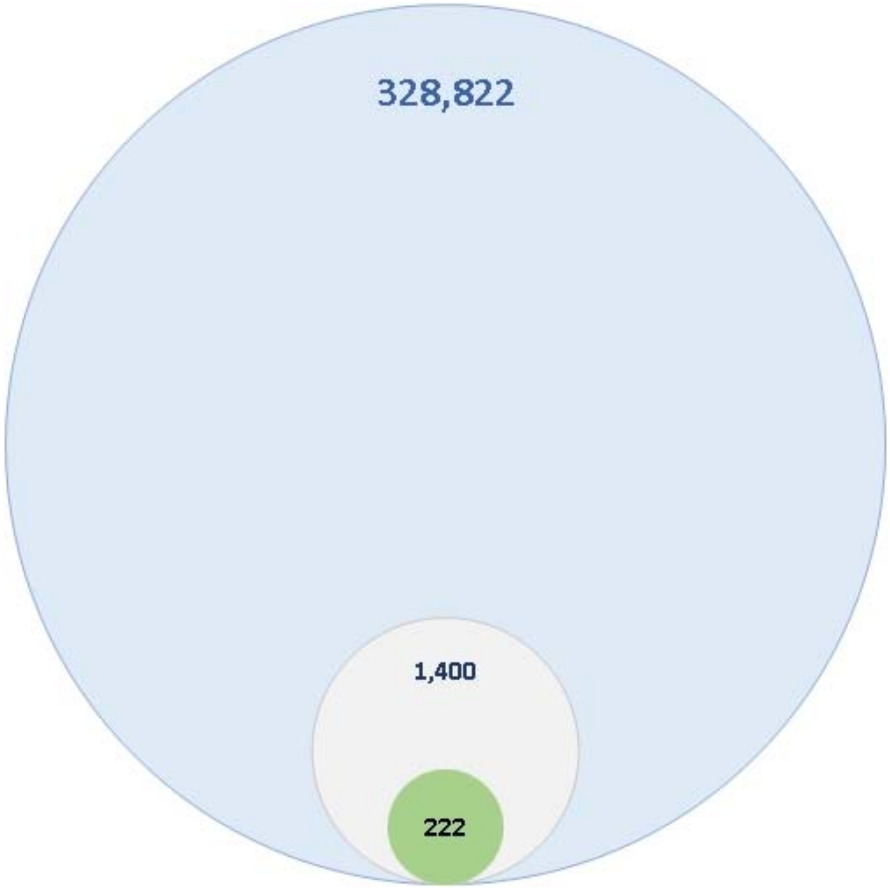


Examples of model's topics with their most dominant words:



Document counts

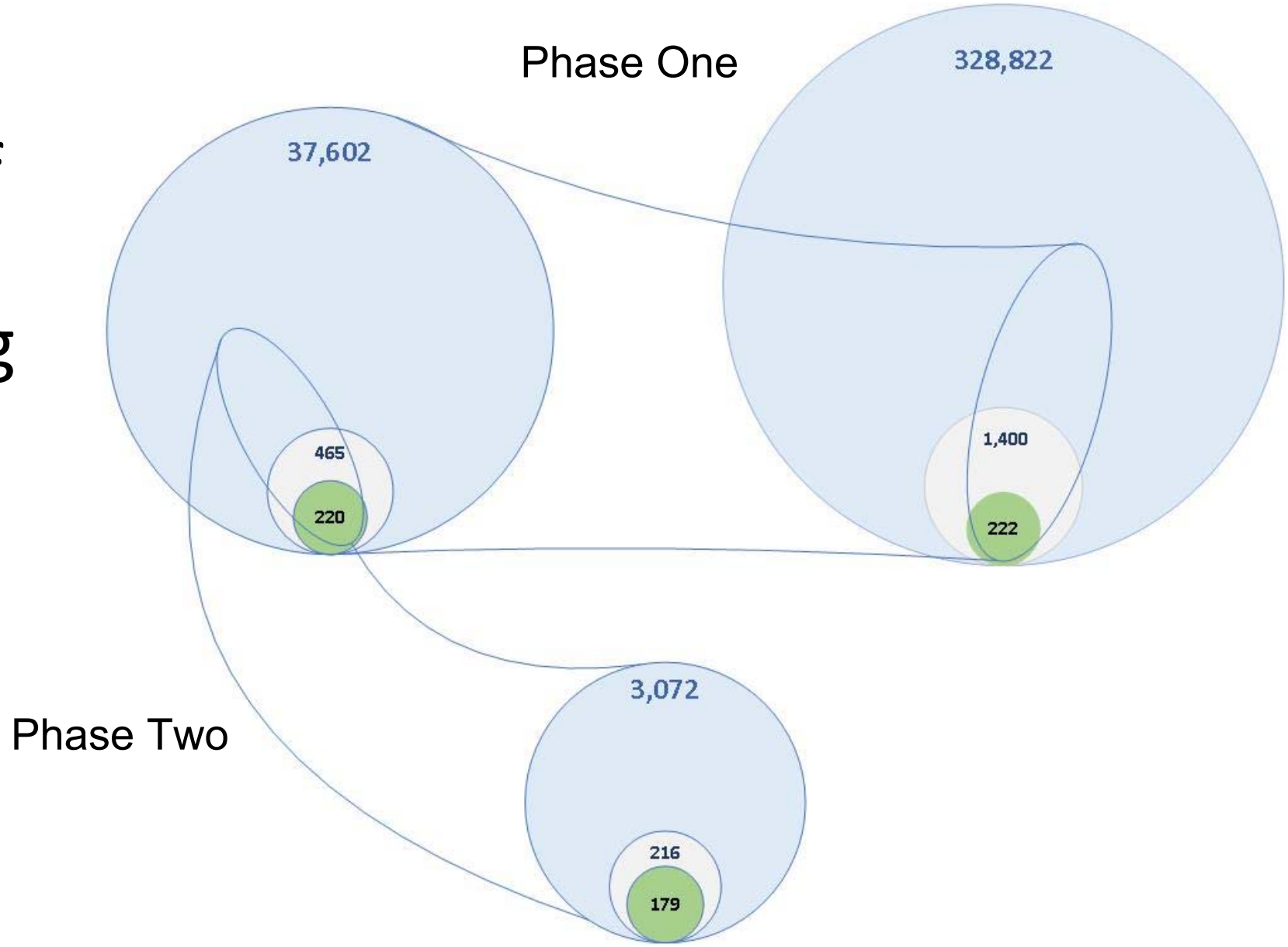
- 328 thousand documents
- 1400 labelled
- 222 safety signals



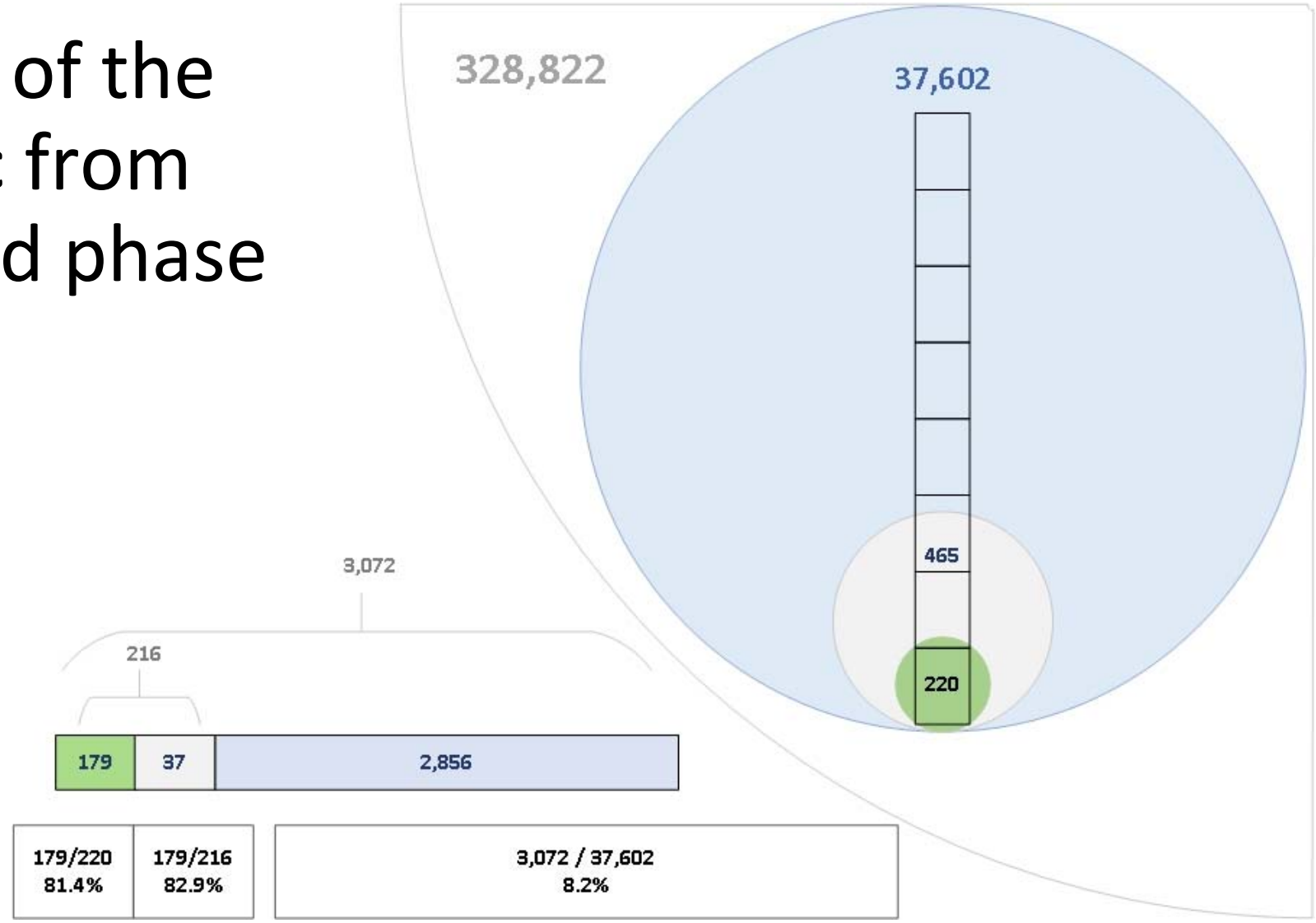
Summary of first phase calculations over labelled data

Topic	Safety signal	Other labels	Labelled Total	Un-labelled	Grand Total	
get_flu_shot - topic 13	220	265	485	37,117	37,602	**45.4%**
Other topics	2	913	915	290,305	291,220	
Grand Total	**222**	**1,178**	**1,400**	**327,422**	**328,822**	
	99.1%					

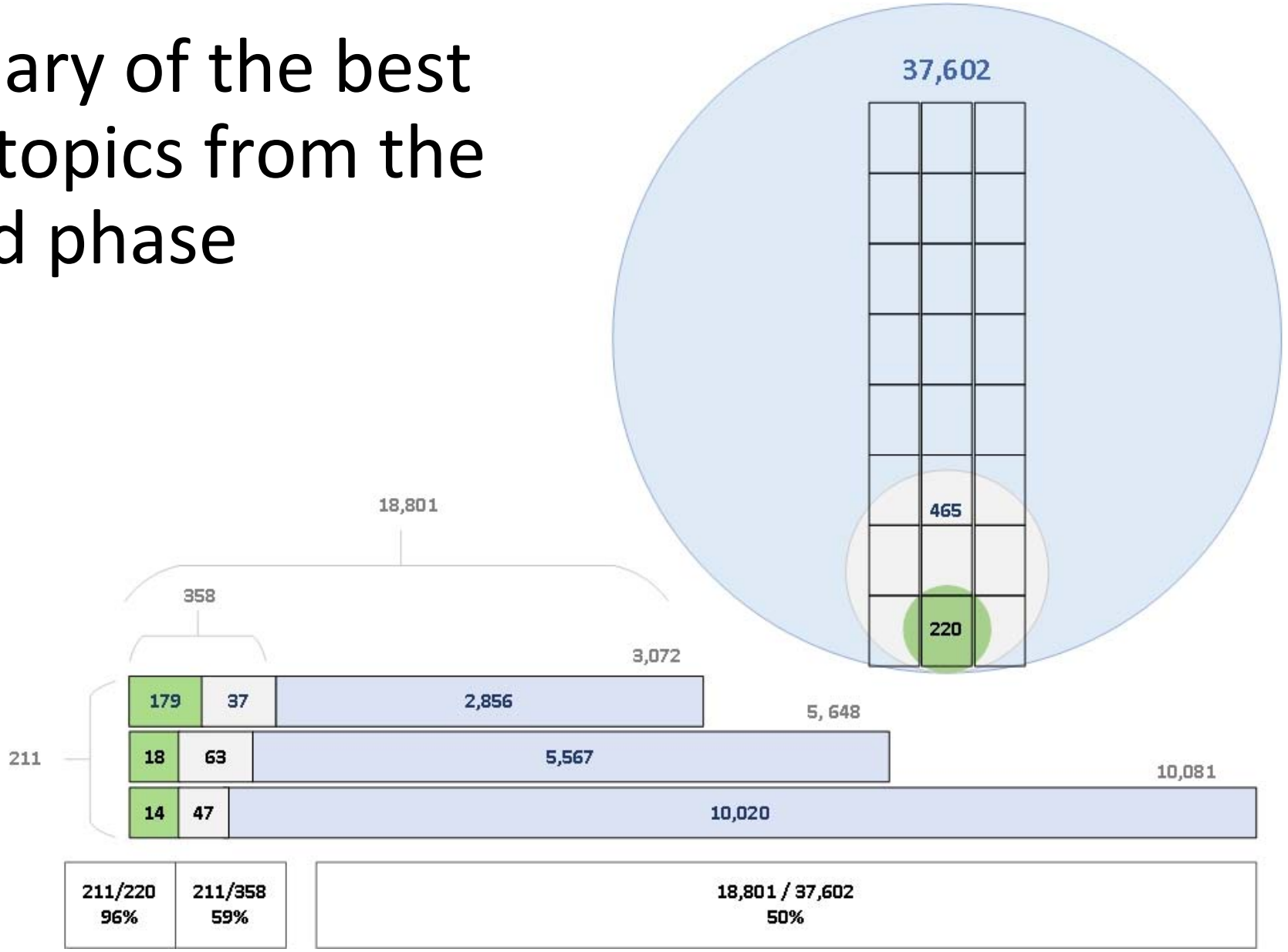
Two phases of topic modelling



Summary of the best topic from the second phase



Summary of the best three topics from the second phase



Summary of second phase calculations

Topic	Safety signal	Other labels	Labelled Total	Un-labelled	Grand Total	
Topic 8	179	37	216	2,856	3,072	**82.9%**
Topics 9 and 1	32	110	142	15,587	15,729	
Other topics	9	118	127	18,674	18,801	
Grand Total	**220**	**265**	**485**	**37,117**	**37,602**	
	81.4%					

Final calculations after manual labelling

Topic	Safety signal	Other labels	Grand Total	
Topic 8	1,300	1,772	3,072	**42.3%**
Topics 9 and 1	432	15,297	15,729	
Grand Total	**1,732**	**17,069**	**18,801**	
	75.1%			

So far:

Established the efficacy of topic modelling for isolating potential vaccine safety signals

Where to:

- Apply classification methods to improve the detection of documents containing safety signals
- Combine the topic and classification models into the social media data stream download process
- Integrate the process into the Adverse Events Following Immunisation – Clinical Assessment Network (AEFI-CAN)