Allied Data Science Communities

DON'T PANIC!

WE ARE DATA SCIENTISTS

Allied Data Science Communities

VDSG
VIENNA DATA SCIENCE GROUP

BUDAPEST.AI

FRANKFURT
DATA SCIENCE

DON’T PANIC, WE ARE DATA SCIENTISTS!

KAGGLE COVID-19 CHALLENGE INTRO

FRANKFURT DATA SCIENCE

VIENNA DATA SCIENCE GROUP

BUDAPEST.AI

BUDAPEST DEEP LEARNING READING SEMINAR

INTRO

LEVENTE SZABADOS

"...originally Buddhist theologian and programmer, AI professional (NLP), lead of research, lecturer, startupper, ex-CTO

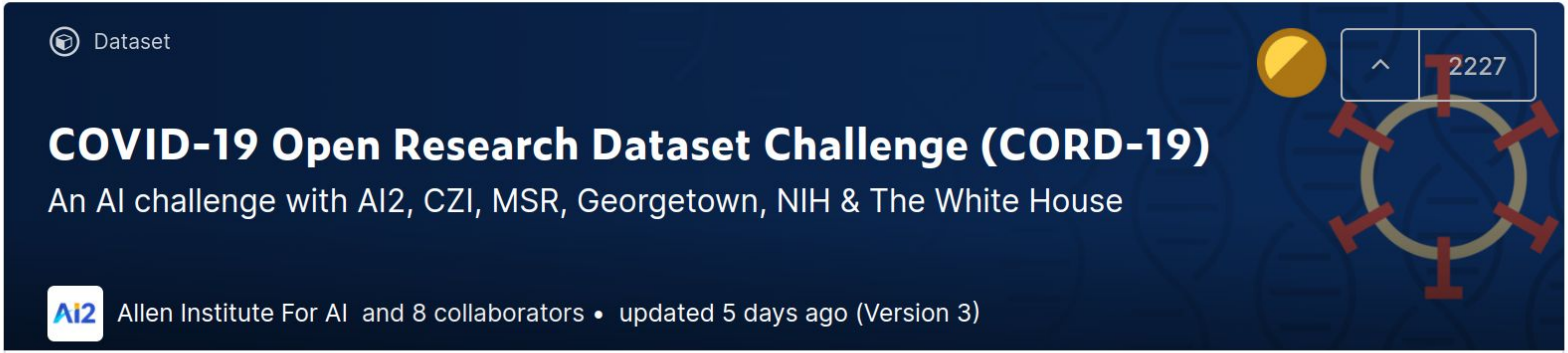
Presently:
Lecturer: Frankfurt School of Finance and Management,
Specialization leader: KÜRT Academy,
Senior Consultant: Neuron Solutions,
Chief organizer: Budapest.AI."

CONTACT

FRANKFURT
DATA SCIENCE

VDSG
VIENNA DATA SCIENCE GROUP

THE CHALLENGE



Dataset

2227

COVID-19 Open Research Dataset Challenge (CORD-19)

An AI challenge with AI2, CZI, MSR, Georgetown, NIH & The White House

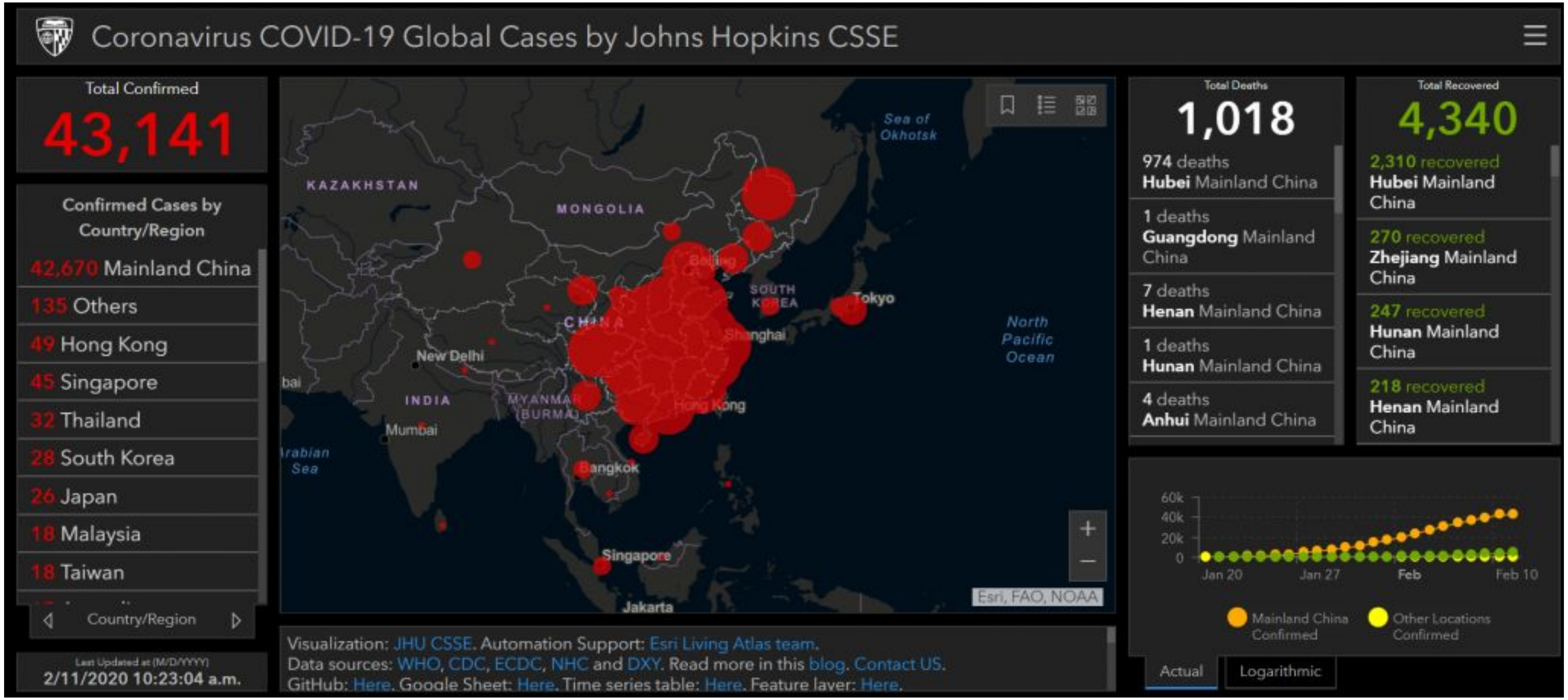
Ai2 Allen Institute For AI and 8 collaborators • updated 5 days ago (Version 3)

https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge

Challenge link:

This presentation:

THERE IS UP TO DATA CASE DATA AVAILABLE



Coronavirus COVID-19 Global Cases by Johns Hopkins CSSE

Total Confirmed

43,141

Confirmed Cases by Country/Region

42,670 Mainland China
135 Others
49 Hong Kong
45 Singapore
32 Thailand
28 South Korea
26 Japan
18 Malaysia
18 Taiwan

Country/Region

Last Updated at (M/D/YYYY)
2/11/2020 10:23:04 a.m.

Total Deaths

1,018

974 deaths
Hubei Mainland China

1 deaths
Guangdong Mainland China

7 deaths
Henan Mainland China

1 deaths
Hunan Mainland China

4 deaths
Anhui Mainland China

Total Recovered

4,340

2,310 recovered
Hubei Mainland China

270 recovered
Zhejiang Mainland China

247 recovered
Hunan Mainland China

218 recovered
Henan Mainland China

Visualization: JHU CSSE. Automation Support: Esri Living Atlas team.
Data sources: WHO, CDC, ECDC, NHC and DXY. Read more in this blog. Contact US.
GitHub: Here. Google Sheet: Here. Time series table: Here. Feature layer: Here.

VISUALIZATION BY JHU: https://coronavirus.jhu.edu/map.html

DATA ON GITHUB:

https://github.com/CSSEGISandData/COVID-19

BIT HIGH LEVEL

PROBLEM FORMULATION

- There is a huge amount of literature being produced in a stream about COVID-19 and it's broader context.

- From a scientific perspective, the rapid pace of produced information can easily be overwhelming.

- We, from the AI and NLP community should lend a helping hand by mining knowledge from research papers and answer specific task questions.

DETAILS:
Challenge description

MORE LIKE COOPERATION

“COMPETITION”

- Kaggle spirit, open and cooperative

- Prize is 1000USD per task, but can be offered for charity

- Very few restrictions apply (eg. over 18 years of age)

- Tasks are evaluated individually, according to task specific (somewhat vague) criteria (understandably)

DETAILS:
Challenge description

PILE OF TEXT + METADATA

THE DATASET

All in all 29000 articles + metadata

Full text, in json: **13202**

"Majority of files are in json format.

The files are grouped in 4 folders and 4 tar archives."

Approx. 2GB JSON

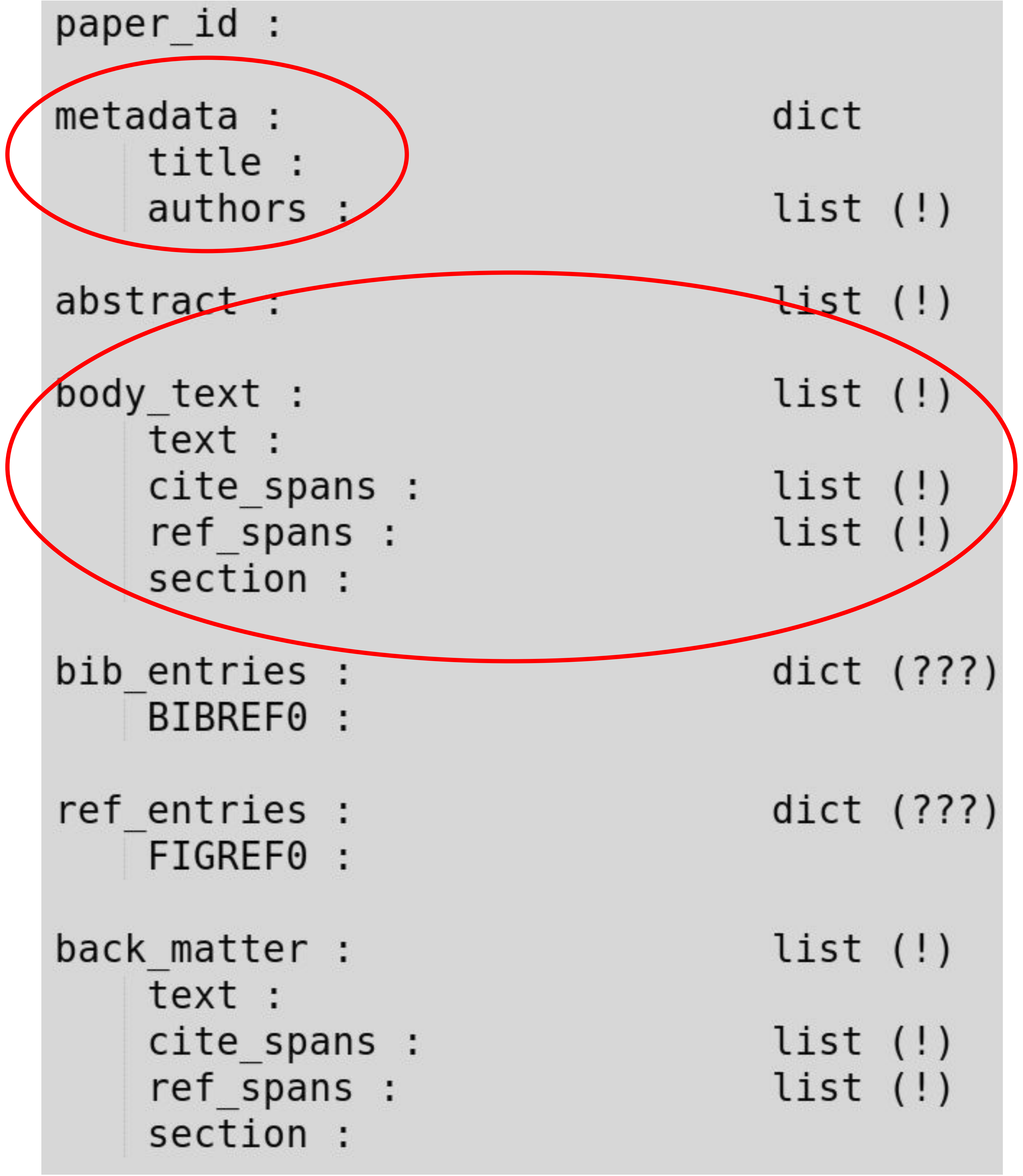
Pretty decent corpus with ~4.5M (?) tokens, ~385k vocabulary, but beware, multiple domains!

DETAILS:
Challenge description

One notebook with some descriptives

STRUCTURE OF THE JSONS

THE DATASET



```
paper_id :

metadata :                    dict
    title :
    authors :                 list (!)

abstract :                    list (!)

body_text :                   list (!)
    text :
    cite_spans :              list (!)
    ref_spans :               list (!)
    section :

bib_entries :                 dict (???)
    BIBREF0 :

ref_entries :                 dict (???)
    FIGREF0 :

back_matter :                 list (!)
    text :
    cite_spans :              list (!)
    ref_spans :               list (!)
    section :
```

Observations:

- Metadata, like author, citation (citation graph can be built)

- Text is not in one chunk, needs to be merged

- Also rich reference information is mapped "onto" the text

- One has to pay attention to dict vs. list usage

DETAILS:

Challenge description

One notebook with some descriptives

Notebook with a citation network iomplementation

VERY BROAD QU ESTIONS!

THE TASKS

10 "Tasks" in total:

1. What is known about transmission, incubation, and environmental stability?
2. What do we know about COVID-19 risk factors?
3. What do we know about virus genetics, origin, and evolution?
4. Sample task with sample submission - Help us understand how geography affects virality.
5. What do we know about non-pharmaceutical interventions?
6. What do we know about vaccines and therapeutics?
7. What do we know about diagnostics and surveillance?
8. What has been published about information sharing and inter-sectoral collaboration?
9. What has been published about ethical and social science considerations?
10. What has been published about medical care?

Observations:

- Very broadly defined tasks!

- Some hint towards very specific propositions (or entities, relations) to mine for (Like eg. 6)

- Some are more open ended, general, like "What has been said about...", hint to "summaries" or textual parts (Like eg. 8)

- No easy mapping between taks and NLP methods, so requires thought to re-formulate task as well as to reason about the results

DETAILS:

Task definitions

AND NOT JUST EVEN ONE!

THE MAIN CHALLENGES - SPECIALIZED DOMAIN(S!)

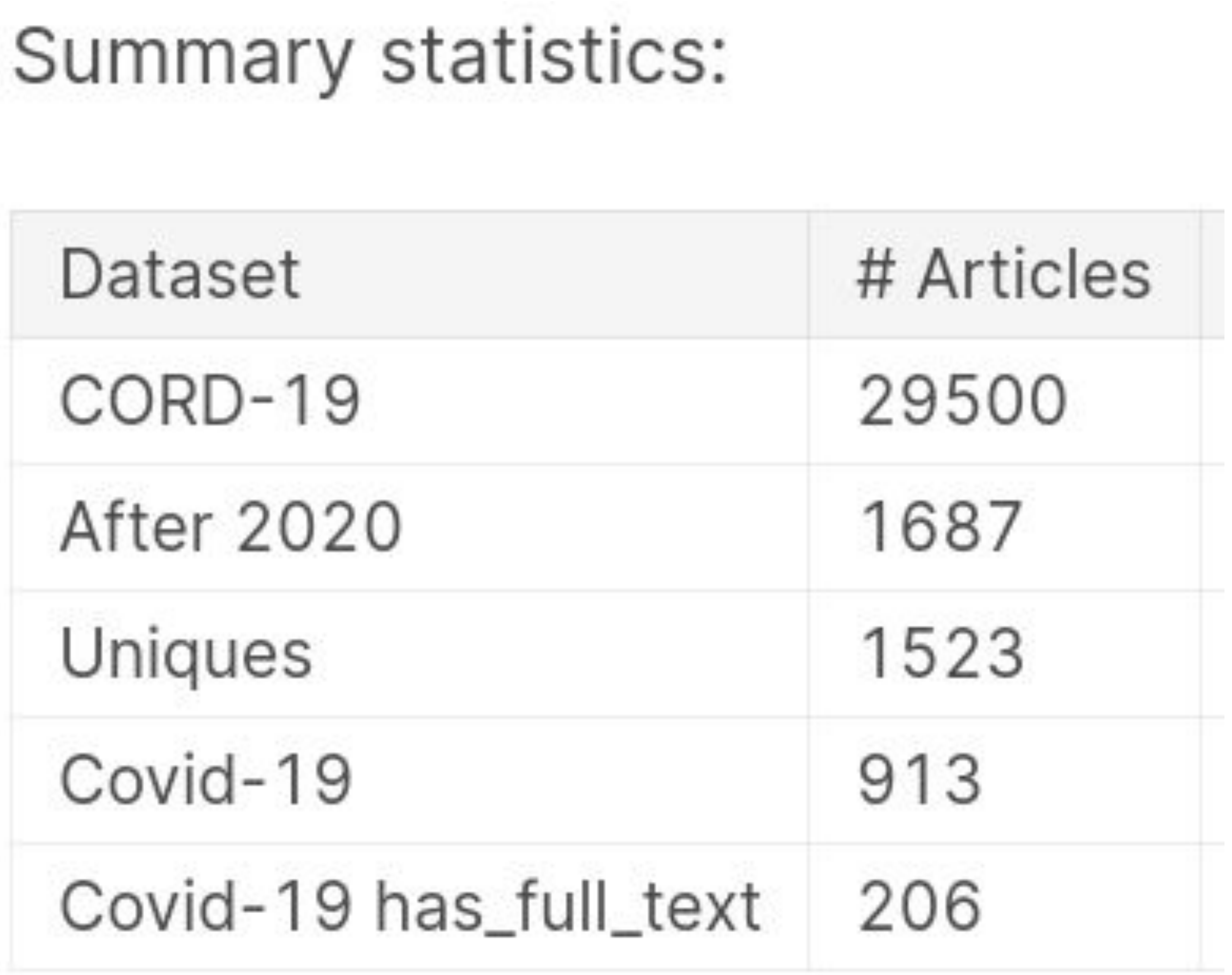
Observations:

- Specialized domain(s!!!)

    - Different from general language, so out of the box external resources have to be adapted (think word vectors, or something like SciBERT?)

    - Specialized taxonomy, complex Multi-Word Expressions
        - If available, taxonomical resources have to be adapted, but maybe even unavailable
        - This is maybe the most interesting data, so good Multi-Word Expression / keyword handling is of importance

- Multiple domains, not just strict epidemIology!



Summary statistics:

Dataset	# Articles
CORD-19	29500
After 2020	1687
Uniques	1523
Covid-19	913
Covid-19 has_full_text	206

DETAILS:

Notebook source for the table right
Notebook with SciBERT embeddings

WHAT CAN WE REALLY SAY?

THE MAIN CHALLENGES - SPECIFIC ANSWERS?

Observations:

- It can be, that we are mining for "specifics"

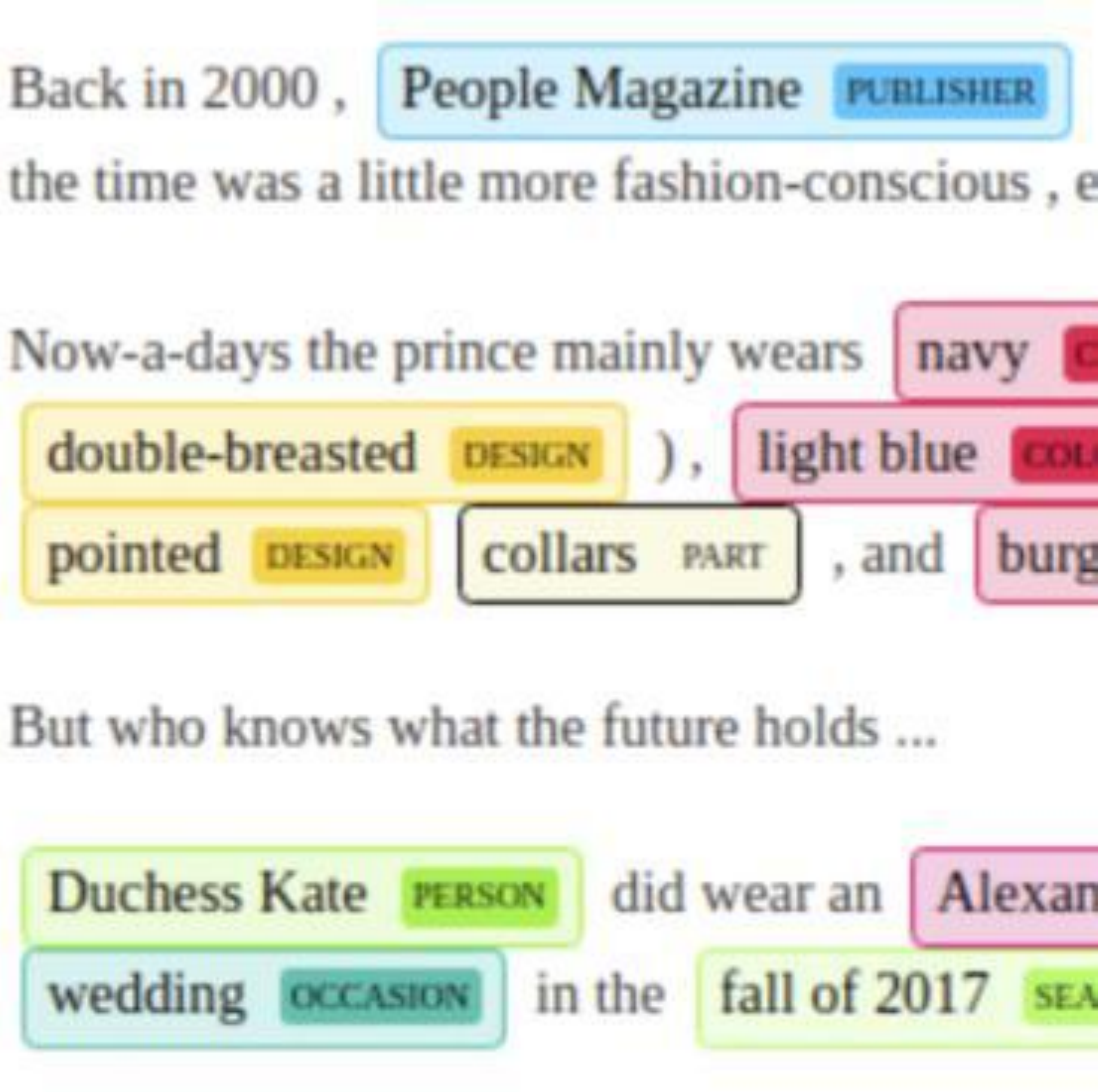
- Single propositions may be of strong interest

("The administration of ...(drug)... had significant effect.")

- Sometimes these may come from tabular like parts of the text

(Have to be checked!)

- Parameters and numeric values can also be of interest ("X dose of Y")



Back in 2000 , People Magazine PUBLISHER

the time was a little more fashion-conscious , e

Now-a-days the prince mainly wears navy C

double-breasted DESIGN) , light blue COL

pointed DESIGN collars PART , and burg

But who knows what the future holds ...

Duchess Kate PERSON did wear an Alexan

wedding OCCASION in the fall of 2017 SEA

DETAILS:

One notebook with NER training in SpaCy

WHAT ELSE CAN WE REALLY SAY?

THE MAIN CHALLENGES - OTHER CONSIDERATIONS

Observations:

- Merging of the citation graph information with the textual can be valuable (think:weighting of a texts's importance?), but non-trivial

- The temporal order of incoming information is highly relevant, it can be, that later findings override the validity of previous propositions.

- Quantification of uncertainity regarding information can be of high value

WHAT SHOULD WE TRY?

POSSIBLE AVENUES OF “ATTACK”

Directions it might be worth pondering:

- Topic mining (+topic changes in time?)
- Text summarization (extractive?) methods for a relevant topic
- Search solutions or Q&A models with supporting evidence
    - Domain adapted semantic vectors

- Visualization of co-occurrence graph and frequency RELEVANT information

- Custom trained Named Entity Recognizer on entities of interest
- Multi-Word Expression detection and keyword extraction techniques
- Knowledge Graph Mining, Fact Extraction and it's relevant versions

WHAT SHOULD WE TRY?

WHY PARTICIPATE?

- Learn (yourself and from others) & practice

- A chance to learn about different industries (not just data science)

- Be part of the global community

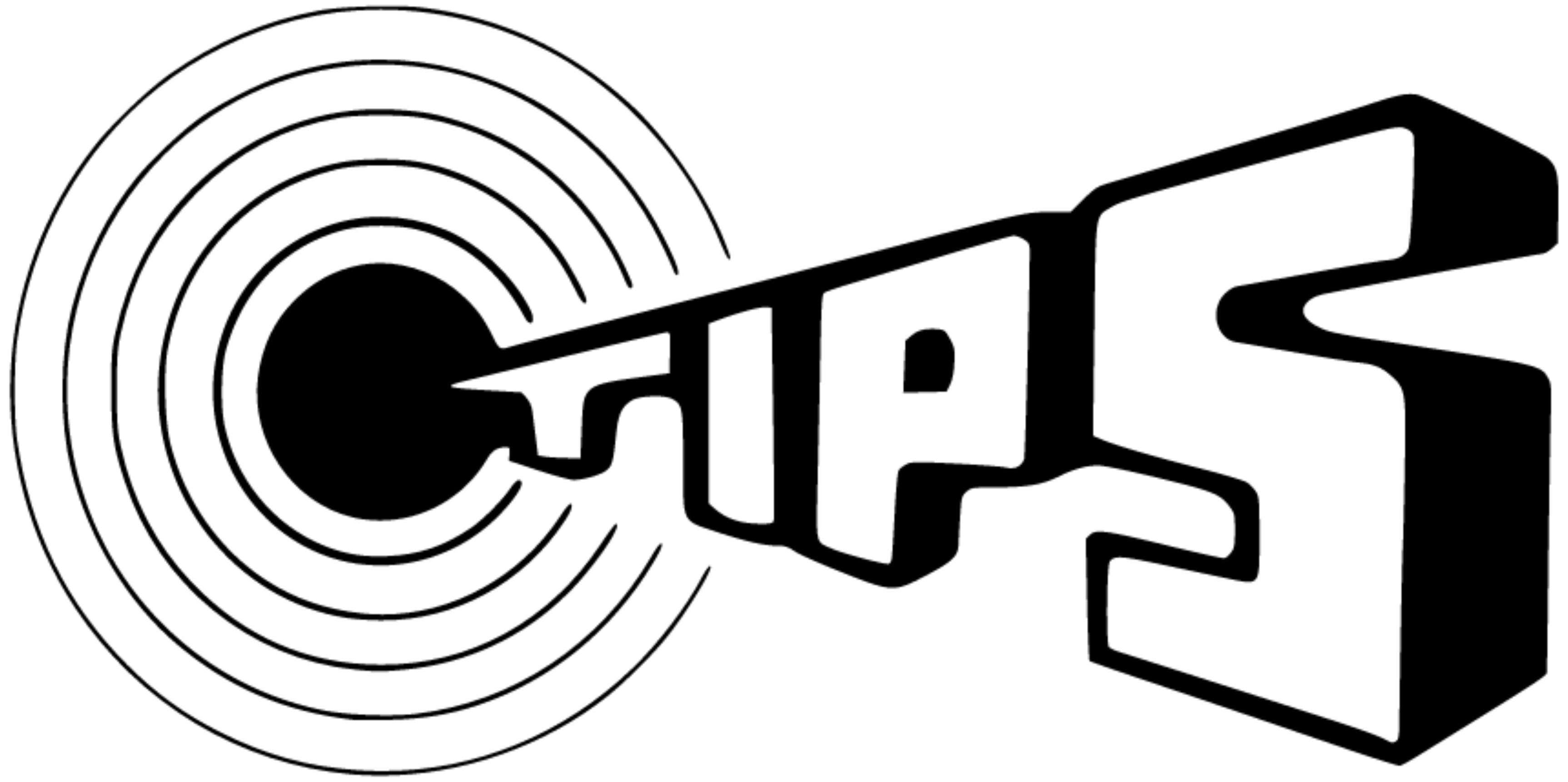
- A perk to your CV or even get hired

- Get some cash too!

WHAT SHOULD WE TRY?

TIPS FOR BEGINNERS

- Set incremental goals
- Review most voted kernels
- Asks questions on the forums
- Work solo to develop core skills
- Team up to learn more from others
- Don't worry about low ranks



Source

LET'S DO IT!

NEXT STEPS:

- **Read collected papers, repos and contribute!***

- **Join our Slack!**

- **Brainstorm!**

- **Form teams?**

- **Save the world!**

*(Meta Warning! These are papers we use to process papers!)

THANKS FOR THE ATTENTION!

LET’S CONTINUE!

data-hackers-zone.slack.com

PRESENTATION

COMMUNITY

LEV

ELDAR