From Words to Culture: Natural Language Processing for Social Sciences

Dirk Hovy
Bocconi University
Milan, Italy

www.dirkhovy.com
mail@dirkhovy.com
@dirkhovy.bsky.social

INTEGRATOR
European Research Council

A Sad Question

Is it still worth studying NLP?

Goals

- Answer emphatically in the positive
- Convince you we’re just at the beginning
- Show 3 studies why CSS is a great area for NLP

Hovy & Purschke (2018)
Purschke & Hovy (2019)
Hovy et al. (2020)

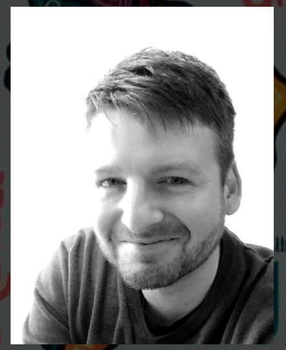
MilaNLP

Language Maps on Social Media

with

Tim Baldwin
University of Melbourne
Australia

Julian Brooke
University of British Columbia
Canada

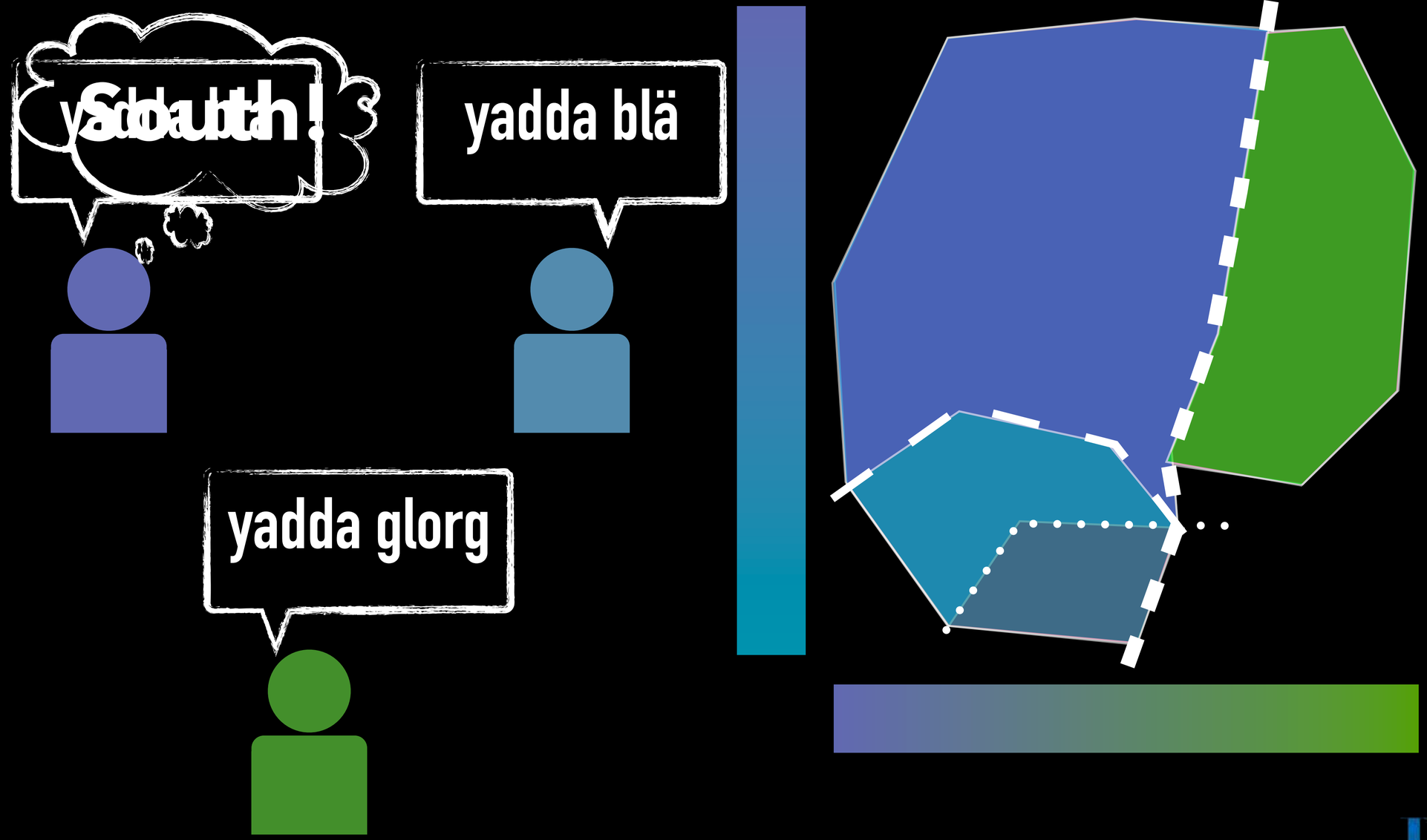


Christoph Purschke
University of Luxembourg
Luxembourg

Afshin Rahimi
University of Queensland
Australia

Bocconi

Local Differences



Getting Data

MilaNLP

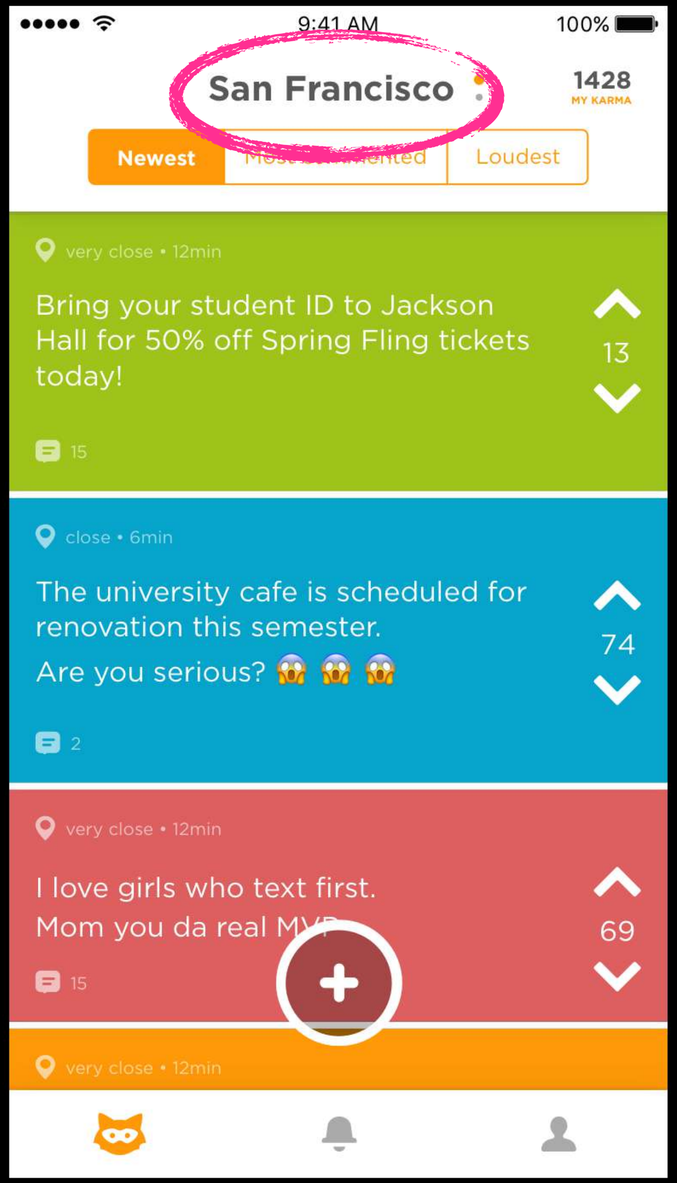
Bocconi

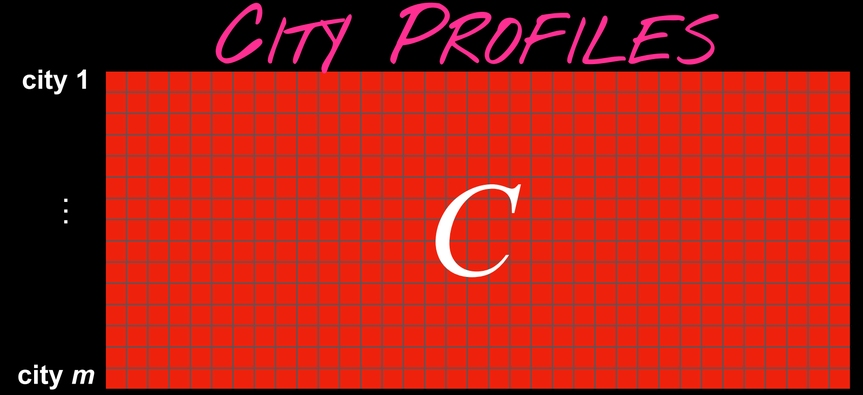
Data

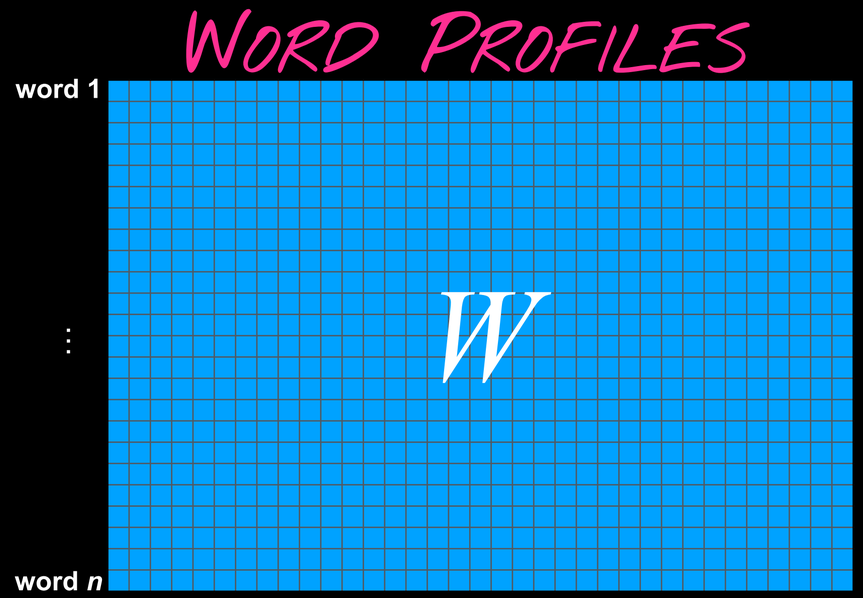
408 German-speaking cities

2.3M threads

16.8M posts



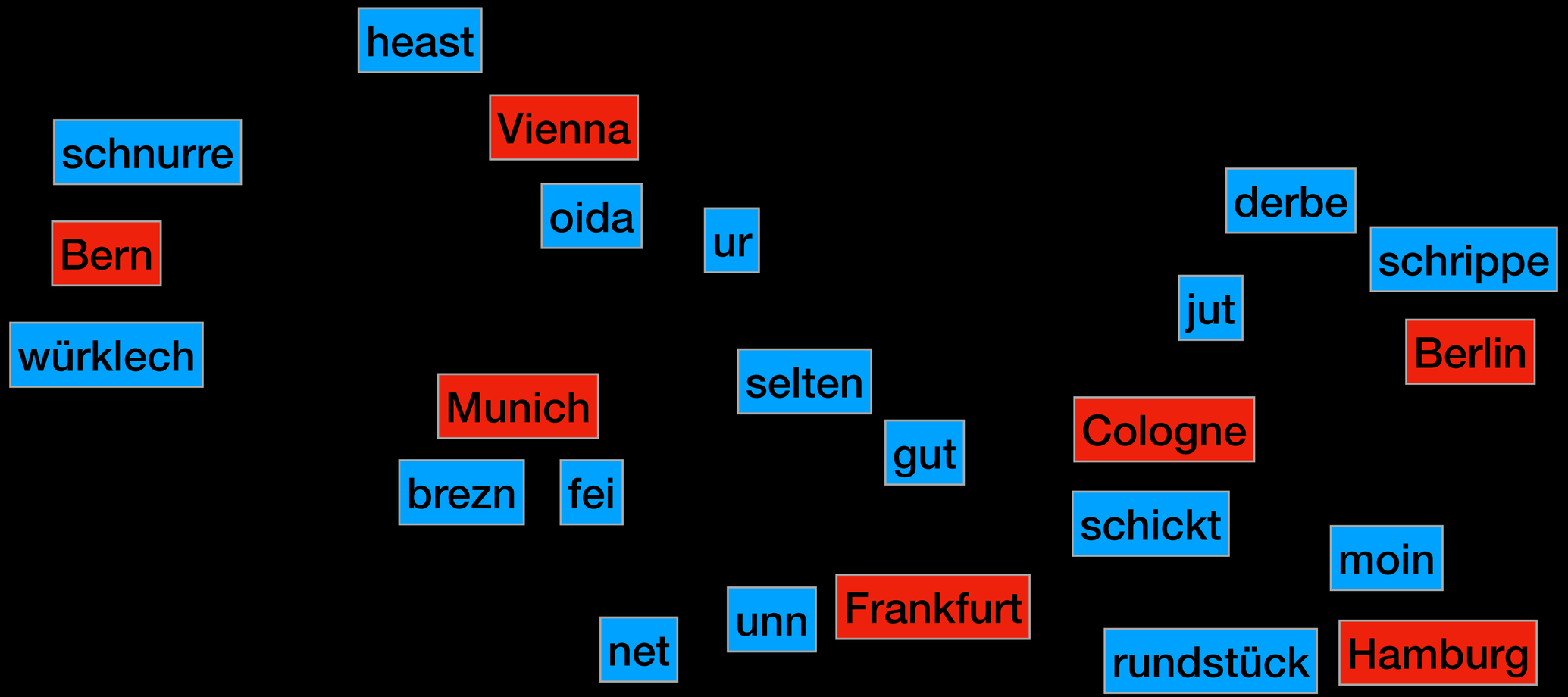




Dialects vs. Topics

Eisenstein et al. (2010)
Bamman et al. (2014)
Salehi et al. (2017)

Representations in Space



I.E., CITIES, REGIONS, PEOPLE, ...

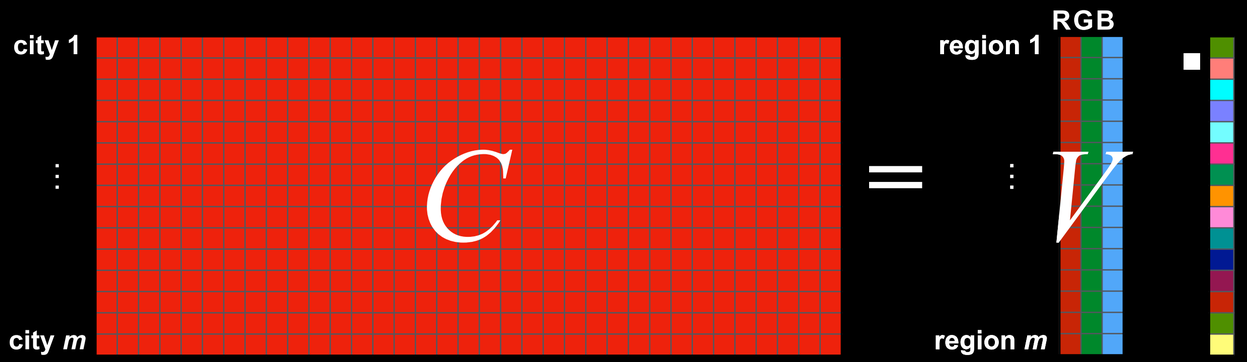
FOR ILLUSTRATION PURPOSES ONLY

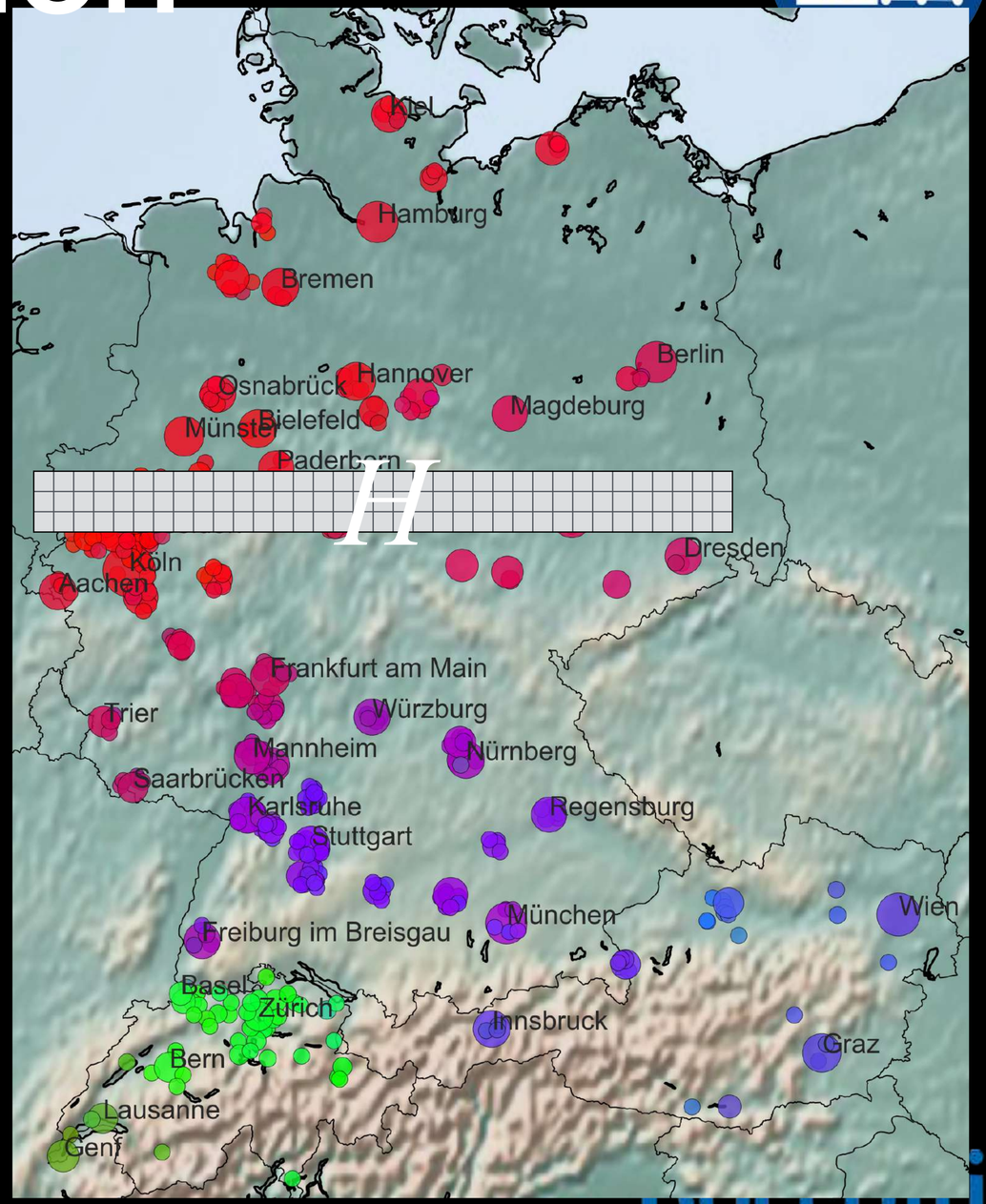
Where does that Leave Us?

- Vector representations allow us to
  - compare **cities** to **cities** WHICH CITIES ARE LINGUISTICALLY MOST SIMILAR?
  - compare **words** to **cities** WHICH WORDS ARE MOST INDICATIVE OF A CITY?
  - find latent features of **words** and **cities** WHAT IS THEIR STYLE, LOCATION, ETC?
  - compare **words** to **words** WHICH WORDS HAVE A SIMILAR MEANING?

Visualization

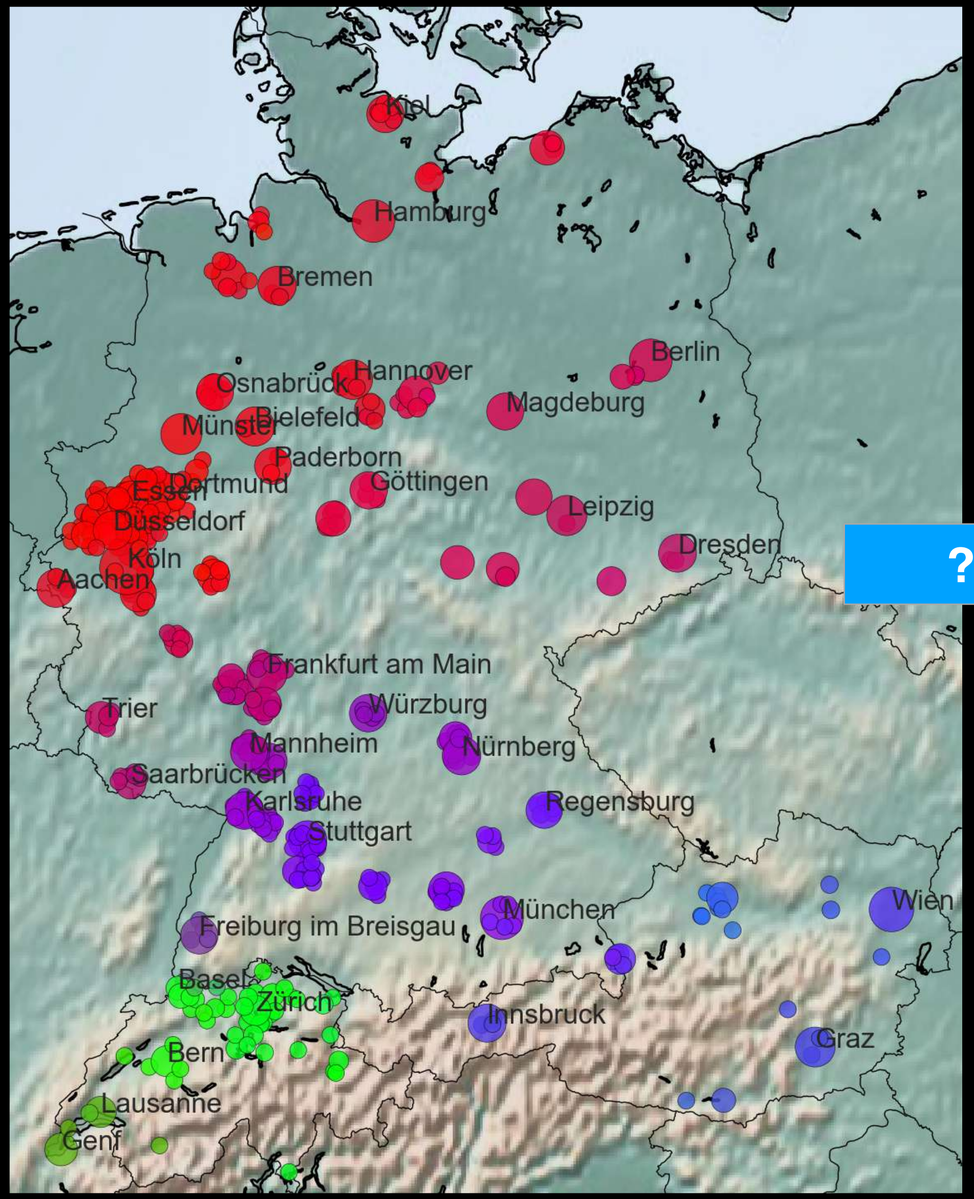
Learned Representations



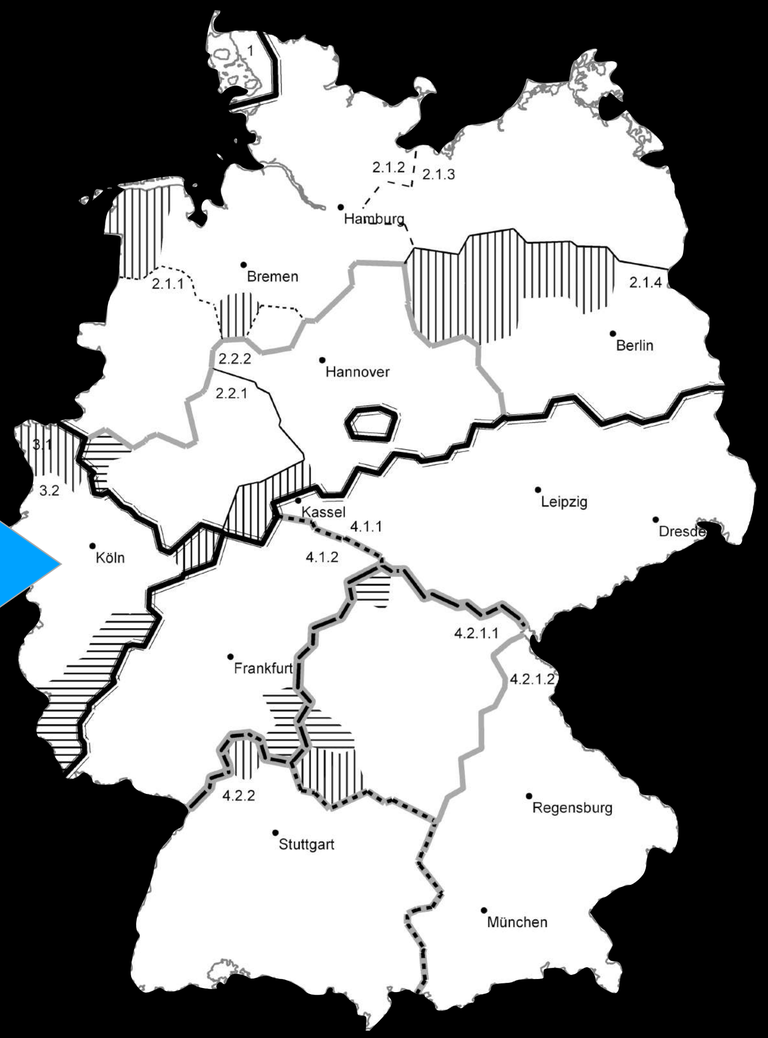


Alignment with linguistic theory?

Lameli (2013)



???

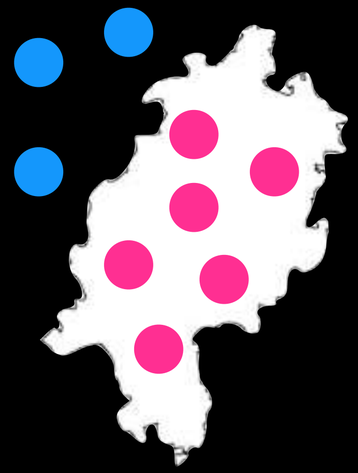


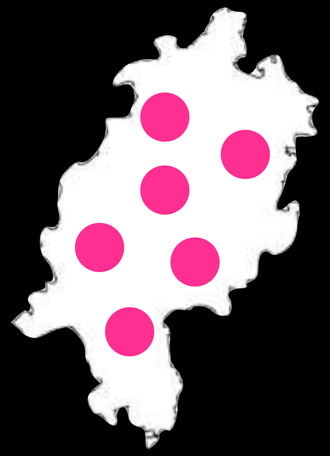
Evaluation Metrics

Homogeneity
cluster in only 1 region

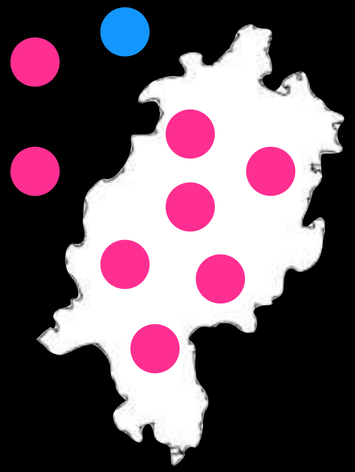
Completeness
region has only 1 cluster

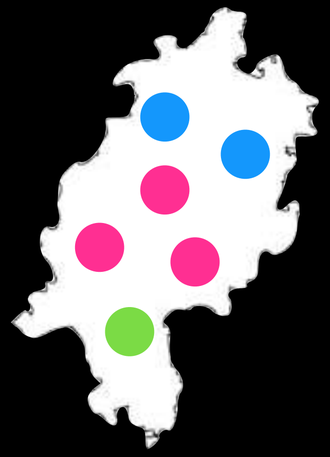
Good:





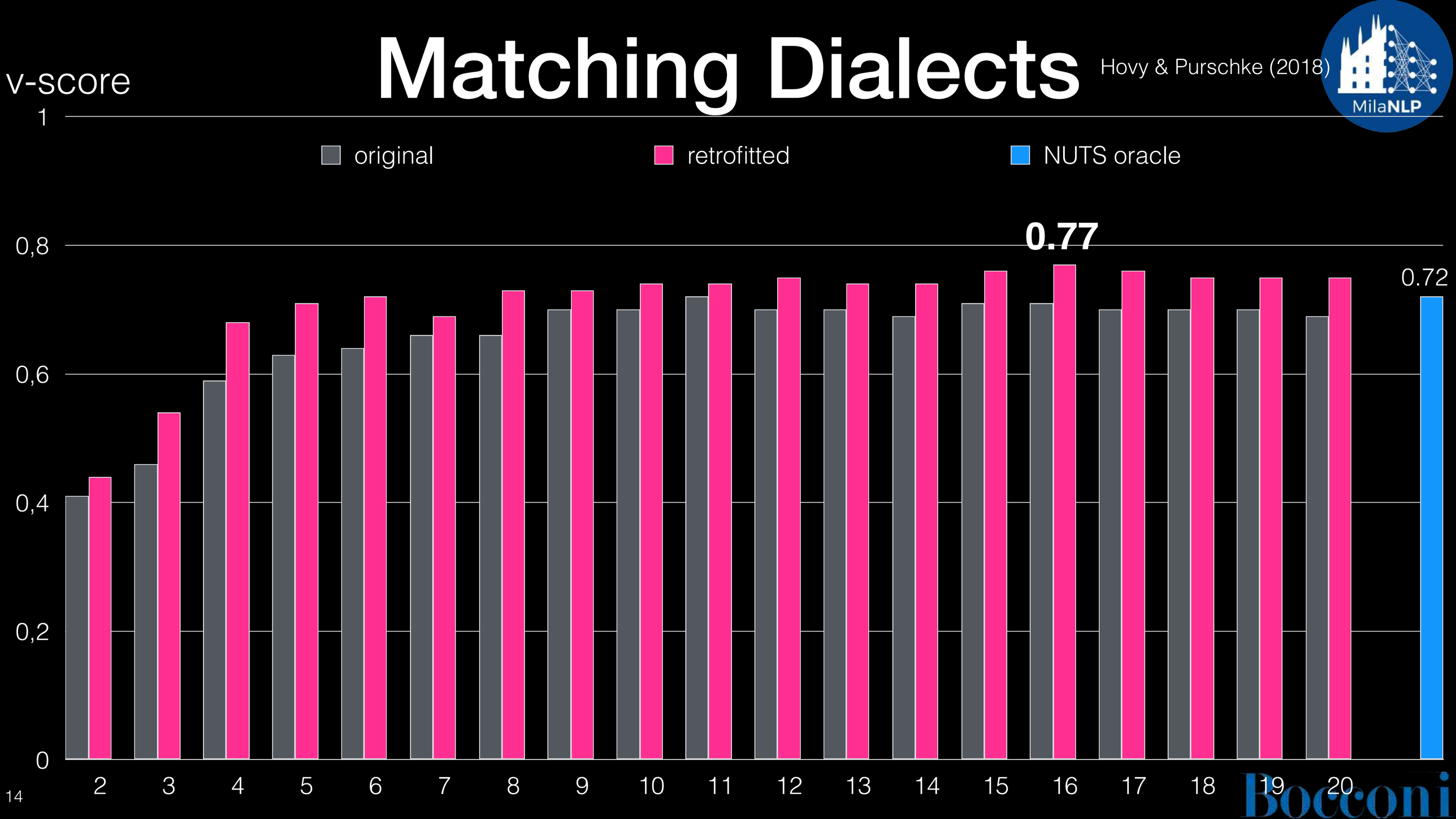
Bad:

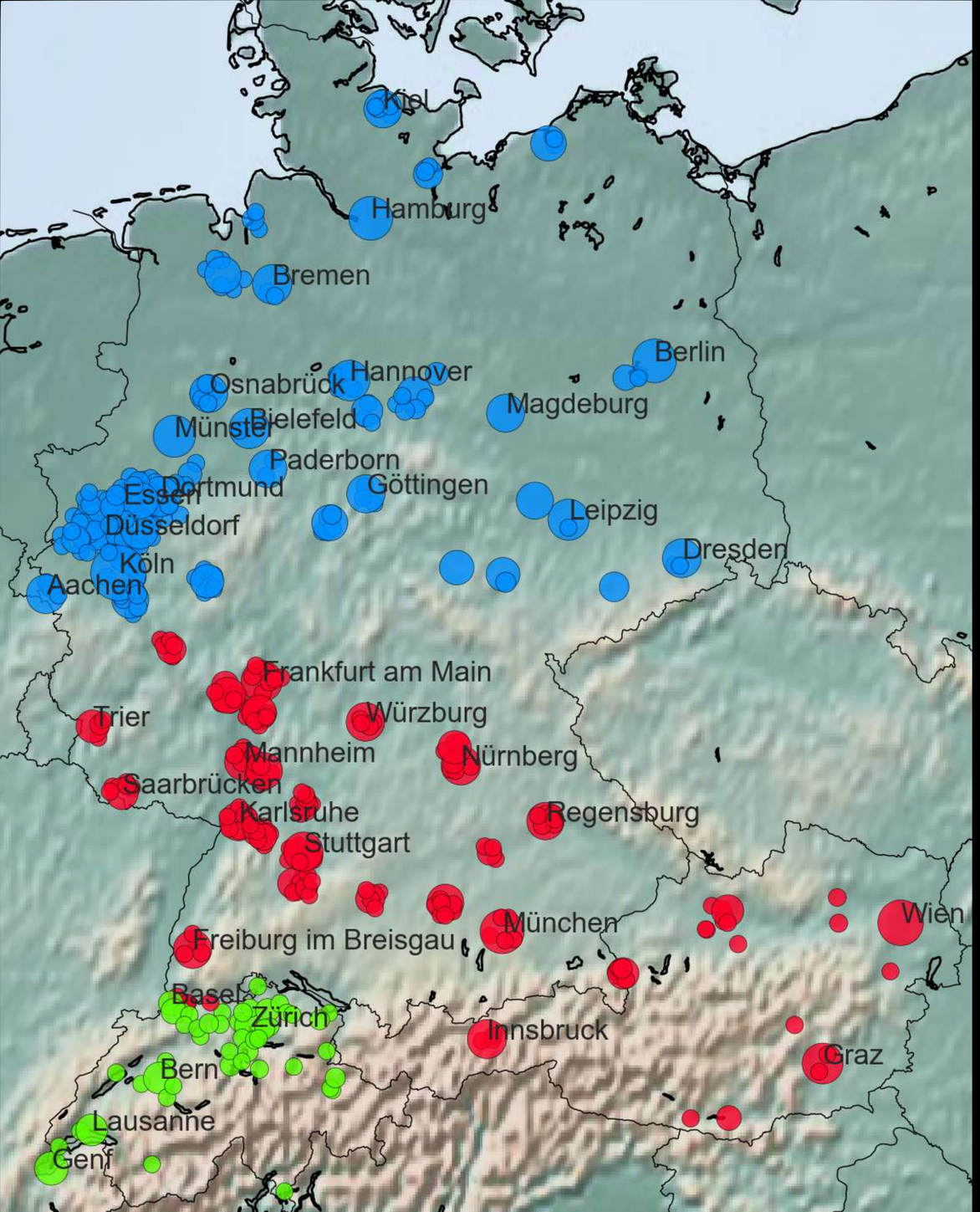




Matching Dialects

Hovy & Purschke (2018)

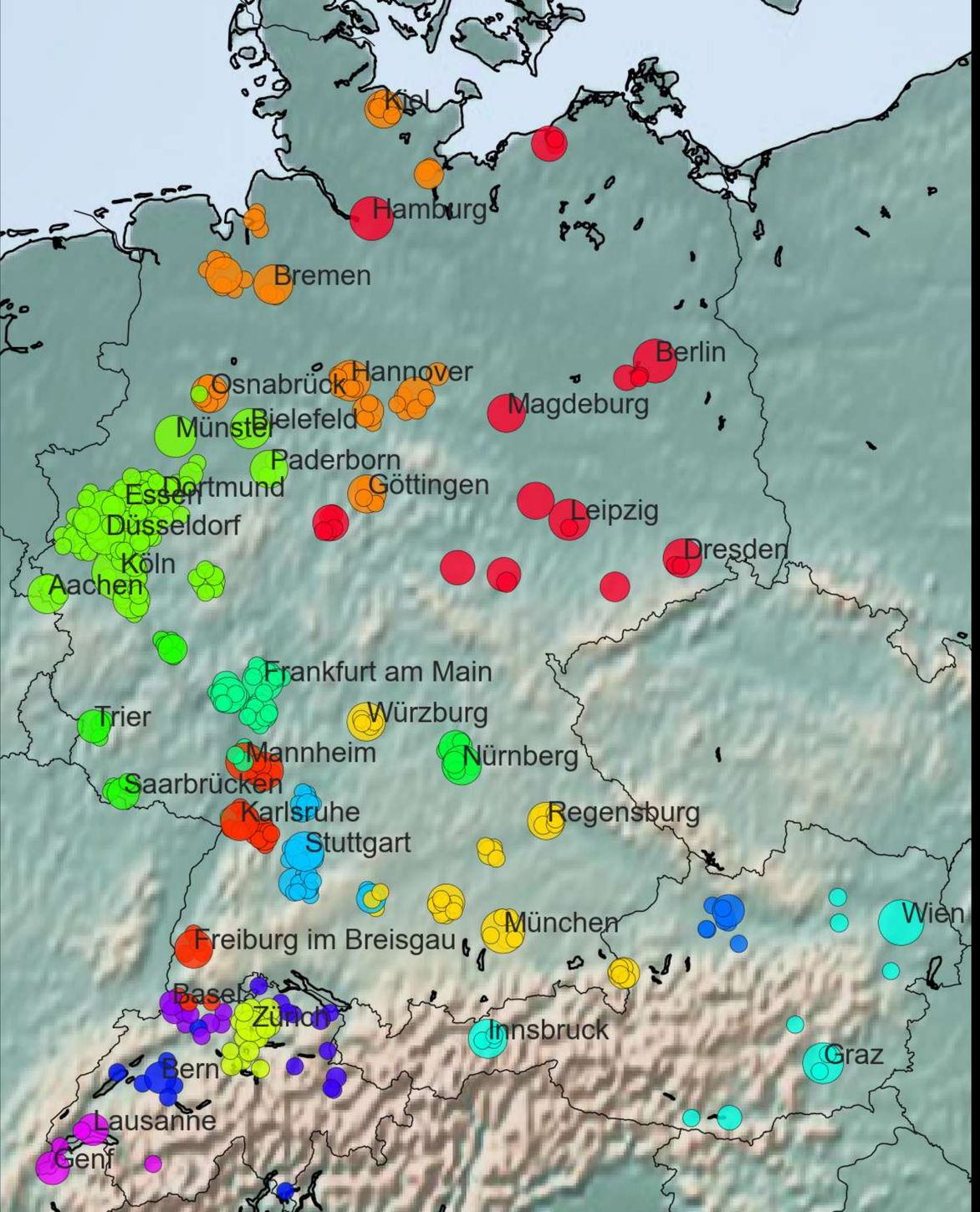




3 Clusters

- Separates out Switzerland
- Distinction of Low and Upper German varieties

16 Clusters

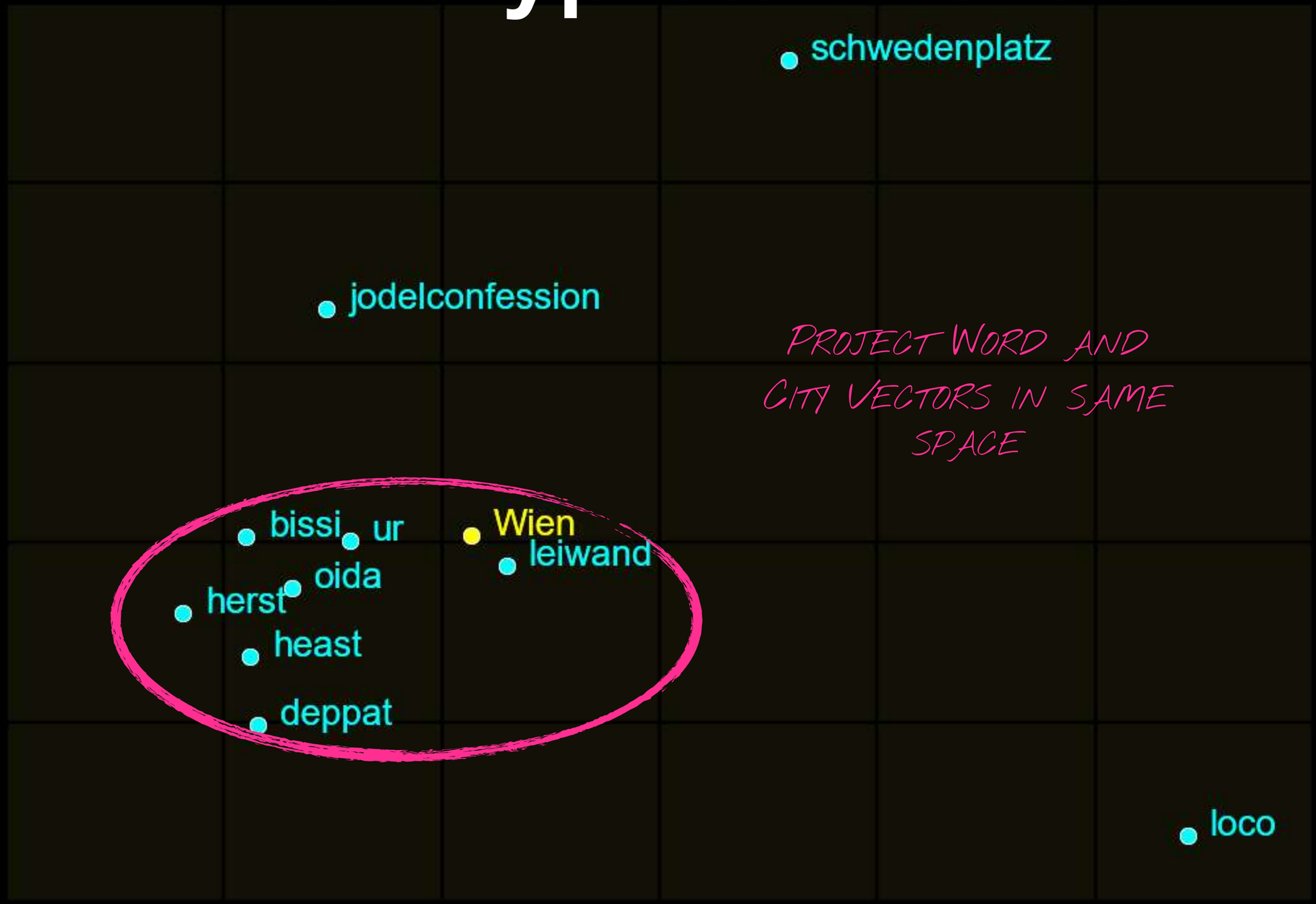


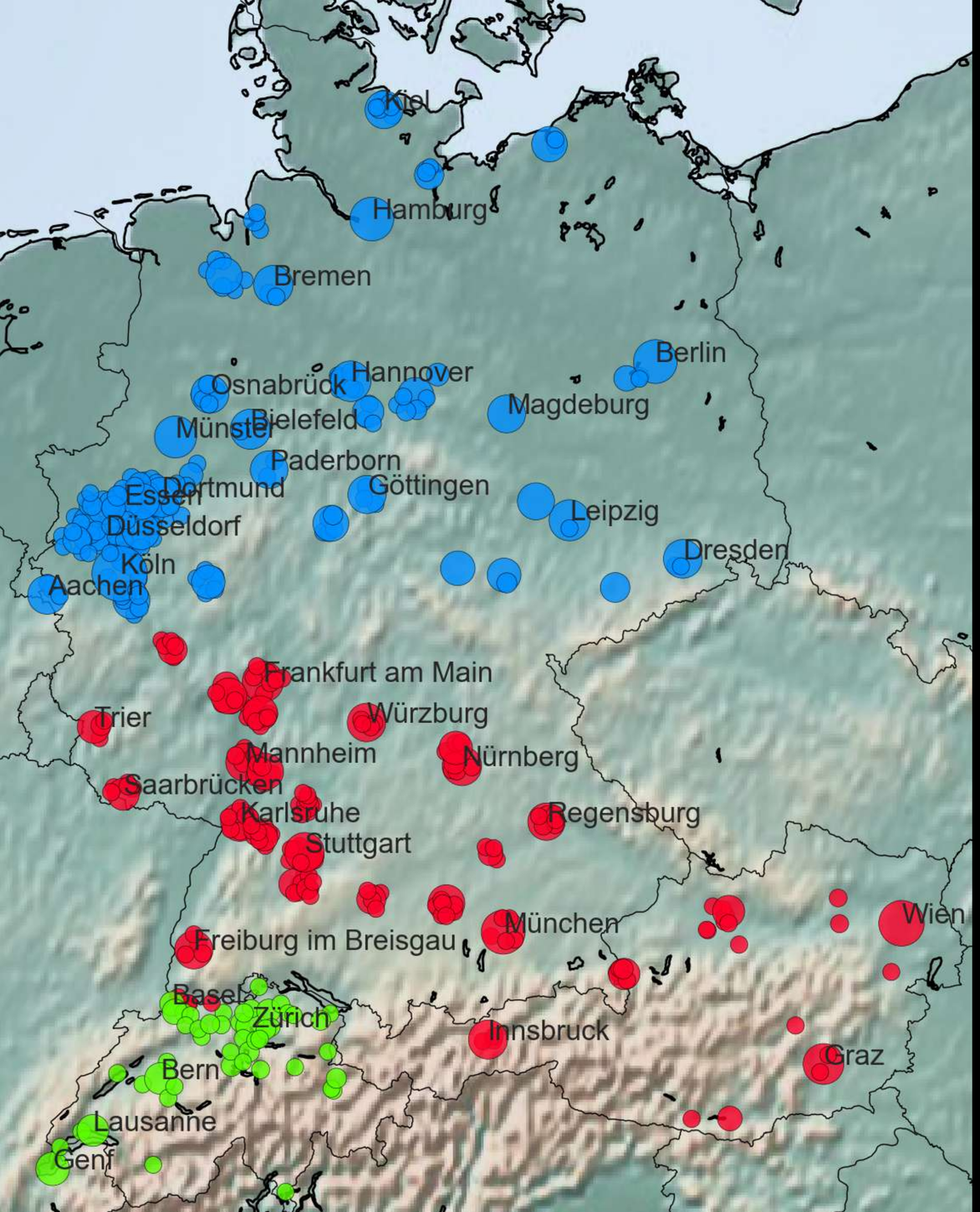
- Fine-grained region in German SW
- Austria separated
- Distinctions within Switzerland & Austria

MilaNLP

Bocconi

Prototypical Words





Words by Cluster

ja gut, erstmal, sieht, drauf, vielleicht (*well yes, first, sees, onto, maybe*)

afoch, voi, nd, i a, oda
(*easy, full, and, me too, or*)

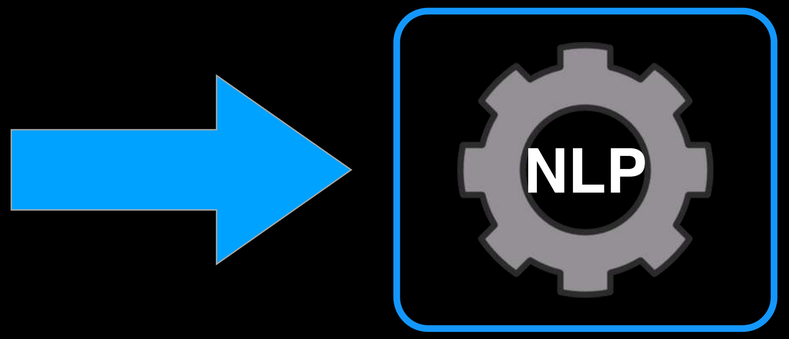
esch, ond, vell, gaht, wüki, nöd (*is, and, many, goes, really*)

Classifying Varieties

1. Manual Labeling

gerne
oifach
@vj
citylife
:)
jamais
möppes

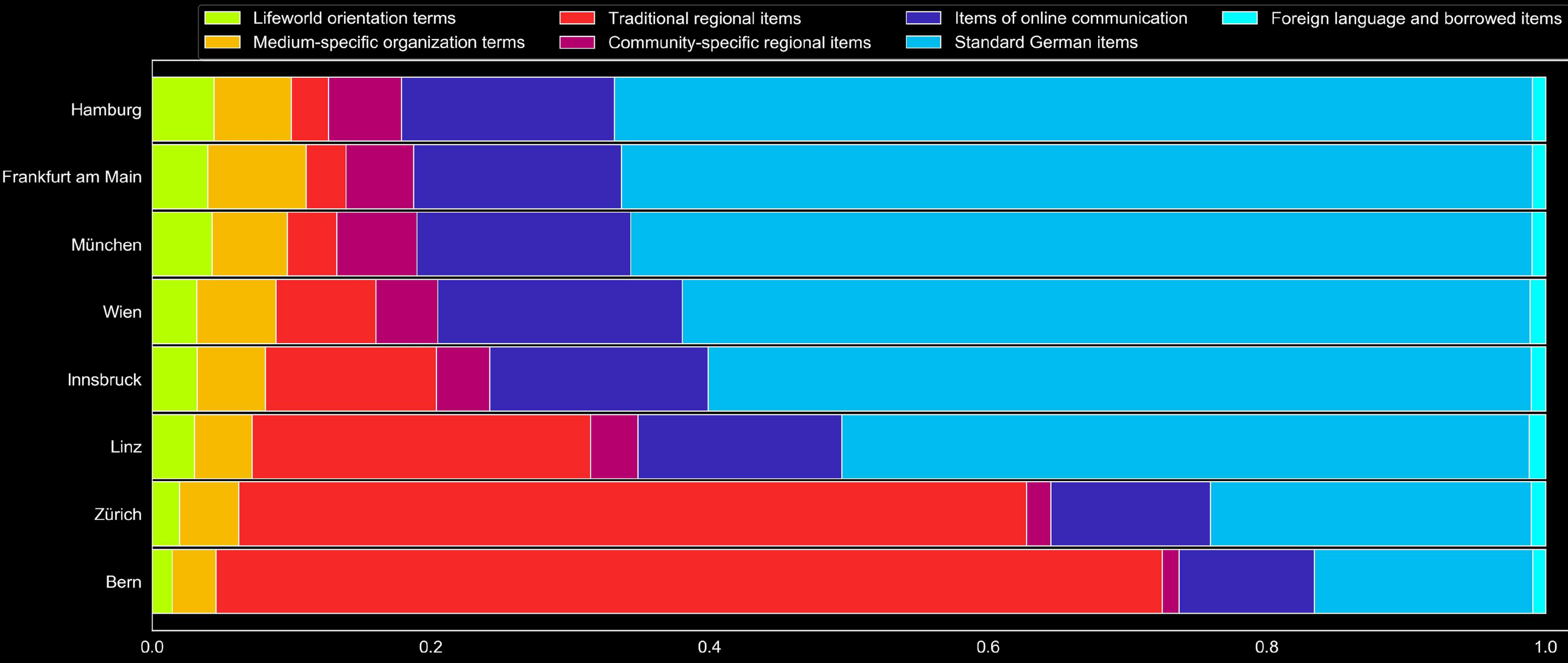
Standard German
Traditional Regiolect
Medium-Specific Term
Lifeworld Orientation Term
Online Communication
Foreign Words
Community-Specific Regional



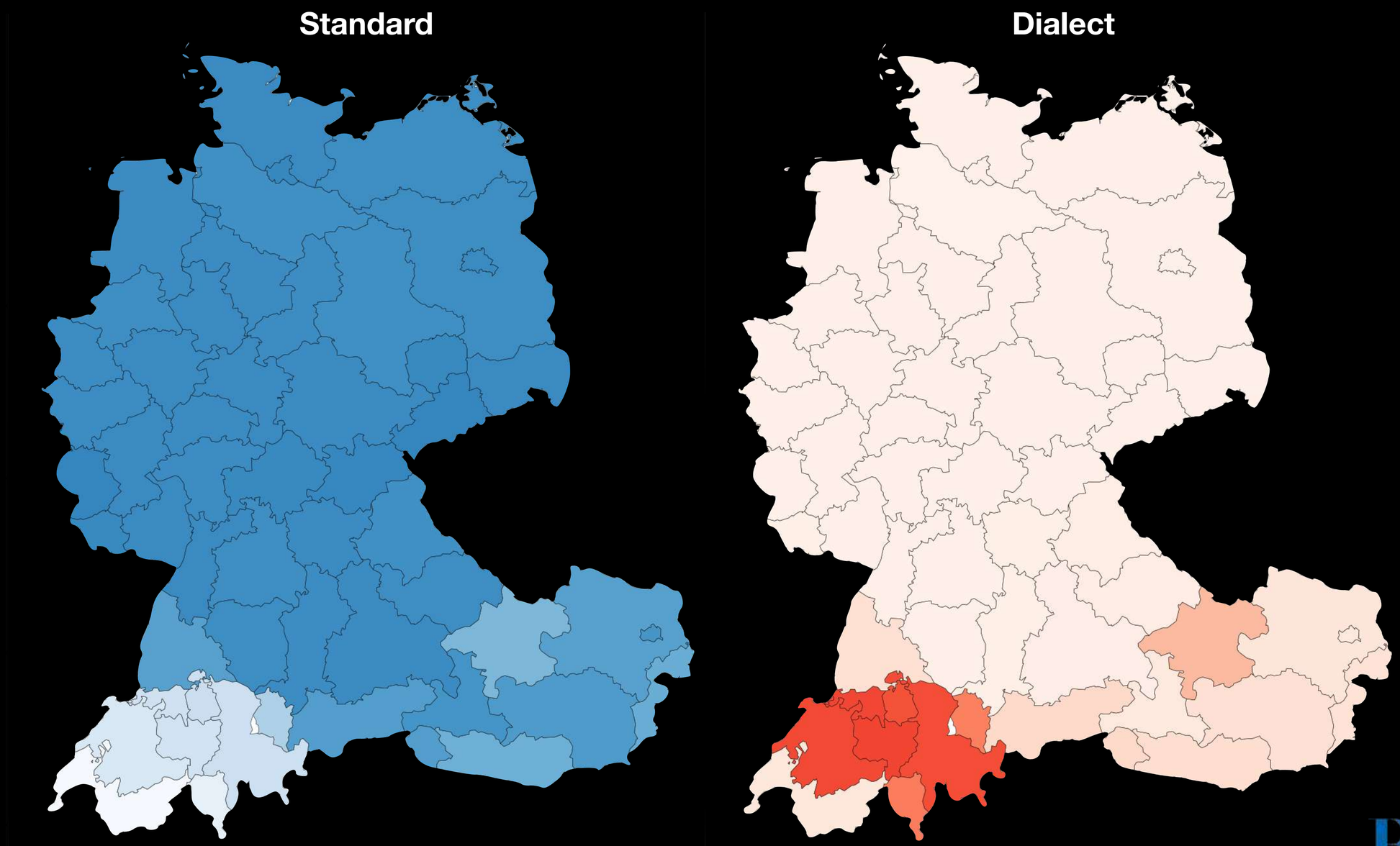
Predict

2. Train an ML Classifier

Varieties by City

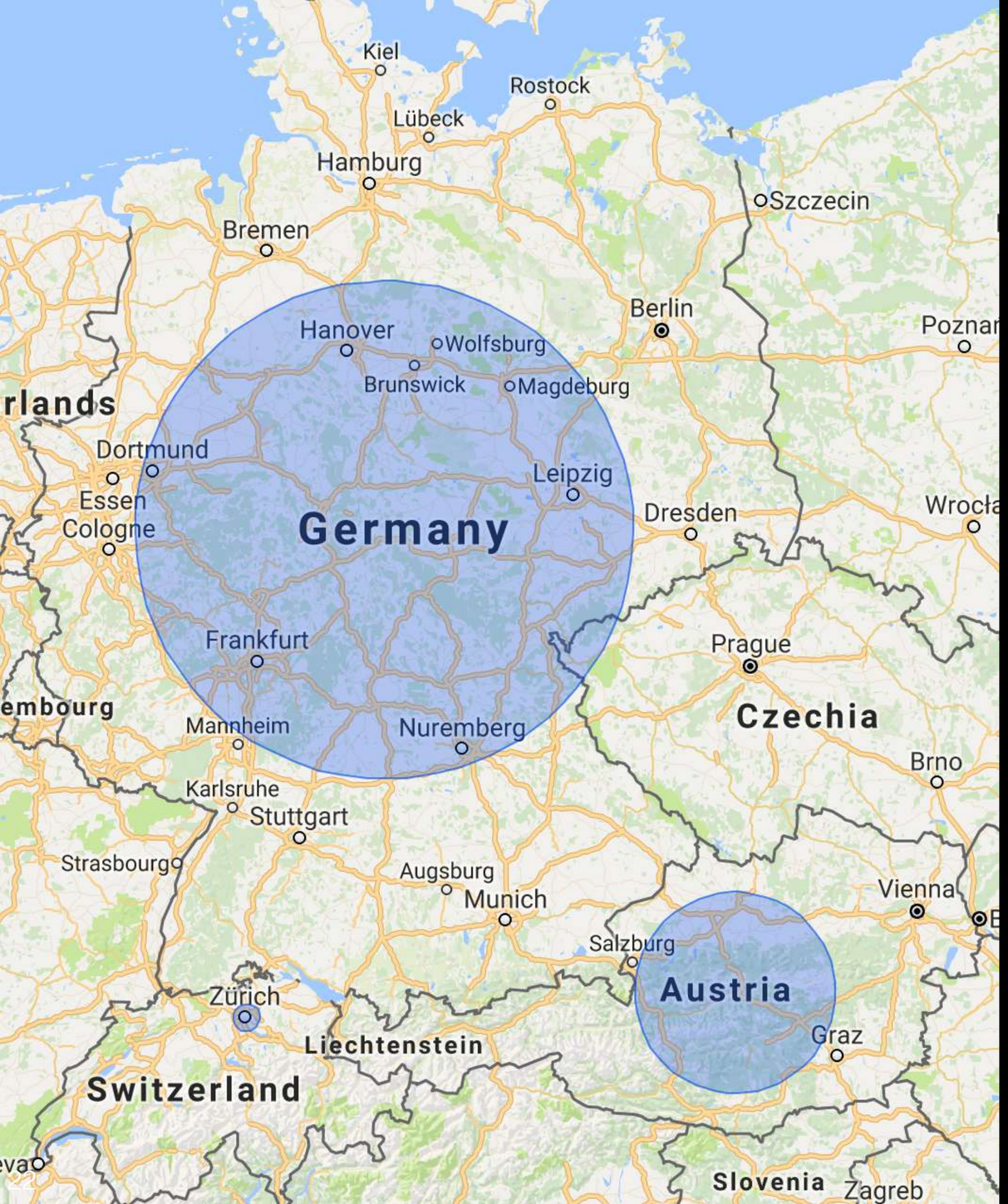


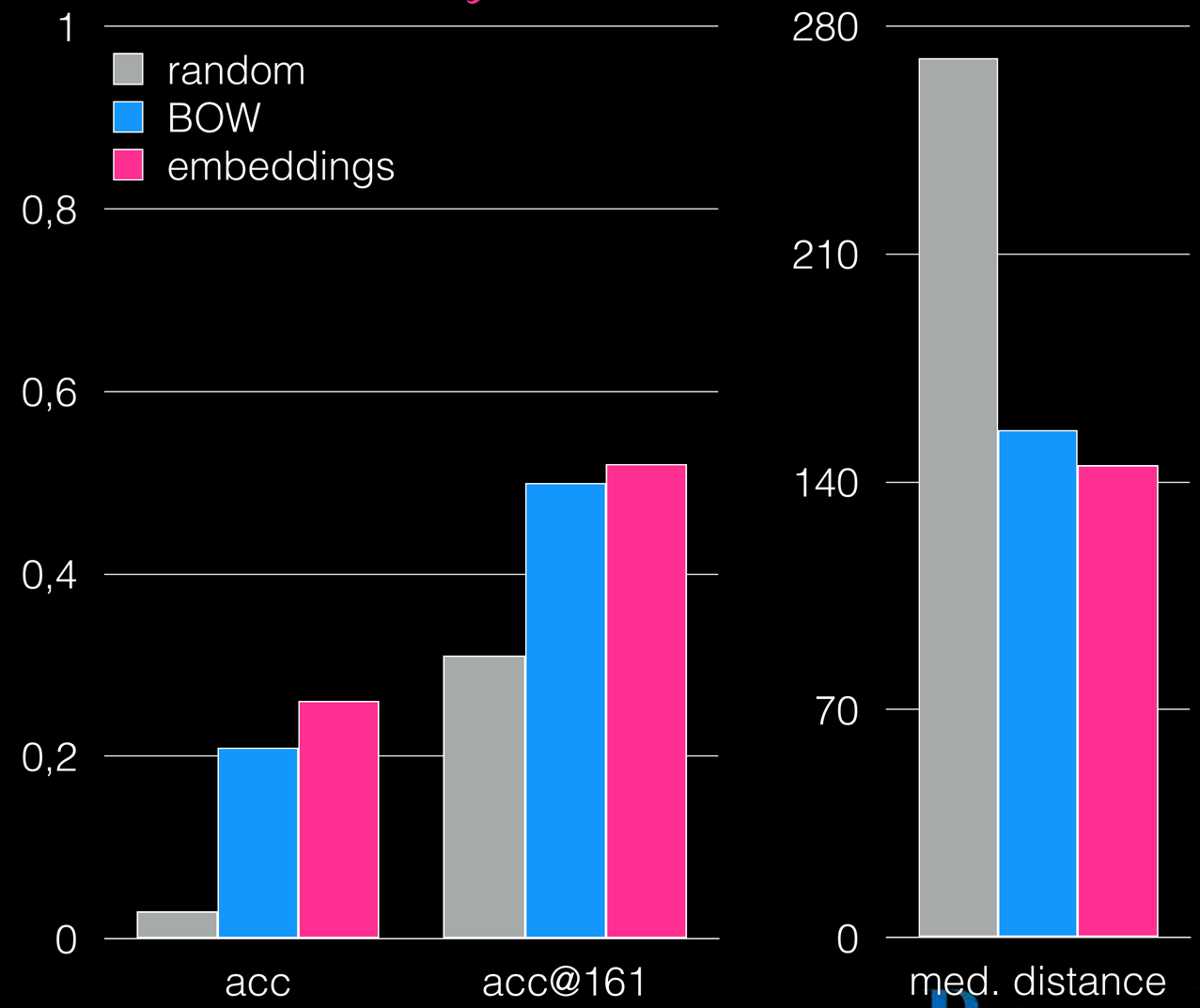
Aggregating in Space



Corollary: Geolocation

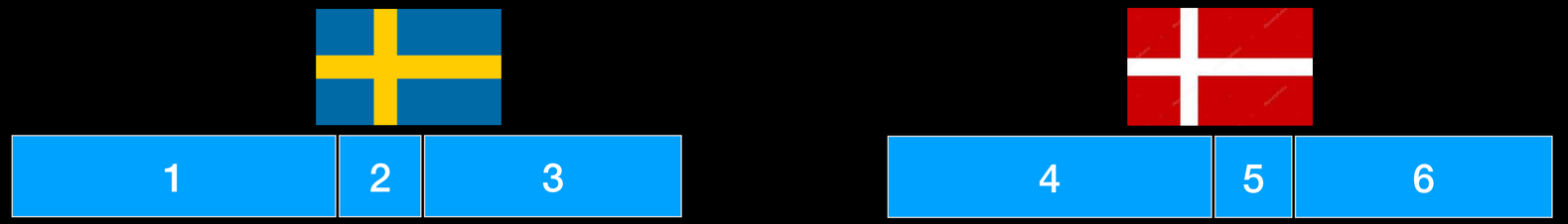
Even After removing most indicative words!





Cross-Lingual Similarity?

"Welcome to Hamburg"



Words:

Overlap

0 OF 6

WHAT A COMPUTER "SEES"

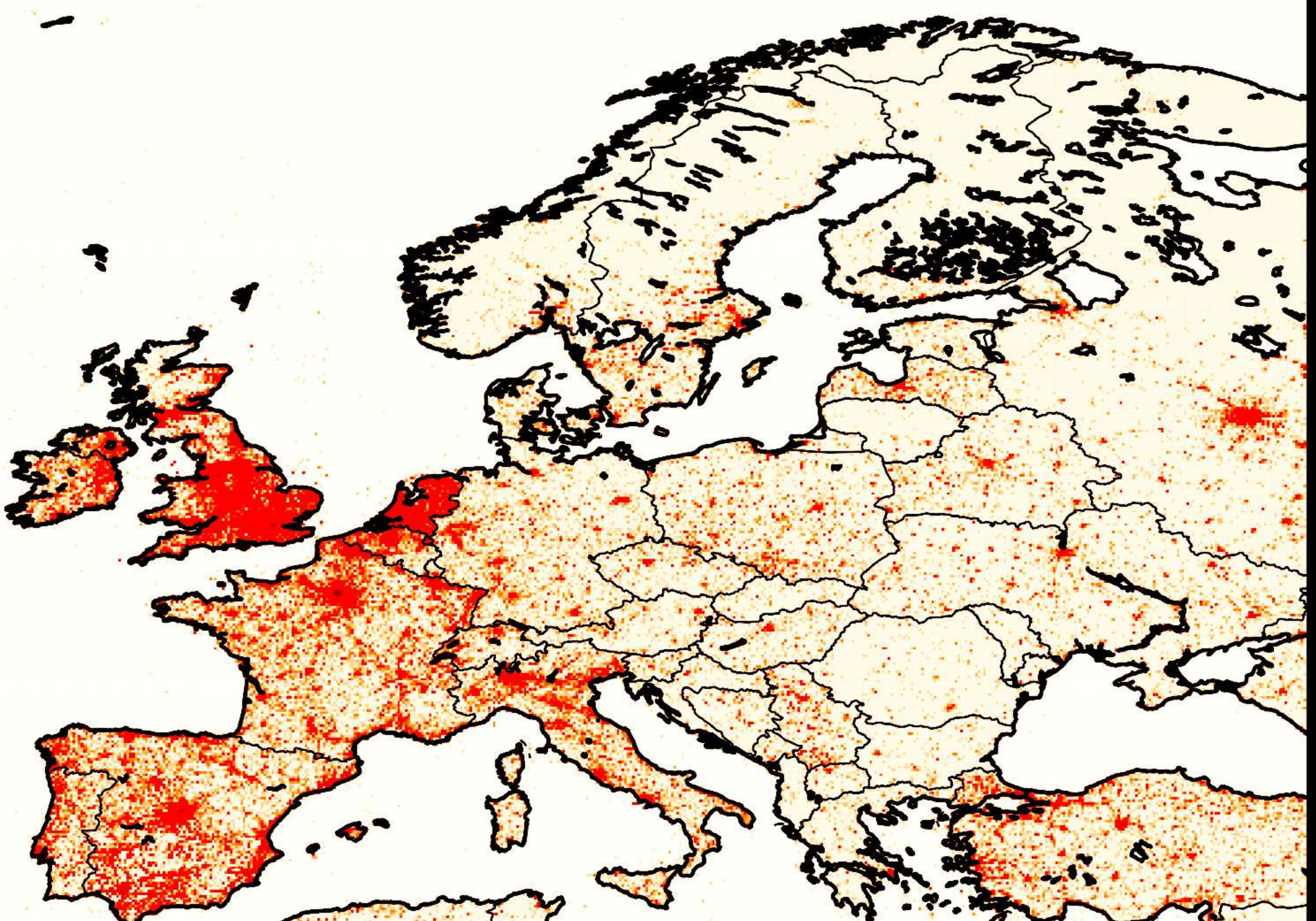
n-grams:

Väl, älk, lko, kom,
omm, mme, men, en ,
n t, ti, til, ill,
ll , l H, Ha, Ham,
amb, mbu, bur, urg

Vel, elk, lko, kom,
omm, mme, men, en ,
n t, ti, til, il ,
l H, Ha, Ham, amb,
mbo, bor, org

26 OF 39

Bocconi

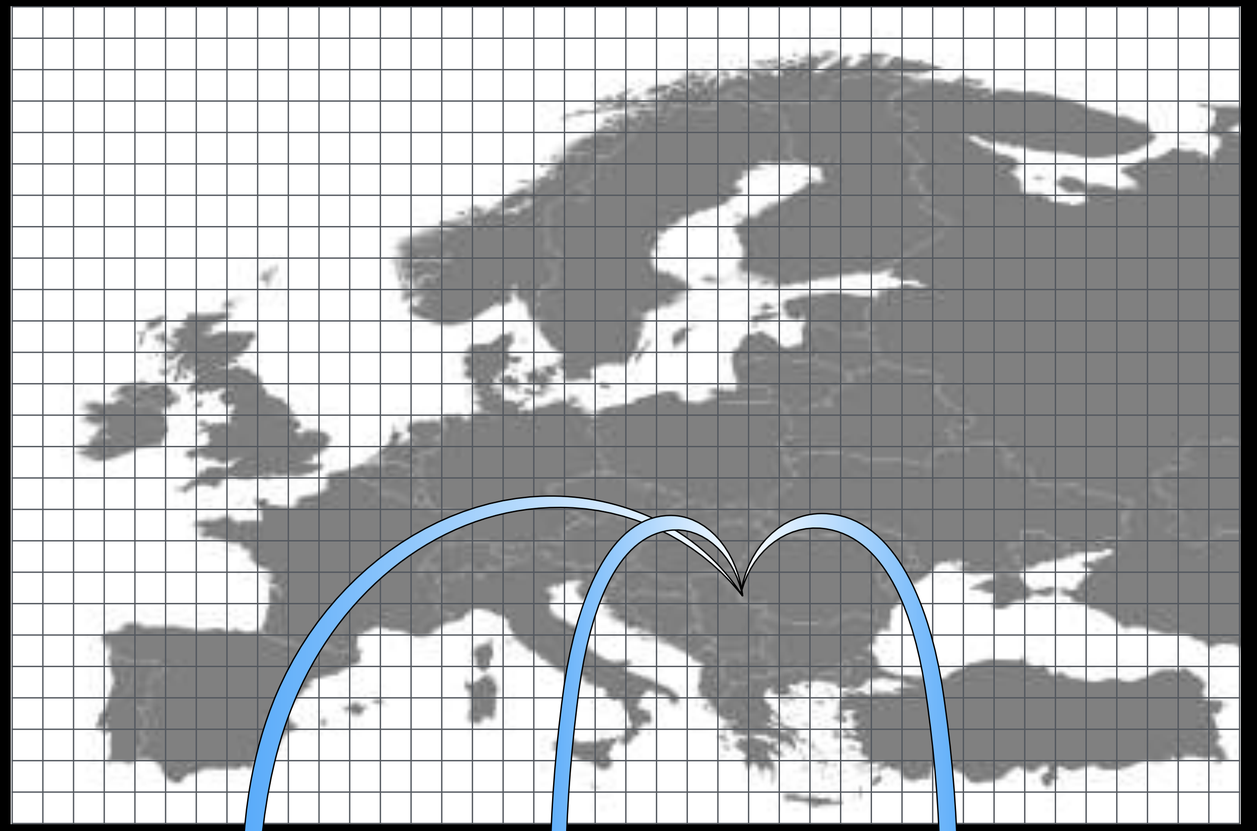


MilaNLP

Collect 50M Tweets

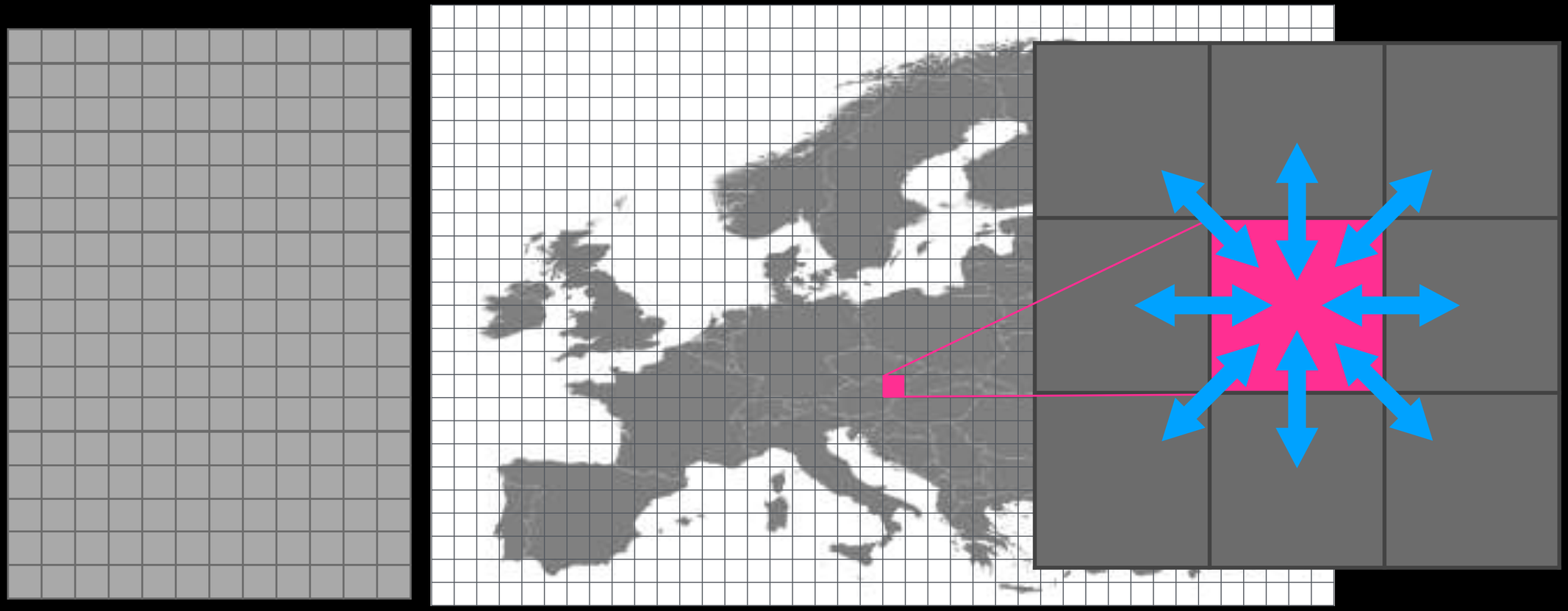
Bocconi

Extract Tweets from Grid



CHOP INTO N-GRAMS

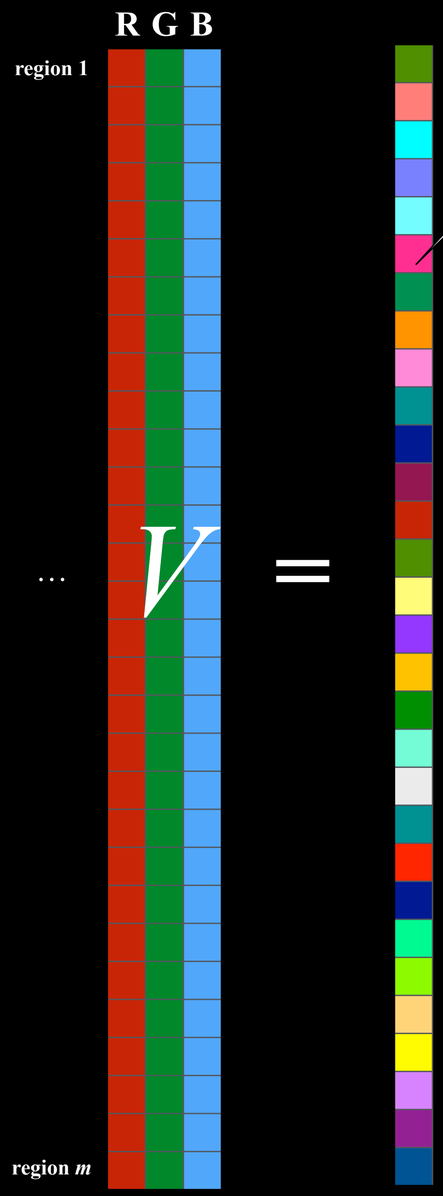
Learn Representations



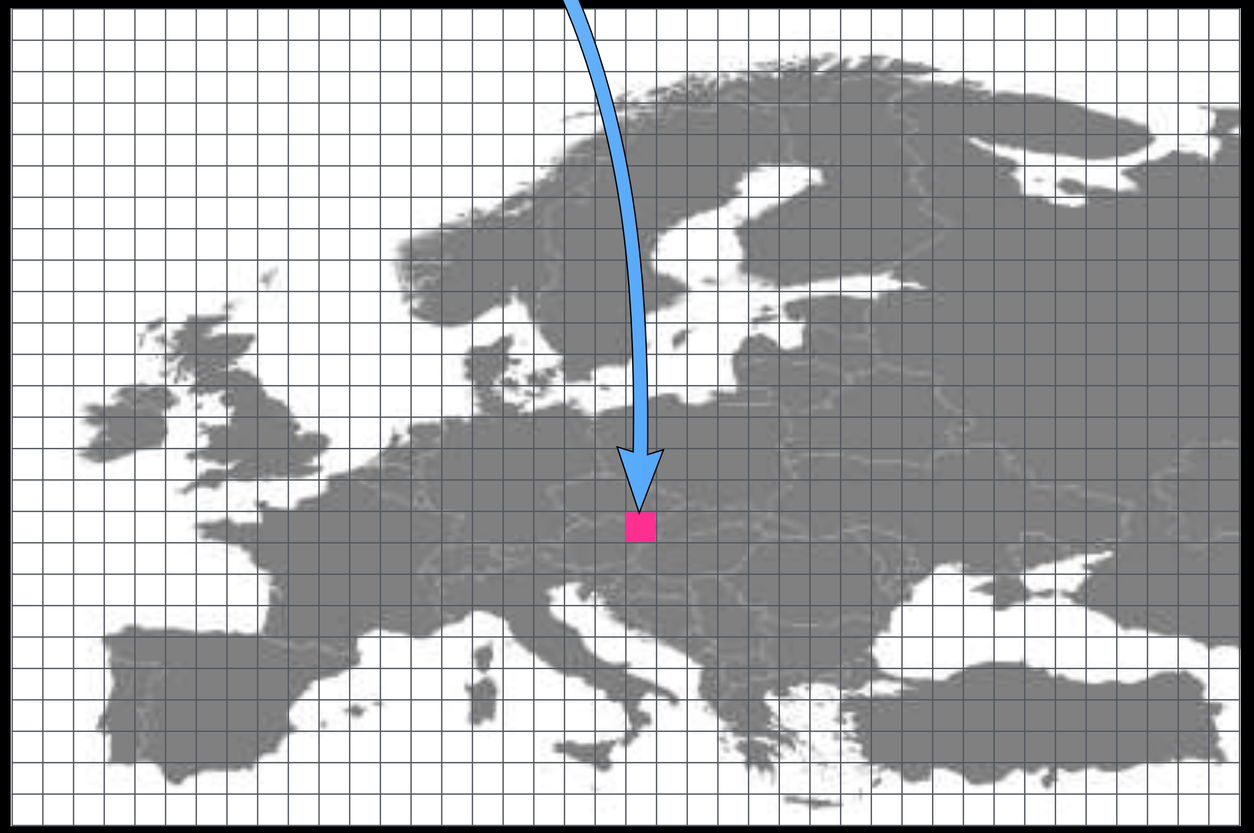
Aggregate n-grams by cell

Interpolate w/ neighbors

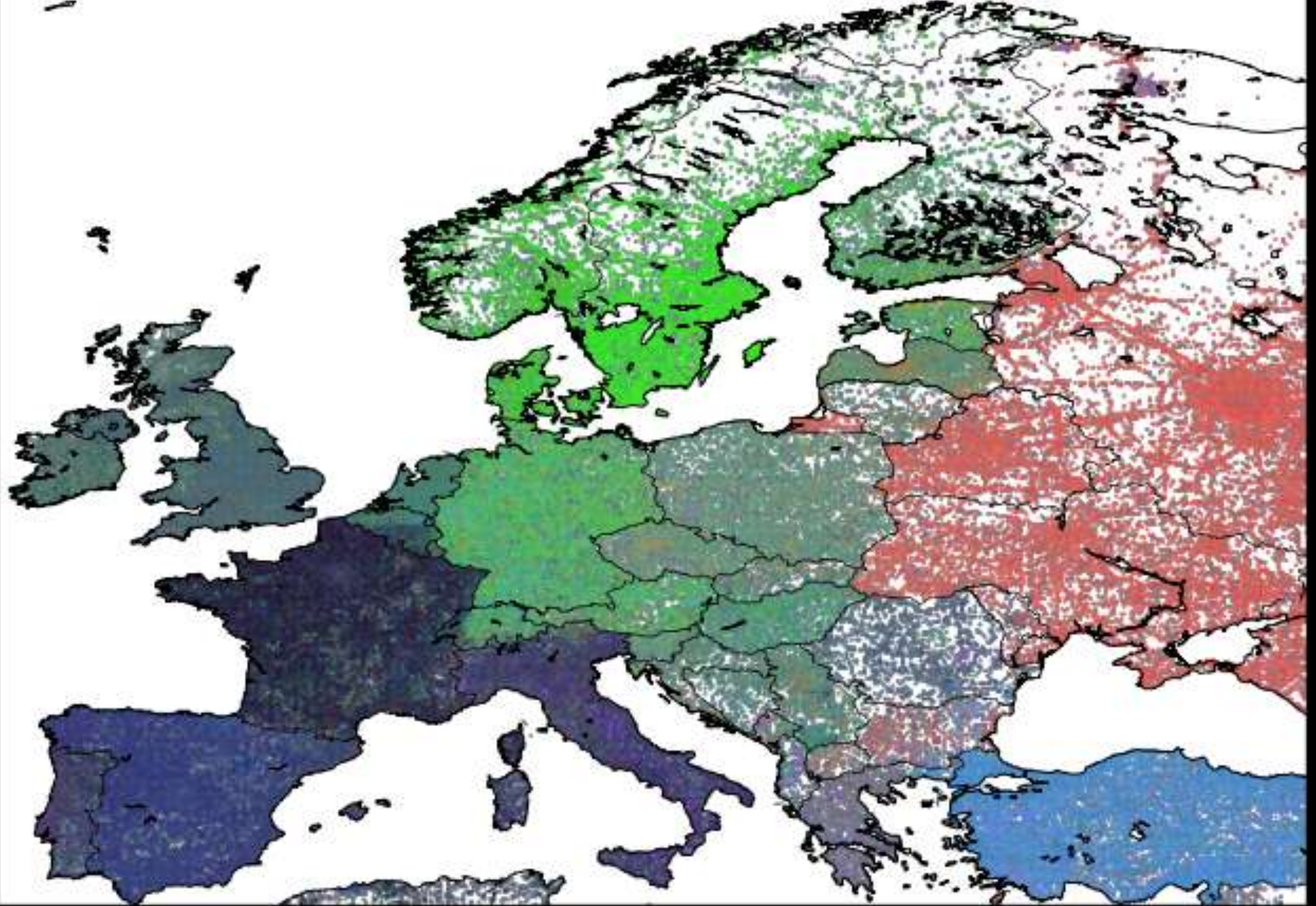
Color Grid



V =



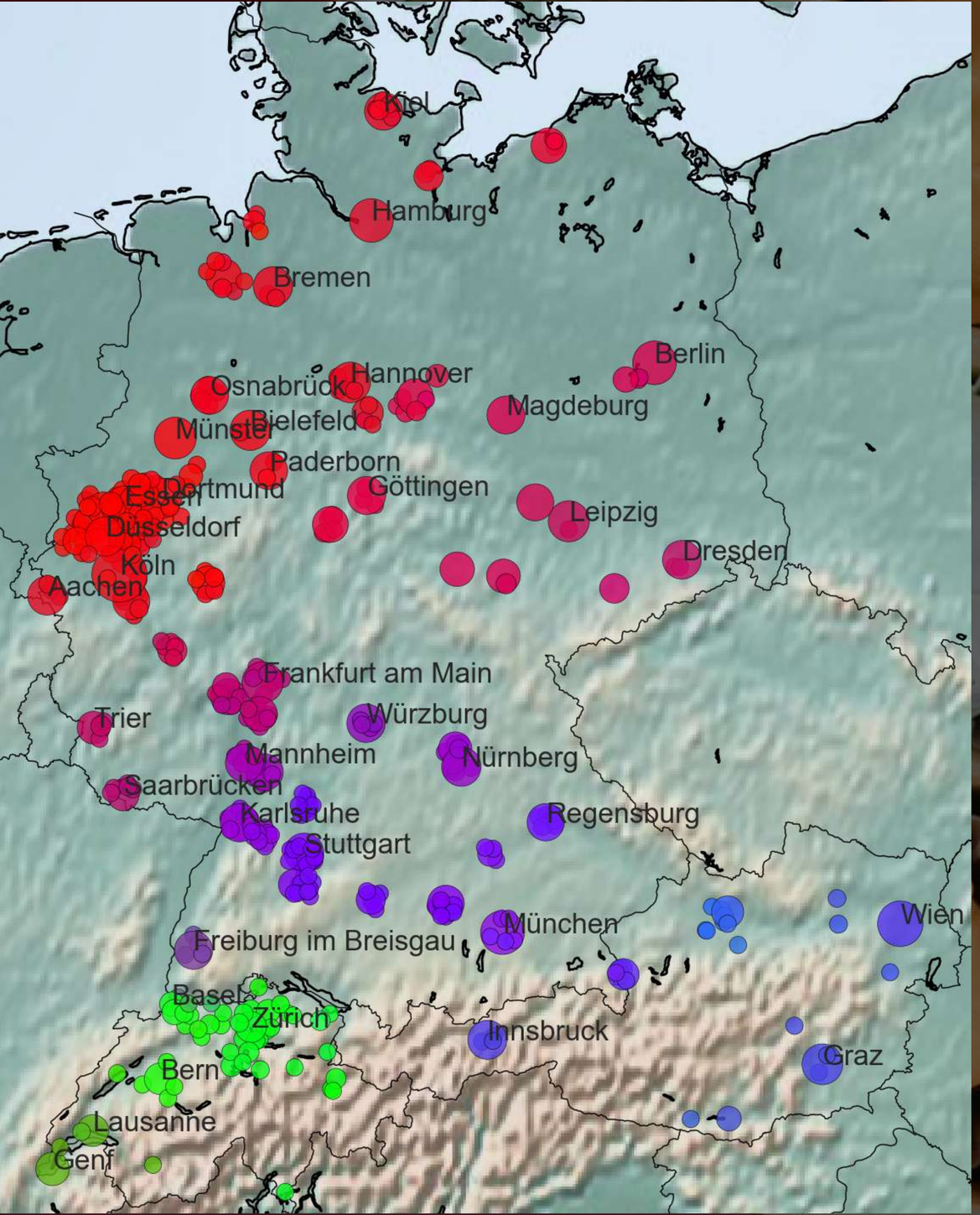
GET RGB VALUES



MilaNLP

Map It

Bocconi



Take-Aways

- Representation learning captures **language continuum**
- Embeddings capture word features:
  - resource/style
  - geolocation
- Enables quantitative analysis

MilaNLP

Bocconi

Habibi et al. (2025)

The Content Moderator's Dilemma:

Online Plurality and the Removal of Toxic Content

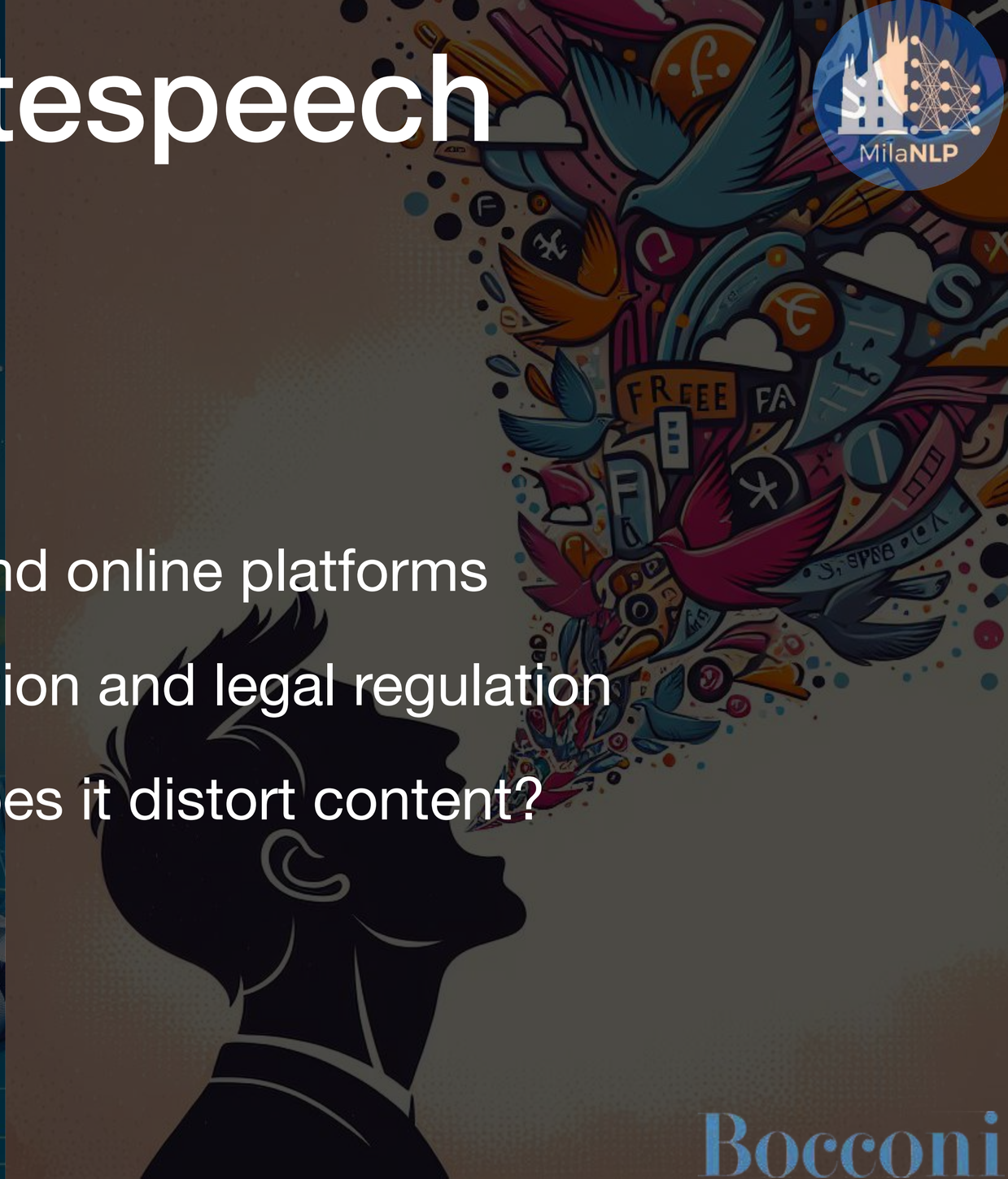
with

Mahyar Habibi
Lyft
Canada

Carlo Schwarz
Bocconi University
Italy

Online Hatespeech

- Widespread proliferation
- Problem for users, policymakers, and online platforms
- Increased calls for content moderation and legal regulation
- BUT: moderation means cutting: does it distort content?



Data

- Quasi-random sample of ~500,000 Twitter Users:
    - Get random user IDs (Siegel et al. 2021)
    - Collect profiles, tweets, and followed accounts (ca. 400 Million tweets)
    - Fine-tune a BERTweet model to predict political content
- Random Samples:
    - 5 million political Tweets
    - 1 million a) English, b) German, c) Italian
- Classify w/ Perspective API, avg 0.19 (≈%people who consider it offensive)

Measuring Moderation Effects

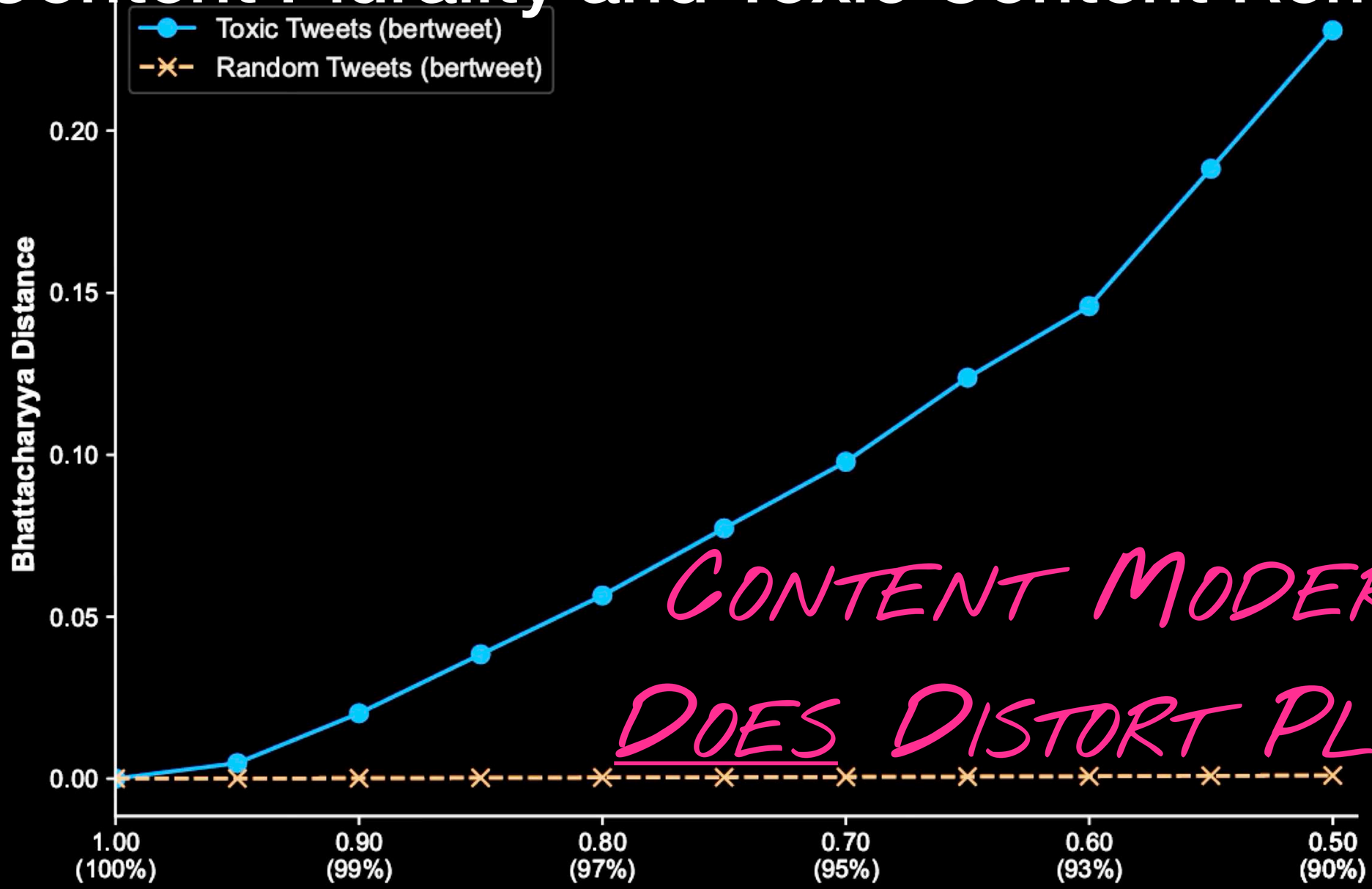
- Measure Embedding distortions with Bhattacharyya distance (BCD):

$$BCD(N_1, N_2) = \frac{1}{8}(\mu_1 - \mu_2)^T \Sigma^{-1} (\mu_1 - \mu_2) + \frac{1}{2}\left(\frac{det\Sigma}{\sqrt{det\Sigma_1 \cdot det\Sigma_2}}\right)$$

where $N_i(\mu_i, \Sigma_i)$ are multivariate normal distributions.

- BCD measures dissimilarity of two distros by detecting shifts in μ and Σ
- BCD is content-agnostic
- Computationally tractable in high-dimensional settings

Content Plurality and Toxic Content Removal

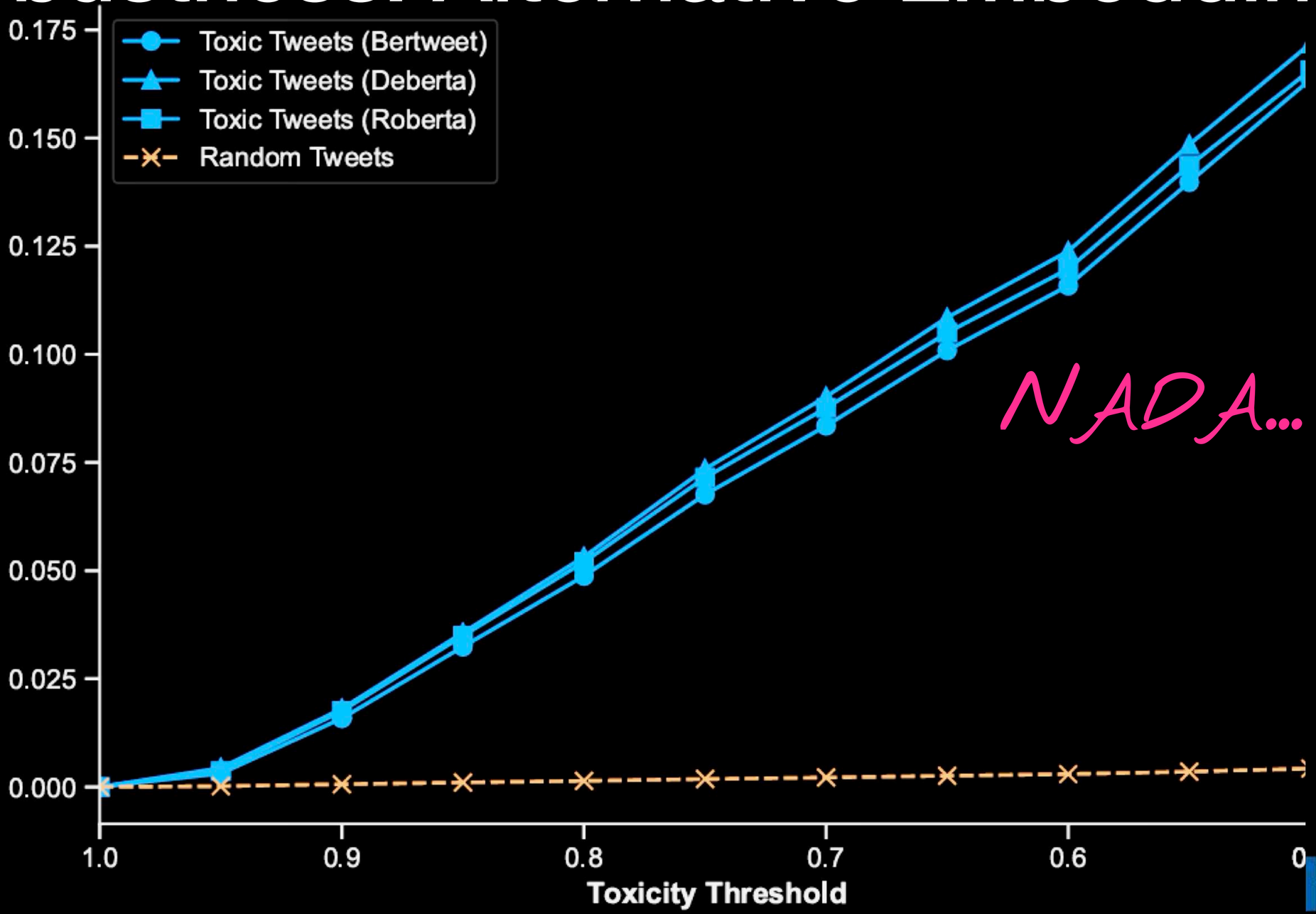


Content Moderation Does Distort Plurality

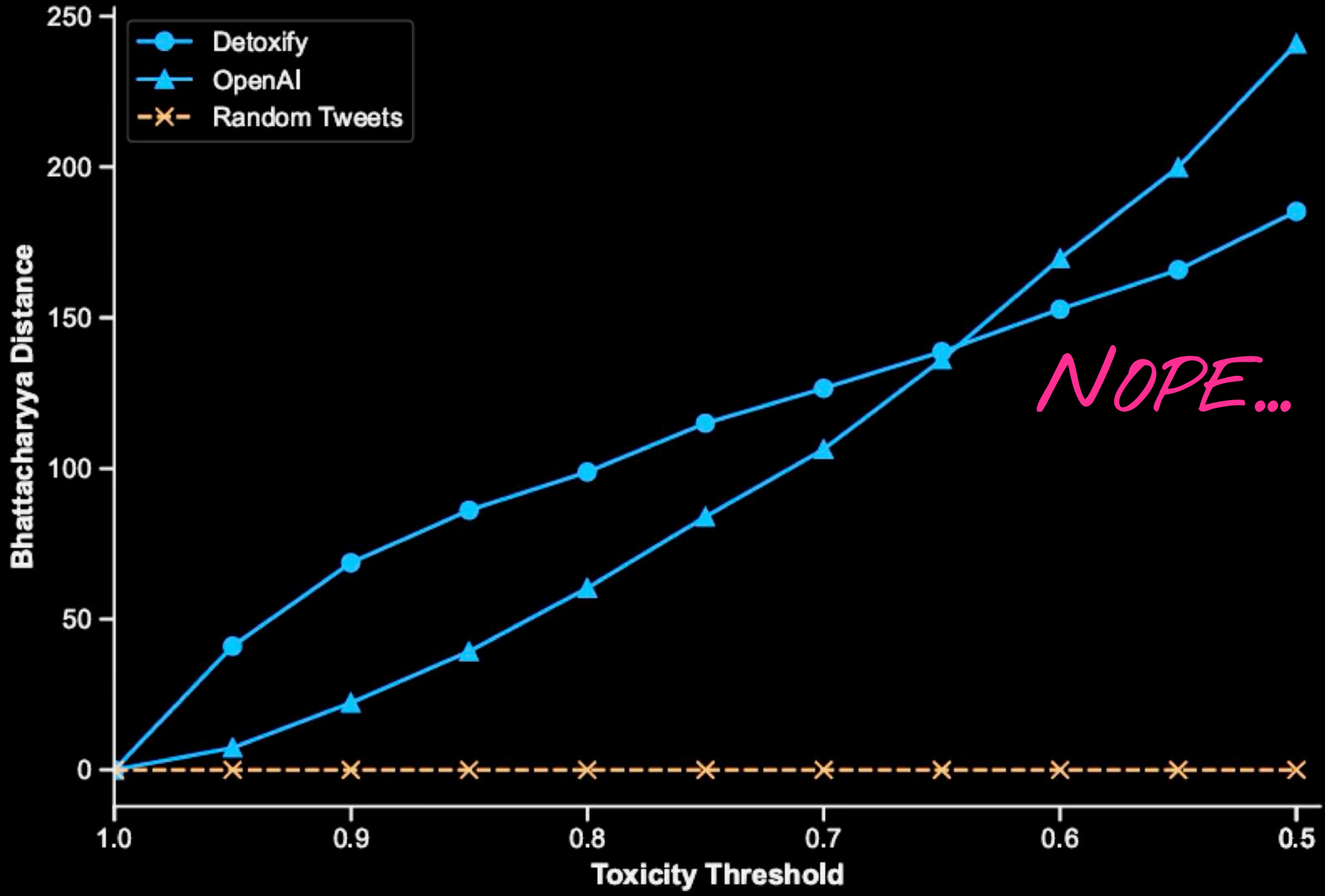
Robustness

- Ok, but: couldn't this be explained by other factors, e.g.,
  - the specific embedding model?
  - the toxicity measure?
  - the language or domain (political or not)?

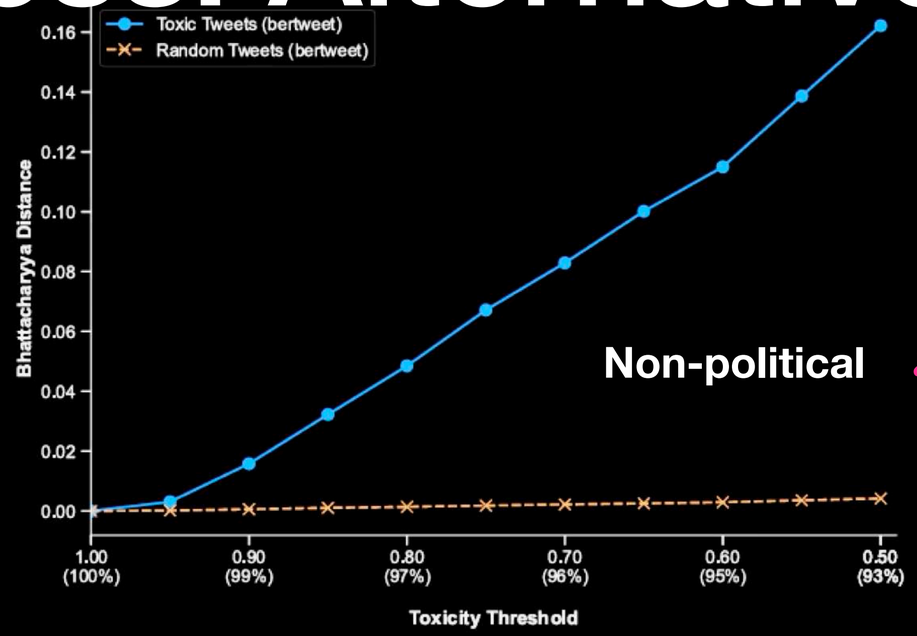
Robustness: Alternative Embeddings



Robustness: Alternative Toxicity Measures

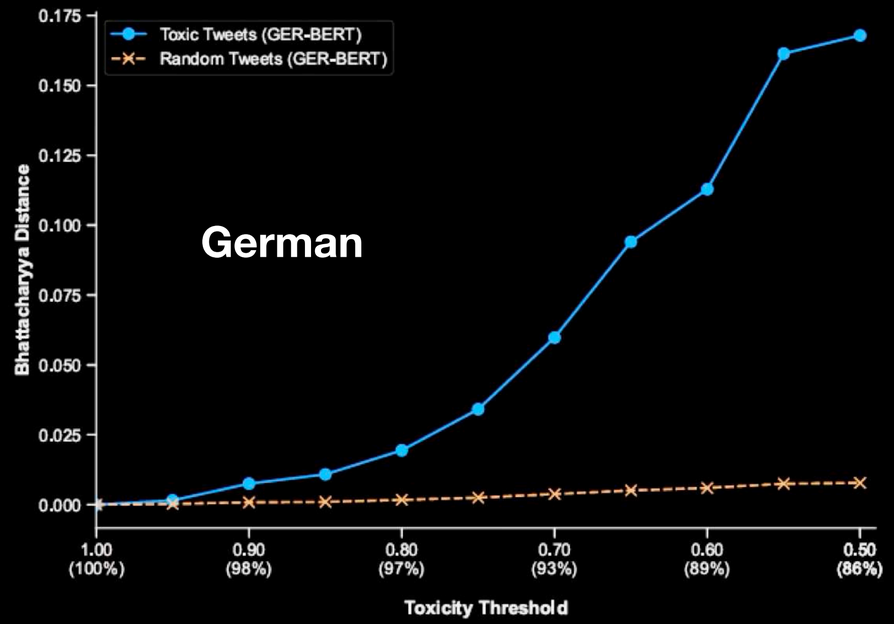


Robustness: Alternative Samples

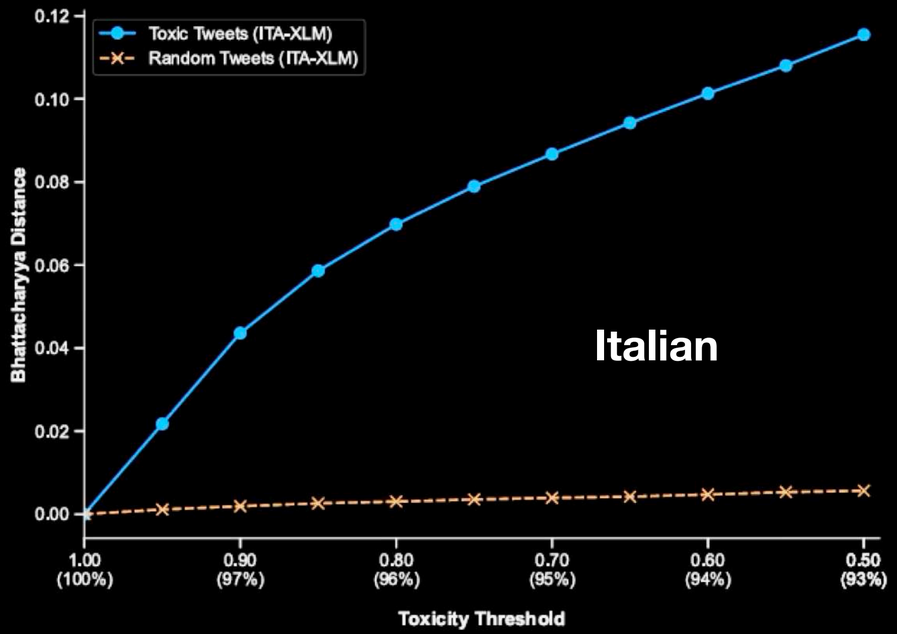


Non-political

MINOR DIFFERENCES, SAME STORY...



German



Italian

Alternative Content Moderation Strategy

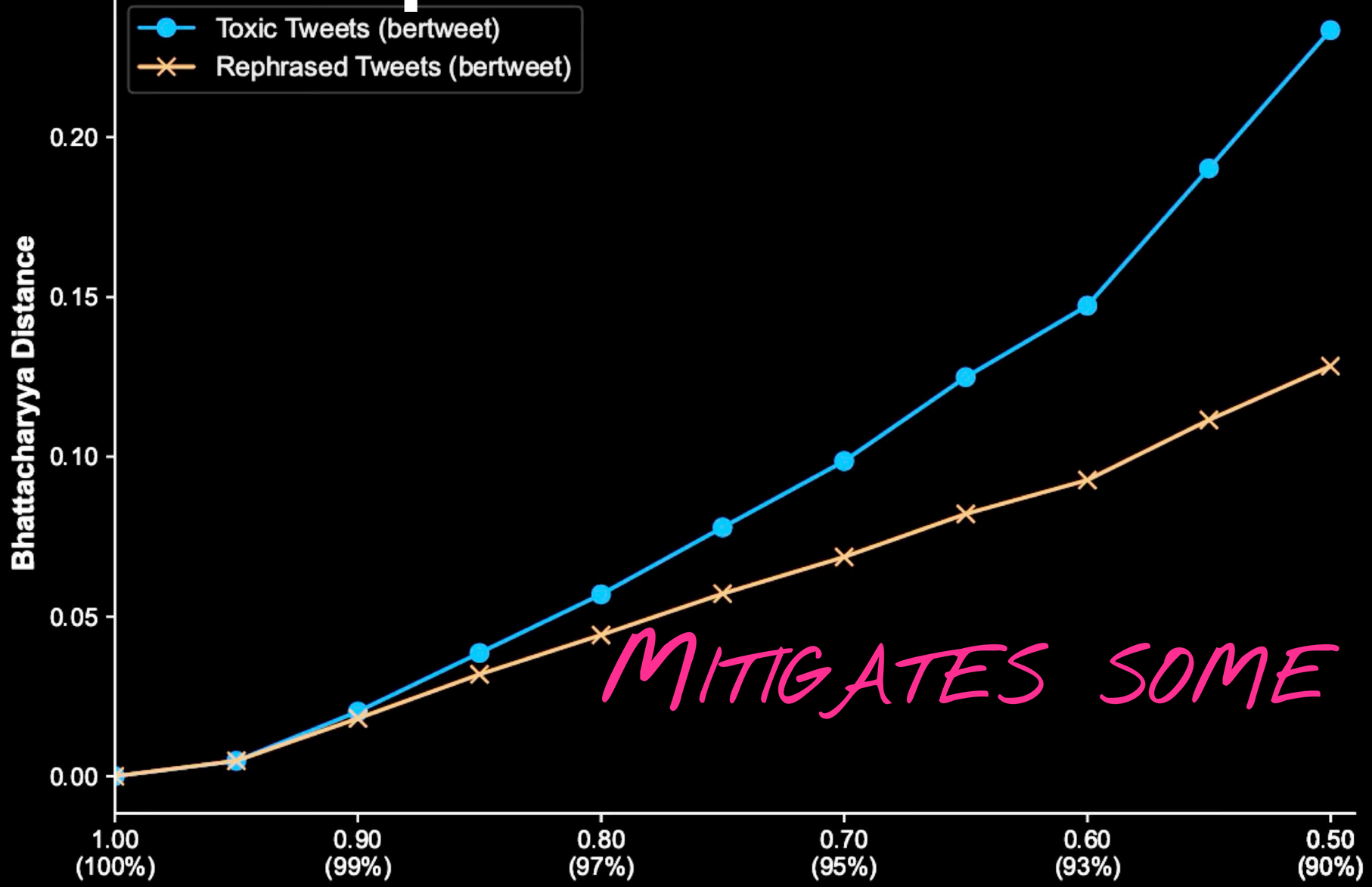
- Can we mitigate moderation-induced distortions?
- Some toxic posts might have salvageable content
- Get rid of toxicity without eliminating content *IDEA!*
- GPT Prompt:

```
Your task is to rephrase a highly toxic tweet and write
a less toxic version of it while aiming to make minimal
changes to the original tweet. It's crucial to preserve
the original wording, content, style, and tone in the
tweet. Keep the Twitter special elements such as RT and
XXX unchanged. Please respond in JSON format with the
key RephrasedText. Here is the tweet to rephrase:
```

Examples

Original	Toxicity	Rephrasing
What is wrong with people!! I don't get what burning the US flag proves. Side note: moron in the cowboy hat is not a real cowboy	(0.71, 0.97, 0.77)	What is wrong with people!! I don't get what burning the US flag proves. Side note: the *person* in the cowboy hat is not a real cowboy
@realDonaldTrump We need to stop having school shootings how bout you focus on that actually try to make a difference and take Obama d**k out your mouth	(0.93, 0.96, 0.97)	@realDonaldTrump We need to stop having school shootings. How *about* you focus on that, actually try to makea difference, and *move on from the past*?
If you're a Christian offended by foul language but not the rejection of refugees, you're not a f**king Christian.	(0.95, 0.98, 0.73)	If you're a Christian offended by foul language but not the rejection of refugees, you're not a *true* Christian.

Rephrased Posts



MITIGATES SOME LOSS...

Take Aways

- Content moderation distorts plurality of online content
- Not driven by toxicity alone (imperfect predictions)
- Bhattacharyya Distance robust measure of embedding plurality
- Rephrasing might mitigate some content loss

Russo et al. (2025)

Value Calibration Improves LLMs' Pluralism and Moral Judgment Abilities

Peppe Russo
EPFL
Switzerland

Paul Röttger
Bocconi University
Italy

Debora Nozza
Bocconi University
Italy

"LLM HAS VALUE X" = "LLM SHOWS BEHAVIOR CONSISTENT WITH A PERSON HAVING VALUE X"

What Moral Values do LLMs Reflect?

- People turn to LLMs for moral guidance/actions
- Models should reflect range of human values
- Evaluating LLMs' moral guidance is critical to prevent real-world harms

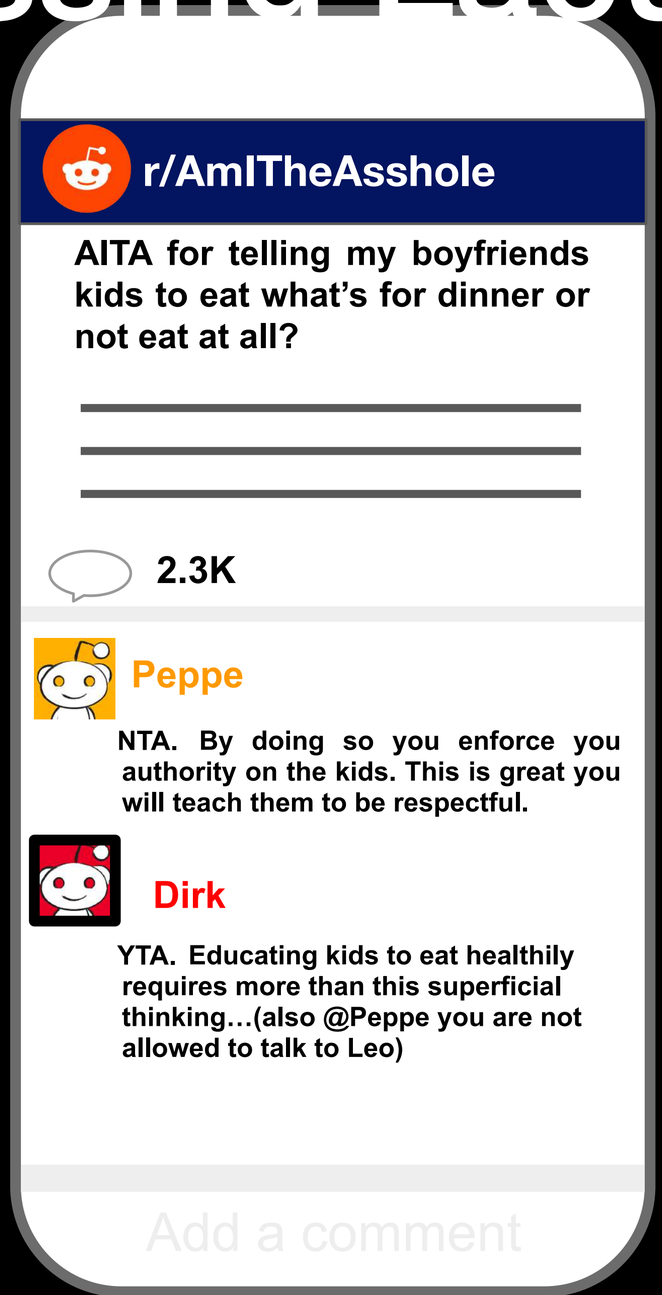
RQ1: How aligned are LLMs' and humans' moral judgments?

RQ2: What values drive LLM advise?

RQ3: How can we steer LLMs values to be more diverse?

Addressing Lacunae

Issue 1: Very **unrealistic** situations



Reply can be NTA or YTA + Rationale

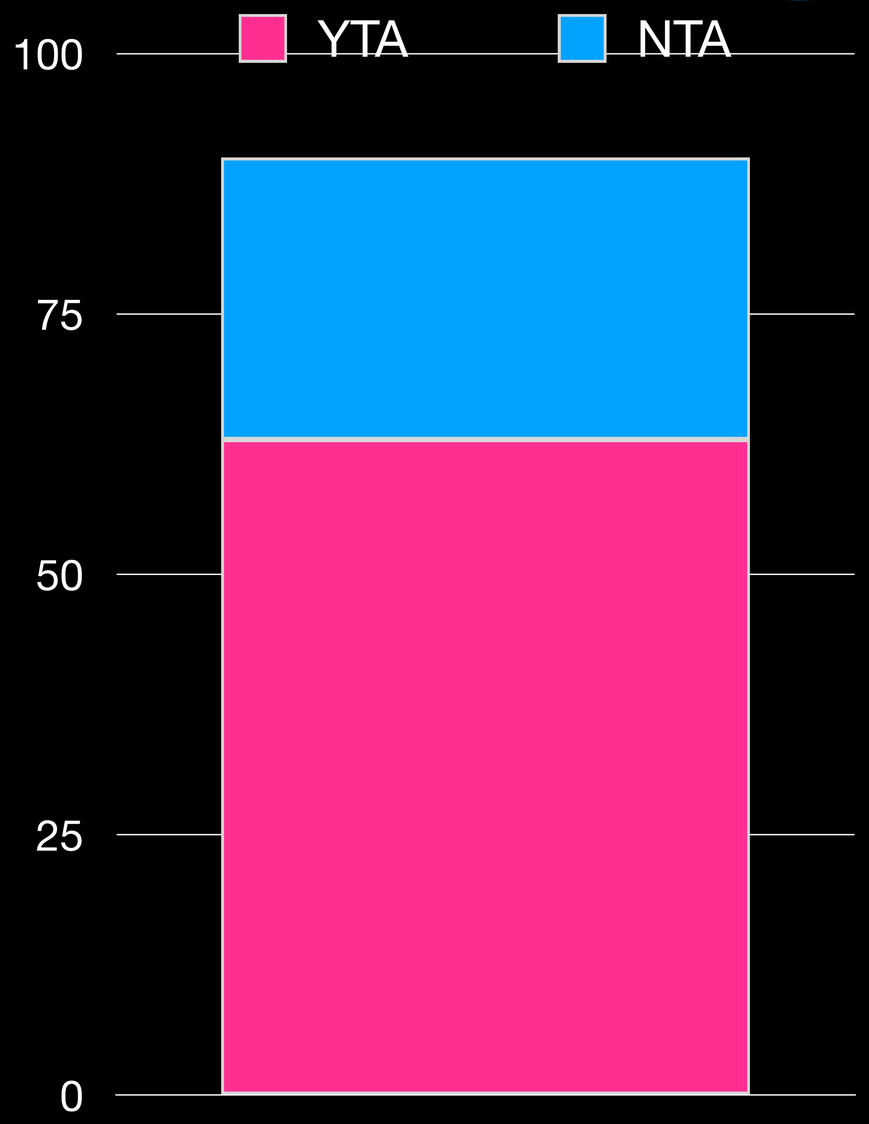
Issue 2: Lack of comparison with human judgments

Issue 3: Typically evaluated over multiple choice questions

Human Opinions

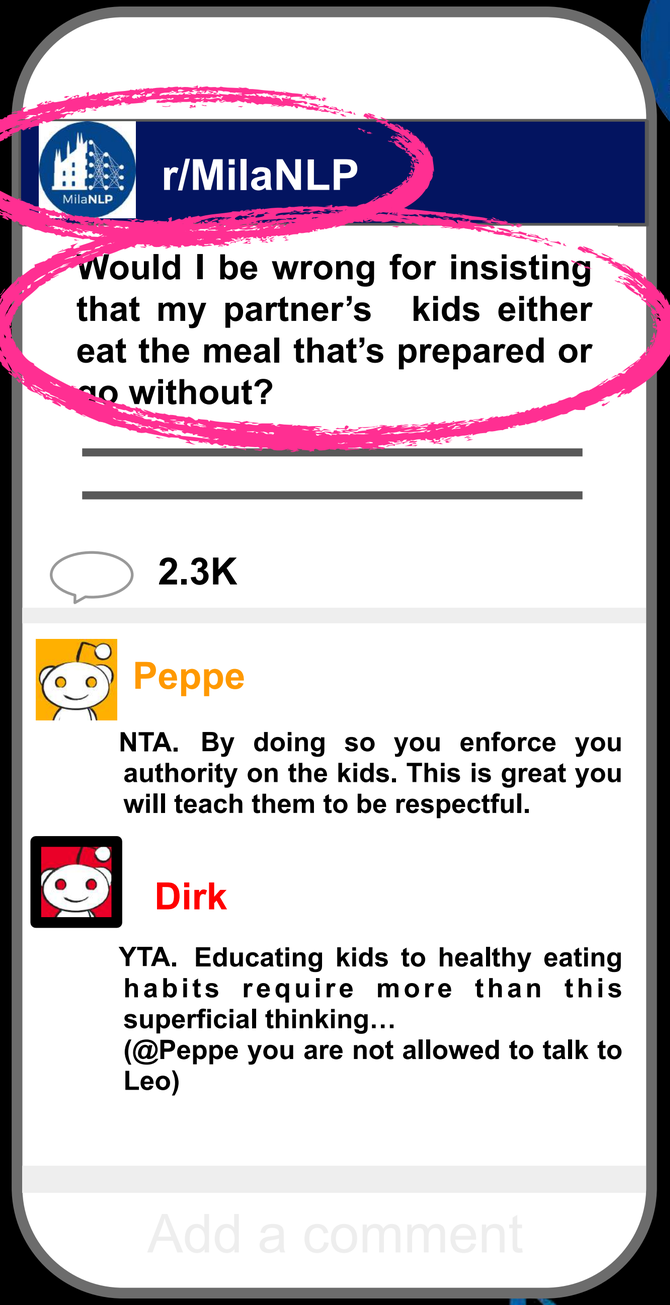
Data

- `r/AmITheAsshole` has 22M active users
- ~150 comments/post (only first level comments)
- Average ~43 judgments per situation
- All posts we collected are *after* GPT cutoff
- (Partial) access to the demographics

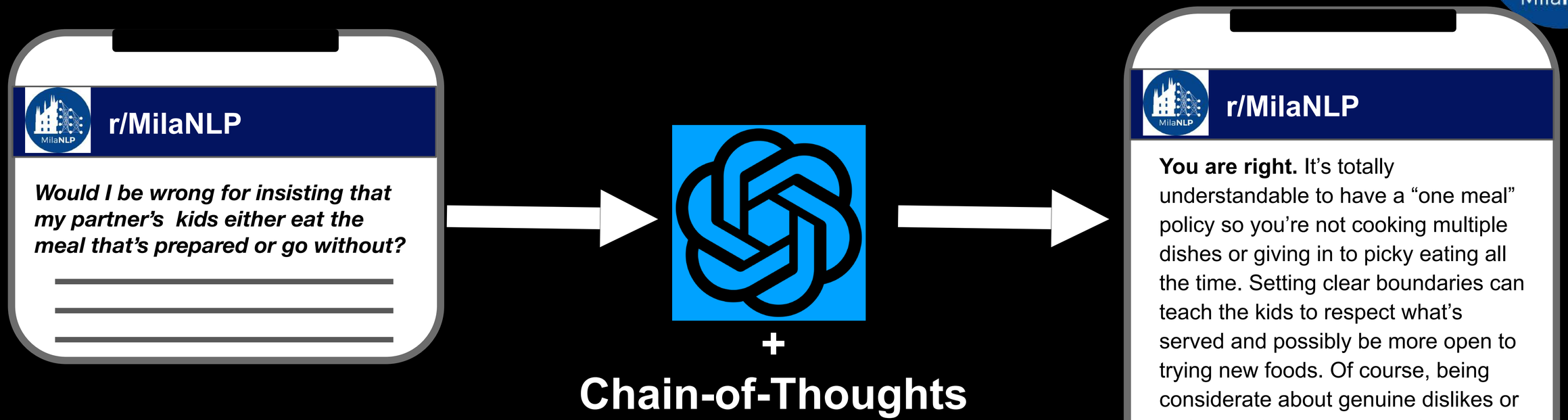


Data Processing

- Prevent data contamination effects:
  - Remove AITA elements from main post
  - Rewrite posts to more abstract situations
- Check if GPT can still recognize the post
- Keep only unrecognized posts:
  - 1618 moral situations
  - ~52k judgments



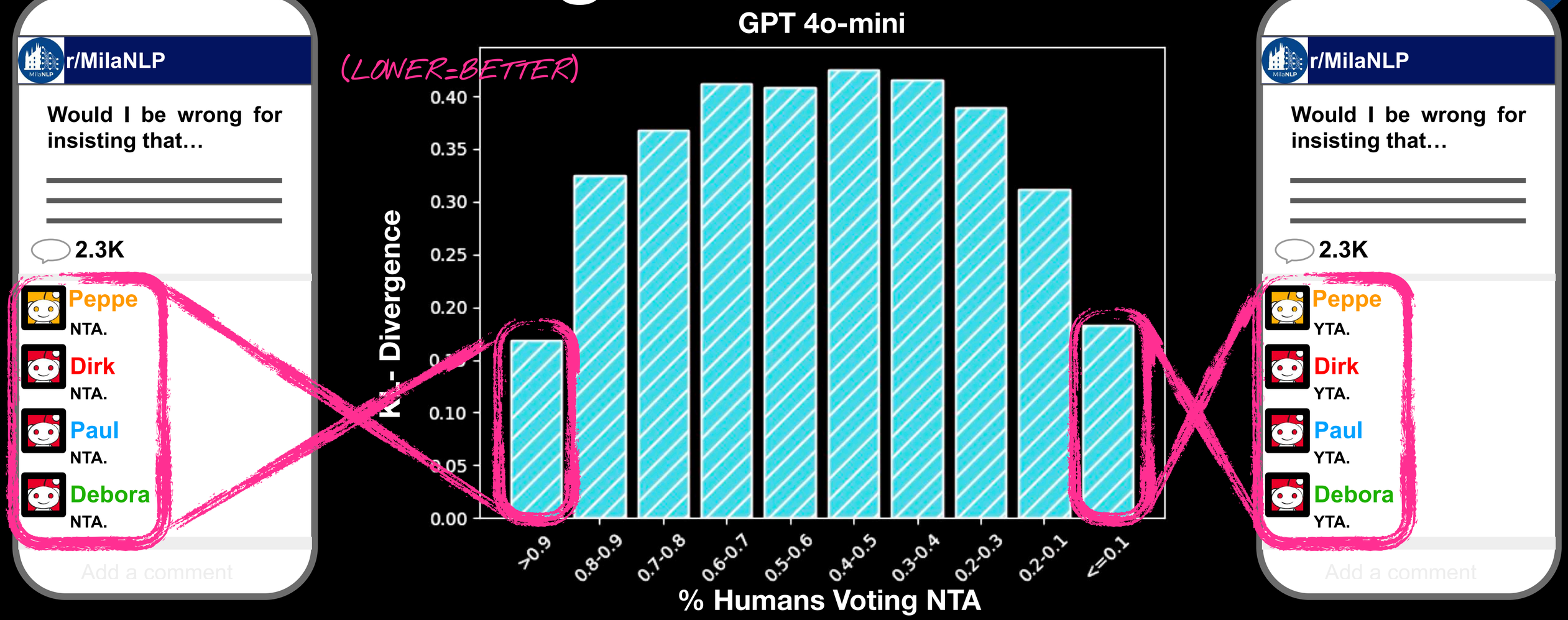
Method



- Run pipeline for all 1618 situations
- Prompt LLMs repeatedly for each
- Compute distribution over right and wrong
- Compare to distribution of human answers

RQ1: How aligned are LLMs' and humans' moral judgments?

Judgment Time

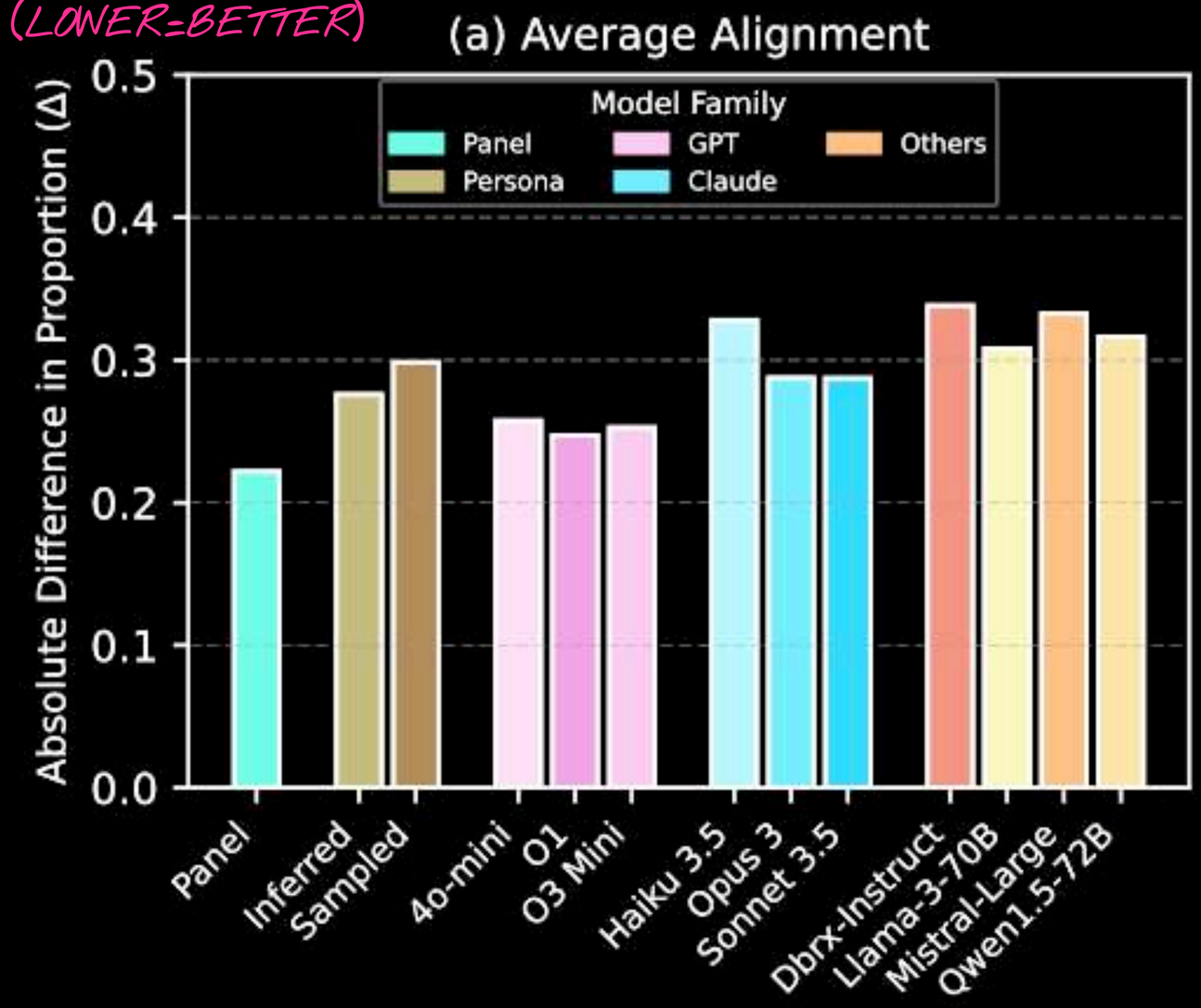


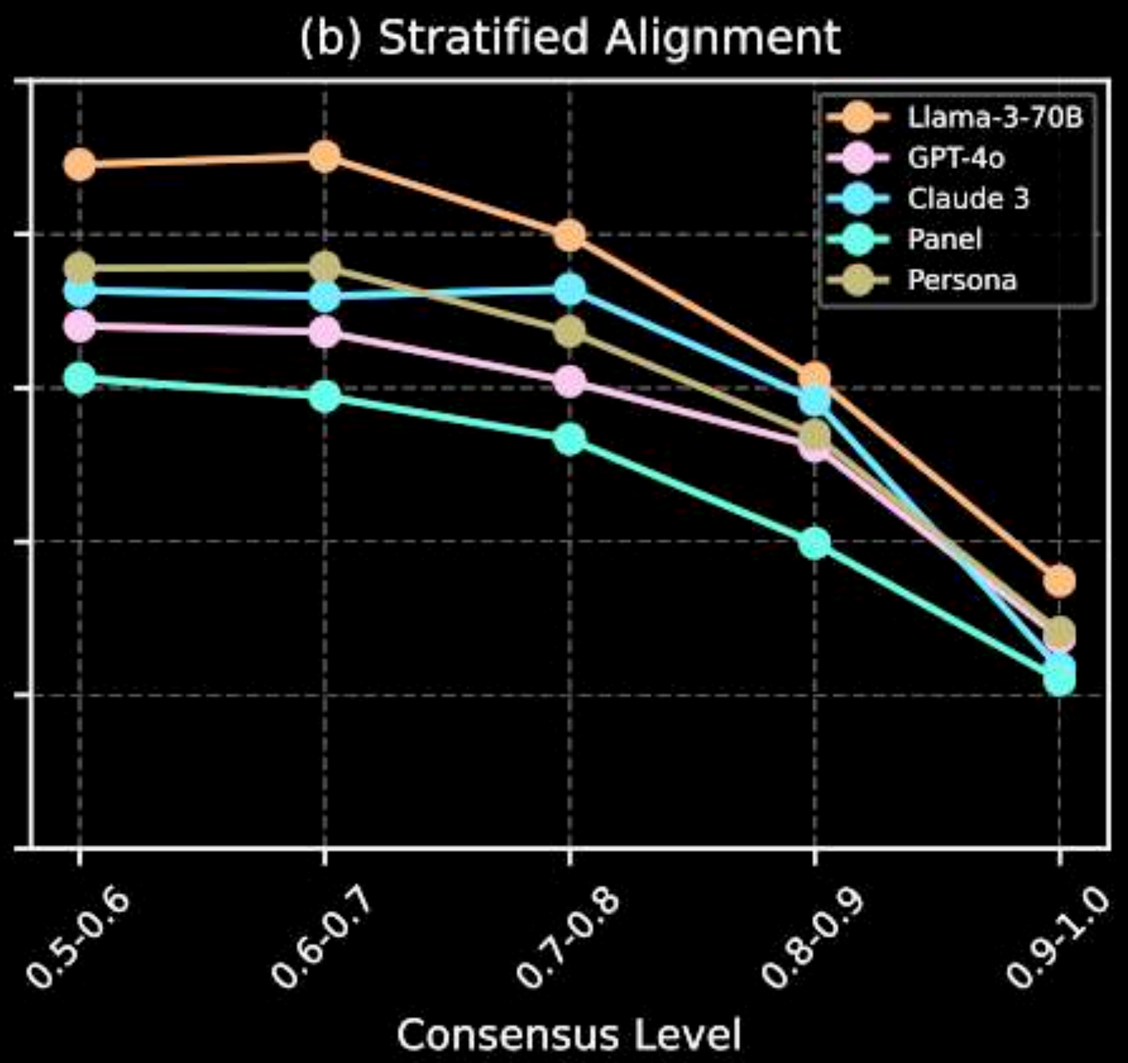
- Models perform well on human high-consensus cases
- Performance drops progressively for controversial situations

Bocconi

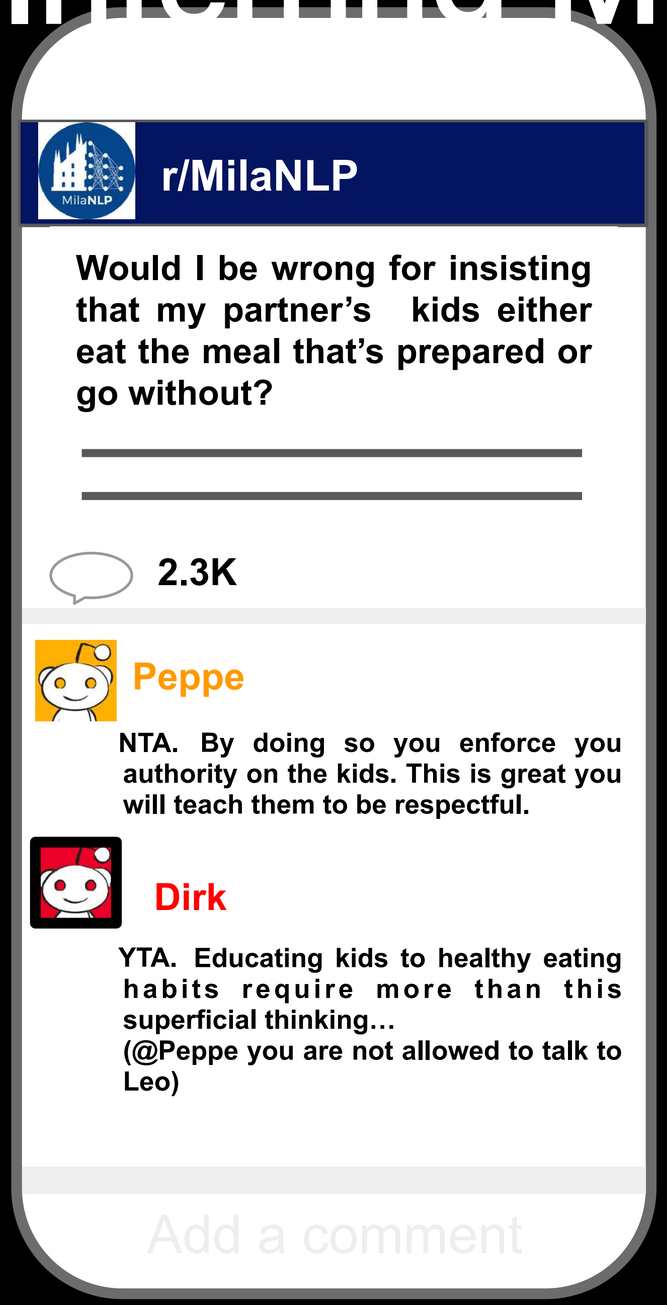
Judgment Time

(LOWER=BETTER)

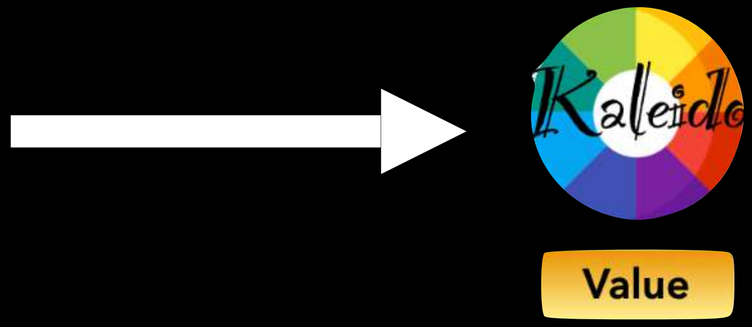




Inferring Moral Dimensions from Data



ADAPT KALEIDO MODEL TO ONLY OUTPUT "VALUES"



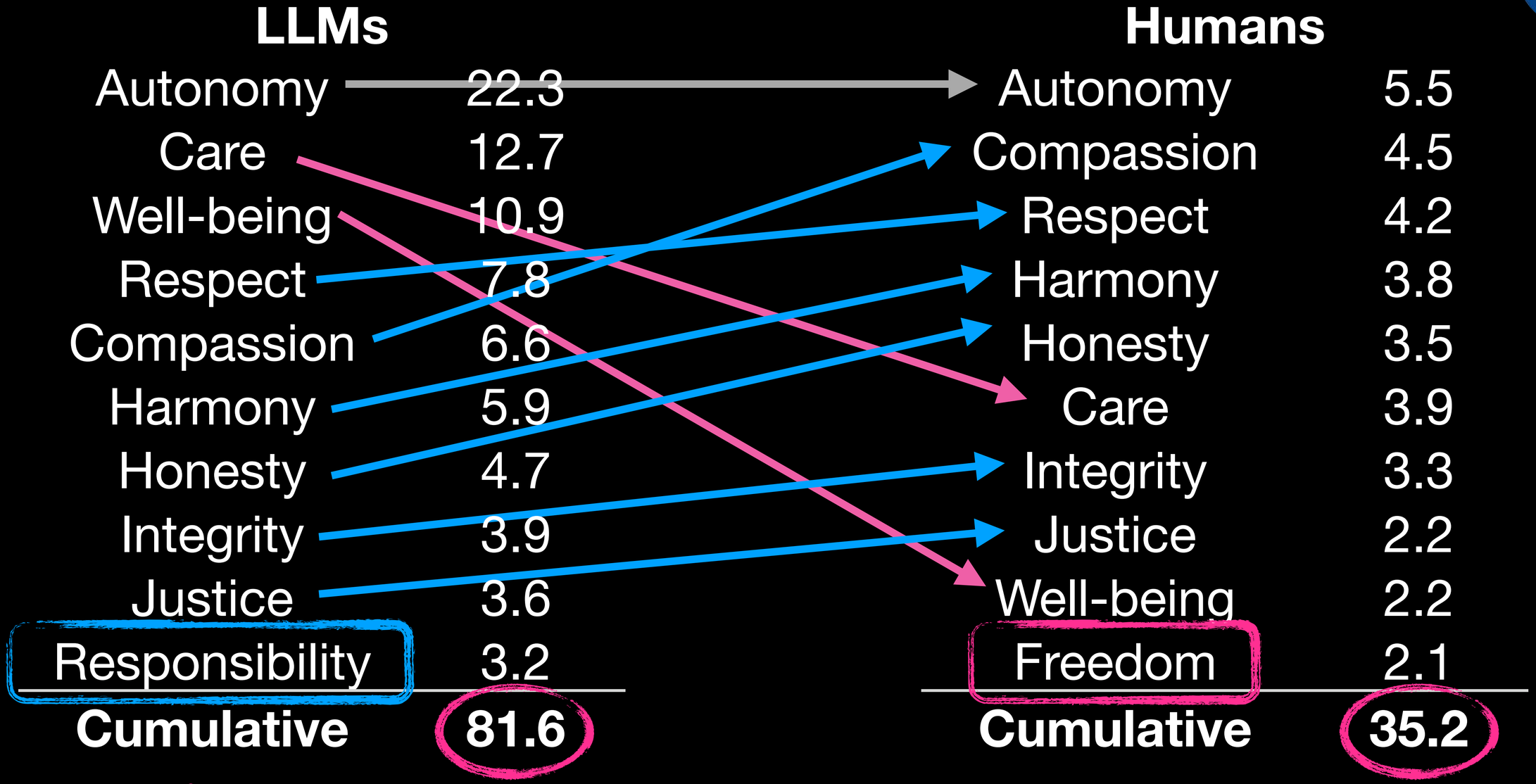
- Well being of the child
- Others well-being
- Education
- …
- Nutrition
- Authority
- Honesty

CONS:

- HIGHLY REPETITIVE
- UNSTRUCTURED
- NOT BASED ON SOCIAL THEORIES

RQ2: What values drive LLM advise?

Value Differences

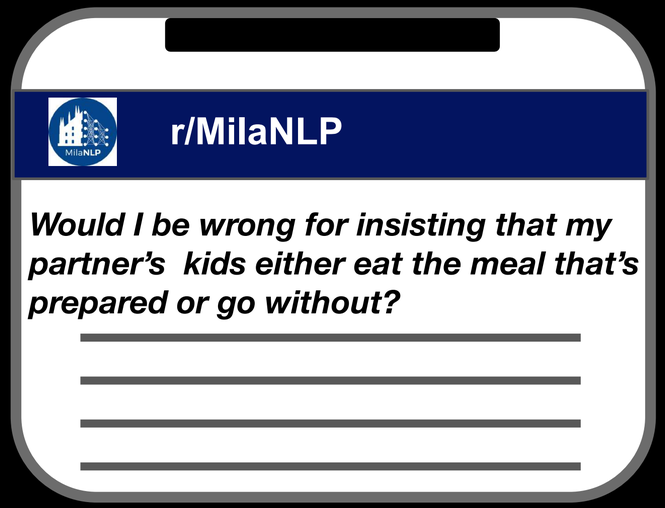


LLMs OVERUSE FEW VALUES, NO PLURALISM

Encouraging Value Plurality

Use DMP, Dirichlet Process to increase diversity of moral values used by LLMs:

$$G_0(v) = \frac{\sum_t \text{Count}(t, v)}{\sum_t \sum_v \text{Count}(t, v)} \log(\alpha) - \frac{1}{\alpha} \sum (G_0(v)) - 1$$

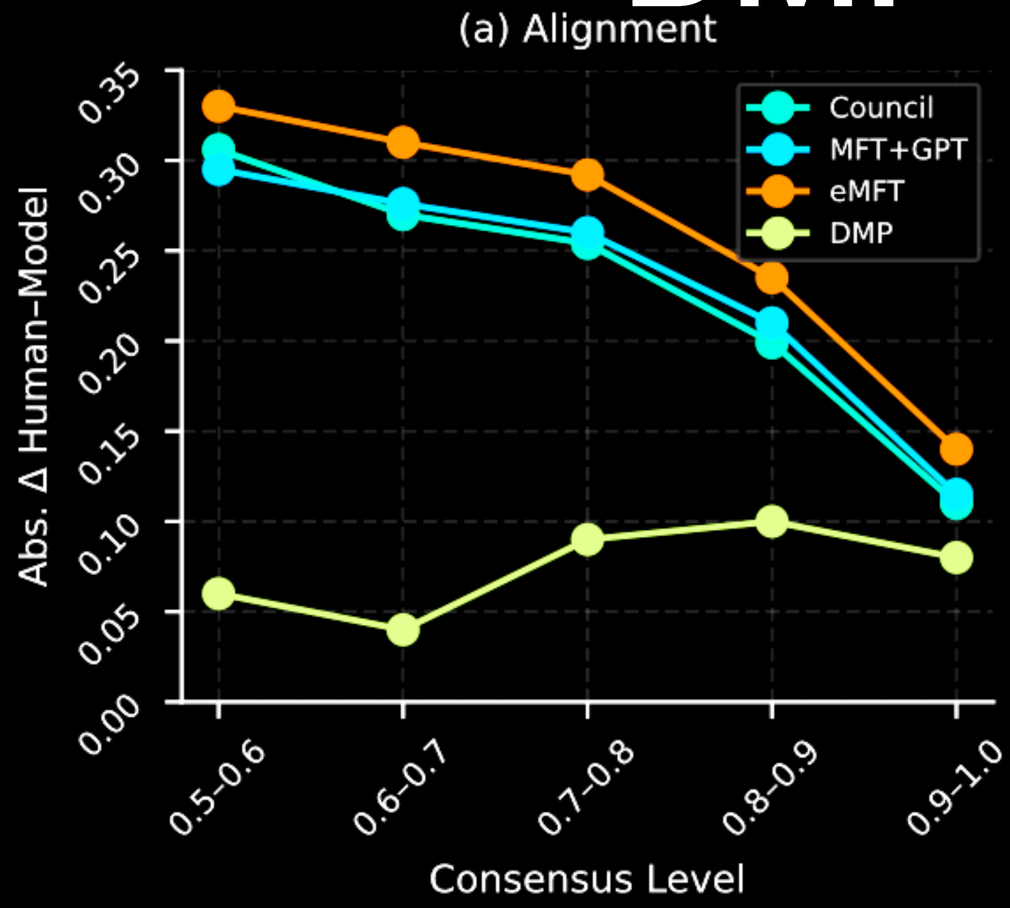


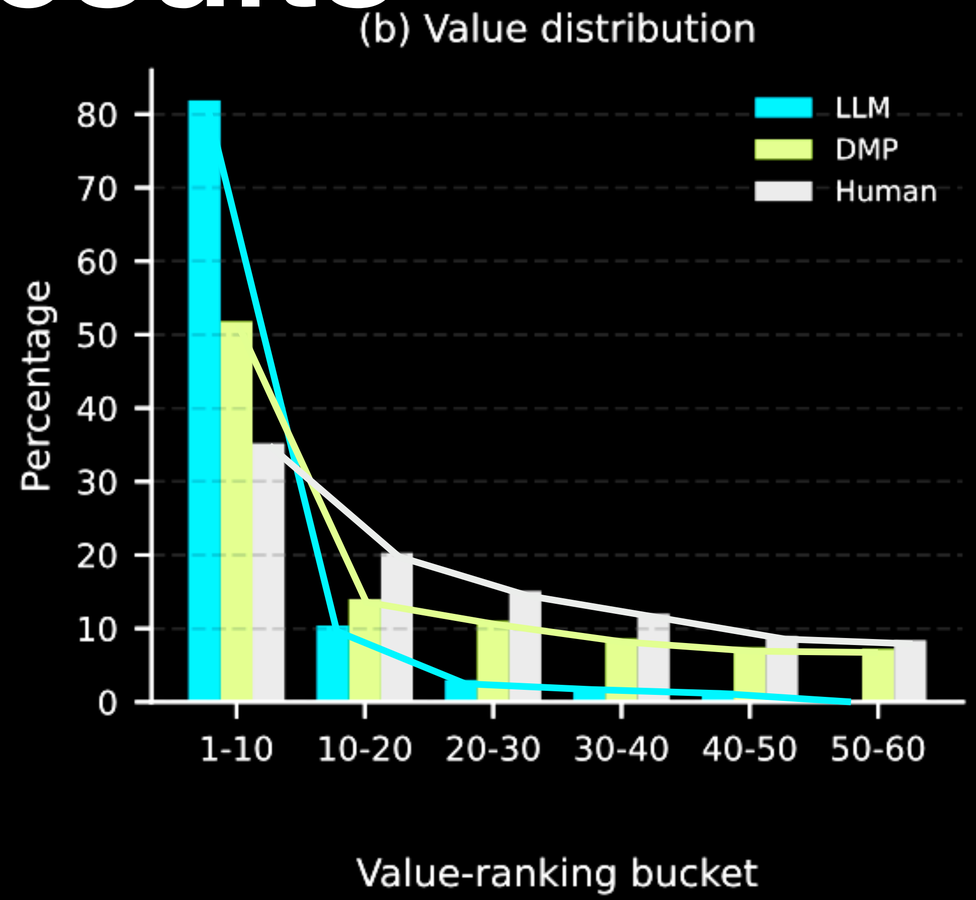
Sample N-moral profile

$$G_v \sim DP(\alpha, G_0)$$

RQ3: How can we steer LLMs values to be more diverse?

DMP Results



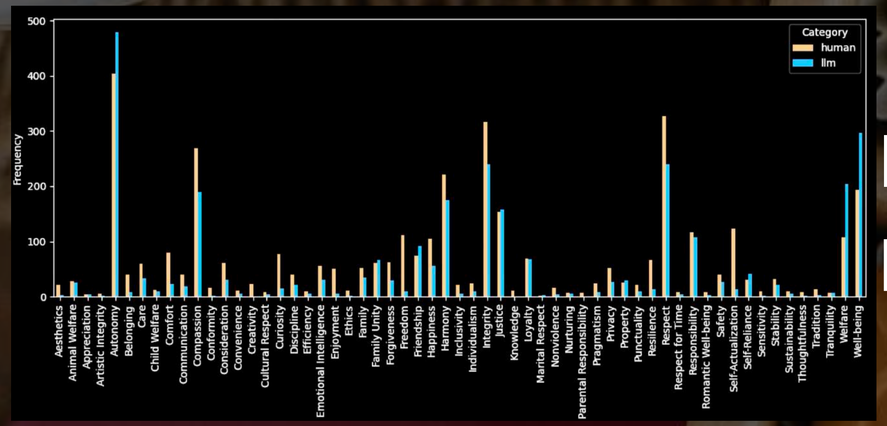


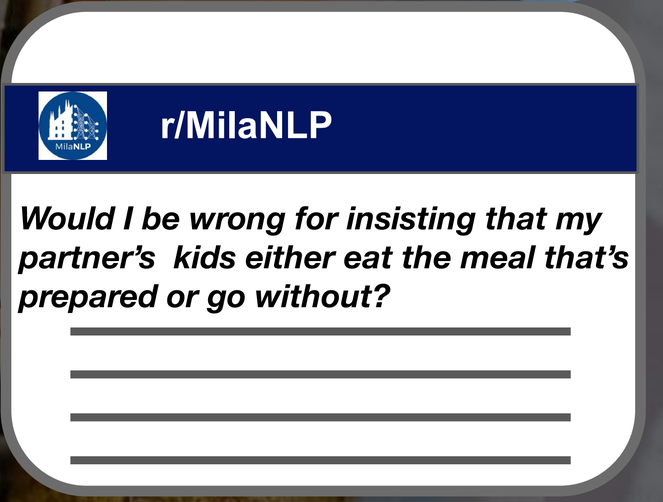
DMP SIGNIFICANTLY IMPROVES VALUE PLURALISM!

RQ3: How can we improve LLMs performance on moral guidance?

Take Aways

Benchmark with highly realistic moral judgment scenarios





Insights into moral values alignment between LLMs and humans

Sample N-moral profile

$$G_v \sim DP(\alpha, G_0)$$

Scalable approach to elicit pluralism in moral judgment

Wrapping up…

Conclusion

- Language is central to social science
- Tons of exciting questions left to answer
- Findings will also inform better NLP models
- Still lots of fun to be had with NLP


MilaNLP
Thank you!
milanlproc.github.io/
@dirkhovy.bsky.social
www.dirkhovy.com
INTEGRATOR
European Research Council
Bocconi

Questions?

milanlproc.github.io/

@dirkhovy.bsky.social
www.dirkhovy.com

INTEGRATOR
European Research Council

MilaNLP

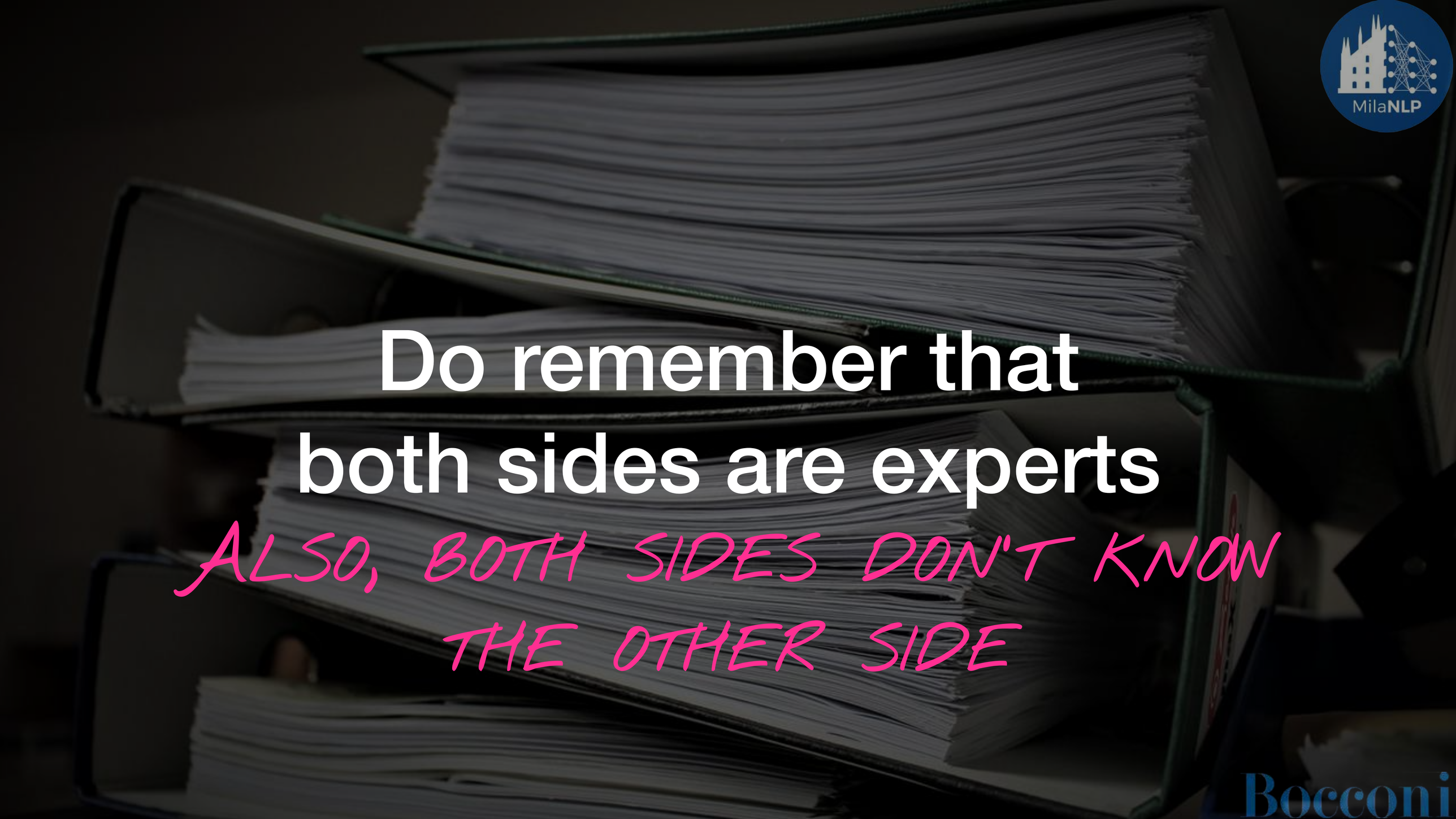
Bocconi

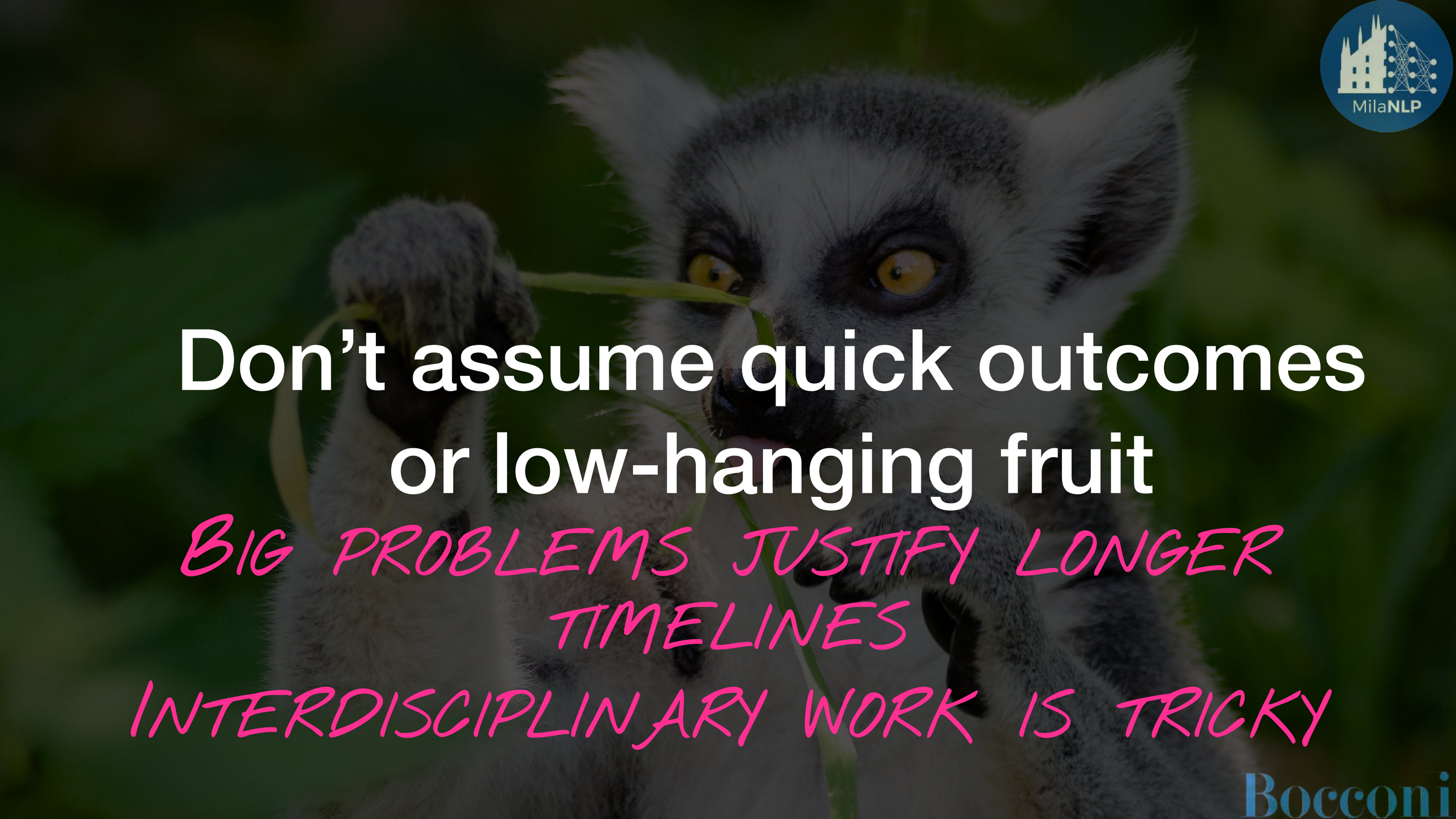
Do’s and Don’ts

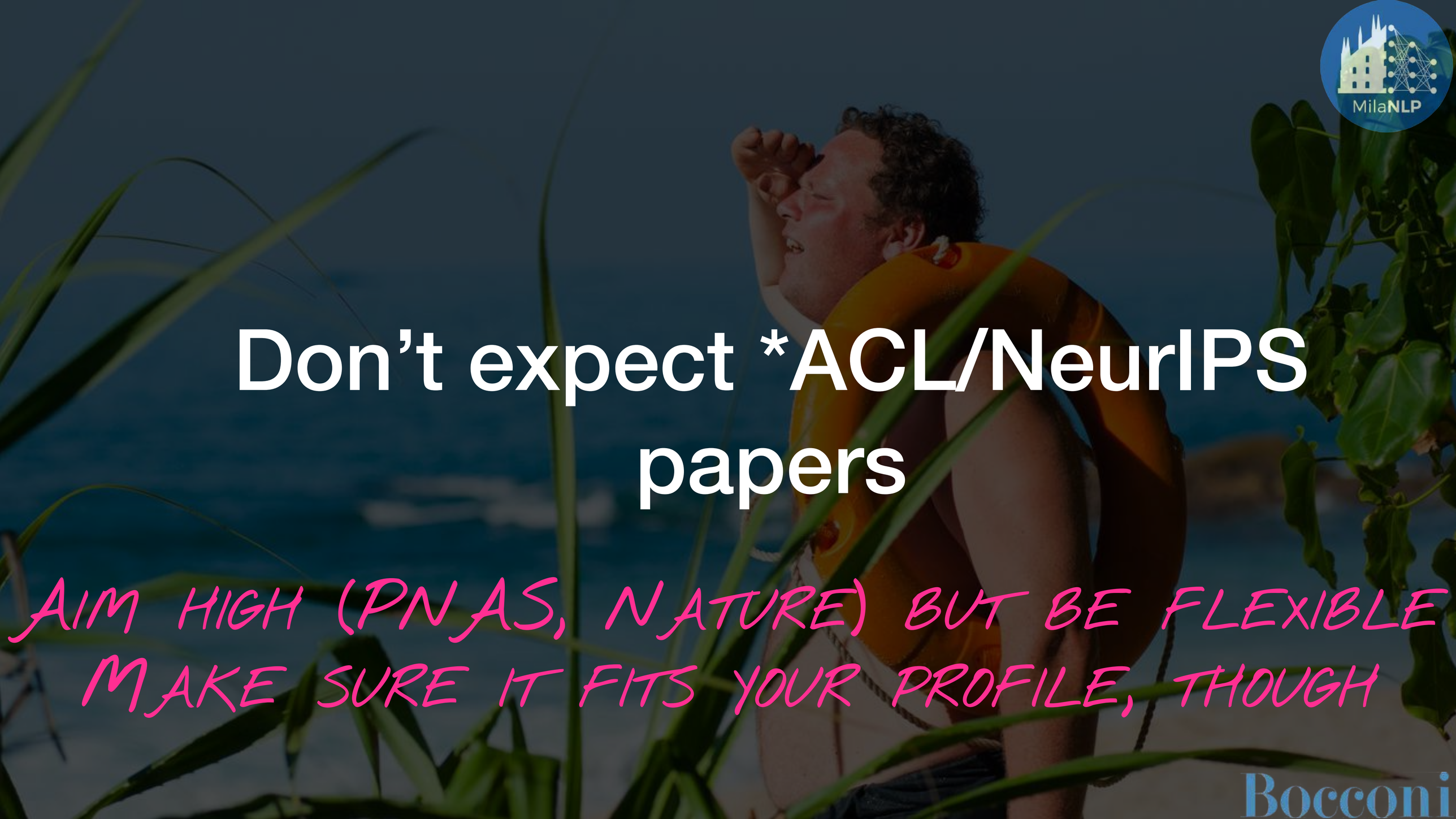
Based on first-hand experience and advise from

Tim Baldwin
University of Melbourne
Australia


Do be explicit
TIMELINES
PUBLICATION EXPECTATIONS
WHAT IS A CONTRIBUTION?
MilaNLP
Bocconi


Do remember that
both sides are experts
ALSO, BOTH SIDES DON'T KNOW
THE OTHER SIDE
MilaNLP
Bocconi


MilaNLP
Don’t assume quick outcomes or low-hanging fruit
BIG PROBLEMS JUSTIFY LONGER TIMELINES
INTERDISCIPLINARY WORK IS TRICKY
Bocconi



Don’t expect *ACL/NeurIPS papers

AIM HIGH (PNAS, NATURE) BUT BE FLEXIBLE
MAKE SURE IT FITS YOUR PROFILE, THOUGH

MilaNLP

Bocconi


MilaNLP
Don’t think algorithms solve everything
MOST ALGOS WORK ON WELL-DEFINED PROBLEMS
THE REAL (SOCIAL) WORLD IS MESSY
Bocconi


MilaNLP
Don’t worry about SOTA
Better performance ≠ more insights
Try Simple Models First
Explainability Counts for a lot
Bocconi

...JUST DO IT...

GIVE IT YOUR BEST SHOT AND HAVE FUN!