# Stochastic Processes for Inference

## An application to Authorship Attribution and Evil

 ←This is actually in the Sapienza template!

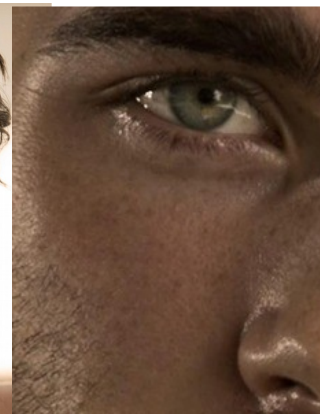
# How to find the author of a text

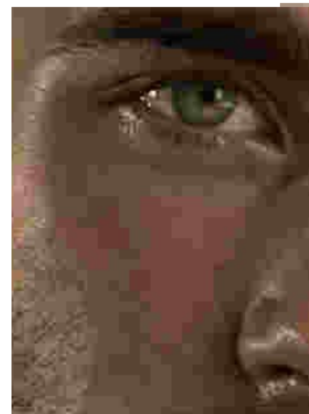
1) Go to the restaurant
2) Find the author
3) Turn evil

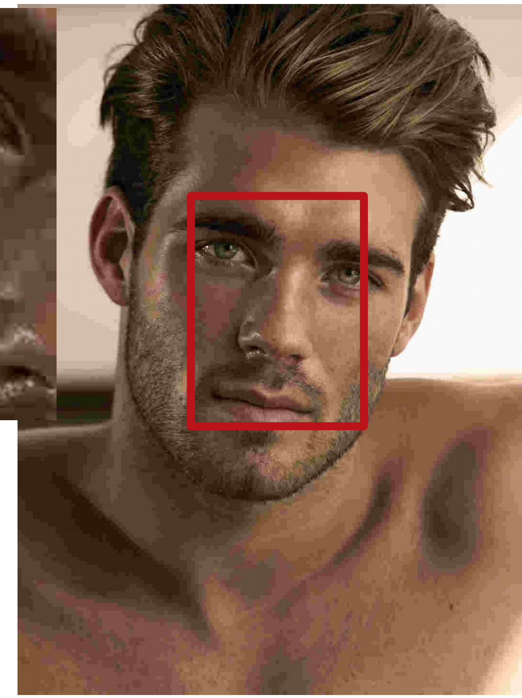
# Why don't we go deep learning?

No compression

High compression







We are not always this lucky

2600 pages of text =
2 x War and Peace

10 pages of text

# So what?

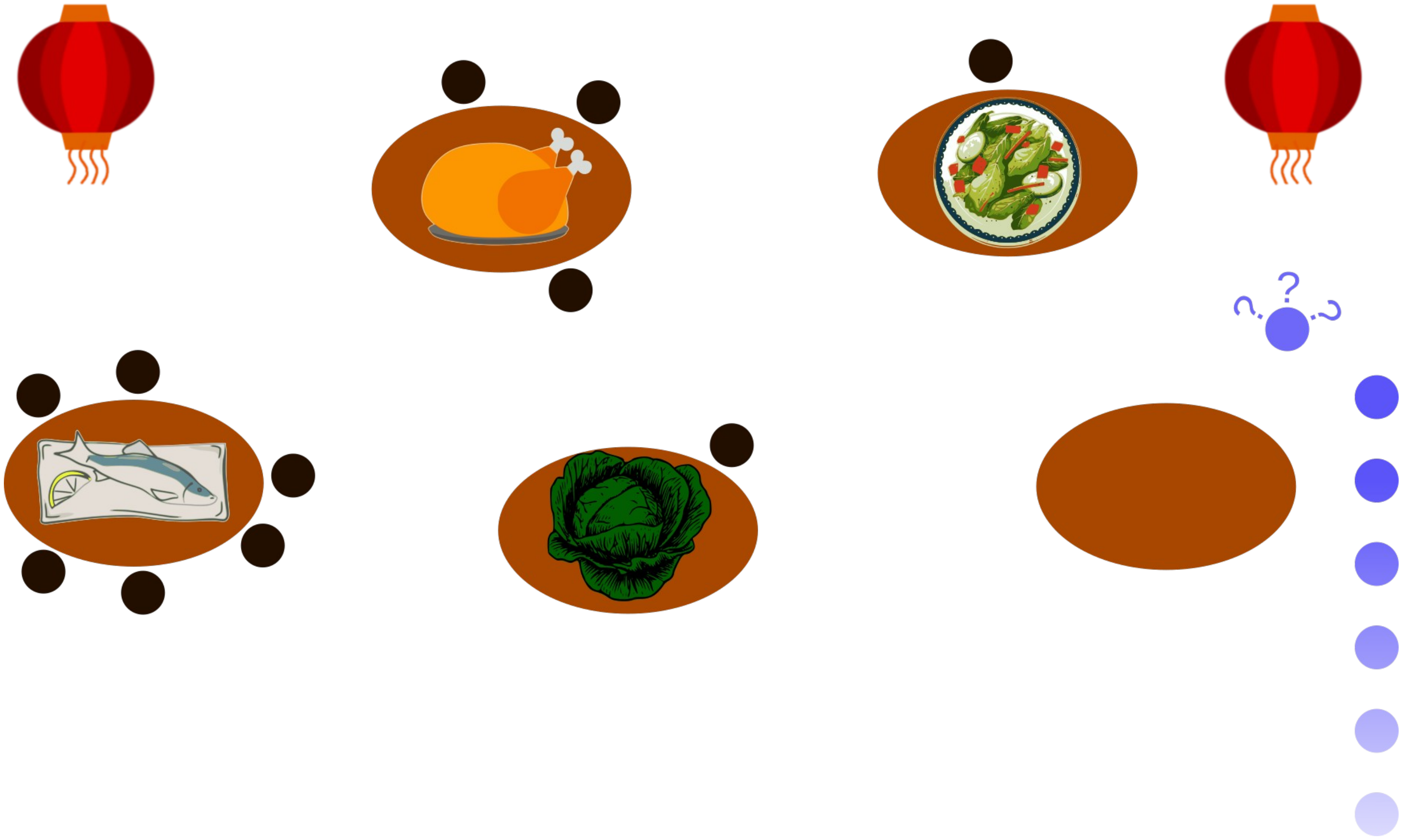
- We may use traditional classifiers
  - We need to learn class boundaries
  - Problems with many (thousands) classes
- We may use statistical inference
  - We need to infer the parameters
  - Used already in the ‘90 for RNA (HMM)

# How to choose a model?

- Must be able to work with an unbounded vocabulary
  - OK, no language has unbounded vocabulary, but then a German names a law: “Rinderkennzeichnungs- und **Rindfleischetikettierungsüberwachungsaufgabenübertragungsgesetz**”… so lets say around $10^{95}$
- Must have as few parameters as possible
  - The fewer the parameters, the less data needed to (roughly) infer them
  - The fewer the parameters, the happier the physicist

**Go to the restaurant!**

# Chinese Restaurant Process



# Chinese Restaurant Process – 2

Probability of the next element:

$$P(x^*_{n+1} = \cdot | x_1, \ldots, x_n, \alpha, \theta, P_0) = \frac{\theta + k_n \alpha}{\theta + n} P_0(\cdot) + \sum_{j=1}^{k_n} \delta_{y_j, \cdot} \frac{n_j - \alpha}{\theta + n}$$

# Poisson—Dirichlet Process

$$P \sim PD(\alpha, \theta, P_0)$$

$$P(\cdot) = \sum_{i=1}^{\infty} p_i \delta_{y_i},.$$

The CRP is a sequential sampling from *P*

Good for inference:

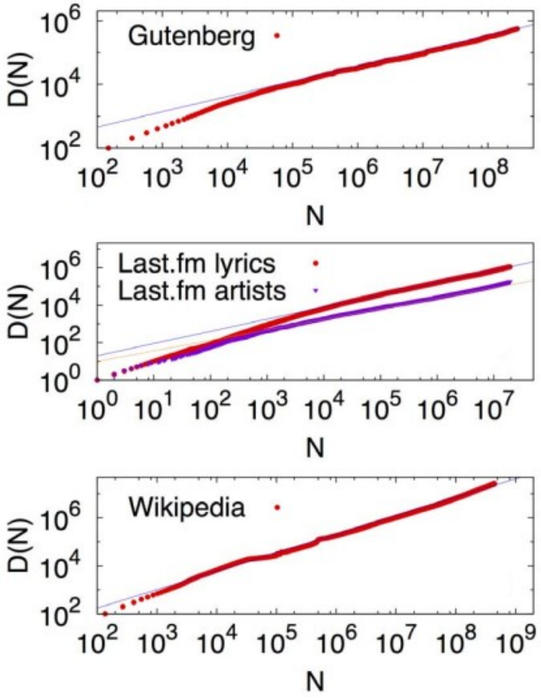
- *Conjugacy*
- *Exchangeability*
- *Statistic properties* → *power-law behaviours*

# Heaps' Law

- Power-law relation between the number of elements and the number of different elements

$$k \propto n^{\beta}$$

$$\beta \leq 1$$



# Zipf's Law

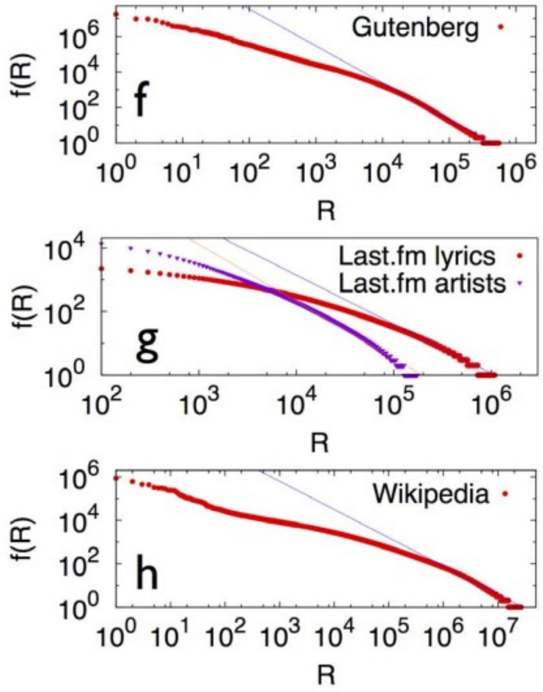
- Power-law relation between the frequency of an element and its rank

$$f \propto R^{-\alpha}$$

Actually holds whenever:

$$P(f) \propto f^{-1-\frac{1}{\alpha}}$$

$$\left( \beta = \frac{1}{\alpha} \right)$$



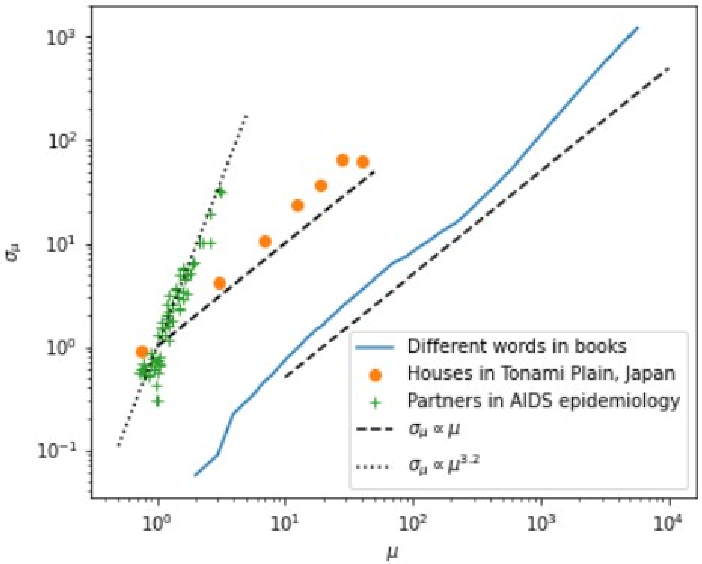
# Taylor's Law

- Relation between different systems
- Relation between the deviation and the mean

$$\sigma \propto \mu^{\gamma}$$

where:

$\gamma = \frac{1}{2}$ = random sampling



# Poisson—Dirichlet Process – 2

$$P(x^*_{n+1} = \cdot | x_1, \ldots, x_n, \alpha, \theta, P_0) = \frac{\theta + k_n \alpha}{\theta + n} P_0(\cdot) + \sum_{j=1}^{k_n} \delta_{y_j, \cdot} \frac{n_j - \alpha}{\theta + n}$$

- Blunt approximation of a language model but:
  - Has only two parameters
  - Doesn't require context (like N-gram models)
  - Doesn't require strange and fragile language tools (lemmers, stemmers, PoS taggers, …)
  - Gets the broad (statistical) picture

# Note on $P_0$

Is $P_0$ continuous or discrete?

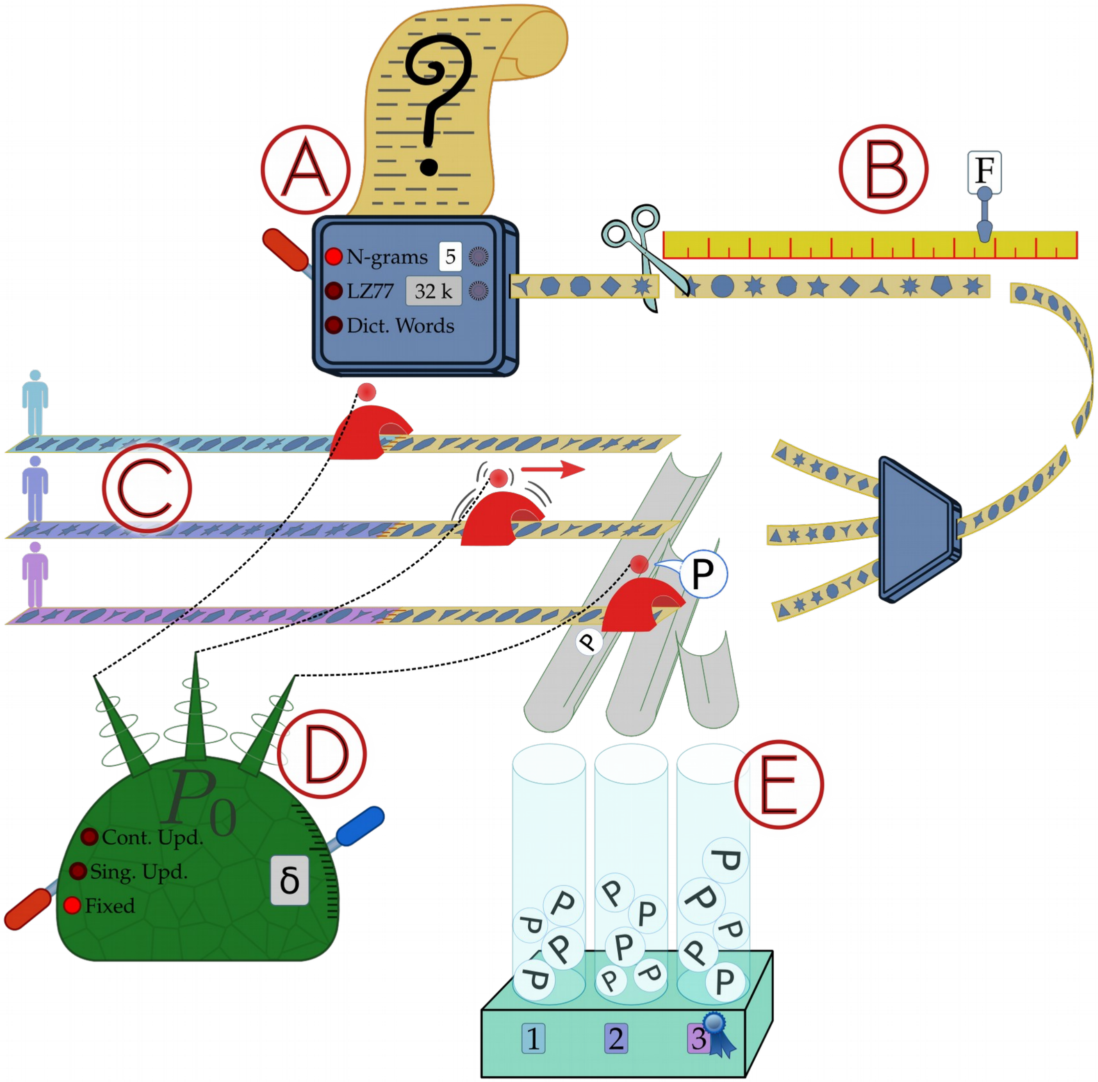
$$P(x^*_{n+1} = \cdot | x_1, \ldots, x_n, \alpha, \theta, P_0) = \frac{\theta + \alpha \sum_{j=1}^{k} t_j}{\theta + n} P_0(\cdot) + \sum_{j=1}^{k_n} \delta_{y_j, \cdot} \frac{n_j - \alpha t_j}{\theta + n}$$

This is terrible!!

We'll use the **C**ontinuous version of the **P**rocess but a **D**iscrete **$P_0$**…

CP—DP!

# Plan for the attribution task



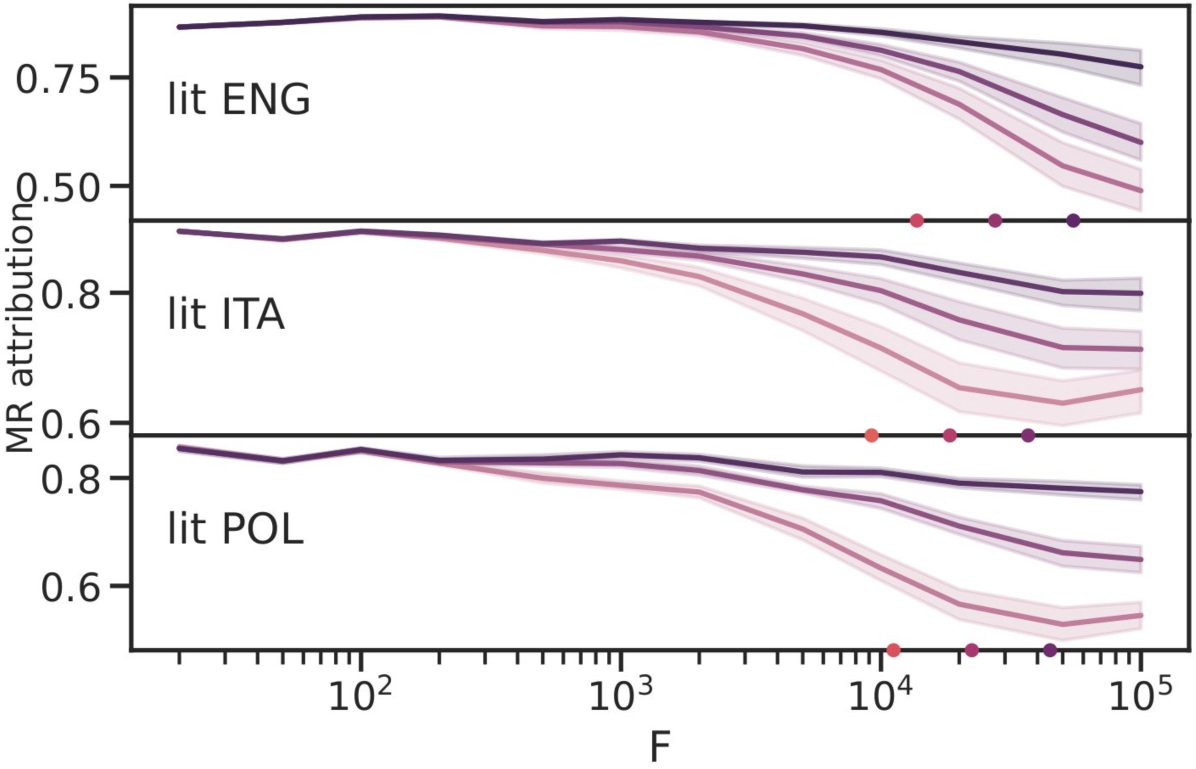
# How to choose the tokens

- Space-separated words
- (Overlapping Space-Free) Character N-grams
- Repeated subsequences (LZ77 algorithm)

No golden rules but some hints

# How to choose the fragment length

- Fragments too long



# How to choose the fragment length

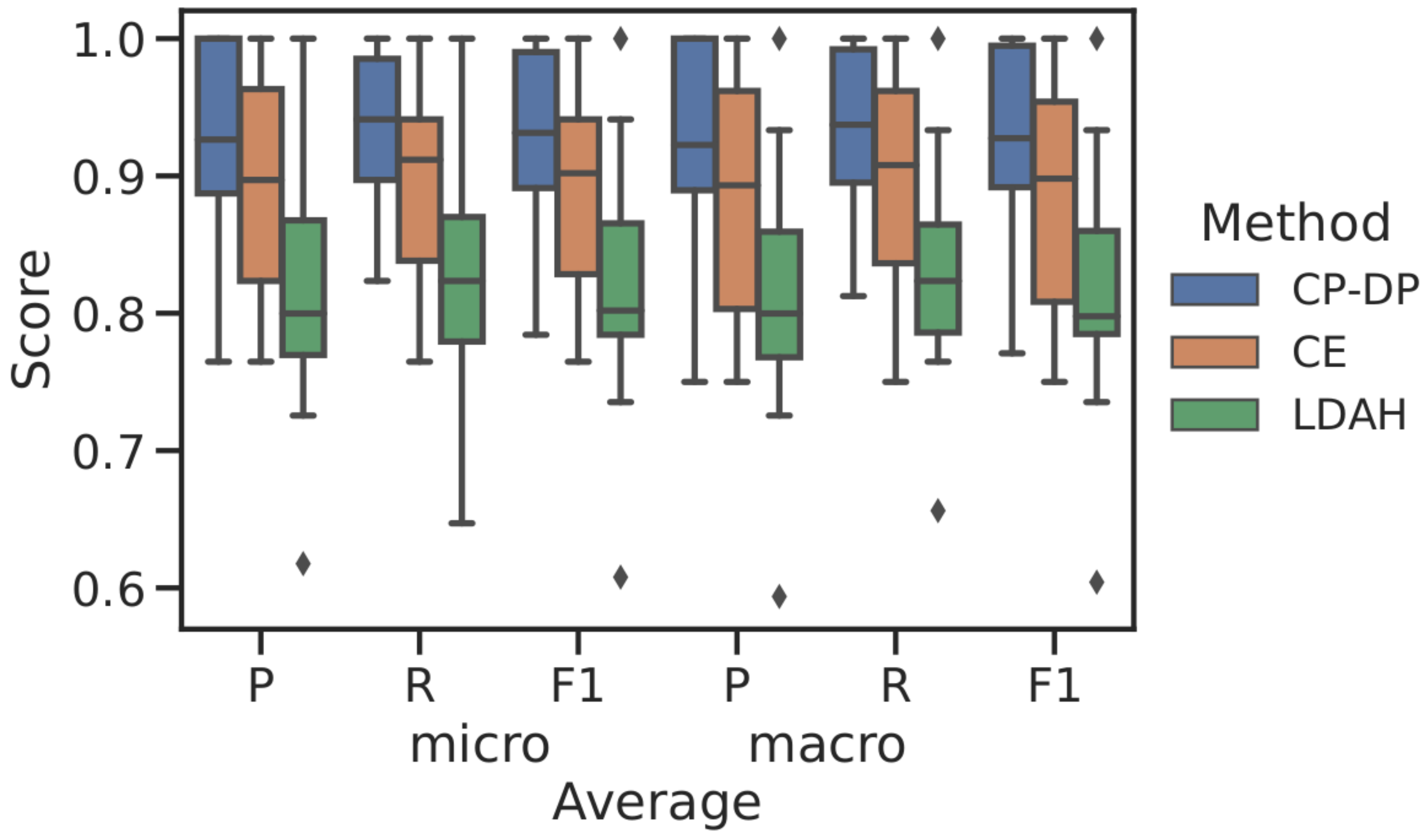
- Fragments too long
- Fragments too short
  - almost Kullback-Leibler Divergence (single token limit)
  - no opportunity to adapt

$$D_{KL}(f \parallel \mathcal{A}) = \sum_{j=1}^{k'+k} \nu_j \log_2 \frac{\nu_j}{\tilde{\nu}'_j} - G(\mathcal{A})$$

$$\tilde{\nu}'_j = \begin{cases} \nu'_j - \frac{\alpha}{n'}, & y_j \in A \\ \frac{(\theta_A + \alpha_A k') P_0(y_j)}{n'}, & y_j \notin A \end{cases}$$

# Now attribute!

- 171 Italian novels, 39 authors: 93.5%



# And attribute shorter texts!

- We are interested <u>only</u> in academic inquiries, we can try with Latin poetry
  - from 6 lines, up to two pages per poem

| | Lygdamo | Ovidio | Properzio | Tibullo |
|---|---|---|---|---|
| Poems | 6 | 41 | 92 | 16 |
| Bytes | 11.7 | 97.8 | 165.9 | 51.8 |

| Corpus | Book | Author |
|---|---|---|
| Tib-Pro | 63,77% | 100,00% |
| Tib-Pro-Ovi | 66,44% | 98,66% |
| Tib-Pro-Ovi-Lig | 67,74% | 98,71% |

# Does it work with informal texts?

- Enron corpus: 72 authors, 9337 emails

| Method | Attribution | Notes |
|---|---|---|
| Kourtis 2011 | 0.658 | SVM + supporting classifier |
| Seroussi 2014 DADT-P | 0.594 | Infer every author and document |
| **CP-DP 2022** | 0.556 | |
| Yang 2017 TDM | 0.542 | Tracks author evolution across documents |
| Seroussi 2012 LDAH | 0.426 | Lots of inference but conceptually simple |

# Does it work with many authors?

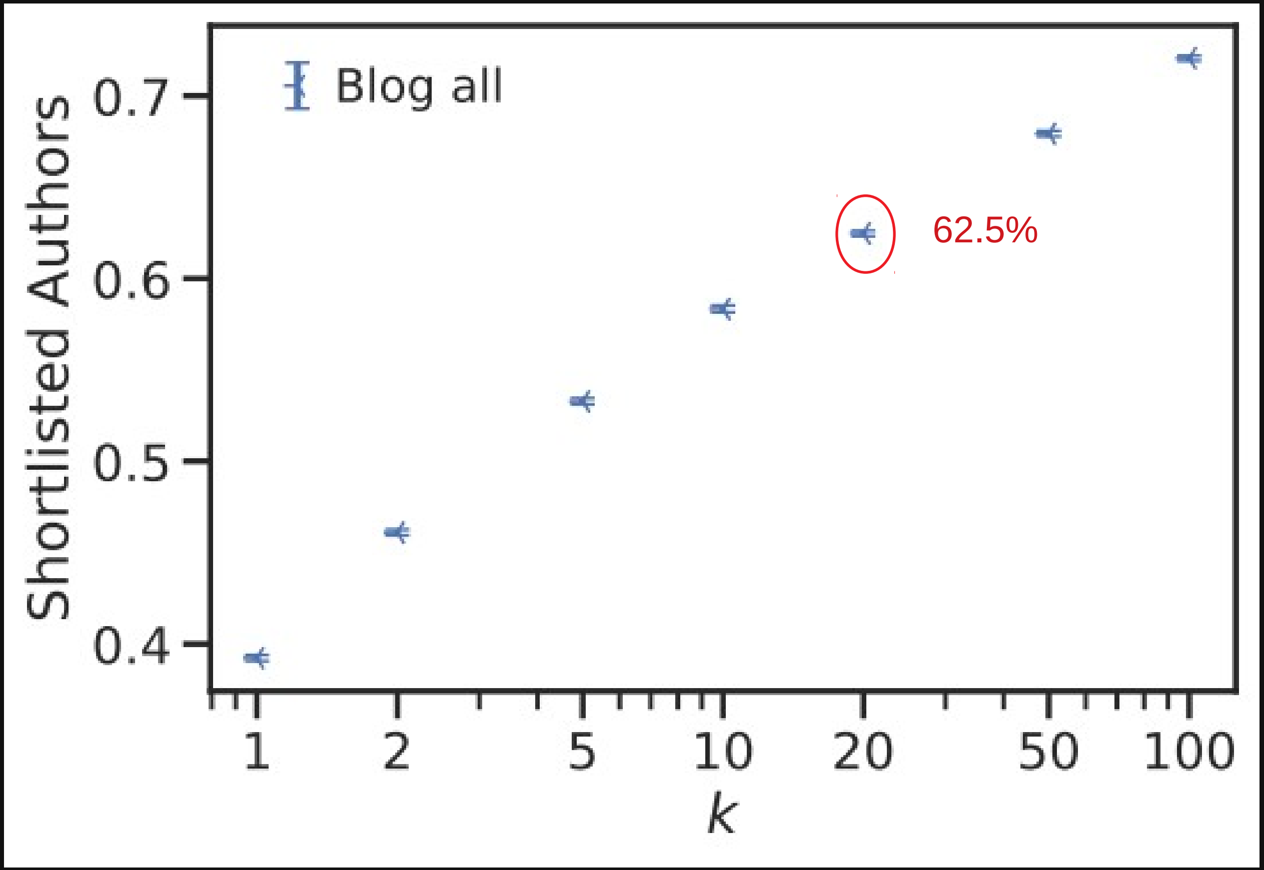
- Blog corpus: 19,320 authors, 678,161 posts
  - for ~40% of the authors, less than 3 pages in total

| Method | Prolific 1000 | All authors |
|---|---|---|
| **CP-DP** | **0.495** | **0.375** |
| Yang 2017 TDM | – | 0.308 |
| Seroussi 2014 DADT-P | 0.437 | 0.286 |
| Seroussi 2012 LDAH | 0.216 | 0.079 |

# Time to turn evil

- No assumption on the topic
  - “*** *****” is enough, for us every post with at least 50 characters (one sentence) is relevant
- We don’t need to find the author, it’s enough to have they in a shortlist.
  - Then we may:
    - Call NSO, buy Pegasus
    - Torture all those in the shortlist

# Time to turn evil – 2



# Obfuscation

- Effective (almost) only against the attribution method they are built for
- Extremely easy (~90%) to detect obfuscation
- Loose effectiveness if the attacker reduces the set of candidates
- Hard to use (semi-automated versions)
- Not preserving “semantics” (automated versions)
  - (way) less than 60% of the time
  - taking back the meaning reduces effectiveness

# Conclusions

- A simple model with a few parameter can go a long way in Authorship Attribution
- Concealing your IP address is clearly not enough
- How can your research (or the technologies developed to make it possible) be used for evil?
  - (maybe in 50 years from now)