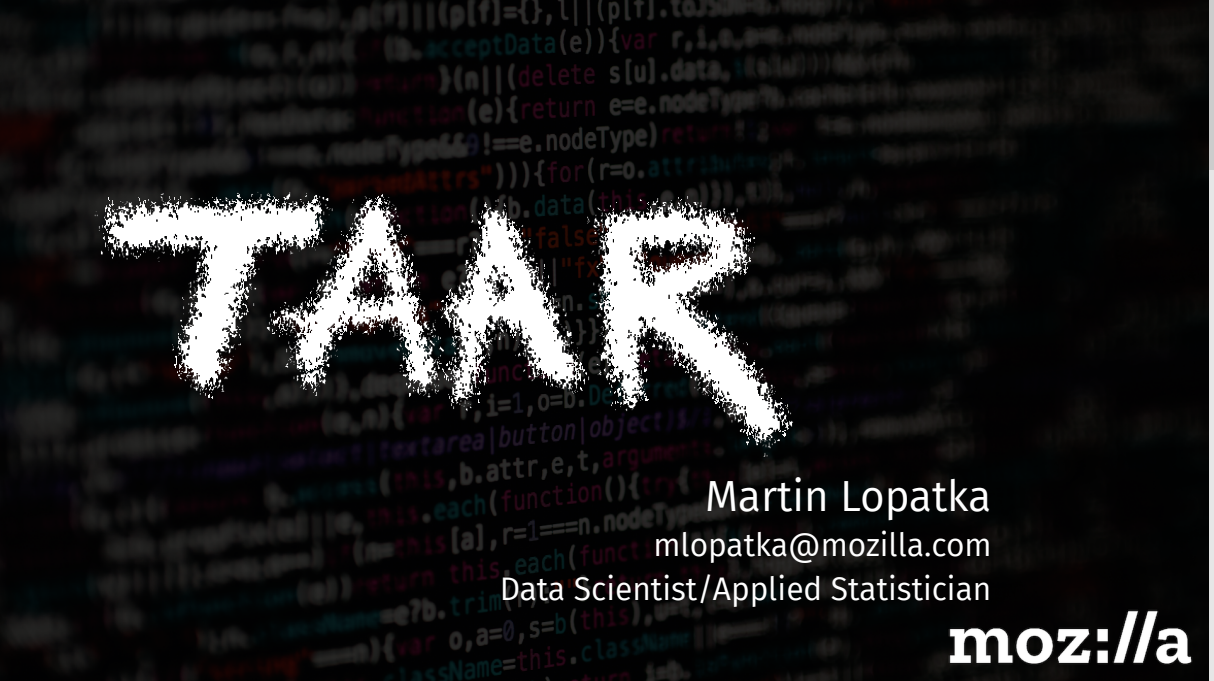
2019-06-10

- Personalised web browsing experience is hard
- Especially with a rigorous and respectful privacy policy
- Ultimately, many of the approaches in UMAP strive to find innovative ways to extract a meaningful signal from very noisy data.



TAAR
Martin Lopatka
mlopatka@mozilla.com
Data Scientist/Applied Statistician

- Martin Lopatka
- Time we have through… Mozilla's approach to recommending browser extensions
- T.A.A.R.
- Curiosity vs. builders -> move **spatial anchor**
- given the time, focus on a very brief overview, and two specific design choices
- Privacy by design and CLLR

Telemetry

Firefox Telemetry (optionally) measures and collects non-personal, performance and usage information.[1]

[1] https://wiki.mozilla.org/Telemetry

moz://a

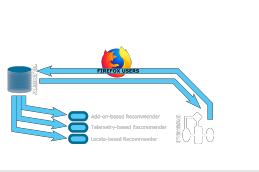
└Telemetry

Firefox Telemetry (optionally) measures and collects non-personal, performance and usage information.

- application localization identifier: (ch-de, br-pt)
- operating system
- subsession length
- bookmark count
- open tab count
- unique TLDs
- add-ons installed

Telemetry-Aware Add-on Recommender

moz://a



- Full system Spec
- Three modules each leveraging different subsets of client information based on availability.
- Individual recommendations combined via linear stacked ensemble
- These are domain specific and specific to our telemetry infra, so lets treat them like black boxes
- more interesting is the comparison of functions for determining individual weighting of the recommendations.

Differential Privacy

- Differentially private release mechanism for frequencies reports an approximate answer to an ***item:count*** distribution.
- Noise must be chosen to preserve the usefulness of the provided answer while protecting the privacy of the more rare counts

Introduction to DP: https://robertovitillo.com/2016/07/29/differential-privacy-for-dummies/

└Differential Privacy

- Formalizes the idea that released data set can not be used to infer whether any one person is present
- Typically generated on the basis of a known distribution and some known noise distribution.
- We adapt this technique to generate add-on installation frequency tables for each locale according to the following procedure
- Guards against Overfitting

Log Likelihood Ratio Cost (cLLR)

```
'zh-CN': [('guid_01', 0.75),...,('guid_02', 0.05)],
'fr-FR': [('guid_03', 0.24),...,('guid_04', 0.01)],
...,
'en-US': [('guid_04', 0.18),...,(guid_05', 0.02)]
```

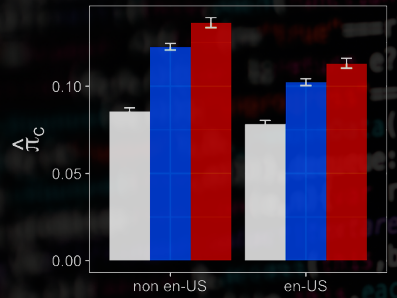
include boxplots vs other metrics.

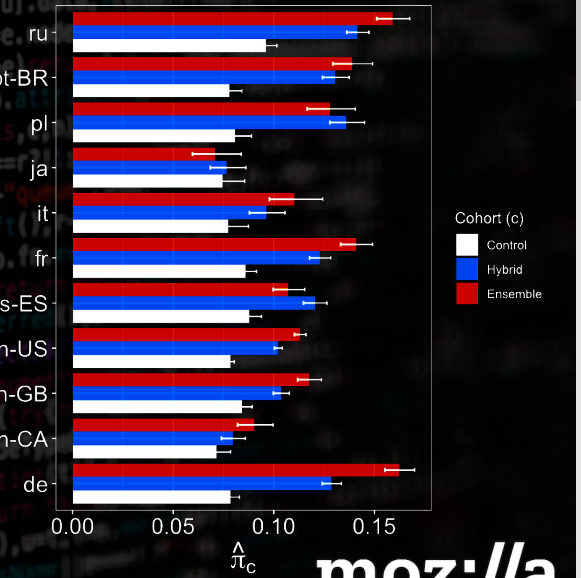
└Log Likelihood Ratio Cost (cLLR)

'zh-CN': [('guid_01', 0.75),...,('guid_02', 0.05)],
'fr-FR': [('guid_03', 0.24),...,('guid_04', 0.01)],
...,
'en-US': [('guid_04', 0.18),...,(guid_05', 0.02)]
include boxplots vs other metrics.

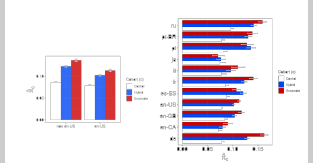
- better usage of full signal if component modules a probabilistic (flavoured)
- Symmetry also accounts for incorrect recommendations (instead of just relative rank fo correct recommendations and the relevance score)
- Versus Discounted Cumulative Gain (DCG), and versus Mean Average Precision (MAP) for including in the recommendation list at all

Performance





moz://a



- variable availability of our data, not only in terms of quantity but in terms of fields
- offer users granular choice in Telemetry
- Performs well and scales with data availability
- And 100% open source
- TAAR Served over 450K recommendations to 348 900 unique clients

Acknowledgements

Victor Ng
Ben Miroglio
David Zeber
Alessio Placitelli
Laura Thomson

Fredrik Wollsén
Jason Thomas
Stuart Colville
Shell Escalante
Scott Devaney

AMO team
InfoSec
QA team
Florian Hartmann
Roberto Vitillo

moz://a

└Acknowledgements

Victor Ng
Ben Miroglio
David Zeber
Alessio Placitelli
Laura Thomson

Fredrik Wollsén
Jason Thomas
Stuart Colville
Shell Escalante
Scott Devaney

AMO team
InfoSec
QA team
Florian Hartmann
Roberto Vitillo

- Thank you all for choosing to come engage with me here
- I'll be happy... questions
- but first... acknowledgements