Vertex Nomination

improved fusion of content and context

Glen A. Coppersmith

Human Language Technology Center of Excellence

Johns Hopkins University

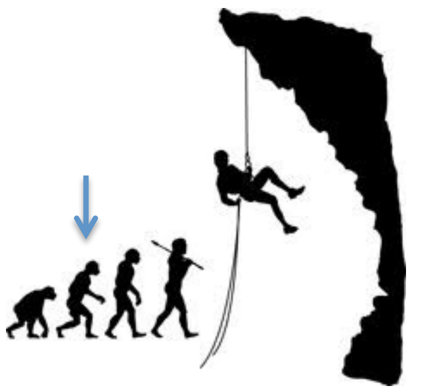
human language technology
center of excellence

JOHNS HOPKINS
UNIVERSITY

Presented at Interface Symposium: May 17, 2012

Vertex Nomination

improved fusion of content and context



Glen A. Coppersmith

Human Language Technology Center of Excellence

Johns Hopkins University

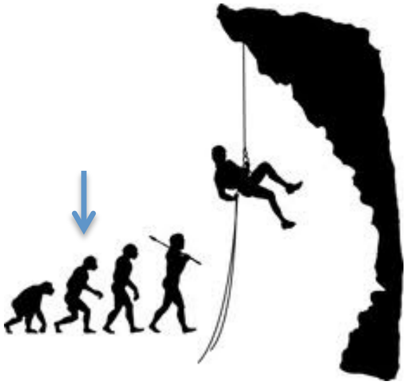
human language technology
center of excellence

JOHNS HOPKINS
UNIVERSITY

Presented at Interface Symposium: May 17, 2012

Vertex Nomination

improved fusion of content and context



Human Language Content

Communications Graph

Glen A. Coppersmith

Human Language Technology Center of Excellence

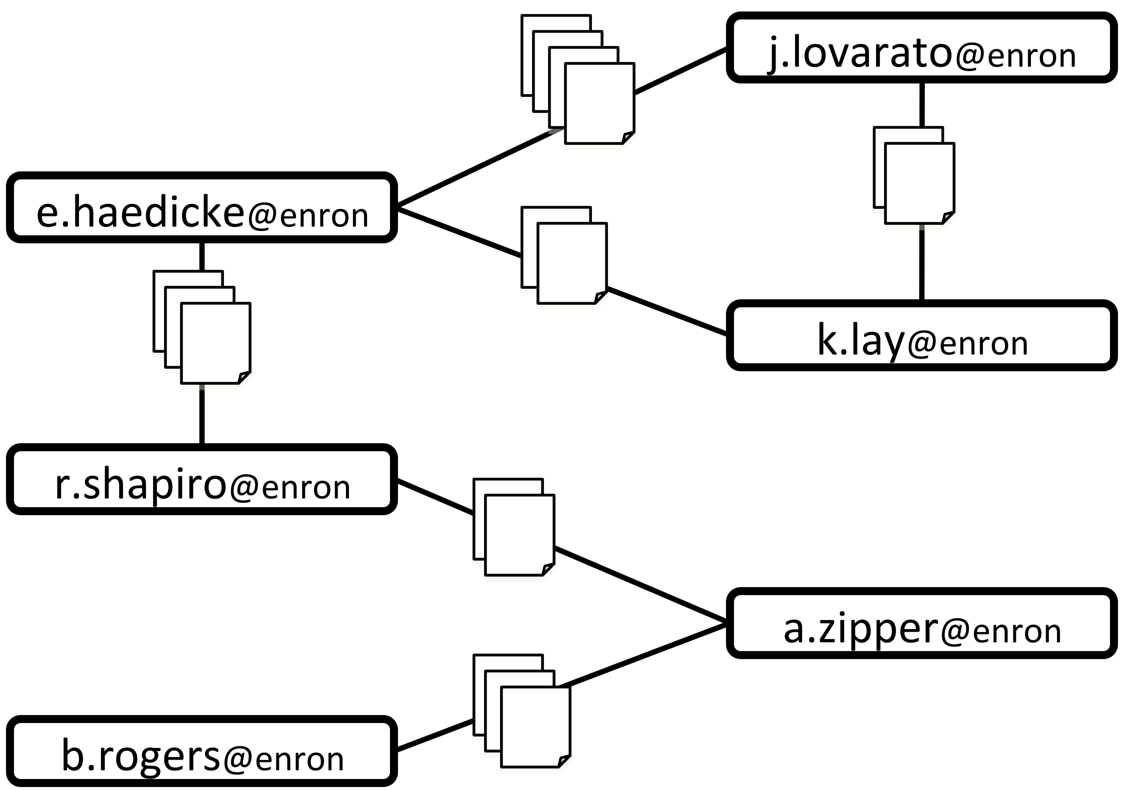
Johns Hopkins University

human language technology
center of excellence

Presented at Interface Symposium: May 17, 2012

JOHNS HOPKINS
UNIVERSITY

Our data: Enron Email Corpus



Motivation and Problem Statement

- We know the identities of a few fraudsters.
- We observe the content and the context of both fraudsters and non-fraudsters.
- We want to know the identities of more fraudsters.

- Inference Task
  - Nominate persons likely to be fraudsters.

Outline

- Introduction
- Method
  - Importance Sampling
  - Evaluation
- Analytics – Content and Context
- Fusions
- Conclusions & Future Directions

Outline

- Introduction
- Method
  - Importance Sampling
  - Evaluation
- Analytics – Content and Context
- Fusions
- Conclusions & Future Directions

Motivation and Problem Statement

- We know the identities of a few fraudsters.
- We observe the content and the context of both fraudsters and non-fraudsters.
- We want to know the identities of more fraudsters.

- Inference Task
  - Nominate persons likely to be fraudsters.

Motivation and Problem Statement

- We know the identities of a few fraudsters.
- [Graph attributed with human language]
- We want to know the identities of more fraudsters.

- Inference Task
  - Nominate persons likely to be fraudsters.

With loss of generality…

- Netflix [See Trevor's Keynote]
- Genomics [Half the talks I've seen at IF12]

- Noted similarities Recommender Systems
  - We focus on those with a graph
  - … and those with human language

With loss of generality…

- Netflix [See Trevor's Keynote]
- Genomics [Half the talks I've seen at IF12]

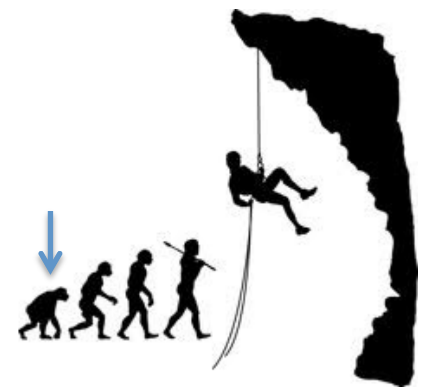
- Noted similarities Recommender Systems
  – We focus on those with a graph
  – … and those with human language

What's already been done?

- Theory
  - (Nam) Lee & Priebe 2011
  - (Dominic) Lee & Priebe (submitted 2012)
- Simulations and Experiments
  - Marchette, Priebe, and Coppersmith (2011)
  - Coppersmith & Priebe (submitted 2011)
- (Content + Context) > (Content | Context)
  - Human Language Technology and Graph Theory
  - Assumptions are valid for the Enron data

What's already been done?

- (Content + Context) > (Content | Context)
  - Human Language Technology and Graph Theory
  - Assumptions are valid for the Enron data
  - Importance Sampling provides sensible partitions



Assumptions

- The fraudsters talk to each other more than expected of a random pair of people.
- The fraudsters talk about different things than expected of a random pair of people.

Fusion of Disparate Information

- Some signal from the communications graph
- Some signal from the human language content

- How do you fuse them?
  - A principled fusion would be nice
  - A useful fusion is more important
    - Robust to real world problems
    - Scalable to real world applications

Questions for this talk

- How should we fuse?
  - (both performance and scalability are important)

- What kind of HLTs should we use?
  - (Do they do different things?)

Outline

- Introduction
- Method
  - Importance Sampling
  - Evaluation
- Analytics – Content and Context
- Fusions
- Conclusions & Future Directions

Mathematical Model: κ graph

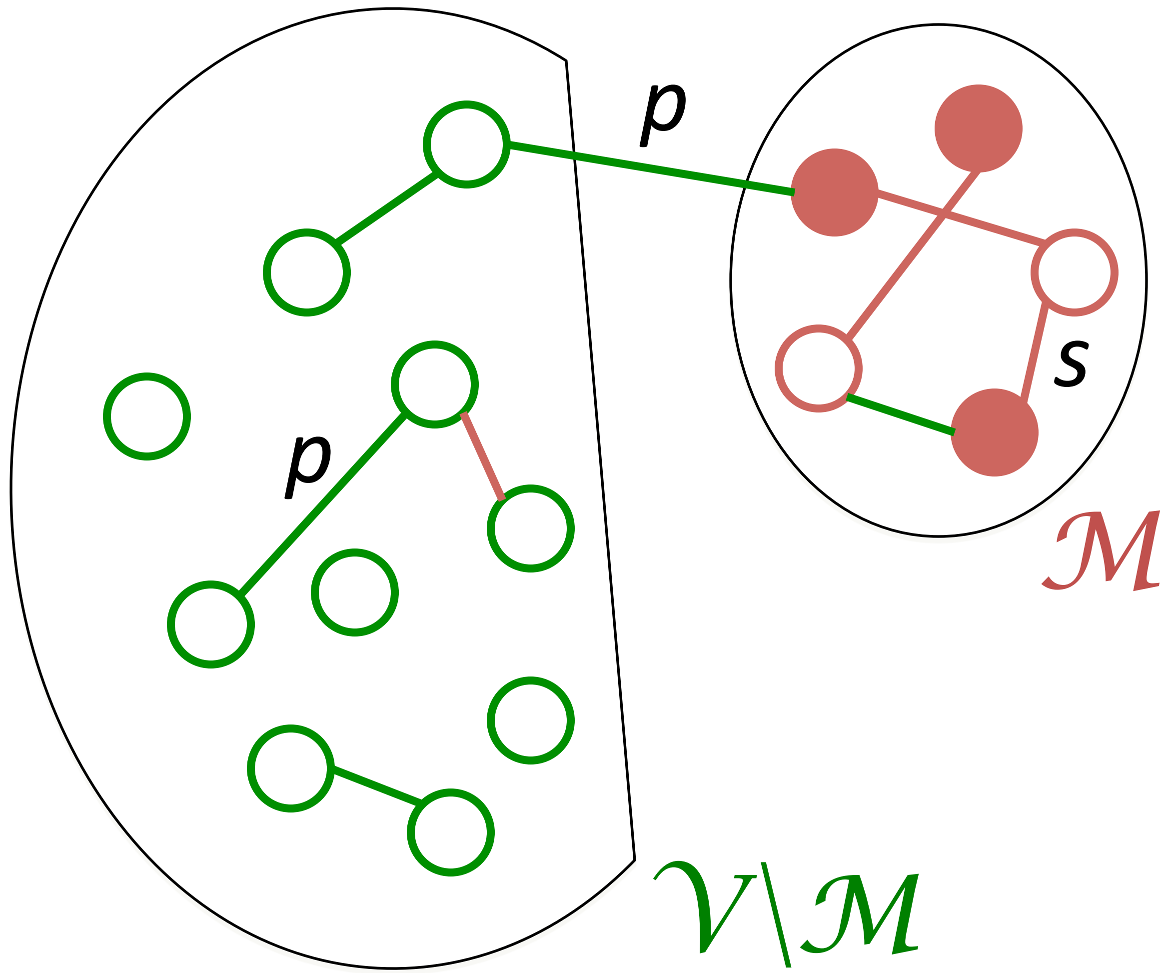
- ●○○ $|\mathcal{V}|$=n
- ○● $|\mathcal{M}|$=m
- ● $|\mathcal{M}'|$=m'
- ○ $|\mathcal{V}\backslash\mathcal{M}|$=n-m

$p = [p_0,p_1]$

$s = [s_0,s_1]$

$p_0 = s_0$

$p_1 < s_1$



$\kappa(n, p, m, m', s)$

human language technology center of excellence

JOHNS HOPKINS UNIVERSITY

Mathematical Model: κ graph

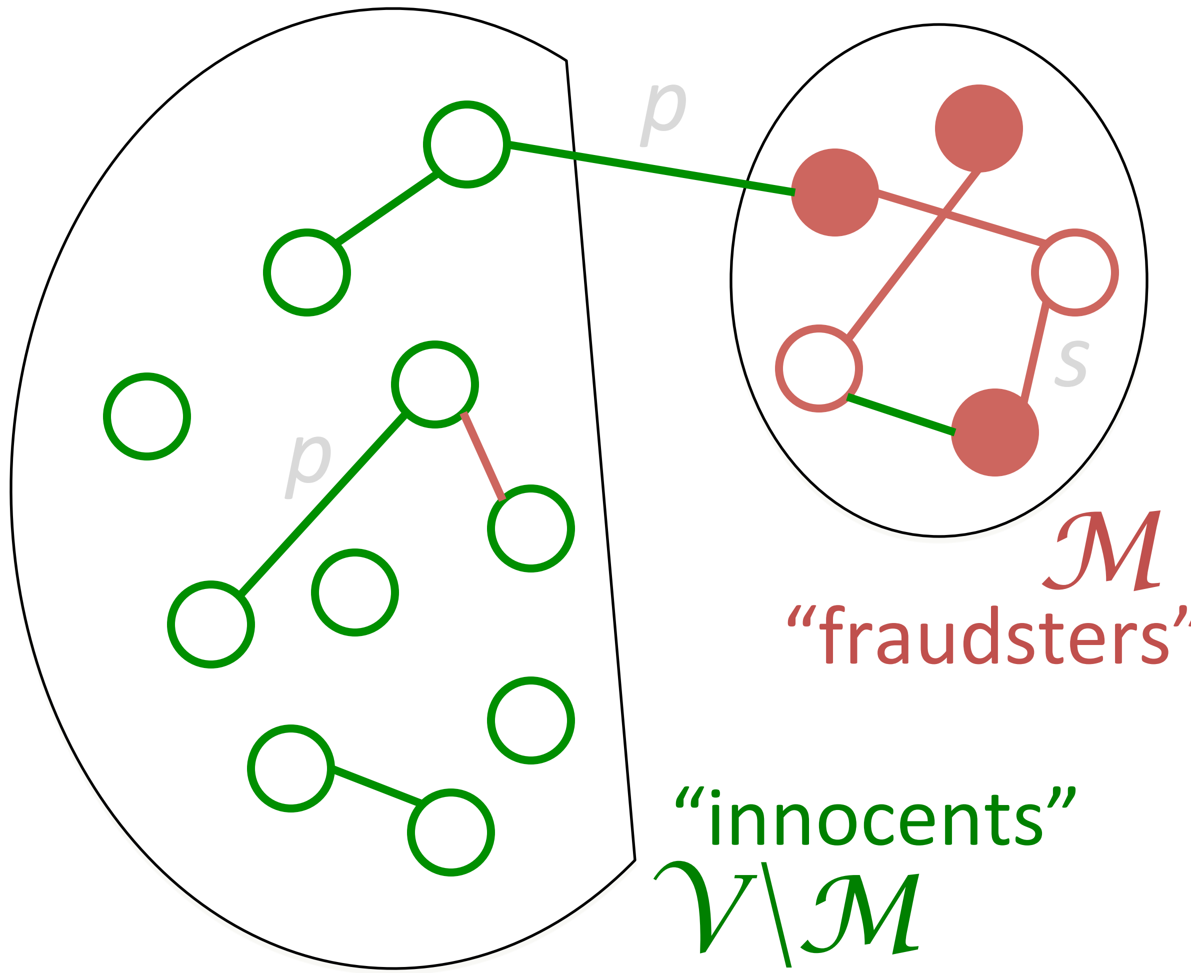
- $|\mathcal{V}|$=n
- $|\mathcal{M}|$=m
- $|\mathcal{M}'|$=m'
- $|\mathcal{V}\backslash\mathcal{M}|$=n-m

$p = [p_0, p_1]$

$s = [s_0, s_1]$

$p_0 = s_0$

$p_1 < s_1$



$\kappa(n, p, m, m', s)$

human language technology
center of excellence

JOHNS HOPKINS
UNIVERSITY

Assumptions

- The fraudsters talk to each other more than expected of a random pair of people.
  - $\mathcal{M}$ is more dense than $\mathcal{V} \backslash \mathcal{M}$
- The fraudsters talk about different things than expected of a random pair of people.
  - $p[p_0,p_1]$ and $s[s_0,s_1]$
  - $p_0 = s_0$, $p_1 < s_1$

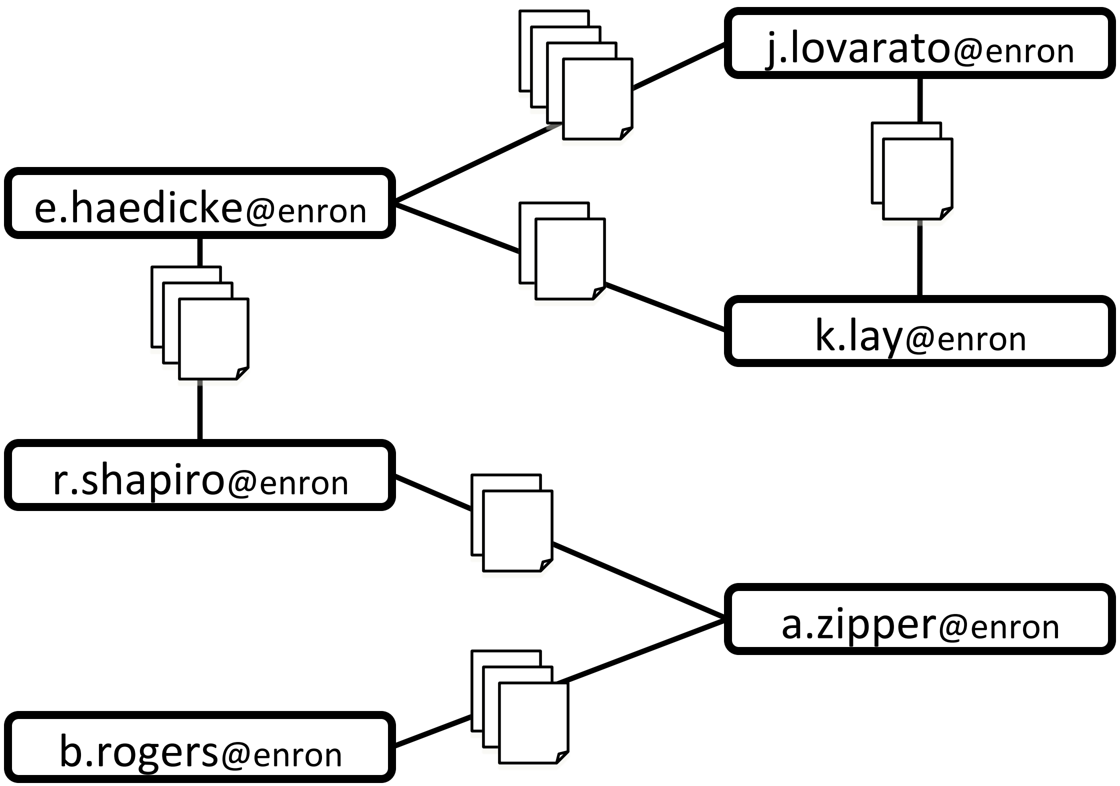
$\mathcal{V} \backslash \mathcal{M}$ ○

$\mathcal{M}$ ○ ●

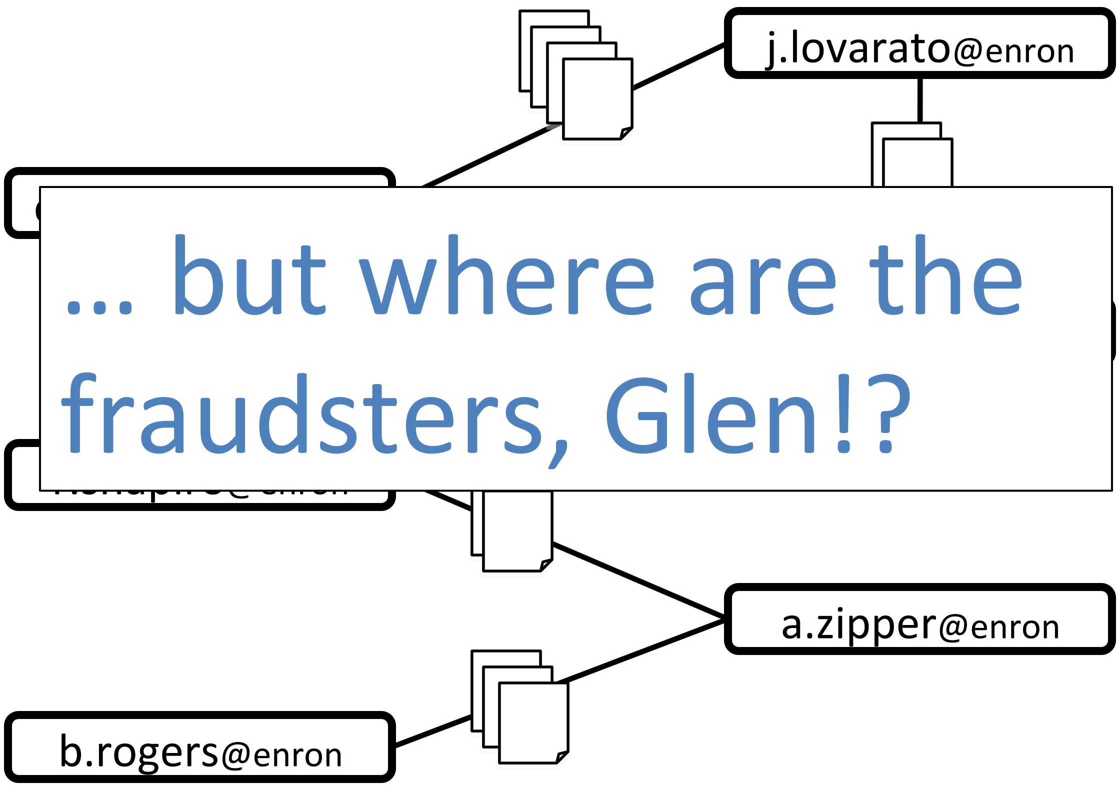
human language technology
center of excellence

JOHNS HOPKINS
UNIVERSITY

Our data: Enron Email Corpus



Our data: Enron Email Corpus



… but where are the fraudsters, Glen!?

Importance Sampling Procedure

- Randomly partition Enron into $\mathcal{M}$ and $\mathcal{V}\setminus\mathcal{M}$
- Question assumptions
  - Density $\mathcal{M}$ > Density $\mathcal{V}\setminus\mathcal{M}$
  - Topic Distribution $\mathcal{M}$ ≠ Topic Distribution $\mathcal{V}\setminus\mathcal{M}$
- Discard partitions that violate assumptions.

- Collect 5000 partitions.

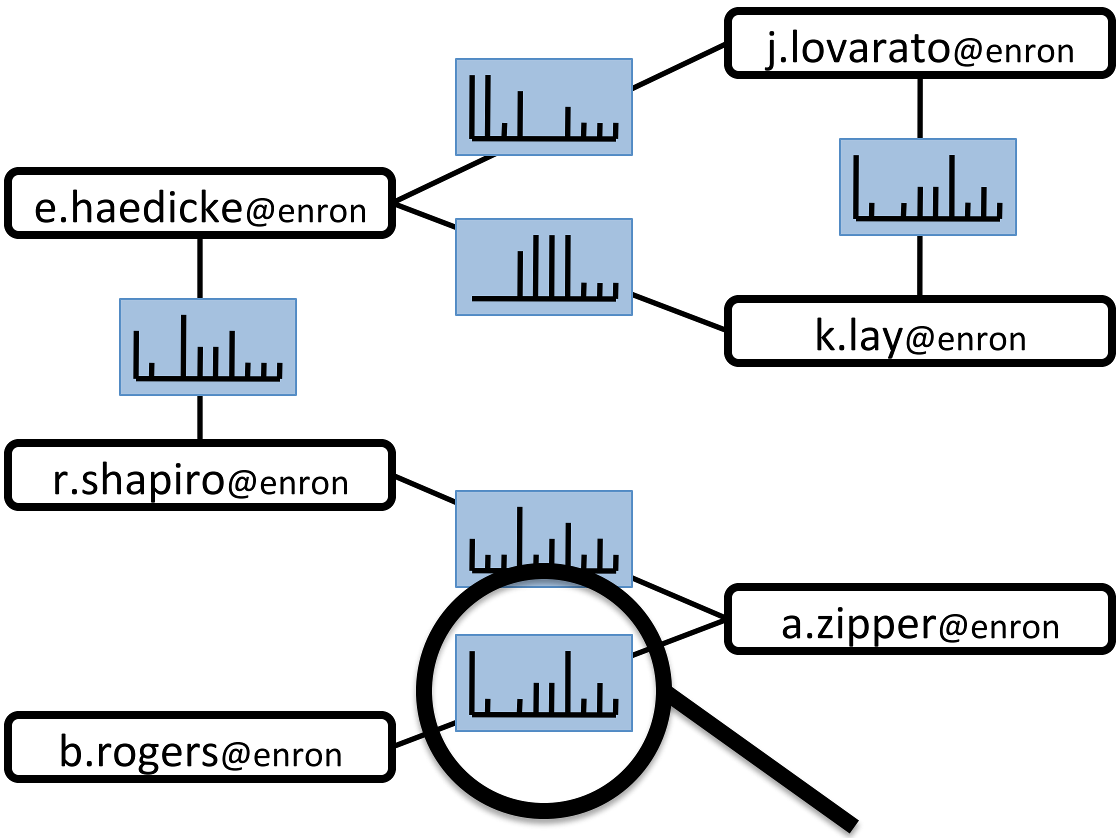
$\mathcal{V}\setminus\mathcal{M}$

$\mathcal{M}$

human language technology
center of excellence

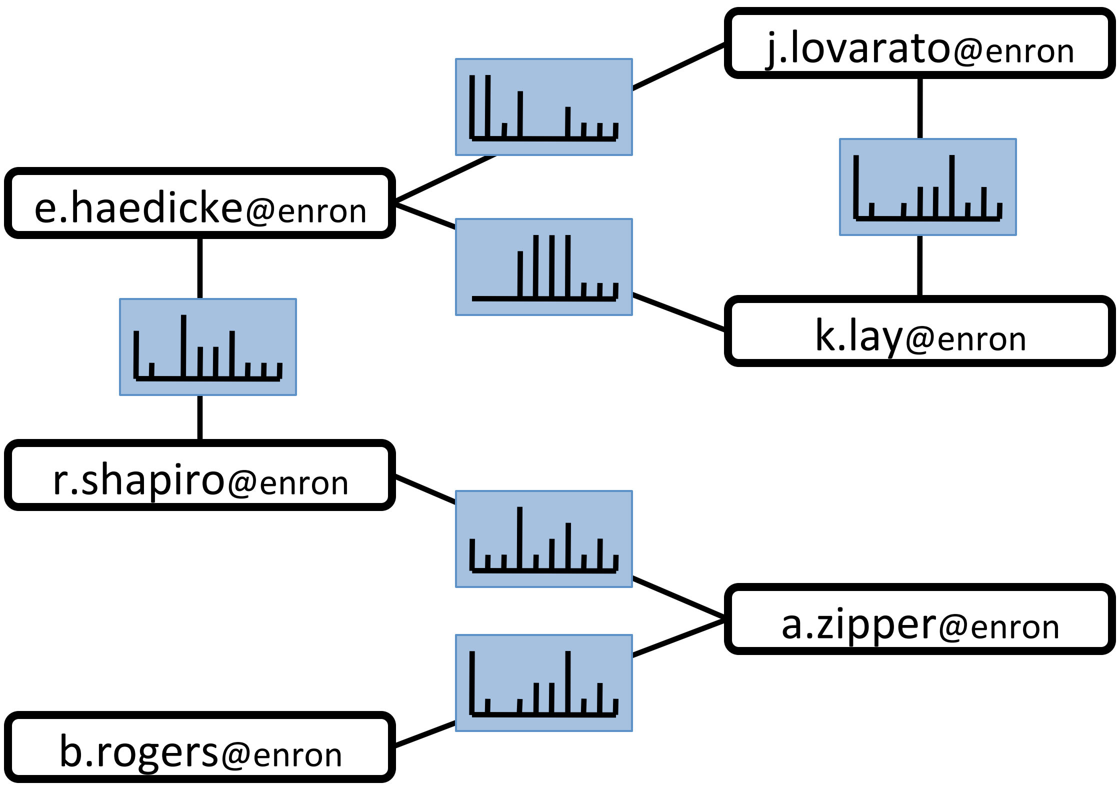
JOHNS HOPKINS
UNIVERSITY

Importance Sampling



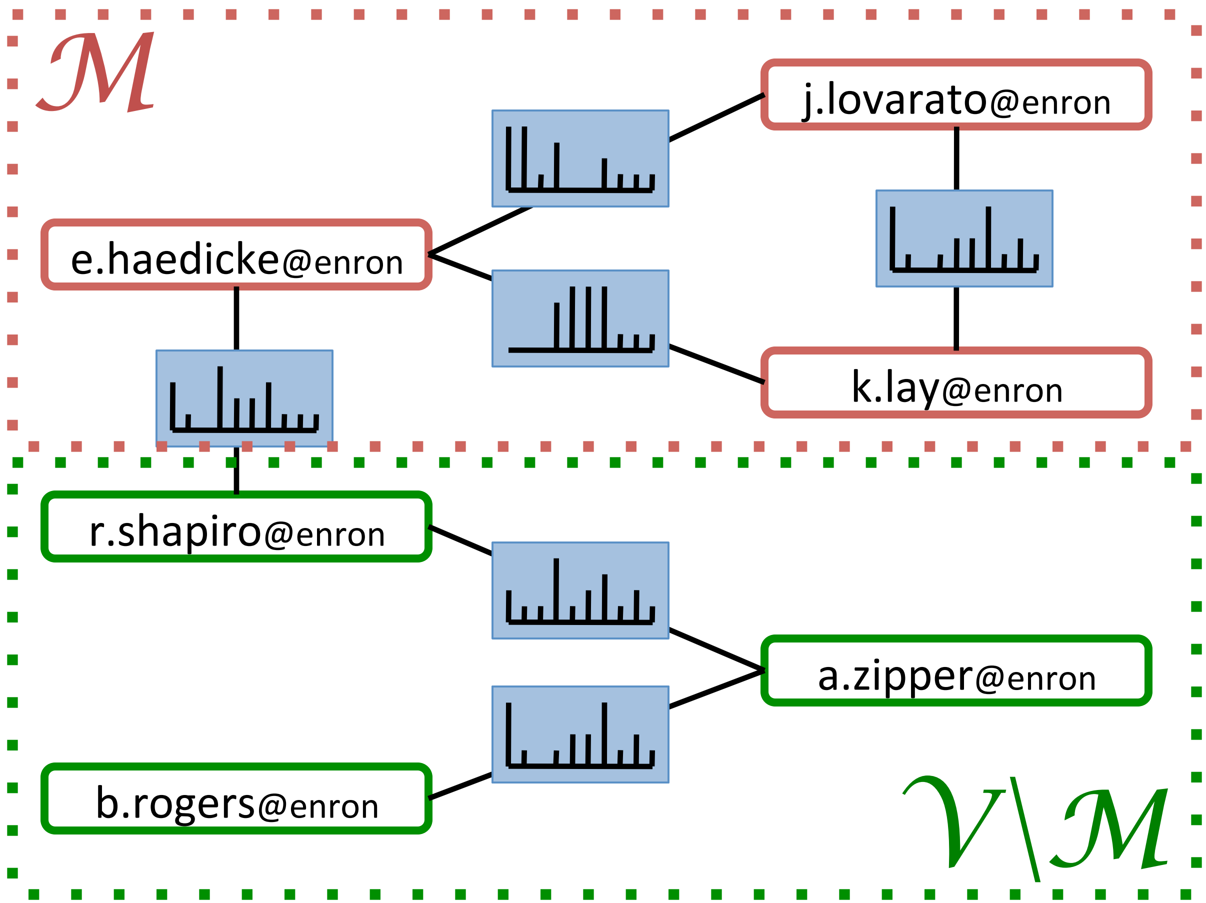
Hand-labels provided by Michael Berry, 2004

Importance Sampling



Hand-labels provided by Michael Berry, 2004

Importance Sampling



Testing Partitions: Density

$$\rho(\mathcal{M}) = \frac{|\text{observed edges in } \mathcal{M}|}{|\text{possible edges in } \mathcal{M}|}$$

$$\Delta\rho = \rho(\mathcal{M}) - \rho(\mathcal{V} \setminus \mathcal{M})$$

$\mathcal{V} \setminus \mathcal{M}$ ○

$\mathcal{M}$ ○●

"The fraudsters talk to each other more than expected of a random pair of people."

Testing Partitions: Topic Distribution

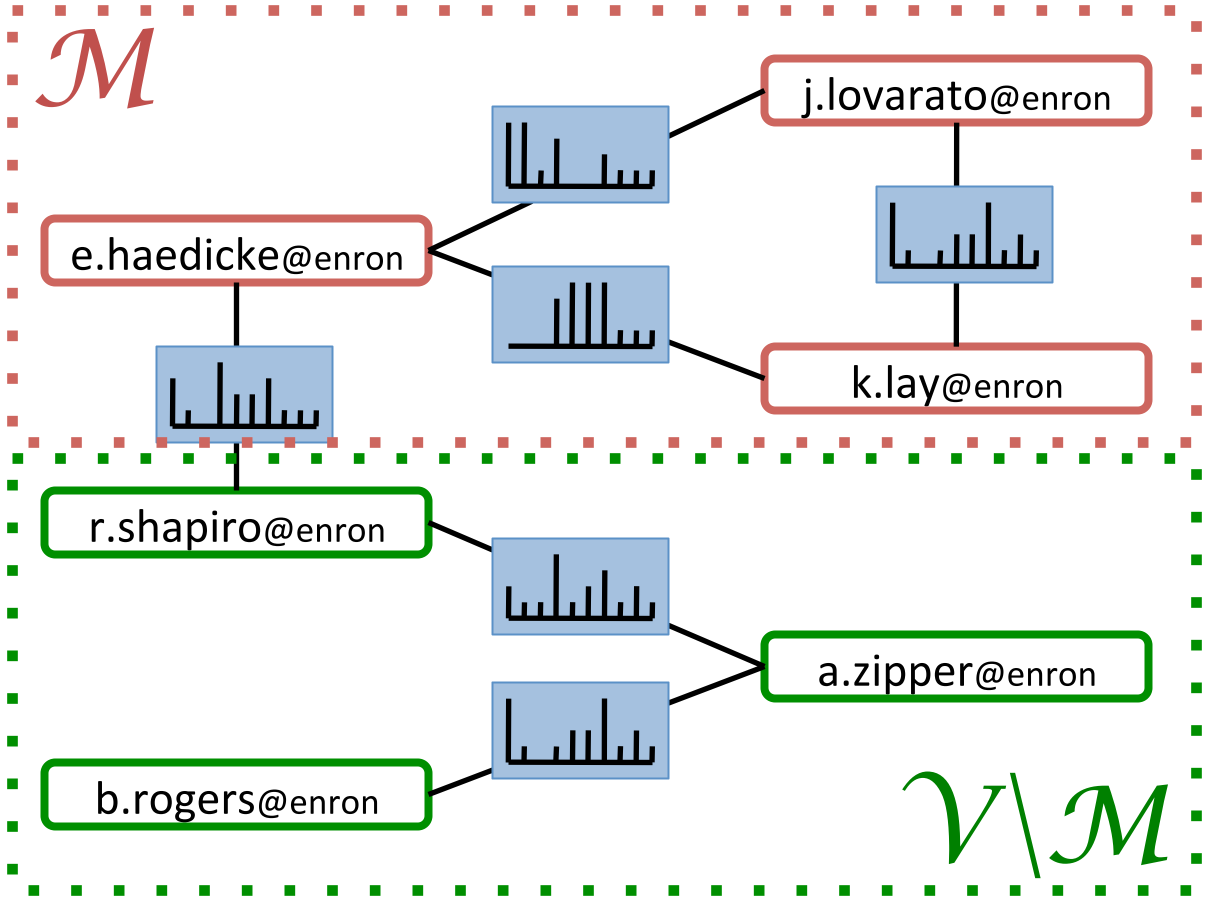
	California Power	Events of 9/11	Pro Football	Weather	…	Energy Legislation
Topic($\mathcal{M}$) =	.2	.1	0	0	…	.25
Topic($\mathcal{V}\backslash\mathcal{M}$) =	.1	.1	.3	.15	…	.1
Δ Topic =	|.1|	|0|	|-.3|	|-.15|	…	|.15|

$\mathcal{V}\backslash\mathcal{M}$ ○

$\mathcal{M}$ ○●

"The fraudsters talk about different things than expected of a random pair of people."

Importance Sampling



human language technology
center of excellence

JOHNS HOPKINS
UNIVERSITY

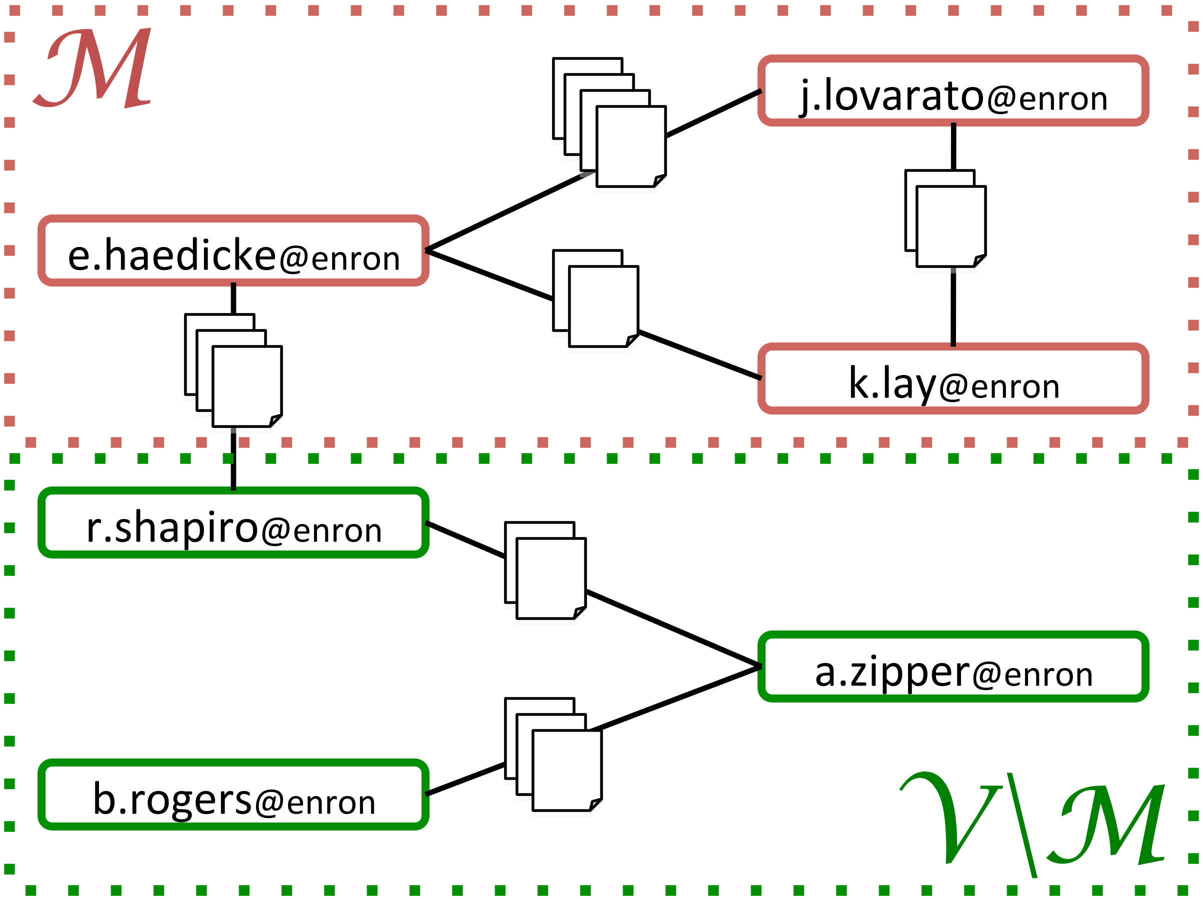
Method

- $|\mathcal{M}|=10$ $|\mathcal{M}'|=5$ $|\mathcal{V}\backslash\mathcal{M}'|=179$

- For each vertex (v) in $\mathcal{V}\backslash\mathcal{M}'$ we calculate each analytic.
- Rank vertices according to each analytic or fusion.
- Evaluate quality of ranked lists.

$\mathcal{M}'$
$\mathcal{M}$
$\mathcal{V}\backslash\mathcal{M}'$

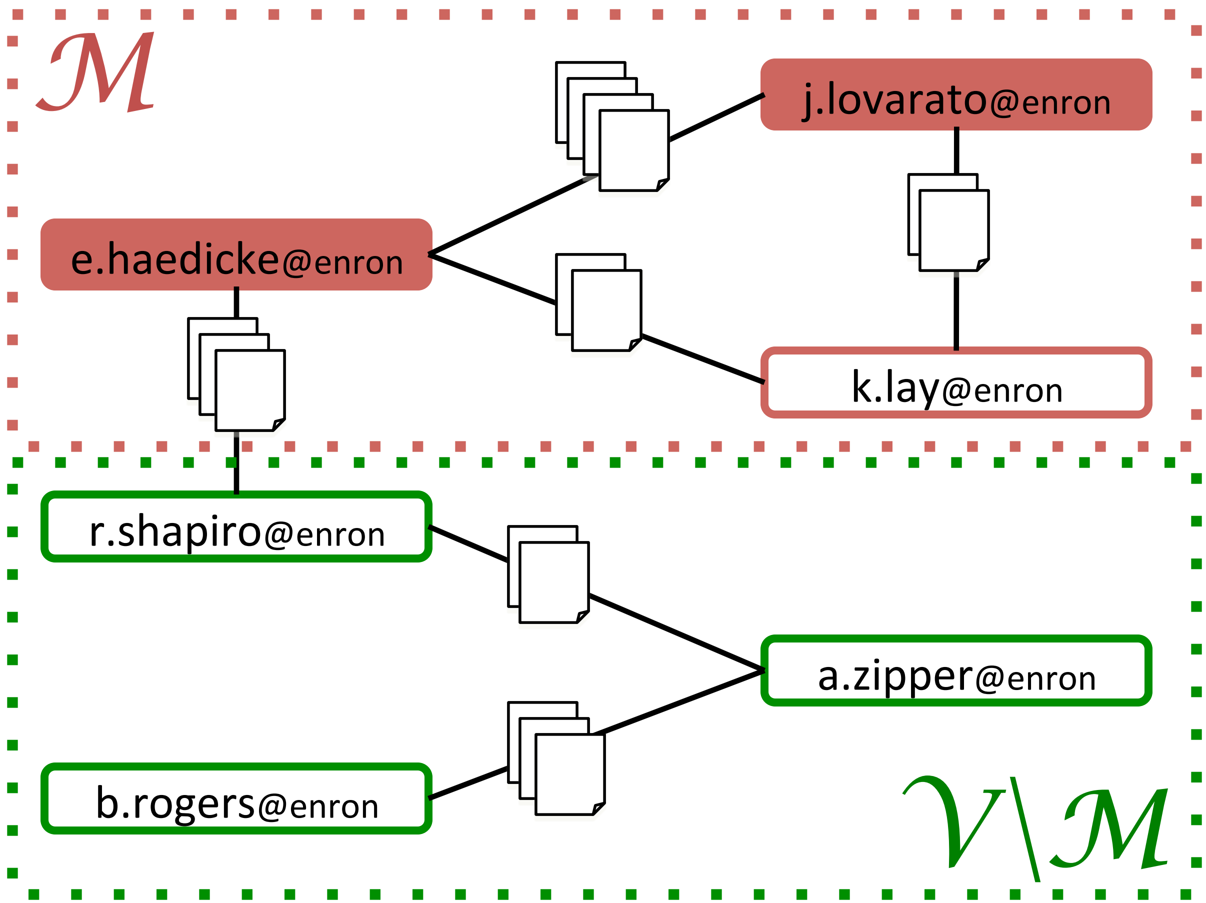
One experiment



human language technology
center of excellence

JOHNS HOPKINS
UNIVERSITY

One experiment



human language technology
center of excellence

JOHNS HOPKINS
UNIVERSITY

Evaluation

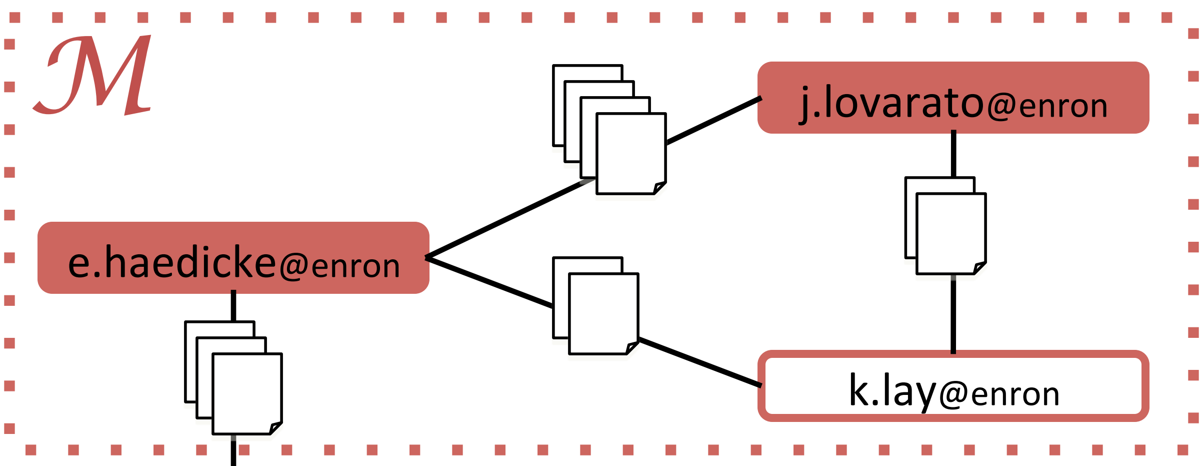
- Ranked lists evaluated by standard Information Retrieval measures
- What is our inference task?
  - Need to find all of them – Mean Average Precision (MAP)
  - Need to find one more of them – Mean Reciprocal Rank (MRR)
  - Can only examine *k* vertices (p@*k*), $k = \{5,10\}$

Outline

- Introduction
- Method
  - Importance Sampling
  - Evaluation
- Analytics – Content and Context
- Fusions
- Conclusions & Future Directions

Context Analytic

- *The fraudsters talk to each other more than expected of a random pair of people.*
- Number of known fraudsters in 1-hop neighborhood of candidate vertex.

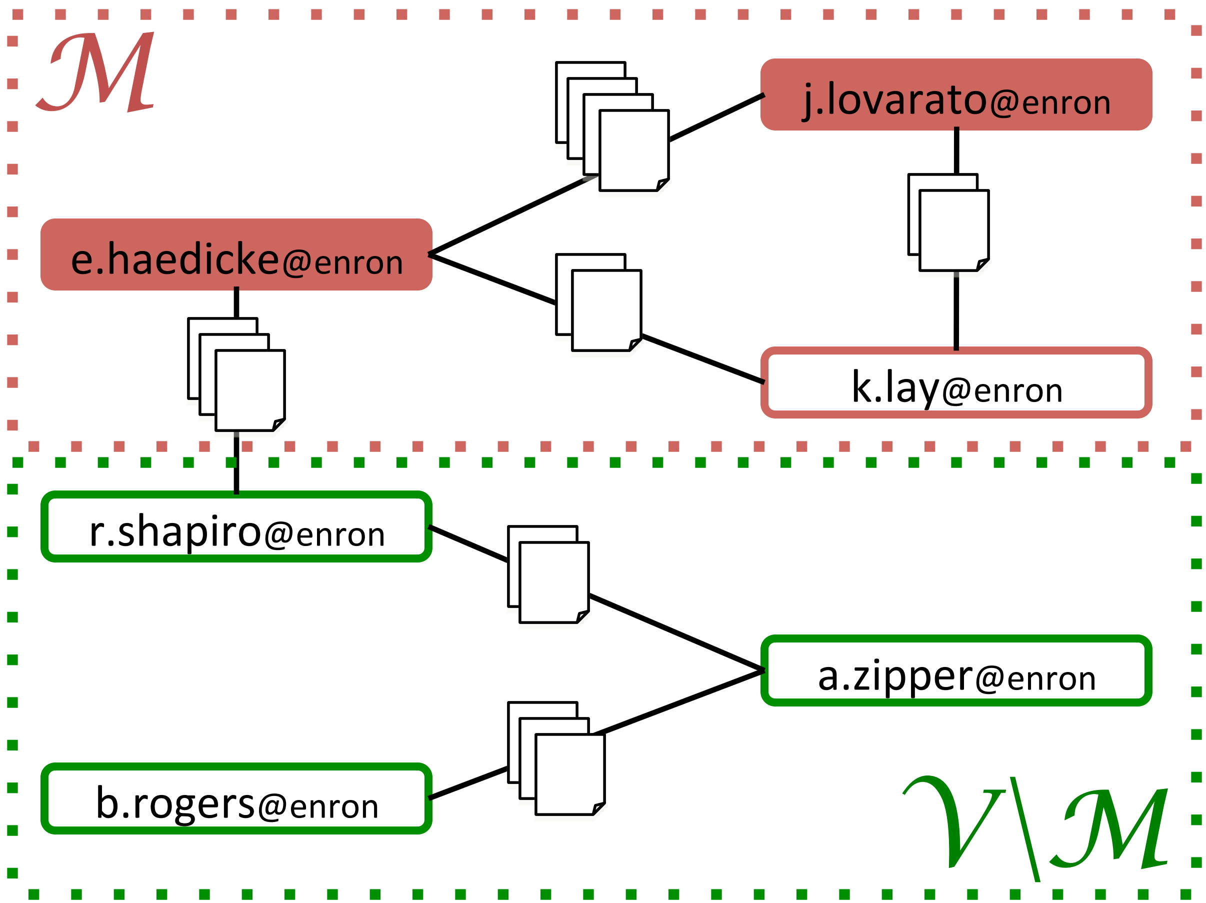


Content Analytics (HLTs)

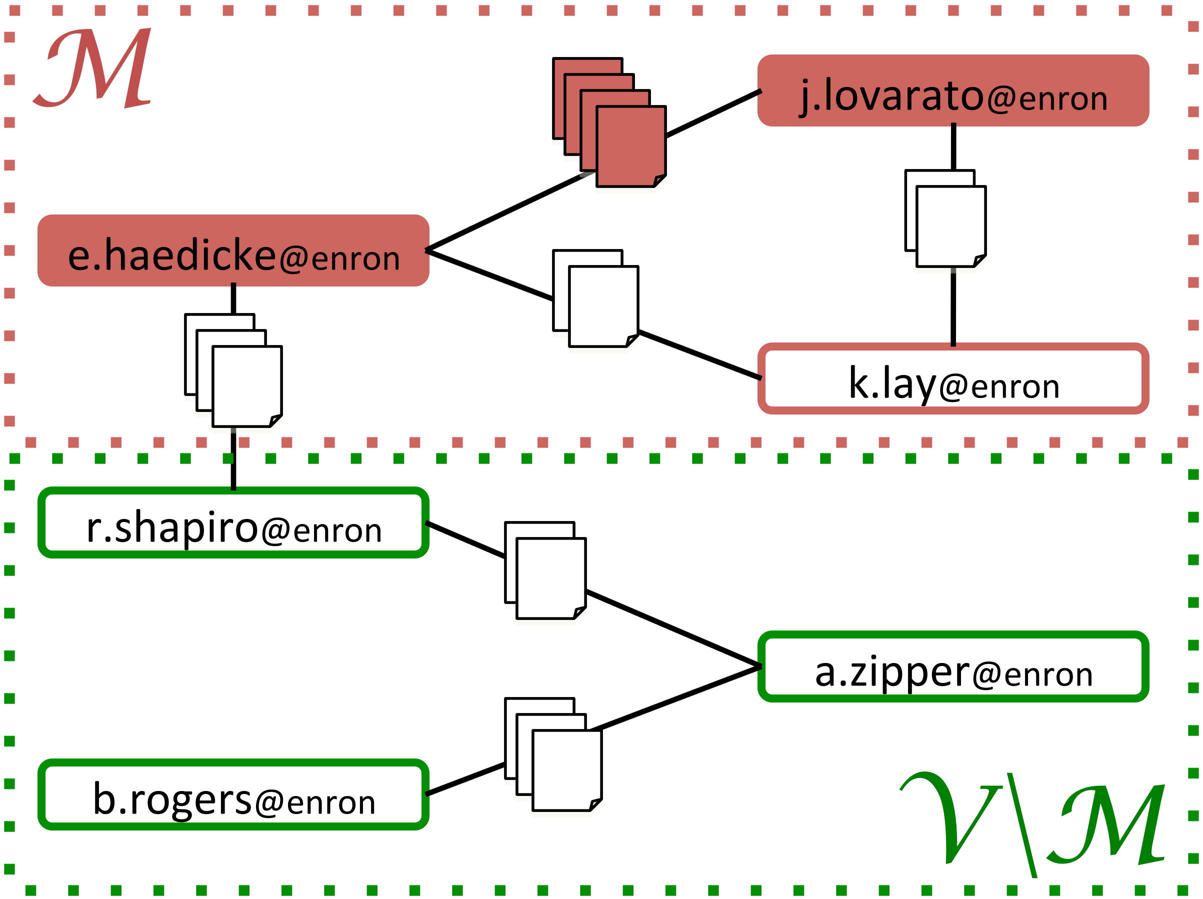
- *The fraudsters talk about different things than expected of a random pair of people.*

- How ‘similar’ is the content of each candidate vertex to the known fraudsters?

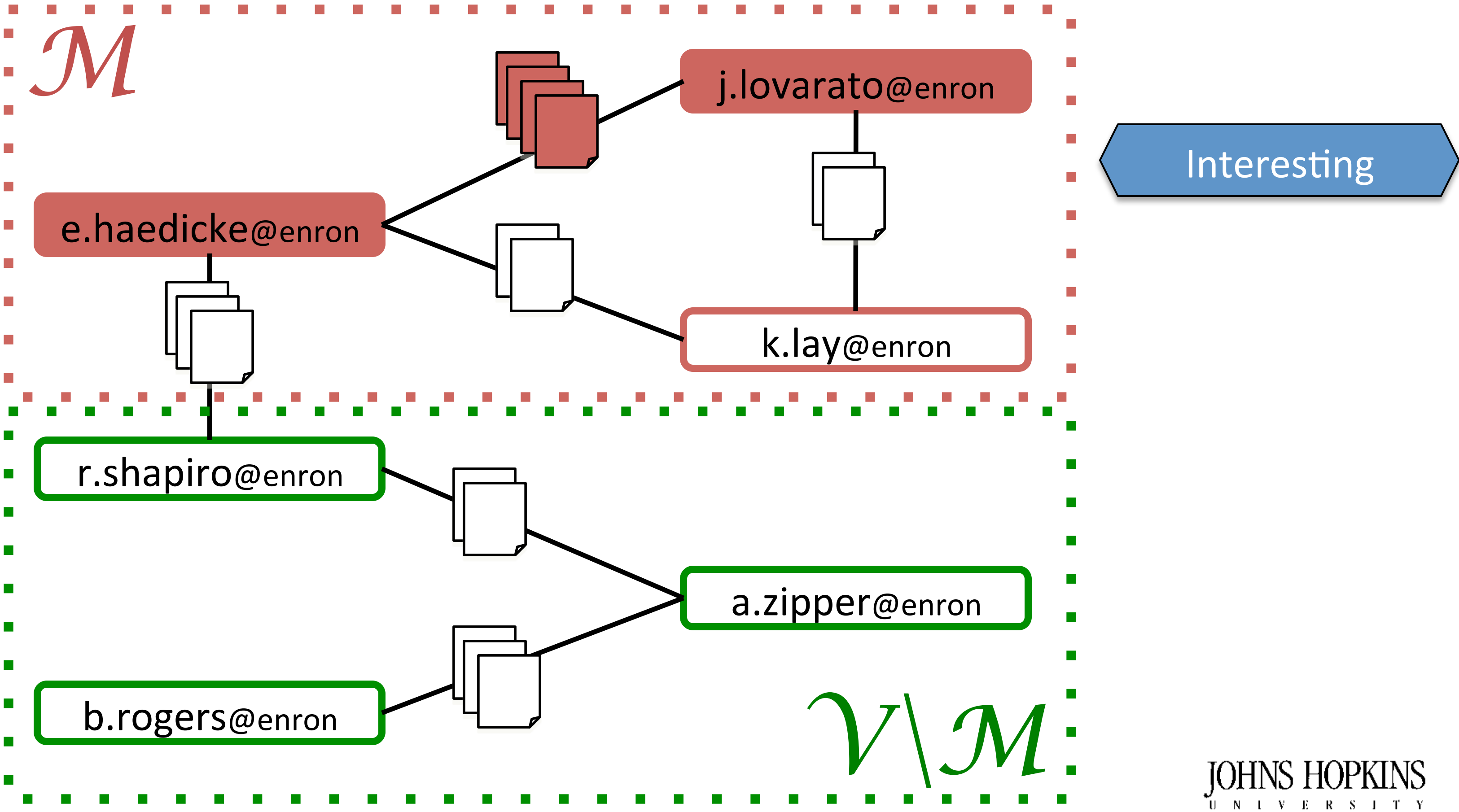
Training HLTs



Training HLTs



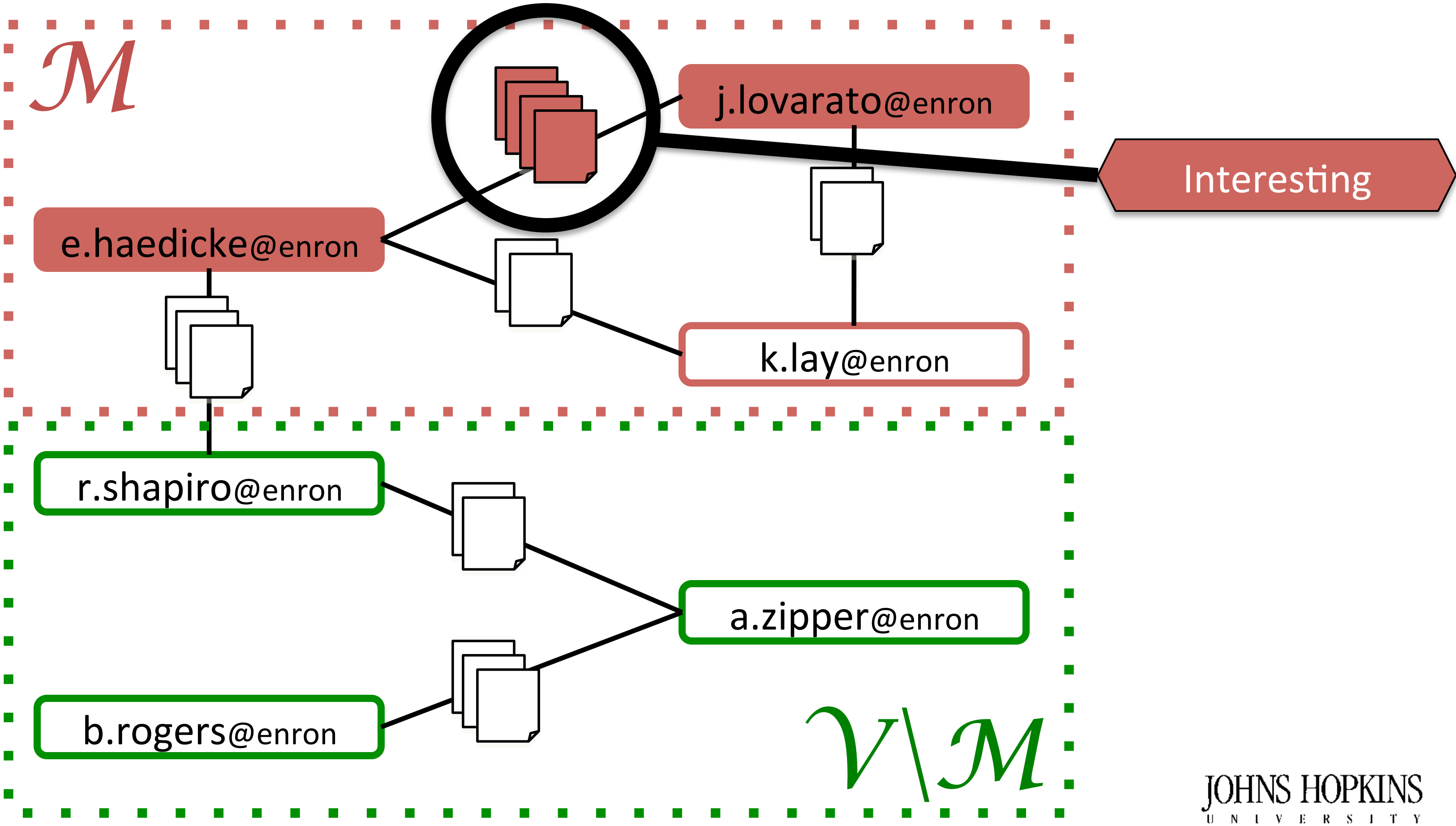
Training HLTs



human language technology
center of excellence

JOHNS HOPKINS
UNIVERSITY

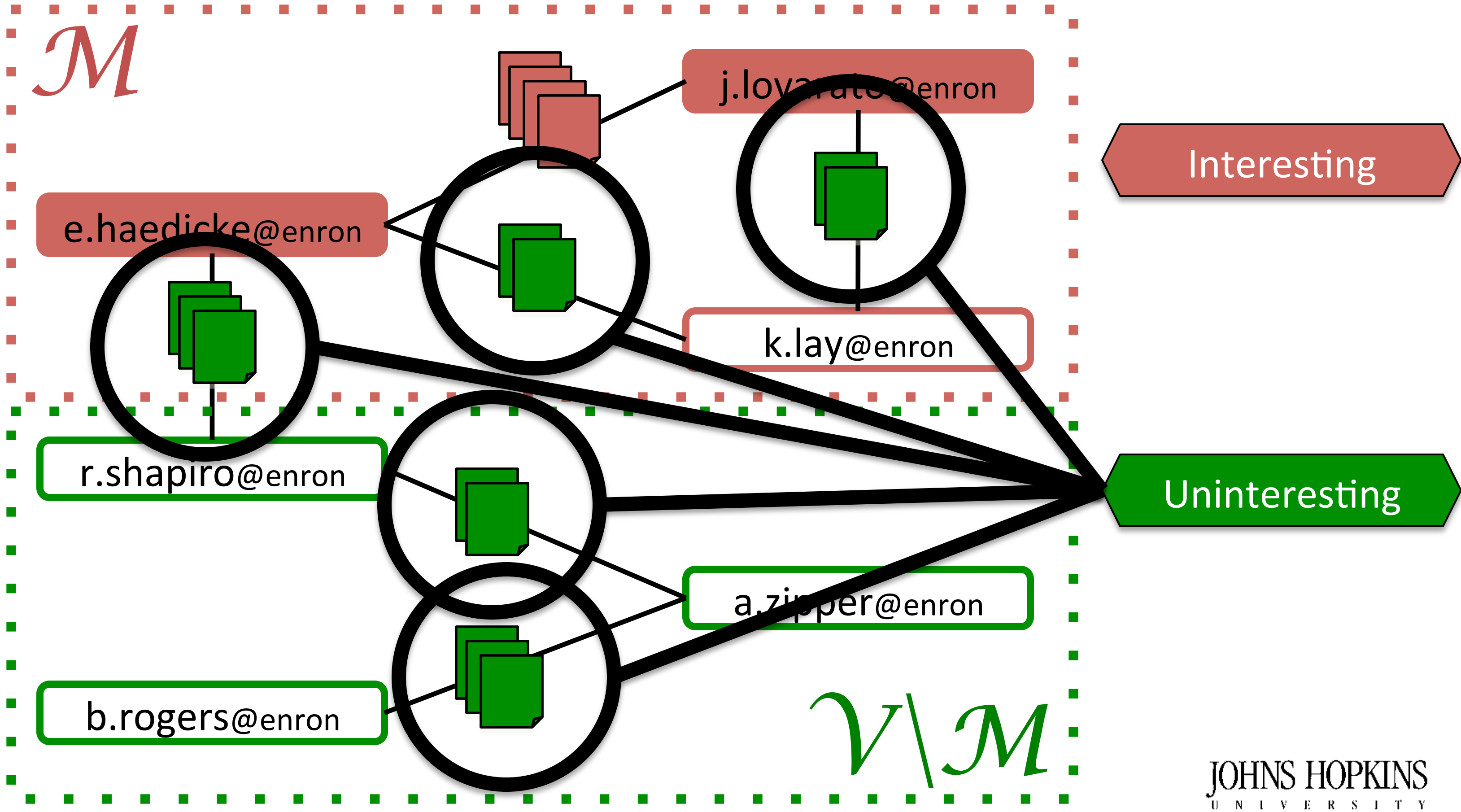
Training HLTs



human language technology
center of excellence

JOHNS HOPKINS
UNIVERSITY

Training HLTs



human language technology
center of excellence

JOHNS HOPKINS
UNIVERSITY

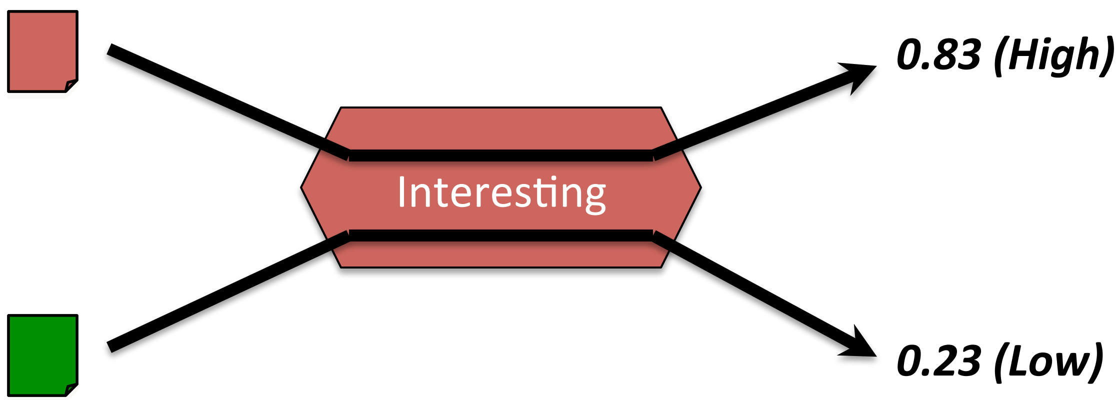
HLT$_1$: Average Word Count Histogram

- What proportion of the document $\boldsymbol{d_i}$ is made up of word $\boldsymbol{w_j}$?
- Each $\boldsymbol{d_i}$ represented as probability vector $\boldsymbol{x_i}$.
- $|\mathbf{x}|$ = $\boldsymbol{W}$, $\boldsymbol{W}$ word types in the corpus.
- Vector $\boldsymbol{I}$ is average of all interesting ( ) $\boldsymbol{x_i}$.

Interesting

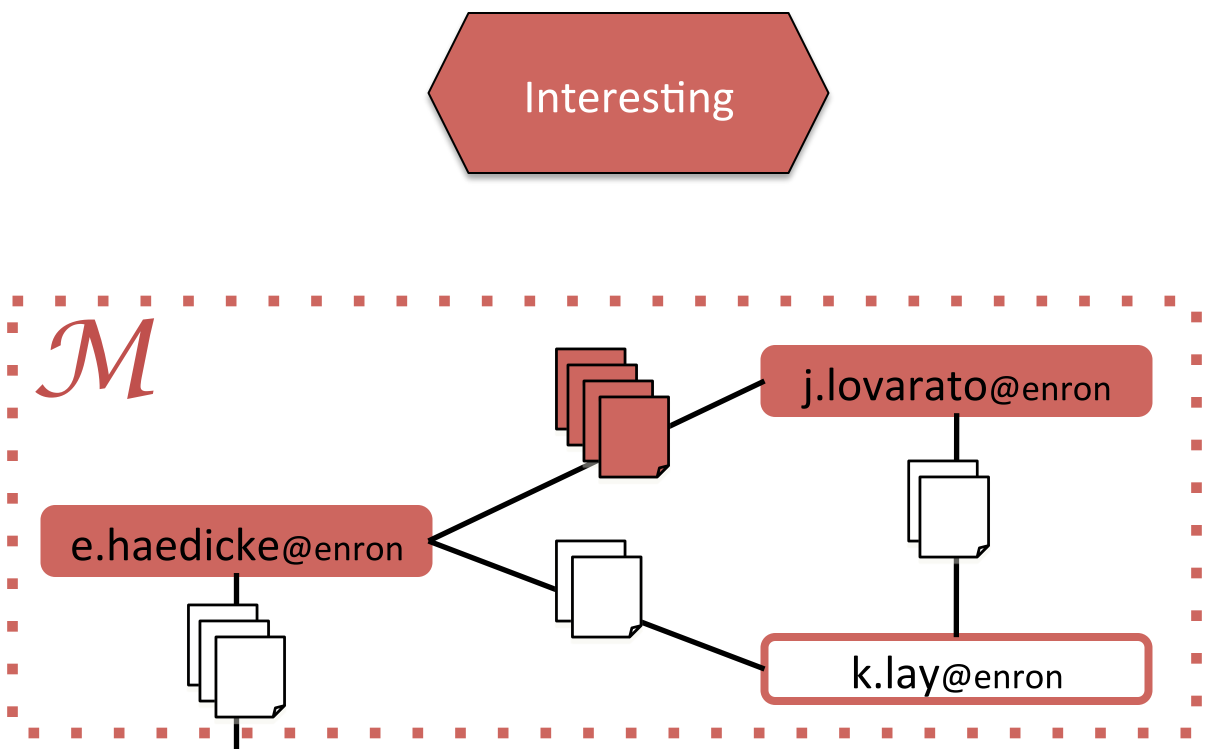
HLT$_1$: Average Word Count Histogram

- Score each $\boldsymbol{d_i}$ by 1-JS($\boldsymbol{x_i}$, $\boldsymbol{I}$)



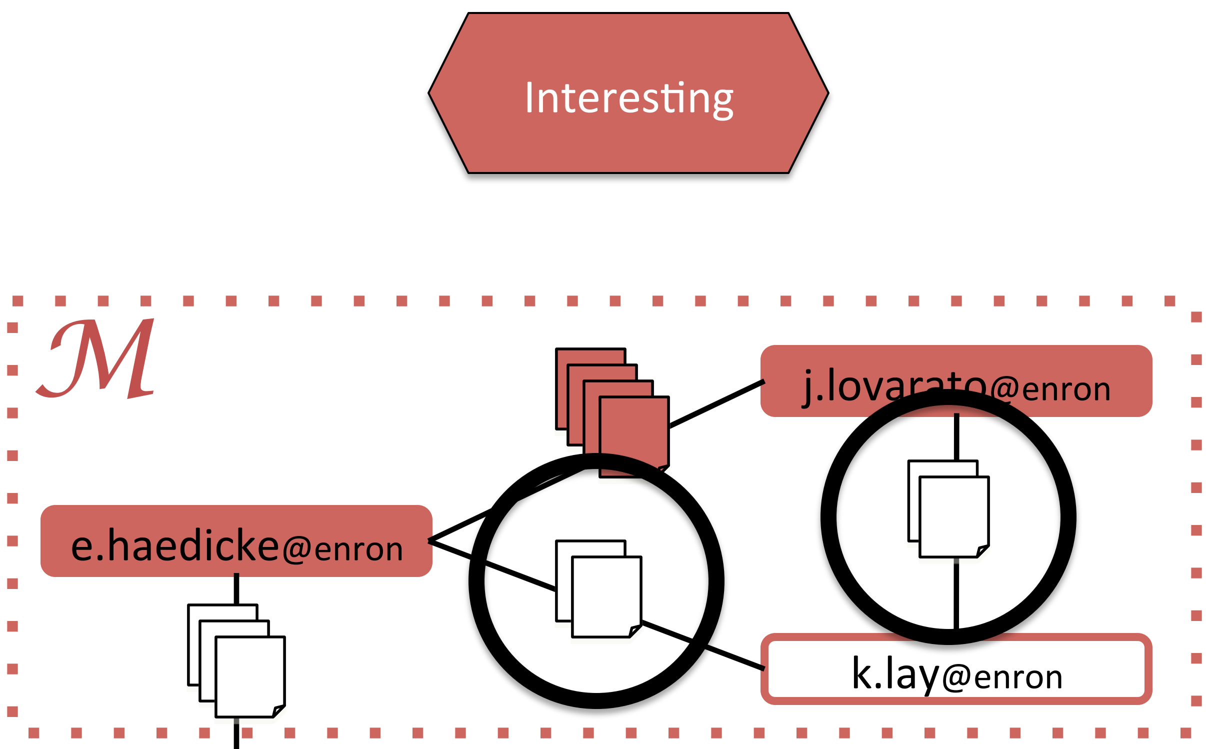
HLT$_1$: Average Word Count Histogram

- Score all d_i for each vertex: (k.lay@enron)
- Average scores



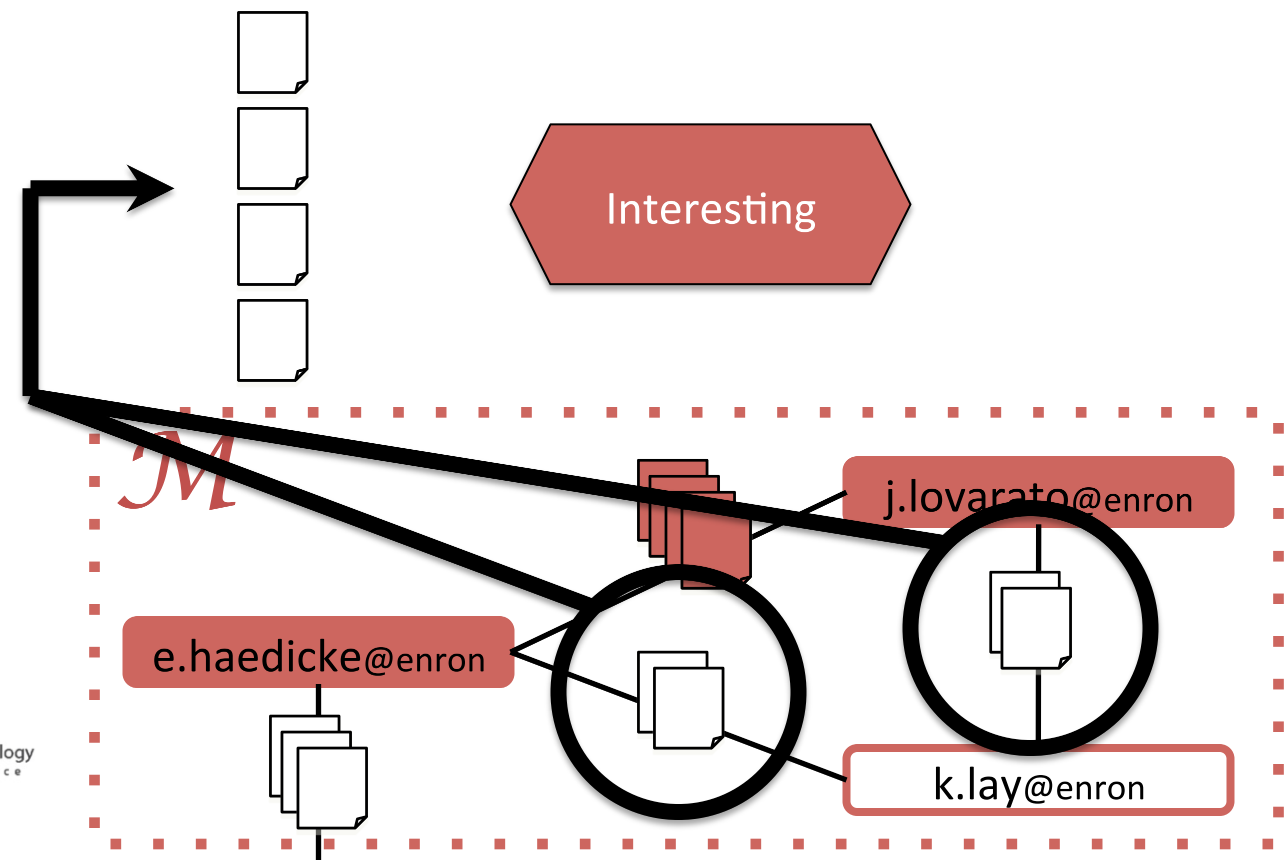
HLT$_1$: Average Word Count Histogram

- Score all d_i for each vertex: (k.lay@enron)
- Average scores



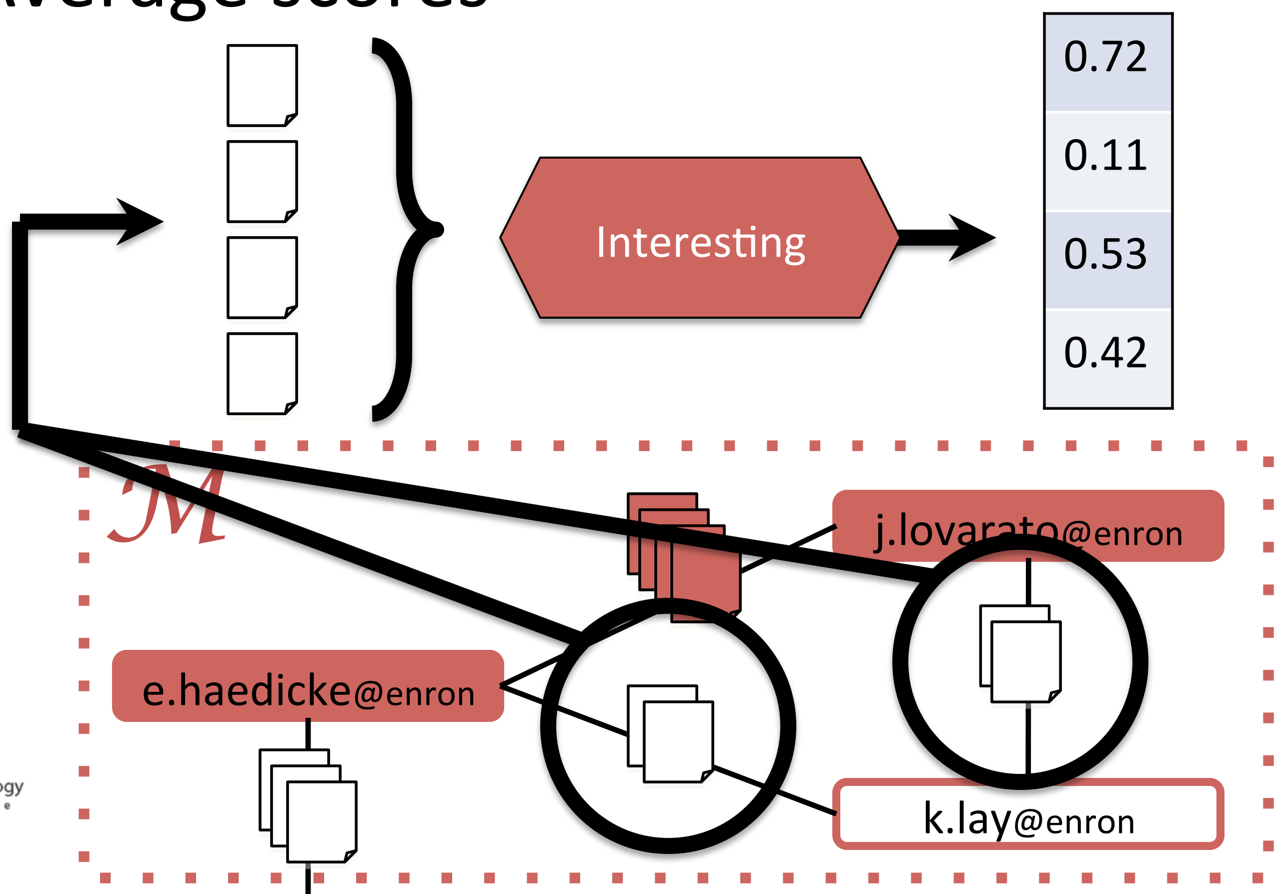
HLT$_1$: Average Word Count Histogram

- Score all $\boldsymbol{d_i}$ for each vertex: (k.lay@enron)
- Average scores



HLT$_1$: Average Word Count Histogram

- Score all d_i for each vertex: (k.lay@enron)
- Average scores

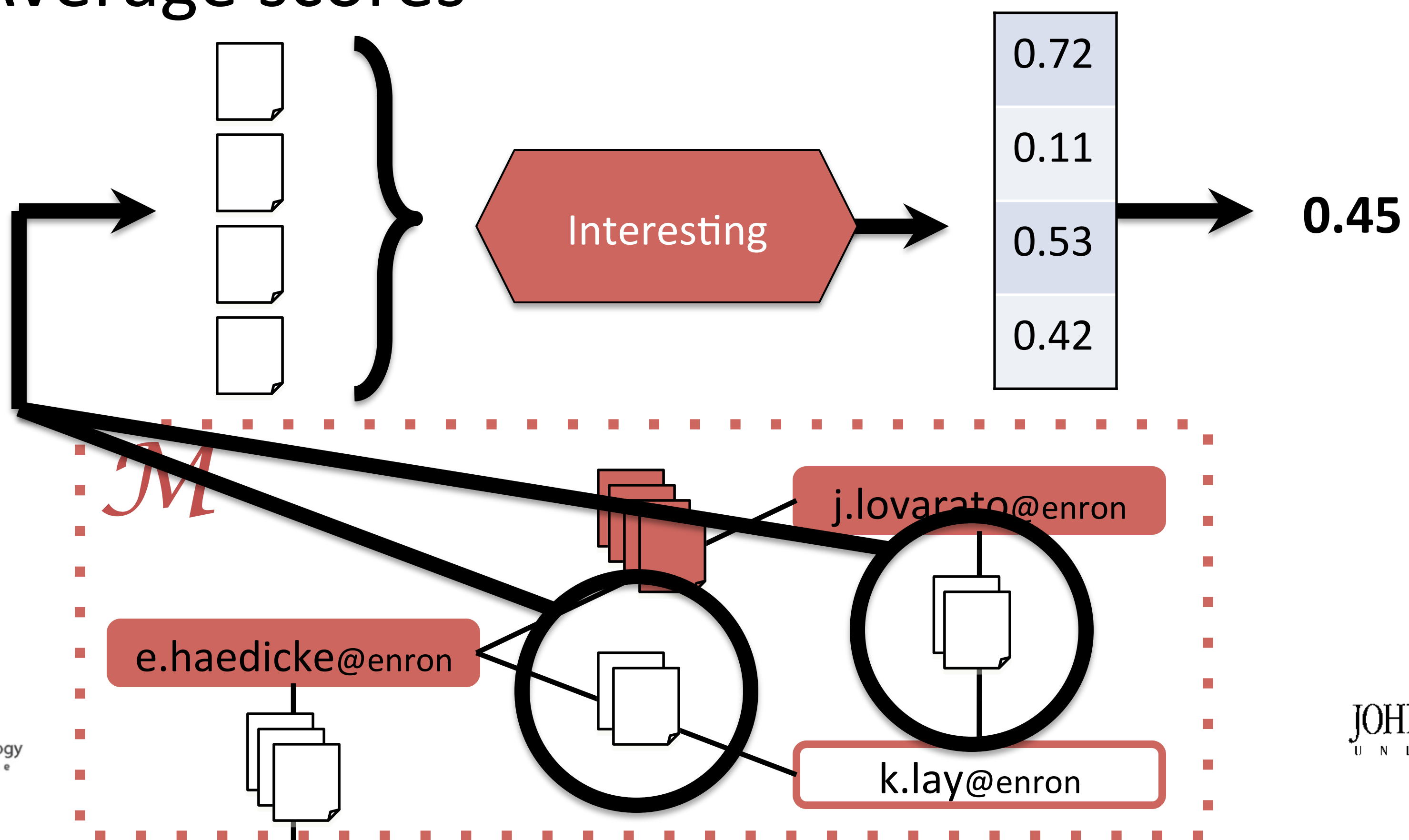


human language technology
center of excellence

JOHNS HOPKINS
UNIVERSITY

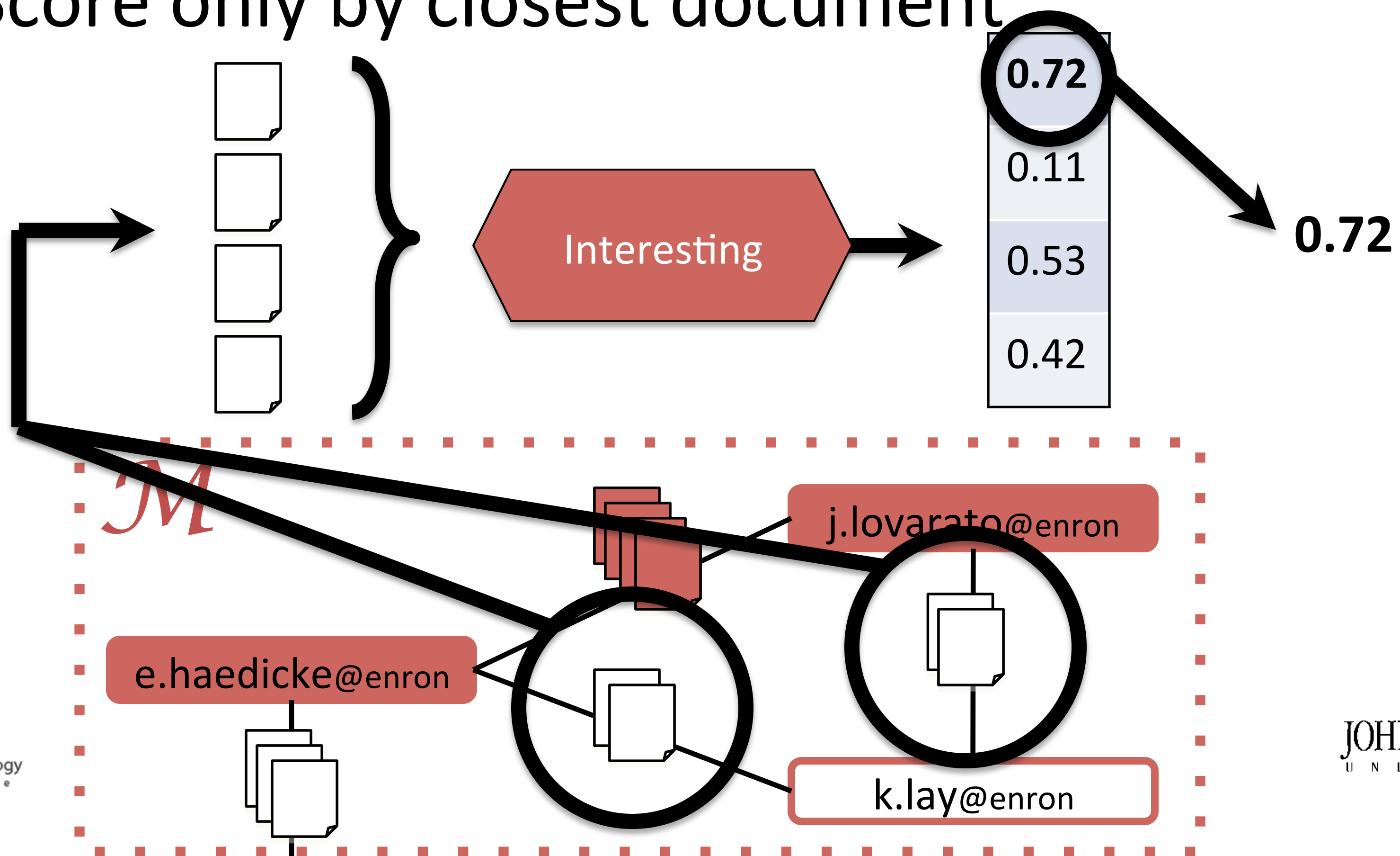
HLT$_1$: Average Word Count Histogram

- Score all $\boldsymbol{d_i}$ for each vertex: (k.lay@enron)
- Average scores



HLT$_2$: Closest Word Count Histogram

- Score all $\boldsymbol{d_i}$ for each vertex: (k.lay@enron)
- Score only by closest document



human language technology
center of excellence

JOHNS HOPKINS
UNIVERSITY

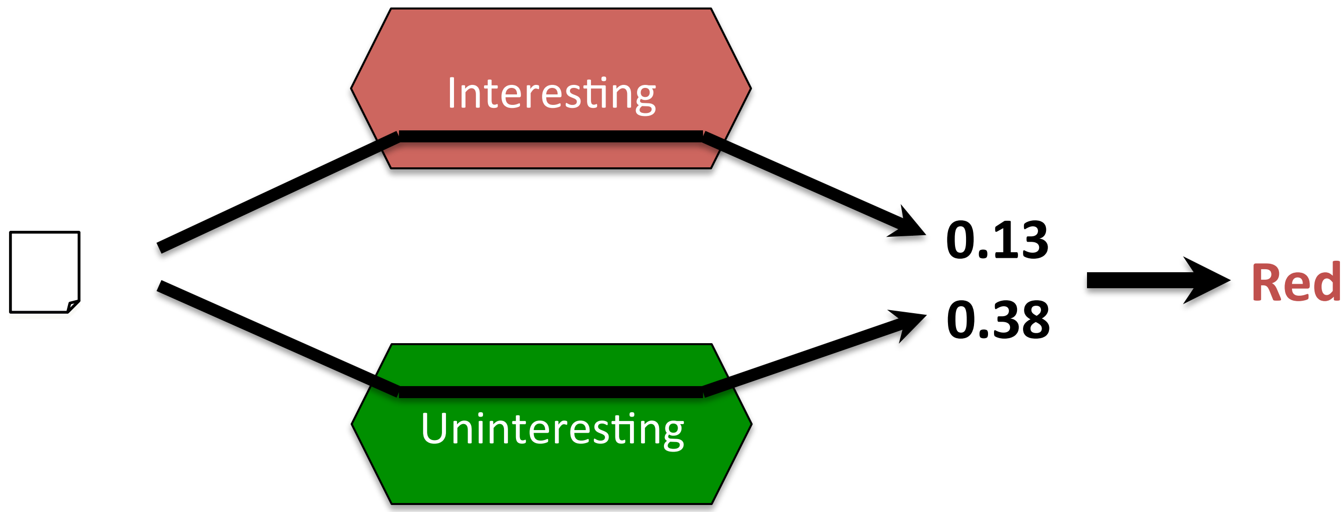
HLT_3: Compression Language Modeling

- How well does a given message compress?

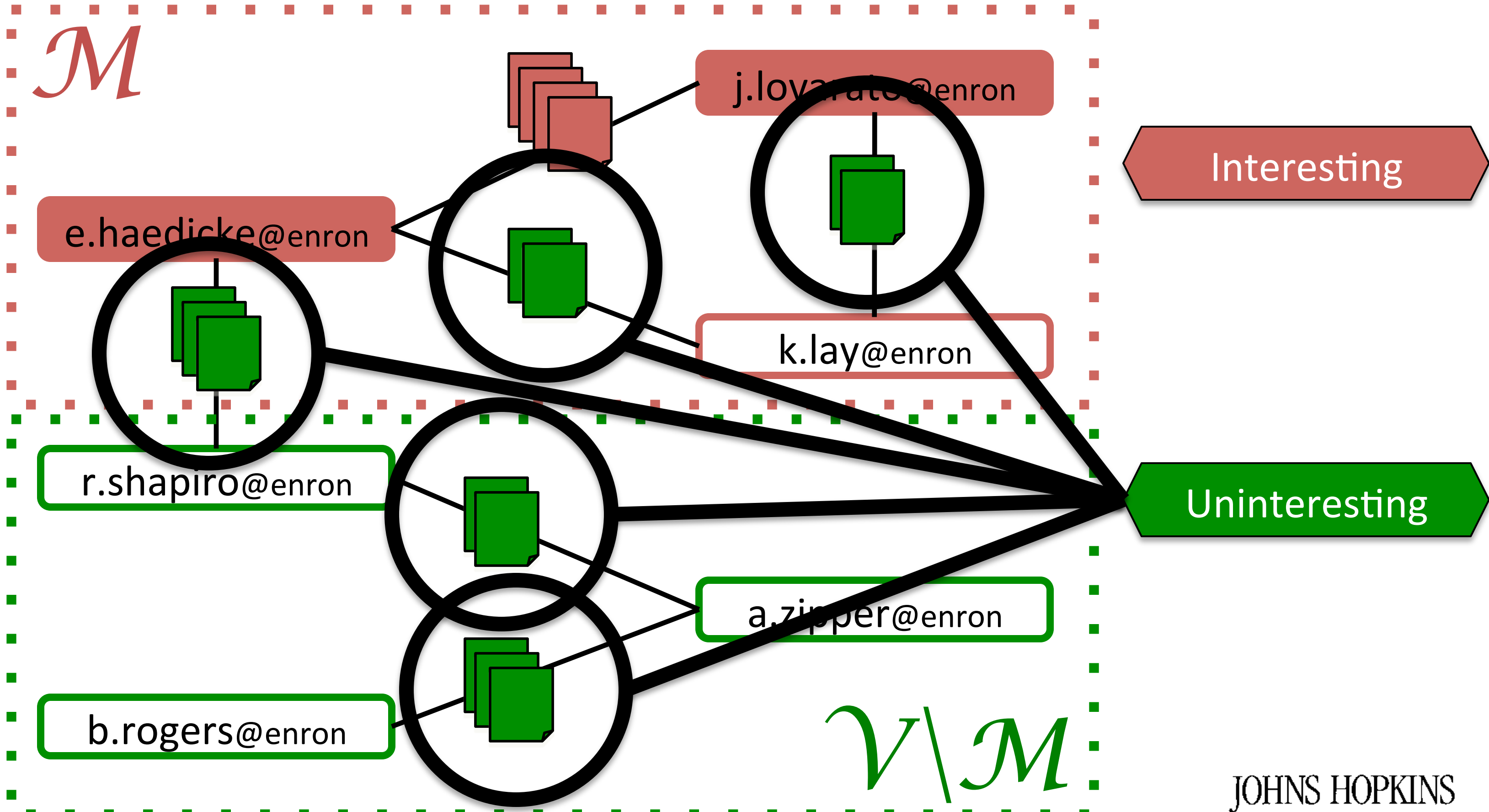
- Repeated sequences compress well
- Novel sequences do not compress well

HLT$_3$: Compression Language Modeling

- Discriminative Model – which model fits better?



HLT$_3$: Compression Language Modeling



human language technology
center of excellence

JOHNS HOPKINS
UNIVERSITY

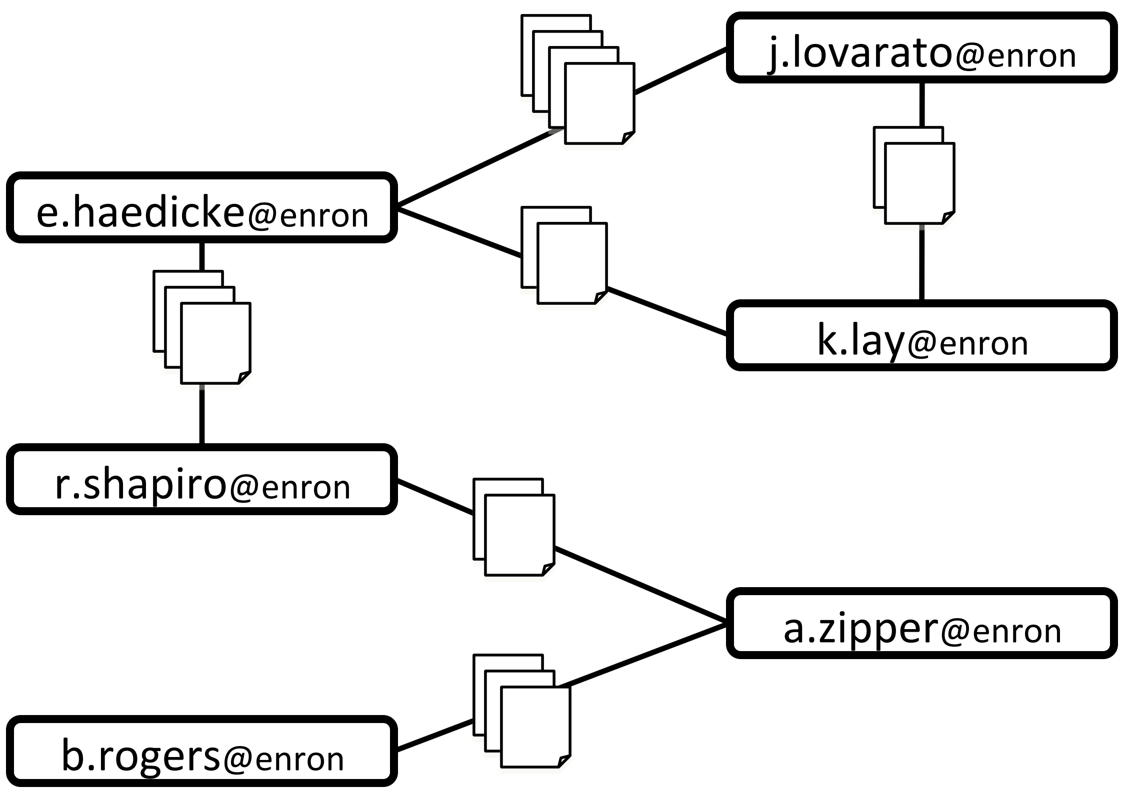
HLT_4: Topic Modeling

- Latent Dirichlet Allocation (ala Blei, Wallach, McCallum, Mimno, ...) [we use *mallet*]

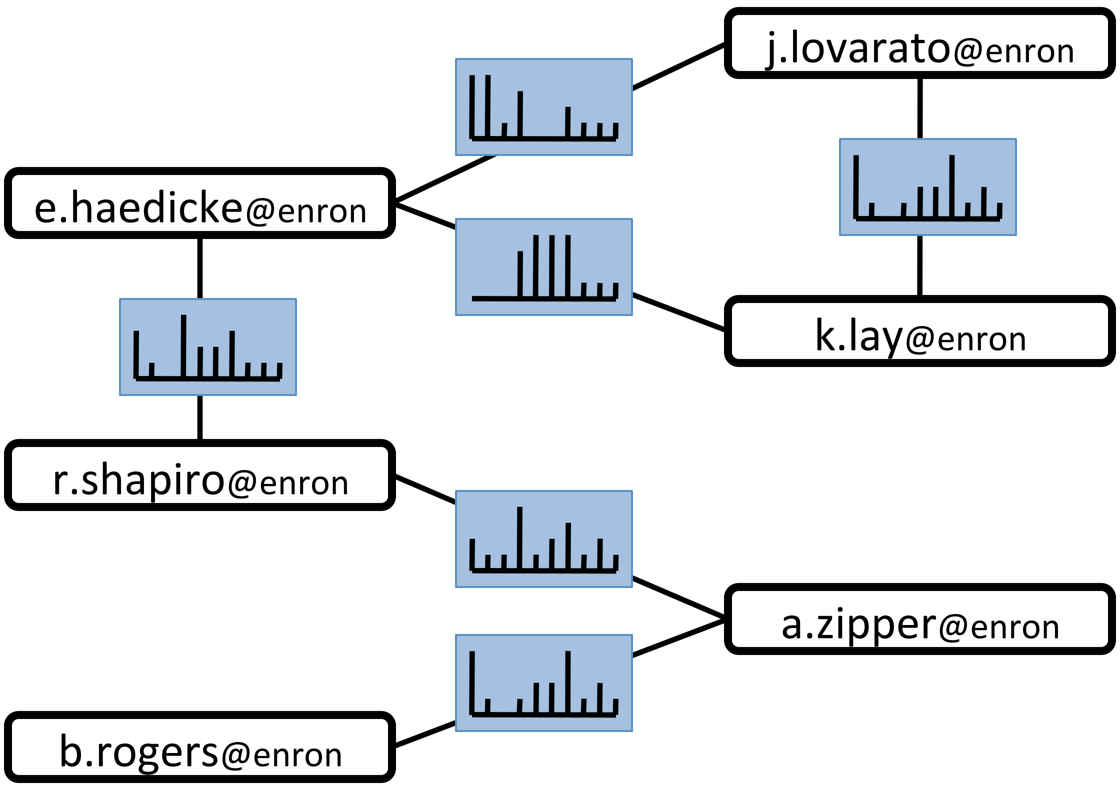
- Each document is a mixture of topics
- Each topic is a probability distribution over ***W***

mallet is available from http://mallet.cs.umass.edu/

HLT$_4$: Topic Modeling



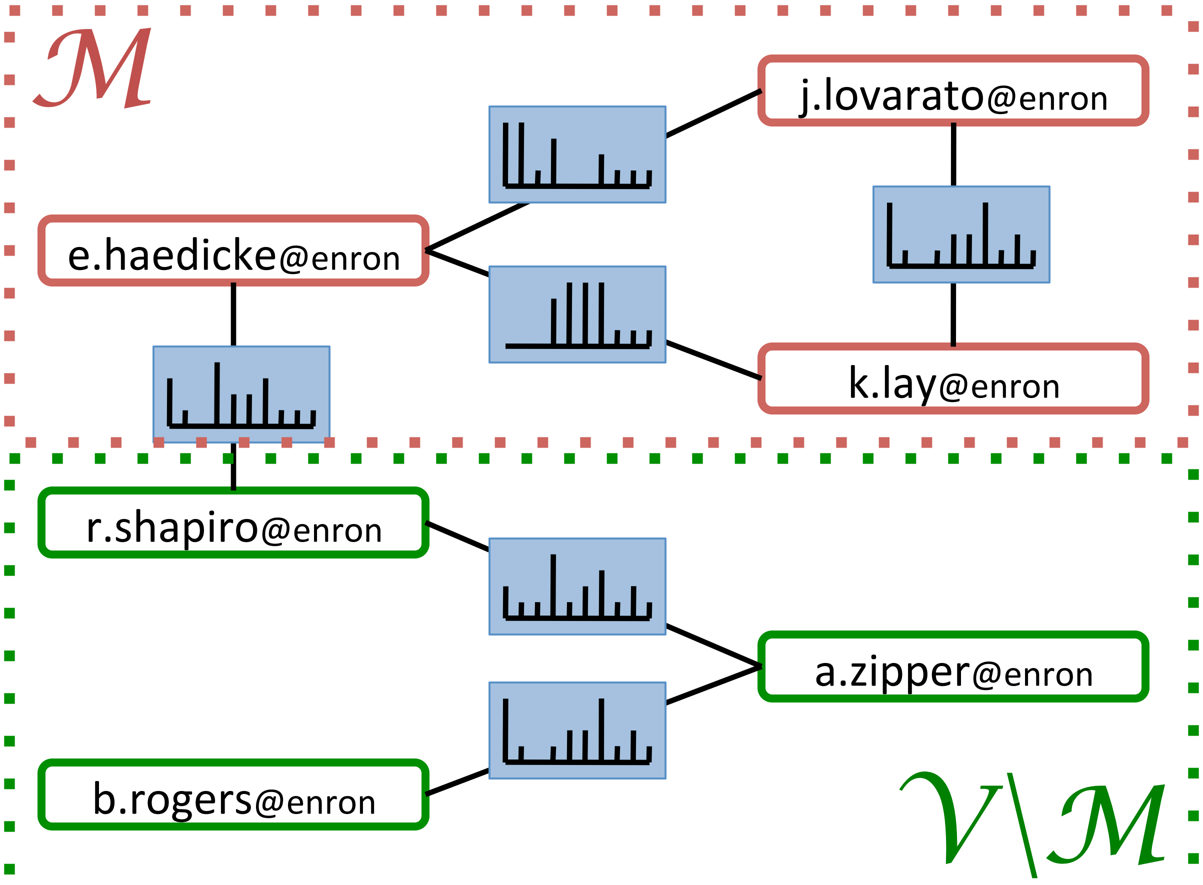
HLT$_4$: Topic Modeling



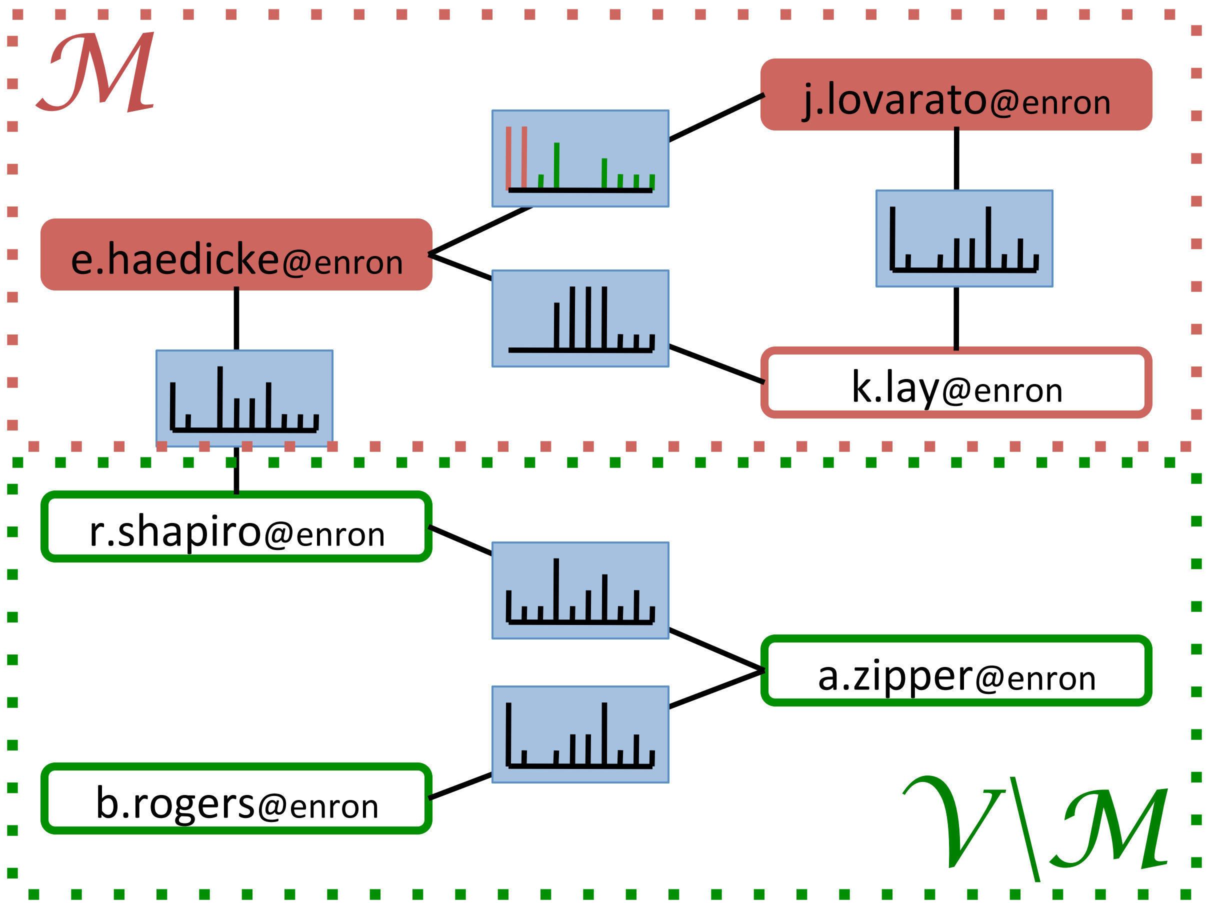
human language technology
center of excellence

JOHNS HOPKINS
UNIVERSITY

HLT$_4$: Topic Modeling



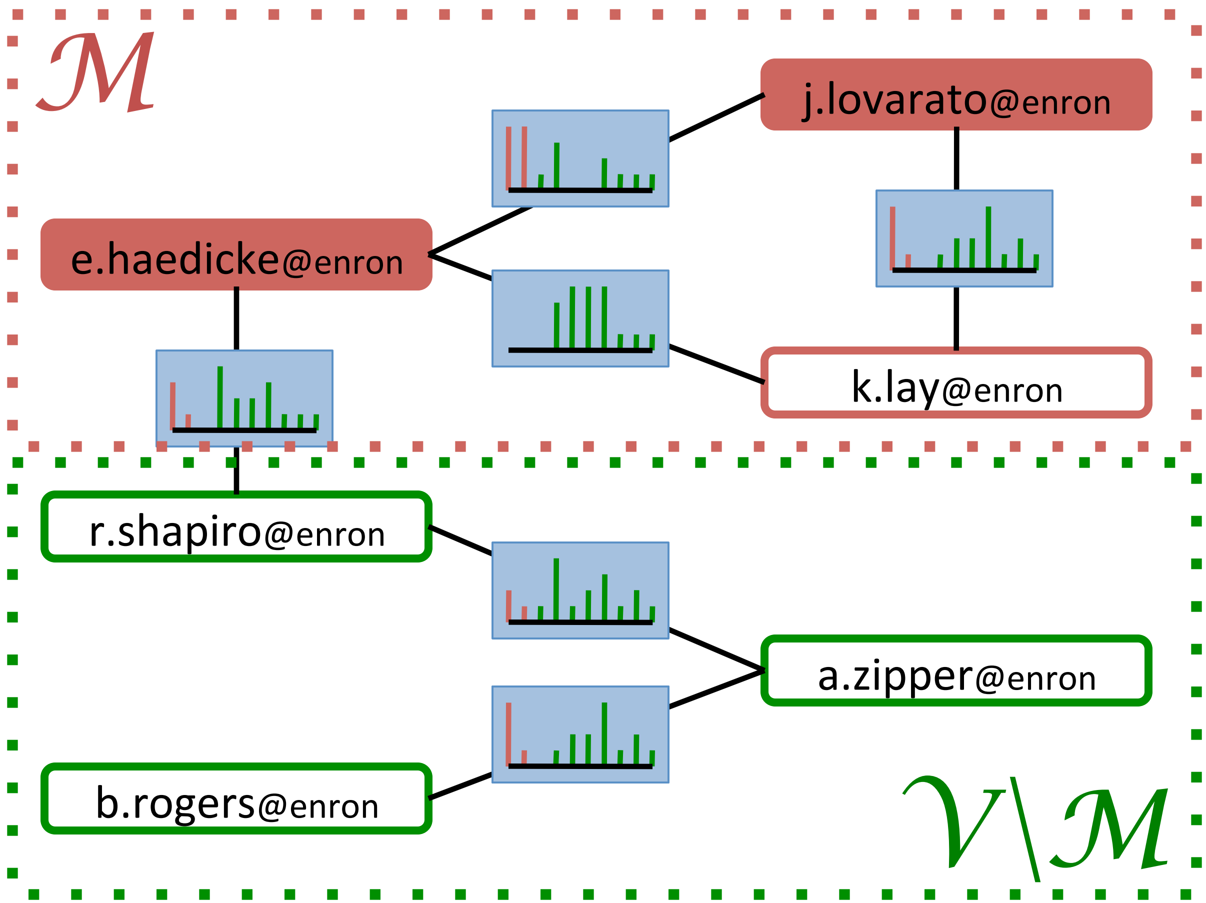
HLT$_4$: Topic Modeling



human language technology
center of excellence

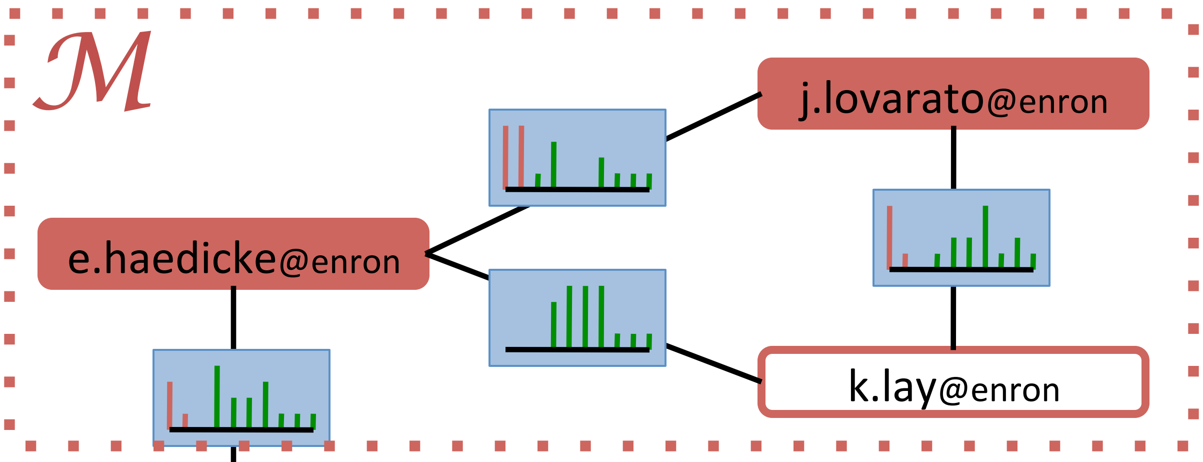
JOHNS HOPKINS
UNIVERSITY

HLT$_4$: Topic Modeling



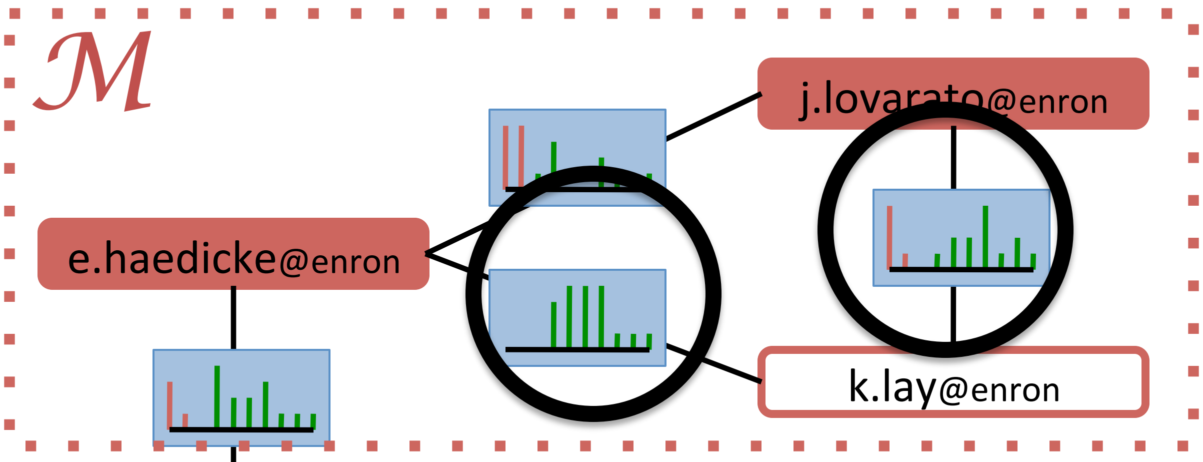
HLT$_4$: Topic Modeling

- Score all $\boldsymbol{d_i}$ for each vertex: (k.lay@enron)
- Sum the weight given to topics of interest



HLT$_4$: Topic Modeling

- Score all $\boldsymbol{d_i}$ for each vertex: (k.lay@enron)
- Sum the weight given to topics of interest

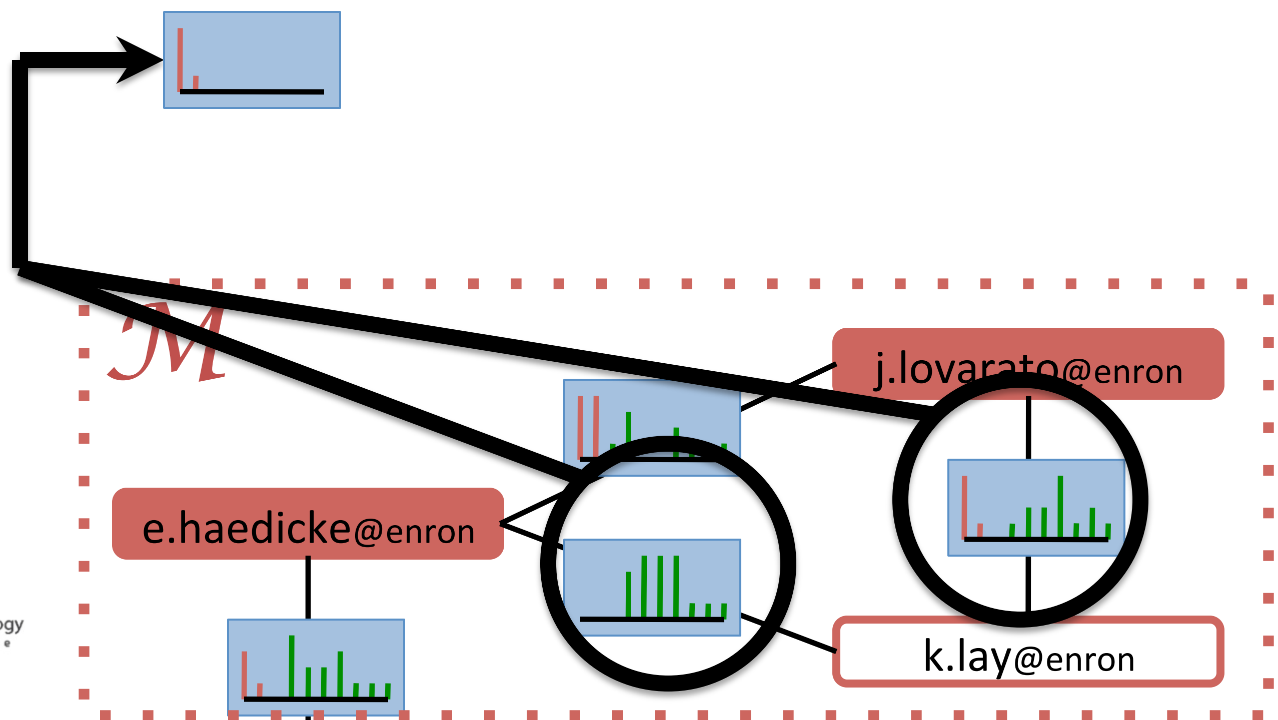


human language technology
center of excellence

JOHNS HOPKINS
UNIVERSITY

HLT$_4$: Topic Modeling

- Score all $\boldsymbol{d_i}$ for each vertex: (k.lay@enron)
- Sum the weight given to topics of interest

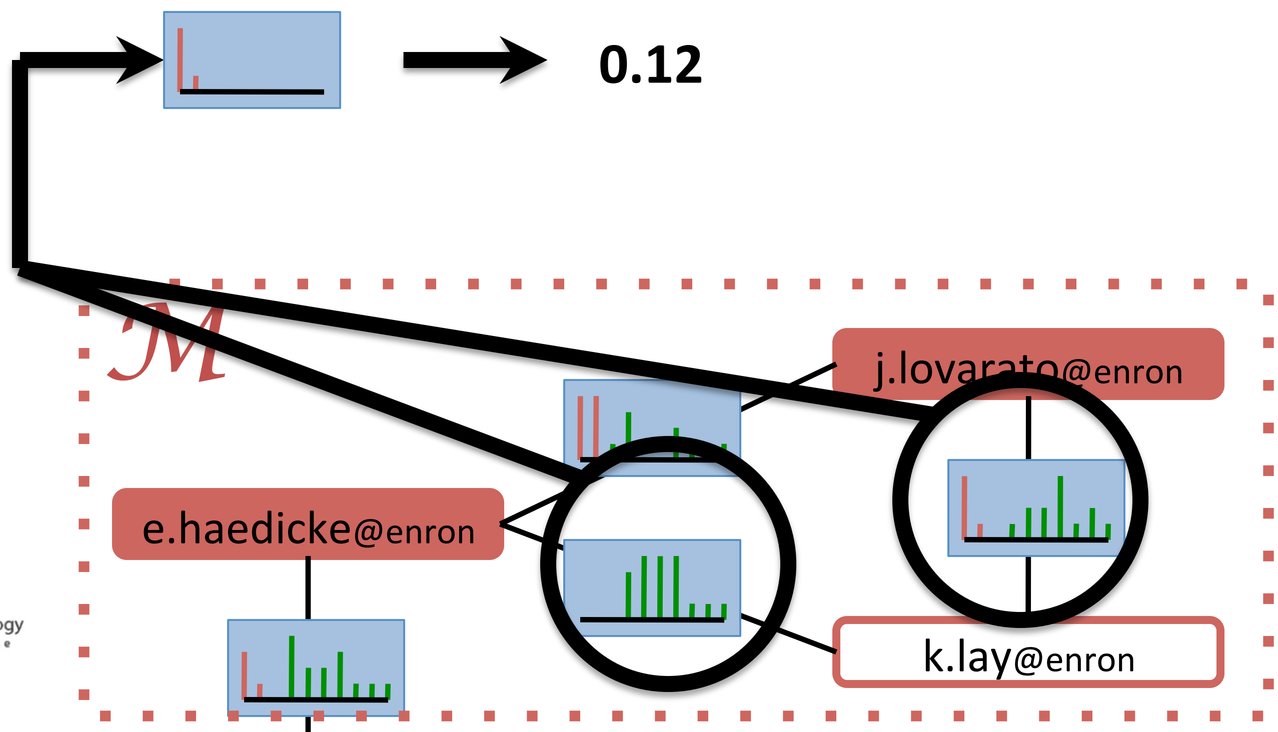


human language technology
center of excellence

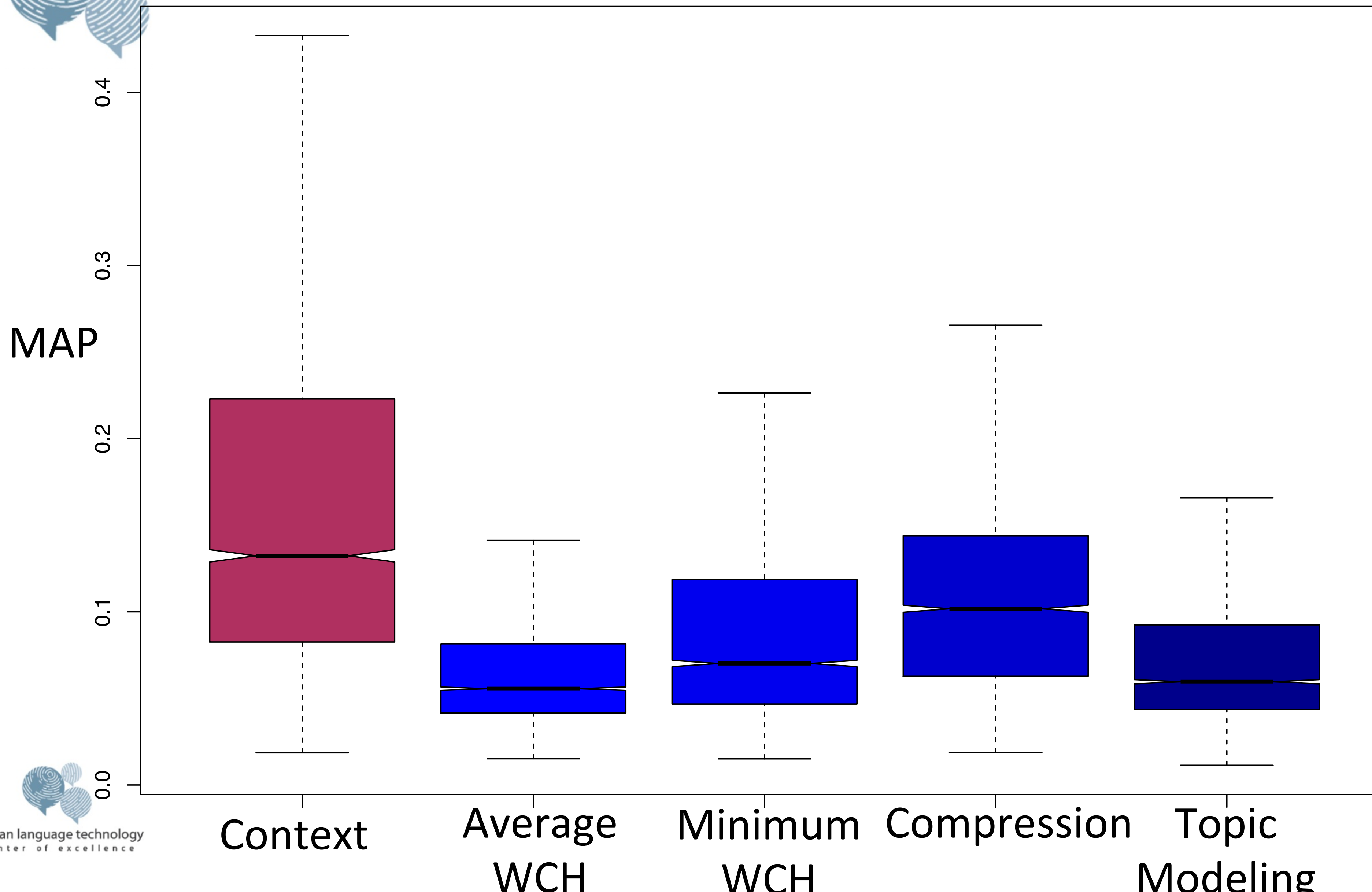
JOHNS HOPKINS
UNIVERSITY

HLT$_4$: Topic Modeling

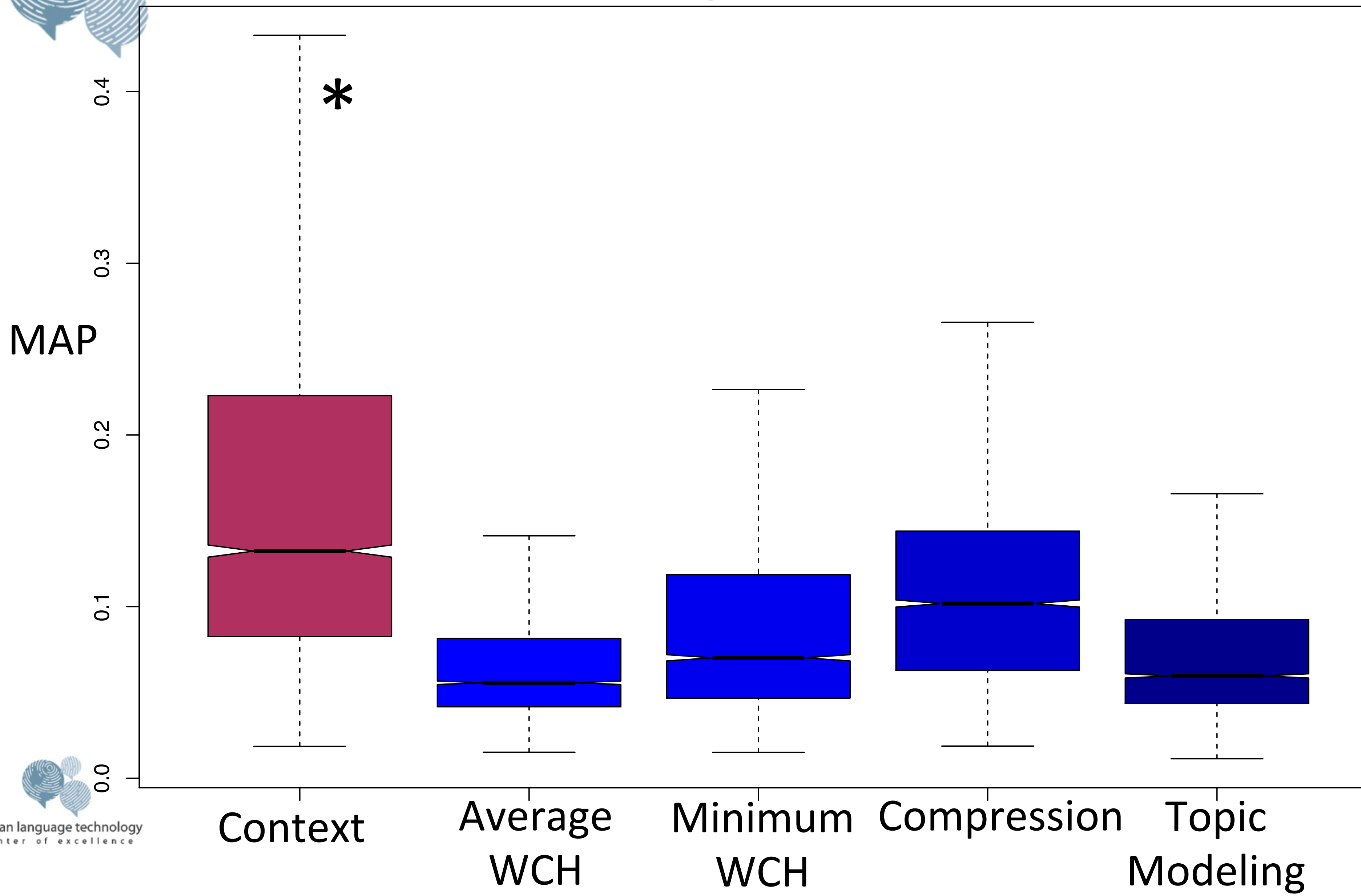
- Score all $\boldsymbol{d_i}$ for each vertex: (k.lay@enron)
- Sum the weight given to topics of interest



Individual Analytic Performance



Individual Analytic Performance



human language technology
center of excellence

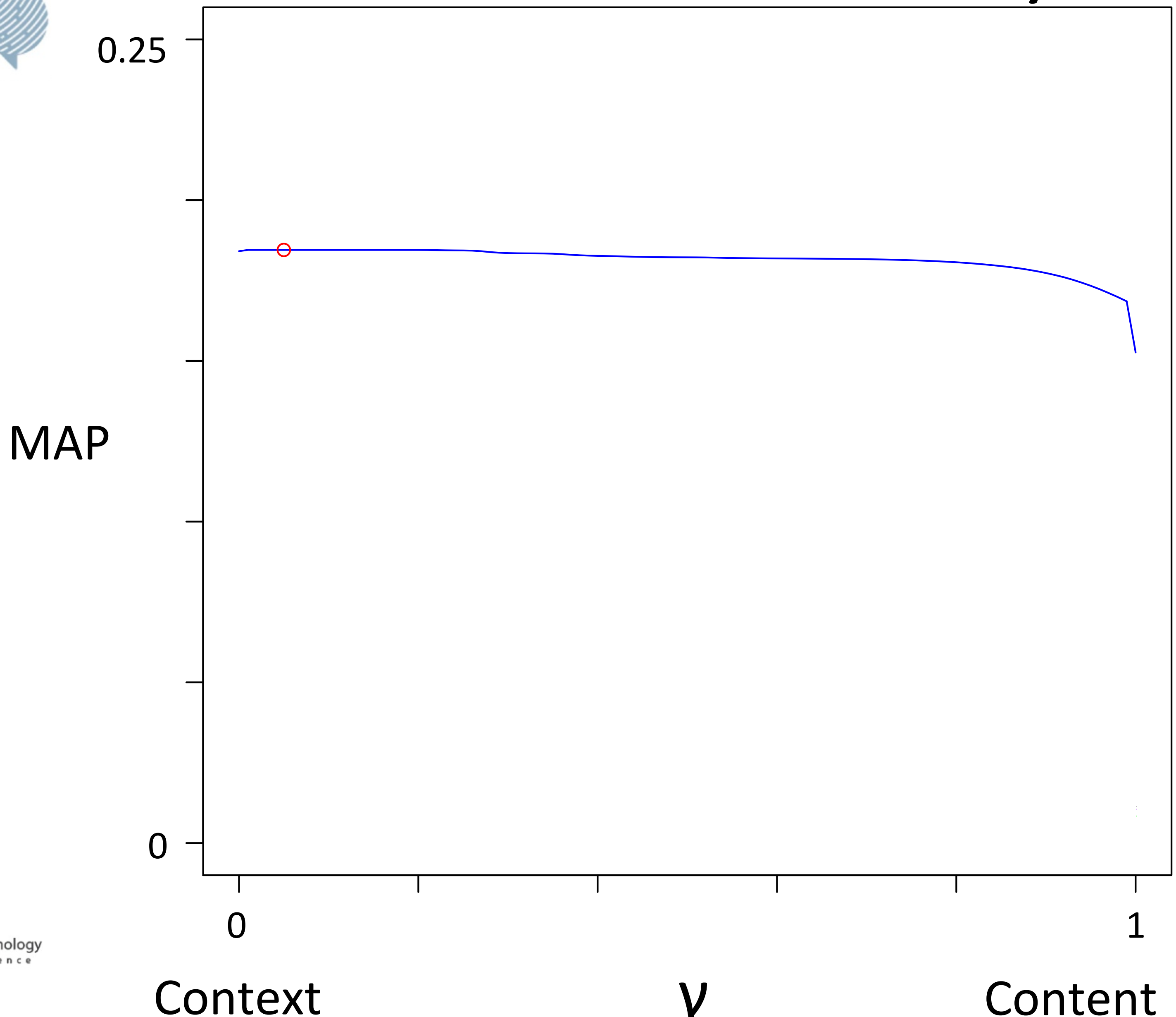
Outline

- Introduction
- Method
  - Importance Sampling
  - Evaluation
- Analytics – Content and Context
- Fusions
- Conclusions & Future Directions

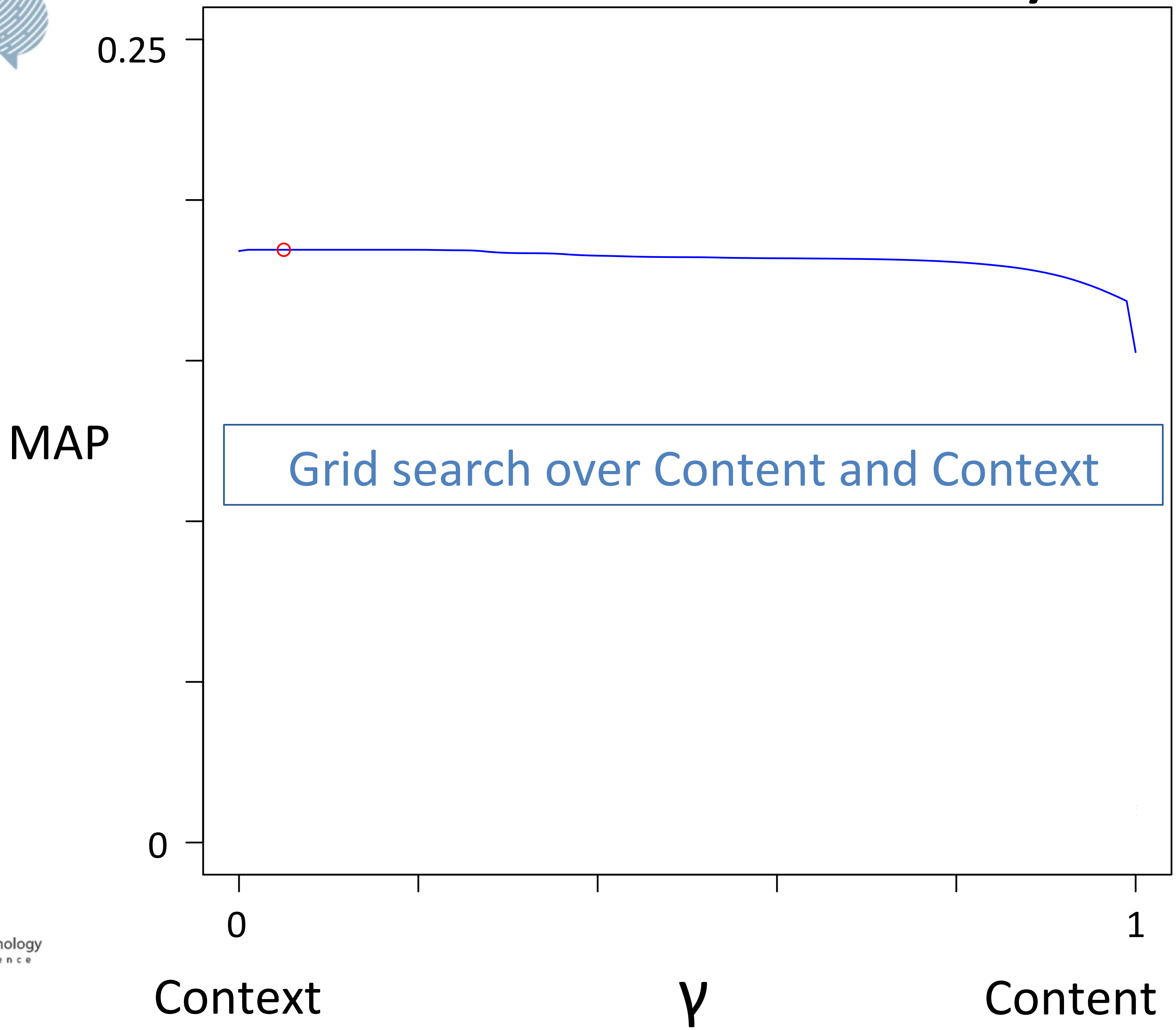
Linear Fusion

- $F = \gamma\ HLT_1 + (1-\gamma)\ Context$
  - 2 Analytics

Linear fusion – 2 Analytics



Linear fusion – 2 Analytics



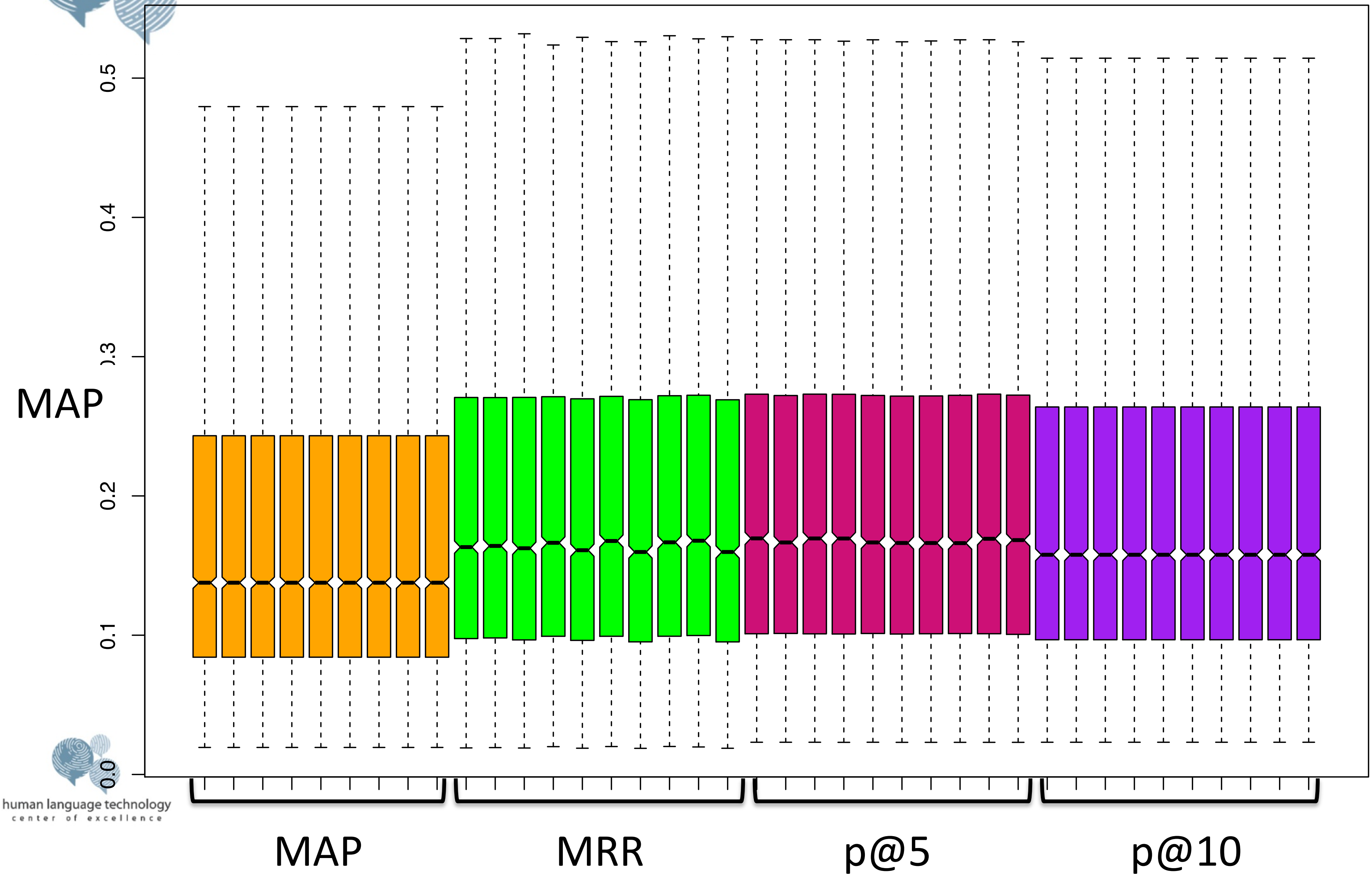
human language technology
center of excellence

JOHNS HOPKINS
UNIVERSITY

Linear Fusion

- $F = \gamma\, HLT_1 + (1-\gamma)\, Context$
  - 2 Analytics
- $F = \sum_i (\gamma_i\, HLT_i) + (1-\sum_i \gamma_i)\, Context$
  - Arbitrary number of Analytics
- NB: Scores must be calibrated.

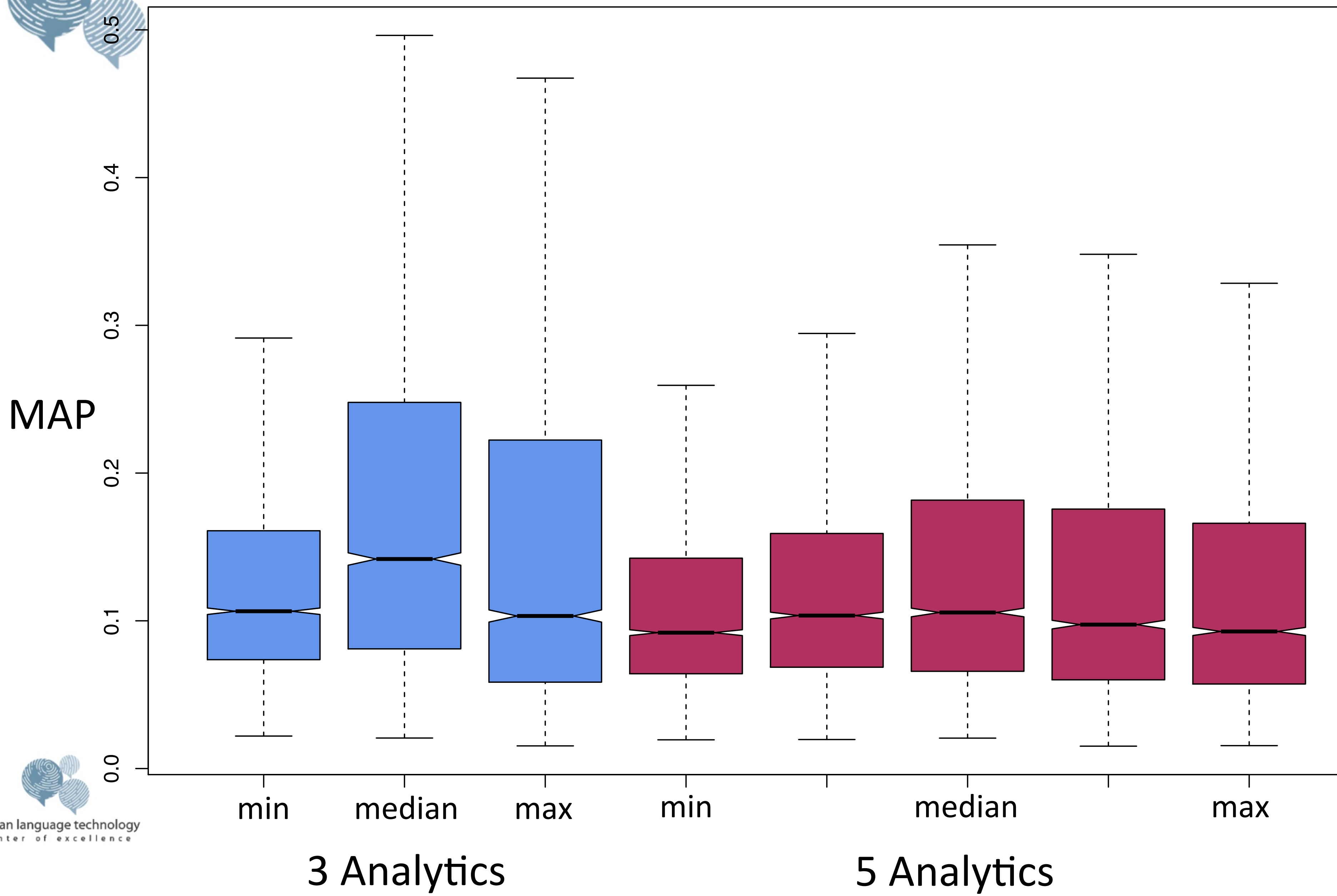
Gridsearch Linear Combinations



Rank Fusion

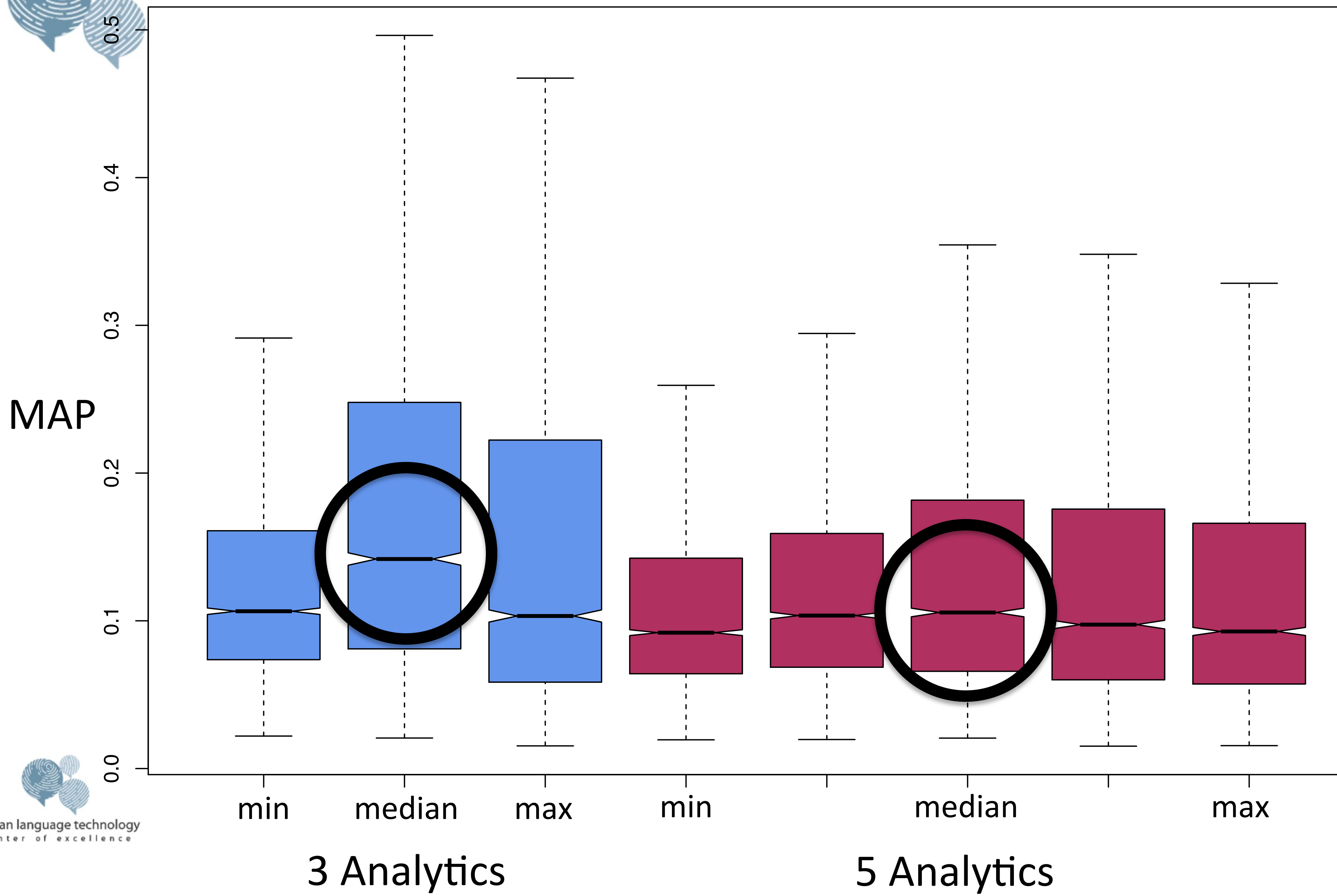
- Fuse *ranks* instead of *scores*.
- Rank vertices by each analytic.
- Each vertex represented by vector of ranks
- Fusion score is a function of that vector
- Min() – One measure can damn you
- Max() – One measure can save you
- Median() – Something in between

Rank Fusion



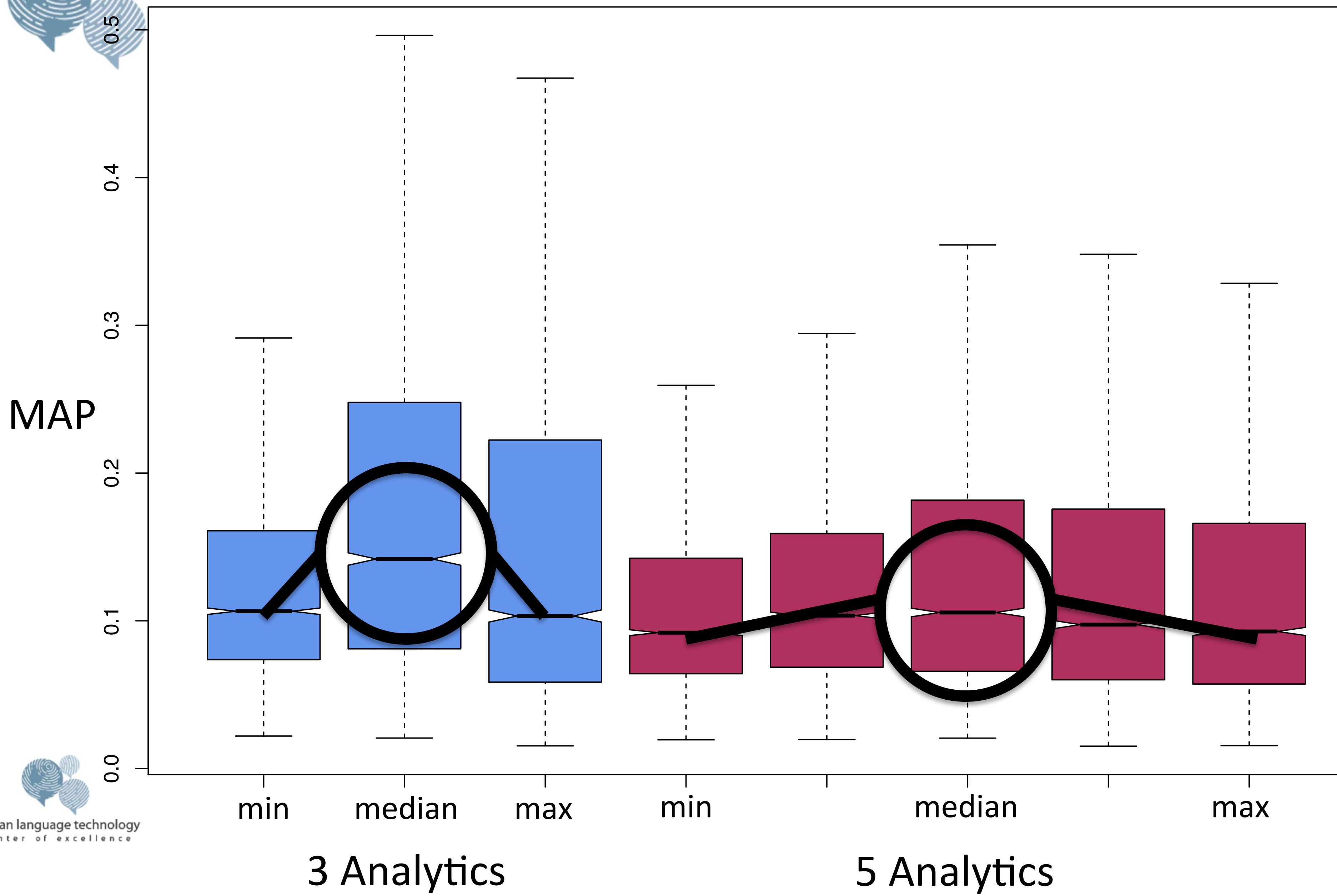
human language technology
center of excellence

Rank Fusion



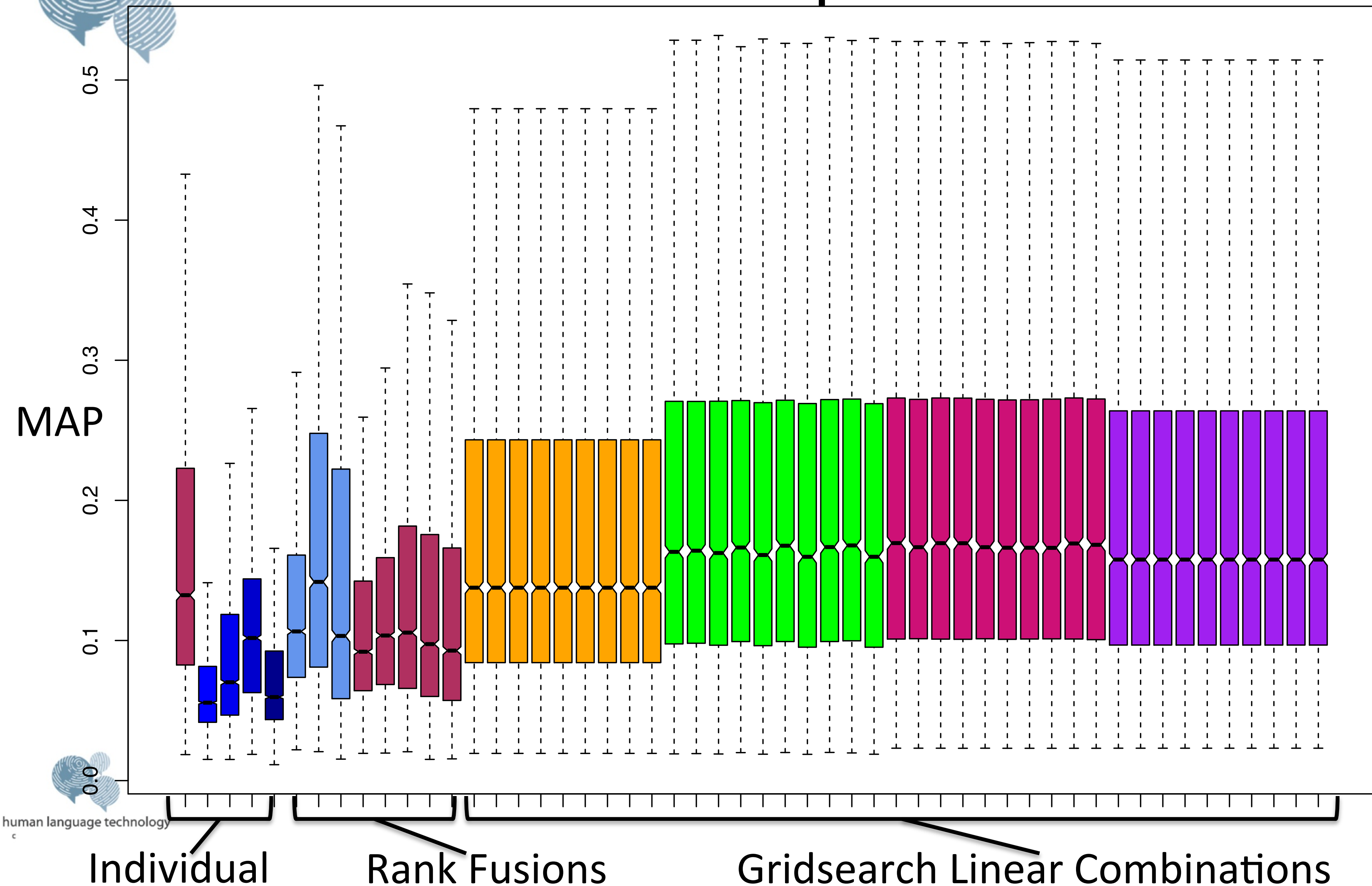
human language technology
center of excellence

Rank Fusion



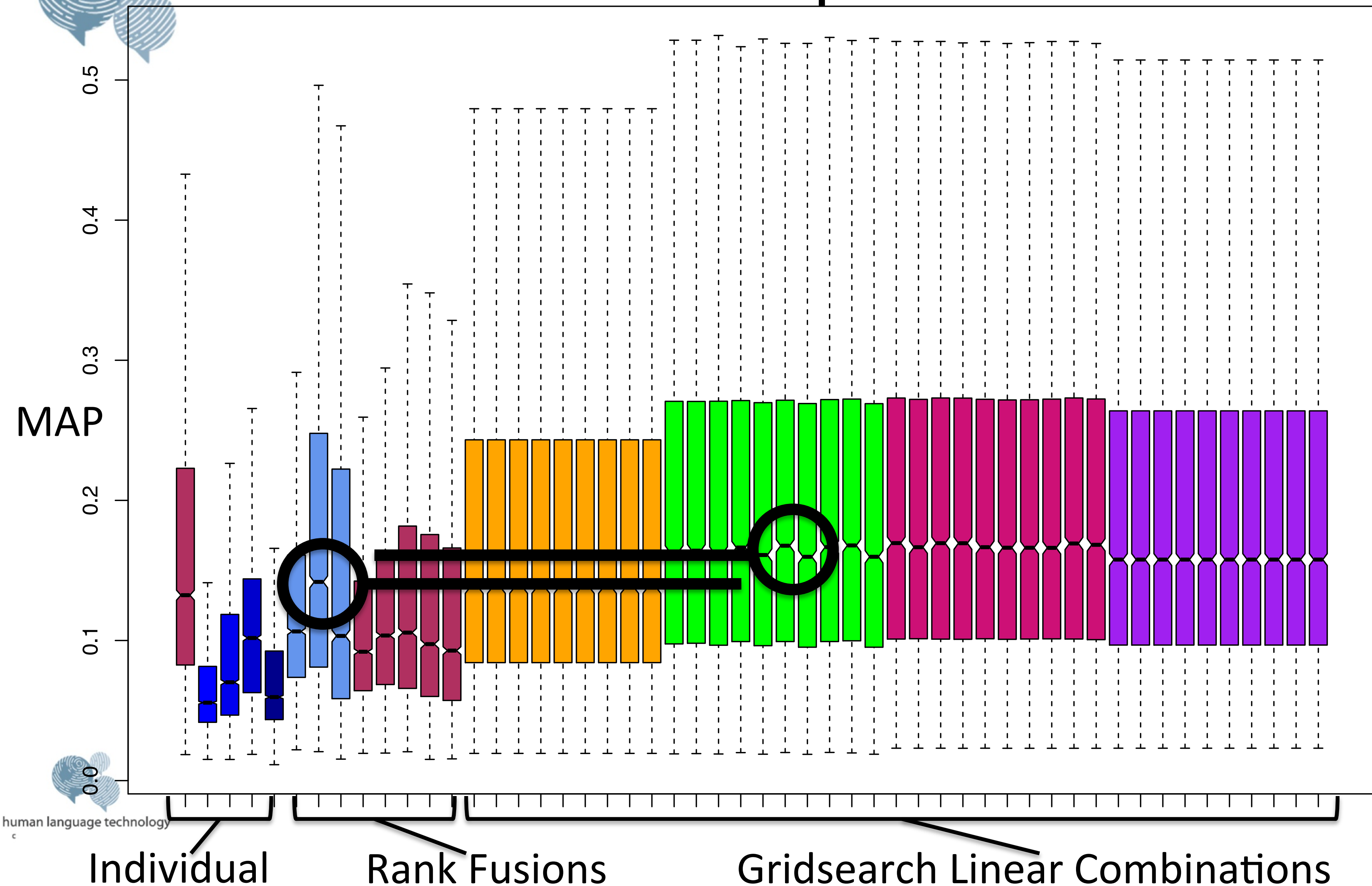
human language technology
center of excellence

Overall Comparisons

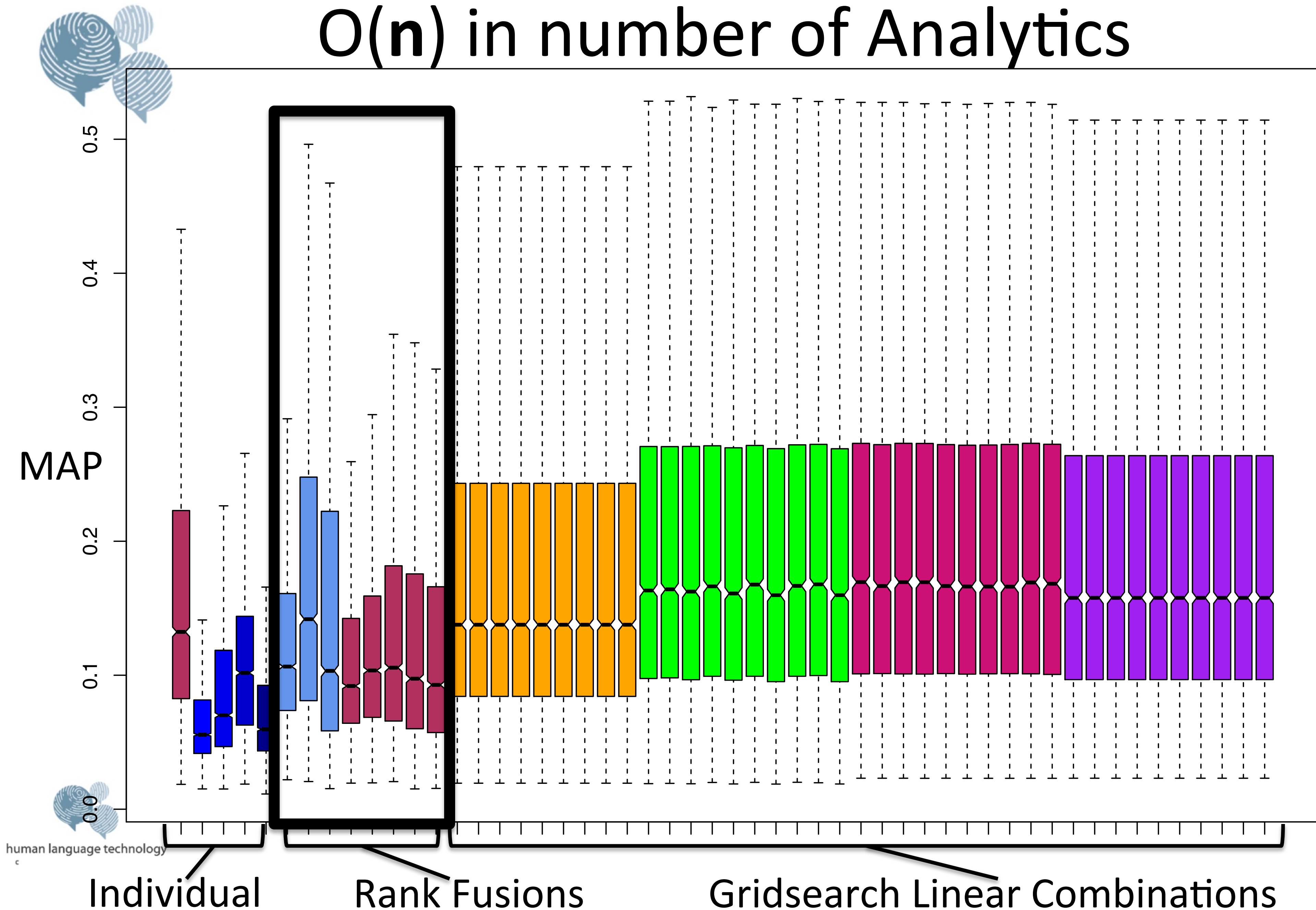


human language technology

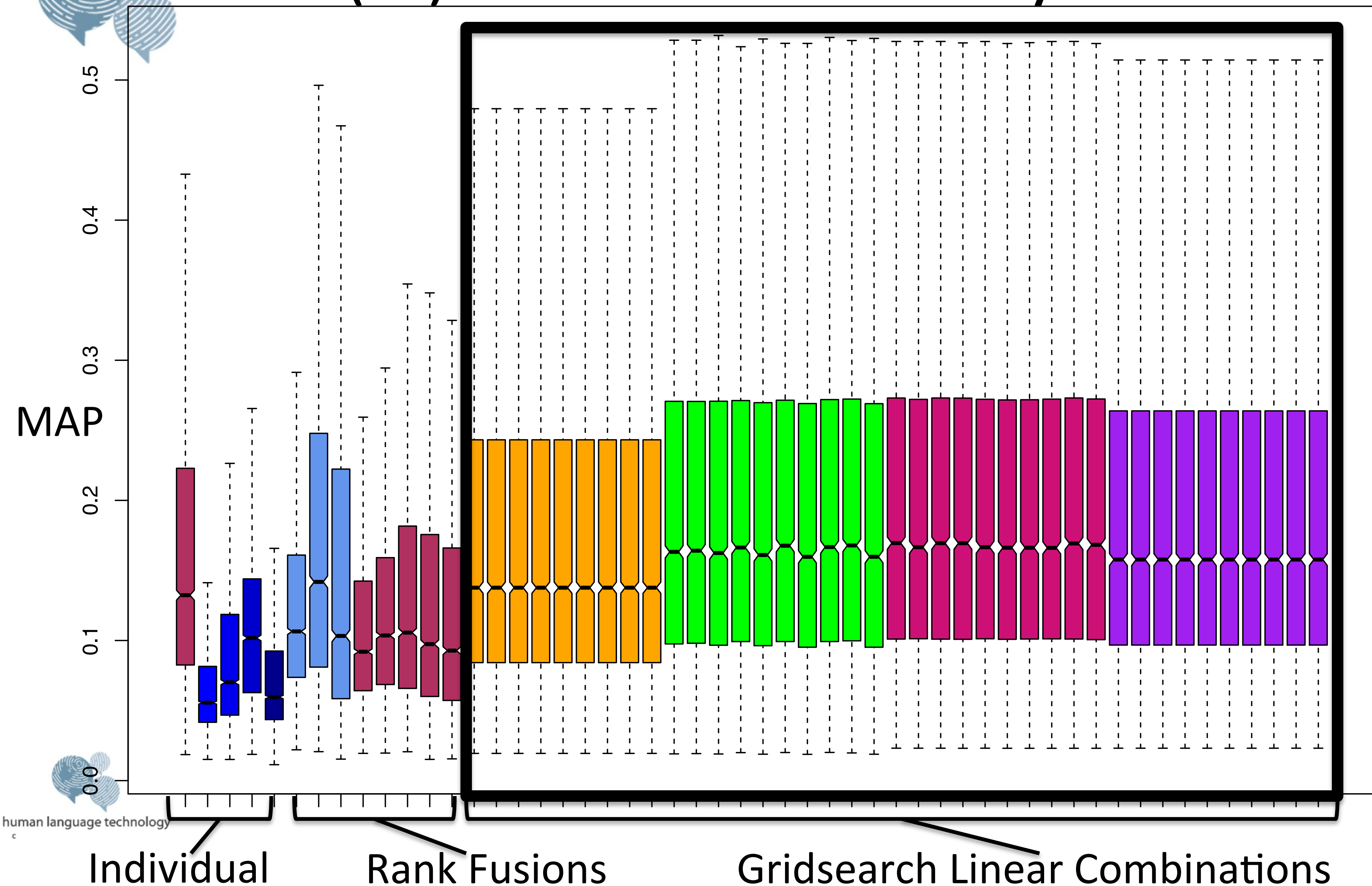
Overall Comparisons



O(**n**) in number of Analytics

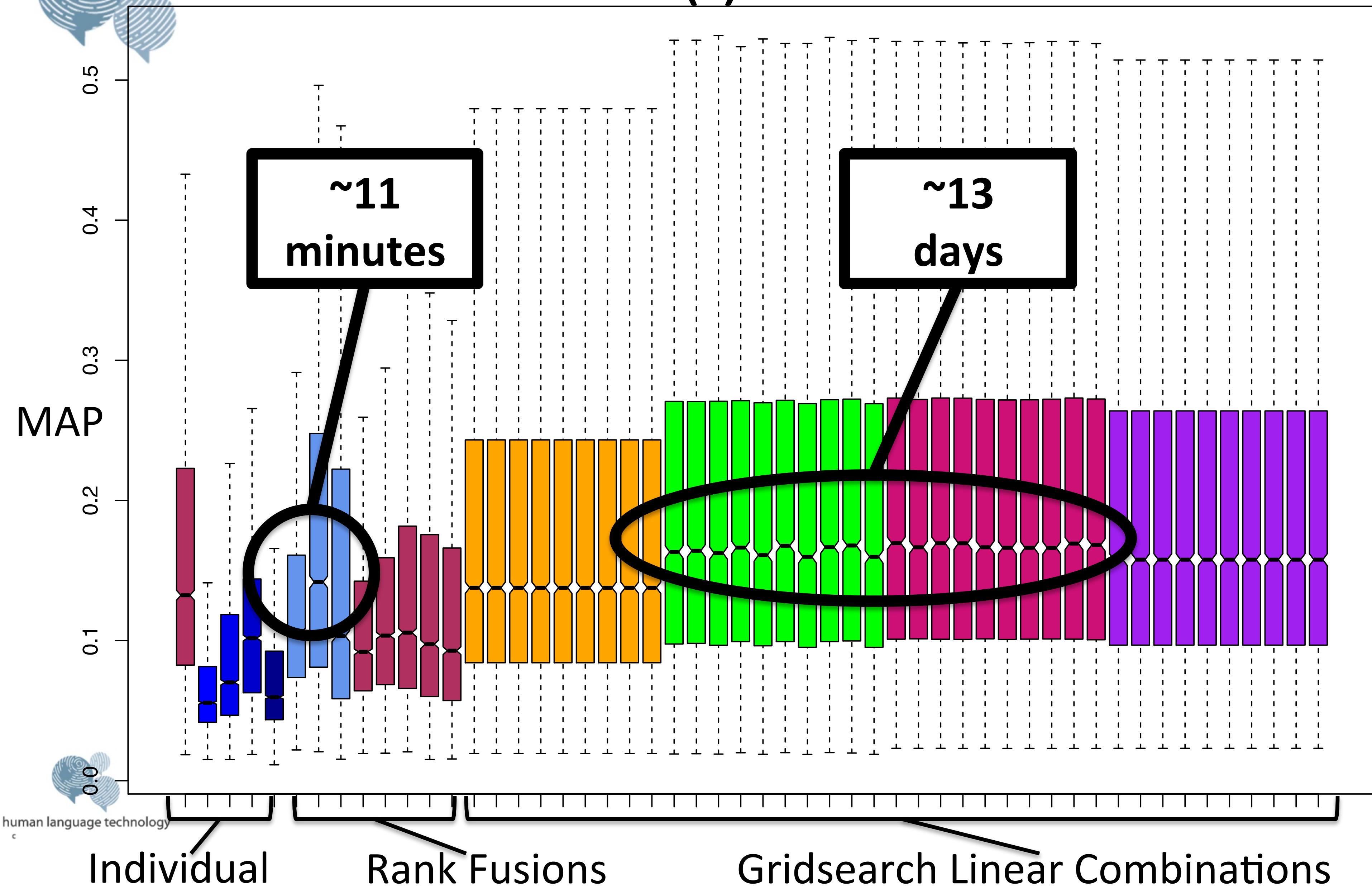


O(x^n) in number of Analytics!!



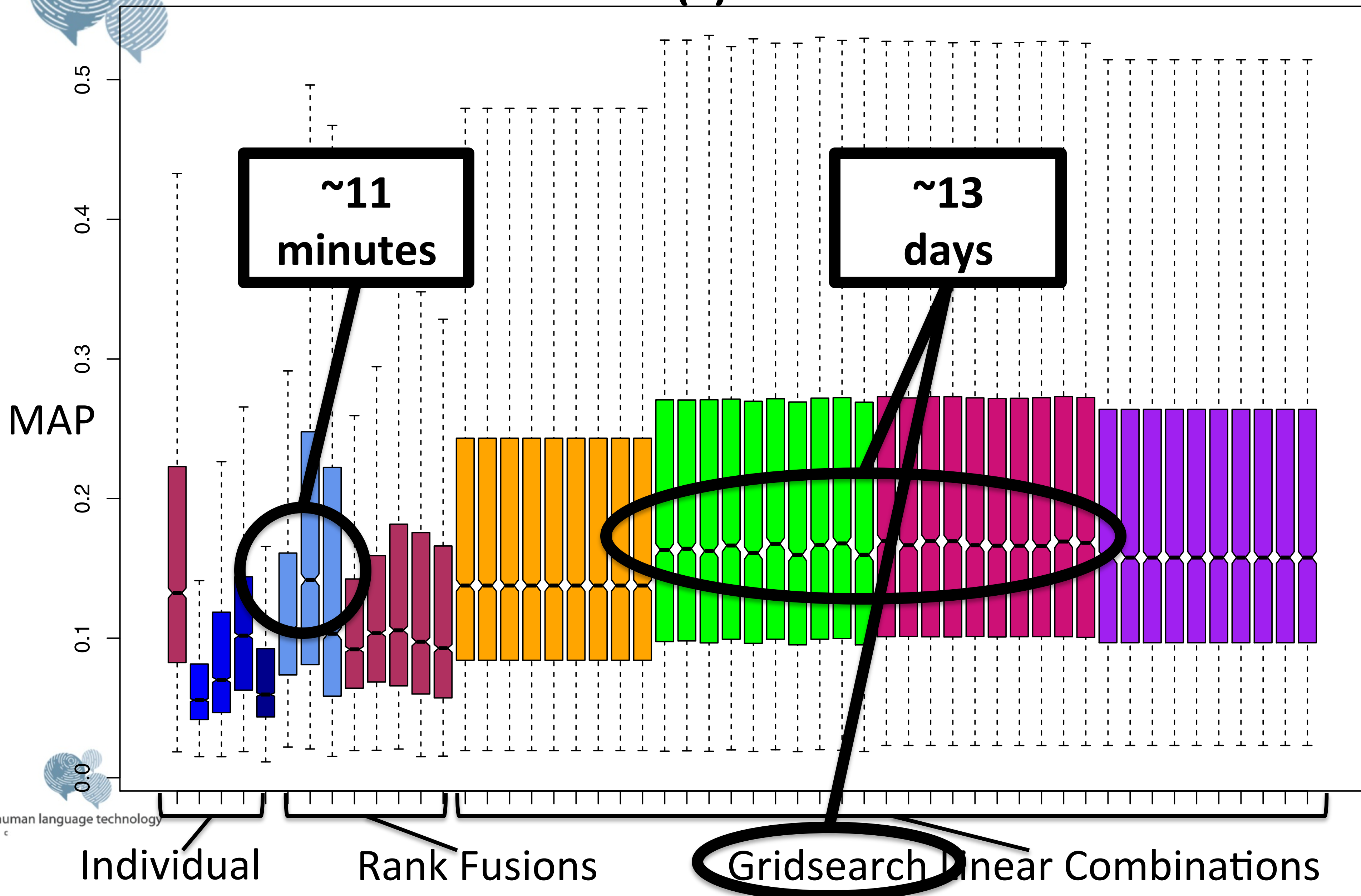
human language technology

Nevermind O(.) for a moment...



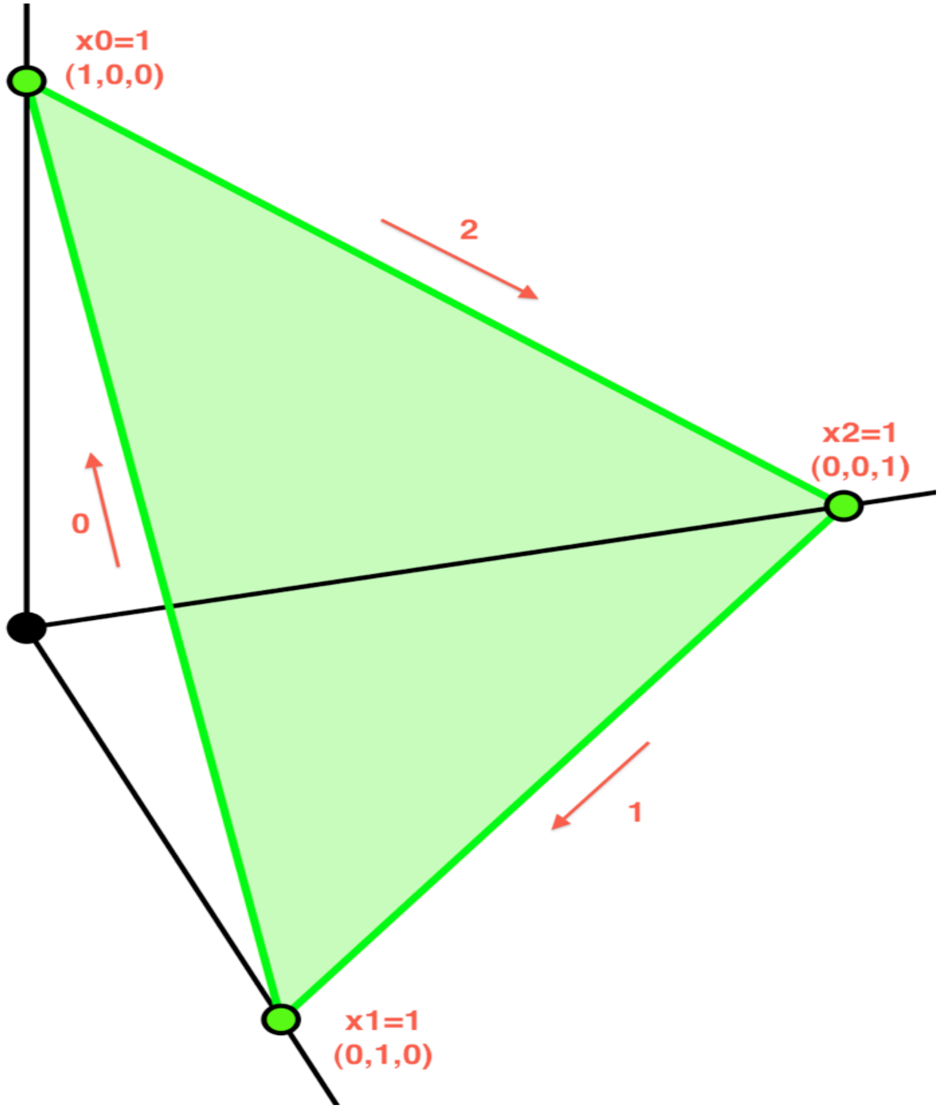
human language technology

Nevermind O(.) for a moment...



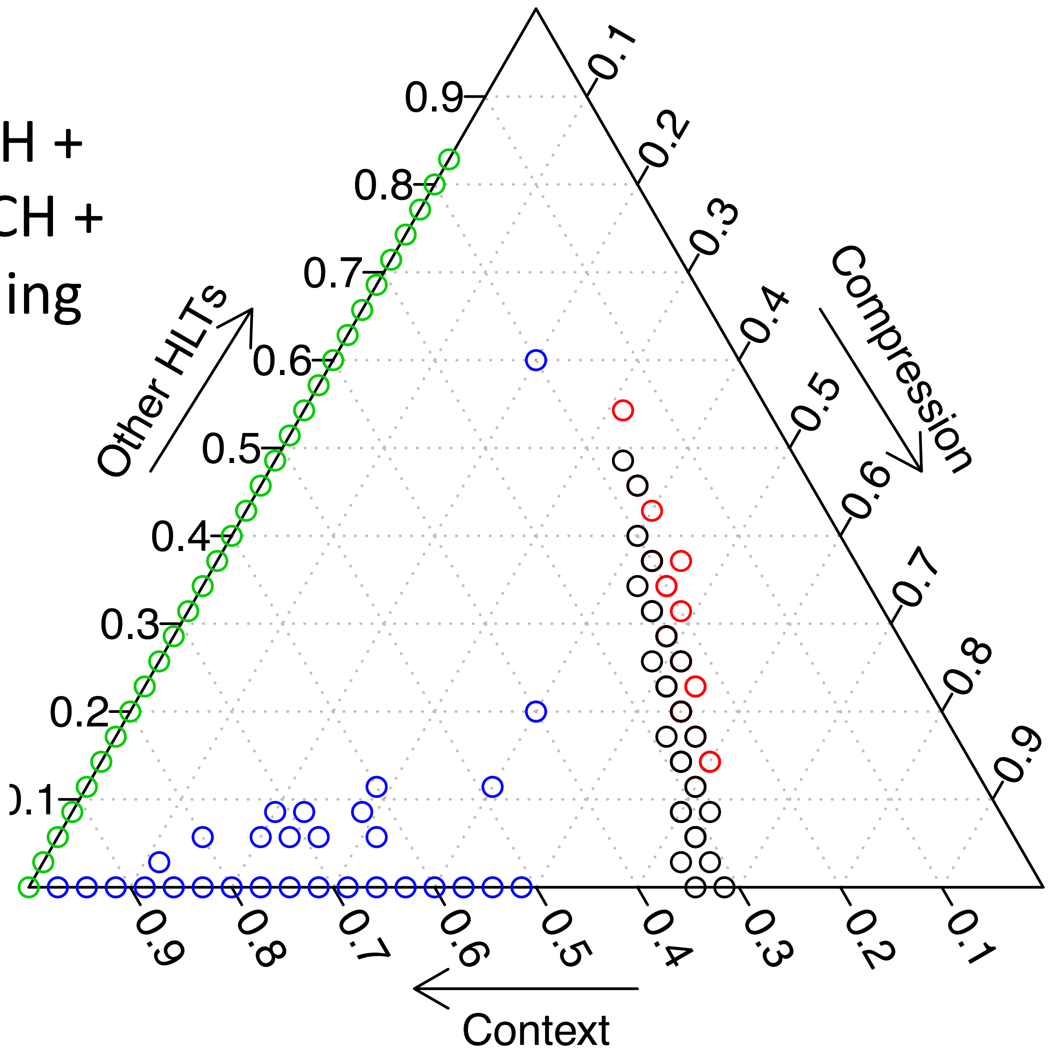
human language technology

Linear fusion – 3 Analytics

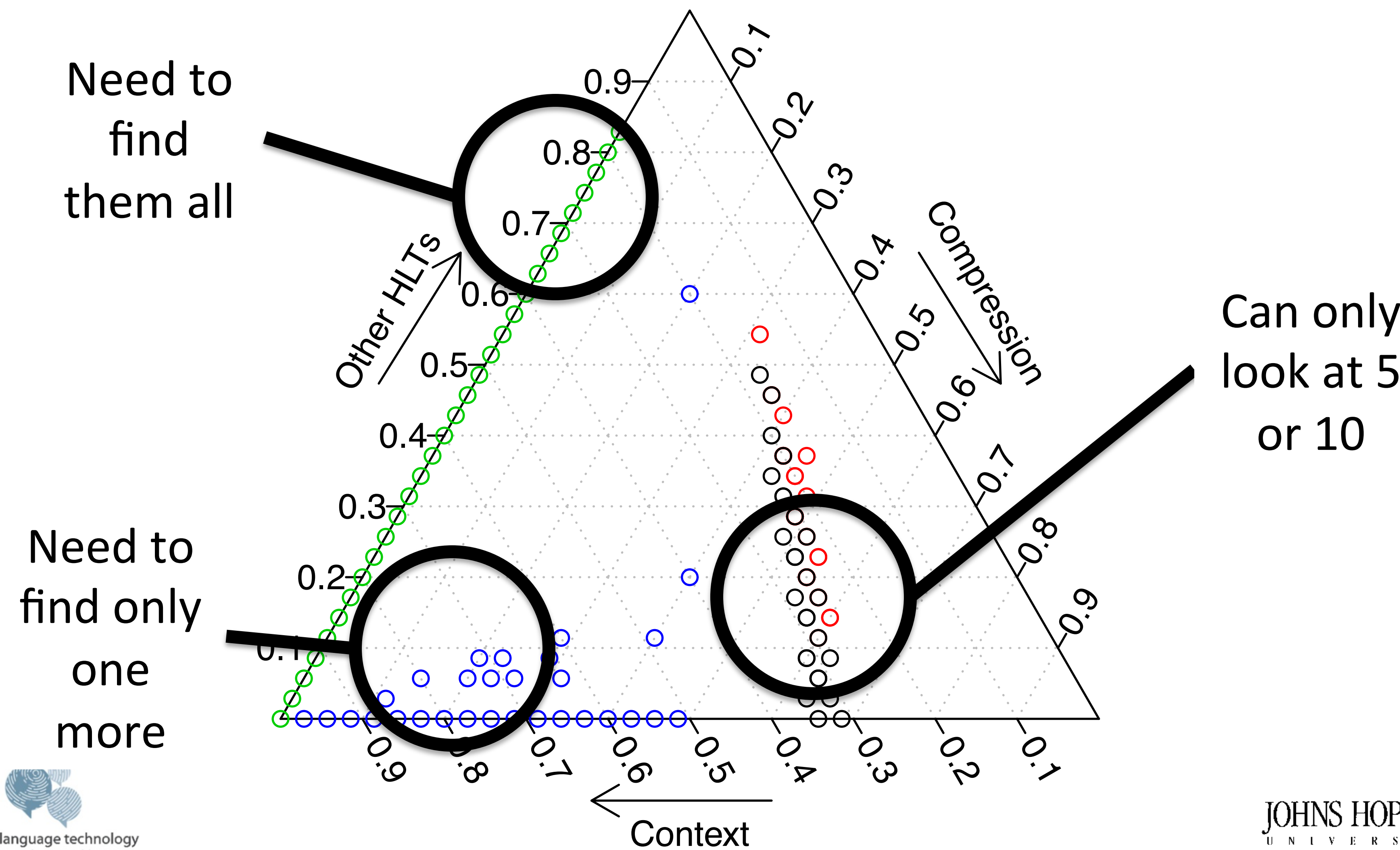


Which HLTs should I use?

Average WCH +
Minimum WCH +
Topic Modeling



Which HLTs should I use?



human language technology center of excellence

JOHNS HOPKINS UNIVERSITY

Outline

- Introduction
- Method
  - Importance Sampling
  - Evaluation
- Analytics – Content and Context
- Fusions
- Conclusions & Future Directions

Questions

- How should we fuse?
  - For small number of analytics – grid search linear
  - For large number of analytics – rank fuse (or think harder)

- What kind of HLTs should we use?
  - Depends on your inference task, of course.
  - We can actually recommend what to use!
  - Insight into relative strengths of HLTs exposed.

Future and Related Work

- This is post-hoc fusion of scores or ranks, what about more native fusion?

- Different data and inference tasks
  - ~~What movie should I watch? (Netflix)~~
  - Who should I cite? (ACL)
  - Who should I work with? (Github)
  - Vote prediction (Congressional Bills)

Collaborators

• Carey Priebe	[JHU HLTCOE]
• Allen Gorin	[JHU HLTCOE]
• Richard Cox	[JHU HLTCOE]
• David Marchette	[Navy]
• Yongser Park	[JHU CIS]
• Minh Tang	[JHU AMS]
• Michael Decerbo	[Raytheon BBN]
• Hanna Wallach	[UMass Amherst]
• Jim Mayfield	[JHU APL/HLTCOE]
• Paul McNamee	[JHU APL/HLTCOE]
• William Szewczyk	[DoD]

Collaborators

• Carey Priebe	[JHU HLTCOE]
• Allen Gorin	[JHU HLTCOE]
• Richard Cox	[JHU HLTCOE]
• David Marchette	[Navy]
• Yongser Park	[JHU CIS]
• Minh Tang	[JHU AMS]
• Michael Decerbo	[Raytheon BBN]
• Hanna Wallach	[UMass Amherst]
• Jim Mayfield	[JHU APL/HLTCOE]
• Paul McNamee	[JHU APL/HLTCOE]
• William Szewczyk	[DoD]

human language technology
center of excellence

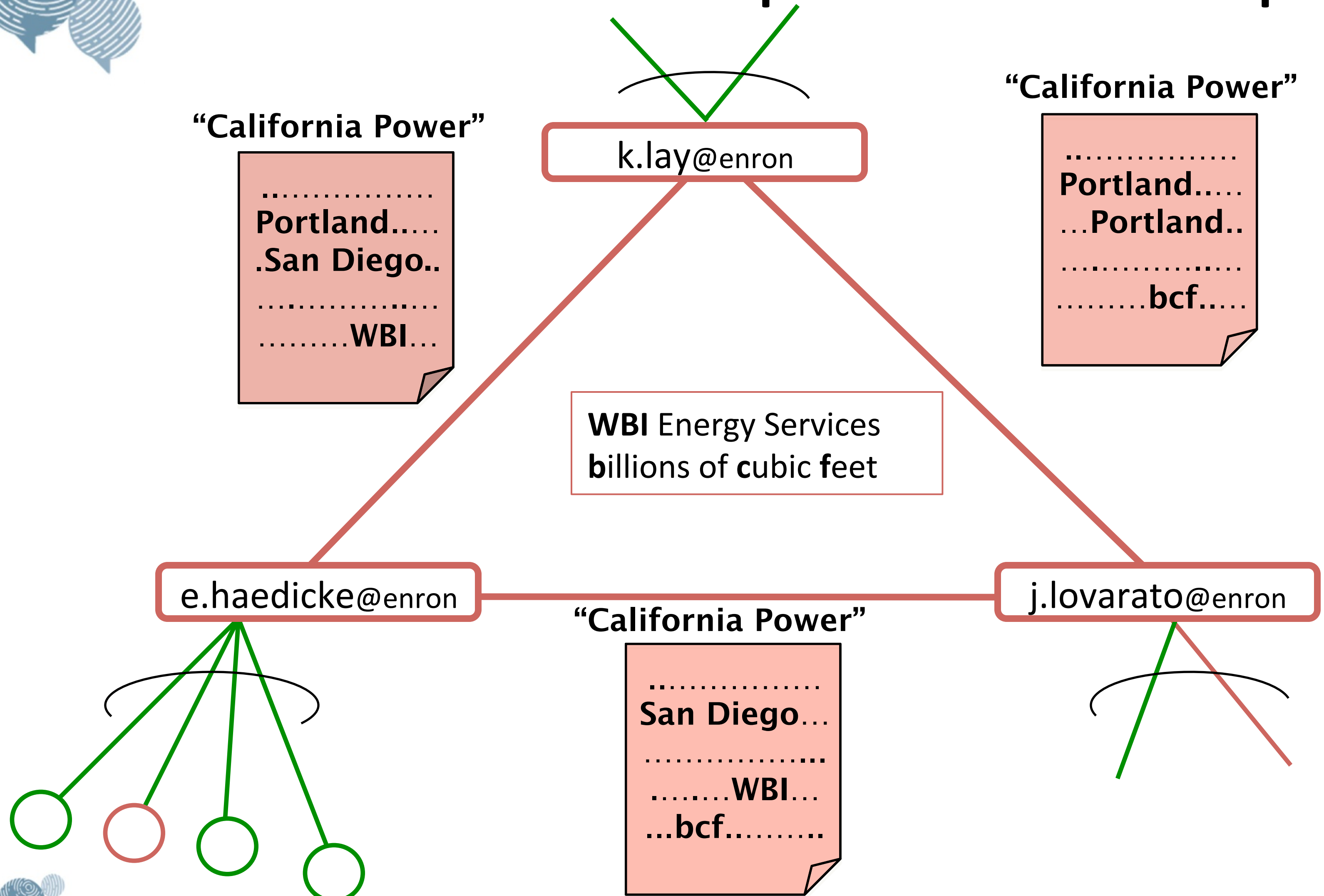
JOHNS HOPKINS
UNIVERSITY

Thank you.

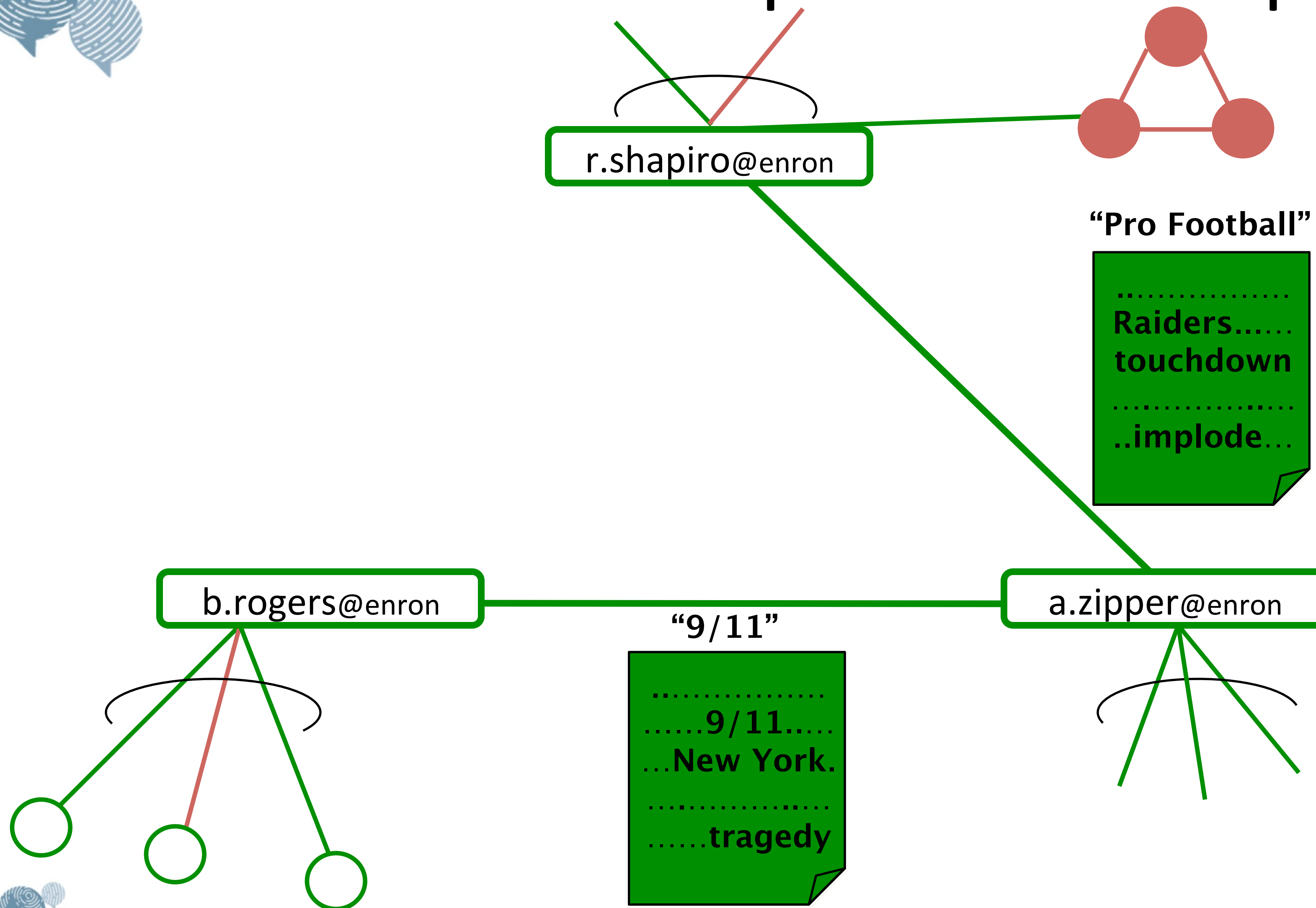
Coppersmith & Priebe (submitted): [arxiv.org/1201.4118]

Latest papers (often) available from glencoppersmith.com

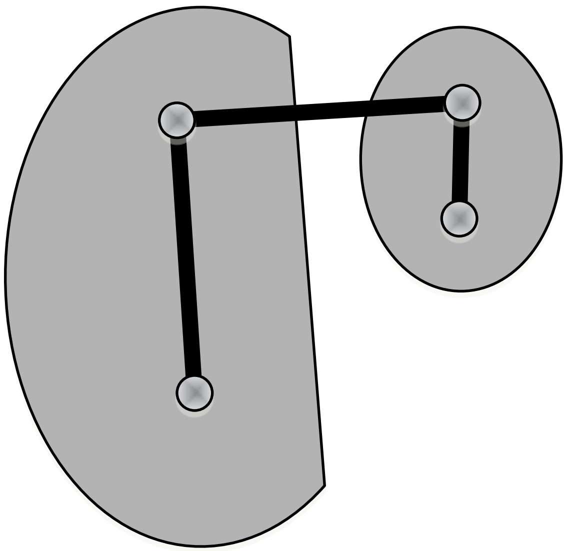
Content of Importance Sample

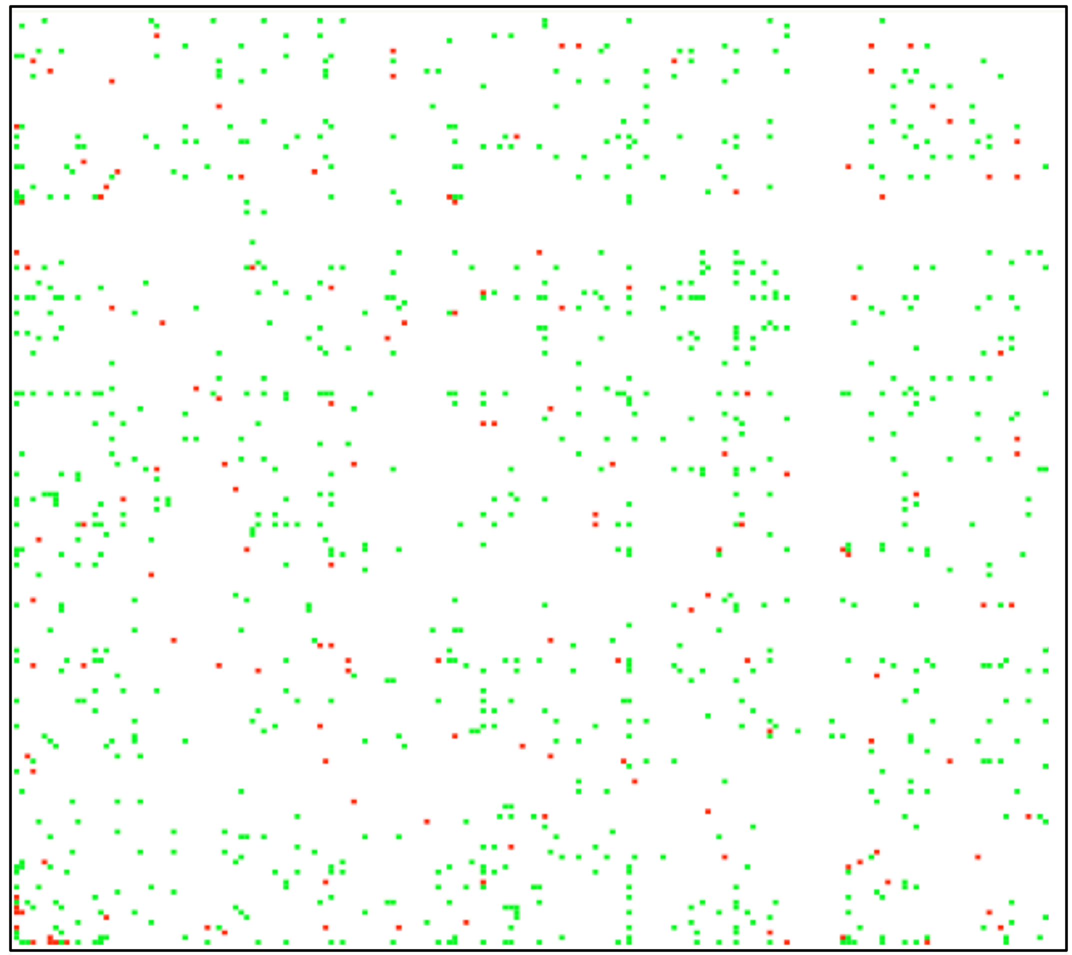


Content of Importance Sample



Context of Importance Sample

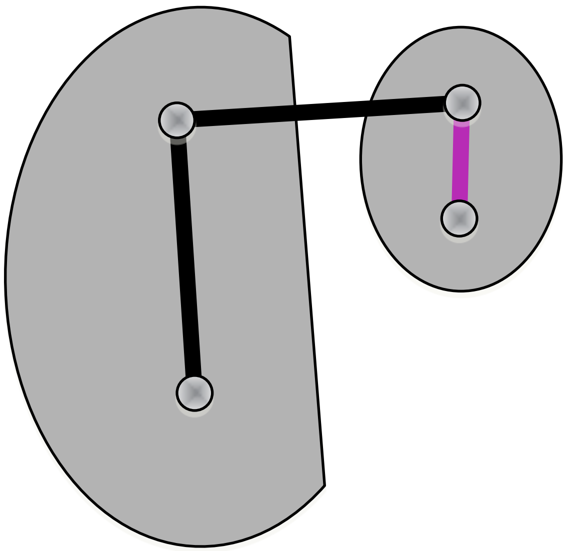


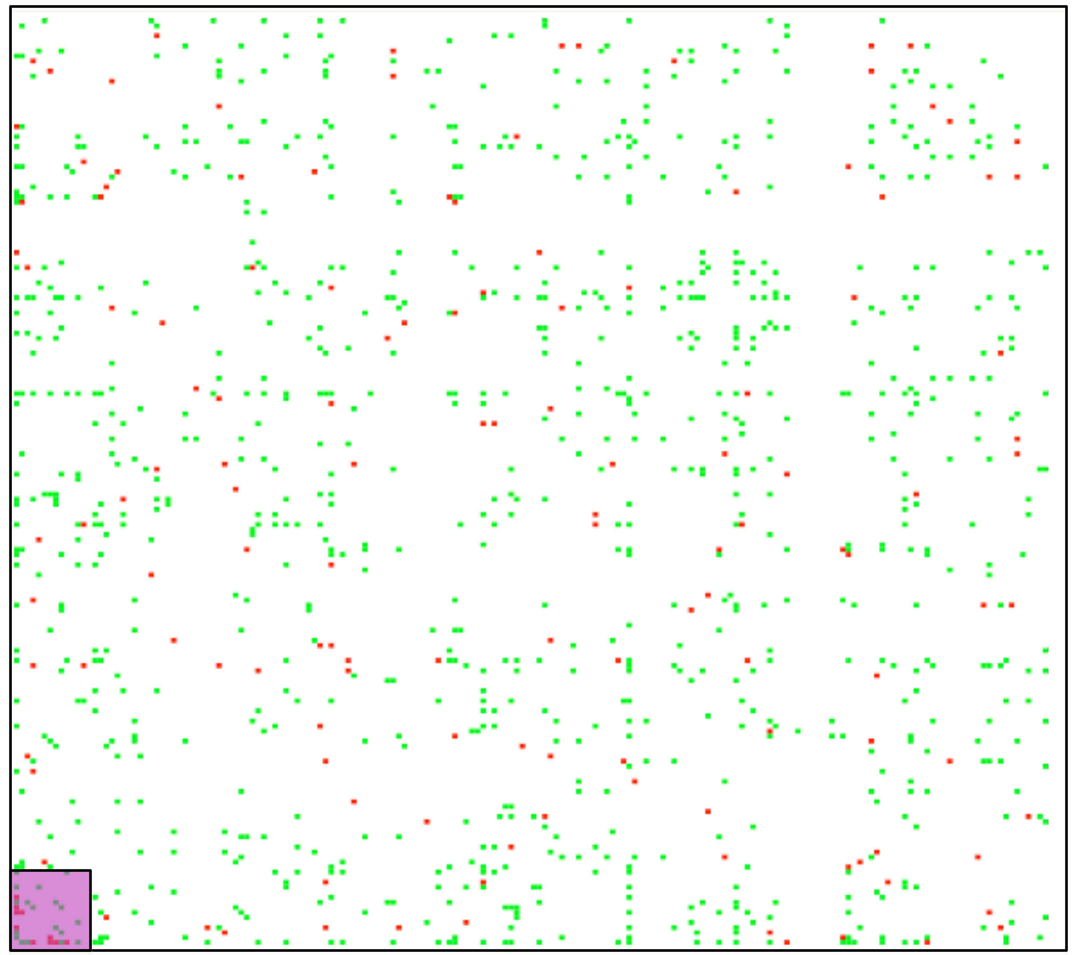


human language technology
center of excellence

JOHNS HOPKINS
UNIVERSITY

Context of Importance Sample



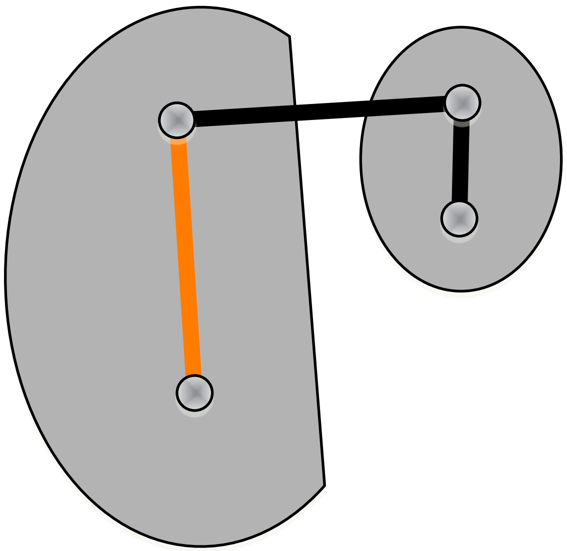


Observed s_1
= **0.45**

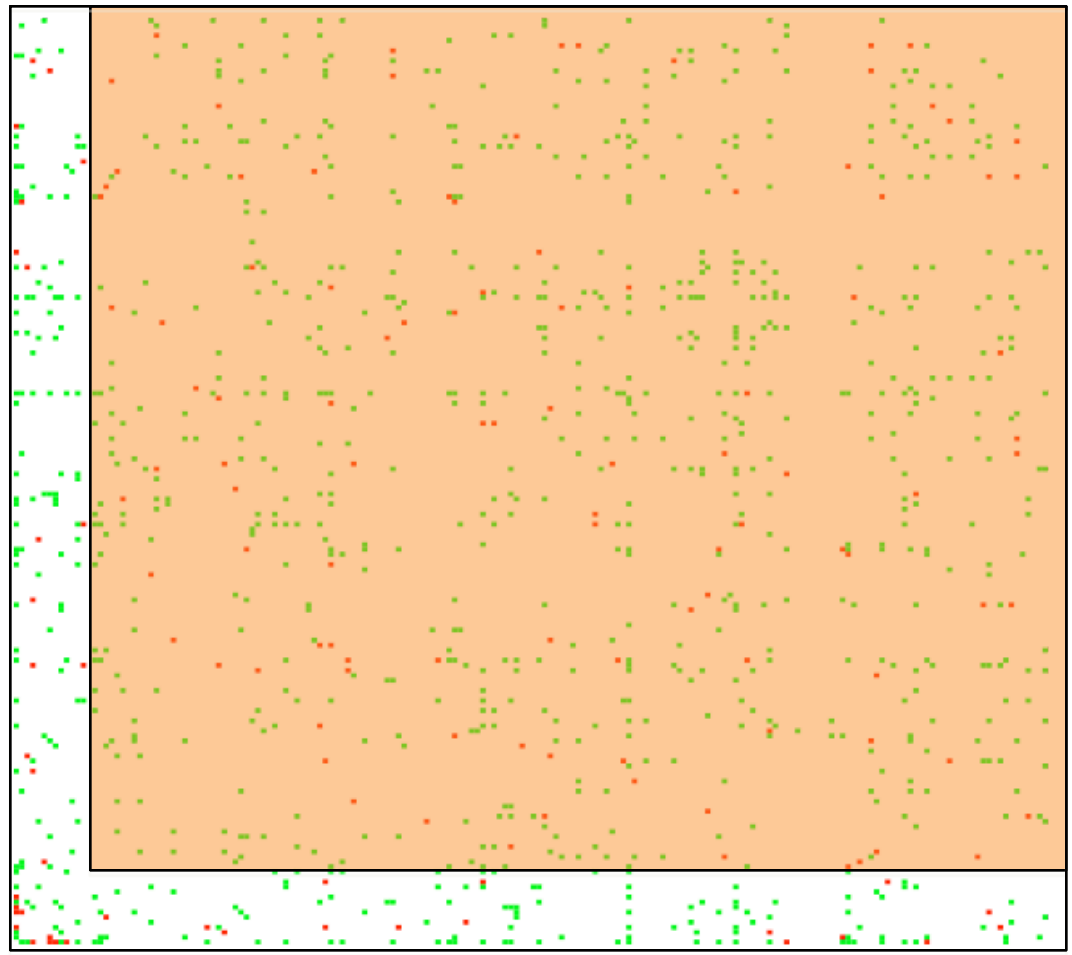
human language technology
center of excellence

JOHNS HOPKINS
UNIVERSITY

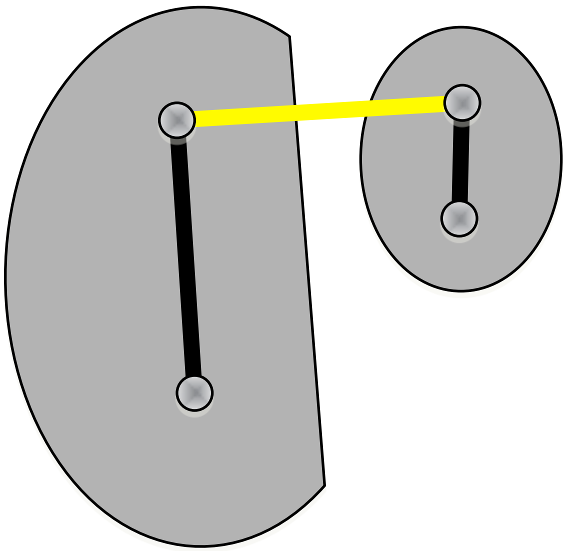
Context of Importance Sample

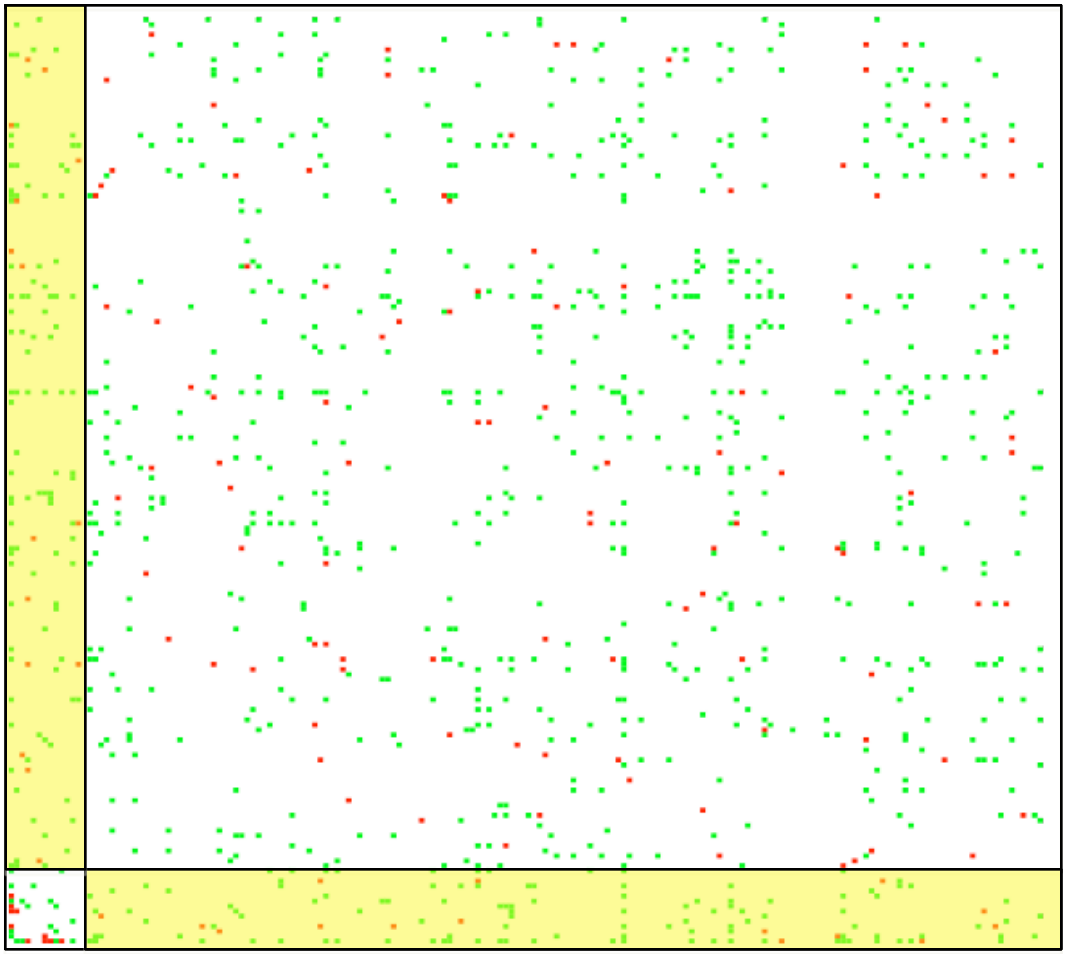


Observed p_1= **0.13**



Context of Importance Sample



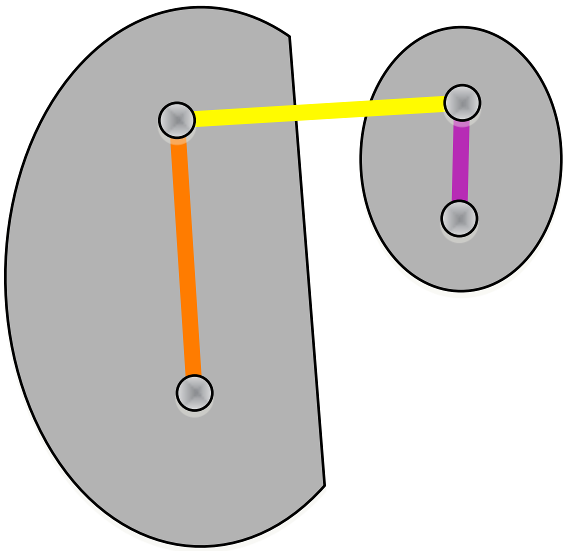


Observed p_1
$= \mathbf{0.12}$

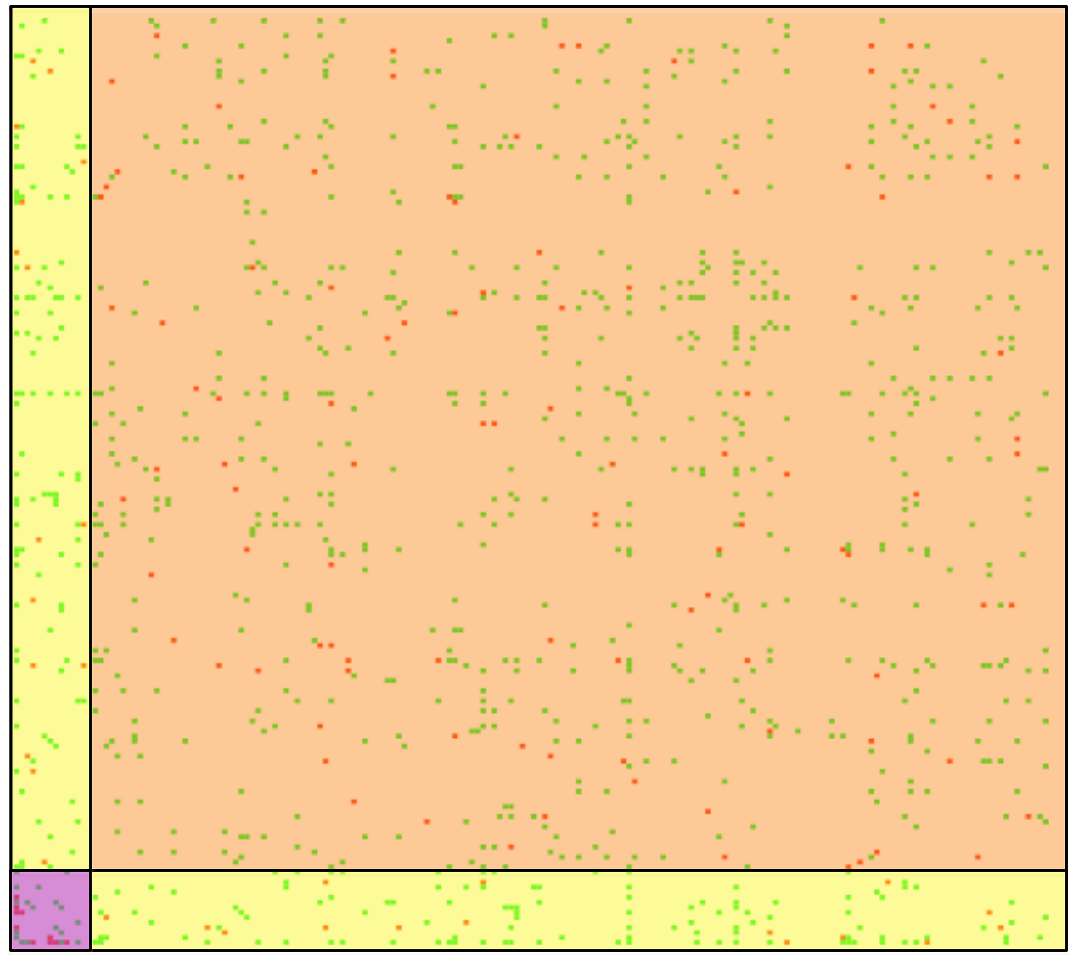
human language technology
center of excellence

JOHNS HOPKINS
UNIVERSITY

Context of Importance Sample



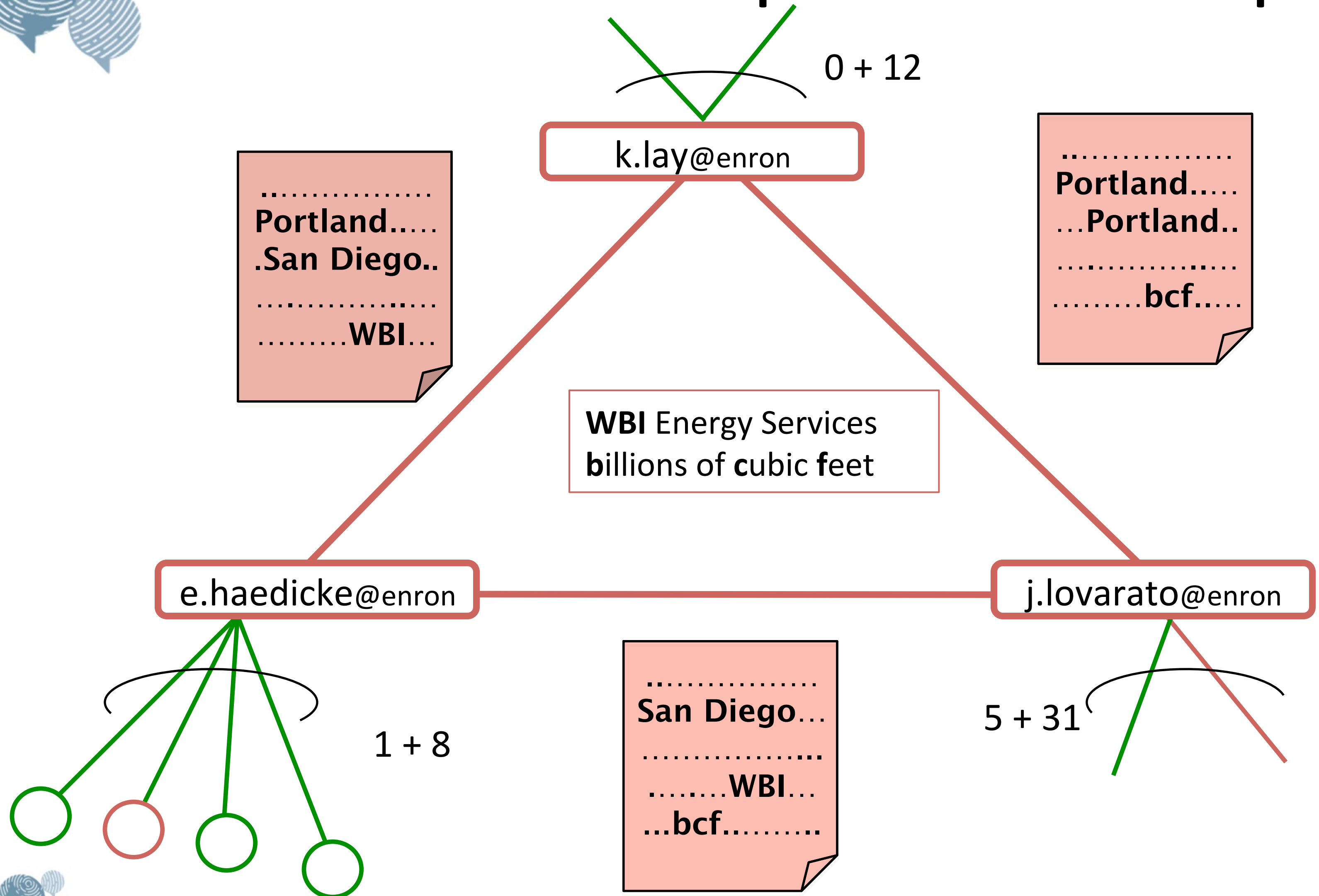
Observed p_1= **0.13**



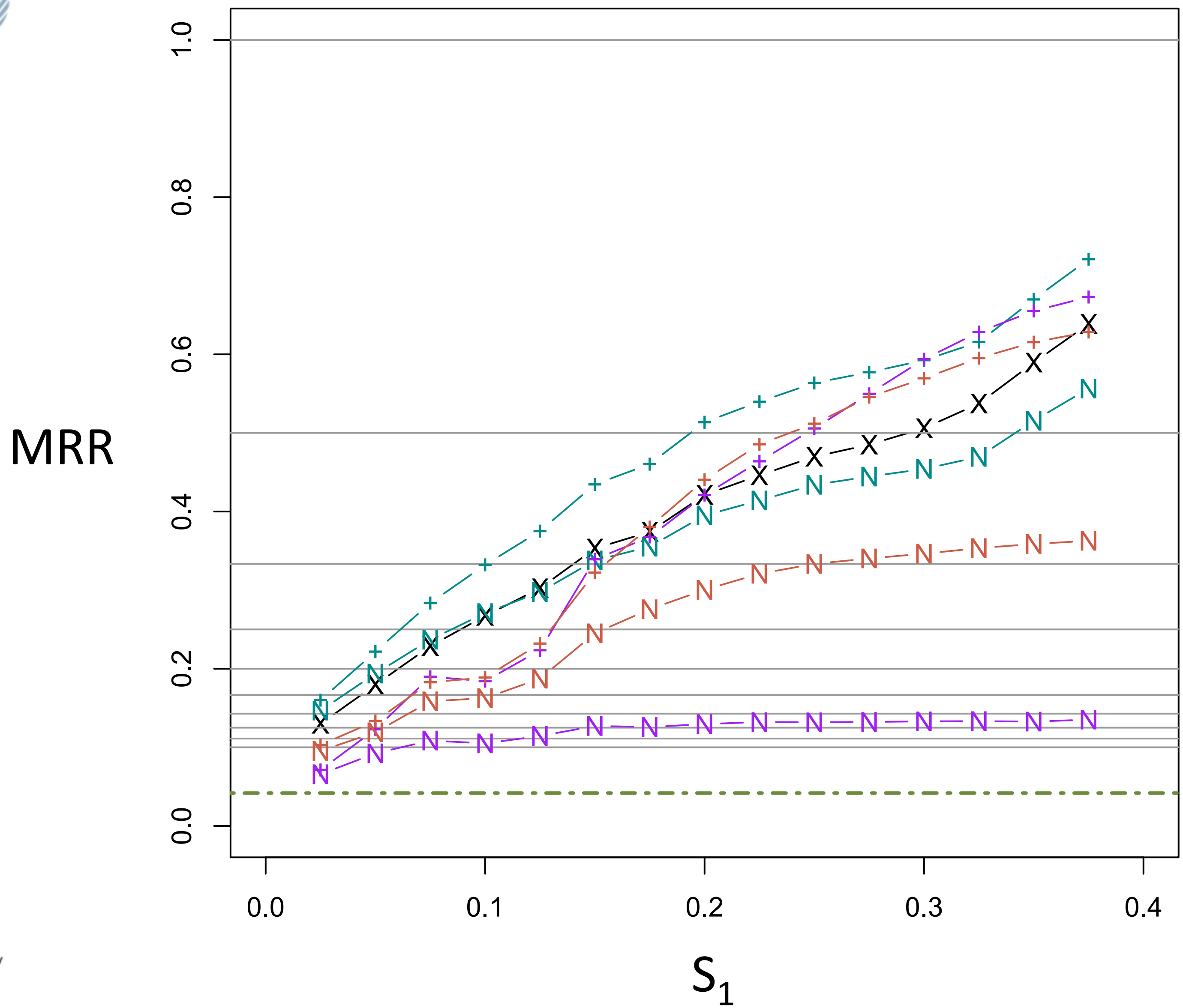
Observed s_1 = **0.45**

Observed p_1 = **0.12**

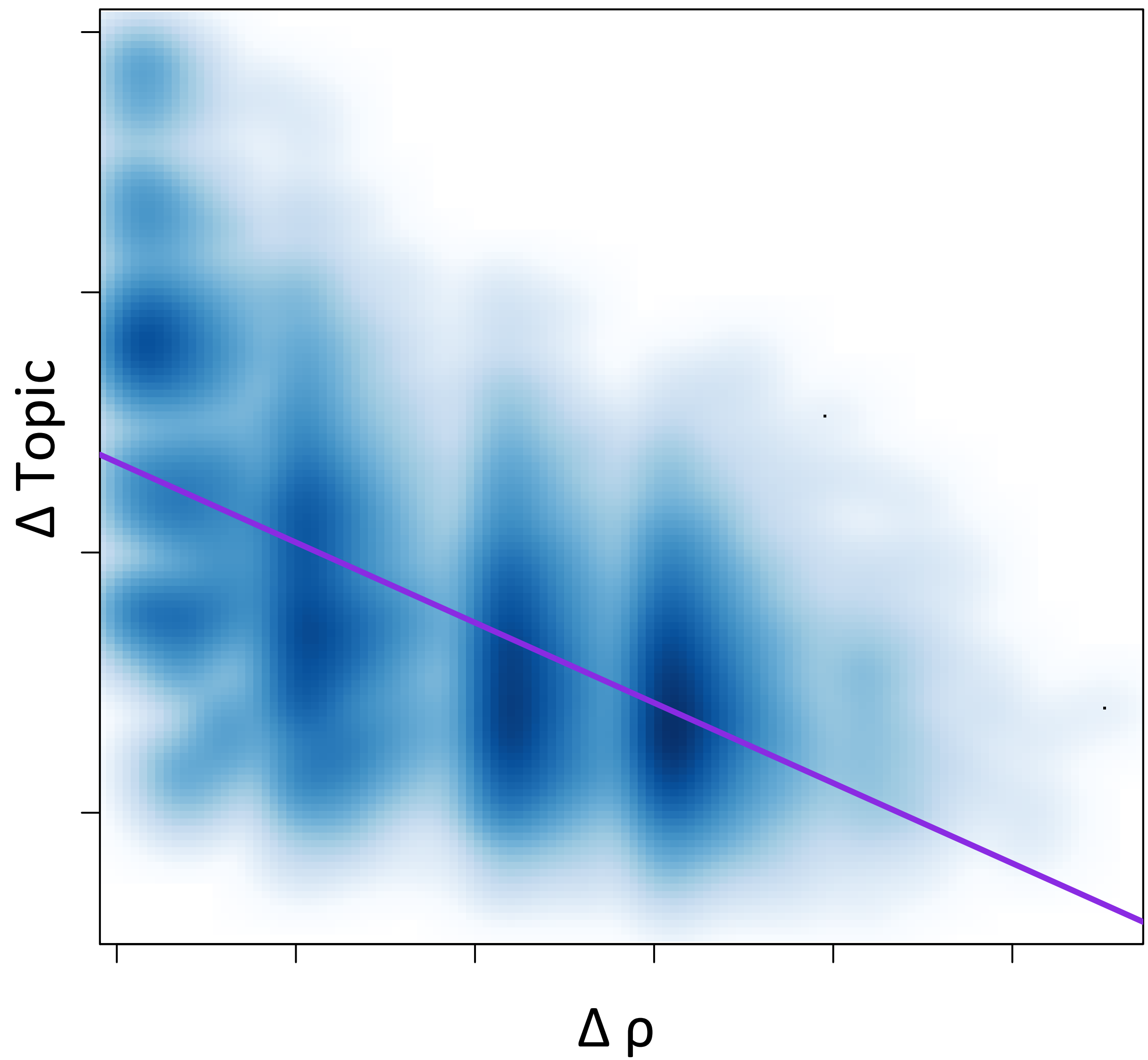
Content of Importance Sample



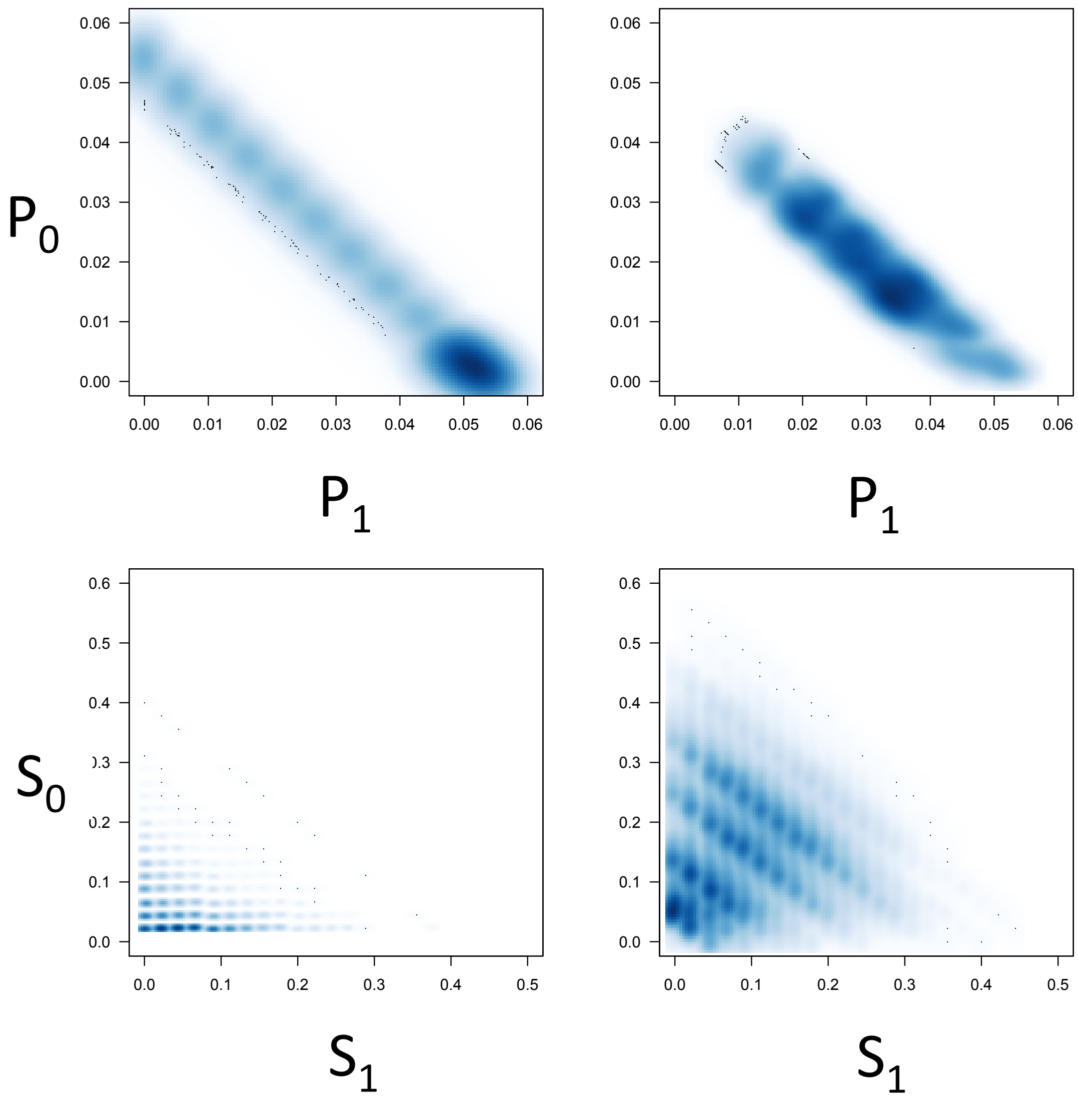
Comparing HLT Methods



Importance Sampled Joint



Injection ≠ Importance



human language technology
center of excellence

JOHNS HOPKINS
UNIVERSITY