Hash Kernels & Linear Models

Carnegie Mellon University

Vector Space Model for Text

A Vector Space Model for Automatic Indexing

G. Salton, A. Wong
and C. S. Yang
Cornell University

Communications
of
the ACM

November 1975
Volume 18
Number 11

In a document retrieval, or other pattern matching environment where stored entities (documents) are compared with each other or with incoming patterns (search requests), it appears that the best indexing (property) space is one where each entity lies as far away from the others as possible; in these circumstances the value of an indexing system may be expressible as a function of the density of the object space; in particular, retrieval performance may correlate inversely with space density. An approach based on space density computations is used to choose an optimum indexing vocabulary for a collection of documents. Typical evaluation results are shown, demonstating the usefulness of the model.

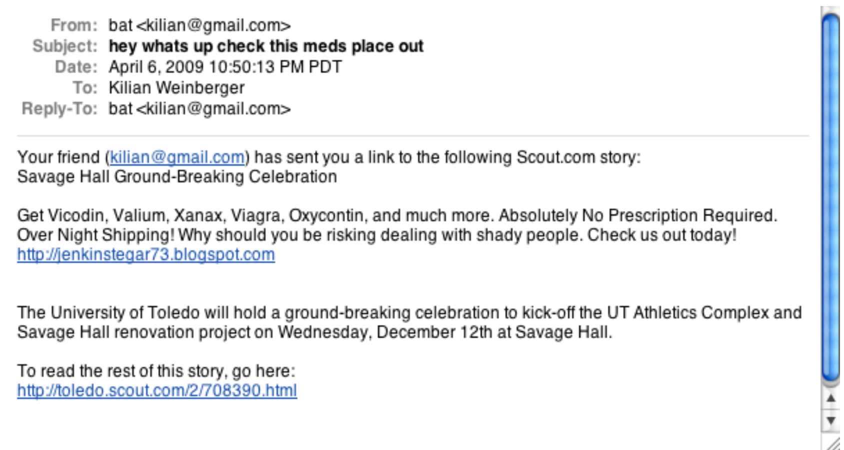
Linear functions

- Function

$$f(x) = \langle w, x \rangle + b = \sum_i w_i x_i + b$$

- Used for spam filtering, internet advertising, ranking, retrieval, summarization, gene finding, stock prediction, user profiling, ...
- Good as long as we don't have more than 1 billion parameters (8GB of memory for `double` precision)
- What if we have more parameters?
  - Lasso (with distributed optimization)
  - Hashing

Personalized Spam Filtering



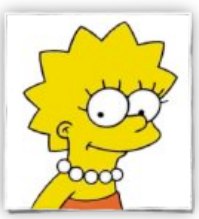
From: bat <kilian@gmail.com>
Subject: **hey whats up check this meds place out**
Date: April 6, 2009 10:50:13 PM PDT
To: Kilian Weinberger
Reply-To: bat <kilian@gmail.com>

Your friend (kilian@gmail.com) has sent you a link to the following Scout.com story:
Savage Hall Ground-Breaking Celebration

Get Vicodin, Valium, Xanax, Viagra, Oxycontin, and much more. Absolutely No Prescription Required. Over Night Shipping! Why should you be risking dealing with shady people. Check us out today!
http://jenkinstegar73.blogspot.com

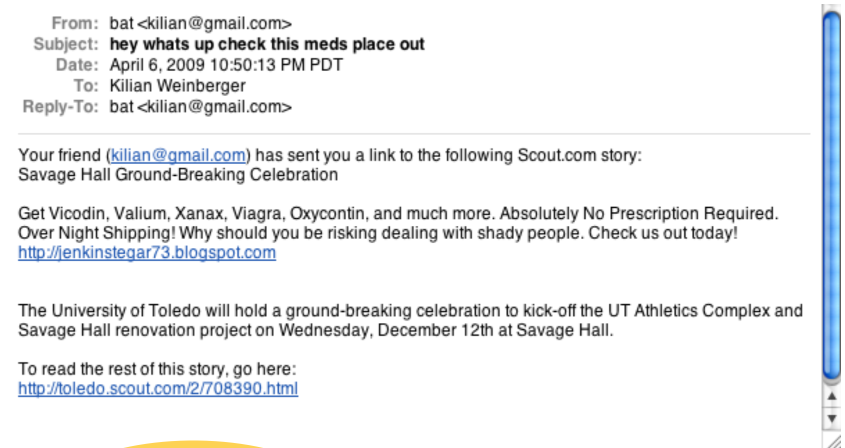
The University of Toledo will hold a ground-breaking celebration to kick-off the UT Athletics Complex and Savage Hall renovation project on Wednesday, December 12th at Savage Hall.

To read the rest of this story, go here:
http://toledo.scout.com/2/708390.html





Personalized Spam Filtering



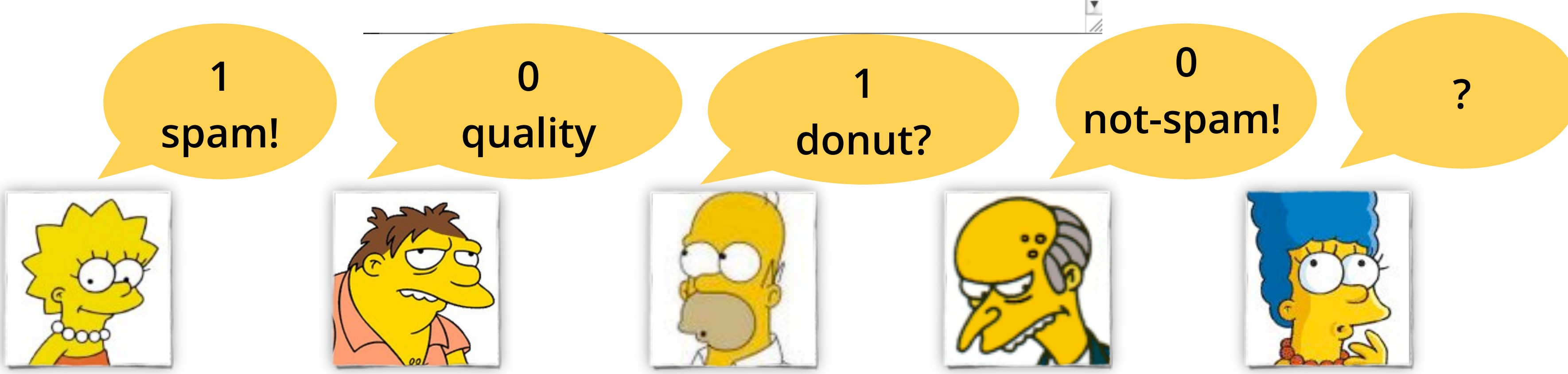
From: bat <kilian@gmail.com>
Subject: **hey whats up check this meds place out**
Date: April 6, 2009 10:50:13 PM PDT
To: Kilian Weinberger
Reply-To: bat <kilian@gmail.com>

Your friend (kilian@gmail.com) has sent you a link to the following Scout.com story:
Savage Hall Ground-Breaking Celebration

Get Vicodin, Valium, Xanax, Viagra, Oxycontin, and much more. Absolutely No Prescription Required. Over Night Shipping! Why should you be risking dealing with shady people. Check us out today!
http://jenkinstegar73.blogspot.com

The University of Toledo will hold a ground-breaking celebration to kick-off the UT Athletics Complex and Savage Hall renovation project on Wednesday, December 12th at Savage Hall.

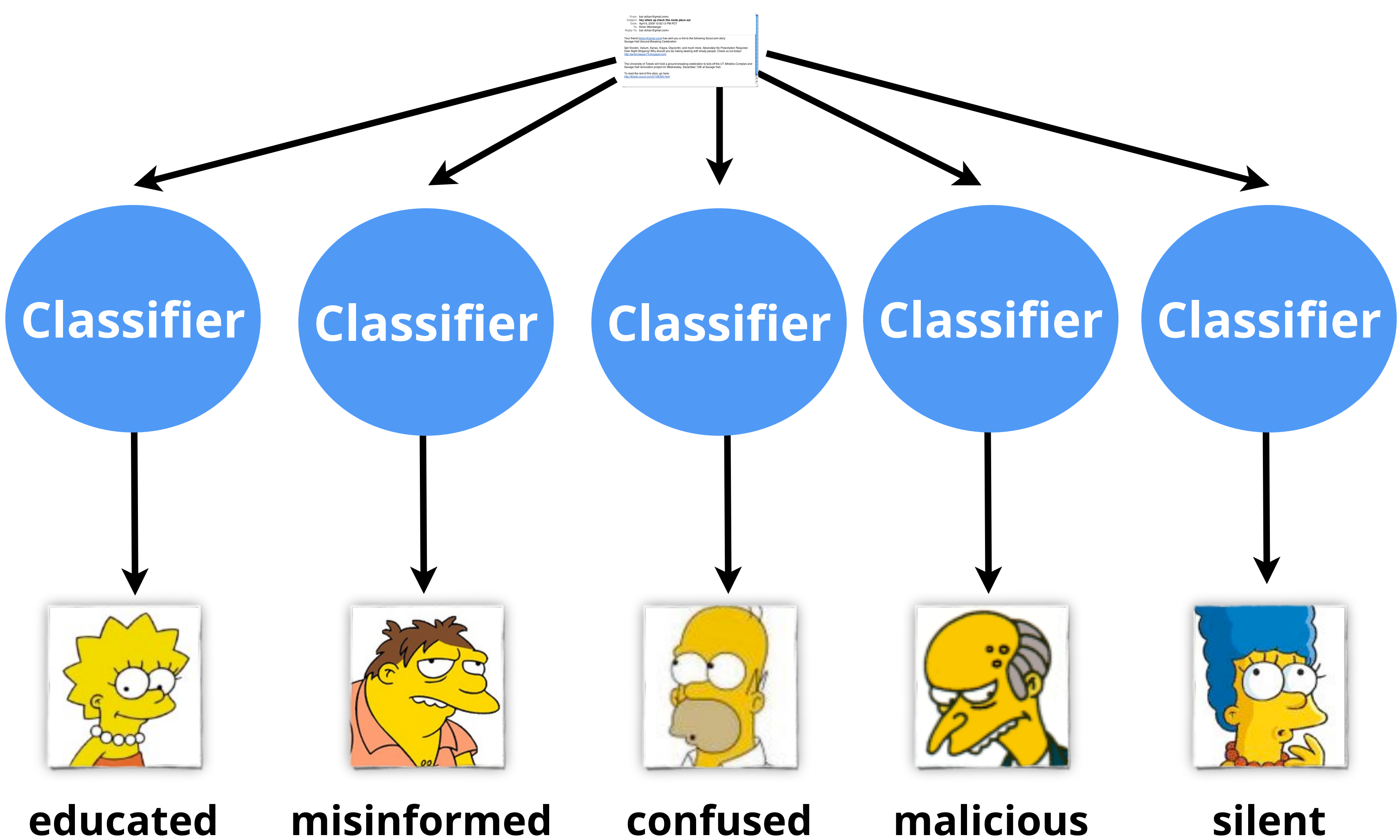
To read the rest of this story, go here:
http://toledo.scout.com/2/708390.html



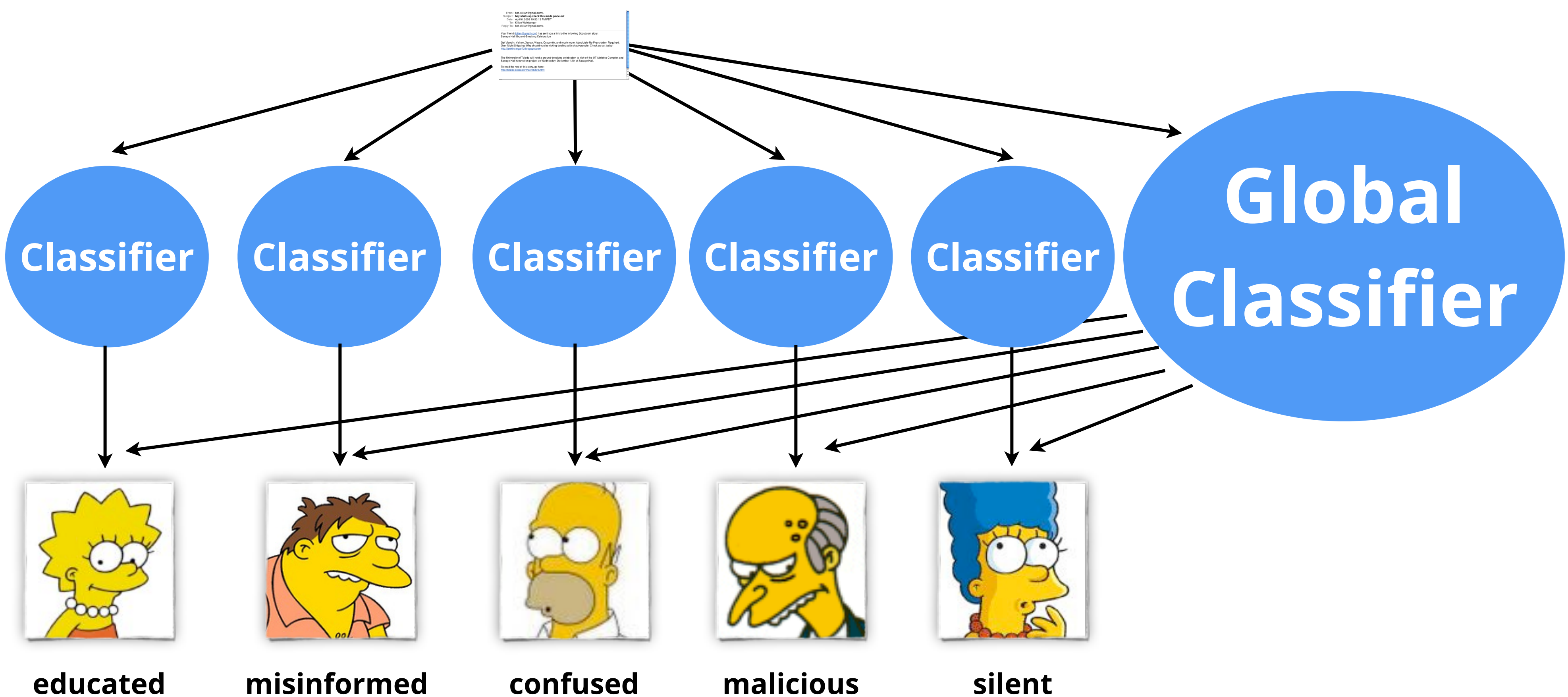
educated misinformed confused malicious silent

Carnegie Mellon University

Personalized Spam Filtering



Multitask Learning



Carnegie Mellon University

Collaborative Classification

- **Primal space representation**

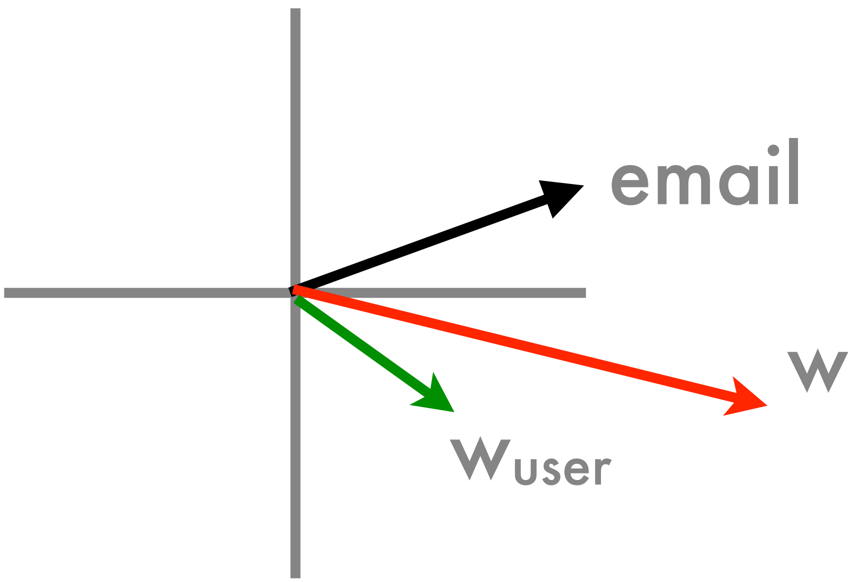
$$f(x,u) = \langle \phi(x), w \rangle + \langle \phi(x), w_u \rangle = \langle \phi(x) \otimes (1 \oplus e_u), w \rangle$$

  Kernel representation

$$k((x,u),(x',u')) = k(x,x')[1 + \delta_{u,u'}]$$

  Multitask kernel (e.g. Pontil & Michelli, Daume). Usually does not scale well ...
- **Problem -** dimensionality is 10^{13}. That is 40TB of space

Collaborative Classification



- **Primal space representation**

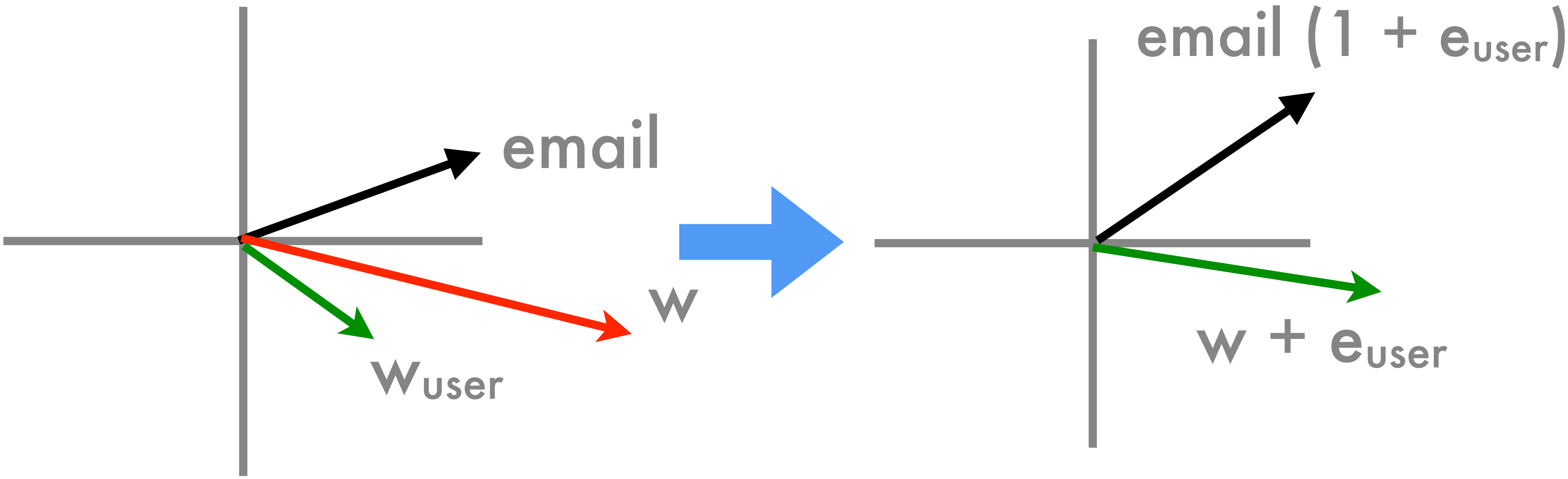
$$f(x,u) = \langle \phi(x), w \rangle + \langle \phi(x), w_u \rangle = \langle \phi(x) \otimes (1 \oplus e_u), w \rangle$$

  Kernel representation

$$k((x,u),(x',u')) = k(x,x')[1 + \delta_{u,u'}]$$

  Multitask kernel (e.g. Pontil & Michelli, Daume). Usually does not scale well ...
- **Problem -** dimensionality is 10^{13}. That is 40TB of space

Collaborative Classification



- **Primal space representation**

$$f(x,u) = \langle \phi(x), w \rangle + \langle \phi(x), w_u \rangle = \langle \phi(x) \otimes (1 \oplus e_u), w \rangle$$

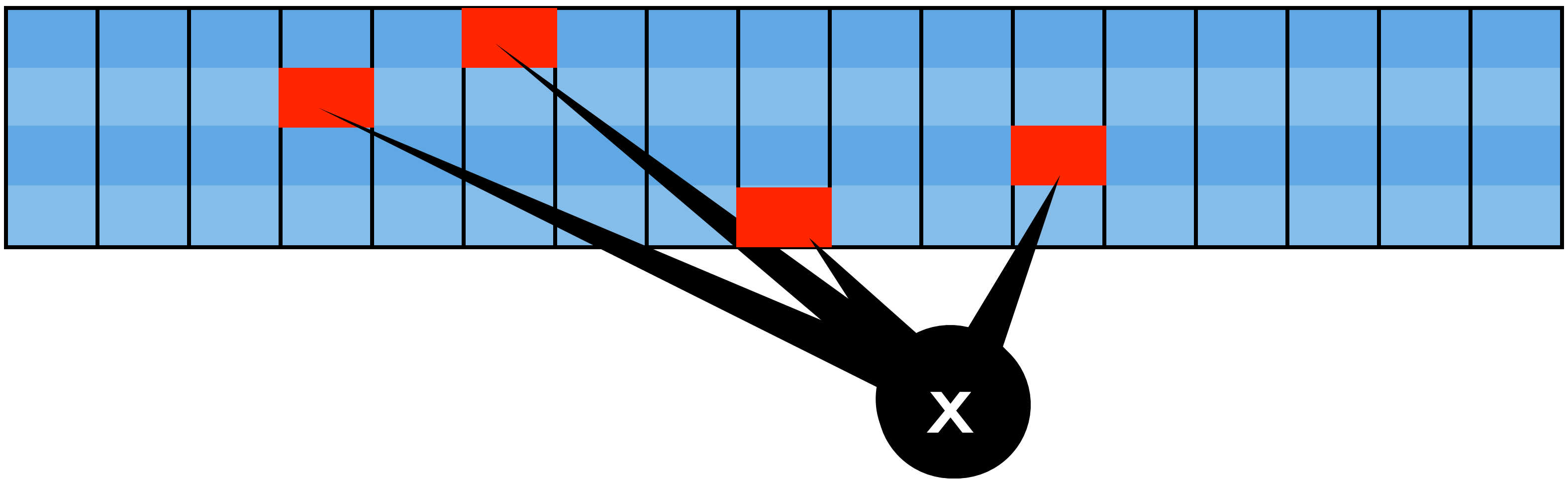
  Kernel representation

$$k((x,u),(x',u')) = k(x,x')[1 + \delta_{u,u'}]$$

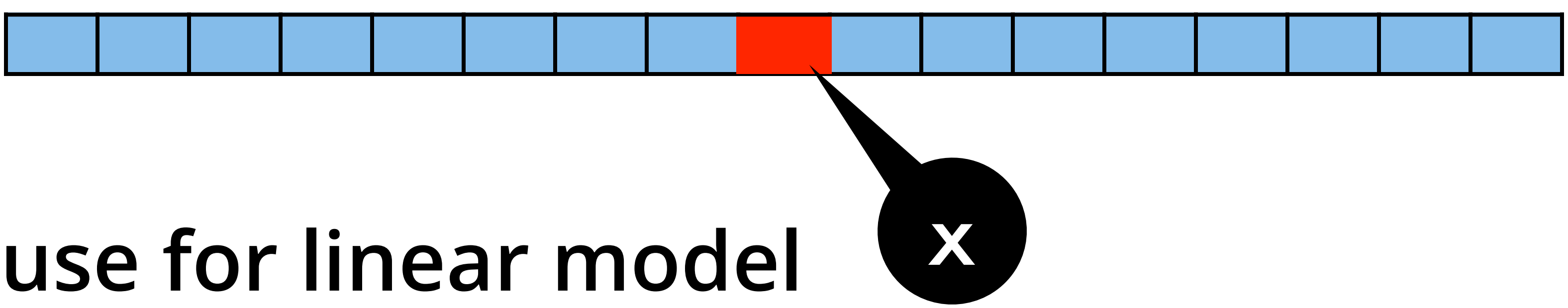
  Multitask kernel (e.g. Pontil & Michelli, Daume). Usually does not scale well ...
- **Problem -** dimensionality is 10^{13}. That is 40TB of space

CountMin Datastructure

- Count Sketch Basic



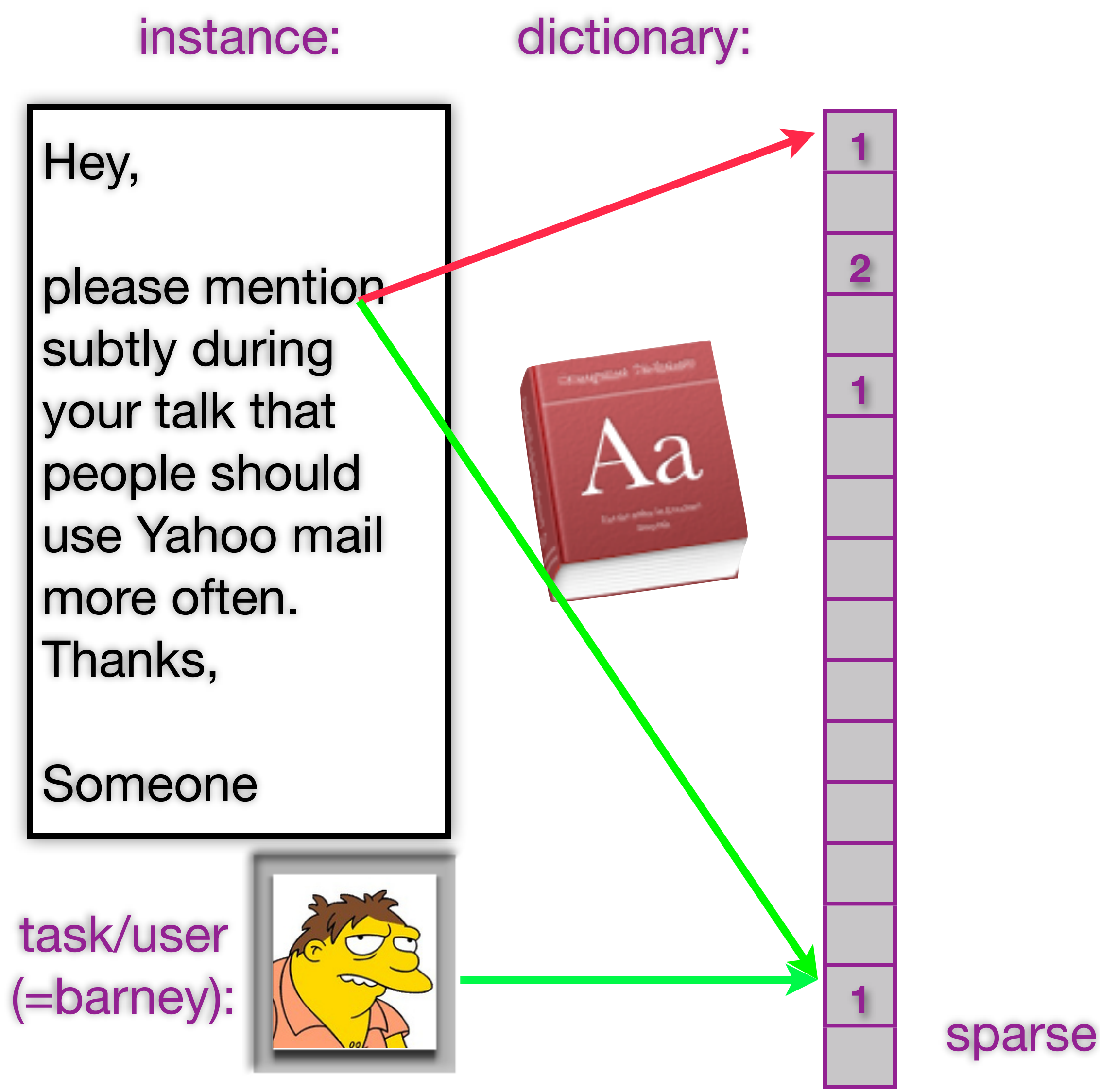
- Making it even more sparse



use for linear model

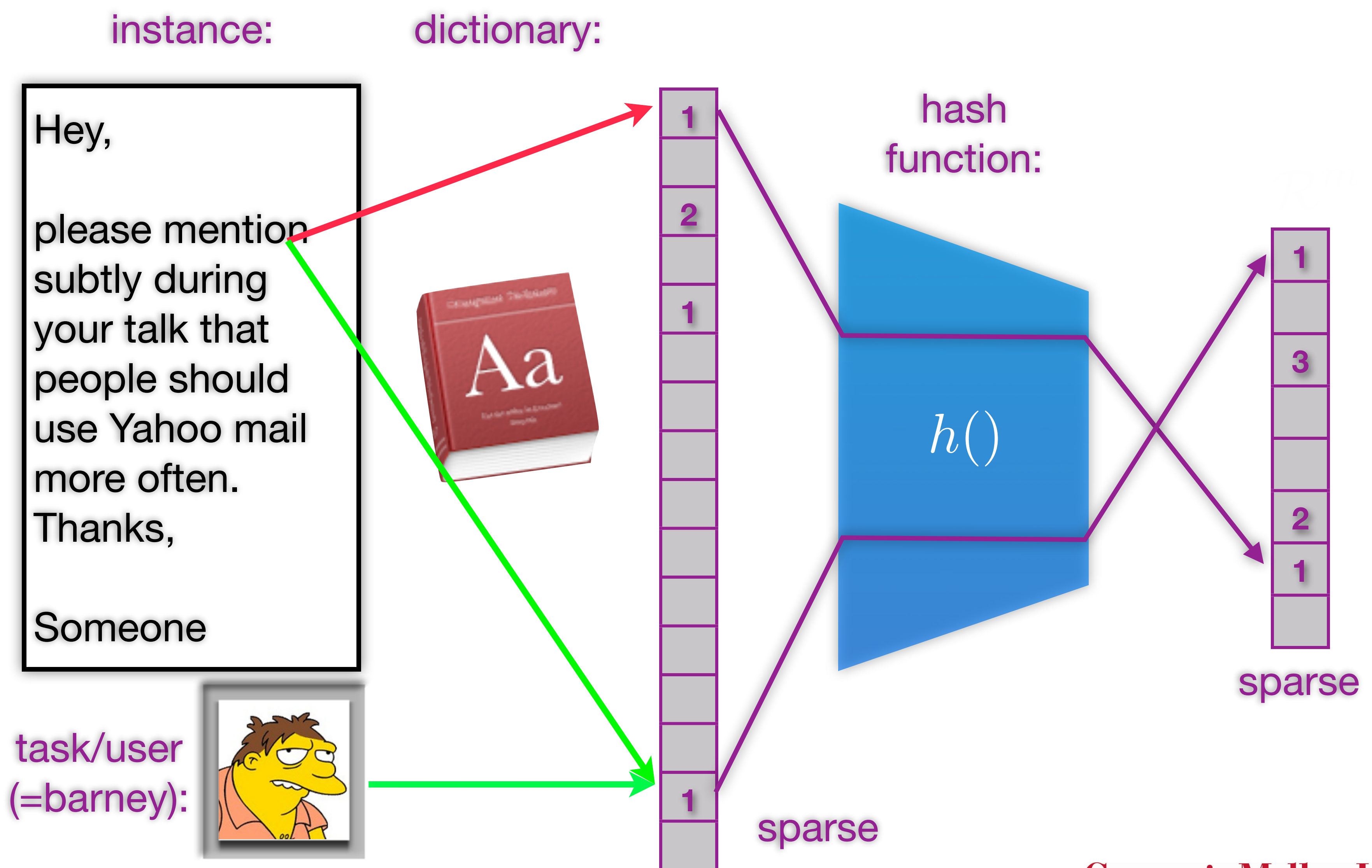
Hash Kernel

Hash Kernel

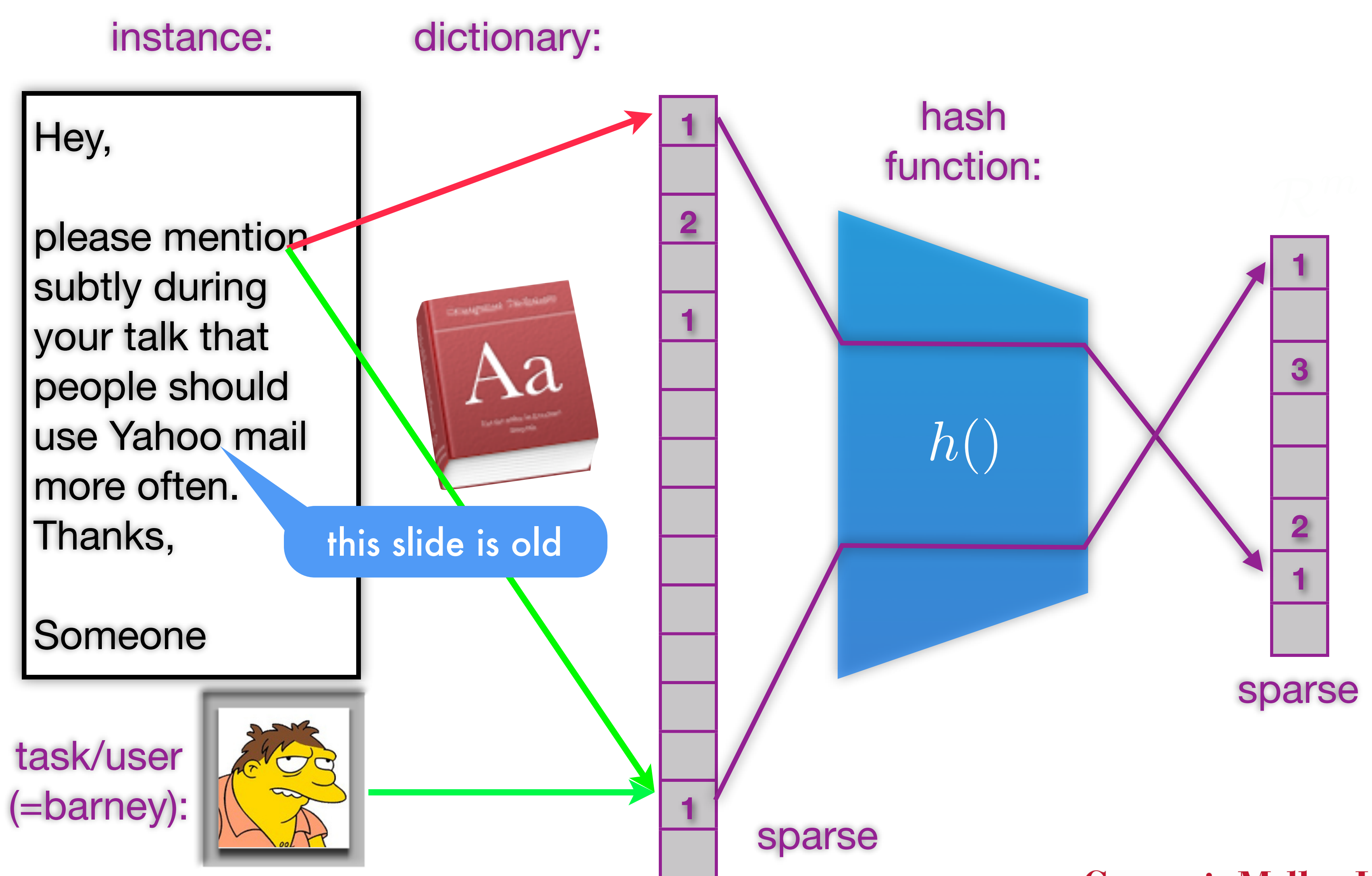


Carnegie Mellon University

Hash Kernel

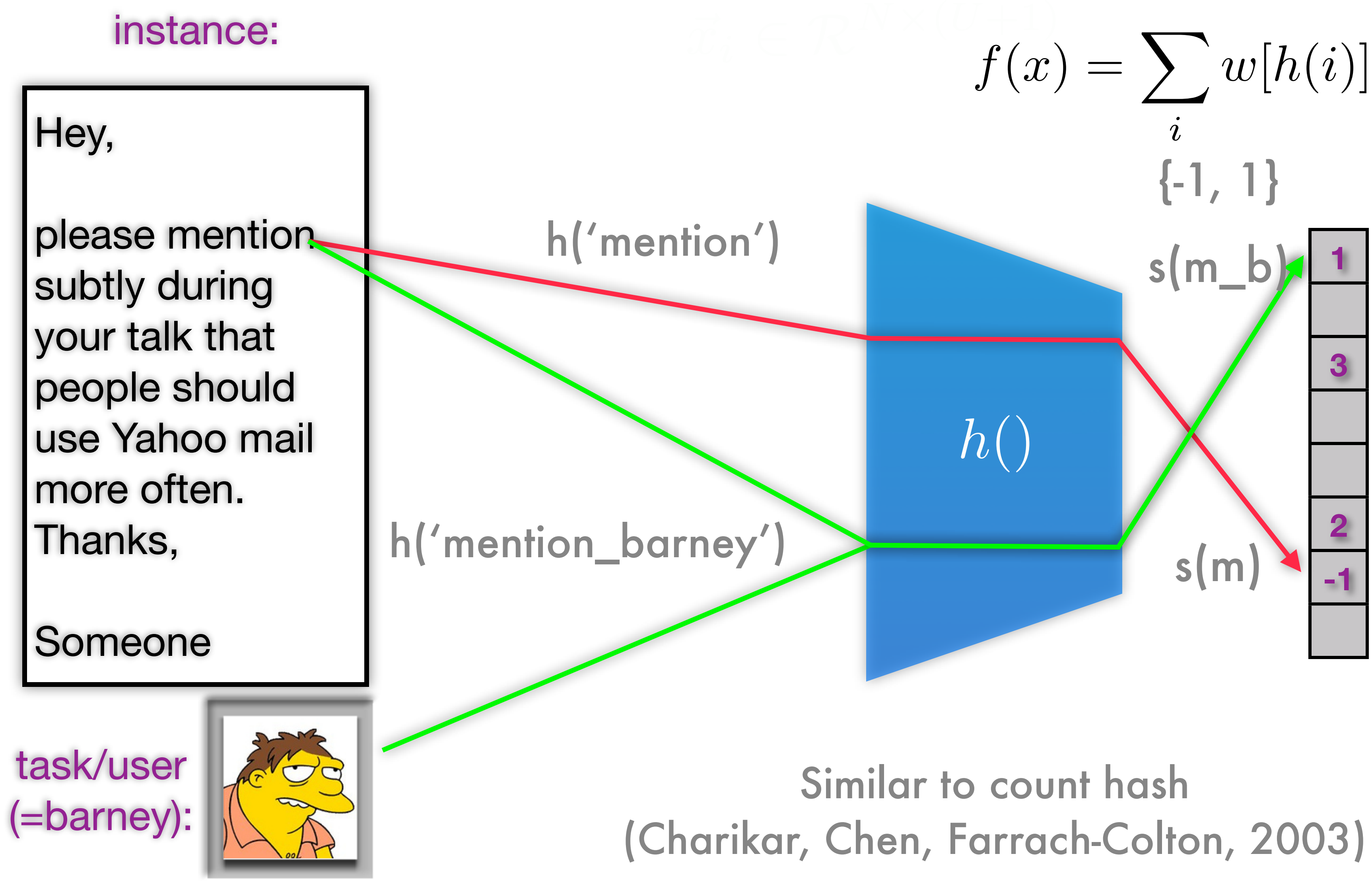

instance:
dictionary:
Hey,
please mention subtly during your talk that people should use Yahoo mail more often.
Thanks,
Someone
task/user (=barney):
1
2
1
1
sparse
hash function:
$h()$
1
3
2
1
sparse

Hash Kernel


instance:
dictionary:
Hey,
please mention subtly during your talk that people should use Yahoo mail more often.
Thanks,
Someone
this slide is old
task/user (=barney):
1
2
1
1
hash function:
$h()$
1
3
2
1
sparse
sparse

Hash Kernel

$$f(x) = \sum_i w[h(i)]\sigma(i)x_i$$



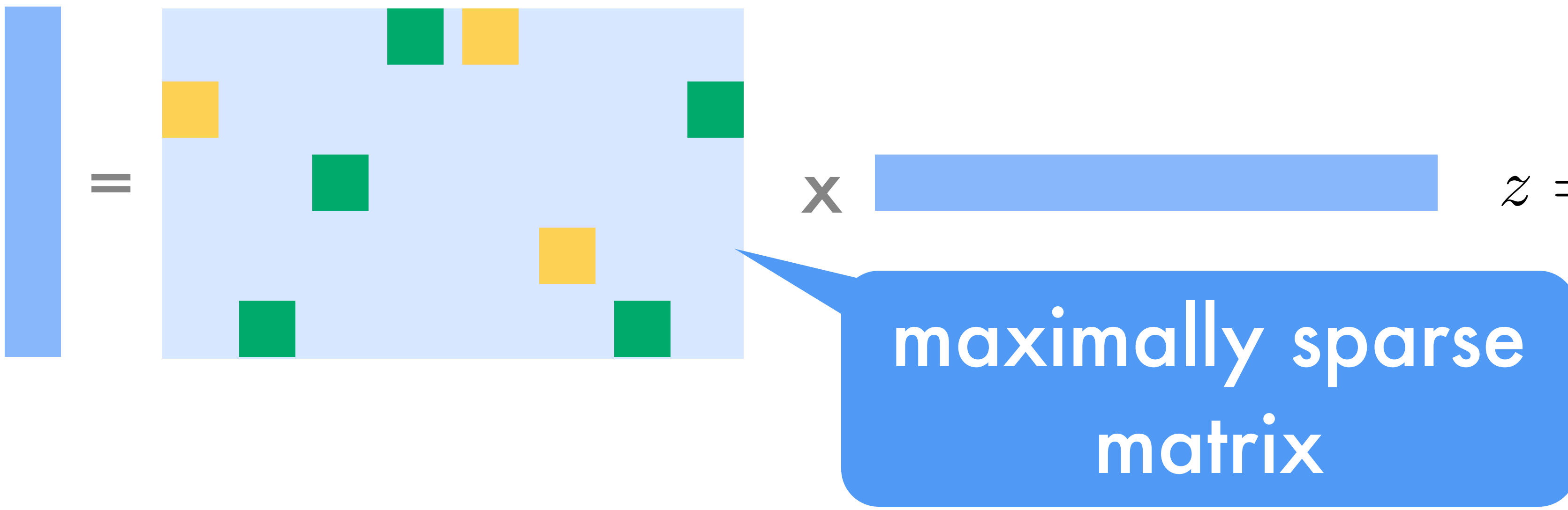
Similar to count hash
(Charikar, Chen, Farrach-Colton, 2003)

Advantages of hashing

- No dictionary!
  - Content drift is no problem
  - All memory used for classification
  - Finite memory guarantee
    (good for online learning)

- No Memory needed for projection (vs. LSH).
- Implicit mapping into high dimensional space!
- Sparsity preserving! (vs LSH)

Hash Kernels - the matrix view



$$z = Px$$

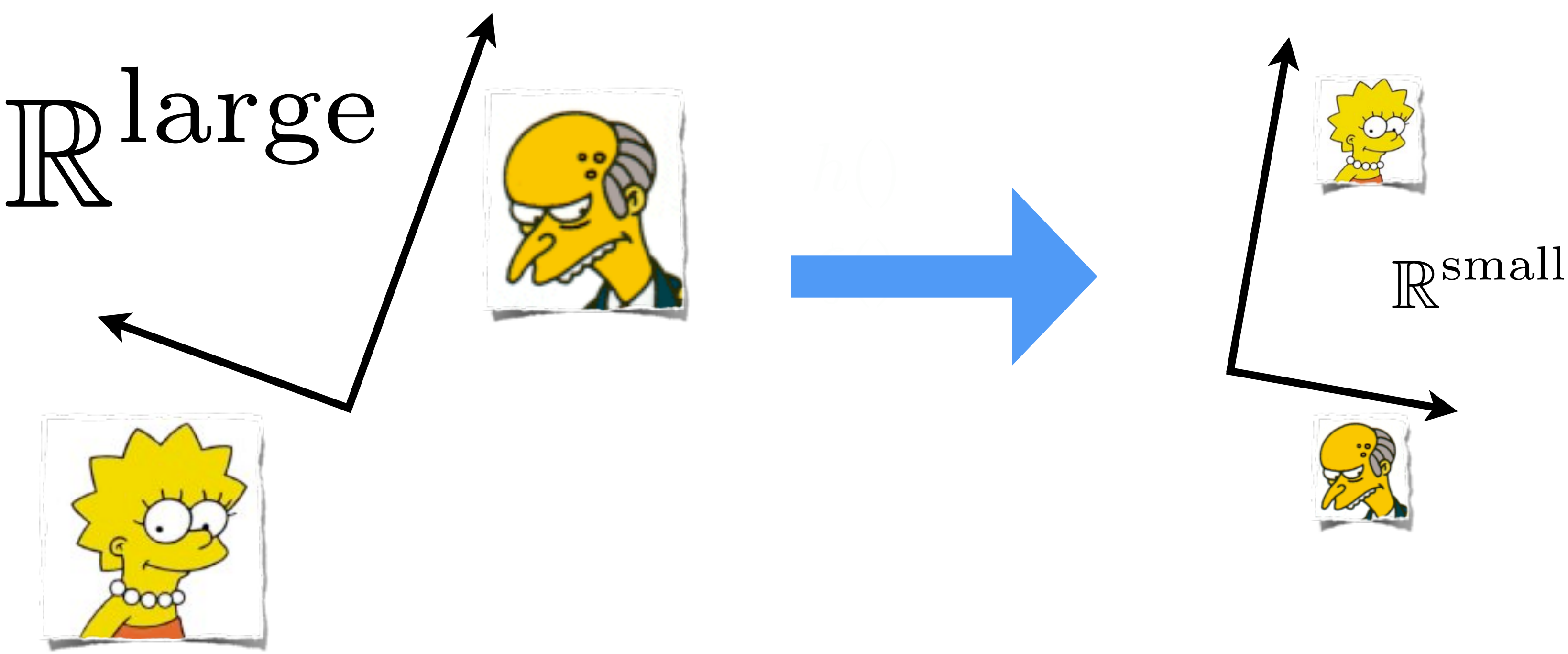
- **Preserves inner product**

$$\langle w, x \rangle = \sum_i w_i x_i \quad \langle \bar{w}, \bar{x} \rangle = \sum_j \left[\sum_{i:h(i)=j} w_i \sigma(i) \right] \left[\sum_{i:h(i)=j} x_i \sigma(i) \right]$$

Rademacher hash

$$\mathbf{E}_\sigma [\sigma(i)\sigma(i')] = \delta_{ii'}$$

Approximate Orthogonality



We can do multi-task learning!

Guarantees

- For a random hash function the inner product vanishes with high probability via

$$\Pr\{|\langle w_v, h_u(x)\rangle| > \epsilon\} \leq 2e^{-C\epsilon^2 m}$$

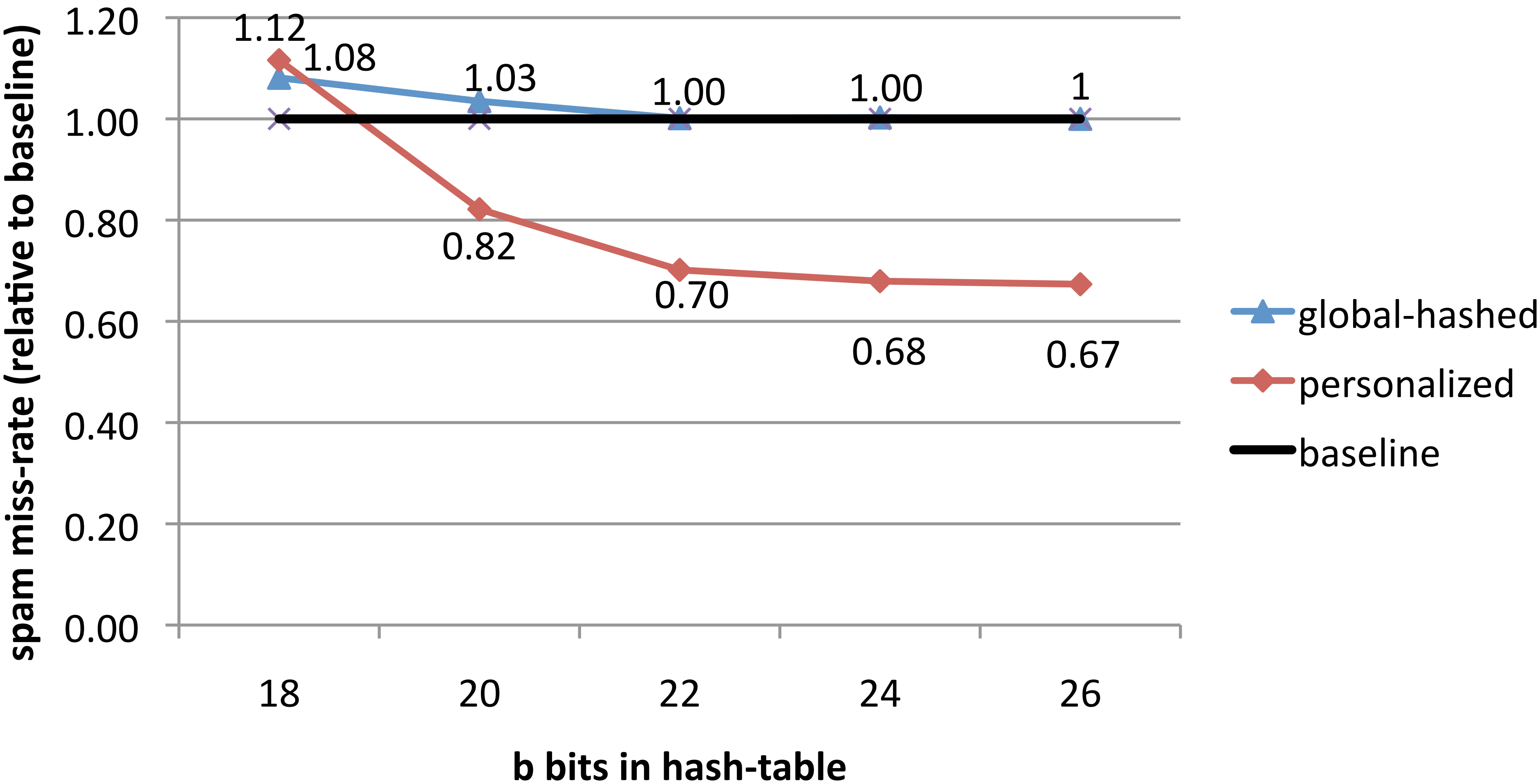
- We can use this for multitask learning

Direct sum in Hilbert Space → **Sum** in Hash Space

- Hashed inner product is unbiased
- Variance is O(1/n)
- Restricted isometry property (Kumar, Sarlos, Dasgupta 2010)

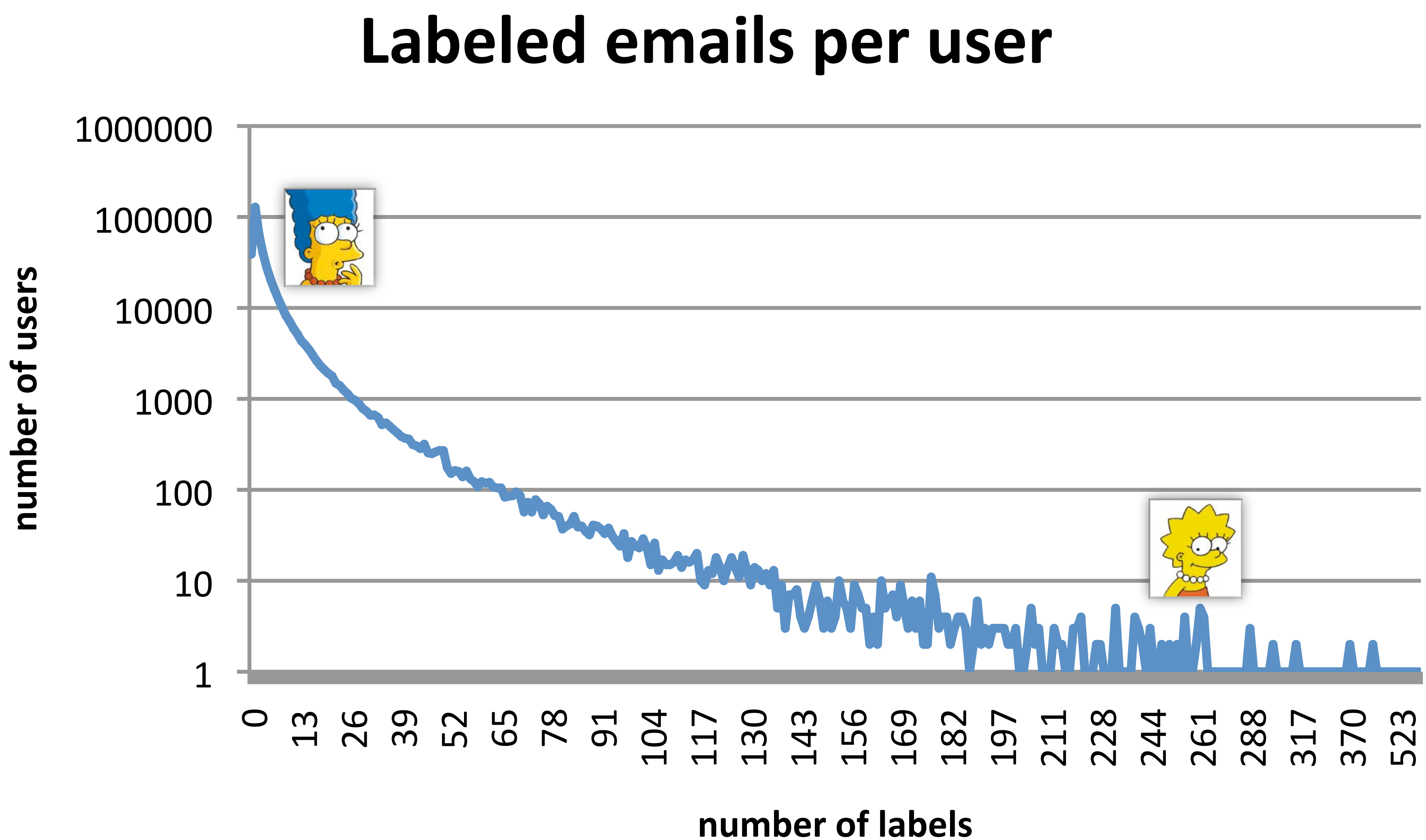
Weinberger, K., Dasgupta, A., Attenberg, J., Langford, J., and Smola, A. J.,Feature Hashing for Large Scale Multitask Learning, International Conference on Machine Learning, 2009. PDF

Spam classification results

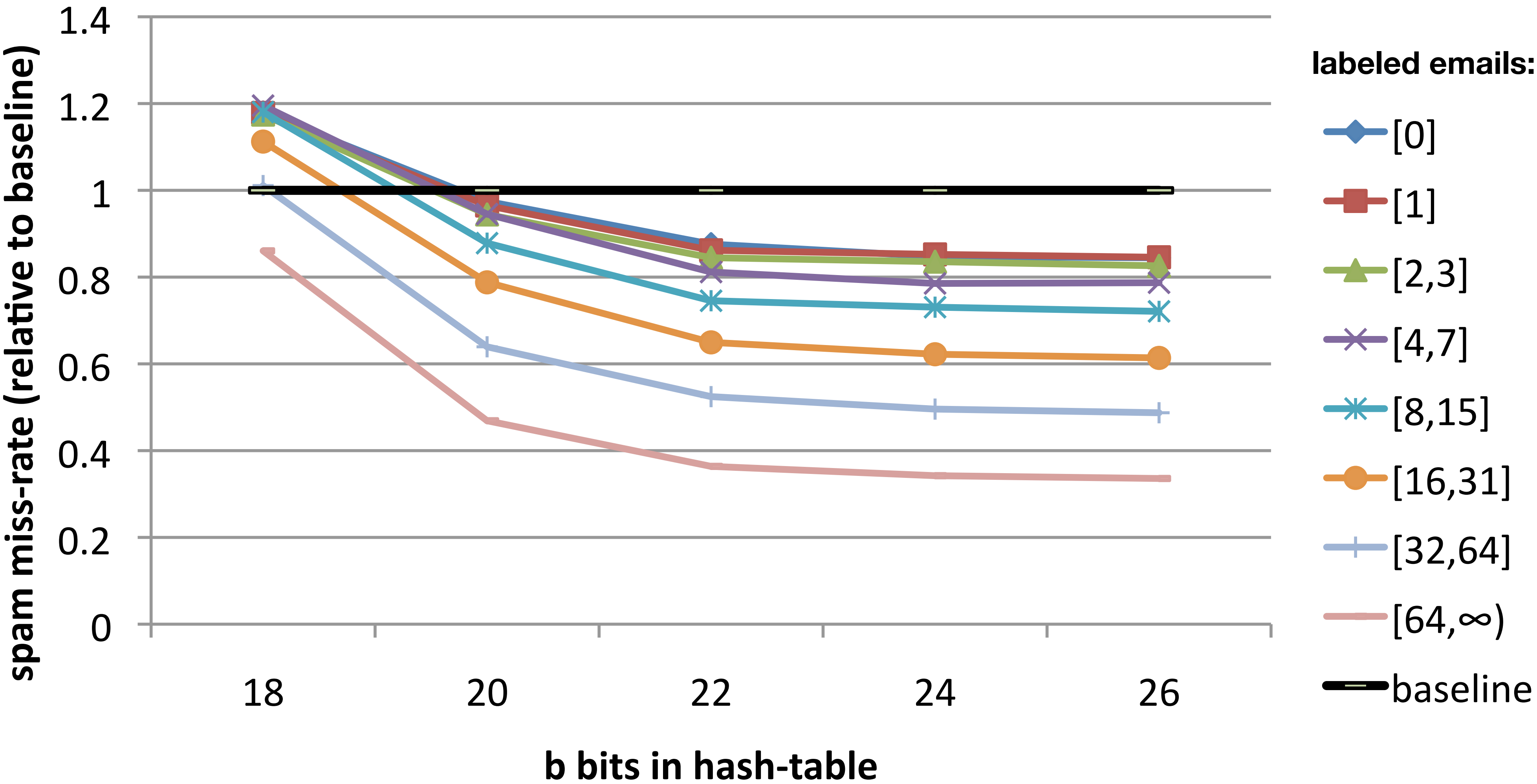


20 million emails, 400,000 users

Lazy users ...

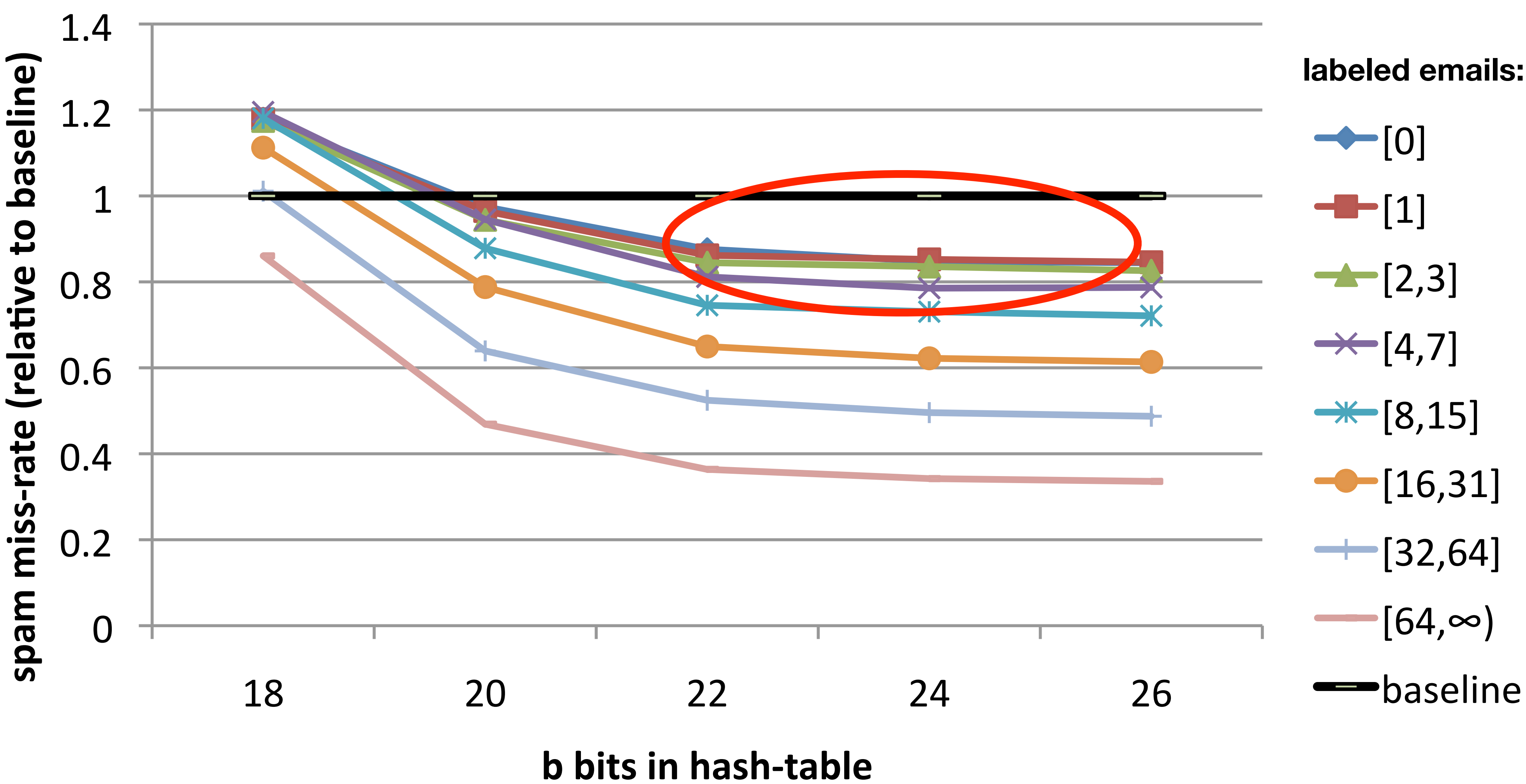

Labeled emails per user
number of users
1000000
100000
10000
1000
100
10
1
0
13
26
39
52
65
78
91
104
117
130
143
156
169
182
197
211
228
244
261
288
317
370
523
number of labels

Results by user group



Carnegie Mellon University

Results by user group



Carnegie Mellon University

Approximate String Matches

- General idea

$$k(x,x') = \sum_{w \in x} \sum_{w' \in x'} \kappa(w,w') \text{ for } |w - w'| \le \delta$$

Carnegie		Carnegie
Carneg1e	catch all	Carneg*e
0arnegie	with wildcards	*arnegie
Canegie		Ca*negie
Carn3gie		Carn*gie

- Map into fragments: `dog -> (*og, d*g, do*)`
- Hash fragments and weigh them based on mismatch amount
- Exponential in number of mismatches. Not alphabet size.

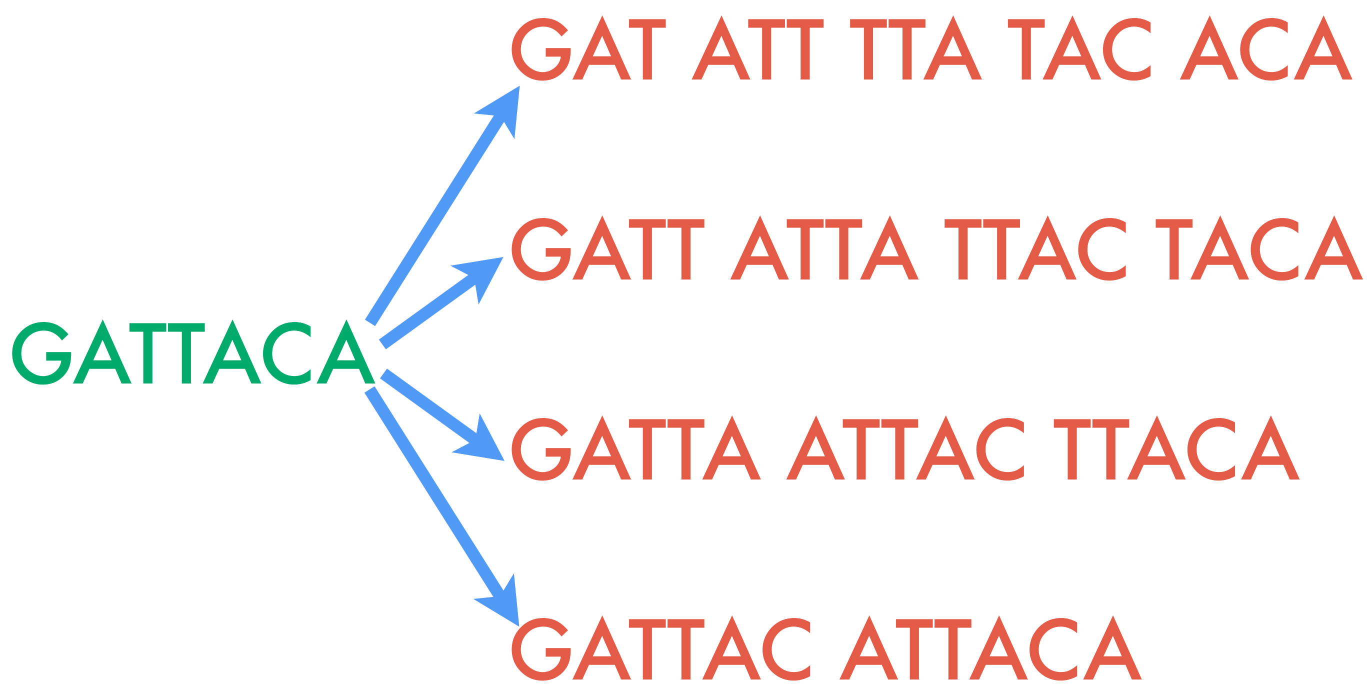
Fast String Kernels

- Example - DNA sequence

GATTACA

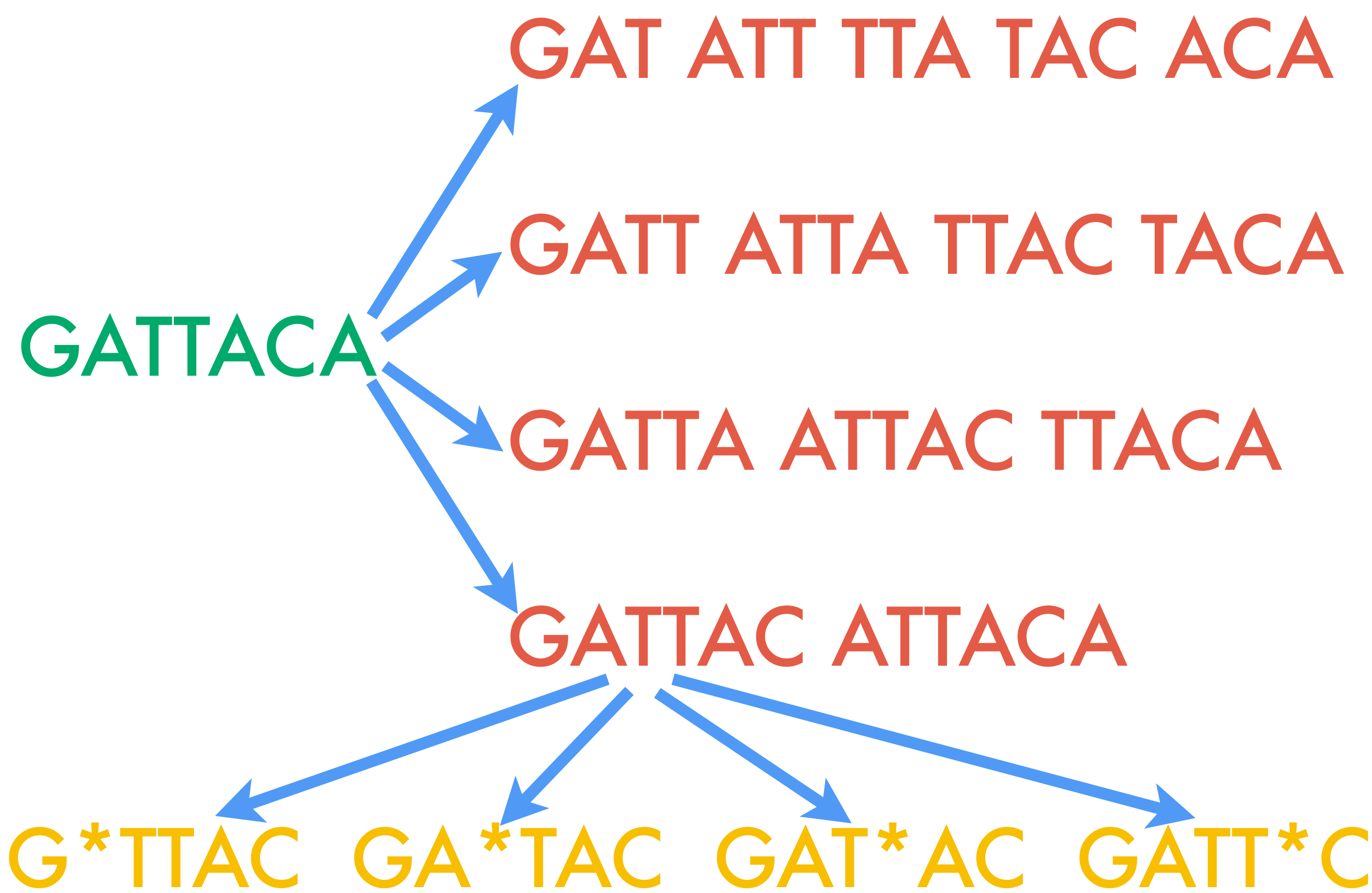
Fast String Kernels

- Example - DNA sequence



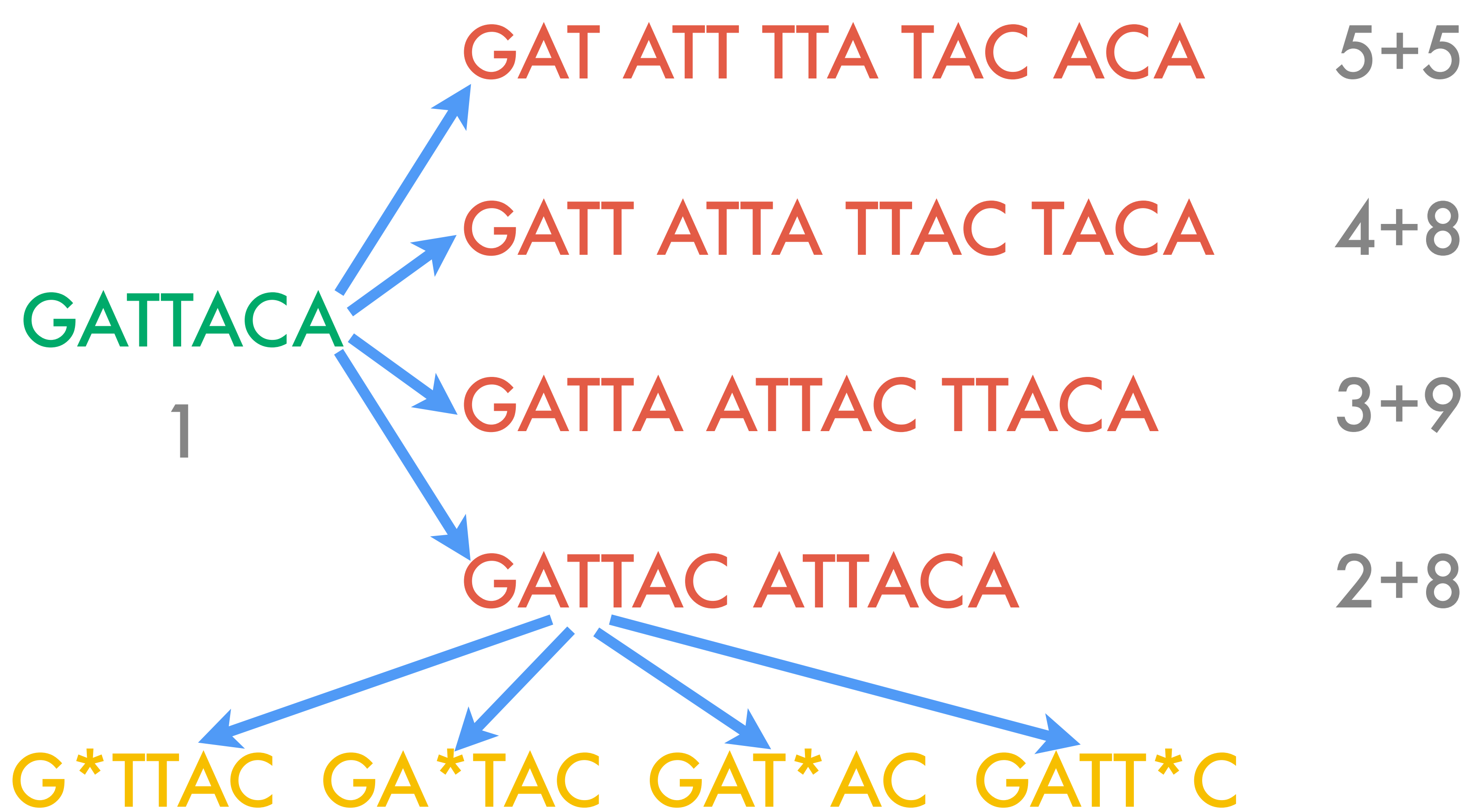
Fast String Kernels

- Example - DNA sequence



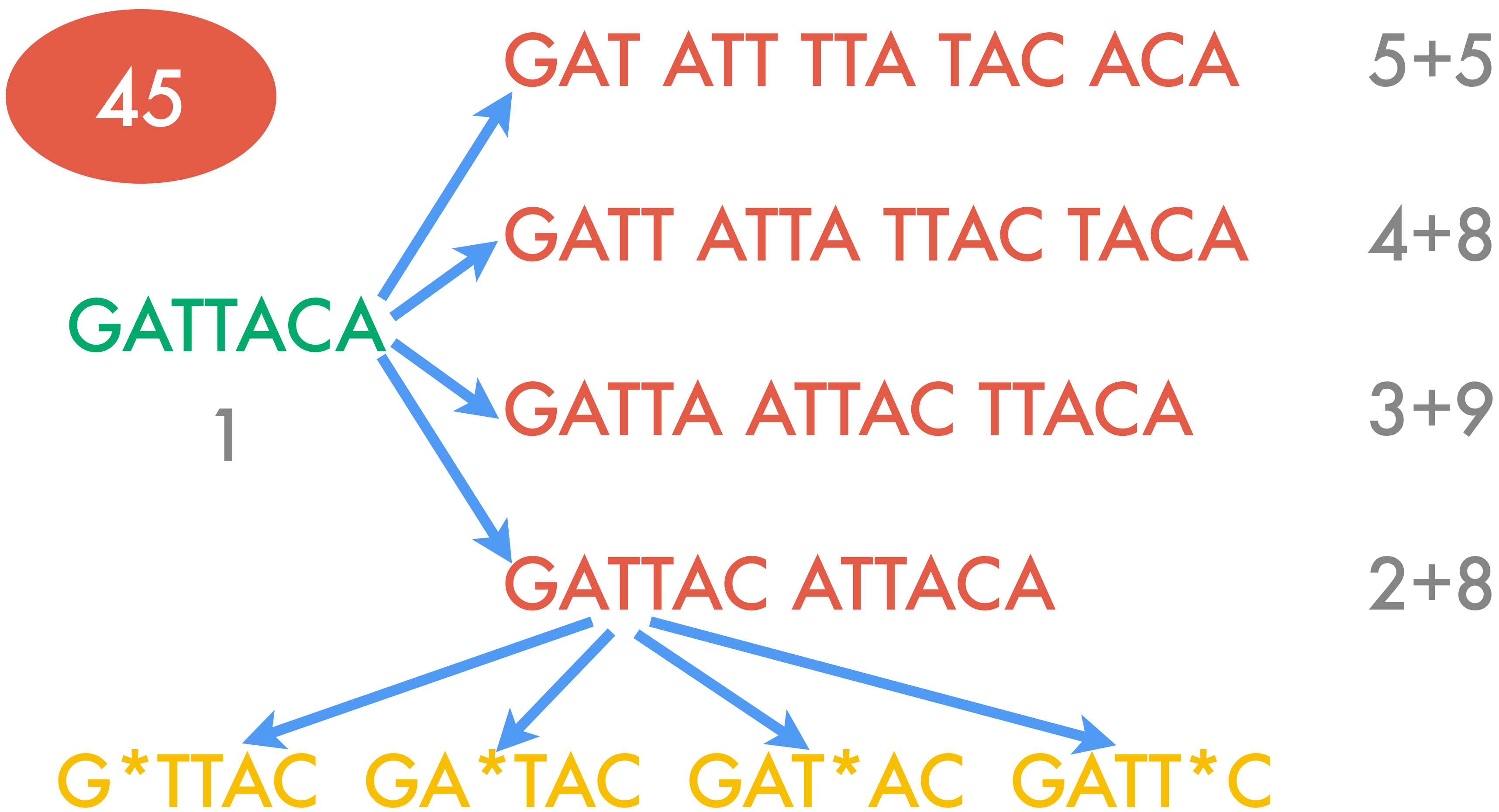
Fast String Kernels

- Example - DNA sequence



Fast String Kernels

- Example - DNA sequence



Carnegie Mellon University

Fast String Kernels

- Store coefficients explicitly (complicated)
- Use hash kernel to update counts (trivial)

GATTACA

GAT ATT TTA TAC ACA

GATT ATTA TTAC TACA

GATTA ATTAC TTACA

GATTAC ATTACA

G*TTAC GA*TAC GAT*AC GATT*C