



# Templates

for scalable data analysis

## 1 Introduction to Big Learning

Amr Ahmed, Alexander J Smola, Markus Weimer  
Yahoo! Research & UC Berkeley & ANU

# Thanks



Mohamed  
Aly



Joey  
Gonzalez



Yucheng  
Low



Qirong  
Ho



Shravan  
Narayananurthy



Amr  
Ahmed



Choon Hui  
Teo



Eric  
Xing



James  
Petterson



Sergiy  
Matyusevich



Jake  
Eisenstein



Shuang Hong  
Yang



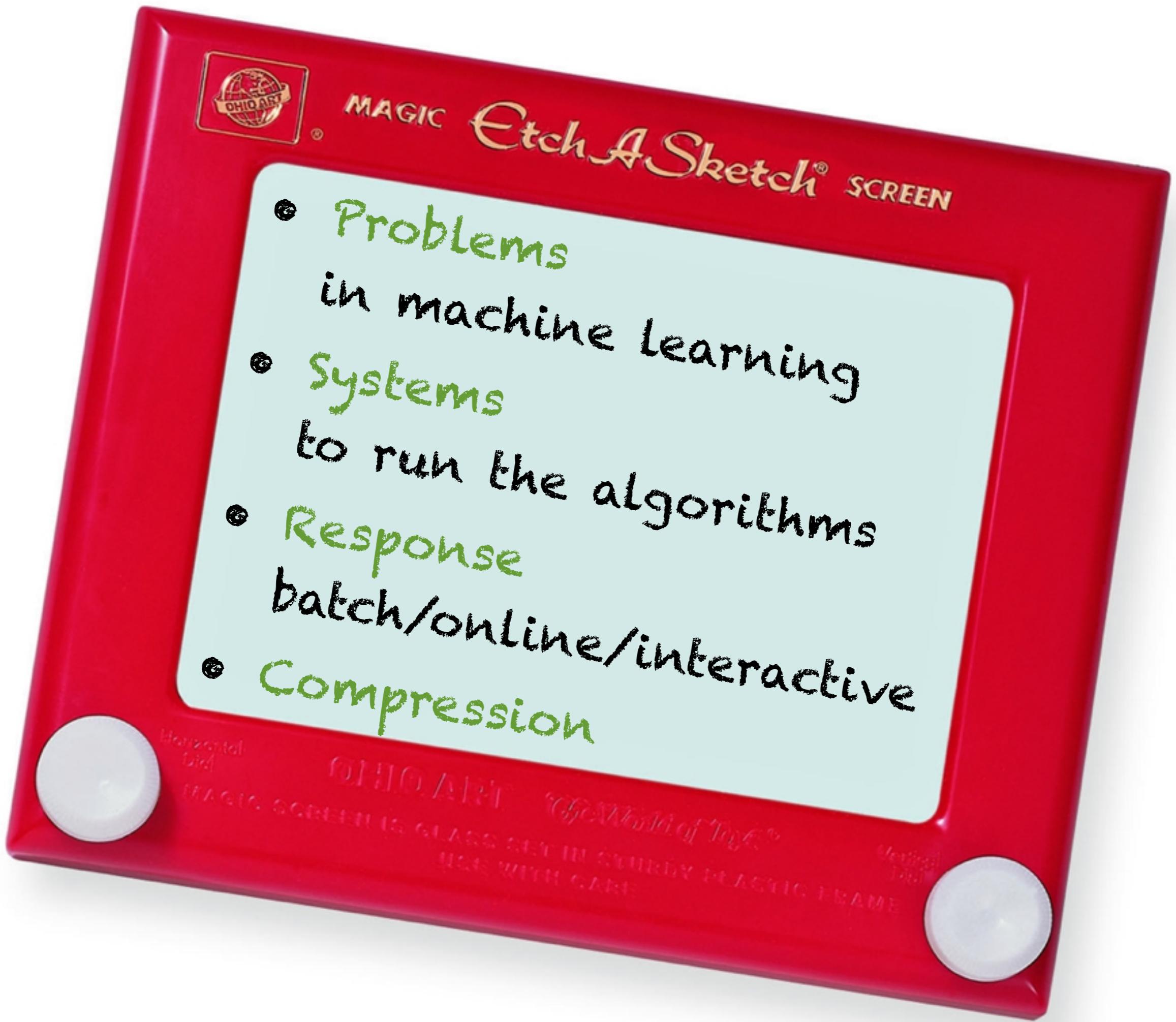
Vishy  
Vishwanathan

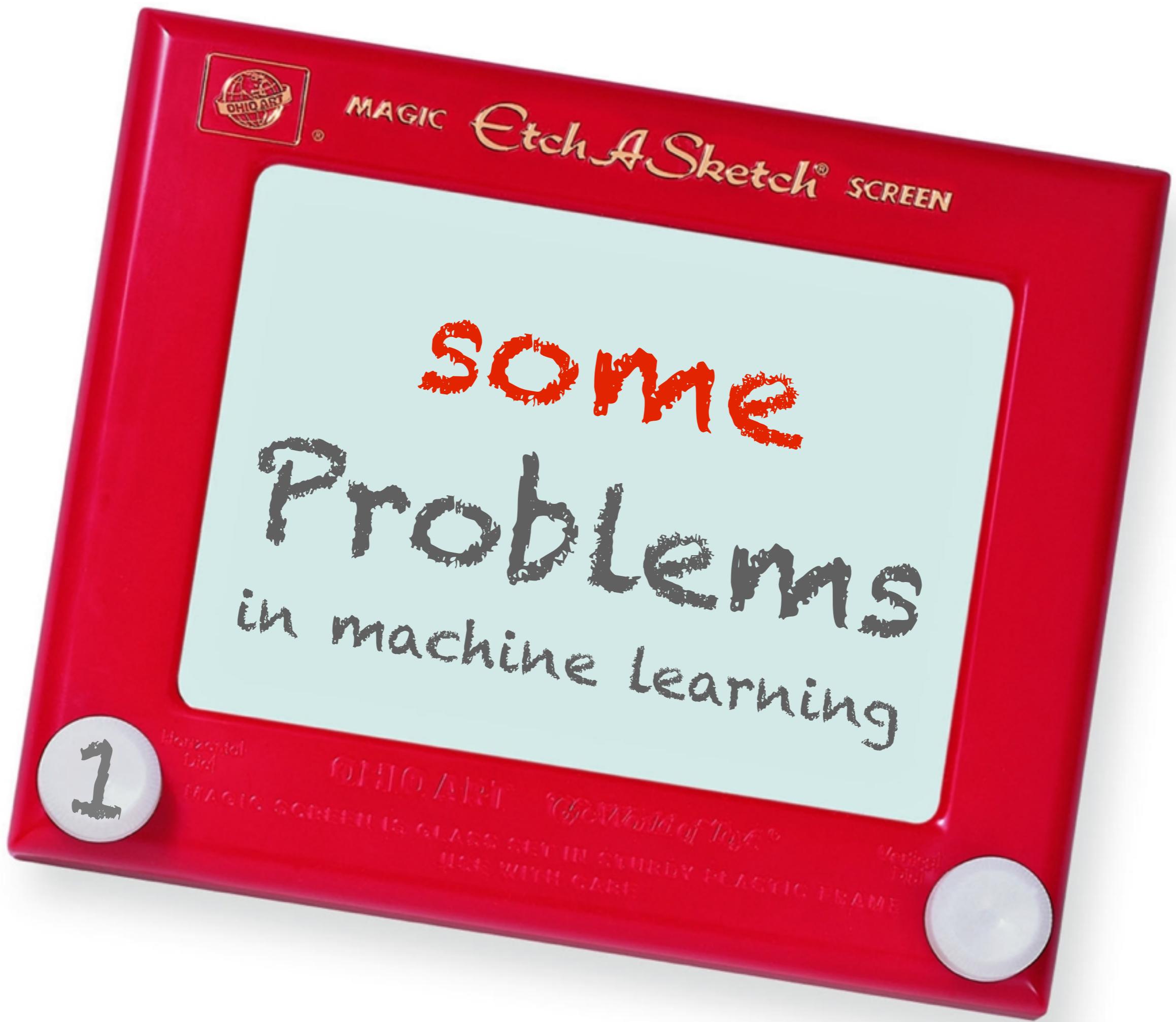


Markus  
Weimer



Vanja  
Josifovski





# Classification



# 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](mailto: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://jenkinstege13.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>



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

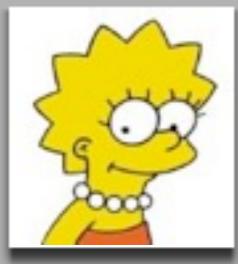
Get Vicodin, Valium, Xanax, Viagra, Oxycontin, and much more. Absolutely No Prescription Required.  
Over Night Shipping! Why should you be risking dealing with shady people. Check us out today!  
<http://jenkinstege.com.blogspot.com>

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>



**1: spam!**



**educated**

**0: quality**



**misinformed**

**1: donut?**



**confused**

**0: not-spam!**



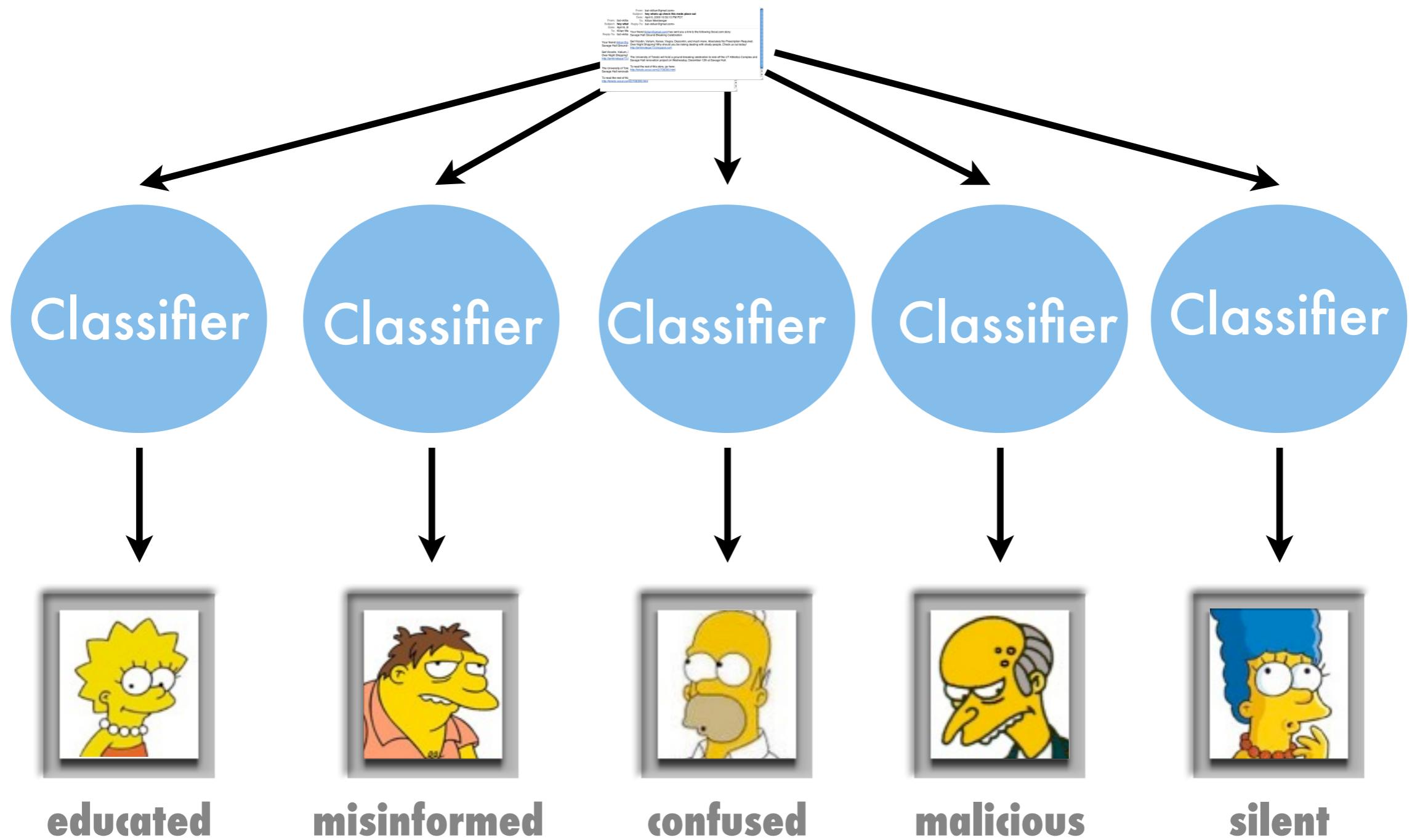
**malicious**

**?**

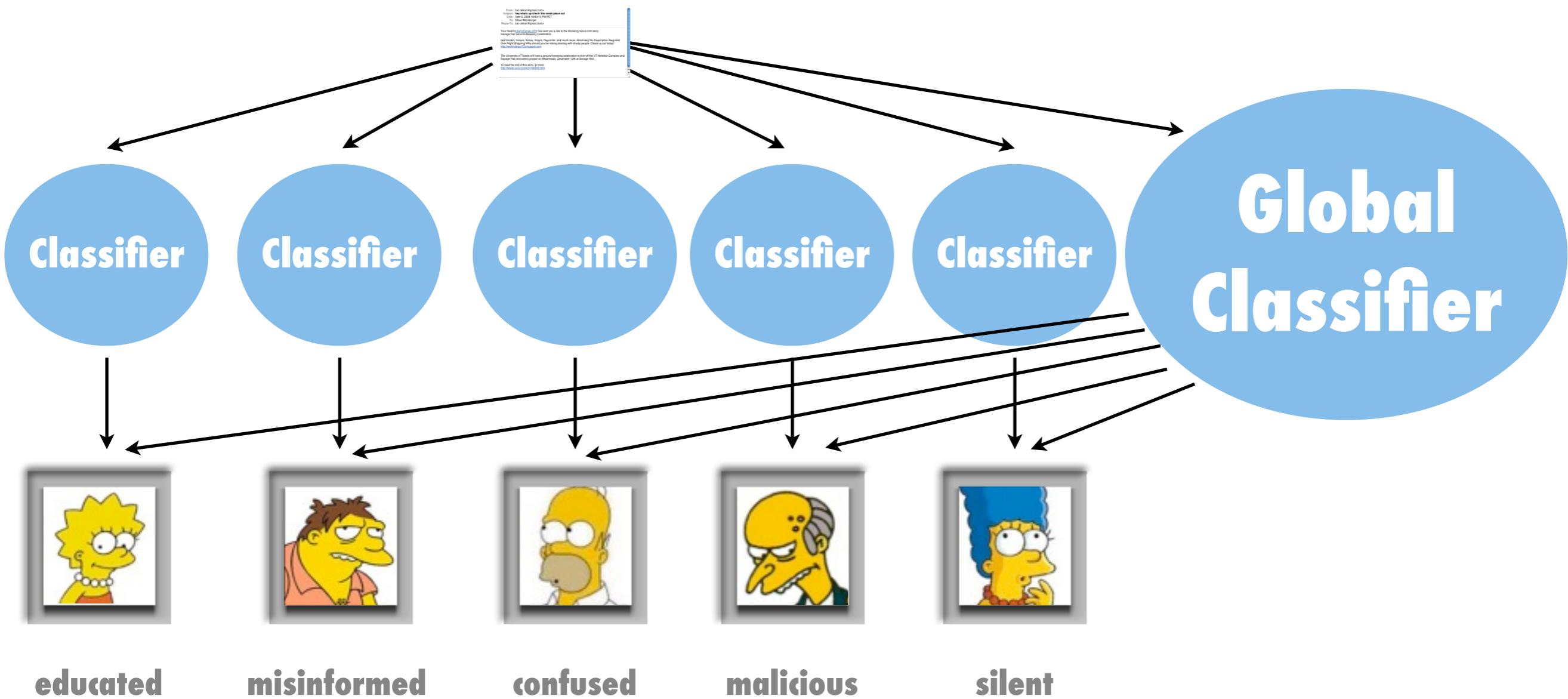


**silent**

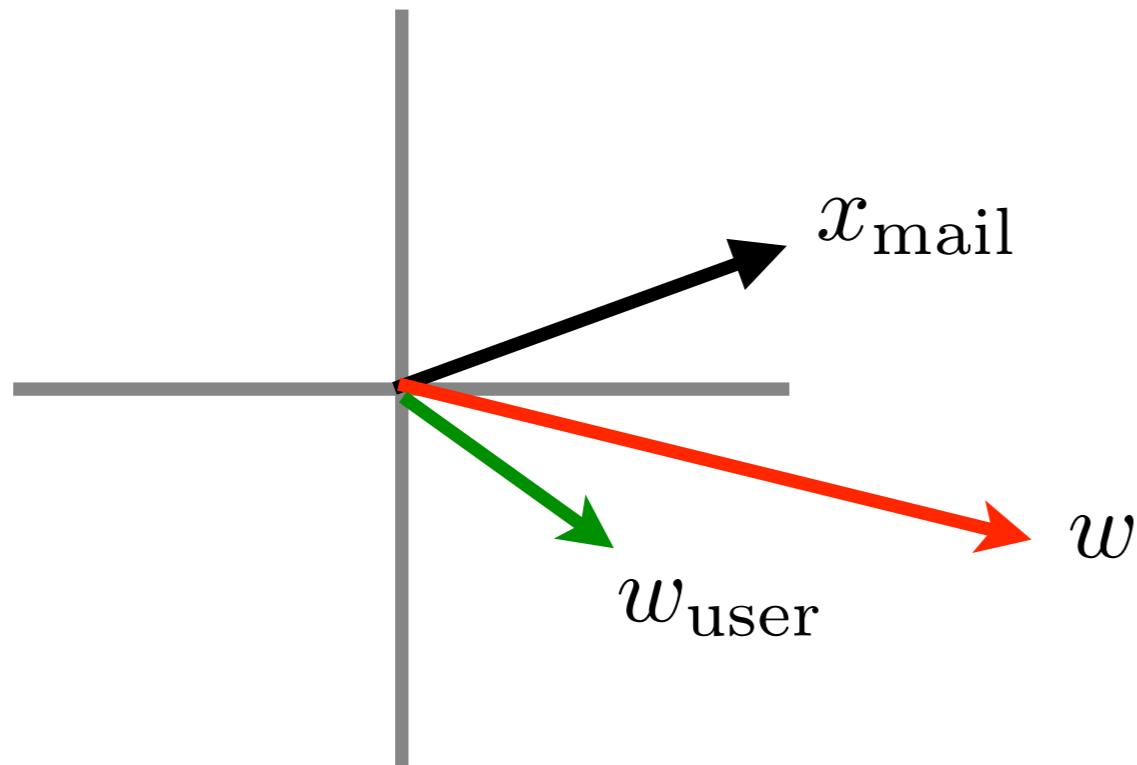
# Spam Filtering



# Personalized Spam Filtering



# Personalized Spam Filtering



- Function representation

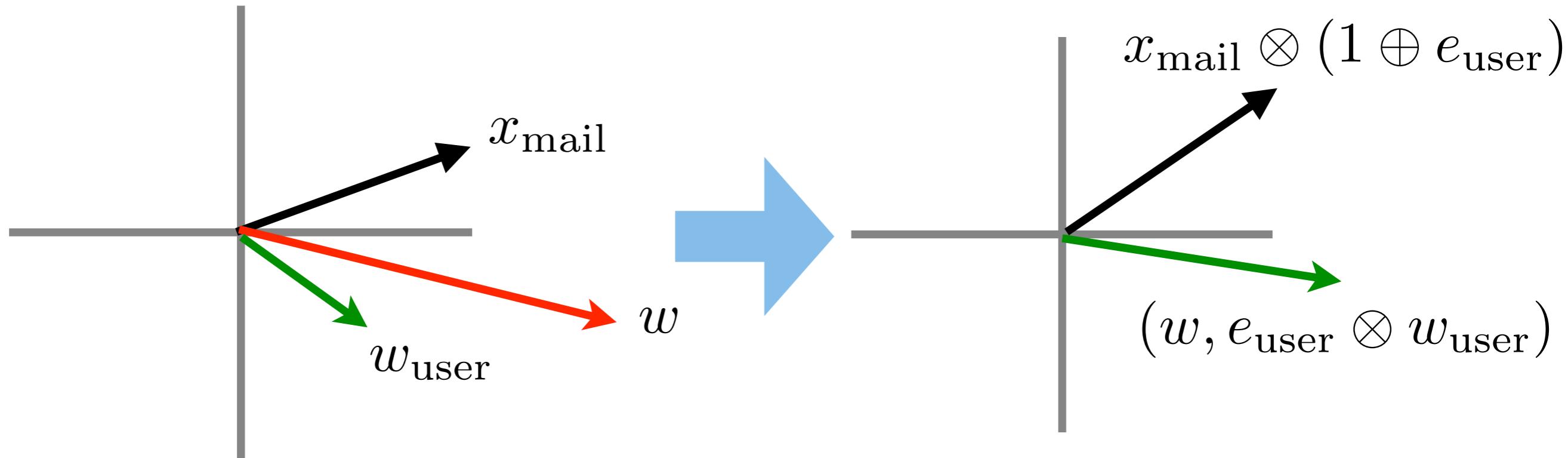
$$f(x, u) = \langle \phi(x), w \rangle + \langle \phi(x), w_u \rangle = \langle \phi(x) \otimes (1 \oplus e_u), w \rangle$$

(corresponds to multitask kernel of Pontil & Michelli, Daume)

- Reduce to binary classification problem and classify with

$$\operatorname{sgn} f(x, u)$$

# Personalized Spam Filtering



- Function representation

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

(corresponds to multitask kernel of Pontil & Michelli, Daume)

- Reduce to binary classification problem and classify with

$$\operatorname{sgn} f(x, u)$$

# Personalized Spam Filtering

			1-50 of 150	<	>	More
Delete all spam messages now (messages that have been in Spam more than 30 days will be automatically deleted)						
<input type="checkbox"/>	<input type="star"/>	<input type="checkbox"/>	吳林慧	性藥品全球-最有效最知名美國.聖品 - 催情藥大王-讓我們.夫妻high到底 每天都在打拼-就該買性藥品讓'我黑皮 ...	3:51 pm	
<input type="checkbox"/>	<input type="star"/>	<input type="checkbox"/>	leomasilqhfq	[moewwx] 可先看貨 再付款 經典&新款&名牌&包夾&名錶&鞋子&特價中IYI1AeU%5EqQ)9\$m]u=yi - 名牌包包,皮夾,鞋子,手錶	11:25 am	
<input type="checkbox"/>	<input type="star"/>	<input checked="" type="checkbox"/>	Penis Growth Sample	Smell sweater below the belt - Girls dig really long ones, yours will be LONGER after you take our organic pills http://biggr...	9:24 am	
<input type="checkbox"/>	<input type="star"/>	<input checked="" type="checkbox"/>	Edward Bell	Re: Re: Mig!%ori boosters ERO on-line - Ogni medicina nel gruppo di disfunzione erektili è qui http://njuzo.velvdoctor.ru	9:10 am	
<input type="checkbox"/>	<input type="star"/>	<input checked="" type="checkbox"/>	hr	Suuri Laina tarjous - Subject: Suuri Laina tarjous Hei, Tarvitsetko lainaa edulliseen korko on 3%. Ota yhteyttä ...	2:52 am	
<input type="checkbox"/>	<input type="star"/>	<input type="checkbox"/>	leomasilqhfq	[moewwx] 可先看貨 再付款 經典&新款&名牌&包夾&名錶&鞋子&特價中*#unaZSv\$*1?FLSahnu#* - 名牌包包,皮夾,鞋子,手錶	Apr 7	
<input type="checkbox"/>	<input type="star"/>	<input checked="" type="checkbox"/>	AOL Mail	AOL Mail notification - Technical E-mail from AOL Mail You can reply to this message by visiting AOL Message Center ...	Apr 7	
<input type="checkbox"/>	<input type="star"/>	<input type="checkbox"/>	Mr. Alan Johnson	Dear Sir/Madam - I write to know if this is your valid email. Please, let me know i want to discuss an important ...	Apr 7	
<input type="checkbox"/>	<input type="star"/>	<input checked="" type="checkbox"/>	超值团购	仅49.8元，多乐士套4盒,跳跳蛋,7件成人用品，1件情趣内衣 - 套餐一：49.8元(多乐士4盒42只+震动环+情趣内衣+跳跳蛋+印度...	Apr 6	
<input type="checkbox"/>	<input type="star"/>	<input type="checkbox"/>	leomasilqhfq	[moewwx] 可先看貨 再付款 經典&新款&名牌&包夾&名錶&鞋子&特價中P>d)ynZ%\$iUMAvq1 - 名牌包包,皮夾,鞋子,手錶,眼...	Apr 6	
<input type="checkbox"/>	<input type="star"/>	<input type="checkbox"/>	K WILL	Good days to you - Good days to you Please kindly accept my apology for sending you this email without your consent ...	Apr 6	

- 100-1000 million users
- 10-1000 messages per user
- Distributed storage and processing
- Real-time response required
- Implicit response

$$\underset{w}{\text{minimize}} \sum_{i=1}^m \max(0, 1 - y \langle w, x \rangle) + \frac{\lambda}{2} \|w\|^2$$

# Ontologies

**dmoz open directory project** In partnership with **AOL Search.**

[about dmoz](#) | [dmoz blog](#) | [suggest URL](#) | [help](#) | [link](#) | [editor login](#)

[advanced](#)

<b>Arts</b> <a href="#">Movies, Television, Music...</a>	<b>Business</b> <a href="#">Jobs, Real Estate, Investing...</a>	<b>Computers</b> <a href="#">Internet, Software, Hardware...</a>
<b>Games</b> <a href="#">Video Games, RPGs, Gambling...</a>	<b>Health</b> <a href="#">Fitness, Medicine, Alternative...</a>	<b>Home</b> <a href="#">Family, Consumers, Cooking...</a>
<b>Kids and Teens</b> <a href="#">Arts, School Time, Teen Life...</a>	<b>News</b> <a href="#">Media, Newspapers, Weather...</a>	<b>Recreation</b> <a href="#">Travel, Food, Outdoors, Humor...</a>
<b>Reference</b> <a href="#">Maps, Education, Libraries...</a>	<b>Regional</b> <a href="#">US, Canada, UK, Europe...</a>	<b>Science</b> <a href="#">Biology, Psychology, Physics...</a>
<b>Shopping</b> <a href="#">Clothing, Food, Gifts...</a>	<b>Society</b> <a href="#">People, Religion, Issues...</a>	<b>Sports</b> <a href="#">Baseball, Soccer, Basketball...</a>
<b>World</b> <a href="#">Català, Dansk, Deutsch, Español, Français, Italiano, 日本語, Nederlands, Polski, Русский, Svenska...</a>		

[Become an Editor](#) Help build the largest human-edited directory of the web

Copyright © 2012 Netscape

5,018,902 sites - 95,017 editors - over 1,010,596 categories

- 10k to 1M categories
- Few instances per category
- Hierarchical structure (top level more important than leaf)
- Category selection arbitrary
- Low entropy on leaves
- Often several ontologies in use



# Ontologies

**dmoz open directory project** In partnership with **Aol Search.**

[about dmoz](#) | [dmoz blog](#) | [suggest URL](#) | [help](#) | [link](#) | [editor login](#)

[advanced](#)

<b>Arts</b> <a href="#">Movies, Television, Music...</a>	<b>Business</b> <a href="#">Jobs, Real Estate, Investing...</a>	<b>Computers</b> <a href="#">Internet, Software, Hardware...</a>
<b>Games</b> <a href="#">Video Games, RPGs, Gambling...</a>	<b>Health</b> <a href="#">Fitness, Medicine, Alternative...</a>	<b>Home</b> <a href="#">Family, Consumers, Cooking...</a>
<b>Kids and Teens</b> <a href="#">Arts, School Time, Teen Life...</a>	<b>News</b> <a href="#">Media, Newspapers, Weather...</a>	<b>Recreation</b> <a href="#">Travel, Food, Outdoors, Humor...</a>
<b>Reference</b> <a href="#">Maps, Education, Libraries...</a>	<b>Regional</b> <a href="#">US, Canada, UK, Europe...</a>	<b>Science</b> <a href="#">Biology, Psychology, Physics...</a>
<b>Shopping</b> <a href="#">Clothing, Food, Gifts...</a>	<b>Society</b> <a href="#">People, Religion, Issues...</a>	<b>Sports</b> <a href="#">Baseball, Soccer, Basketball...</a>
<b>World</b> <a href="#">Català, Dansk, Deutsch, Español, Français, Italiano, 日本語, Nederlands, Polski, Русский, Svenska...</a>		

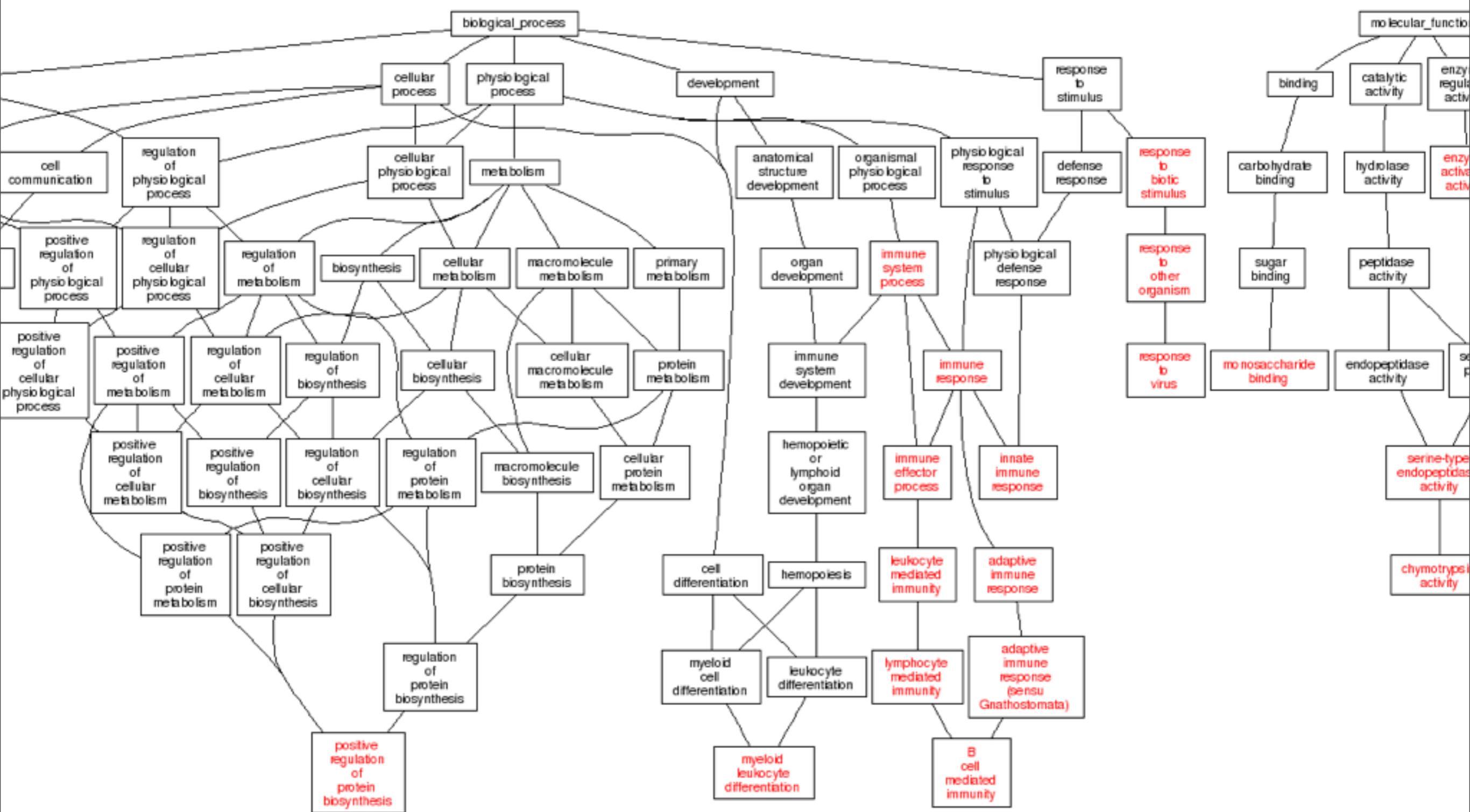
[Become an Editor](#) Help build the largest human-edited directory of the web

Copyright © 2012 Netscape

5,018,902 sites - 95,017 editors - over 1,010,596 categories

- 10k to 1M categories
- Few instances per category
- Hierarchical structure (top level more important than leaf)
- Category selection arbitrary
- Low entropy on leaves
- Often several ontologies in use

# Gene Ontology DAG



# Ontologies

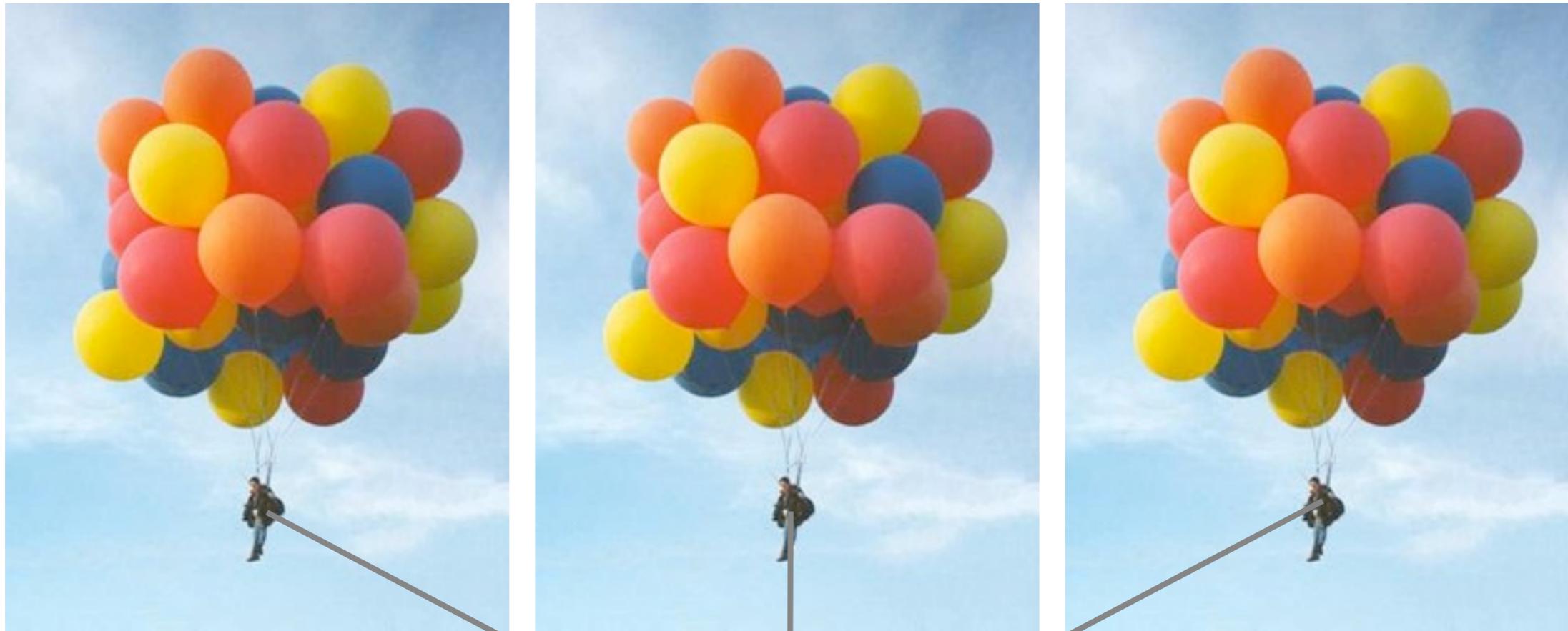
- 1000s of categories
- High error rate (impossible to learn them all)
- Structured loss  
(count common top level categories)
- Good strategy is additive function class

$$f(x, y) = \sum_{y' \in \text{path}(y)} \langle w_{y'}, x \rangle$$

Need efficient decoding on tree

- Alternative - obtain ontology automatically

# Clustering



YAHOO!

# Clustering

YAHOO!

# Clustering

The image displays two distinct web pages side-by-side, illustrating different design approaches and user interface elements.

**Left Side (United Airlines Website):**

- Header:** UNITED logo, My profile | Worldwide sites | Customer service, Planning & booking, Reservations & check-in, Mileage Plus®, Services & information, Search site.
- Banner:** "ON TIME" badge, "United, #1 in on-time arrivals. Details", "Use 30% fewer miles on your next United flight." with a large orange percentage sign icon.
- Flight Booking Tools:** Flights, Check-in, Flight status, BOOK FLIGHT, REDEEM MILES buttons, From (Find airport) To (Find airport) fields, Departing/Returning date pickers, Search by (Schedule & price, Price, Flexible), Adult/Cabin (Economy, Refundable) dropdowns, Promotion code or Electronic certificate input field, Log in link.
- Footer:** About United, Investor relations, Business resources, Careers, Site map, DIRECTIONS, MAILING LIST, Copyright © 2006 Chez Panisse Restaurant & Cafe. All Rights Reserved.

**Right Side (Australian National University Website):**

- Header:** Change Location, Search, You Fly, Loyalty Programmes, Promotions, myEMAIL, IVLE, LIBRARY, MAPS, CALENDAR, SITEMAP, CONTACT, e-CARDS.
- Search Bar:** Search for... in NUS Websites, GO button.
- Navigation:** RESEARCH, ENTERPRISE, CAMPUS LIFE, GIVING, CAREERS@NUS.
- Content:** A large banner image with the text "centred in Asia". Below it are sections for CURRENT STUDENTS, RESEARCH & EDUCATION, ABOUT ANU, and STAFF.
- Footer:** Search ANU..., WEB, CONTACTS, MAP, GO buttons, Copyright © 2012 The Australian National University.

# Clustering

The United Airlines website interface. At the top, there are links for 'My profile', 'Worldwide sites', and 'Customer service'. Below this, a search bar shows 'Flight #1 is online arrival details'. The main content area includes tabs for 'Flights', 'Check-in', and 'Flight status'. A large banner on the left says 'Use 30% fewer miles on your next United flight.' with a large orange percentage sign icon. To the right, there's a 'Log in' form and a 'BOOK FLIGHT' section with dropdown menus for 'From', 'To', 'Departing', and 'Returning'. Below these are sections for 'Search tips', 'Cabin', 'Promotion code or Electronic certificate', and 'Log in to view all seating options'. At the bottom, there are links for 'About United', 'Investor relations', 'Business resources', 'Careers', and 'Site map'. A footer section features 'SIA Holidays' and 'Hotel Bookings'.

The KrisFlyer website interface. It features a 'Log in' form with fields for 'Mileage Plus # or email address' and 'Password'. Below it is a 'Start with' section for 'My Mileage Plus' or 'My reservations'. A 'Start earning miles today! Join Mileage Plus' button is visible. The main content area includes a '6-Digit PIN' input field, a 'Log In Help' link, and a 'Log In+' button. Below this are several flight booking options with prices: Singapore - Bangkok SGD 395\*, Singapore - Hong Kong SGD 546\*, Singapore - Taipei SGD 768\*, Singapore - Tokyo (Haneda) SGD 983\*, and Singapore - London. Each option has a 'Book Now' button.

The Australian National University (ANU) website. At the top, there are links for 'EXPLORE ANU', 'A-Z INDEX', and a search bar. The main header features the 'ANU' logo and 'THE AUSTRALIAN NATIONAL UNIVERSITY'. Below the header, there are navigation tabs for 'HOME', 'FUTURE STUDENTS', 'CURRENT STUDENTS', 'RESEARCH & EDUCATION', 'ABOUT ANU', and 'STAFF'. A large banner on the right shows a close-up of a small plant growing from a tree trunk. The central content area includes a news item about 'Ash forests rise and rise again', a photo of students, and links for 'Forests renew after Black Saturday fires', 'School of Music at Floride', 'Undergraduate studies', and 'Higher Degree Research'. A 'Joint Evacuation Exercises' section is also present.

A comparison of two website designs. On the left is the Chez Panisse website, featuring a dark grey background with white text. It includes sections for 'RESERVATIONS', 'MENUS', 'ABOUT', 'SPECIAL EVENTS', 'STORE', and 'CONTACT'. On the right is the Suntec REIT website, which features a large image of a brightly lit church at night. The page includes links for 'Wining & Dining', 'Contact', 'Sitemap', and 'About Suntec REIT'. The overall design is more visually focused on images than text.

YAHOO!

# Clustering

The screenshot shows the United Airlines website interface. At the top, there are navigation links for 'Planning & booking', 'Reservations & check-in', 'Mileage Plus®', and 'Services & information'. A search bar is present. Below the header, there's a large promotional banner for 'Use 30% fewer miles on your next United flight' featuring a large orange percentage sign. To the right of the banner is a 'Log in' form. Further down, there are sections for 'Flights', 'Check-in', and 'Flight status'. The main search area includes fields for 'From', 'To', and 'Departing/Returning' dates. Below these are sections for 'Search tips', 'Cabin', 'Promotion code or Electronic certificate', and 'Log in to view all seating options'. At the bottom, there are links for 'About United', 'Investor relations', 'Business resources', 'Careers', and 'Site map'. A footer section includes links for 'SIA Holidays' and 'Hotel Bookings'.

airline

The screenshot shows the ANU website. At the top, there are links for 'EXPLORE ANU', 'A-Z INDEX', and a search bar. The main header features the ANU logo and the text 'The Australian National University'. Below the header, there are navigation tabs for 'HOME', 'FUTURE STUDENTS', 'CURRENT STUDENTS', 'RESEARCH & EDUCATION', 'ABOUT ANU', and 'STAFF'. A prominent feature is a large image of a forest with a small plant growing through a log, with the caption 'Ash forests rise and rise again'. Other news items include 'Forests renew after Black Saturday fires' and 'Higher Degree Research'. At the bottom, there are buttons for 'PROSPECTIVE STUDENTS', 'CURRENT STUDENTS', 'STAFF', 'ALUMNI', and 'VISITORS'.

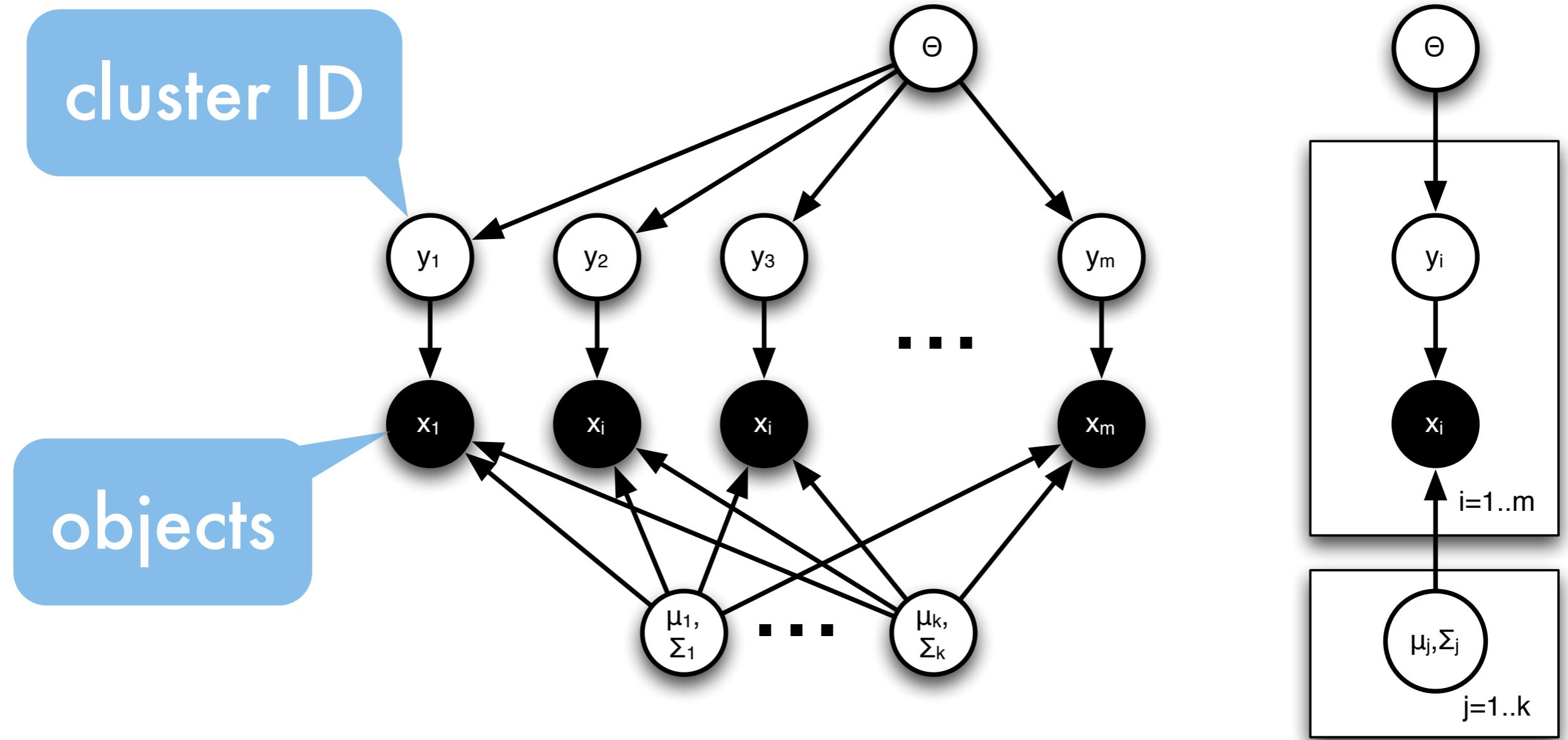
university

The screenshot shows the Chez Panisse website. On the left, there's a sidebar with links for 'RESERVATIONS', 'MENUS', 'ABOUT', 'SPECIAL EVENTS', 'STORE', and 'CONTACT'. The main content area features a large photograph of the restaurant's exterior, which is a rustic building with a sign that reads 'BAR DE LA PECHE'. Below the photo, there are links for 'Wine', 'Wining & Dining', 'Contact', 'Sitemap', and 'About Suntec REIT'. The bottom of the page includes a footer with links for 'Directions', 'Reservations', 'Contact', 'Feedback | Terms & Conditions', and 'About Suntec REIT'.

restaurant

YAHOO!

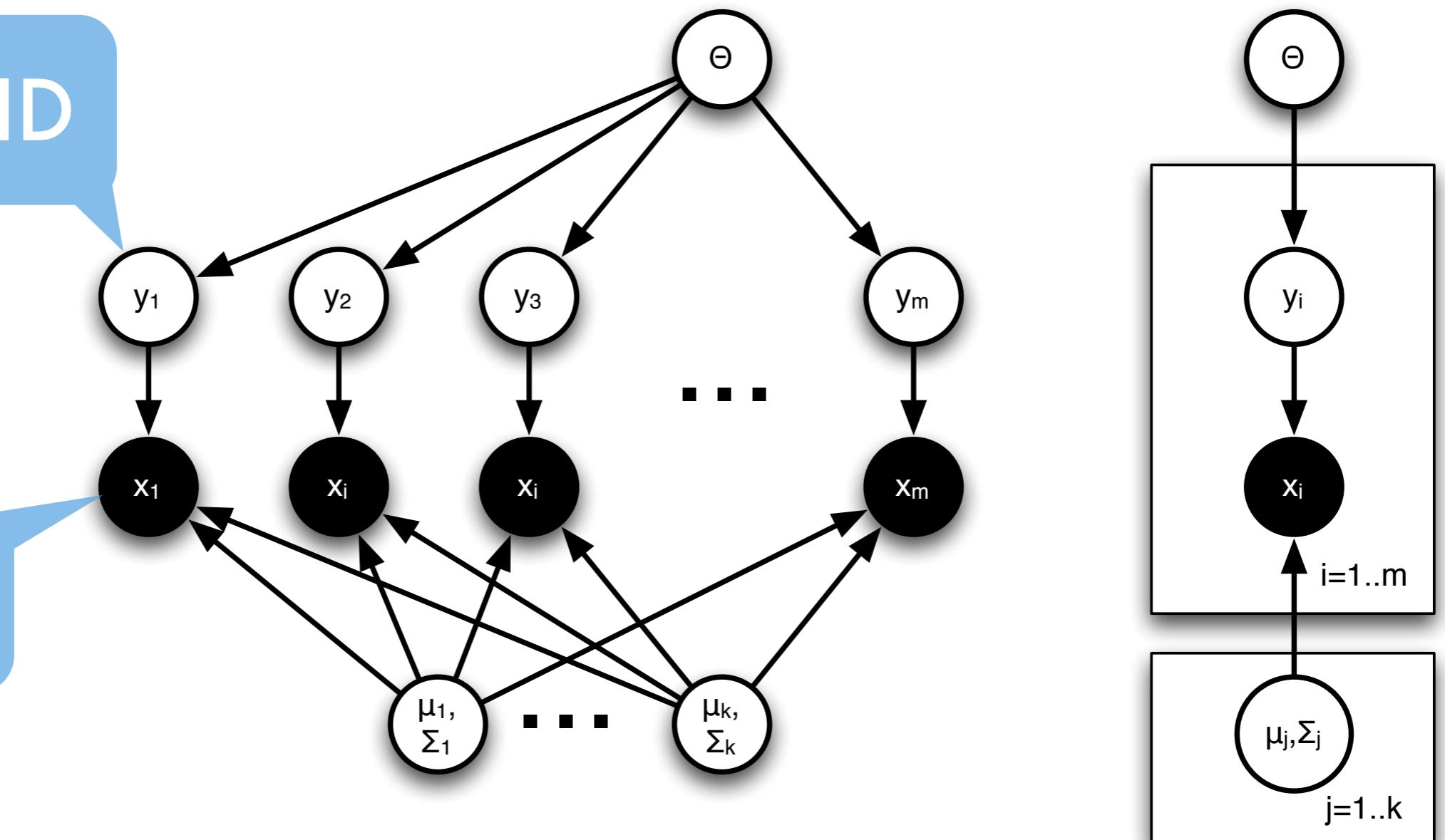
# Generative Model



# Generative Model

cluster ID

objects



$$p(X, Y | \theta, \sigma, \mu) = \prod_{i=1}^n p(x_i | y_i, \sigma, \mu) p(y_i | \theta)$$

# What can we cluster?

# What can we cluster?

The diagram illustrates various entities that can be clustered, arranged in a grid-like structure:

- Row 1:** mails, text, urls, products
- Row 2:** news, users
- Row 3:** queries, locations
- Row 4:** spammers, ads, events
- Row 5:** abuse

# Topic Models

The William Randolph Hearst Foundation will give \$1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. “Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services,” Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center’s share will be \$200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive \$400,000 each. The Juilliard School, where music and the performing arts are taught, will get \$250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual \$100,000 donation, too.

Latent Dirichlet Allocation; Blei, Ng, Jordan, JMLR 2003

# Grouping objects

# Grouping objects

The image displays three distinct web pages, each featuring the word "Singapore" in a prominent location:

- Singapore Airlines:** The top navigation bar includes a "Change Location" dropdown set to "Singapore".
- National University of Singapore (NUS):** The page title is "Singapore" and is enclosed in a large red speech bubble.
- Chijmes:** The page title is "Discover a century of resplendent living history behind the cloistered walls." and features a photograph of a historic building at night.

Below the main content, there are additional sections for staff, alumni, and visitors.

**Singapore Airlines Navigation:**

- Help | Site Map | Contact Us | Singapore | Change Location
- Search

**NUS Navigation:**

- The Experience | Flights & Fares | Before You Fly | Loyalty Programmes | Promotions
- myEMAIL | IVLE | LIBRARY | MAPS | CALENDAR | SITEMAP | CONTACT | e-CARDS
- Book a Flight | Check In
- Round Trip | One Way
- From:
- ABOUT NUS | GLOBAL | ADMISSIONS | ENTERPRISE | CAMPUS LIFE | GIVING | CAREERS@NUS
- Search  in  GO

**Chijmes Navigation:**

- Home | About Us | Services | Events & Promotions | Shopping, Wining & Dining | Contact | Sitemap
- CHIJMES restaurants • bars • shops
- Discover a century of resplendent living history behind the cloistered walls.
- Chijmes, a premier lifestyle destination in Singapore
- Owned by: SUNTEC, Managed by: ARA, Property Manager: PAC
- Feedback | Terms & Conditions

**Yahoo! Logo:**

YAHOO!

# Grouping objects

UNITED

Planning & booking | Reservations & check-in | Mileage Plus® | Services & information | Search site

United, #1 in on-time arrivals. Details

Flights Check-in Flight status

BOOK FLIGHT REDEEM MILES

From (Find airport) To (Find airport)

Search nearby airports  Search nearby airports

Roundtrip  One-way > Multicity

Departing Anytime

Returning Anytime

Search by Schedule & price Price > Flexible

Adult 1 (child or senior?)

Cabin Economy  Refundable

Promotion code or Electronic certificate More info

Log in to view all seating options

> Advanced Search Search >>

Cars Hotels Vacations

Use 30% fewer miles on your next United flight.

Save now on Saver Awards for flights 700 miles or less. Learn more

3 of 6

United news and deals

- > Travel waiver issued due to Hurricane Earl
- > E-Fares: Save on weekend getaways
- > Opt to send your bags ahead
- > Wireless check-in, paperless boarding
- > Receive deal alerts: Follow us on Twitter
- > Take our survey & you could win miles

United-Continental merger Learn more about the merger

© 1998-2010 Chijmes Restaurants

About United | Investor relations | Business resources | Careers | Site map

A STAR ALLIANCE MEMBER

Owned by: SUNTEC Managed by: Property Manager: ARA PAC

Singapore Change Location Search

Before You Fly Loyalty Programmes Promotions

CALENDAR SITEMAP CONTACT e-CARDS

Search for... in NUS Websites GO

GIVING CAREERS@NUS

Search ANU... WEB CONTACTS MAP GO



The Australian National University

CURRENT STUDENTS RESEARCH & EDUCATION ABOUT ANU STAFF

It's the spectacular natural after the Black Saturday fires typical natural

Forests renew after Black Saturday fires

School of Music at Floriade

Undergraduate studies

Higher Degree Research

# Grouping objects

The screenshot shows the United Airlines website's flight booking interface. It includes fields for 'From' and 'To' locations, departure and return dates, and search filters for roundtrip, one-way, and multi-city flights. A large promotional banner on the left offers 30% fewer miles on the next United flight. Below it, another banner for 'Saver Awards' highlights flights 700 miles or less. The page also features sections for 'United news and deals' and 'KIOSKFLYER' flight deals to various destinations like Bangkok, Hong Kong, Taipei, Tokyo, and London.



The screenshot shows the homepage of the Australian National University (ANU). The top navigation bar includes links for 'HOME', 'FUTURE STUDENTS', 'CURRENT STUDENTS', 'RESEARCH & EDUCATION', 'ABOUT ANU', and 'STAFF'. A prominent feature is a large image of a small green plant growing from a tree trunk. Below the image, there are sections for 'Ash forests rise and rise again', 'Forests renew after Black Saturday fires', and 'School of Music at Floride'. There are also tabs for 'PROSPECTIVE STUDENTS', 'CURRENT STUDENTS', 'STAFF', 'ALUMNI', and 'VISITORS'.



**YAHOO!**

# Grouping objects

The screenshot shows the United Airlines website's homepage. It features a search bar at the top with fields for 'From' and 'To'. Below the search bar, there are sections for 'BOOK FLIGHT', 'ROOM MILES', and 'Flight status'. A large orange speech bubble containing the word 'airline' is overlaid on the left side of the page. Promotional banners include one for 'Use 30% fewer miles on your next United flight.' and another for 'Save now on Saver Awards for flights 700 miles or less.' The footer contains links for 'About United', 'Investor relations', 'Business resources', 'Careers', and 'Site map'.

The screenshot shows the ANU website's homepage. It features a navigation bar with links for 'HOME', 'FUTURE STUDENTS', 'CURRENT STUDENTS', 'RESEARCH & EDUCATION', 'ABOUT ANU', and 'STAFF'. A large orange speech bubble containing the word 'university' is overlaid on the right side of the page. The main content area includes a banner for 'Ash forests rise and rise again' and a photo of students smiling.

The screenshot shows the Chez Panisse website on the left and a photograph of a restaurant interior on the right. The website features sections for 'RESERVATIONS', 'MENUS', 'ABOUT', 'SPECIAL EVENTS', 'STORE', and 'CONTACT'. The interior photo shows a rustic restaurant with wooden tables and chairs, and a sign that reads 'BAR DE LA PECHE'.

YAHOO!

# Grouping objects

The screenshot shows the United Airlines website's homepage. It features a top navigation bar with links like 'Planning & booking', 'Reservations & check-in', 'Mileage Plus', 'Services & information', and a search bar. Below this is a large promotional banner for 'Mileage Plus' with the text 'Use 30% fewer miles on your next United flight.' and a large orange percentage sign icon. To the right is a login form for 'My profile' and 'Customer service'. On the left, there's a sidebar with sections for 'RESERVATIONS', 'RESTAURANT & CAFÉ', 'MENUS', 'ABOUT', 'SPECIAL EVENTS', 'STORE', and 'CONTACT'. The main content area shows a photograph of a restaurant terrace at night.

USA

This screenshot shows the Chijmes website. At the top, there's a flight booking interface from Singapore Airlines with fields for 'From', 'To', and travel dates. Below this is a banner for 'Chijmes' with the text 'Discover living history' and 'Chijmes, a premier lifestyle destination in Singapore'. The Chijmes building is shown in a night photograph. The footer contains copyright information and links to 'Feedback | Terms & Conditions'.

Singapore

The screenshot shows the homepage of the Australian National University (ANU). It features the ANU logo and the text 'The Australian National University'. A large orange speech bubble on the right side contains the word 'Australia'. The page includes a search bar, links for 'HOME', 'FUTURE STUDENTS', and various news articles. The footer has links for 'Forests renew after Black Saturday fires', 'School of Music at Floriade', 'Undergraduate studies', and 'Higher Degree Research'.

Australia

YAHOO!

# Topic Models

USA  
airline

The United Airlines website interface includes a header with the United logo, navigation links for Planning & Booking, Reservations & check-in, Mileage Plus, Services & Information, and a search bar. A prominent banner on the left offers 30% fewer miles on flights. The main content area shows flight search fields for departure and return dates, cabin class, and search criteria like round trip or one-way.

Australia  
university

The homepage of The Australian National University (ANU) features the ANU logo and a banner headline "Ash forests rise and rise again". Below this, there are sections for Future Students, Current Students, and About ANU. A central image shows a small green plant growing from a tree trunk. Navigation links include HOME, FUTURE STUDENTS, CURRENT STUDENTS, and ABOUT ANU.

Singapore  
airline

The Singapore Airlines website displays a search interface for "Book a Flight" with fields for departure and destination cities, travel dates, and passenger information. A sidebar lists various international flight options with prices, such as Singapore to Bangkok (SGD 395), Hong Kong (SGD 546), Taipei (SGD 768), Tokyo (SGD 983), and London (SGD 1,288). The page also includes a "Member Log-in" section and a "KrisFlyer" rewards program summary.

Singapore  
university

The NUS website features the NUS logo and a search bar. The main menu includes links for ABOUT NUS, GLOBAL, ADMISSIONS, EDUCATION, RESEARCH, ENTERPRISE, CAMPUS LIFE, GIVING, and CAREERS@NUS. A large banner image shows students smiling outdoors. Below the banner, there are tabs for PROSPECTIVE STUDENTS, CURRENT STUDENTS, STAFF, ALUMNI, and VISITORS. A news item about a joint evacuation exercise is visible on the right.

Chez Panisse

The Chez Panisse website includes sections for RESERVATIONS, MENUS, ABOUT, SPECIAL EVENTS, STORE, and CONTACT. A large image of the restaurant's exterior is on the left. The menu section shows a list of items under categories like RESTAURANT & CAFE and MONDAY NIGHTS.

USA  
food

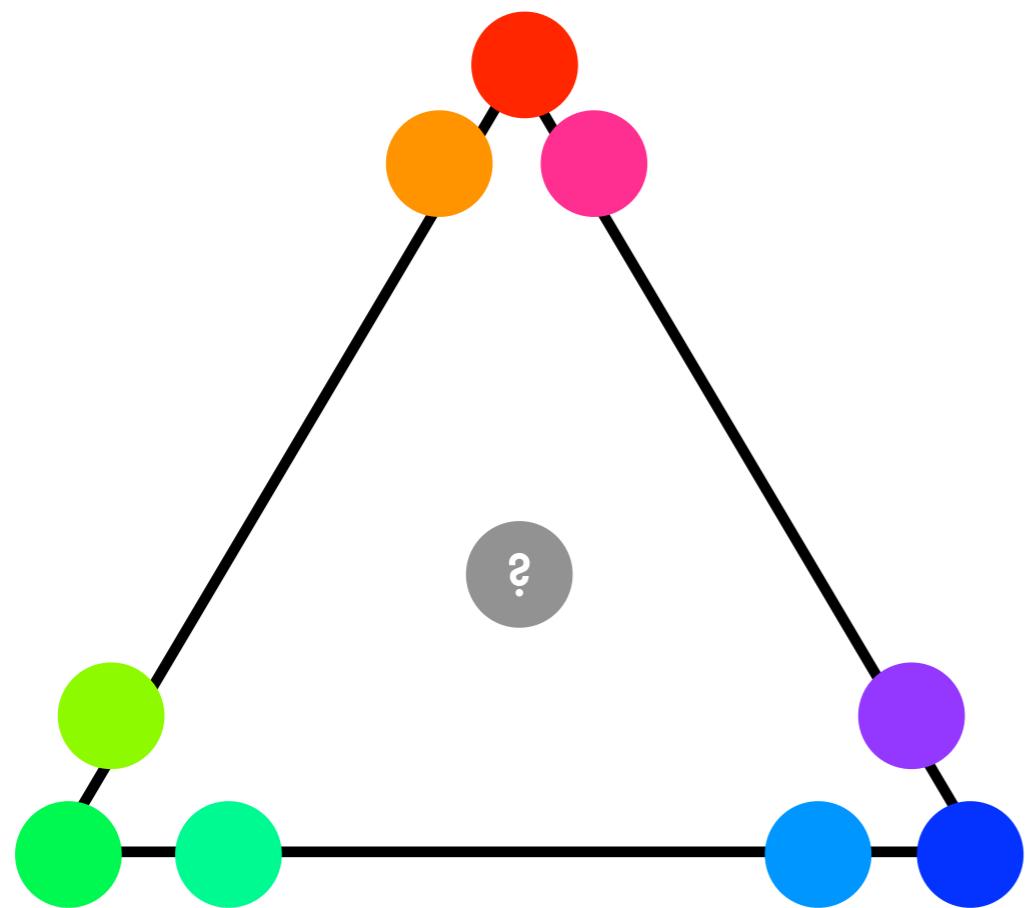
This part of the Chez Panisse website shows a reservation form with fields for Name, Email, Phone, and Date. Below the form, there are images of the restaurant's interior, including a view of the kitchen and a couple in the dining room. The URL "www.chezpanisse.com/reserve" is visible at the bottom.

Singapore  
food

The Chijmes website highlights "Discover a century of resplendent living history behind the cloisters" and "Chijmes, a premier lifestyle destination in Singapore". It lists partners including SUNTEC, ARA, and APC. The page includes a footer with copyright information and links to Feedback and Terms & Conditions.

# Clustering & Topic Models

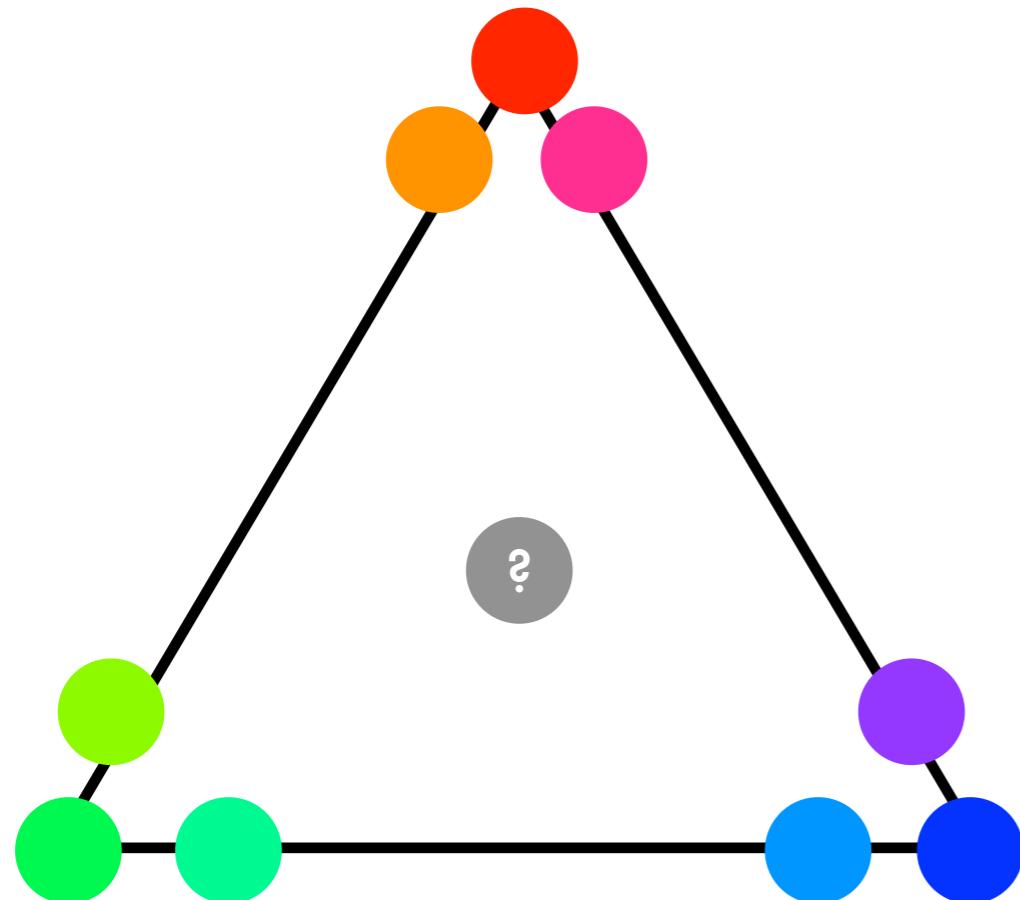
## Clustering



group objects  
by prototypes

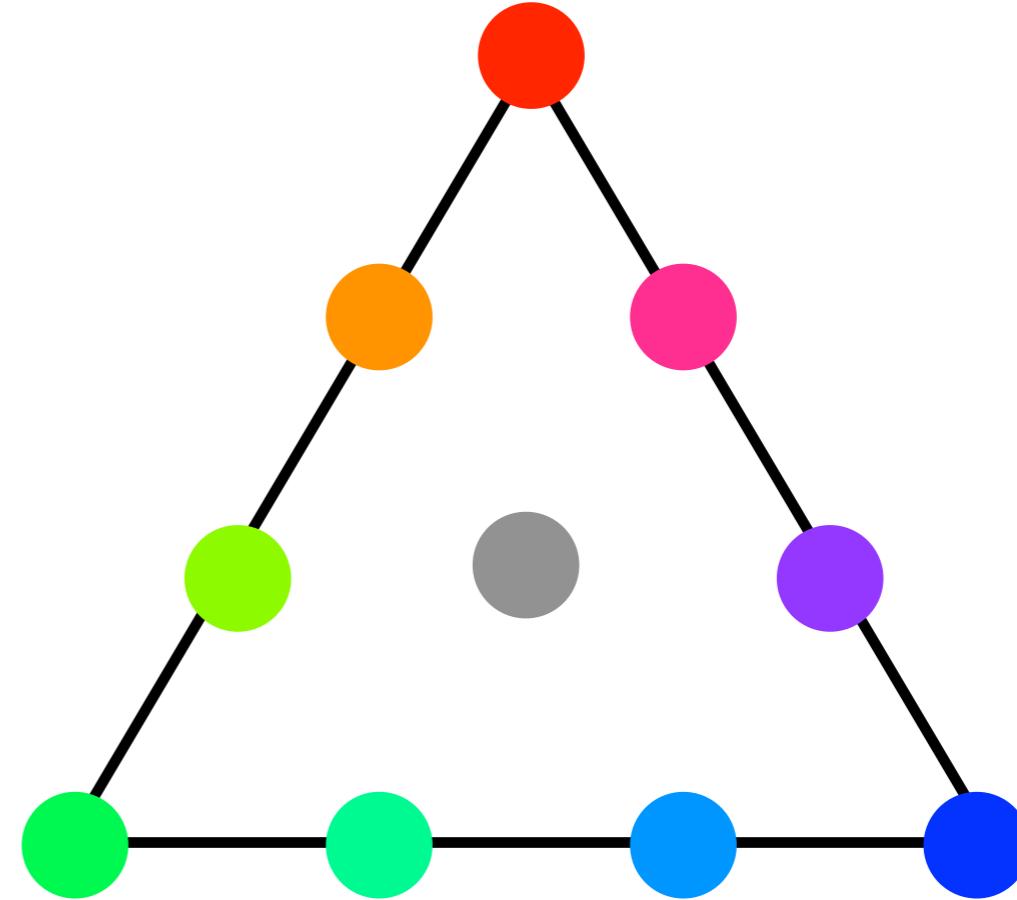
# Clustering & Topic Models

Clustering



group objects  
by prototypes

Topics

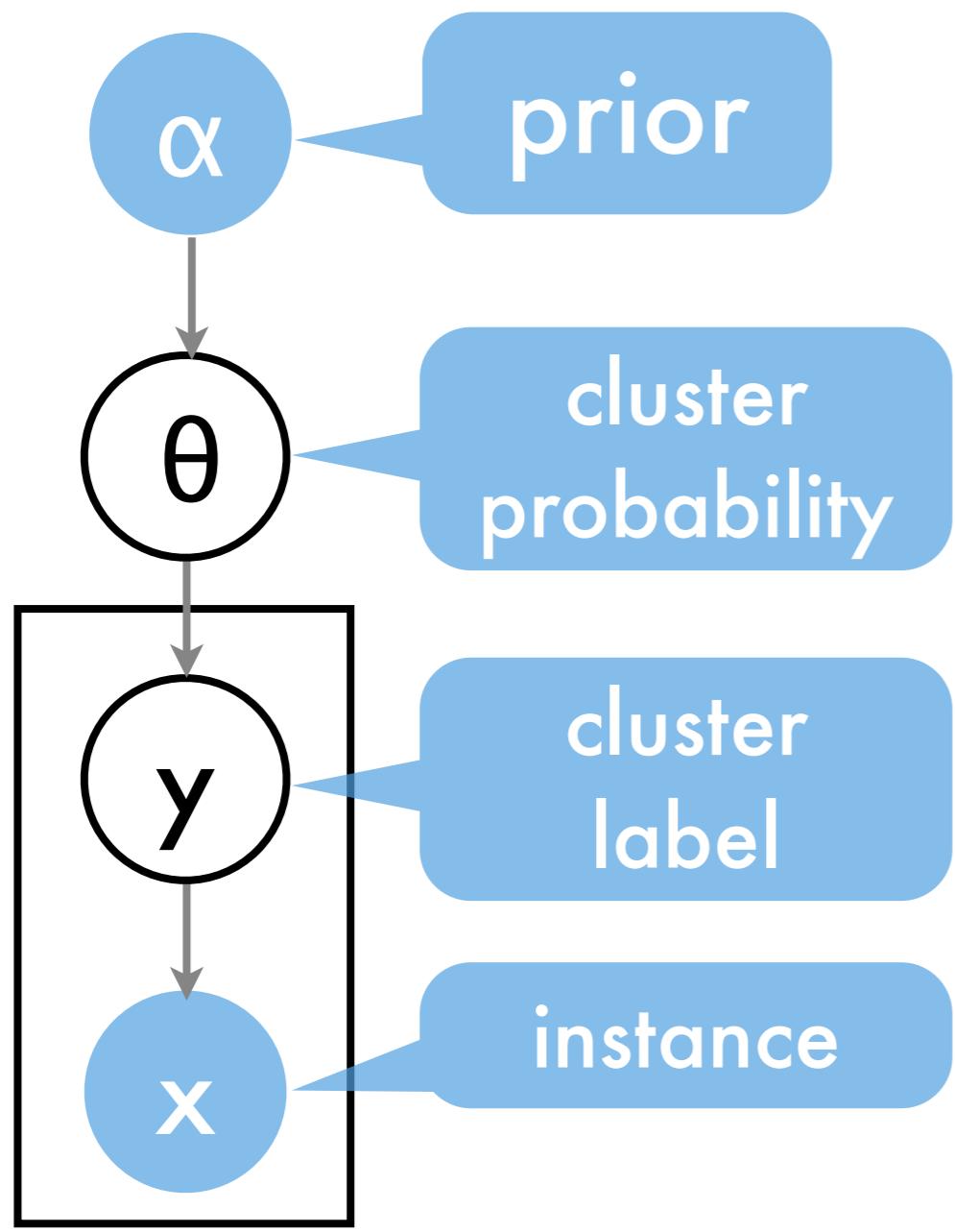


decompose objects  
into prototypes

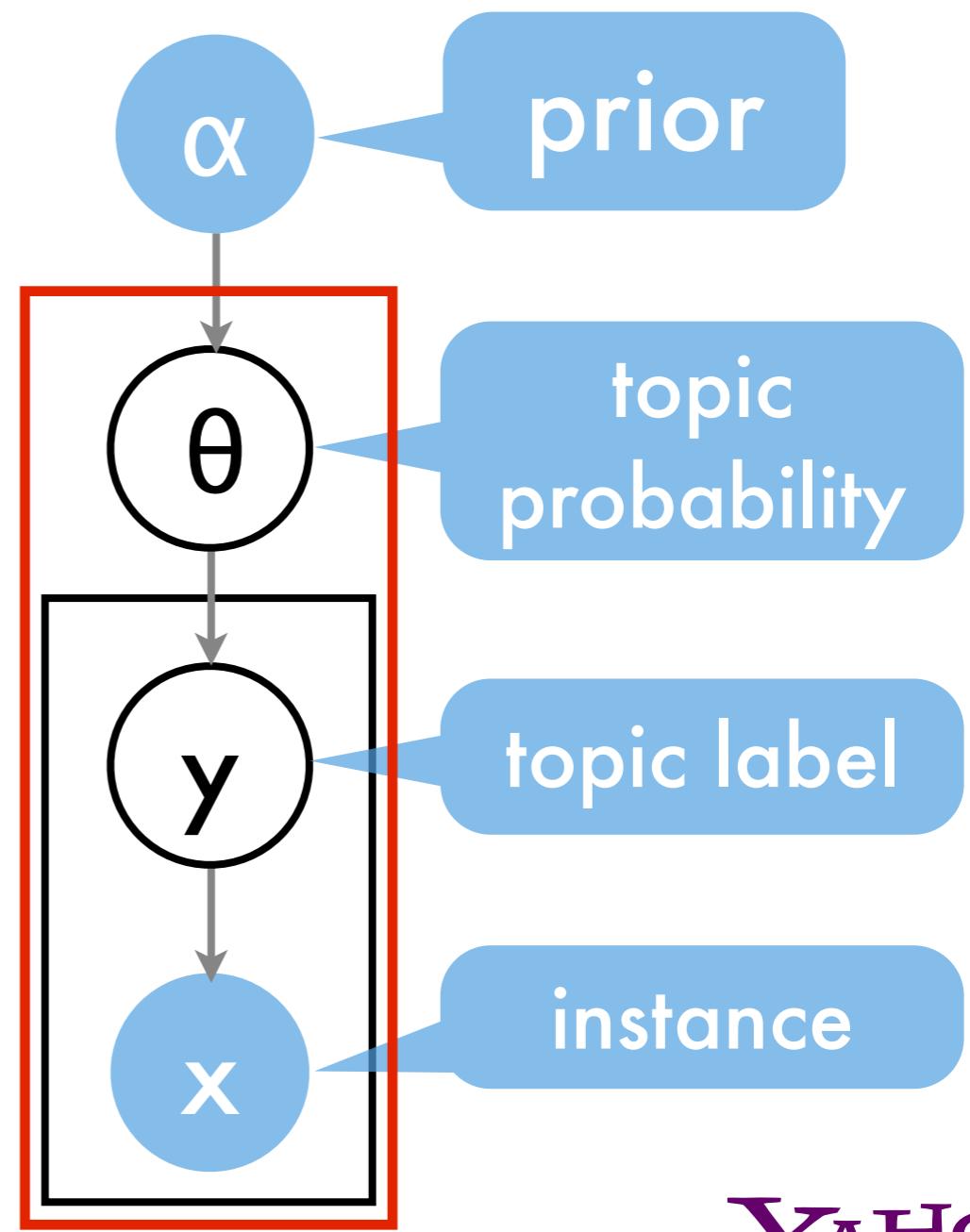
YAHOO!

# Clustering & Topic Models

clustering

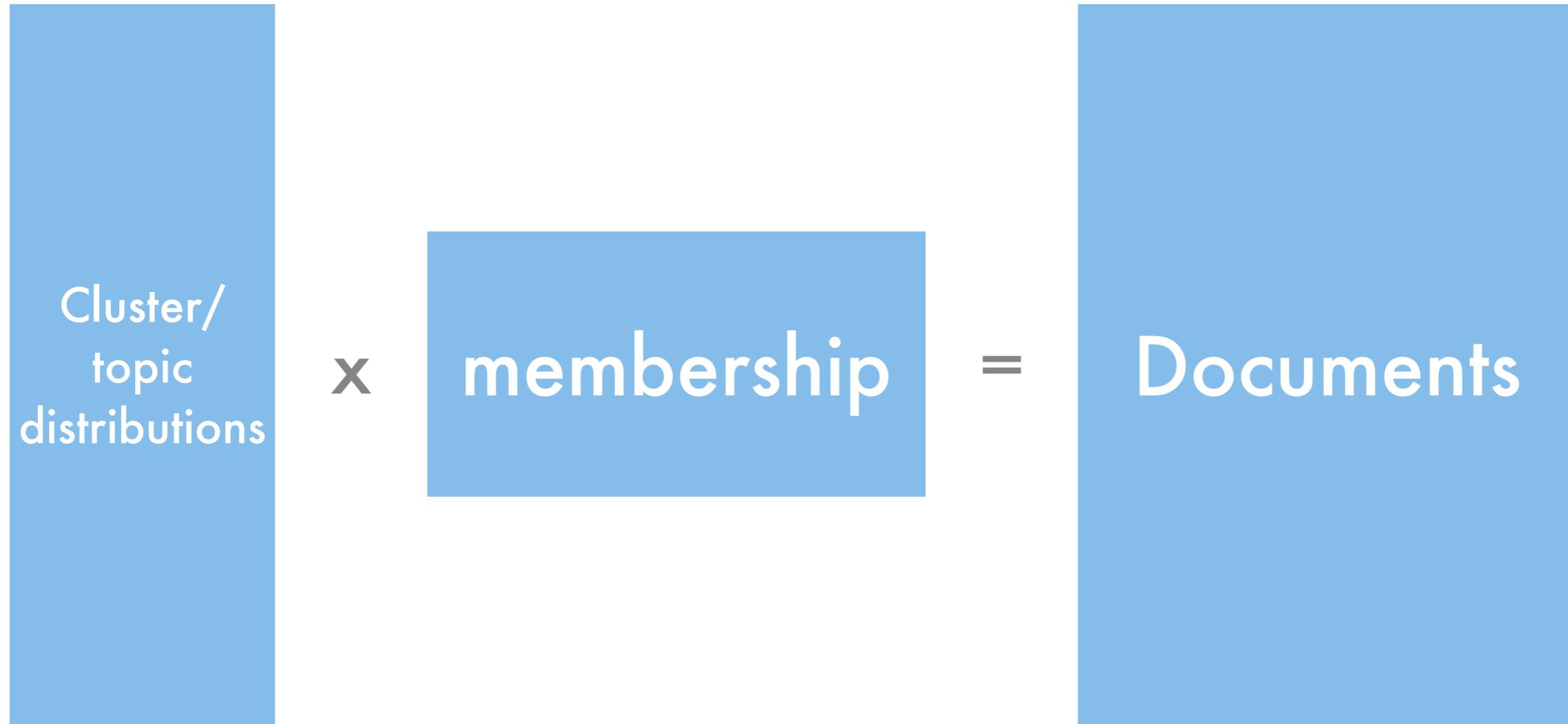


Latent Dirichlet Allocation



YAHOO!

# Clustering & Topic Models



clustering: (0, 1) matrix

topic model: stochastic matrix

LSI: arbitrary matrices

YAHOO!

# Many more

- Regression  
inventory, traffic, reserve price, elasticity
- Novelty detection  
abuse, change in traffic, server farm
- Entity tagging  
keywords, named entities, segmentation
- Collaborative filtering  
recommend related movies, books, songs
- Inferring structure from data  
trees, DAGs, segmentation boundaries, user models

# Optimization & inference problems (horrible oversimplification)

- Supervised problems

$$\underset{w}{\text{minimize}} \sum_{i=1}^m l(x_i, y_i, w) + \lambda \|w\|^\alpha$$

goodness of fit

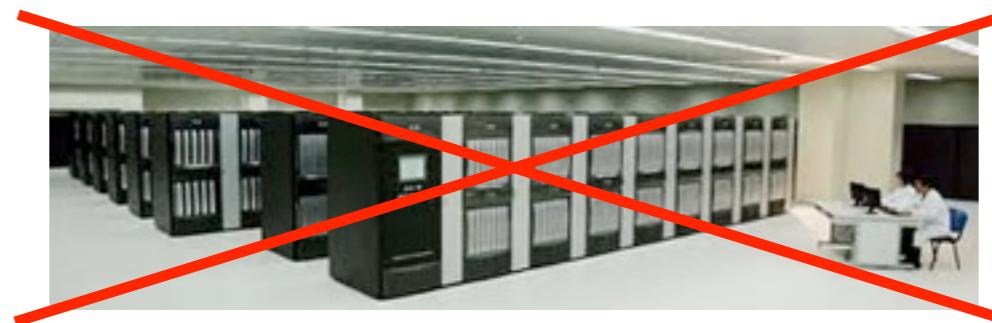
complexity penalty

- convex problem
- solve subproblem and merge works well
- Unsupervised problems
  - nonconvex problem (looks similar)
  - fast synchronization required



# Hardware

- NOT High Performance Computing



- Consumer hardware  
**Cheap, efficient, not very reliable**



# The Joys of Real Hardware

Typical first year for a new cluster:

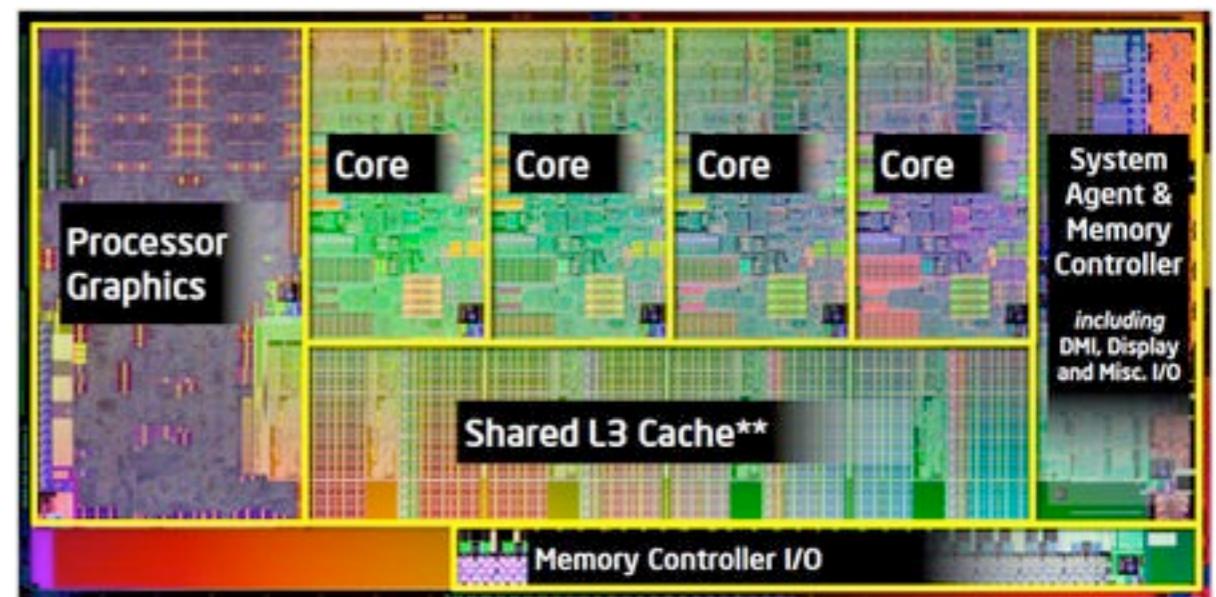
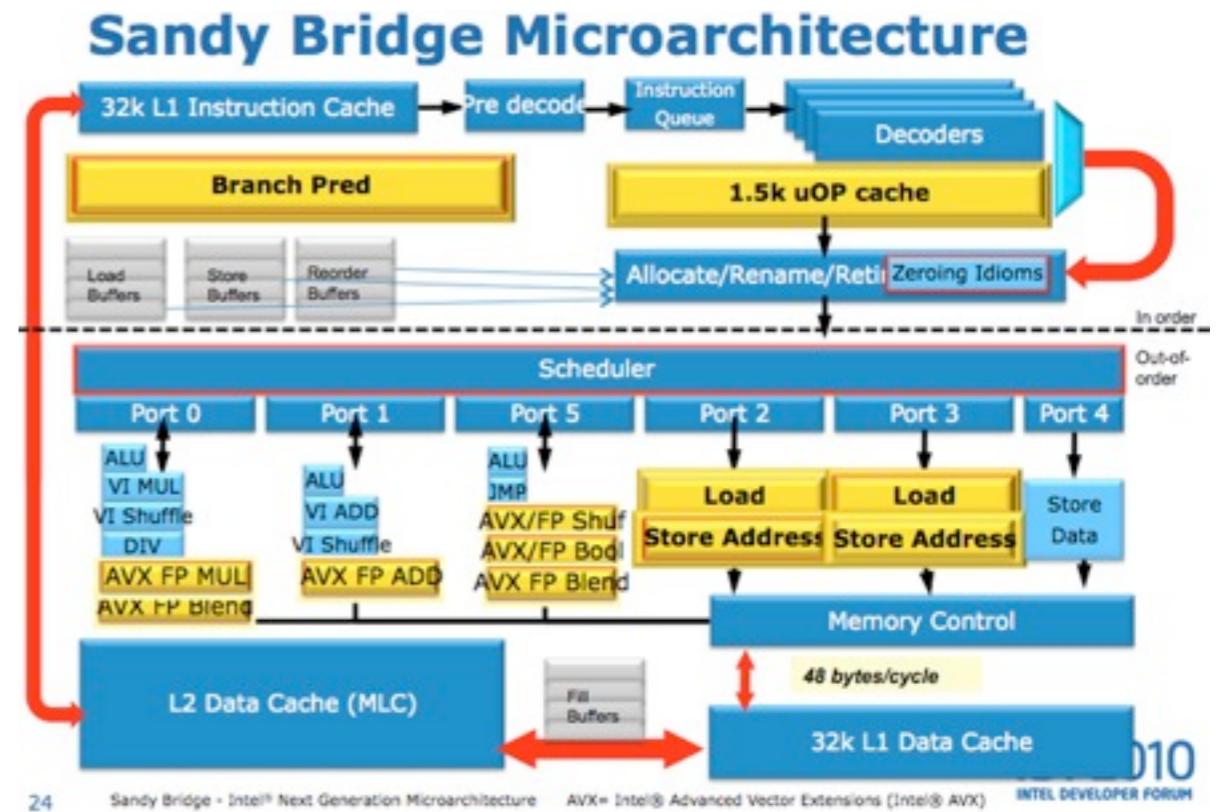
- ~0.5 overheating (power down most machines in <5 mins, ~1-2 days to recover)
  - ~1 PDU failure (~500-1000 machines suddenly disappear, ~6 hours to come back)
  - ~1 rack-move (plenty of warning, ~500-1000 machines powered down, ~6 hours)
  - ~1 network rewiring (rolling ~5% of machines down over 2-day span)
  - ~20 rack failures (40-80 machines instantly disappear, 1-6 hours to get back)
  - ~5 racks go wonky (40-80 machines see 50% packetloss)
  - ~8 network maintenances (4 might cause ~30-minute random connectivity losses)
  - ~12 router reloads (takes out DNS and external vips for a couple minutes)
  - ~3 router failures (have to immediately pull traffic for an hour)
  - ~dozens of minor 30-second blips for dns
  - ~1000 individual machine failures
  - ~thousands of hard drive failures
- slow disks, bad memory, misconfigured machines, flaky machines, etc.

Slide from talk of Jeff Dean



# CPU

- 8-32 cores
- Memory interface  
20-60GB/s
- Internal bandwidth  
>100GB/s
- >100 GFlops for matrix  
matrix multiply
- Integrated low end GPU



# RAM

- High latency (100ns for DDR3)
- High burst data rate (>10 GB/s)



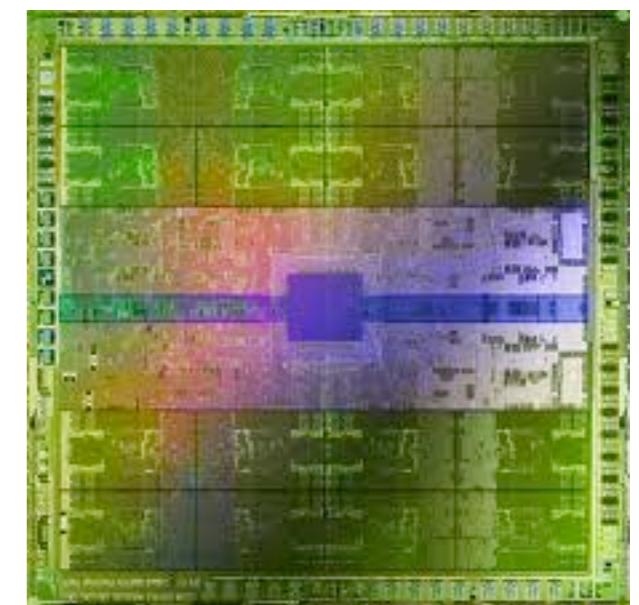
- Avoid random access in code if possible.
- Memory align variables
- Know your platform (FBDIMM vs. DDR)



<http://www.anandtech.com/show/3851/everything-you-always-wanted-to-know-about-sdram-memory-but-were-afraid-to-ask>

# GPU

- Up to 512 cores / **200W**
- Tricky to synchronize threads
- 1-3GB memory (Tesla 6GB)
- 1 TFlop
- Memory bandwidth > 100GB/s
- **4GB/s PCI bus bottleneck**



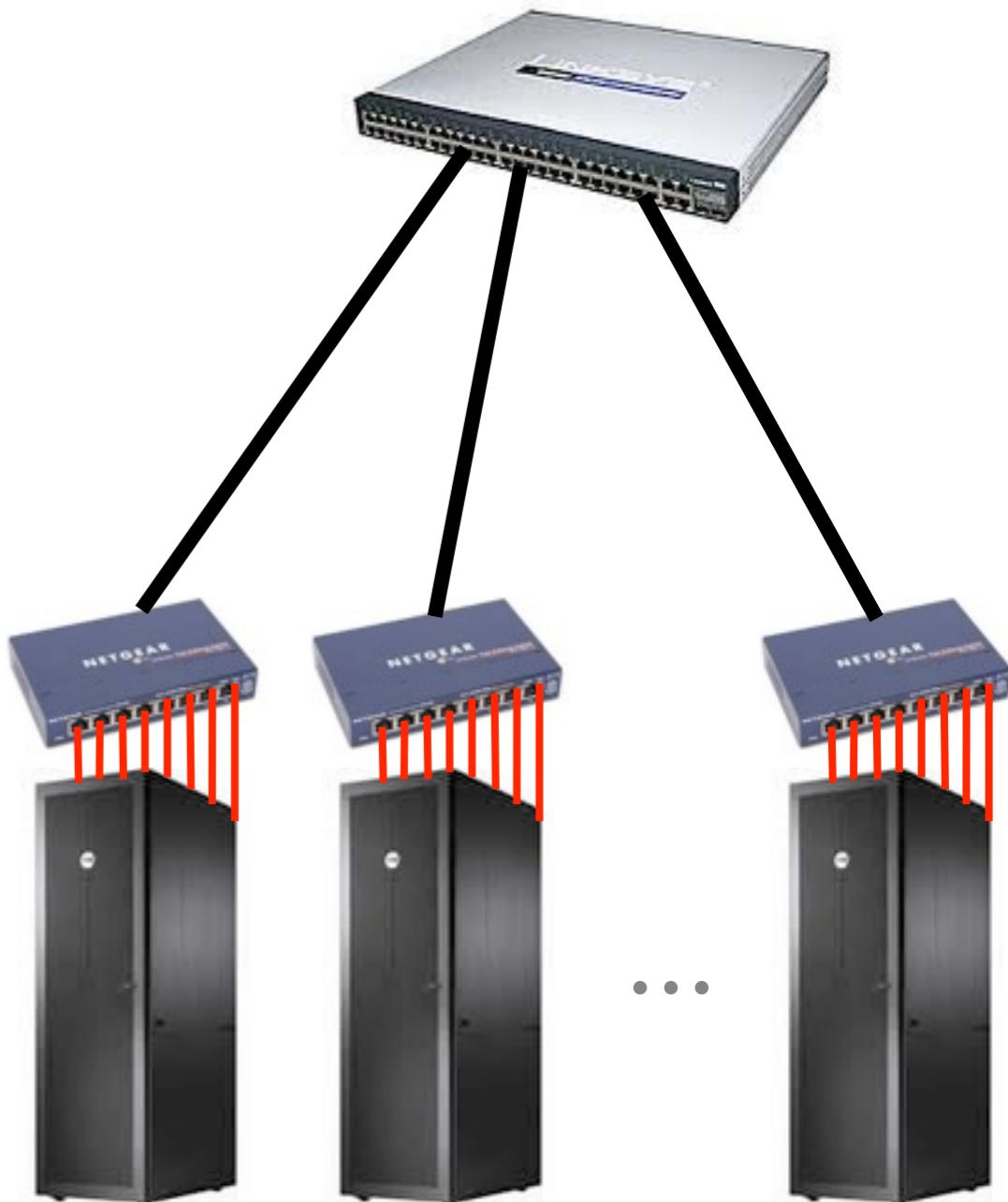
# Storage

- Harddisks
  - 3TB of storage (30MB/\$)
  - 100 MB/s bandwidth (sequential)
  - 5 ms seek (200 IOPS)
- SSD
  - 100-500 MB storage (1MB/\$)
  - 300 MB/s bandwidth (sequential)
  - 50,000 IOPS / 1 ms seek (queueing)



# Switches & Colos

- Big switches are expensive
- Switches have finite buffers
  - many connections to single machine
  - dropped packets / collisions
- Hierarchical structure
  - more bandwidth within rack
  - lower latency within rack
  - lots of latency between colos



recent development on 'flat' networks

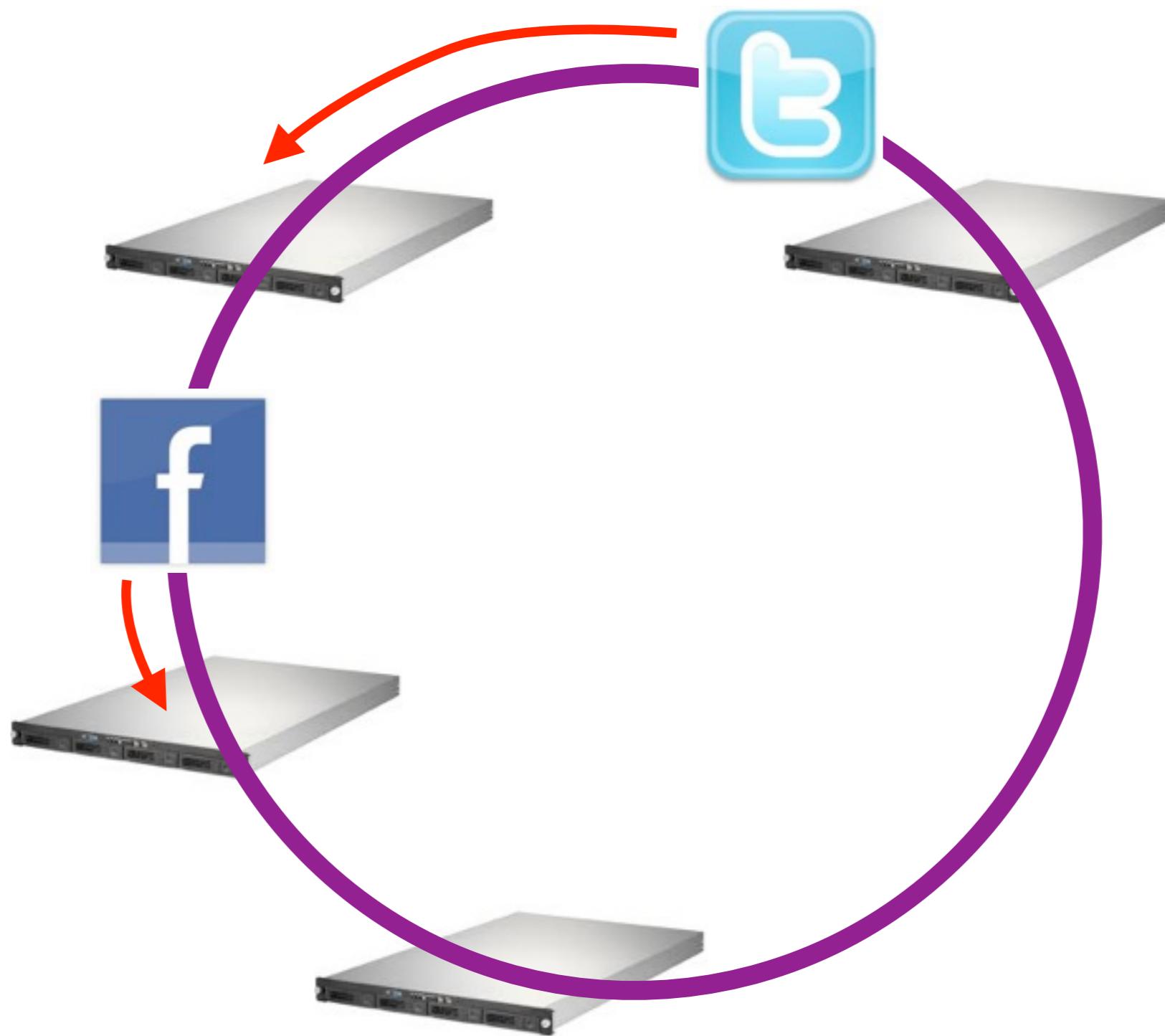
# Numbers Everyone Should Know

L1 cache reference	0.5 ns
Branch mispredict	5 ns
L2 cache reference	7 ns
Mutex lock/unlock	100 ns
Main memory reference	100 ns
Compress 1K bytes with Zippy	10,000 ns
Send 2K bytes over 1 Gbps network	20,000 ns
Read 1 MB sequentially from memory	250,000 ns
Round trip within same datacenter	500,000 ns
Disk seek	10,000,000 ns
Read 1 MB sequentially from network	10,000,000 ns
Read 1 MB sequentially from disk	30,000,000 ns
Send packet CA->Netherlands->CA	150,000,000 ns

Slide from talk of Jeff Dean



# Distribution and Balancing



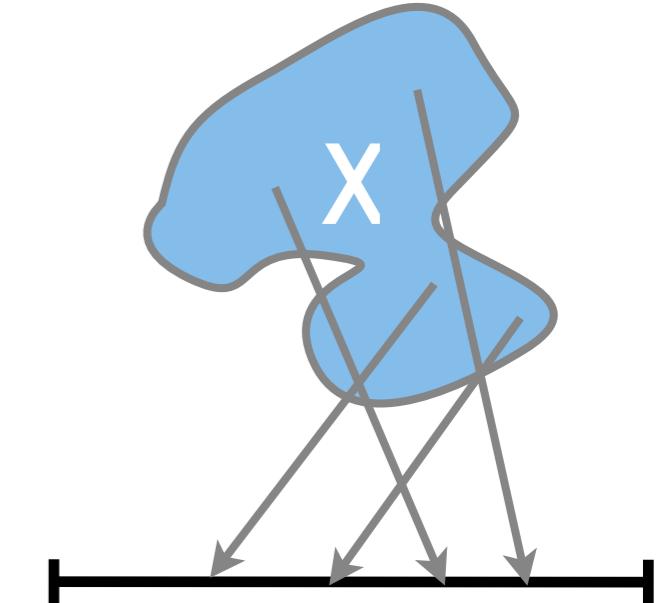
# Concepts

- Large number of objects (a priori unknown)
- Large pool of machines (often faulty)
- Assign objects to machines such that
  - Object goes to the same machine (if possible)
  - Machines can be added/fail dynamically
- Consistent hashing (elements, sets, proportional)

**symmetric (no master), dynamically scalable, fault tolerant**

# Hash function

- Mapping from domain  $X$  to integer range  $[1..N]$
- Indistinguishable from uniform distribution
- $n$ -ways independent hash function
  - Draw  $h$  from set hash functions  $H$  at random
  - For  $n$  instances in  $X$  their hash  $[h(x_1), \dots h(x_n)]$  is essentially indistinguishable from  $n$  random draws from  $[1 \dots N]$
- For many cases we only need 2-ways independence



$$\text{for all } x, y \Pr_{y \in H} \{h(x) = h(y)\} = \frac{1}{N}$$

- In practice use MD5 or Murmur Hash for high quality  
<https://code.google.com/p/smhasher/>

# Argmin Hash

- Consistent hashing

$$m(\text{key}) = \operatorname{argmin}_{m \in \mathcal{M}} h(\text{key}, m)$$

- Uniform distribution over machine pool  $\mathcal{M}$
- Fully determined by hash function  $h$ . No need to ask master
- If we add/remove machine  $m'$  all but  $O(1/m)$  keys remain

$$\Pr \{m(\text{key}) = m'\} = \frac{1}{m}$$

- Consistent hashing with  $k$  replications

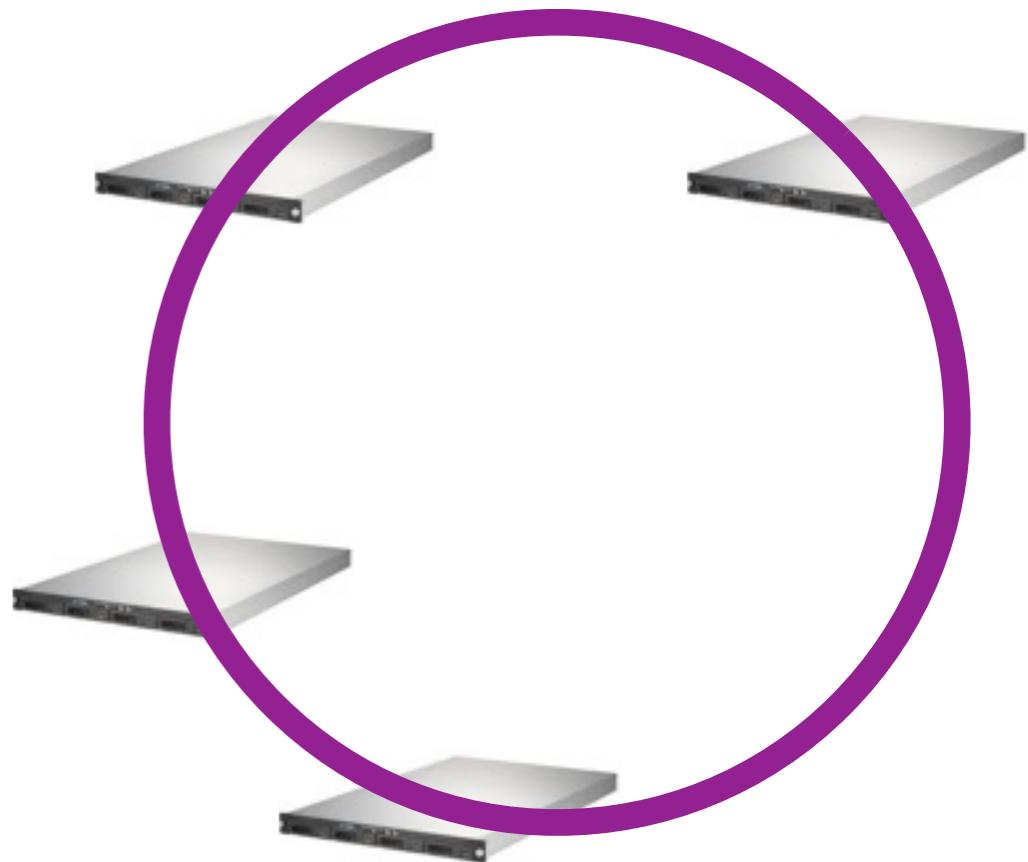
$$m(\text{key}, k) = k \operatorname{smallest}_{m \in \mathcal{M}} h(\text{key}, m)$$

- If we add/remove a machine only  $O(k/m)$  need reassigning
- Cost to assign is  $O(m)$ . This can be expensive for 1000 servers

# Distributed Hash Table

- Fixing the  $O(m)$  lookup
  - Assign machines to ring via hash  $h(m)$
  - Assign keys to ring
  - Pick machine nearest to key to the left
- $O(\log m)$  lookup
- Insert/removal only affects neighbor  
**(however, big problem for neighbor)**
- Uneven load distribution  
**(load depends on segment size)**
- Insert machine more than once to fix this
- For  $k$  term replication, simply pick the  $k$  leftmost machines (skip duplicates)

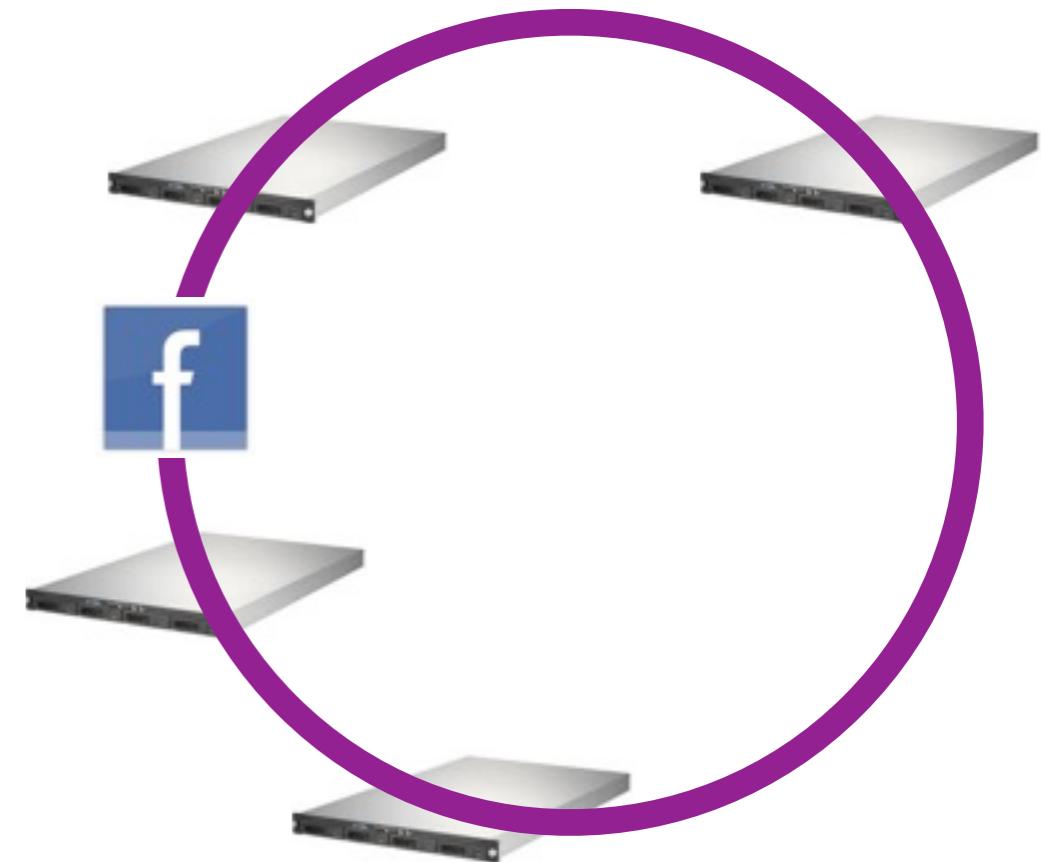
ring of  $N$  keys



# Distributed Hash Table

- Fixing the  $O(m)$  lookup
  - Assign machines to ring via hash  $h(m)$
  - Assign keys to ring
  - Pick machine nearest to key to the left
- $O(\log m)$  lookup
- Insert/removal only affects neighbor  
**(however, big problem for neighbor)**
- Uneven load distribution  
**(load depends on segment size)**
- Insert machine more than once to fix this
- For  $k$  term replication, simply pick the  $k$  leftmost machines (skip duplicates)

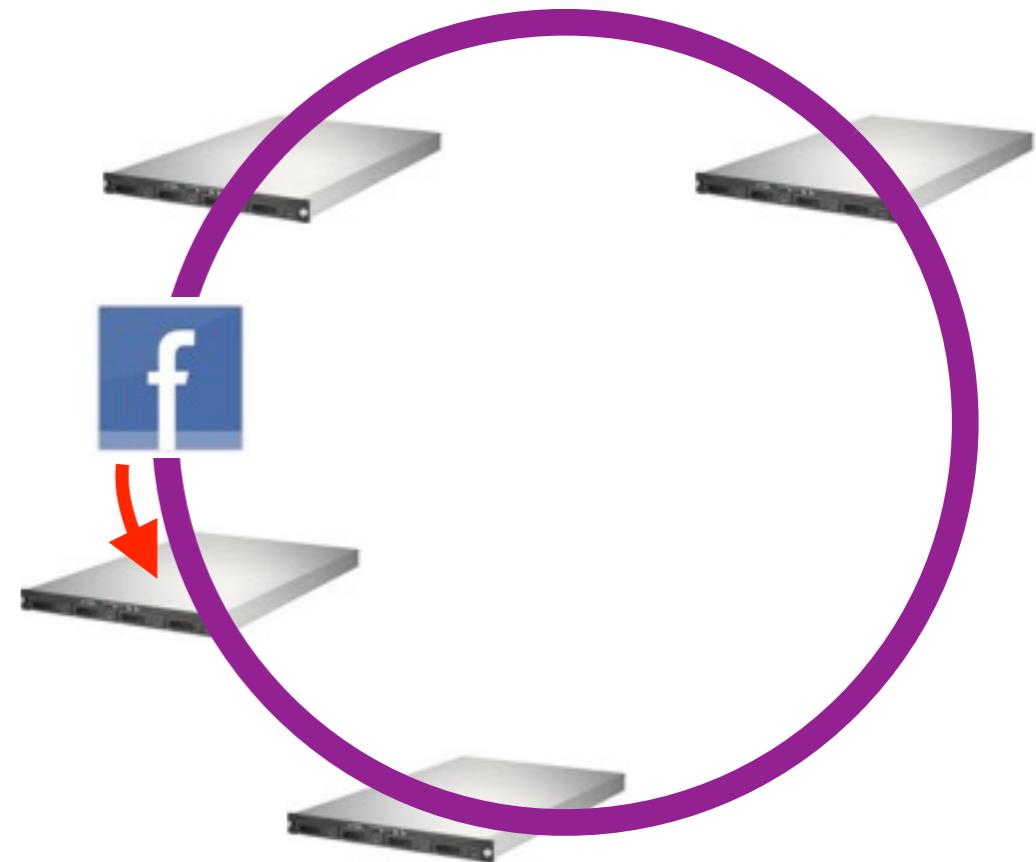
ring of  $N$  keys



# Distributed Hash Table

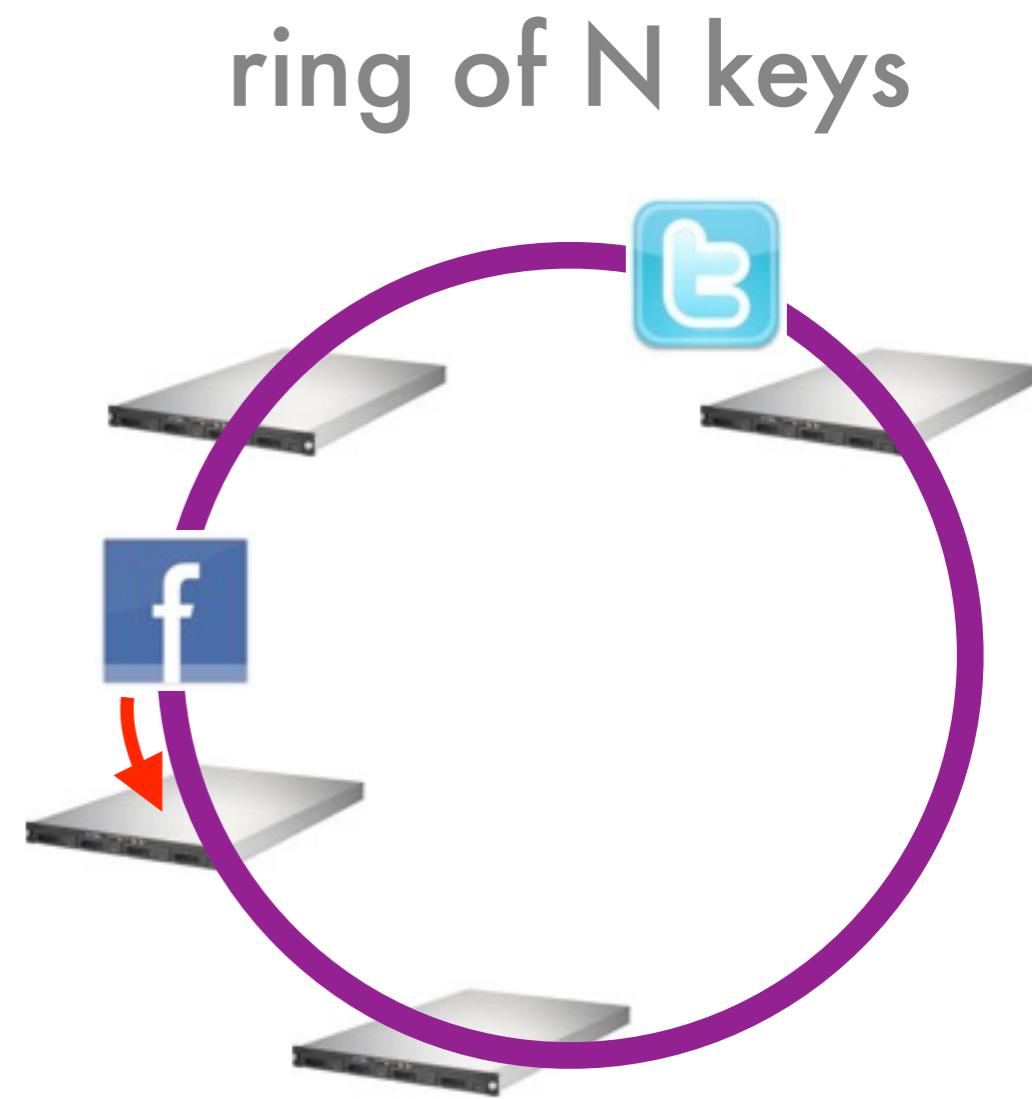
- Fixing the  $O(m)$  lookup
  - Assign machines to ring via hash  $h(m)$
  - Assign keys to ring
  - Pick machine nearest to key to the left
- $O(\log m)$  lookup
- Insert/removal only affects neighbor  
**(however, big problem for neighbor)**
- Uneven load distribution  
**(load depends on segment size)**
- Insert machine more than once to fix this
- For  $k$  term replication, simply pick the  $k$  leftmost machines (skip duplicates)

ring of  $N$  keys



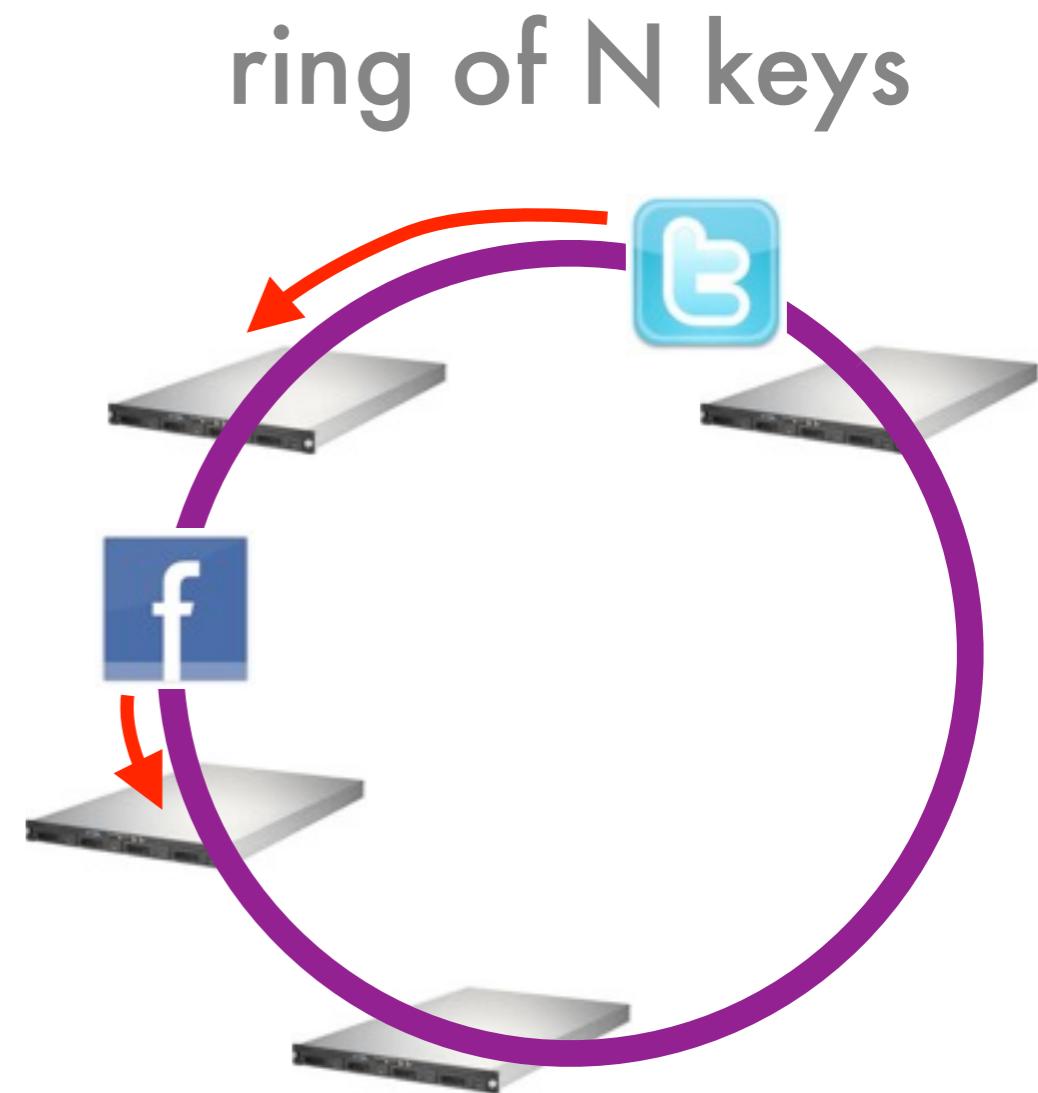
# Distributed Hash Table

- Fixing the  $O(m)$  lookup
  - Assign machines to ring via hash  $h(m)$
  - Assign keys to ring
  - Pick machine nearest to key to the left
- $O(\log m)$  lookup
- Insert/removal only affects neighbor  
**(however, big problem for neighbor)**
- Uneven load distribution  
**(load depends on segment size)**
- Insert machine more than once to fix this
- For  $k$  term replication, simply pick the  $k$  leftmost machines (skip duplicates)



# Distributed Hash Table

- Fixing the  $O(m)$  lookup
  - Assign machines to ring via hash  $h(m)$
  - Assign keys to ring
  - Pick machine nearest to key to the left
- $O(\log m)$  lookup
- Insert/removal only affects neighbor  
**(however, big problem for neighbor)**
- Uneven load distribution  
**(load depends on segment size)**
- Insert machine more than once to fix this
- **For  $k$  term replication, simply pick the  $k$  leftmost machines (skip duplicates)**



# D2 - Distributed Hash Table

- For arbitrary node segment size is minimum over  $(m-1)$  independent uniformly distributed random variables

$$\Pr \{x \geq c\} = \prod_{i=2}^m \Pr \{s_i \geq c\} = (1 - c)^{m-1}$$

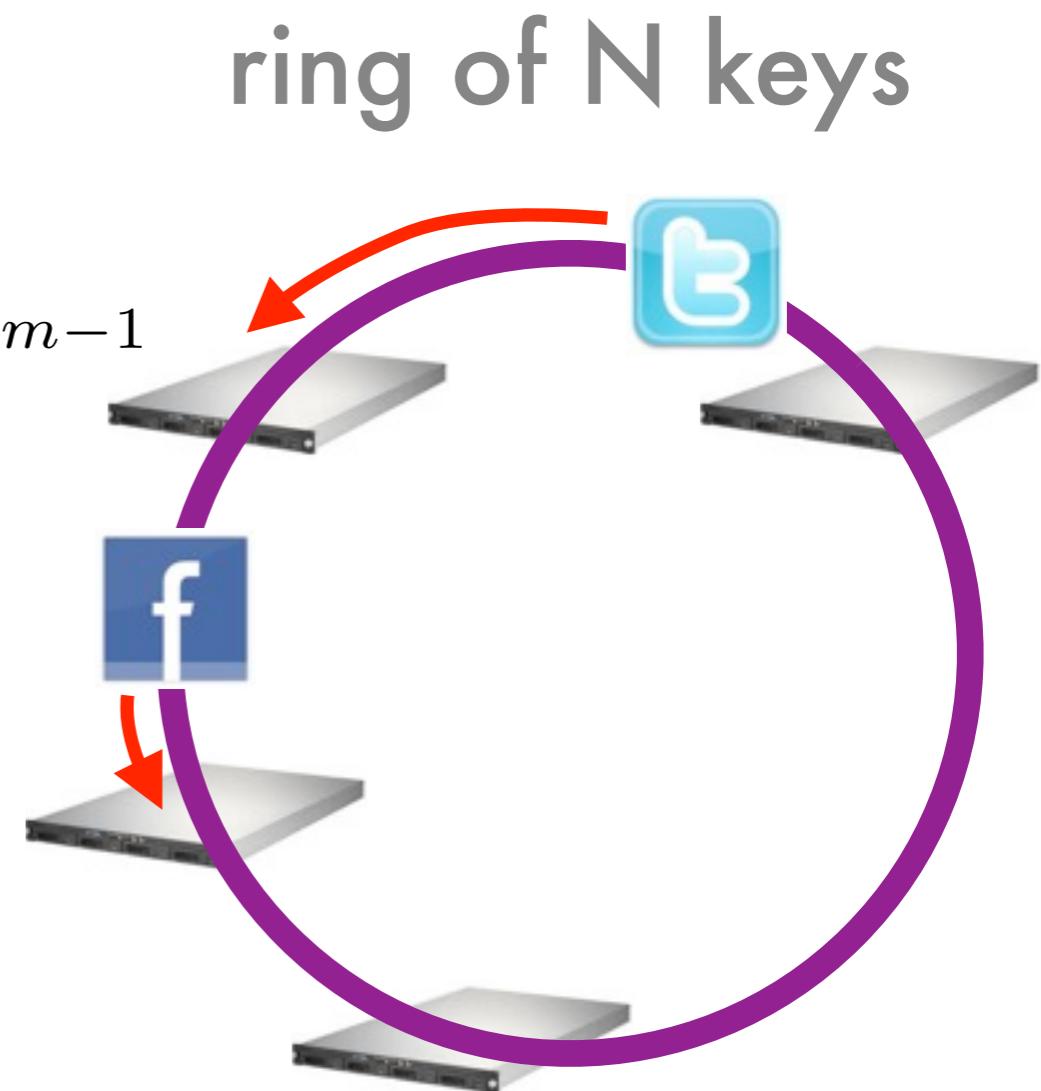
- Density is given by derivative

$$p(c) = (m-1)(1 - c)^{m-2}$$

- Expected segment length is  $c = \frac{1}{m}$  (follows from symmetry)

- Probability of exceeding expected segment length (for large m)

$$\Pr \left\{ x \geq \frac{k}{m} \right\} = \left( 1 - \frac{k}{m} \right)^{m-1} \rightarrow e^{-k}$$



# Proportional Allocation Table

- Assign items according to machine capacity
  - Create allocation table with segments proportional to capacity
  - Leave space for additional machines
  - Hash key  $h(x)$  and pick machine covering it
  - If failure, re-hash the hash until it hits a bin
  - For replication hit  $k$  bins in a row
- 
- Proportional load distribution
  - Limited scalability
  - Need to distribute and update table
  - Limit peak load by further delegation  
(SPOCA - Chawla et al., USENIX 2011)

1

2

3

4

# Proportional Allocation Table

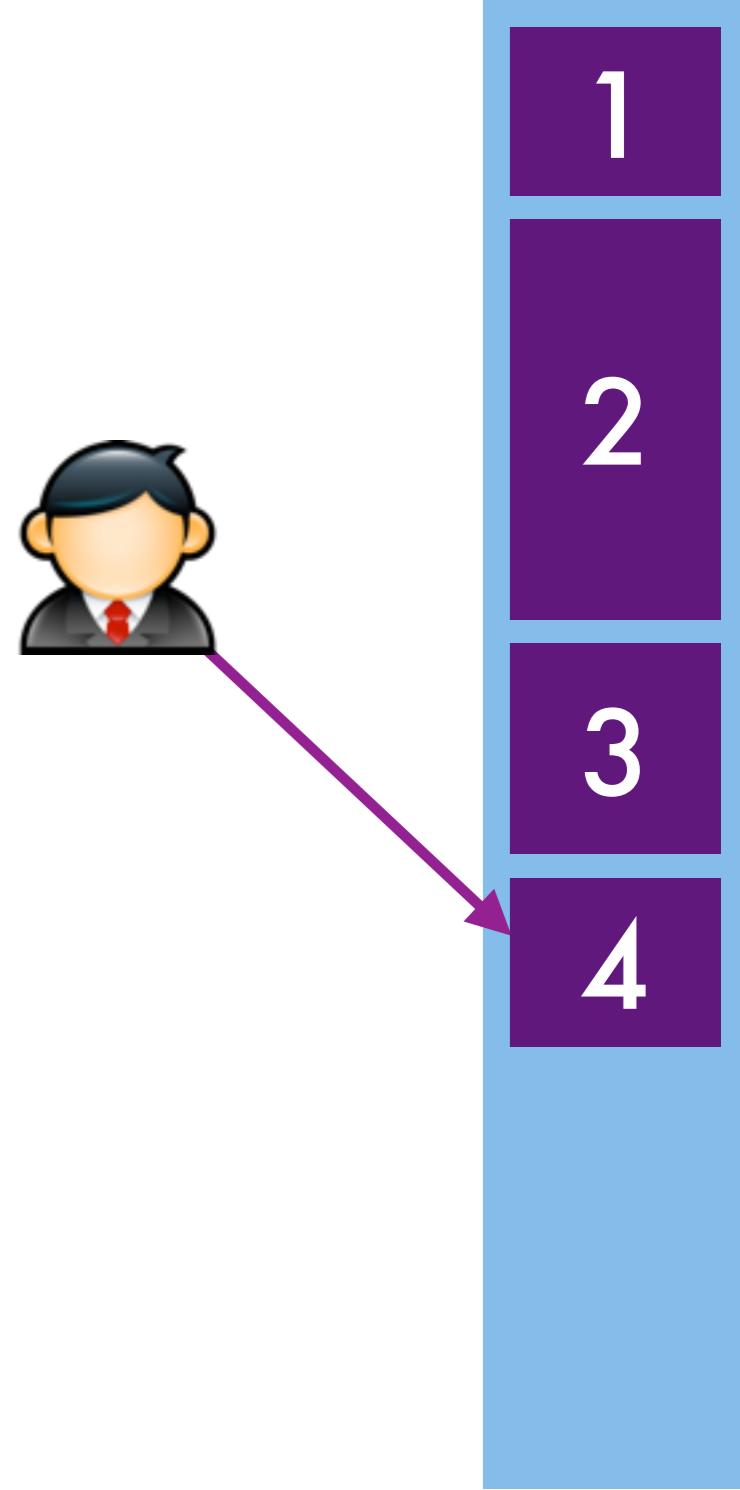
- Assign items according to machine capacity
  - Create allocation table with segments proportional to capacity
  - Leave space for additional machines
  - Hash key  $h(x)$  and pick machine covering it
  - If failure, re-hash the hash until it hits a bin
  - For replication hit  $k$  bins in a row
- 
- Proportional load distribution
  - Limited scalability
  - Need to distribute and update table
  - Limit peak load by further delegation  
(SPOCA - Chawla et al., USENIX 2011)



1  
2  
3  
4

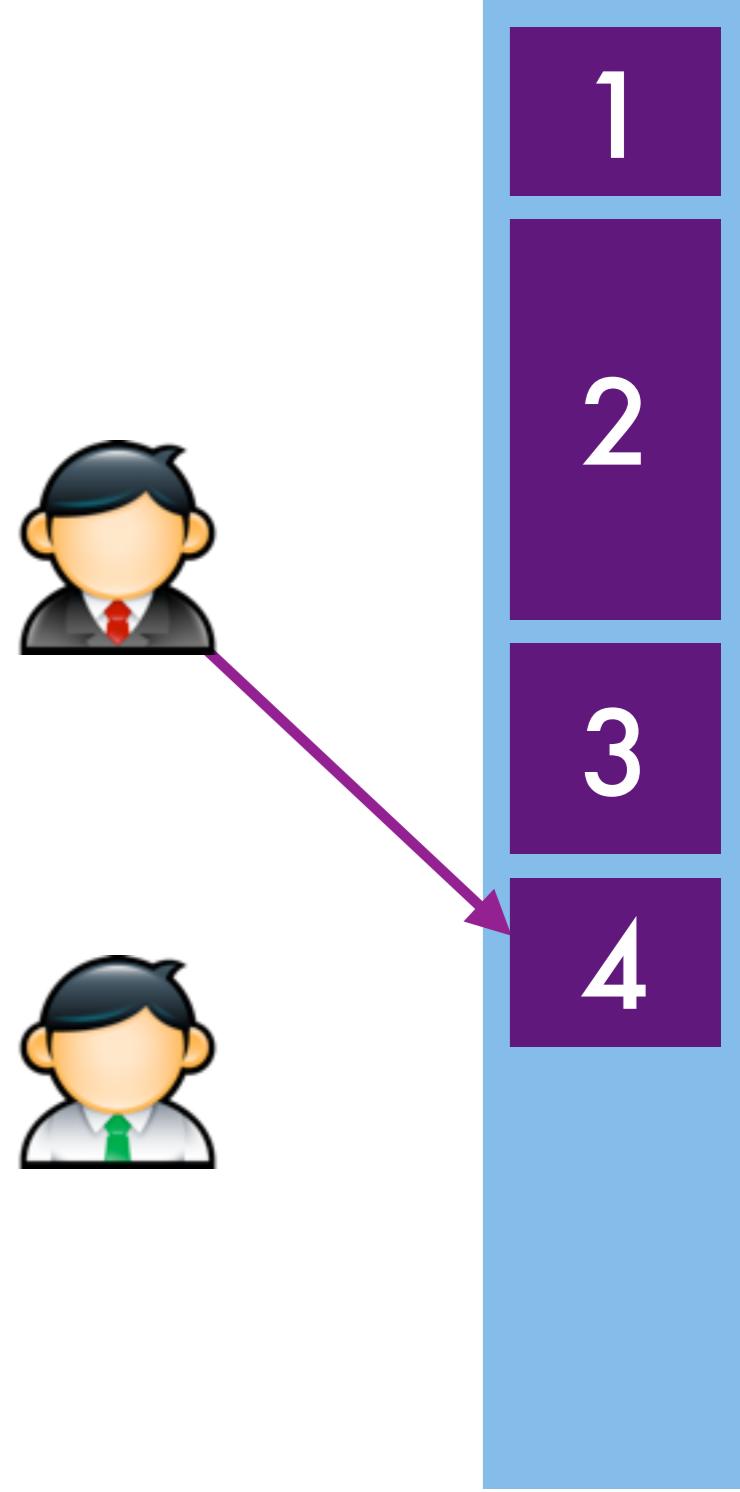
# Proportional Allocation Table

- Assign items according to machine capacity
  - Create allocation table with segments proportional to capacity
  - Leave space for additional machines
  - Hash key  $h(x)$  and pick machine covering it
  - If failure, re-hash the hash until it hits a bin
  - For replication hit  $k$  bins in a row
- 
- Proportional load distribution
  - Limited scalability
  - Need to distribute and update table
  - Limit peak load by further delegation  
(SPOCA - Chawla et al., USENIX 2011)



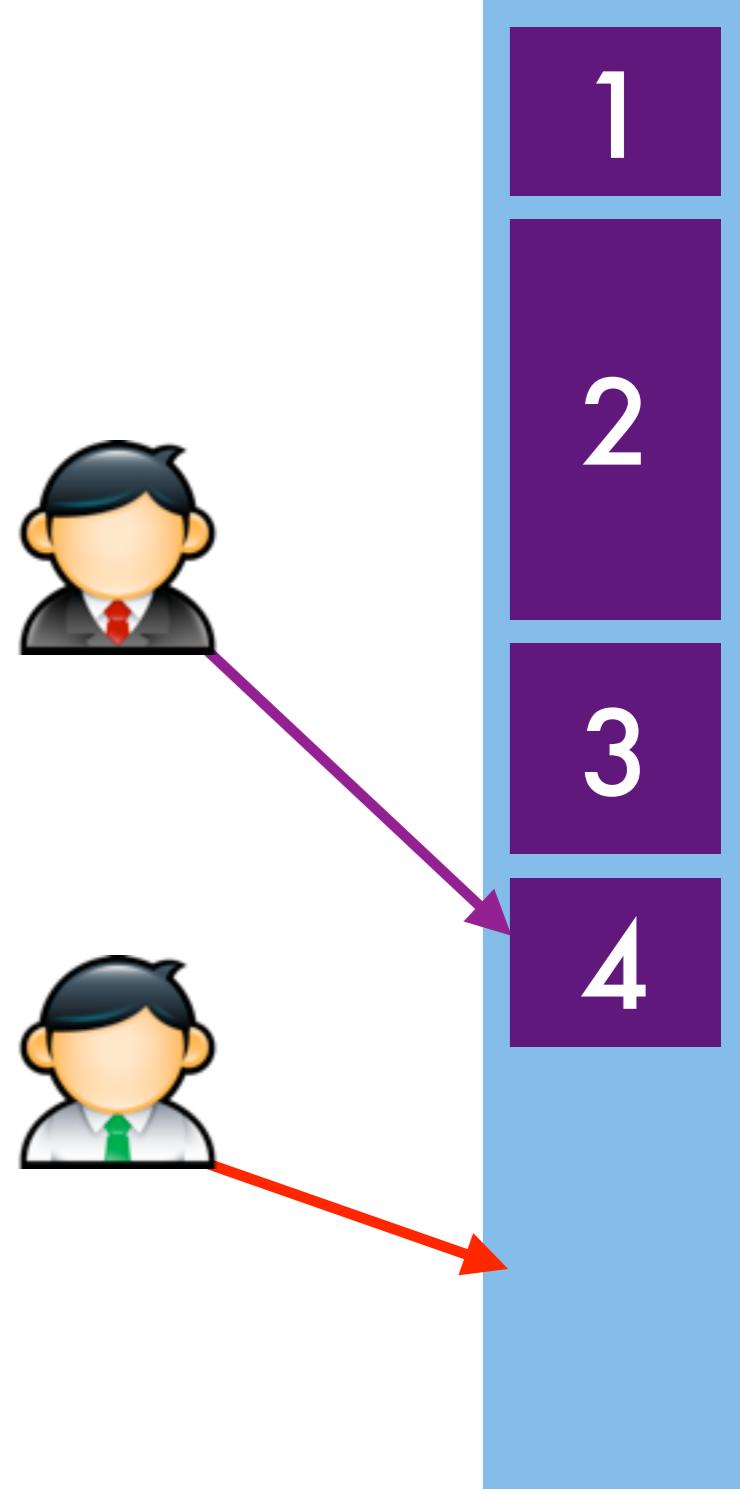
# Proportional Allocation Table

- Assign items according to machine capacity
  - Create allocation table with segments proportional to capacity
  - Leave space for additional machines
  - Hash key  $h(x)$  and pick machine covering it
  - If failure, re-hash the hash until it hits a bin
  - For replication hit  $k$  bins in a row
- 
- Proportional load distribution
  - Limited scalability
  - Need to distribute and update table
  - Limit peak load by further delegation  
(SPOCA - Chawla et al., USENIX 2011)



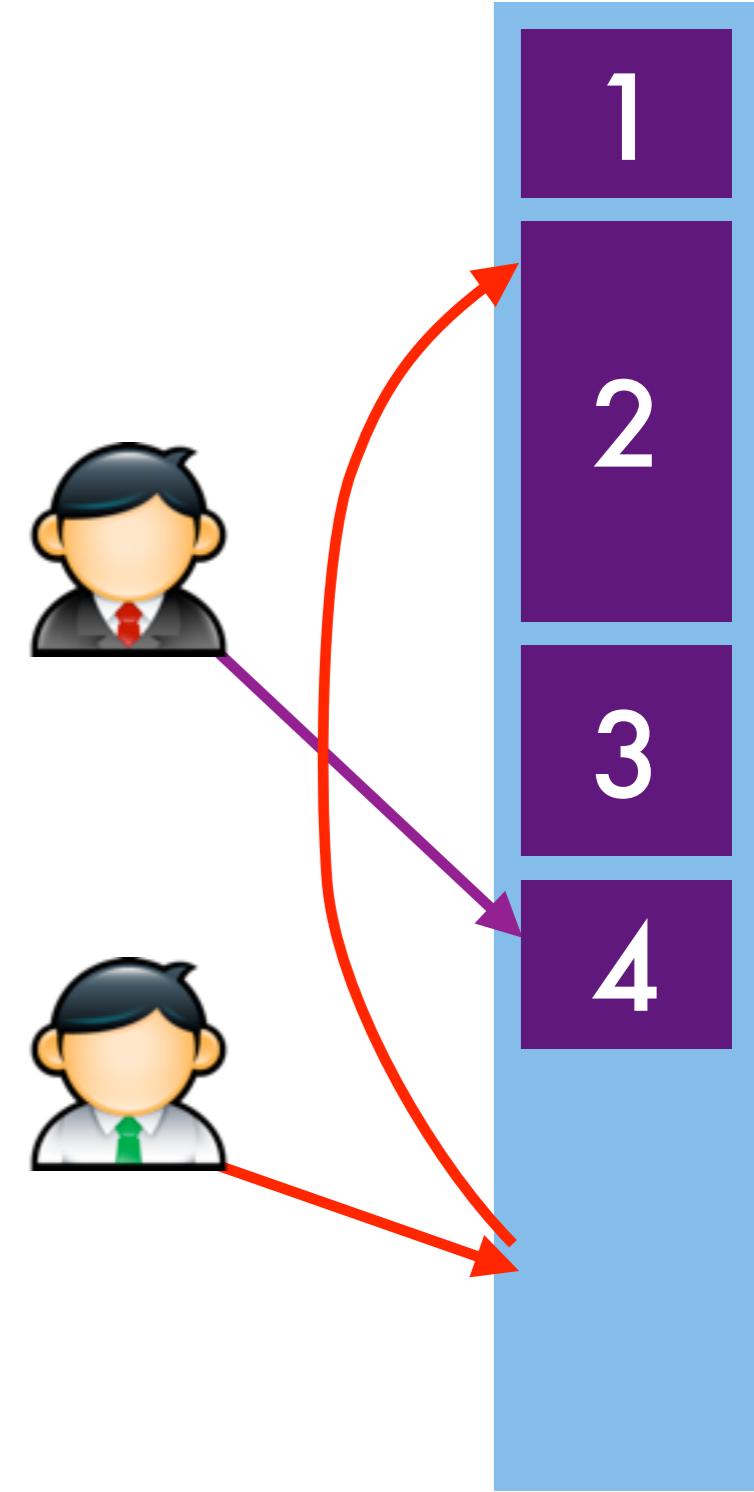
# Proportional Allocation Table

- Assign items according to machine capacity
  - Create allocation table with segments proportional to capacity
  - Leave space for additional machines
  - Hash key  $h(x)$  and pick machine covering it
  - If failure, re-hash the hash until it hits a bin
  - For replication hit  $k$  bins in a row
- 
- Proportional load distribution
  - Limited scalability
  - Need to distribute and update table
  - Limit peak load by further delegation  
(SPOCA - Chawla et al., USENIX 2011)



# Proportional Allocation Table

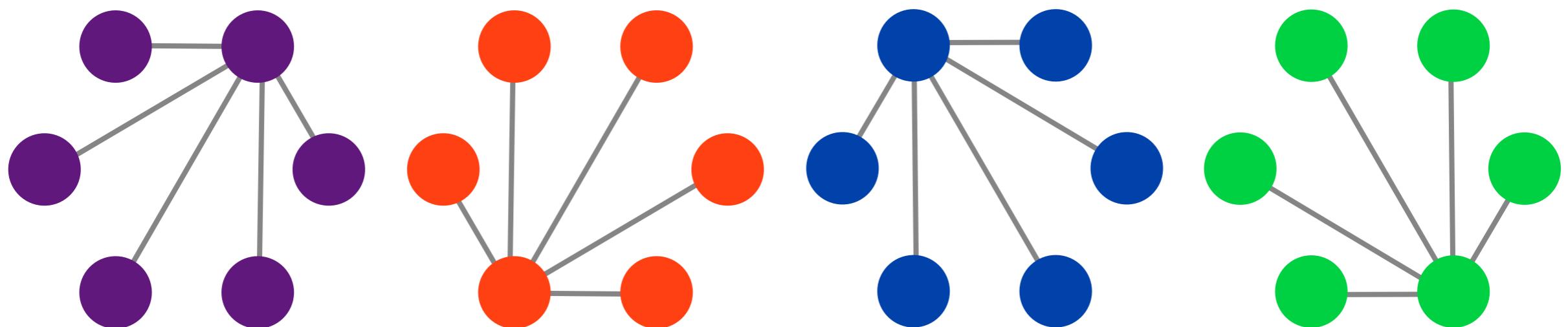
- Assign items according to machine capacity
  - Create allocation table with segments proportional to capacity
  - Leave space for additional machines
  - Hash key  $h(x)$  and pick machine covering it
  - If failure, re-hash the hash until it hits a bin
  - For replication hit  $k$  bins in a row
- 
- Proportional load distribution
  - Limited scalability
  - Need to distribute and update table
  - Limit peak load by further delegation  
(SPOCA - Chawla et al., USENIX 2011)



# Random Caching Trees

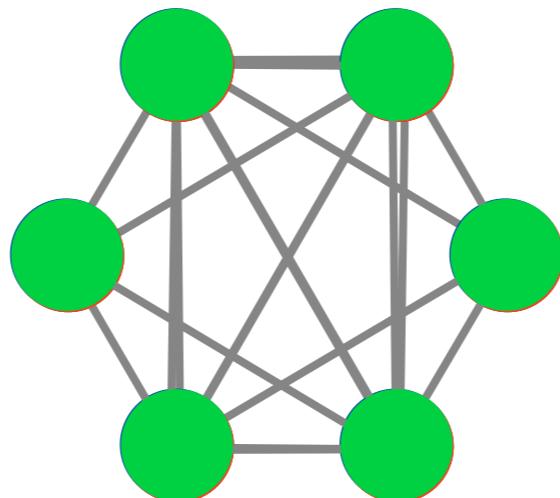
## (Karger et al. 1999, Akamai paper)

- Cache / synchronize an object
  - Uneven load distribution
  - Must not generate hotspot
- 
- For given key, pick random order of machines
  - Map order onto tree / star via BFS ordering



# Random Caching Trees

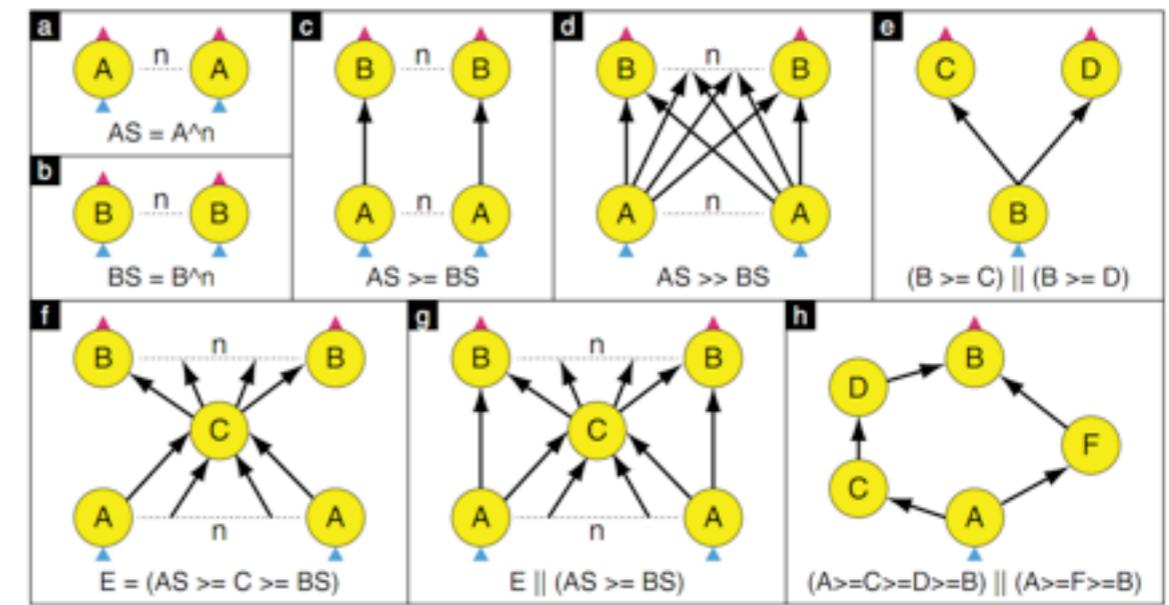
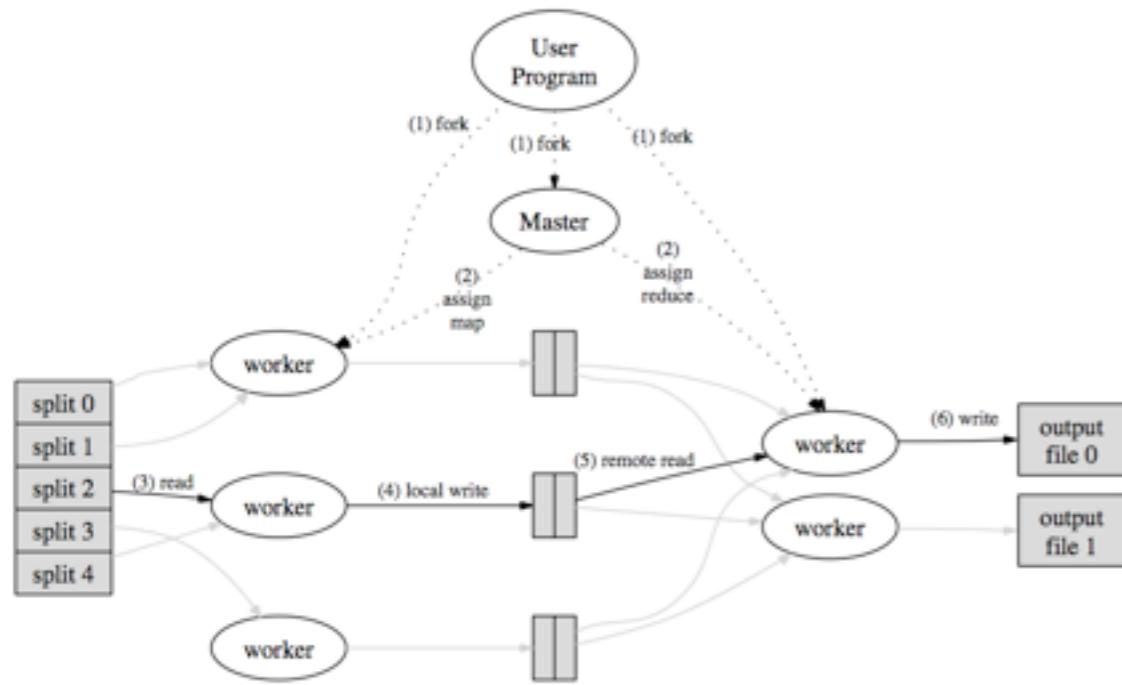
- Cache / synchronize an object
  - Uneven load distribution
  - Must not generate hotspot
- 
- For given key, pick random order of machines
  - Map order onto tree / star via BFS ordering

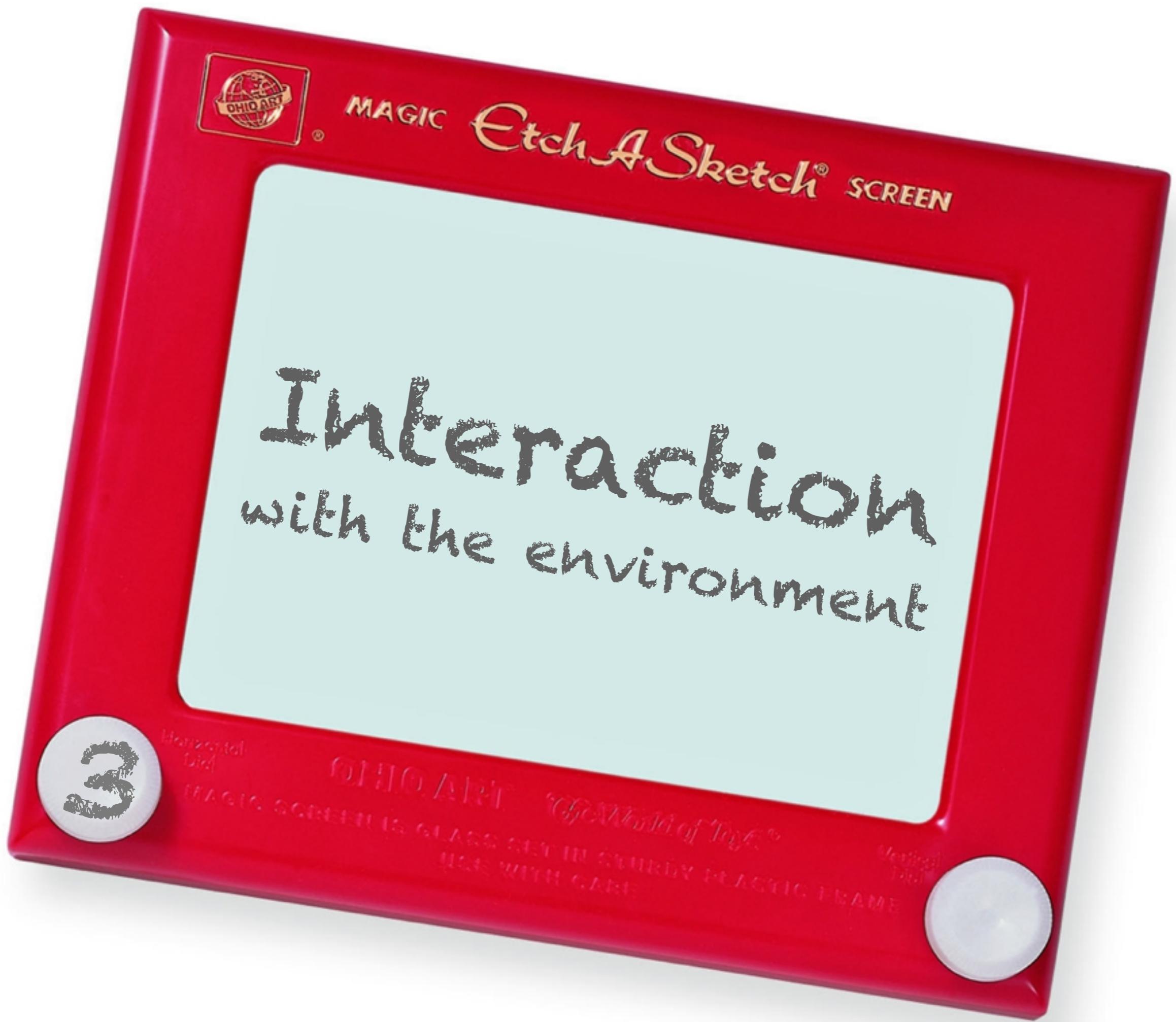


e.g. memcached

# More stuff

- Map reduce (e.g. Hadoop)
- Online streaming (e.g. S4, Dryad, Storm)
- NoSQL Database (e.g. pnut, bigtable)
- Fault tolerant (key,value) storage (e.g. dynamo)
- Smart file system layout (e.g. ceph, GFS2)





# Batch

- Data generated independently
  - Editors label data
  - Recorded log files
- Learning algorithm
  - Often invoked from scratch
  - No influence on data source
- Deployment
  - No direct influence on learning
  - Ignores influence on source



# Online

- Data generated independently
  - Editors label data
    - Incoming log files
  - Learning algorithm
    - Update happens in (near) realtime
    - Adapts to changing data source (good for spam, attacks, news)
  - Deployment
    - No direct influence on learning
    - Ignores influence on source



# Interactive / Explore & Exploit

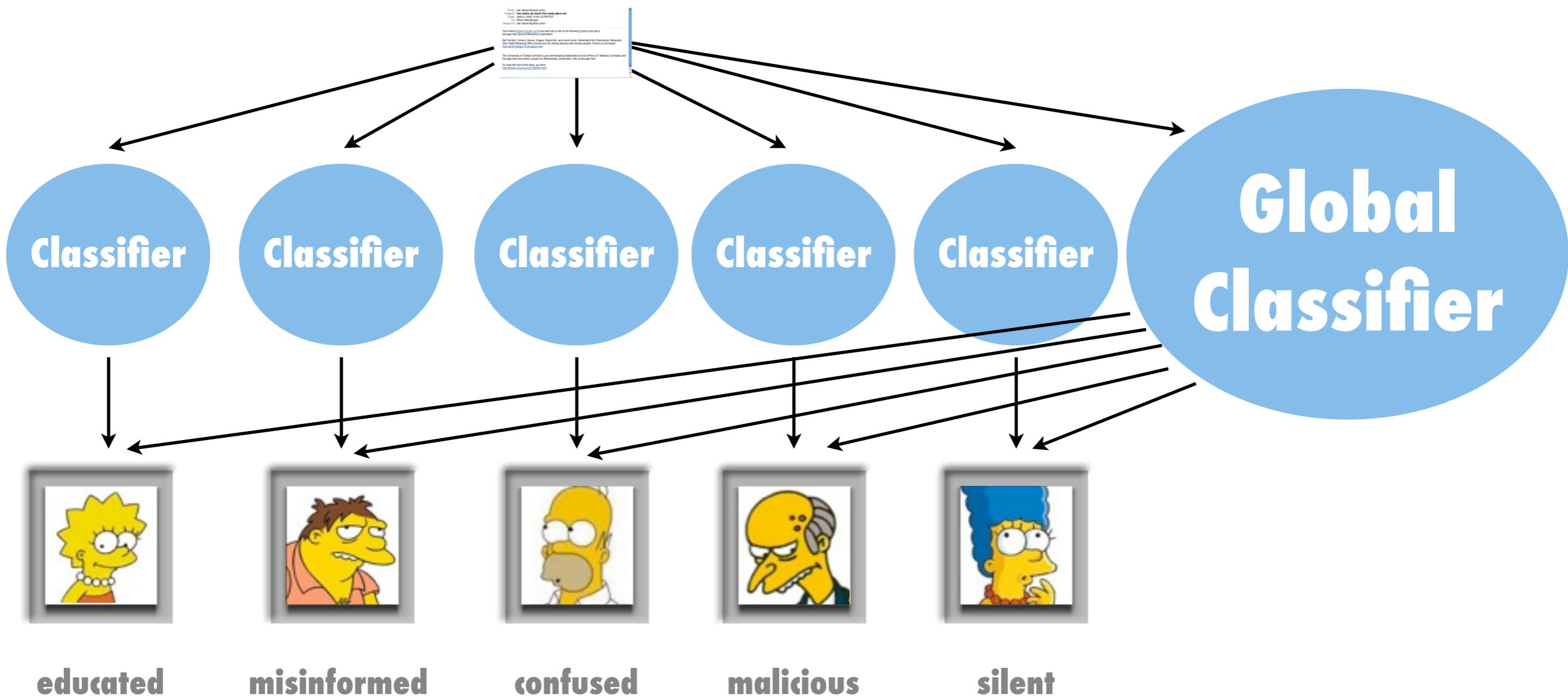
- Data is response to current model
  - Story recommendations
  - Personalized news ranking
- Learning algorithm
  - Update happens in (near) realtime
  - Adapts to changing data source
- Deployment
  - Predictive uncertainty influences exploration
  - Value of information & current payoff



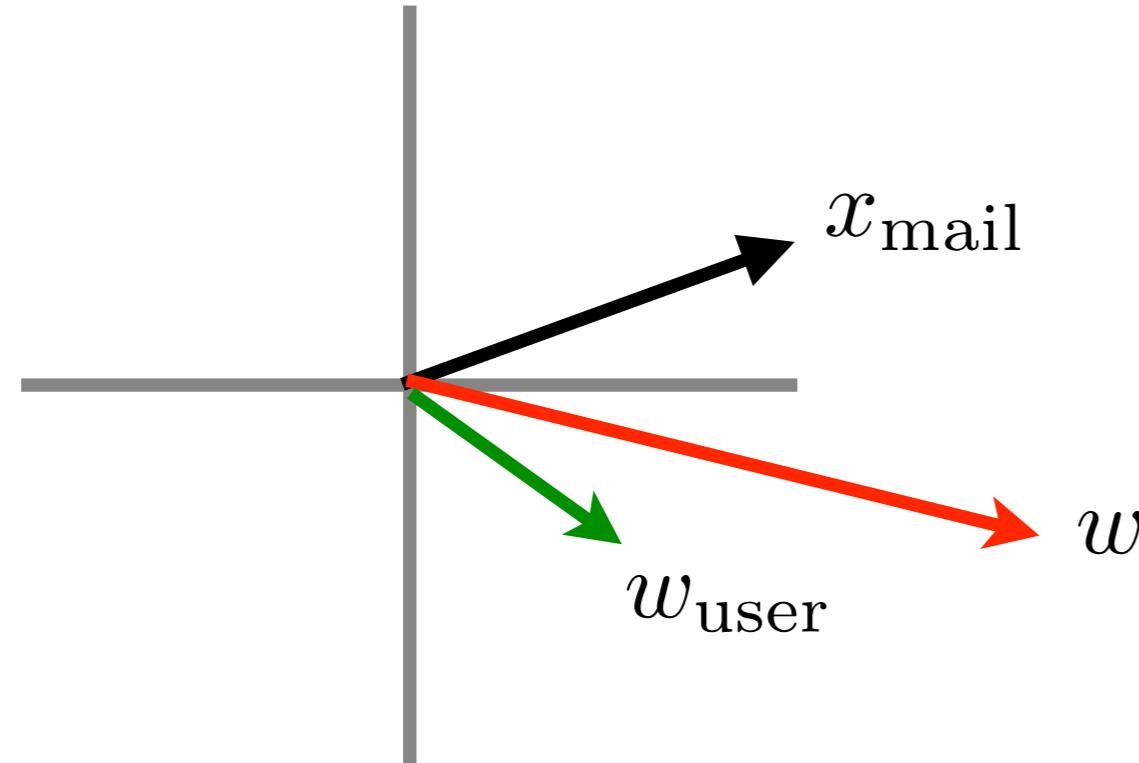




# Personalized Spam Classification



# Personalized Spam Classification



- **Primal representation**

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

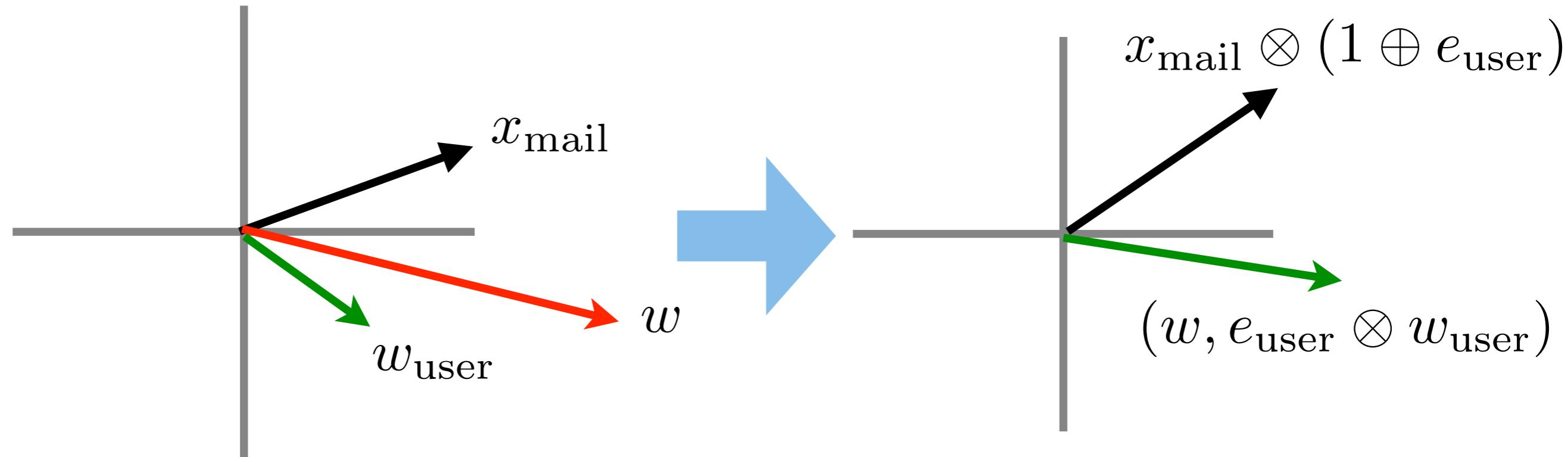
- **Kernel representation**

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

Multitask kernel (e.g. Pontil & Michelli, Daume). Usually does not scale well ...

- **Problem** - dimensionality is  $10^6 \times 10^8$ . That is 400TB of space

# Personalized Spam Classification



- **Primal 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  $10^6 \times 10^8$ . That is 400TB of space**

# Hash Kernels

# Hash Kernels

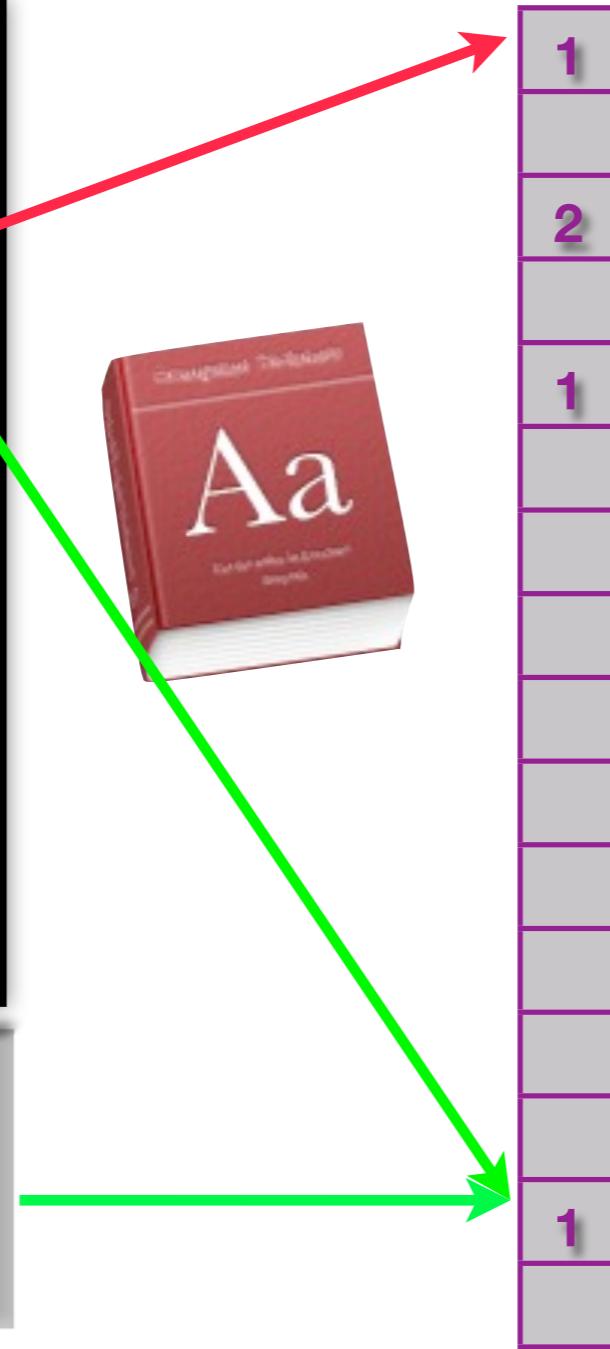
## instance:

Hey,  
please mention  
subtly during your  
talk that people  
should use Yahoo  
products more  
often.  
Thanks,

# dictionary:



task/user  
(=barney):



# sparse

# Hash Kernels

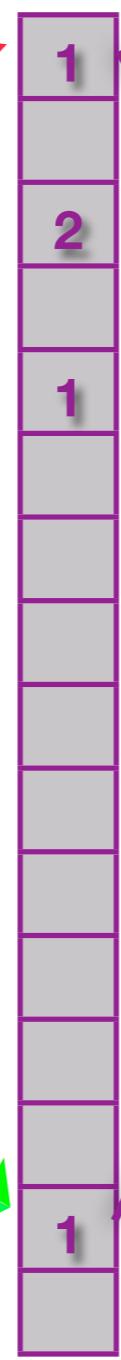
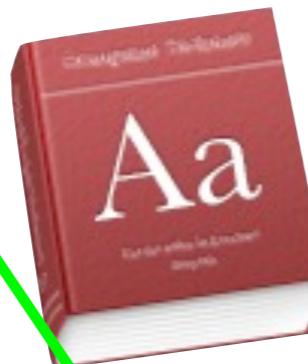
## instance:

Hey,  
please mention  
subtly during your  
talk that people  
should use Yahoo  
products more  
often.

Thanks,

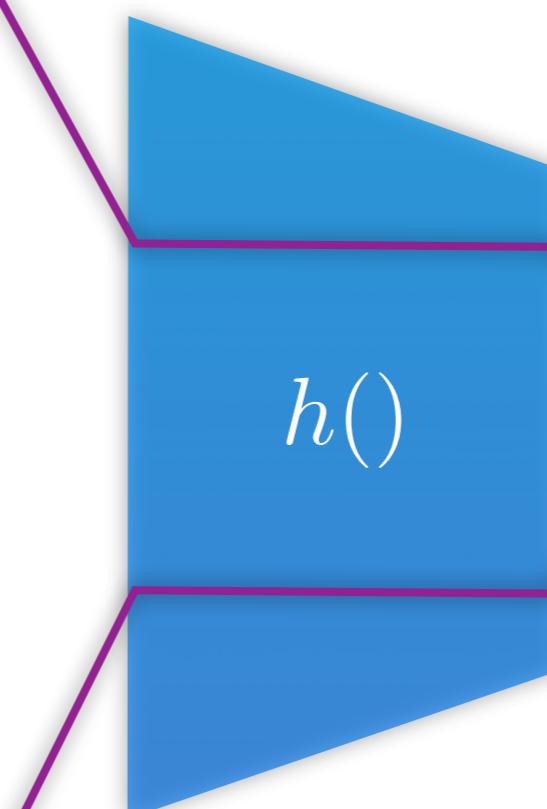
# Someone important

## dictionary:



## hash function:

$h()$



	1
	3
	2
	1

## sparse

task/user  
(=barney):



## sparse

# Hash Kernels

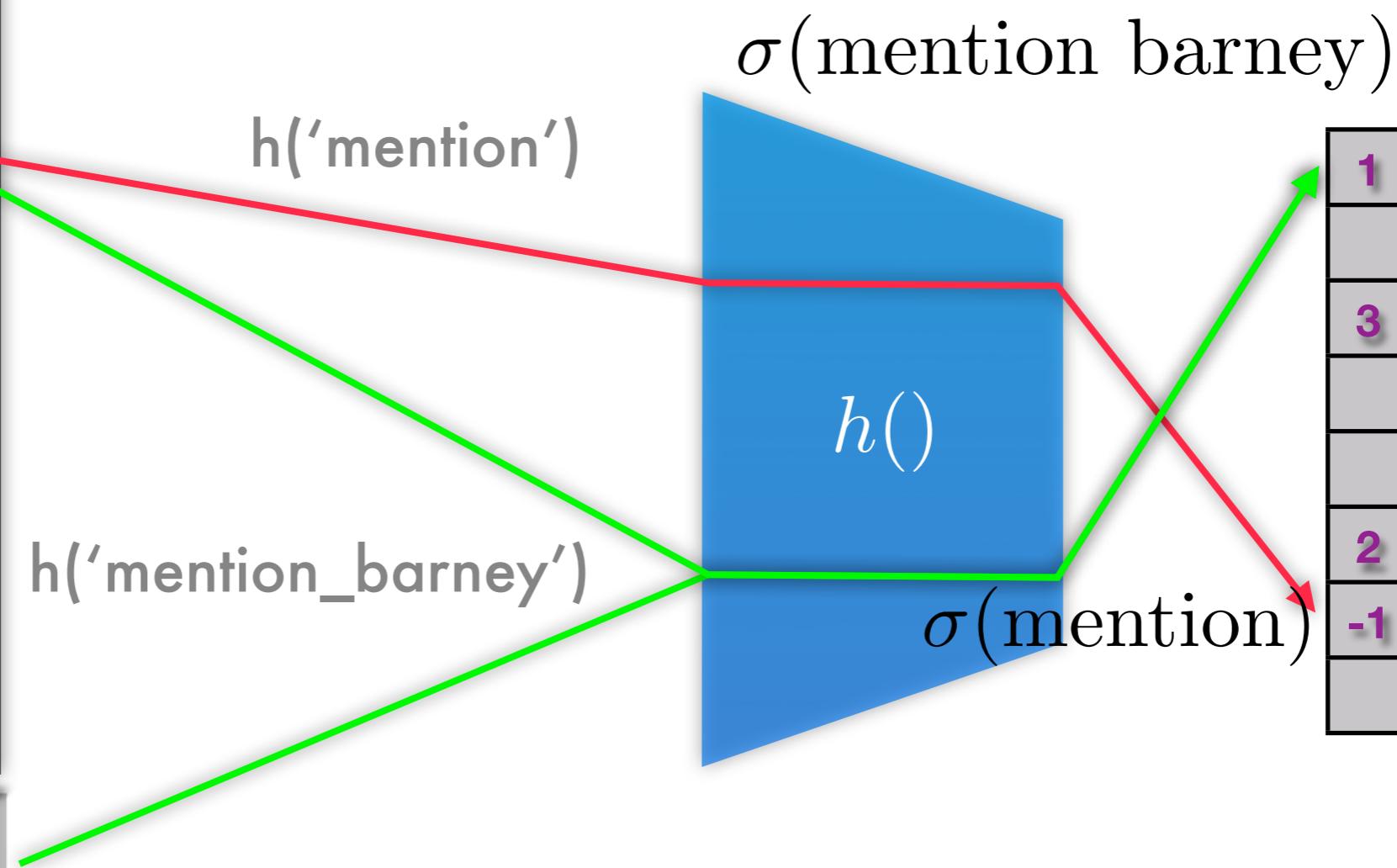
instance:

sparsity preserving, dictionary free

Hey,  
please mention  
subtly during your  
talk that people  
should use Yahoo  
products more  
often.  
Thanks,  
Someone important



task/user  
= barney



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

# Hash Kernels

- Function evaluation

$$f(x) = \sum_i w_i x_i + b$$

$$f_{\text{hash}}(x) = \sum_i \sigma(i) w[h(i)] x_i + b$$

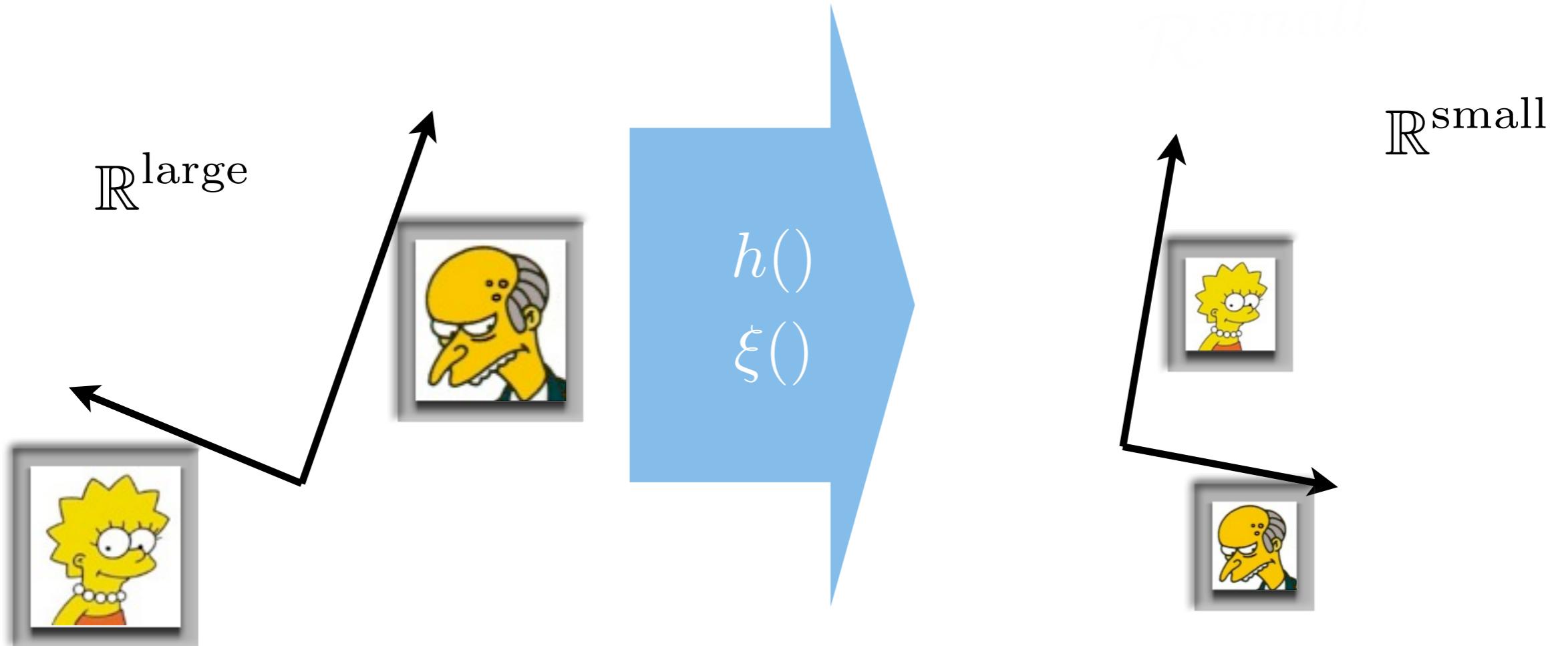
- Kernel

$$k(x, x') = \sum_i x_i x'_i$$

collisions

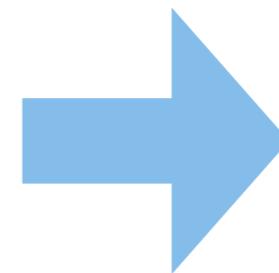
$$k_{\text{hash}}(x, x') = \sum_{j=1}^n \left[ \sum_{i:h(i)=j} x_i \sigma(i) \right] \left[ \sum_{i:h(i)=j} x'_i \sigma(i) \right]$$

# Approximate Orthogonality



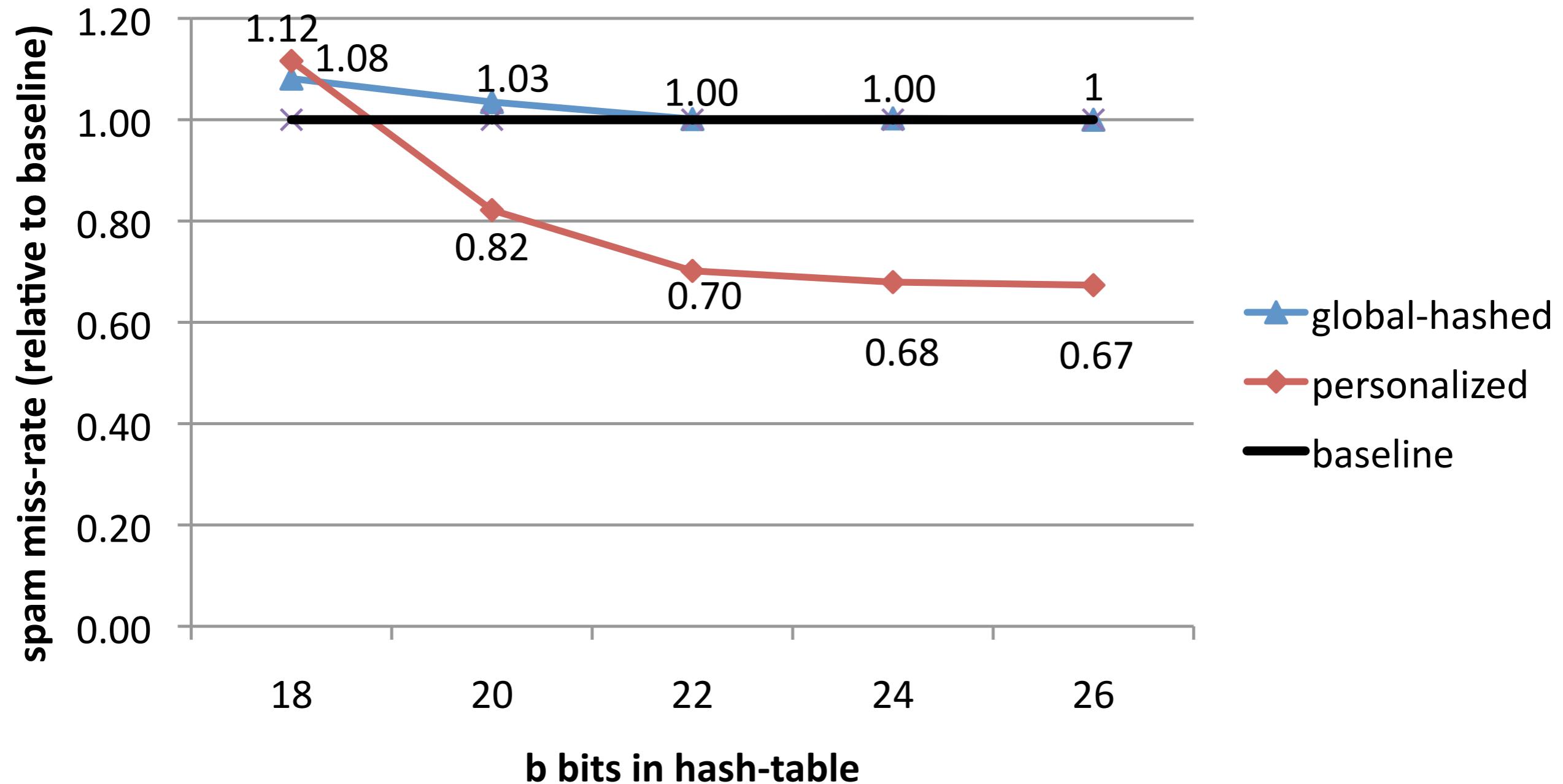
We can do multi-task learning!

**Direct sum in  
Hilbert Space**



**Sum in  
Hash Space**

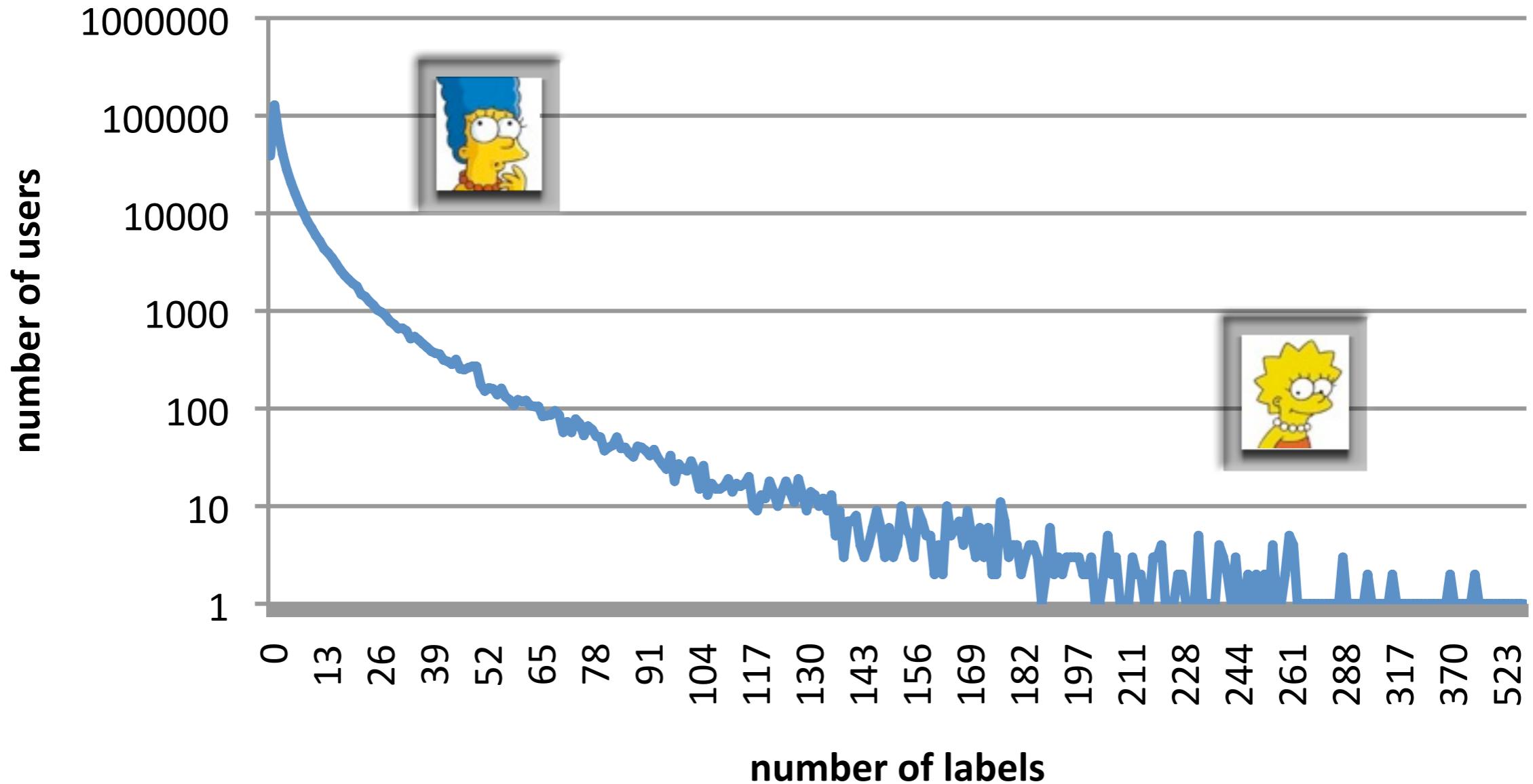
# Spam classification results



N=20M, U=400K

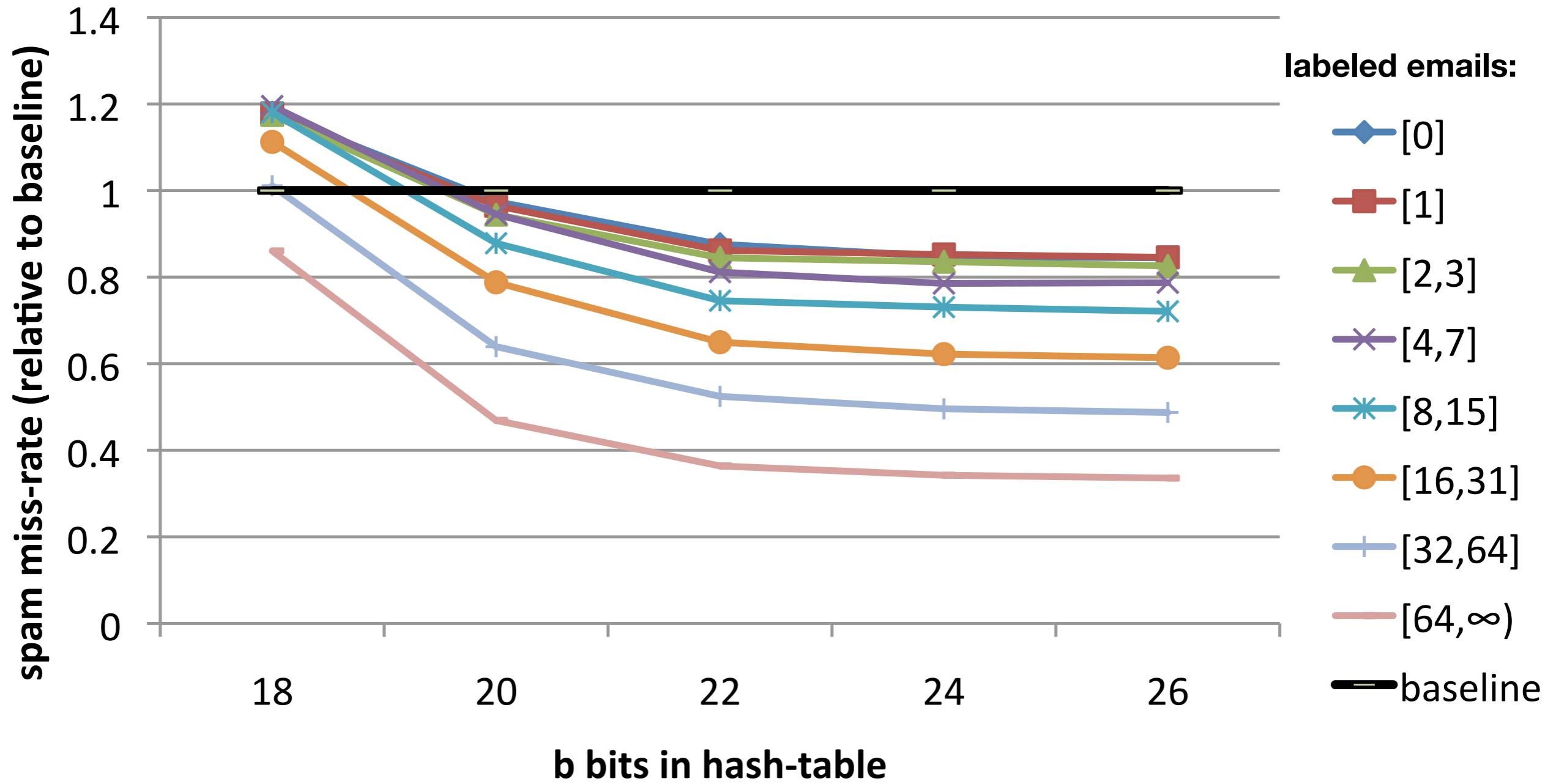
# Lazy users ...

Labeled emails per user

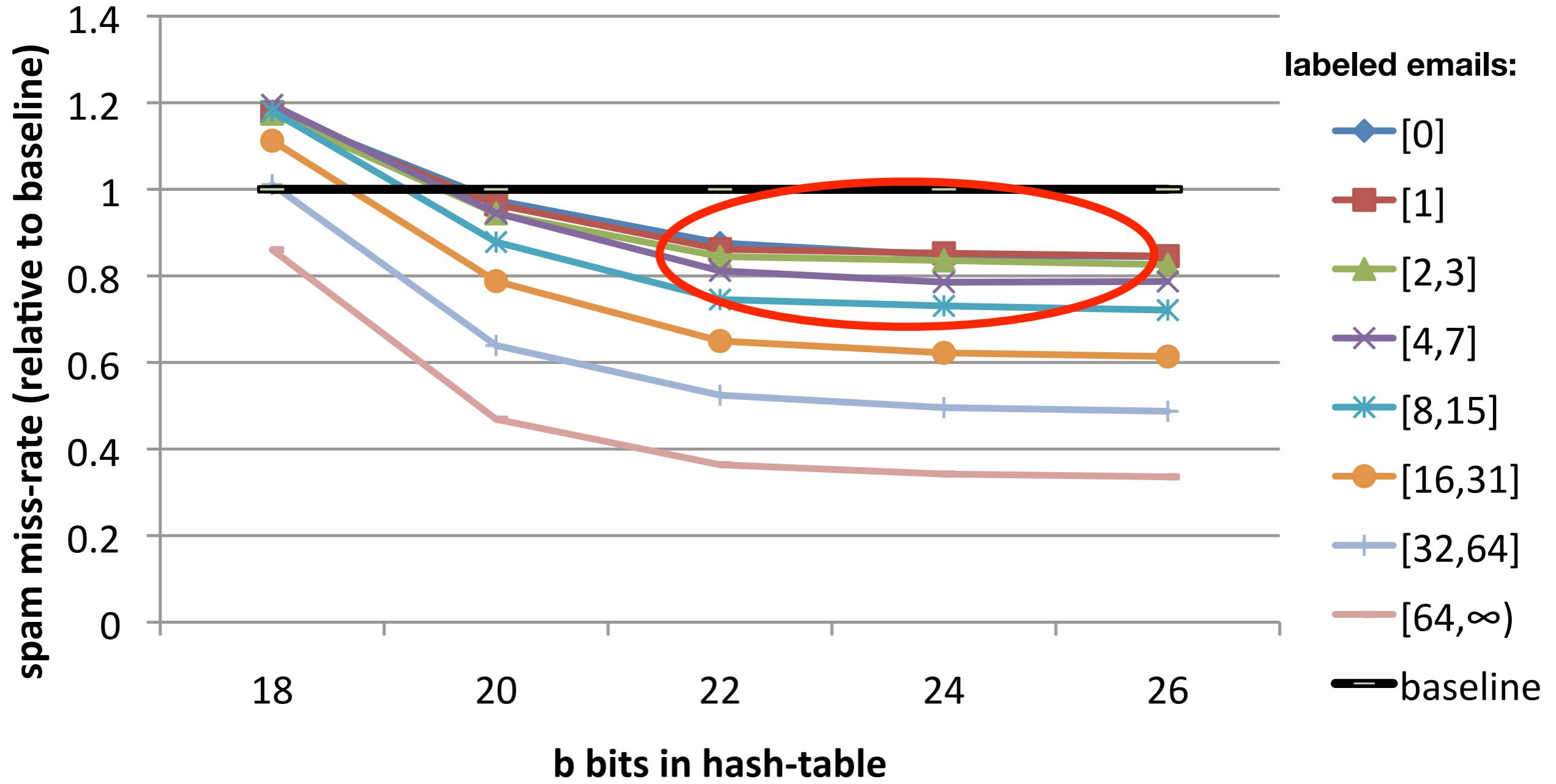


# Results by user group

# Results by user group



# Results by user group



# Even more

- Fast graph comparison
  - Extract subgraph signatures
  - Avoiding to implement dynamic data structures
    - Ontologies (hash ontology path labels)
    - Hierarchical factorization (hash context)
    - Content personalization (hash source, user, context)
  - Collaborative filtering
    - Compress many users into common parameter vector
    - String comparison (kernels)
      - Generate sequence with mismatches, hash and weight  
e.g. dog becomes {(dog,1), (\*og, 0.5), (d\*g, 0.5), (do\*, 0.5)}
  - Replace w[complicated key] by w[h(complicated key)]