



# Templates

for scalable data analysis

## 1 Introduction to Big Learning

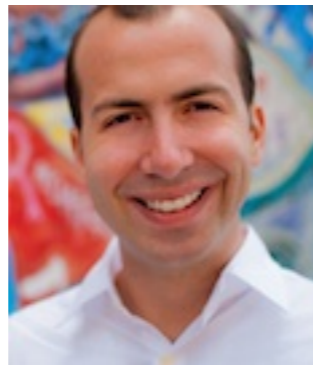
**Amr Ahmed, Alexander J Smola, Markus Weimer**

Yahoo! Research & UC Berkeley & ANU

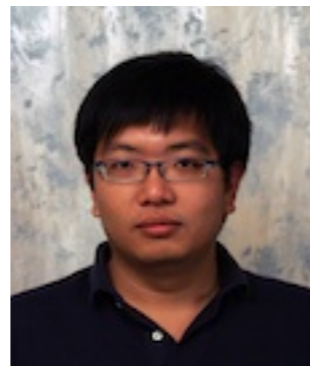
# Thanks



Mohamed  
Aly



Joey  
Gonzalez



Yucheng  
Low



Qirong  
Ho



Shravan  
Narayanamurthy



Amr  
Ahmed



Choon Hui  
Teo



Eric  
Xing



James  
Petterson



Sergiy  
Matyusevich



Jake  
Eisenstein



Shuang Hong  
Yang



Vishy  
Vishwanathan



Markus  
Weimer



Vanja  
Josifovski



MAGIC Etch A Sketch<sup>®</sup> SCREEN

- Problems in machine learning
- Systems to run the algorithms
- Response batch/online/interactive
- Compression

Horizontal  
Lid

OHIO ART "A World of Toys"

MAGIC SCREEN IS GLASS SET IN STURDY PLASTIC FRAME  
USE WITH CARE

Vertical  
Lid



MAGIC Etch A Sketch<sup>®</sup> SCREEN

Some  
Problems  
in machine learning

1

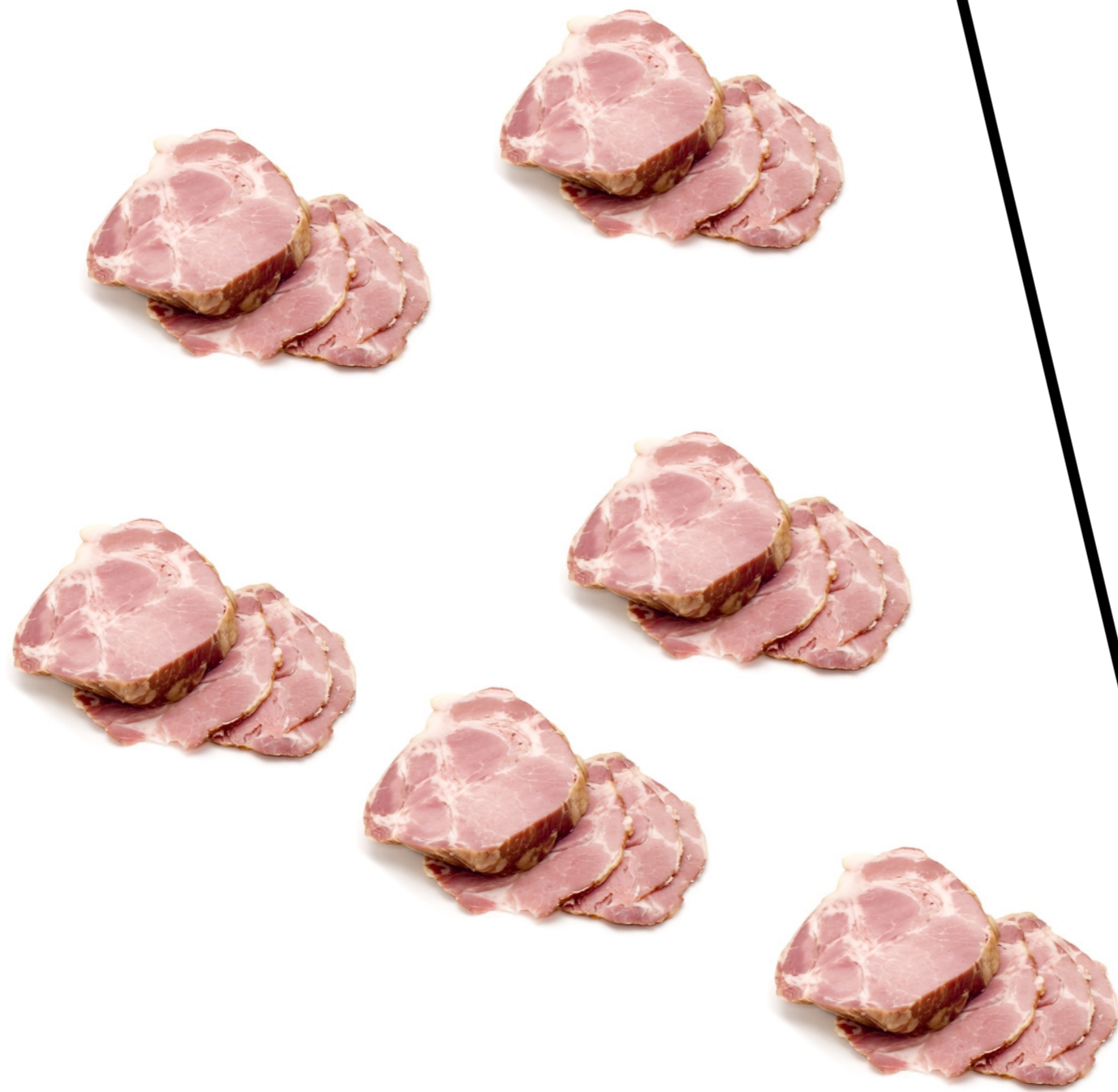
Horizontal  
Lid

OHIO ART "World of Toys"<sup>®</sup>

MAGIC SCREEN IS GLASS SET IN STURDY PLASTIC FRAME  
USE WITH CARE

Vertical  
Lid

# Classification



# Spam Filtering

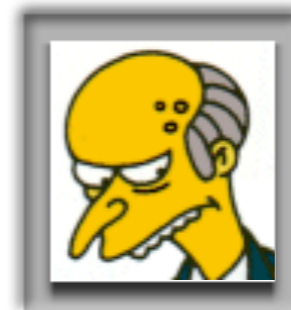
From: bat <kilian@gmail.com>  
Subject: **hey whats up check this meds place out**  
Date: April 6, 2009 10:50:13 PM PDT  
To: Kilian Weinberger  
Reply-To: bat <kilian@gmail.com>

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://jenkinste3f.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://jenkinste3f.3.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>

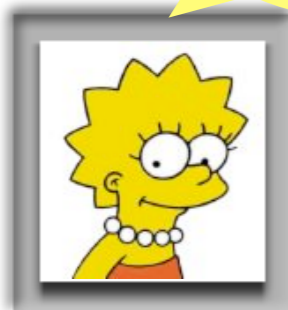
**1: spam!**

**0: quality**

**1: donut?**

**0: not-spam!**

**?**



**educated**



**misinformed**



**confused**

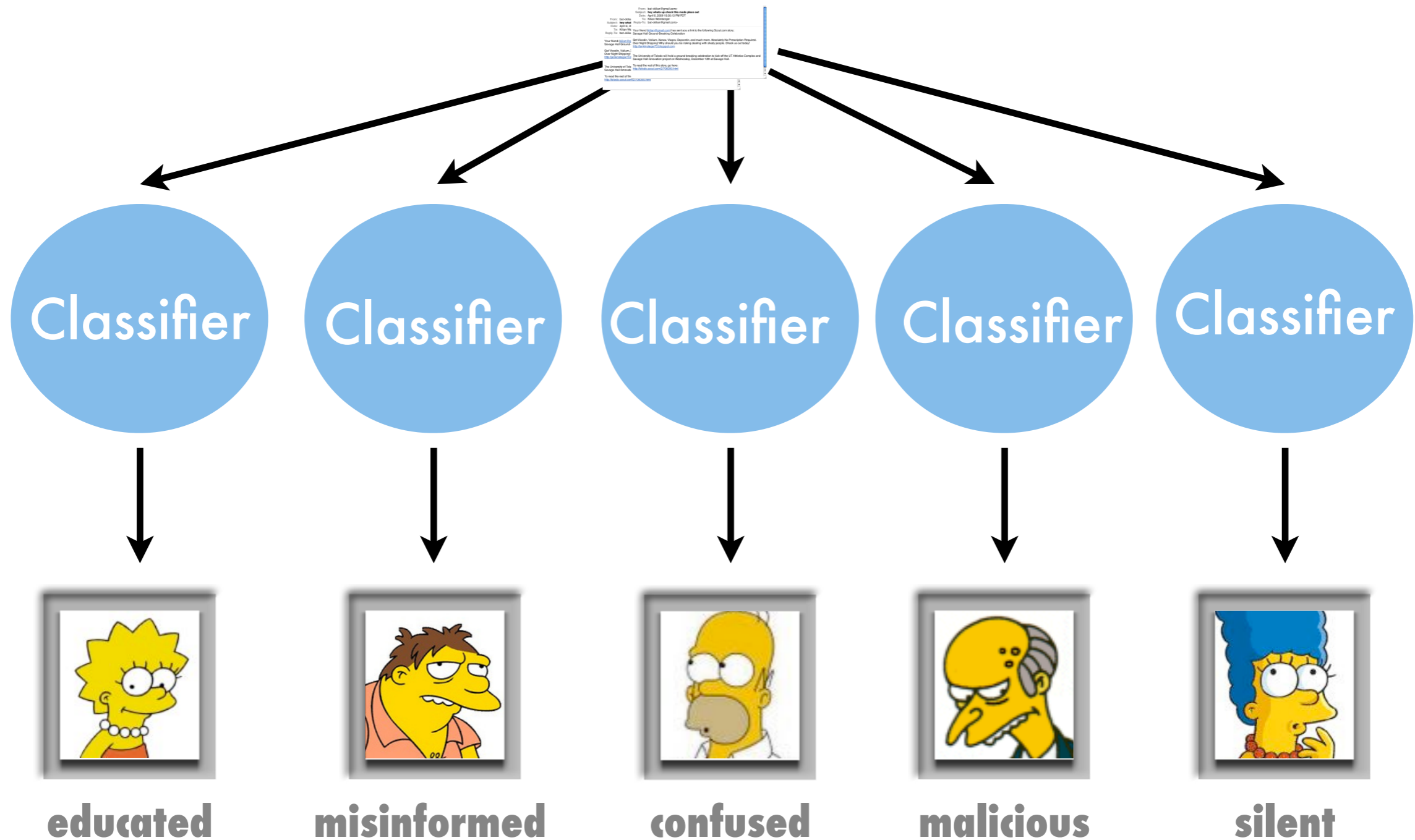


**malicious**



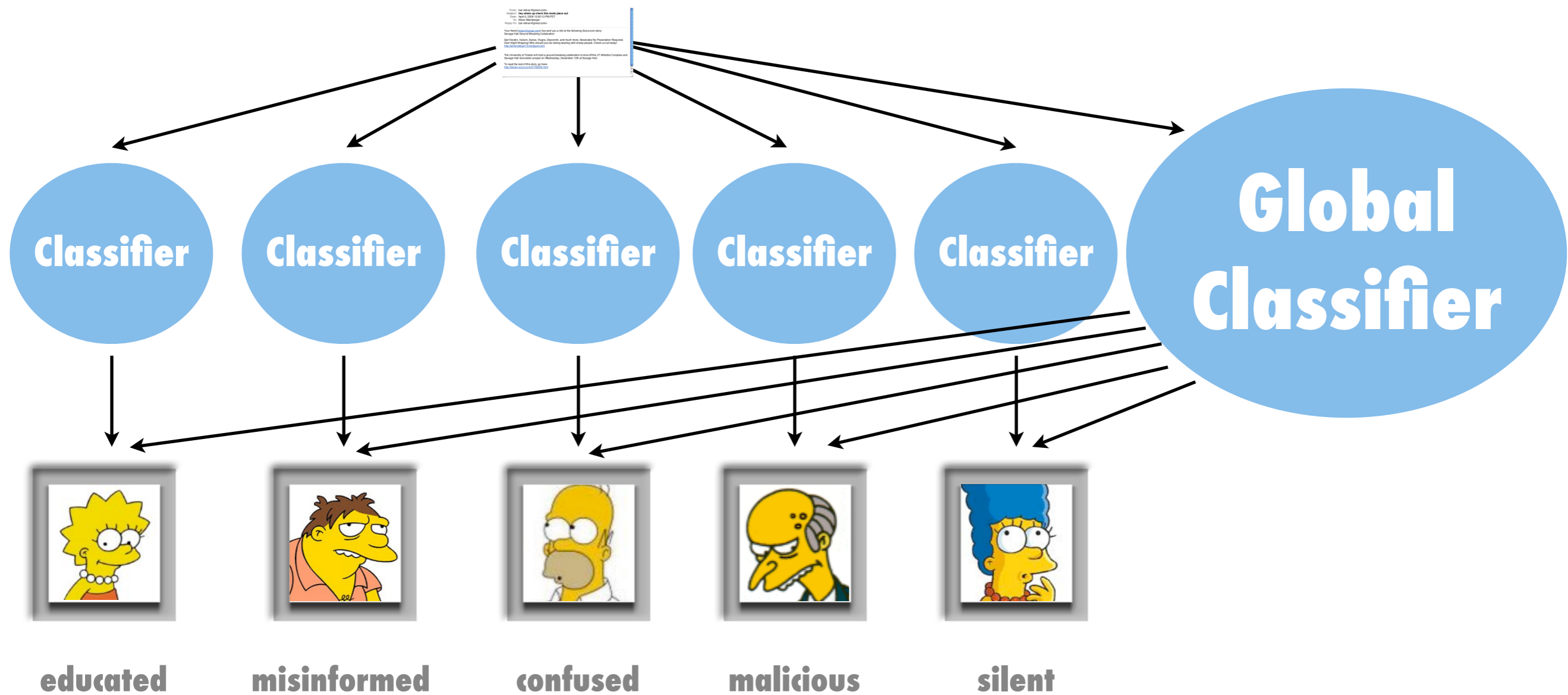
**silent**

# Spam Filtering

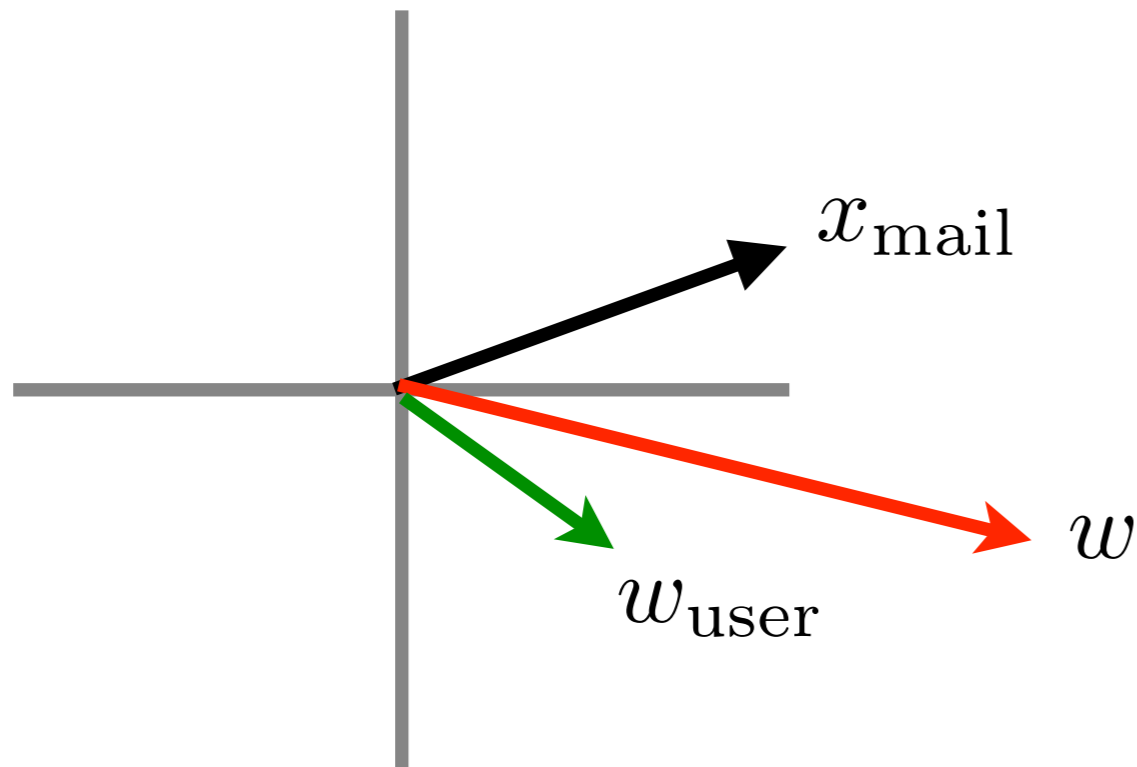




# Personalized Spam Filtering



# Personalized Spam Filtering



- **Function representation**

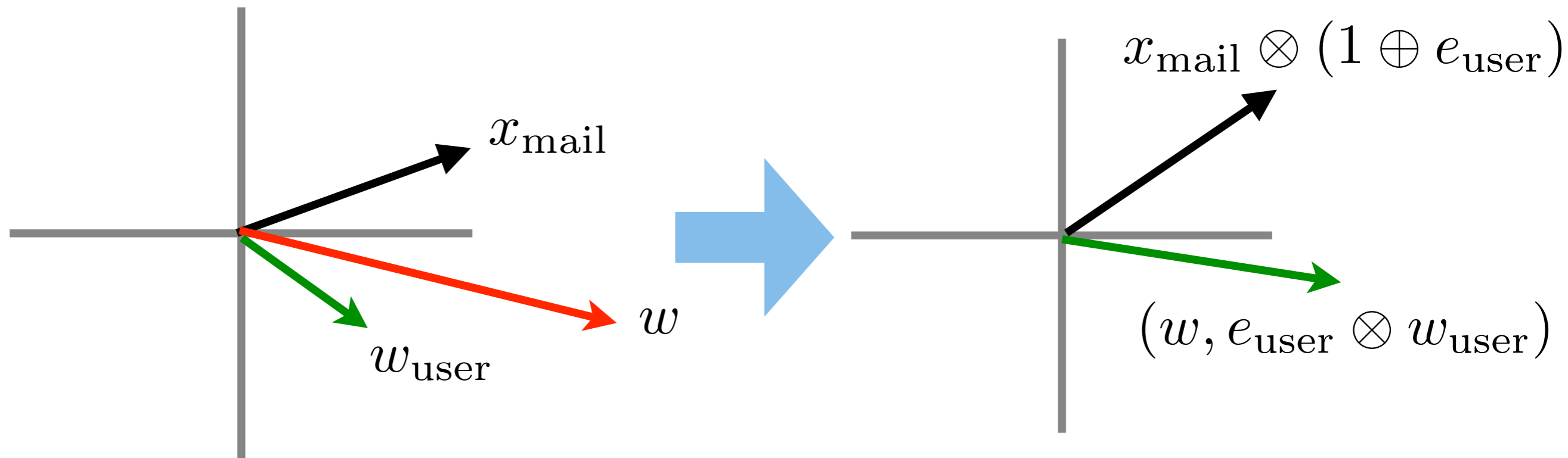
$$f(x, u) = \langle \phi(x), w \rangle + \langle \phi(x), w_u \rangle = \langle \phi(x) \otimes (1 \oplus e_u), w \rangle$$

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

- **Reduce to binary classification problem and classify with**

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

# 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

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

# Personalized Spam Filtering

1-50 of 150

Delete all spam messages now (messages that have been in Spam more than 30 days will be automatically deleted)

<input type="checkbox"/>	<input type="checkbox"/>	<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="checkbox"/>	<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="checkbox"/>	<input checked="" type="checkbox"/>	Penis Growth Sample	Smell sweeter 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="checkbox"/>	<input checked="" type="checkbox"/>	Edward Bell	Re: Re: Migl%ori boosters ERO on-line - Ogni medicina nel gruppo di*disfunzione erettile è qui http://njuzo.velvdoctor.ru	9:10 am
<input type="checkbox"/>	<input type="checkbox"/>	<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="checkbox"/>	<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="checkbox"/>	<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="checkbox"/>	<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="checkbox"/>	<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="checkbox"/>	<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%\$iUMAavq1 - 名牌包包,皮夾,鞋子,手錶,眼	Apr 6
<input type="checkbox"/>	<input type="checkbox"/>	<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><u>Arts</u></b> <a href="#">Movies</a> , <a href="#">Television</a> , <a href="#">Music</a> ...	<b><u>Business</u></b> <a href="#">Jobs</a> , <a href="#">Real Estate</a> , <a href="#">Investing</a> ...	<b><u>Computers</u></b> <a href="#">Internet</a> , <a href="#">Software</a> , <a href="#">Hardware</a> ...
<b><u>Games</u></b> <a href="#">Video Games</a> , <a href="#">RPGs</a> , <a href="#">Gambling</a> ...	<b><u>Health</u></b> <a href="#">Fitness</a> , <a href="#">Medicine</a> , <a href="#">Alternative</a> ...	<b><u>Home</u></b> <a href="#">Family</a> , <a href="#">Consumers</a> , <a href="#">Cooking</a> ...
<b><u>Kids and Teens</u></b> <a href="#">Arts</a> , <a href="#">School Time</a> , <a href="#">Teen Life</a> ...	<b><u>News</u></b> <a href="#">Media</a> , <a href="#">Newspapers</a> , <a href="#">Weather</a> ...	<b><u>Recreation</u></b> <a href="#">Travel</a> , <a href="#">Food</a> , <a href="#">Outdoors</a> , <a href="#">Humor</a> ...
<b><u>Reference</u></b> <a href="#">Maps</a> , <a href="#">Education</a> , <a href="#">Libraries</a> ...	<b><u>Regional</u></b> <a href="#">US</a> , <a href="#">Canada</a> , <a href="#">UK</a> , <a href="#">Europe</a> ...	<b><u>Science</u></b> <a href="#">Biology</a> , <a href="#">Psychology</a> , <a href="#">Physics</a> ...
<b><u>Shopping</u></b> <a href="#">Clothing</a> , <a href="#">Food</a> , <a href="#">Gifts</a> ...	<b><u>Society</u></b> <a href="#">People</a> , <a href="#">Religion</a> , <a href="#">Issues</a> ...	<b><u>Sports</u></b> <a href="#">Baseball</a> , <a href="#">Soccer</a> , <a href="#">Basketball</a> ...
<b><u>World</u></b> <a href="#">Català</a> , <a href="#">Dansk</a> , <a href="#">Deutsch</a> , <a href="#">Español</a> , <a href="#">Français</a> , <a href="#">Italiano</a> , <a href="#">日本語</a> , <a href="#">Nederlands</a> , <a href="#">Polski</a> , <a href="#">Русский</a> , <a href="#">Svenska</a> ...		

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



Copyright © 2012 Netscape

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

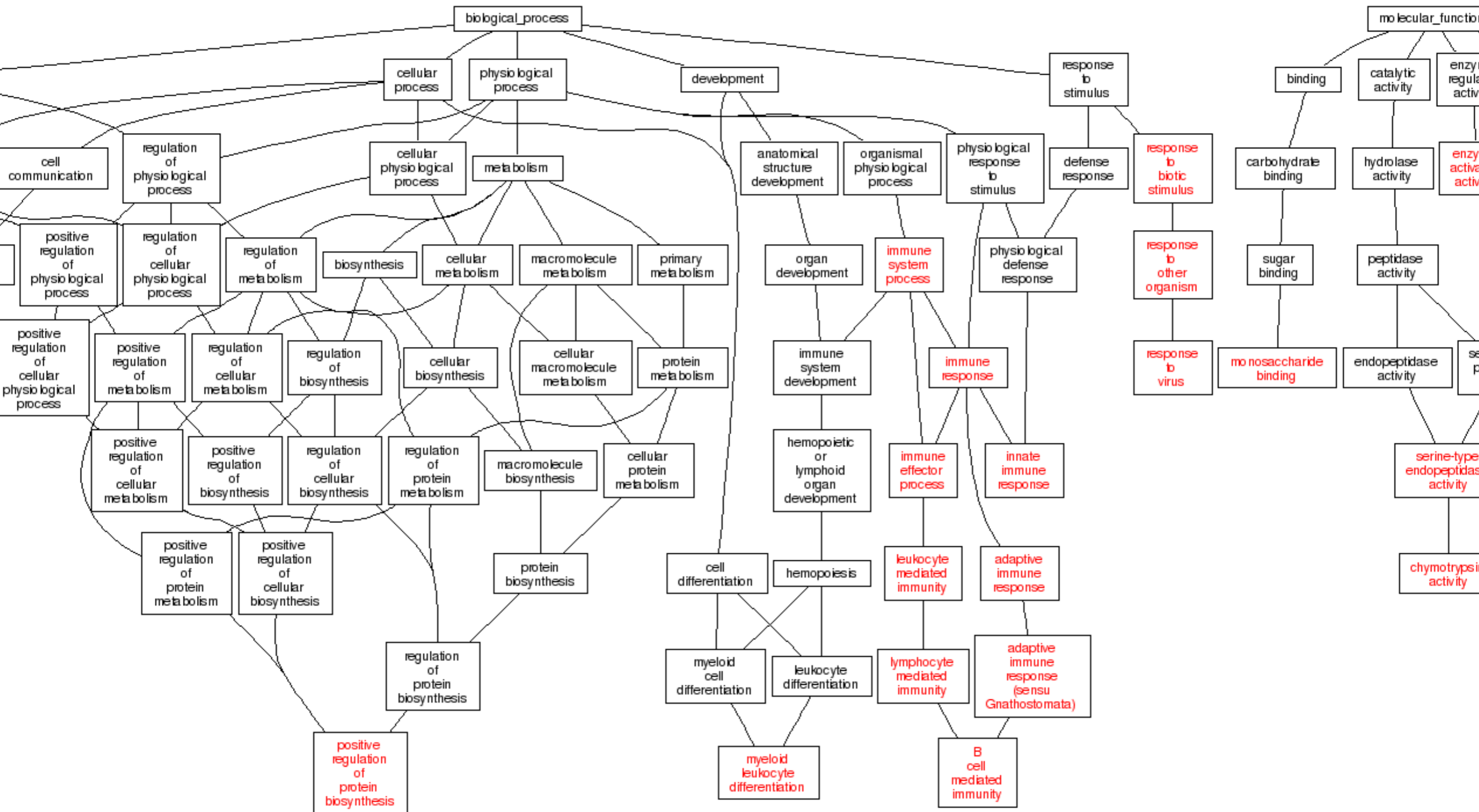
[Become an Editor](#) Help build the largest human-edited directory of the web 

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

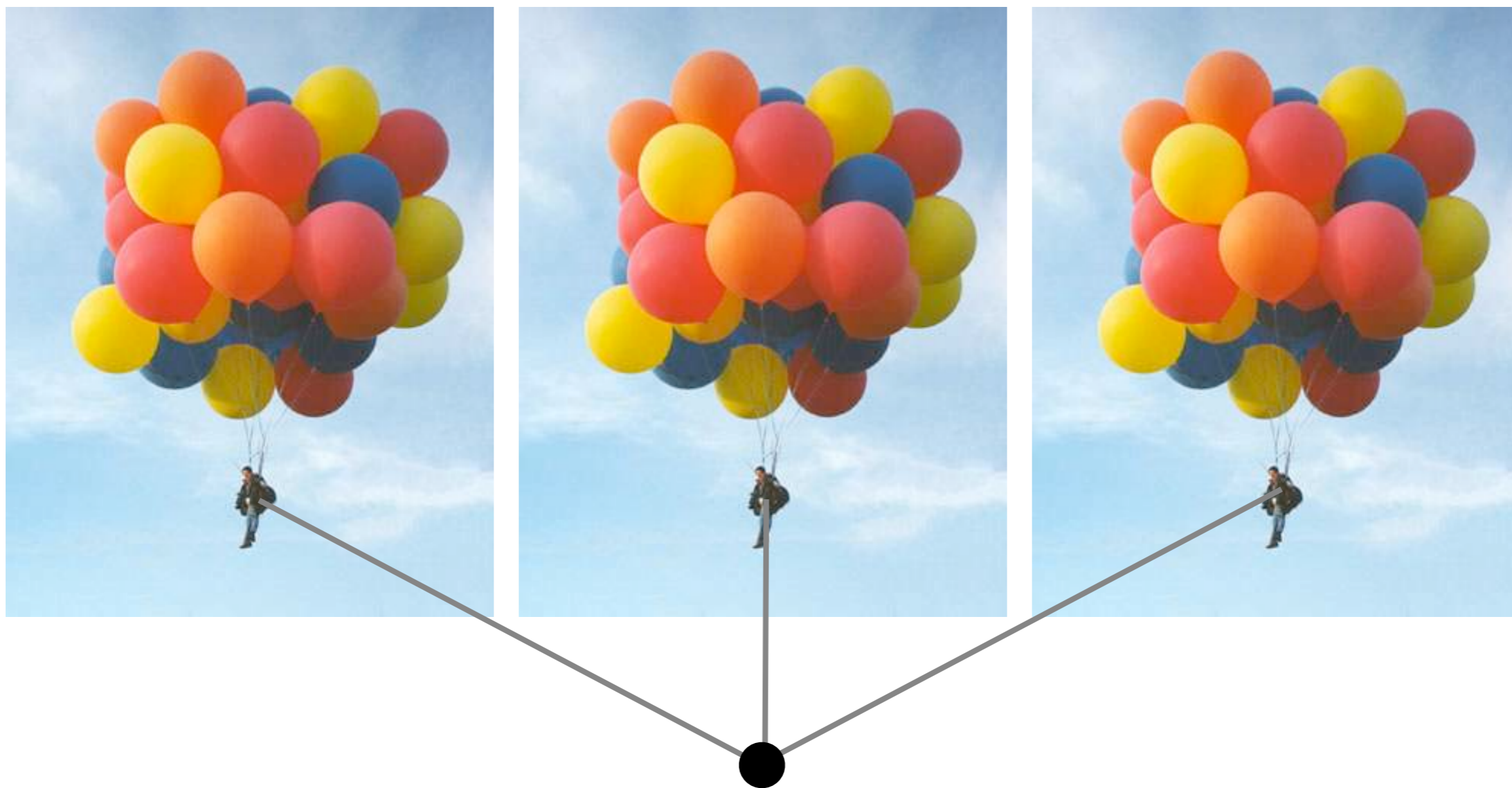
$$f(x, y) = \sum_{y' \in \text{path}(y)} \langle w_{y'}, x \rangle$$

Need efficient decoding on tree

- **Alternative - obtain ontology automatically**



# Clustering



# Clustering

# Clustering

The screenshot shows the United Airlines website interface. At the top, there's the United logo and navigation links like 'My profile', 'Worldwide sites', and 'Customer service'. Below that are dropdown menus for 'Planning & booking', 'Reservations & check-in', 'Mileage Plus', and 'Services & information'. A search bar is also present. The main content area features a large promotional banner for 'Use 30% fewer miles on your next United flight.' with an image of a person holding a large orange balloon shaped like a percentage sign. To the left of the banner is a 'BOOK FLIGHT' form with fields for 'From', 'To', 'Departing', and 'Returning'. Below the form are options for 'Search by' (Schedule & price, Price, Flexible), 'Adult' count, and 'Cabin' type. A 'Search' button is at the bottom of the form. To the right of the banner is a 'Log in' section with fields for 'Mileage Plus # or email address' and 'Password', and a 'Log in' button. Below the login section are links for 'Start earning miles today', 'united.com benefits and features', and 'Travel information'. At the bottom of the page, there are links for 'About United', 'Investor relations', 'Business resources', 'Careers', and 'Site map'.

The screenshot shows the website for The Australian National University (ANU). At the top, there's a 'Change Location' button and a search bar. Below that are navigation links for 'You Fly', 'Loyalty Programmes', and 'Promotions'. A blue navigation bar contains links for 'myEMAIL', 'IVLE', 'LIBRARY', 'MAPS', 'CALENDAR', 'SITEMAP', 'CONTACT', and 'e-CARDS'. Below this is another search bar with the text 'Search search for...' and a 'GO' button. Further down, there are more navigation links: 'RESEARCH', 'ENTERPRISE', 'CAMPUS LIFE', 'GIVING', and 'CAREERS@NUS'. A banner for 'entred in Asia' is visible. Below the banner is a search bar for 'Search ANU...' and links for 'WEB', 'CONTACTS', and 'MAP'. The main heading 'The Australian National University' is prominently displayed. Below that are navigation links for 'CURRENT STUDENTS', 'RESEARCH & EDUCATION', 'ABOUT ANU', and 'STAFF'. The bottom of the page features a large image of a tree trunk with a small plant growing from it, and a navigation bar with links for 'Forests renew after Black Saturday fires', 'School of Music at Floriade', 'Undergraduate studies', and 'Higher Degree Research'.

# Clustering

The screenshot shows the United Airlines website interface. At the top, there's a navigation bar with 'UNITED' logo and links for 'My profile', 'Worldwide sites', and 'Customer service'. Below this is a search bar and a menu for 'Planning & booking', 'Reservations & check-in', 'Mileage Plus', and 'Services & information'. The main content area is divided into several sections: 'Flights' with a 'BOOK FLIGHT' and 'REDEEM MILES' tab, a 'Log in' section with fields for Mileage Plus # or email address and password, and a 'Travel information' section. A large promotional banner on the left says 'Use 30% fewer miles on your next United flight.' with a large percentage sign graphic. Below this, there are sections for 'United news and deals' and 'United-Continental merger'. At the bottom, there are links for 'Need Help?', 'SIA Holidays', and 'Hotel Bookings'.

The screenshot shows the Australian National University (ANU) website. The top navigation bar includes 'EXPLORE ANU', 'A-Z INDEX', and a search bar. The main header features the ANU logo and the text 'The Australian National University'. Below this is a secondary navigation bar with links for 'HOME', 'FUTURE STUDENTS', 'CURRENT STUDENTS', 'RESEARCH & EDUCATION', 'ABOUT ANU', and 'STAFF'. The main content area features a large banner with the headline 'Ash forests rise and rise again' and a sub-headline 'A new book that graphically documents the spectacular natural recovery of Victoria's ash forests after the Black Saturday bushfires also argues that wildfires are typical natural disturbances in these environments.' Below the banner are four featured articles: 'Forests renew after Black Saturday fires', 'School of Music at Floriade', 'Undergraduate studies', and 'Higher Degree Research'. At the bottom, there are five orange buttons: 'PROSPECTIVE STUDENTS', 'CURRENT STUDENTS', 'STAFF', 'ALUMNI', and 'VISITORS'.

The screenshot shows the Chez Panisse website. The top navigation bar includes 'Home', 'Wining & Dining', 'Contact', 'Sitemap', and 'About Suntec REIT'. The main content area features a large image of the restaurant's interior. Below the image is a white box with the text 'Chez Panisse' in a cursive font, followed by 'RESERVATIONS RESTAURANT & CAFÉ', 'MENUS RESTAURANT • CAFÉ MONDAY NIGHTS • WINE LIST', 'ABOUT CHEZ PANISSE • ALICE WATERS OUR CHEFS • FRIENDS • PRESS FOUNDATION & MISSION', 'SPECIAL EVENTS CALENDAR', 'STORE BOOKS • POSTERS • GIFTS', 'CONTACT INFORMATION DIRECTIONS • MAILING LIST'.



Home | Wining & Dining | Contact | Sitemap | About Suntec REIT



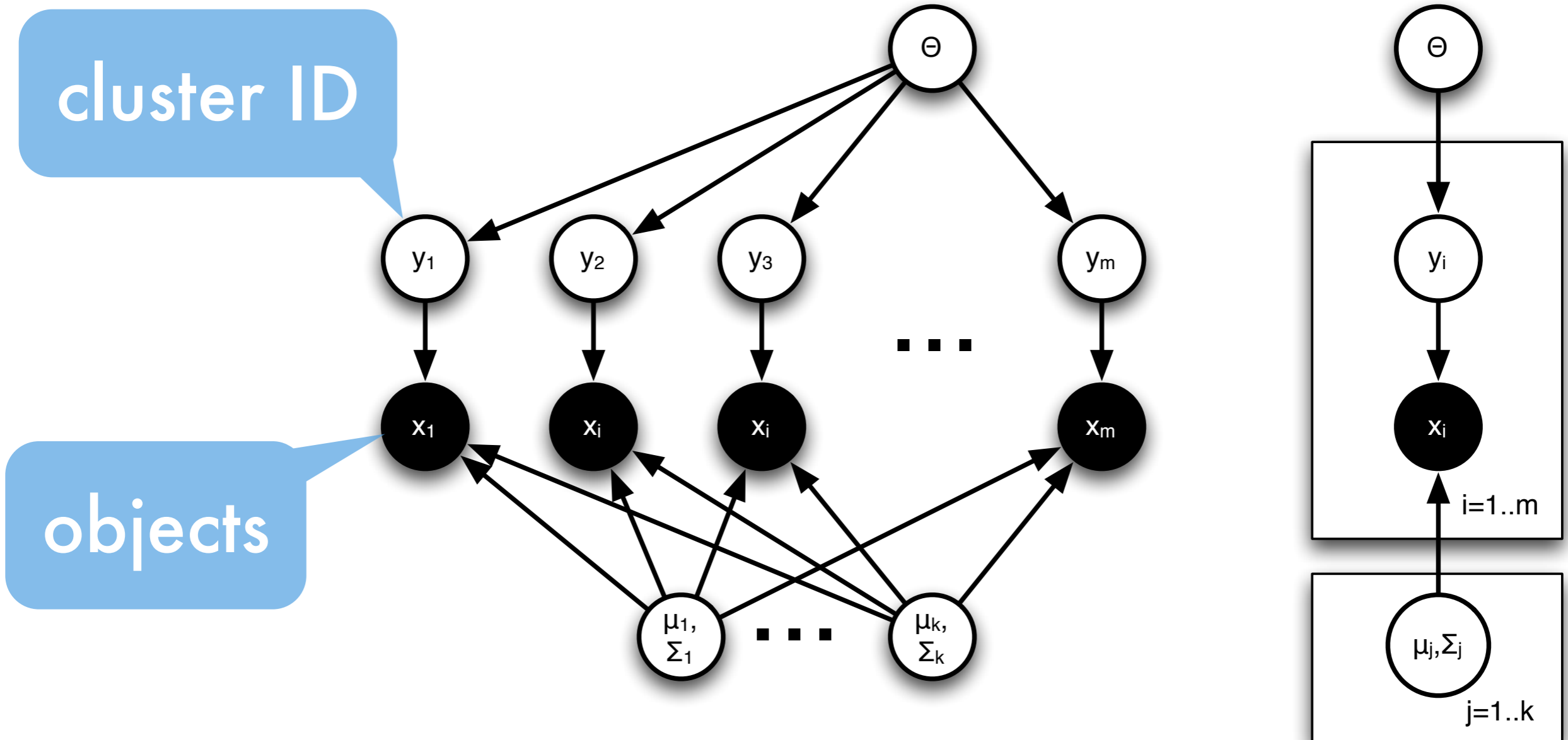
# Clustering

The screenshot shows the United Airlines website interface. At the top, there are navigation links for "My profile", "Worldwide sites", and "Customer service". Below that, there are tabs for "Planning & booking", "Reservations & check-in", "Mileage Plus", "Services & information", and a search bar. The main content area features a "Use 30% fewer miles on your next United flight" promotion with a large percentage sign graphic. There are sections for "Log in", "Travel information", and "United news and deals". A red speech bubble with the word "airline" is overlaid on the right side of the page.

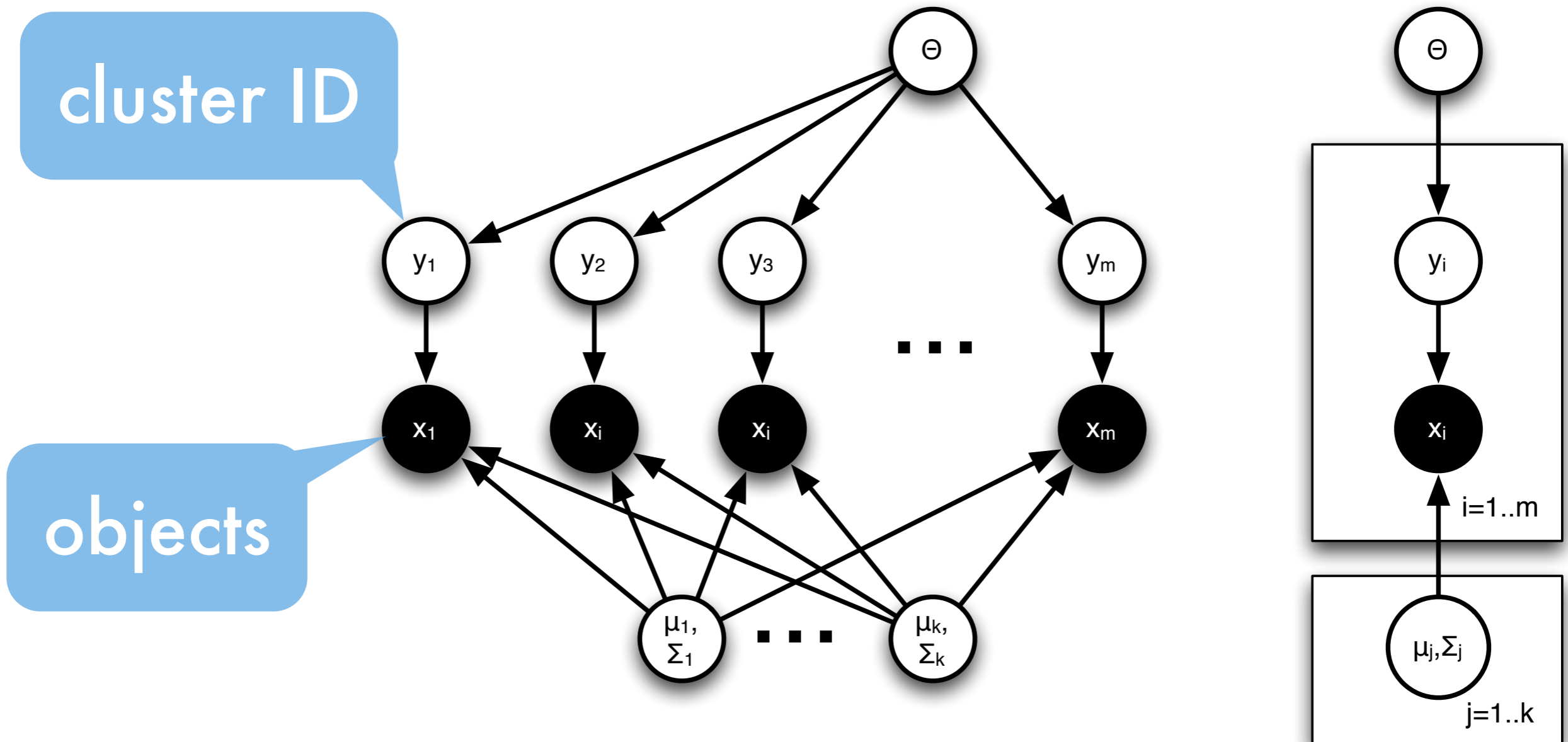
The screenshot shows the Australian National University (ANU) website. At the top, there are navigation links for "EXPLORE ANU", "A-Z INDEX", and a search bar. The main content area features a large banner image of a forest with a red speech bubble containing the word "university" overlaid on it. Below the banner, there are sections for "ABOUT NUS" and "Forests renew after Black Saturday fires". At the bottom, there are navigation buttons for "PROSPECTIVE STUDENTS", "CURRENT STUDENTS", "STAFF", "ALUMNI", and "VISITORS".

The screenshot shows the Chez Panisse restaurant website. At the top, there are navigation links for "Home", "Wining & Dining", "Contact", "Sitemap", and "About Suntec REIT". The main content area features a large banner image of a restaurant interior with a red speech bubble containing the word "restaurant" overlaid on it. Below the banner, there are sections for "RESERVATIONS", "MENUS", "ABOUT", "SPECIAL EVENTS", "STORE", and "CONTACT".

# Generative Model



# Generative Model



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

# What can we cluster?



# What can we cluster?

mails      text      urls      products

news      queries      users

spammers      ads      locations

abuse      events

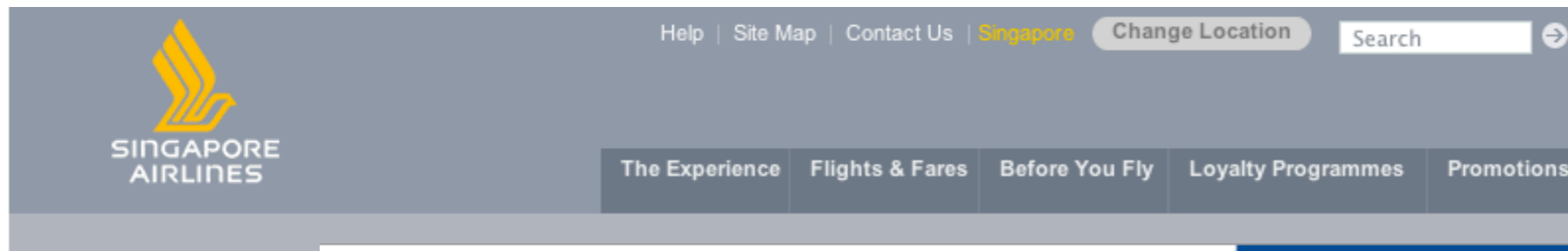
# Topic Models

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



SINGAPORE AIRLINES

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

The Experience | Flights & Fares | Before You Fly | Loyalty Programmes | Promotions

Book a Flight | Check In

Round Trip  One Way

From:



myEMAIL | IVLE | LIBRARY | MAPS | CALENDAR | SITEMAP | CONTACT | e-CARDS

Search  in  GO

ABOUT NUS | GLOBAL | ADMISSIONS | ENTERPRISE | CAMPUS LIFE | GIVING | CAREERS@NUS

Home | About Us | Services | Events & Promotions | Shopping, Wining & Dining | Contact | Sitemap

Singapore

CHIJMES  
restaurants • bars • shops

Discover a century of resplendent living history behind the cloistered walls.

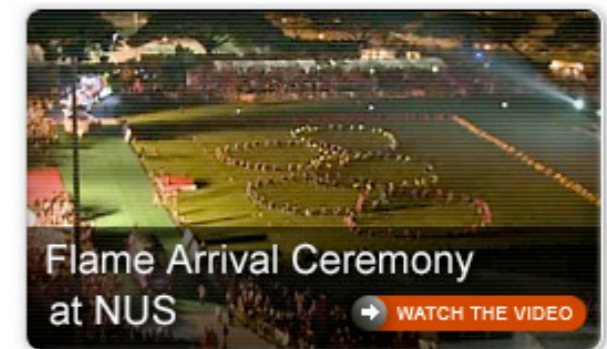
Chijmes, a premier lifestyle destination in Singapore

Owned by: Managed by: Property Manager:



Copyright © 2006 Chijmes. All rights reserved.

Feedback | Terms & Conditions



Flame Arrival Ceremony at NUS

WATCH THE VIDEO



Joint Evacuation Exercises

- 7 & 14 Sept 2010
- 10am - 12pm
- Heng Mui Keng Terrace & vicinity

MORE DETAILS

STAFF | ALUMNI | VISITORS

YAHOO!

# Grouping objects

The screenshot shows the United Airlines website interface. At the top, there's the United logo and navigation links like 'My profile', 'Worldwide sites', and 'Customer service'. Below that are menu items for 'Planning & booking', 'Reservations & check-in', 'Mileage Plus', and 'Services & information'. A search bar is also present. The main content area features a 'BOOK FLIGHT' section with fields for 'From', 'To', 'Departing', and 'Returning'. A large promotional banner in the center reads 'Use 30% fewer miles on your next United flight.' with an illustration of a person holding a large orange percentage sign. To the right is a 'Log in' section with fields for 'Mileage Plus # or email address' and 'Password'. Below the flight booking section, there are links for 'United news and deals' and 'United-Continental merger'.

The screenshot shows the Australian National University website. At the top, there's a search bar and navigation links like 'Change Location', 'SITEMAP', 'CONTACT', and 'e-CARDS'. Below that are more navigation links: 'Before You Fly', 'Loyalty Programmes', 'Promotions', 'GIVING', and 'CAREERS@NUS'. A search bar for 'Search ANU...' is visible. The main banner features the text 'The Australian National University' in a large, light blue font. Below the banner are navigation links for 'CURRENT STUDENTS', 'RESEARCH & EDUCATION', 'ABOUT ANU', and 'STAFF'. The background of the banner shows a blurred image of a building and trees.

The footer section of the United Airlines website includes the following text: 'Owned by: SUNTEC', 'Managed by: ARA', and 'Property Manager: APC'. Below the logos, there is a small image of a building at night.

# Grouping objects

The screenshot shows the United Airlines website interface. At the top, there's a navigation bar with 'UNITED' logo and links for 'My profile', 'Worldwide sites', and 'Customer service'. Below this is a search bar and a menu with categories like 'Planning & booking', 'Reservations & check-in', 'Mileage Plus', and 'Services & information'. The main content area is divided into several sections: a flight search form with fields for 'From', 'To', 'Departing', and 'Returning'; a 'Log in' section with fields for 'Mileage Plus # or email address' and 'Password'; a 'Use 30% fewer miles on your next United flight.' promotional banner; a 'United news and deals' section with various offers; and a 'KrisFlyer' section with flight prices for routes like Singapore-Bangkok (SGD 395\*), Singapore-Hong Kong (SGD 546\*), Singapore-Taipei (SGD 768\*), Singapore-Shanghai (SGD 824\*), Singapore-Sydney, Singapore-Tokyo (Haneda) (SGD 983\*), and Singapore-London. There are also 'Book Now' buttons for each route.

The screenshot shows the Australian National University (ANU) website. At the top, there's a navigation bar with 'EXPLORE ANU', 'A-Z INDEX', and a search bar. Below this is the ANU logo and the text 'The Australian National University'. The main content area features a large banner with the headline 'Ash forests rise and rise again' and a sub-headline 'A new book that graphically documents the spectacular natural recovery of Victoria's ash forests after the Black Saturday bushfires also argues that wildfires are typical natural disturbances in these environments.' Below the banner are several navigation links: 'HOME', 'FUTURE STUDENTS', 'CURRENT STUDENTS', 'RESEARCH & EDUCATION', 'ABOUT ANU', and 'STAFF'. There are also buttons for 'PROSPECTIVE STUDENTS', 'CURRENT STUDENTS', 'STAFF', 'ALUMNI', and 'VISITORS'. A 'Joint Evacuation Exercises' banner is visible at the bottom right.

The screenshot shows the Chez Panisse website. The main content area is a vertical list of navigation links: 'RESERVATIONS', 'MENUS', 'ABOUT', 'SPECIAL EVENTS', 'STORE', and 'CONTACT'. Each link is accompanied by a small icon or text describing the page content. The background features a photograph of the restaurant's interior.



ng, Wining & Dining | Contact | Sitemap | About Suntec REIT



# Grouping objects

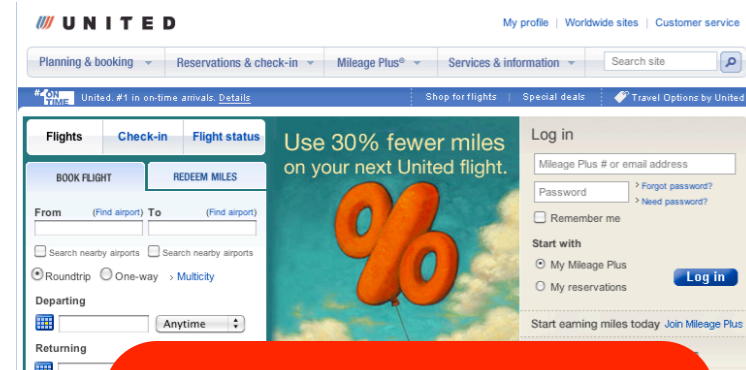
The image shows a screenshot of the United Airlines website. The page features a navigation bar with 'UNITED' and links for 'My profile', 'Worldwide sites', and 'Customer service'. Below the navigation, there are sections for 'Flights', 'Check-in', and 'Flight status'. A prominent red speech bubble with the word 'airline' is overlaid on the page. The website content includes a search form for flights, a 'Use 30% fewer miles' promotion, and a list of flight routes with prices.

Route	Price (SGD)
Singapore - Bangkok	395*
Singapore - Hong Kong (Ball)	546*
Singapore - Taipei	768*
Singapore - Shanghai	824*
Singapore - Tokyo (Haneda)	983*
Singapore - Sydney	
Singapore - London	

The image shows a screenshot of the Australian National University (ANU) website. The page features a navigation bar with 'EXPLORE ANU', 'A-Z INDEX', and a search bar. Below the navigation, there are sections for 'HOME', 'FUTURE STUDENTS', 'CURRENT STUDENTS', 'RESEARCH & EDUCATION', 'ABOUT ANU', and 'STAFF'. A prominent red speech bubble with the word 'university' is overlaid on the page. The website content includes a news article titled 'Ash forests rise and rise again' and a list of navigation buttons for 'PROSPECTIVE STUDENTS', 'CURRENT STUDENTS', 'STAFF', 'ALUMNI', and 'VISITORS'.

The image shows a screenshot of the Chez Panisse restaurant website. The page features a navigation bar with 'Home', 'Wining & Dining', 'Contact', 'Sitemap', and 'About Suntec REIT'. Below the navigation, there is a large image of the restaurant's exterior at night. A prominent red speech bubble with the word 'restaurant' is overlaid on the page. The website content includes a list of navigation buttons for 'RESERVATIONS', 'MENUS', 'ABOUT', 'SPECIAL EVENTS', 'STORE', and 'CONTACT'.

# Grouping objects



UNITED My profile | Worldwide sites | Customer service

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

Use 30% fewer miles on your next United flight.

Log in

Mileage Plus # or email address

Password  Forgot password? Need password?

Remember me

Start with

My Mileage Plus

My reservations

Start earning miles today Join Mileage Plus

USA



RESERVATIONS RESTAURANT & CAFÉ

MENUS RESTAURANT • CAFÉ MONDAY NIGHTS • WINE LIST

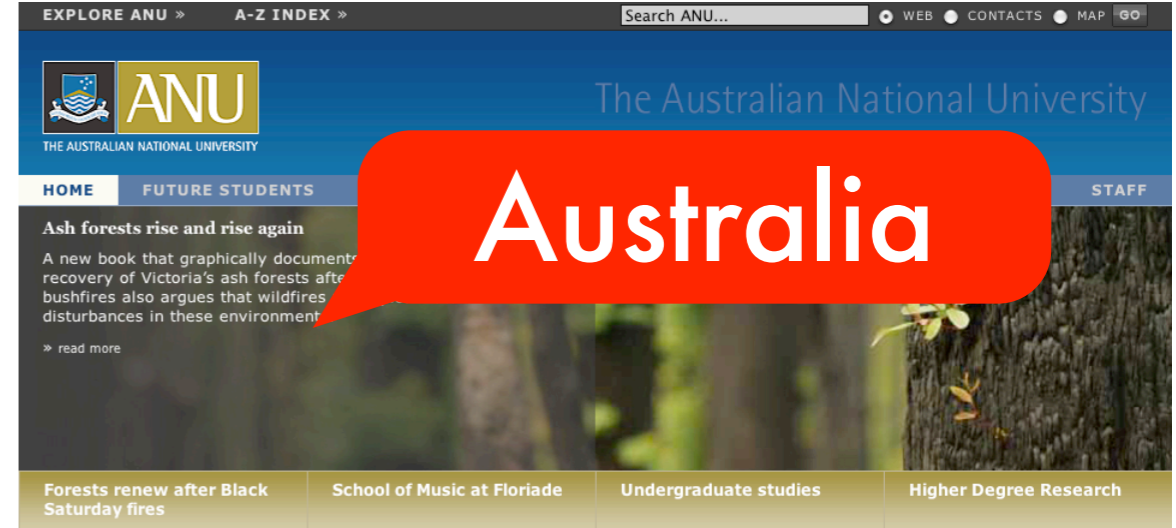
ABOUT CHEZ PANISSE • ALICE WATERS OUR CHEFS • FRIENDS • PRESS FOUNDATION & MISSION

SPECIAL EVENTS CALENDAR

STORE BOOKS • POSTERS • GIFTS

CONTACT INFORMATION DIRECTIONS • MAILING LIST

© 1998-2010 Chez Panisse Restaurant & Café. All Rights Reserved.



EXPLORE ANU » A-Z INDEX » Search ANU... WEB CONTACTS MAP GO

ANU THE AUSTRALIAN NATIONAL UNIVERSITY

The Australian National University

HOME FUTURE STUDENTS STAFF

Ash forests rise and rise again

A new book that graphically documents recovery of Victoria's ash forests after bushfires also argues that wildfires disturbances in these environment

read more

Forests renew after Black Saturday fires School of Music at Floriade Undergraduate studies Higher Degree Research

Australia



SINGAPORE AIRLINES

Book a Flight | Check in | Flight Status | My Bookings

Round Trip One Way Stopover/Multi-city

From: Departure City To: Destination City

Must travel on these dates

Adults: 1 Children (2-11): 0 Infants: 0

Need Help? View Book A Flight G

SIA Holidays Hotel Bookings

NUS National University of Singapore

myEMAIL IVLE LIBRARY MAPS CALENDAR SITEMAP CONTACT e-CARDS

Search search for... in NUS Websites GO

ABOUT NUS GLOBAL ADMISSIONS EDUCATION RESEARCH ENTERPRISE CAMPUS LIFE GIVING CAREERS@NUS

A Leading Global University Centred in Asia

Home | About Us | Services | Events & Promotions | Shopping, Wining & Dining | Contact | Sitemap | About Suntec REIT

Flame Arrival Ceremony at NUS WATCH THE VIDEO

Joint Evacuation Exercises

7 & 14 Sept 2010

10am - 12pm

Heng Mui Keng Terrace & vicinity

MORE DETAILS

ALUMNI VISITORS

Singapore



CHIJMES restaurant

Discover living in Singapore

Chijmes, a premier lifestyle destination in Singapore

Owned by: Managed by: Property Manager:

SUNTEC ARA PC

Copyright © 2006 Chijmes. All rights reserved. Feedback | Terms & Conditions

YAHOO!



# Topic Models

UNITED My profile | Worldwide sites | Customer service

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

Use 30% fewer miles on your next United flight.

BOOK FLIGHT REDEEM MILES

From (Find airport) To (Find airport)

Roundtrip One-way Multicity

Departing Anytime

Returning Anytime

Search by Schedule & price Price & Flex

Adult (child or senior?)

Cabin Economy Refundable

Promotion code or Electronic certificate

Log in to view all seating options

Advanced Search

Cars Hotels Vacations

Learn more

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

A STAR ALLIANCE MEMBER

USA  
airline

EXPLORE ANU » A-Z INDEX » Search ANU... WEB CONTACTS

ANU THE AUSTRALIAN NATIONAL UNIVERSITY

HOME FUTURE STUDENTS CUR ABOUT ANU

Ash forests rise and rise again

A new book that graphically documents the recovery of Victoria's ash forests after the bushfires also argues that wildfires are typical disturbances in these environments.

Forests renew after Black Saturday fires

School of Music at Monash

Undergraduate studies

Higher Degree Research

Australia  
university

SINGAPORE AIRLINES

The Experience | Flights & Fares | Before You Fly | Loyalty Programmes | Promotions

Book a Flight | Check In | Flight Status | My Bookings | Member Log-in

Round Trip One Way Stopover/Multi-city

From: Depart: Departure City

To: Return: Destination City

Must travel on these dates

Adults: Children (2-11): Infants:

Need Help? View Book A Flight

SIA Holidays Hotel Bookings

Singapore - Bangkok SGD 395\*

Singapore - Hong Kong SGD 546\*

Singapore - Taipei SGD 768\*

Singapore - Tokyo (Haneda) SGD 983\*

Singapore - Sydney

Singapore - London

Singapore  
airline

NUS National University of Singapore

myEMAIL IVLE LIBRARY MAPS CALENDAR SITEMAP CONTACT CARDS

Search search for... in NUS Websites GO

ABOUT NUS GLOBAL ADMISSIONS EDUCATION RESEARCH ENTERPRISE CAMPUS LIFE GIVING CAREERS@NUS

A Leading Global University

Game Arrival Ceremony

Joint Evacuation Exercises

7 & 14 Sept 2010

10am - 12pm

Heng Mui Keng Terrace & vicinity

PROSPECTIVE STUDENTS CURRENT STUDENTS STAFF ALUMNI VISITORS

Singapore  
university

Chez Panisse

RESERVATIONS RESTAURANT & CAFÉ

MENUS RESTAURANT • CAFÉ MONDAY NIGHTS • WINE LIST

ABOUT CHEZ PANISSE • ALICE WATERS OUR CHEFS • FRIENDS • PRESS FOUNDATION & MISSION

SPECIAL EVENTS CALENDAR

STORE BOOKS • POSTERS • GIFTS

CONTACT INFORMATION DIRECTIONS • MAILING LIST

© 1998-2010 Chez Panisse Restaurant & Café. All Rights Reserved.

Directions Reservations Contact

USA  
food

Services | Events & Promotions | Shopping, Wining & Dining | Contact | Sitemap | About Suntec REIT

Chijmes

restaurants • bars • shops

Discover a century of resplendent living history behind the cloisters

Chijmes, a premier lifestyle destination in Singapore

Owned by: SUNTEC

Managed by: ARA

Property Manager: APC

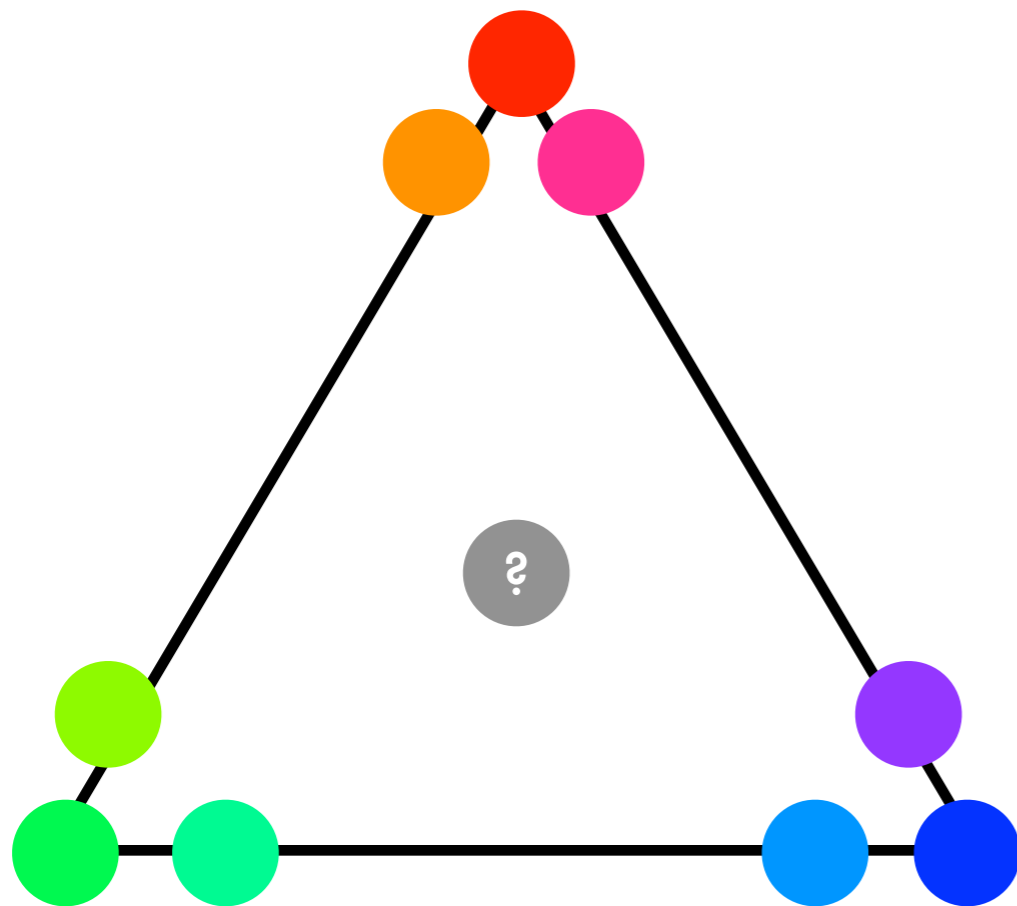
Copyright © 2006 Chijmes. All rights reserved.

Feedback | Terms & Conditions

Singapore  
food

# Clustering & Topic Models

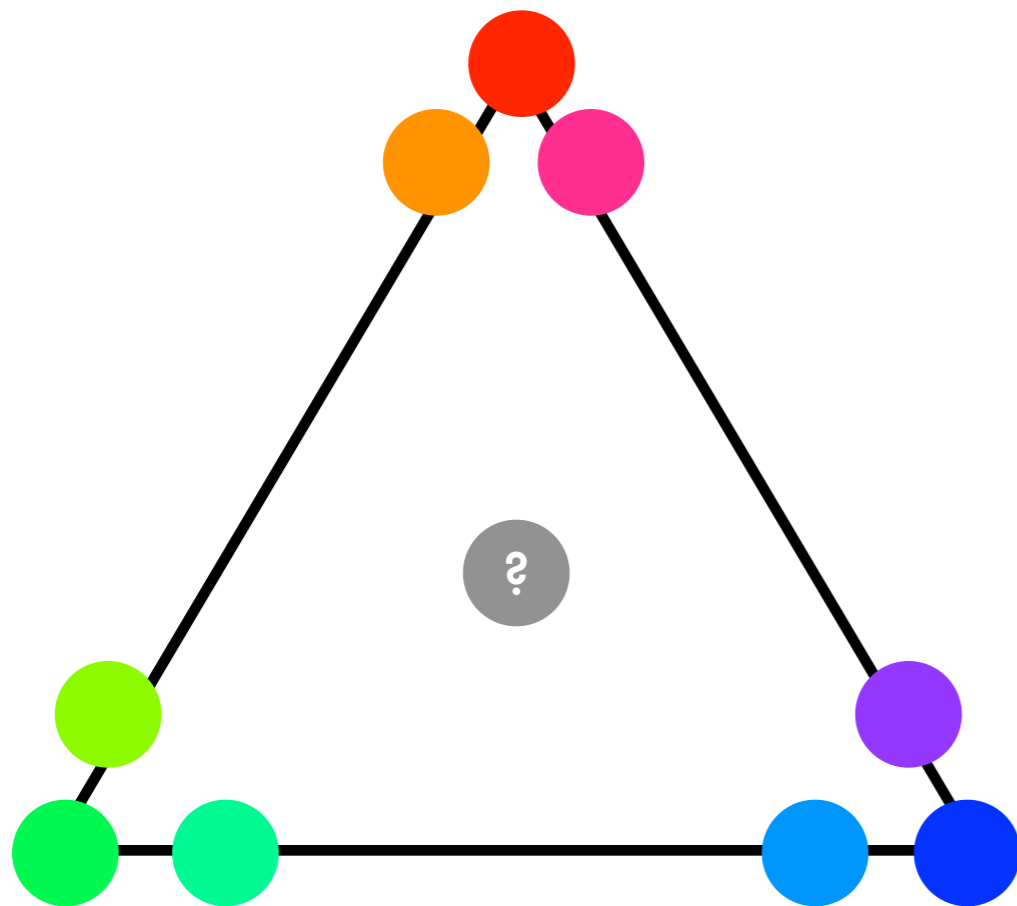
Clustering



group objects  
by prototypes

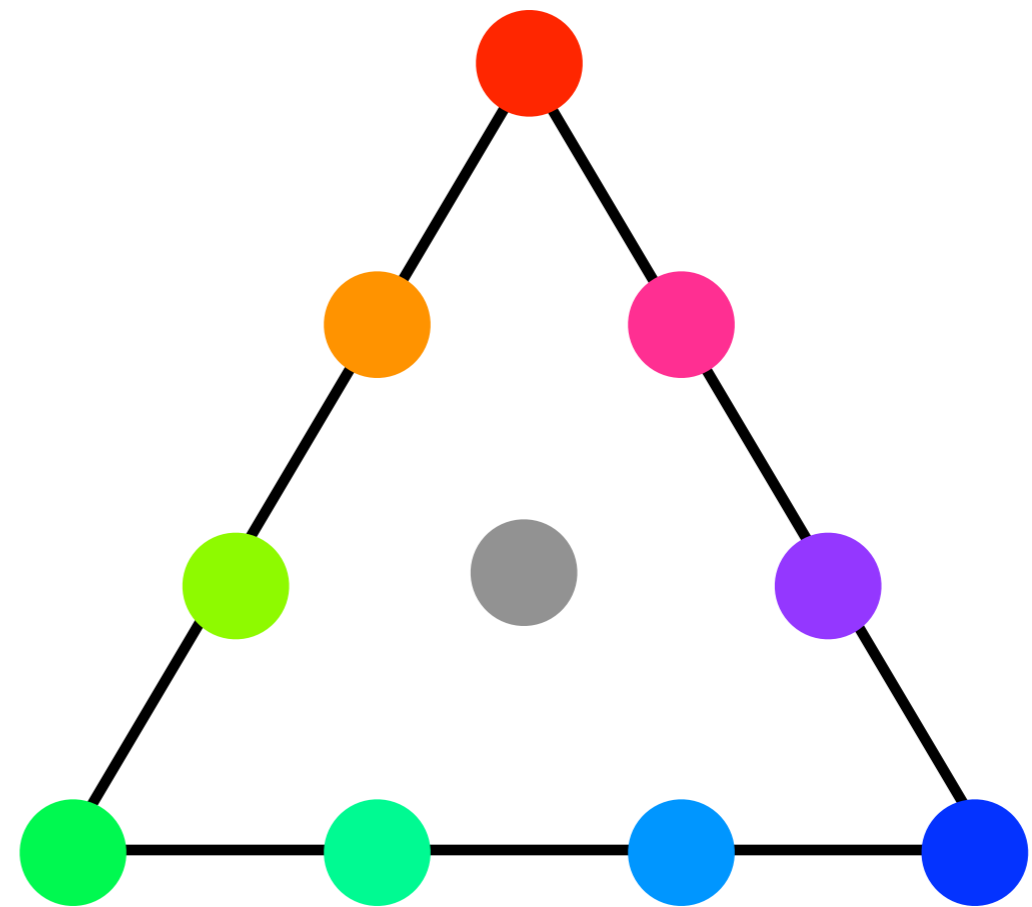
# Clustering & Topic Models

Clustering



group objects  
by prototypes

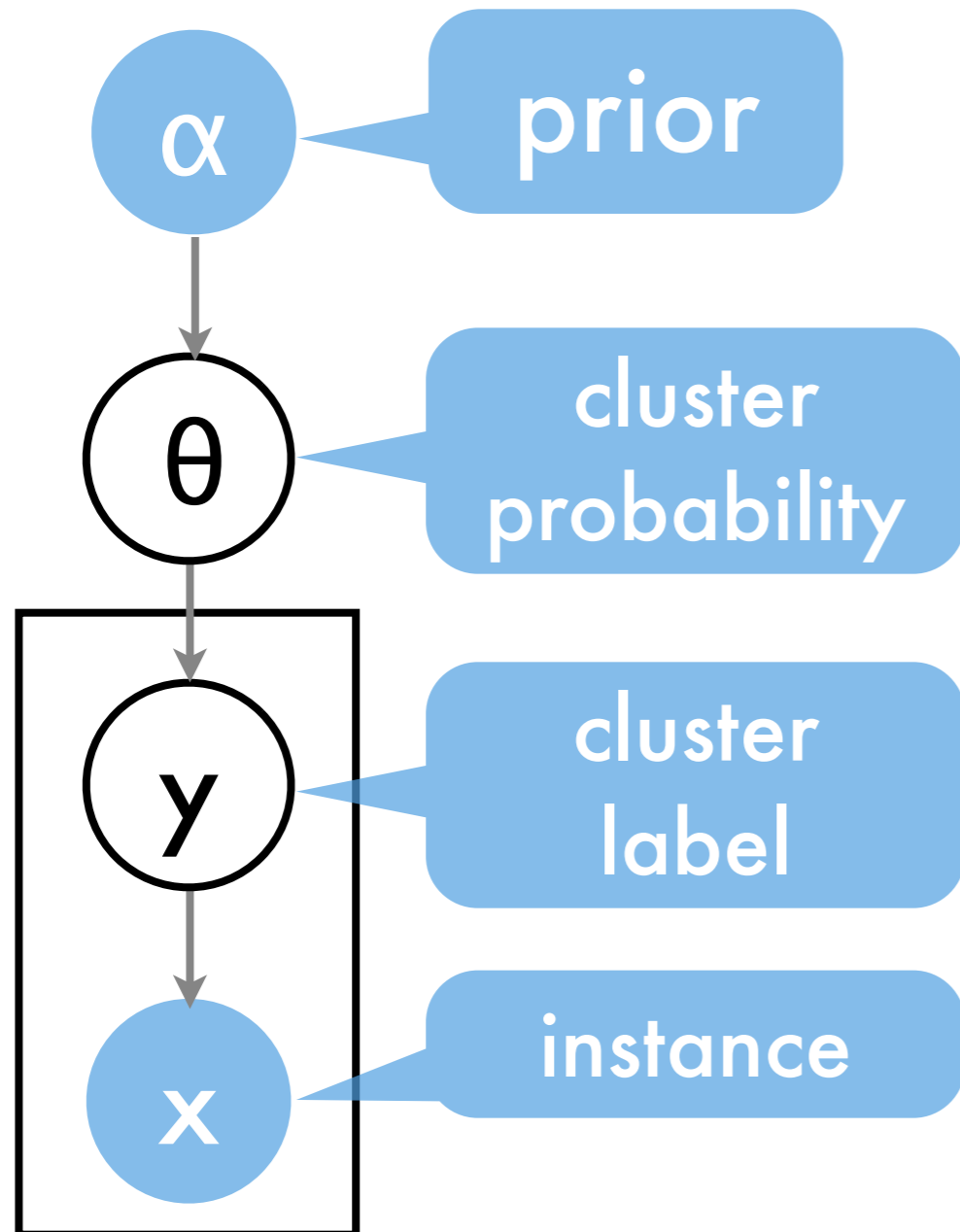
Topics



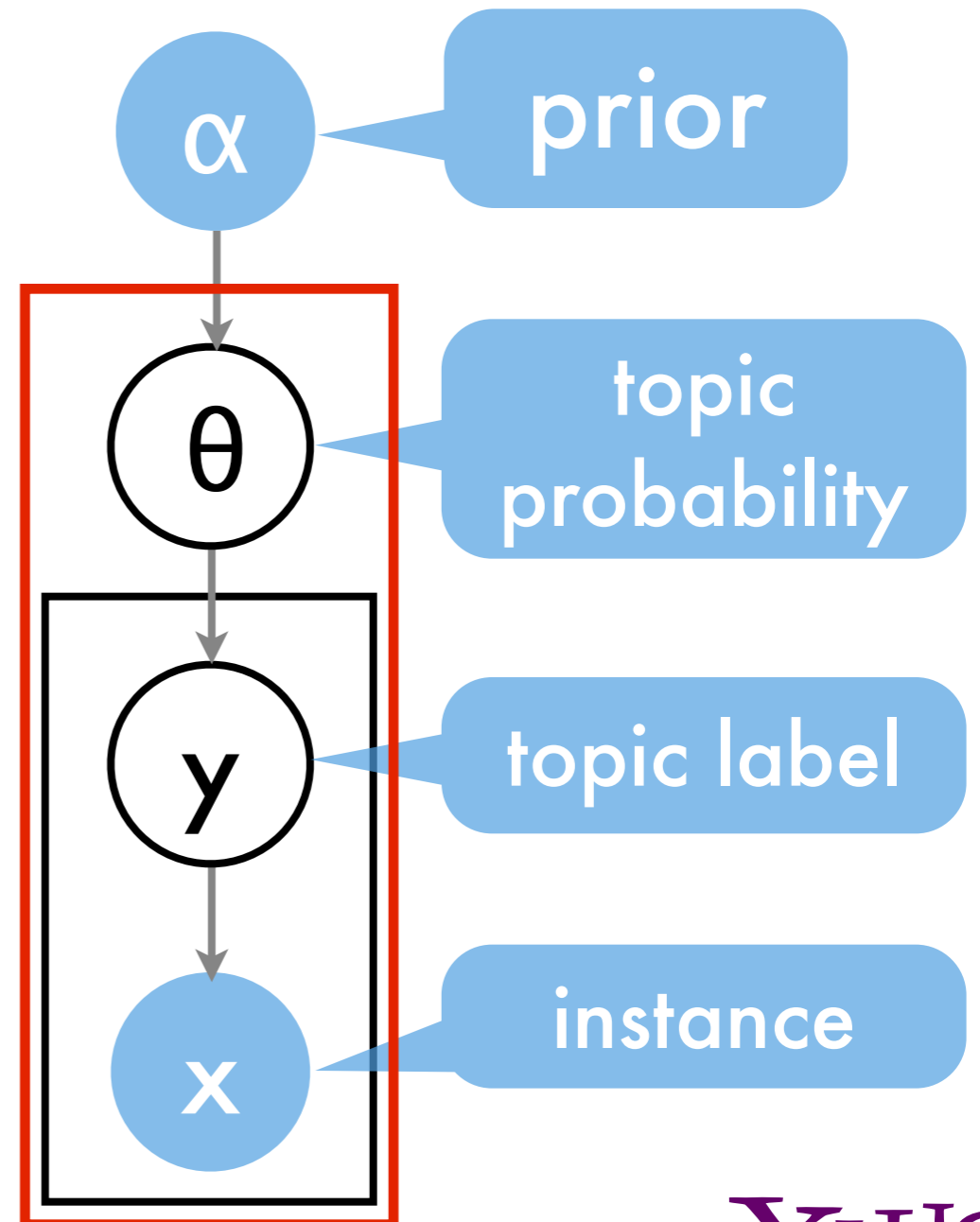
decompose objects  
into prototypes

# Clustering & Topic Models

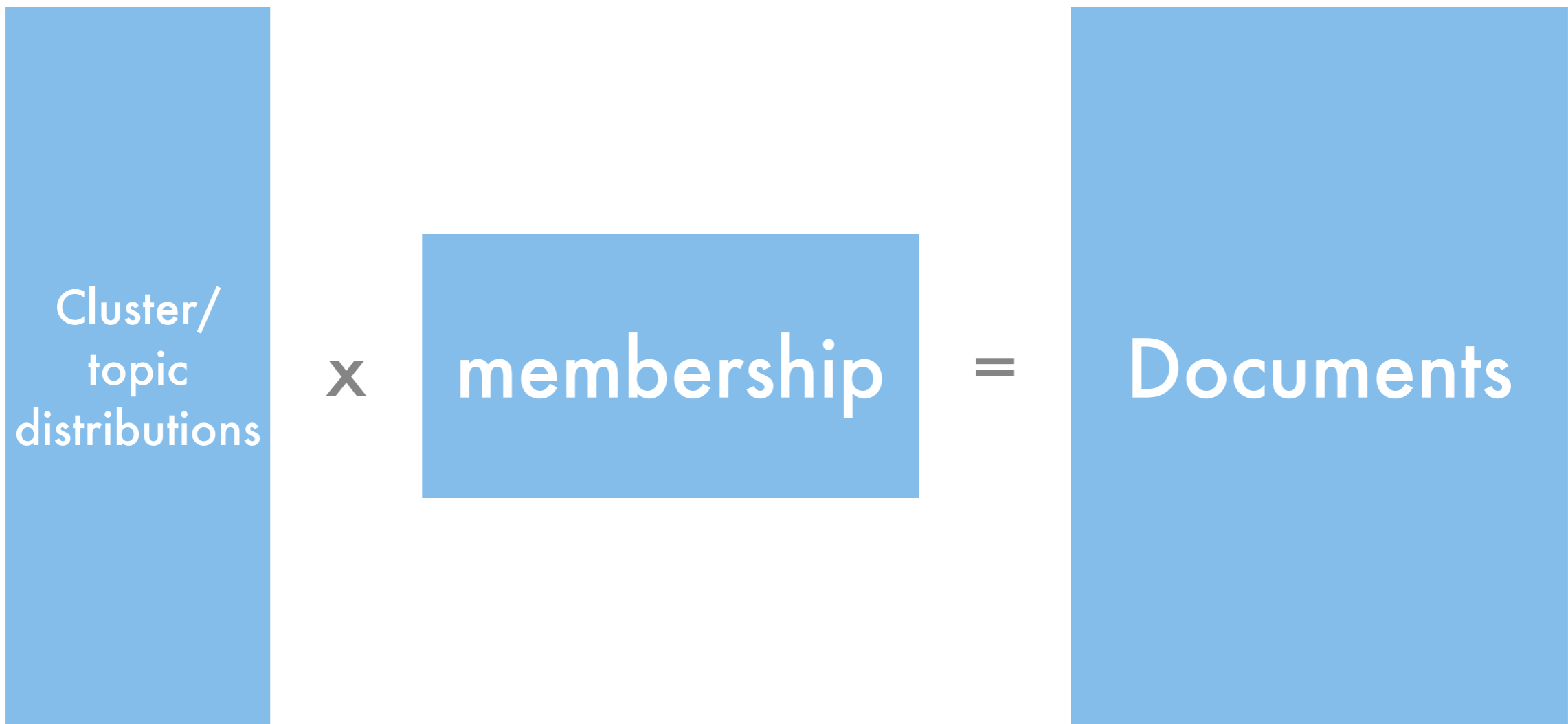
clustering



Latent Dirichlet Allocation



# Clustering & Topic Models



clustering: (0, 1) matrix  
topic model: stochastic matrix  
LSI: arbitrary matrices

# 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**

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

goodness of fit

complexity penalty

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



MAGIC Etch A Sketch<sup>®</sup> SCREEN

Systems  
to run our algorithms on



Horizontal  
Lid

OHIO ART "A World of Toys"

MAGIC SCREEN IS GLASS SET IN DURABLE PLASTIC FRAME  
USE WITH CARE

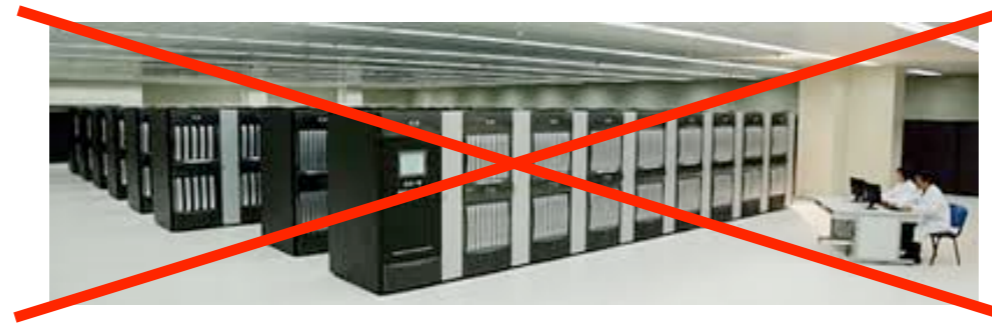
Vertical  
Lid





# 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**

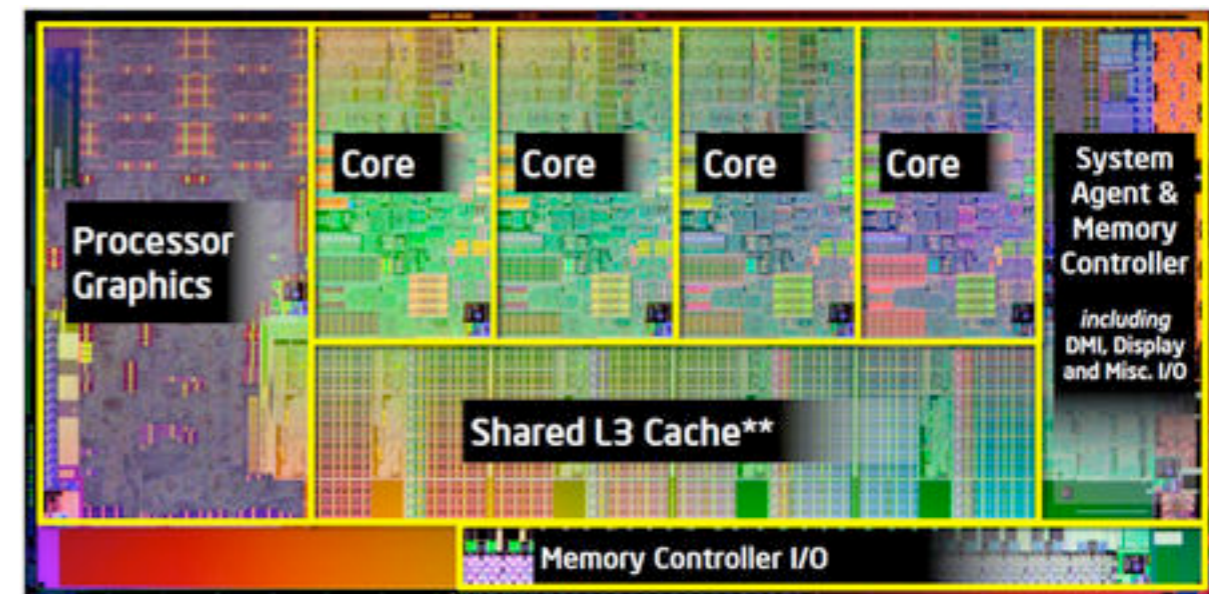
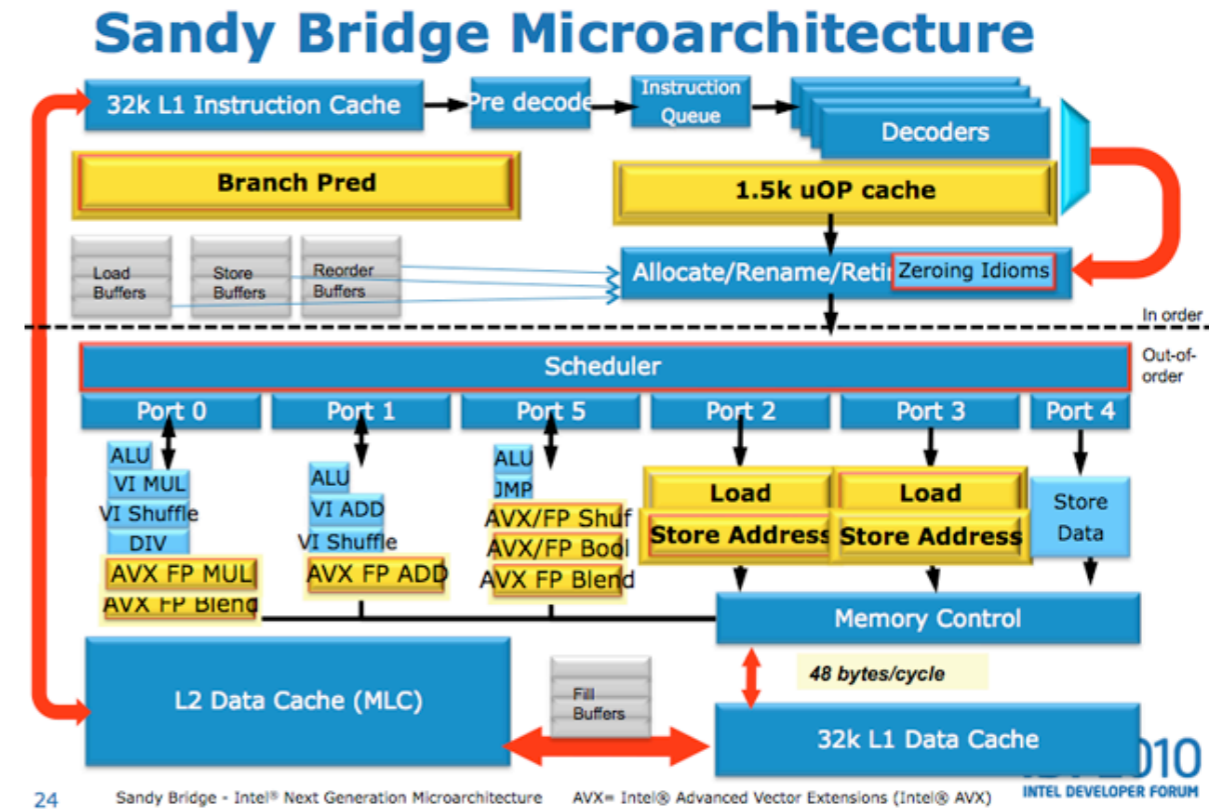
slow disks, bad memory, misconfigured machines, flaky machines, etc.

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



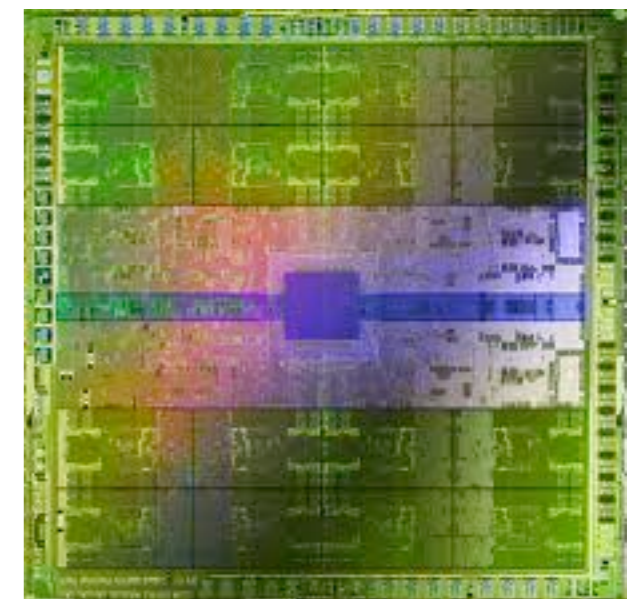
# RAM

- 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)



# GPU

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



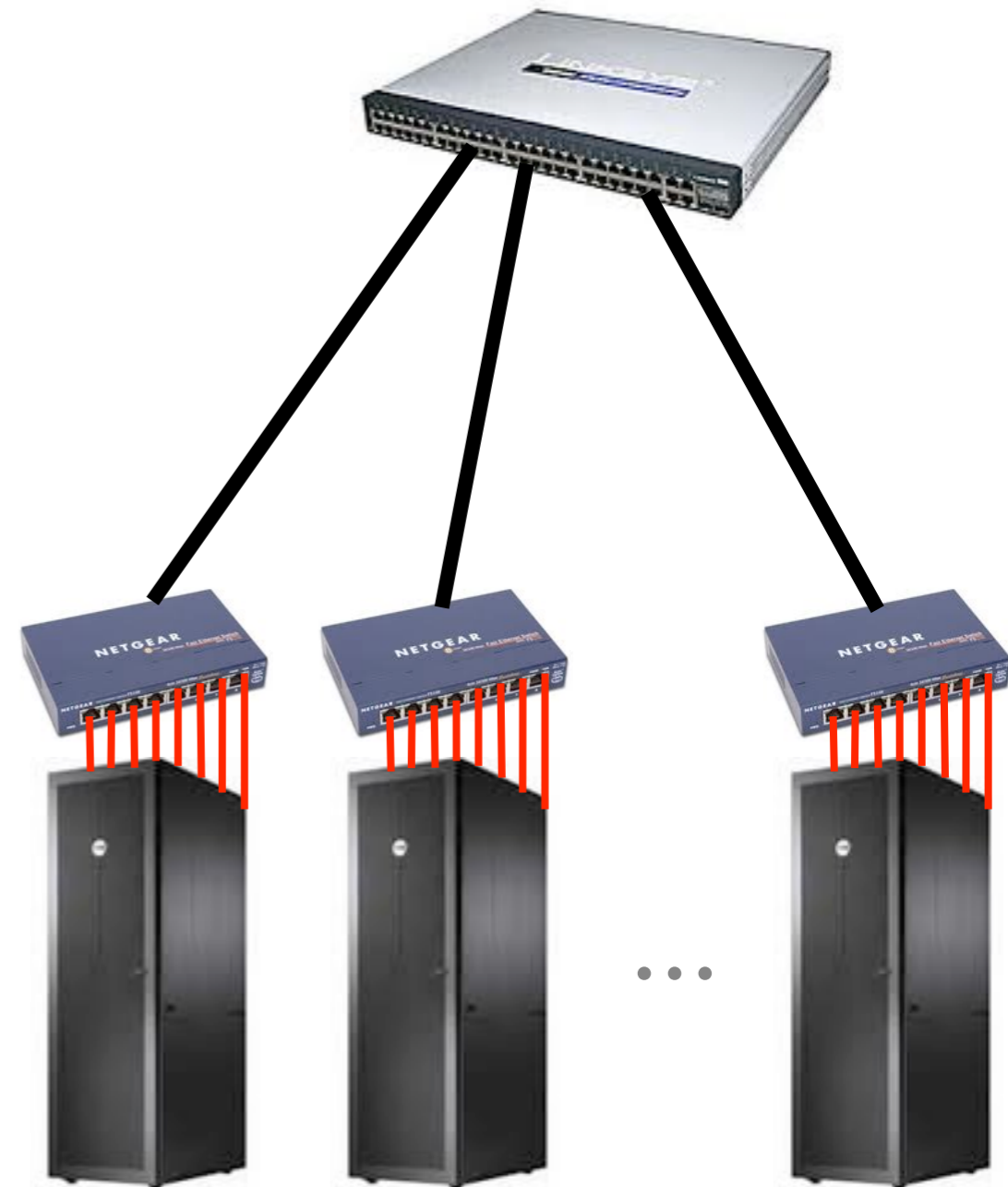
# Storage

- 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

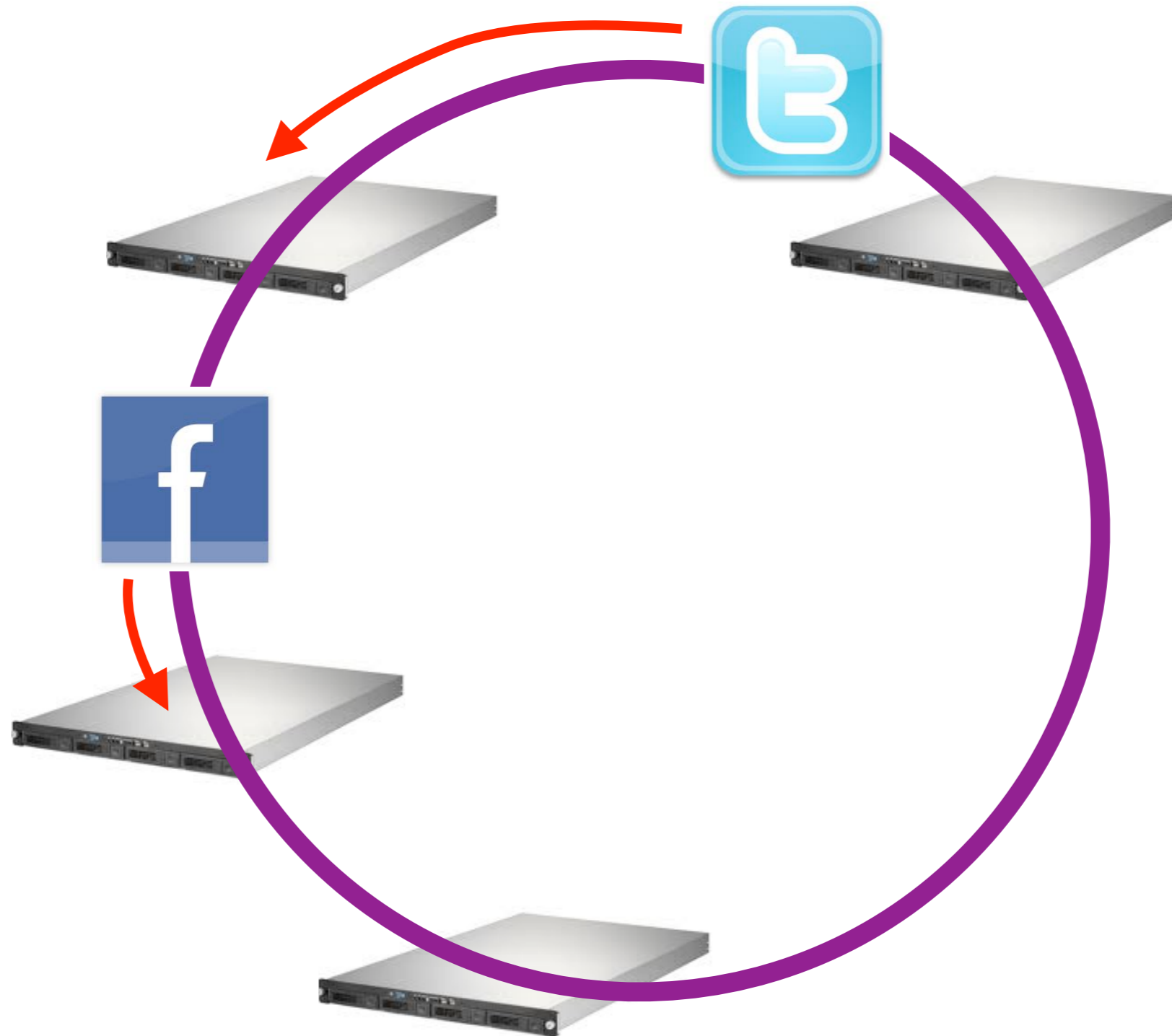
L1 cache reference	0.5 ns
Branch mispredict	5 ns
L2 cache reference	7 ns
Mutex lock/unlock	100 ns
Main memory reference	100 ns
Compress 1K bytes with Zippy	10,000 ns
Send 2K bytes over 1 Gbps network	20,000 ns
Read 1 MB sequentially from memory	250,000 ns
Round trip within same datacenter	500,000 ns
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**

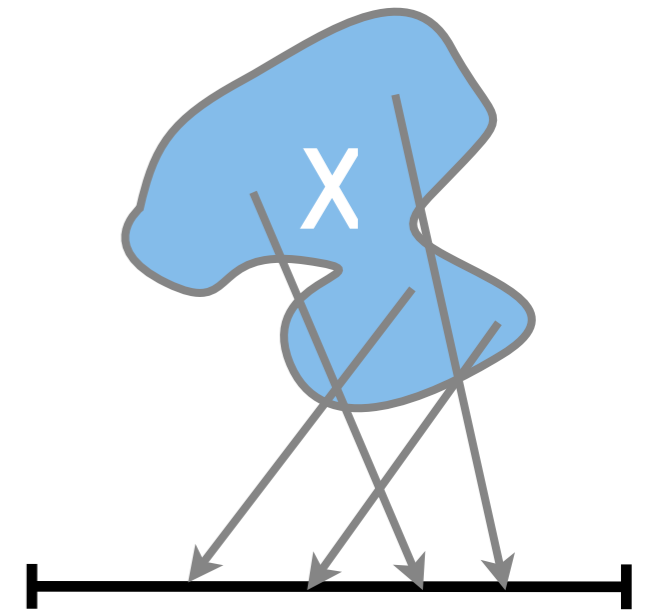
# Hash function

- 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

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

- 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  $M$
- 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

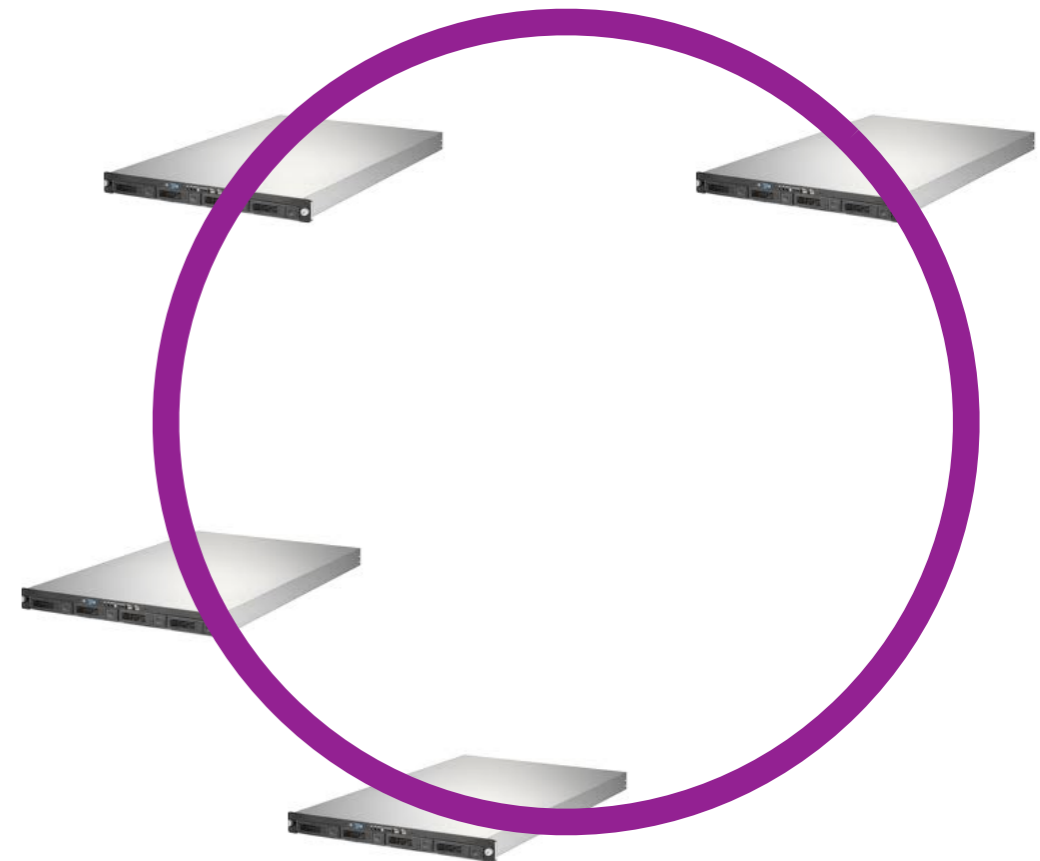
$$m(\text{key}, k) = k \text{ smallest } h(\text{key}, m)_{m \in \mathcal{M}}$$

- 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)

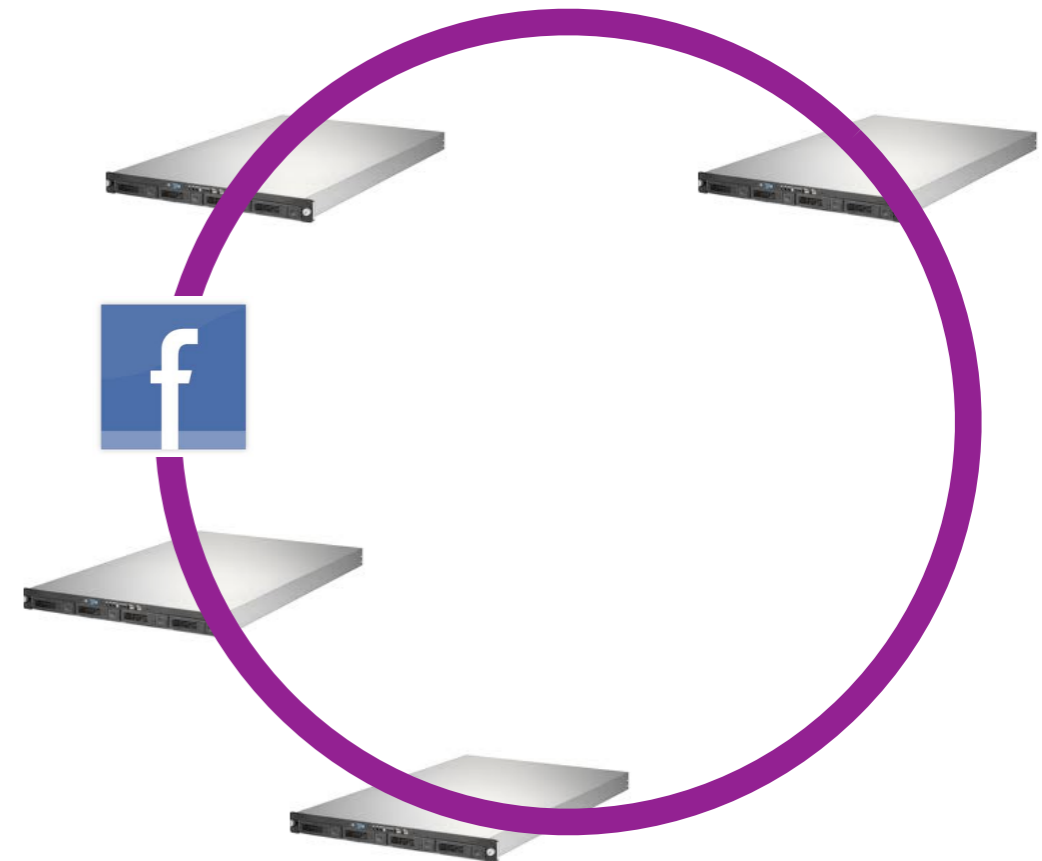
ring of  $N$  keys



# 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)

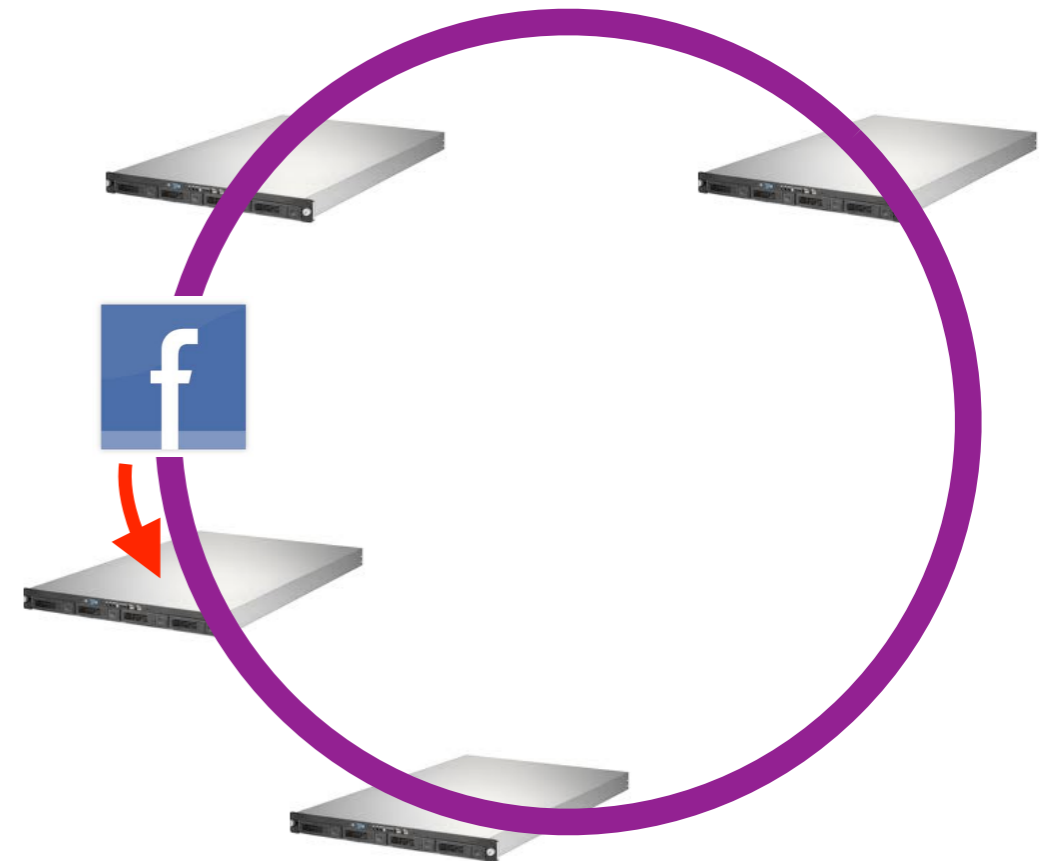
ring of  $N$  keys



# Distributed Hash Table

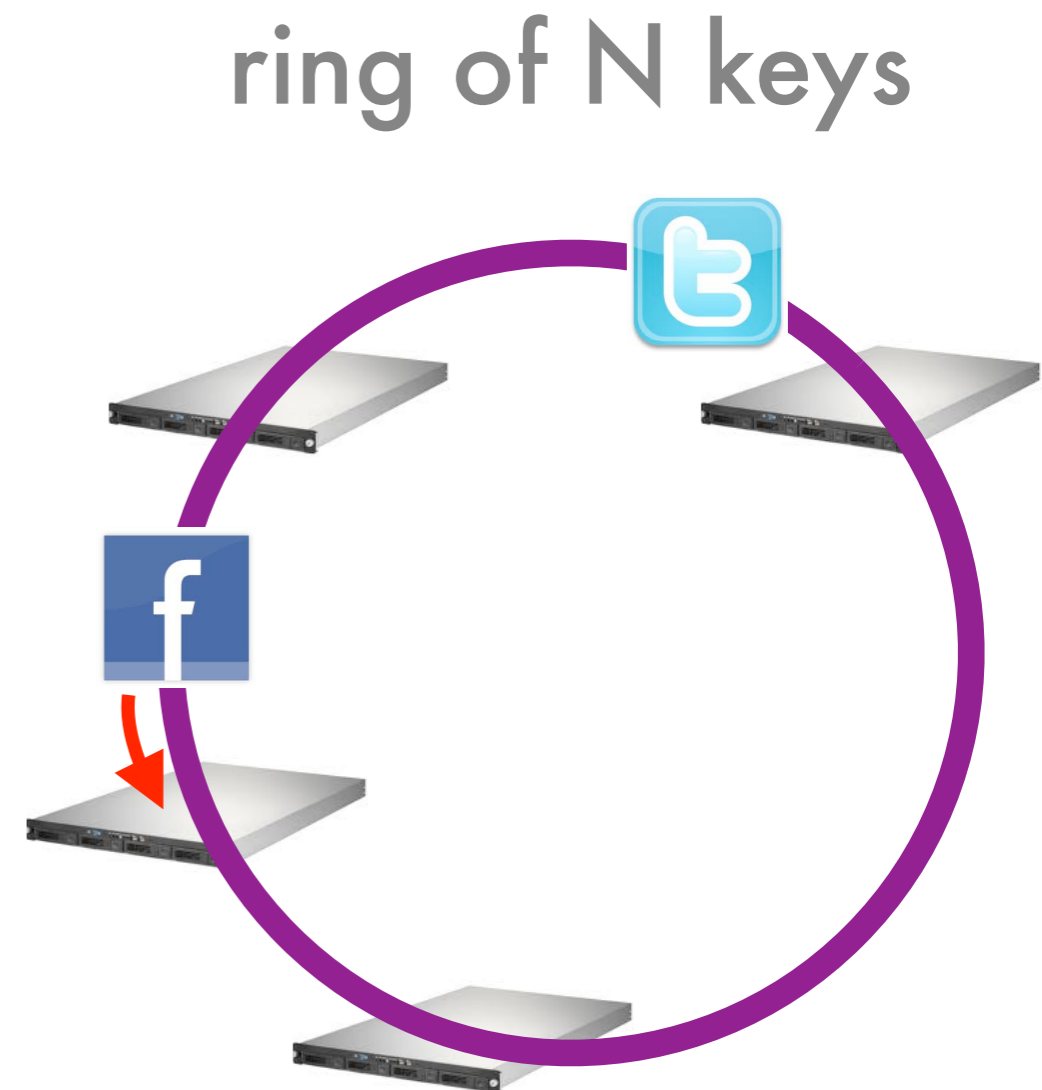
- Fixing the  $O(m)$  lookup
  - Assign machines to ring via hash  $h(m)$
  - Assign keys to ring
  - Pick machine nearest to key to the left
- $O(\log m)$  lookup
- Insert/removal only affects neighbor (however, big problem for neighbor)
- Uneven load distribution (load depends on segment size)
- Insert 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

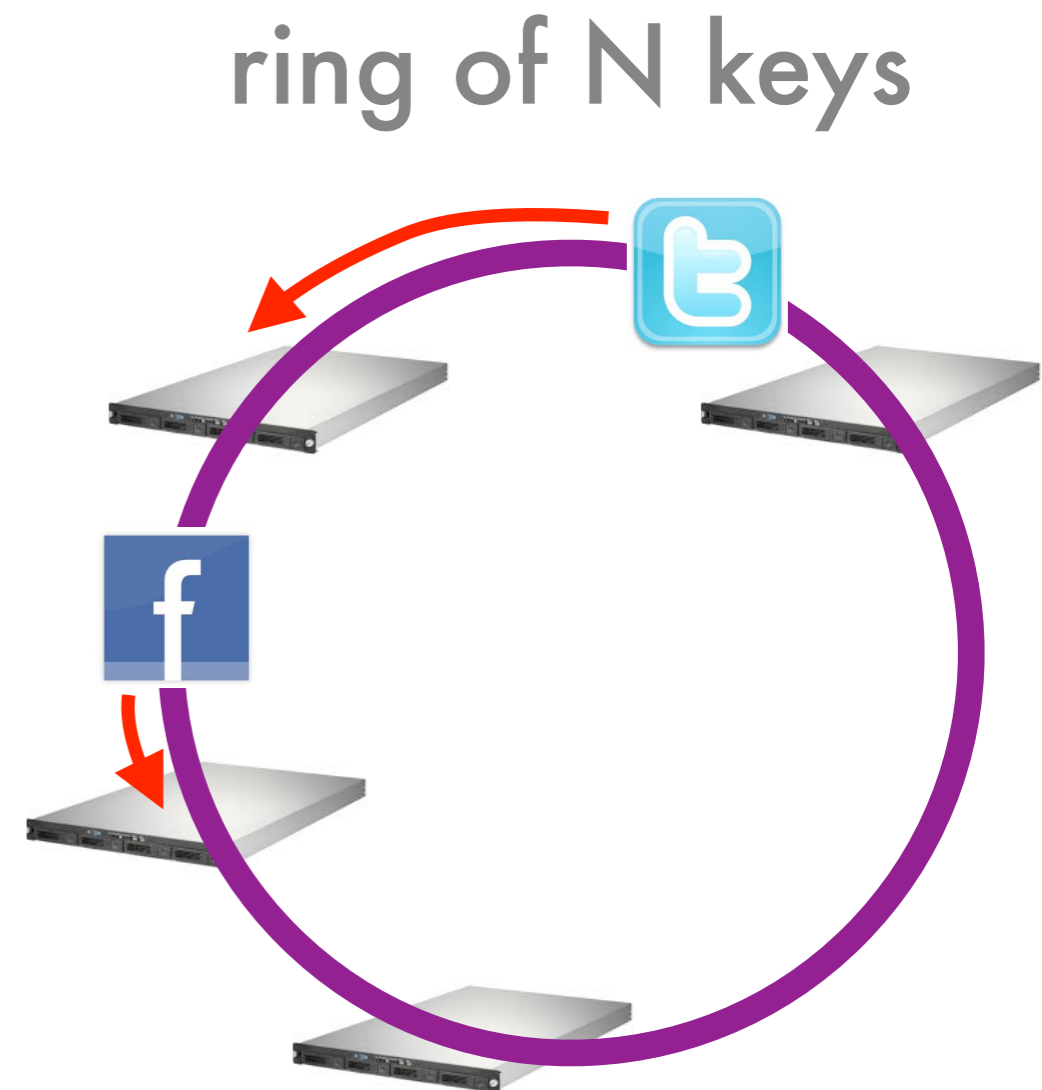
- Fixing the  $O(m)$  lookup
  - Assign machines to ring via hash  $h(m)$
  - Assign keys to ring
  - Pick machine nearest to key to the left
- $O(\log m)$  lookup
- Insert/removal only affects neighbor (however, big problem for neighbor)
- Uneven load distribution (load depends on segment size)
- Insert 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

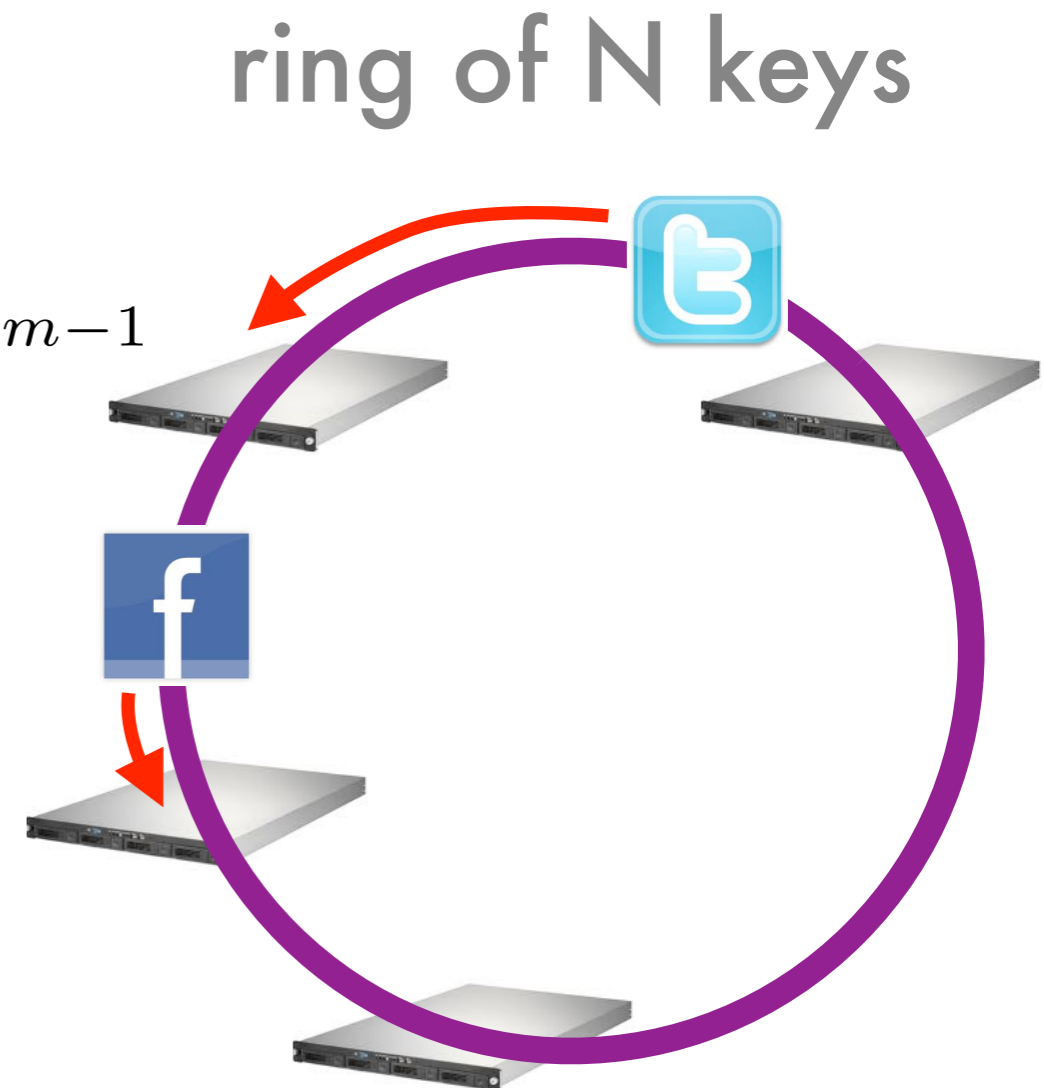
$$\Pr \{x \geq c\} = \prod_{i=2}^m \Pr \{s_i \geq c\} = (1 - c)^{m-1}$$

- 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  $m$ )

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



# 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

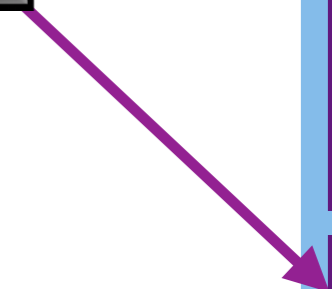
# Proportional Allocation Table

- 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)



# 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

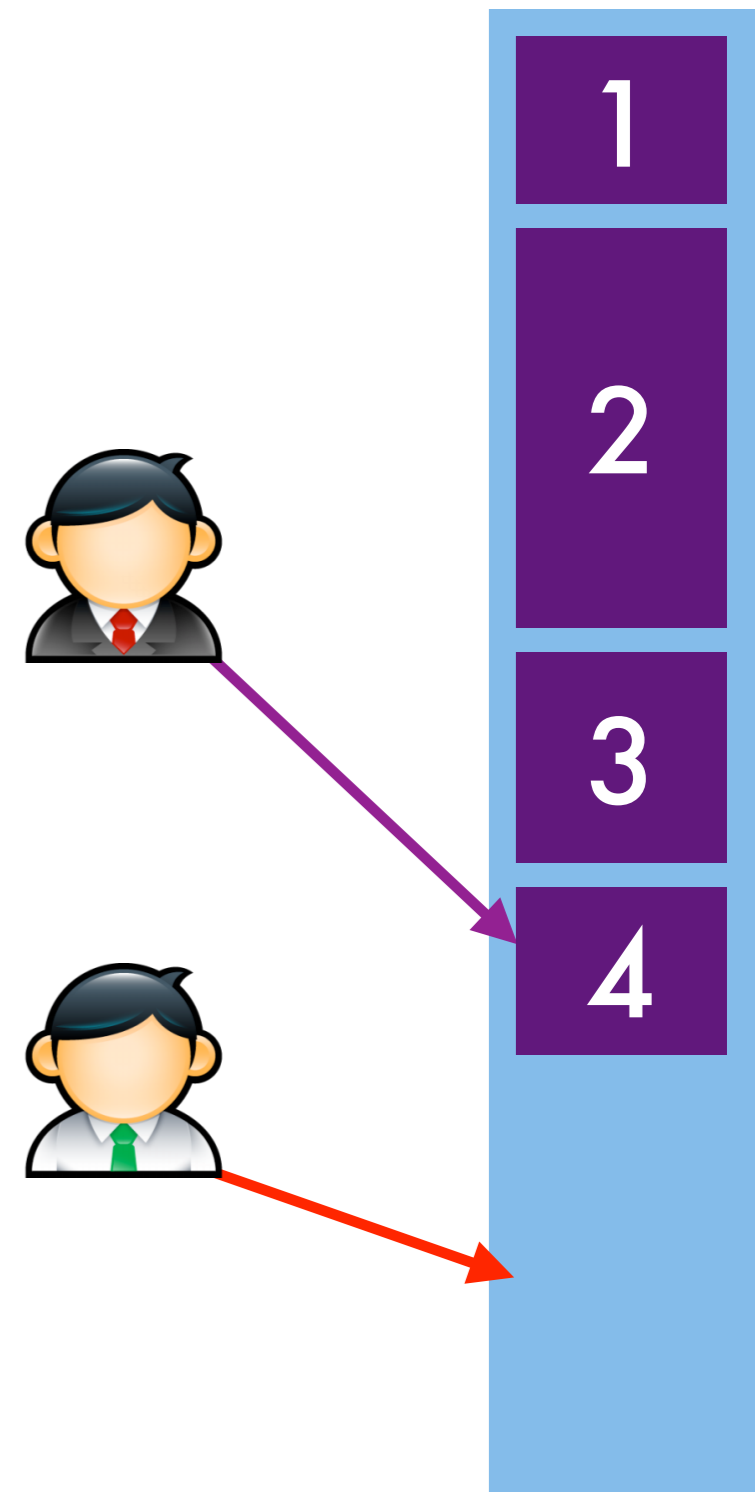
# Proportional Allocation Table

- 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)



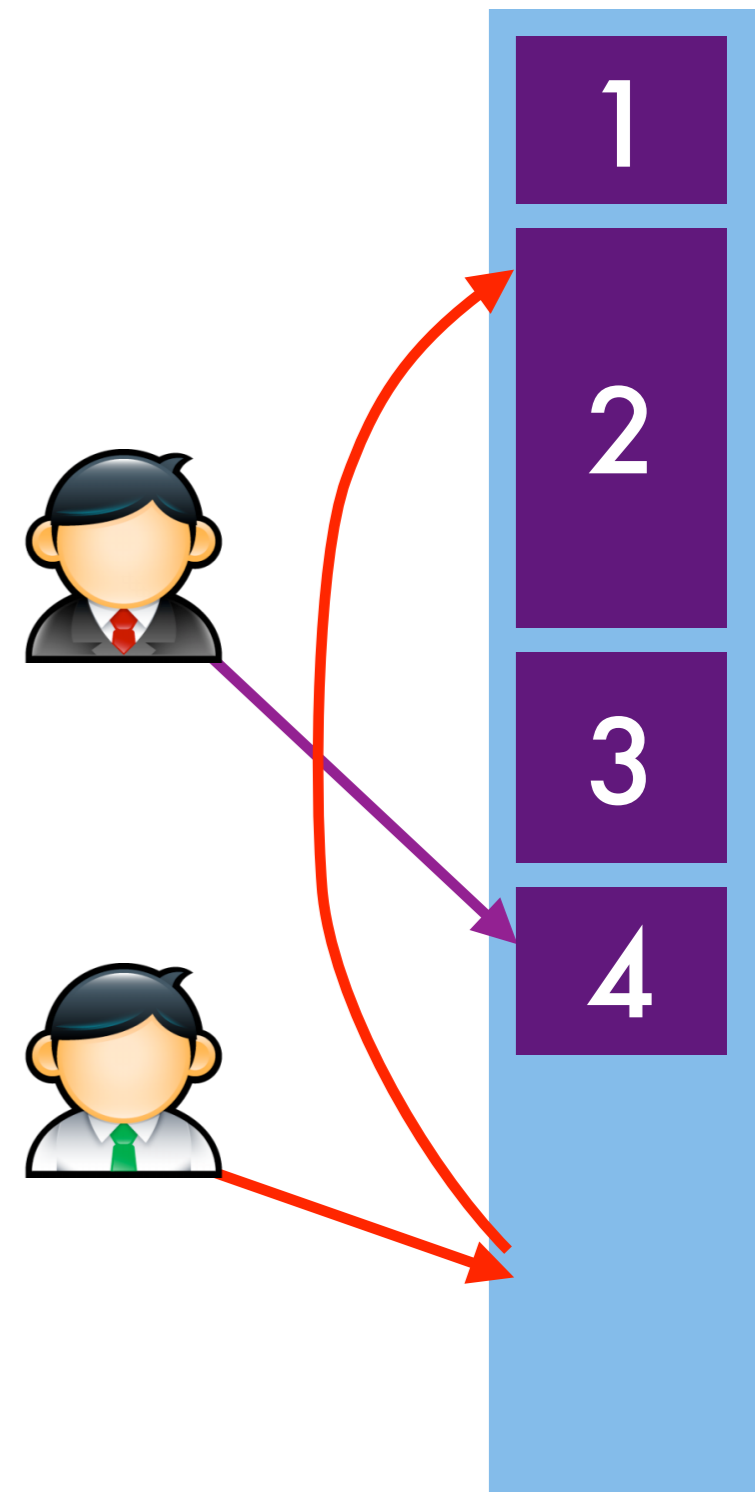
# Proportional Allocation Table

- 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)

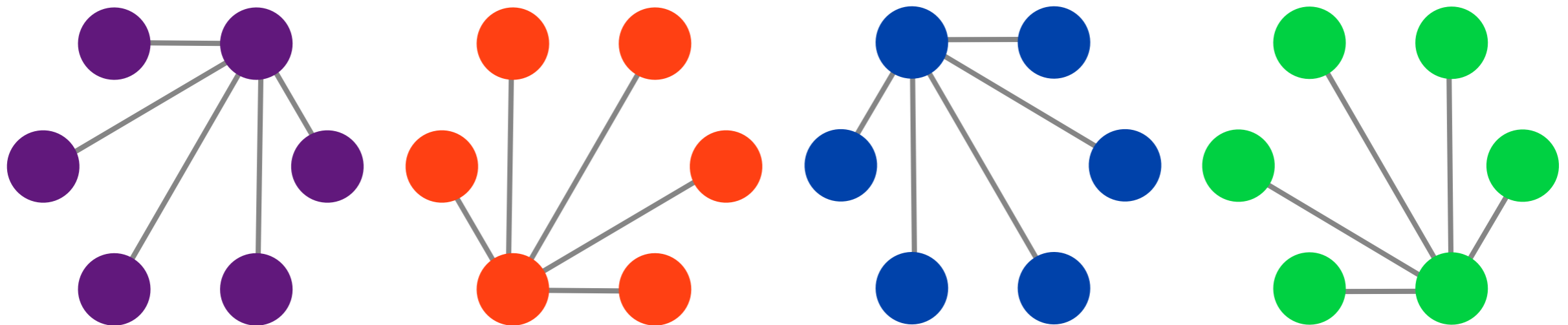




# Random Caching Trees

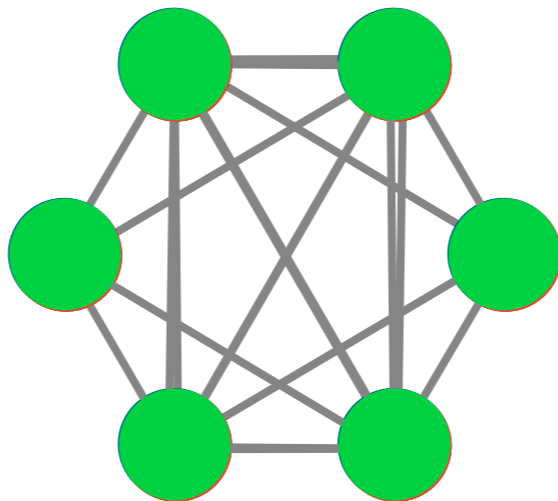
(Karger et al. 1999, Akamai paper)

- 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

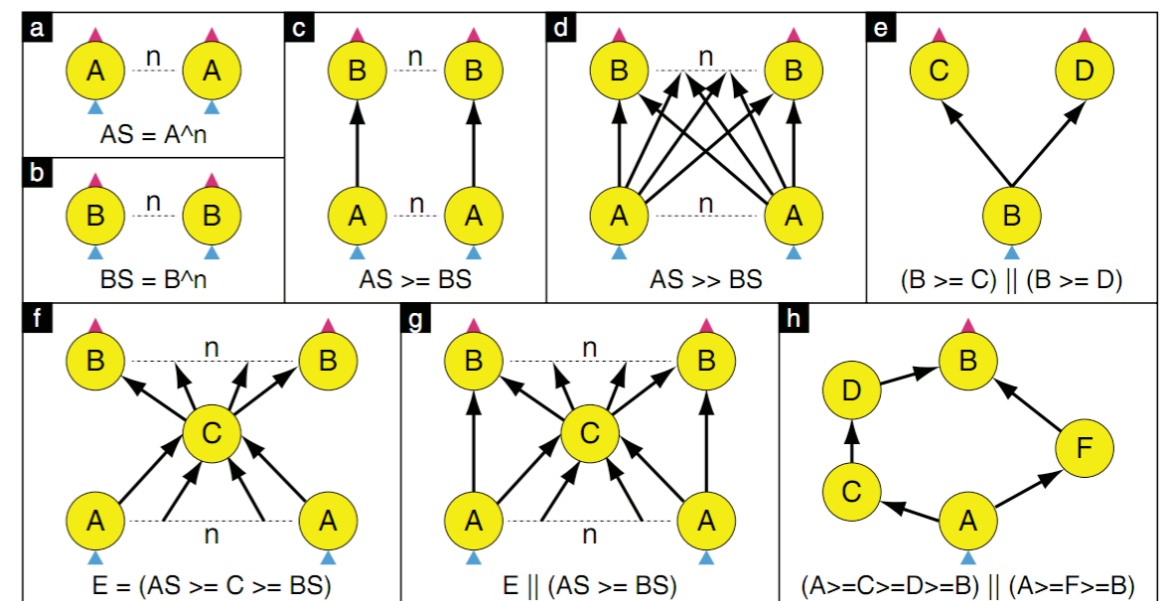
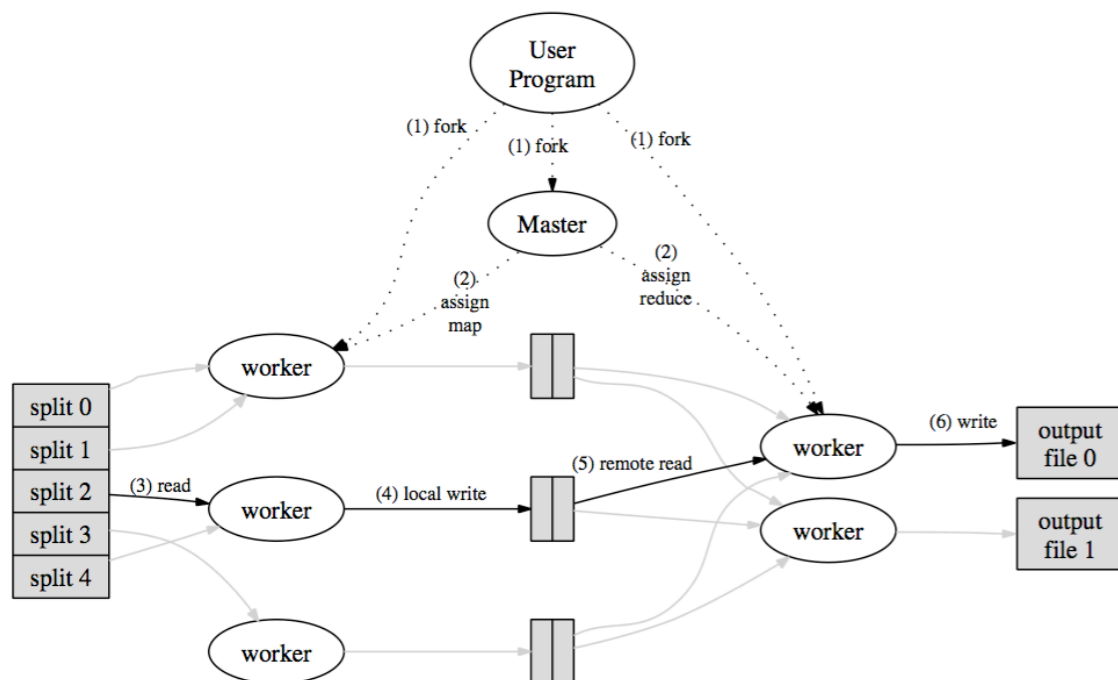
- Cache / synchronize an object
- Uneven load distribution
- Must not generate hotspot
- For given key, pick random order of machines
- Map order onto tree / star via BFS ordering



e.g. memcached

# More stuff

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





MAGIC Etch A Sketch<sup>®</sup> SCREEN

Interaction  
with the environment



Horizontal  
Lid

OHIO ART "A World of Toys"

MAGIC SCREEN IS GLASS SET IN STURDY PLASTIC FRAME  
USE WITH CARE

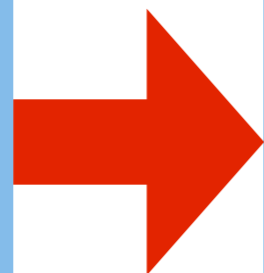
Vertical  
Lid



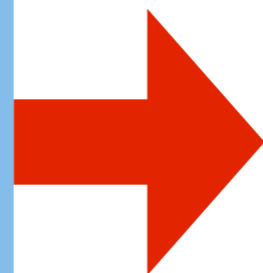
# Batch

- 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

data source



inference  
& learning



deployment

# 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





MAGIC Etch A Sketch<sup>®</sup> SCREEN

- Problems in machine learning
- Systems to run the algorithms
- Response batch/online/interactive
- Compression

Horizontal  
Lid

OHIO ART "A World of Toys"

MAGIC SCREEN IS GLASS SET IN STURDY PLASTIC FRAME  
USE WITH CARE

Vertical  
Lid





MAGIC Etch A Sketch<sup>®</sup> SCREEN

Compression  
hashing for limited memory

4

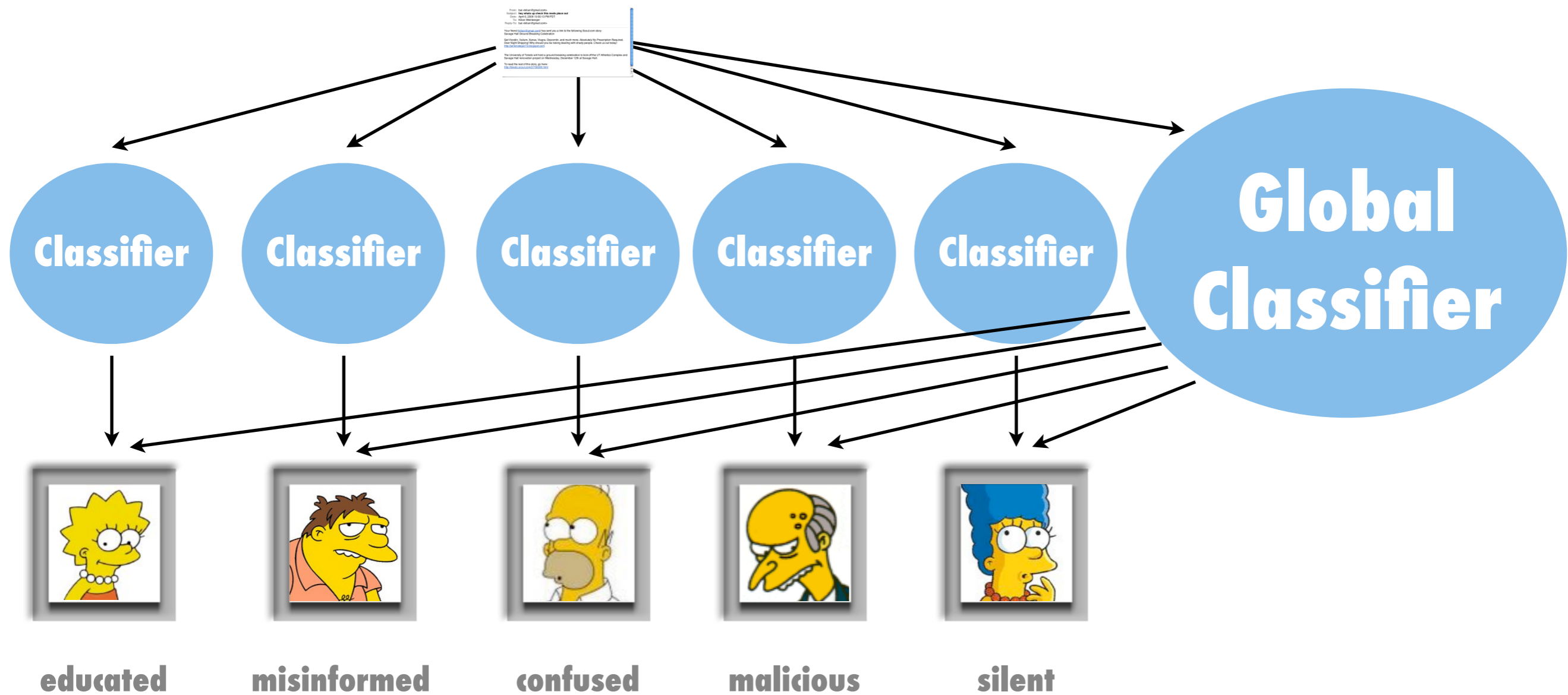
Horizontal  
Lid

OHIO ART "A World of Toys"

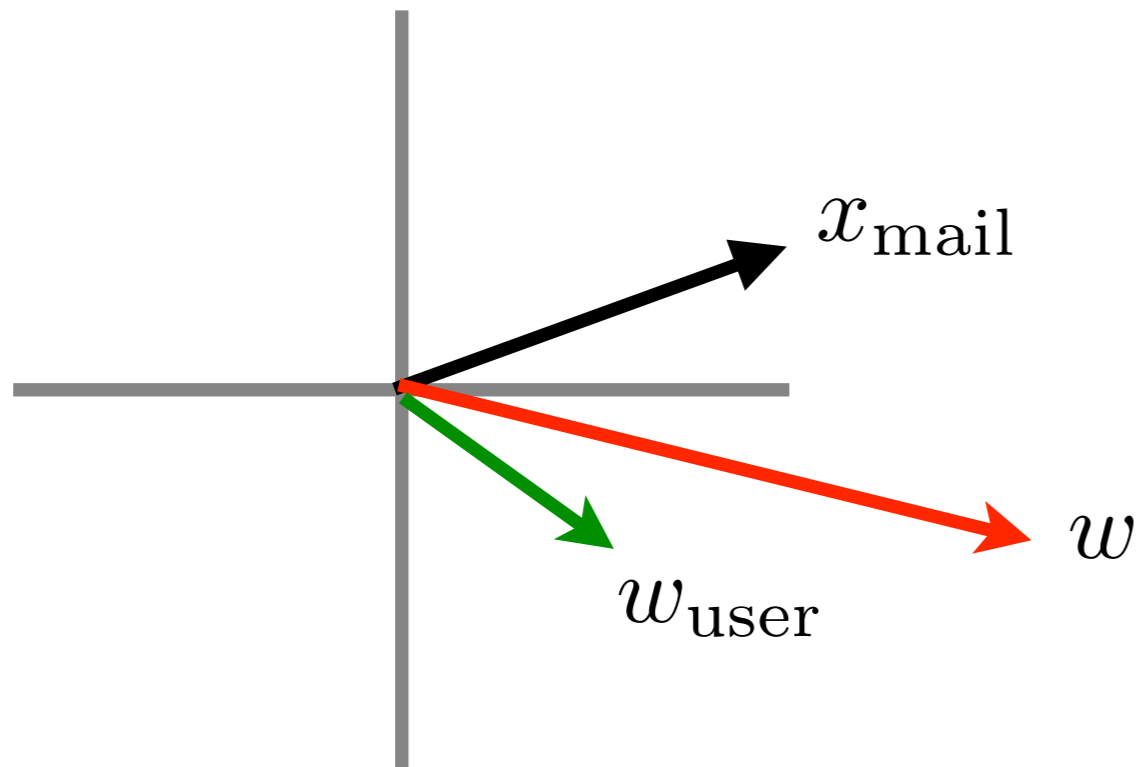
MAGIC SCREEN IS GLASS SET IN STURDY PLASTIC FRAME  
USE WITH CARE

Vertical  
Lid

# Personalized Spam Classification



# Personalized Spam Classification



- **Primal representation**

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

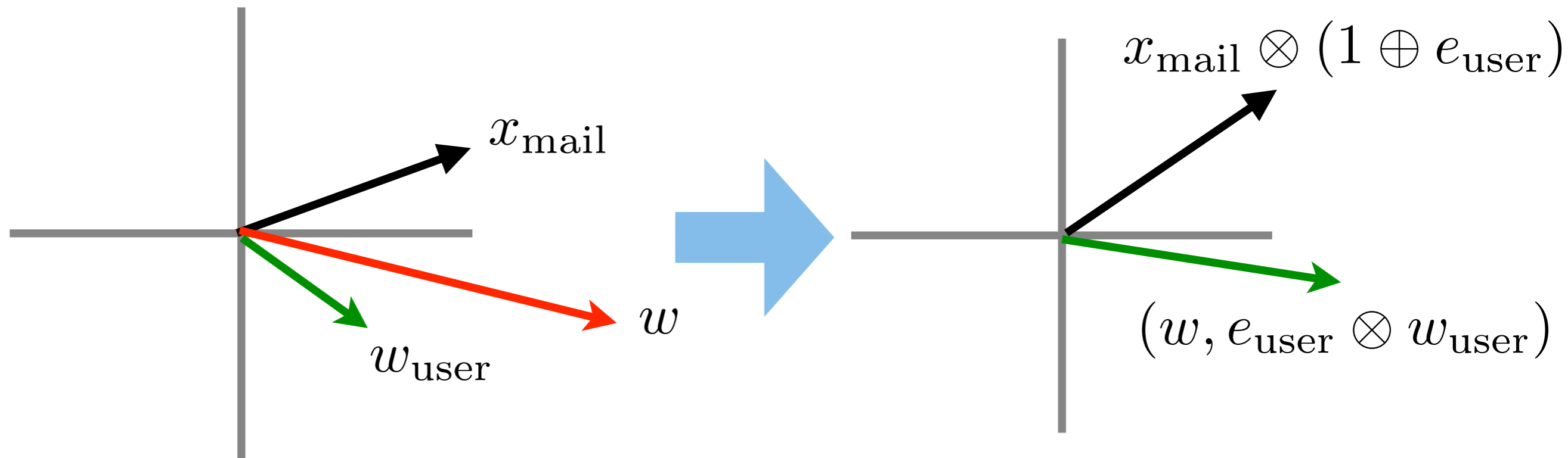
## Kernel representation

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

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,  
Someone important

dictionary:

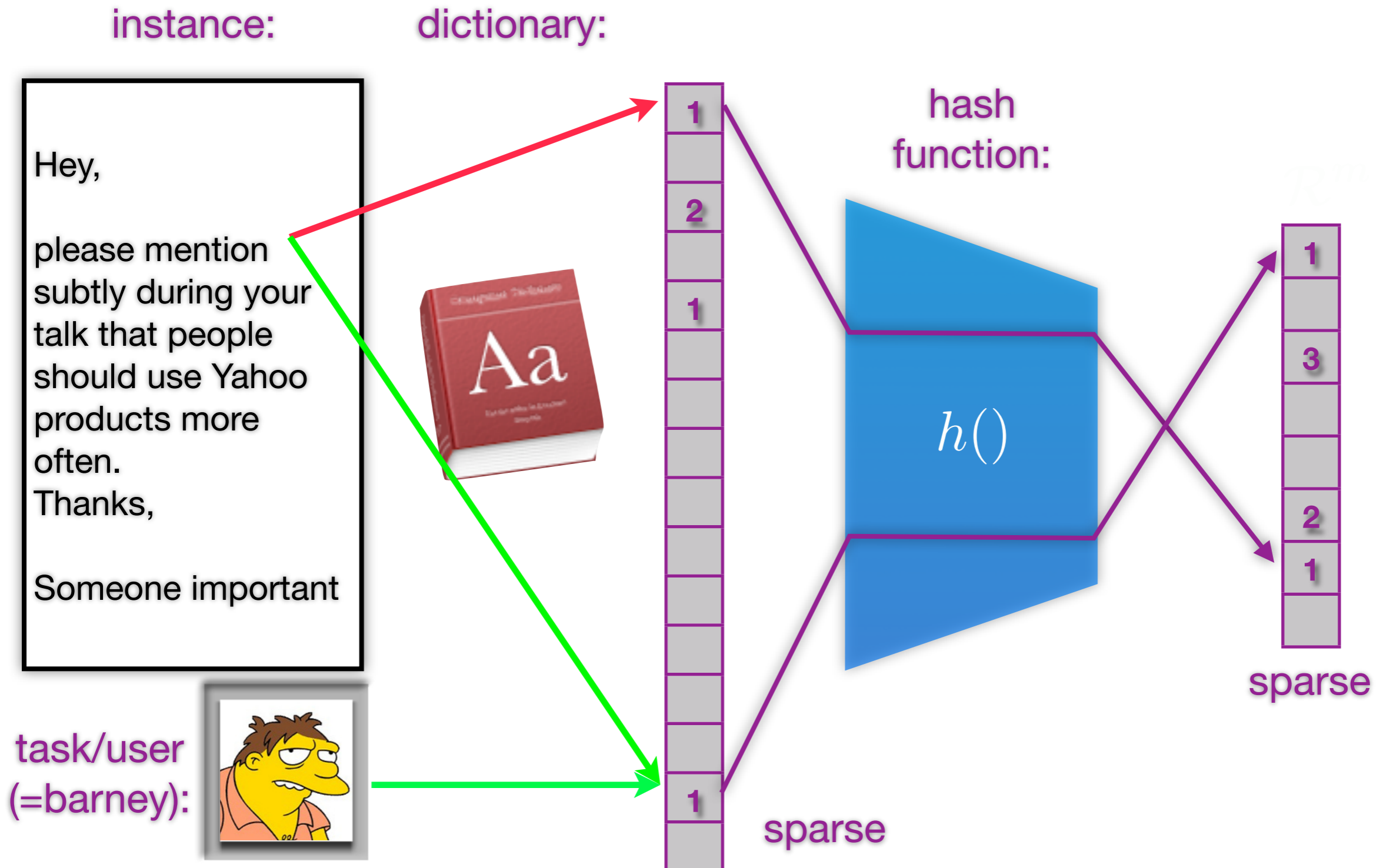


task/user  
(=barney):



sparse

# Hash Kernels



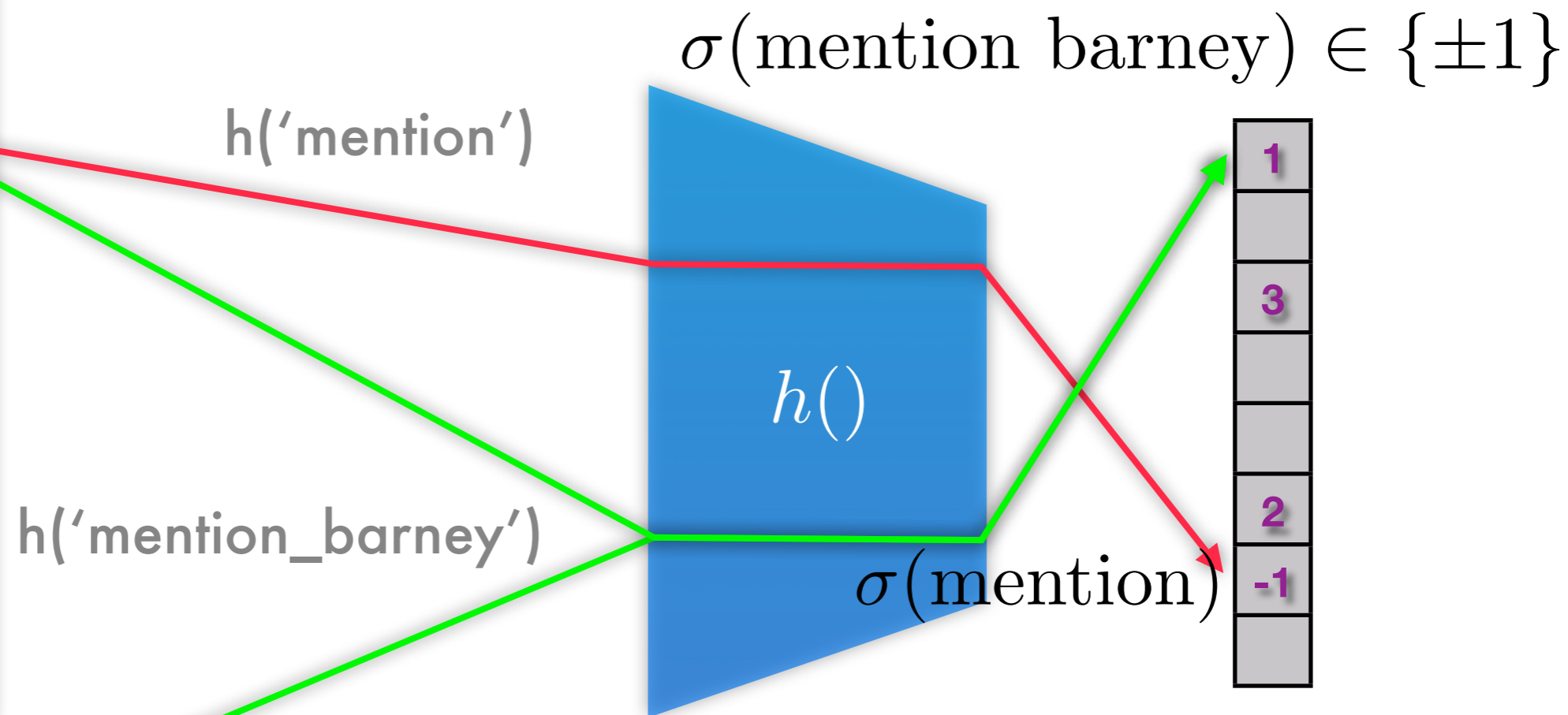
# Hash Kernels

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**

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

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

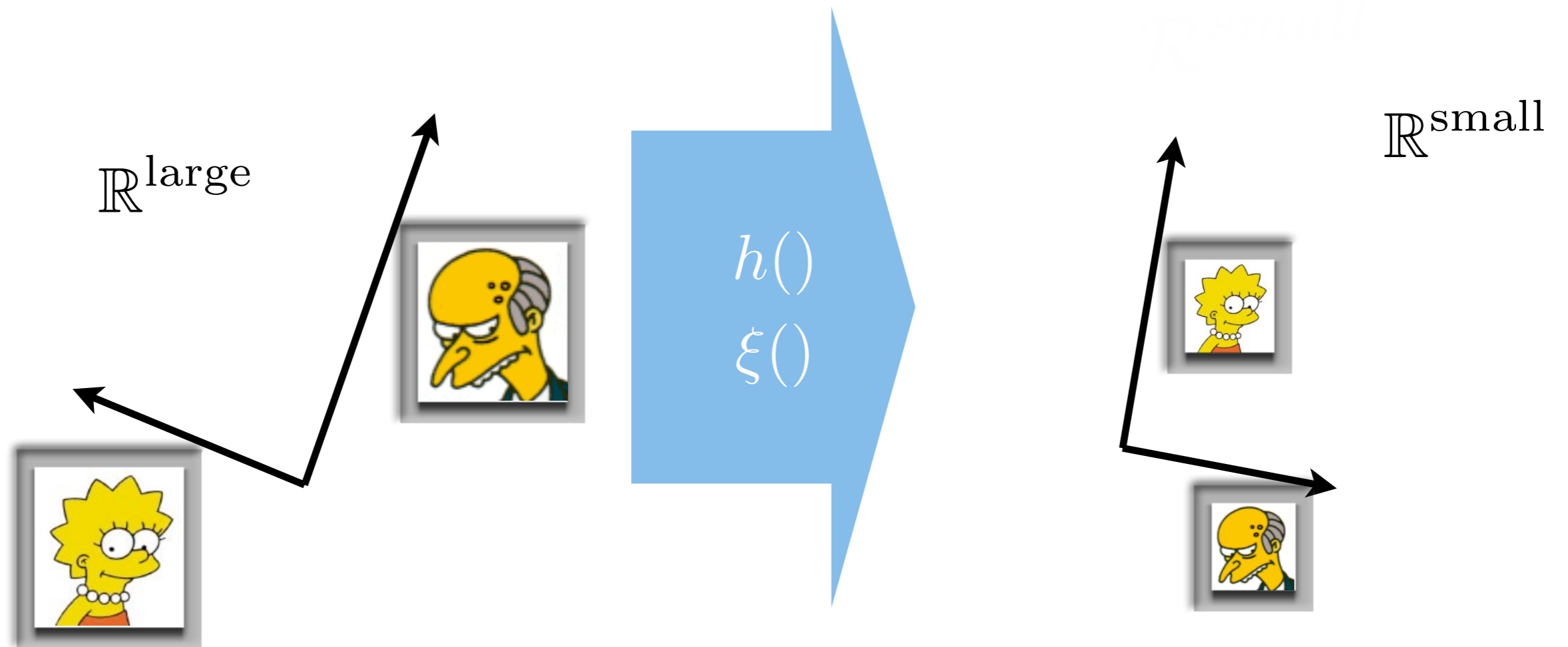
- **Kernel**

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

collisions

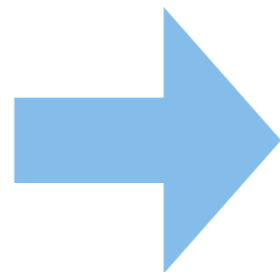
$$k_{\text{hash}}(x, x') = \sum_{j=1}^n \left[ \sum_{i:h(i)=j} x_i \sigma(i) \right] \left[ \sum_{i:h(i)=j} x'_i \sigma(i) \right]$$

# Approximate Orthogonality



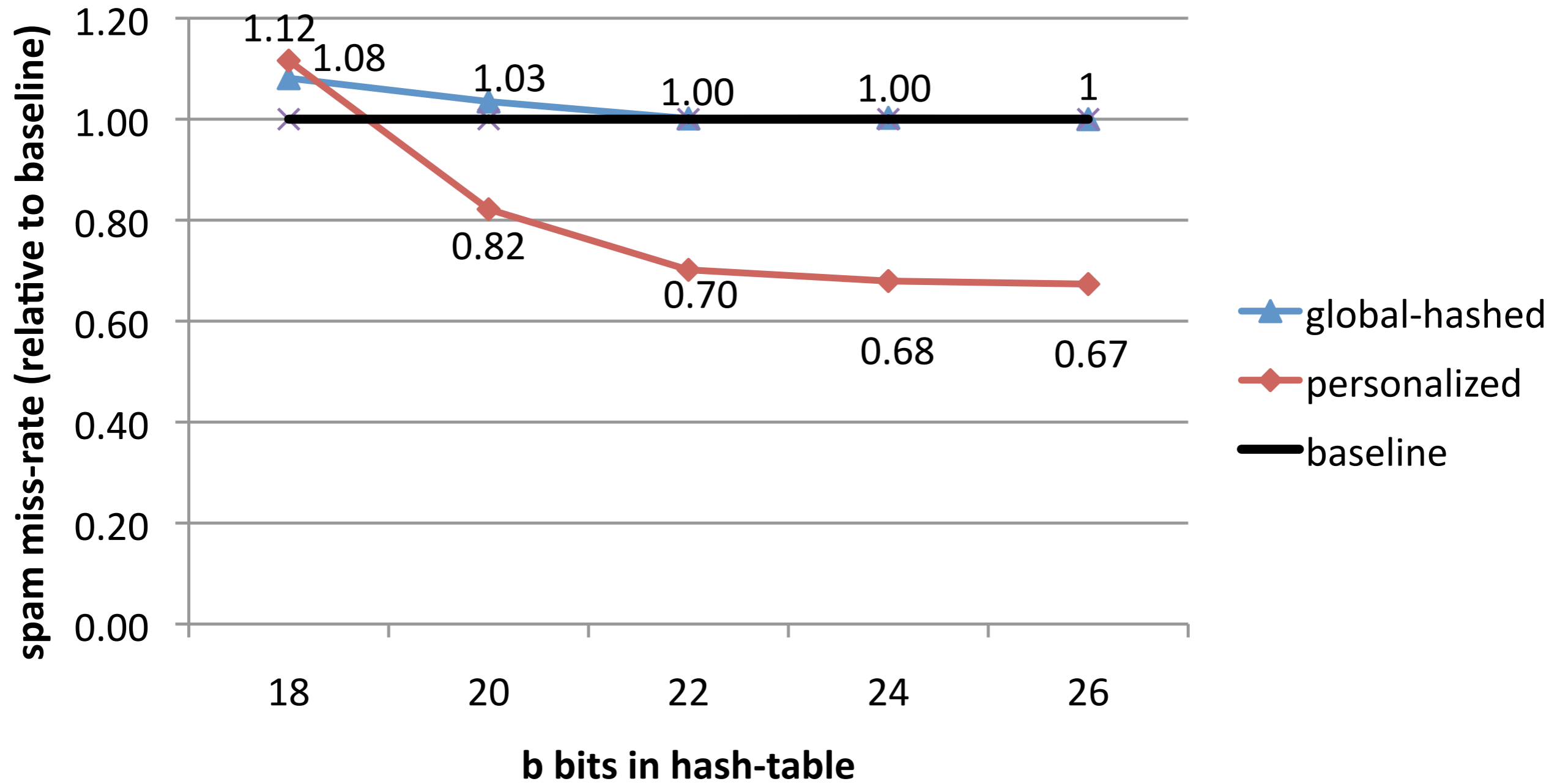
We can do multi-task learning!

**Direct sum** in  
Hilbert Space



**Sum** in  
Hash Space

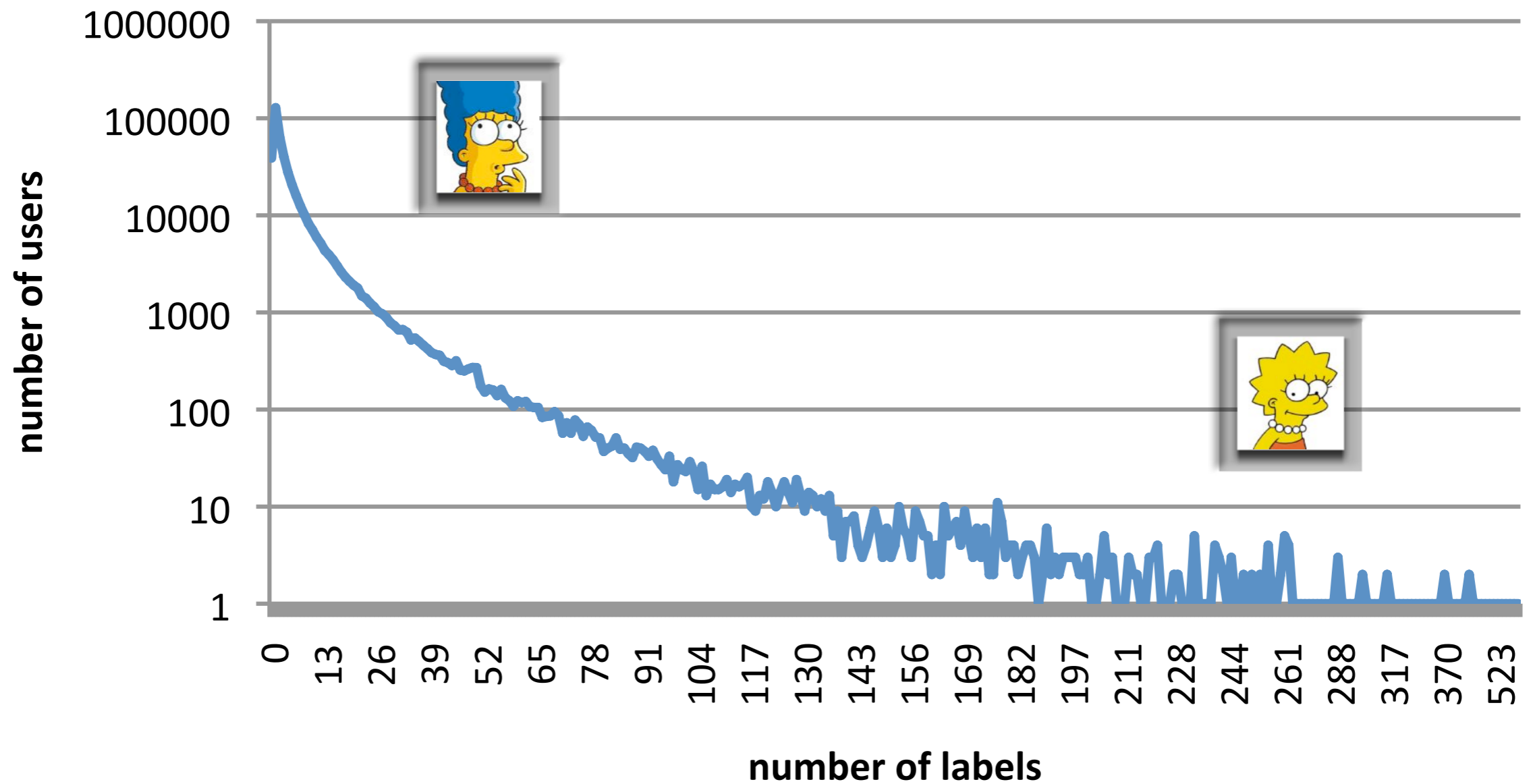
# Spam classification results



$N=20M, U=400K$

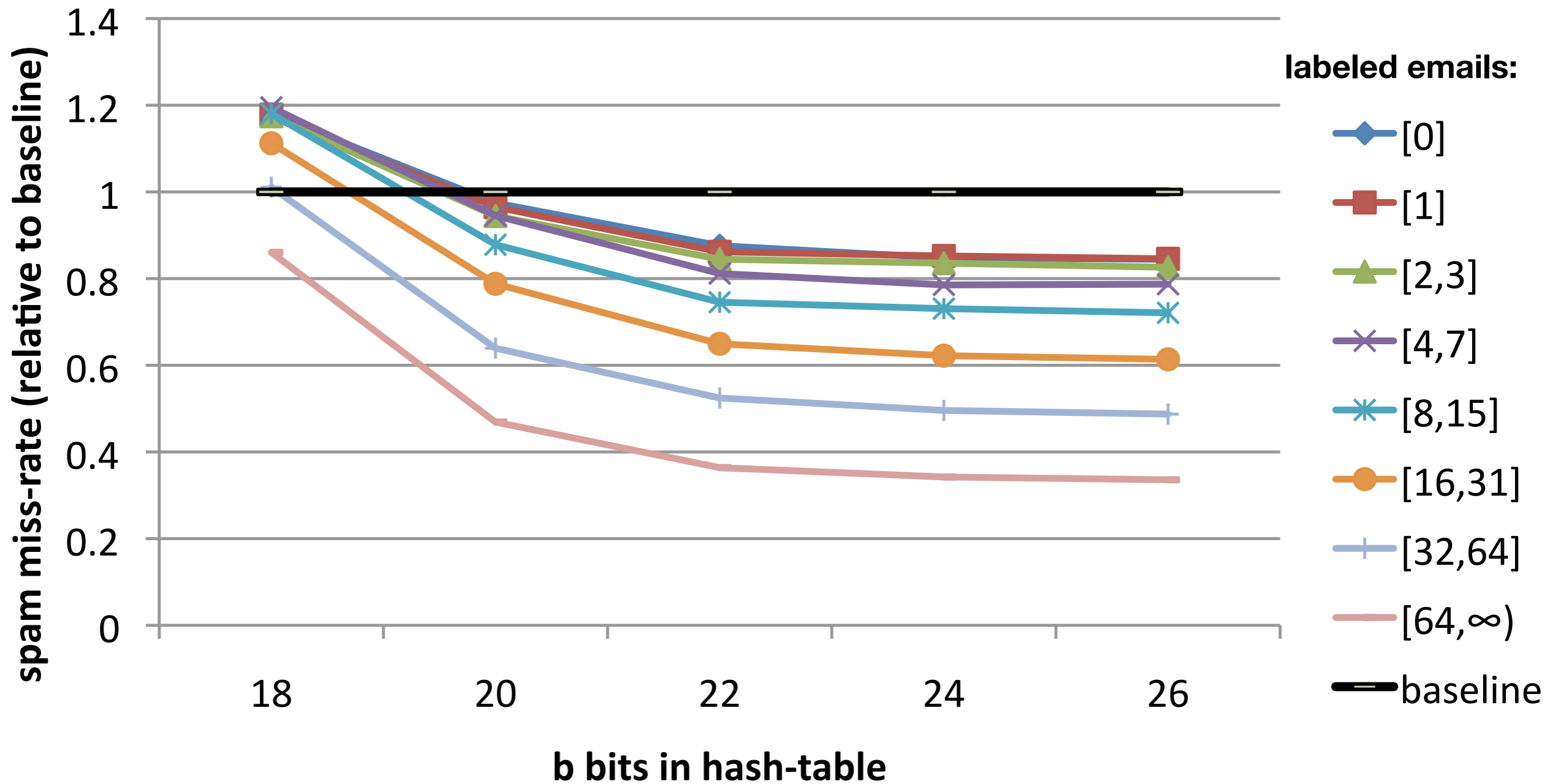
# Lazy users ...

## Labeled emails per user

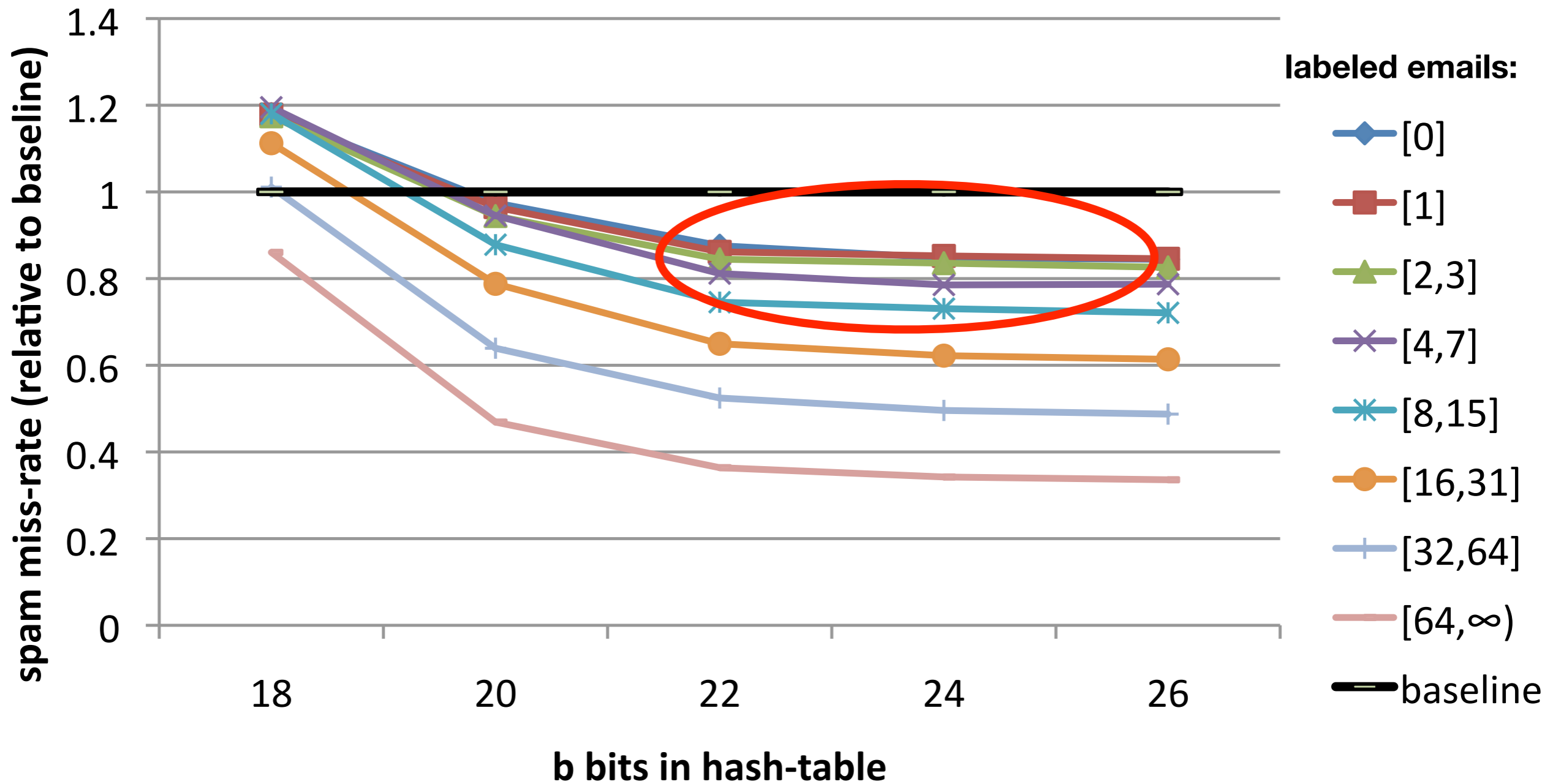


# Results by user group

# 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  $\{(\text{dog}, 1), (*\text{og}, 0.5), (\text{d}^*\text{g}, 0.5), (\text{do}^*, 0.5)\}$
- **Replace  $w[\text{complicated key}]$  by  $w[h(\text{complicated key})]$**