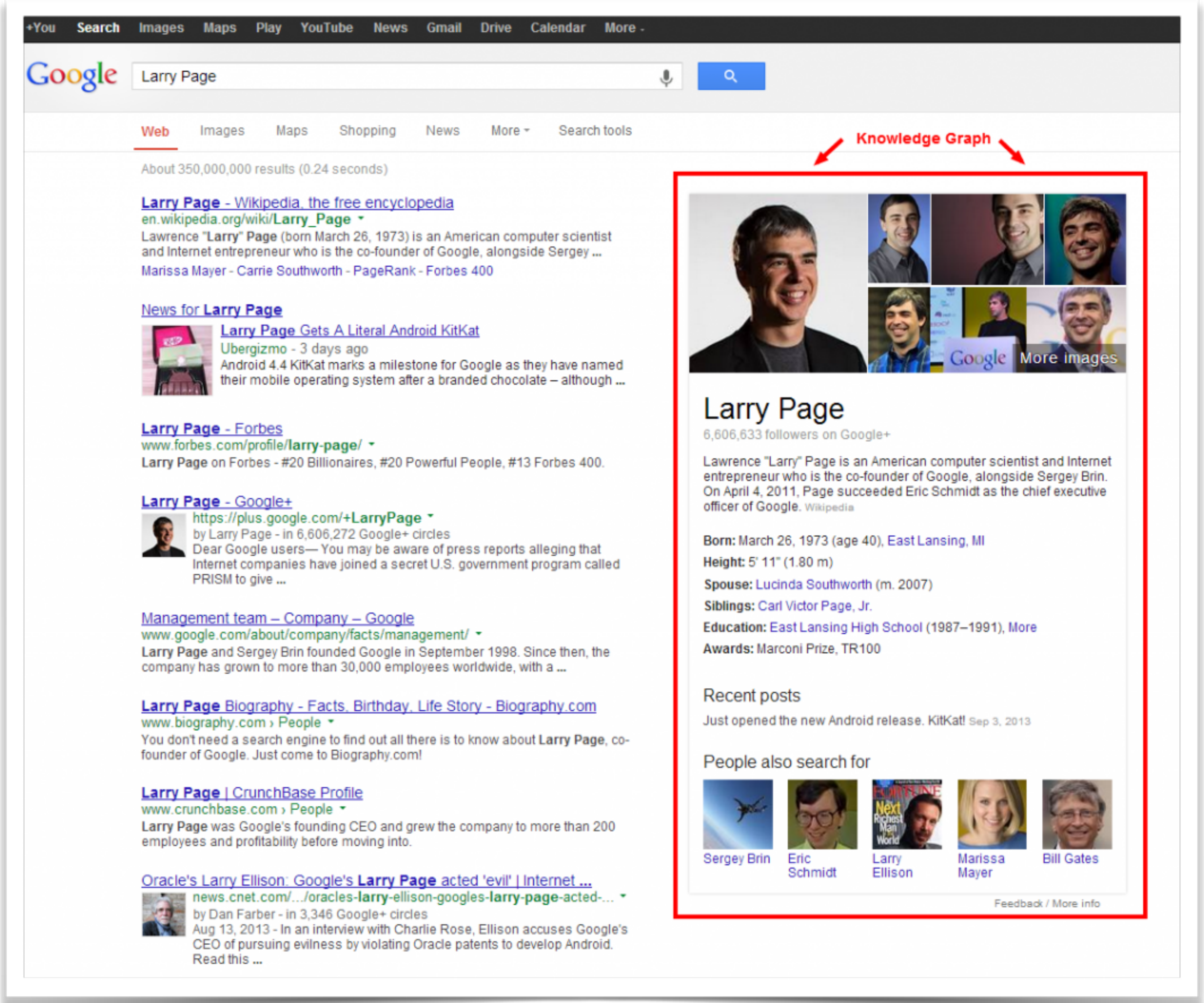
Optimizing Sampling-based Entity Resolution over Streaming Documents

Christan Grant and Daisy Zhe Wang
University of Florida

SIAM BSA Workshop 2015

Knowledge Bases are important structure for organizing and categorizing information.

Knowledge Bases are important structure for organizing and categorizing information.



Larry Page

About 350,000,000 results (0.24 seconds)

Larry Page - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/**Larry_Page**
Lawrence "**Larry**" **Page** (born March 26, 1973) is an American computer scientist and Internet entrepreneur who is the co-founder of Google, alongside Sergey ...
Marissa Mayer - Carrie Southworth - PageRank - Forbes 400

News for **Larry Page**
Larry Page Gets A Literal Android KitKat
Ubergizmo - 3 days ago
Android 4.4 KitKat marks a milestone for Google as they have named their mobile operating system after a branded chocolate – although ...

Larry Page - Forbes
www.forbes.com/profile/**larry-page**/
Larry Page on Forbes - #20 Billionaires, #20 Powerful People, #13 Forbes 400.

Larry Page - Google+
https://plus.google.com/+**LarryPage**
by Larry Page - in 6,606,272 Google+ circles
Dear Google users— You may be aware of press reports alleging that Internet companies have joined a secret U.S. government program called PRISM to give ...

Management team – Company – Google
www.google.com/about/company/facts/management/
Larry Page and Sergey Brin founded Google in September 1998. Since then, the company has grown to more than 30,000 employees worldwide, with a ...

Larry Page Biography - Facts, Birthday, Life Story - Biography.com
www.biography.com › People
You don't need a search engine to find out all there is to know about **Larry Page**, co-founder of Google. Just come to Biography.com!

Larry Page | CrunchBase Profile
www.crunchbase.com › People
Larry Page was Google's founding CEO and grew the company to more than 200 employees and profitability before moving into.

Oracle's Larry Ellison: Google's **Larry Page** acted 'evil' | Internet ...
news.cnet.com/.../oracles-**larry**-ellison-googles-**larry-page**-acted-...
by Dan Farber - in 3,346 Google+ circles
Aug 13, 2013 - In an interview with Charlie Rose, Ellison accuses Google's CEO of pursuing evilness by violating Oracle patents to develop Android. Read this ...

Knowledge Graph

Larry Page
6,606,633 followers on Google+

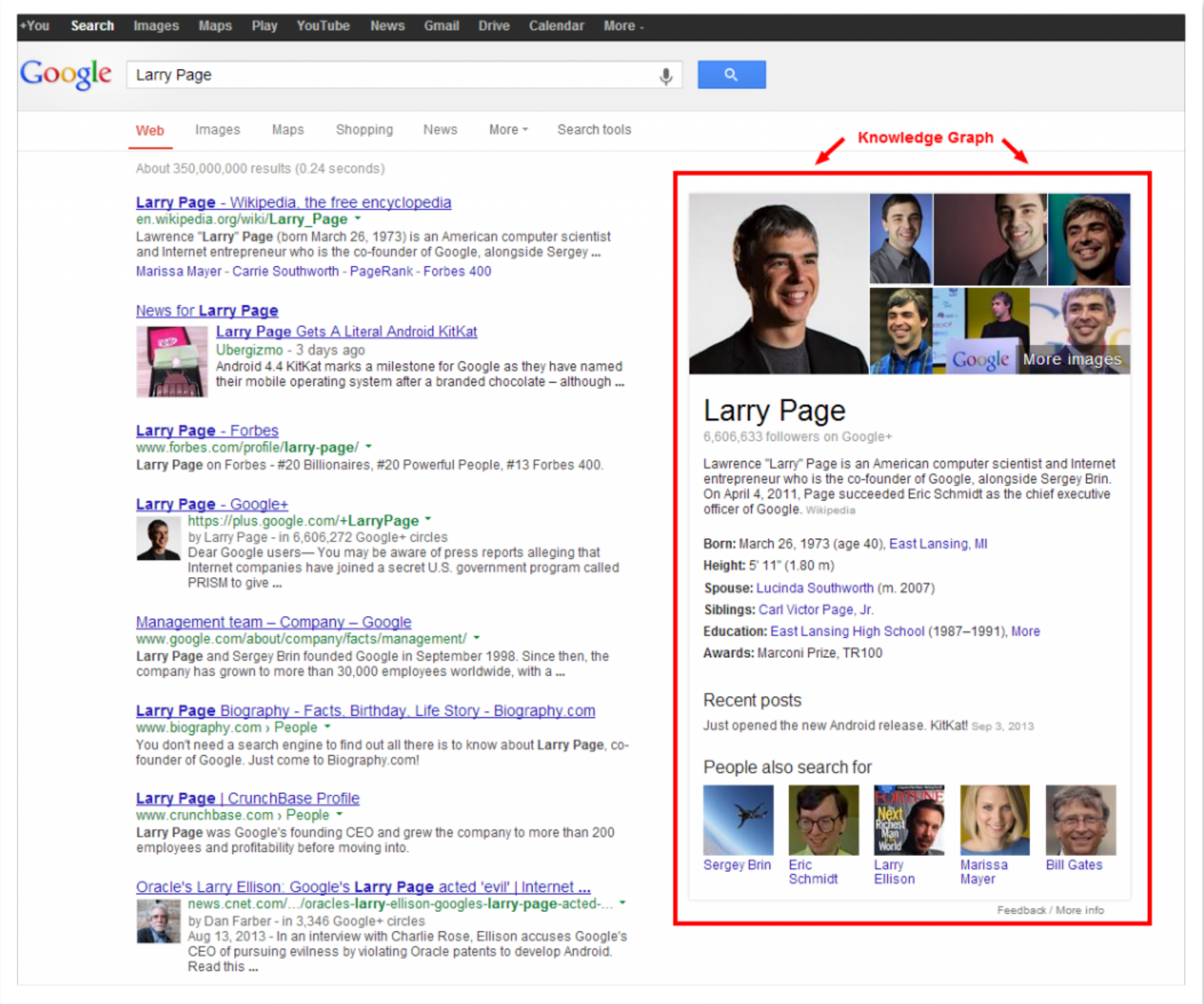
Lawrence "Larry" Page is an American computer scientist and Internet entrepreneur who is the co-founder of Google, alongside Sergey Brin. On April 4, 2011, Page succeeded Eric Schmidt as the chief executive officer of Google. Wikipedia

Born: March 26, 1973 (age 40), East Lansing, MI
Height: 5' 11" (1.80 m)
Spouse: Lucinda Southworth (m. 2007)
Siblings: Carl Victor Page, Jr.
Education: East Lansing High School (1987–1991), More
Awards: Marconi Prize, TR100

Recent posts
Just opened the new Android release. KitKat! Sep 3, 2013

People also search for
Sergey Brin
Eric Schmidt
Larry Ellison
Marissa Mayer
Bill Gates

Knowledge Bases are important structure for organizing and categorizing information.



About 350,000,000 results (0.24 seconds)

Larry Page - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/**Larry_Page**
Lawrence "**Larry**" **Page** (born March 26, 1973) is an American computer scientist and Internet entrepreneur who is the co-founder of Google, alongside Sergey ...
Marissa Mayer - Carrie Southworth - PageRank - Forbes 400

News for **Larry Page**
Larry Page Gets A Literal Android KitKat
Ubergizmo - 3 days ago
Android 4.4 KitKat marks a milestone for Google as they have named their mobile operating system after a branded chocolate – although ...

Larry Page - Forbes
www.forbes.com/profile/**larry-page**/
Larry Page on Forbes - #20 Billionaires, #20 Powerful People, #13 Forbes 400.

Larry Page - Google+
https://plus.google.com/+**LarryPage**
by Larry Page - in 6,606,272 Google+ circles
Dear Google users— You may be aware of press reports alleging that Internet companies have joined a secret U.S. government program called PRISM to give ...

Management team – Company – Google
www.google.com/about/company/facts/management/
Larry Page and Sergey Brin founded Google in September 1998. Since then, the company has grown to more than 30,000 employees worldwide, with a ...

Larry Page Biography - Facts, Birthday, Life Story - Biography.com
www.biography.com › People
You don't need a search engine to find out all there is to know about **Larry Page**, co-founder of Google. Just come to Biography.com!

Larry Page | CrunchBase Profile
www.crunchbase.com › People
Larry Page was Google's founding CEO and grew the company to more than 200 employees and profitability before moving into.

Oracle's Larry Ellison: Google's **Larry Page** acted 'evil' | Internet ...
news.cnet.com/.../oracles-**larry**-ellison-googles-**larry-page**-acted-...
by Dan Farber - in 3,346 Google+ circles
Aug 13, 2013 - In an interview with Charlie Rose, Ellison accuses Google's CEO of pursuing evilness by violating Oracle patents to develop Android. Read this ...

Larry Page
6,606,633 followers on Google+

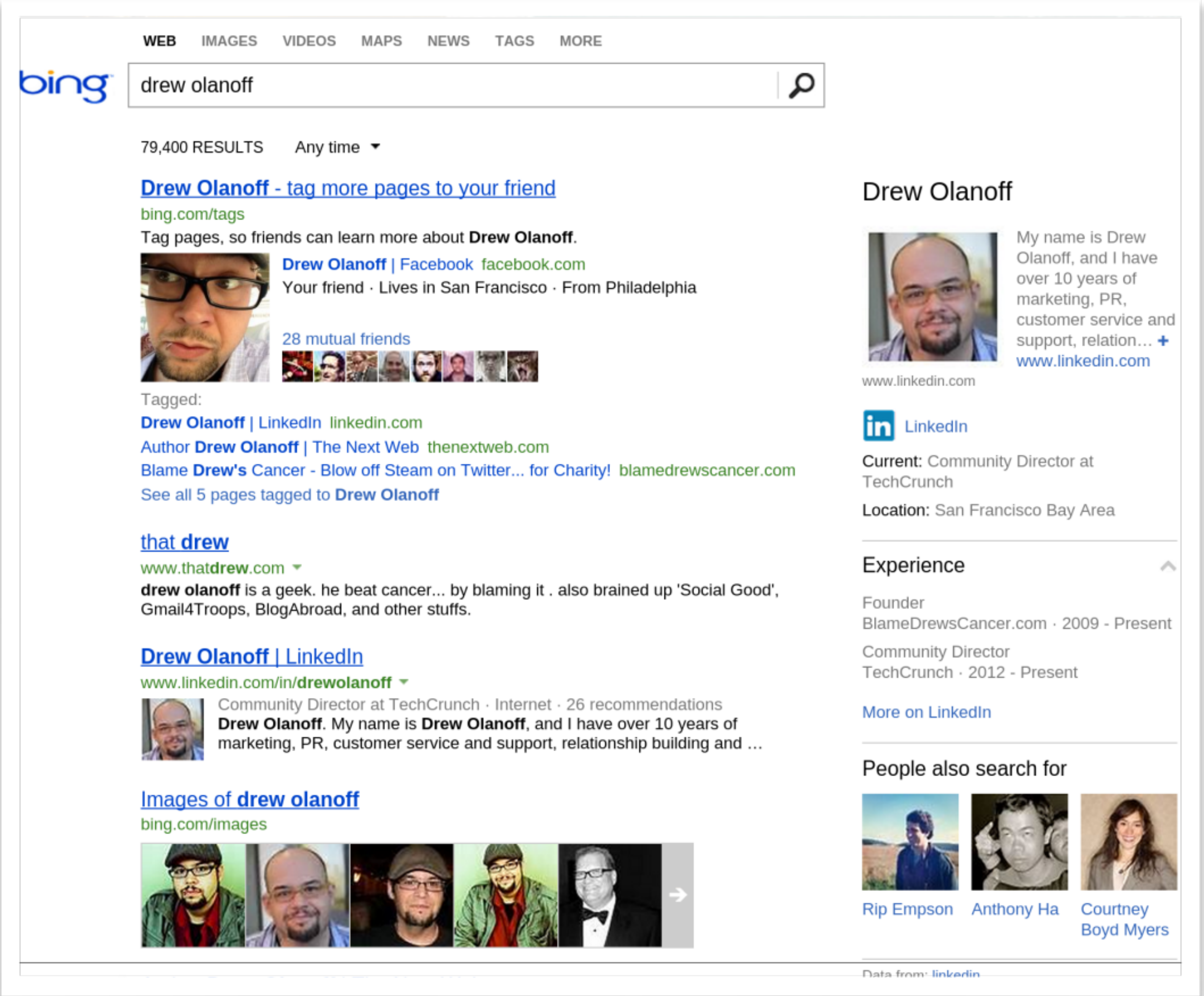
Lawrence "Larry" Page is an American computer scientist and Internet entrepreneur who is the co-founder of Google, alongside Sergey Brin. On April 4, 2011, Page succeeded Eric Schmidt as the chief executive officer of Google. Wikipedia

Born: March 26, 1973 (age 40), East Lansing, MI
Height: 5' 11" (1.80 m)
Spouse: Lucinda Southworth (m. 2007)
Siblings: Carl Victor Page, Jr.
Education: East Lansing High School (1987–1991), More
Awards: Marconi Prize, TR100

Recent posts
Just opened the new Android release. KitKat! Sep 3, 2013

People also search for
Sergey Brin
Eric Schmidt
Larry Ellison
Marissa Mayer
Bill Gates

Feedback / More info



79,400 RESULTS

Drew Olanoff - tag more pages to your friend
bing.com/tags
Tag pages, so friends can learn more about **Drew Olanoff**.
Drew Olanoff | Facebook facebook.com
Your friend · Lives in San Francisco · From Philadelphia
28 mutual friends
Tagged:
Drew Olanoff | LinkedIn linkedin.com
Author **Drew Olanoff** | The Next Web thenextweb.com
Blame **Drew's** Cancer - Blow off Steam on Twitter... for Charity! blamedrewscancer.com
See all 5 pages tagged to **Drew Olanoff**

that **drew**
www.that**drew**.com
drew olanoff is a geek. he beat cancer... by blaming it . also brained up 'Social Good', Gmail4Troops, BlogAbroad, and other stuffs.

Drew Olanoff | LinkedIn
www.linkedin.com/in/**drewolanoff**
Community Director at TechCrunch · Internet · 26 recommendations
Drew Olanoff. My name is **Drew Olanoff**, and I have over 10 years of marketing, PR, customer service and support, relationship building and ...

Images of **drew olanoff**
bing.com/images

Drew Olanoff
My name is Drew Olanoff, and I have over 10 years of marketing, PR, customer service and support, relation... +
www.linkedin.com
www.linkedin.com
LinkedIn
Current: Community Director at TechCrunch
Location: San Francisco Bay Area

Experience
Founder
BlameDrewsCancer.com · 2009 - Present
Community Director
TechCrunch · 2012 - Present
More on LinkedIn

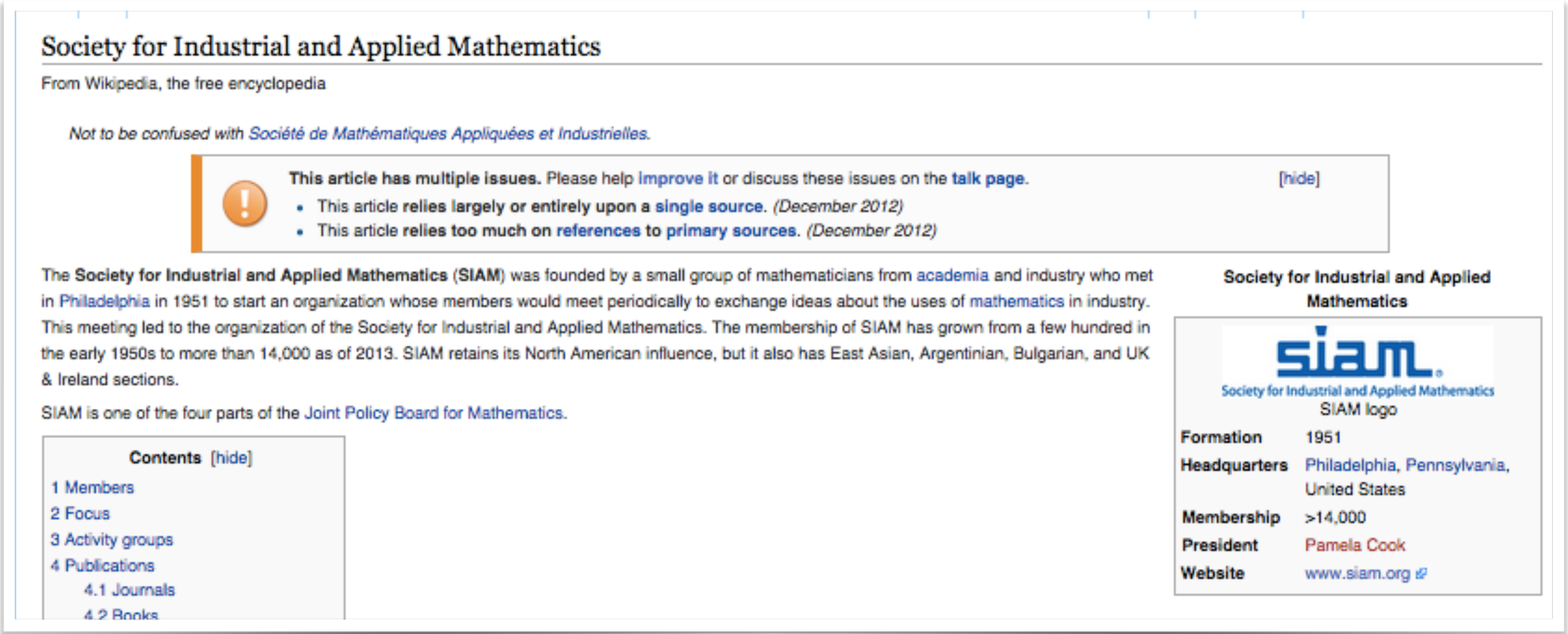
People also search for
Rip Empson
Anthony Ha
Courtney Boyd Myers

Data from: linkedin

- Many of these knowledge bases and new knowledge bases are bootstrapped using **Wikipedia**/Freebase.

- Many of these knowledge bases and new knowledge bases are bootstrapped using **Wikipedia**/Freebase.

- All Wikipedia information is based on facts from (reputable?) web sources.



Society for Industrial and Applied Mathematics

From Wikipedia, the free encyclopedia

Not to be confused with Société de Mathématiques Appliquées et Industrielles.

This article has multiple issues. Please help improve it or discuss these issues on the talk page. [hide]

- This article **relies largely or entirely upon a single source.** *(December 2012)*
- This article **relies too much on references to primary sources.** *(December 2012)*

The **Society for Industrial and Applied Mathematics** (**SIAM**) was founded by a small group of mathematicians from academia and industry who met in Philadelphia in 1951 to start an organization whose members would meet periodically to exchange ideas about the uses of mathematics in industry. This meeting led to the organization of the Society for Industrial and Applied Mathematics. The membership of SIAM has grown from a few hundred in the early 1950s to more than 14,000 as of 2013. SIAM retains its North American influence, but it also has East Asian, Argentinian, Bulgarian, and UK & Ireland sections.

SIAM is one of the four parts of the Joint Policy Board for Mathematics.

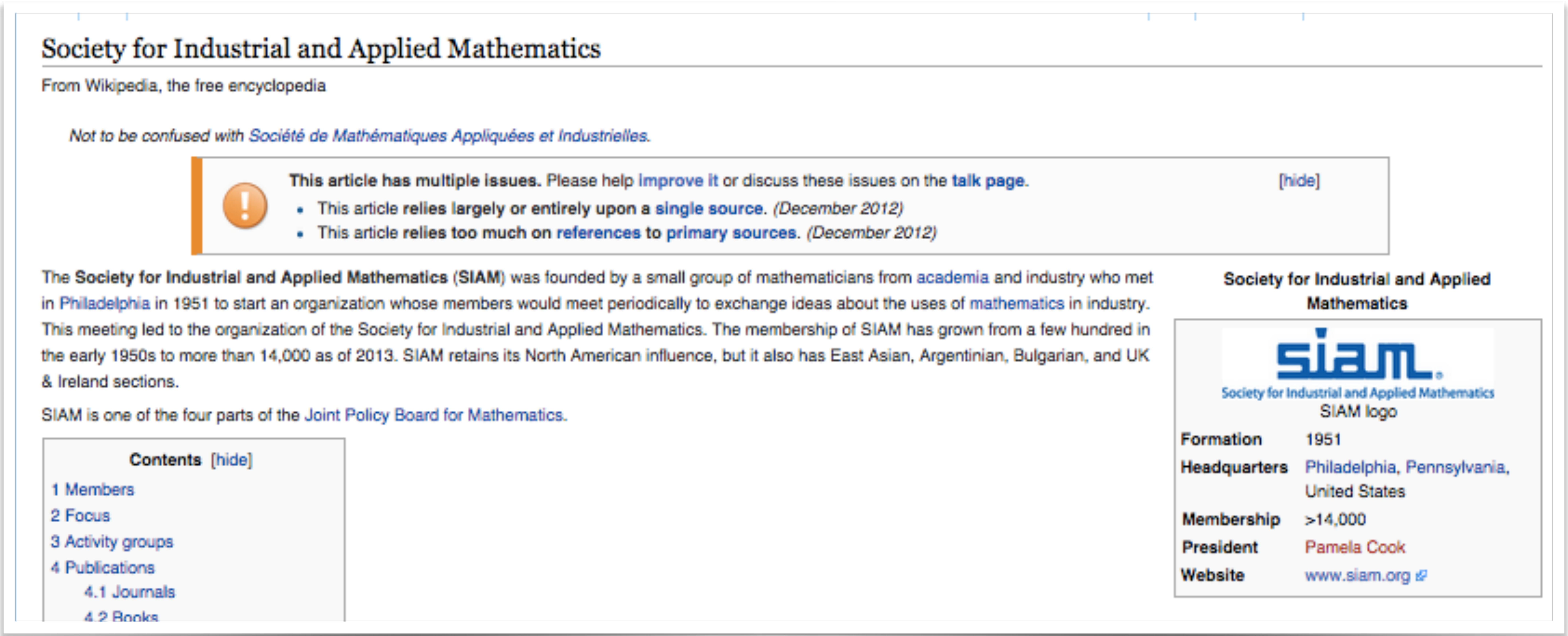
Contents [hide]

Society for Industrial and Applied Mathematics

siam. Society for Industrial and Applied Mathematics

SIAM logo

Formation	1951
Headquarters	Philadelphia, Pennsylvania, United States
Membership	>14,000
President	Pamela Cook
Website	www.siam.org



Society for Industrial and Applied Mathematics

From Wikipedia, the free encyclopedia

Not to be confused with Société de Mathématiques Appliquées et Industrielles.

This article has multiple issues. Please help improve it or discuss these issues on the talk page. [hide]

- This article **relies largely or entirely upon a single source.** *(December 2012)*
- This article **relies too much on references to primary sources.** *(December 2012)*

The **Society for Industrial and Applied Mathematics** (**SIAM**) was founded by a small group of mathematicians from academia and industry who met in Philadelphia in 1951 to start an organization whose members would meet periodically to exchange ideas about the uses of mathematics in industry. This meeting led to the organization of the Society for Industrial and Applied Mathematics. The membership of SIAM has grown from a few hundred in the early 1950s to more than 14,000 as of 2013. SIAM retains its North American influence, but it also has East Asian, Argentinian, Bulgarian, and UK & Ireland sections.

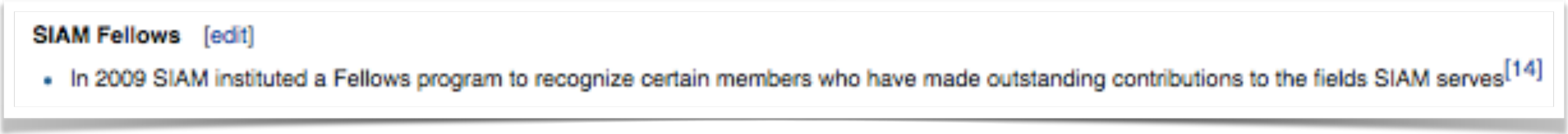
SIAM is one of the four parts of the Joint Policy Board for Mathematics.

Society for Industrial and Applied Mathematics

SIAM logo

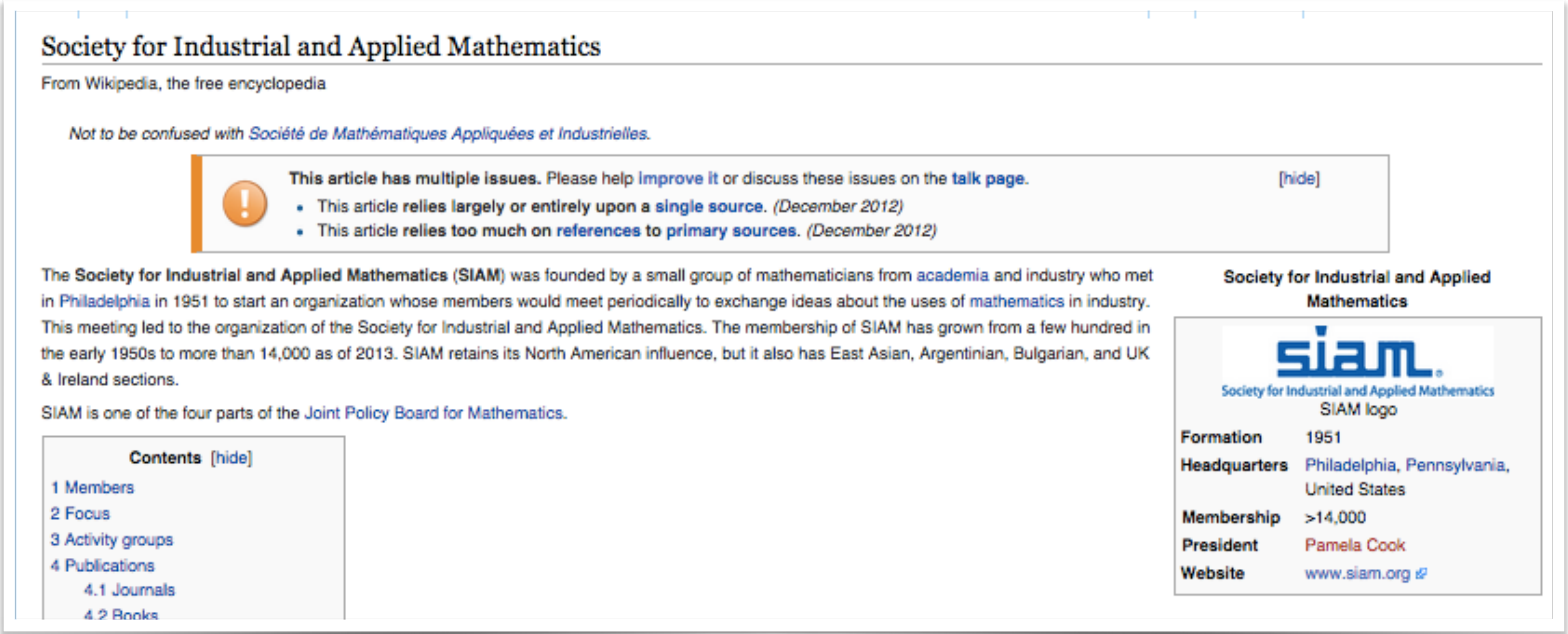
Formation	1951
Headquarters	Philadelphia, Pennsylvania, United States
Membership	>14,000
President	Pamela Cook
Website	www.siam.org

Contents [hide]



SIAM Fellows [edit]

- In 2009 SIAM instituted a Fellows program to recognize certain members who have made outstanding contributions to the fields SIAM serves[14]



Society for Industrial and Applied Mathematics

From Wikipedia, the free encyclopedia

Not to be confused with Société de Mathématiques Appliquées et Industrielles.

This article has multiple issues. Please help improve it or discuss these issues on the talk page. [hide]

- This article **relies largely or entirely upon a single source.** *(December 2012)*
- This article **relies too much on references to primary sources.** *(December 2012)*

The **Society for Industrial and Applied Mathematics** (**SIAM**) was founded by a small group of mathematicians from academia and industry who met in Philadelphia in 1951 to start an organization whose members would meet periodically to exchange ideas about the uses of mathematics in industry. This meeting led to the organization of the Society for Industrial and Applied Mathematics. The membership of SIAM has grown from a few hundred in the early 1950s to more than 14,000 as of 2013. SIAM retains its North American influence, but it also has East Asian, Argentinian, Bulgarian, and UK & Ireland sections.

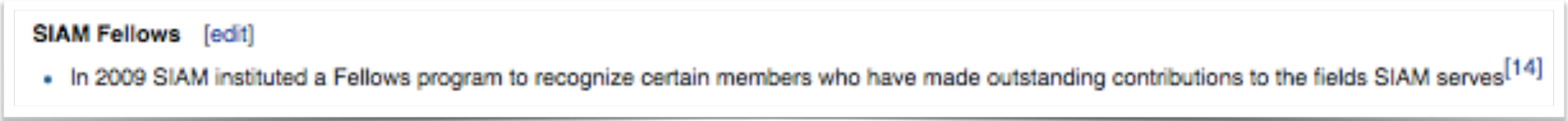
SIAM is one of the four parts of the Joint Policy Board for Mathematics.

Society for Industrial and Applied Mathematics

siam. Society for Industrial and Applied Mathematics

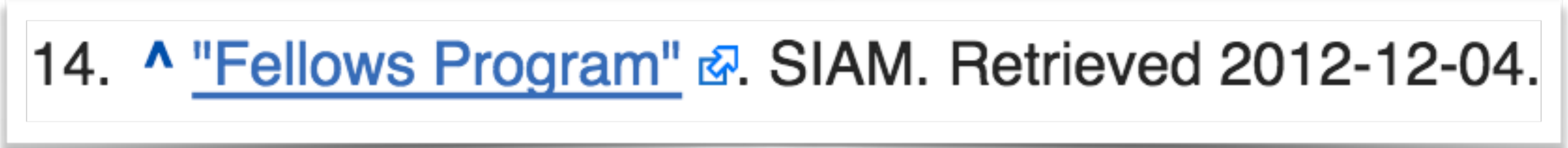
SIAM logo

Formation	1951
Headquarters	Philadelphia, Pennsylvania, United States
Membership	>14,000
President	Pamela Cook
Website	www.siam.org

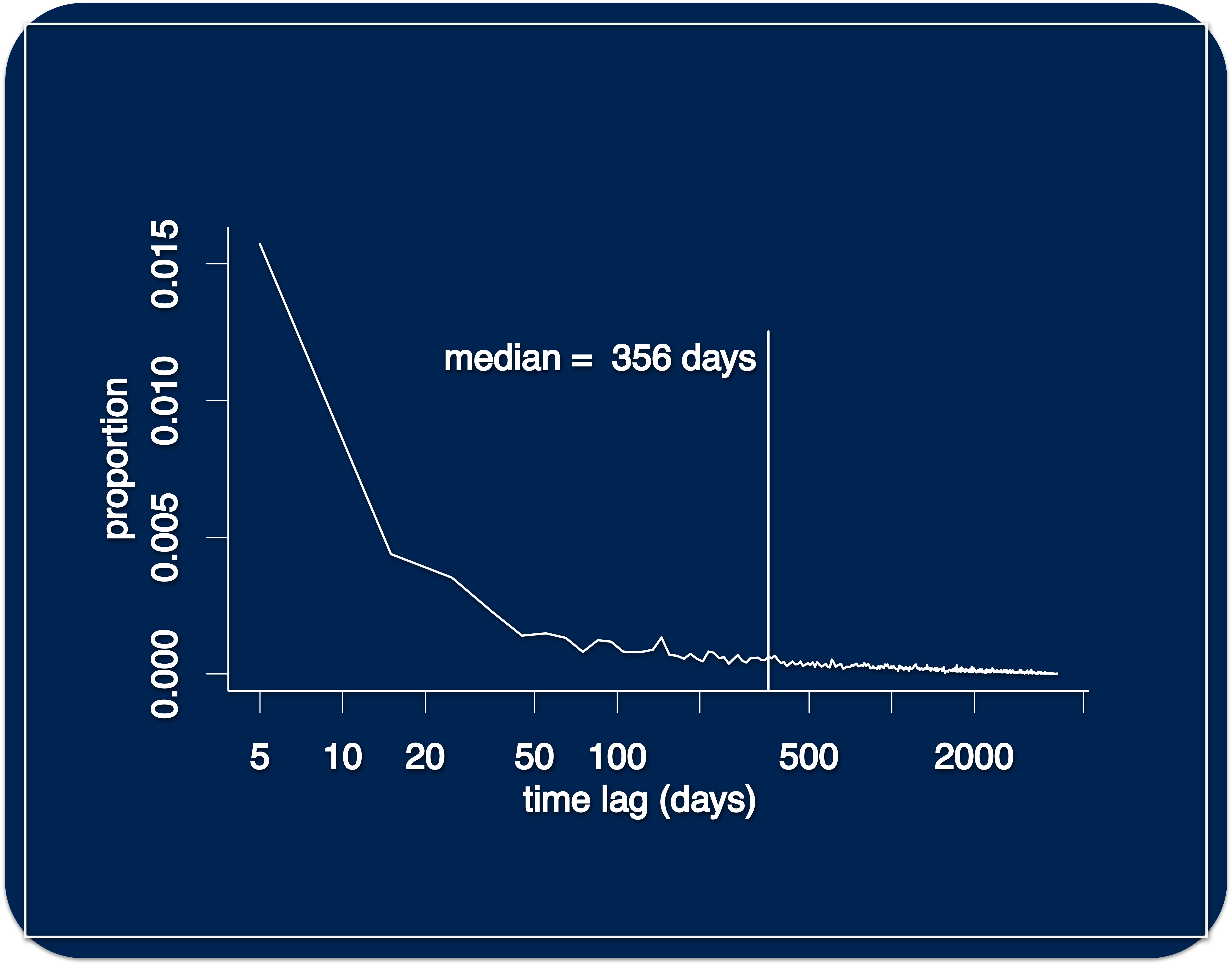


SIAM Fellows [edit]

- In 2009 SIAM instituted a Fellows program to recognize certain members who have made outstanding contributions to the fields SIAM serves[14]



14. ^ "Fellows Program". SIAM. Retrieved 2012-12-04.



The average time between an event and its appearance on Wikipedia is **356** days.

J. Frank et al. 2012

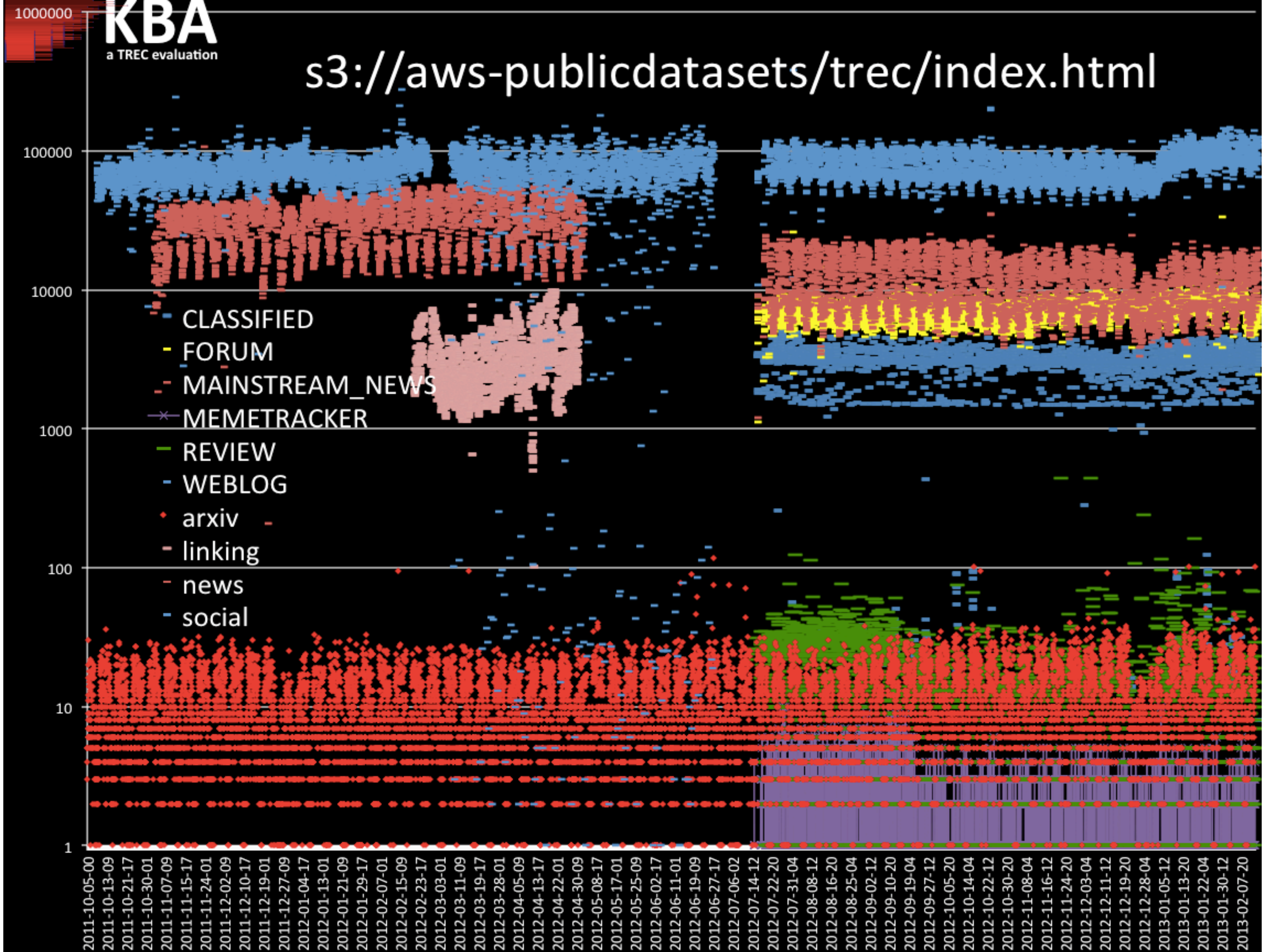
Knowledge Base Acceleration

Knowledge Base Acceleration

NIST TREC created a track that reads in streaming documents and a set of entities and suggests citations for wikipedia entities.

KBA
a TREC evaluation

s3://aws-publicdatasets/trec/index.html



Knowledge Base Acceleration

NIST TREC created a track that reads in streaming documents and a set of entities and suggests citations for wikipedia entities.

Knowledge Base Acceleration

NIST TREC created a track that reads in streaming documents and a set of entities and suggests citations for wikipedia entities.

Challenges:

Knowledge Base Acceleration

NIST TREC created a track that reads in streaming documents and a set of entities and suggests citations for wikipedia entities.

Challenges:

1) A large amount of documents

Knowledge Base Acceleration

NIST TREC created a track that reads in streaming documents and a set of entities and suggests citations for wikipedia entities.

Challenges:

1) A large amount of documents

2) Ambiguous text

Knowledge Base Acceleration

NIST TREC created a track that reads in streaming documents and a set of entities and suggests citations for wikipedia entities.

Challenges:

1) A large amount of documents

2) Ambiguous text

3) Ambiguous Entities

Knowledge Base Acceleration

NIST TREC created a track that reads in streaming documents and a set of entities and suggests citations for wikipedia entities.

Challenges:

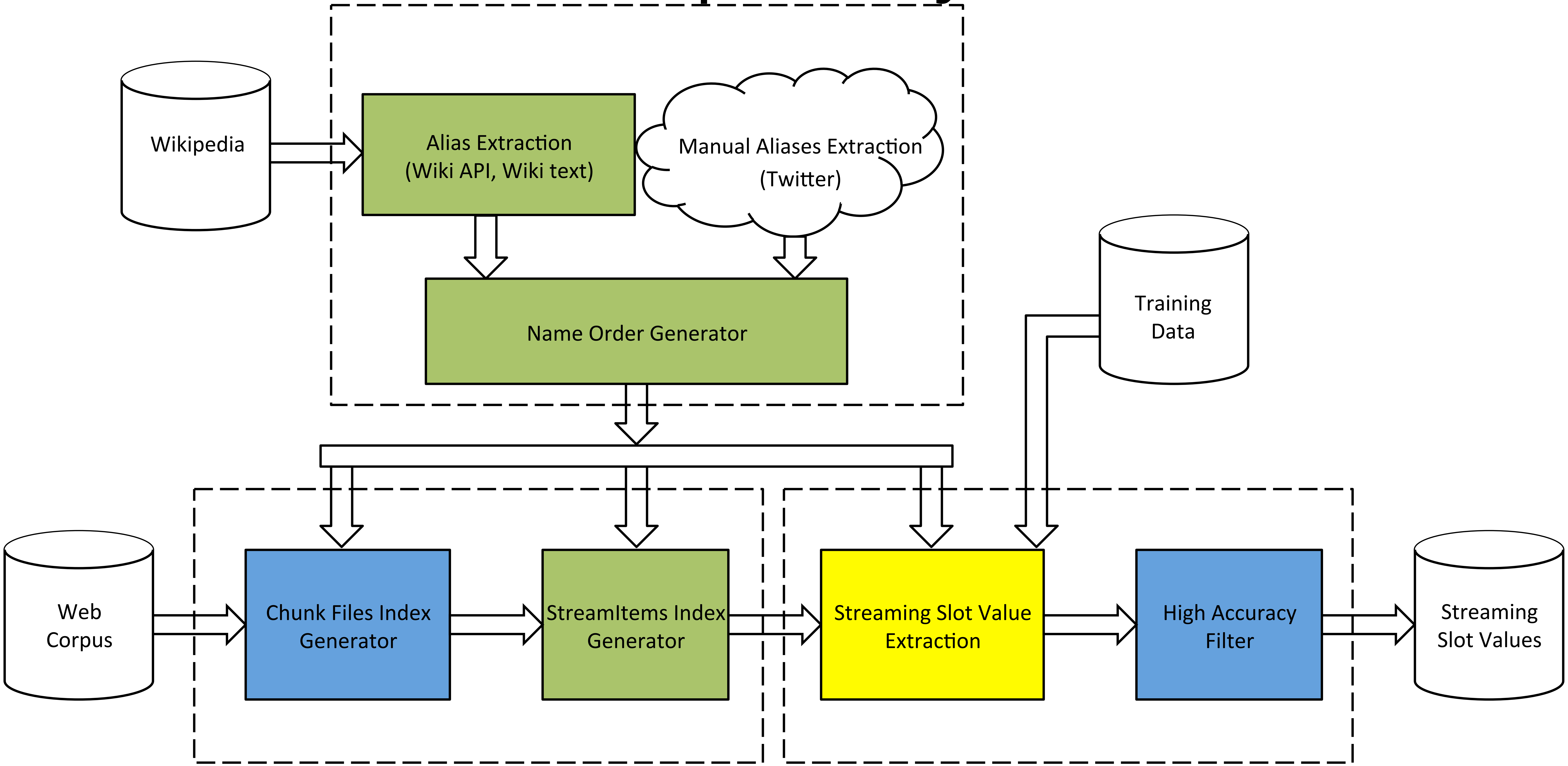
1) A large amount of documents

2) Ambiguous text

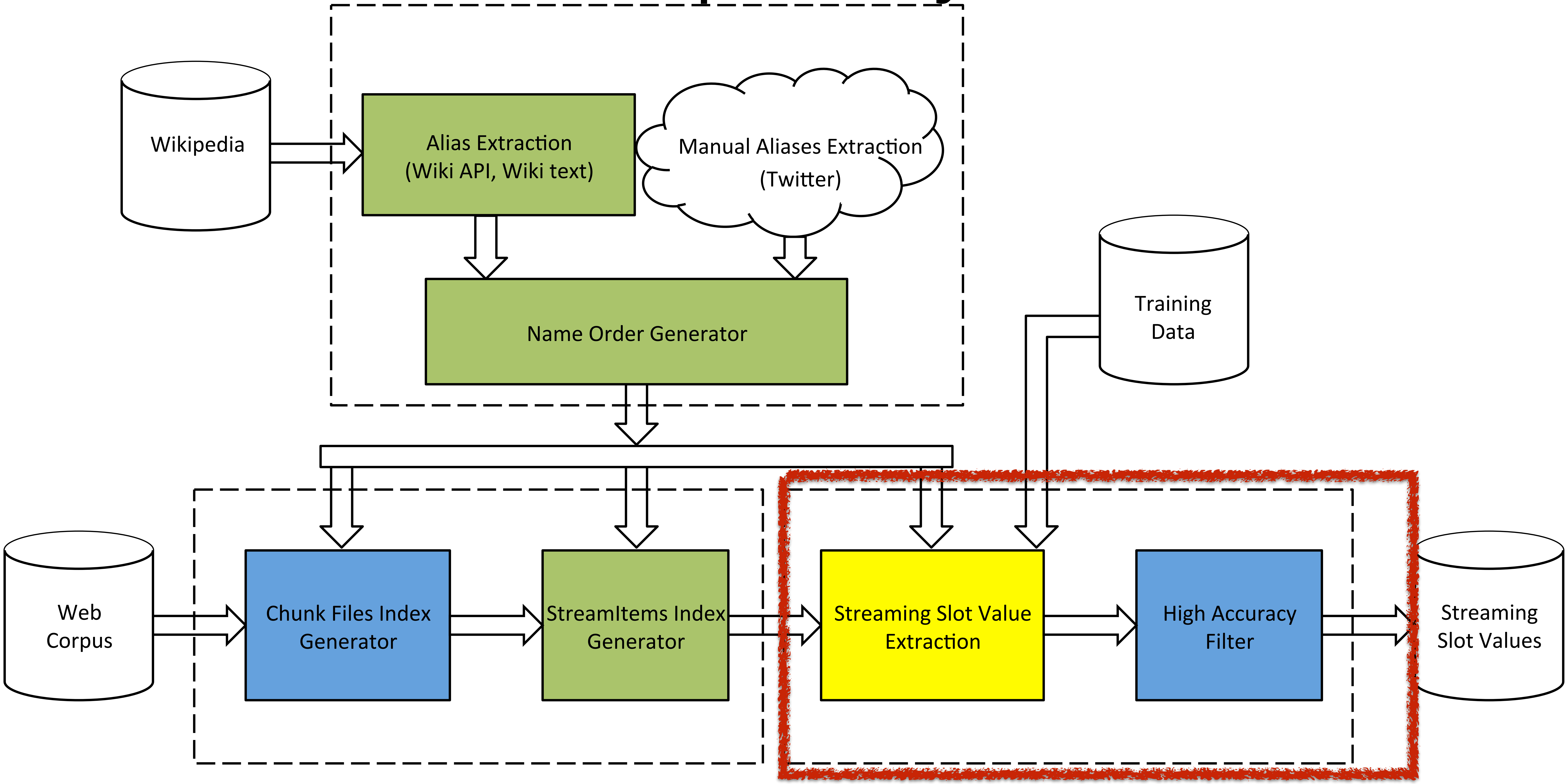
3) Ambiguous Entities

4) Finding *relevant* facts

Example System



Example System



Entity Resolution

- Entity resolution is the process of identifying and clustering different manifestations (e.g., mentions, noun phrases, named entities) of the same real world object.

Entity Resolution

- Entity resolution is the process of identifying and clustering different manifestations (e.g., mentions, noun phrases, named entities) of the same real world object.

- Difficult because of ambiguity

Entity Resolution

- Entity resolution is the process of identifying and clustering different manifestations (e.g., mentions, noun phrases, named entities) of the same real world object.

- Difficult because of ambiguity

Same Name, Different Person

Entity Resolution

- Entity resolution is the process of identifying and clustering different manifestations (e.g., mentions, noun phrases, named entities) of the same real world object.

- Difficult because of ambiguity

Same Name, Different Person

Different Name, Same Person

Entity Resolution

- Entity resolution is the process of identifying and clustering different manifestations (e.g., mentions, noun phrases, named entities) of the same real world object.

- Difficult because of ambiguity

Same Name, Different Person

Different Name, Same Person

Entity Resolution

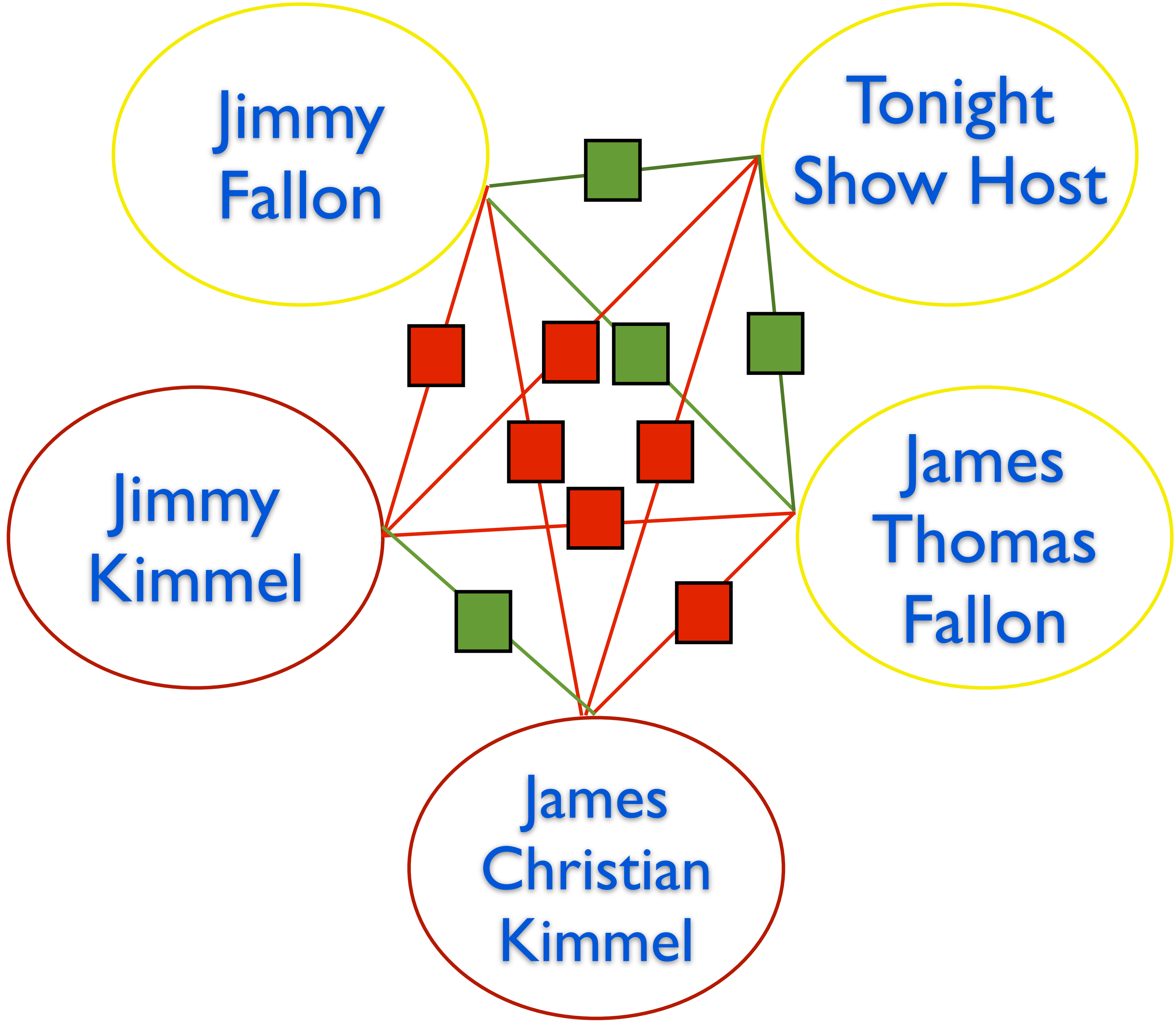
- Entity resolution is the process of identifying and clustering different manifestations (e.g., mentions, noun phrases, named entities) of the same real world object.

  - Difficult because of ambiguity

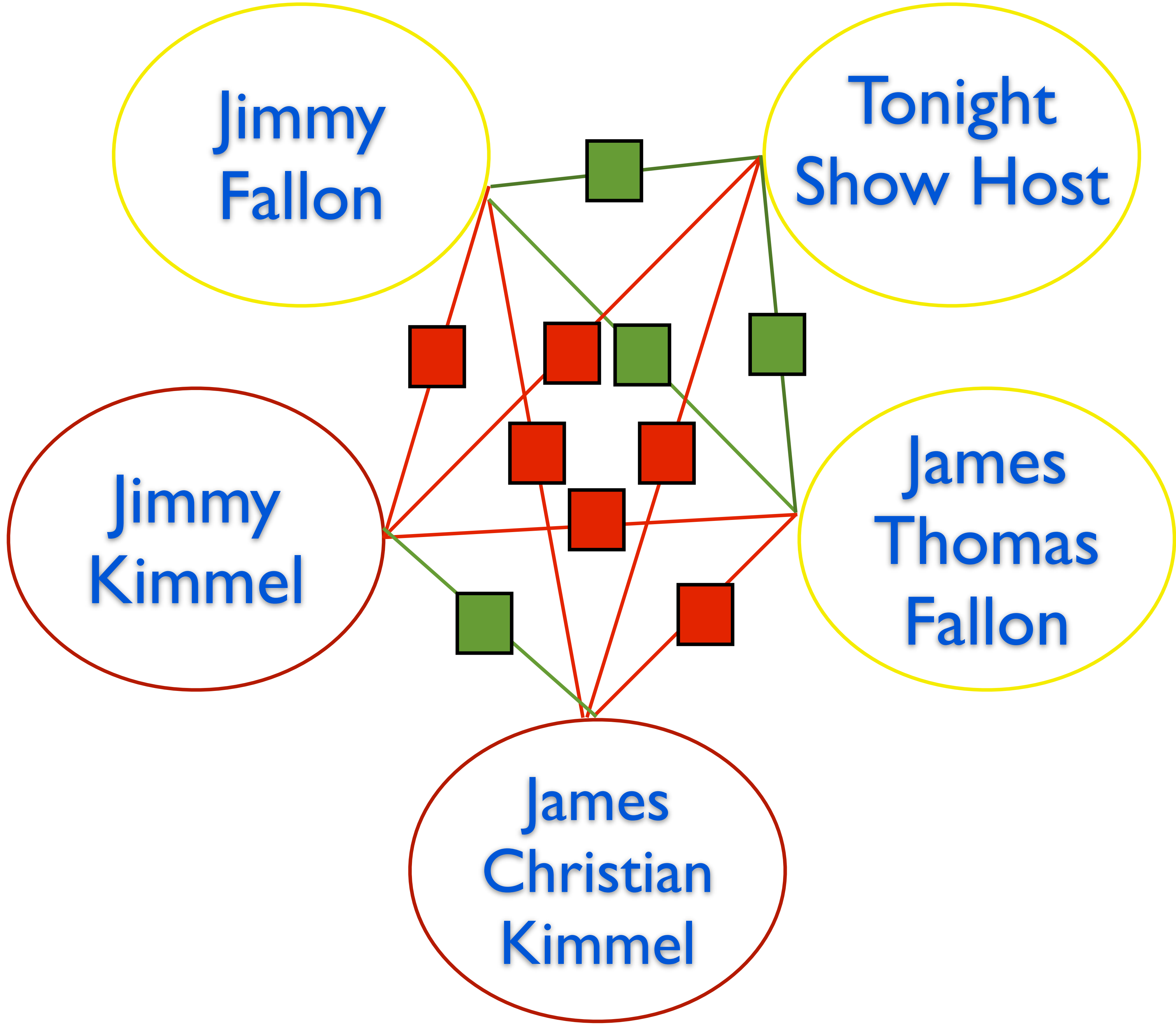
Same Name, Different Person

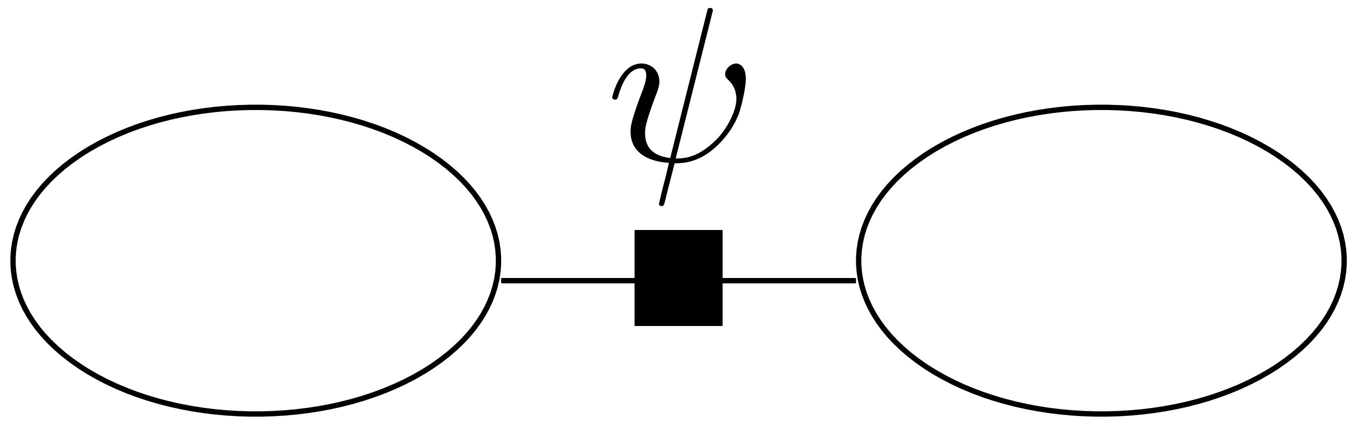
Different Name, Same Person

Entity Resolution Model



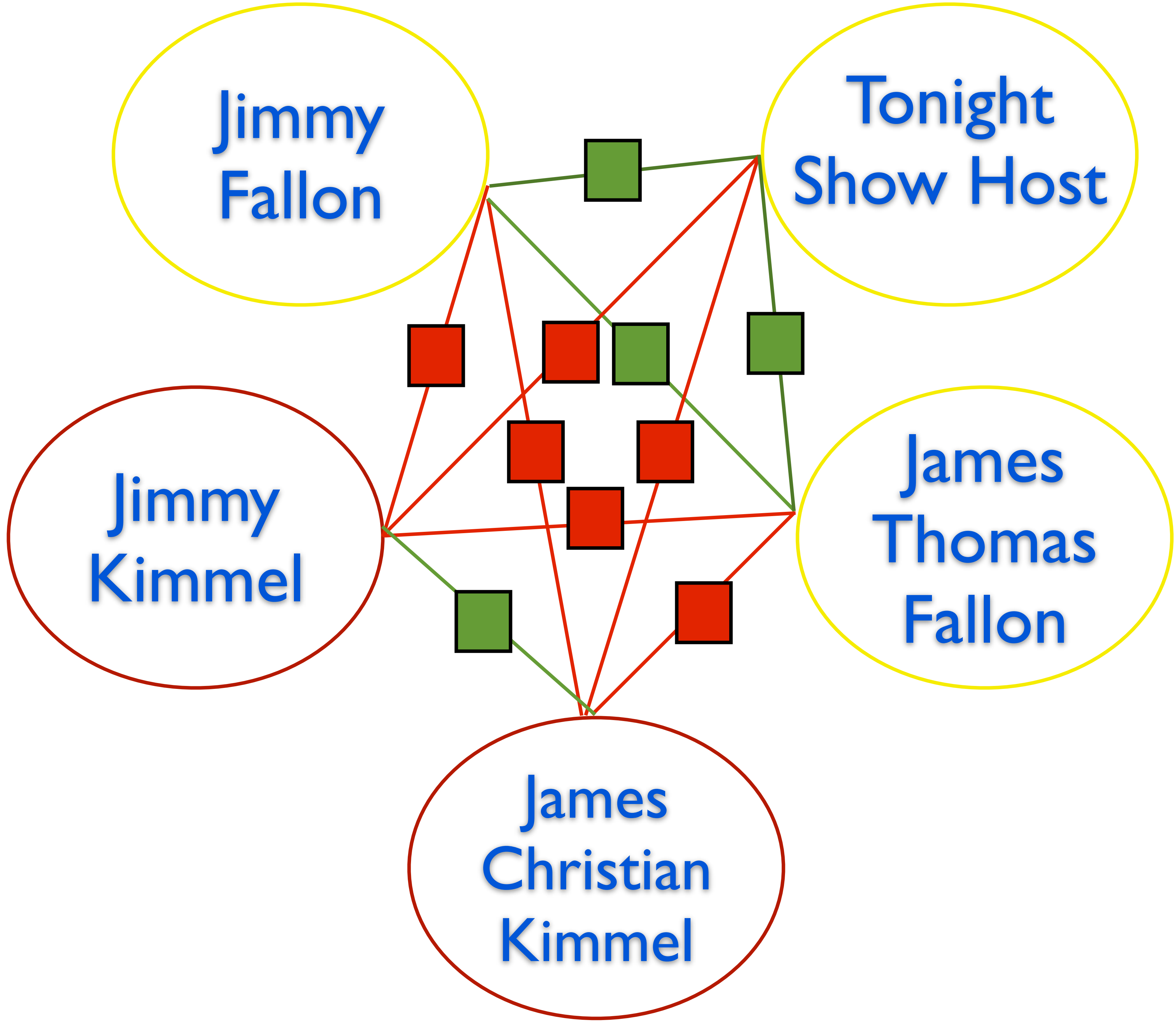
Entity Resolution Model

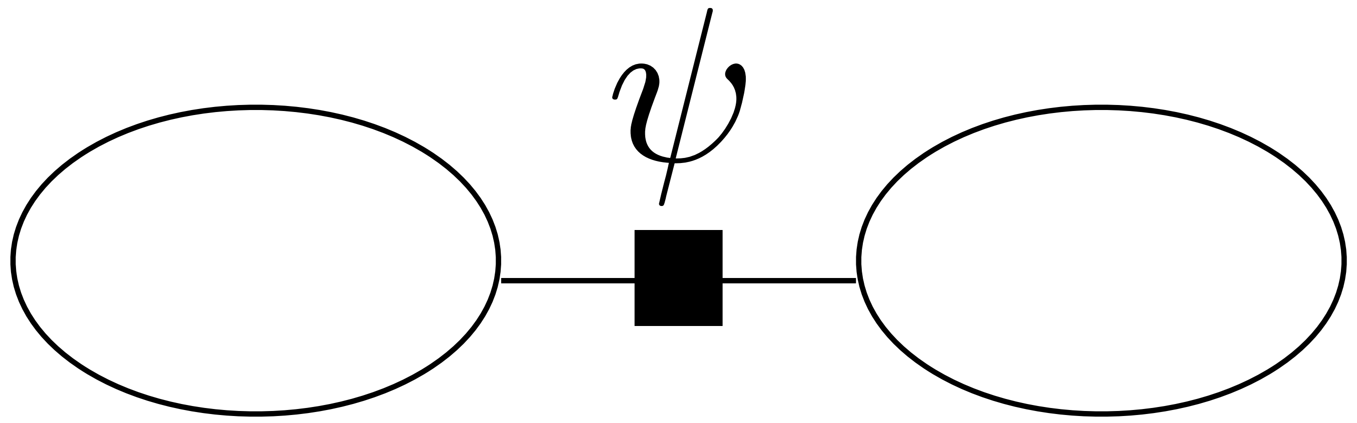




Entity Resolution Model

$$\begin{aligned}\hat{\mathbf{e}} &= \arg\max_{\mathbf{e}} p(\mathbf{e}) \\ &= \arg\max_{\mathbf{e}} \sum_{e\in\mathbf{e}} \Big\{ \sum_{m,n\in e} \psi_a^{mn} \\ &\qquad + \sum_{m\in e, n\notin e} \psi_r^{mn} \Big\}\end{aligned}$$

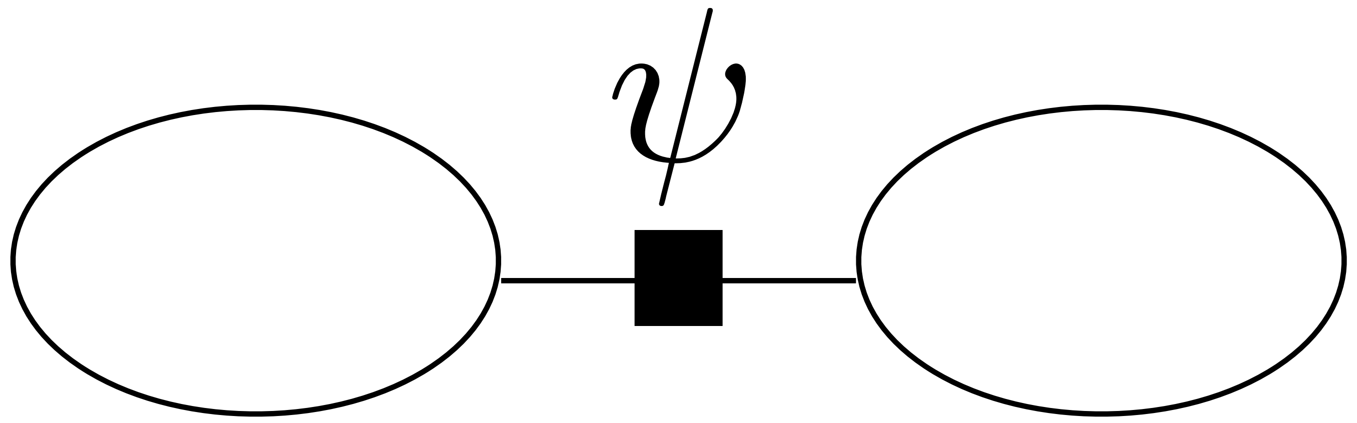


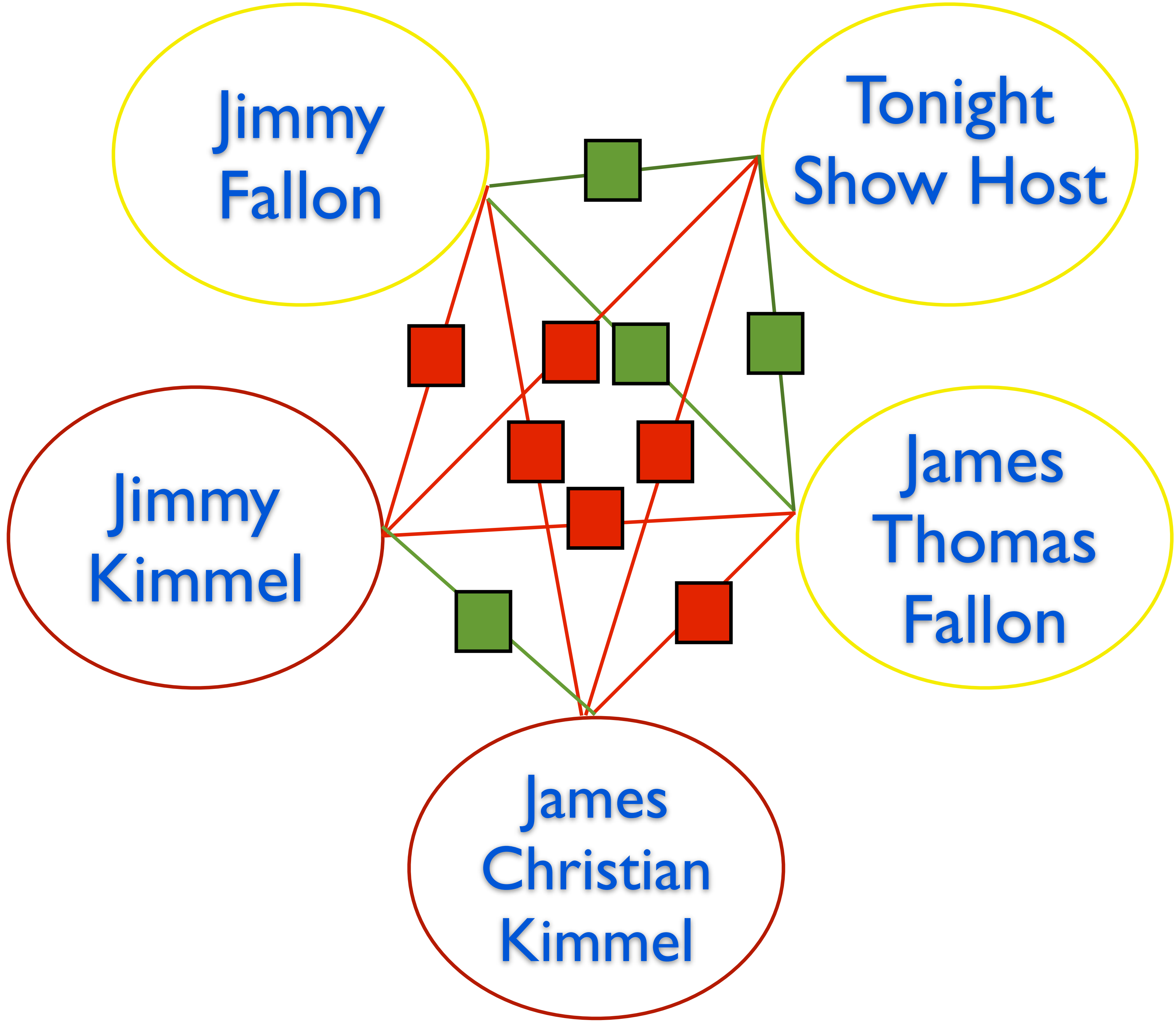


Entity Resolution Model

Find the best arrangement.

$$\begin{aligned}\hat{\mathbf{e}} &= \arg\max_{\mathbf{e}} p(\mathbf{e}) \\ &= \arg\max_{\mathbf{e}} \sum_{e\in\mathbf{e}} \Big\{ \sum_{m,n\in e} \psi_a^{mn} \\ &\qquad + \sum_{m\in e, n\notin e} \psi_r^{mn} \Big\}\end{aligned}$$

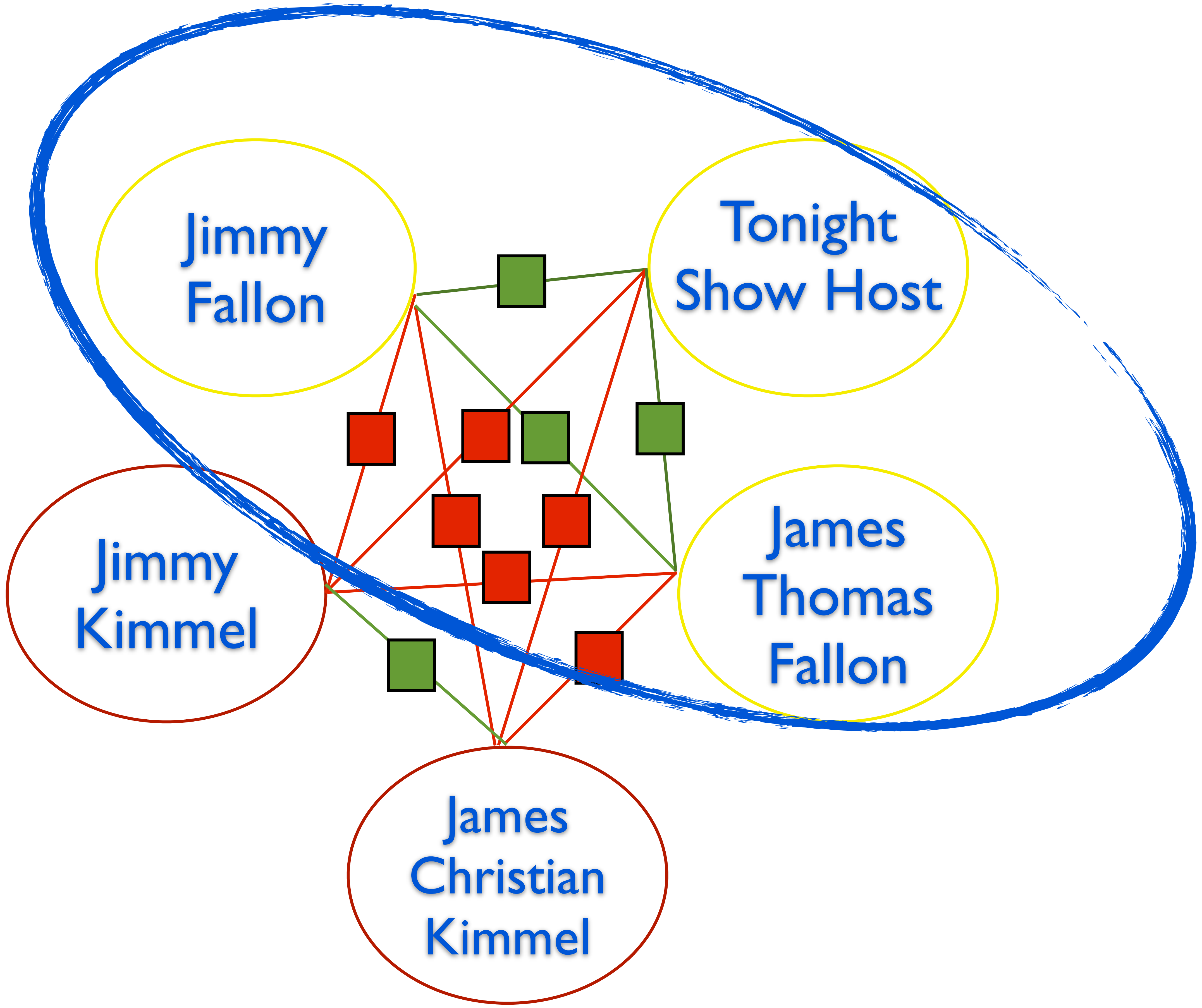


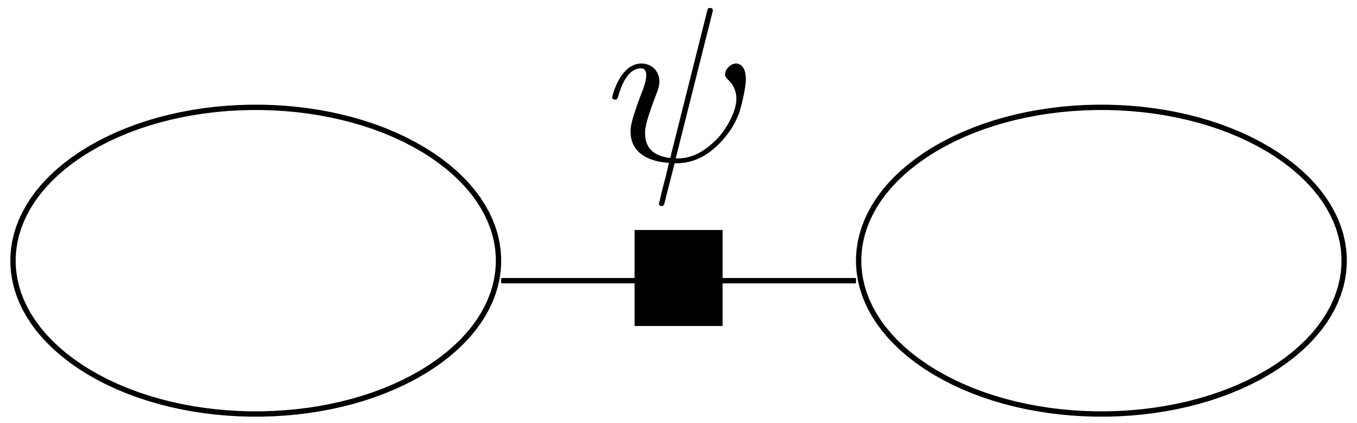


Entity Resolution Model

Find the best arrangement.

$$\begin{aligned}\hat{\mathbf{e}} &= \arg\max_{\mathbf{e}} p(\mathbf{e}) \\ &= \arg\max_{\mathbf{e}} \sum_{e\in\mathbf{e}} \Big\{ \sum_{m,n\in e} \psi_a^{mn} \\ &\qquad + \sum_{m\in e, n\notin e} \psi_r^{mn} \Big\}\end{aligned}$$

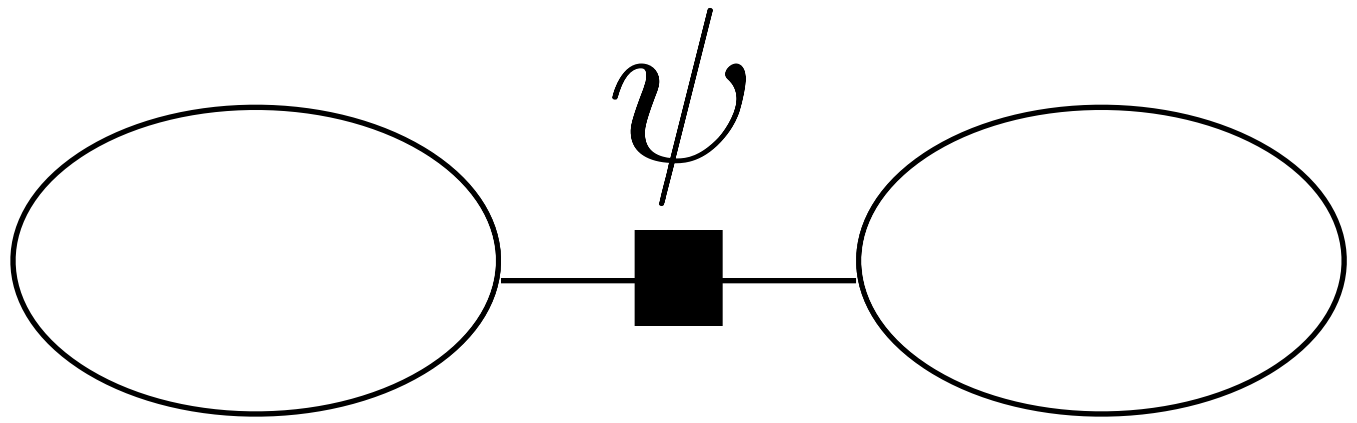


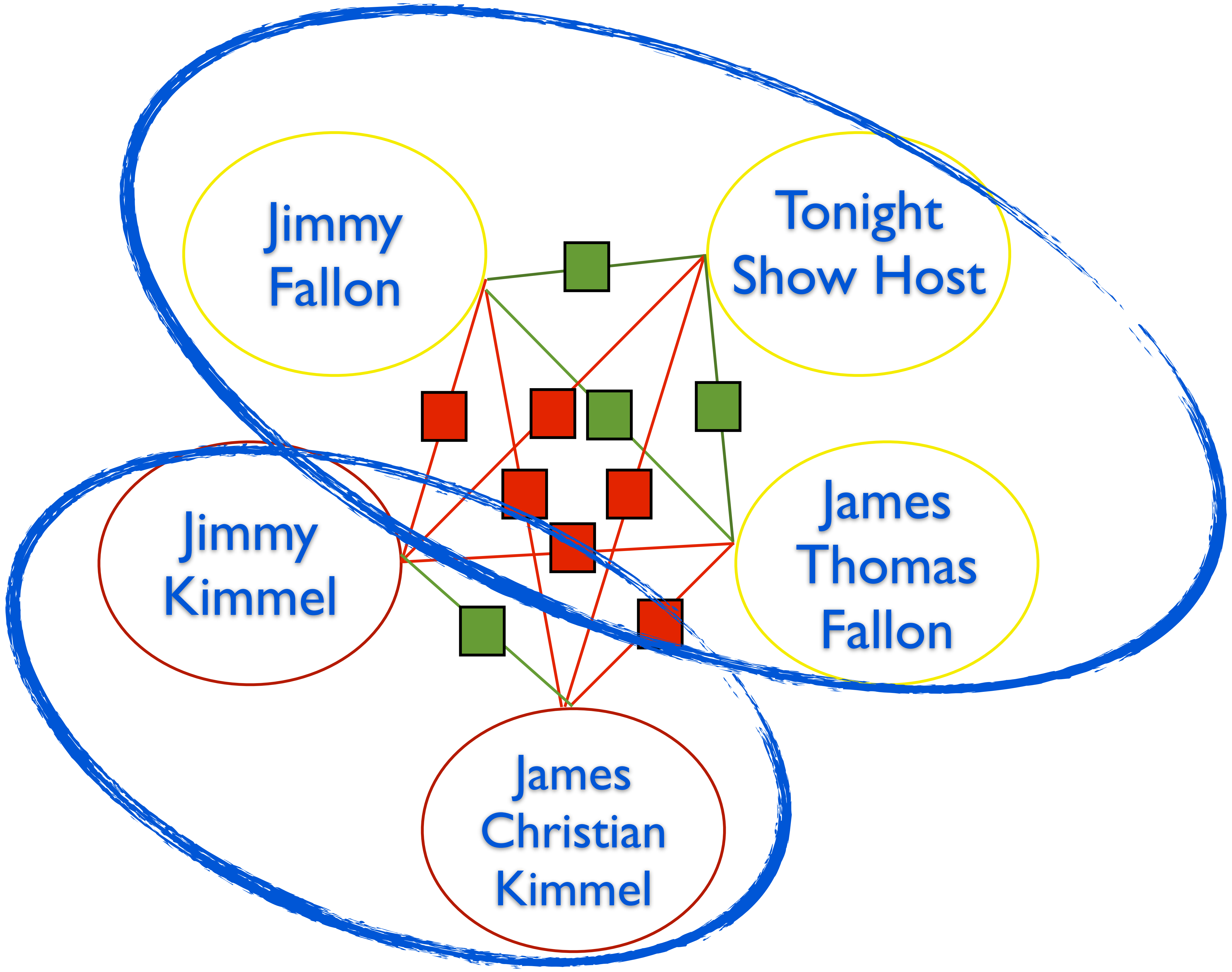


Entity Resolution Model

Find the best arrangement.

$$\begin{aligned}\hat{\mathbf{e}} &= \arg\max_{\mathbf{e}} p(\mathbf{e}) \\ &= \arg\max_{\mathbf{e}} \sum_{e\in\mathbf{e}} \Big\{ \sum_{m,n\in e} \psi_a^{mn} \\ &\qquad + \sum_{m\in e,n\notin e} \psi_r^{mn} \Big\}\end{aligned}$$





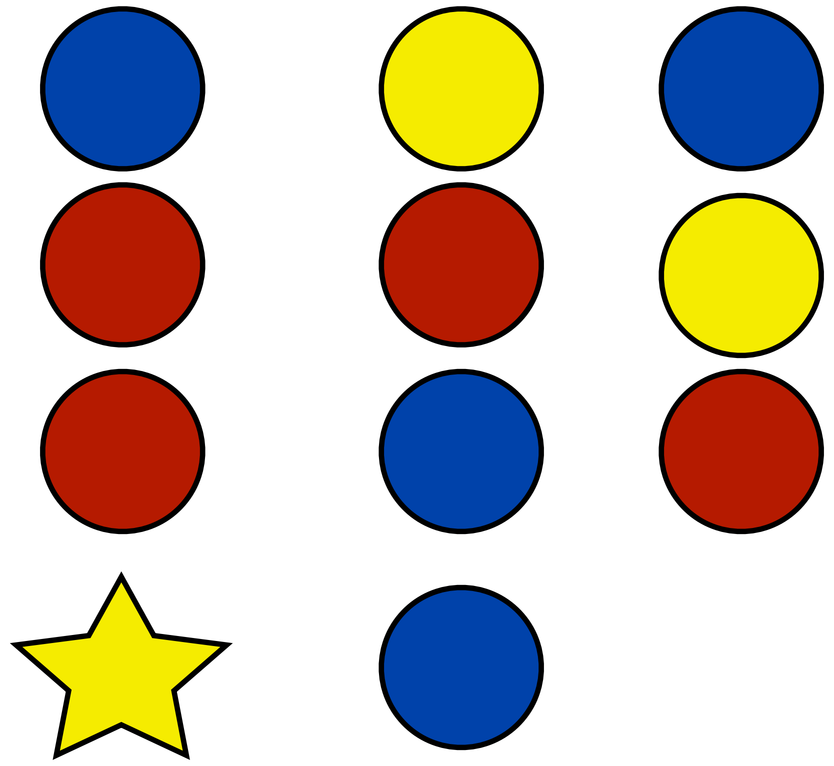
Entity Resolution Algorithm

The Baseline ER metropolis hastings takes a random mention and adds it to a random entity.

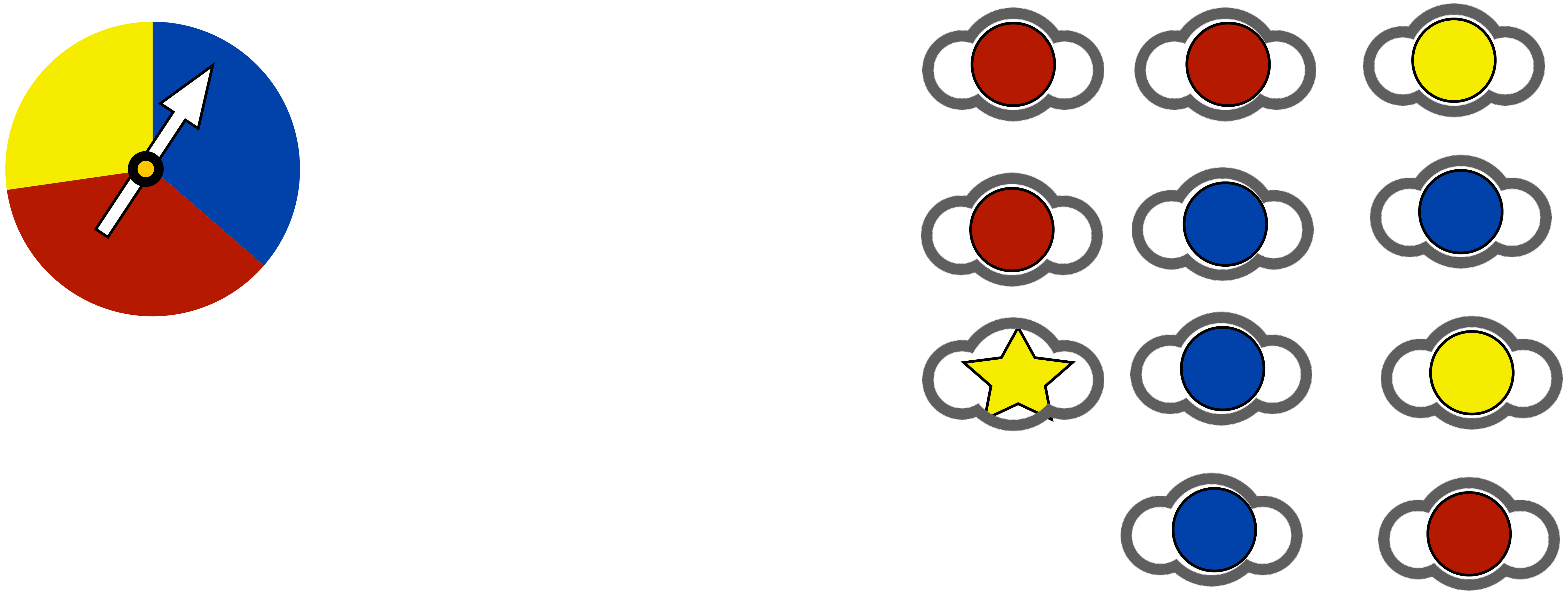
Entity Resolution Algorithm

The Baseline ER metropolis hastings takes a random mention and adds it to a random entity.

Random Number Generator

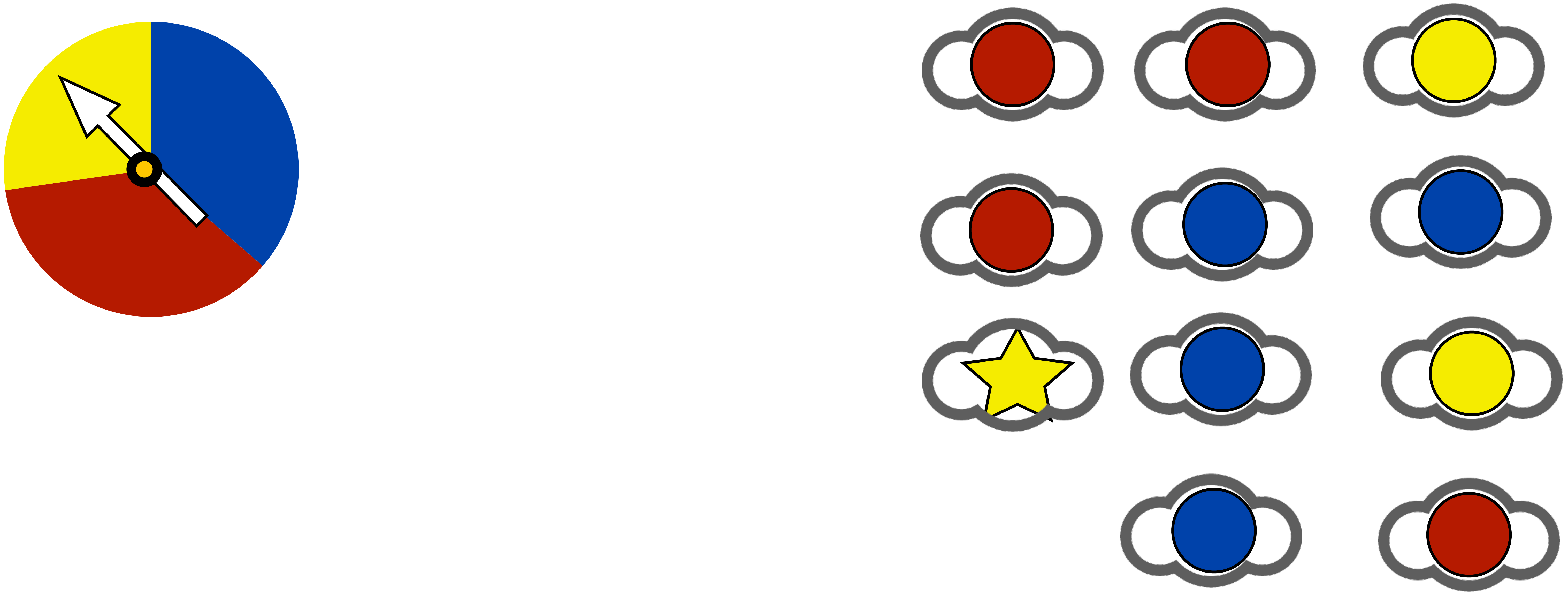


Entity Resolution Algorithm



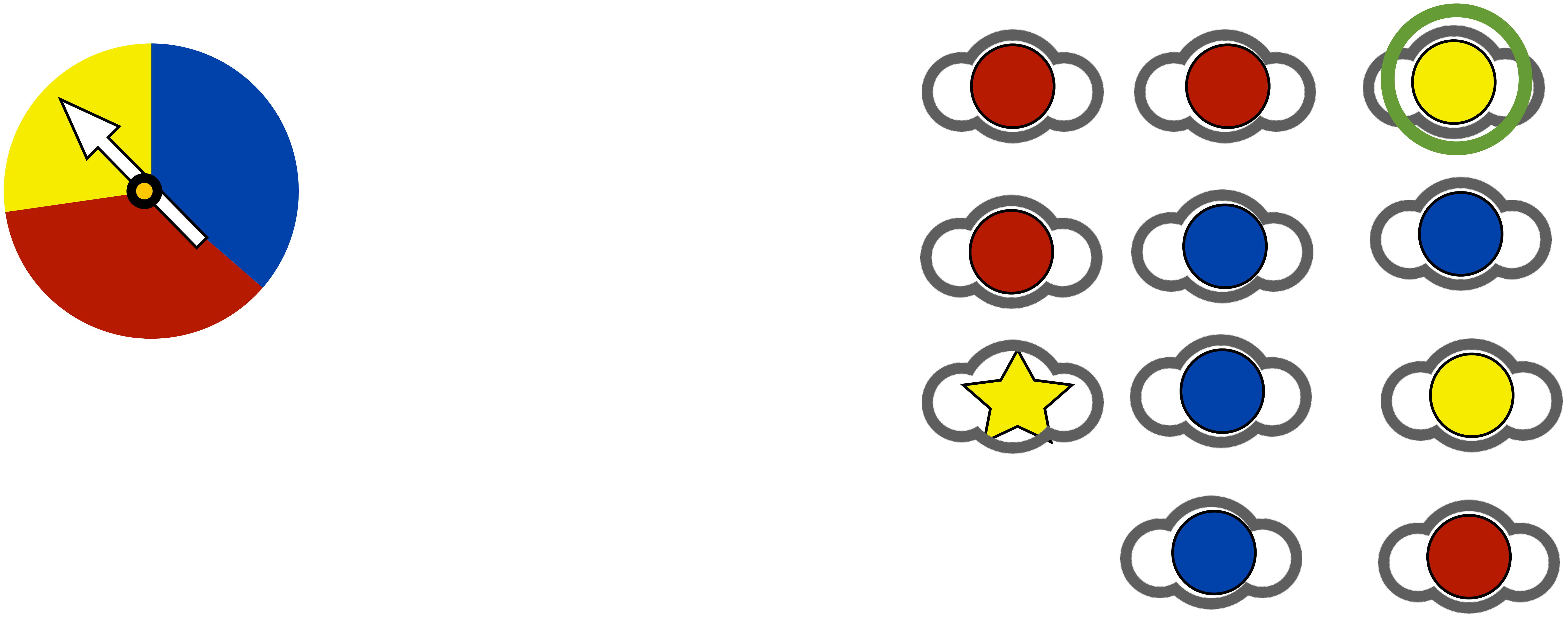
Entity Resolution Algorithm

1. Select a source mention at *random*.



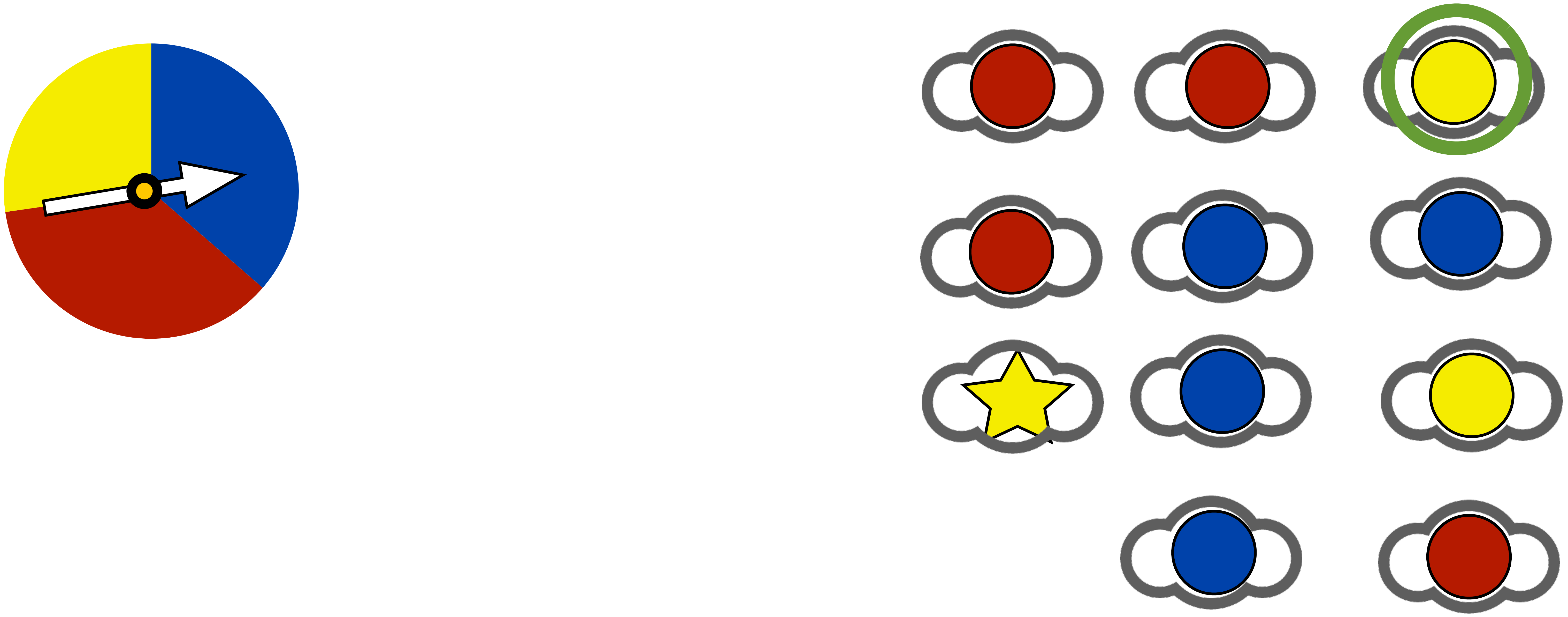
Entity Resolution Algorithm

1. Select a source mention at *random*.



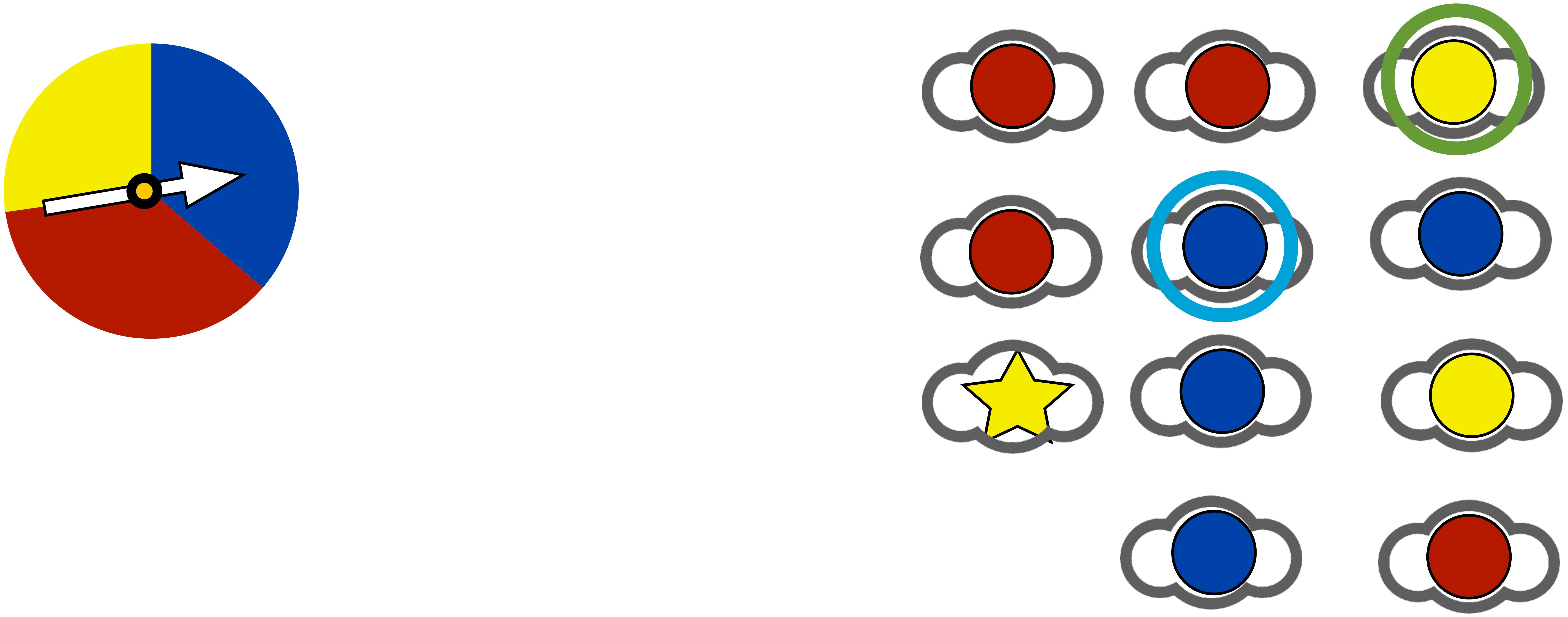
Entity Resolution Algorithm

1. Select a source mention at *random*.
2. Select a destination mention at *random*.



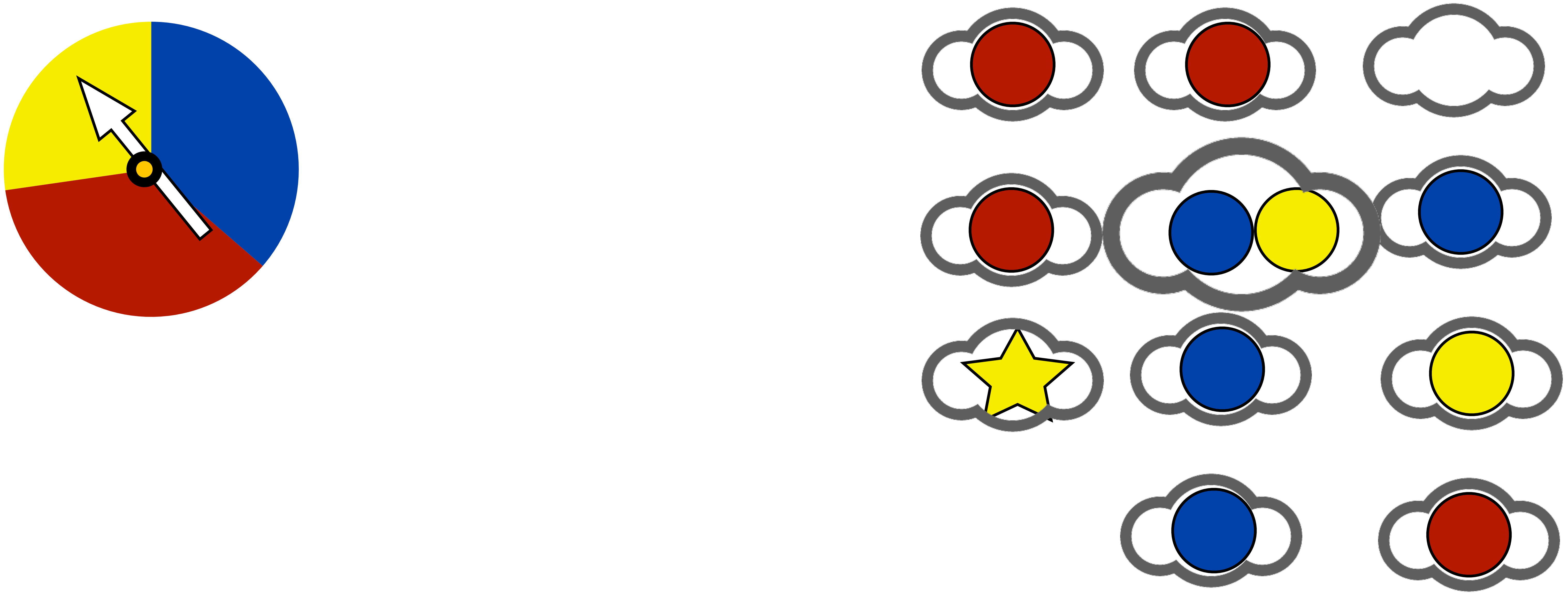
Entity Resolution Algorithm

1. Select a source mention at *random*.
2. Select a destination mention at *random*.



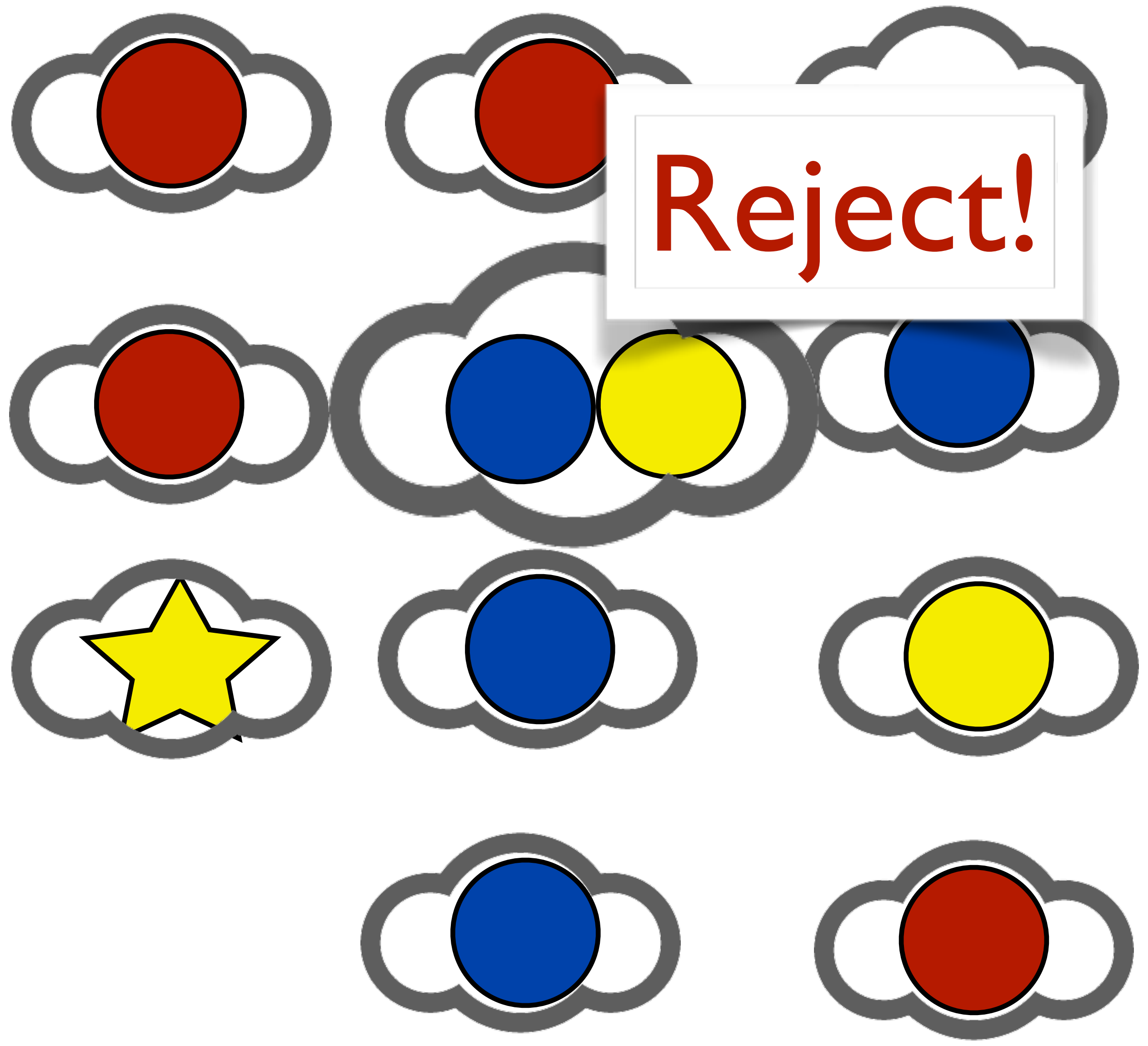
Entity Resolution Algorithm

1. Select a source mention at *random*.
2. Select a destination mention at *random*.
3. Propose a merge.



Entity Resolution Algorithm

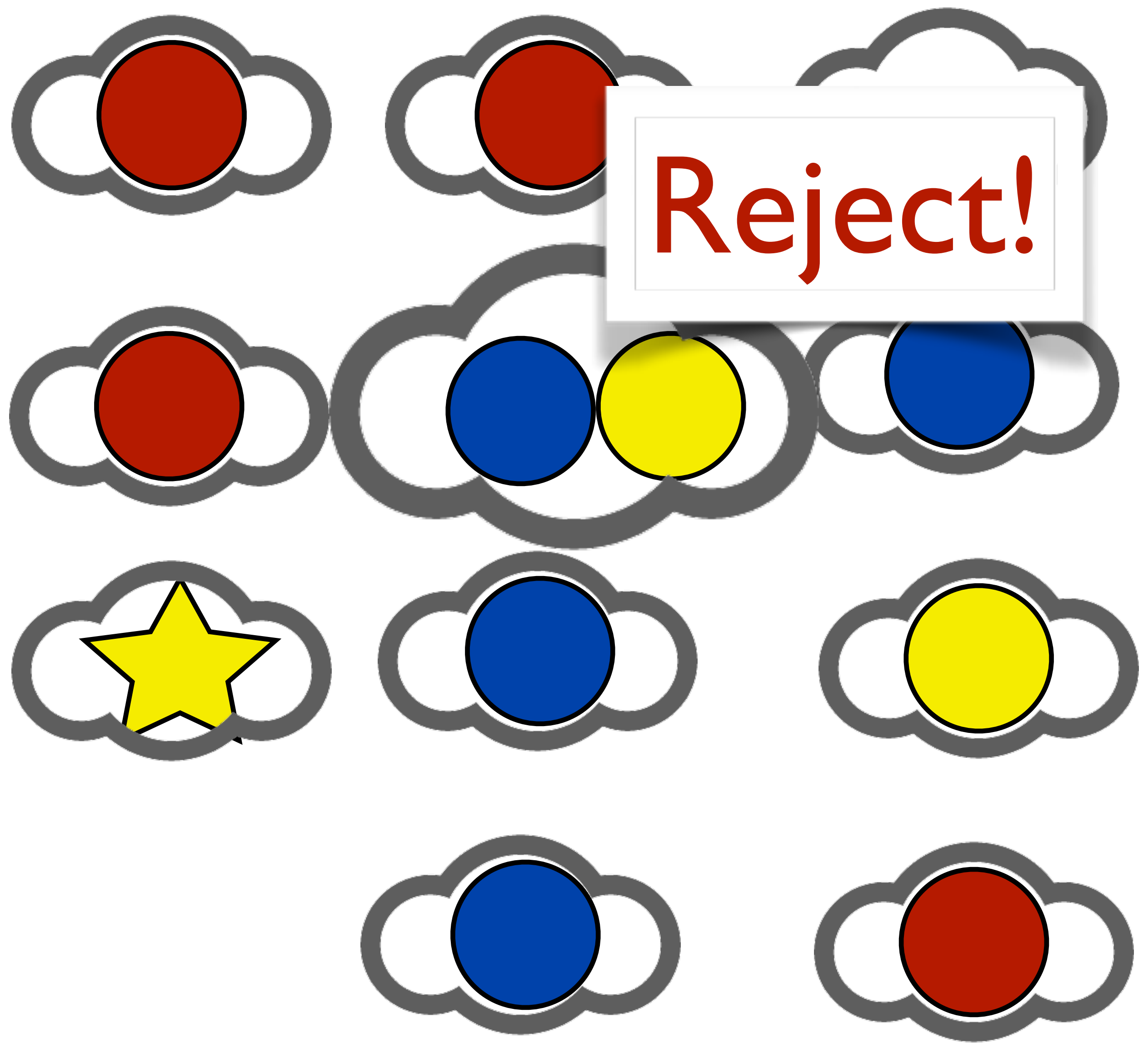
1. Select a source mention at *random*.
2. Select a destination mention at *random*.
3. Propose a merge.



Reject!

Entity Resolution Algorithm

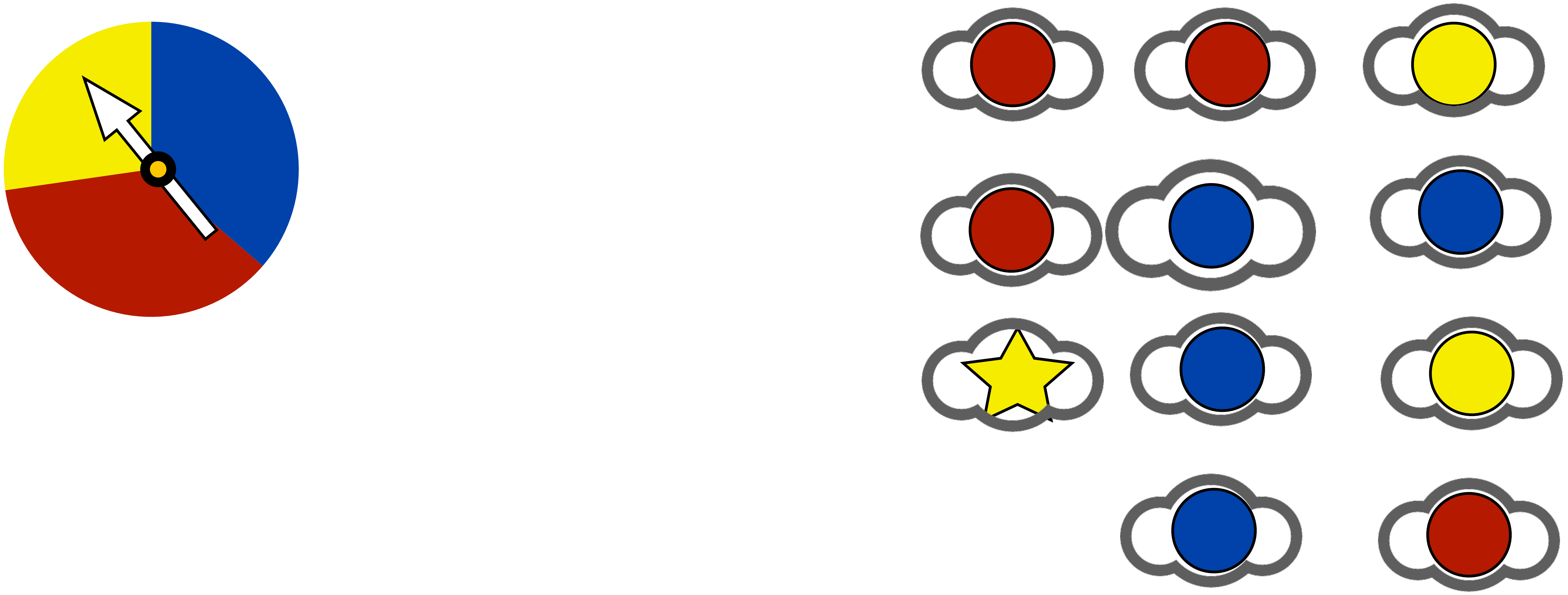
1. Select a source mention at *random*.
2. Select a destination mention at *random*.
3. Propose a merge.



$$\alpha(\mathbf{e}, \mathbf{e}') = \min\left(1, \frac{p(\mathbf{e}')}{p(\mathbf{e})}\right)$$

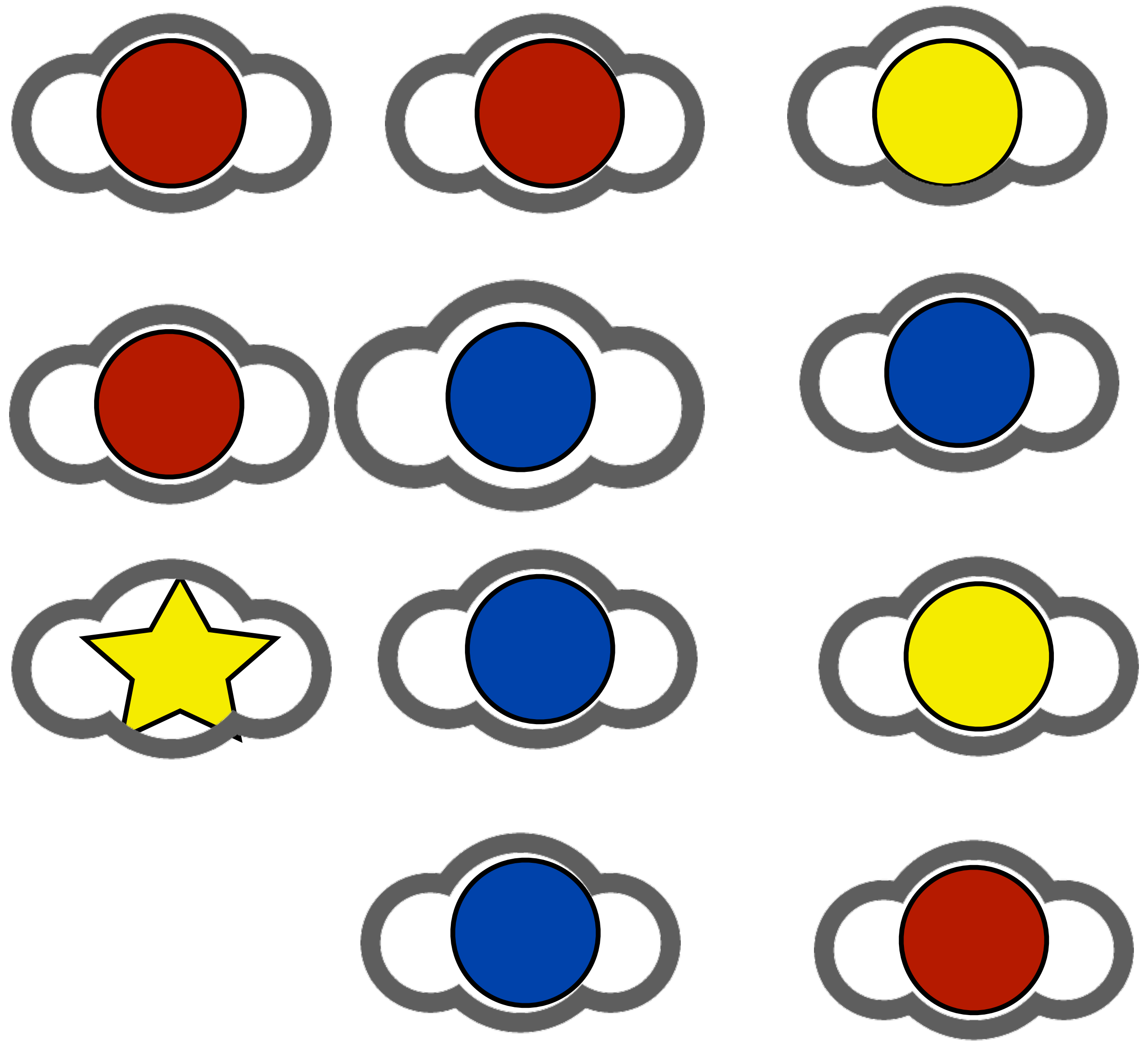
Entity Resolution Algorithm

1. Select a source mention at *random*.
2. Select a destination mention at *random*.
3. Propose a merge.



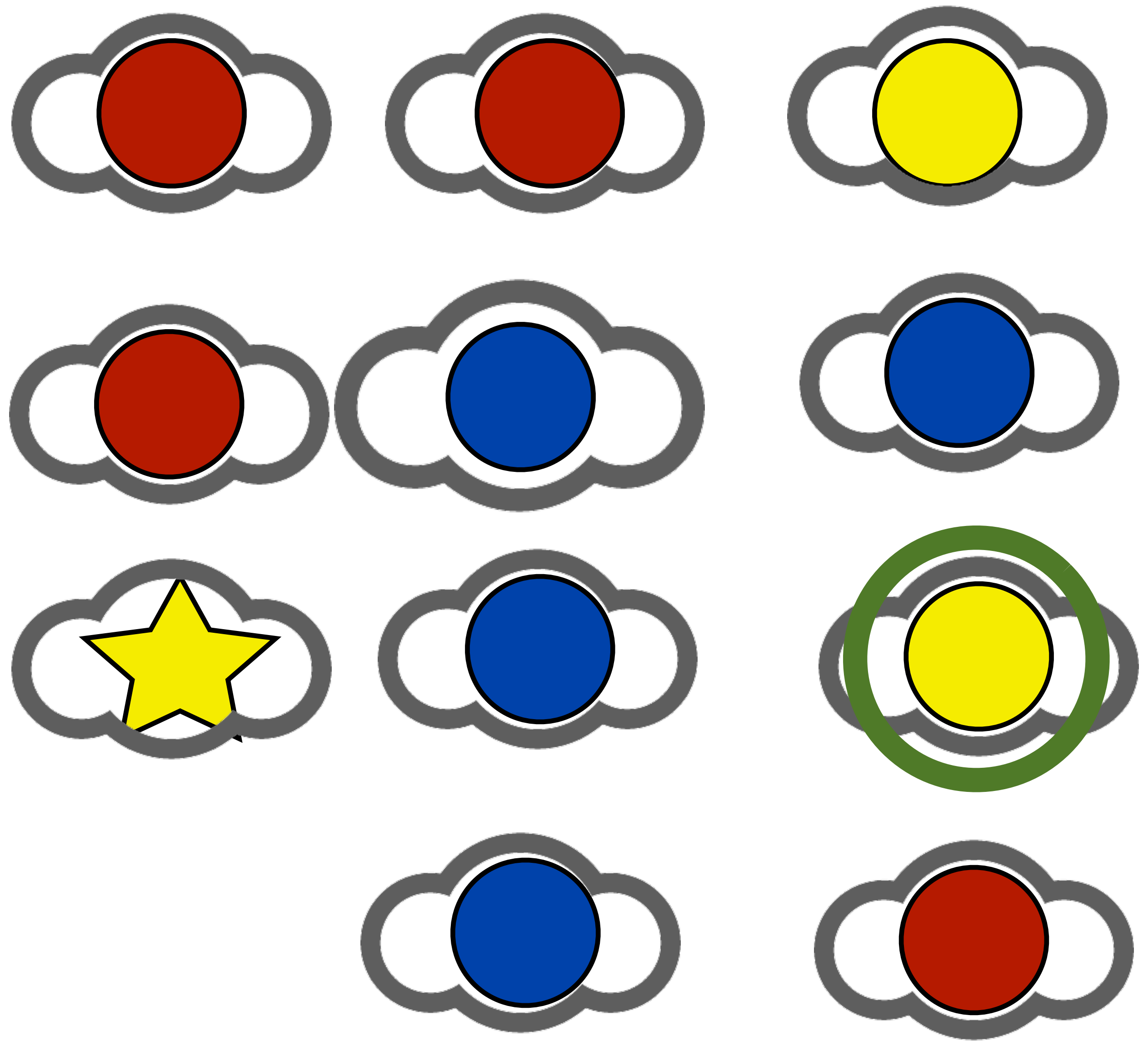
Entity Resolution Algorithm

1. Select a source mention at *random*.
2. Select a destination mention at *random*.
3. Propose a merge.



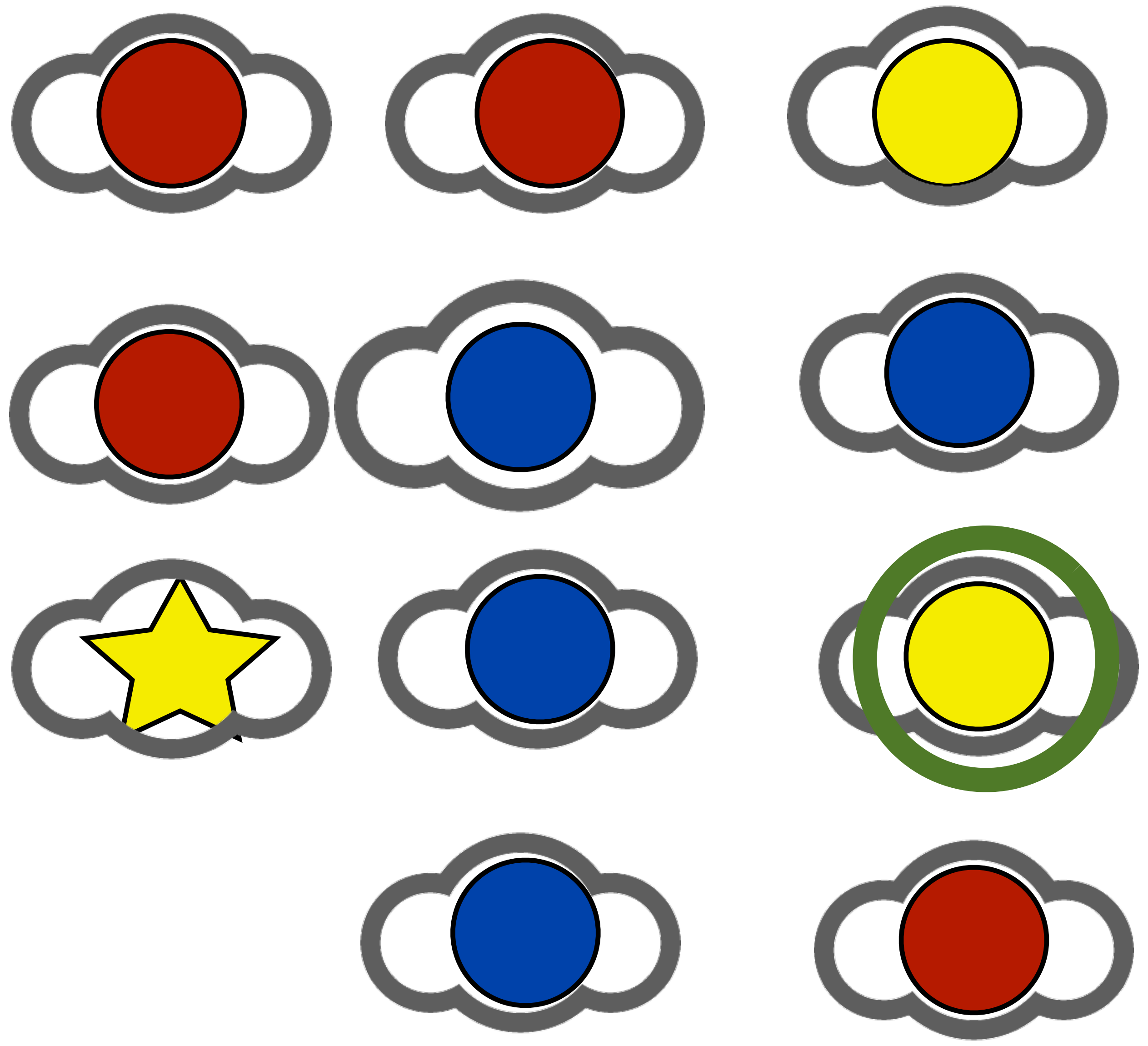
Entity Resolution Algorithm

1. Select a source mention at *random*.
2. Select a destination mention at *random*.
3. Propose a merge.



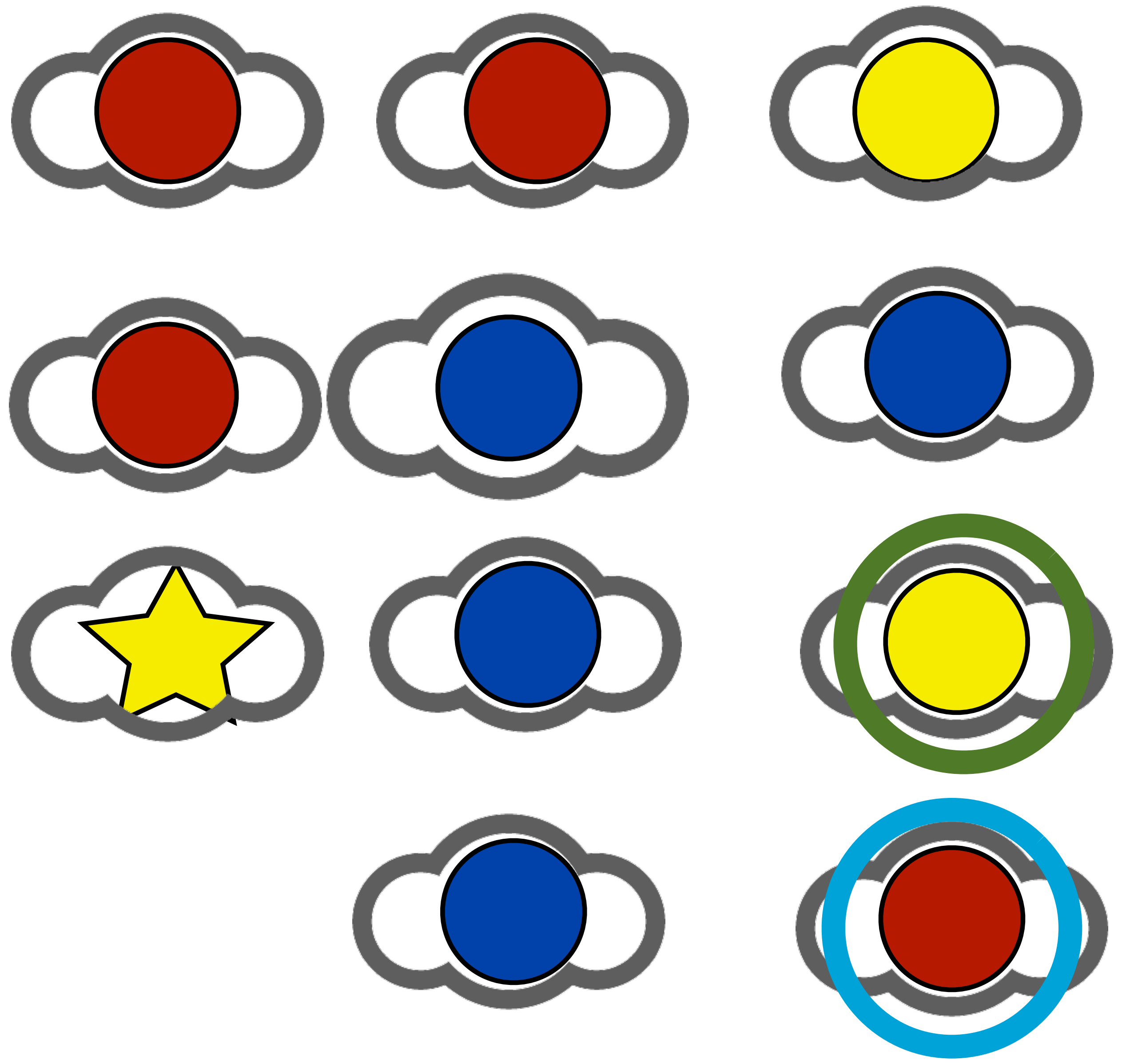
Entity Resolution Algorithm

1. Select a source mention at *random*.
2. Select a destination mention at *random*.
3. Propose a merge.



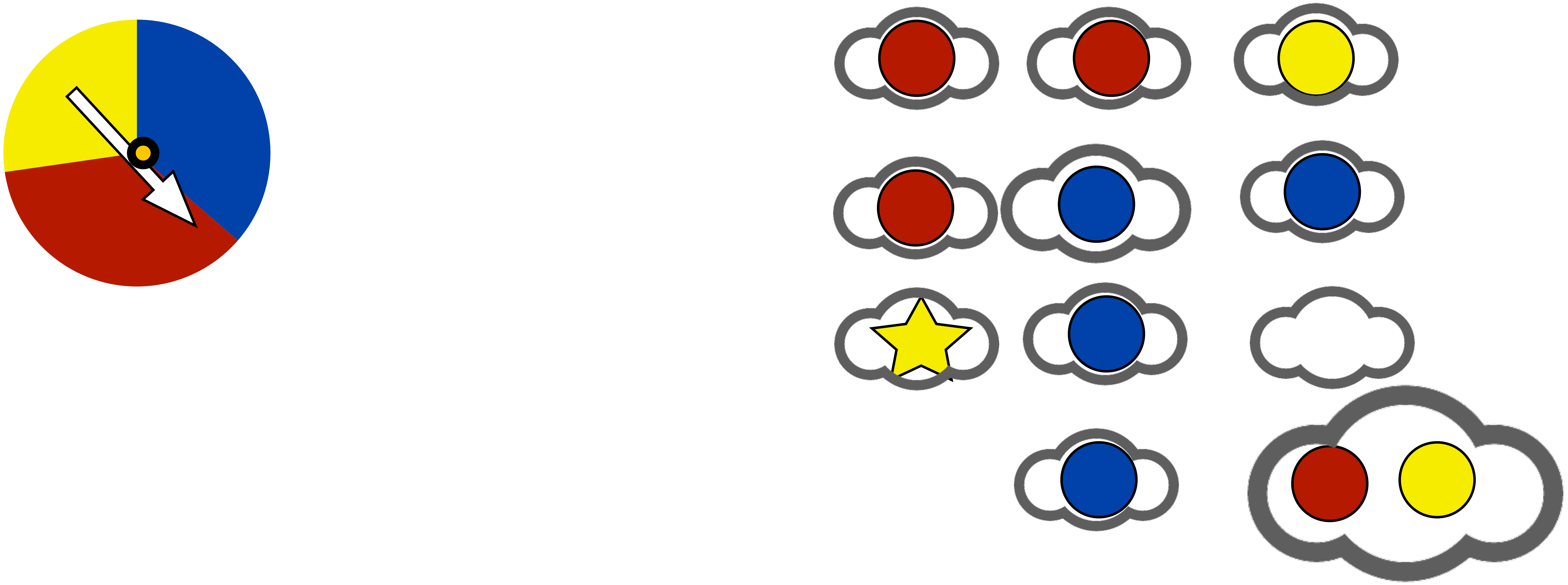
Entity Resolution Algorithm

1. Select a source mention at *random*.
2. Select a destination mention at *random*.
3. Propose a merge.



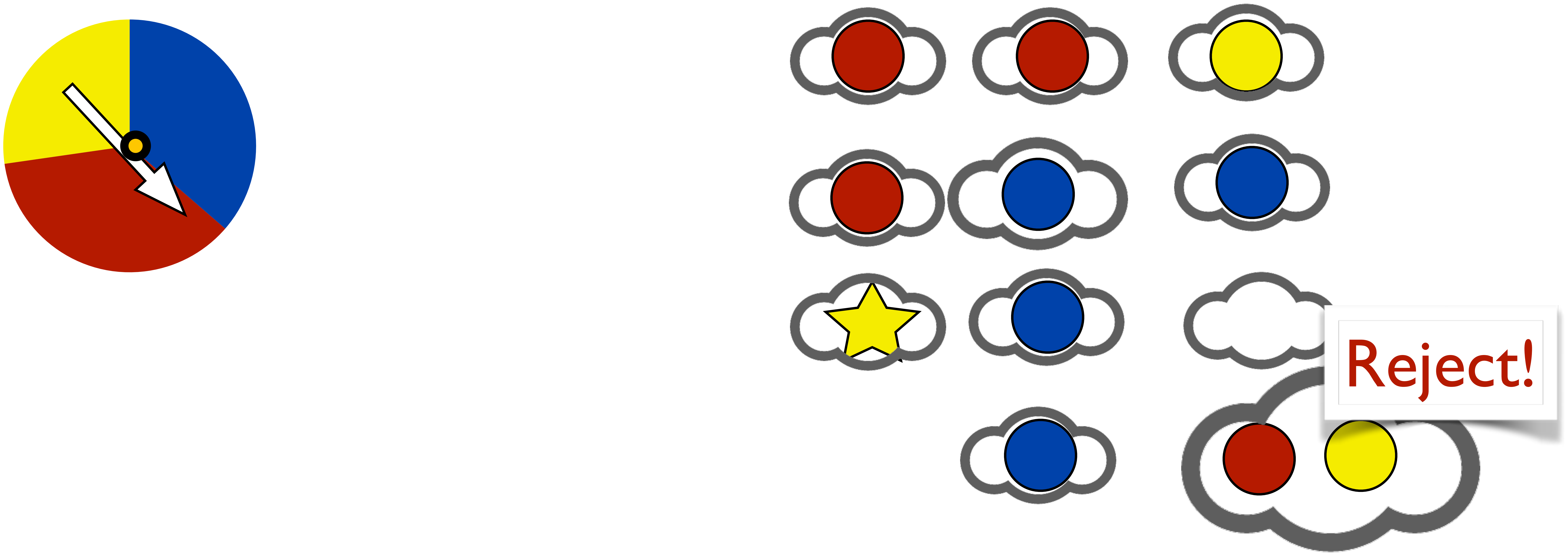
Entity Resolution Algorithm

1. Select a source mention at *random*.
2. Select a destination mention at *random*.
3. Propose a merge.



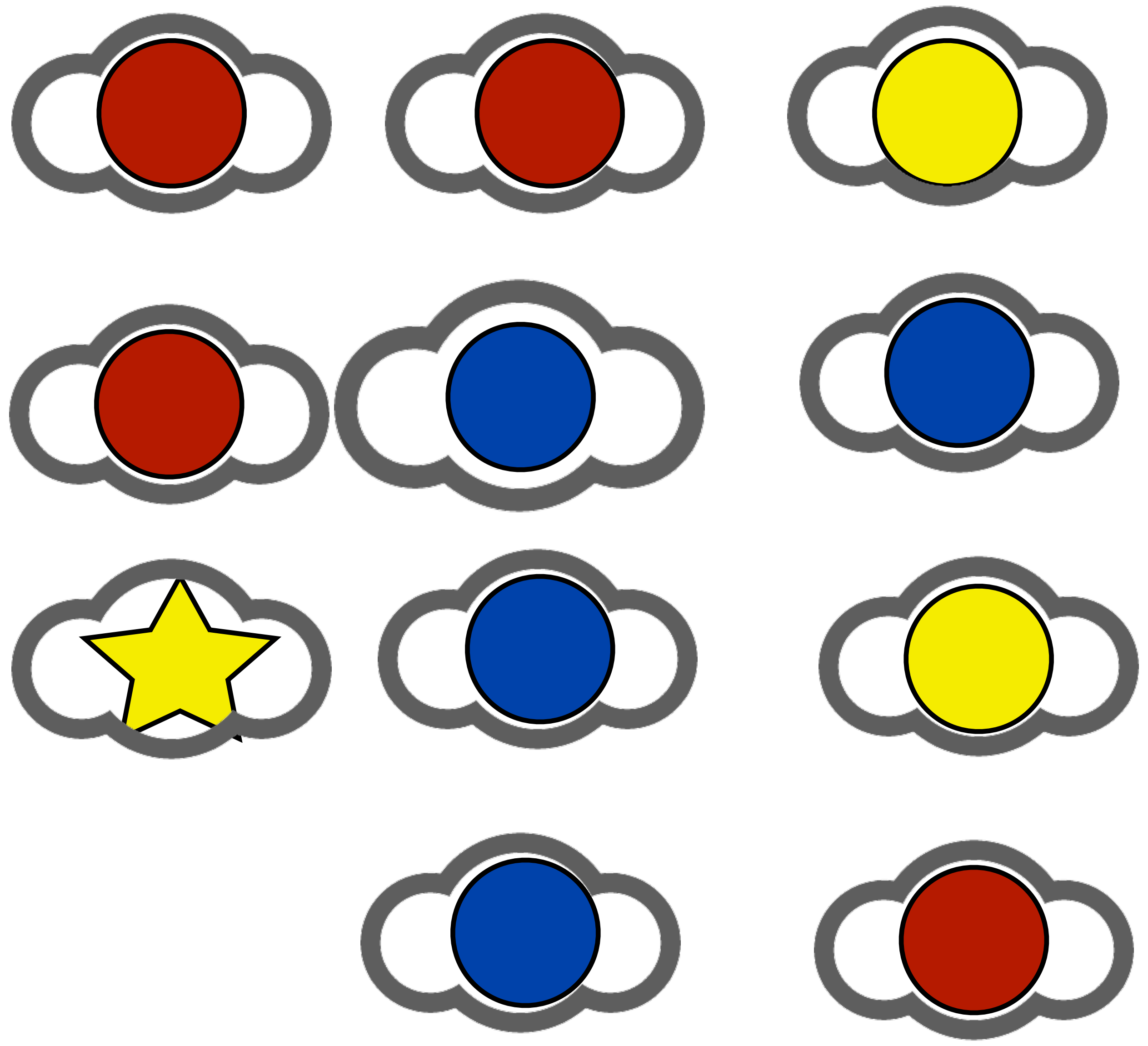
Entity Resolution Algorithm

1. Select a source mention at *random*.
2. Select a destination mention at *random*.
3. Propose a merge.



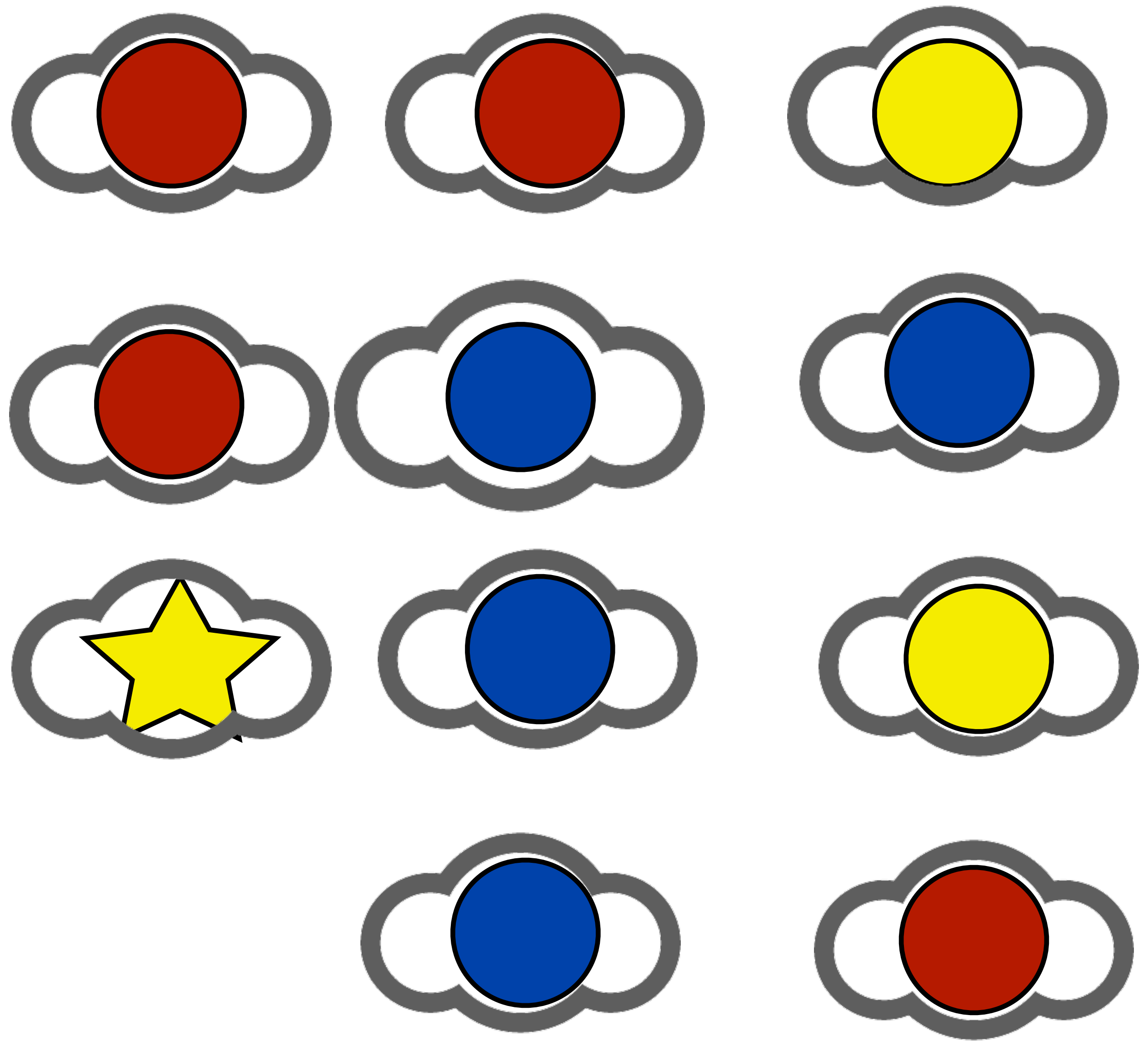
Entity Resolution Algorithm

1. Select a source mention at *random*.
2. Select a destination mention at *random*.
3. Propose a merge.



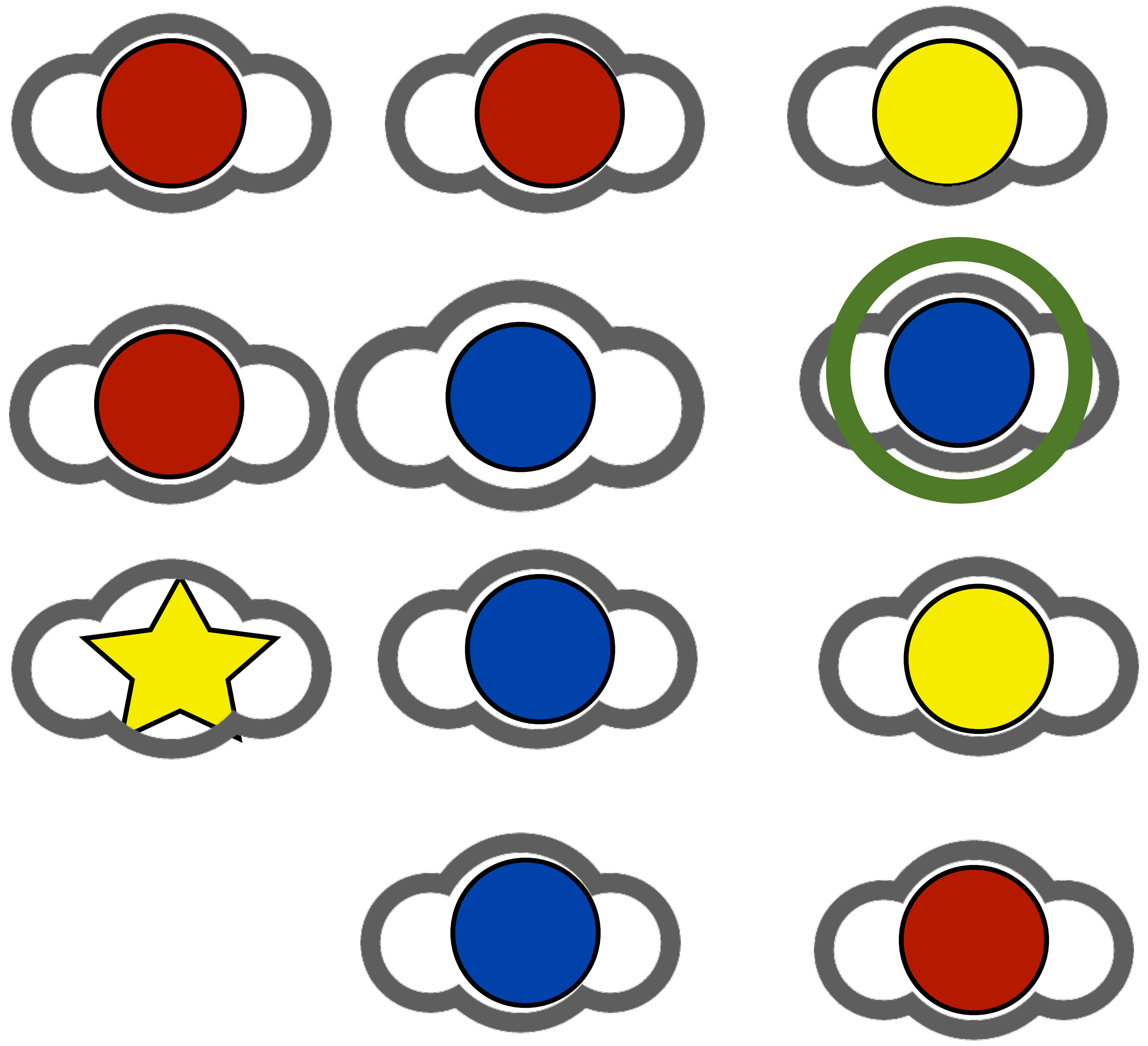
Entity Resolution Algorithm

1. Select a source mention at *random*.
2. Select a destination mention at *random*.
3. Propose a merge.



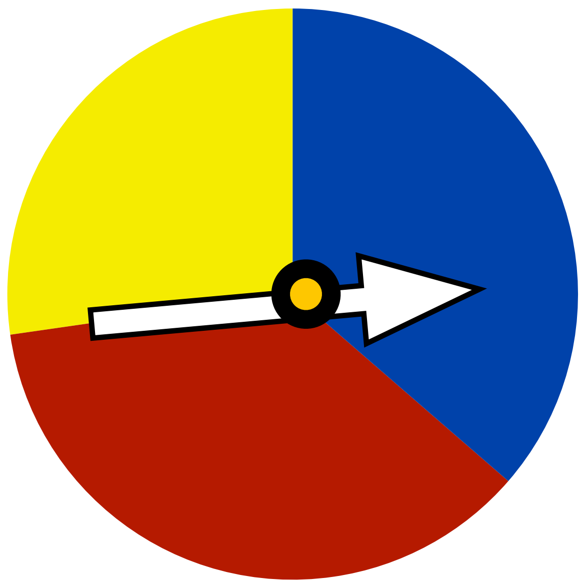
Entity Resolution Algorithm

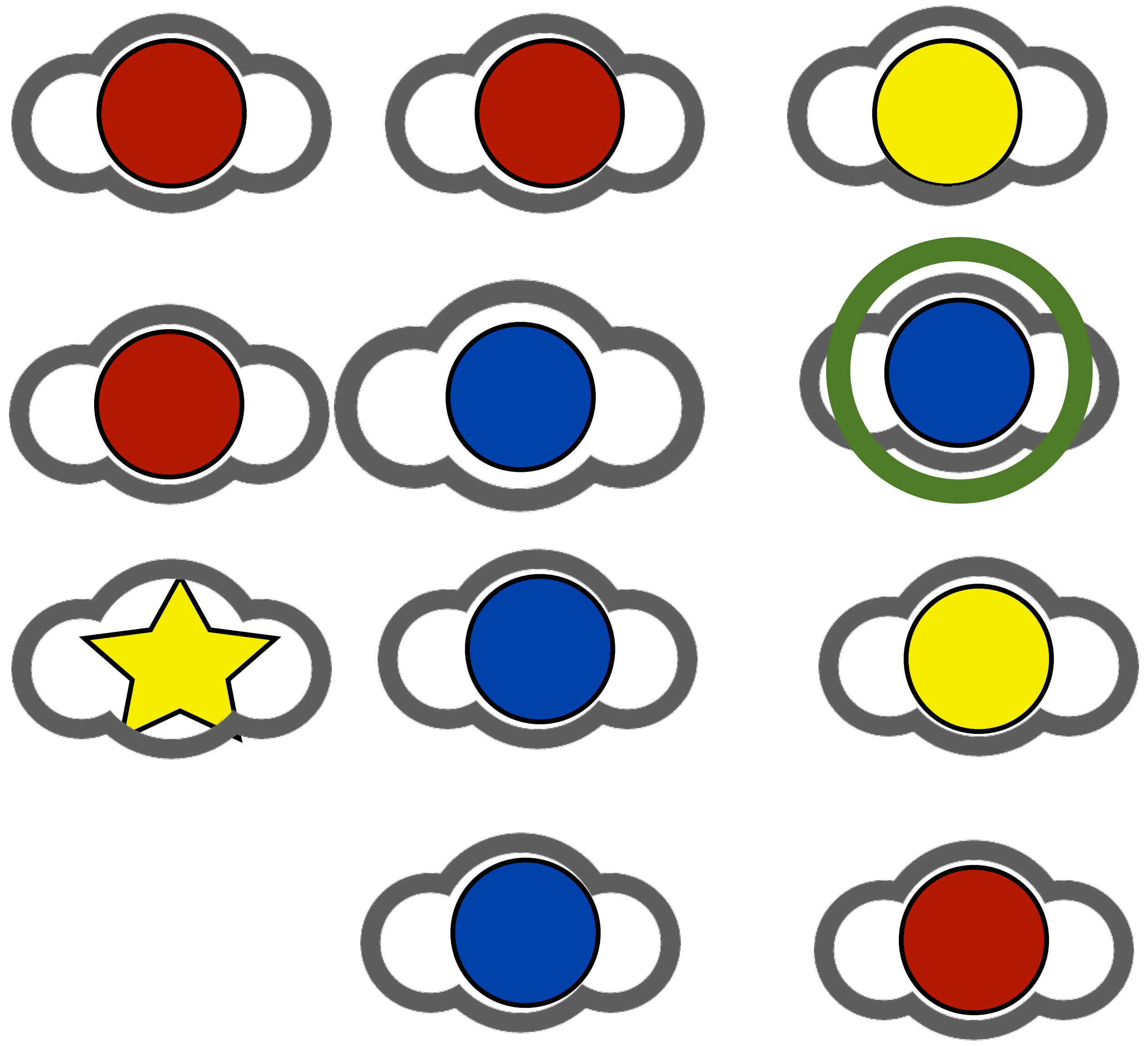
1. Select a source mention at *random*.
2. Select a destination mention at *random*.
3. Propose a merge.



Entity Resolution Algorithm

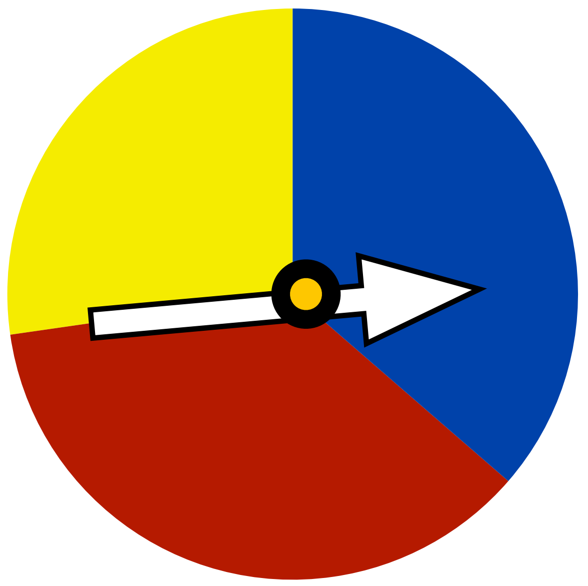
1. Select a source mention at *random*.
2. Select a destination mention at *random*.
3. Propose a merge.

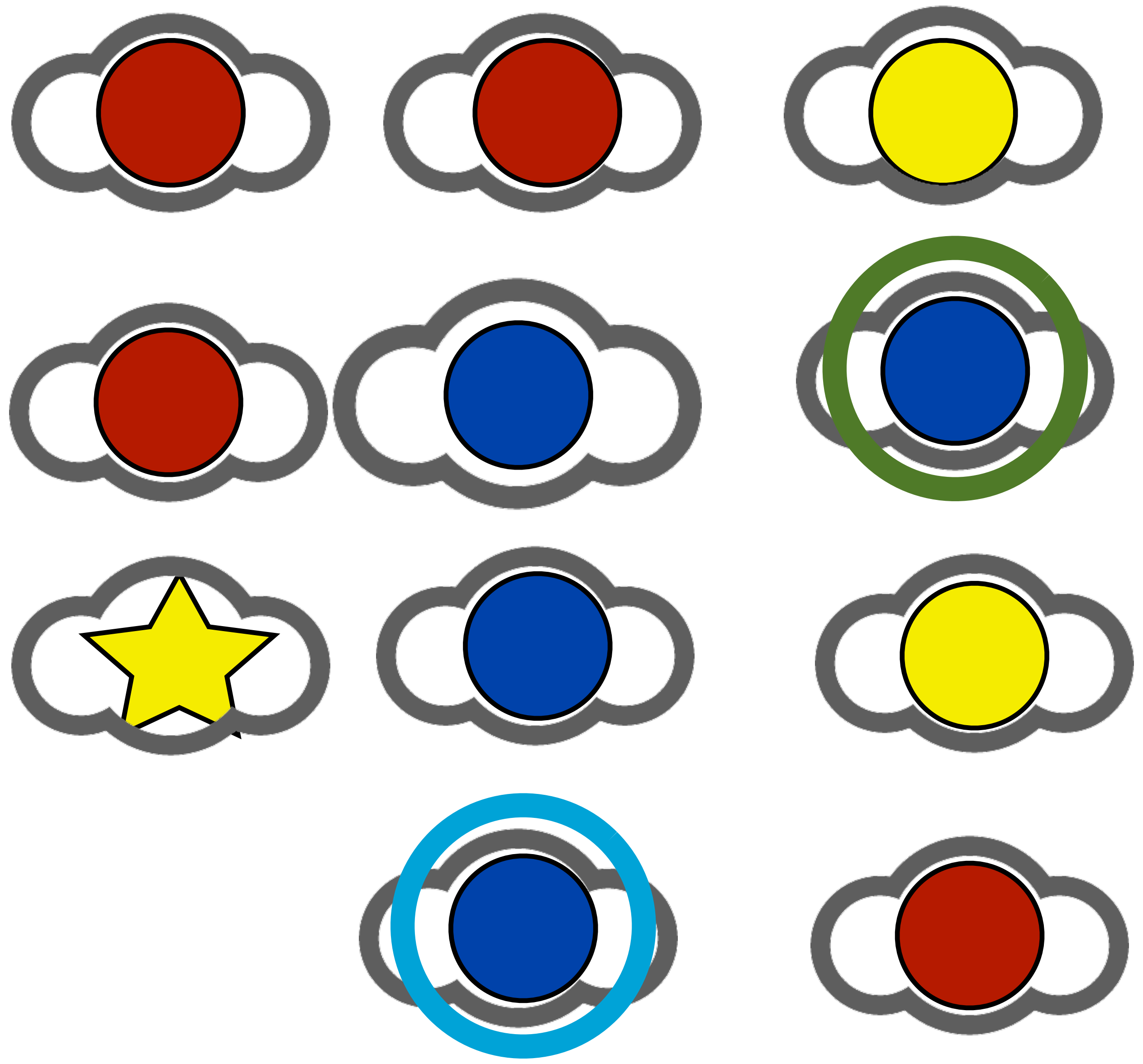




Entity Resolution Algorithm

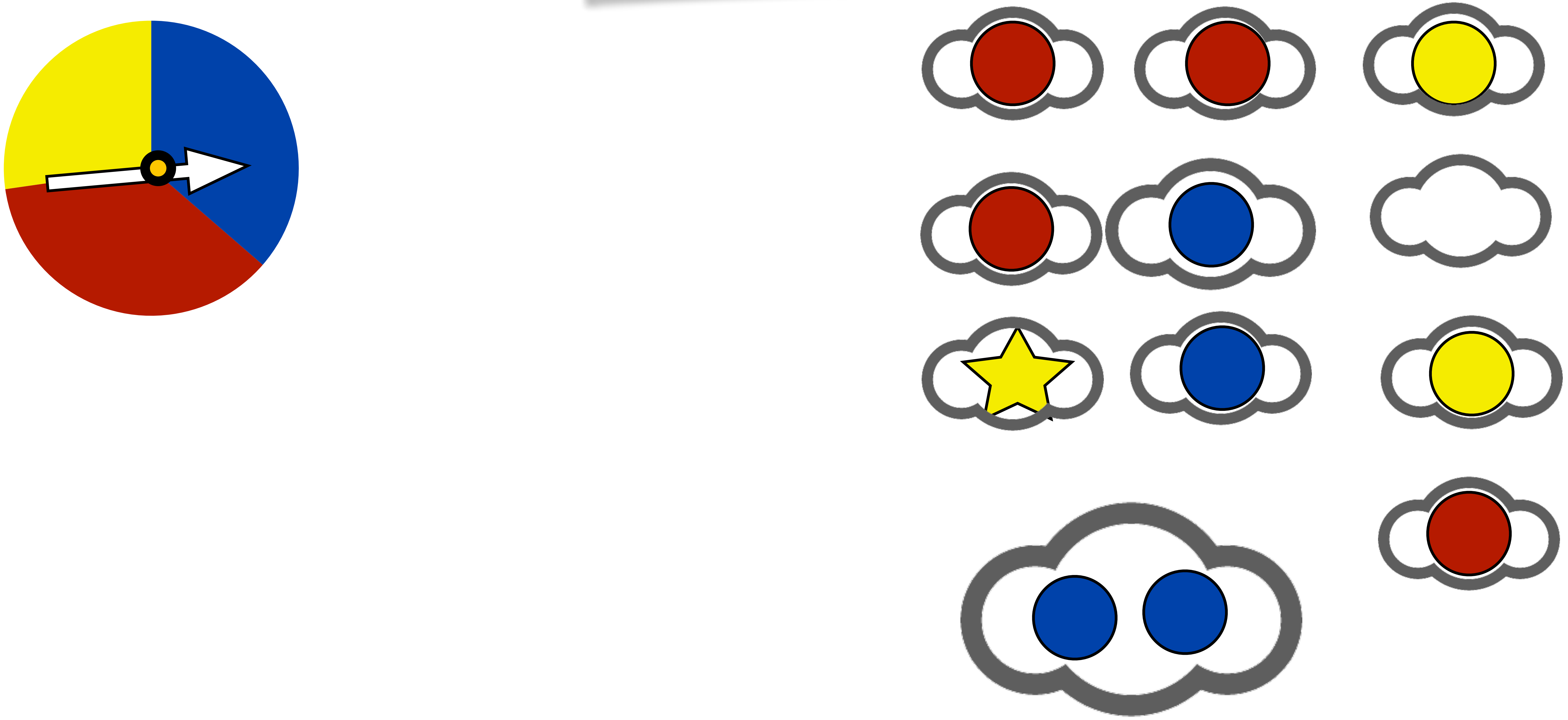
1. Select a source mention at *random*.
2. Select a destination mention at *random*.
3. Propose a merge.





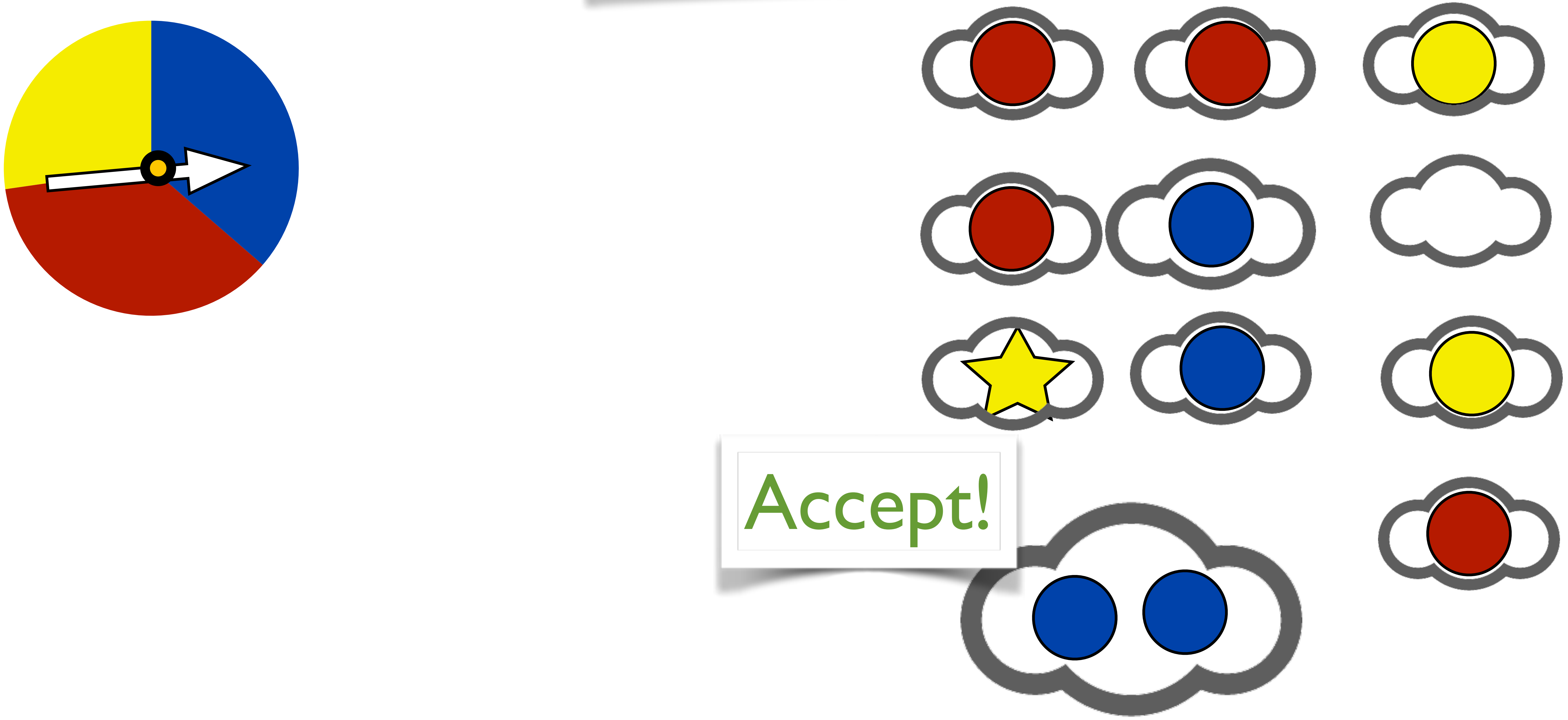
Entity Resolution Algorithm

1. Select a source mention at *random*.
2. Select a destination mention at *random*.
3. Propose a merge.
4. Accept when it improves the state.



Entity Resolution Algorithm

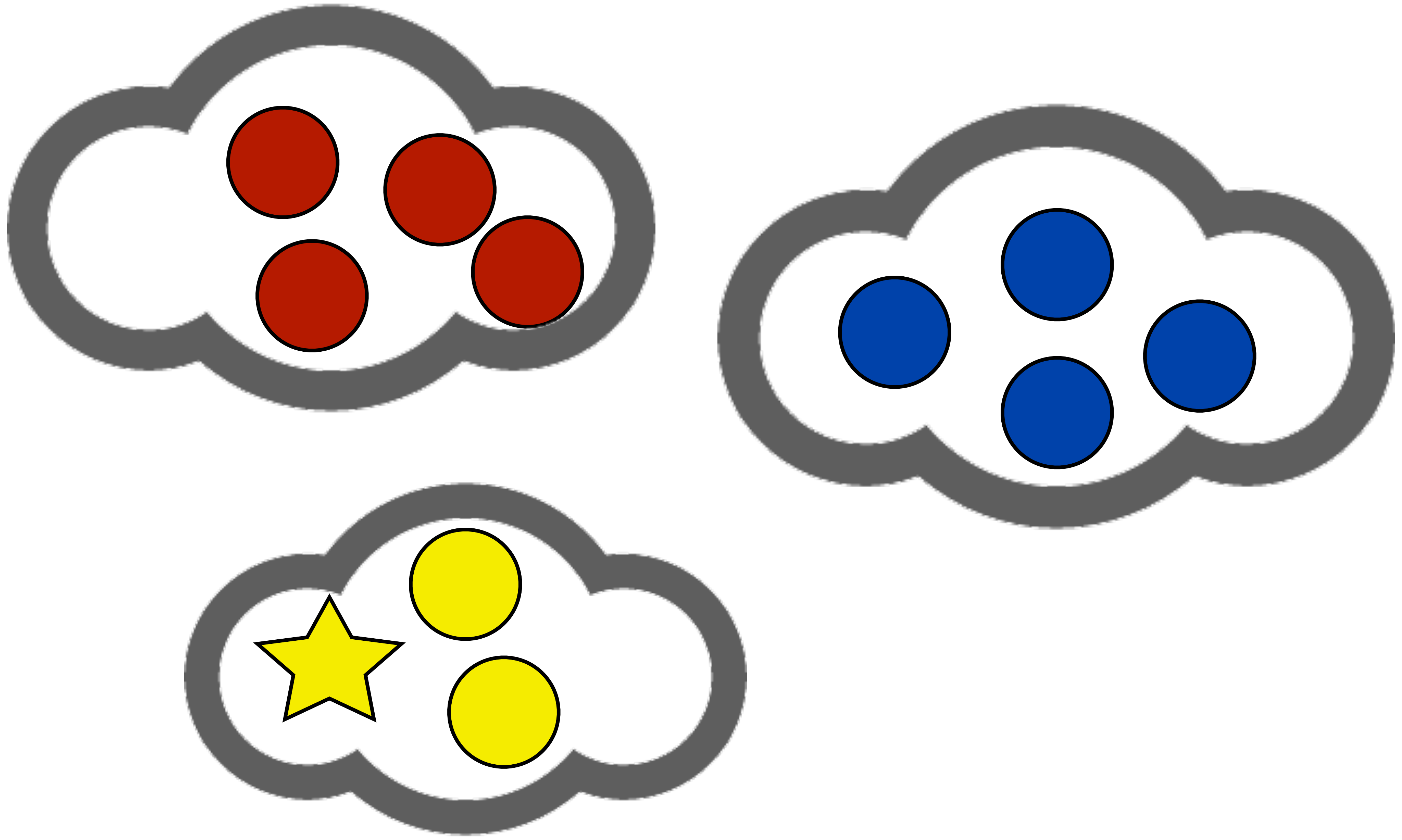
1. Select a source mention at *random*.
2. Select a destination mention at *random*.
3. Propose a merge.
4. Accept when it improves the state.



Accept!

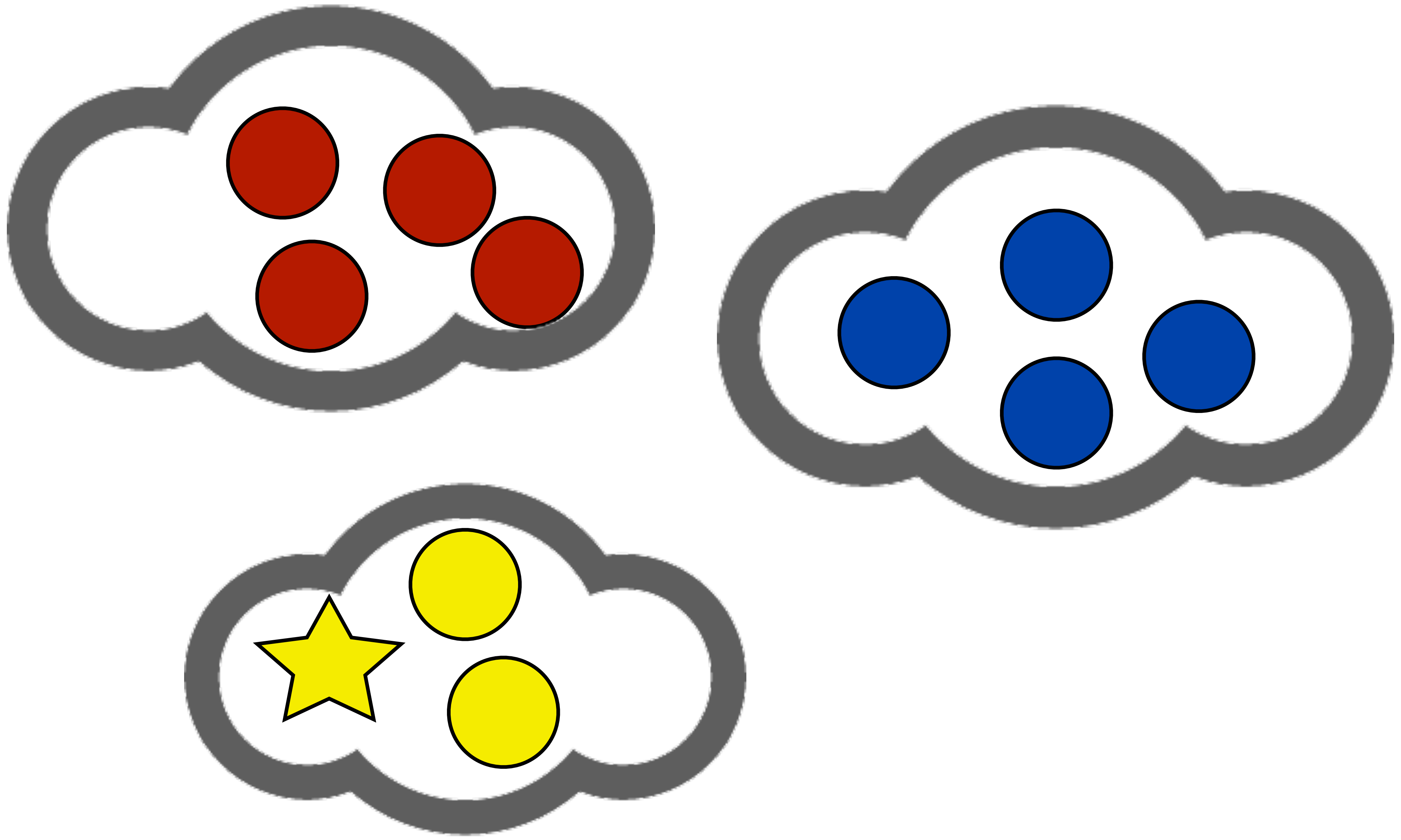
Entity Resolution Algorithm

Eventually **converges**. (State does not oscillate or vary)



Entity Resolution Algorithm

Eventually **converges**. (State does not oscillate or vary)



Markov Chain Monte Carlo Metropolis Hastings!

Sampling Optimizations

Distributed Computations (Singh et al. 2011)

Sampling Optimizations

Distributed Computations (Singh et al. 2011)

Query-Driven Computation (Grant et al. 2015)

Sampling Inefficiencies

Sampling Inefficiencies

1. Large clusters are the slowest.

Sampling Inefficiencies

1. Large clusters are the slowest.
    - Pairwise comparisons are expensive.

Sampling Inefficiencies

1. Large clusters are the slowest.
    - Pairwise comparisons are expensive. $\Theta(n^2)$

Sampling Inefficiencies

1. Large clusters are the slowest.

   Pairwise comparisons are expensive. $\Theta(n^2)$

2. Excessive computation on unambiguous entities

Sampling Inefficiencies

1. Large clusters are the slowest.

   Pairwise comparisons are expensive. $\Theta(n^2)$

2. Excessive computation on unambiguous entities

   Entities such as *Carnegie Mellon* are relatively unambiguous.

Sampling Inefficiencies

1. Large clusters are the slowest.

   Pairwise comparisons are expensive. $\Theta(n^2)$

2. Excessive computation on unambiguous entities

   Entities such as *Carnegie Mellon* are relatively unambiguous.

Streaming documents exacerbates these problems.

Optimizer for MCMC Sampling ☆

Database style optimizer for streaming MCMC.

Optimizer for MCMC Sampling ☆

Database style optimizer for streaming MCMC.

This optimizer makes two decisions:

Optimizer for MCMC Sampling

Database style optimizer for streaming MCMC.

This optimizer makes two decisions:

1.Can I approximate the state score calculation?

Optimizer for MCMC Sampling

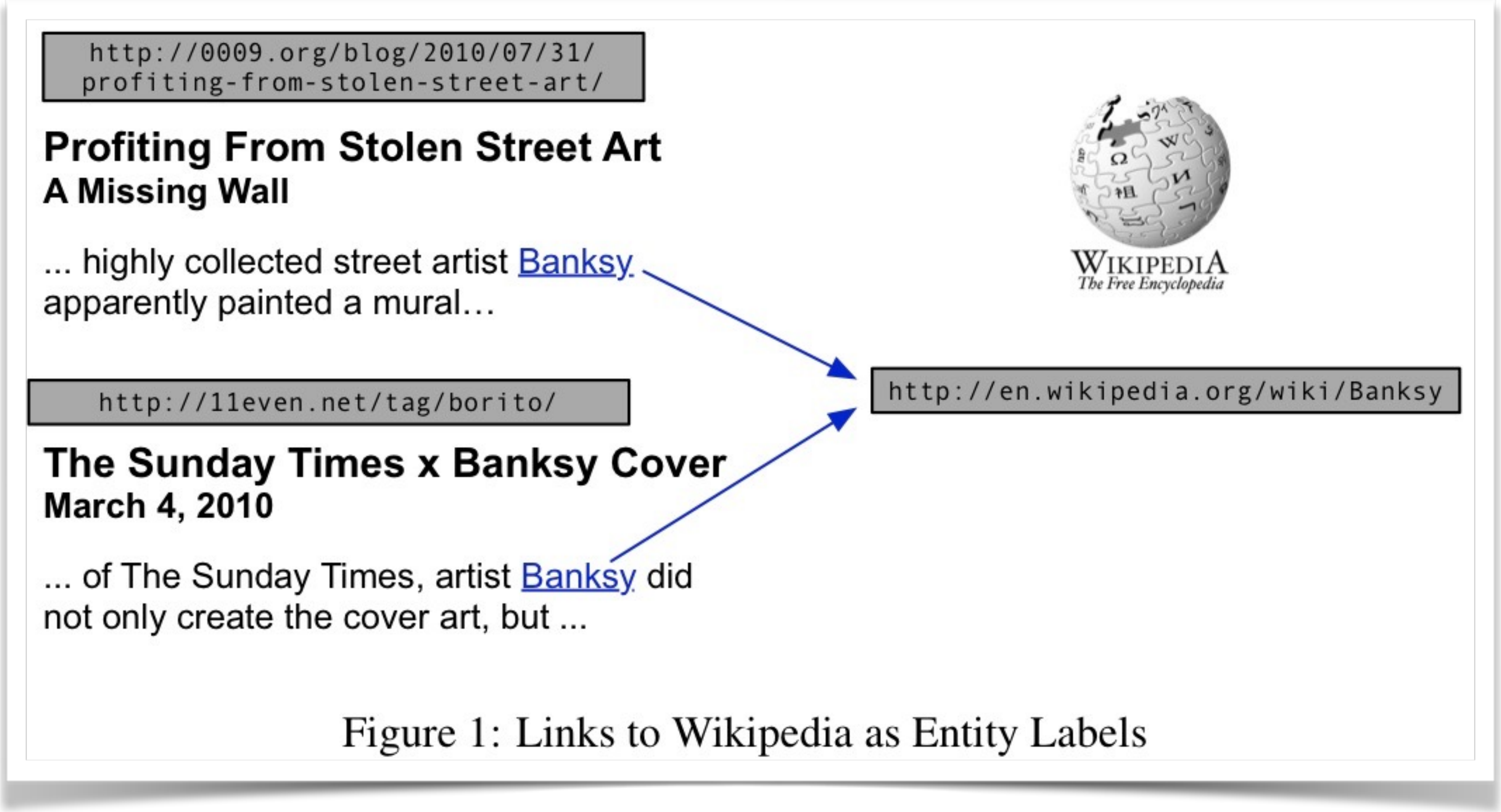
Database style optimizer for streaming MCMC.

This optimizer makes two decisions:

1. Can I approximate the state score calculation?
2. Should I compress an Entity?

Experiments

- **Wikilink Data Set** *(Singh, Subramaniya, Pereira, McCallum, 2011)*
  - Largest fully-labeled data set
  - 40 Million Mentions
  - 180 GBs of data



http://0009.org/blog/2010/07/31/profiting-from-stolen-street-art/

Profiting From Stolen Street Art
A Missing Wall

... highly collected street artist Banksy apparently painted a mural…

http://11even.net/tag/borito/

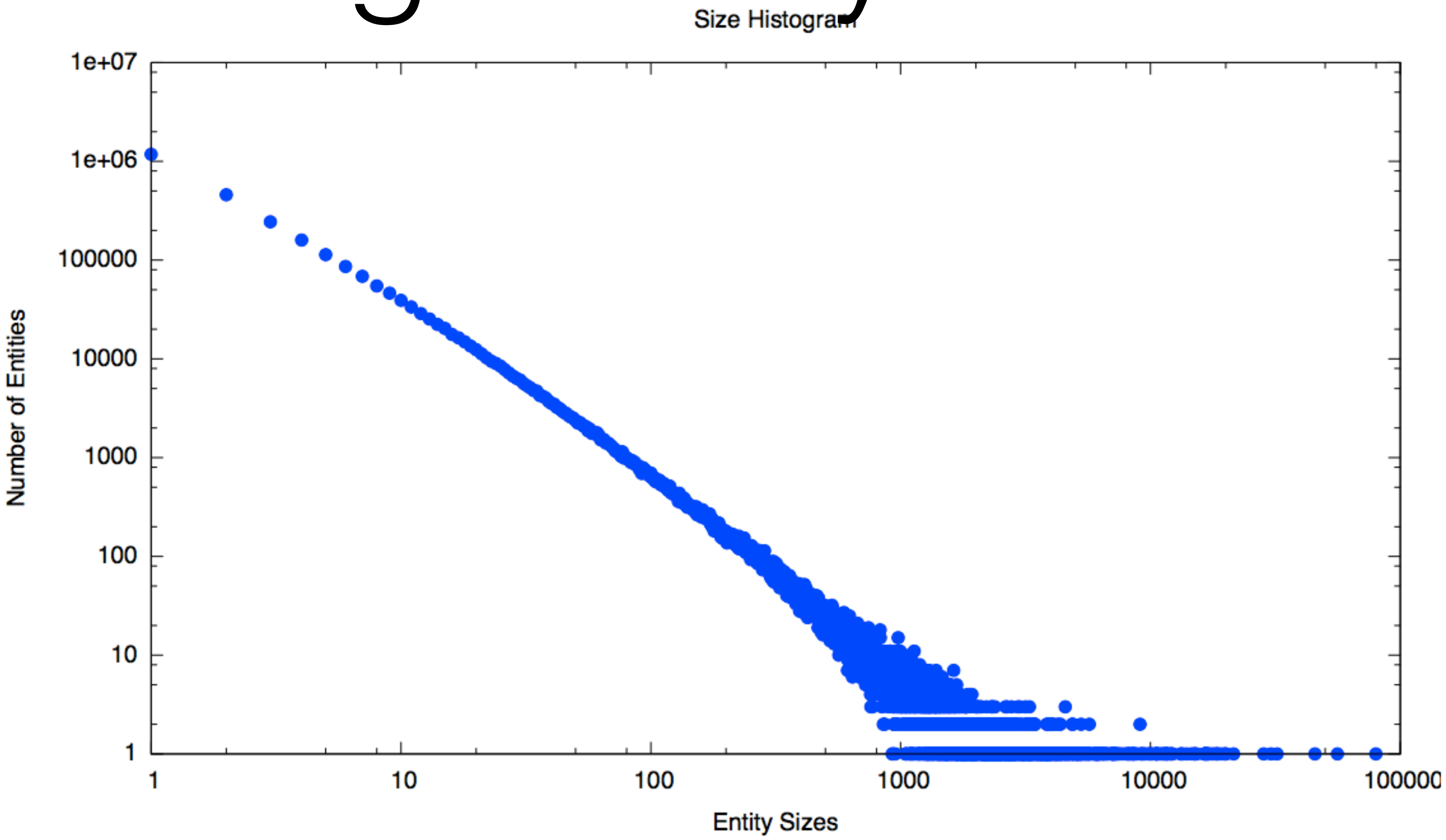
The Sunday Times x Banksy Cover
March 4, 2010

... of The Sunday Times, artist Banksy did not only create the cover art, but ...

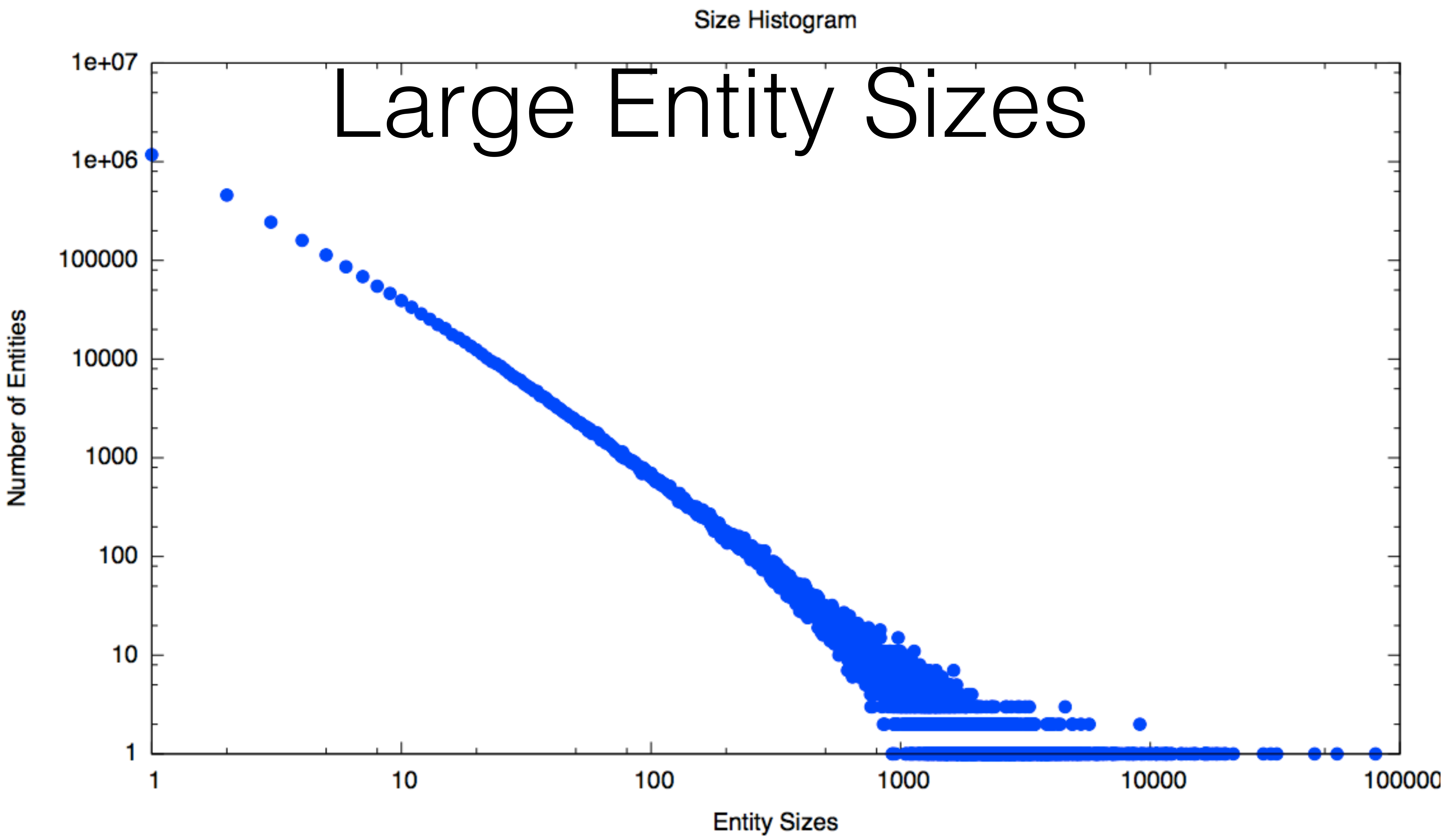
http://en.wikipedia.org/wiki/Banksy

Figure 1: Links to Wikipedia as Entity Labels

Large Entity Sizes



Large Entity Sizes



Entity Compression

Entity Compression

- Known matches can be compressed into a representative mention.

Entity Compression

- Known matches can be compressed into a representative mention.
- Entity compression can reduce the number of mentions (***n***).

Entity Compression

- Known matches can be compressed into a representative mention.
- Entity compression can reduce the number of mentions (***n***).
- Compression of large and *popular* entities is **costly**.

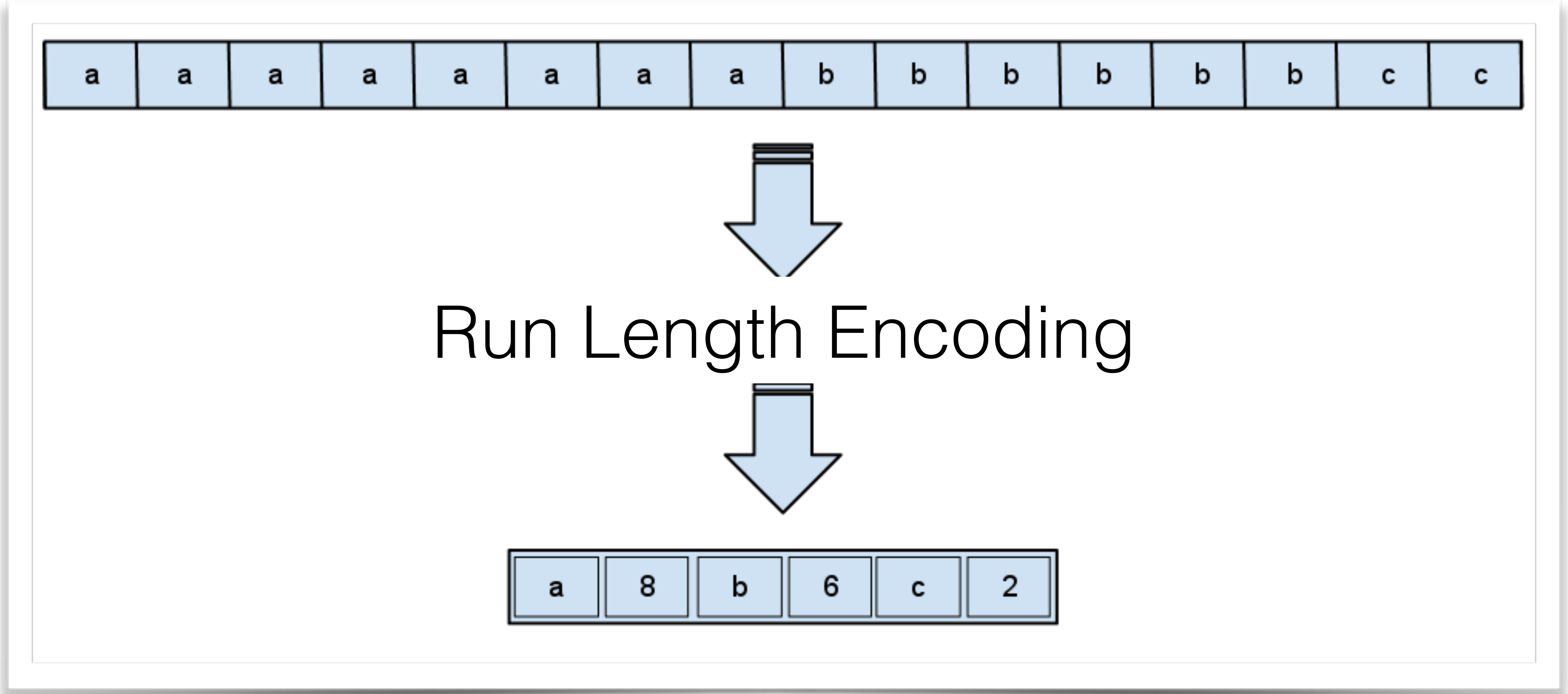
Entity Compression

- Known matches can be compressed into a representative mention.
- Entity compression can reduce the number of mentions (***n***).
- Compression of large and *popular* entities is **costly**.
- Compression errors are permanent.

Compression Types

- Run-Length Encoding

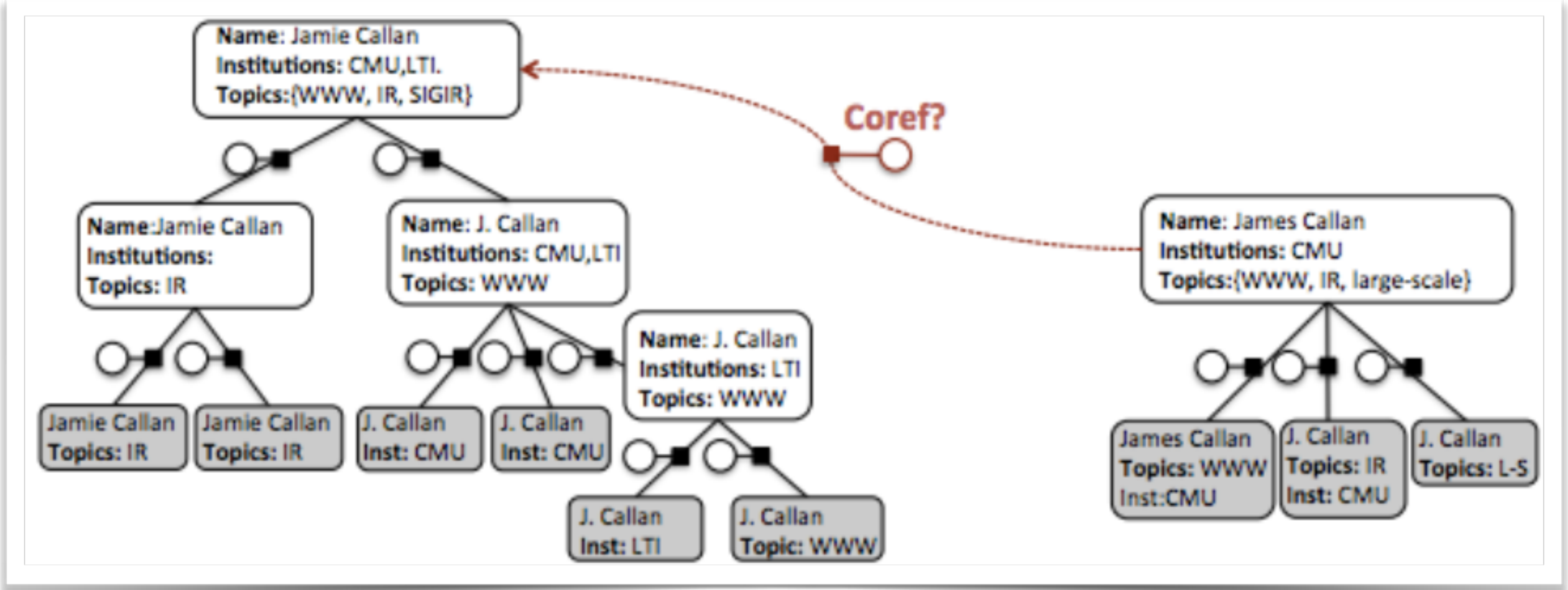
- Hierarchical Compression (Wick et al.)



a	a	a	a	a	a	a	a	b	b	b	b	b	b	c	c

Run Length Encoding

a	8	b	6	c	2



Name: Jamie Callan
Institutions: CMU,LTI.
Topics:{WWW, IR, SIGIR}

Name:Jamie Callan
Institutions:
Topics: IR

Name: J. Callan
Institutions: CMU,LTI
Topics: WWW

Name: J. Callan
Institutions: LTI
Topics: WWW

Jamie Callan
Topics: IR

Jamie Callan
Topics: IR

J. Callan
Inst: CMU

J. Callan
Inst: CMU

J. Callan
Inst: LTI

J. Callan
Topic: WWW

Coref?

Name: James Callan
Institutions: CMU
Topics:{WWW, IR, large-scale}

James Callan
Topics: WWW
Inst:CMU

J. Callan
Topics: IR
Inst: CMU

J. Callan
Topics: L-S

Early Stopping

- Can we estimate the computation of the features?

Singh et al. EMNLP'12

Early Stopping

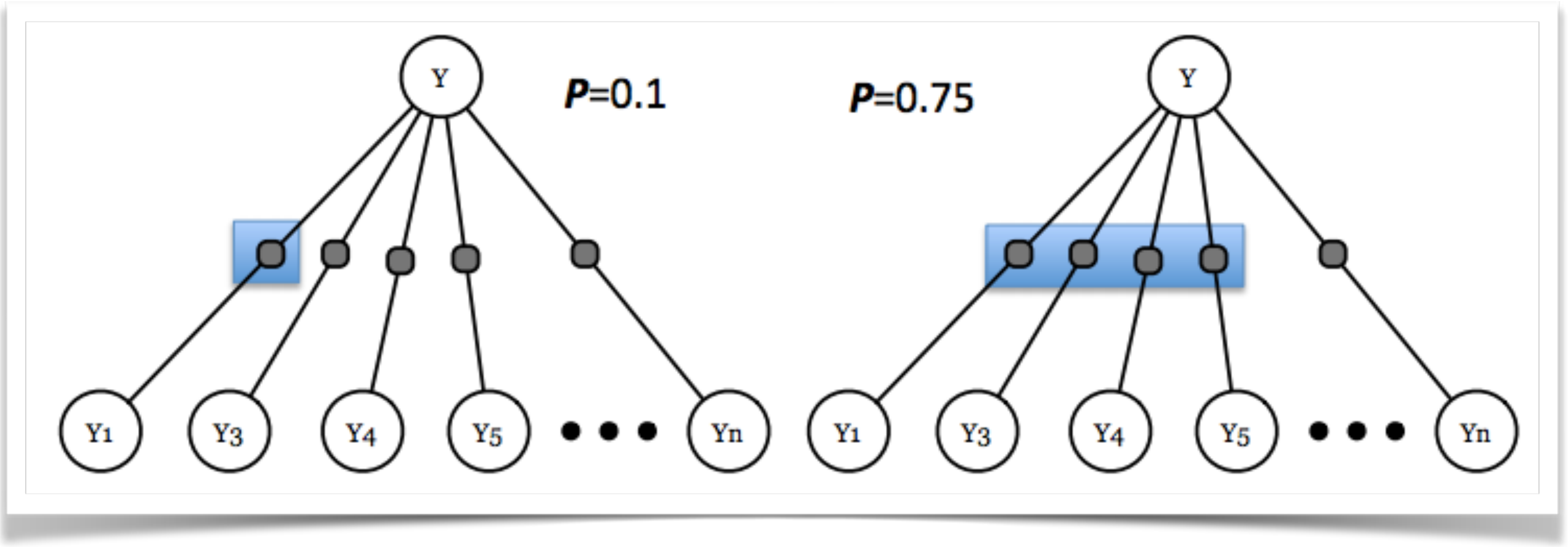
- Can we estimate the computation of the features?

- Given a ***p*** value, randomly select less values.

Singh et al. EMNLP’12

Early Stopping

- Can we estimate the computation of the features?
- Given a ***p*** value, randomly select less values.



Singh et al. EMNLP'12

Optimizer

Current work

1. Classifier for deciding when to perform *early stopping*.
2. Classifier for the decision to *compress*.

When should it compress?

When should it compress?

Power law says there are only a small number of very large clusters.

When should it compress?

Power law says there are only a small number of very large clusters.

We can treat these in a special way.

When should it compress?

Power law says there are only a small number of very large clusters.

We can treat these in a special way.

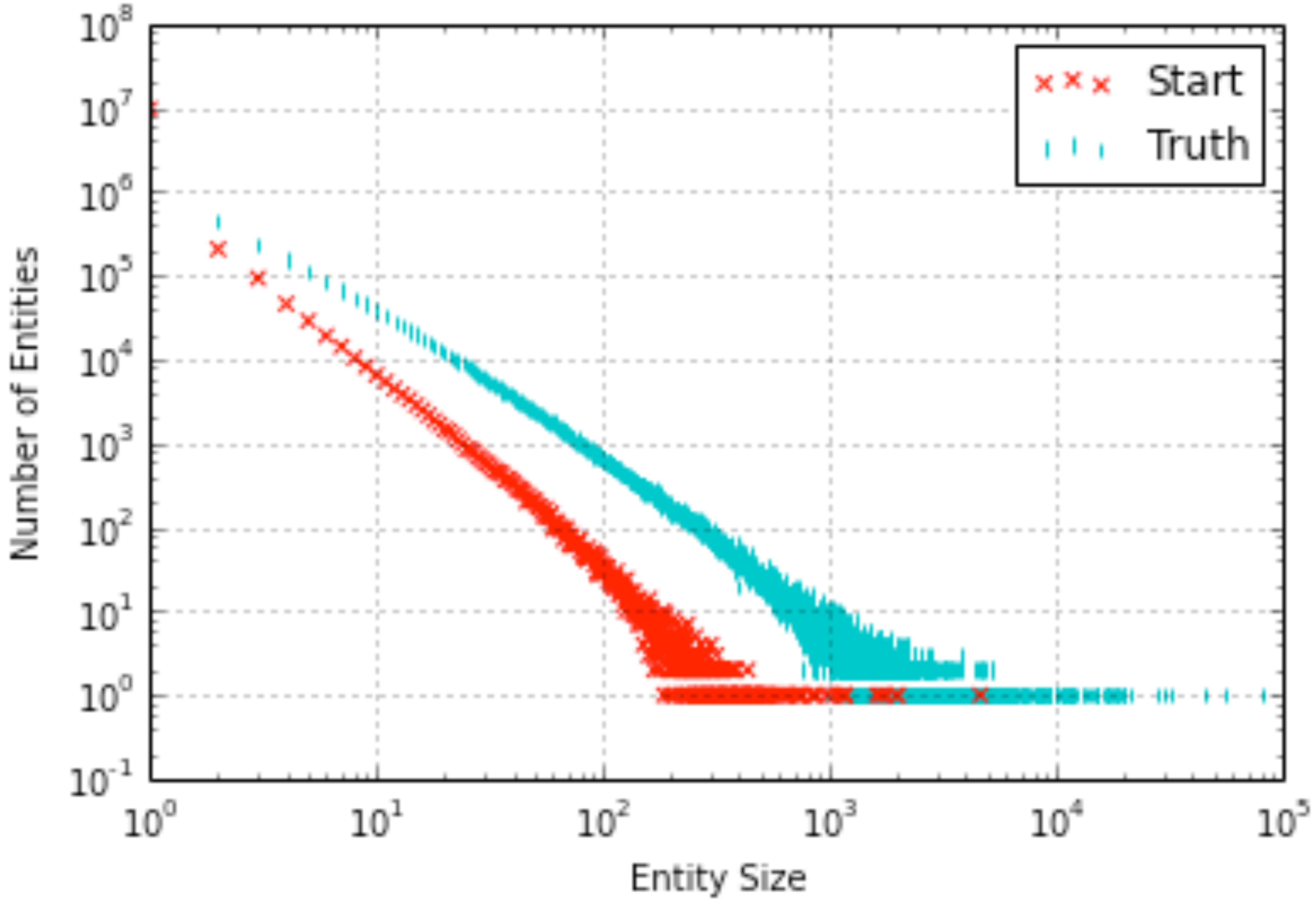
Examining the Wiki Link data set.

When should it compress?

Power law says there are only a small number of very large clusters.

We can treat these in a special way.

Examining the Wiki Link data set.



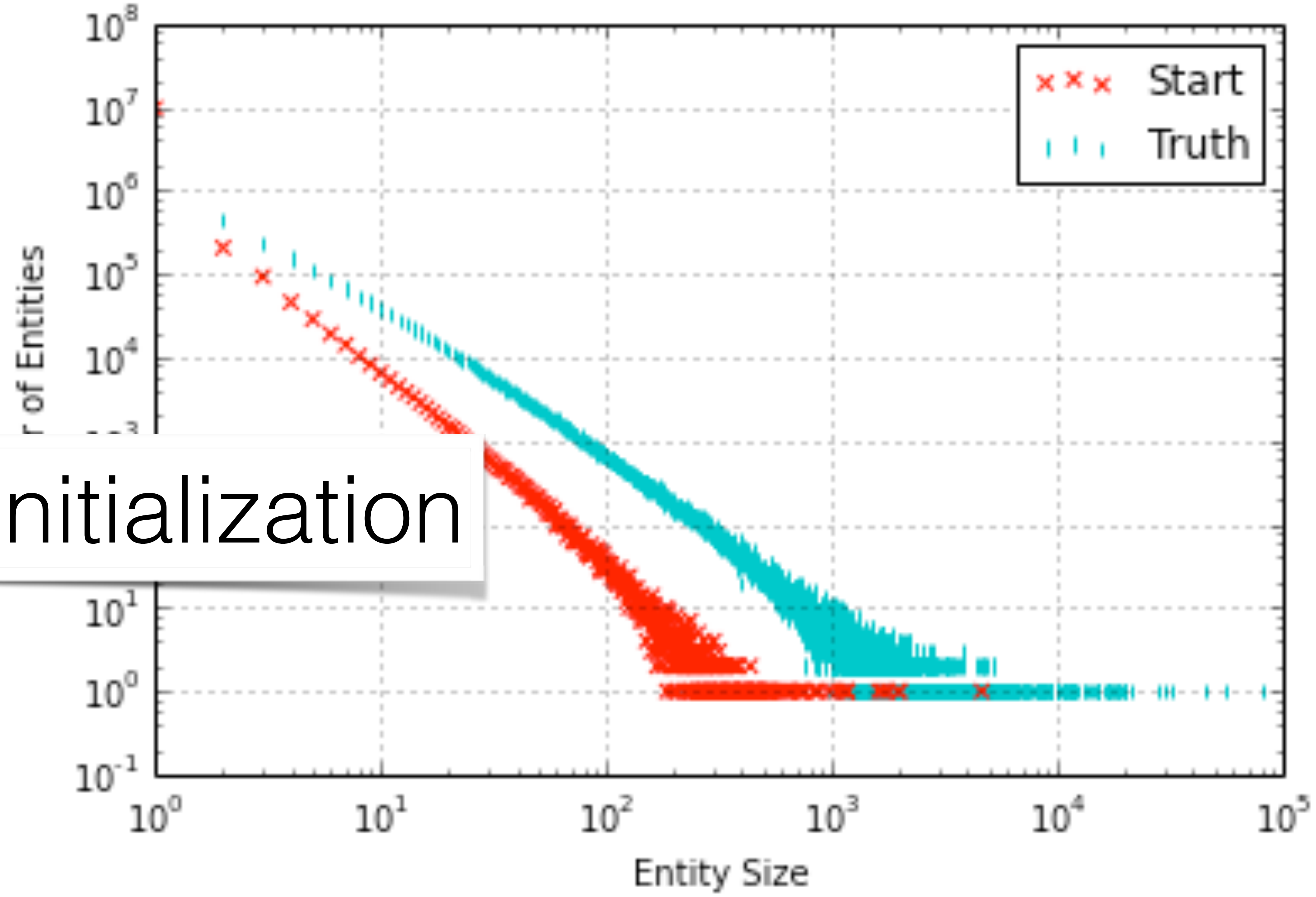
When should it compress?

Power law says there are only a small number of very large clusters.

We can treat these in a special way.

Examini
set.

Exact String Match Initialization



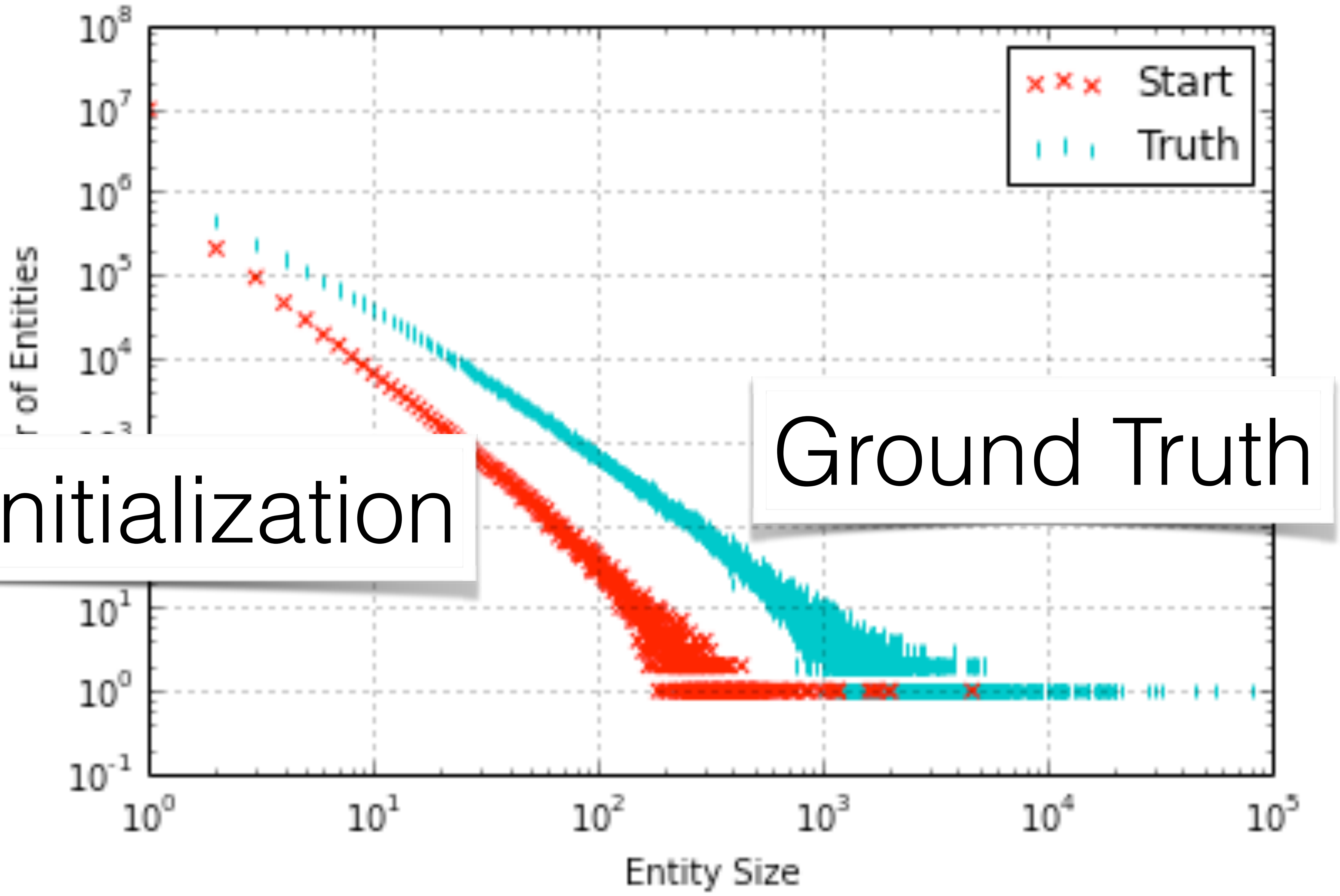
When should it compress?

Power law says there are only a small number of very large clusters.

We can treat these in a special way.

Examini

set.



Exact String Match Initialization

Ground Truth

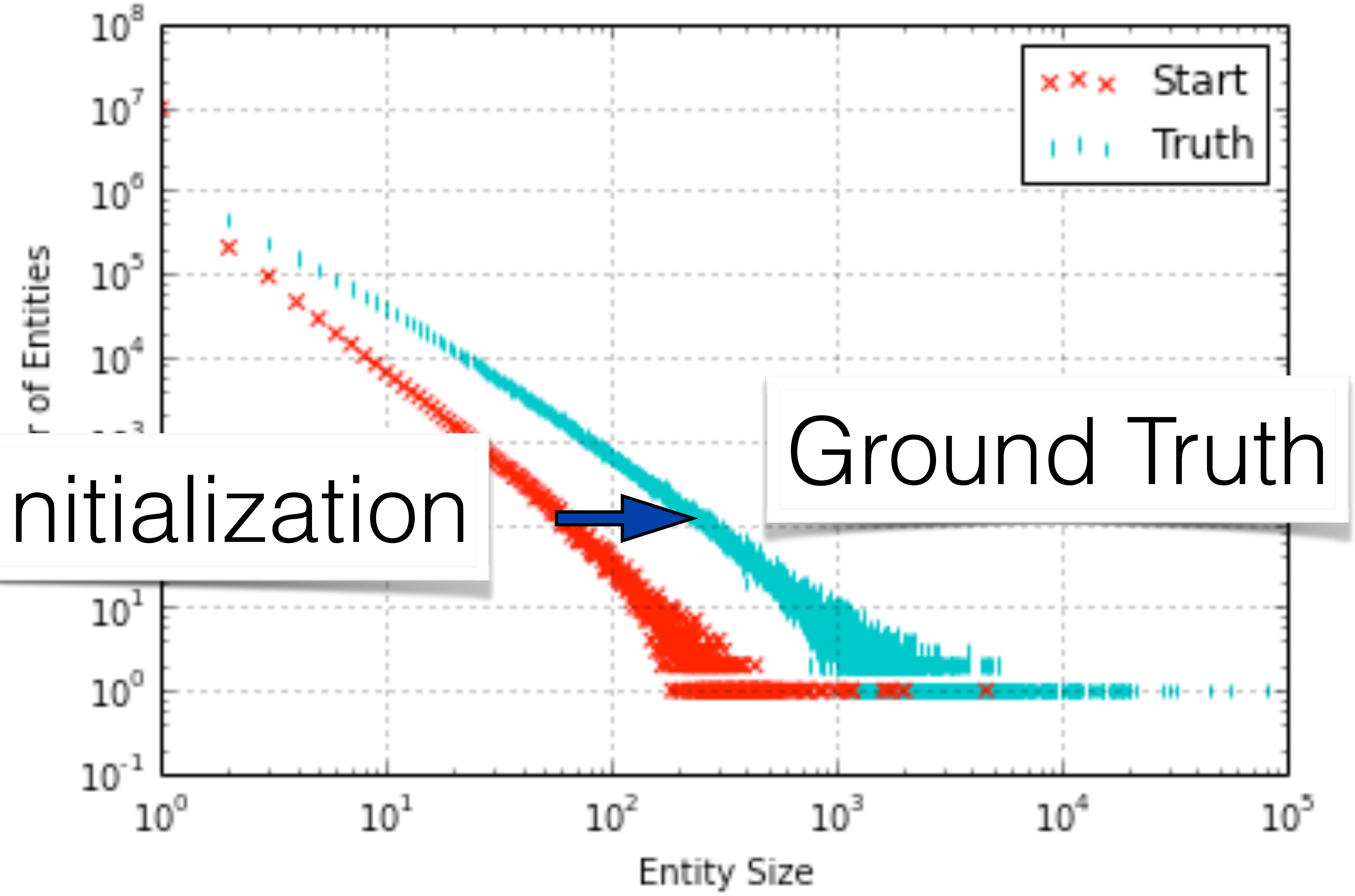
When should it compress?

Power law says there are only a small number of very large clusters.

We can treat these in a special way.

Examini

set.



Exact String Match Initialization

Ground Truth

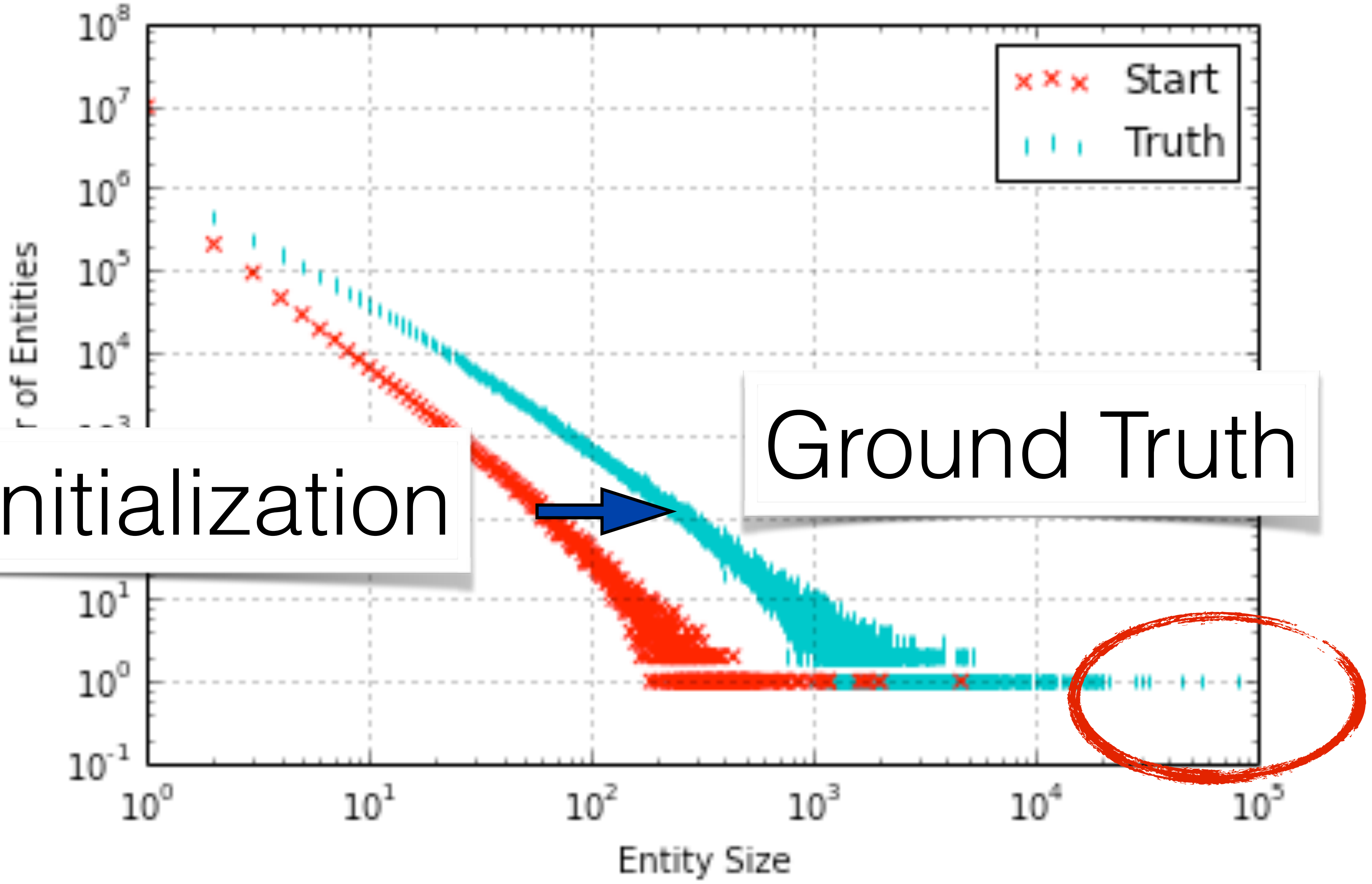
When should it compress?

Power law says there are only a small number of very large clusters.

We can treat these in a special way.

Examini

set.

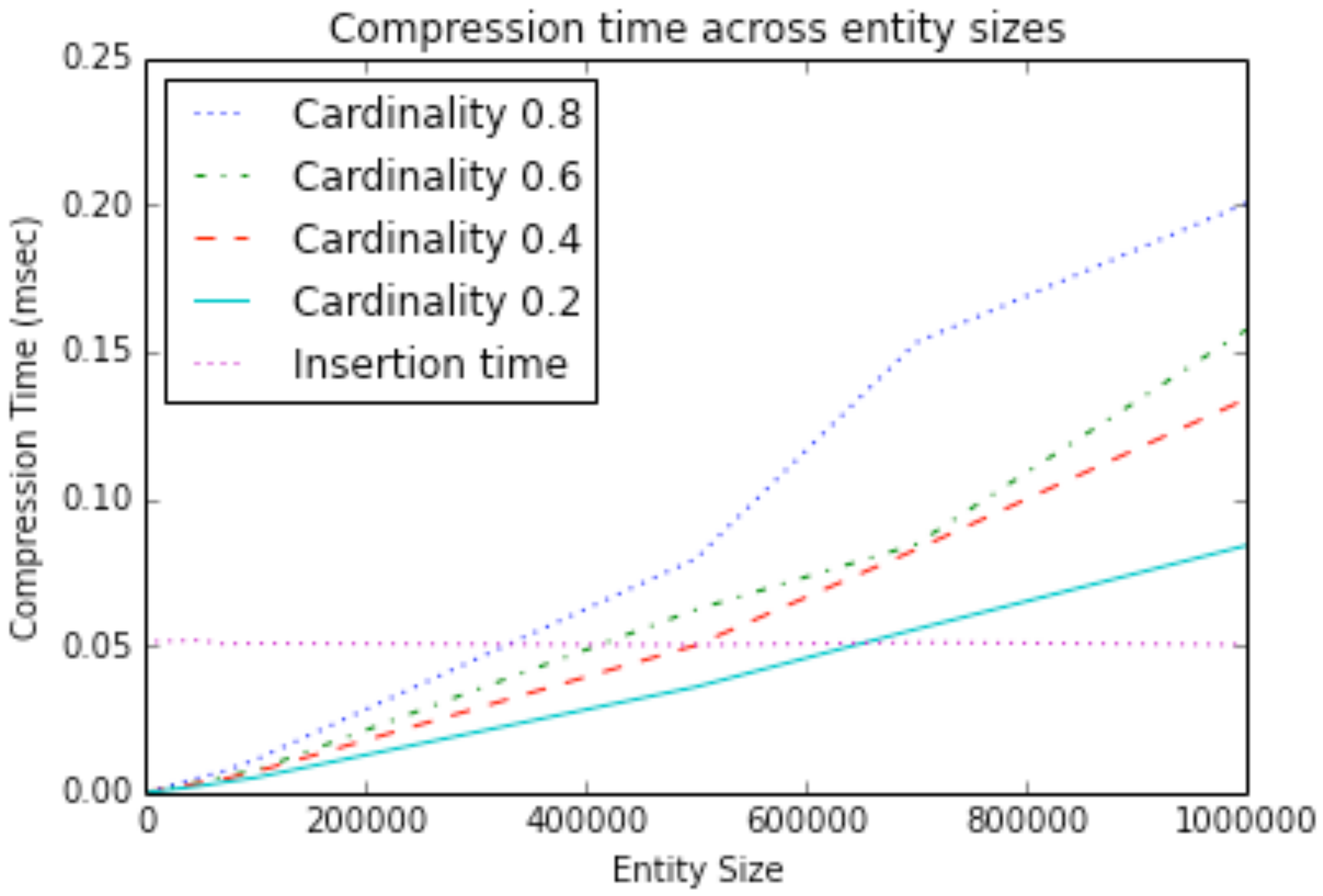


Exact String Match Initialization

Ground Truth

When should it compress?

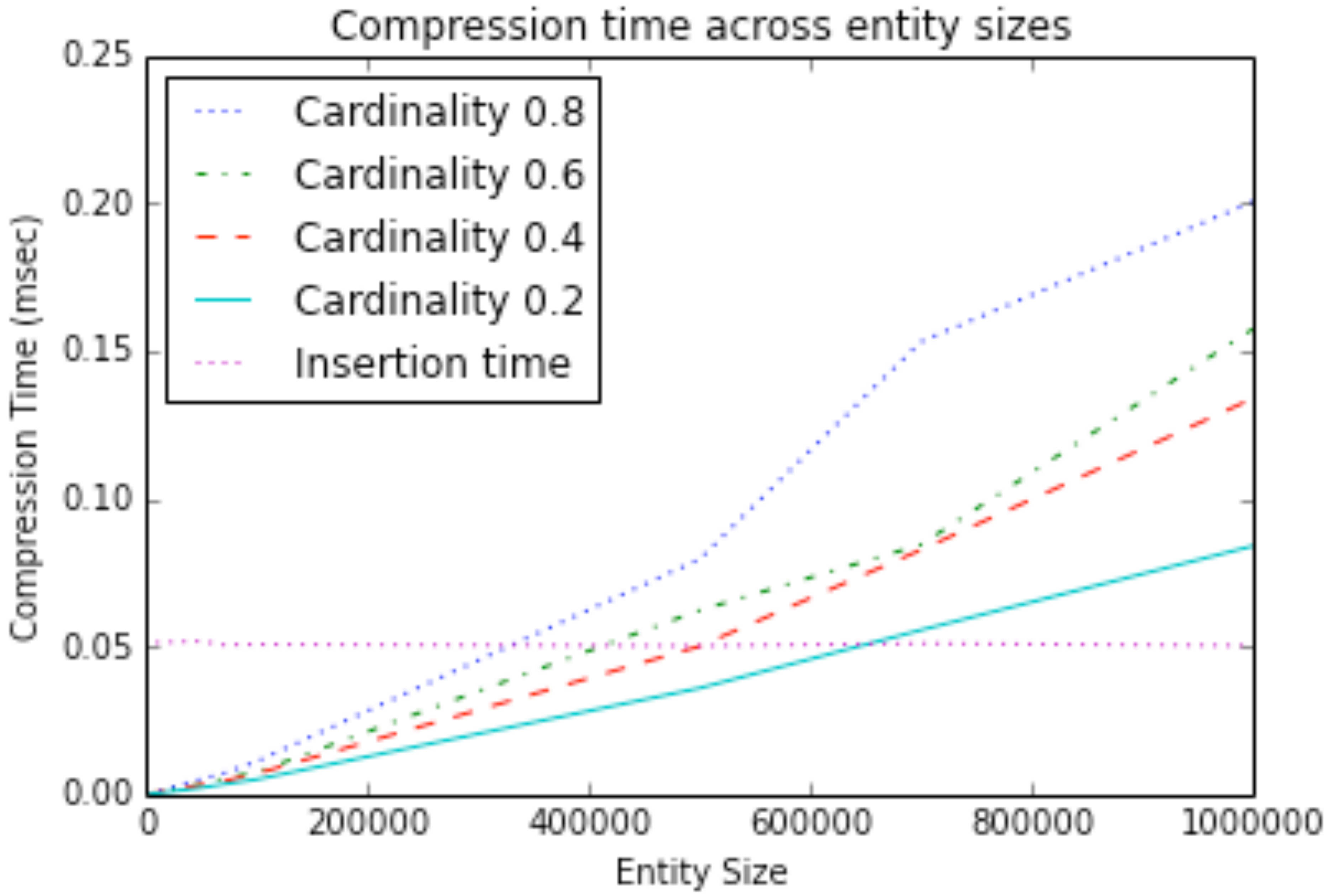
We could make 100,000 insertions in the time it take to to compress a 300K mention cluster.



When should it compress?

We could make 100,000 insertions in the time it take to to compress a 300K mention cluster.

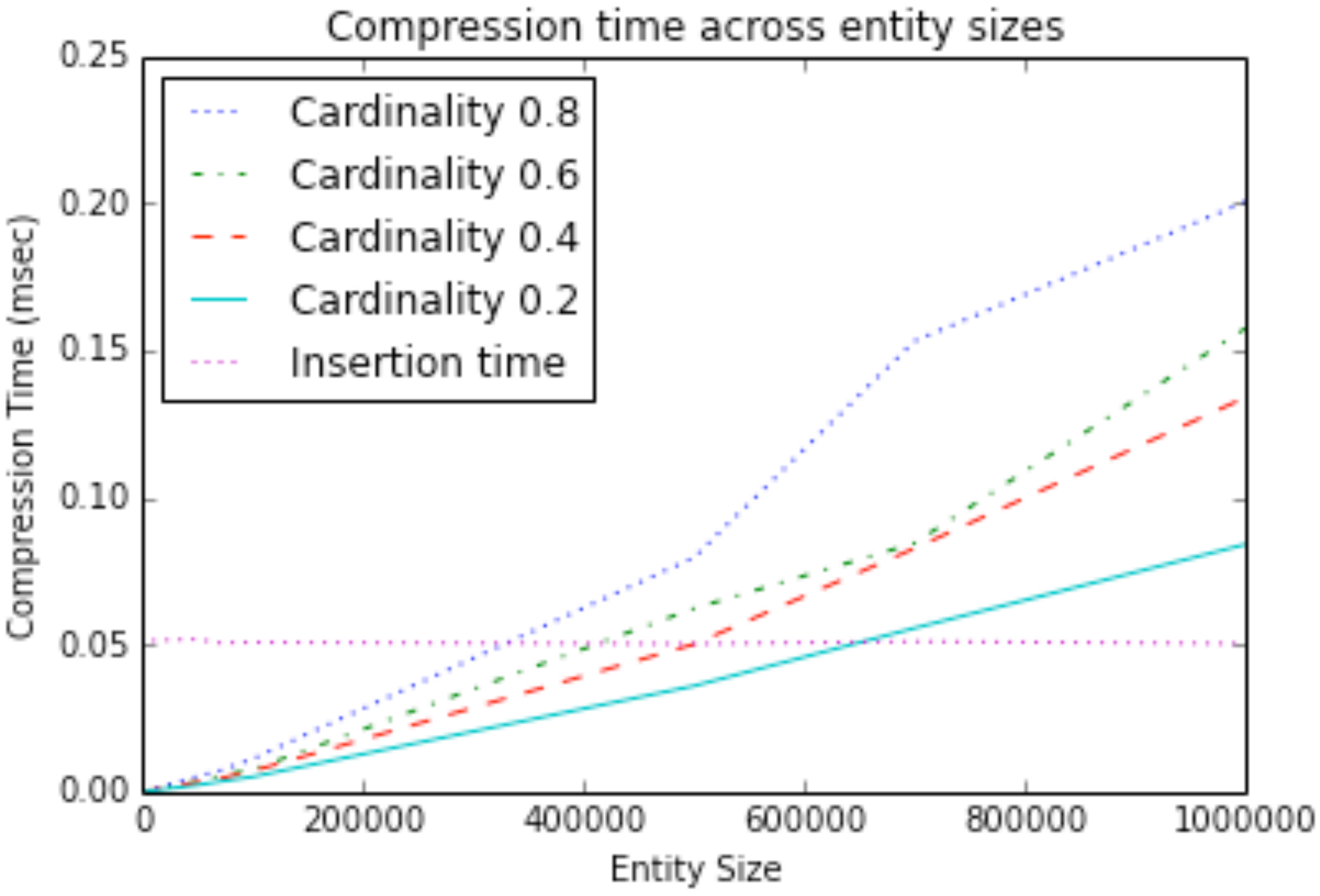
Compression must be worth it.



When should it compress?

We could make 100,000 insertions in the time it take to to compress a 300K mention cluster.

Compression must be worth it.



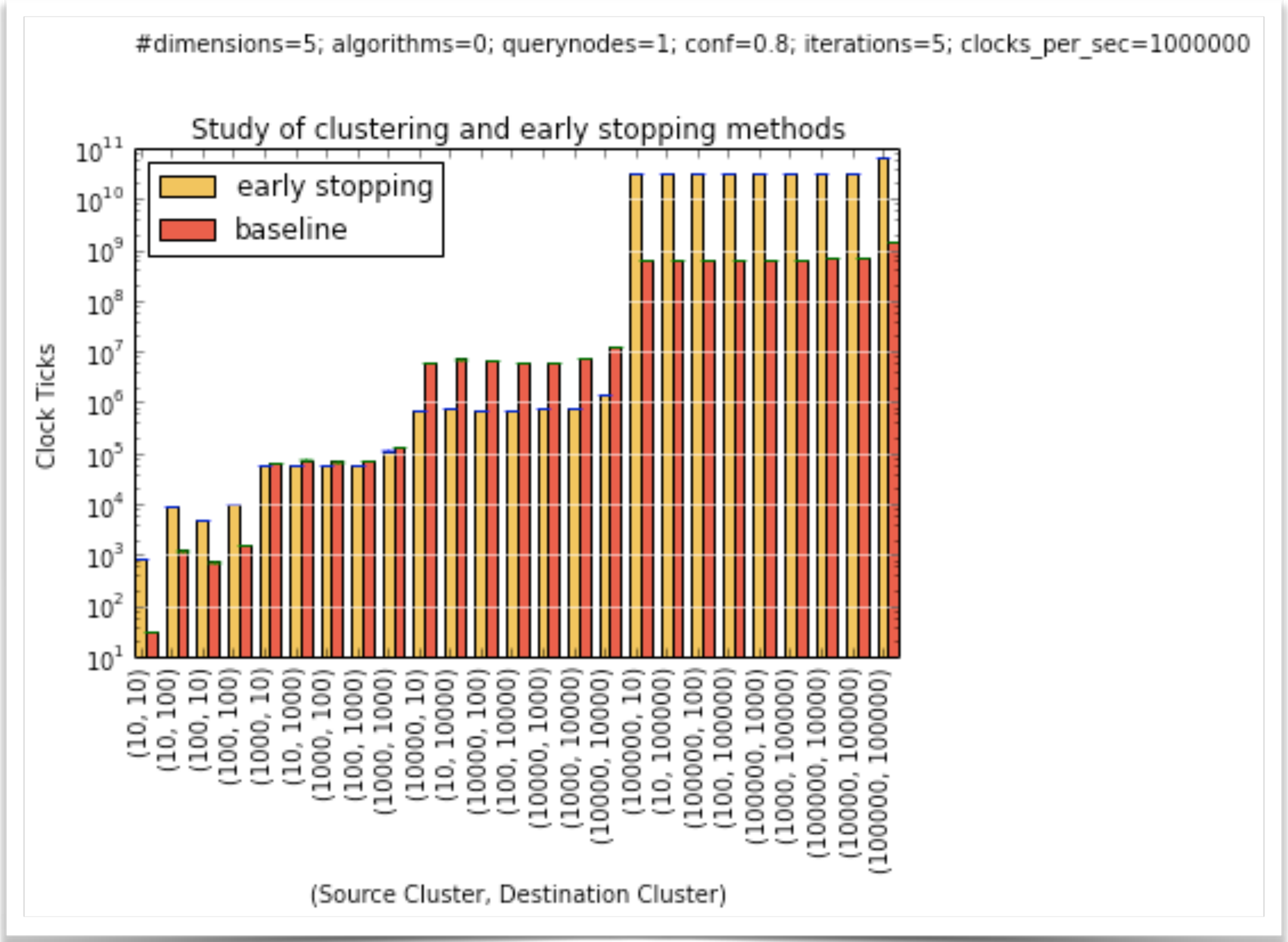
When should we approximate?

When should we approximate?

- Early stopping only makes sense for clusters of medium size.
- It is better to do full comparison for small and large cluster sizes.

When should we approximate?

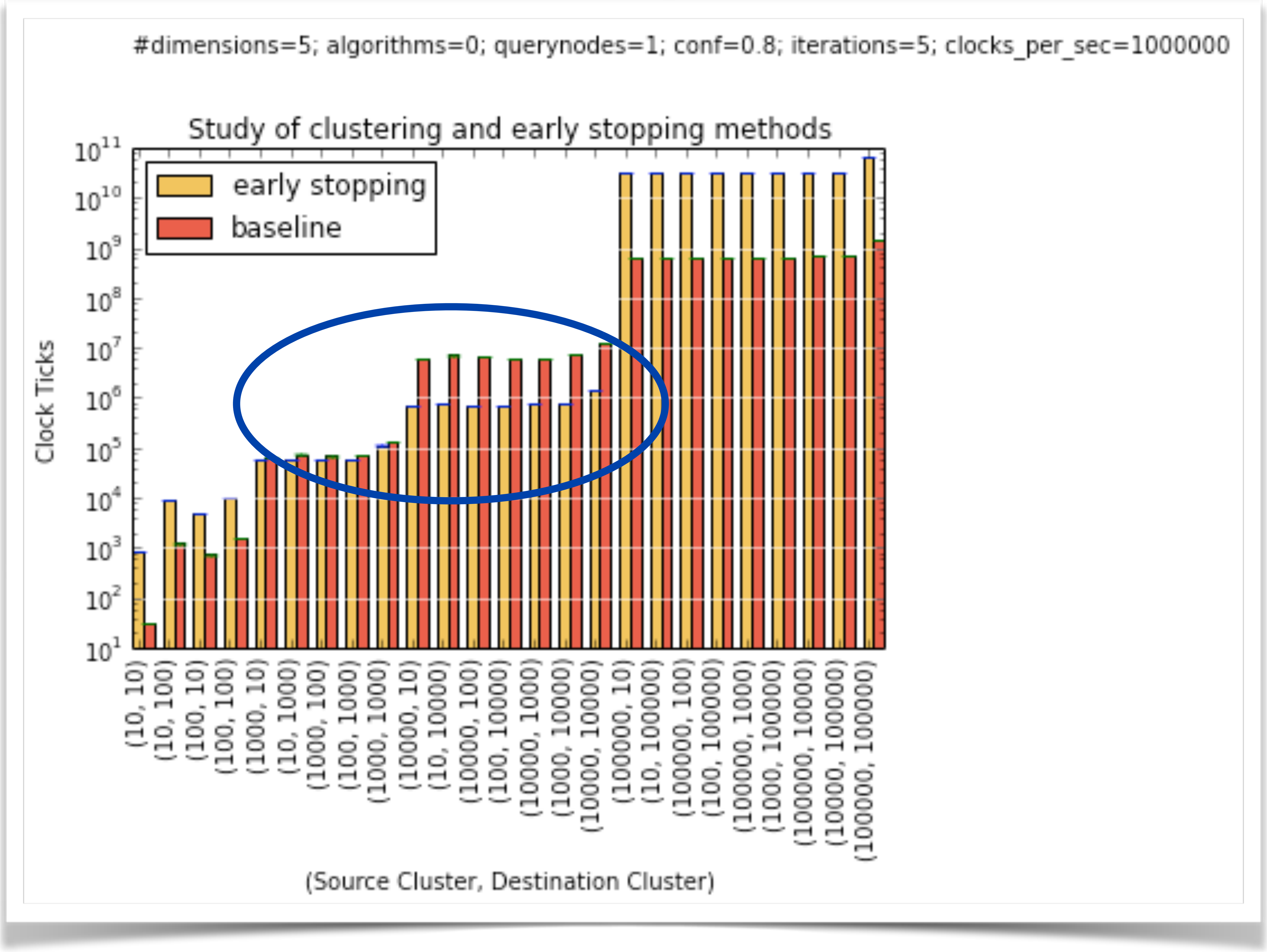
- Early stopping only makes sense for clusters of medium size.
- It is better to do full comparison for small and large cluster sizes.



When should we approximate?

- Early stopping only makes sense for clusters of medium size.

- It is better to do full comparison for small and large cluster sizes.



Optimizer for Query-Driven Sampling

Optimizer needs to know:

- Current Cardinality of Items in each entity.
- Memory/CPU configuration for estimating baseline time

```
while samples--  > 0:
    m ~ Mentions
    e ~ Entities
    state’ = move(state, m, e)
    o = Optimize(state, state’, m, e)
    if (!score(state’,state, o)):
      state = state’
    doCompress(state, m, e, o)
```

Summary

- We motivated the need and discussed the open space for optimization of MCMC sampling methods.
- We plan to use the newly released labeled TREC stream corpus.
- Want to collaborate?!
- Lets talk if you want to do a Ph.D. at the University of Oklahoma!

Thank you!

NSF

DARPA