





Data Science of Love

Vaclav Petricek @petricek

The eHarmony Difference ›

Compatibility Matching System®

The eHarmony Difference >

Compatibility Matching System®



**Compatibility Matching  
System®**



**Compatibility Matching  
System®**



**Compatibility  
Matching**



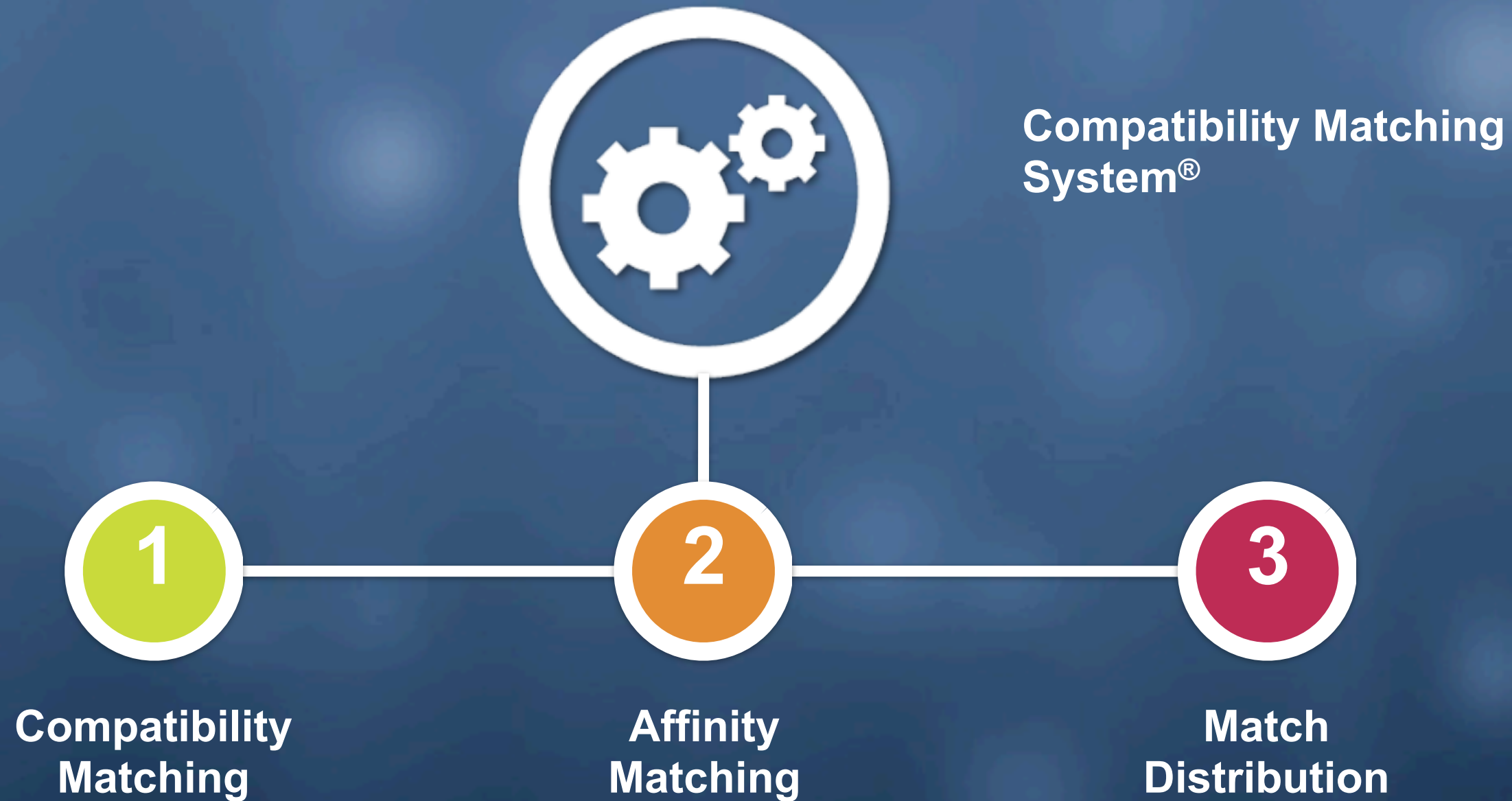
Compatibility Matching System®



Compatibility Matching



Affinity Matching



# The eHarmony Difference







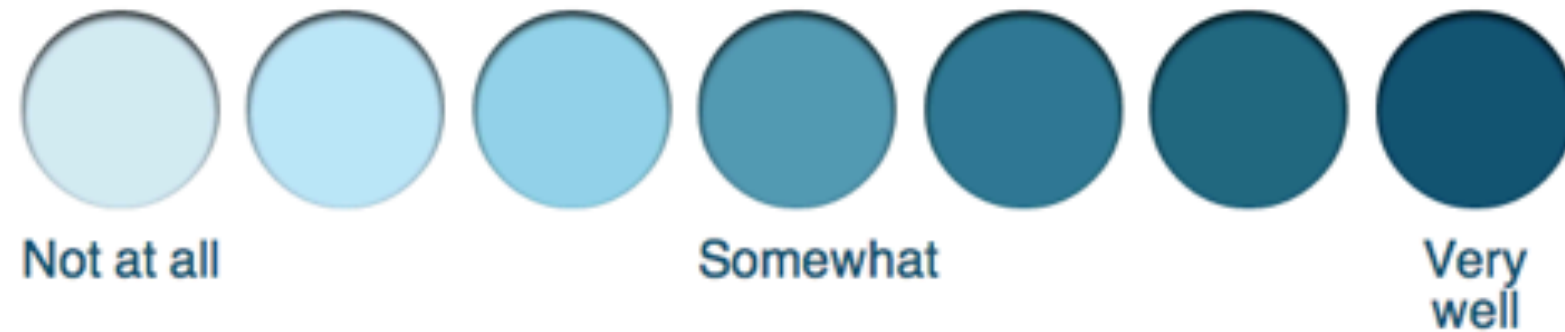




How well does each of the following describe you?

---

**I try to accommodate the other person's position**



Not at all

Somewhat

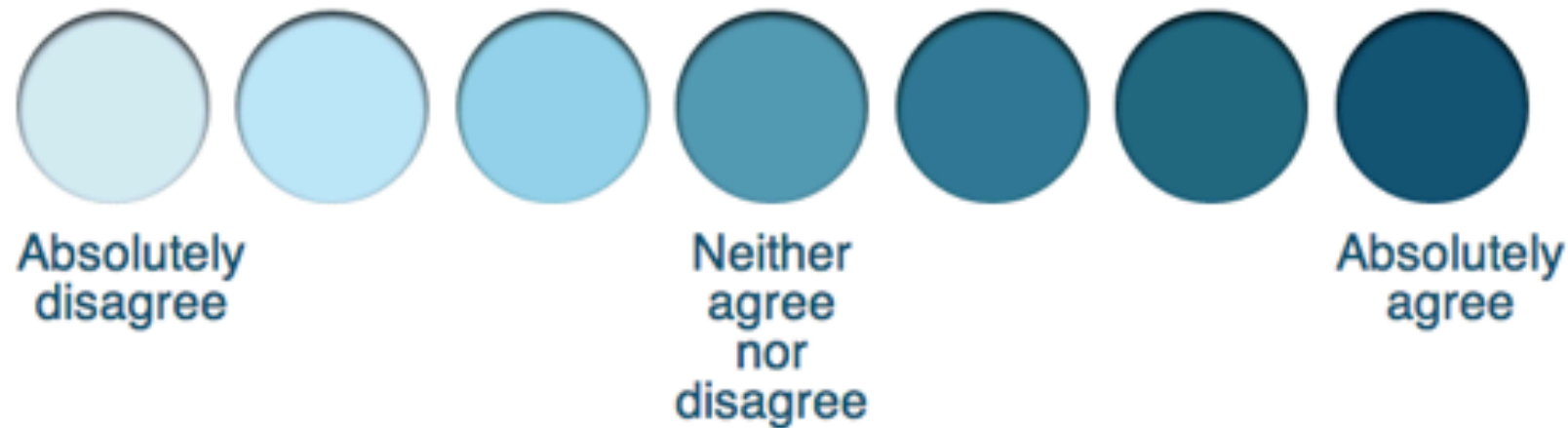
Very well



How strongly do you agree or disagree with...?

---

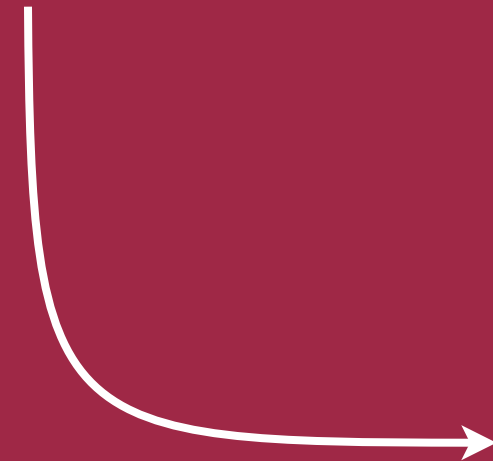
**People often let you down if you depend on them**



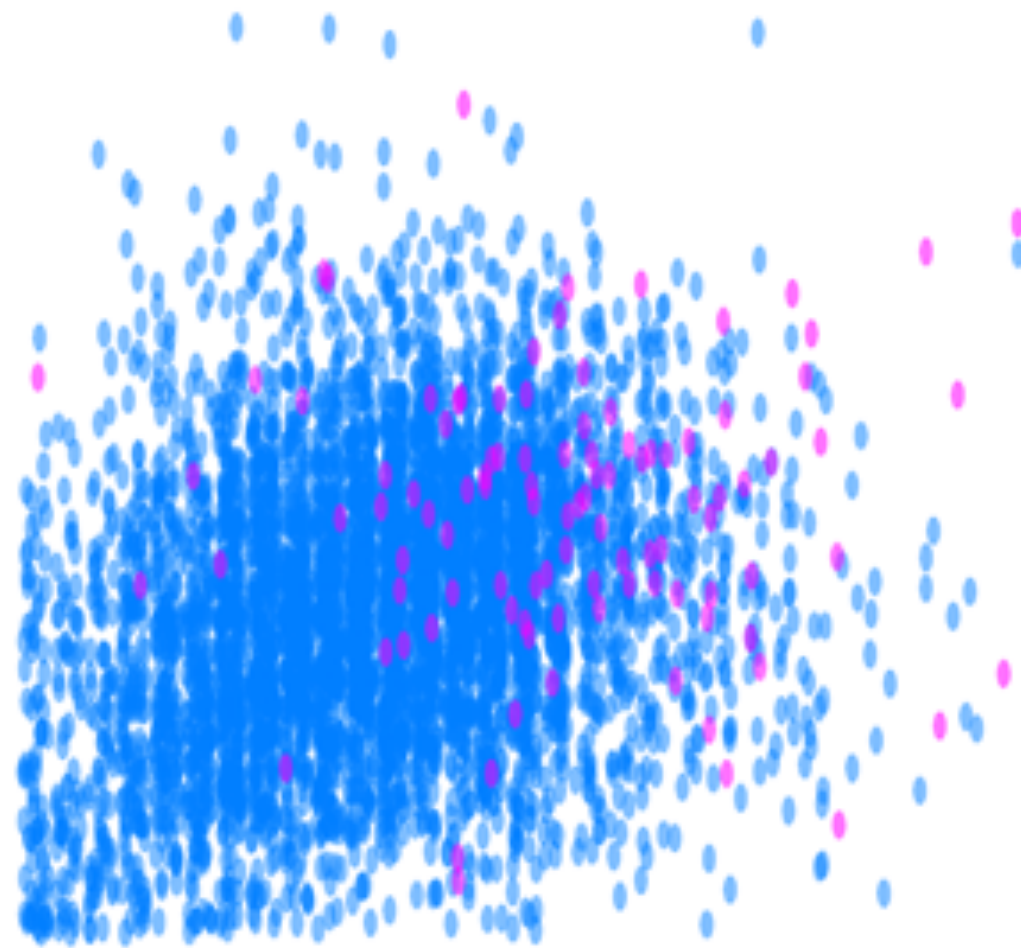
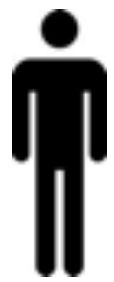
**150**  
**questions**

**150**

**questions**



**Personality  
Values  
Attributes  
Beliefs**



**ob·strep·er·ous**

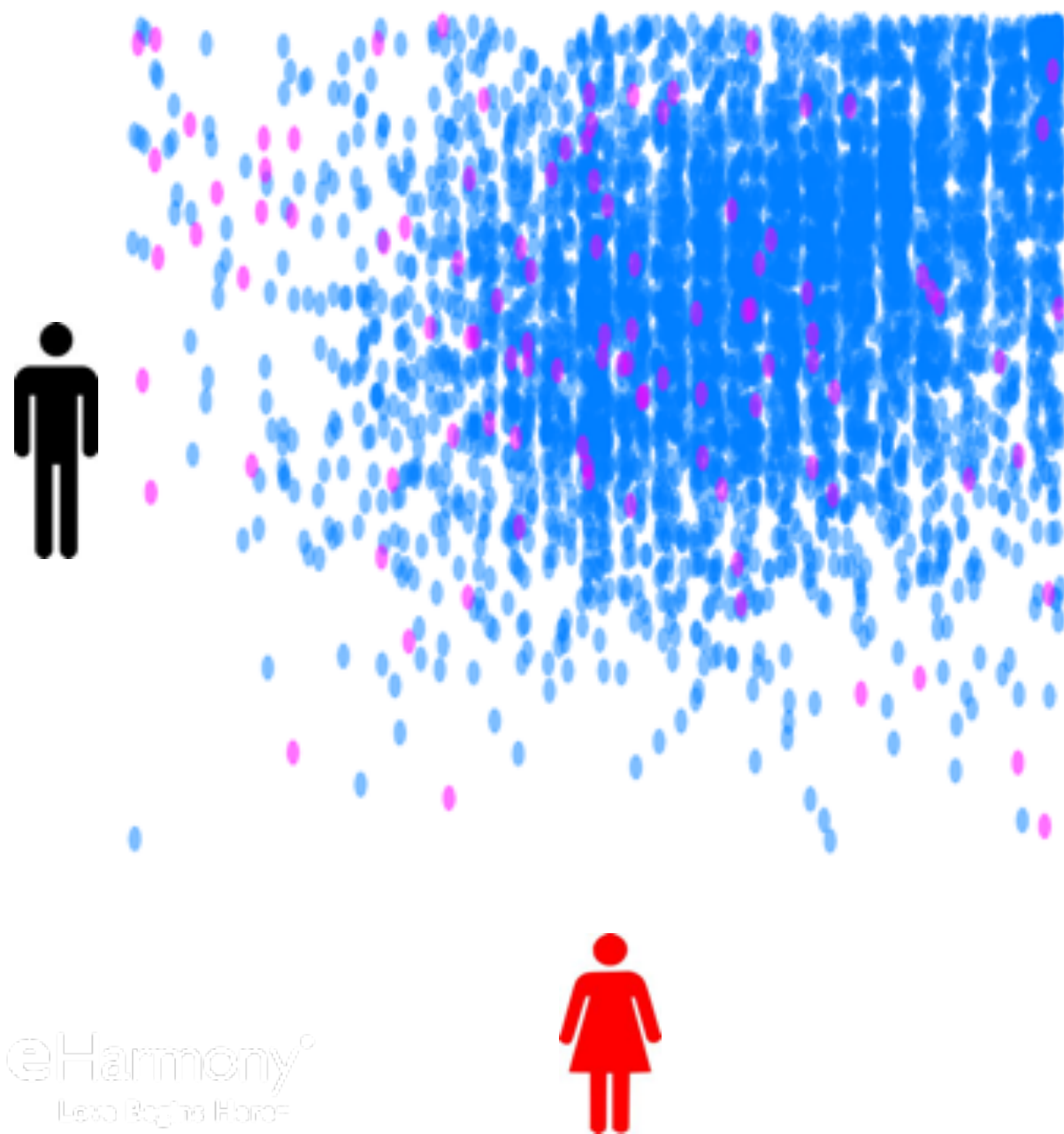
/əb'strepərəs/ 

Adjective


Noisy and difficult to control: "the boy is cocky and obstreperous".

Synonyms

noisy - loud - clamorous - rumbustious - boisterous



## ro·man·tic

/rō'mantik/ 

### Adjective

Inclined toward or suggestive of the feeling of excitement and mystery associated with love.

### Noun

A person with romantic beliefs or attitudes.

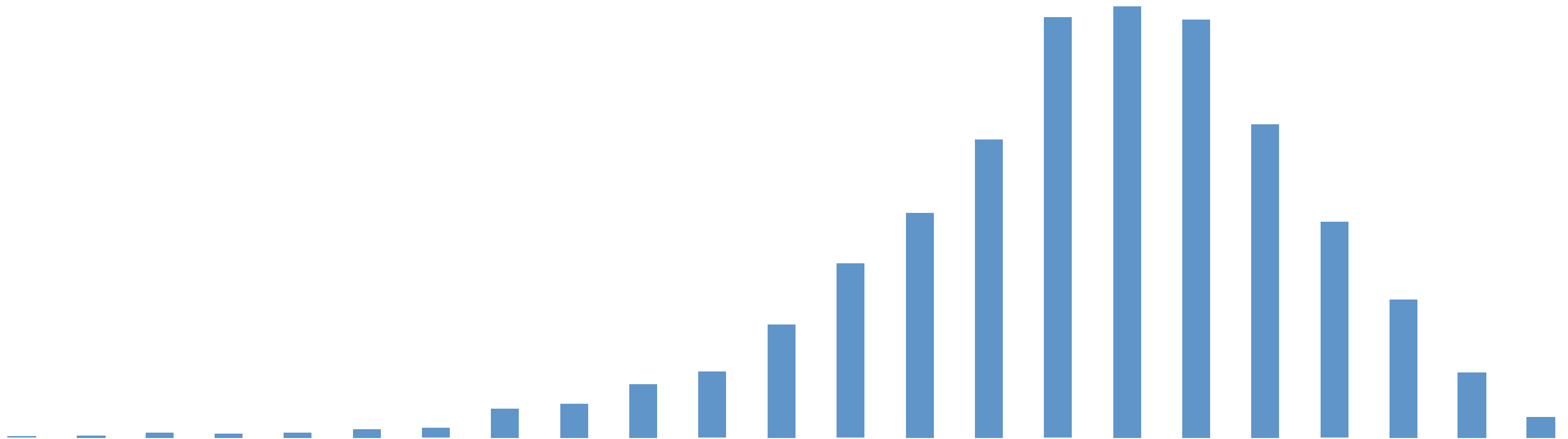
### Synonyms

romanticist



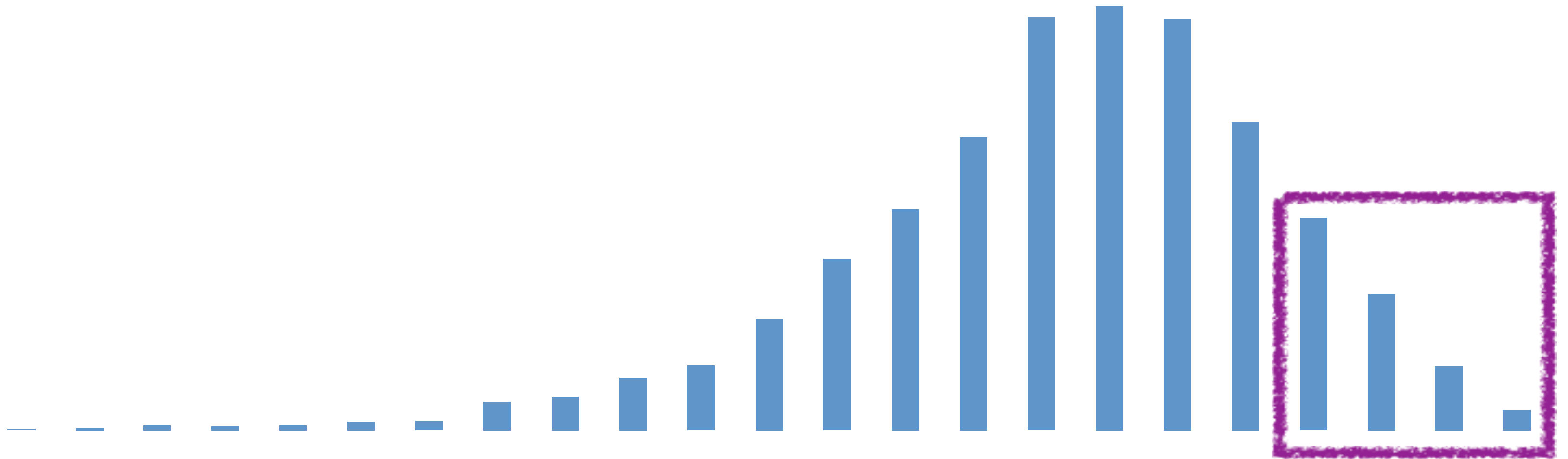
# Compatibility Matching ›

# Marital satisfaction



# Compatibility Matching >

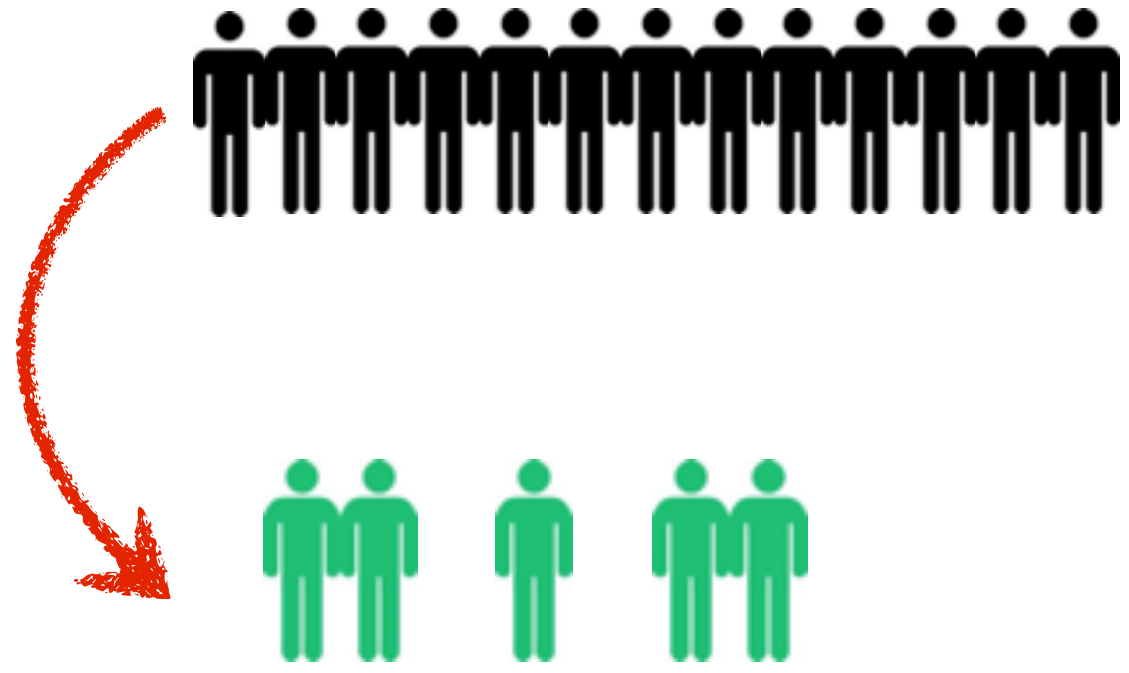
# Marital satisfaction



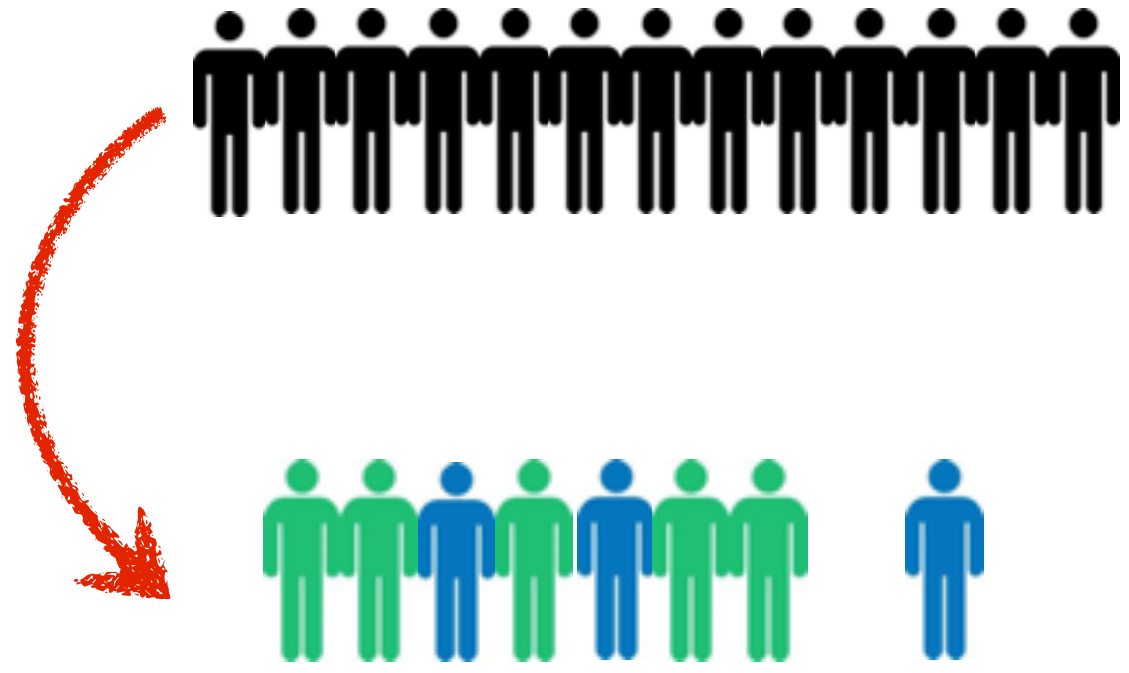
# CMP (CMP Makes Pairings)



# CMP (CMP Makes Pairings)



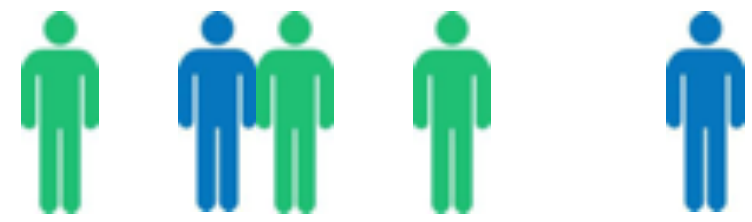
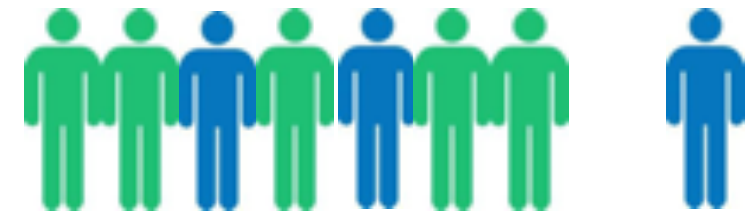
# CMP (CMP Makes Pairings)



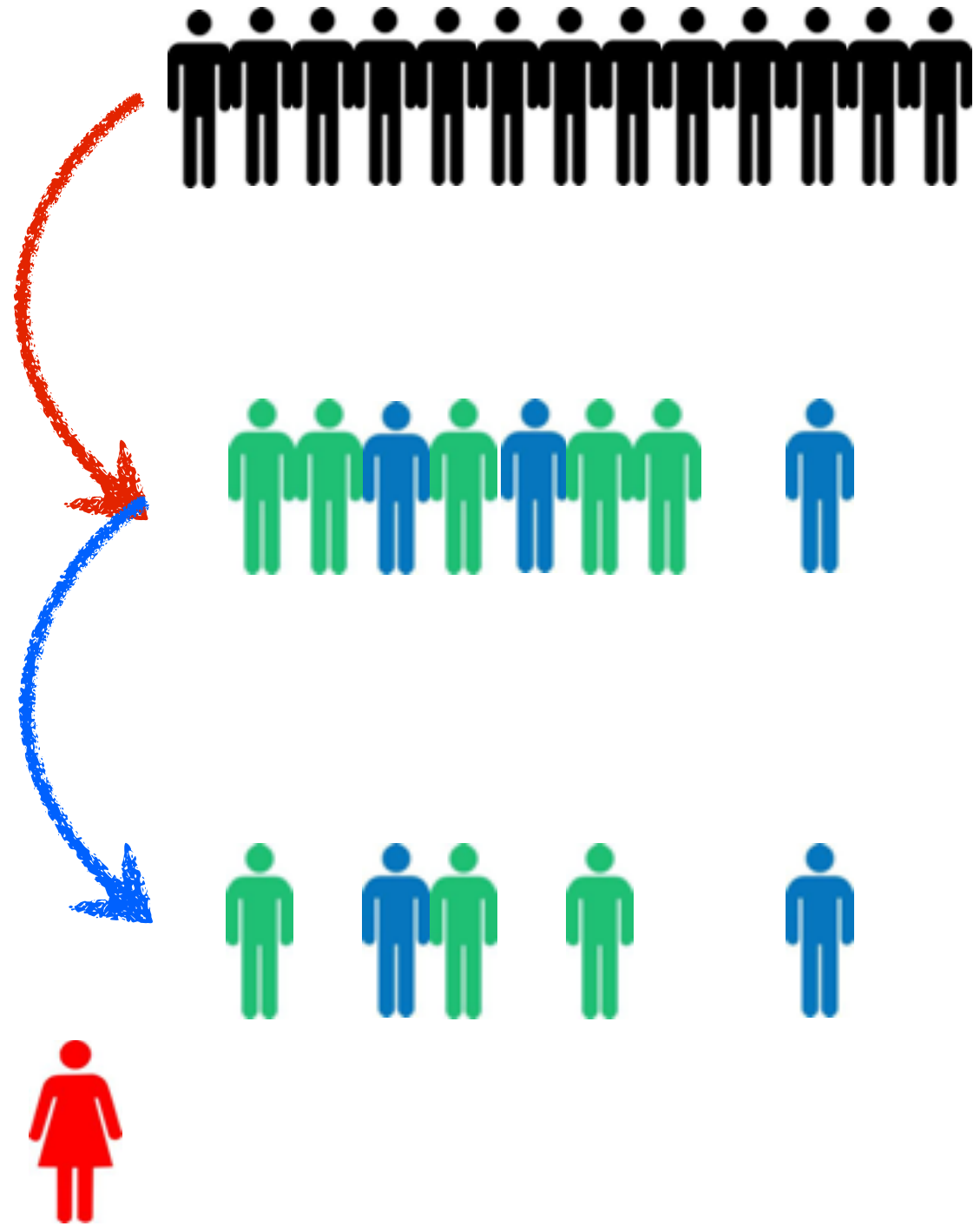
# CMP (CMP Makes Pairings)



**Compatibility Models**



# CMP (CMP Makes Pairings)

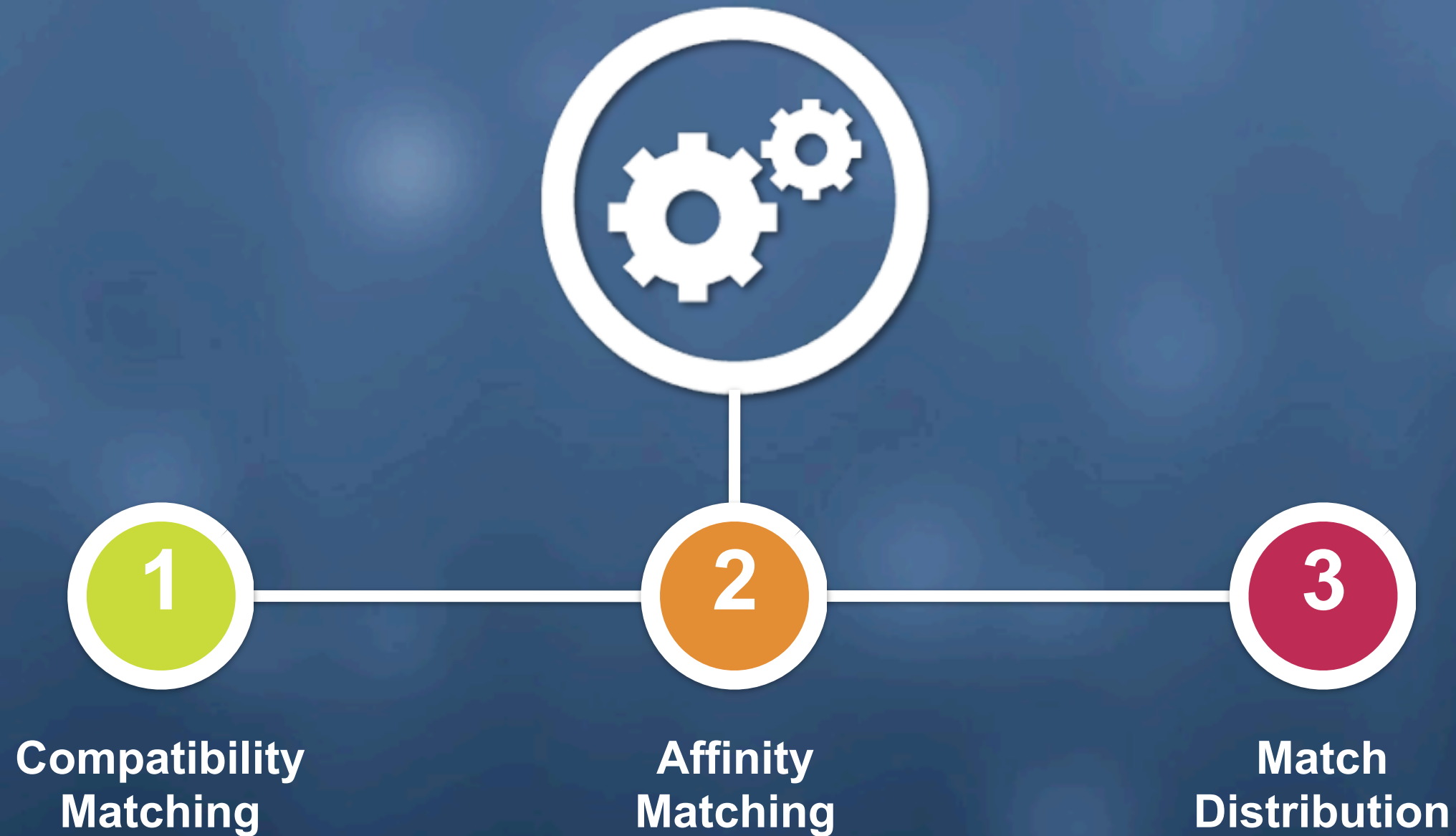


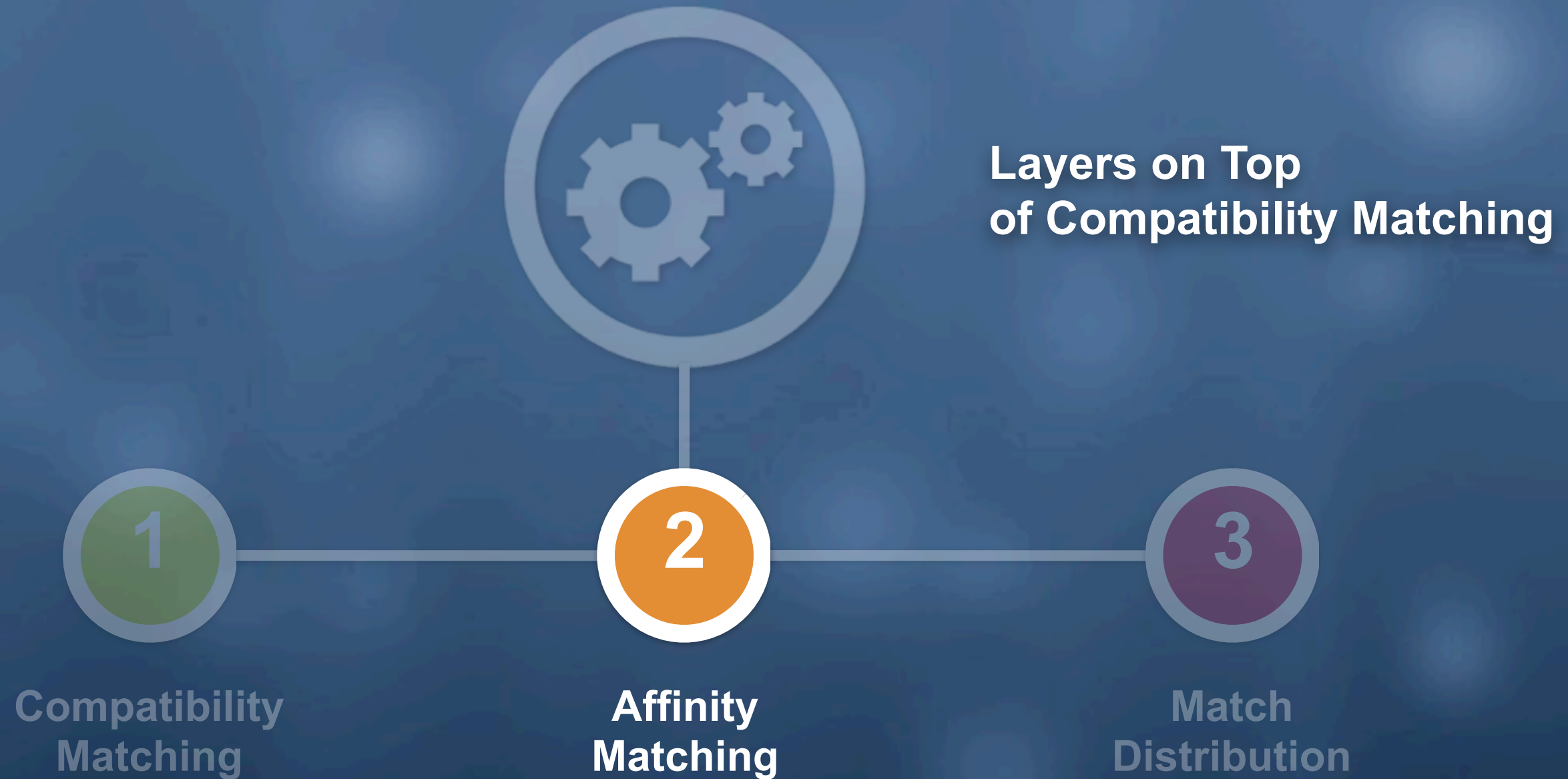
# Compatibility Matching ›



# Compatibility Matching ›







# Affinity Matching ›



# Affinity Matching ›



**61**



**21**



# Affinity Matching ›



**61**



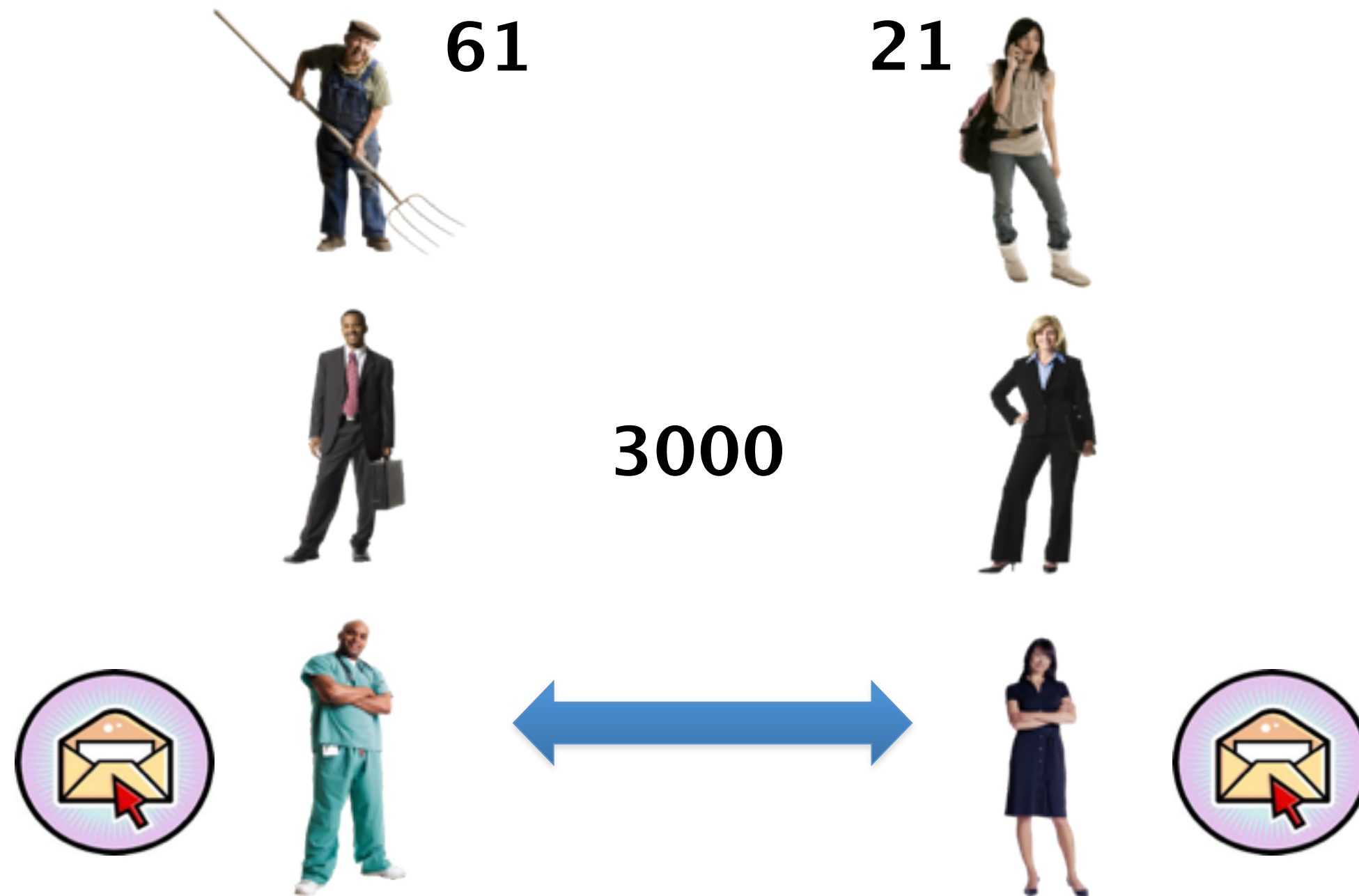
**21**



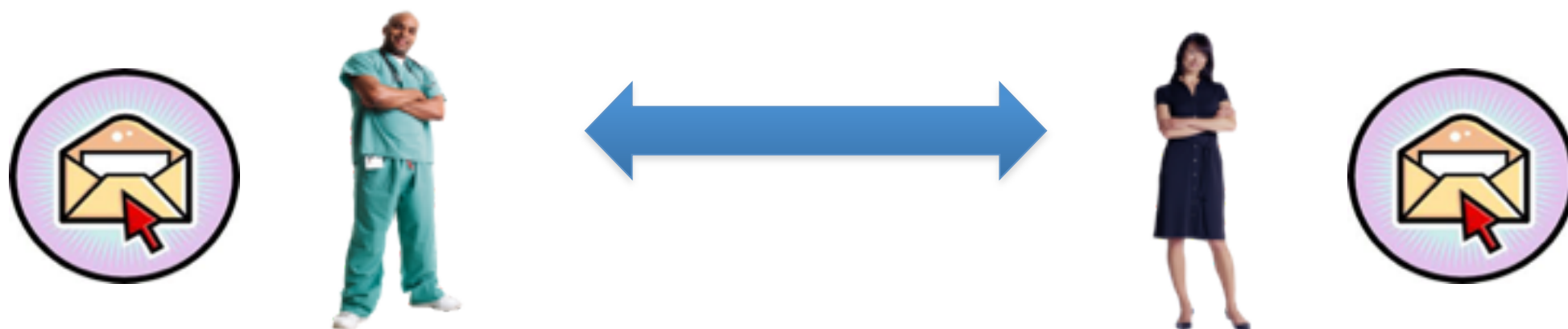
**3000**



# Affinity Matching >

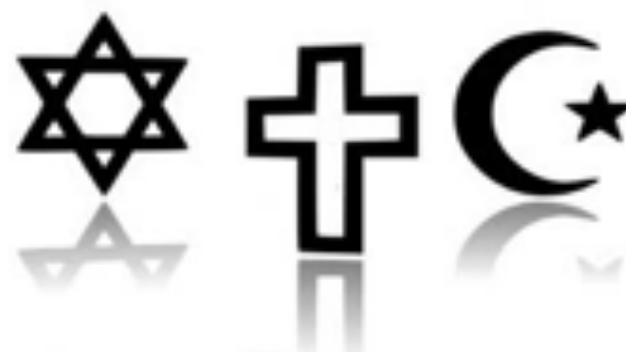


# Affinity Matching ›





# Affinity Matching >



...

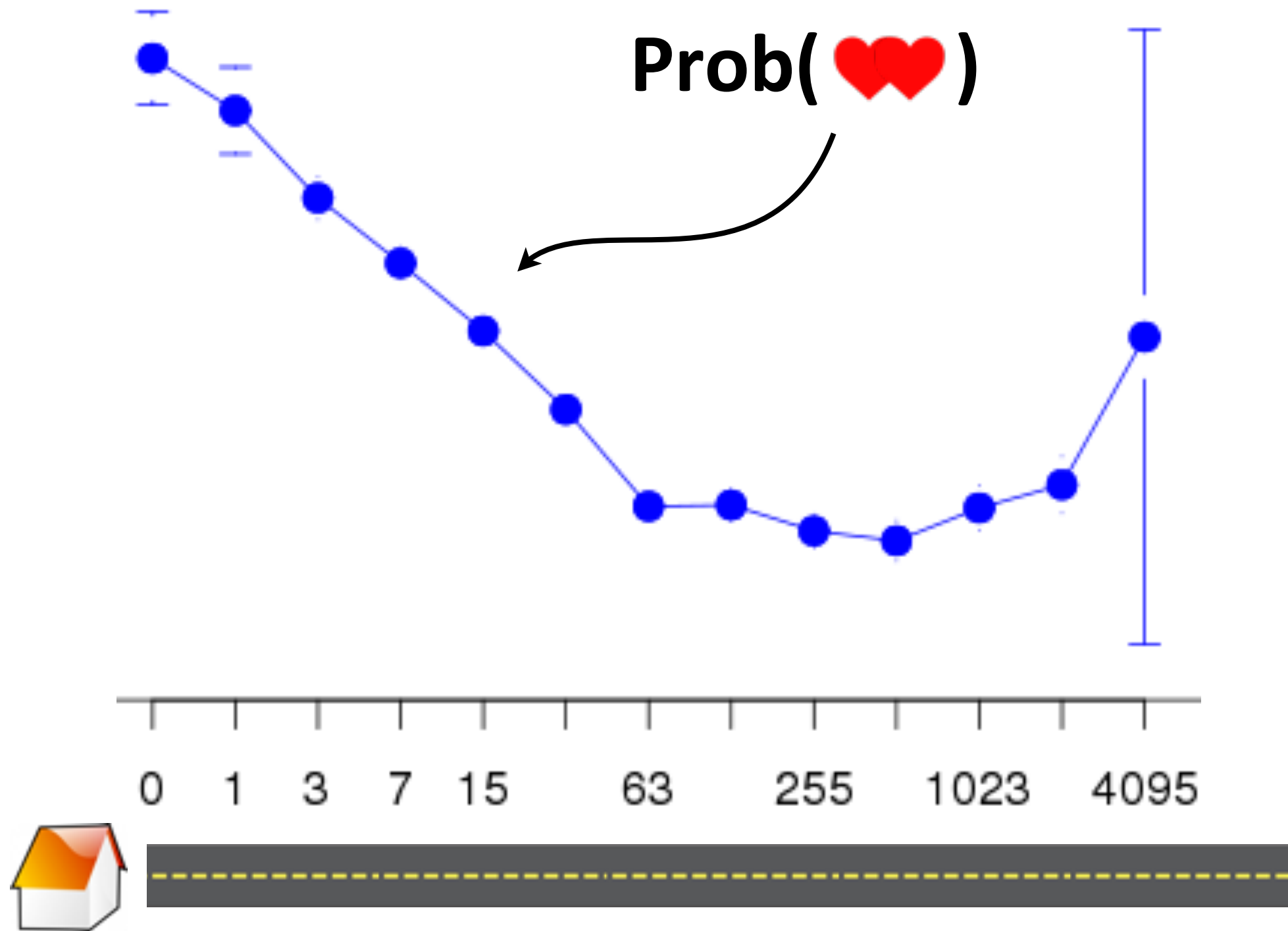


...



...





Affinity Matching ›

Distance

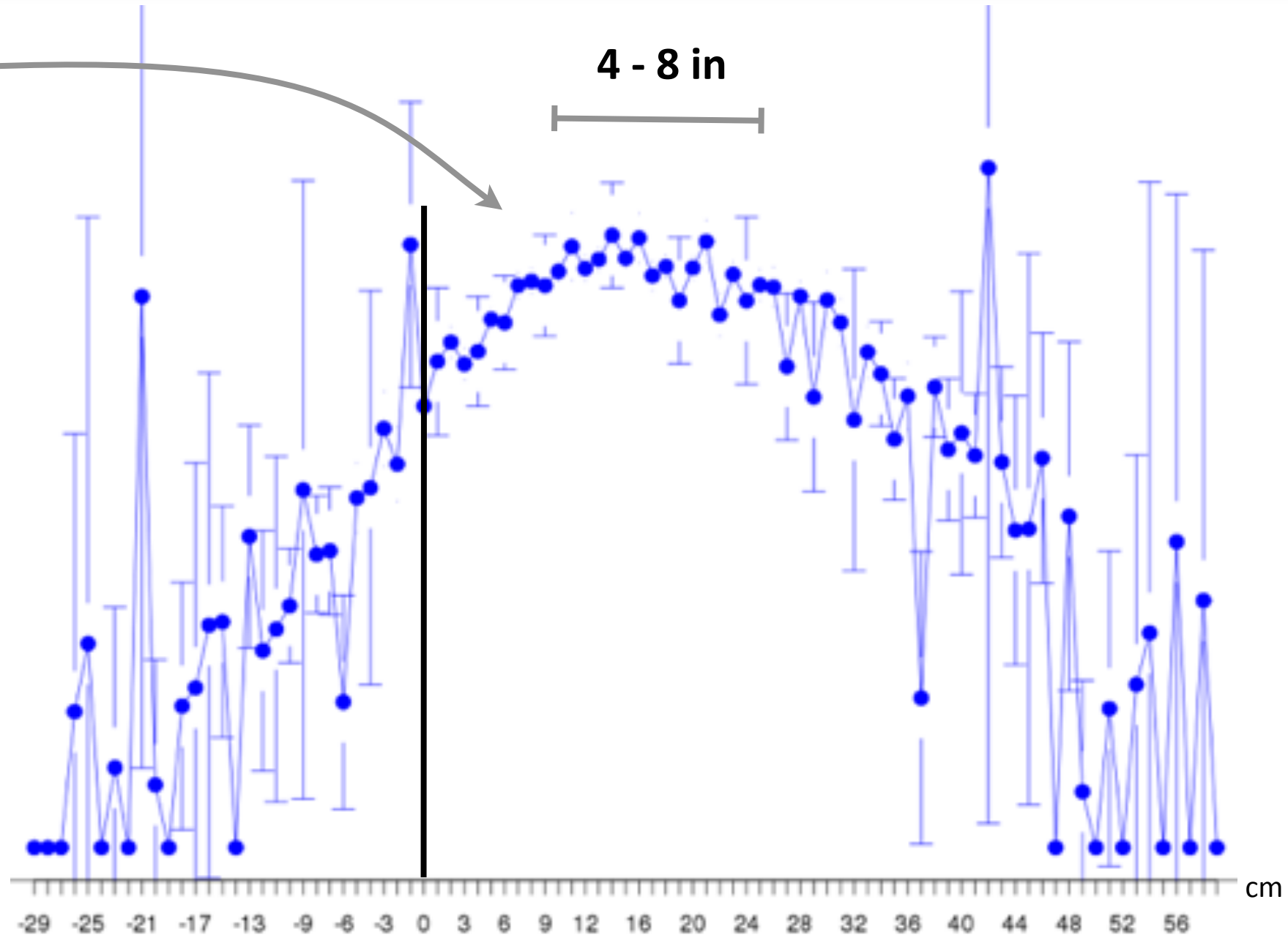
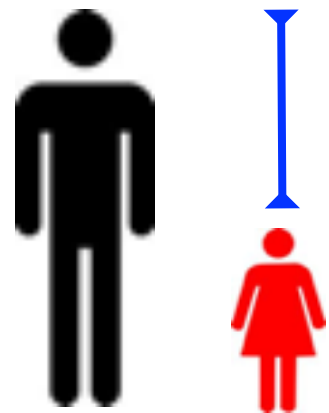


# Affinity Matching >

# Height difference

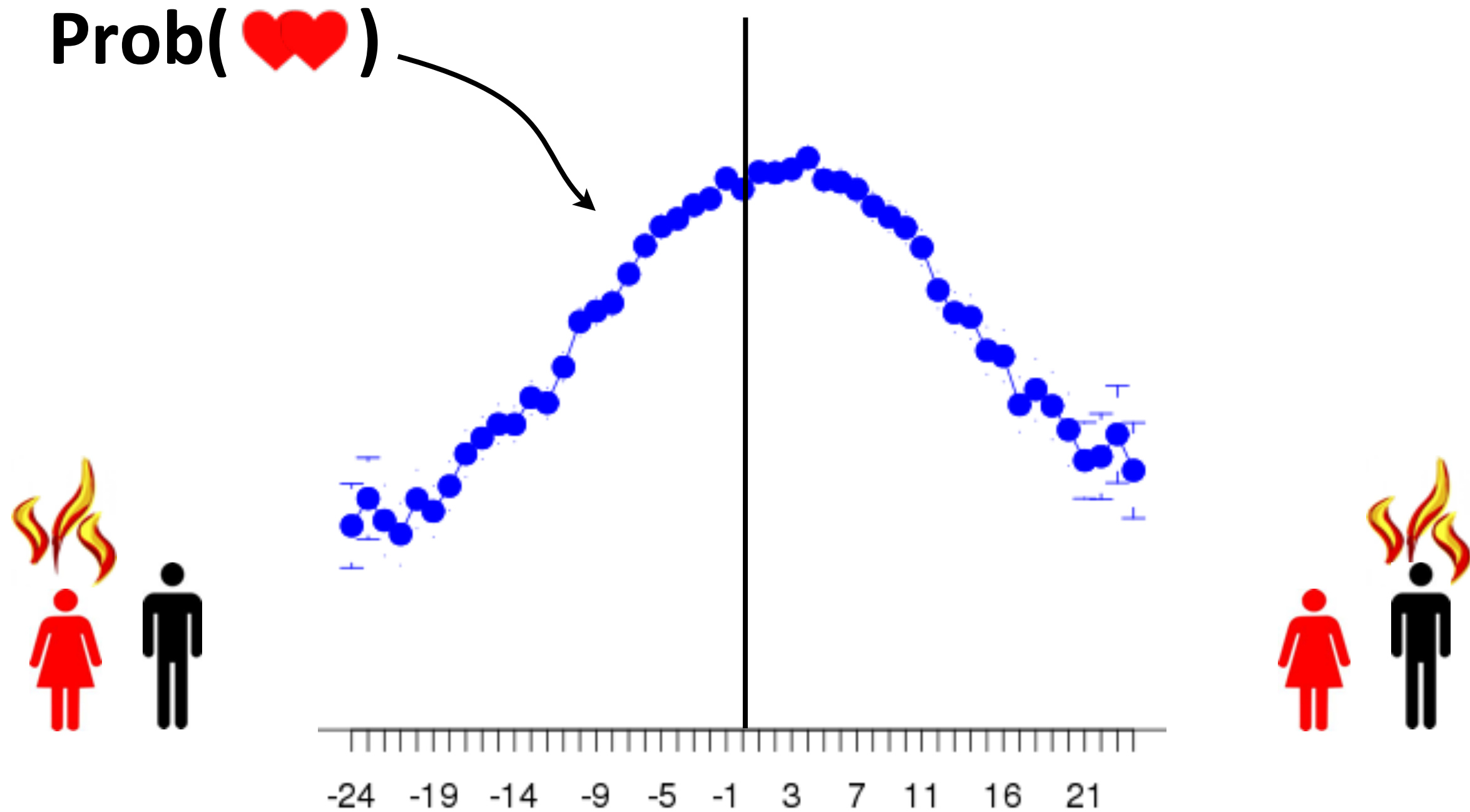
Prob(♥♥)

4 - 8 in



# Affinity Matching >

“Attractiveness”



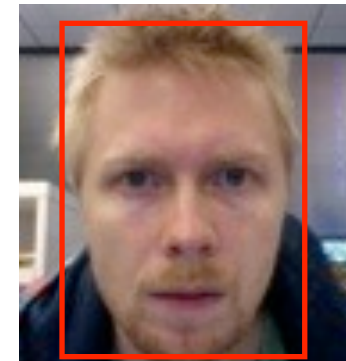
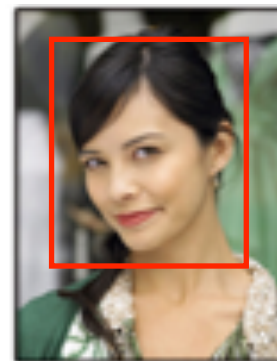
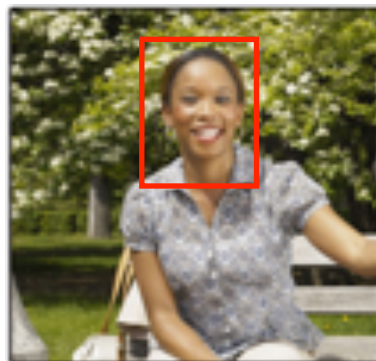
# Affinity Matching ›

## Photo features



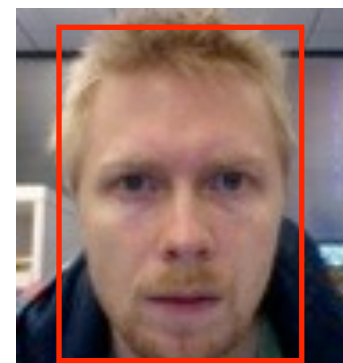
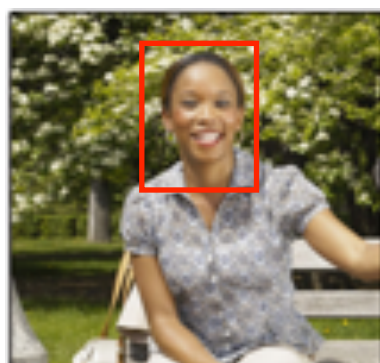
# Affinity Matching >

# Photo features



# Affinity Matching >

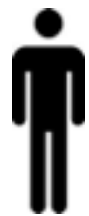
# Photo features
















# Affinity Matching >

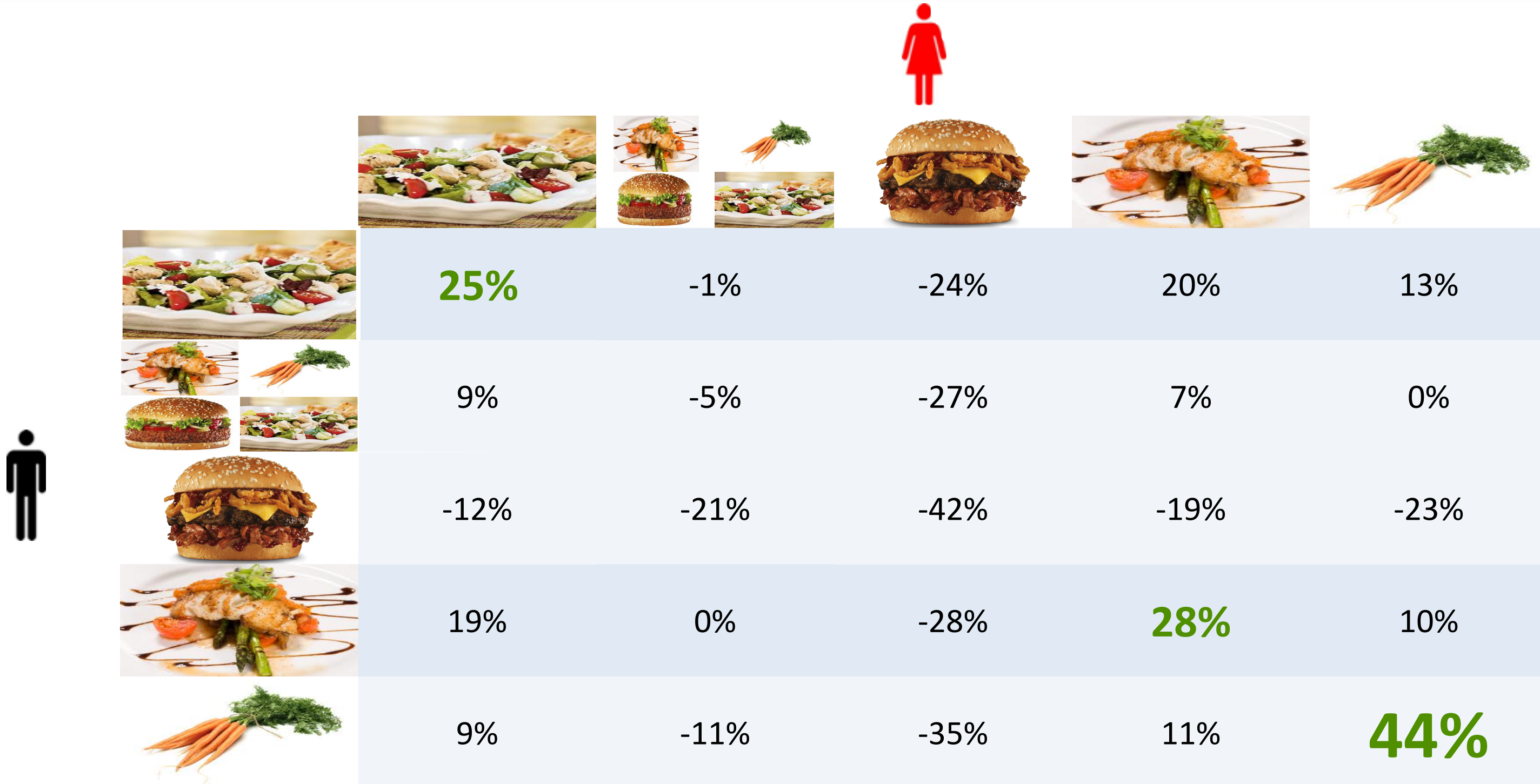
# Food preference



						
	25%	-1%	<b>-24%</b>	20%	13%	
	9%	-5%	<b>-27%</b>	7%	0%	
	<b>-12%</b>	<b>-21%</b>	<b>-42%</b>	<b>-19%</b>	<b>-23%</b>	
	19%	0%	<b>-28%</b>	28%	10%	
	9%	-11%	<b>-35%</b>	11%	44%	

# Affinity Matching >

# Food preference



# MR. NOISY

By Roger Hargreaves



# LITTLE MISS SHY

By Roger Hargreaves



# MR. QUIET

By Roger Hargreaves



# LITTLE MISS CHATTERBOX

By Roger Hargreaves



# MR. NOISY

By Roger Hargreaves



# LITTLE MISS SHY

By Roger Hargreaves



Copyrighted Material

# LITTLE MISS CHATTERBOX

By Roger Hargreaves



# MR. QUIET

By Roger Hargreaves



# MR. NOISY

By Roger Hargreaves



# LITTLE MISS SHY

By Roger Hargreaves



# MR. QUIET

By Roger Hargreaves



# LITTLE MISS CHATTERBOX

By Roger Hargreaves



# MR. NOISY

By Roger Hargreaves



# LITTLE MISS SHY

By Roger Hargreaves



# MR. QUIET

By Roger Hargreaves



# LITTLE MISS CHATTERBOX

By Roger Hargreaves



# MR. NOISY

By Roger Hargreaves



# LITTLE MISS SHY

By Roger Hargreaves



Copyrighted Material

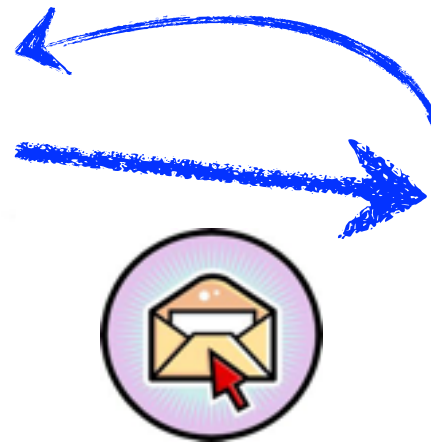
# MR. QUIET

By Roger Hargreaves



# LITTLE MISS CHATTERBOX

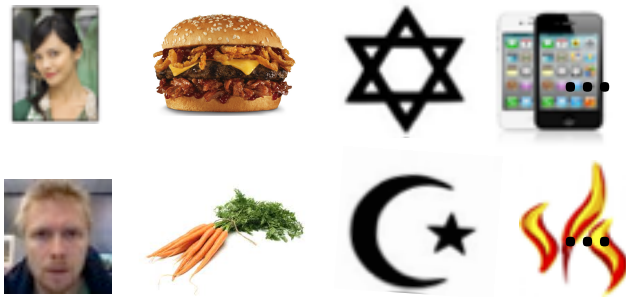
By Roger Hargreaves



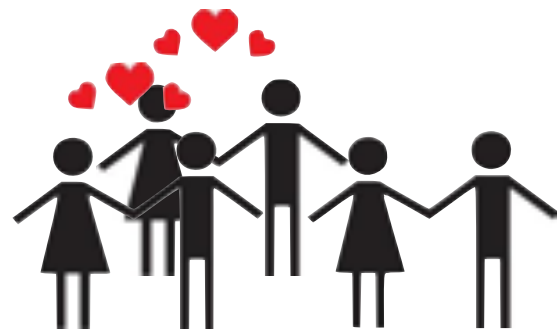
# Affinity Matching >

Prob(❤️❤️ | data)

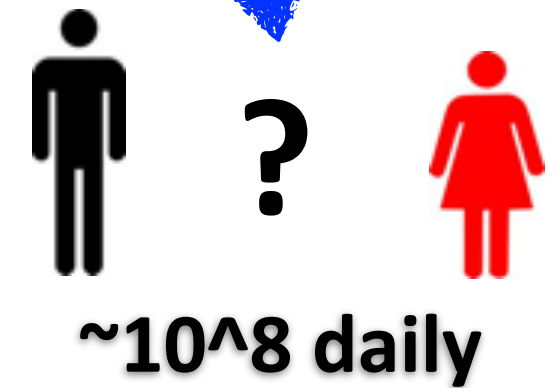
~40M registered users



~10<sup>3</sup> attributes



~10<sup>7</sup> matches per day



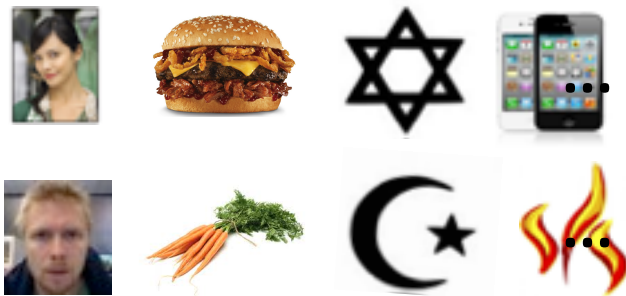
Prob(❤️❤️ | features)



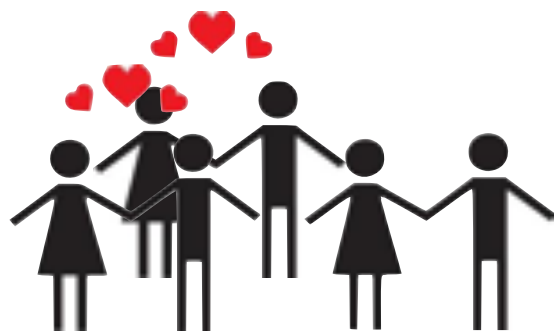
# Affinity Matching >

Prob(❤️❤️ | data)

~40M registered users

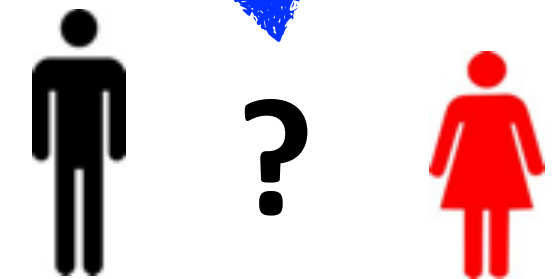


~10<sup>3</sup> attributes



~10<sup>7</sup> matches per day

Constructed features  
Unsupervised features  
(LDA, classifiers)



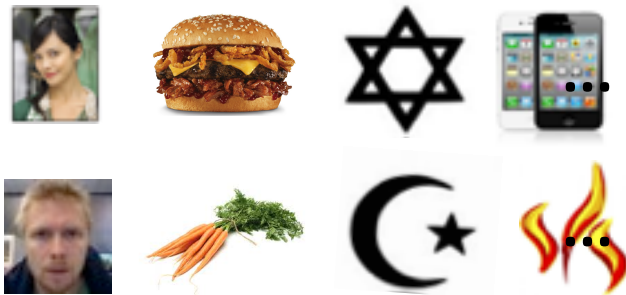
~10<sup>8</sup> daily

Prob(❤️❤️ | features)

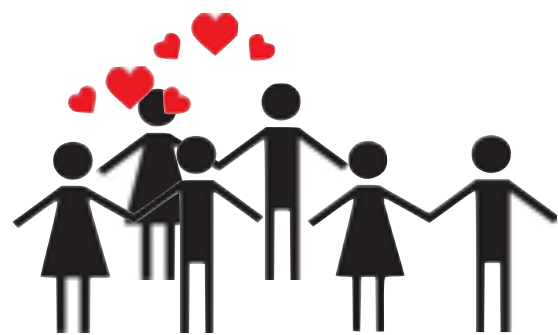
# Affinity Matching >

Prob(❤️❤️ | data)

~40M registered users



~10<sup>3</sup> attributes

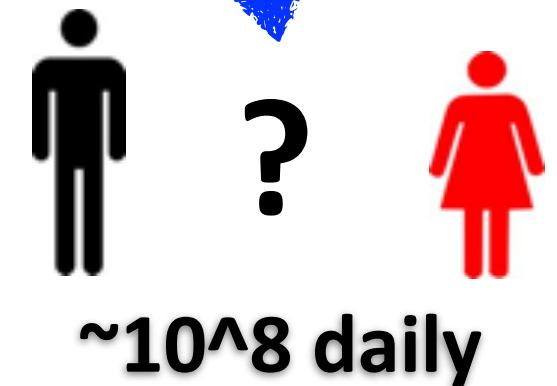


~10<sup>7</sup> matches per day

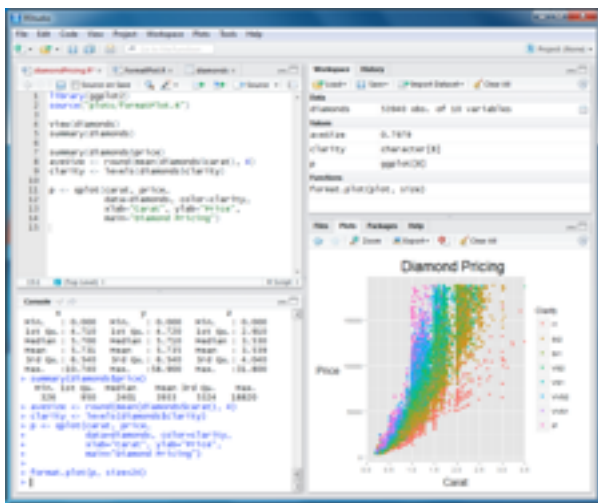
Constructed features  
Unsupervised features  
(LDA, classifiers)



L1 regularization  
transfer learning  
holdout validation  
subsampling  
calibration



~10<sup>8</sup> daily  
Prob(❤️❤️ | features)



RStudio server

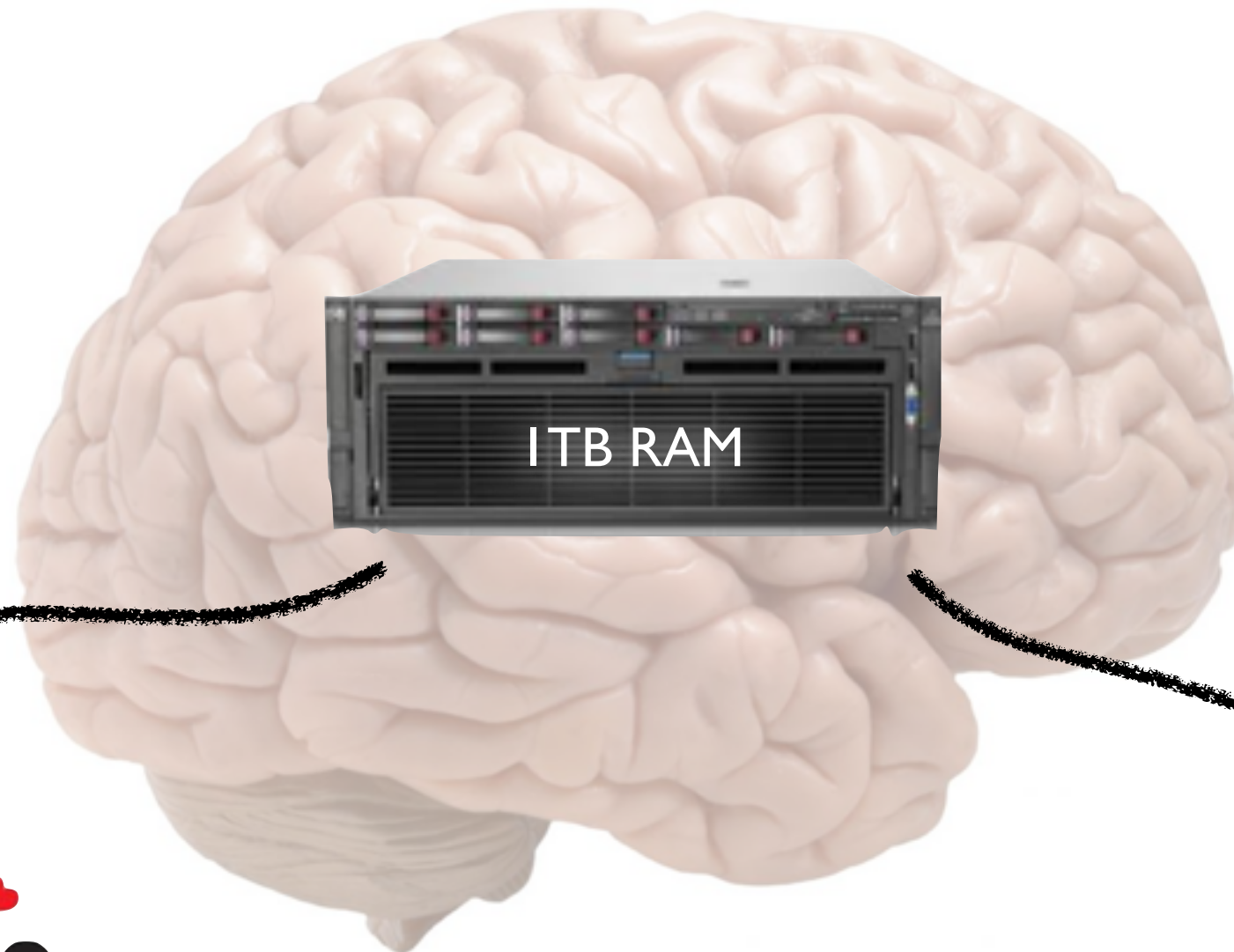


# Eureqa

genetic algorithms

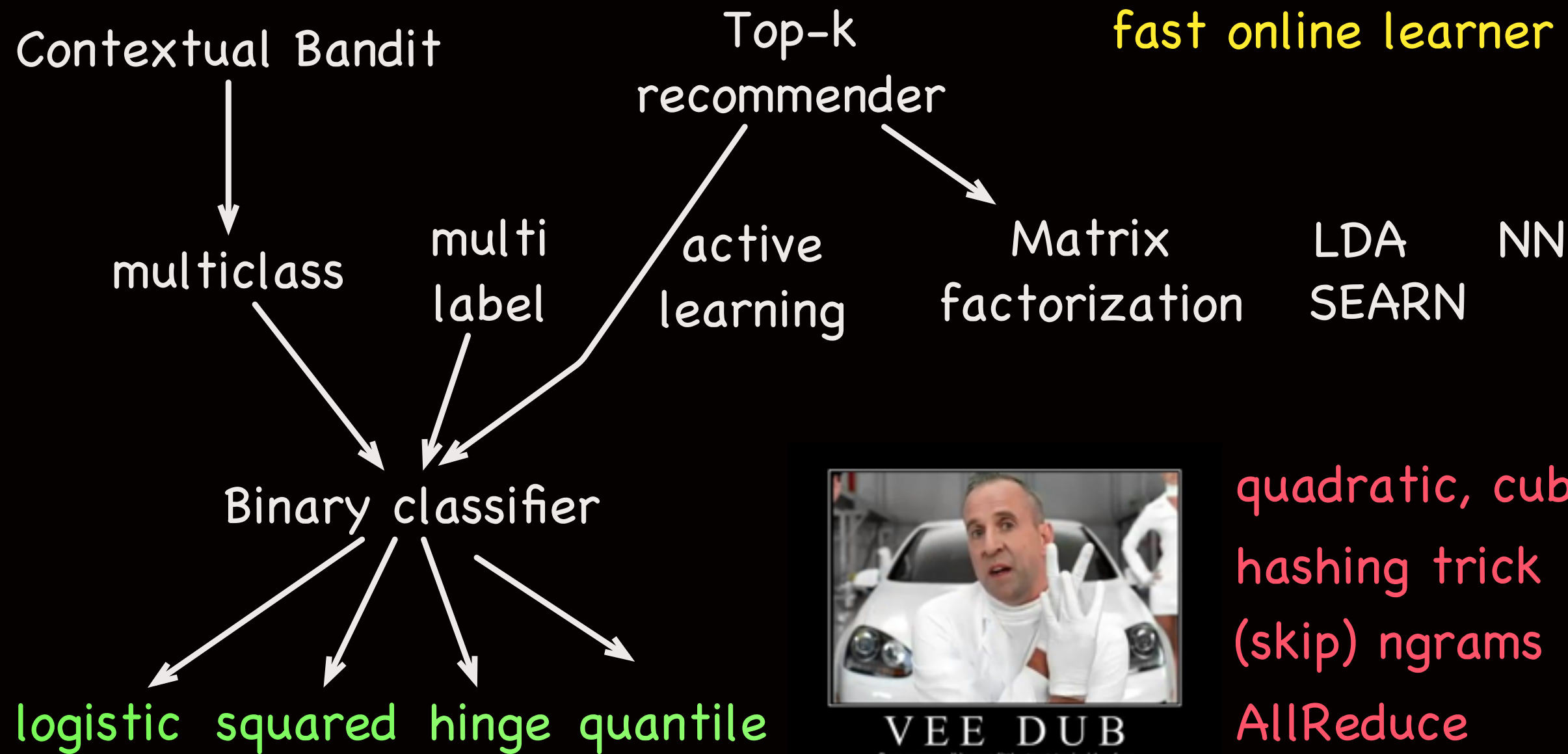


vowpal wabbit



YARN cluster

# Affinity Matching › Vowpal Wabbit (aka “vee-dub”)



VEE DUB  
Because we all have a little gangster inside of us

quadratic, cubic  
hashing trick  
(skip) ngrams  
AllReduce



# Holdout validation

#1

Progressive validation loss:



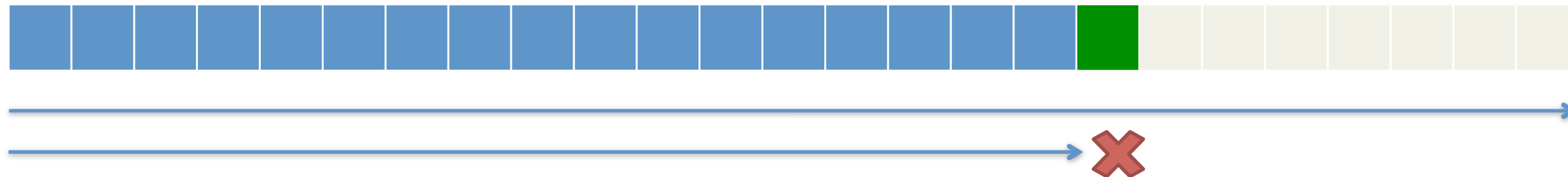
**NEW**

# Holdout validation

#1

Progressive validation loss:

Meaningful for 1 pass only



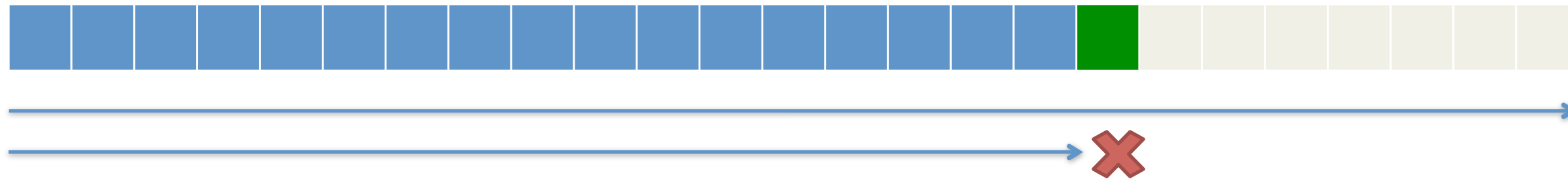


# Holdout validation

#1

Progressive validation loss:

Meaningful for 1 pass only



Holdout loss:



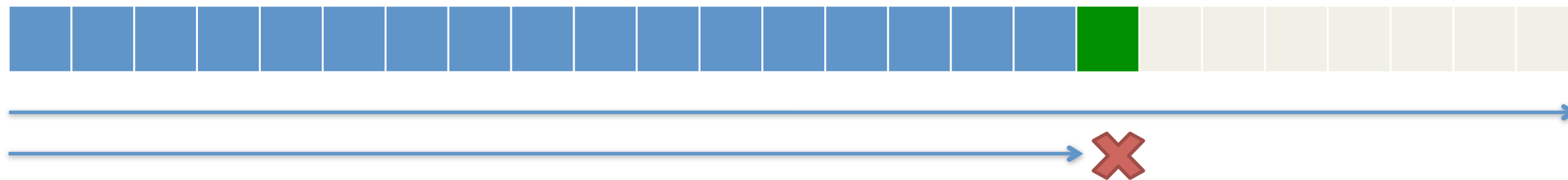


# Holdout validation

#1

Progressive validation loss:

Meaningful for 1 pass only



Holdout loss:





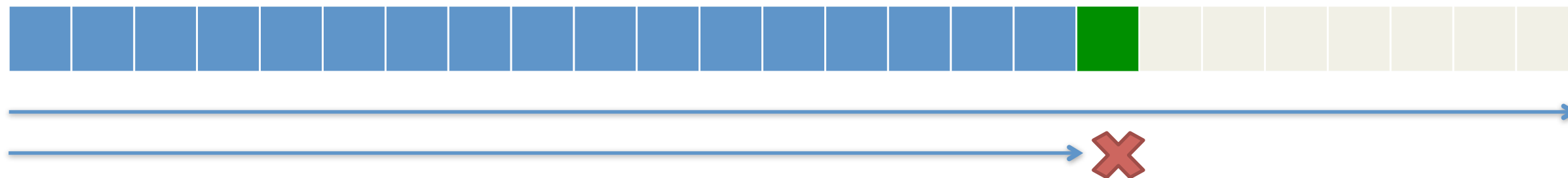


# Holdout validation

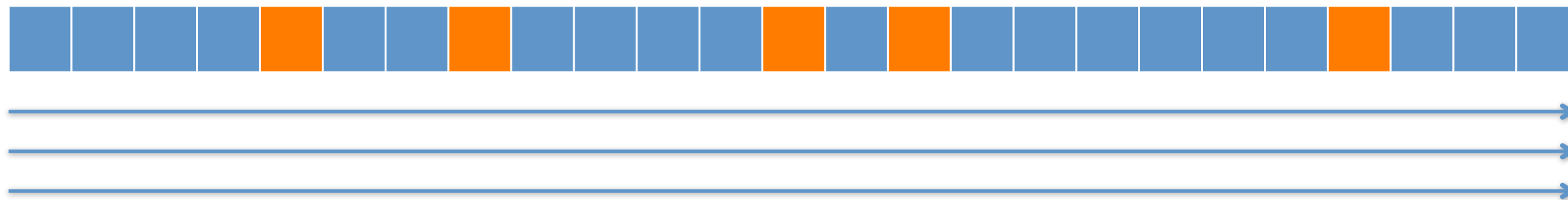
# #1

Progressive validation loss:

Meaningful for 1 pass only



Holdout loss:



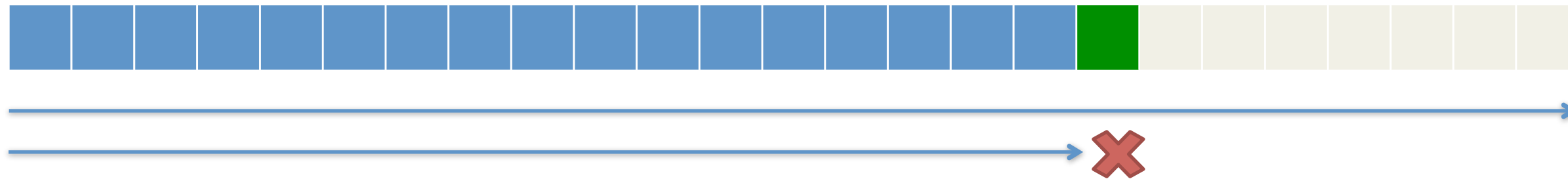


# Holdout validation

#1

Progressive validation loss:

Meaningful for 1 pass only



Holdout loss:

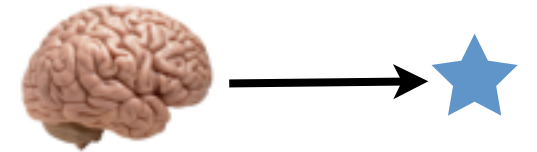


**Now: default behavior for multipass**

**NEW**

--bs: bootstrapping

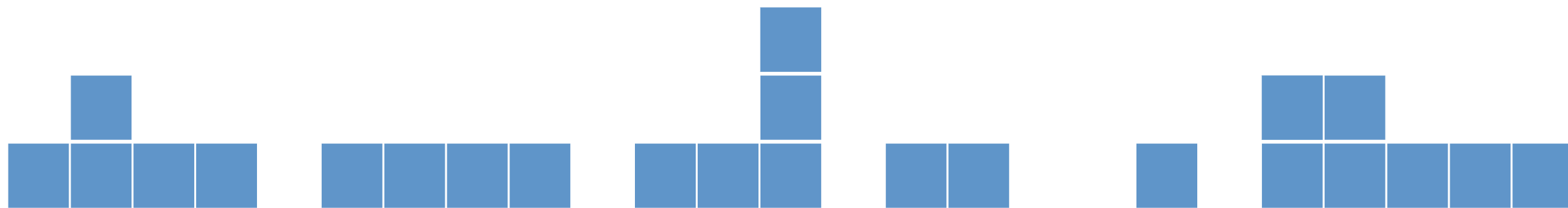
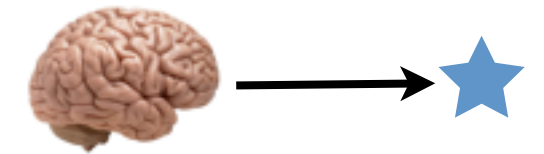
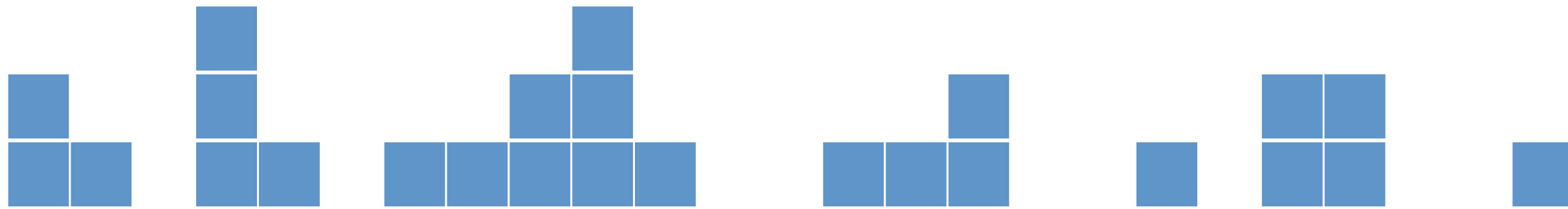
**#2**



**NEW**

--bs: bootstrapping

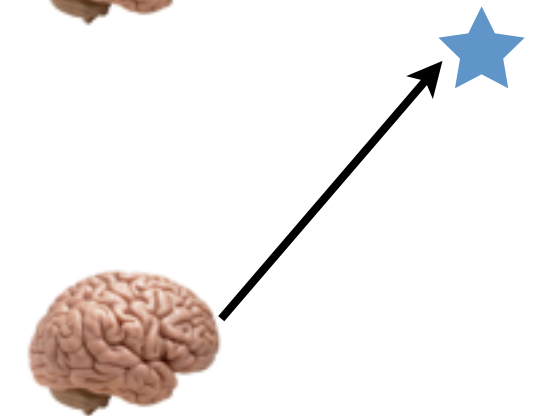
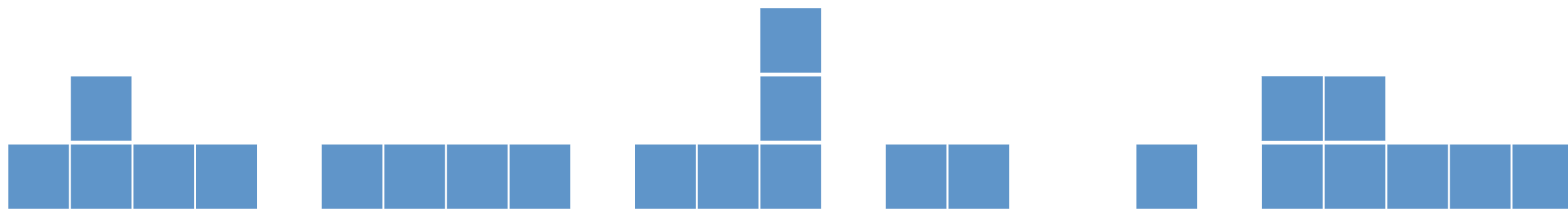
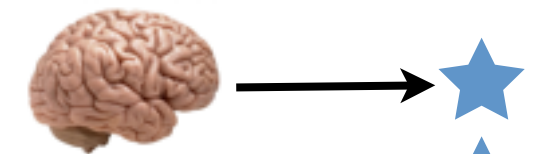
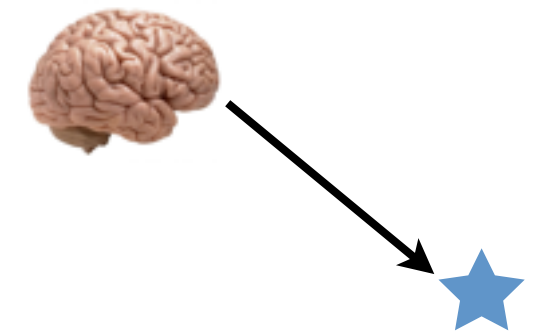
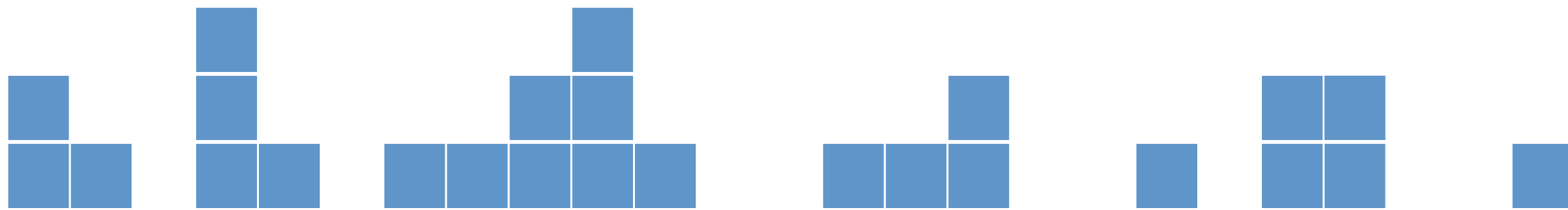
#2



**NEW**

--bs: bootstrapping

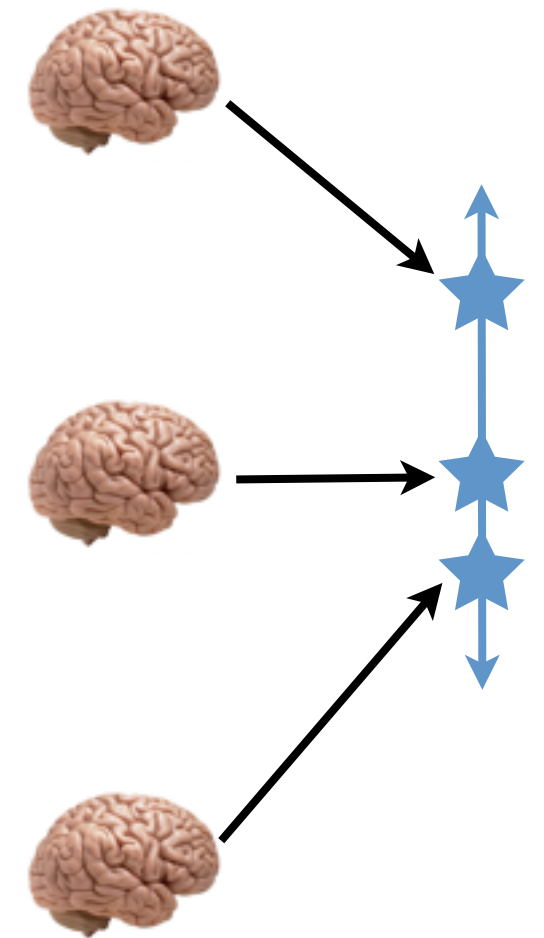
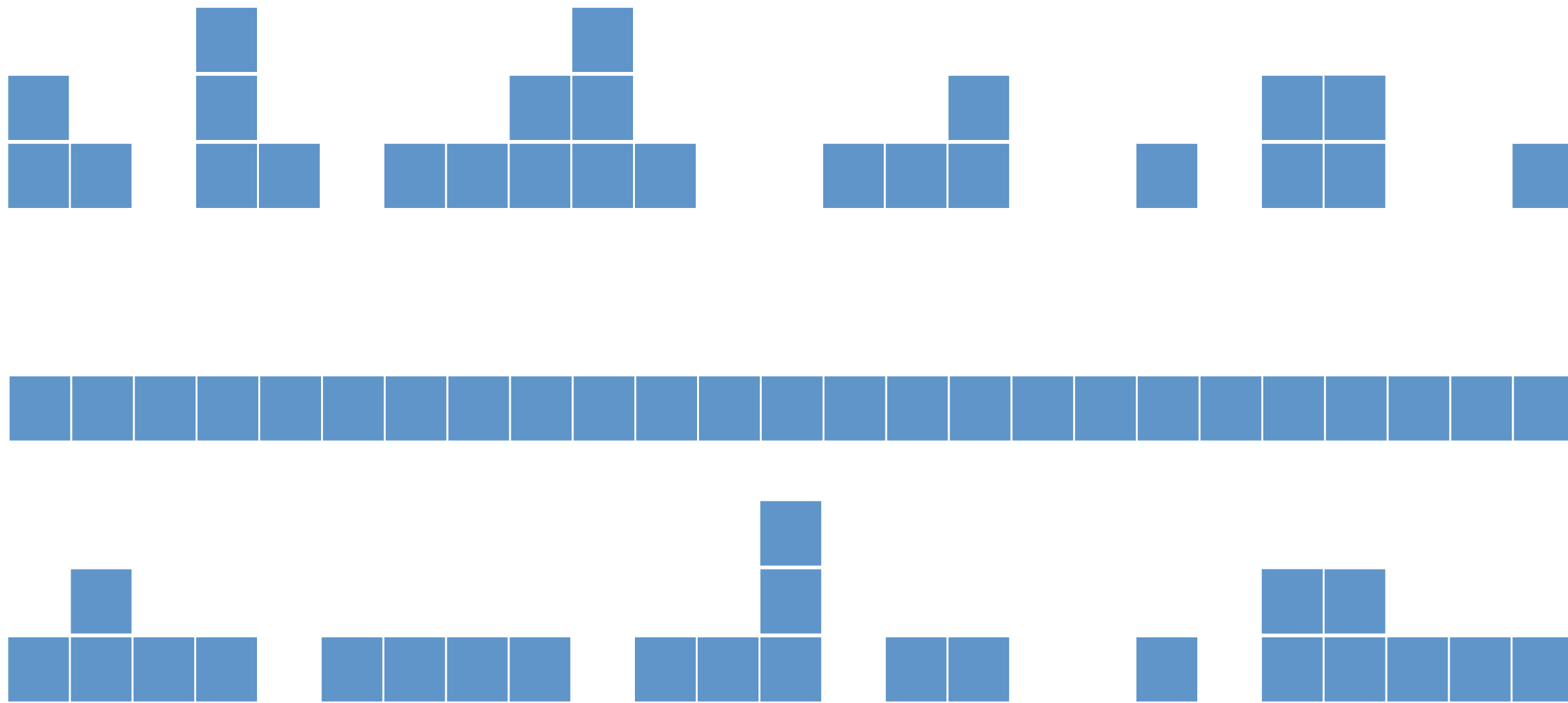
**#2**



**NEW**

--bs: bootstrapping

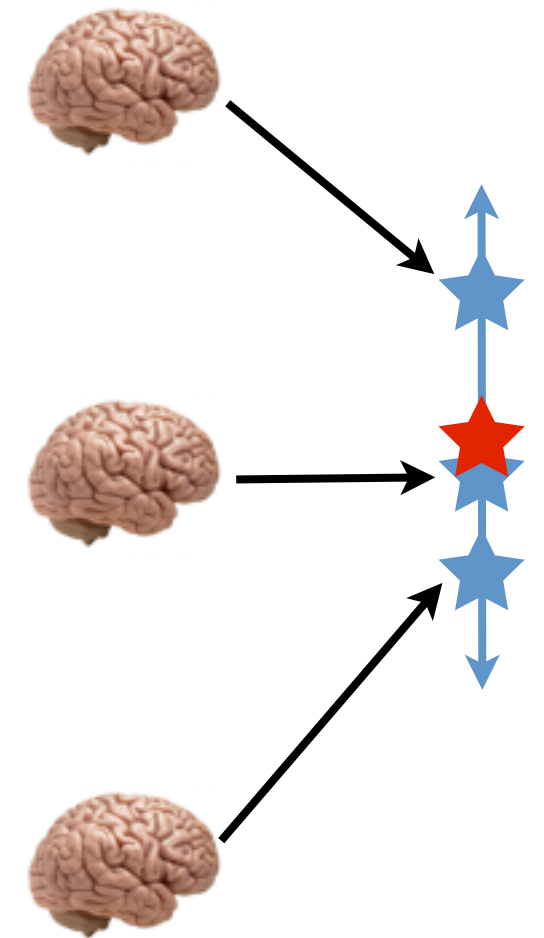
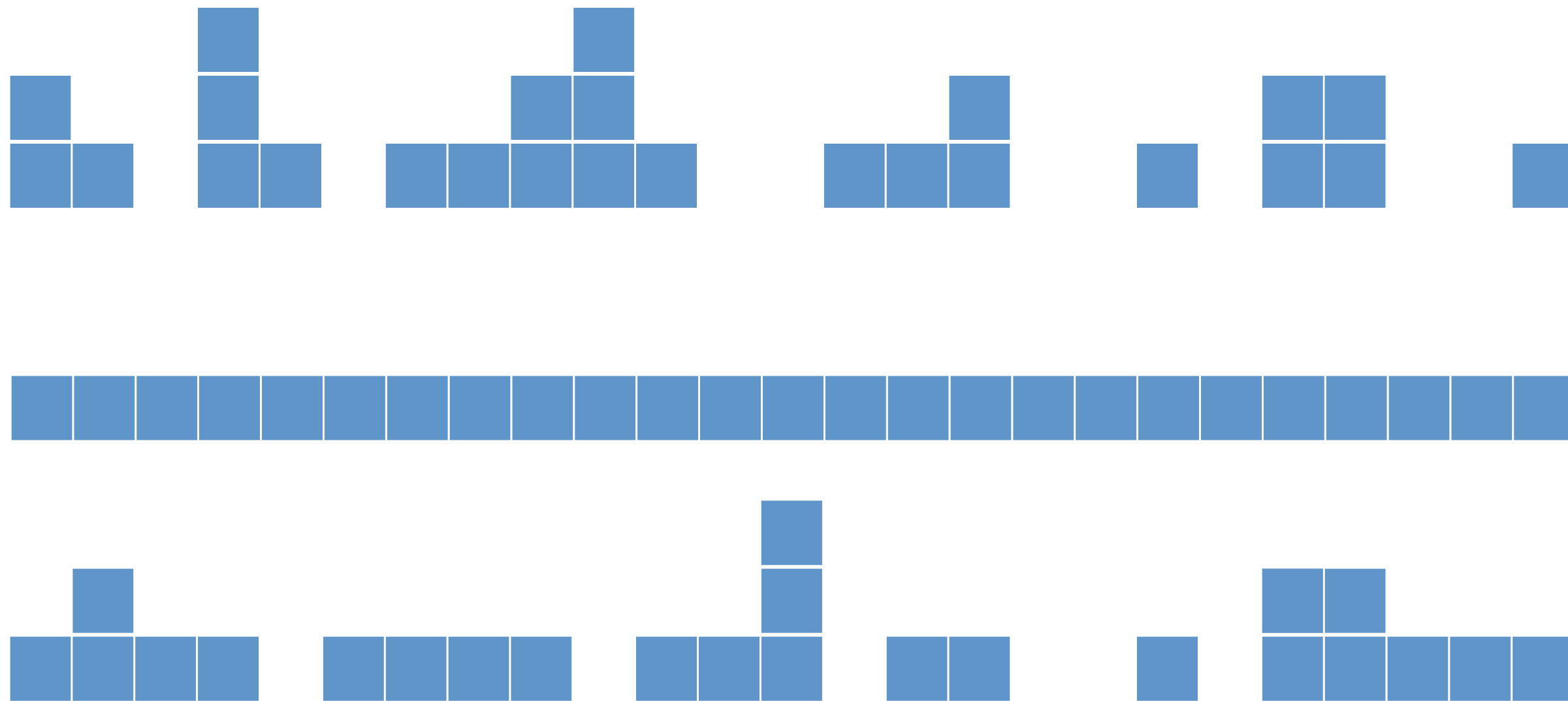
**#2**



**NEW**

--bs: bootstrapping

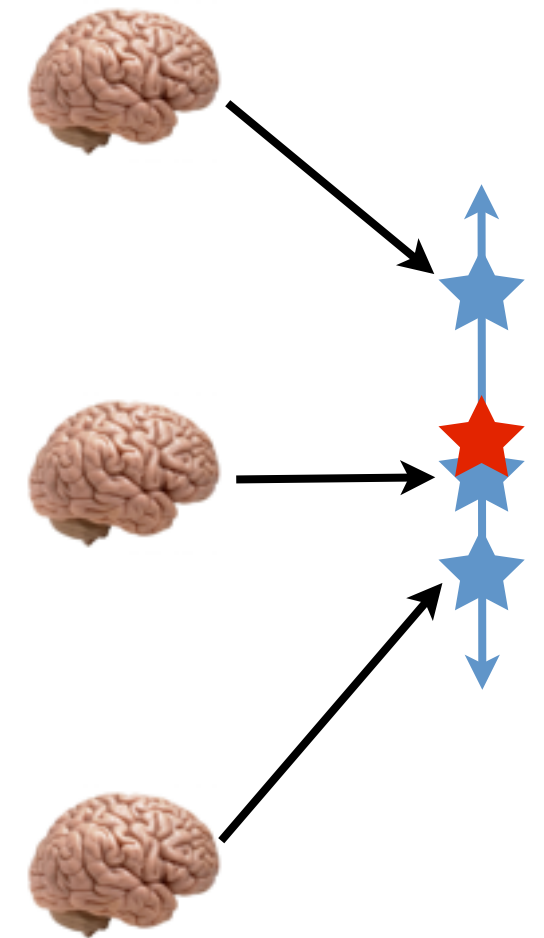
**#2**



**NEW**

--bs: bootstrapping

**#2**

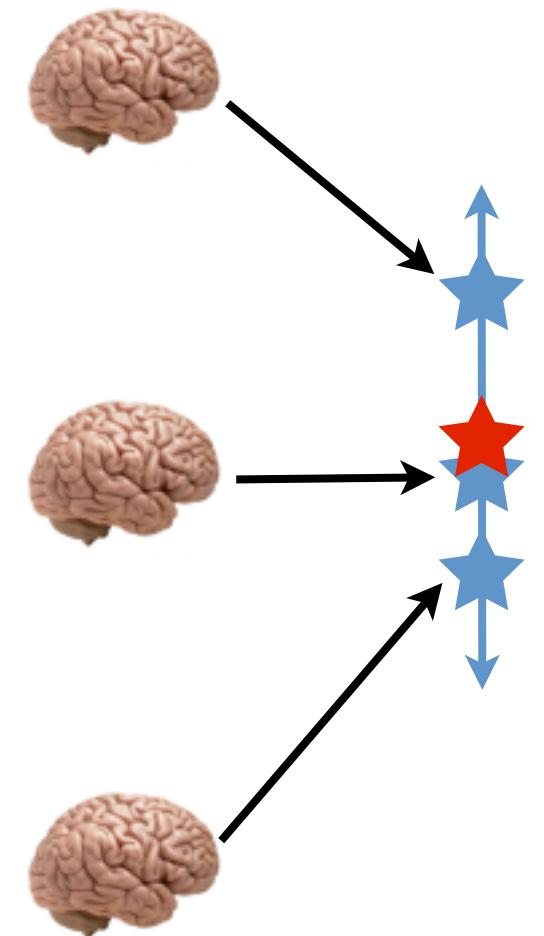
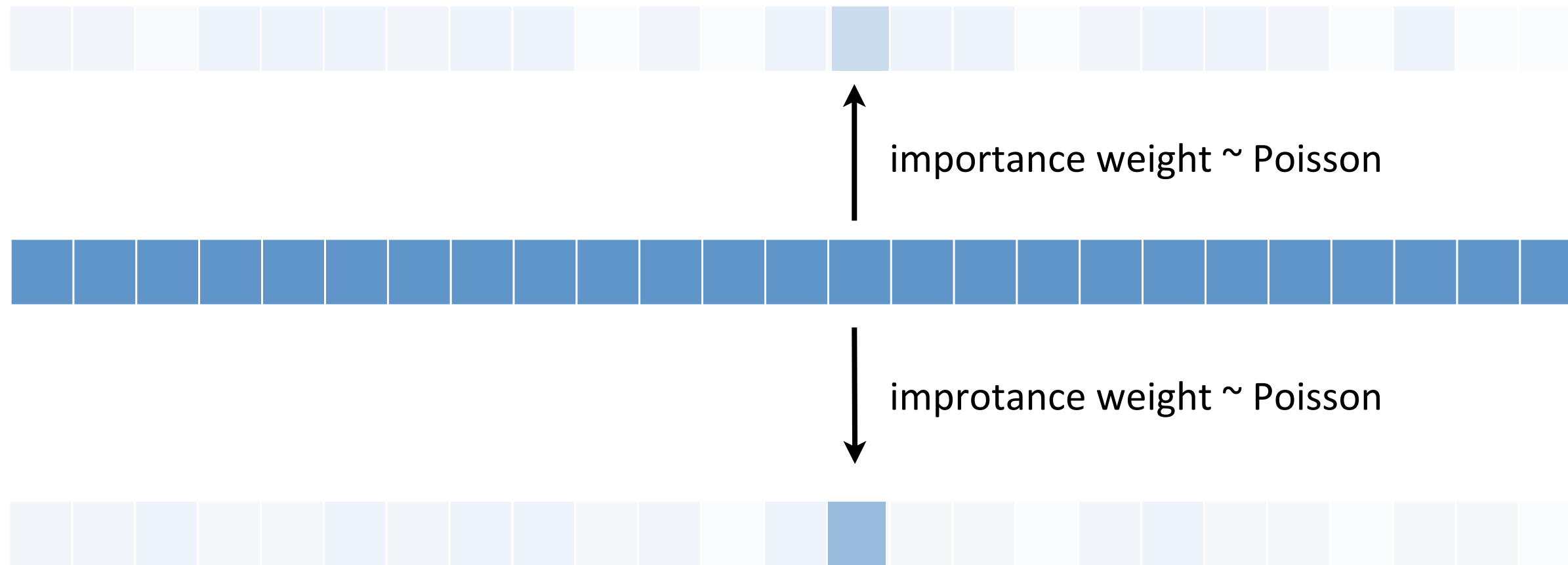




**NEW**

--bs: bootstrapping

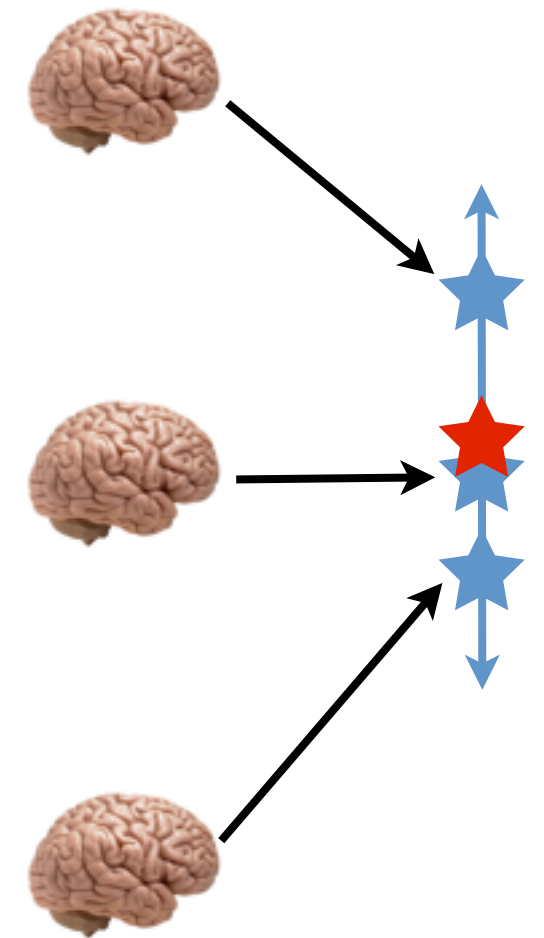
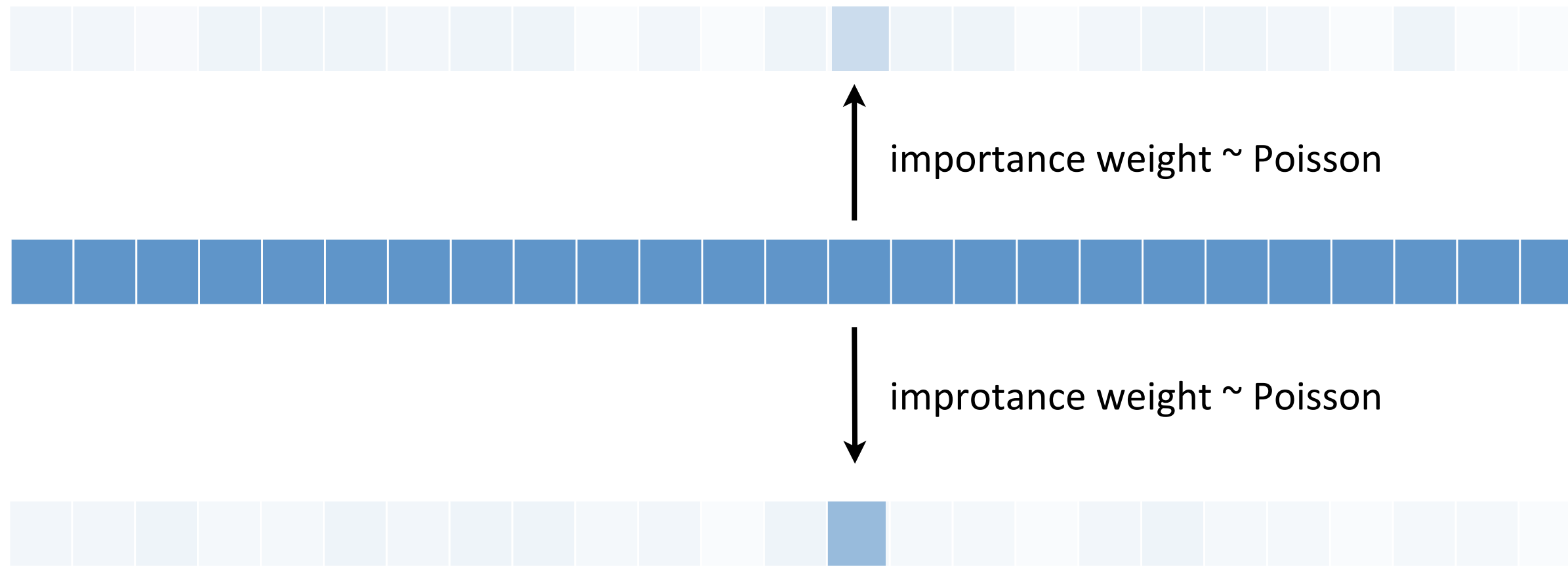
**#2**



**NEW**

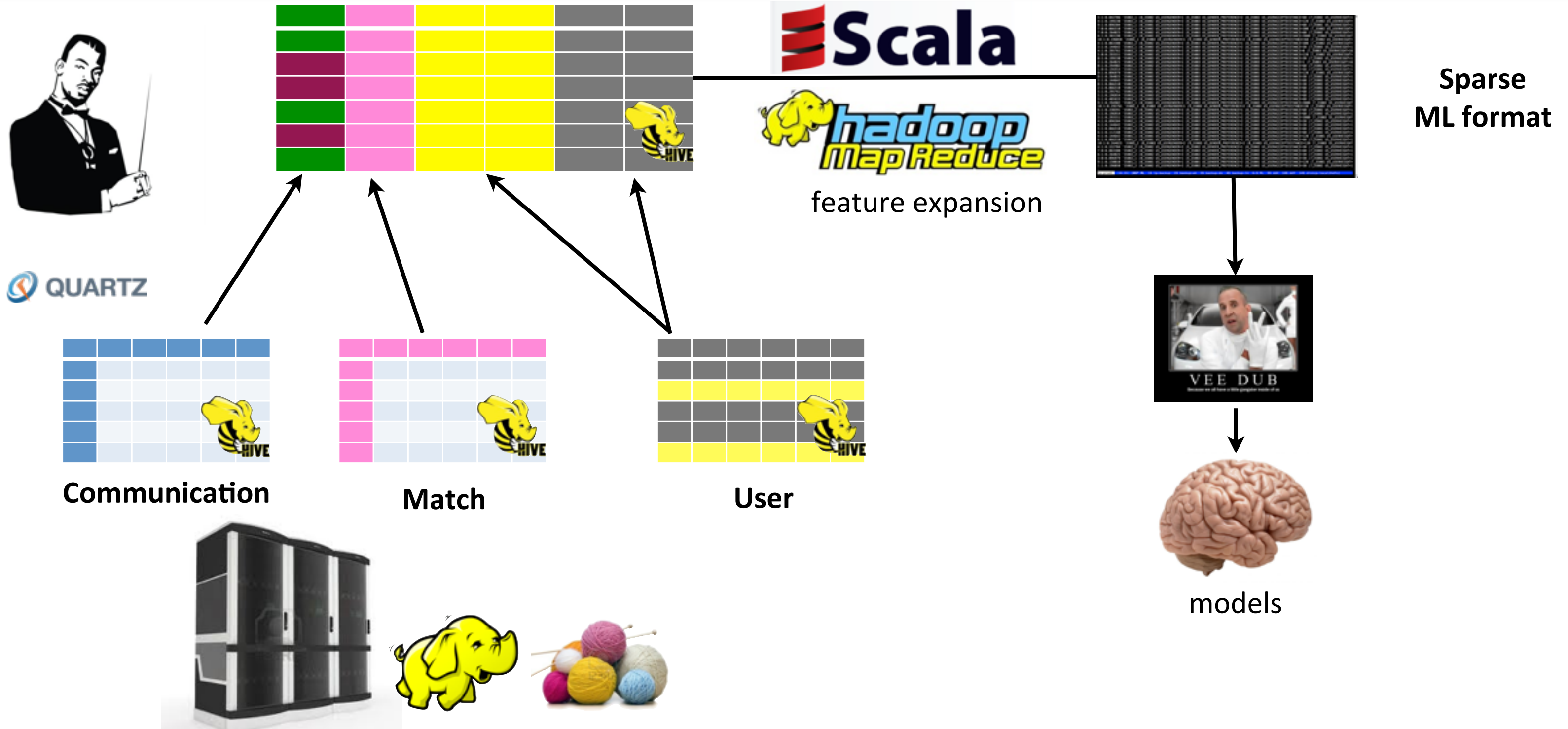
--bs: bootstrapping

**#2**



**better models. measure of uncertainty. superlinear speedup.**

# Modeling: Maestro

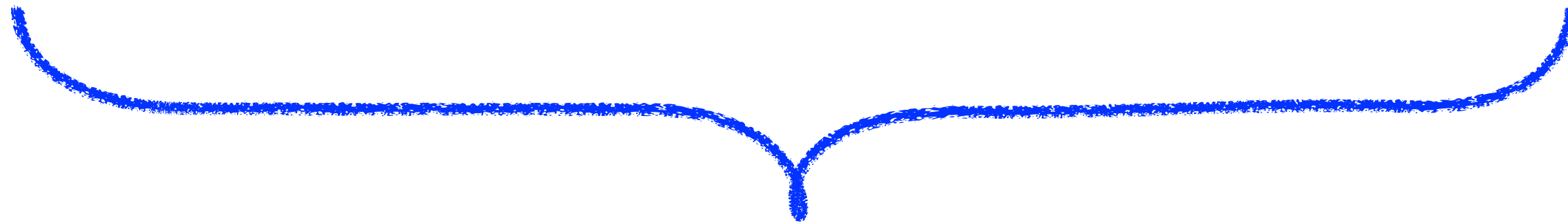
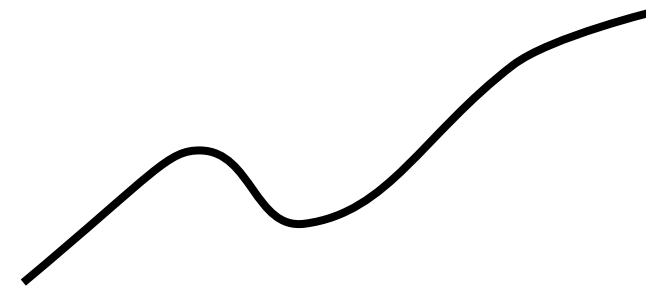


# Modeling: Model parametrizations

## Model parameters

features  
weights  
tree splits

## Calibration Spline



# Modeling: Model parametrizations

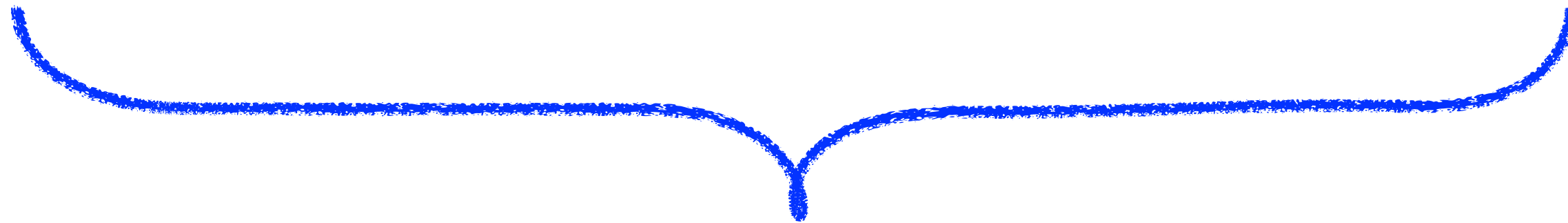
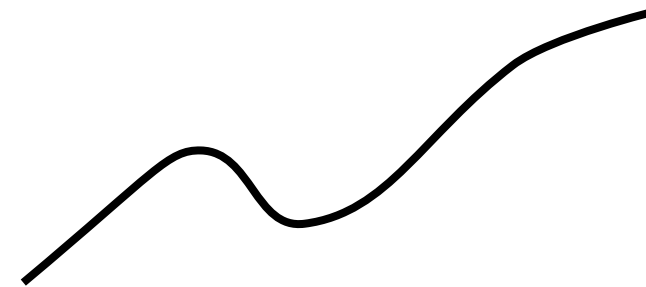
## Model parameters

features  
weights  
tree splits



Scala  
DSL

## Calibration Spline



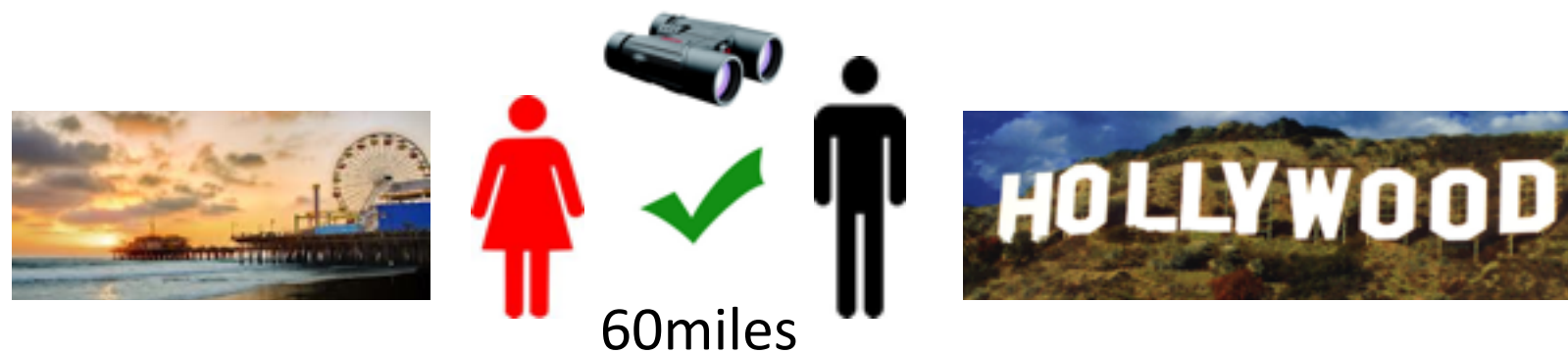
# Affinity Matching: Scala DSL



“same\_religion”:”`${user.profile.religion}=={cand.profile.religion}`”



“cmp\_drinking”:”`cmp(${user.profile.drinking},{cand.profile.drinking})`”



60miles

“strict\_distance\_u”:”`${user.profile.accepted_distance}<={pairing.distance}`”

# Production: Spring Conductor



Spring Batch Admin

Home Jobs Executions Files

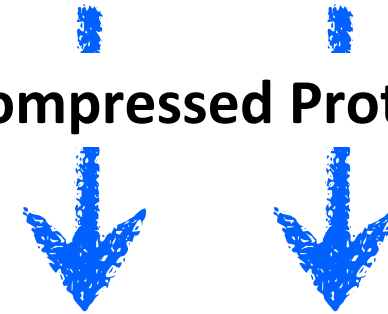
Job Names Registered

Name	Description	Execution Count	Launchable	Incrementable
dataReport	No description	17	true	true
consolidateMtsFiles	No description	1,213	true	true
quartzManualTrigger	No description	12	true	true
userProtoToTax	No description	56	true	true
queueModelTraining	No description	18	true	true
sqoopDataLoading	No description	343	true	true
queueUserProtoToTax	No description	4	true	true
modelInteroperabilityTest	No description	2	true	true
offlineDataWarehousing	No description	304	true	true
netezzaDataWarehousing	No description	257	true	true
modelRetrainingJob	No description	66	true	true
warehouseActData	No description	4	true	true
createCompetitionDataset	No description	5	true	true

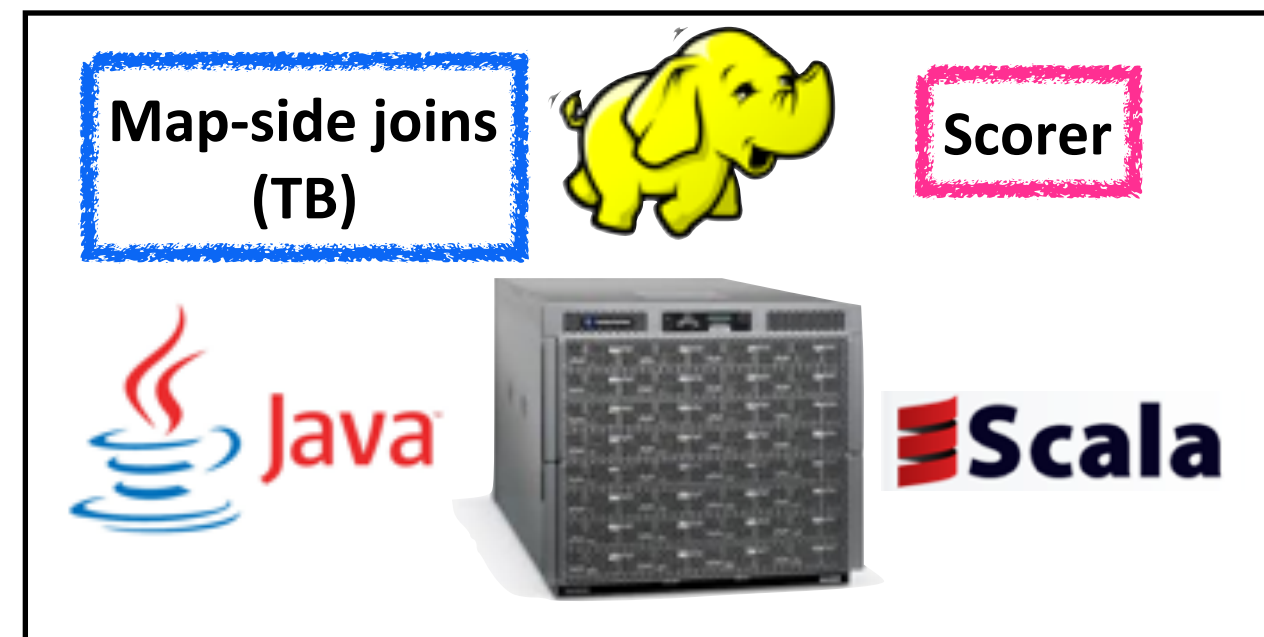
## Matching User Service



1+G Compressed Protocol Buffers



750M Compressed Protocol Buffers



## Pairings Browser Service





```
[User[Demographic][Photo][Activity][FX]]  
[Cand[Demographic][Photo] [Activity][FX]]  
[Pairing[Distance][Flags]]
```

```
[User[Demographic][Photo][Activity][FX]]  
[Cand[Demographic][Photo] [Activity][FX]]  
[Pairing[Distance][Flags]]
```

```
[User[Demographic][Photo][Activity][FX]]  
[Cand[Demographic][Photo] [Activity][FX]]  
[Pairing[Distance][Flags]]
```

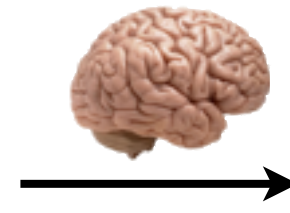
...

```
[User[Demographic][Photo][Activity][FX]]  
[Cand[Demographic][Photo] [Activity][FX]]  
[Pairing[Distance][Flags]]
```



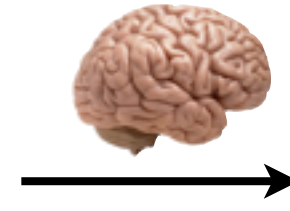


[User[Demographic][Photo][Activity][FX]]  
[Cand[Demographic][Photo] [Activity][FX]]  
[Pairing[Distance][Flags]]



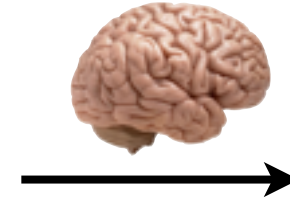
Prob( | data)

[User[Demographic][Photo][Activity][FX]]  
[Cand[Demographic][Photo] [Activity][FX]]  
[Pairing[Distance][Flags]]



Prob( | data)

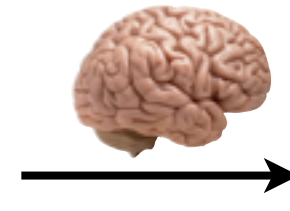
[User[Demographic][Photo][Activity][FX]]  
[Cand[Demographic][Photo] [Activity][FX]]  
[Pairing[Distance][Flags]]



Prob( | data)

...

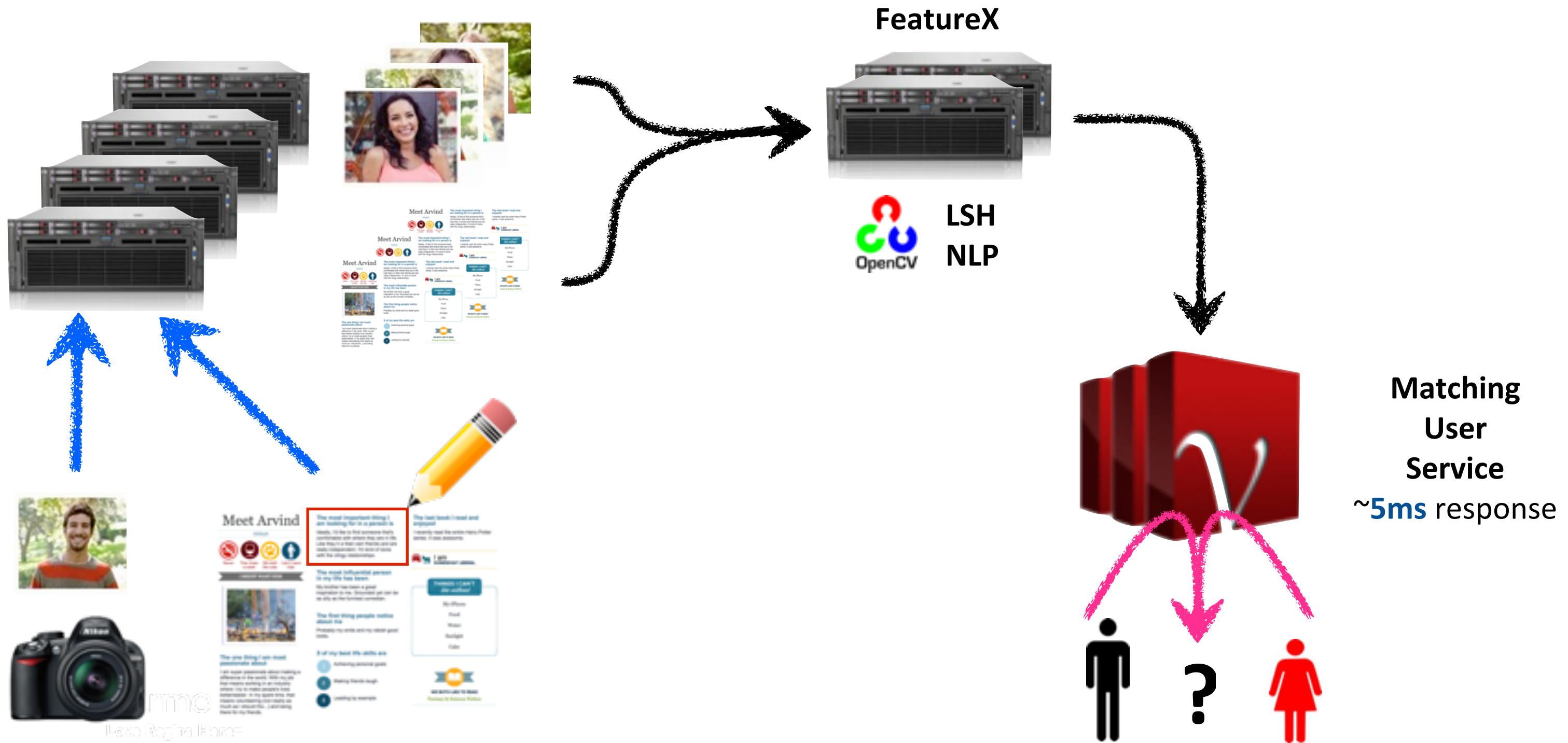
[User[Demographic][Photo][Activity][FX]]  
[Cand[Demographic][Photo] [Activity][FX]]  
[Pairing[Distance][Flags]]



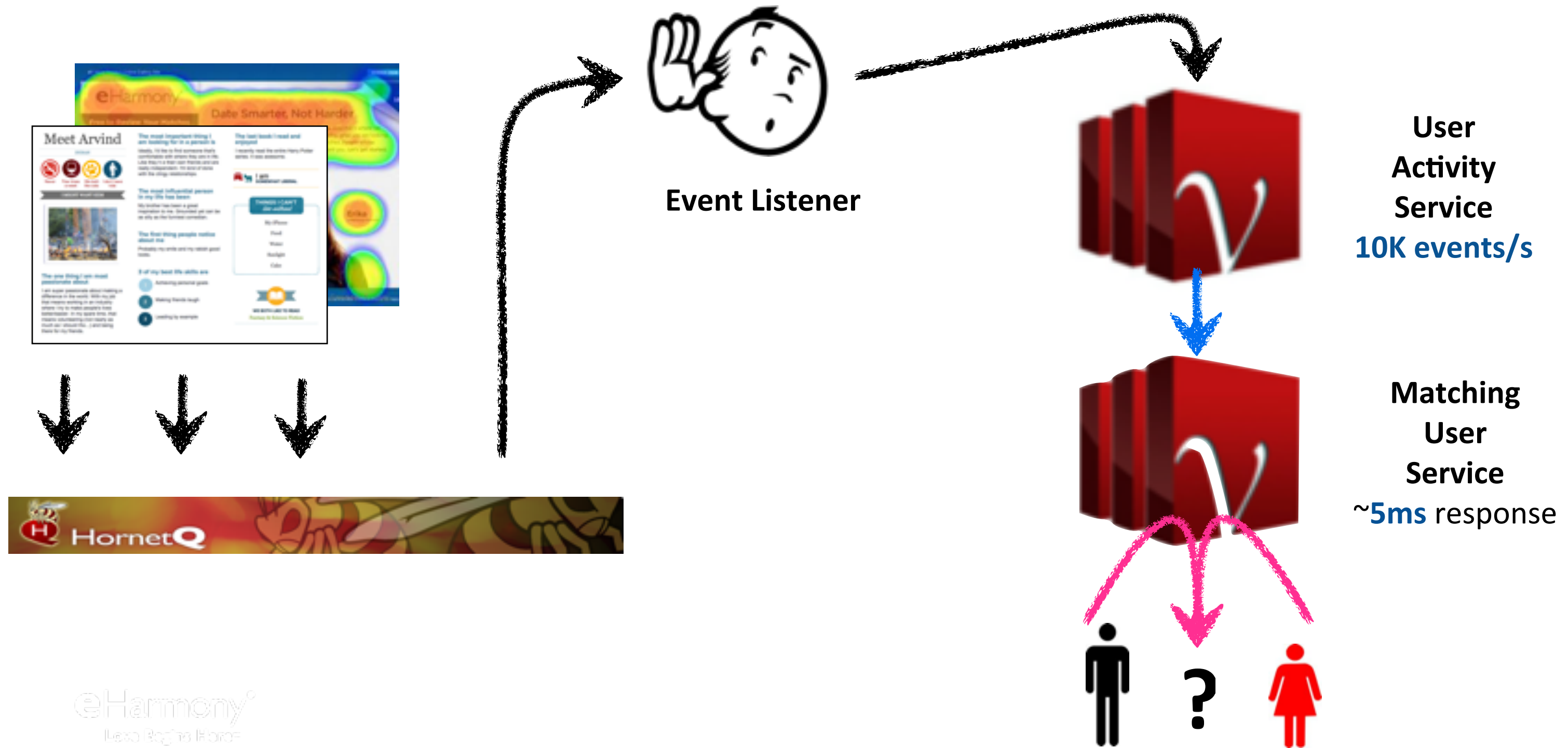
Prob( | data)



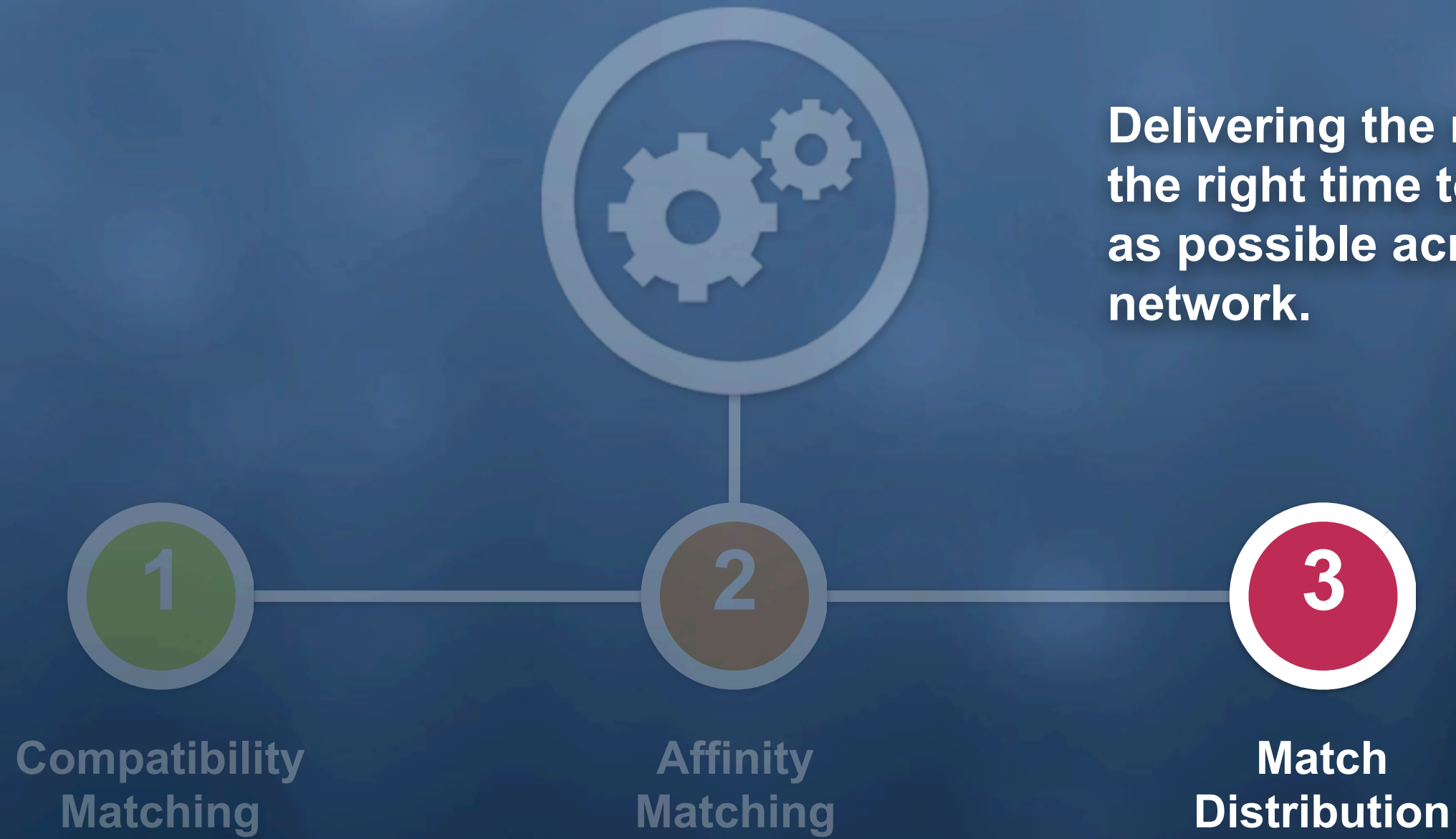
# Production: FeatureX for expensive features



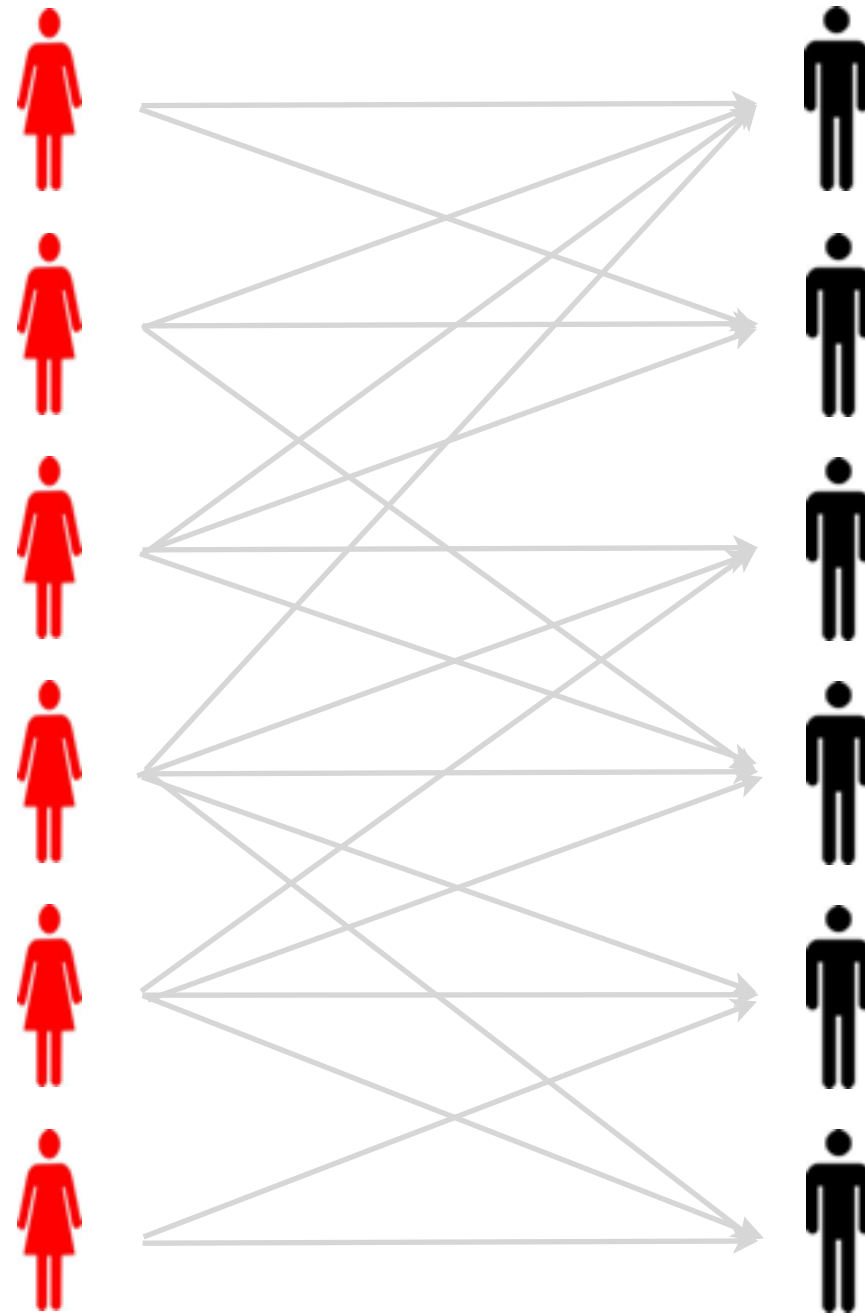
# Production: User Activity Service





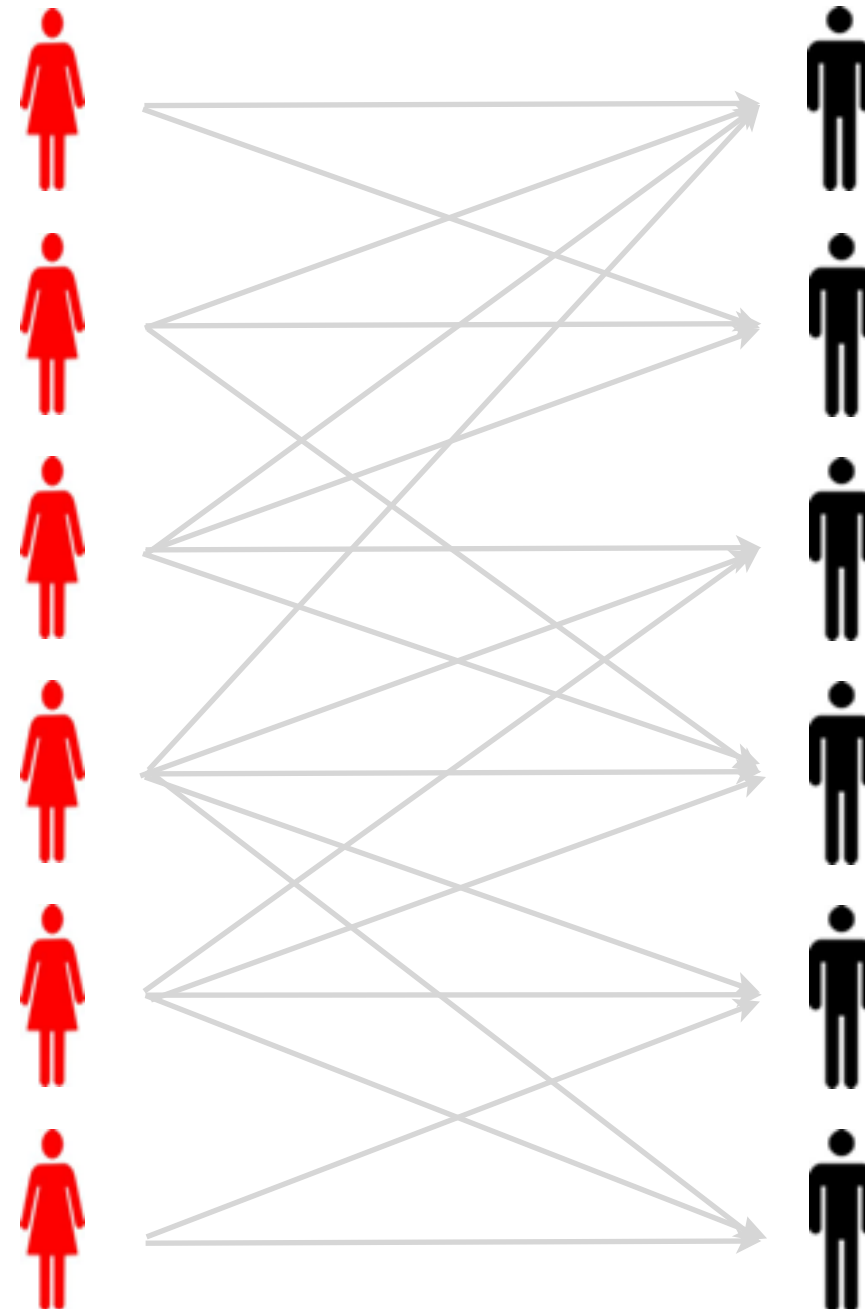








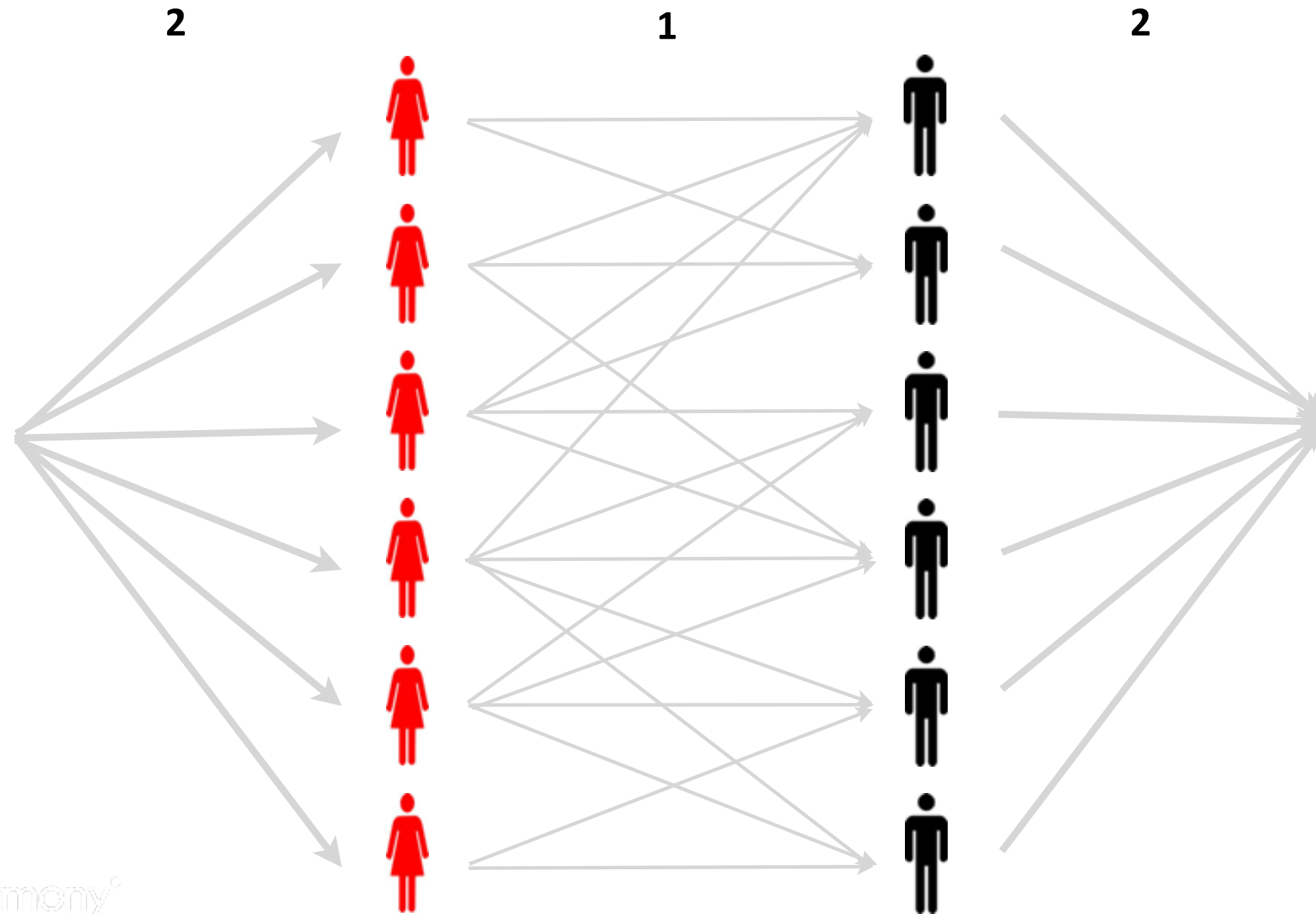
2



2

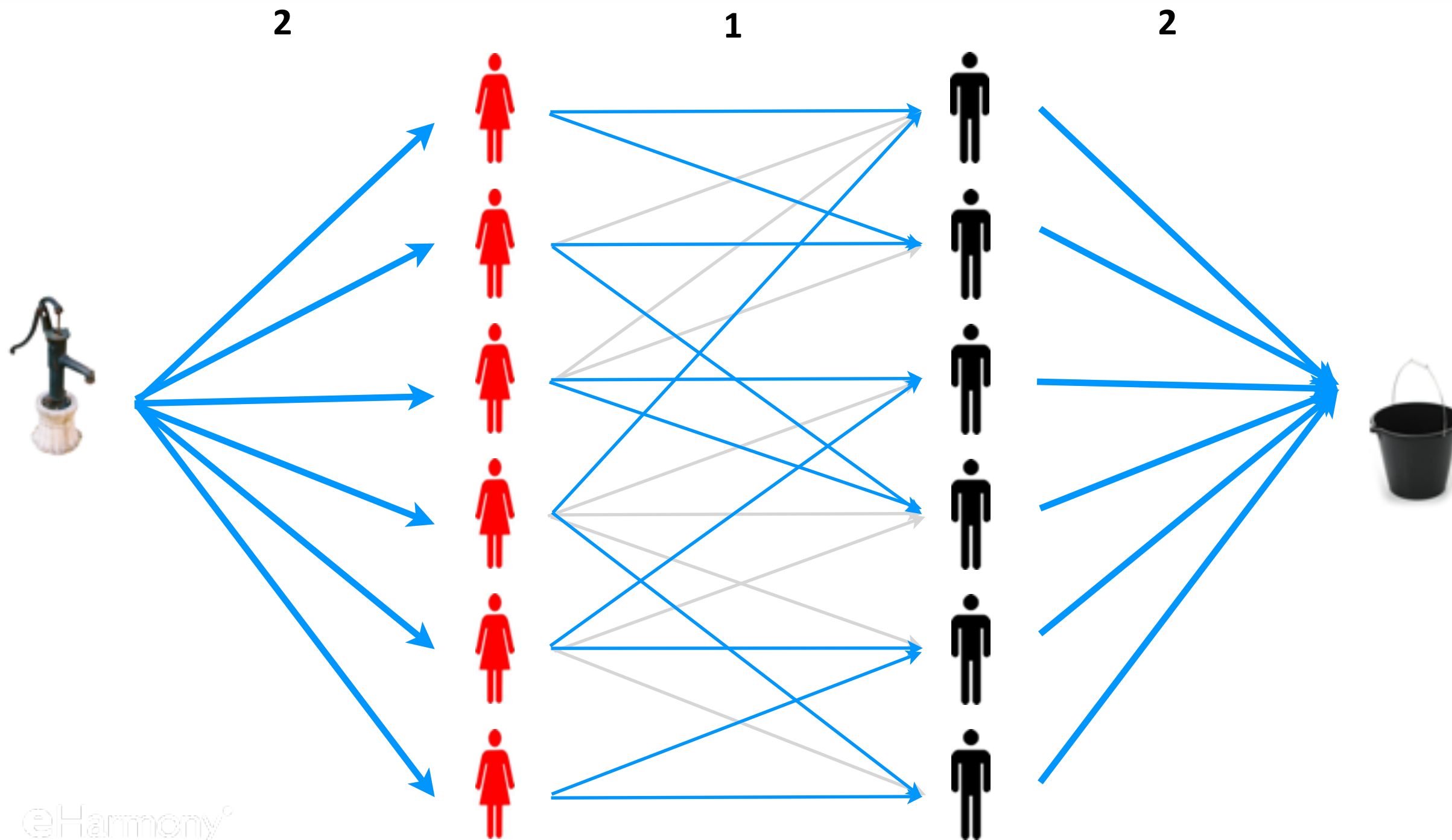
# Match Distribution >

Graph optimization



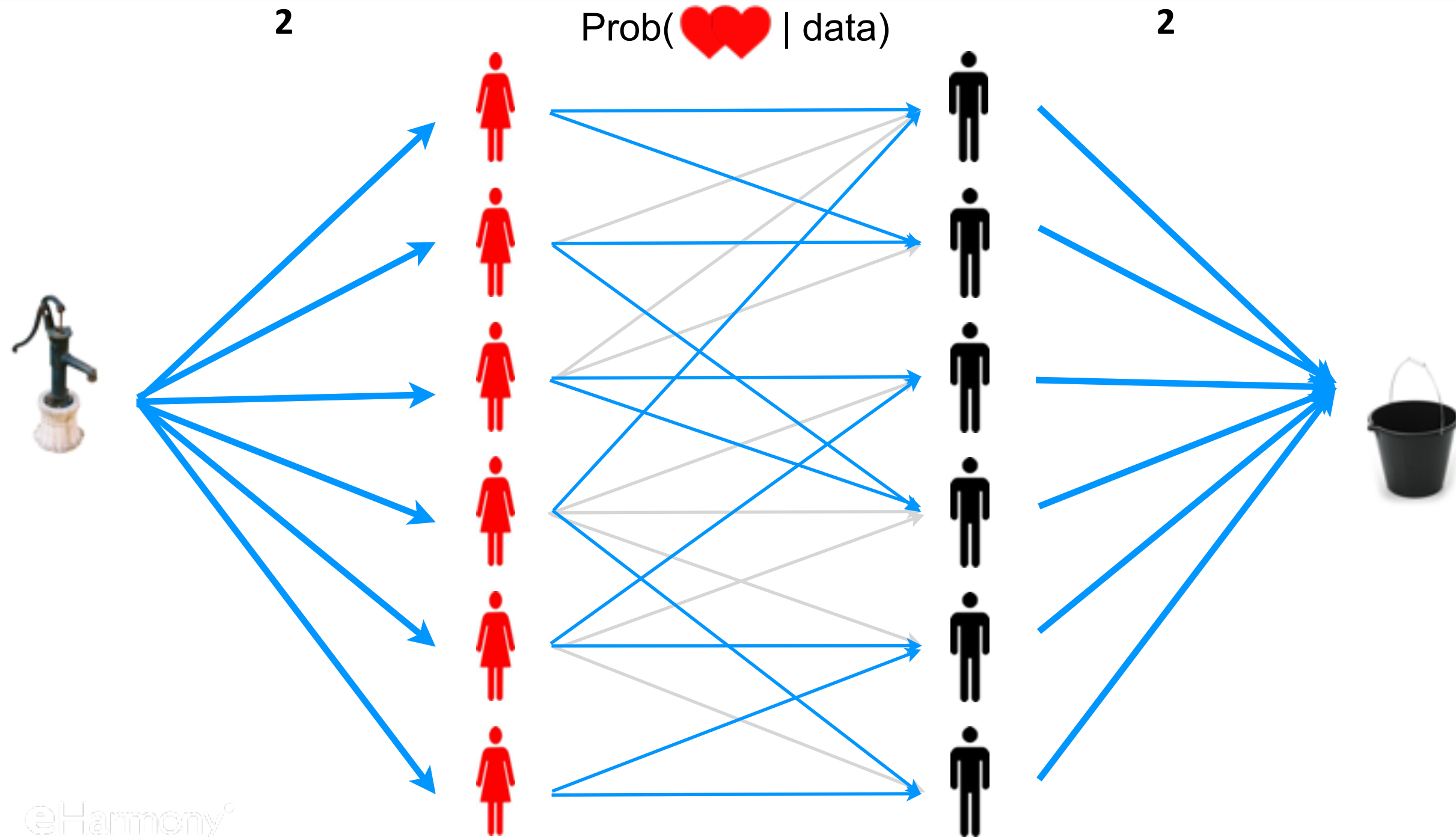
# Match Distribution >

Graph optimization

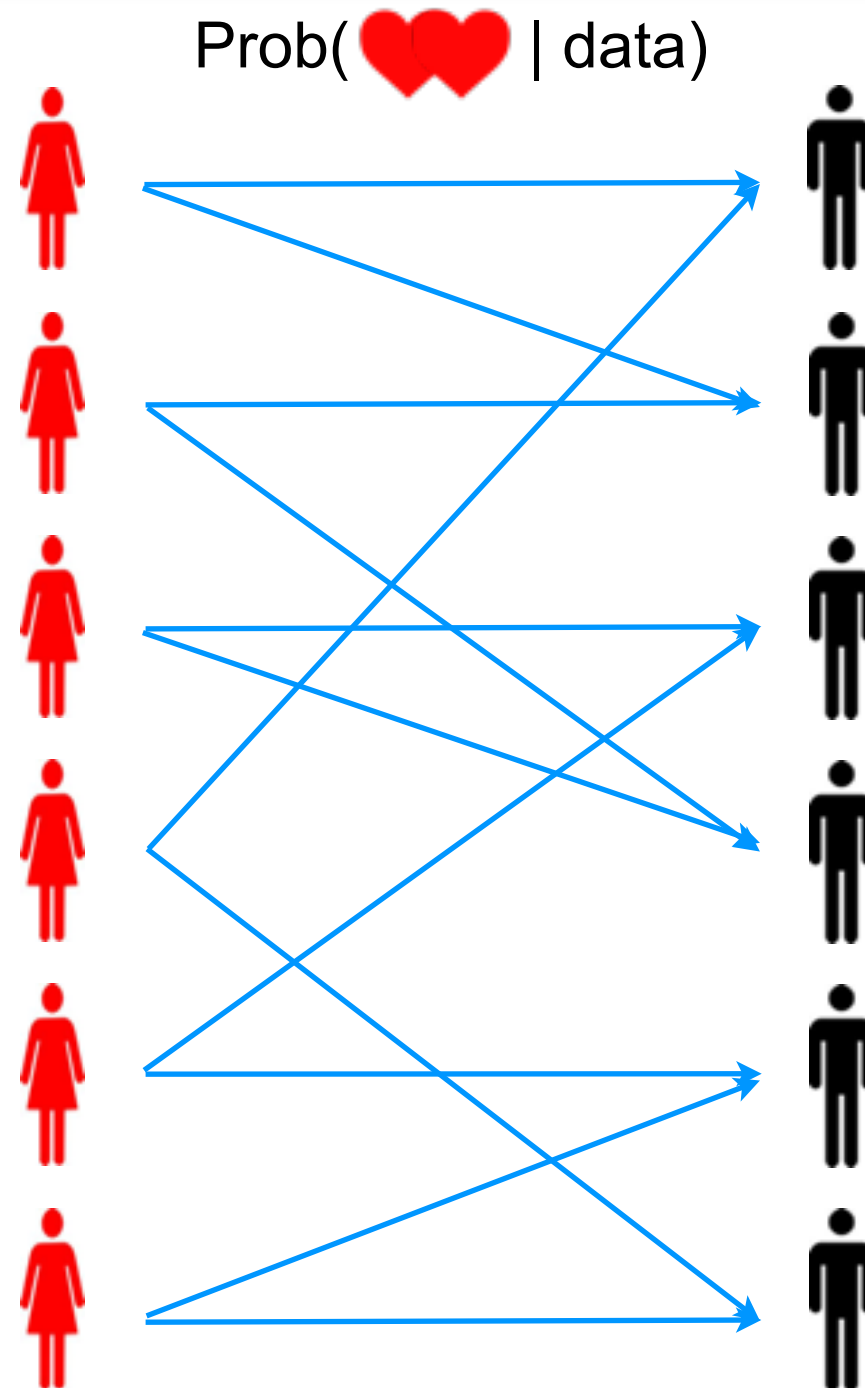


# Match Distribution >

Graph optimization



2



2

Resulting Customer Experience ›

Guided Communication

Resulting Customer Experience >

Guided Communication



Resulting Customer Experience ›

Guided Communication





Resulting Customer Experience ›

Success!



# Resulting Customer Experience ›

Success!





2005



90

eHarmony Members  
Married Every Day



2005

2007



**236**

**eHarmony Members  
Married Every Day**





2005

2007

2009



**542**

**eHarmony Members  
Married Every Day**



## Marital satisfaction and break-ups differ across on-line and off-line meeting venues

John T. Cacioppo<sup>a,1</sup>, Stephanie Cacioppo<sup>a</sup>, Gian C. Gonzaga<sup>b</sup>, Elizabeth L. Ogburn<sup>c</sup>, and Tyler J. VanderWeele<sup>c</sup>

<sup>a</sup>Department of Psychology, Center for Cognitive and Social Neuroscience, University of Chicago, Chicago, IL 60637; <sup>b</sup>Gestalt Research, Santa Monica, CA 90403; and <sup>c</sup>Department of Epidemiology, Harvard University, Boston, MA 02115

Edited by Linda M. Bartoshuk, University of Florida, Gainesville, FL, and approved May 1, 2013 (received for review December 24, 2012)

Marital discord is costly to children, families, and communities. The advent of the Internet, social networking, and on-line dating has affected how people meet future spouses, but little is known about the prevalence or outcomes of these marriages or the demographics of those involved. We addressed these questions in a nationally representative sample of 19,131 respondents who married between 2005 and 2012. Results indicate that more than one-third of marriages in America now begin on-line. In addition, marriages that began on-line, when compared with those that began through traditional off-line venues, were slightly less likely to result in a marital break-up (separation or divorce) and were associated with slightly higher marital satisfaction among those respondents who remained married. Demographic differences were identified between respondents who met their spouse through on-line vs. traditional off-line venues, but the findings for marital break-up and marital satisfaction remained significant after statistically controlling for these differences. These data suggest that the Internet may be altering the

because on-line venues have tended to be treated as a homogenous terrain (2) despite on-line venues having grown in number, variety, and complexity.

### Results

The demographic characteristics of the respondents who married between 2005 and 2012 as well as US Census data for married individuals indicated that the weighted sample of 19,131 respondents was generally representative (Table S1). For each marriage, participants were asked the month and year of the marriage and, if the most recent marriage ended in divorce, the month and year of the divorce. As summarized in Fig. 1A, 92.01% of the sample reported being currently married, 4.94% reported being divorced, 2.50% reported being separated from their spouse, and 0.55% reported being widowed (7). As in prior research (2), marital break-ups were defined as separated or divorced and constituted 7.44% of the sample.

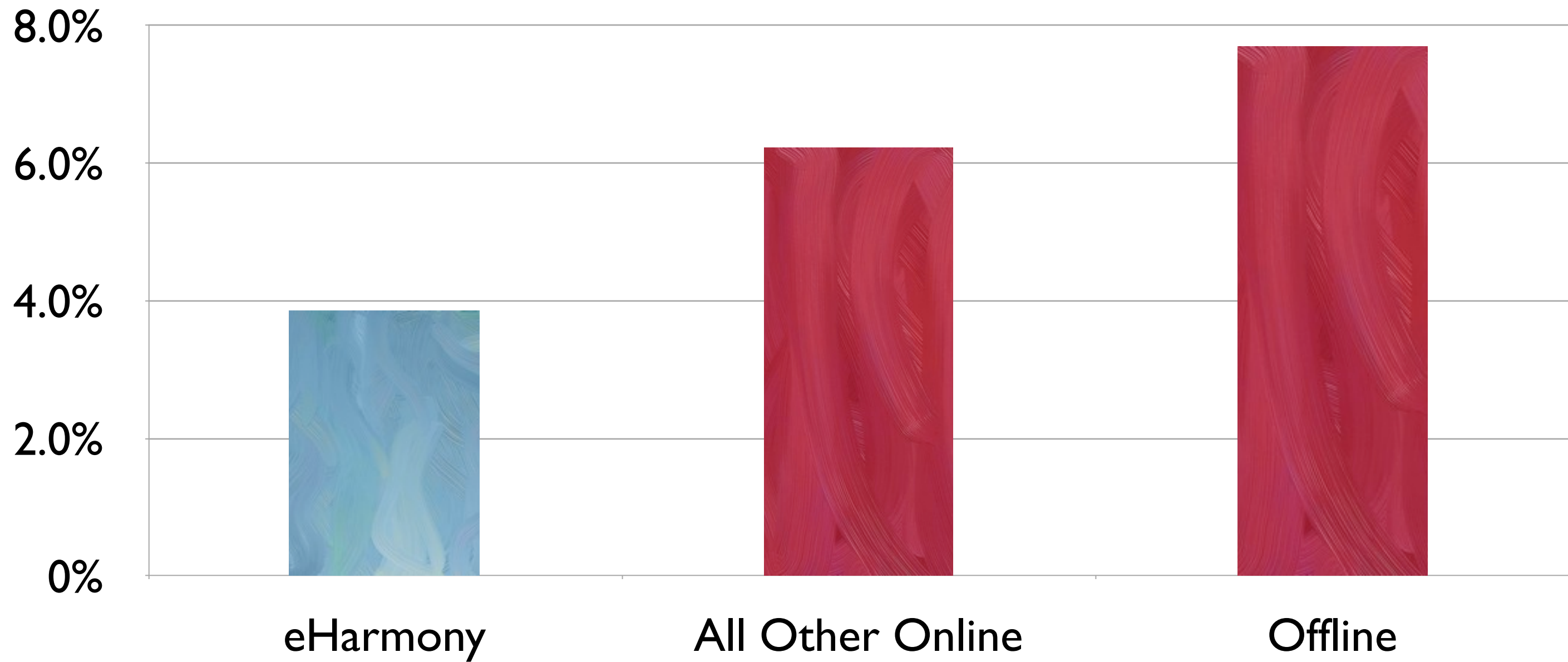


Since 2005, about **1 in 3** couples  
who have married in the US  
have met online (35%)



The largest number  
of marriages surveyed  
who met via online dating  
had met on **eHarmony (25%)**

# Rates of breakup or divorce



\* according to survey of couples married between 2005-2012 by Harris Interactive for eHarmony

[bit.ly/jobateharmony](http://bit.ly/jobateharmony)



\* according to survey of couples married between 2005-2012 by Harris Interactive for eHarmony