



Java and Data Science in Digital Advertising

Fabiane Bizinella Nardon

@fabianenardon

Chief Data Scientist at TailTarget



What do you know about digital advertising?

1. I have no idea how it works
2. I know something about it, but I'm not in the field
3. I deal with DSPs, SSPs and DMPs everyday



**“The best minds of my generation
are thinking about
how to make people click ads.
That sucks.”**

Jeff Hammerbacher



Modern Digital Advertising

Huge Datasets +

(+/-Easy to Measure) Metrics +

Velocity



Digital Advertising



Advertiser

Direct Orders

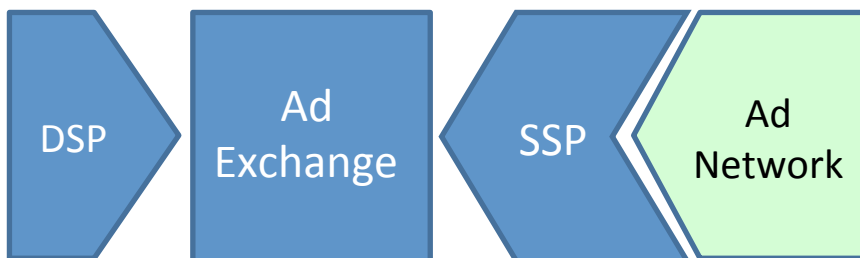




Modern Digital Advertising



Advertiser

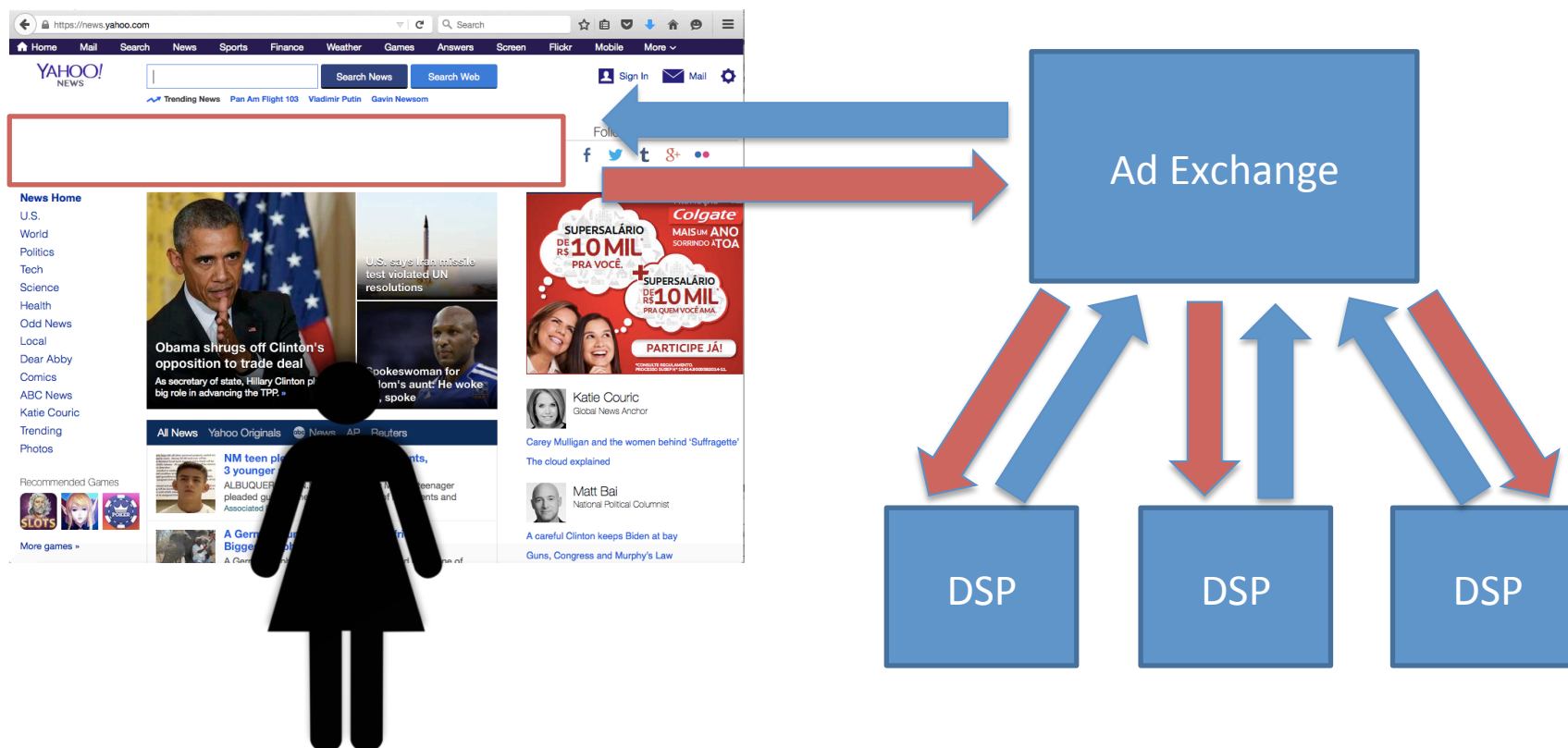


Programmatic Media





Real Time Bidding

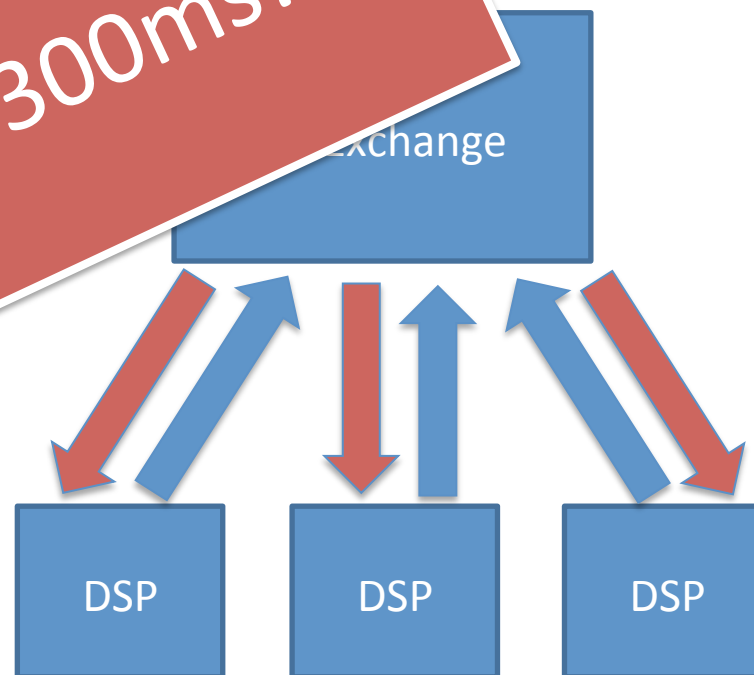




Real Time Bidding



Less than 300ms!





Scalability

Data Science



Architecting for Scalability

Architect for at least 3

Assume something will fail

Isolate your services

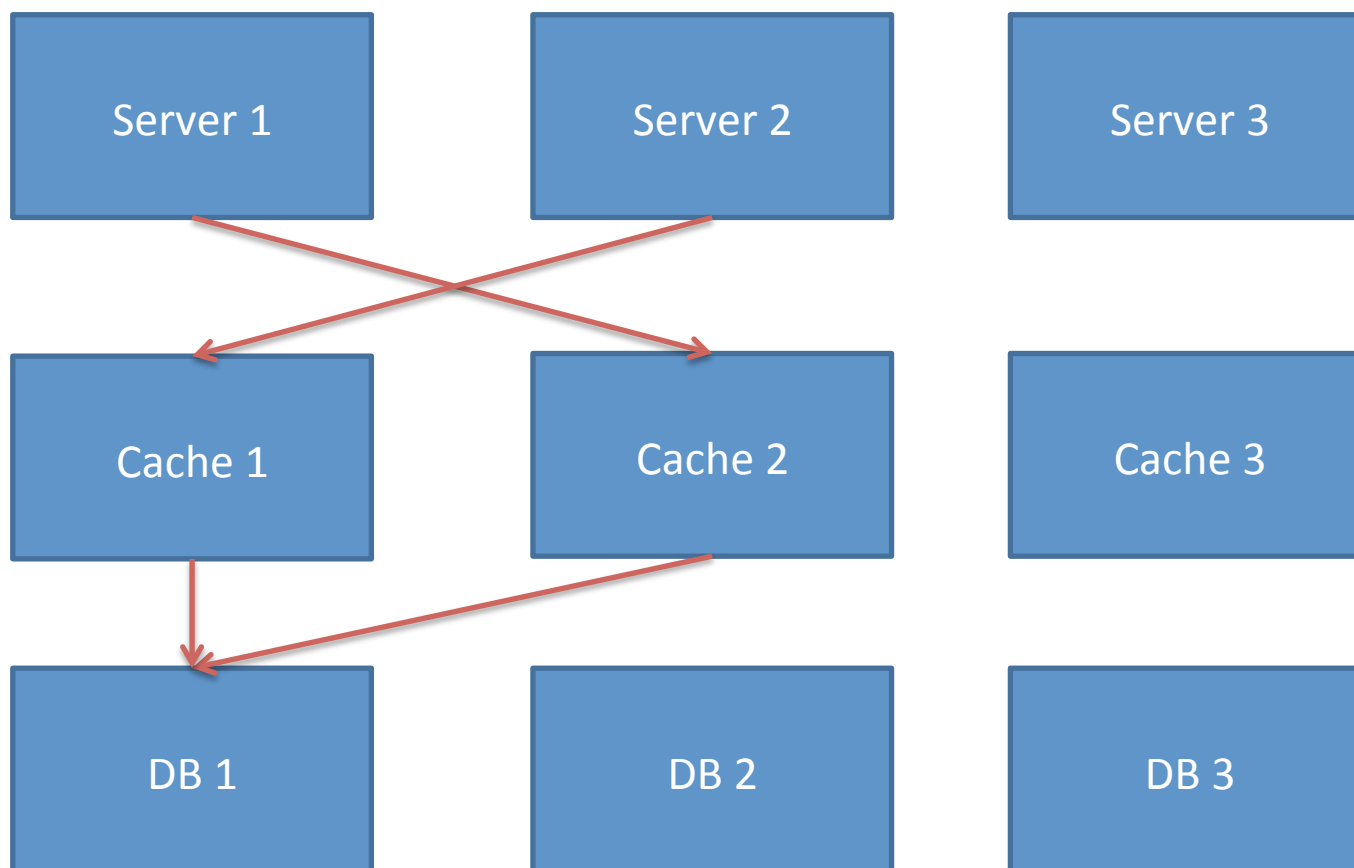


Architecting for Scalability

SHARED NOTHING

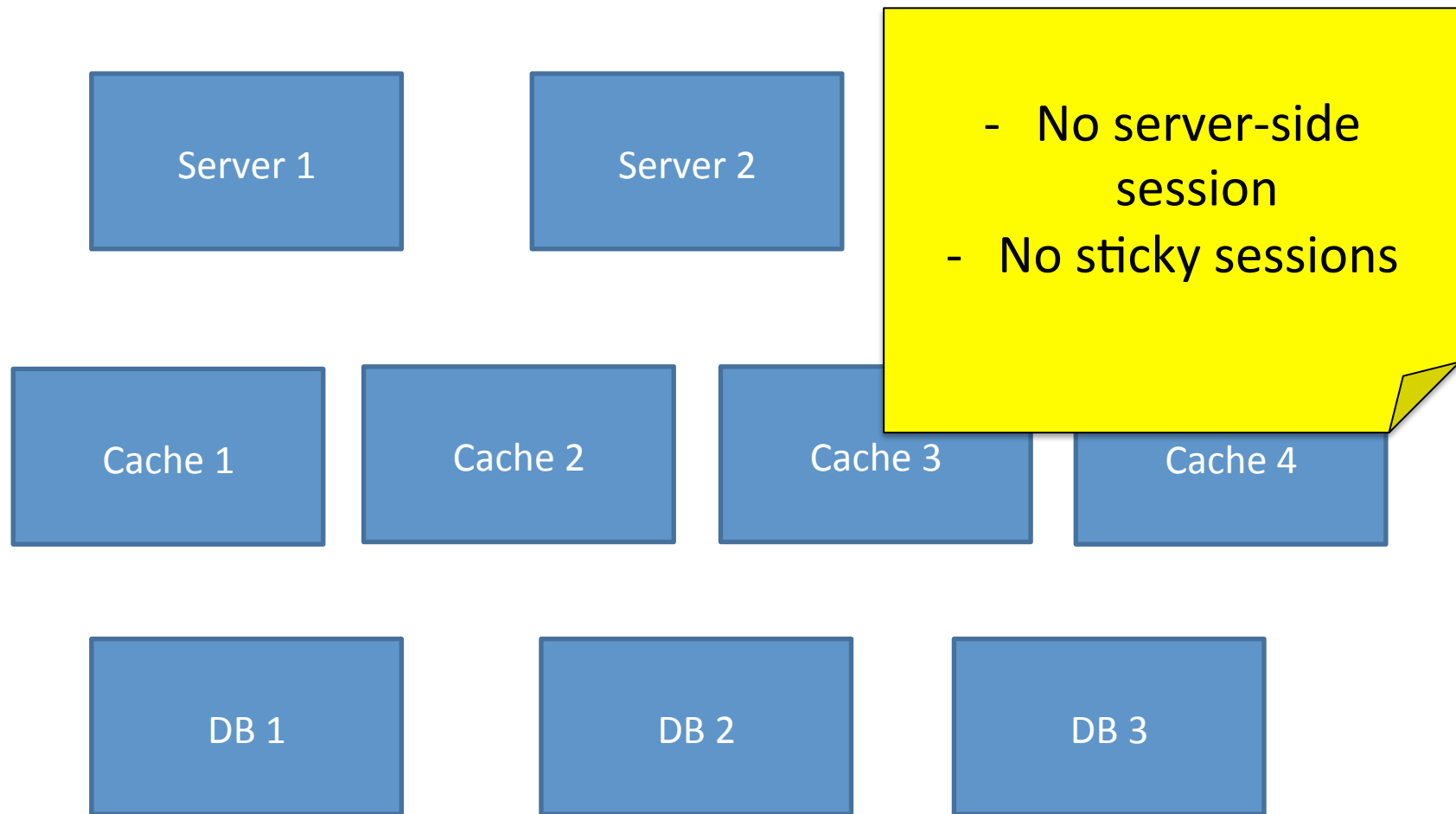


Shared Nothing



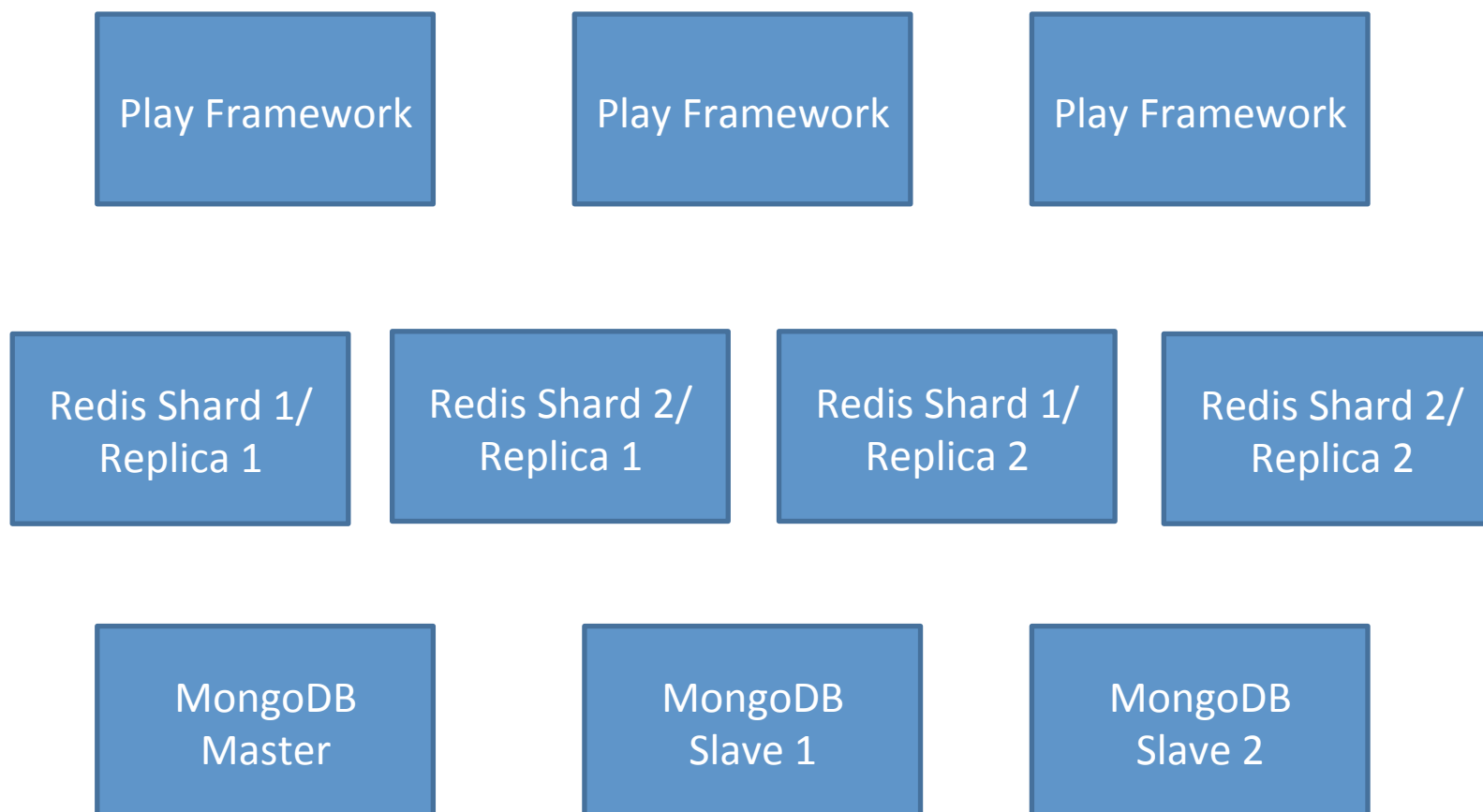


Shared Nothing





Shared Nothing





Scaling

Use the power of
the cloud

but use your
DATA!



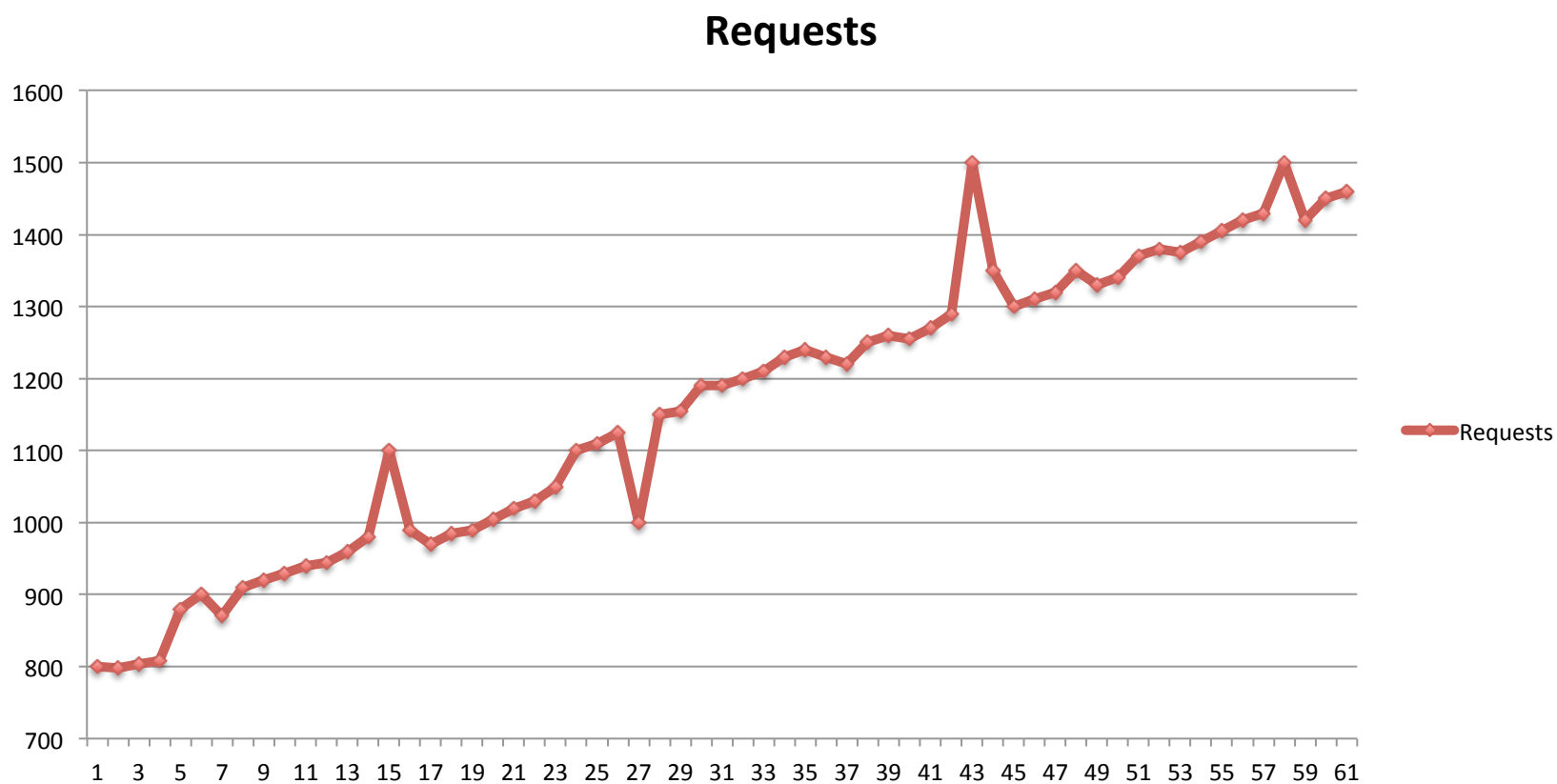
Traffic Prediction – Why?

It can take 10 to 20 min to have a new machine running. Can you wait that much?

Avoid false traffic drops

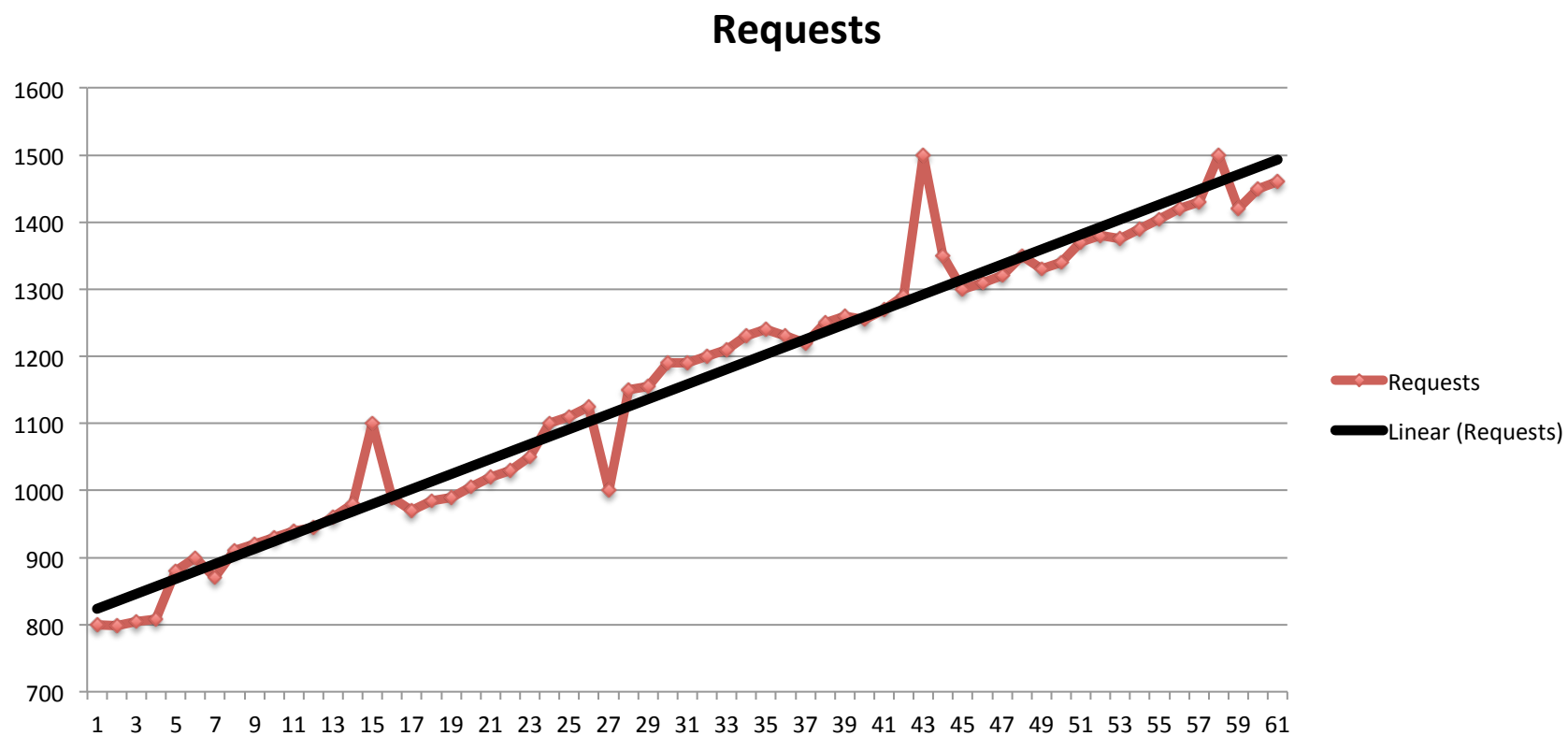


Traffic Prediction



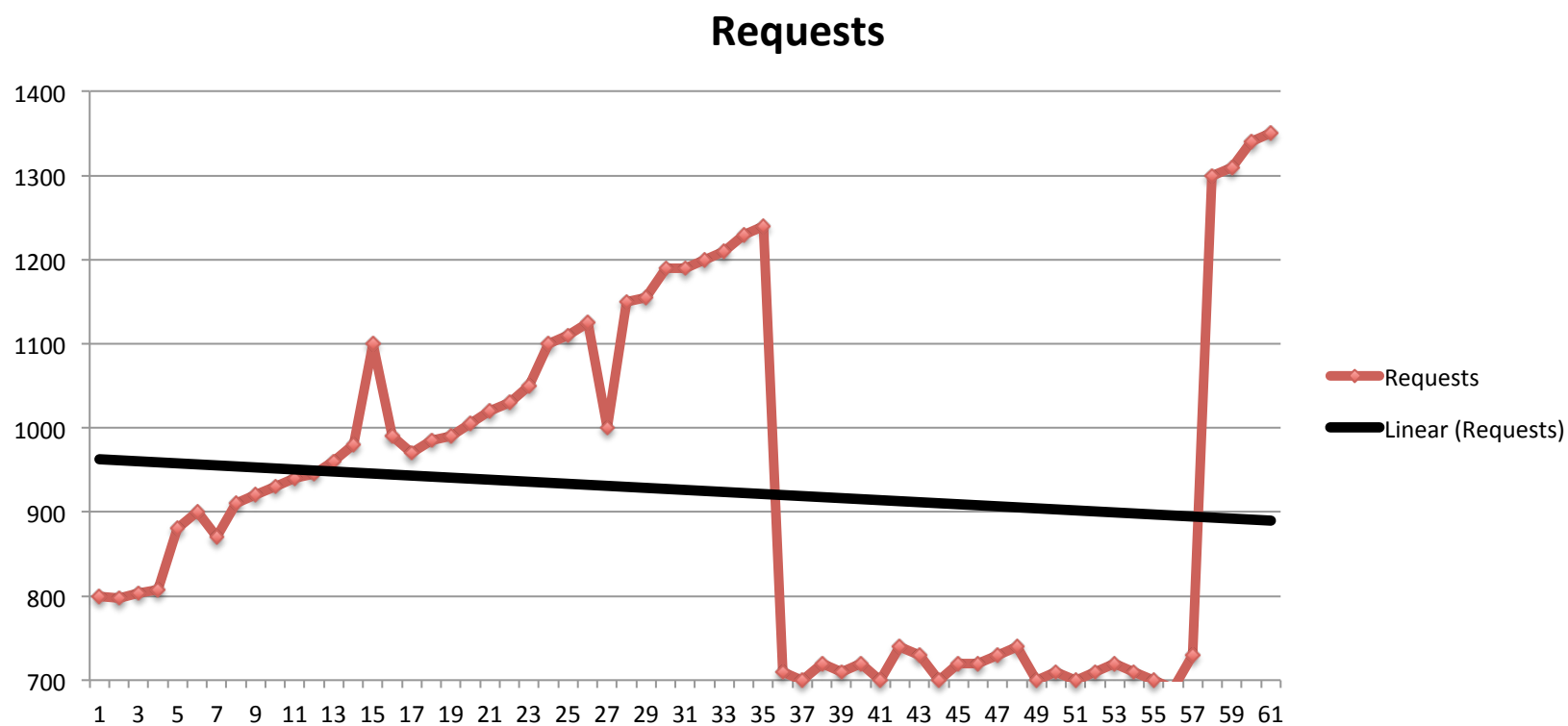


Traffic Prediction



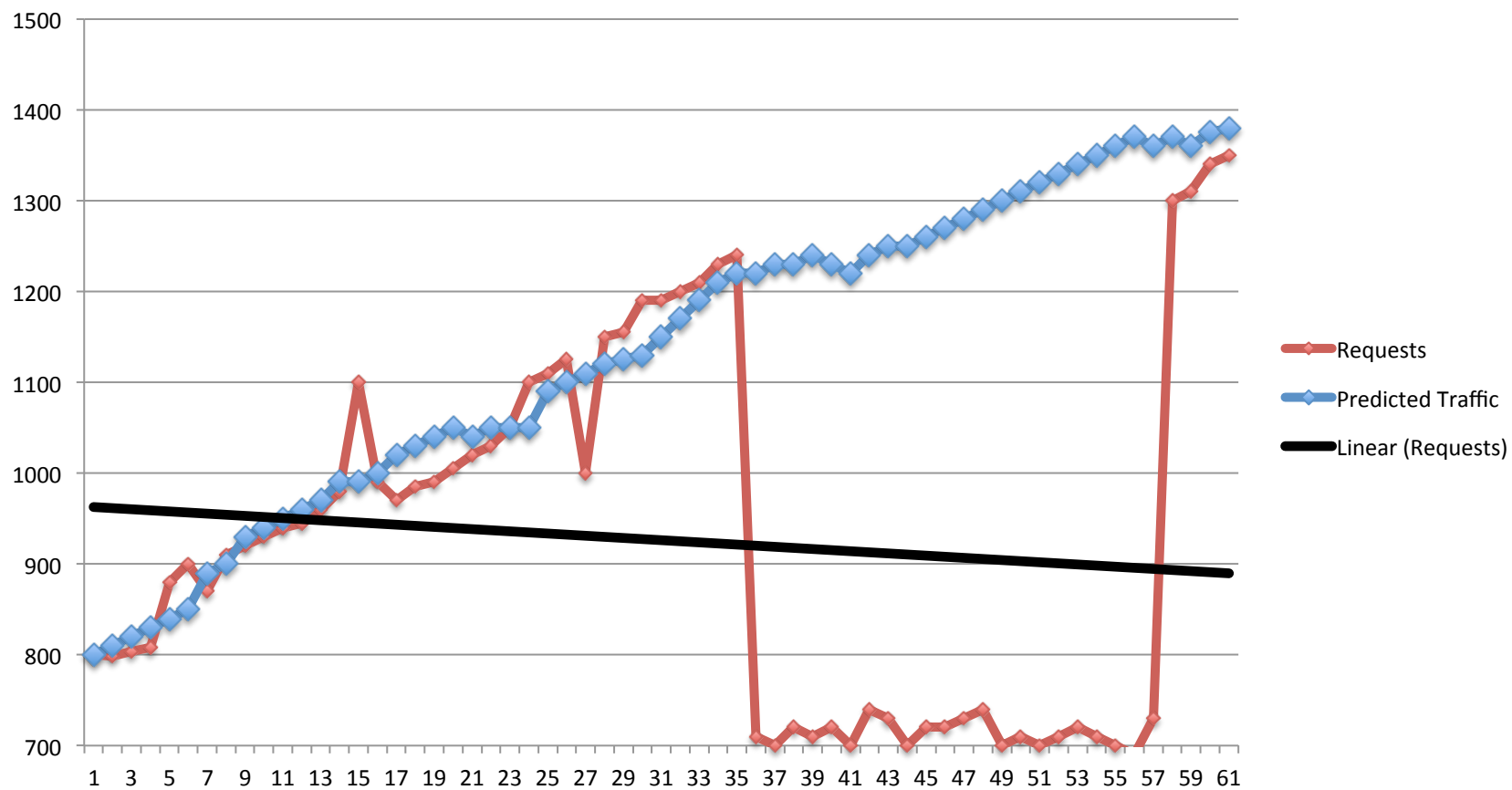


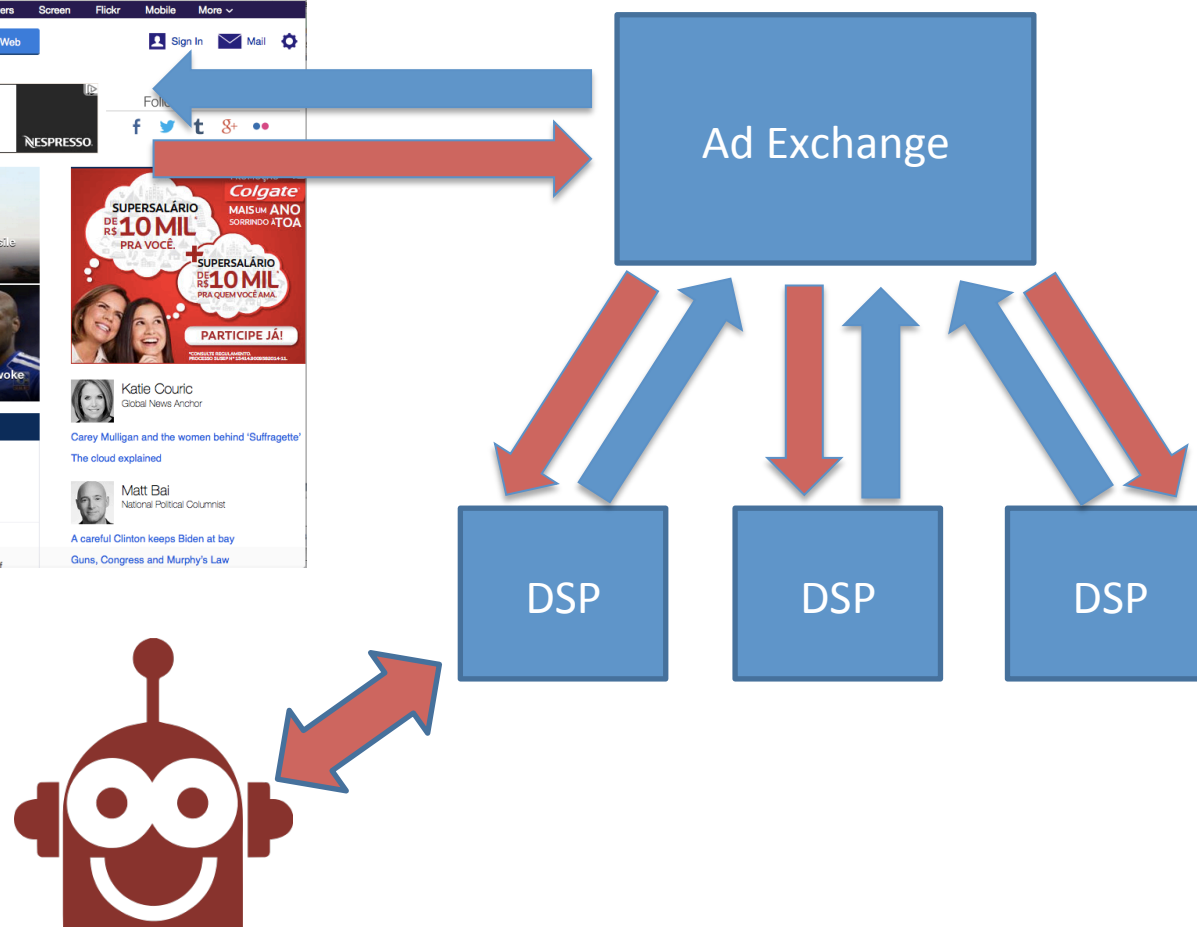
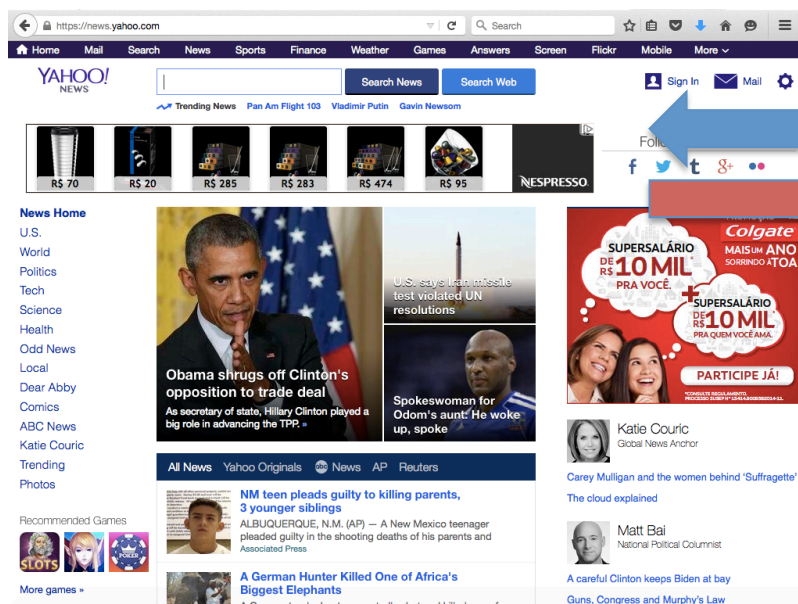
Traffic Prediction





Traffic Prediction







What are the main interests of male Java developers?

- 1. Cars**
- 2. Politics**
- 3. Science**
- 4. Finance**
- 5. Soap Operas**



Data Science

BIG DATA

+

STATISTICS

+

ARTIFICIAL INTELLIGENCE



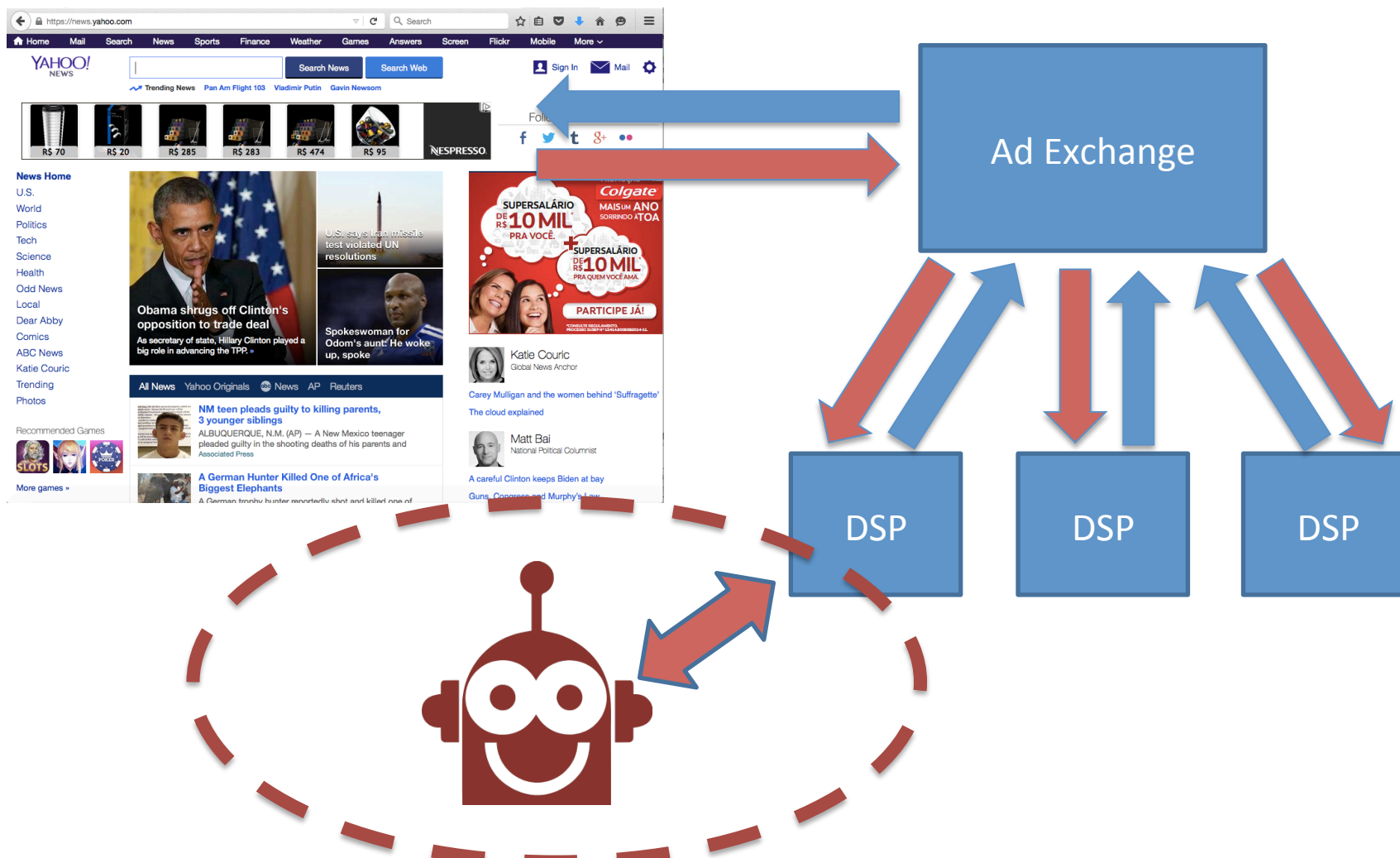
Data Science and Java



DL4J Deep Learning for Java

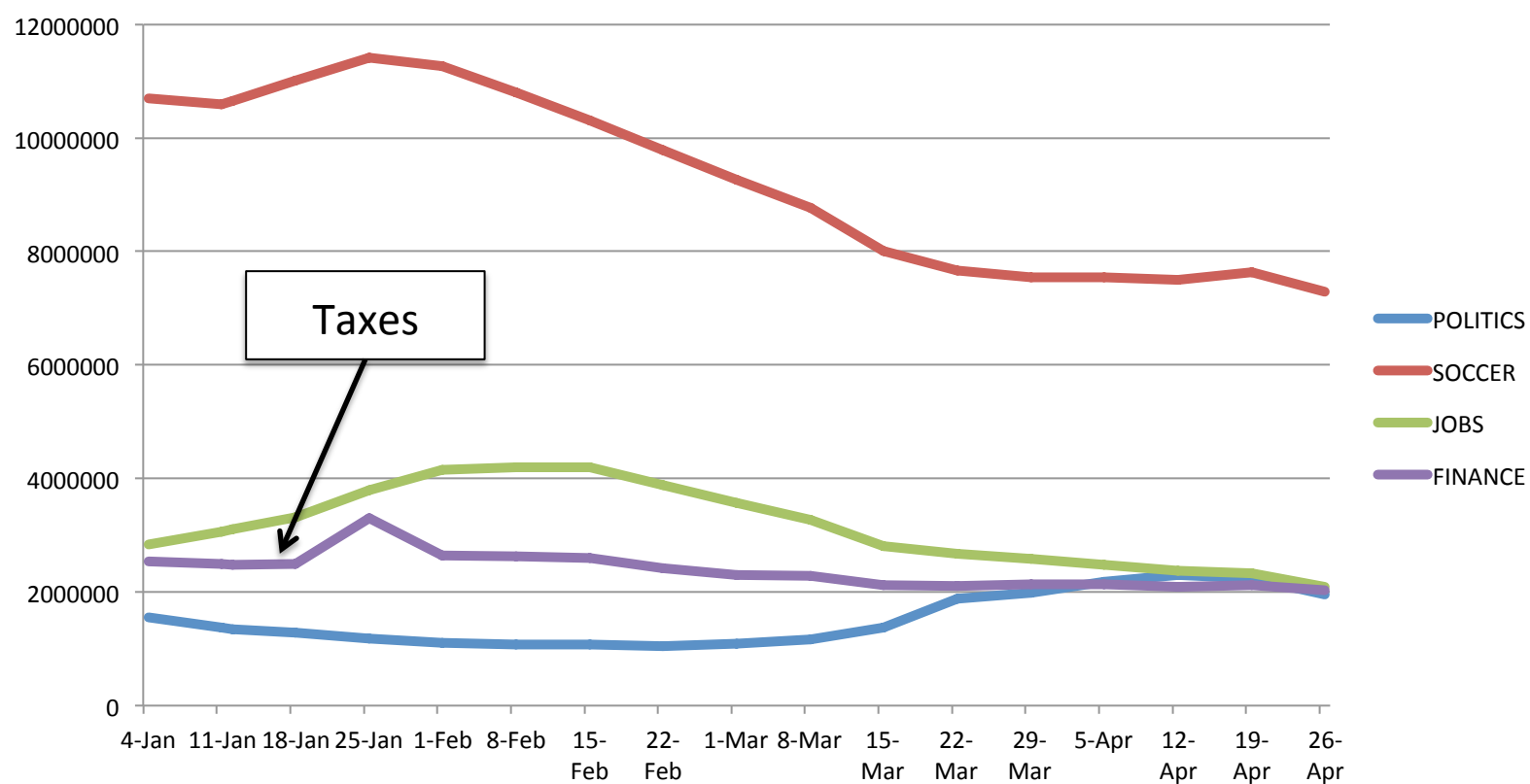


Predicting User Behaviour



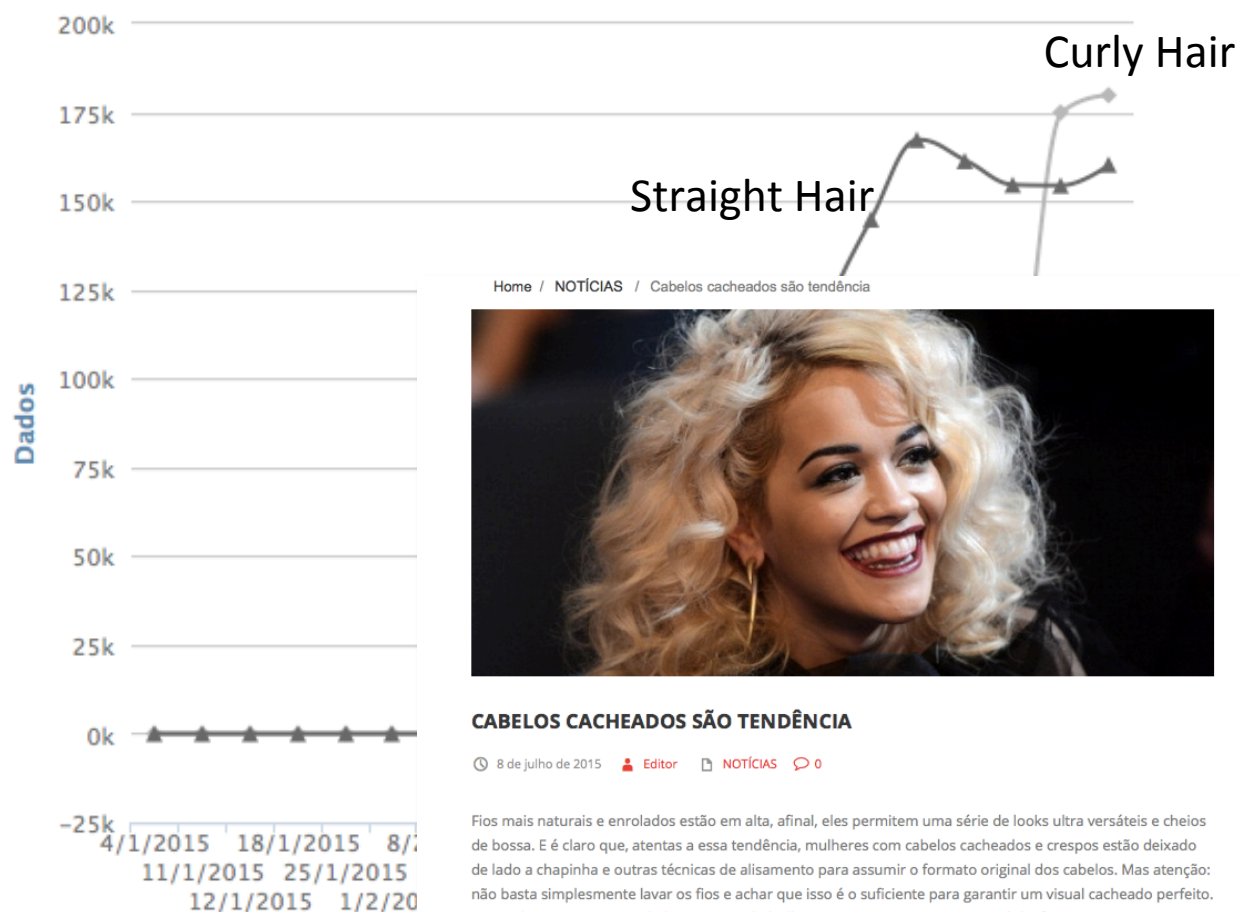


Predicting User Behavior





Predicting User Behavior



Home / NOTÍCIAS / Cabelos cacheados são tendência



CABELOS CACHEADOS SÃO TENDÊNCIA

8 de julho de 2015 Editor NOTÍCIAS 0

Fios mais naturais e enrolados estão em alta, afinal, eles permitem uma série de looks ultra versáteis e cheios de bossa. E é claro que, atentas a essa tendência, mulheres com cabelos cacheados e crespos estão deixando de lado a chapinha e outras técnicas de alisamento para assumir o formato original dos cabelos. Mas atenção: não basta simplesmente lavar os fios e achar que isso é o suficiente para garantir um visual cacheado perfeito. Para valorizar os encaracolados, garantindo brilho, movimento e textura na medida, é preciso investir no corte certo e escolher produtos adequados aos cuidados que esse tipo de fio demanda.

Inspire-se na seleção de estilos de cabelos cacheados das celebridades e confira as dicas essenciais para usar e abusar da beleza natural dos seus fios:

Buscar ...

BUSCAR

NOVO TRESemmé
SELAGEM CAPILAR



EXPERIMENTE

RECENTES

POPULARES



Babylights: para clarear e trazer..

24 de julho de 2015 0

Cuidados com os cabelos durante a gravidez

Cuidados com os cabelos durante ..

22 de julho de 2015 0

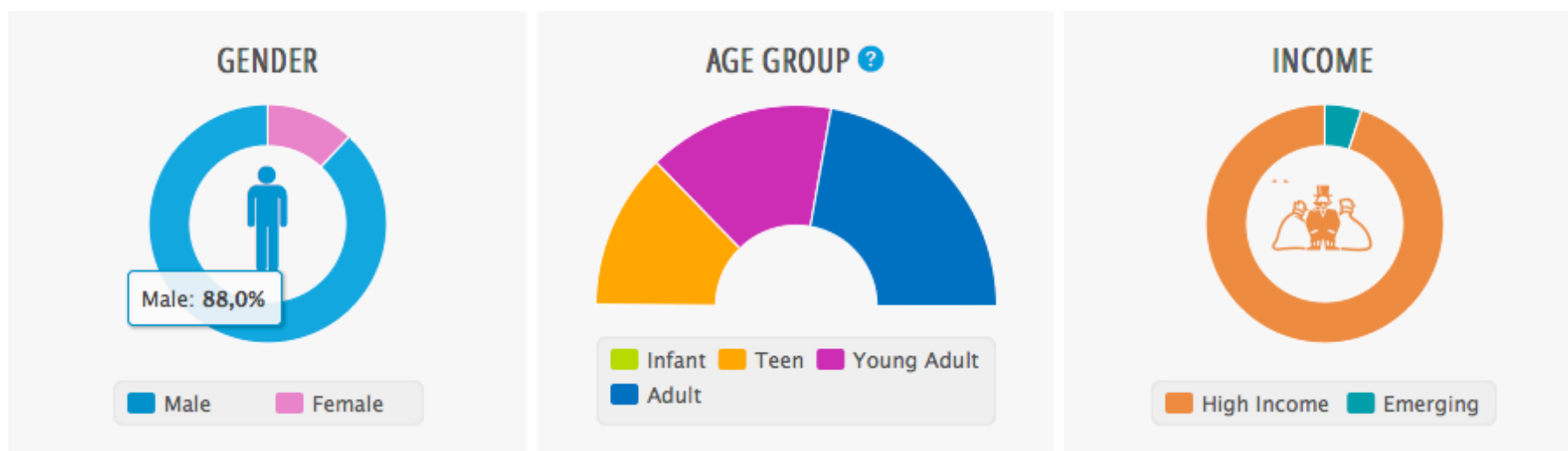


Cinco dicas essenciais do expert..

20 de julho de 2015 0



Predicting User Behavior

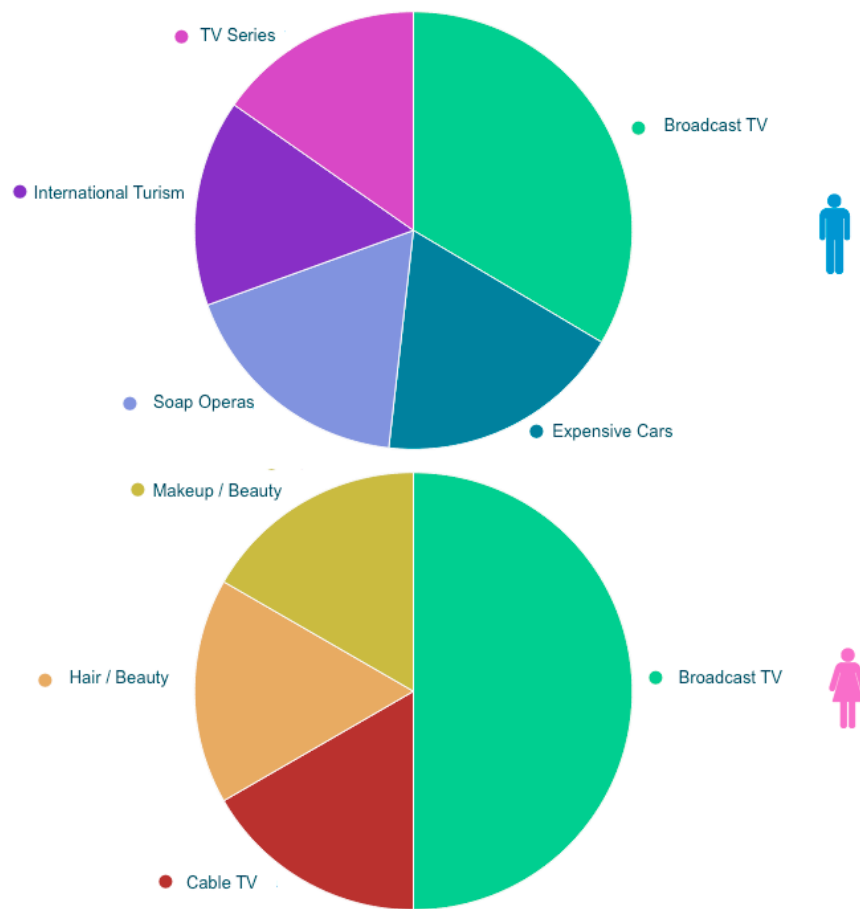


Brazilian Java Developers

<http://www.tailtarget.com/blog/getting-to-know-the-java-developer-using-data-science/>

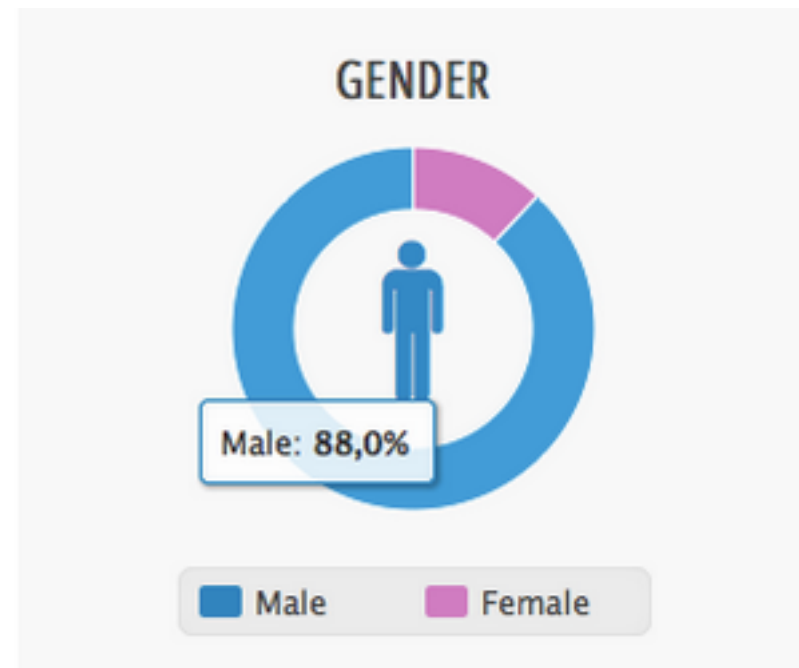
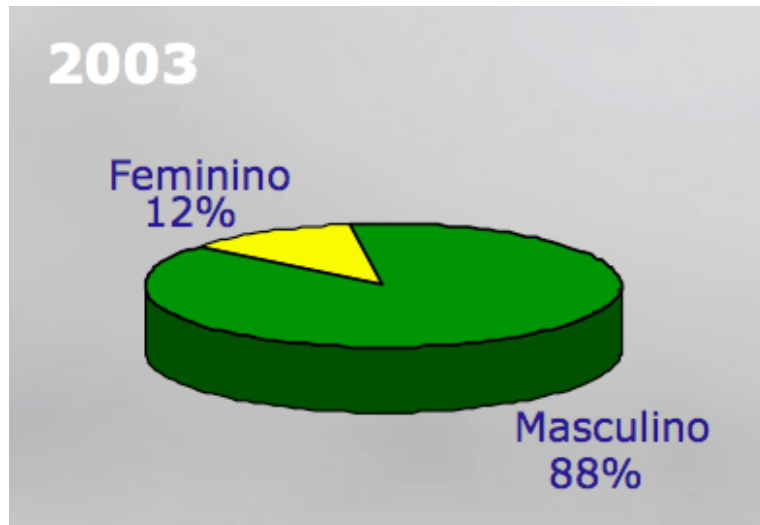


Predicting User Behavior



Brazilian Java Developers

Does Behavior Data Work?





Predicting User Behavior

Classifiers Clustering

Recommendation Systems

Deep Learning

Detecting User Behavior



SOCCER

User Profile

uid=123 : [SOCCER: 1 view, MALE: 1 view]



Detecting User Behavior



uid=123 : [SOCCER,
MALE]



SOCCER

User Profile

uid=123 : [SOCCER: 10 views, MALE: 20
views]

Implementing a Classifier

SUPERVISED LEARNING

1. DATA PREPARATION

Segment	Country	Seed
SOCCER	BRAZIL	http://esportes.terra.com.br/futebol/ http://placar.abril.com.br/materia/ http://esporte.uol.com.br/futebol/
FASHION	BRAZIL	http://www.fhits.com.br http://www.dolcemoda.com.br
POLITICS	BRAZIL	http://eleicoes.uol.com.br http://noticias.terra.com.br/brasil/politica/ http://ultimosegundo.ig.com.br/politica/
FAMILY	BRAZIL	http://guiadobebe.uol.com.br/ http://mdemulher.abril.com.br/familia/

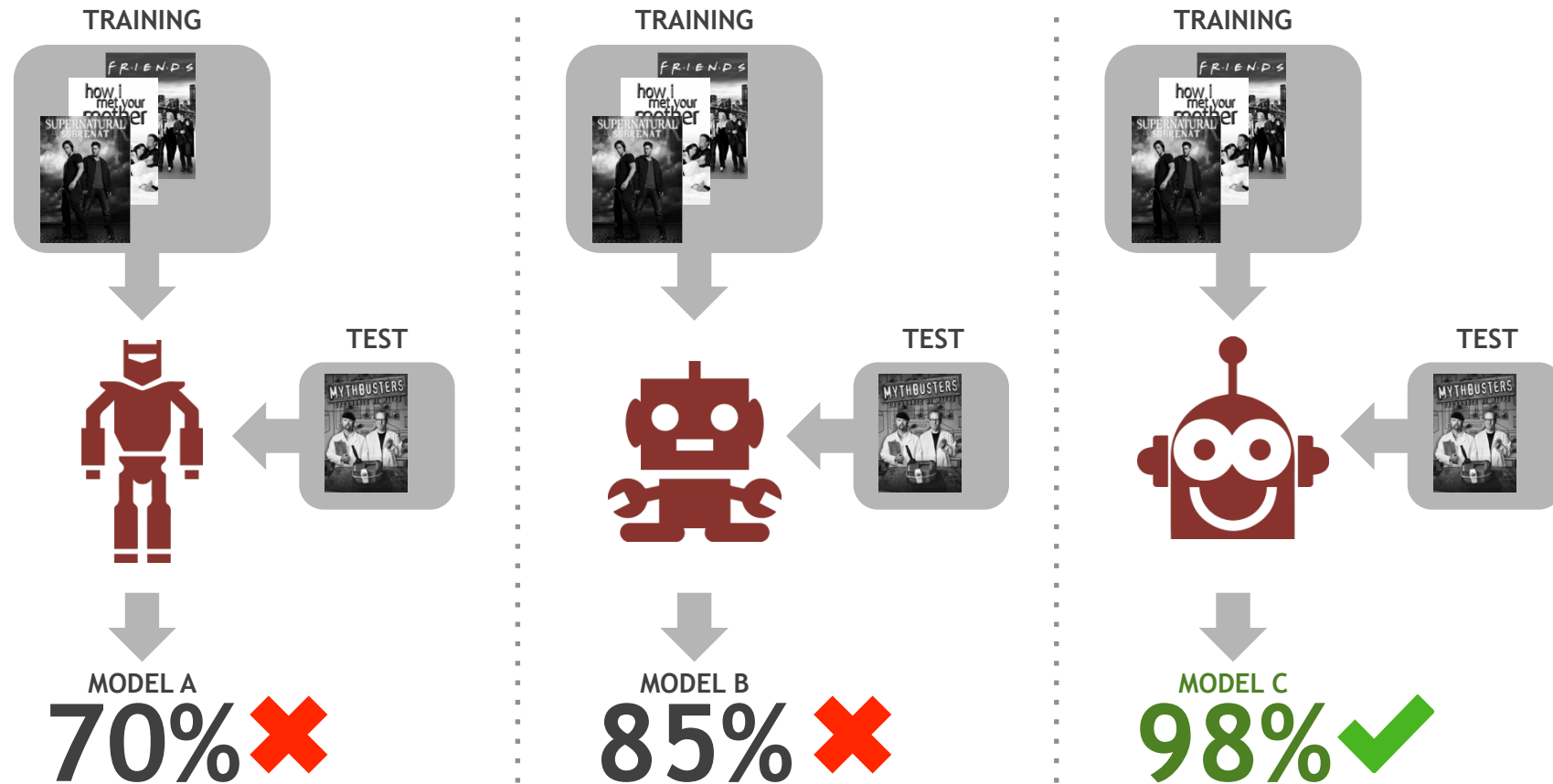
1. DATA PREPARATION (CRAWLER)

Segment	Country	Pages (3,000)
SOCCER	BRAZIL	1. http://esportes.terra.com.br/futebol/internacional/mourinho-diz-que-rafa-benitez-destruiu-a-inter-de-milao-em-2010,3652d311ab33c95fb07b01f6ee27168ew7uvRCRD.html 2. http://esportes.terra.com.br/futebol/brasileiro-serie-a/pato-que-voltar-a-selecao-e-lamenta-nao-enfrentar-corinthians,ac2efed99a4ed4ed83198389e01961f7sih9RCRD.html
FASHION	BRAZIL	(...)
POLITICS	BRAZIL	(...)
FAMILY	BRAZIL	(...)

2. SUPERVISED LEARNING

	Segment	Country	Pages (3,000)
80% TRAINING	SOCCER	BRAZIL	1. http://esportes.terra.com.br/futebol/internacional/mourinho-diz-que-rafa-benitez-destruiu-a-inter-de-milao-em-2010,3652d311ab33c95fb07b01f6ee27168ew7uvRCRD.html
			2. http://esportes.terra.com.br/futebol/brasileiro-serie-a/pato-que-voltar-a-selecao-e-lamenta-nao-enfrentar-corinthians,ac2efed99a4ed4ed83198389e01961f7sih9RCRD.html
	FASHION	BRAZIL	(...)
20% TEST	POLITICS	BRAZIL	(...)
	FAMILY	BRAZIL	(...)

3. FINDING THE MODEL



4. TESTING THE MODEL

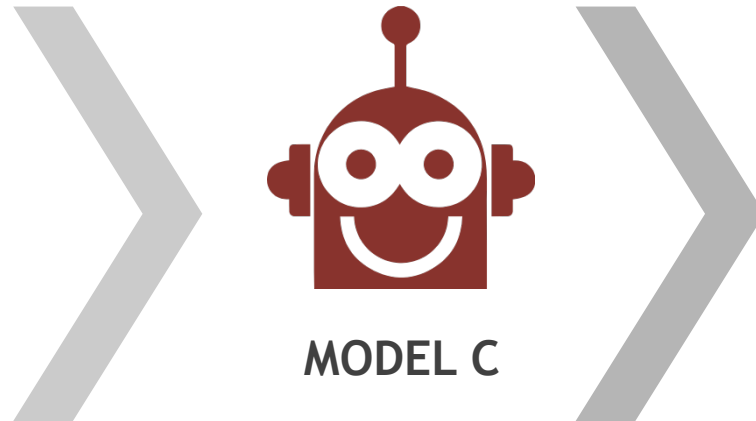
CONFUSION MATRIX

	SOCCER	FASHION	POLITICS	FAMILY
SOCCER	997	0	1	2
FASHION	1	999	0	0
POLITICS	3	1	995	1
FAMILY	2	1	0	997

4. MODEL IN ACTION



Page



MODEL C

TV

SUBJECT

RETRAIN 

APPLYING THE MODEL

Colorado ZH

oBoticário Emocione no Dia das Mães com presentes O Boticário.

Publicidade

Gre-Nal decisivo
Comentaristas opinam: o que o Inter precisa para não sofrer gols no Beira-Rio

Se Alisson não for vazado, o máximo que pode acontecer é disputa por pênaltis

27/04/2015 | 17h09



Foto: Carlos Macedo / Agência RBS

Após **empatar em 0 a 0** com o Grêmio no primeiro jogo da final do Gauchão no último domingo, na Arena, o Inter tem a vantagem de decidir em casa. Contudo, é imprescindível ao time de Diego Aguirre que não sofra gols no Beira-Rio, pois pode correr o risco de perder o título mesmo com um empate — o regulamento do campeonato prevê gol qualificado fora de casa.

Opinião: O Gre-Nal 405 pelos cascudos do jornalismo gaúcho

Se o placar da Arena for repetido, o troféu será erguido após uma disputa por pênaltis. Aguirre deve sair no atacante para garantir a vitória rapidamente? Ou segurar atrás e tentar pegar o Grêmio no contra-ataque? Foi o que *Zero Hora* perguntou aos colunistas Diogo Olivier e Luiz Zini Pires e ao comentarista da **Rádio Gaúcha**, Cléber Grabauska. Confira as opiniões:

Gre-Nal da zoeira: o clássico nas redes sociais

Diogo Olivier, colunista de Zero Hora

"O Inter tem se defendido bem no 4-2-3-1 inclusive quando usa atacantes de velocidade pelos lados, como Sasha e Valdivia, mais Nilmar lá na frente. Todos são disciplinados e ajudam a marcar sem a bola. No Chile, contra La U, foi assim tanto com Jorge Henrique (até este se lesionar) quanto depois, a partir da entrada de Valdivia. Alisson foi mero espectador.

Publicidade

Compartilhar: TWITTER FACEBOOK EMAIL CORRIGIR

Publicidade

Publicidade

Publicidade

O MELHOR DA ZH

De Maio

Dilma não fará pronunciamento na TV em 1º de Maio

Guarda Nacional acionada

Prefeito de Baltimore declara toque de recolher e chama a Guarda Nacional

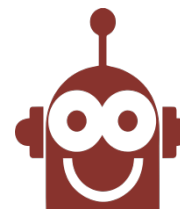
Escassez de recursos

Caixa reduz teto de financiamento para imóveis usados

Segurança

com.cybozu.labs
langdetect

Portuguese



MODEL



SOCCKER

COUNTRY DETECTION



Reinvente-se com a

DietaDukan

MINUTO 30.COM

Aquí tienes toda la información LIGA AGUILA

Home Medellín Bogotá Antioquia Nacional Política Internacional Entretenimiento Deportes

Medellín Innovadora Ciudad Crónica Denuncia Ciudadana Ciencia y Tecnología Sexualidad Horóscopo Solo Sexy's Liga Aguila I - 2015

Home » Política » Archive by Category "Medellín"

Concejales de Medellín rechazaron el cruel asesinato de una edil en San Antonio de Prado
By Sandra Hoyos
Los concejales de Medellín rechazaron en bloque durante la sesión plenaria de este viernes su enérgico rechazo al asesinato de la vicepresidente de la Junta Administradora Local de San Antonio de Prado...

EN VIDEO: Con algunos inconvenientes para los militantes del Centro Democrático avanzan las consultas internas en Antioquia
By Minuto30.com
Minuto30.com.- La Registradora Delegada para Antioquia, Mónica Larduy, aseguró que con algunos inconvenientes para los militantes del Centro Democrático, partido que realiza una consulta cerrada para...

Si votas en los puestos censo de Medellín, ten en cuenta que éstos fueron trasladados
By SoloDuque
La Registraduría Nacional anunció la reubicación de los dos puestos censo de Medellín para las consultas de este domingo 19 de abril. El puesto censo que originalmente estaba ubicado en el Estado...

Centro Democrático elegirá listas al Concejo de Medellín, Jal y delegados para la convención en consulta interna
By Cristian Cartagena
Minuto30.com.- El Centro Democrático definió participar en la jornada electoral de consultas internas del próximo 19 de abril que habilitó la Registraduría Nacional del Estado Civil para que los partidos...

EN VIDEO: Ana Cristina Moreno, la apuesta de Álvaro Uribe para el Concejo de Medellín
By Cristian Cartagena
Minuto30.com.- Después de integrar la lista cerrada a la Cámara de Representantes del Centro Democrático por Antioquia, en el renglón 10, en el uribismo quedó algo claro: Ana Cristina Moreno debía...

EN VIDEO: De la mano del ramismo, Nicolás López busca llegar al Concejo de Medellín
By Cristian Cartagena
Minuto30.com.- Con el objetivo de trabajar en cuatro propuestas en torno a las problemáticas de seguridad, educación, movilidad y desarrollo regional, el exsecretario de Minas de Antioquia, Nicolás...

EN VIDEO: Concejales se disculpan entre ellos y ofrecen excusas a la ciudadanía
By Cristian Cartagena
Minuto30.com.- Tras el bochornoso episodio de este jueves, protagonizado por concejales Rober Bohórquez, Carlos Mario Mejía, Carlos Bayer, Miguel Quintero y Juan Felipe Campuzano disculparon, en el...

LO MEJOR DE LA MODA HASTA 60% OFF

COMPRAR dafiti cyberlunes

¿Quiénes Somos? ¿Cómo anunciar? Envíenos Noticias Contáctenos Spot Publicitario

CIUDAD S

Liga Aguila

Ahora los sábados son el mejor día para jugar y ganar.

Ver más

decolar.com

OFERTAS MALUCAS!

Passagem para

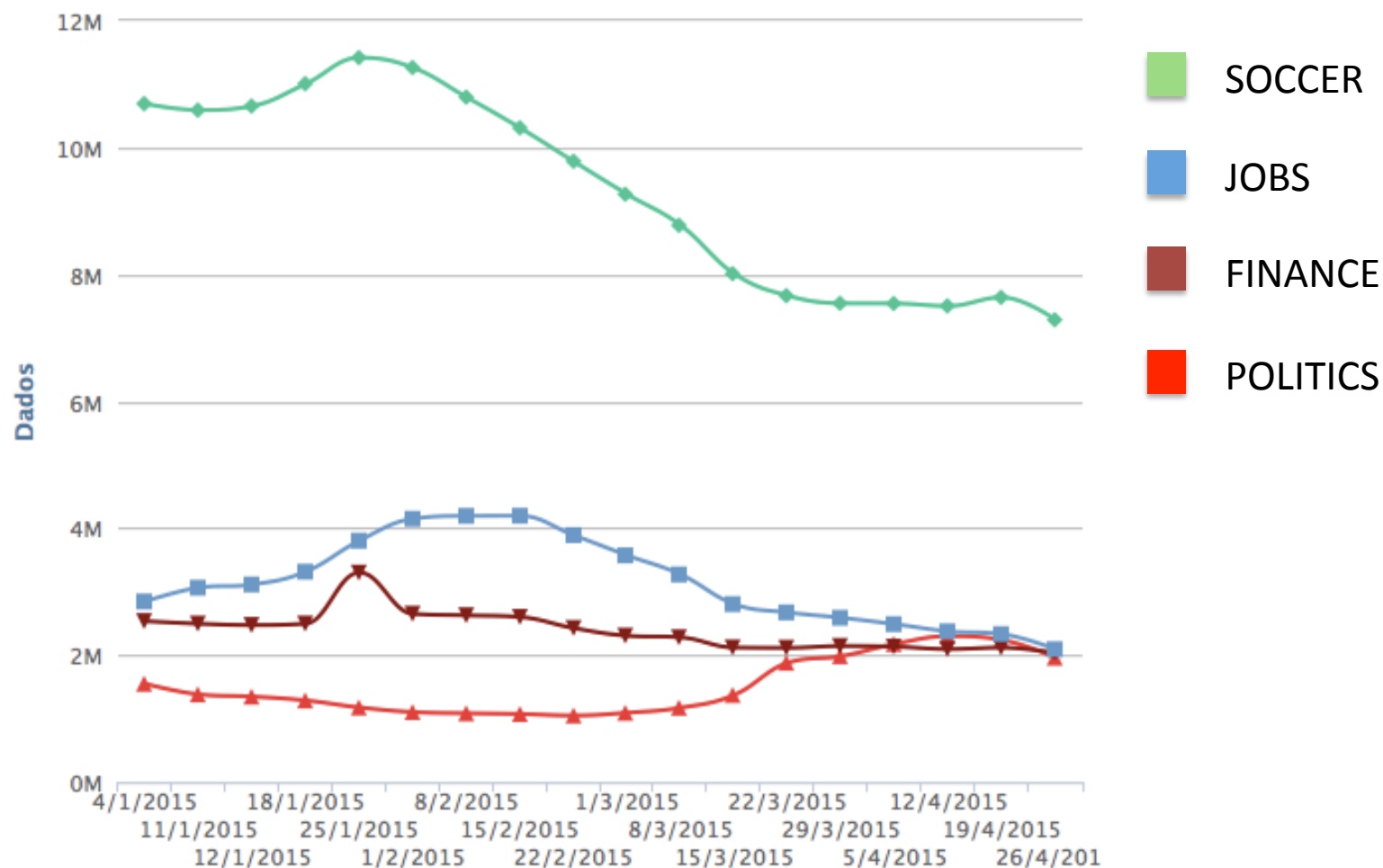
Colombia: 8.999
Argentina: 100
Mexico: 87
Venezuela: 50
USA: 45



**Many times, you don't need to
process all your data**

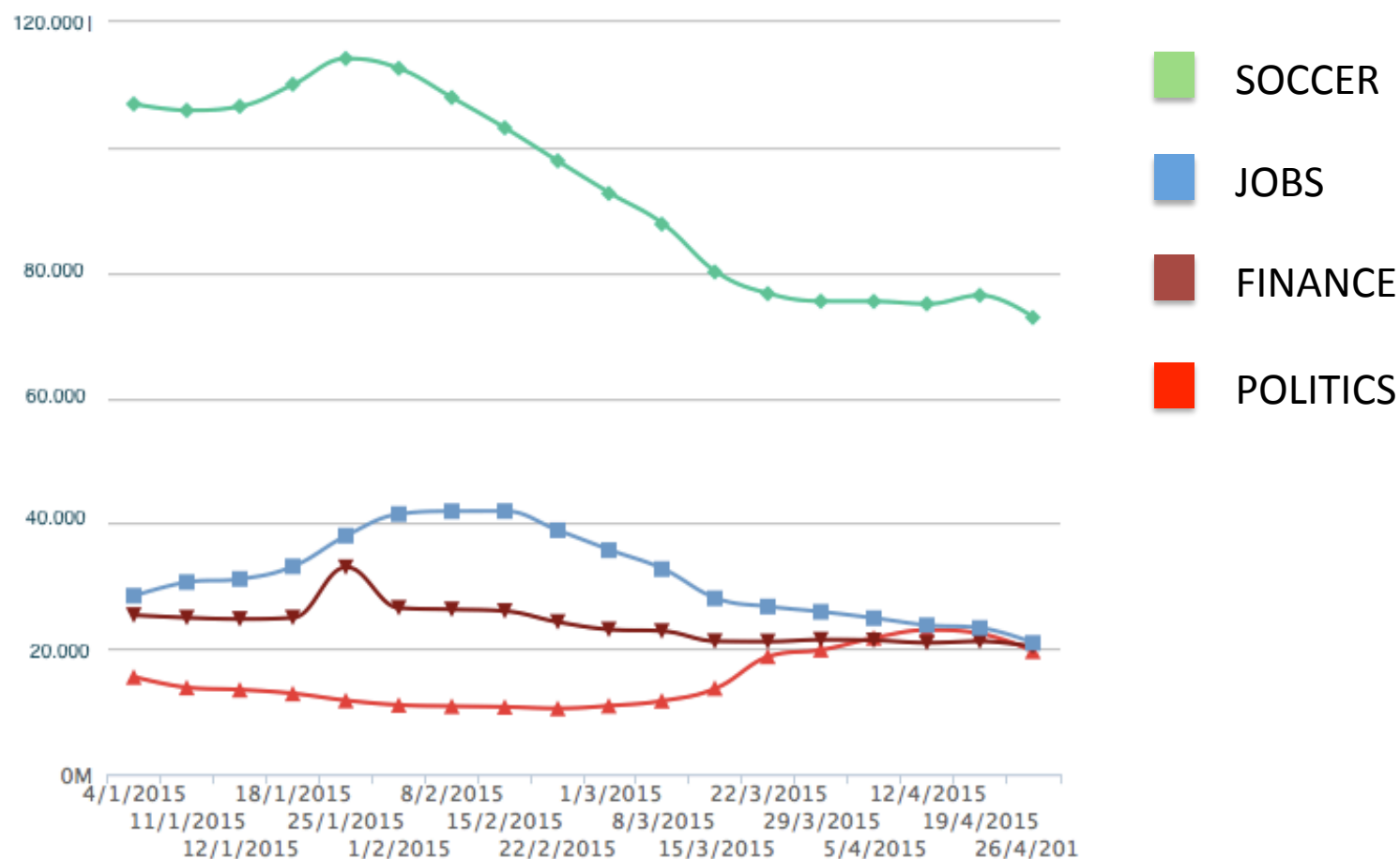


Users by Interest (ALL)





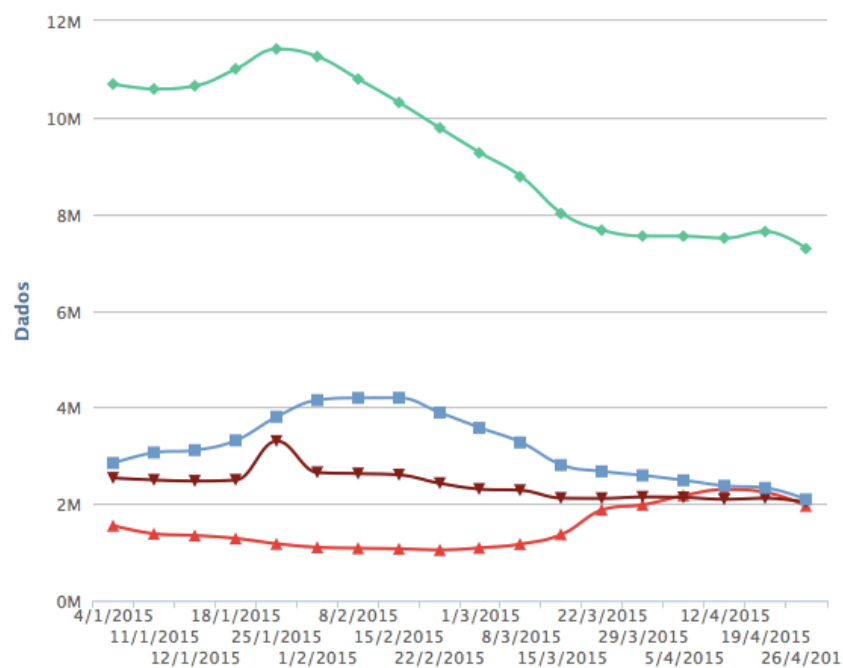
Users by Interest (1%)



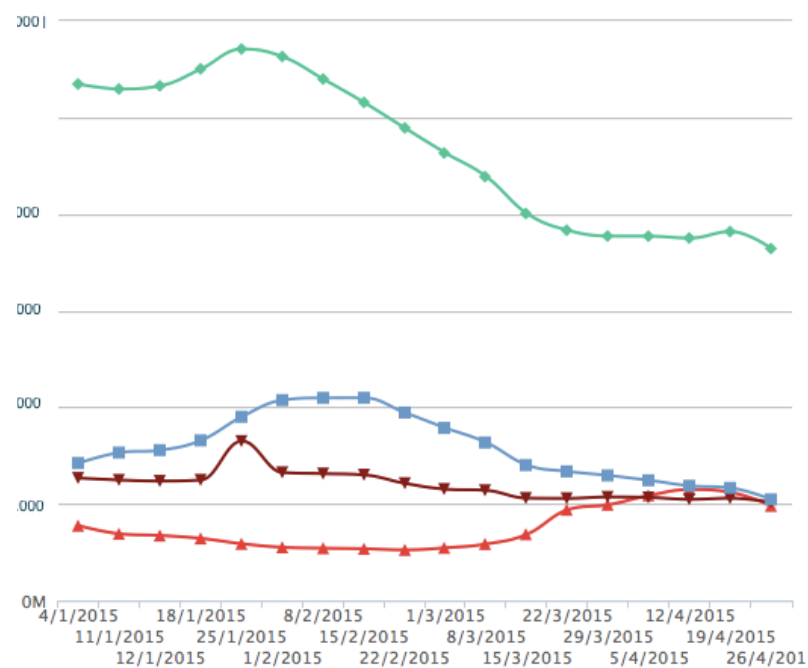


Users by Interest

ALL



1%



How to select a good sampling

Size has to be representative

Homogeneous distribution

Sampling Size

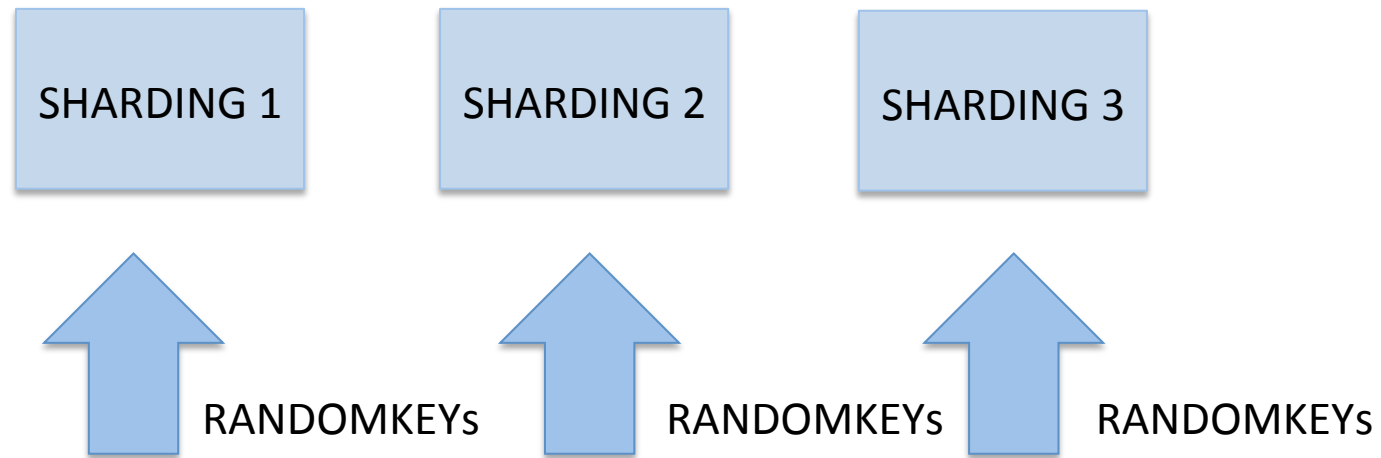
BIGGER SAMPLING = MORE ACCURACY

Sample Size

Find the minimum sample size for your problem.

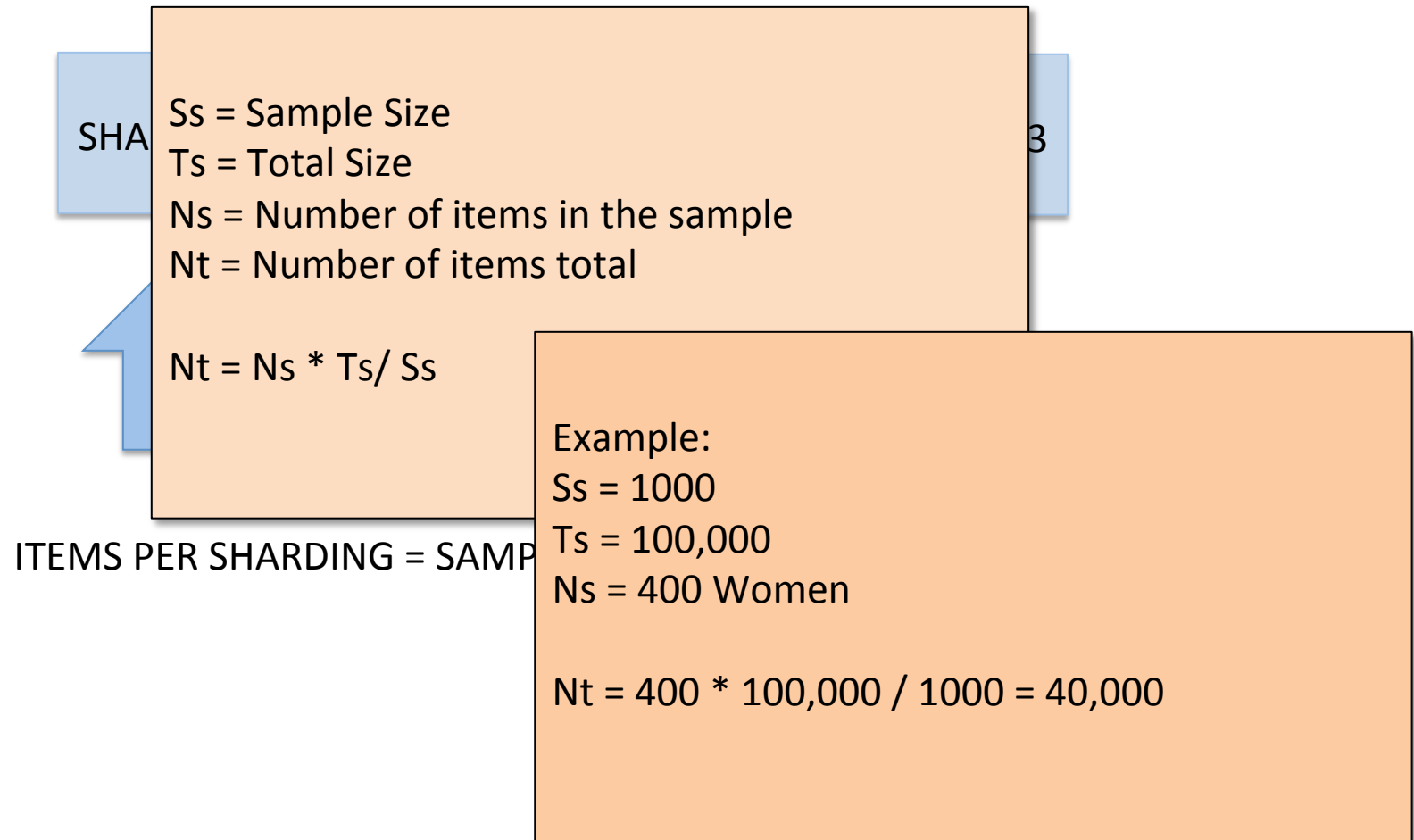
If the database has less than the minimum,
use all your data

Sample Distribution (E.g.: Redis)

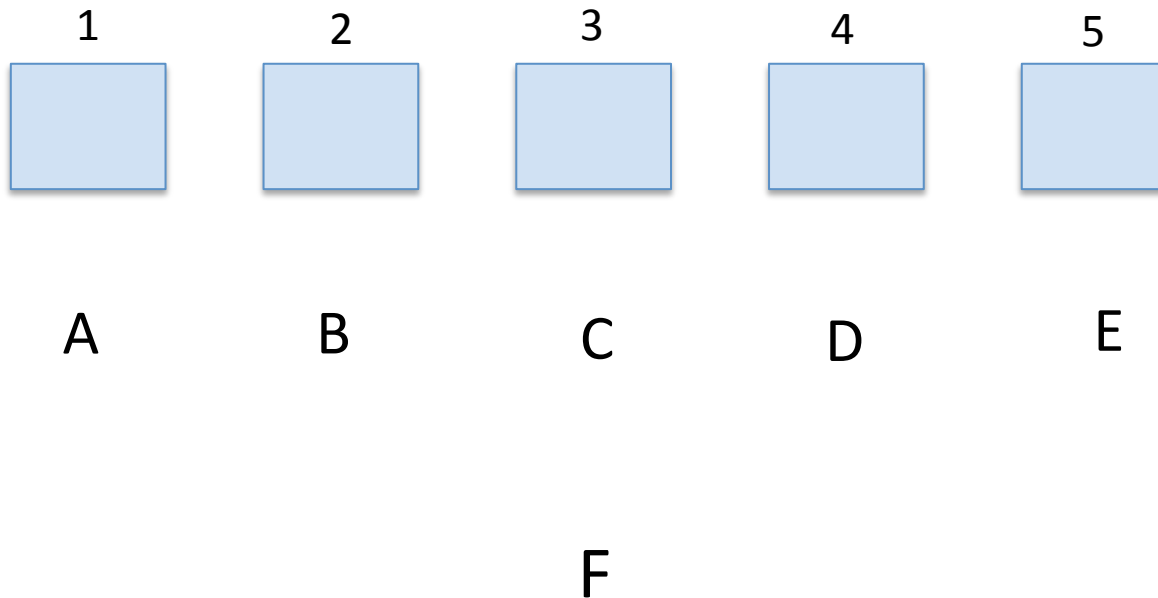


ITEMS PER SHARDING = SAMPLE SIZE / NUMBER OF SHARDINGS

Sample Distribution (E.g.: Redis)



And when you don't know how many items you have? Reservoir Sampling

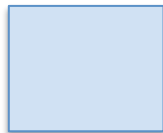


Random (0..1): 0.7
 $K = S_s / i$
 $K = 5 / 6 = 0.83$
Se $K > \text{Random} \Rightarrow \text{CHANGE!}$

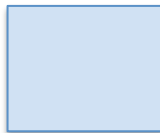
Distributed Reservoir Sampling

A B C D E F G H I J K L M N O P Q R S T U V X Y W Z

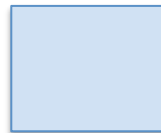
1



2



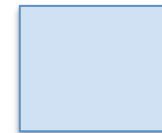
3



4



5



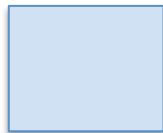
Distributed Reservoir Sampling

A B C D E F G H I

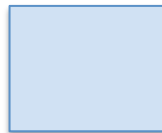
J K L M N O P Q R

S T U V X Y W Z

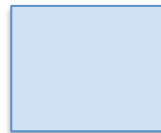
1



2



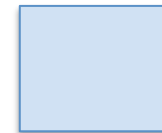
3



4



5



Distributed Reservoir Sampling

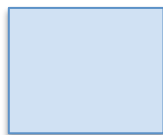
A:0.1 B:0.3 C:0.2 D:0.7 E:0.9 F:0.11
G:0.4 H:0.6 I:0.76

J:0.8 K:0.2 L:0.54 M:0.4 N:0.21 O:0.33
P:0.56 Q:0.32 R:0.23

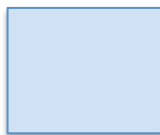
S:0.21 T:0.32 U:0.22 V:0.7 X:0.12 Y:
0.23 W:0.3 Z:0.76

```
private SortedMap<Double, MyObject> reservoir;  
  
...  
if (reservoir.size() < SAMPLE_SIZE) {  
    reservoir.put(score, myObject);  
} else if (score > reservoir.firstKey()) {  
    reservoir.remove(reservoir.firstKey());  
    reservoir.put(score, myObject);  
}
```

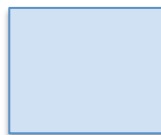
1



2



3



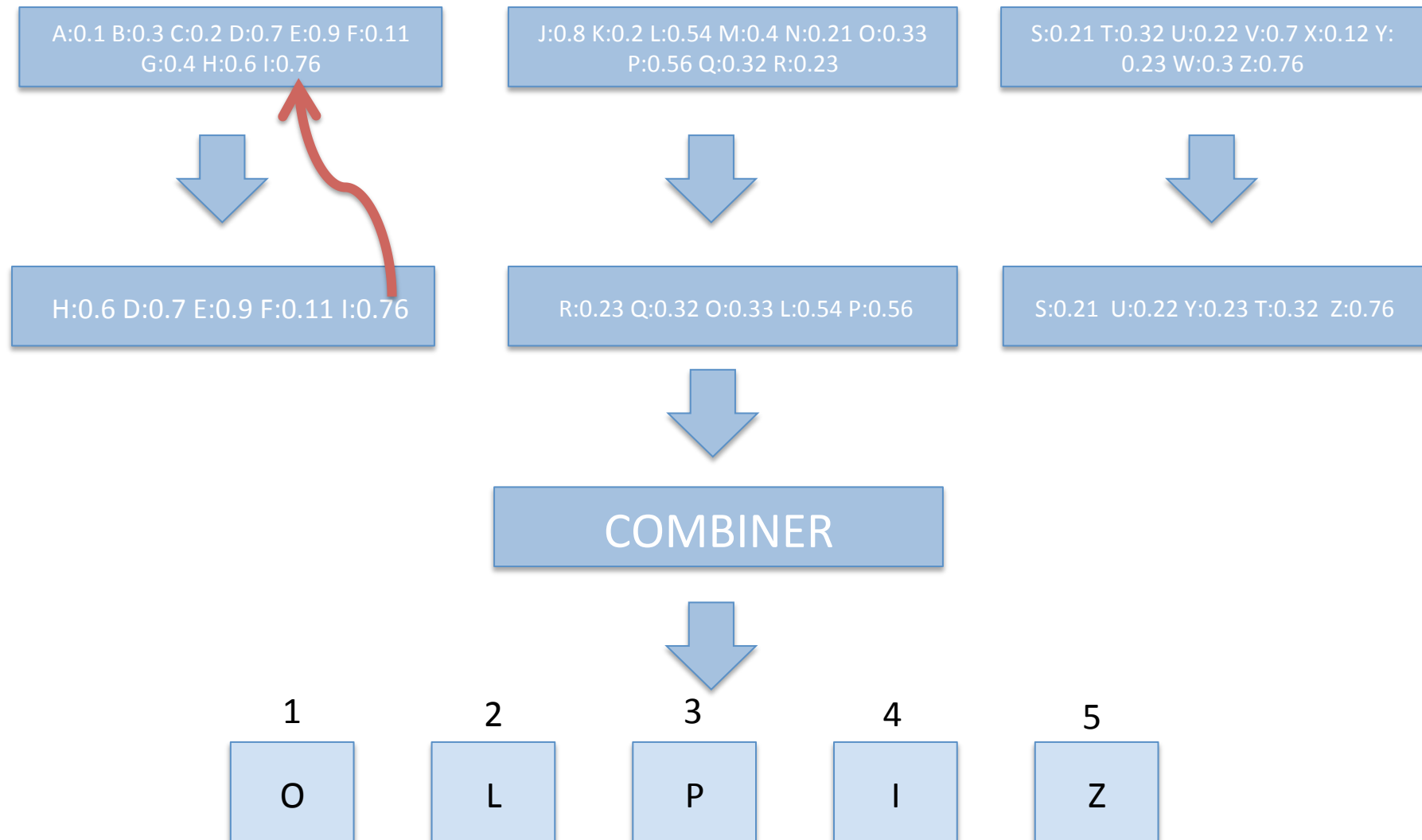
4



5



Distributed Reservoir Sampling



Distributed Reservoir Sampling

Apache Crunch:

```
import org.apache.crunch.lib.Sample;  
  
Sample.reservoirSample(PCollection<T> input, int sampleSize)
```

Why Java?

Machine learning libraries

Drivers NoSQL

Performance

Connectors to Distributed Frameworks



Java and Data Science in Digital Advertising

Fabiane Bizinella Nardon

@fabianenardon

Chief Data Scientist at TailTarget

