

Challenges and Big Opportunities for Data Scientist at VCCORP

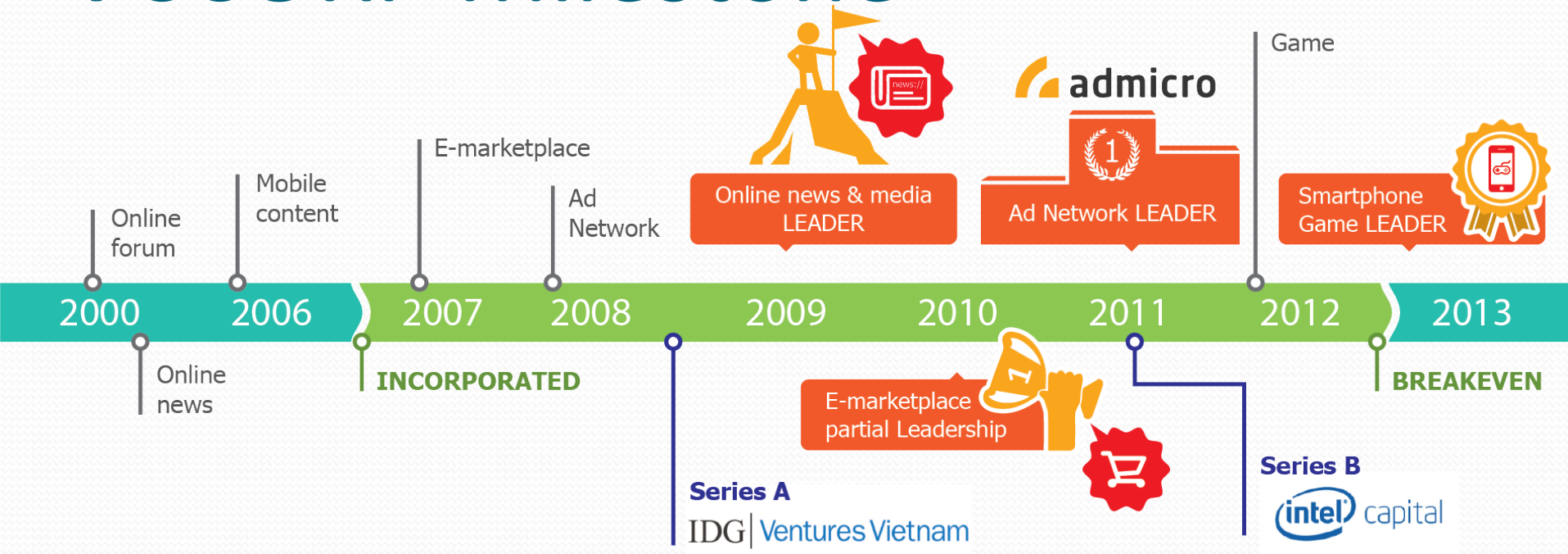
Hoang Anh Tuan
CTO Admicro - VCCORP
tuanhoanganh@vccorp.vn

Content

- Company Overview
- Big Data at VCCORP
- Our main challenges

COMPANY OVERVIEW

VCCORP Milestone



FOUNDERS

Mr. Vuong Vu Thang is the Founder and Chairman of VCCorp

His success in developing VCC to be the leading new-media company in Vietnam began when he founded his first online community, TTVNOL, from a garage start-up in 2000. After founding this network, he founded the first online media & news portal Tintucvietnam in 2002. In 2005, he founded the first private company in mobile value added services and internet content, which formed the initial foundation of the current-day VCC.

As a natural entrepreneur, Thang has been the first mover in all the emerging internet sectors in Vietnam including media content, social network, mobile content & services, and ecommerce. Naturally, VCC became an innovative leader of Vietnam's internet & media industry. As the technology strategist, he has been the chief architect of disruptive technologies in Vietnam in the last 10 years including CMS, key portal technologies, and cloud computing to name a few.



Mr. Vuong Vu Thang
Founder, Chairman

Mr. Nguyen The Tan is the Co-Founder and CEO of VCCorp

With unparalleled know-how in internet monetization, Nguyen The Tan's leadership is driving the growth of VCC. His results are apparent as he oversaw 100% annual growth in advertising, mobile services and ecommerce revenue for each of his last 6 years at VCC. As a strategic visionary within the industry, Nguyen The Tan's work has made a tremendous impact within the Vietnamese marketplace.

Before joining VCC, Tan was Vice-Director of one of the largest telco companies in Vietnam, Viettel Fixed Telecom. Before Viettel Telecom, he was the division director at a leading software and system integration company, CMC. There he built a e-library solutions platform and expanded the distribution of CMC's software and solutions towards broader segments within the market.



Mr. Nguyen The Tan
Co-Founder, CEO

VCCORP Overview



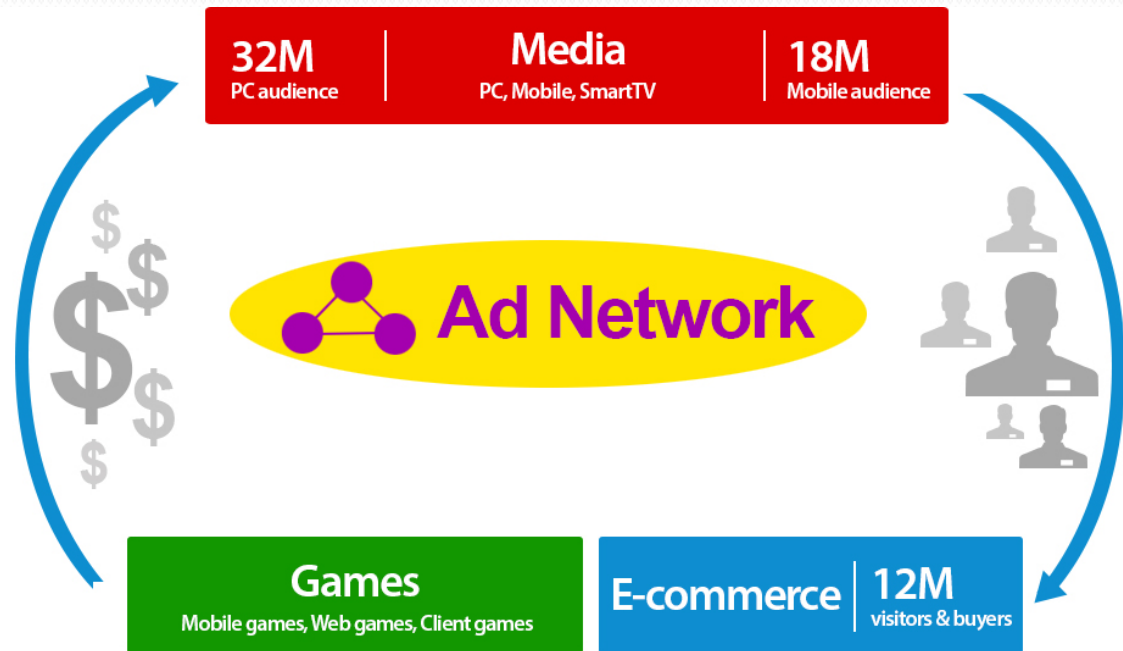
Leading Internet Company

Overview

- ✓ **First mover** DNA
- ✓ **50%** YoY Growth
- ✓ **43M** web audience
- ✓ **38M** mobile audience
- ✓ **1,700** employees

Investors

IDG | Ventures Vietnam



Payment Platform | **Cloud Computing & BigData Analysis**
VCCorp - Innovation. Non-stop!

VCCORP MARKET COVERAGE

- 43M internet user reach (97.6% of VN internet population)
- 38M mobile user reach (95% of VN Smartphone population)
- 10,000+ online & mobile advertisers
- 100,000+ small business merchants
- 12M e-marketplace visitors & buyers
- Largest Ad network in Vietnam with 1000+ publishers, including 200+ top-publishers, 30 of them are exclusive
- 22 leading products with presence in 20+ verticals; 14 sites are in top 100 websites in Vietnam (*news, finance, family, teenage, auto, high-tech, online advertising, B2C and C2C, content consumption mobile*)

BIG DATA AT VCCORP

Big Data in VCCORP

- In the 2007, Big Data was applied early in Baamboo Search Engine
- Since 2009, Big Data platform have been installed for serving ad system in VCCORP
- Currently, Big Data platform is being developed and improved in major areas.
 - Advertisement
 - Digital Content
 - Ecommerce
 - Game
- Current staffs: 100 Data Engineers

The challenges in VCCORP

- Big Data skillsets in-house
- The large-scale data
- The huge amount of specific problems, spreading over many areas which is required creative problem-solving, self-motivated person
- Human resource is not enough

System Info

1,5 Billion

records per day

300 Gb

data per day

20 Tb

processing data per day



30 TB



20.000



4 PB

OUR MAIN CHALLENGES

- User behaviors
- Ad Optimization
- Core NLP and its application
- News Distribution
- Recommendation Engine
- Vccorp Analytic

USER BEHAVIOR

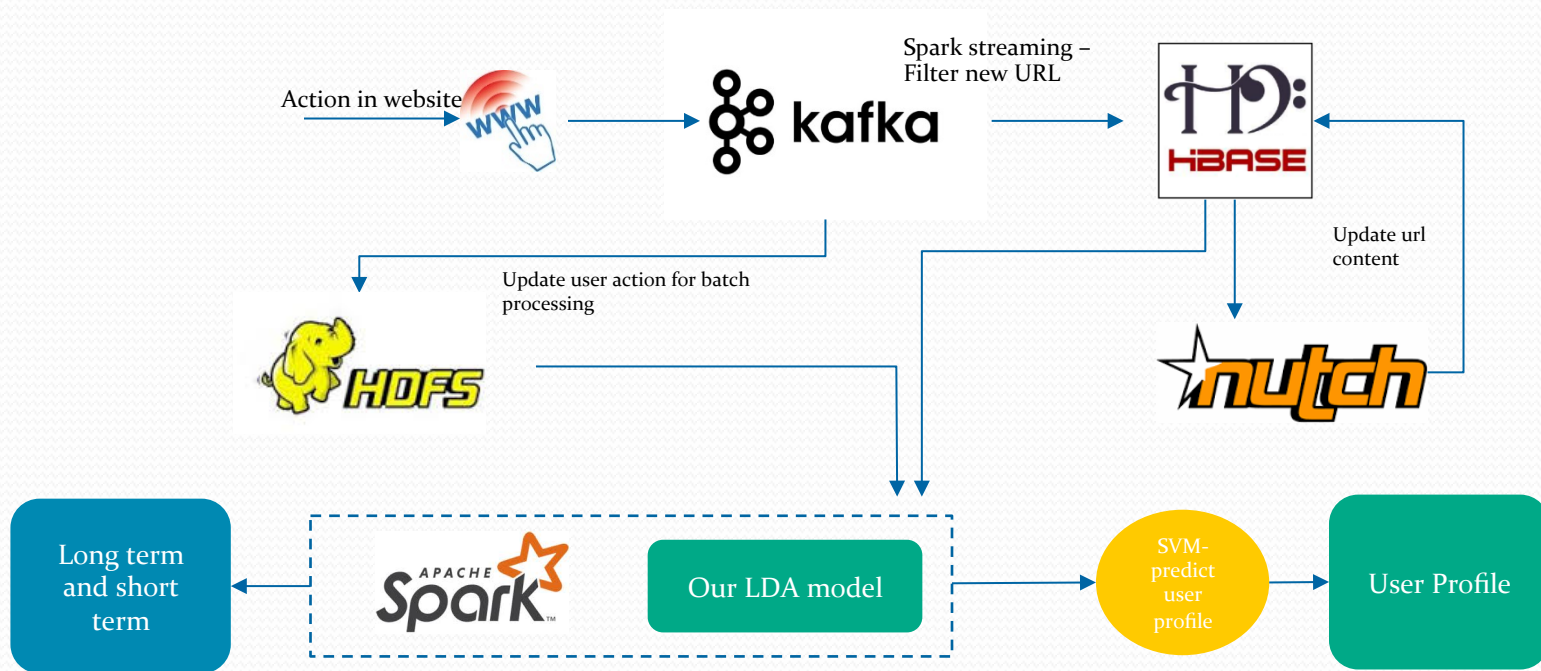
User behavior analytics

- Three main projects:
 - **Demographic:** gender, age
 - **User profile:** behavior, interest
 - **Cross devices:** tracking user on multiple devices

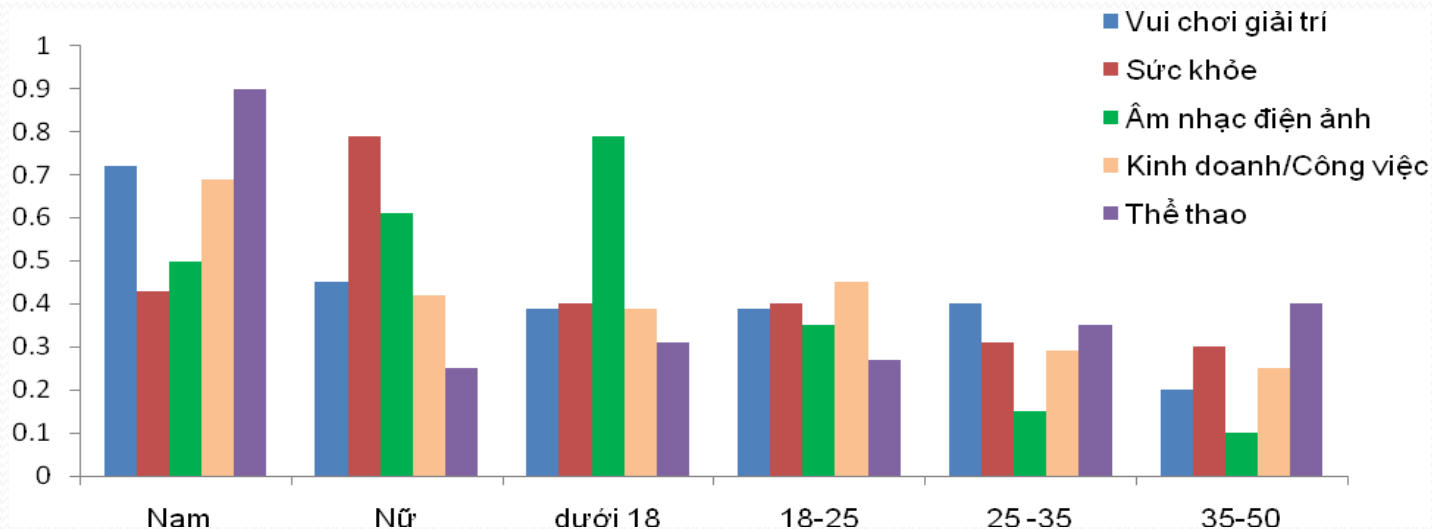
Demographic - User profiling

- Detect user profile including gender, age, user interesting (12-based interests), long term, short term. Based on data
 - Browsing history
 - Keyword search history
 - Time usage, time onsite
- Data:
 - 43 M users in pc
 - 38 M users in mobile
 - 1 Terabyte logging data
- Result – accuracy :
 - Gender : 82.5%
 - Age : 67.5%

System overview



Demographic - Behavior



Benchmark

Benchmarking data: 43M users, 200M documents, 30000 * 10⁶ actions,
230 topics

VCCORP cluster : 20 nodes, 640 cores, 640 GB ram

Our Model

Time : 18h,

Accuracy:

- Gender : 82.5%
- Age : 67.5%

Recall : 92 %

Old classification model

Time : 16h,

Accuracy:

- Gender : 79.5%
- Age : 63.4%

Recall : 92 %

LDA with Spark MLLIB

Time : 36h,

Accuracy:

- Gender : 75.1%
- Age : 60.1%

Recall : 91 %

Cross Device



Cross device

- We used information :
 - Both User-IP and timestamp in their devices
 - Website and categories history
 - User demographic and user interest
- Result:
 - Accuracy : 60%
 - Number detected users : 11M

AD OPTIMIZATION

Admicro Overview

- #1 ad network in Vietnam: cover 38% market share
- 200+ top publishers in Vietnam
- 10,000+ advertisers
- 4B page views per month
- 1,5B impressions per day
- 22 leading ad products
- 43M internet user reach (97.6% of VN internet population)
- 38M mobile user reach (95% of VN Smartphone population)

Ad Optimization

- The advanced techniques were implemented:
 - Personalization
 - Audience Targeting Platform
 - Real Time Bidding
 - Retargeting
 - Contextual Targeting
 - SSP/DSP/DMP

Personalization

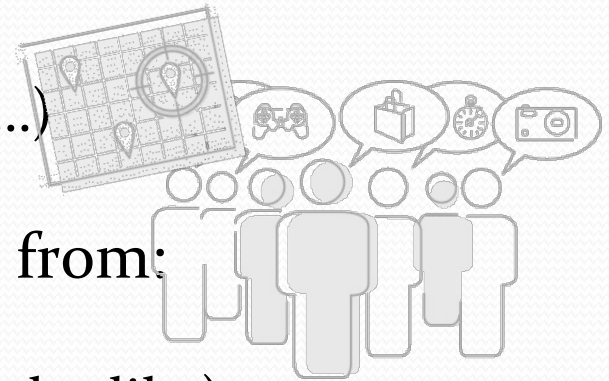
- In traditional advertising, ads are displayed to everyone in the fixed location
- By contrast, personalization technique will choose the best fit ads for each user:
 - 43M internet user
 - 10.000 ads
- Using multiple technologies:
 - High load capacity web server
 - Optimization algorithms
 - Estimate and prediction algorithms

} 430B estimated operations for each time



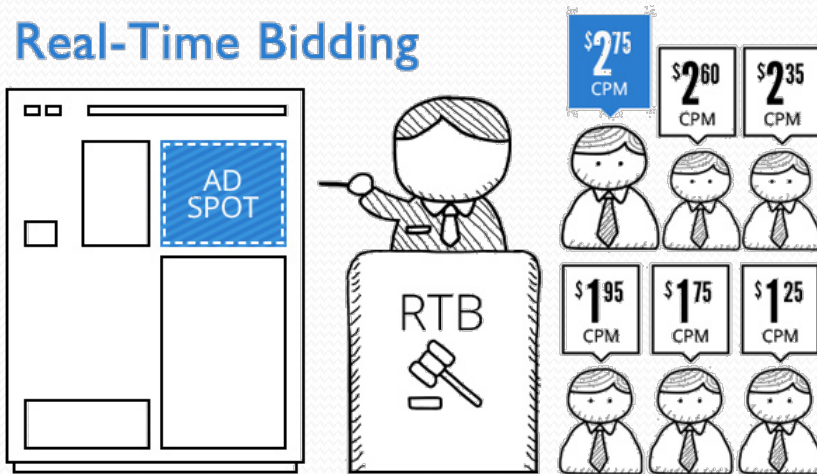
Audience Targeting Platform

- Advertisers can target their particular audience
- Subset of the audience can be prebuilt by “set operators” of user properties
 - Location
 - Demographic (gender, age, relationship...)
 - Interest/Behavior
- Especially, an audience can be made up from:
 - List of email / phone number
 - Automatically find similar audiences (look-alike)



Real Time Bidding

- A transaction (sell/ purchase) of ad impressions is immediately proceeded when an audience trigger the ad zones



Real Time Bidding

- A transaction (sell/ purchase) of ad impressions is immediately proceeded when an audience trigger the ad zones
- Challenges:
 - 80 ms is the maximum time of a transaction
 - 1000 sites in VN
 - 4.5 billion request / day
 - Number of Transaction: \$ 200,000 / day

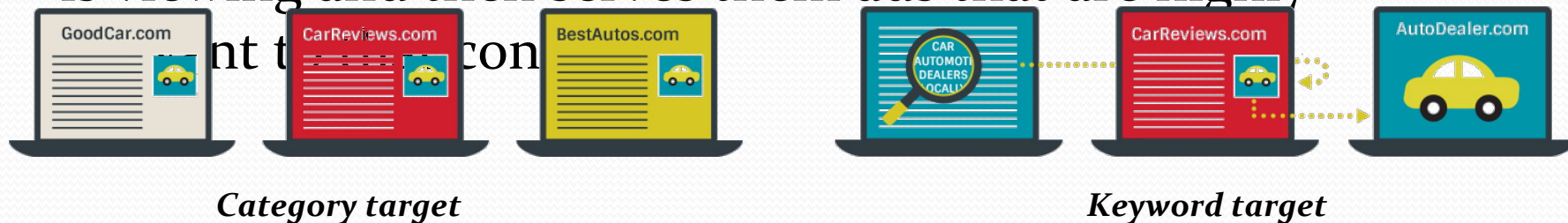
Retargeting

- Retargeting is a powerful branding and conversion optimization tool
- Ads will follow customers after they do the shopping
- Ads will be displayed in
 - Any Web pages
 - Multiple devices



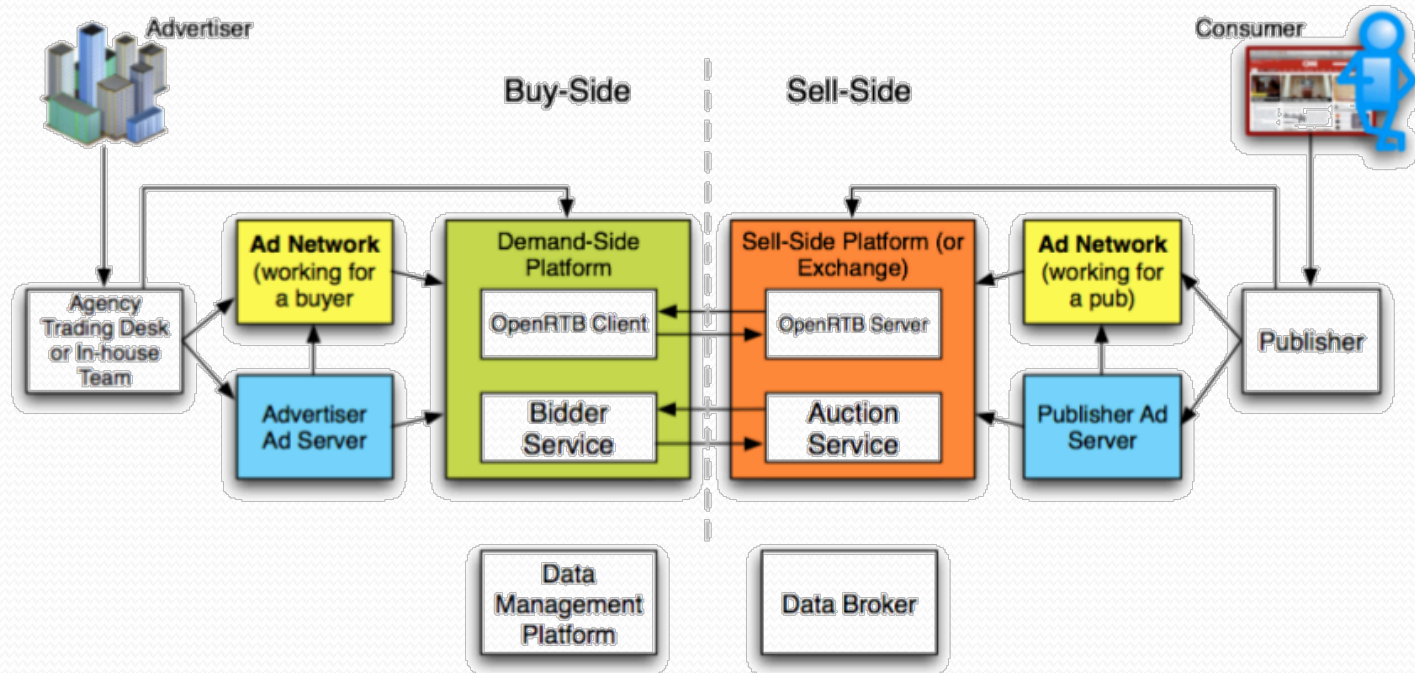
Contextual Targeting

- Contextual targeting looks at the **category** or **keywords** of the current page a consumer is viewing and then serves them ads that are highly



- We implemented:
 - Content classification systems (LDAP)
 - Keyword index and search engine system

SSP/DSP/DMP



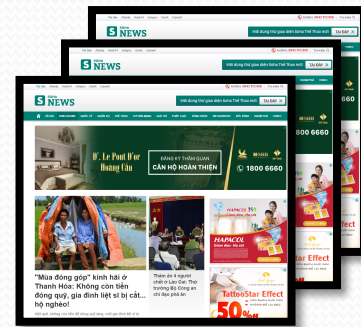
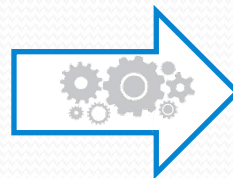
NEWS DISTRIBUTION

News distribution

- VCCORP possesses many large online newspapers in VN
- We are proud of becoming the first company in Vietnam able to implement an automatic method for publishing news



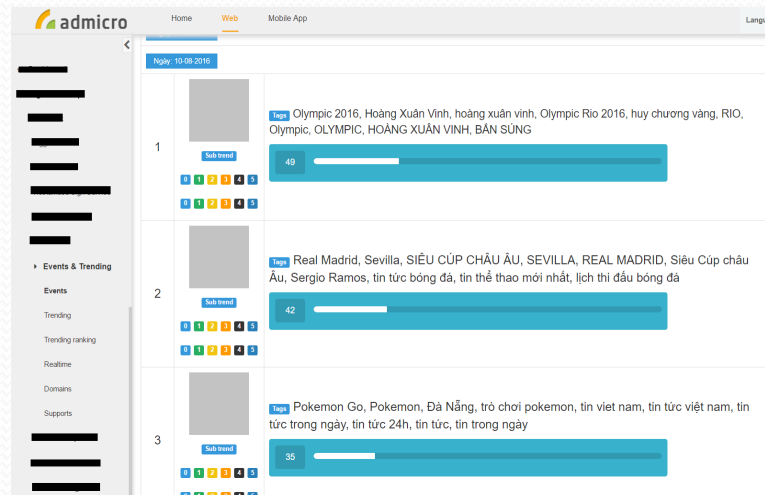
Analytic System



Auto publish

News distribution

- A key challenge of news websites is to help users find the articles that are interesting to read
- Many techniques are applied as :
 - Real-time engagement statistic
 - Personalization
 - NLP
 - Event detection
 - Trending detection
 - Breaking news detection





"Mô hình đội" Suicide Squad, vì sao DC mới hút hút trên màn ảnh rộng?



500 triệu đồng của khách hàng bay khỏi thẻ ATM Vietcombank



Chứng phải ai ci và không nên nư



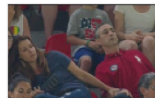
Bà mẹ trẻ dứt ruột bỏ con vì thai nhi "mọc" đuôi cá

La & Cool - 20 phút trước
Không chỉ bị dị tật con quái, em bé bất hạnh này còn có tỷ lệ sống chỉ 48%. Vì thế, sản phụ này đã đành lòng bỏ đi đứa con có 48% mạng sống trong nhiều tháng qua.



Cô gái Hà Nội đi bar thuê vệ sĩ để ra oai với bạn bè

Xã hội - 1 giờ trước
Câu chuyện đùa như thật - tuyến vệ sĩ với mức lương hấp dẫn dẫn buổi họp tập của một cô gái tên Facebook nhận được nhiều ý kiến trái chiều từ dân mạng.



Phản ứng "bá đạo" của cặp phụ huynh Mỹ khi xem con gái thi đấu ở Olympic

Sport - 1 giờ trước
Bố mẹ của Aly Raisman trở thành hiện tượng trên mạng nhờ phản ứng hài hước khi xem cô con gái thi đấu.



Netizen: Khác biệt duy nhất giữa Black Pink và 2NE1? Black Pink xinh hơn!

Huawei - 1 giờ trước
Nhiều ý kiến cho rằng Black Pink và 2NE1 giống nhau về hình tượng và phong cách âm nhạc.



Tai nạn hy hữu: Người dân ông mất thăng bằng, lao thẳng vào cửa kính xe khách

Thể thao - 2 giờ trước
Trong khi xe khách đang di chuyển trên đường, một hành khách ở tỉnh Hải Phòng, Trung Quốc đột nhiên lao về phía trước và bị mắc kẹt vào cửa kính của chiếc xe.

ĐÁNG CHÚ Ý



Người phát tán tin nhắn riêng tư của Vũ Cát Tường có thể bị





Xa hội

Sau 200 xe tăng T-90MS, Việt Nam sẽ mua T-72B3 với số lượng lớn hơn nhiều?

41 phút trước

T-72B3 được xem là sự kết hợp hài hòa giữa giá thành sản xuất và tính năng kỹ chiến thuật, nó đang giữ vai trò xương sống của lực lượng tăng thiết giáp Quân đội Nga.

Sân thần được chiêu chống, quý bà ăn quả đắng

1 giờ trước

Với mức giá chỉ hơn 300 ngàn đồng/kg thì chỉ là ba kích trắng, không phải ba kích rừng. Còn ba kích Tây Bắc, được dân buôn quăng cáo là "thần dược phòng thê", chỉ là cây rút gà, có hình dạng giống ba kích.

Muốn "ước gì được nấy", đừng đặt bất cứ thứ gì dưới gầm giường, ngoại trừ thứ này

1 giờ trước

Có thể nhiều người sẽ không tin nhưng nhìn từ góc độ phong thủy, nếu đặt thứ này dưới gầm giường, cuộc sống của bạn có thể sẽ gặp được những điều như ý.

Tai nạn ly kỳ, hàng loạt ô tô bất ngờ "diễn xiếc" khó hiểu trên đường phố

2 giờ trước

Đang đi chuyển bình thường, những chiếc ô tô bỗng nhiên "nhảy nhur", rồi đổ nghiêng xuống mặt đường. Cảnh tượng này khiến người xem thoát đầu vô cùng khó hiểu.

Báo cáo Mỹ khuyến TQ tham khảo kịch bản này trước khi gây chiến với Washington

RAND Corporation, công ty nghiên cứu quốc phòng quan trọng của Mỹ, mới đây đã công bố một báo cáo lớn, tiêu đề "Chiến tranh với Trung Quốc: Cảnh giá từ những điều không tưởng".

Soha

© 2016 Soha - 1587 lượt xem

Sự kiện

SỰ KIẾN HOT

Olympic 2016 tại Rio, Brazil

Đảo chính quân sự ở Thổ Nhĩ Kỳ

Những bài n

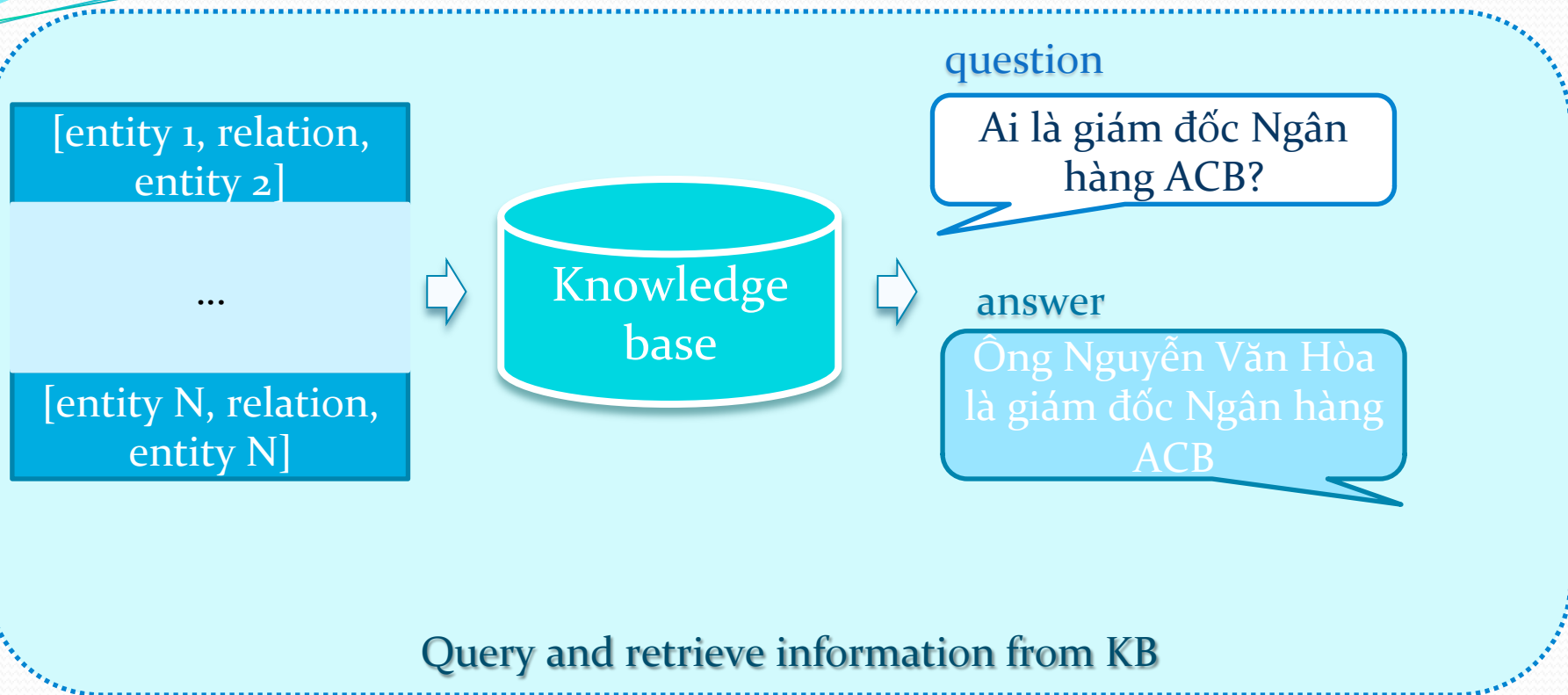


CORE NLP

CORE NLP

System	Accuracy (%) (VCCorp)	Others	Speed(/s)
Word Segmentation	98.8	97.0 (VLSp)	47,855 tokens
POS tagging	94.5	93 (VLSp)	~50k tokens
NER	87.0	85.0 (Baomoi.com)	~22k tokens
Chunking	84.0	81.0 (VLSp)	800 tokens
Universal Dependency Parser	72.0(UAS), 66.0 (LAS)	68.28% (UAS) , 66.30% (LAS)	1200 sentences
Co-reference resolution	57.0	N/A	106 docs

Entity linking Application



Input sentence

Dependency tree

Output links

"Ông Dũng là Thạc_sĩ ngành cơ_điện Trường Đại_học New_York (Mỹ) và Thạc_sĩ Quan_hệ quốc_tế Trường Đại_học Georgetown (Mỹ) . "

1	Ông	-	N	Nc	-	3	nsubj
2	Dũng	-	N	Np	-	1	compound
3	là	-	V	V	-	0	root
4	Thạc_sĩ	-	N	N	-	3	dobj
5	ngành	-	N	N	-	4	compound
6	cơ_điện	-	N	N	-	5	compound
7	Trường	-	N	N	-	5	compound
8	Đại_học	-	N	Np	-	7	compound
9	New_York	-	N	Np	-	7	compound
10	(-	((-	3	punct
11	Mỹ	-	N	Np	-	3	dobj
12)	-))	-	3	dep
13	và	-	C	CC	-	3	cc
14	Thạc_sĩ	-	N	N	-	22	nsubj
15	Quan_hệ	-	N	N	-	14	compound
16	quốc_tế	-	N	N	-	14	compound
17	Trường	-	N	N	-	14	compound
18	Đại_học	-	N	N	-	17	compound
19	Georgetown	-	N	Np	-	18	compound
20	(-	((-	14	punct
21	Mỹ	-	N	Np	-	14	compound
22)	-))	-	3	conj
23	.	-	.	.	-	3	punct

[Ông Dũng, là, Thạc_sĩ]
[Ông Dũng, là, Thạc_sĩ ngành cơ_điện]
[Ông Dũng, là, Thạc_sĩ Quan_hệ quốc_tế]
...

<Raw text>

<CoNLL format>

<Entity 1, relation, Entity 2>

"Trong năm 2015, lãi trước thuế của BIDV đạt 7.944 tỷ đồng (tăng 26,16 %), lãi sau thuế đạt 6.382 tỷ đồng (tăng hơn 28%), vốn điều lệ của BIDV tăng lên 34.187 tỷ đồng, tổng tài sản đạt 850.748 tỷ đồng (tăng 30,8%).

<Raw text>

1	Trong	-	E	E	-	10	prep
2	năm	-	N	N	-	1	pobj
3	2015	-	M	M	-	2	nummod
4	,	-	,	,	-	10	punct
5	lãi	-	N	N	-	10	nsubj
6	trước	-	E	E	-	5	prep
7	thuế	-	N	N	-	6	pobj
8	của	-	E	E	-	7	prep
9	BIDV	-	N	Ny	-	8	pobj
10	đạt	-	V	V	-	0	root
11	7.944	-	M	M	-	12	nummod
12	tỷ	-	N	N	-	10	dobj
13	đồng	-	N	Nu	-	12	compound
14	(-	((-	12	nmod
15	tăng	-	V	V	-	12	acl
16	26,16	-	M	M	-	15	dep
17	%	-	%	%	-	12	punct
18)	-))	-	12	compound
19	,	-	,	,	-	12	punct
20	lãi	-	N	N	-	12	compound
21	sau	-	E	E	-	12	prep
22	thuế	-	N	N	-	21	pobj
23	đạt	-	V	V	-	12	acl
24	6.382	-	M	M	-	25	nummod
25	tỷ	-	N	N	-	23	dobj
26	đồng	-	N	Nu	-	25	compound
27	(-	((-	25	compound
28	tăng	-	V	V	-	25	acl
29	hơn	-	A	A	-	28	amod
30	28	-	M	M	-	31	nummod
31	%	-	%	%	-	28	dobj
32)	-))	-	25	compound
33	,	-	,	,	-	25	punct
34	vốn điều lệ	-	N	N	-	25	compound
35	của	-	E	E	-	25	prep
36	tăng	-	V	V	-	35	pcomp
37	lên	-	V	V	-	36	xcomp
38	34.187	-	M	M	-	39	nummod
39	tỷ	-	N	N	-	37	dobj
40	đồng	-	N	Nu	-	39	compound
41	,	-	,	,	-	39	punct
42	tổng	-	N	N	-	39	compound
43	tài sản	-	N	N	-	42	compound
44	đạt	-	V	V	-	37	xcomp
45	850.748	-	M	M	-	46	nummod
46	tỷ	-	N	N	-	44	dobj
47	đồng	-	N	Nu	-	46	compound
48	(-	((-	46	nmod
49	tăng	-	V	V	-	46	acl
50	30,8	-	M	M	-	49	dep
51	%	-	%	%	-	25	npadvmod
52)	-))	-	25	compound
53	.	-	.	.	-	10	punct

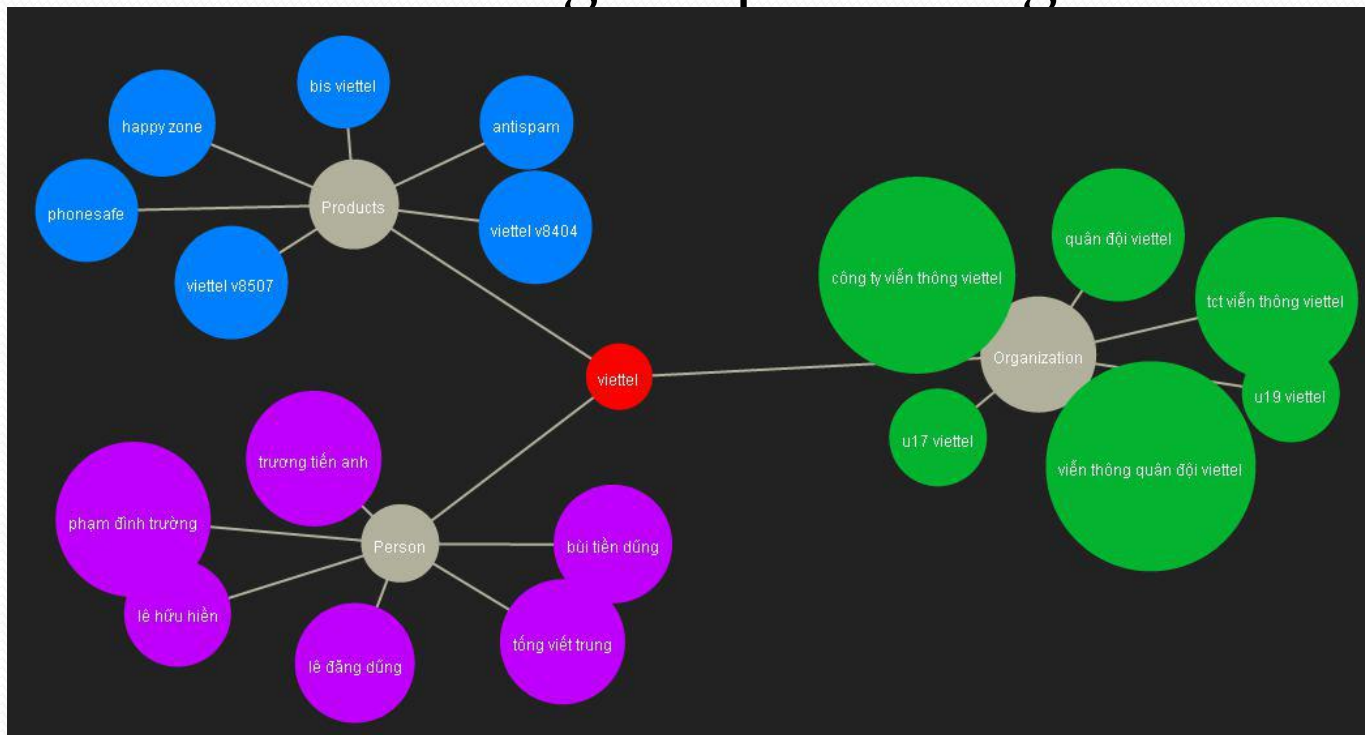
<CoNLL format>

[lãi trước thuế BIDV, đạt 7.944 tỷ trong, năm 2015]
[lãi sau thuế BIDV, đạt 6.381 tỷ đồng trong, năm 2015]
[vốn điều lệ BIDV, tăng 34.187 tỷ đồng trong, năm 2015]
[tổng tài sản BIDV, đạt 850.748 tỷ đồng trong, năm 2015]
...

<Entity 1, relation, Entity 2>

CORE NLP - Knowledge Network

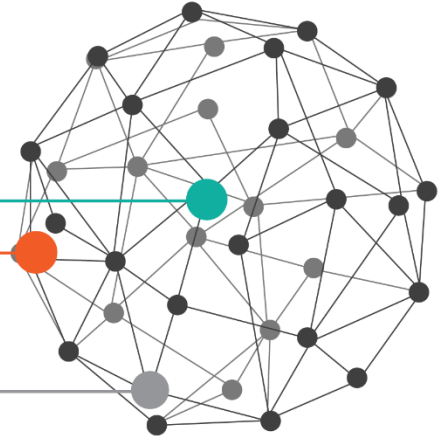
- Building Relevant Brand using Deep Learning and Core NLP



CORE NLP - NER

Named Entity Recognition

iphone 6 và iphone 6 plus là những điện thoại thông minh được thiết kế bởi tập đoàn Apple. Những thiết bị này là 1 phần trong chuỗi các sản phẩm Iphone và được ra mắt lần đầu tiên vào ngày 09/09/2014 và chính thức lên kệ vào ngày 19/09/2014. Iphone 6 và iphone 6 plus chính là thế hệ tiếp theo của iphone 5c và iphone 5s.



iPhone 6 và iPhone 6 Plus



Apple là tập đoàn công nghệ đa quốc gia của Mỹ, được sáng lập bởi Steve Job.

Sentiment Analysis



Sentiment Analysis

- Level: Doc, sentence, entity, aspect level
- Data: ~ 1 billion records , 1 TB processing data
 - Facebook: 5M pages, 500k groups
 - News : 500
 - Forums: 200
- Approach: Using NLP + Topic Modeling + Deep Learning
- Accuracy: ~70%

Sentiment Lexicon (Social)

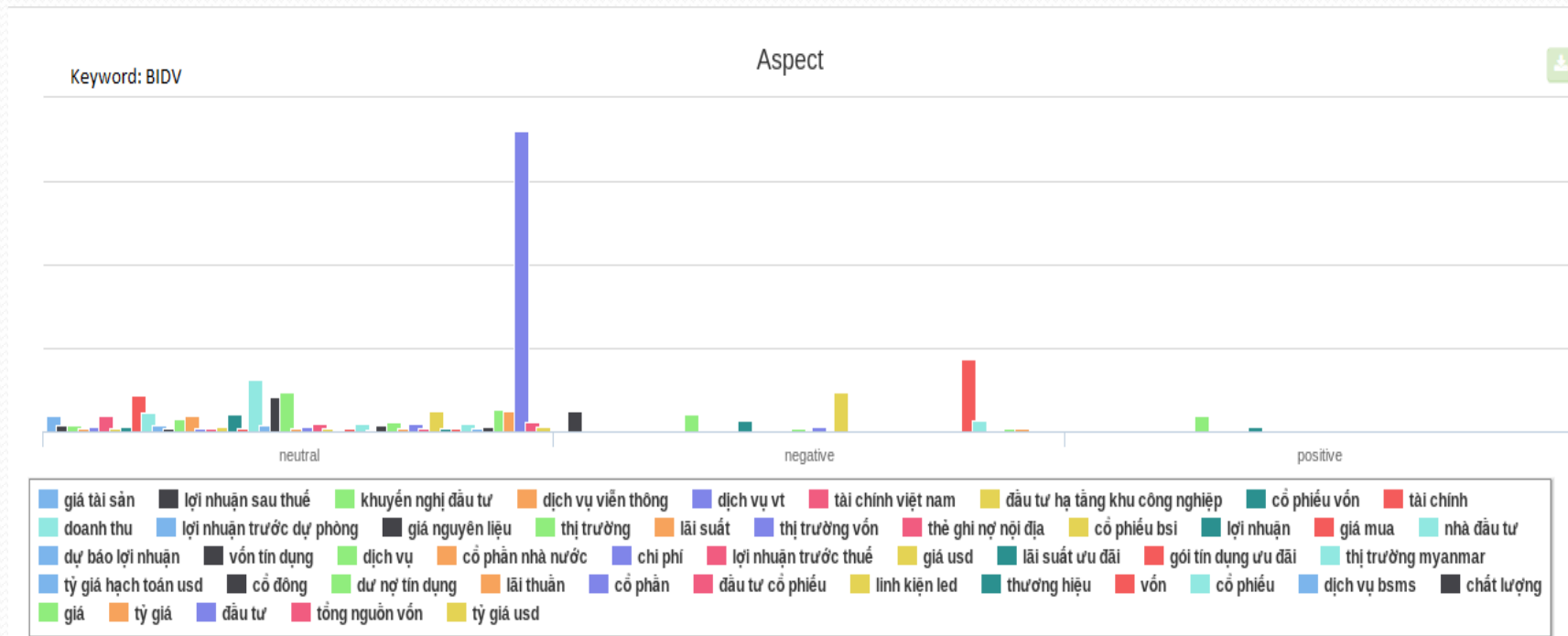
Word: đẹp Position in vocabulary: 134

Word	Cosine distance
xinh	0.720612
Đẹp	0.642578
đẹp	0.641975
đẹp	0.615595
dễ_thương	0.598402
nuột	0.598391
xinhhh	0.585283
ngầu	0.579991
đáng_yêu	0.579326
bảnh	0.570445
Xinh	0.548631
đẹp_trai	0.545112
xinhhhh	0.540463
kute	0.534398
đẹpj	0.529948
ep	0.528245
hotttt	0.516584
hoành_tráng	0.505882
manly	0.505736
đẹppppp	0.504885
dễ_thương	0.504518
xinhhhhh	0.500797

Word: kì_thị Position in vocabulary: 20300

Word	Cosine distance
kỳ_thị	0.695137
xúc_phạm	0.540247
khinh_rẻ	0.534342
hắt_hủi	0.532162
đả_kích	0.531202
coi_thường	0.525783
ngược_đãi	0.522725
chê_trách	0.519394
bắt_nạt	0.517683
ăn_hiếp	0.514011
tiêm_nhiễm	0.508612
cô_lập	0.494751
ghế_lạnh	0.488464
bỏ_rơi	0.487330
sỉ_nhục	0.485879
chỉ_trích	0.484081
nhồi_sọ	0.473971
khinh_thường	0.471812
chọc_tức	0.469445
hoang_tưởng	0.469351
chê_cười	0.467271
lừa	0.464821
lừa_dối	0.464500
gán_ghép	0.463428

Aspect based sentiment analysis



RECOMMENDATION ENGINE

Recommendation Engine

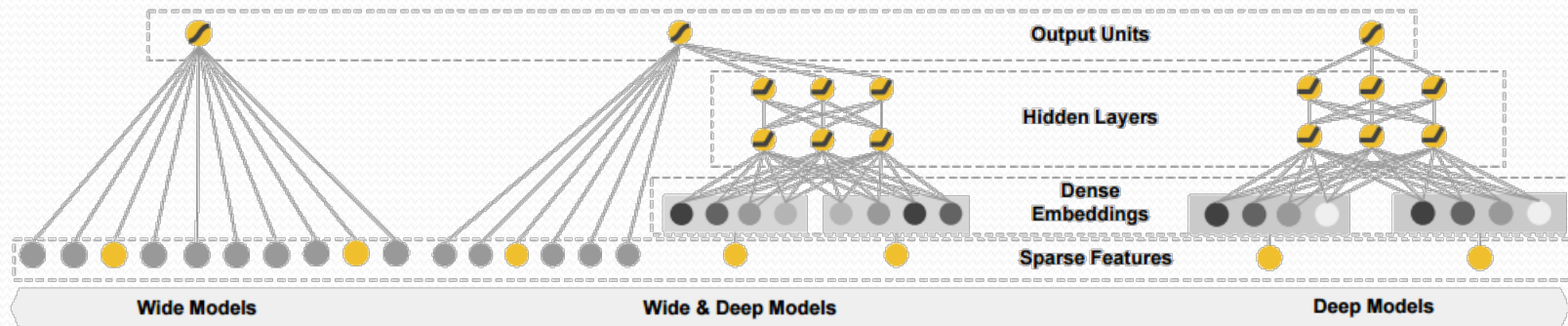
- Building purchase recommendation system for e-commerce sites
- Our suggestion based on information
 - Purchase History and web-browser history
 - Product and buyers knowledge



Recommendation Engine

- The algorithm applied:
 - NER + Deep Neural Network
 - Network and Product knowledge
 - Collaborative filtering
 - F-CTR: combines collaborative algorithm and product knowledge

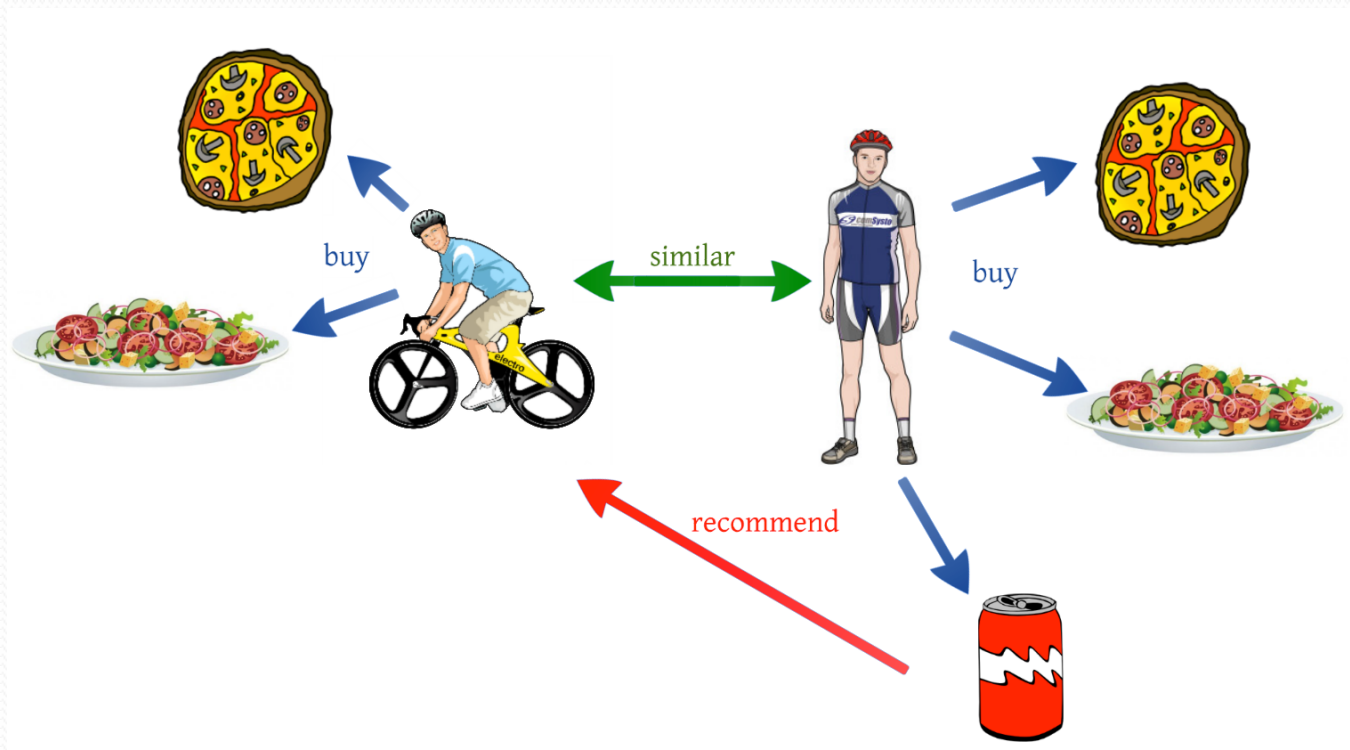
Recommendation Engine – deep learning



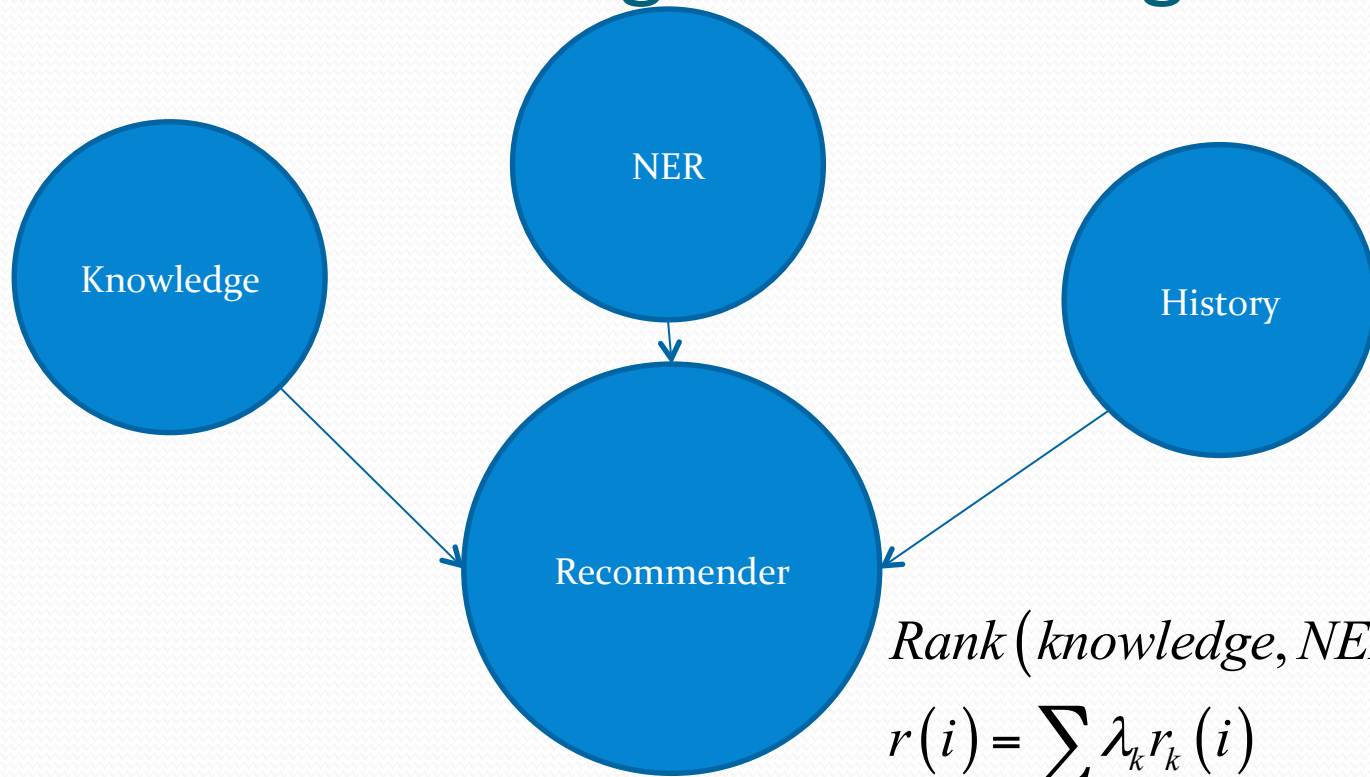
Recommendation Engine



Recommendation Engine – Collaborative Filtering



Recommendation Engine – Ranking

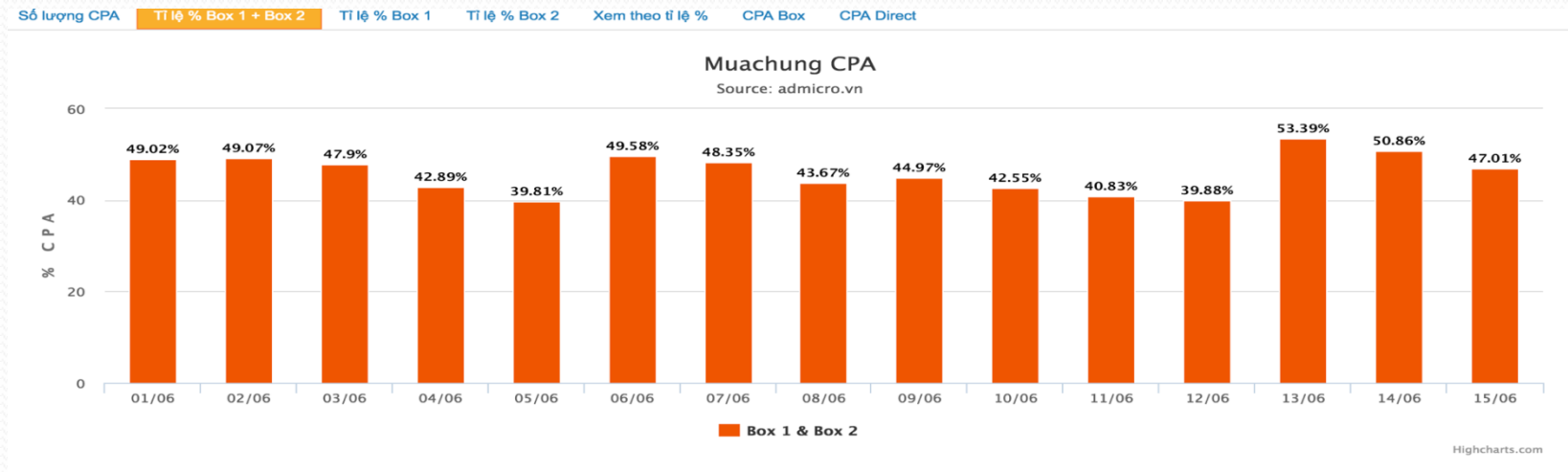


$Rank(knowledge, NER, history)$

$$r(i) = \sum_k \lambda_k r_k(i)$$

REC Performance

Increase 45% traffic from the Recommend Engine boxes



Recommendation Engine - News

LinkHay Media | 2872 Tin mới | Bạn bè 360 Gửi tin Gửi media Gửi quick

18

Hay

Tin mới vụ cao tốc Pháp Vân – Cầu Giẽ thu chênh 700 triệu đồng/ngày • 60 tỷ/tháng báo cáo nộp 32 tỷ/tháng
– ăn cắp (375 clicks)



phonglee2404 gửi • Thời sự • tienphong.vn

Hot 4 giờ trước • Loan tin • CONGTM09

Gợi ý đọc từ bot www.tienphong.vn

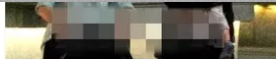
↑

Cao tốc Pháp Vân – Cầu Giẽ thu chênh 700 triệu đồng/ngày: Cơ quan công an cần vào cuộc • tienphong.vn • 30.7.2016

Giật mình chênh lệch thu phí Cao tốc Pháp Vân – Cầu Giẽ • tienphong.vn • 27.7.2016


Thanh tra Chính phủ vào cuộc vụ 2 tuyến buýt đầu thầu bị tố sai phạm • tienphong.vn • 26.7.2016

Cao tốc Pháp Vân – Cầu Giẽ: 10 ngày thu gần 20 tỷ tiền vé • tienphong.vn • 22.7.2016




Bây là lần đầu Đại S lọt vào ống kính phóng viên kể từ sau khi có hạ sinh quý tử "mẹ tròn con vuông" cho nhà họ Ưng.


@ DÀNH CHO BẠN




Bỏ việc đi, thử nộp hồ sơ vào vị trí nhân viên chăm sóc nhím lương 690 triệu đồng/ năm



Thanh Thảo mang 200kg hành lý ra Hà Nội để phục vụ cho liveshow



Con đường đi



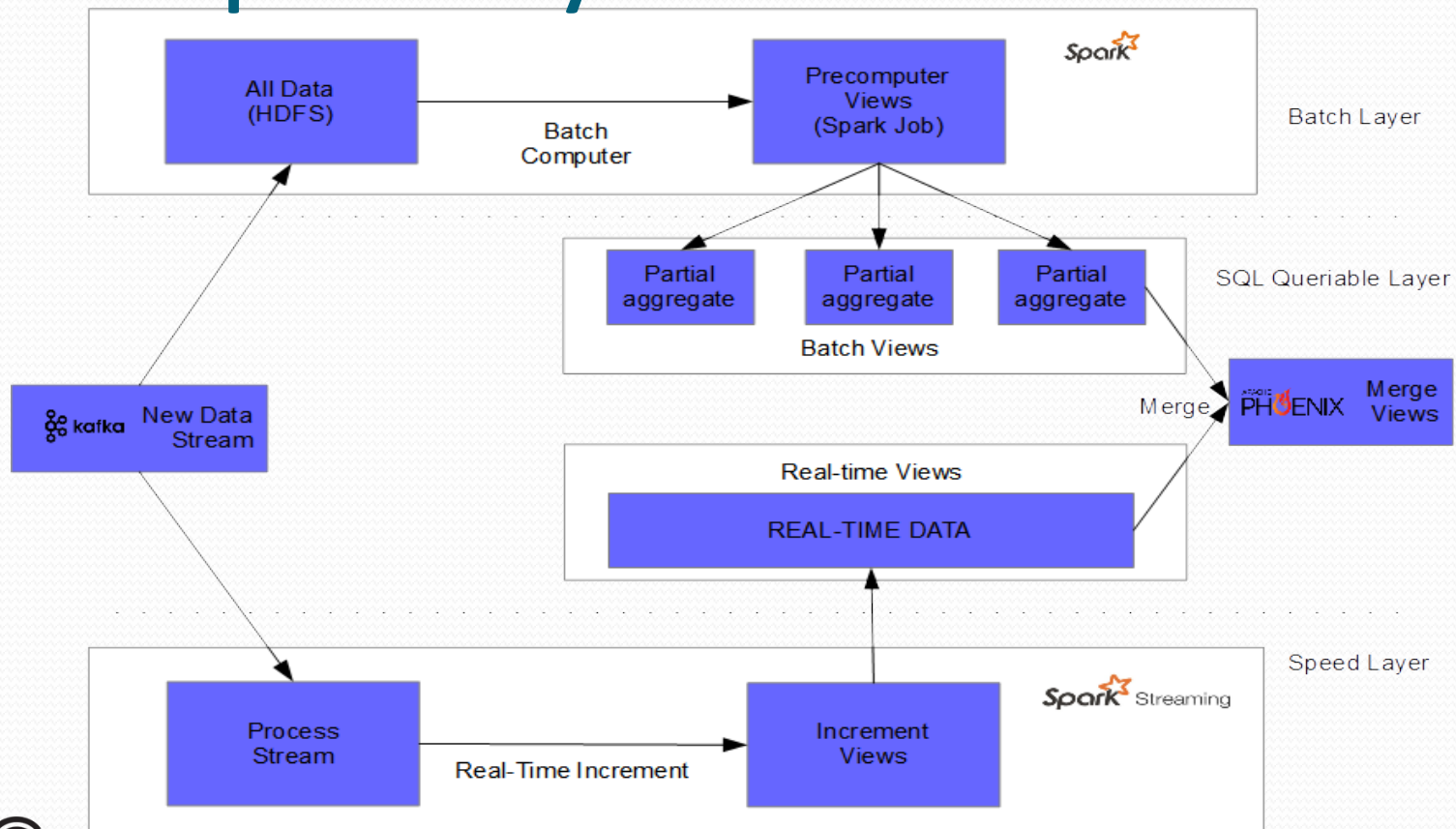
iPhone 7 sẽ là chiếc smartphone không sợ nước

VCCORP ANALYTIC

Vccorp Analytic

- Developing Analytic Tool for websites, showing good performance in compared with Google Analytic(GA) in Vietnam
- Technologies:
 - No-SQL selected as data-warehouse for large-scale data analysis
 - Real-time analytic : Streaming logging


Vccorp Analytic - architecture



Vccorp Analytic – Framework

APACHE
PHOENIX

 **kafka**

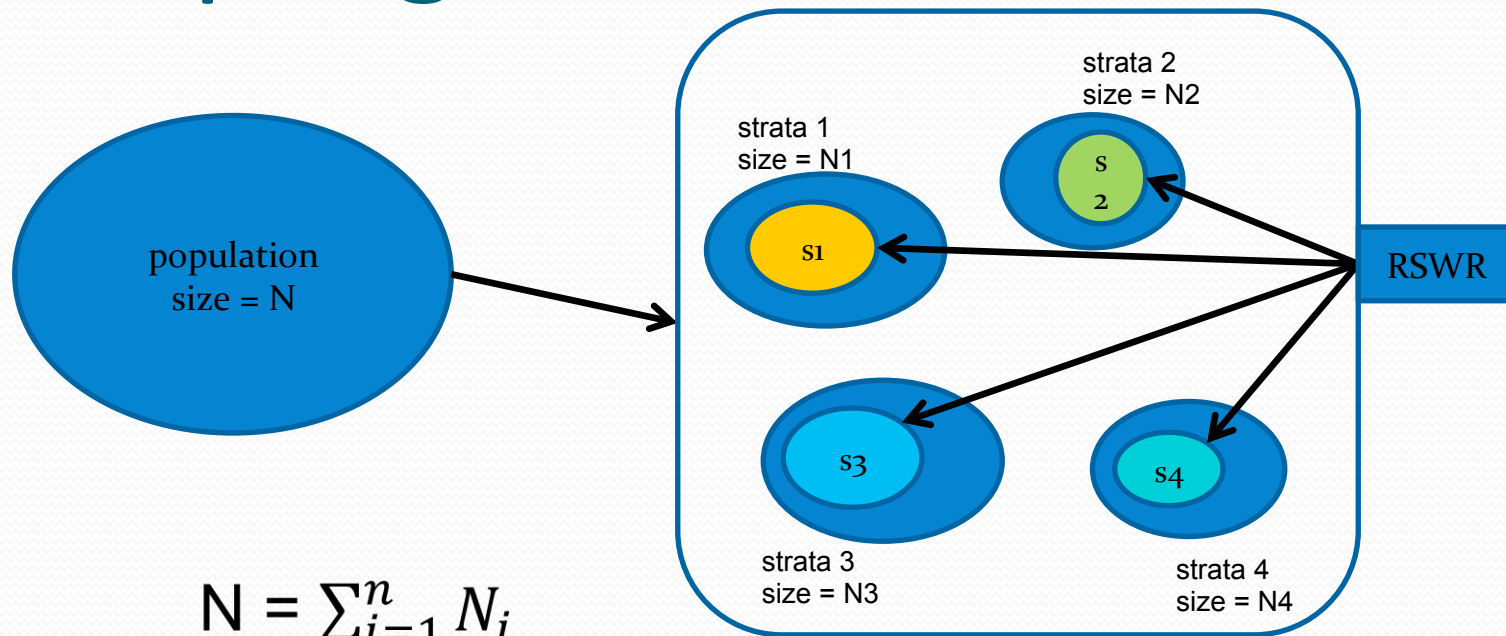
Apache
Spark &
 **Scala**


cassandra

Vccorp Analytic

- The algorithm applied:
 - Sampling data
 - Abnormal detection (remove fault clicks/sessions)
- Results: a good candidate to replace GA, ensuring both accuracy and performance

Sampling Data

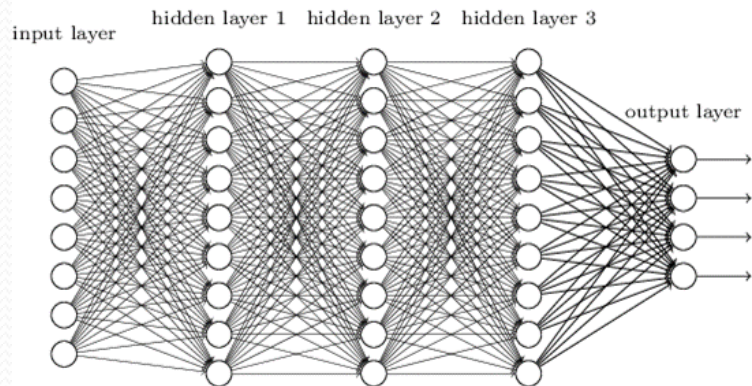


$$N = \sum_{i=1}^n N_i$$

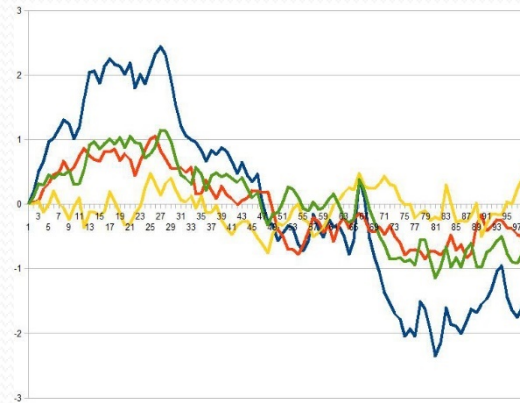
- N: Size of the sampling data
- n: Total strata
- N_i : Size of the i^{th} strata

Vccorp Analytic – abnormal detection

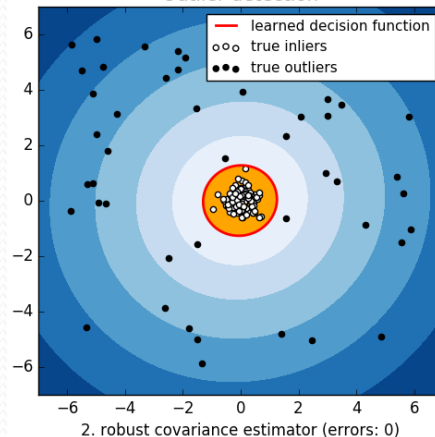
Deep neural network



Regression



Outlier detection



Vccorp Analytic

Right now
261913

active users on site

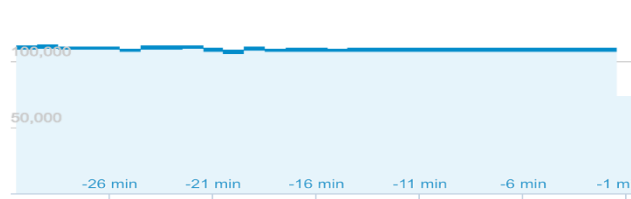


Top Referrals

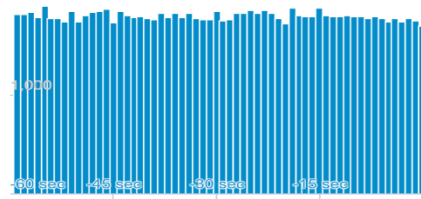
Domain	Active Users	
-1	89,448	34.15%
google.com.vn	41,263	15.75%
m.facebook.com	11,397	4.35%
facebook.com	7,751	2.96%
coccoc.com	3,239	1.24%
yan.vn	2,966	1.13%
google.com	2,820	1.08%
com.google.android.googlequicksearchbox	2,527	0.96%
hdonline.vn	2,322	0.89%
...

Page Views

Per minute



Per second



Top Active Pages Top Active Users

#	Active Page	Page Views
1	...	18,138
2	...	15,588
3	...	11,598
4	...	10,467
5	...	10,357
6	...	10,273
7	...	2,743
8	...	2,526
9	...	2,137
10

One More Thing...





Thanks