VigLink

Content-Driven Commerce @VigLink

text-analytics for content monetization at scale

Adam Pingel @pingel

Katrin Tomanek

Gabor Melli @gmelli

Introduction

This talk is about

- Business and Architecture
- Link Insertion in Depth
- Product Categorization

Affiliate Marketing

VigLink

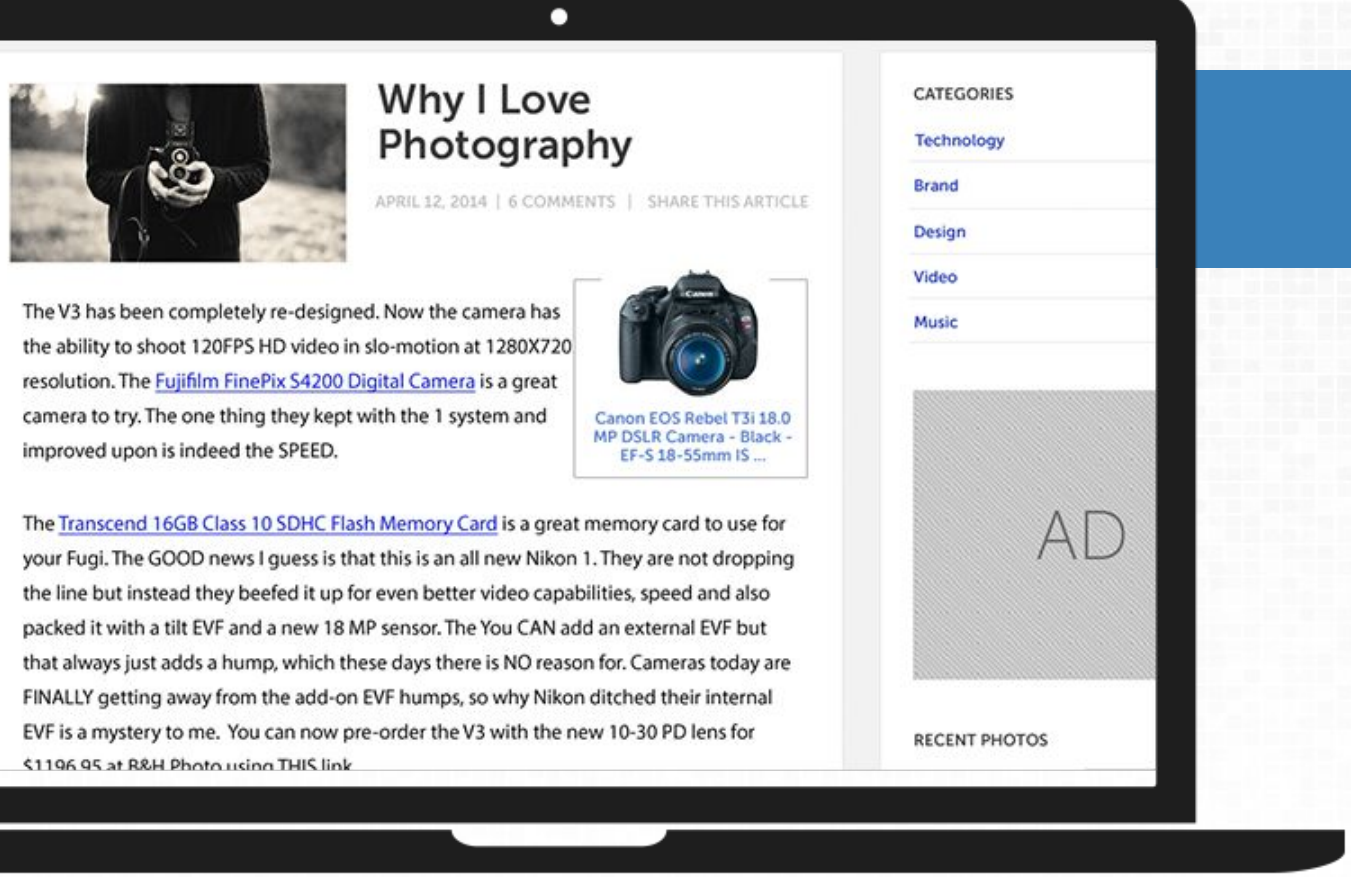
http://affiliate.acme.com/?

sub_id_1=<*publisherId*>&

sub_id_2=<*clickId*>&

url=*encode*(http://acme.com/products/t8000-headphones)

LINK INSERTION



Business

VigLink

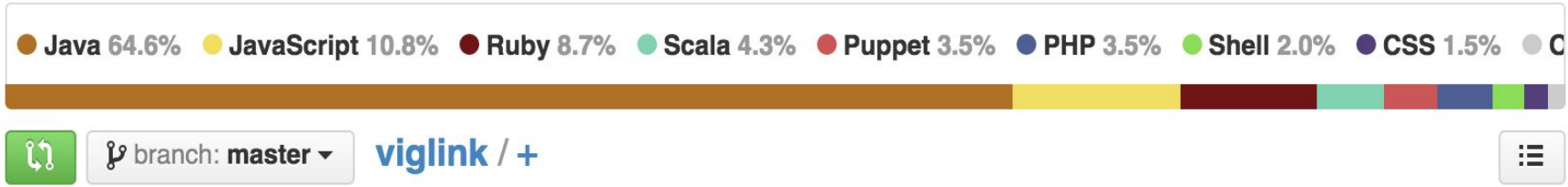
- Founded in late 2008
- Profitable Q4 2013
- $20 million Series C in 2014
- Emergence, Google Ventures, First Round Capital, RRE, Foundry, Silicon Valley Bank
- 45 employees

Metrics

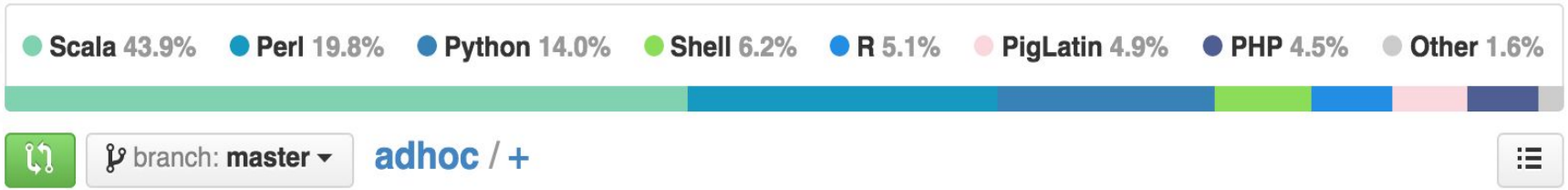
- Hundreds of thousands of publisher campaigns
- ~50k merchants (eBay & Amazon + networks like Commission Junction)
- > 10 billion page views per month
- Millions of monetized clicks/month
- Growing portion of clicks from Link Insertion

Programming Language Usage

VigLink

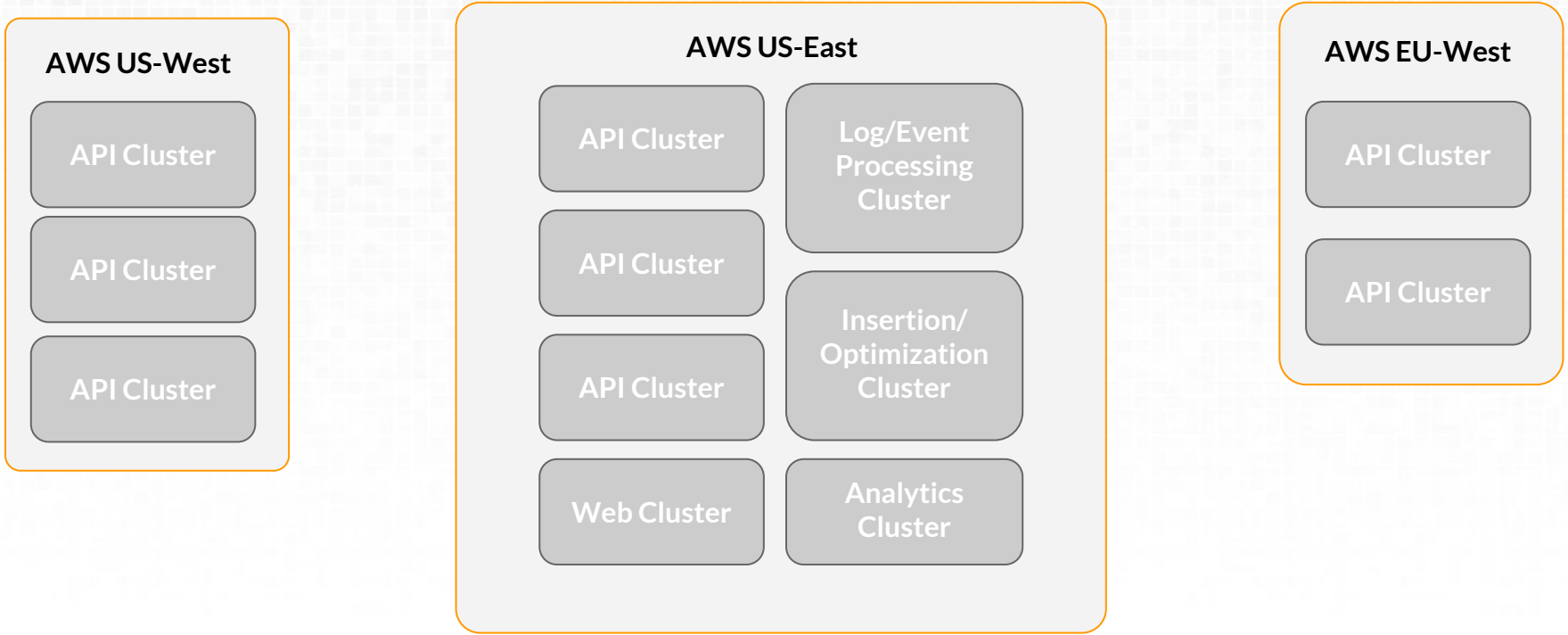


Ad-hoc Analytics

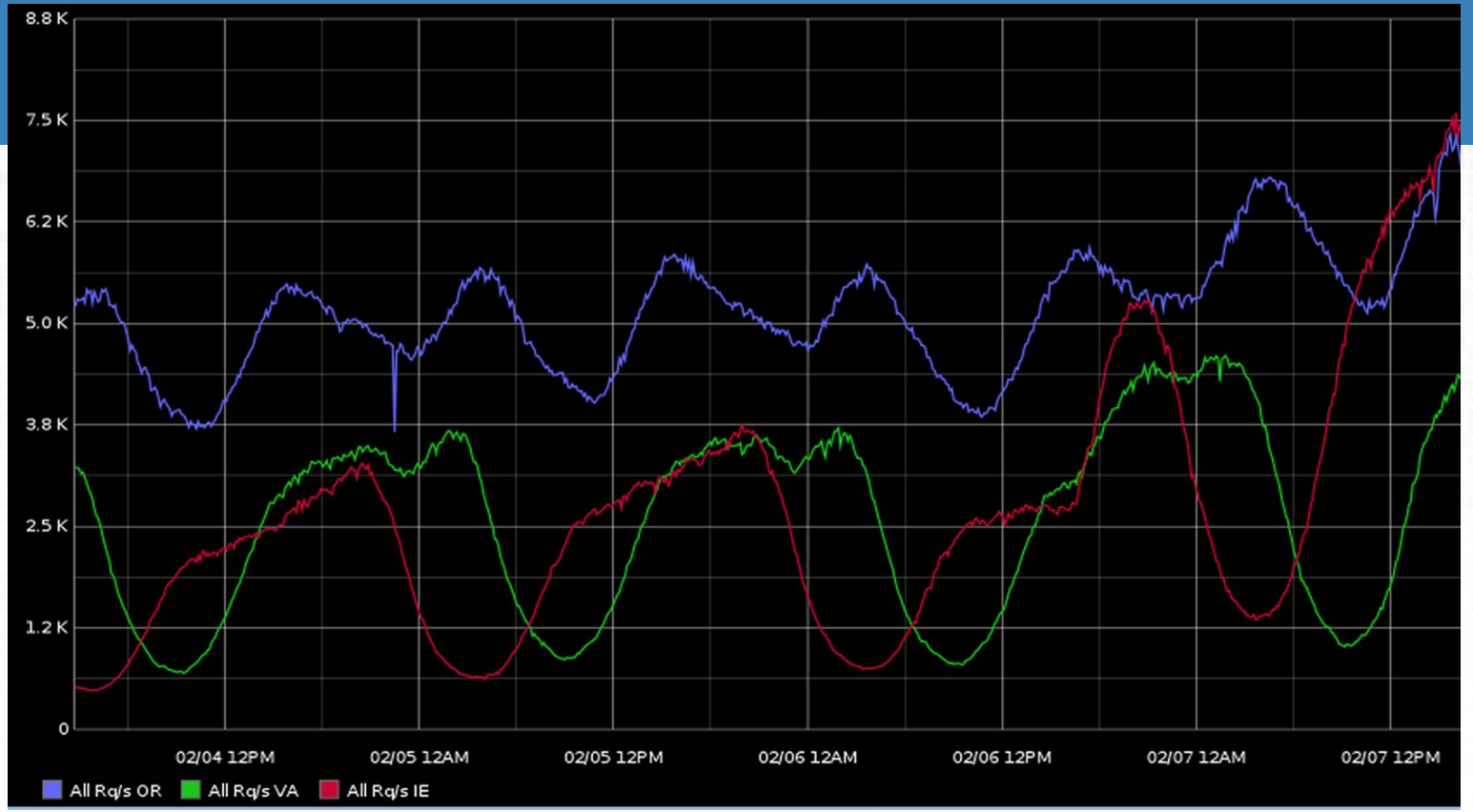


High-level AWS Architecture

VigLink

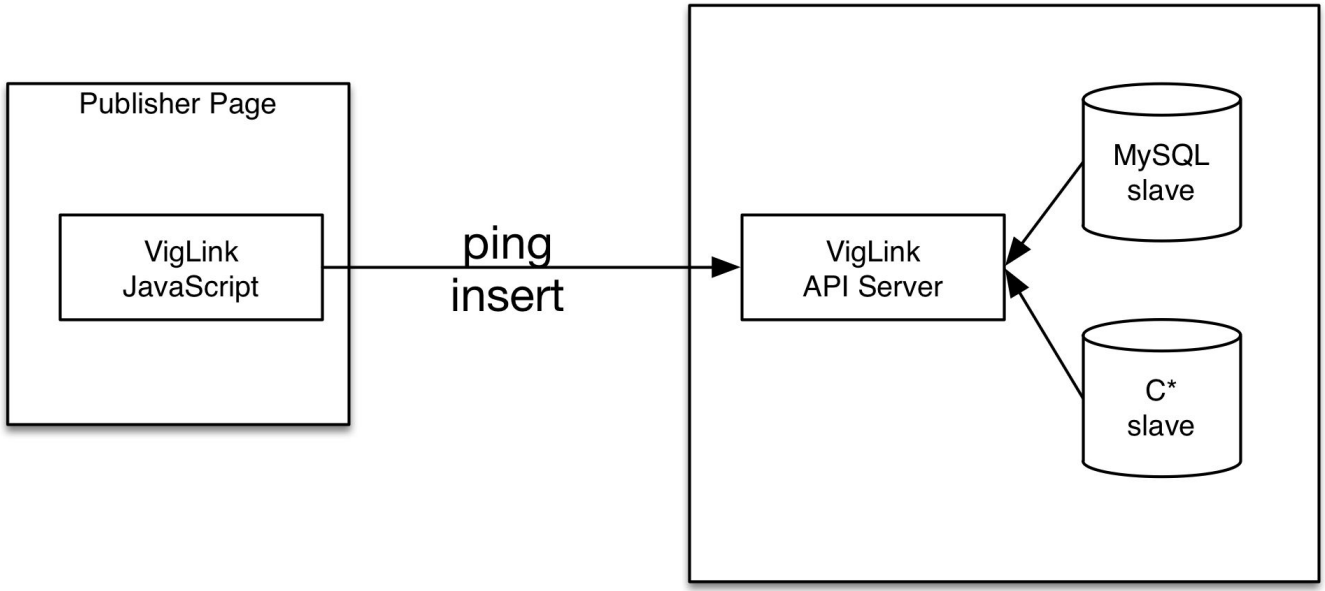


RPS
BY DC


8.8 K
7.5 K
6.2 K
5.0 K
3.8 K
2.5 K
1.2 K
0
02/04 12PM
02/05 12AM
02/05 12PM
02/06 12AM
02/06 12PM
02/07 12AM
02/07 12PM
All Rq/s OR
All Rq/s VA
All Rq/s IE

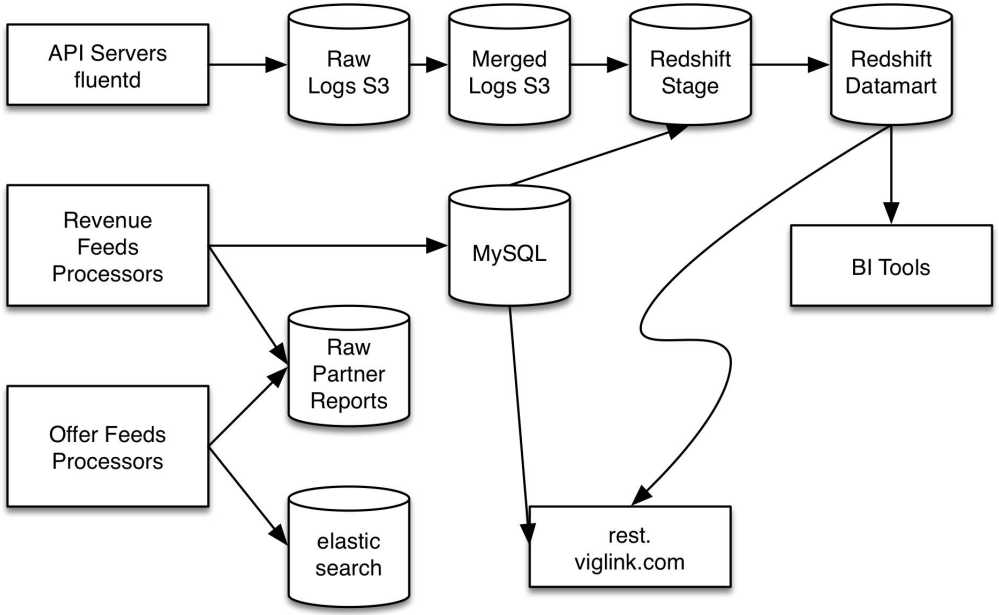
JavaScript and API

VigLink



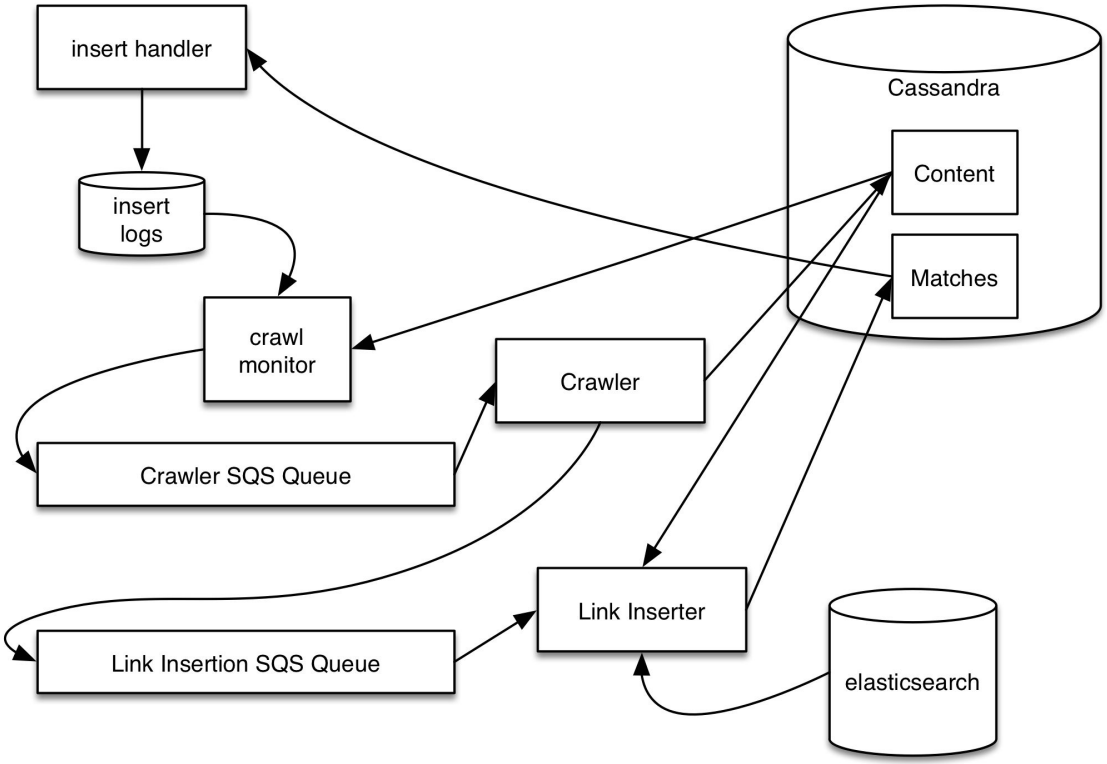
Log and Feed Processing

VigLink



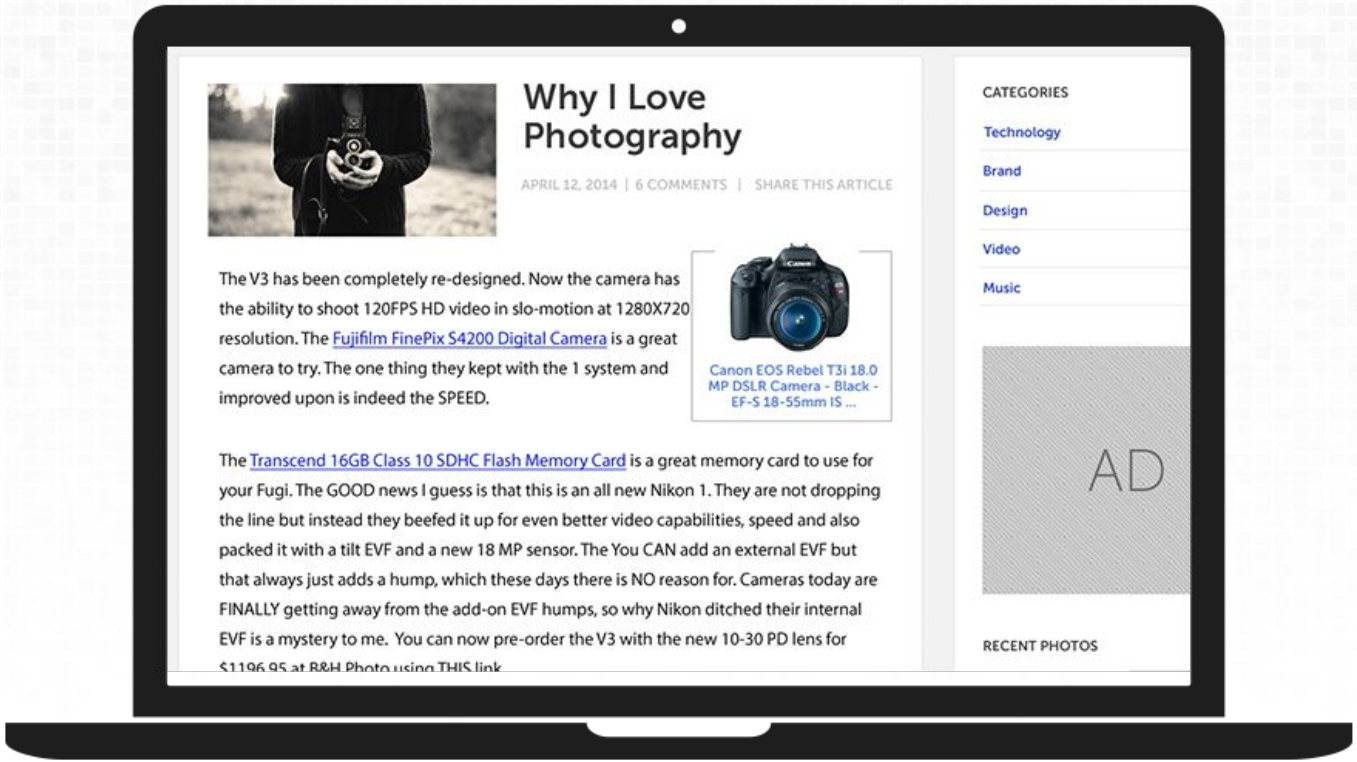
Crawling and Link Insertion

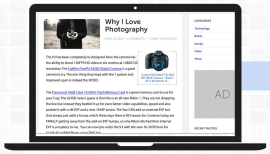
VigLink



Link Insertion in Depth

VigLink





product recognition

Nikon 1

Transcend 16GB Class 10 SDHC Flash Memory Card

Fujifilm FinePix S4200 Digital Camera

statistical modelling/ dictionaries

product-catalog linking

candidate set 1

candidate set 2

acceptable offers

semi-structured query

business logic

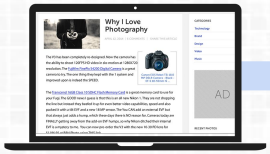
semantic filtering

auction

final offer

$

cassandra



product recognition

Nikon 1

Transcend 16GB Class 10 SDHC Flash Memory Card

Fujifilm FinePix S4200 Digital Camera

statistical modelling/ dictionaries

Fuji FilmÂ® Fujifilm Finepix S4200 Wide-Angle Camera
...
Fujifilm Finepix S Series S4200 14.0 Mp Digital Camera
...
Fujifilm Finepix S4200 14mp 24x Optical Zoom
...
Mithra Camera Bag for Fujifilm FinePix S4200

candidate set 1

candidate set 2

acceptable offers

semi-structured query

business logic

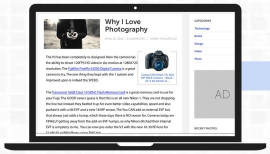
semantic filtering

selection

final offer

$

cassandra



product recognition

Nikon 1

Transcend 16GB Class 10 SDHC Flash Memory Card

Fujifilm FinePix S4200 Digital Camera

statistical modelling/ dictionaries

Fuji FilmÂ® Fujifilm Finepix S4200 Wide-Angle Camera

...

~~**Fujifilm Finepix S Series S4200 14.0 Mp Digital Camera**~~

...

Fujifilm Finepix S4200 14mp 24x Optical Zoom

...

Mithra Camera Bag for Fujifilm FinePix S4200

candidate set 1

candidate set 2

acceptable offers

semi-structured query

business logic

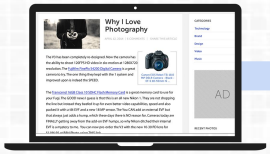
semantic filtering

selection

final offer

$

cassandra



product recognition

Nikon 1

Transcend 16GB Class 10 SDHC Flash Memory Card

Fujifilm FinePix S4200 Digital Camera

statistical modelling/ dictionaries

Fuji FilmÂ® Fujifilm Finepix S4200 Wide-Angle Camera

...

~~Fujifilm Finepix S Series S4200 14.0 Mp Digital Camera~~

...

Fujifilm Finepix S4200 14mp 24x Optical Zoom

...

~~Mithra Camera Bag for Fujifilm FinePix S4200~~

candidate set 1

candidate set 2

acceptable offers

semi-structured query

business logic

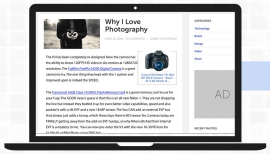
semantic filtering

selection

final offer

$

cassandra



product recognition

Nikon 1

Transcend 16GB Class 10 SDHC Flash Memory Card

Fujifilm FinePix S4200 Digital Camera

statistical modelling/ dictionaries

product-catalog linking

candidate set 1

candidate set 2

acceptable offers

semi-structured query

business logic

semantic filtering

Fuji FilmÂ® Fujifilm Finepix S4200 Wide-Angle Camera (CJ/Sears, 190$, ...)

final offer

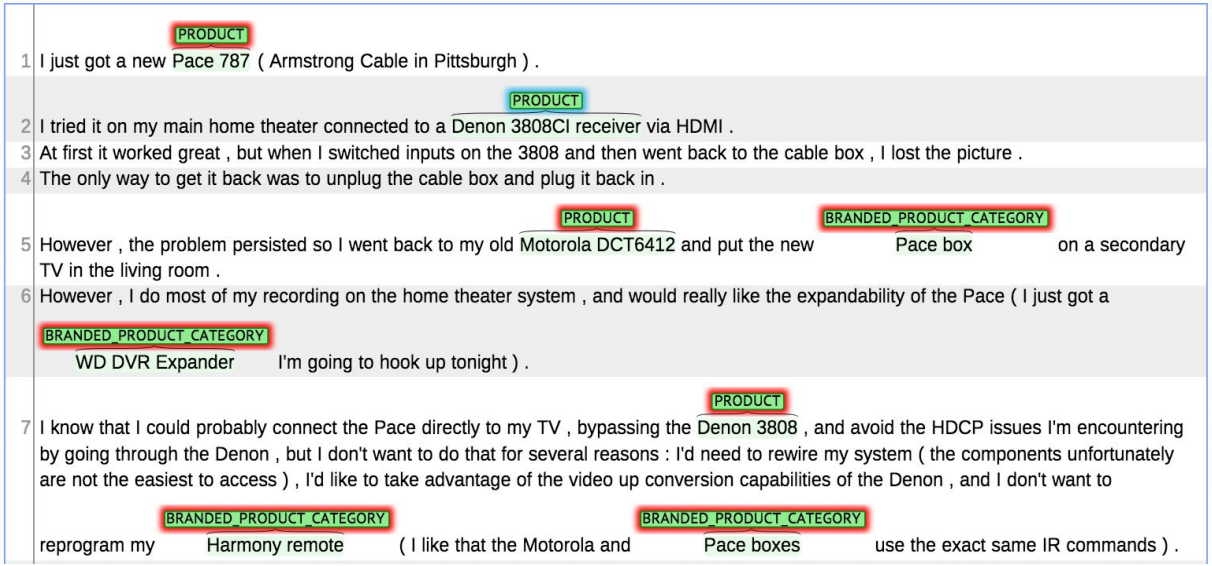
$

cassandra

NLP - related tasks

- product recognition
- linking product to offers
- offer database creation
- understanding offers
- taxonomy

Product Recognition



product-like mentions

- products
- product lines
- (branded) product categories

Product Recognition

VigLink

- typical NER problem
  - sequential tagging
  - impossible to keep all product terms in dictionary
    - spelling variations
    - new products, sub-products…
    - ambiguity, context-dependent
- hybrid approach
  - statistical models
  - dictionary
  - rule-based
  - → ensembles
    - thresholding (R-P lever)

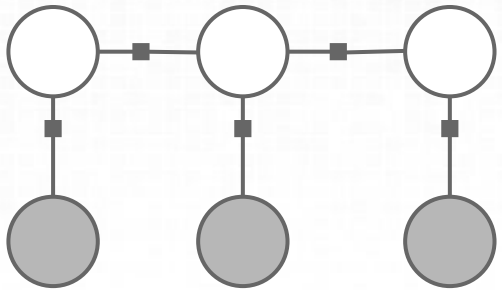
Sequence Labeling for Product Recognition

VigLink

- **Conditional Random Fields**
  - sequence labeling model (linear chain CRF)

	O	O	B-P	I-P	I-P	O	O	O	O	B-BPC	I-BPC	
...	my	old	Motorola	DCT	6412	and	put	the	new	Pace	box	...

- **features**
  - word identity
    - token (stemmed)
  - orthographical/morphological
    - word pattern
    - numbers, dashes, upper/lower case
  - contextual
    - features of neighboring tokens
    - word modifiers (pronouns, articles, quantifiers...)
  - lexicon lookups
    - brand names, typical product features, etc



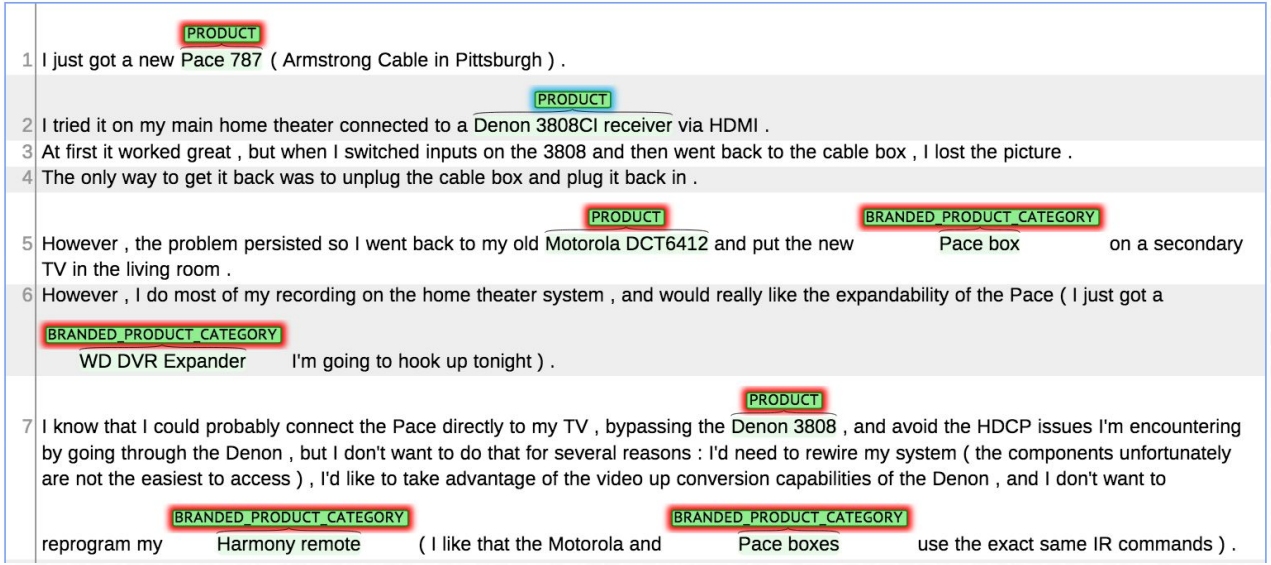
Training (& Test) Data

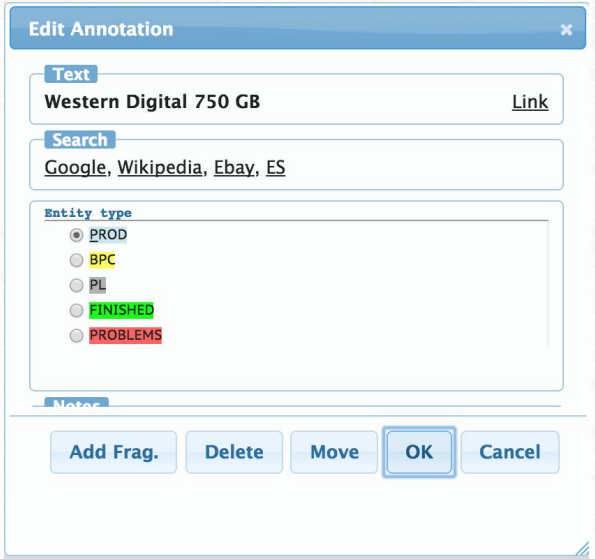
VigLink

- CPROD 1 & 2
  - manually annotated data

	AU	CE	FS	misc
# docs	709	2006	156	76
# entities	465	1572	724	1008
% P	**74%**	**77%**	36%	76%
% PL	18%	16%	**57%**	10%
% BPC	8%	8%	7%	14%

- click-logs
  - additional training data





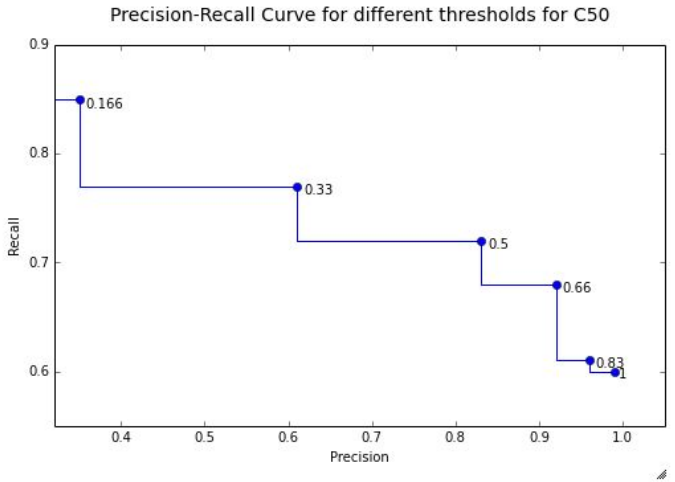
- BRAT (brat rapid annotation tool), http://brat.nlplab.org/
- highly configurable/flexible
- good for sequence labeling, relations, ...
- open-source, web-based, collaborative annotations (diffs)
- pre-labeling & correct
- extensions for IAA tooling

Performance Evaluation

VigLink

- on CPROD (subset)

	R	P	F
dict	75%	91%	82%
rules	69%	100%	82%
M11	76%	51%	61%
M12	78%	48%	60%
M16	72%	83%	77%
MP3	88%	98%	93%
ensemble	92%	43%	59%



- AB tests for CTR
  - sloppy models need to be in (“interesting” terms)

Linking Product Mentions to Offers

VigLink

Content

I"ll wait till tire rack has it up on their site to see how much it is. Everybody around me at local tire stores want $700+ for m **295/30R19**

Matches

Experiment Id	Matcher Type	Country	Phrase	Product Name	Image URL	Source	Match Date	Expiration Date
35	C52C	us	295/30r19	**LE SPORT KU39 TIRE 295/30R19 100Y BW 295-30-19**	/d3vv6xw699rjh3.cloudfront.net/e4de40-963760086_1_260.jpg	PG	2015-02-01T21:09:24.352Z	2015-02-07T00:00:00.000Z

Challenges

- unstructured offers
- accessories
  - ~ 50% in Consumer Electronics
- overloaded offer titles
- nonsensical/uninformative offer titles
- diversity
  - 33 top-level industry types
  - ~50K merchants
  - ~20 feed sources
- volume
  - ~400 million offers

Polk Audio Blackstone TL1600 Speaker sys - home theater - 5.1-CH - wired

iPhone 5 Power Button Replacement

Apple iPhone 5, Space Gray, 16 GB (Unlocked), with cover, free shipping, like new

Challenges

"iphone 5" →

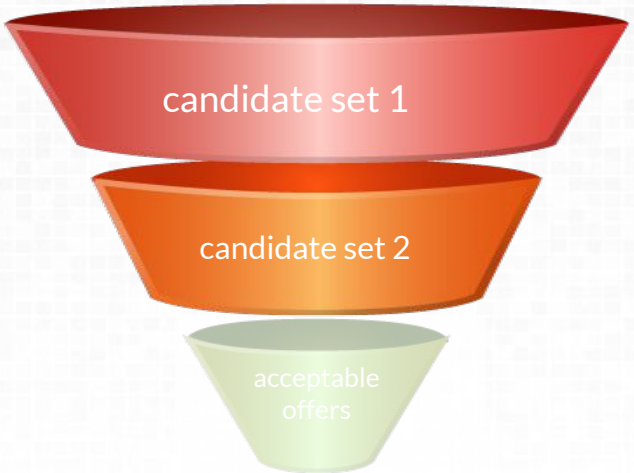
→ 650K offers!

- Armband for **iPhone 5** / 5s / 5c with a slim case (Also fits slim cases for iPhone 5, iPhone 5c, iPhone 4S and iPhone 4)
- Genuine Porsche Martini Racing **iPhone 5** iPhone 5S Case Cover
- iPhone case with your own photo **Iphone 5** Covers
- **iPhone 5** Power Button Replacement
- Apple **iPhone 5**, Space Gray, 16 GB (Unlocked), with cover, free shipping
- …

Candidate Set Construction

VigLink

- semi-structured ES query
  - normalization of product mention
  - restrictions (industry type)
- business logic
  - restrictions based on user settings
  - affiliable?



Semantic Offer Candidate Filtering

VigLink

evaluate a tuple

<product mention,offer candidate>

by these decision criteria

1. validity

"denim jeans" // white denim jeans jacket

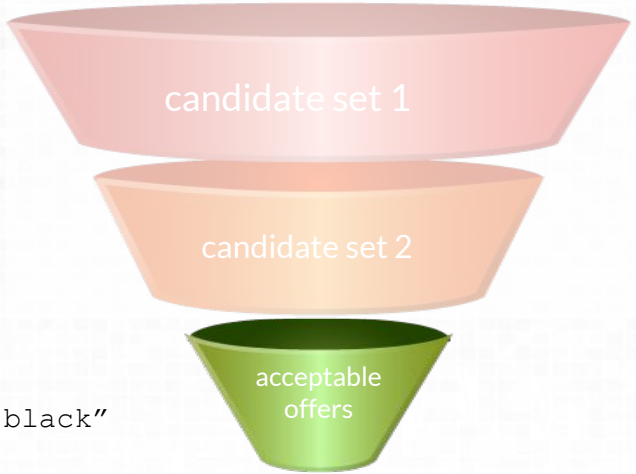
2. centrality

"matrix amp" -> "matrix mini-i 24 384 dac headphone amp black"

"vudu" // "blu-ray disc player - streams vudu videos"

3. accessory

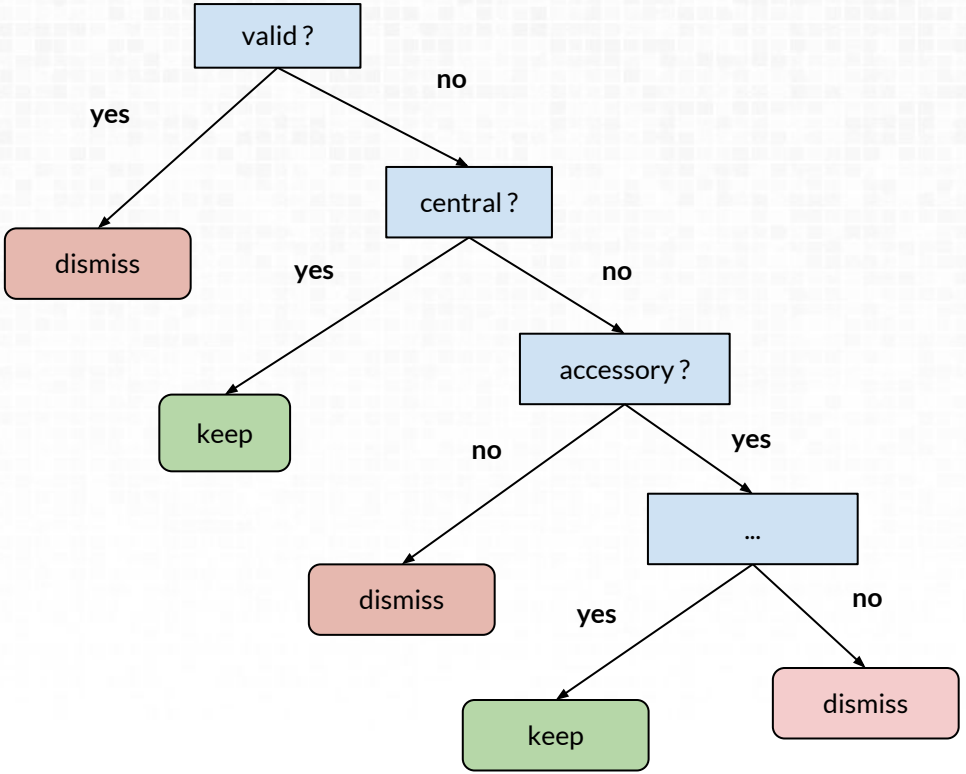
"iphone 4" -> in-ear earbuds headset for iphone 4



Semantic Offer Candidate Filtering

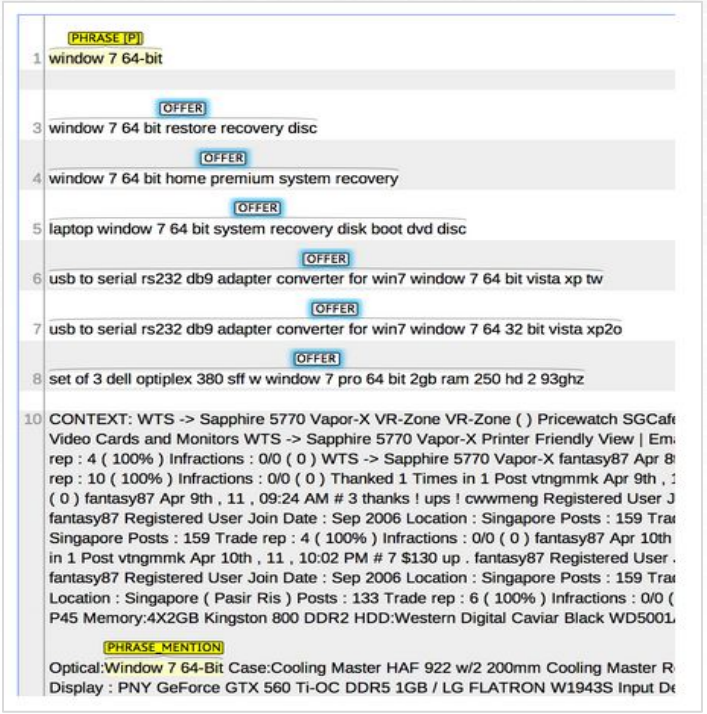
VigLink

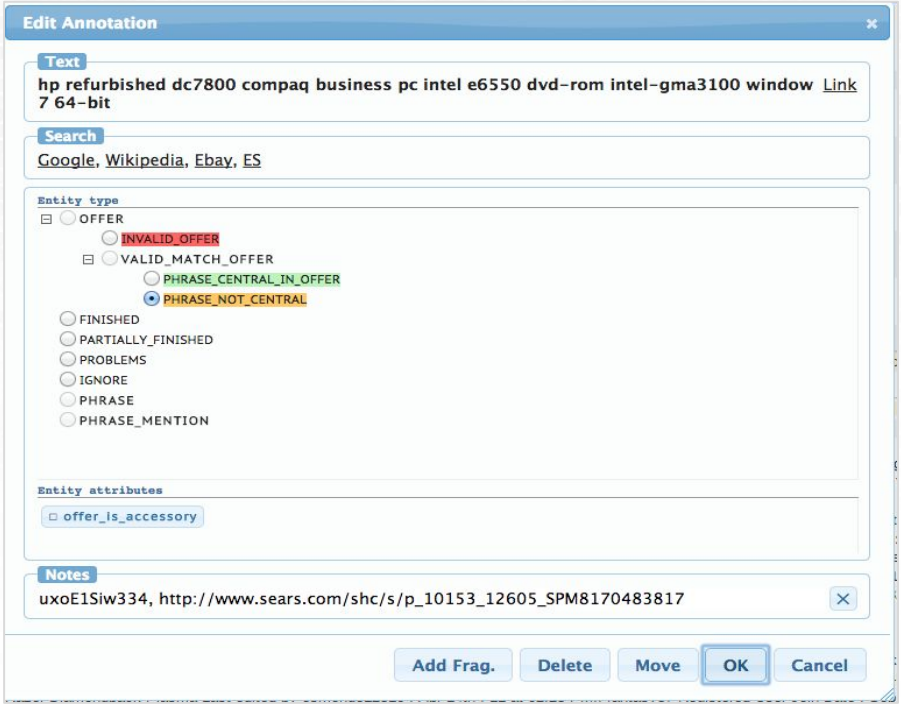
- decision logic based on these criteria
- heuristics & learning-based approaches to recognize aspects for tuples



Manually Annotated Test Set

VigLink





Manually Annotated Test Set

VigLink

- **for 450 product mentions**

	AU	CE	FS	avg
valid	68%	79%	86%	80%
central	**21%**	37%	**86%**	59%
accessory	**79%**	51%	**3%**	31%

- **recognition performance**
  - for final decision (keep/dismiss)
  - ~ 70%

Offer Feeds

- ~20 feeds (some of which come as hundreds of files)
- ~400 million offers
- pulled daily

Offer Store

v1: Tomcat & Lucene

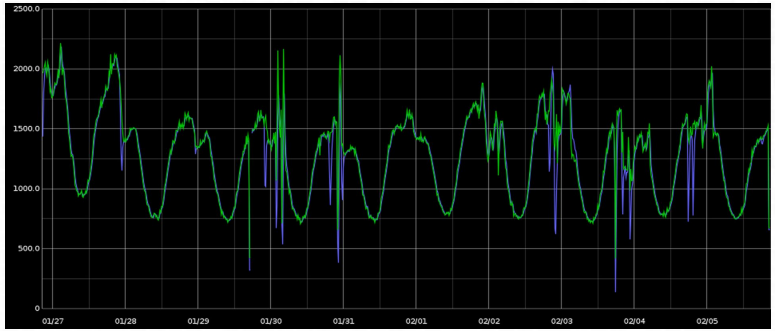
v2: Solr 3.x (x < 5) on a single box with lots of memory

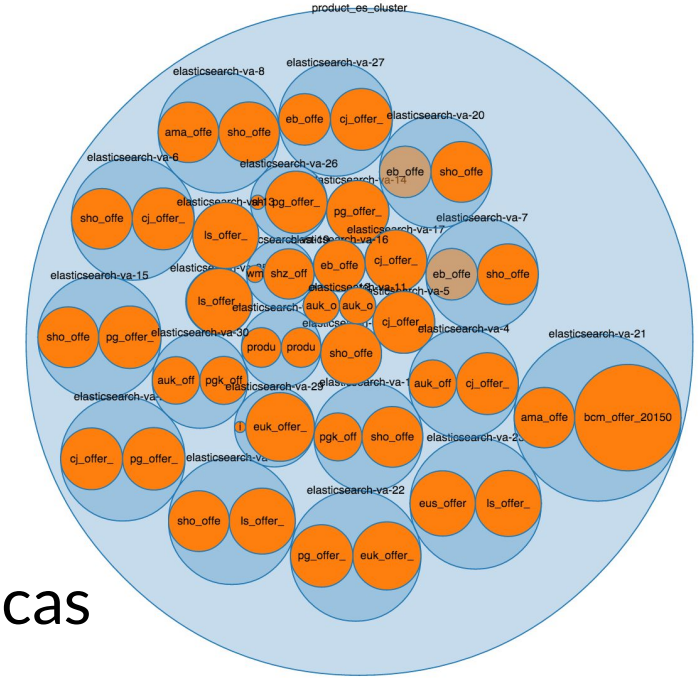
v3: Solr 3.5 with (productId mod 5) sharding and ELB load balancing. Early 2013 - early 2014

Offer Store v4: elasticsearch

VigLink

2-10k qps peak





30 m3.2xlarge nodes

1-8 shards/index. no replicas

indices unioned with alias

Raw Offer Example

VigLink

```
<item_data>
<item_basic_data>
<item_unique_id>B00D4INHD2</item_unique_id>
<item_sku>B00D4INHD2</item_sku>
<parent_asin>B008UO4Y4Q</parent_asin>
<item_upc>848447006694</item_upc>
<item_mpn>MH8U2AM/A</item_mpn>
<item_brand>Beats</item_brand>
<item_manufacturer>Beats Electronics, LLC</item_manufacturer>
<item_name>Powerbeats by Dr. Dre In-Ear Headphone (Neon Pink)</item_name>
<item_category>CE</item_category>
<item_short_desc>&lt;div class="aplus"&gt; &lt;div class="leftImage" style="width: 32
ss="leftImage" style="width: 325px"&gt;  &lt;img src="http://g-ecx.images-amazon.com/
http://g-ecx.images-amazon.com/images/G/01/aplus/detail-page/B00D4INHD2_3-sm.jpg" alt
ts.&lt;/p&gt; &lt;h4&gt;Features&lt;/h4&gt; &lt;ul&gt; &lt;li&gt;Powerful bass and hi
s for a perfectly sealed fit&lt;/li&gt; &lt;li&gt;Actual cord length: 4.3ft/1.3m&lt;/
```

Normalized Offer

VigLink


```
"_index": "euk_offer_20150209_wqm6y",
"_type": "offer",
"_id": "IDLkHGw8IyY",
"_score": 7.1824985,
▼ "_source": {
    "offerId": "IDLkHGw8IyY",
    "productId": null,
    "displayName": "Dr Dre Powerbeats Pink",
    "name": "dr dre powerbeats pink",
    "targetUrl": "http://rover.ebay.com/rover/1/710-53481-19255-0/1?
    toolid=10029&campid=CAMPAIGNID&customid=CUSTOMID&catId=293&type=2&ext=161
    "imageUrl": "http://i.ebayimg.com/00/s/NTMyWDYwMA==/z/YX4AAOSw7ThUjwQJ/$_1.JP
    "expirationDate": "201501140000",
    "source": "EUK",
    ▼ "categories": {
        "1": "CE"
    },
    "listingId": "161522090101",
    "feedMerchantId": null,
    "networkProductId": null,
    "merchantName": "ebay.co.uk",
    "epc": null,
    "price": 83,
    "country": "uk",
    "category": "CE",
    "_timestamp": 1418848860000,
    ▼ "asin": [ ],
    ▼ "upc": [ ],
    "_ttl": 2344740000
```


Offer Import Streams

- parse (csv, xml, …)
- extract fields
- write to elasticsearch as _write
- write to s3 as json
- swap to _read when complete
- drop old index
- 3-day TTL

Akka Streams coming soon (async; backpressure)

Offer Understanding

VigLink

"Polk Audio Blackstone TL1600 Speaker sys- home theater - 5.1-CH - wired"

Brand Name (BN): *Polk Audio*

Product Line (PL): *Blackstone*

Product (PI): *TL 1600*

Product Category (PC): *speaker sys*

Product Feature (PF): *home theater*

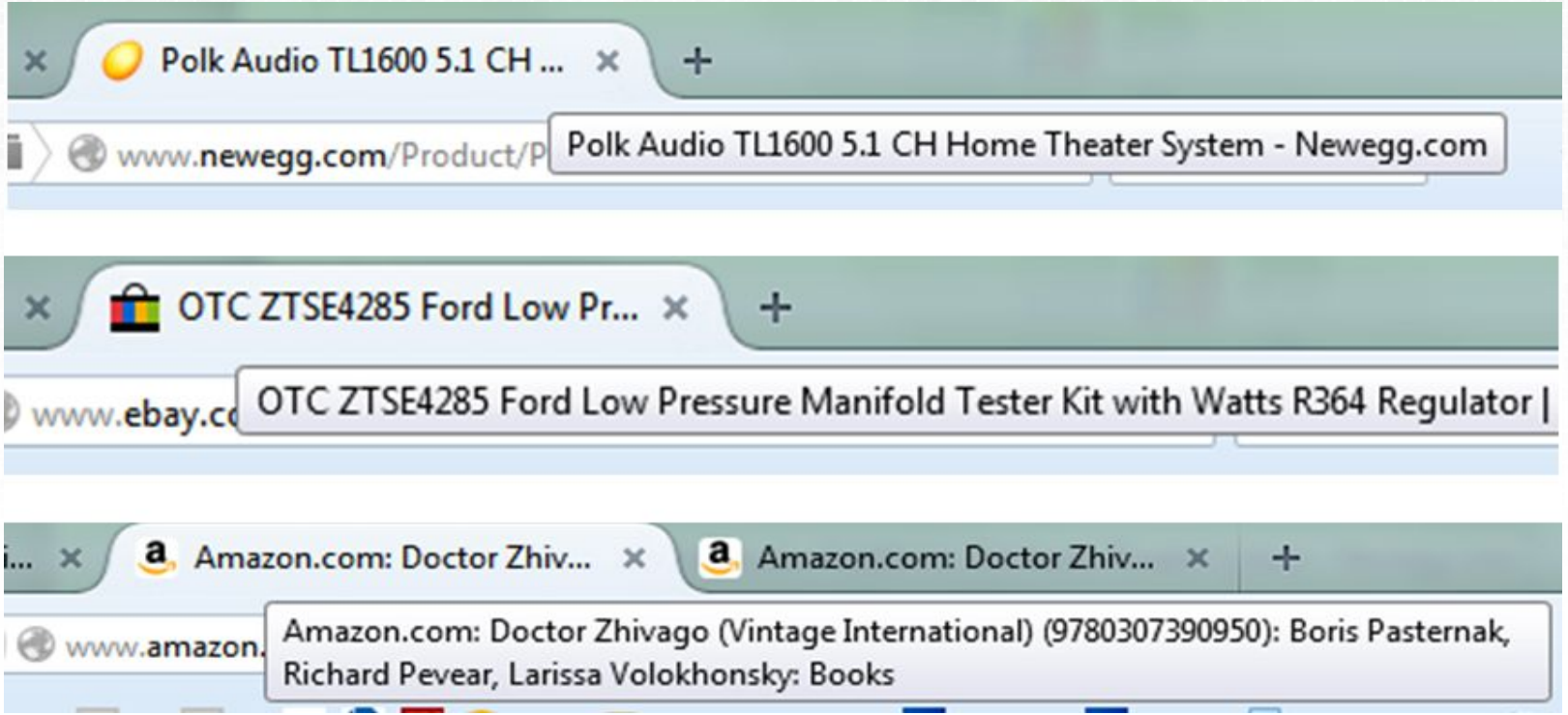
Product Feature (PF): *5.1-CH*

Product Feature (PF): *wired*

[Polk Audio]^{BN} Blackstone^{PL} TL1600^{PL} [Speaker sys]^{PC} [home theater]^{PF} [5.1-CH]^{BN} [wired]^{PF}

>~ 1B Offers w/ Unstructured Titles

VigLink



Gabor Melli's Research Knowledge Base

2014 ShallowSemanticParsingofProduct

- **(Melli, 2014)** ⇒ Gabor Melli. (2014). "Shallow Semantic Parsing of Product Offering Titles (for Better Automatic Hyperlink Insertion)." In: Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining. ISBN 978-1-4503-2956-9 doi:10.1145/2623330.2623343

Subject Headings: Automated Annotation, Product Offer Title.

Notes

- Presentation slides: File:2014 ShallowSemanticParsingofProduct.kdd2014.pdf

Navigation

Main page
Recent changes
Random page

Tools

What links here

Manually Annotated Records

VigLink

Annotated product offering title	Industry
Tripp[PI] *LS606M*[PF] *[600 watt]*[PC] *[Line Conditioner]*[PF] *[6 Outlet]*[PF] *[120 volt]*[QQ]	Electronics
Samsung[PF] *51-Inch*[PF] *720p*[PF] *600Hz*[PF] *Plasma*[PF] *HDTV*[PC] *(Black*[PF] *)*	Electronics
[Wheelskins][PF] *[Genuine Deerskin]*[PC] *[DRIVING GLOVES]*[PC] *- Tan*[PF] *([Size Large]*[PF] *)*	Automotive
[1990-1997][PF] *[Mazda]*[BN] *[Miata]*[PL] *AxleBack*[PL] *[Exhaust Bolt]*[PC] *on*[FT] *Muffler*[PC]	Automotive
[Cloudy with a Chance of Meatballs][PI] *(Two-Disc*[PF] *[Blu-ray / DVD Combo]*[PF] *)*	Entertainment
[A Pius Man : A Holy Thriller][PI] *([The Pius Trilogy]*[PL] *) ([PF|Volume 1]*[PF] *)*	Books
[Amazing Herbs][BN] *[Black Seed]*[PF] *[Cold-Pressed Oil]*[PC] *- 32oz*[PF]	Health & Beauty
([180 days wrty.][OF] *) ZeroLemon*[BN] *LG*[BN] *Nexus*[PL] *4*[PI] *Juicer*[PL] *[Battery Case]*[PC]	Mobile
[WORLDS ONLY][OF] *NEXUS*[PL] *4*[PI] *[BATTERY CASE]*[PC] *(LG-N4-BattCase-black*[PI] *)*	Mobile
[Kenneth Cole][BN] *REACTION*[PL] *men's*[PF] *Cufflinks*[PC]*, Silver*[PF]*, [One Size]*[PF]	Fashion

token	First Char	Char Count		Label
ViewSonic	FRSTCHR_UC	CHARCNT_9		B-BN
3D	FRSTCHR_num	CHARCNT_2		B-PF
HD	FRSTCHR_UC	CHARCNT_2		B-PF
Video	FRSTCHR_UC	CHARCNT_5		B-PF
Processor	FRSTCHR_UC	CHARCNT_9		I-PF
(	FRSTCHR_oth	CHARCNT_1		O
VP3D1	FRSTCHR_UC	CHARCNT_5		B-P
)	FRSTCHR_oth	CHARCNT_1		O
Netgear	FRSTCHR_UC	CHARCNT_7		B-BN
85Mbps	FRSTCHR_num	CHARCNT_6		B-PF
Powerline	FRSTCHR_UC	CHARCNT_9		B-PL
Network	FRSTCHR_UC	CHARCNT_7		B-PC
Adapter	FRSTCHR_UC	CHARCNT_7		I-PC
Kit	FRSTCHR_UC	CHARCNT_3		B-PF
=-	FRSTCHR_oth	CHARCNT_1		O
XETB1001	FRSTCHR_UC	CHARCNT_8		B-P

Training Set	Test Set Category										
training set	All	CE	BK	HG	CM	HO	JW	HB	FS	AU	AE
All	57.7	64.3	41.2	54.6	58.3	48.0	60.3	51.6	55.0	46.7	48.8
CE	47.1	63.3	6.6	46.6	48.7	34.8	35.8	34.6	36.6	42.7	27.2
BK	11.1	4.0	82.0	2.8	4.2	2.6	2.1	3.5	3.8	2.8	37.5
HG	34.9	41.4	7.5	50.0	28.9	32.8	36.1	32.3	32.3	32.9	28.1
CM	41.2	48.9	9.1	39.7	56.5	29.8	32.5	33.1	31.0	35.7	31.5
HO	27.3	30.4	4.1	37.3	22.3	43.5	28.7	29.6	25.7	27.7	18.9
JW	27.1	30.2	7.2	31.8	19.5	23.9	56.2	29.3	31.9	26.1	15.9
HB	27.9	30.9	4.7	35.1	23.3	29.2	27.3	45.8	28.1	24.3	23.6
FS	27.0	32.0	5.8	31.7	22.1	23.1	32.1	27.6	52.3	25.0	27.8
AU	31.1	40.0	4.8	33.9	24.8	26.8	31.4	25.0	26.7	38.2	19.2
AE	13.3	9.0	59.2	6.3	7.5	7.4	4.3	9.7	11.7	4.6	61.3

Training Set	Test Set Category										
training set	All	CE	BK	HG	CM	HO	JW	HB	FS	AU	AE
All	**57.7**	**64.3**	41.2	**54.6**	**58.3**	48.0	**60.3**	**51.6**	55.0	46.7	48.8
CE	47.1	63.3	6.6	46.6	48.7	34.8	35.8	34.6	36.6	42.7	27.2
BK	11.1	4.0	**82.0**	2.8	4.2	2.6	2.1	3.5	3.8	2.8	37.5
HG	34.9	41.4	7.5	50.0	28.9	32.8	36.1	32.3	32.3	32.9	28.1
CM	41.2	48.9	9.1	39.7	56.5	29.8	32.5	33.1	31.0	35.7	31.5
HO	27.3	30.4	4.1	37.3	22.3	43.5	28.7	29.6	25.7	27.7	18.9
JW	27.1	30.2	7.2	31.8	19.5	23.9	56.2	29.3	31.9	26.1	15.9
HB	27.9	30.9	4.7	35.1	23.3	29.2	27.3	45.8	28.1	24.3	23.6
FS	27.0	32.0	5.8	31.7	22.1	23.1	32.1	27.6	52.3	25.0	27.8
AU	31.1	40.0	4.8	33.9	24.8	26.8	31.4	25.0	26.7	38.2	19.2
AE	13.3	9.0	59.2	6.3	7.5	7.4	4.3	9.7	11.7	4.6	61.3
BK&AE	12.9	6.2	65.2	3.6	6.0	4.0	3.0	5.2	4.9	3.3	**65.3**
not BK|AE	55.8	**64.6**	18.1	**54.3**	**59.0**	**49.3**	58.4	**52.1**	**56.3**	**49.2**	34.9

	graphics card	directional antenna	sleeping bag	am radio	messenger bag	gaming mouse	inkjet printer
9/1/2013	6						
9/2/2013	10						
9/3/2013	3						
9/4/2013	7						
9/5/2013	5						
9/6/2013	4						
9/7/2013	6						
9/8/2013	7						
9/9/2013	9			1			
9/10/2013	9					1	
9/11/2013	6						
9/12/2013	8						
9/13/2013	4	2					
9/14/2013	8						
9/15/2013	7	2	2				1
9/16/2013	5	6		1			
9/17/2013	5		3				1
9/18/2013	16	6	6	3	3		2
9/19/2013	28	2	3		3		2
9/20/2013	28	4	5	1	2	1	
9/21/2013	33	10	3	1	1	2	1
9/22/2013	35			1	2	4	
9/23/2013	16	6	6	3	1	3	

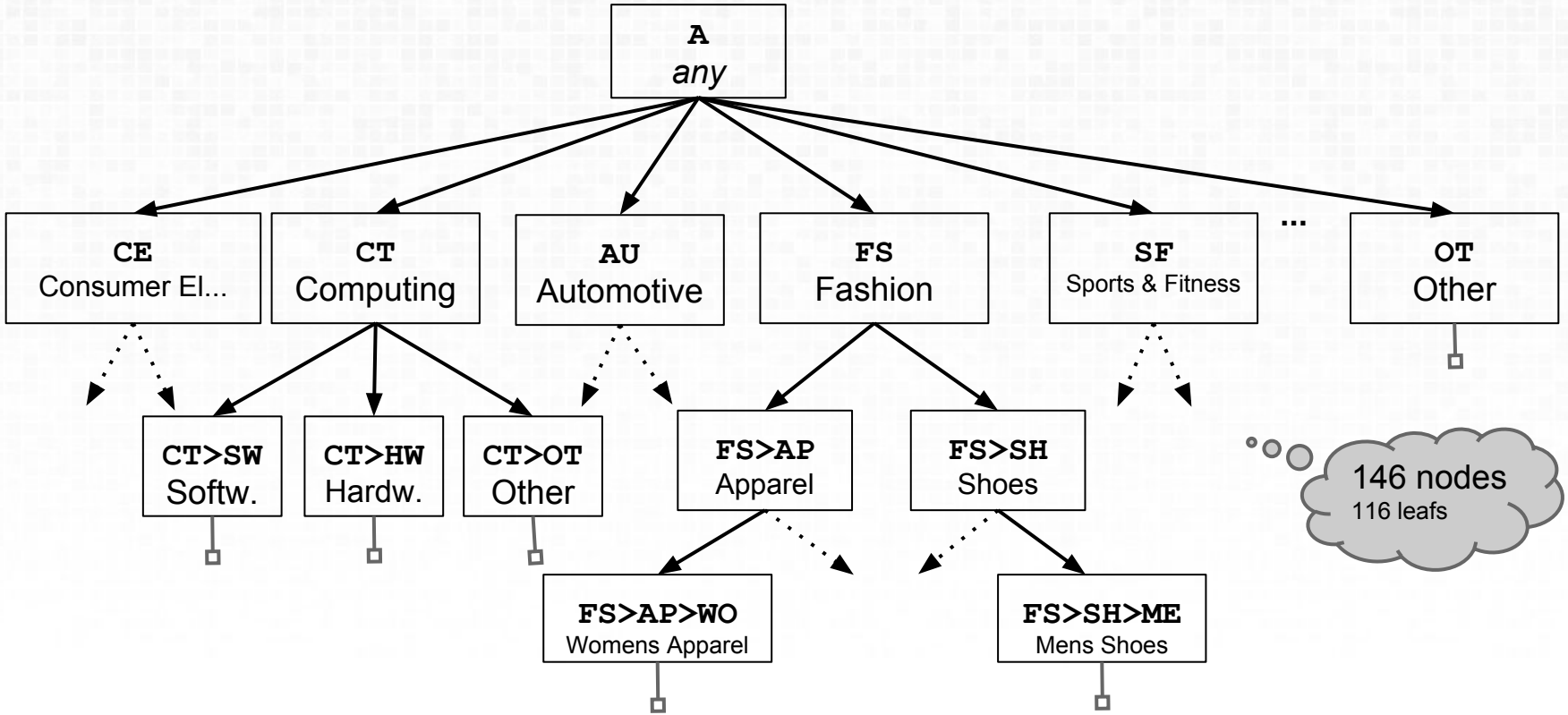
Active Iterative (Re) Training

VigLink

- Add examples with low confidence predictions.
- Preannotate new examples so that the annotator can correct:
  - [token1 token2]^{PF} instead of [token1]^{PF} [token2]^{PF}
  - `PF` instead of PC

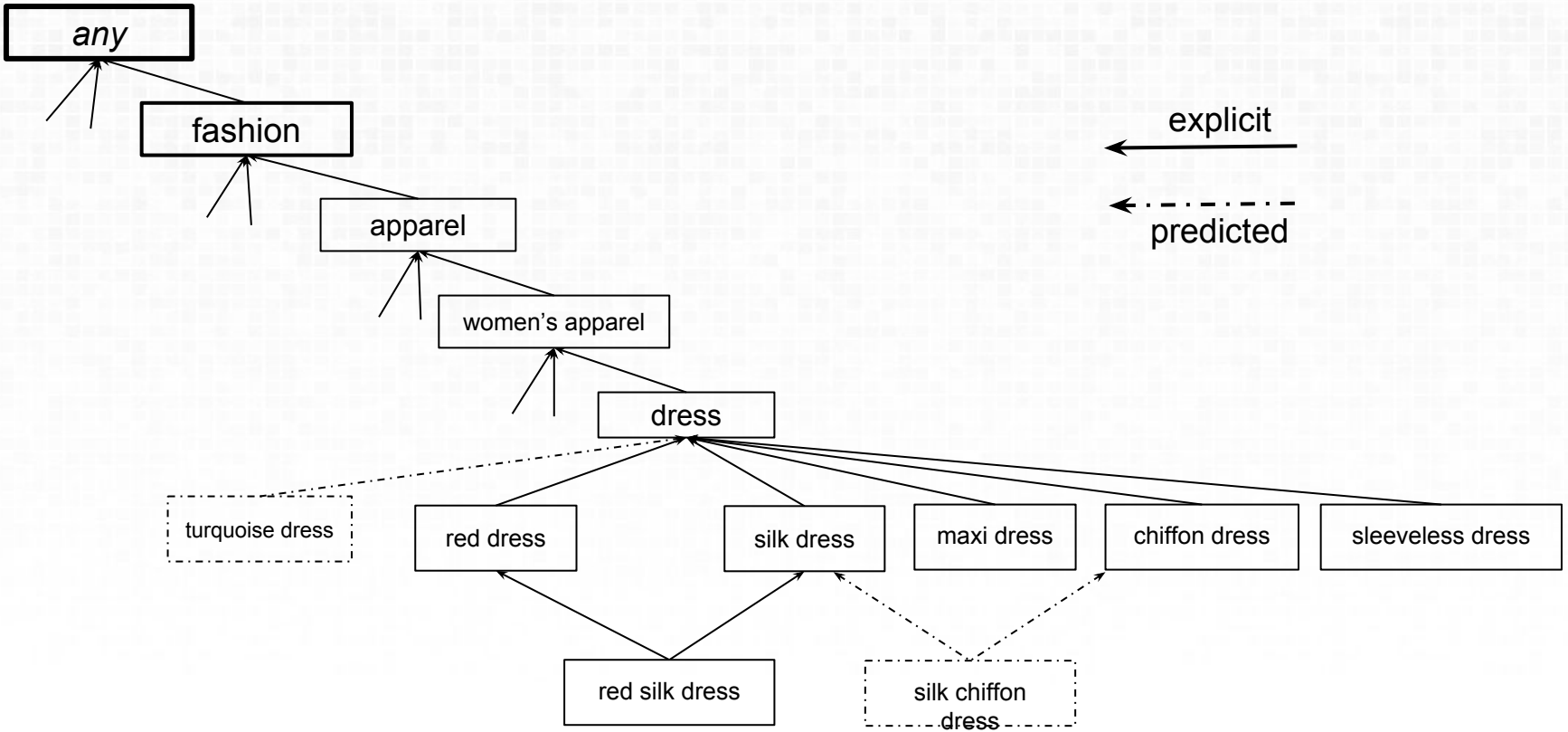
Product Category Hierarchy

VigLink



Product Category Future | Subsumption (isA) Lattice

VigLink



Concept Discovery (word2vec)

VigLink

Cpb-400
Cpb-400b
Cpb-400c
Cpb-403b
Cph-485
Cph-489
Cph-496
Cph-514
Cph-516b
Cpl-536
Cpp-529z
Ct-11
Cx7525
Cxai5198
Cxai5698
…

Pwrc51
Pwrc61
Pwrc62
R-5502-w
R-5650-s
R-5650-w
R5800
R-5800-w
R5800wii
R800
Rb-41
Rb-51
Rb-61
Rc-52
Rc60i
…

Thank You!

VigLink

Interested in working for VigLink?

See open positions at

http://www.viglink.com/about/careers/openings/