OPEN
DATA
SCIENCE
CONFERENCE_

#ODSC

BOSTON 2015

@opendatasci

MACHINE-IN-THE-LOOP FOR KNOWLEDGE DISCOVERY

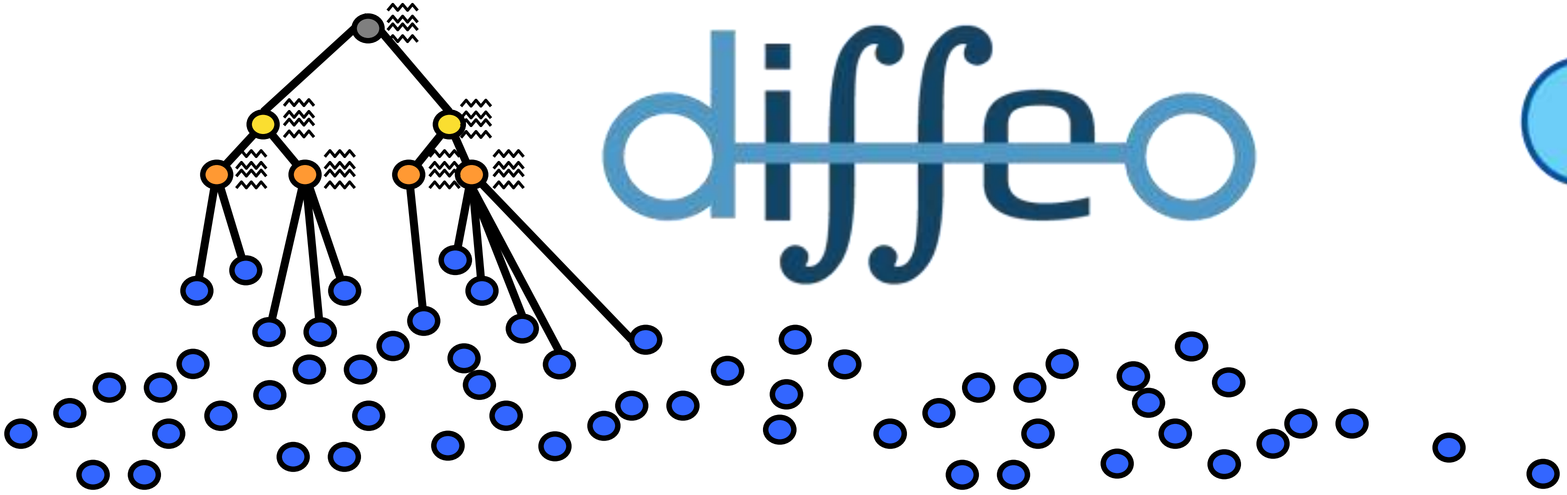
Max Kleiman-Weiner

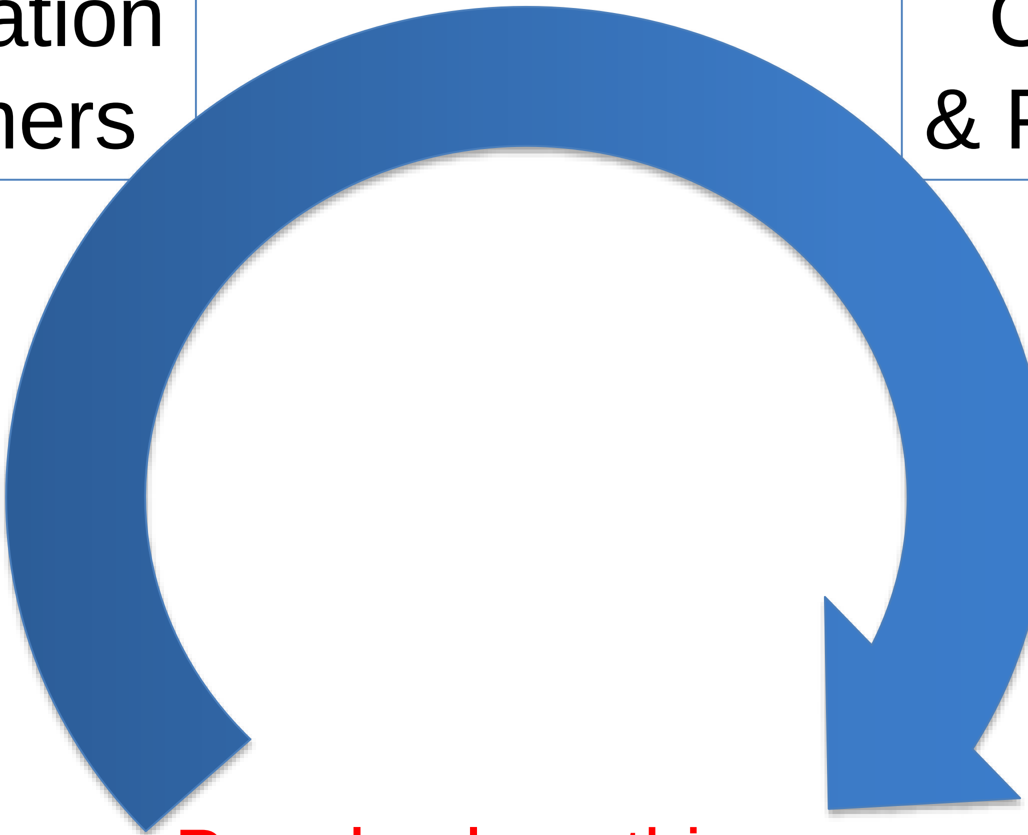
MACHINE-IN-THE-LOOP FOR KNOWLEDGE DISCOVERY

BOSTON 2015
@opendatasci

#ODSC

Max Kleiman-Weiner
Co-founder of Diffeo
Co-organizer of TREC-KBA



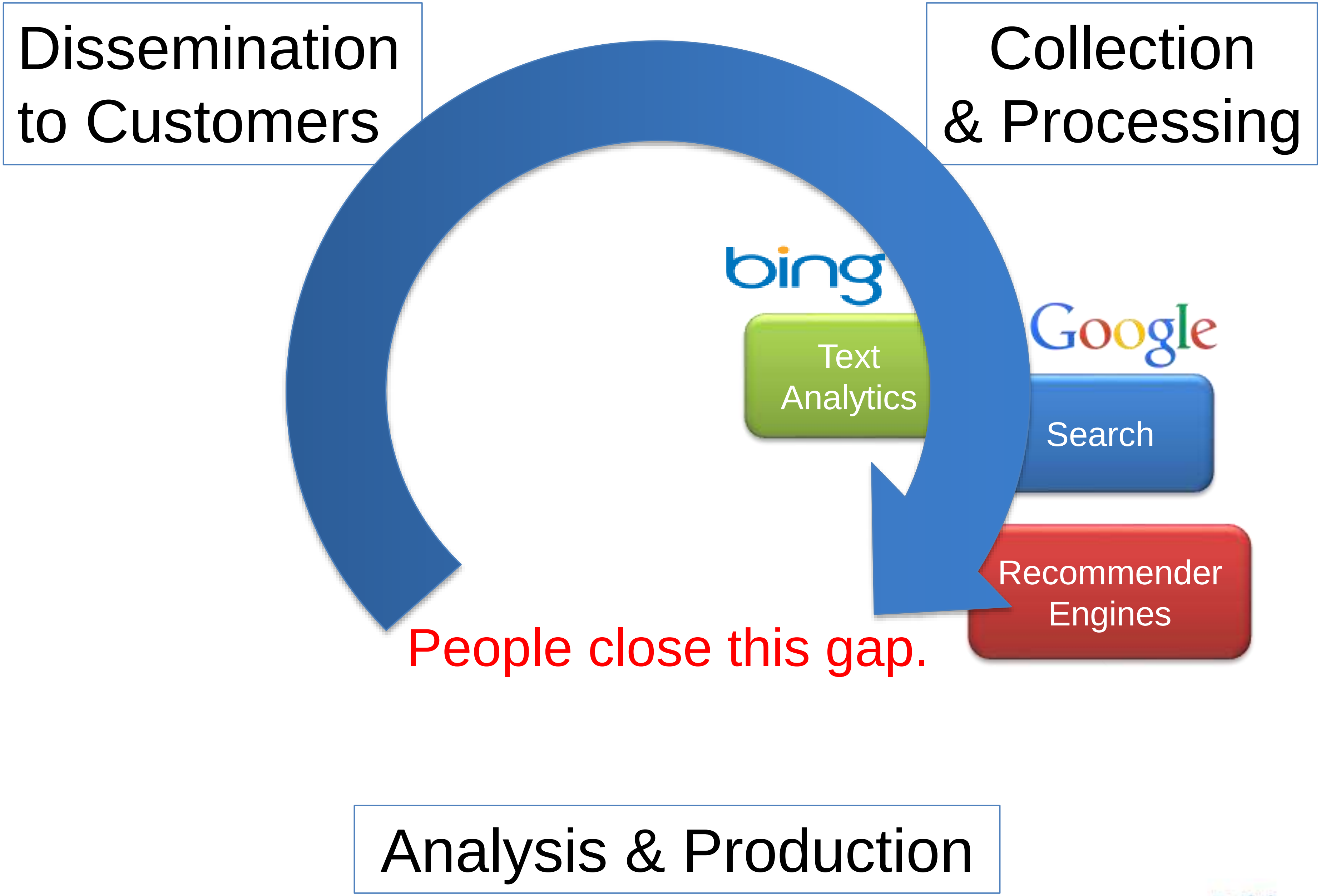


Dissemination to Customers

Collection & Processing

People close this gap.

Analysis & Production



Dissemination to Customers

Collection & Processing

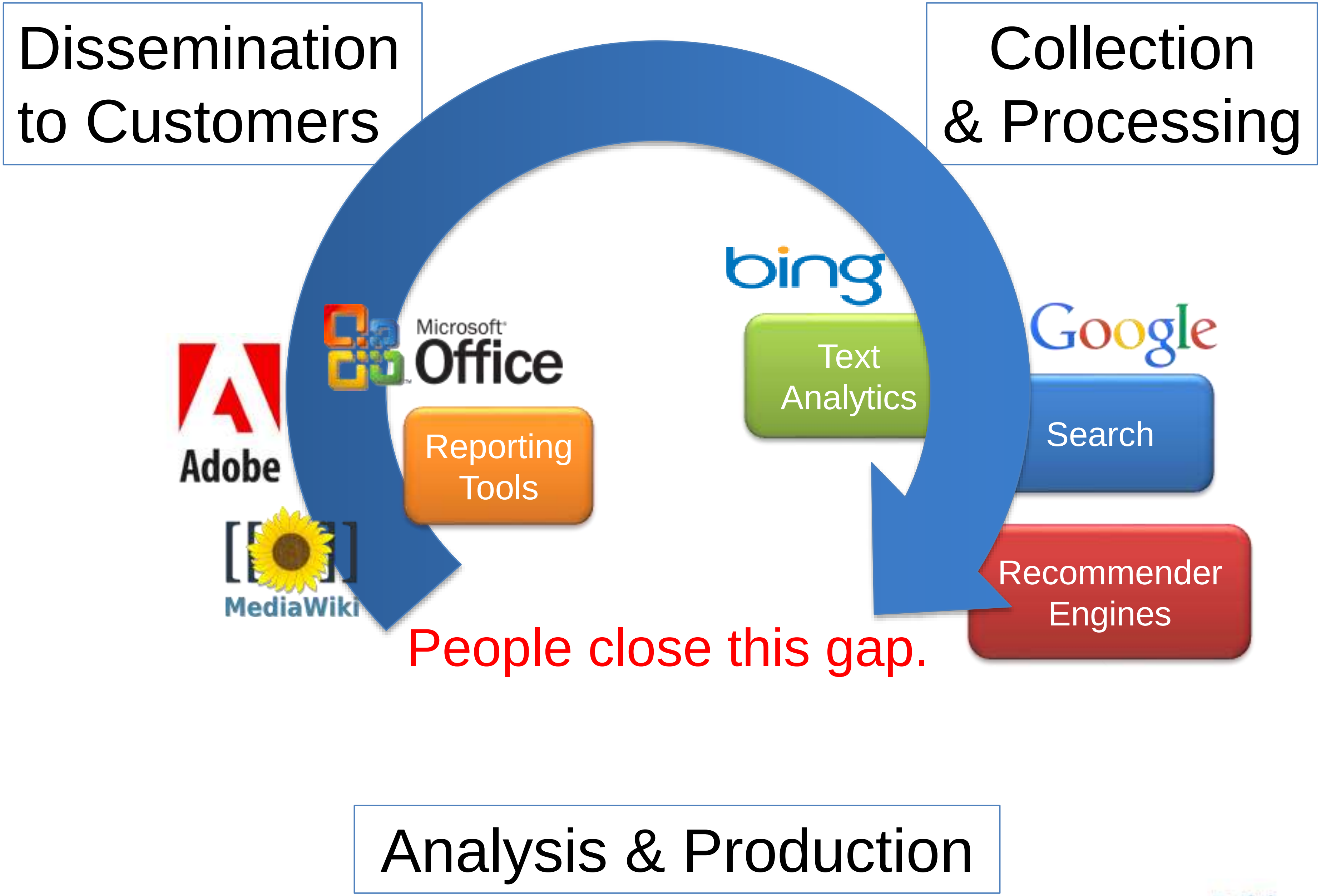
Text Analytics

Search

Recommender Engines

People close this gap.

Analysis & Production



Dissemination to Customers

Collection & Processing

Text Analytics

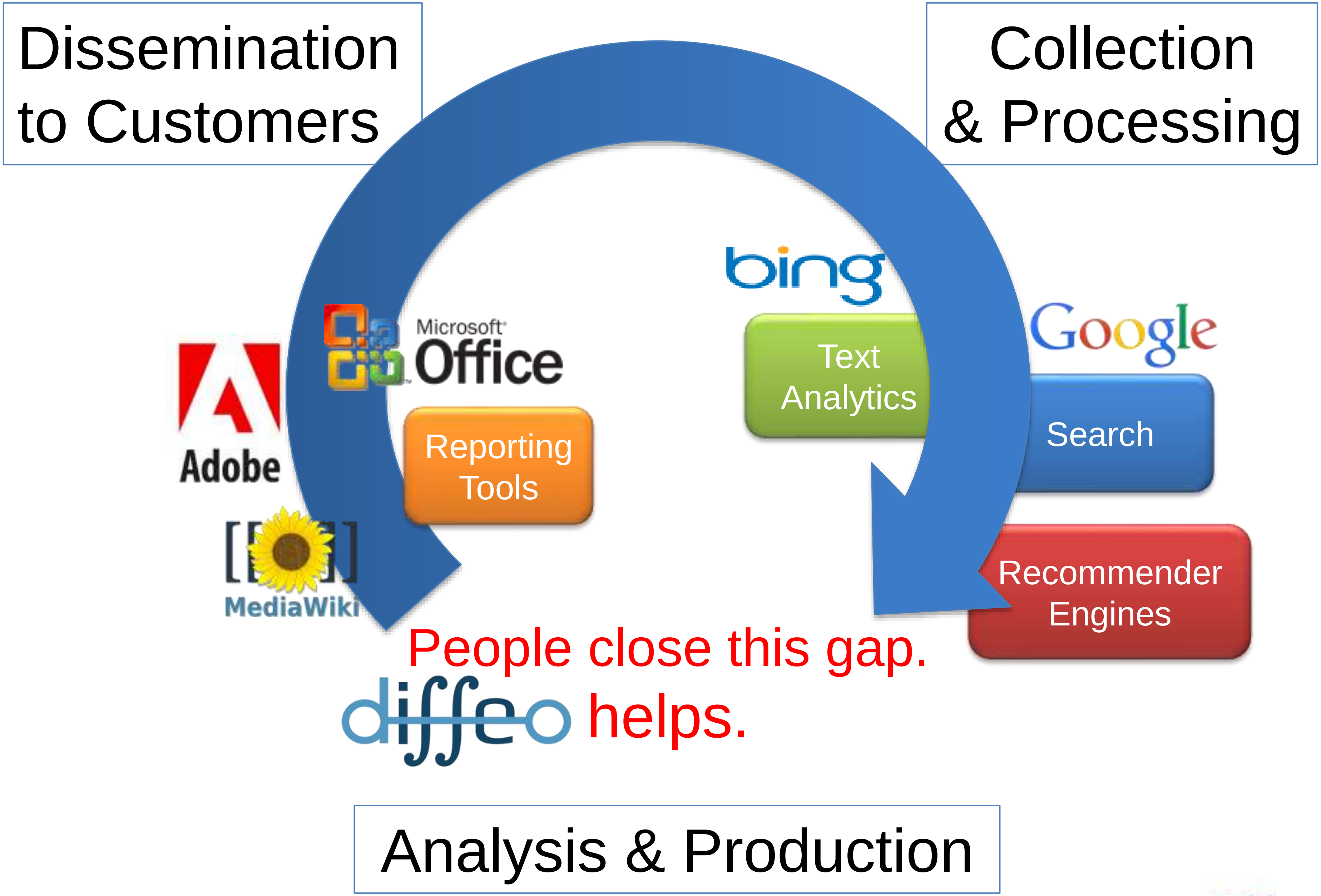
Search

Recommender Engines

Reporting Tools

People close this gap.

Analysis & Production



Dissemination to Customers

Collection & Processing

Text Analytics

Search

Recommender Engines

Reporting Tools

People close this gap.

diffeo helps.

Analysis & Production

From: Google Alerts <googlealerts-noreply@google.com>

Subject: Google Alert - "John R. Frank”

=== Web - 2 new results for **["John R. Frank"]** ===

John R. Frank

SPOKANE, Wash. - **John R. Frank**, **55**, died March 4, 2012, in **Coeur d' Alene, Idaho**. Survivors include: his wife, Miki; daughter, Patricia Frank; ...

<http://www.hutchnews.com/obituaries/Frank--John-CP>

In Memory of **John R Frank**

Biography. **John R. Frank**, **age 55**, passed away at **Sacred Heart Medical Center in Spokane, WA**, on March 4, 2012. **John** was born in **Hutchison, KS**, ...

<http://www.englishfuneralchapel.com/sitemaker/sites/Englis1/obit.cgi?user=583335Frank>

From: Google Alerts <googlealerts-noreply@google.com>

Subject: Google Alert - "John R. Frank”

=== Web - 2 new results for **["John R. Frank"]** ===

John R. Frank

SPOKANE, Wash. - **John R. Frank**, **55**, died March 4, 2012, in **Coeur d' Alene, Idaho**. Survivors include: his wife, Miki; daughter, Patricia Frank; ...

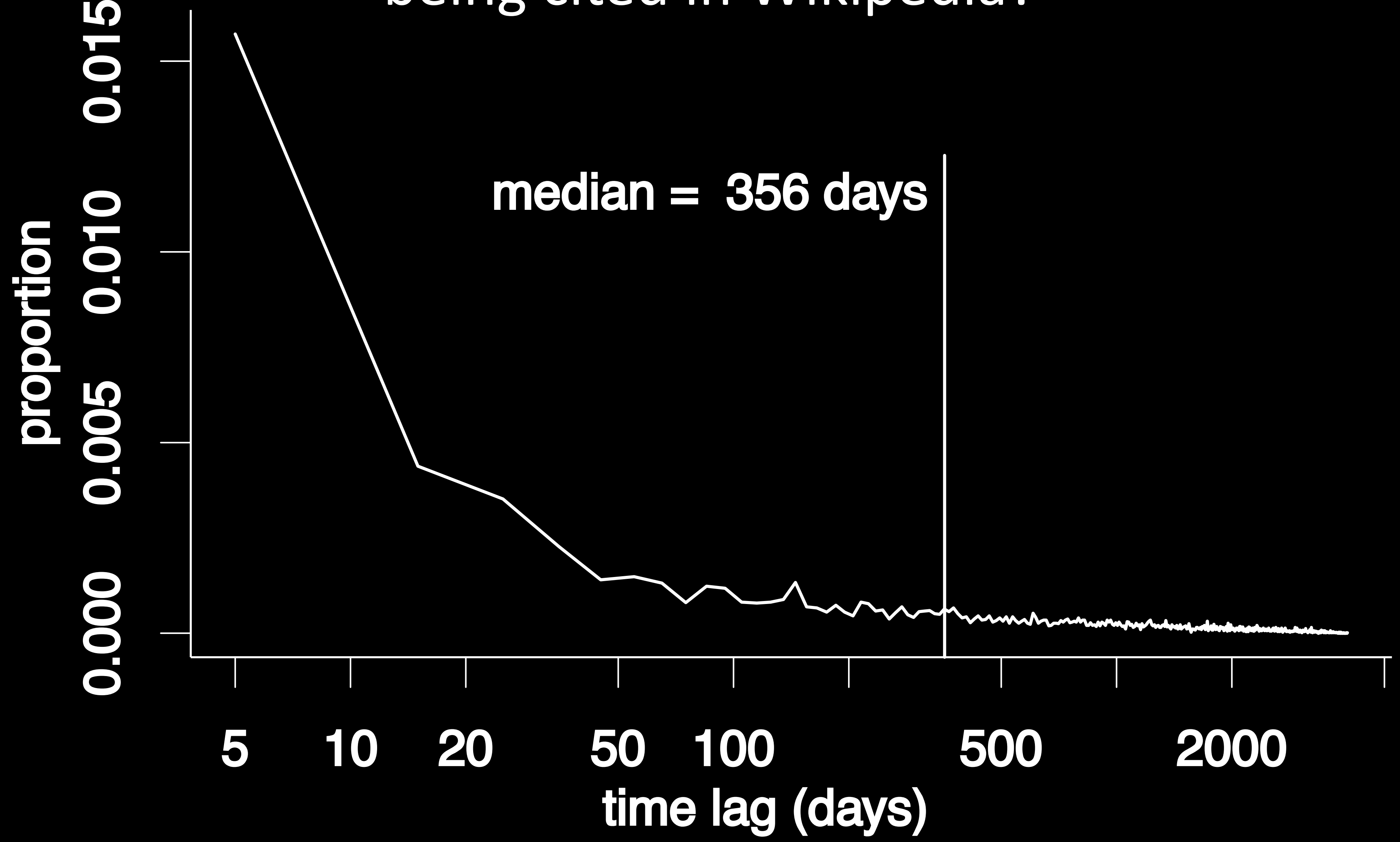
<http://www.hutchnews.com/obituaries/Frank--John-CP>

In Memory of **John R Frank**

Biography. **John R. Frank**, **age 55**, passed away at **Sacred Heart Medical Center in Spokane, WA**, on March 4, 2012. **John** was born in **Hutchison, KS**, ...

<http://www.englishfuneralchapel.com/sitemaker/sites/Englis1/obit.cgi?user=583335Frank>

How many days must a news article wait before being cited in Wikipedia?



Knowledge Base Acceleration?

rate of assimilation << stream size

# editors << # entities << # mentions

(definition of a “large” KB)

TREC KBA 2014

Stream Filtering to Recommend Citations and Fill Slots

Entities in Wikipedia or another Knowledge Base

Takashi Murakami (村上 隆 *Murakami Takashi*, born in Tokyo) is an internationally prolific contemporary Japanese artist. He works in fine arts media—such as painting and sculpture—as well as what is conventionally considered commercial media—fashion, merchandise, and animation— a is known for blurring the line between hi and low art. He coined the term sup

Automatically recommend new edits

Your KBA System

1) Initialize with **target entities**
   - Start with citations before dateX (20%)

1) Iterate over stream of text items
   - Predict citations after dateX (80%)
   - Citation = “vital”, i.e. changes profile

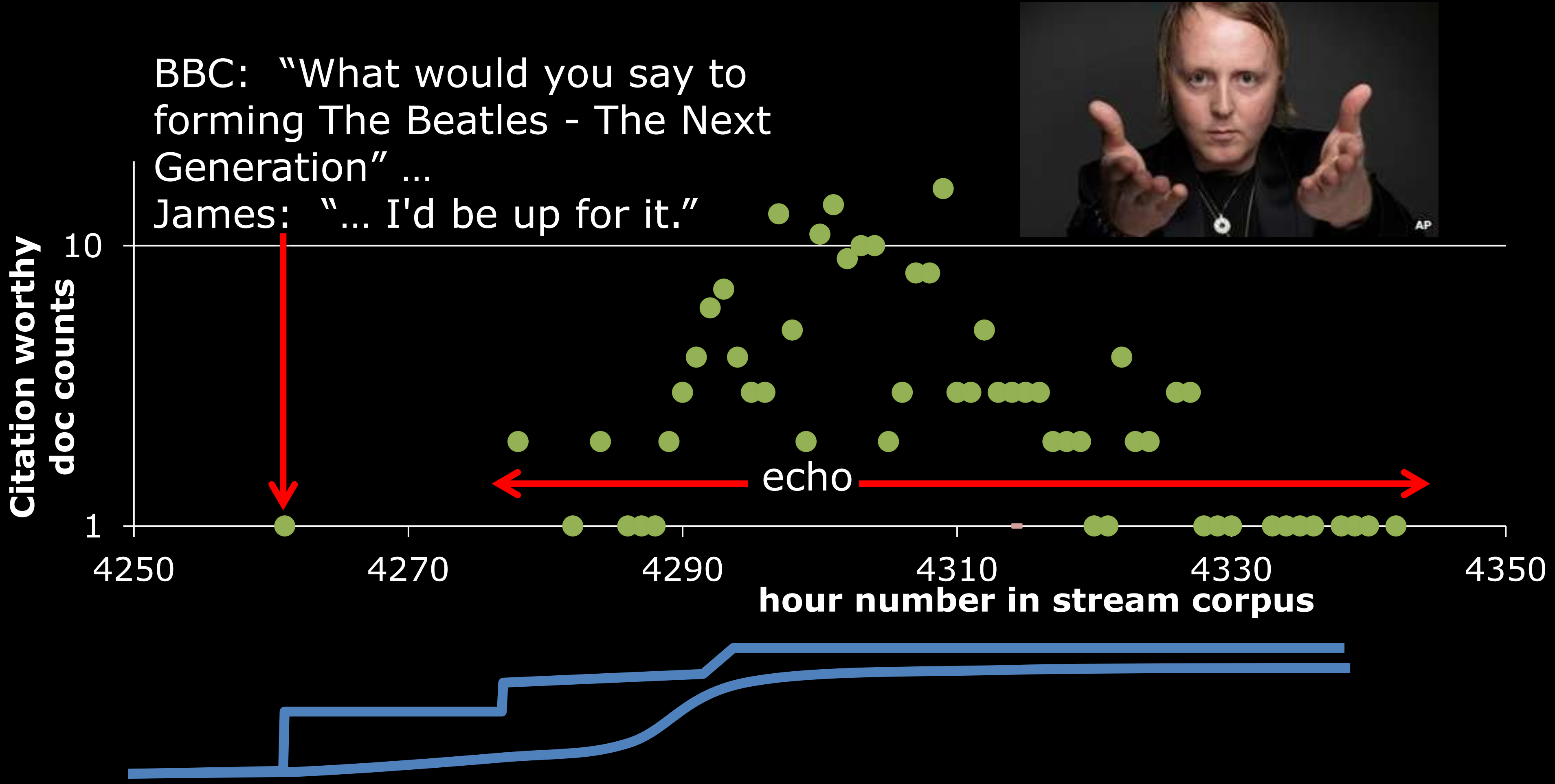
2) Identify infobox slot values

Content Stream

- 1.2bn texts, 50% English
- >500M tagged by Serif
- 13,663 hours of data (19 months)
- 10^5 docs-per-hour
- News, blogs, forums, preprints, and more

SSF example:

“FounderOf” slot on James McCartney



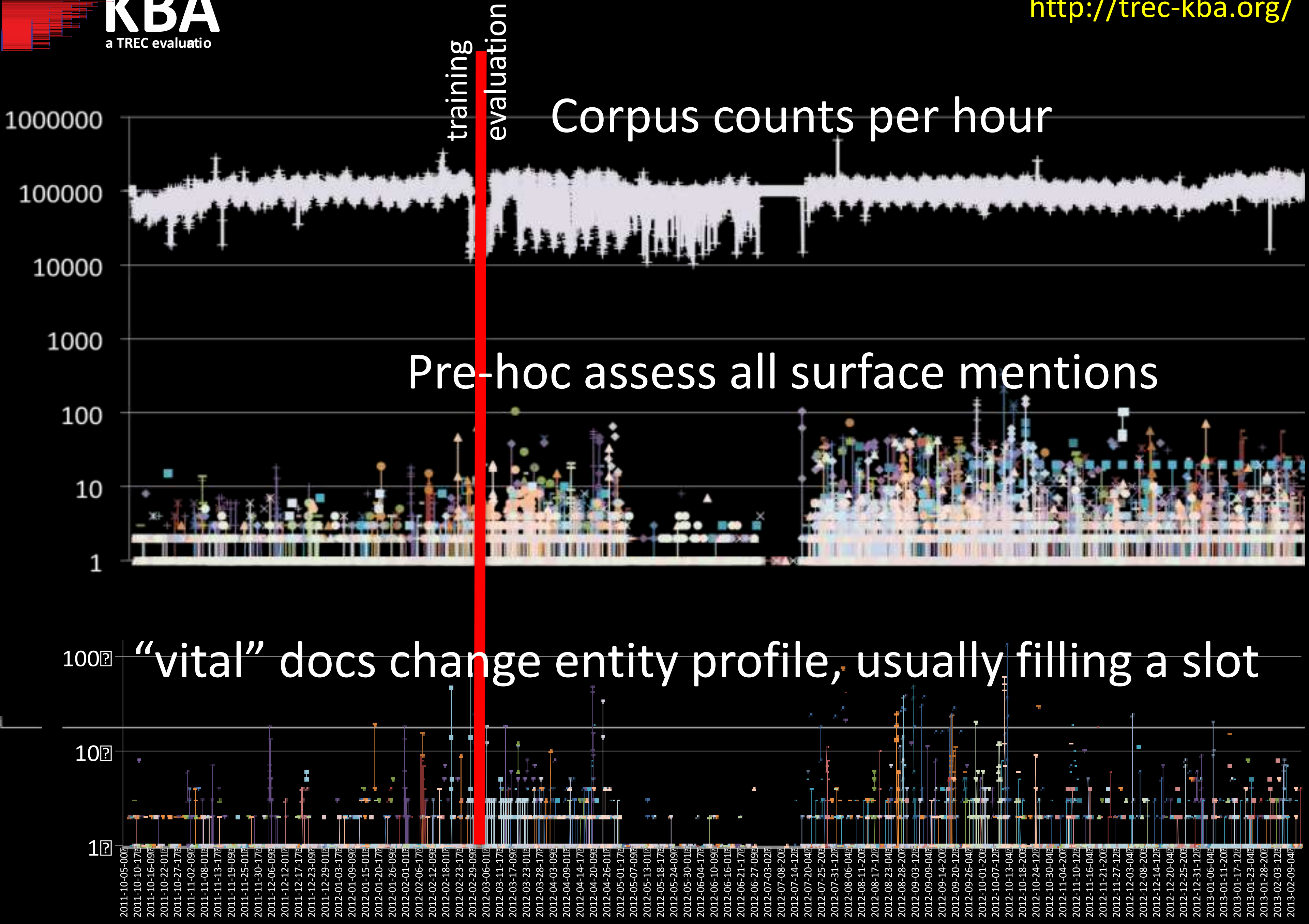
BBC: “What would you say to forming The Beatles - The Next Generation” …
James: “… I'd be up for it.”

Entity profiles evolve over time!

Exploring these issues in new track: **TREC Dynamic Domain**

KBA
a TREC evaluatio

http://trec-kba.org/



Corpus counts per hour

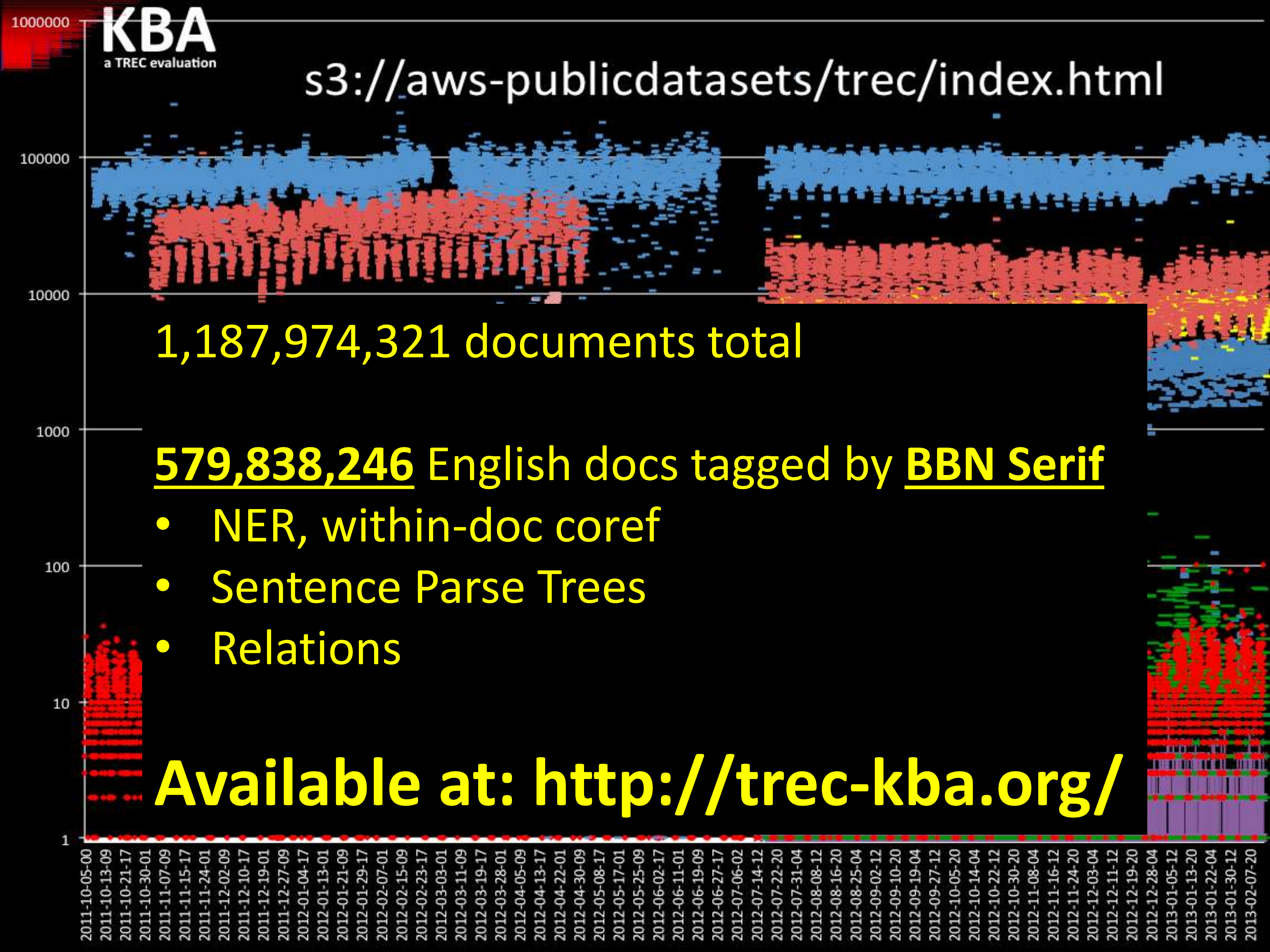
Pre-hoc assess all surface mentions

“vital” docs change entity profile, usually filling a slot

KBA

a TREC evaluation

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



1,187,974,321 documents total

579,838,246 English docs tagged by **BBN Serif**

- NER, within-doc coref
- Sentence Parse Trees
- Relations

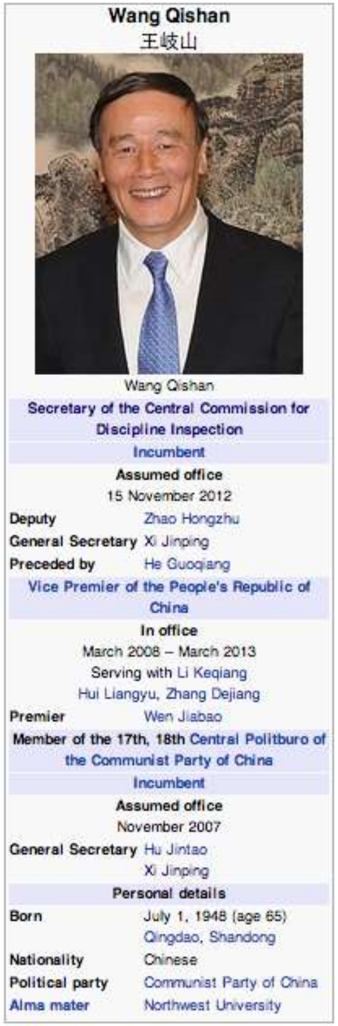
Available at: http://trec-kba.org/

What entity resolvers do easily is boring.

Wang Qishan, China's top graft buster and a member of the Politburo Standing Committee, making him one of the seven most powerful people in China

Wang Qishan, male, Han nationality, is a native of Tianzhen County, Shanxi province. Director, Rural Development Research Center of the State Council, Communication Office. **He** served as a member of the CPC Beijing Municipal Committee in 2003, and he was appointed mayor of Beijing in the same year.

Wang, 64, is member of the Political Bureau of the 17th CPC Central Committee and vice premier. **He** was also elected into the new Central Commission for Discipline Inspection.



Wang Qishan
王岐山

Wang Qishan

Secretary of the Central Commission for Discipline Inspection

Incumbent

Assumed office
15 November 2012

Deputy	Zhao Hongzhu
General Secretary	Xi Jinping
Preceded by	He Guoqiang

Vice Premier of the People's Republic of China

In office
March 2008 – March 2013
Serving with Li Keqiang
Hui Liangyu, Zhang Dejiang

Premier	Wen Jiabao

Member of the 17th, 18th Central Politburo of the Communist Party of China

Incumbent

Assumed office
November 2007

General Secretary	Hu Jintao Xi Jinping

Personal details

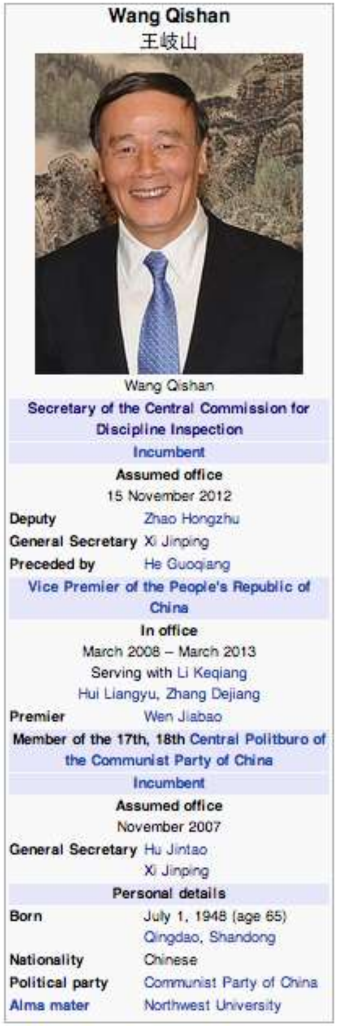
Born	July 1, 1948 (age 65) Qingdao, Shandong
Nationality	Chinese
Political party	Communist Party of China
Alma mater	Northwest University

What entity resolvers *should* do is exciting.

Wang also announced that Trinidad and Tobago would receive a 40M yuan grant (approximately TT$40M) from the Chinese government to fund projects mutually agreed upon by both governments.

Do you realize that Harry Reid and his son Rory have a deal for that land with Chinese Vice Premier **Wang Qishan** for a 5 billion solar plant. Its allover social media…

The U.S. Chamber of Commerce co-hosted a dinner last night at the JW Marriot in Washington to honor Chinese Vice Premier **Wang Qishan**, who is in the United States to co-chair the 23rd China - U.S. Joint Commission on Commerce and Trade



Wang Qishan
王岐山

Wang Qishan

Secretary of the Central Commission for Discipline Inspection
Incumbent
Assumed office
15 November 2012

Deputy	Zhao Hongzhu
General Secretary	Xi Jinping
Preceded by	He Guoqiang

Vice Premier of the People's Republic of China
In office
March 2008 – March 2013
Serving with Li Keqiang
Hui Liangyu, Zhang Dejiang

Premier	Wen Jiabao

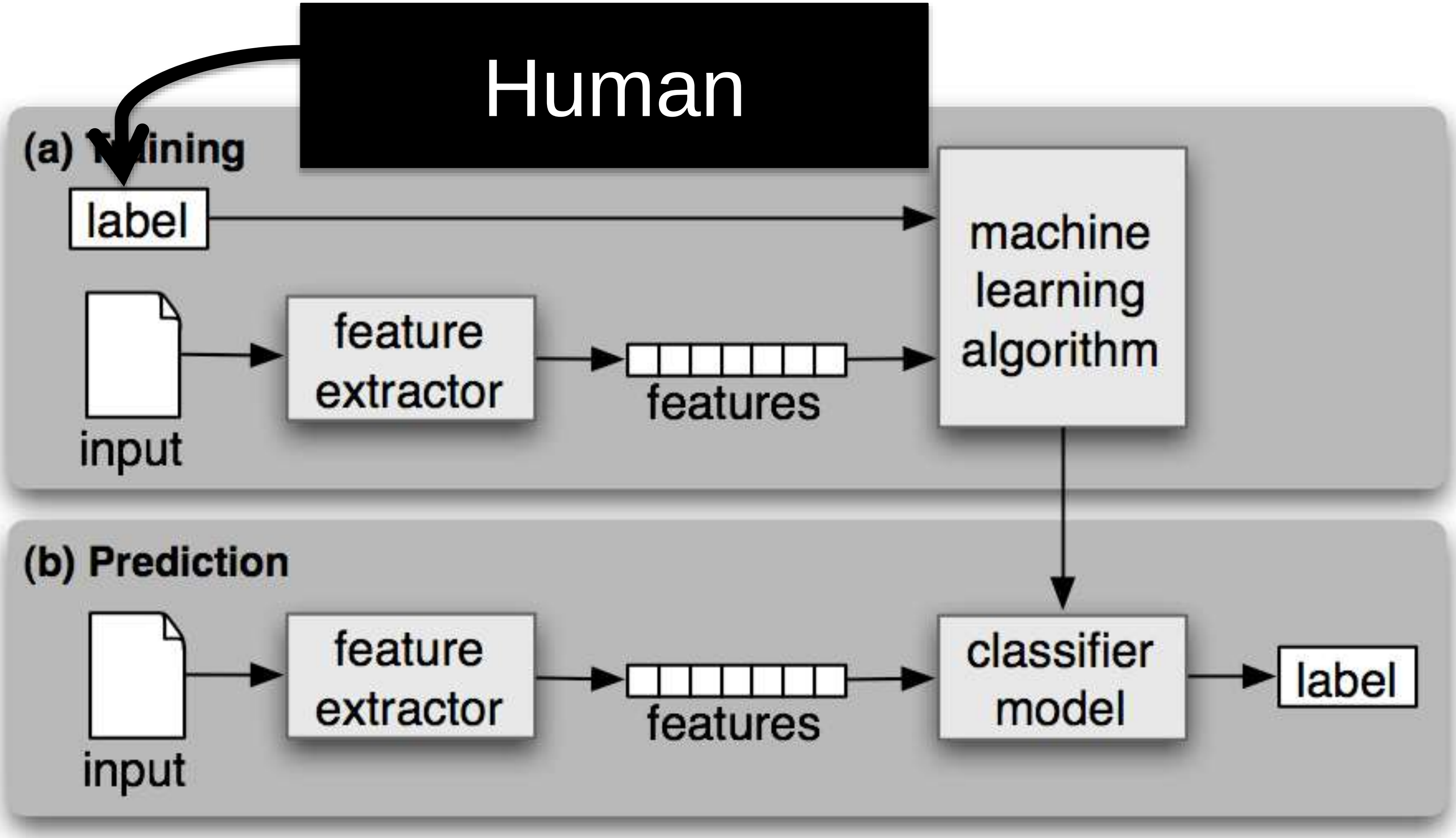
Member of the 17th, 18th Central Politburo of the Communist Party of China
Incumbent
Assumed office
November 2007

General Secretary	Hu Jintao Xi Jinping

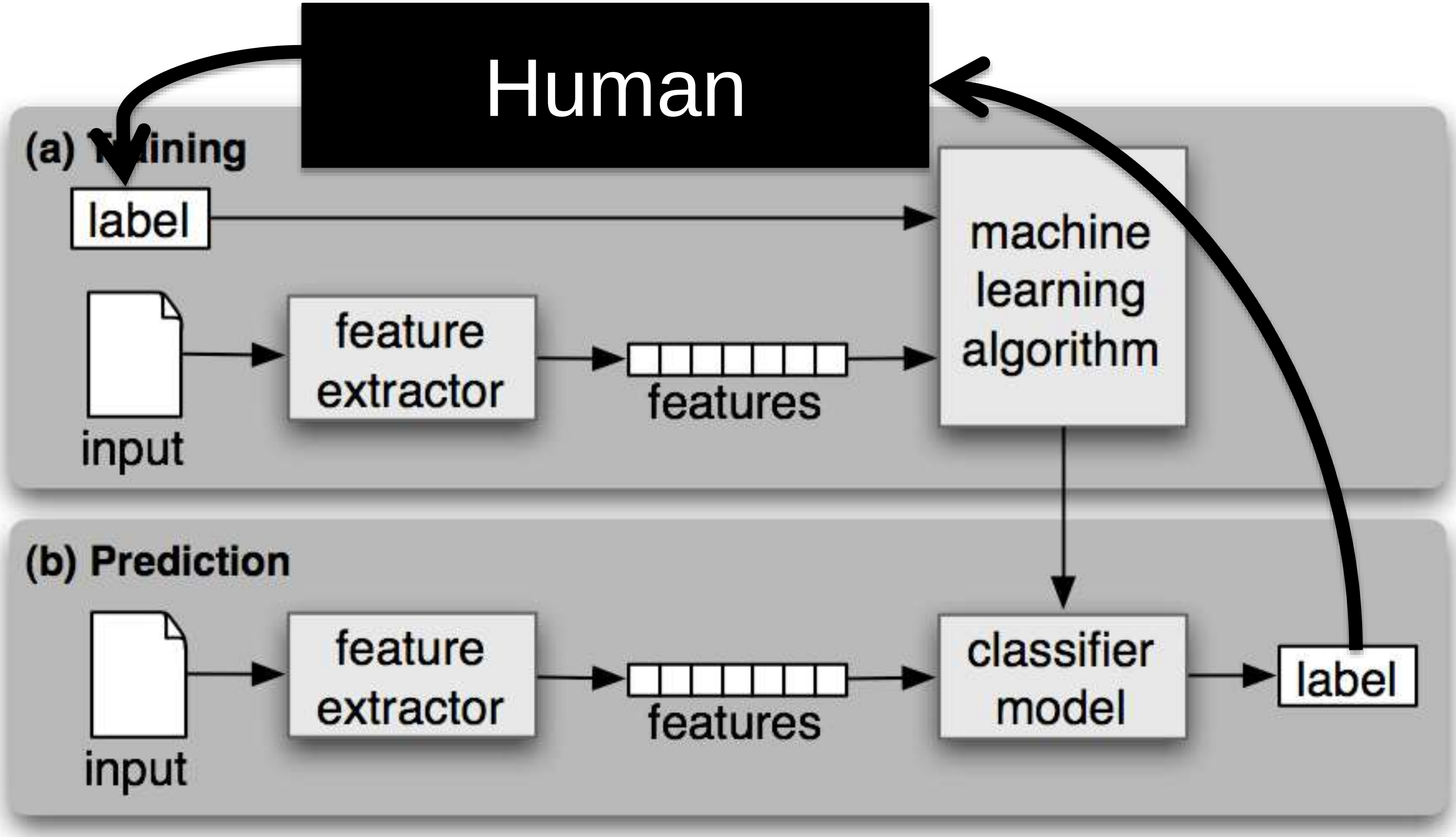
Personal details

Born	July 1, 1948 (age 65) Qingdao, Shandong
Nationality	Chinese
Political party	Communist Party of China
Alma mater	Northwest University

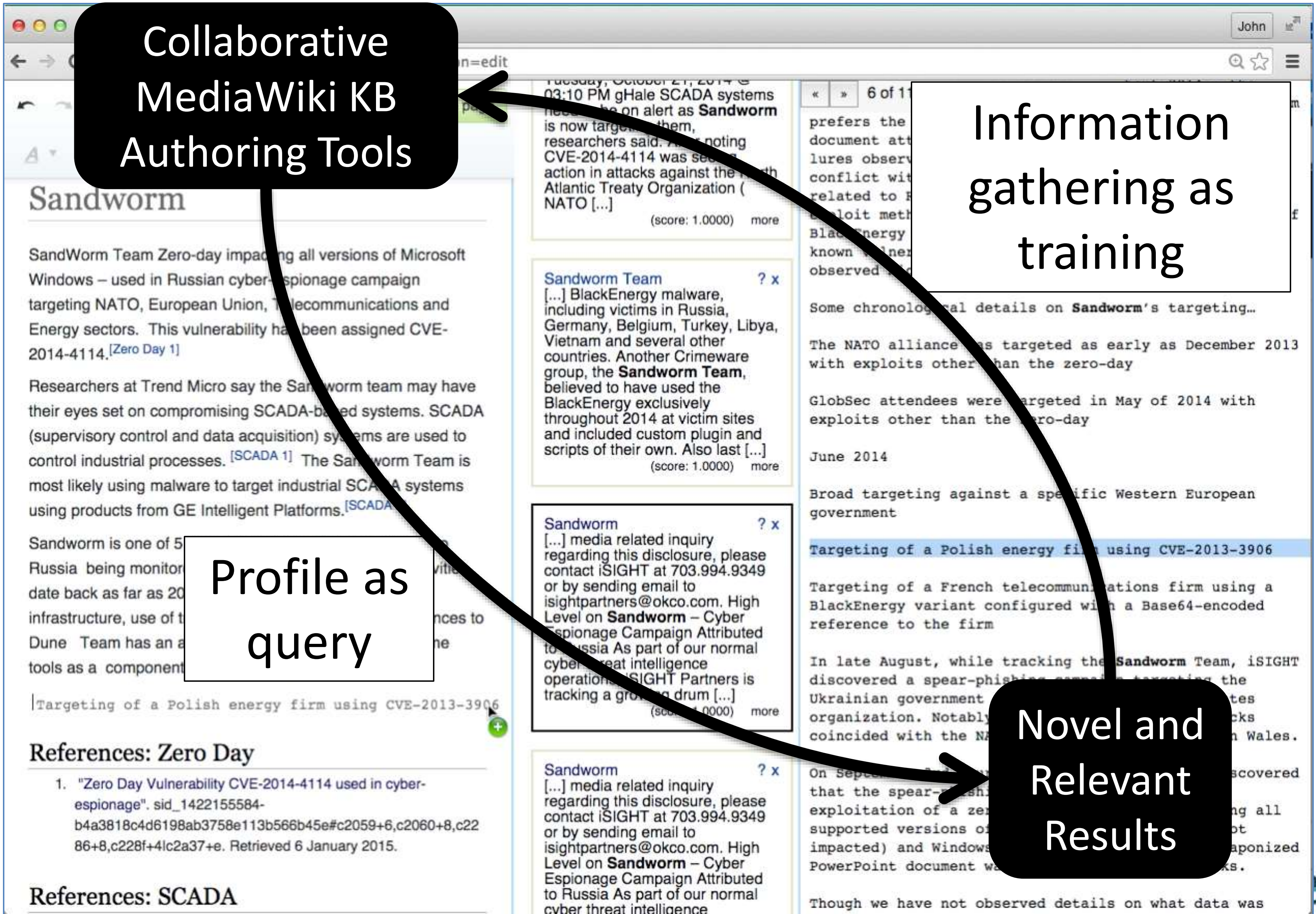
Traditional Machine Learning



Human-in-the-loop (active learning)



Machine-in-the-Loop (active-ranking)



Collaborative MediaWiki KB Authoring Tools

Information gathering as training

Profile as query

Novel and Relevant Results

Sandworm

SandWorm Team Zero-day impacting all versions of Microsoft Windows – used in Russian cyber-espionage campaign targeting NATO, European Union, Telecommunications and Energy sectors. This vulnerability has been assigned CVE-2014-4114.[Zero Day 1]

Researchers at Trend Micro say the Sandworm team may have their eyes set on compromising SCADA-based systems. SCADA (supervisory control and data acquisition) systems are used to control industrial processes. [SCADA 1] The Sandworm Team is most likely using malware to target industrial SCADA systems using products from GE Intelligent Platforms.[SCADA

Sandworm is one of 5
Russia being monitor … vit
date back as far as 20
infrastructure, use of t … nces to
Dune Team has an a … he
tools as a component

Targeting of a Polish energy firm using CVE-2013-3906

References: Zero Day

1. "Zero Day Vulnerability CVE-2014-4114 used in cyber-espionage". sid_1422155584-b4a3818c4d6198ab3758e113b566b45e#c2059+6,c2060+8,c2286+8,c228f+4lc2a37+e. Retrieved 6 January 2015.

References: SCADA

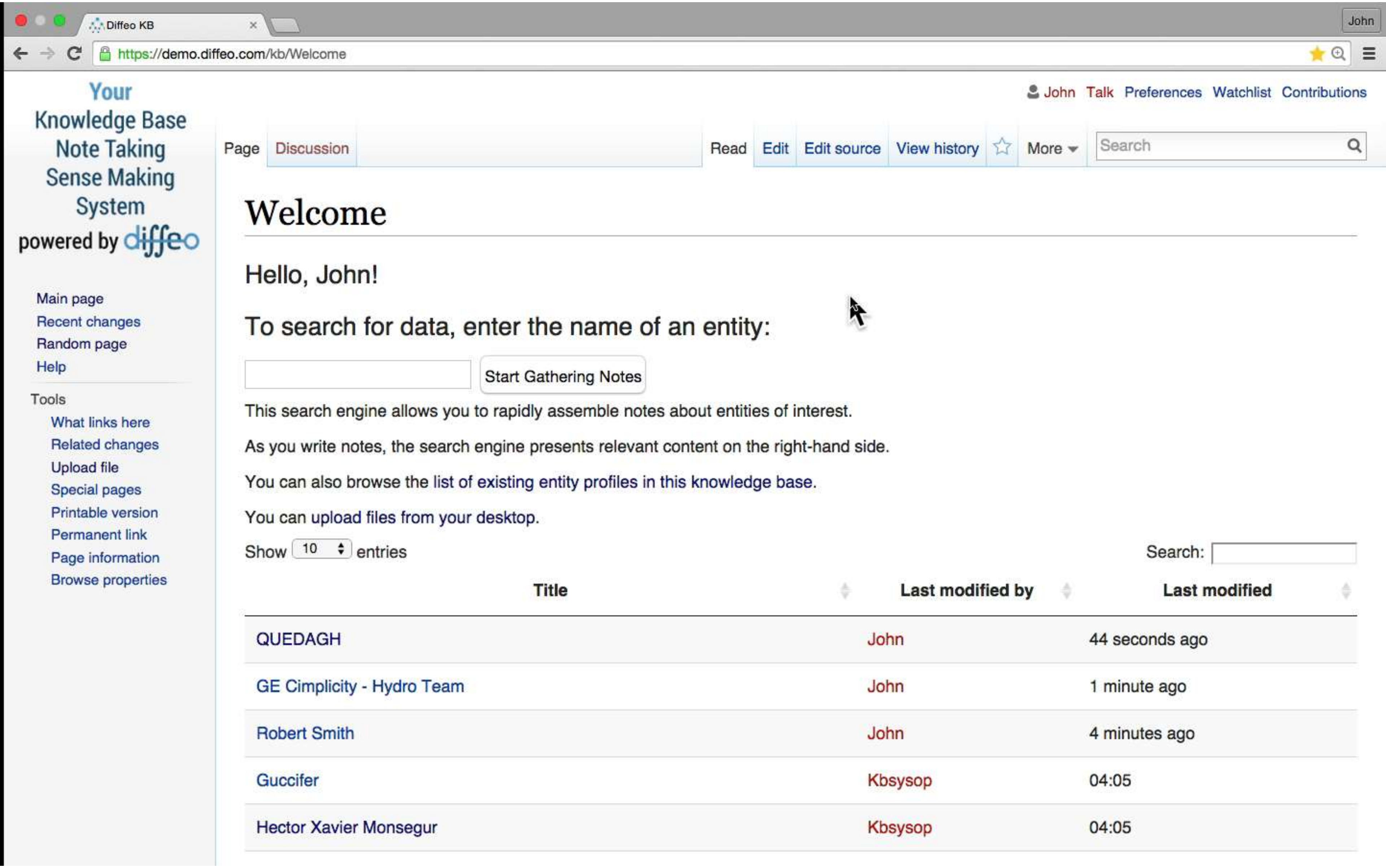
Tuesday, October 21, 2014 @ 03:10 PM gHale SCADA systems need to be on alert as **Sandworm** is now targeting them, researchers said. After noting CVE-2014-4114 was seeing action in attacks against the North Atlantic Treaty Organization (NATO [...]
(score: 1.0000) more

Sandworm Team ? x
[...] BlackEnergy malware, including victims in Russia, Germany, Belgium, Turkey, Libya, Vietnam and several other countries. Another Crimeware group, the **Sandworm Team**, believed to have used the BlackEnergy exclusively throughout 2014 at victim sites and included custom plugin and scripts of their own. Also last [...]
(score: 1.0000) more

Sandworm ? x
[...] media related inquiry regarding this disclosure, please contact iSIGHT at 703.994.9349 or by sending email to isightpartners@okco.com. High Level on **Sandworm** – Cyber Espionage Campaign Attributed to Russia As part of our normal cyber threat intelligence operations iSIGHT Partners is tracking a growing drum [...]
(score: 1.0000) more

Sandworm ? x
[...] media related inquiry regarding this disclosure, please contact iSIGHT at 703.994.9349 or by sending email to isightpartners@okco.com. High Level on **Sandworm** – Cyber Espionage Campaign Attributed to Russia As part of our normal cyber threat intelligence

```
prefers the
document att
lures observ
conflict wit
related to R
exploit meth                                                f
BlackEnergy
known vulner
observed

Some chronological details on Sandworm's targeting…

The NATO alliance was targeted as early as December 2013
with exploits other than the zero-day

GlobSec attendees were targeted in May of 2014 with
exploits other than the zero-day

June 2014

Broad targeting against a specific Western European
government

Targeting of a Polish energy firm using CVE-2013-3906

Targeting of a French telecommunications firm using a
BlackEnergy variant configured with a Base64-encoded
reference to the firm

In late August, while tracking the Sandworm Team, iSIGHT
discovered a spear-phishing campaign targeting the
Ukrainian government                                 tes
organization. Notably                                cks
coincided with the NA                                n Wales.

On Sep                                            scovered
that the spear-phishi
exploitation of a zer                              ng all
supported versions o                               ot
impacted) and Windows                              aponized
PowerPoint document wa                             ks.

Though we have not observed details on what data was
```



Welcome

Hello, John!

To search for data, enter the name of an entity:

This search engine allows you to rapidly assemble notes about entities of interest.

As you write notes, the search engine presents relevant content on the right-hand side.

You can also browse the list of existing entity profiles in this knowledge base.

You can upload files from your desktop.

Show 10 entries

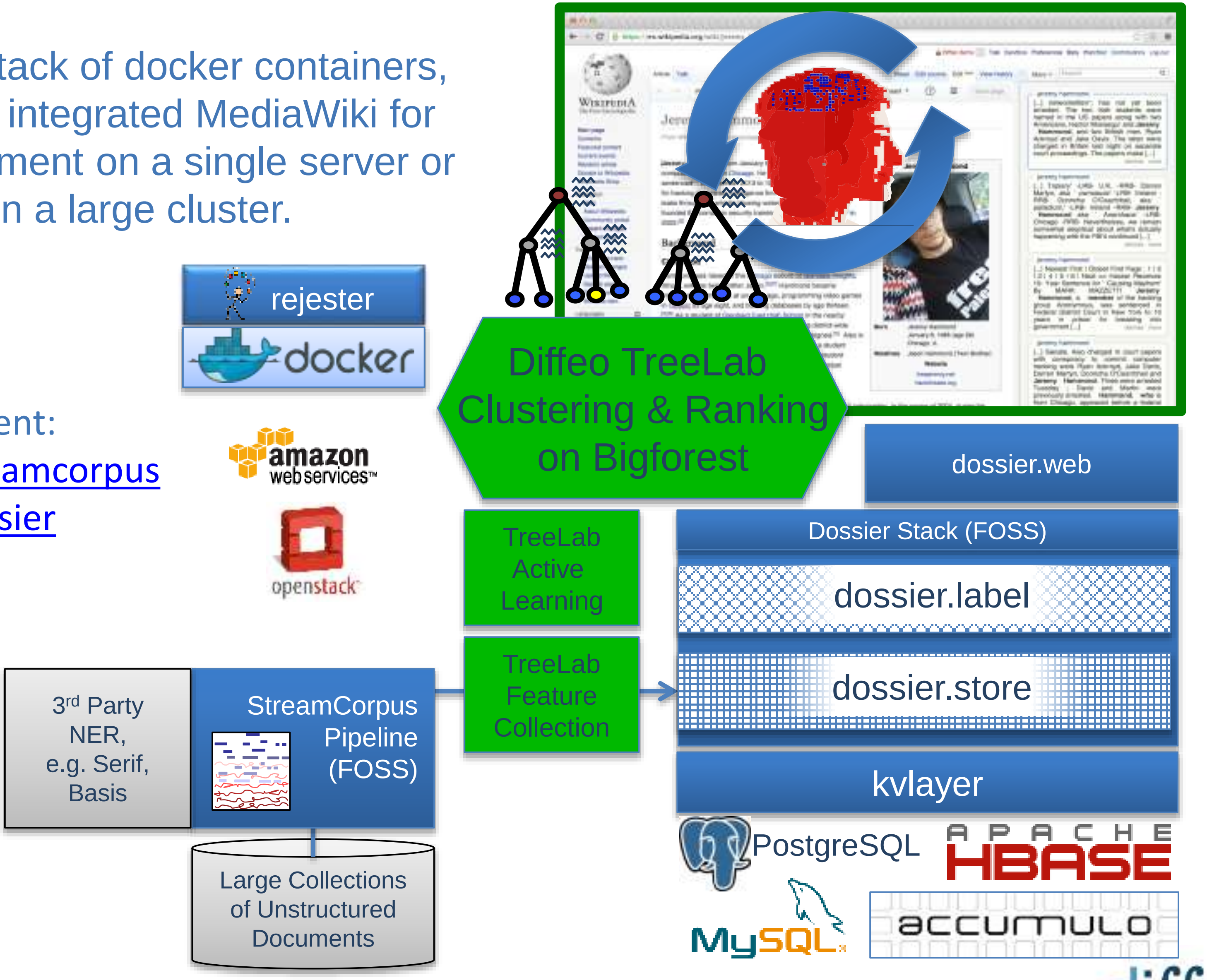
Search:

Title	Last modified by	Last modified
QUEDAGH	John	44 seconds ago
GE Cimplicity - Hydro Team	John	1 minute ago
Robert Smith	John	4 minutes ago
Guccifer	Kbsysop	04:05
Hector Xavier Monsegur	Kbsysop	04:05

Diffeo: Built for Clouds

Diffeo ships a stack of docker containers, including a fully integrated MediaWiki for turn-key deployment on a single server or customization on a large cluster.

FOSS development:
github.com/streamcorpus
github.com/dossier



rejester

Diffeo TreeLab Clustering & Ranking on Bigforest

dossier.web

TreeLab Active Learning

TreeLab Feature Collection

Dossier Stack (FOSS)

dossier.label

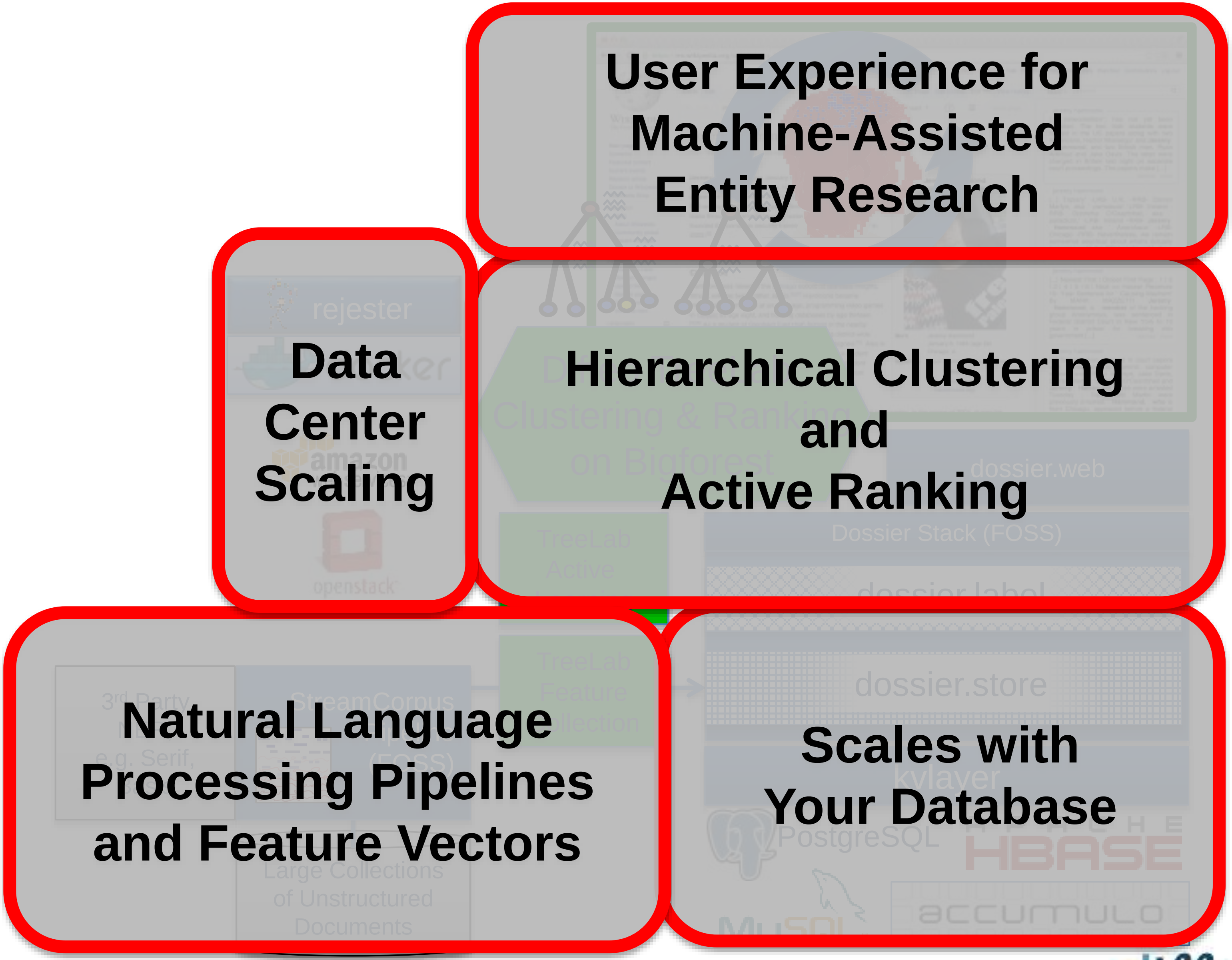
dossier.store

kvlayer

3rd Party NER, e.g. Serif, Basis

StreamCorpus Pipeline (FOSS)

Large Collections of Unstructured Documents



User Experience for Machine-Assisted Entity Research

Data Center Scaling

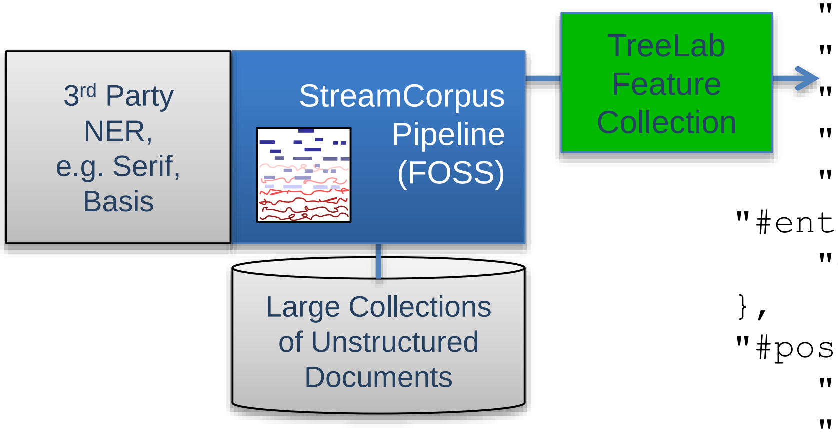
Hierarchical Clustering and Active Ranking

Natural Language Processing Pipelines and Feature Vectors

Scales with Your Database

Natural Language

The **Syrian Electronic Army** (SEA) is a group of computer hackers who support the government of Syrian President Bashar al-Assad. Using spamming,[2] defacement, malware (including the Blackworm tool),[3] phishing, and denial of service attacks, it mainly targets political opposition groups and western websites including news organizations and human rights groups. The Syrian Electronic Army is the first public, virtual army in the Arab world to openly launch cyber attacks on its opponents.[4]


3rd Party NER, e.g. Serif, Basis
StreamCorpus Pipeline (FOSS)
TreeLab Feature Collection
Large Collections of Unstructured Documents

Feature Vectors

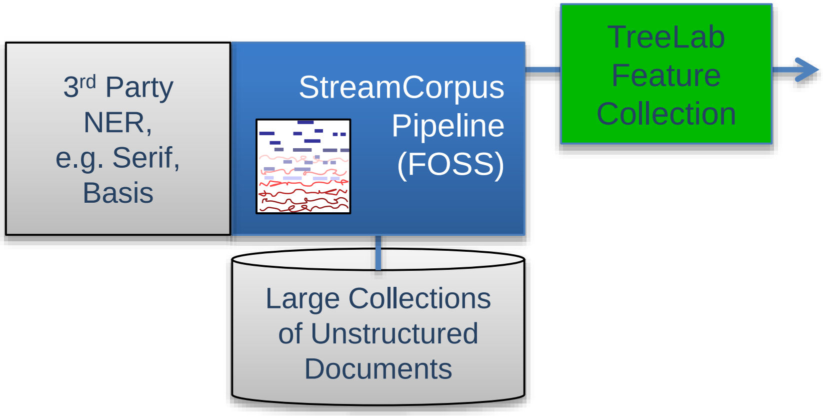
```
"#both_bol_JJ_0": {
     "24 minute": 1,
     "american": 5,
     "linux": 2,
     "pro government": 1,

 "#both_co_LOC_3": {
     "arab": 2,
     "east": 4,
     "iran": 5,
     "kingdom": 2,
     "lebanon": 2,
     "london": 2
"#both_co_MAGIC_VALUE_0": {
     "pagewanted all r 0accessdate22": 1,
     "q cache 9qiv27 ymm8j": 1
 },
 "#both_co_NATIONALITY_2": {
     "american": 6,
     "indians": 2,
     "iranian": 5,
     "saudi": 2,
     "syrian": 93
 "#entity_type": {
     "ORG": 6
 },
 "#post_bow_RB_2": {
     "allegedly": 4,
     "falsely": 1,
     "openly": 2,
```

Feature Vectors

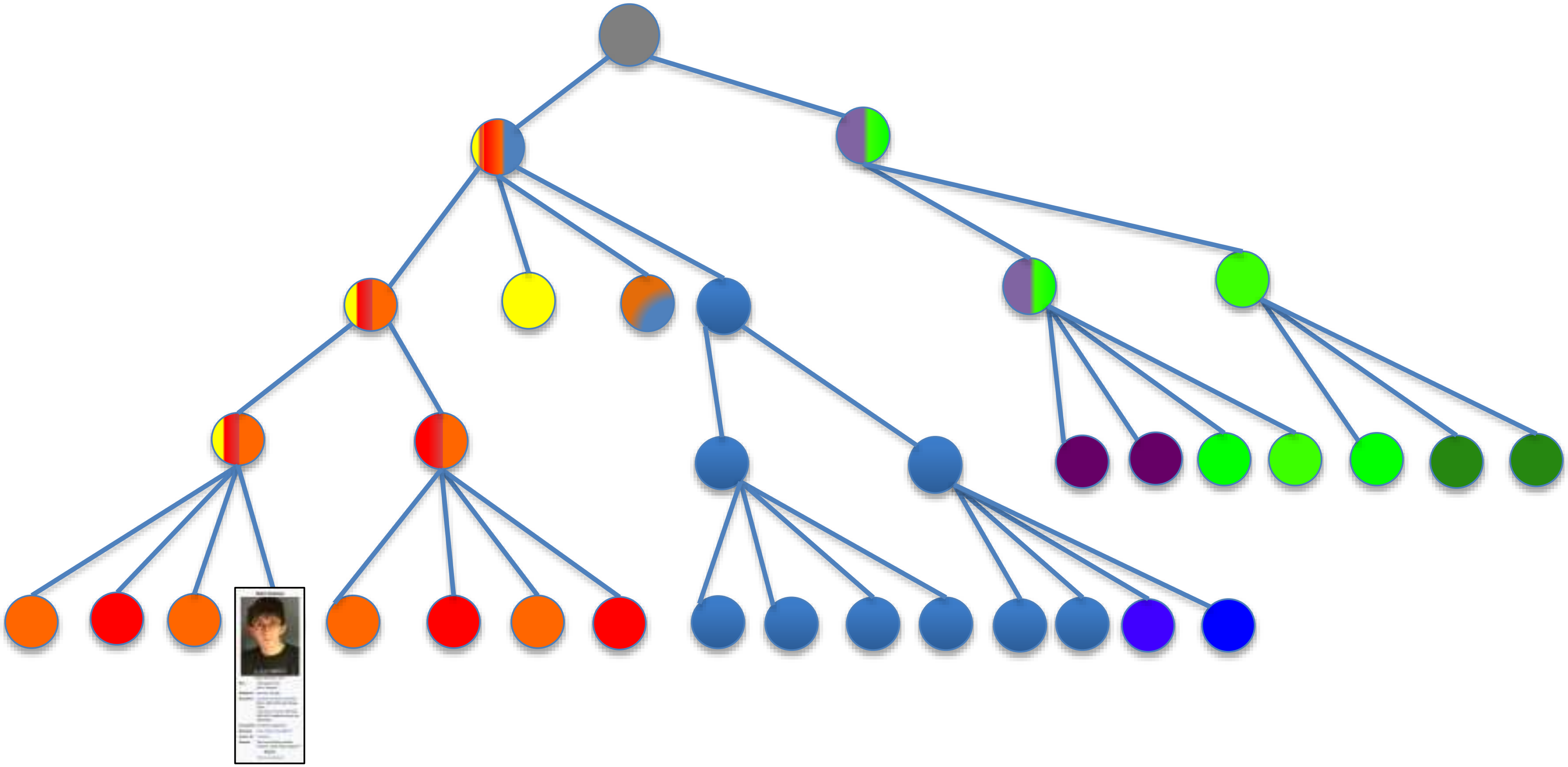
```
"#post_bow_VB_3": {
    "quote": 6,
    "infiltrate": 4,
    "disclose": 2,

"#post_co_DATE_2": {
    "2011": 2,
    "2011 3 15": 1,
    "2012": 4,
    "2013": 26,
    "2014": 12,
    "2015": 4,
    "april": 11,
    "august": 6,
    "february": 17,
    "january": 10,
    "july": 29,
    "june": 8,
    "november": 21,
    "october": 8,
    "september": 10
},
```

```
"#post_co_PER_1": {
         "al assad": 2,
         "barack": 2,
         "bashar": 3,
         "buscemi": 1,
         "crook": 1,
         "foster": 1,
         "fowler": 1,
         "jordan": 1,
         "monde": 1,
         "obama": 6,
         "sabari": 2,
```

```
"post_co_PER_1": {
       "37246009": -1,
       "131492056": 2,
       "437879322": 1,
       "664353967": -2,
       "680387639": -1,
       "681138746": 1,
       "740105633": -1,
       "792534747": 2,
       "973559166": 3,
       "1146544765": 6,
       "1273592341": 2,
```


3rd Party NER, e.g. Serif, Basis
StreamCorpus Pipeline (FOSS)
TreeLab Feature Collection
Large Collections of Unstructured Documents

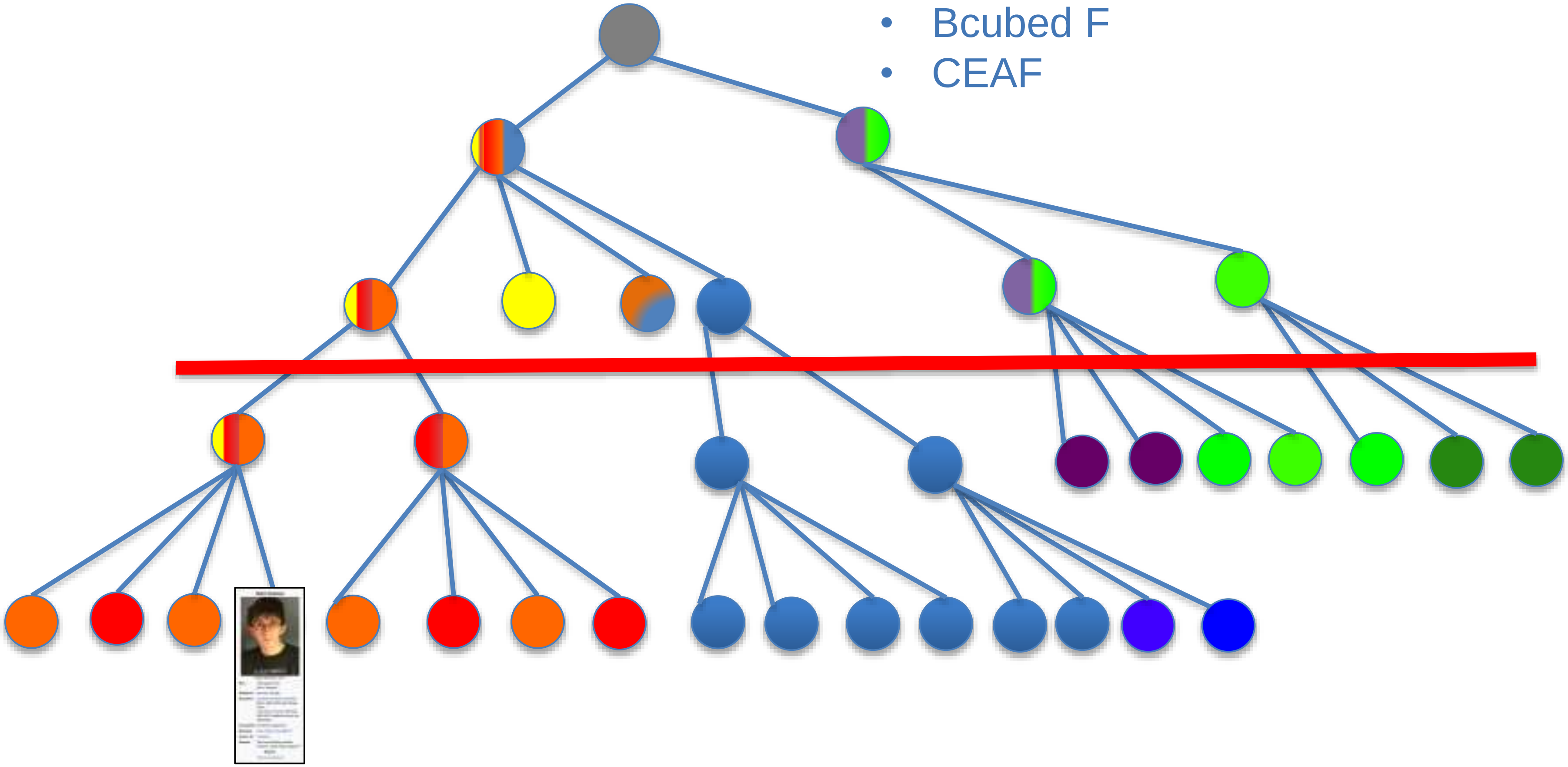
Distantly supervised models → not perfect



Distantly supervised models → not perfect

Can flatten into a set of sets for Cross Doc Coref Resolution (CDCR) with set-based metrics, e.g.,

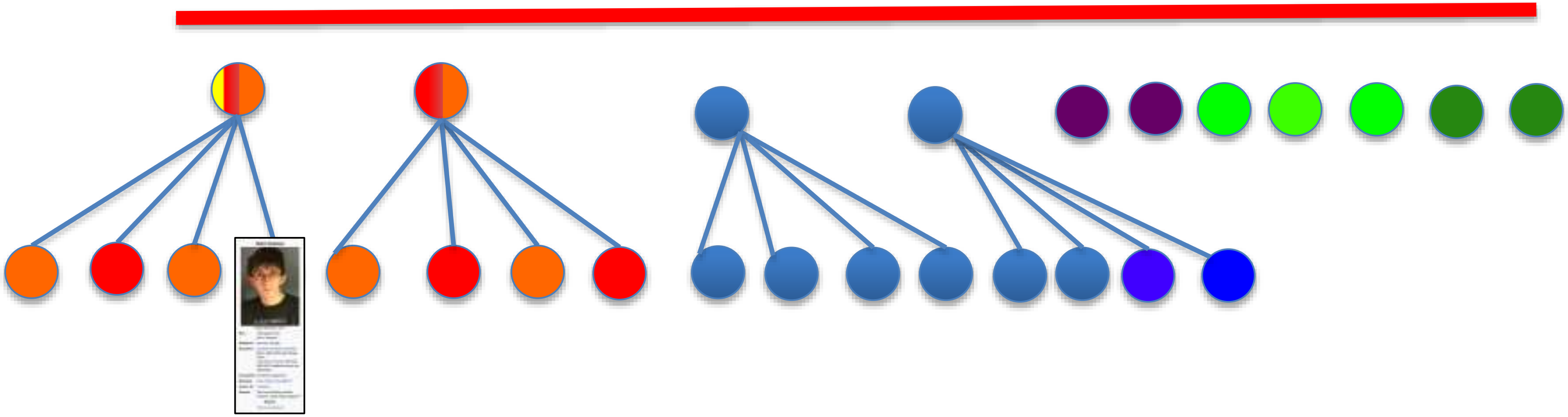
- Bcubed F
- CEAF



Distantly supervised models → not perfect

Can flatten into a set of sets for Cross Doc Coref Resolution (CDCR) with set-based metrics, e.g.,

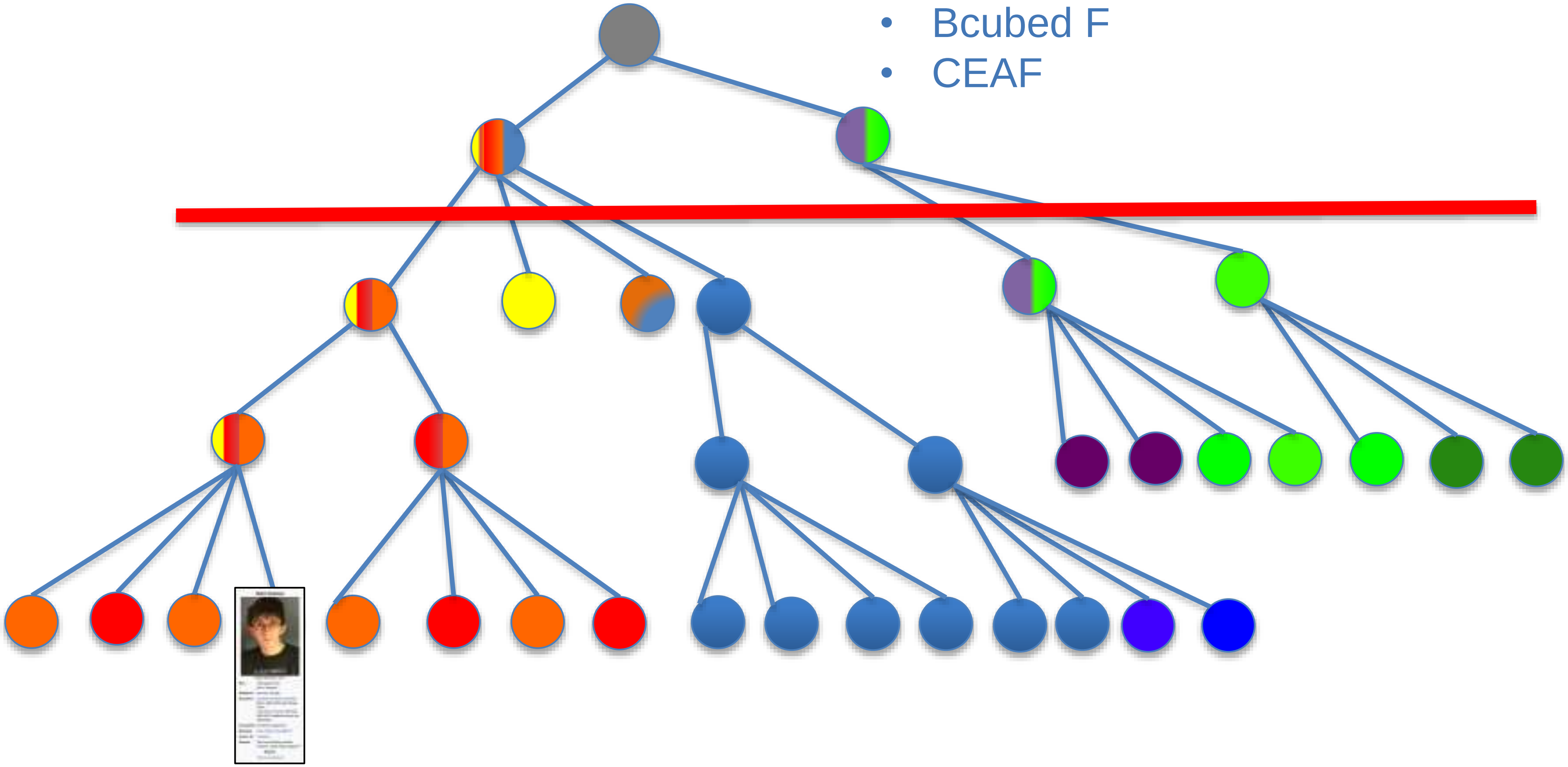
- Bcubed F
- CEAF



Distantly supervised models → not perfect

Can flatten into a set of sets for Cross Doc Coref Resolution (CDCR) with set-based metrics, e.g.,

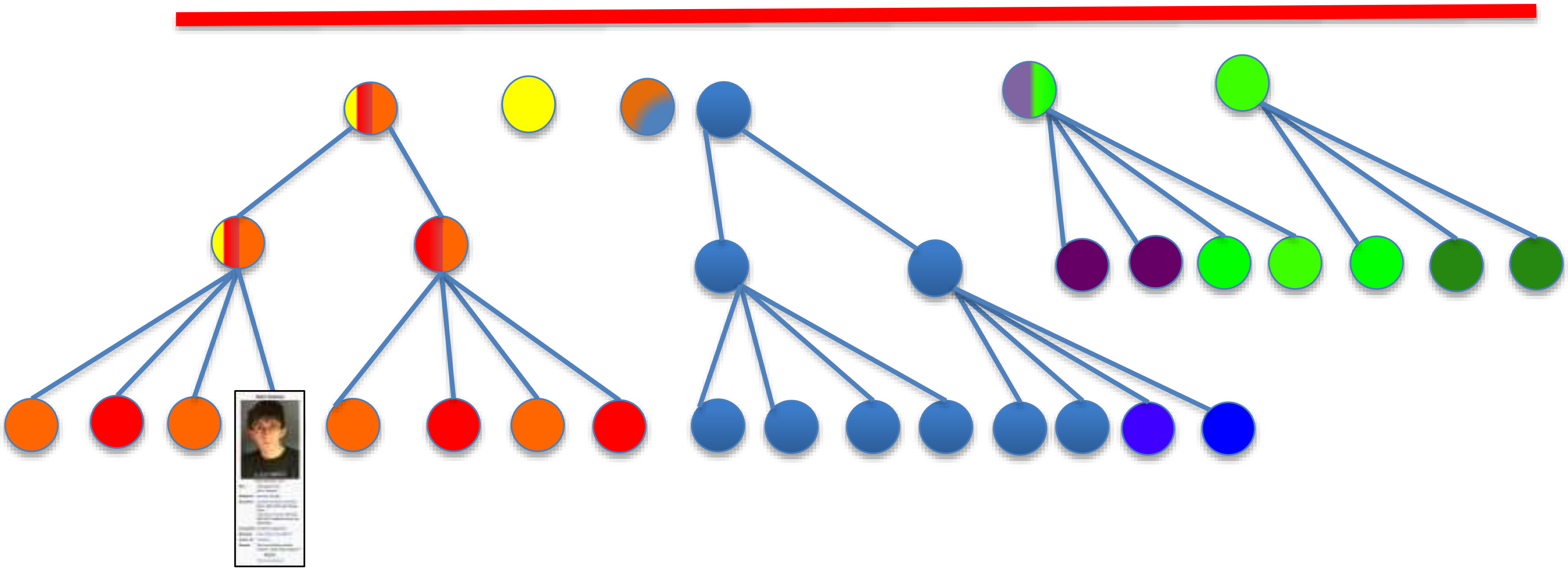
- Bcubed F
- CEAF



Distantly supervised models → not perfect

Can flatten into a set of sets for Cross Doc Coref Resolution (CDCR) with set-based metrics, e.g.,

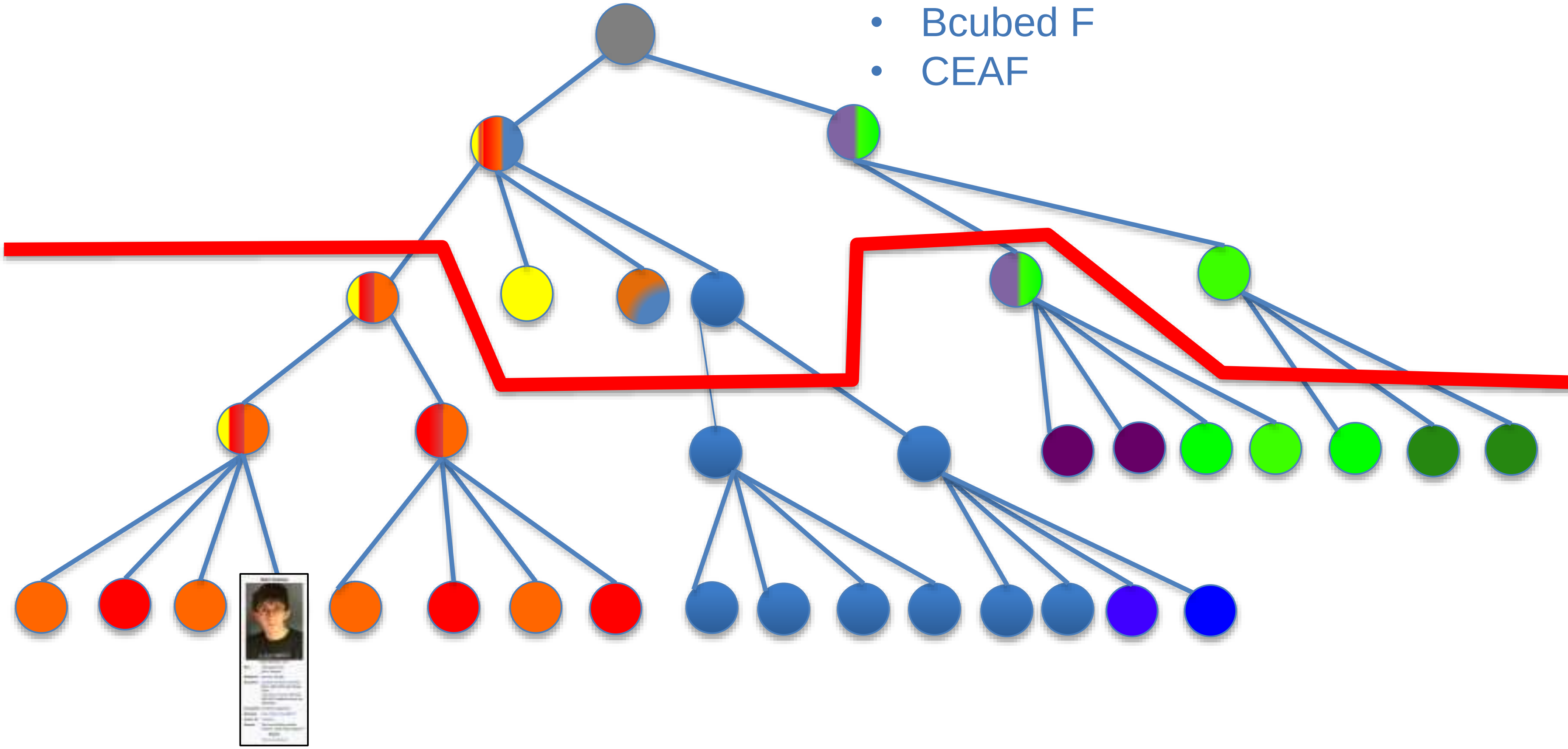
- Bcubed F
- CEAF



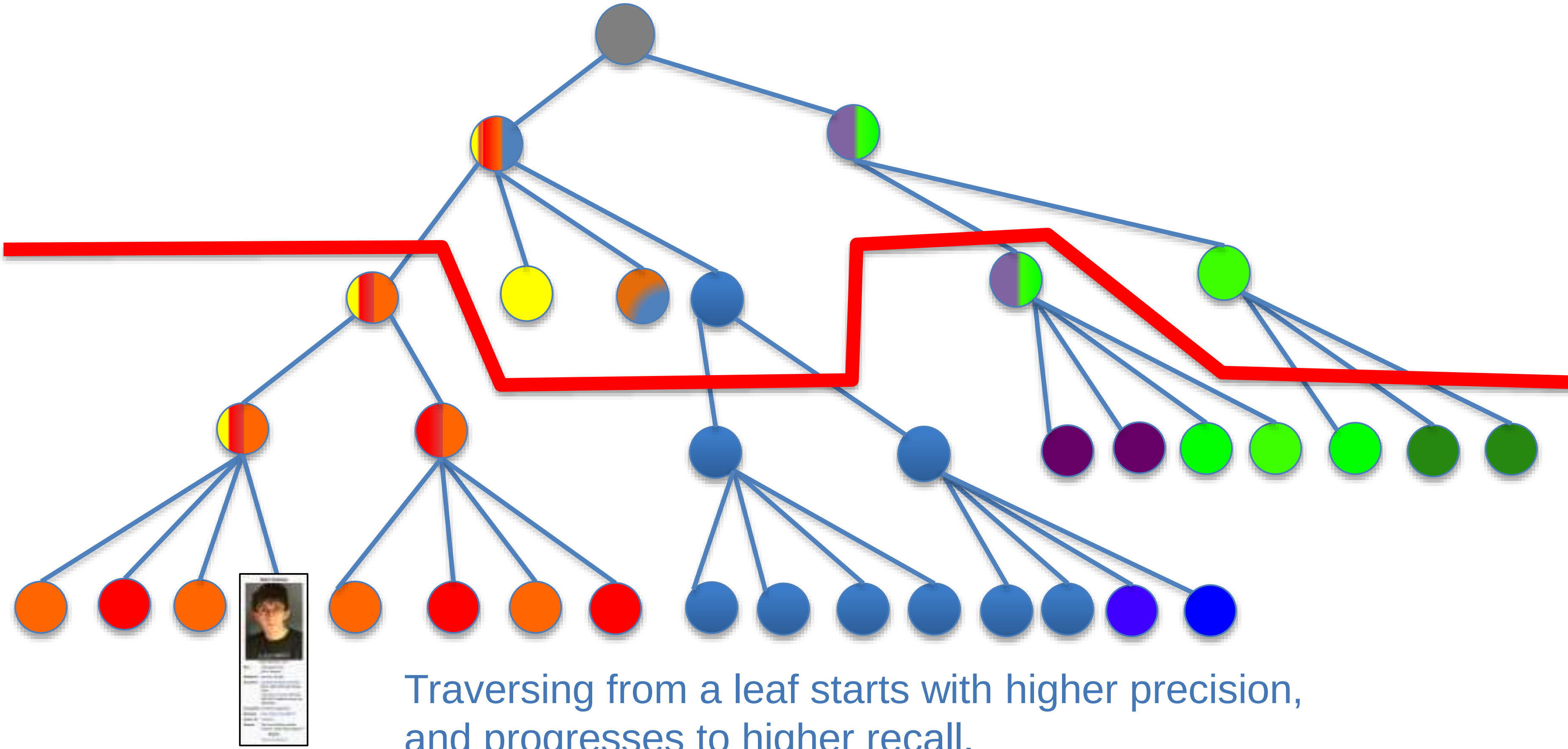
Distantly supervised models → not perfect

Can flatten into a set of sets for Cross Doc Coref Resolution (CDCR) with set-based metrics, e.g.,

- Bcubed F
- CEAF

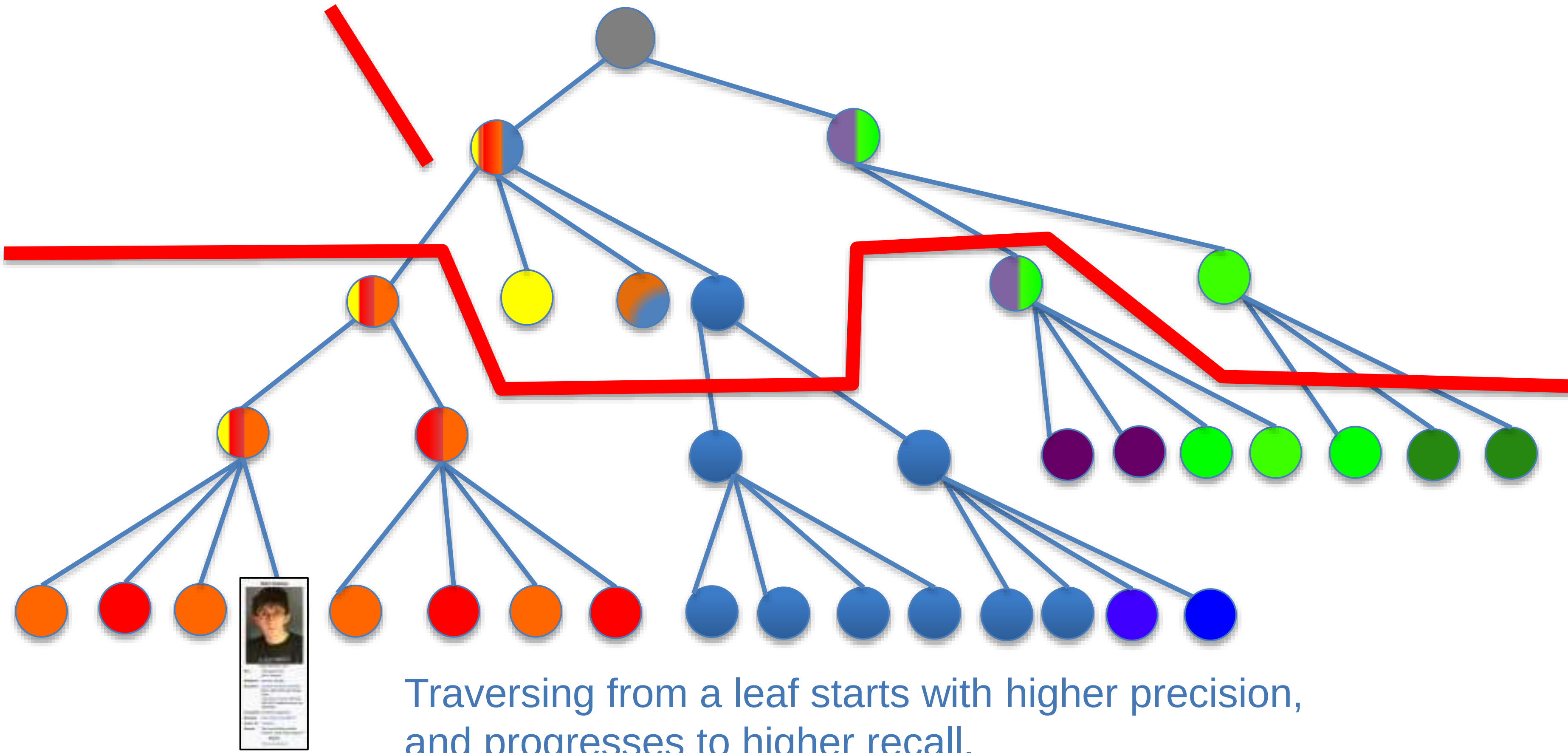


Distantly supervised models → not perfect



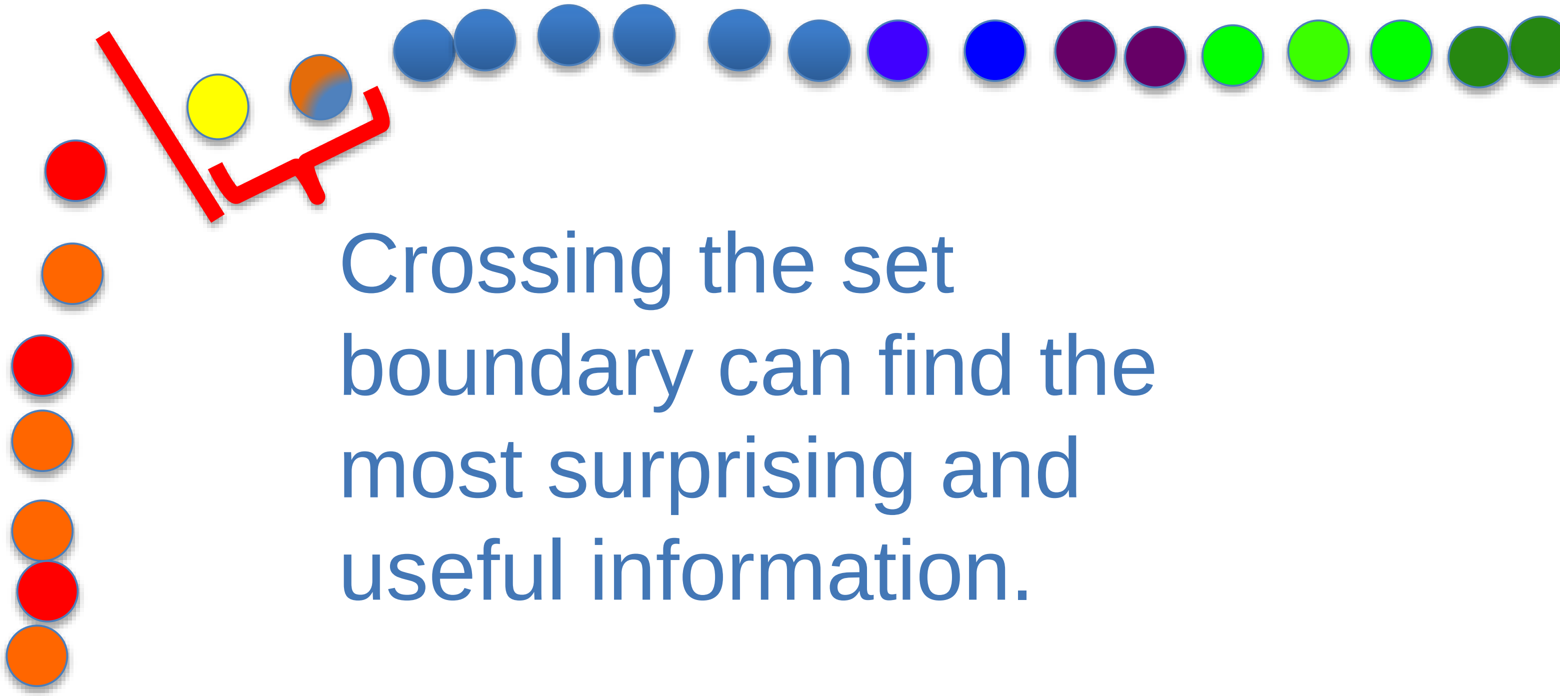
Traversing from a leaf starts with higher precision, and progresses to higher recall.

Ranking models → just different



Traversing from a leaf starts with higher precision, and progresses to higher recall.

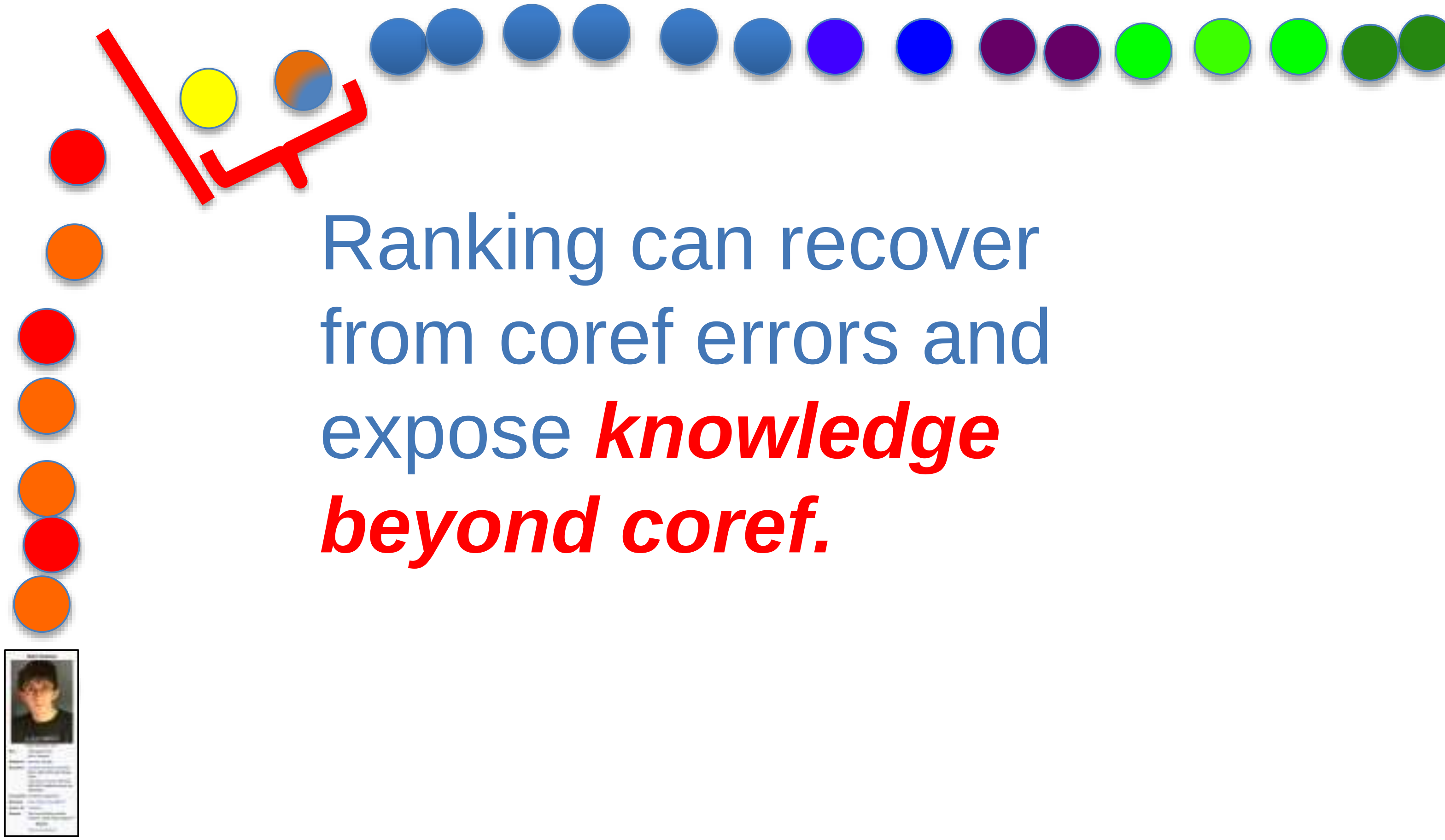
Ranking models → just different



Crossing the set boundary can find the most surprising and useful information.

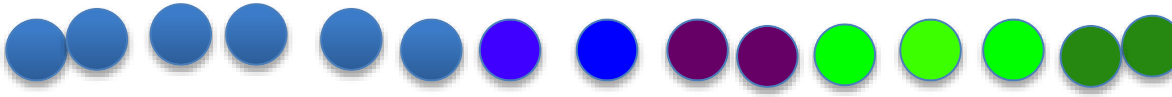
Traversing from a leaf starts with higher precision, and progresses to higher recall.

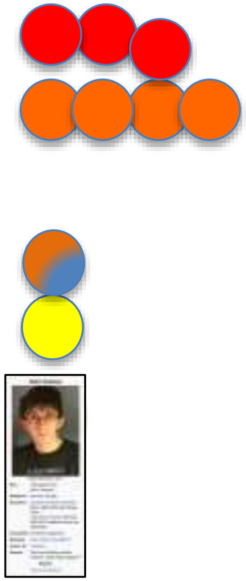
Ranking models
→ beyond coref



Ranking can recover from coref errors and expose ***knowledge beyond coref.***

Ranking models → beyond coref





Ranking can recover from coref errors and expose ***knowledge beyond coref.***

Measurement Methodologies

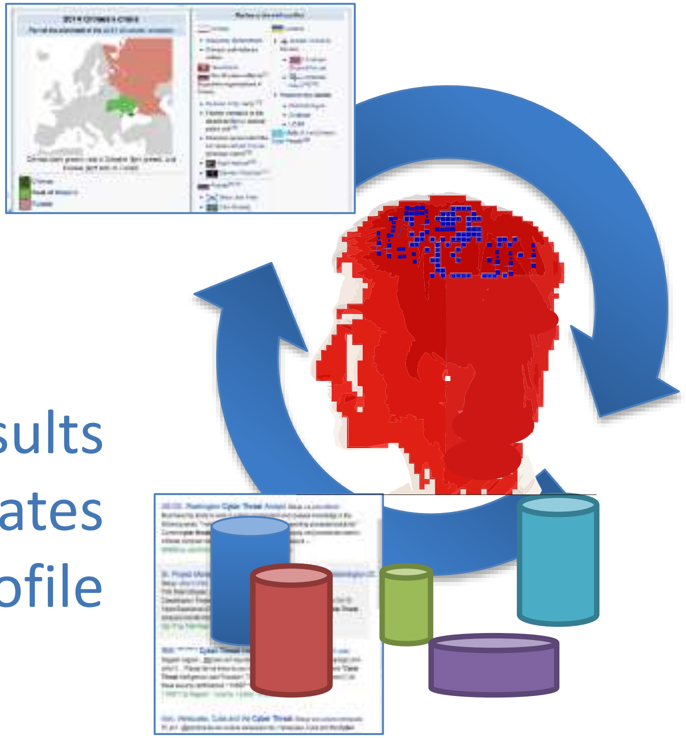
Unexpected Expected Reciprocal Rank u-ERR

```
1.  Submit profile as query to search engine
2.  Receive items
3.  Start reading through the list:
    4.  User examines position  n
    5.  If user finds new knowledge:
    6.          Update profile
    7.          Go to 1 with updated profile as query
    8.  else
    9.          n += 1
    10.         Go to 4
```

Figure of merit: depth in the list where user discovers new knowledge

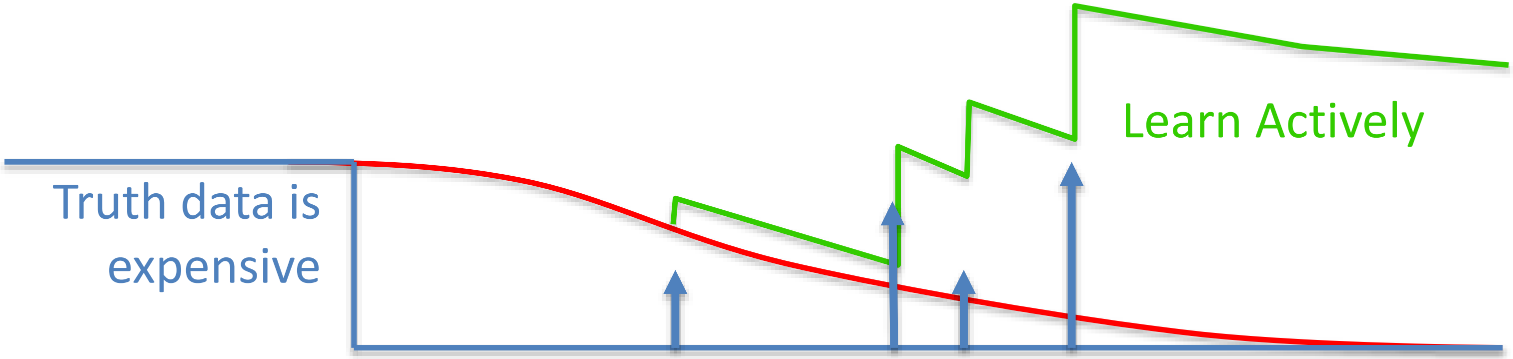
u-ERR = 1 / (expected list position of surprise)

Entity profiles
act as queries for
retrieval algorithms



Retrieval results
suggest updates
to the profile

User interactions
steer the system
beyond coref to
find surprises



Learn Actively

Truth data is
expensive

Observe user actions

Thank you!

diffeo.com

trec-kba.org