INFO116

Semantics for Social Data

Semantic Social Web

- Chapter 7
- SSocW is a key component of Web3.0
- Organize and analyze social data
  - Ontologies
  - Document level metadata

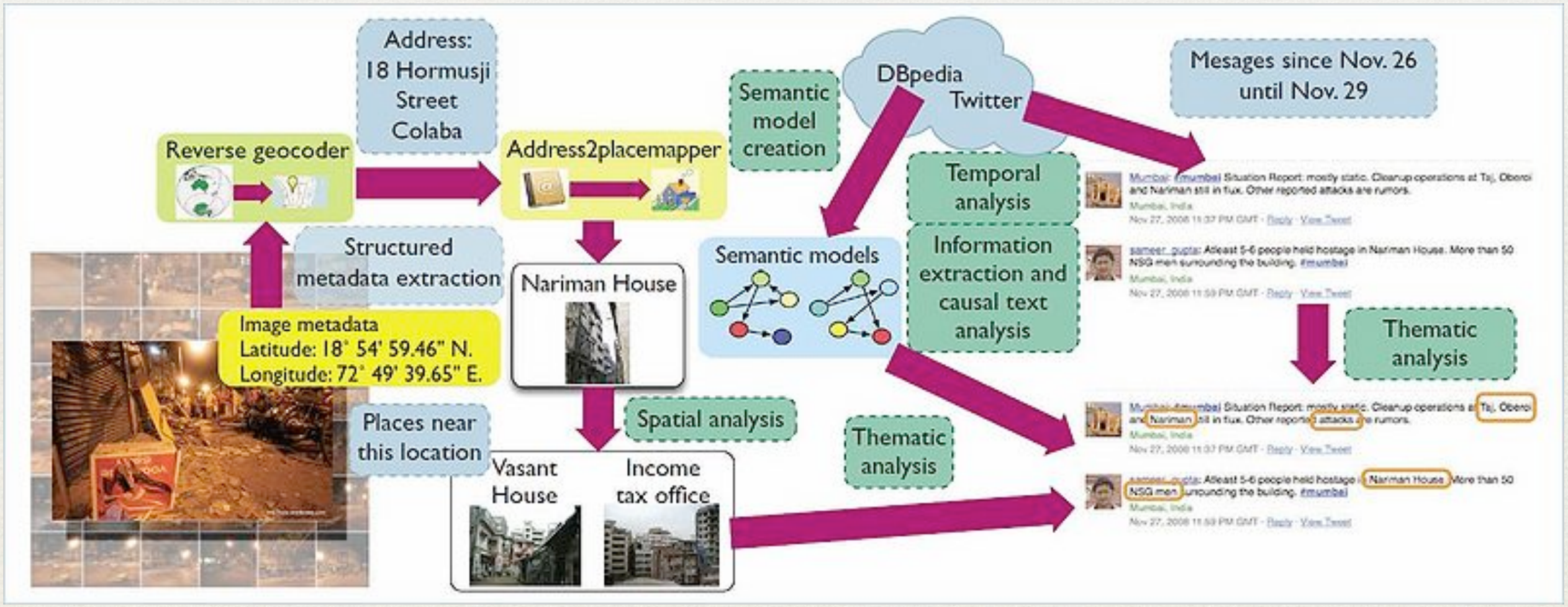
Social Web

- User Generated Content
  - “Informal” Natural Language
    - “Off topic discussions” - context muddled
- Links to other web resources
- Links to other people
- Links to social networks

Social Web Data

- Often commentary on current events from different perspectives
  - What is being said? (theme)
  - Where is it being said? (spatial)
  - When is it being said? (temporal)
- All are naturally dynamic

Situation Awareness

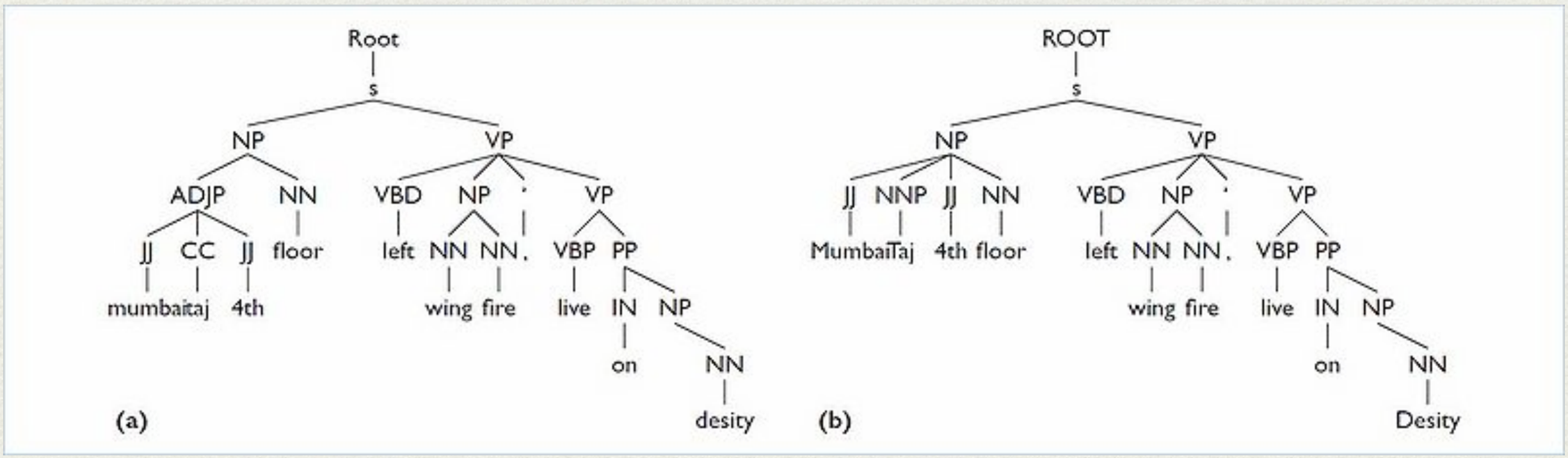


Thematic Analysis

- Semantic annotation of content refers to the process of making data more meaningful through labels (via marking up, tagging, or annotating) that conform to an agreed-upon reference model
- The key to semantically annotating content is the process of identifying and disambiguating named entities.

NLP techniques

- NLP can go wrong, better with named entity recognition
- "mumbai taj 4th floor left wing fire, live on desitv"



http://nlp.cs.berkeley.edu/

jahendler
@jahendler

@csabaveres missing the article - should be "this not a sentence" :-)

18 August 2011 at 15:54
via TweetDeck

In reply to...

Csaba Veres @csabaveres 774d
This not sentence, **@jahendler**

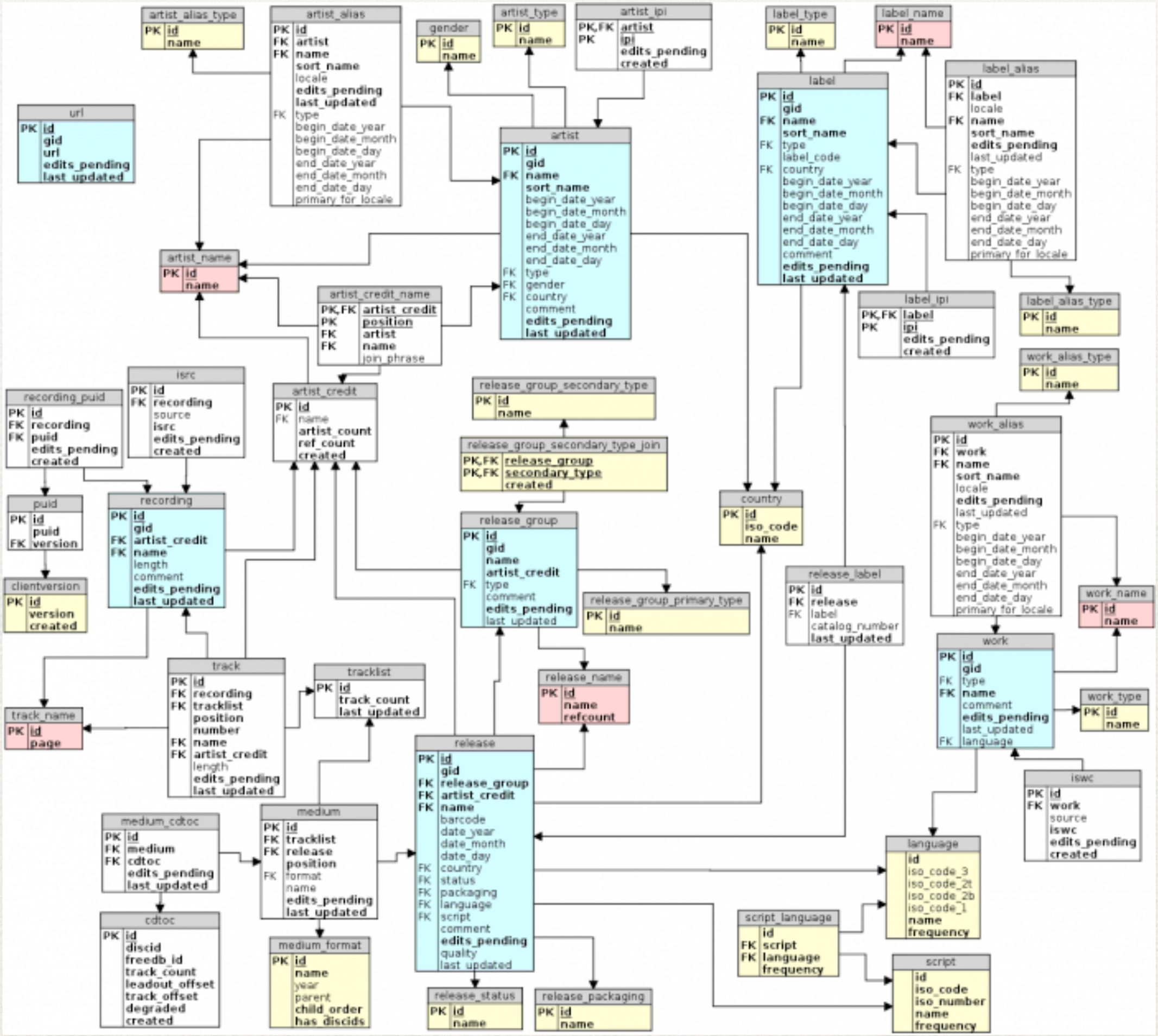
jahendler @jahendler 774d
this sentence, no verb.

Semantic support (1/3)

- An ontology of types and relations would help, but what ontology?
- Terrorism (explosions, blasts, incidents, ...), current affairs, locations, people, ...
- Taj Mahal Palace = taj
- How???
  - "Ontology fragments"

Semantic support (2/3)

- “Lily I loved your cheryl tweedy do … heart Amy.”
- “Lils smile so rocks,”
- http://musicbrainz.org


artist_alias_type
PK id
name
artist_alias
PK id
FK artist
FK name
sort_name
locale
edits_pending
last_updated
FK type
begin_date_year
begin_date_month
begin_date_day
end_date_year
end_date_month
end_date_day
primary_for_locale
gender
PK id
name
artist_type
PK id
name
artist_ipi
PK,FK artist
PK ipi
edits_pending
created
label_type
PK id
name
label_name
PK id
name
label_alias
PK id
FK label
locale
FK name
sort_name
edits_pending
last_updated
FK type
begin_date_year
begin_date_month
begin_date_day
end_date_year
end_date_month
end_date_day
primary_for_locale
label
PK id
gid
FK name
sort_name
FK type
label_code
FK country
begin_date_year
begin_date_month
begin_date_day
end_date_year
end_date_month
end_date_day
comment
edits_pending
last_updated
label_ipi
PK,FK label
PK ipi
edits_pending
created
label_alias_type
PK id
name
work_alias_type
PK id
name
work_alias
PK id
FK work
FK name
sort_name
locale
edits_pending
last_updated
FK type
begin_date_year
begin_date_month
begin_date_day
end_date_year
end_date_month
end_date_day
primary_for_locale
work_name
PK id
name
work
PK id
gid
FK type
FK name
comment
edits_pending
last_updated
FK language
work_type
PK id
name
iswc
PK id
FK work
source
iswc
edits_pending
created
language
id
iso_code_3
iso_code_2t
iso_code_2b
iso_code_1
name
frequency
script_language
id
FK script
FK language
frequency
script
id
iso_code
iso_number
name
frequency
url
PK id
gid
url
edits_pending
last_updated
artist
PK id
gid
FK name
sort_name
begin_date_year
begin_date_month
begin_date_day
end_date_year
end_date_month
end_date_day
FK type
FK gender
FK country
comment
edits_pending
last_updated
artist_name
PK id
name
artist_credit_name
PK,FK artist_credit
PK position
FK artist
FK name
join_phrase
artist_credit
PK id
FK name
artist_count
ref_count
created
isrc
PK id
FK recording
source
isrc
edits_pending
created
recording_puid
PK id
FK recording
FK puid
edits_pending
created
puid
PK id
puid
FK version
clientversion
PK id
version
created
recording
PK id
gid
FK artist_credit
FK name
length
comment
edits_pending
last_updated
track_name
PK id
page
track
PK id
FK recording
FK tracklist
position
number
FK name
FK artist_credit
length
edits_pending
last_updated
tracklist
PK id
track_count
last_updated
release_group_secondary_type
PK id
name
release_group_secondary_type_join
PK,FK release_group
PK,FK secondary_type
created
release_group
PK id
gid
name
artist_credit
FK type
comment
edits_pending
last_updated
release_group_primary_type
PK id
name
release_name
PK id
name
refcount
release_label
PK id
FK release
FK label
catalog_number
last_updated
country
PK id
iso_code
name
release
PK id
gid
FK release_group
FK artist_credit
FK name
barcode
date_year
date_month
date_day
FK country
FK status
FK packaging
FK language
FK script
comment
edits_pending
quality
last_updated
medium_cdtoc
PK id
FK medium
FK cdtoc
edits_pending
last_updated
cdtoc
PK id
discid
freedb_id
track_count
leadout_offset
track_offset
degraded
created
medium
PK id
FK tracklist
FK release
position
FK format
name
edits_pending
last_updated
medium_format
PK id
name
year
parent
child_order
has_discids
release_status
PK id
name
release_packaging
PK id
name

Semantic support (3/3)

- “Cheryl Tweedy” is a track by artist ‘Lily Allen.’
- Amy Winehouse’ and ‘Lily Allen’ are different artists from different genres—Pop and Jazz respectively
- “Smile” is a track by “Lily Allen” (with a high string similarity between “Lily” and “Lils”) and is a possible entity of interest.

Off topic noise

- I NEED HELP WITH SONY VEGAS PRO 8!! Ugh and i have a video project due tomorrow for merrill lynch :(all i need to do is simple: Extract several scenes from a clip, insert captions, transitions and thats it. really. omgg i can't figure out anything!! help!! and i got food poisoning from eggs. its not fun. Pleasssse, help? :(
- Semantic model needs to focus on related words

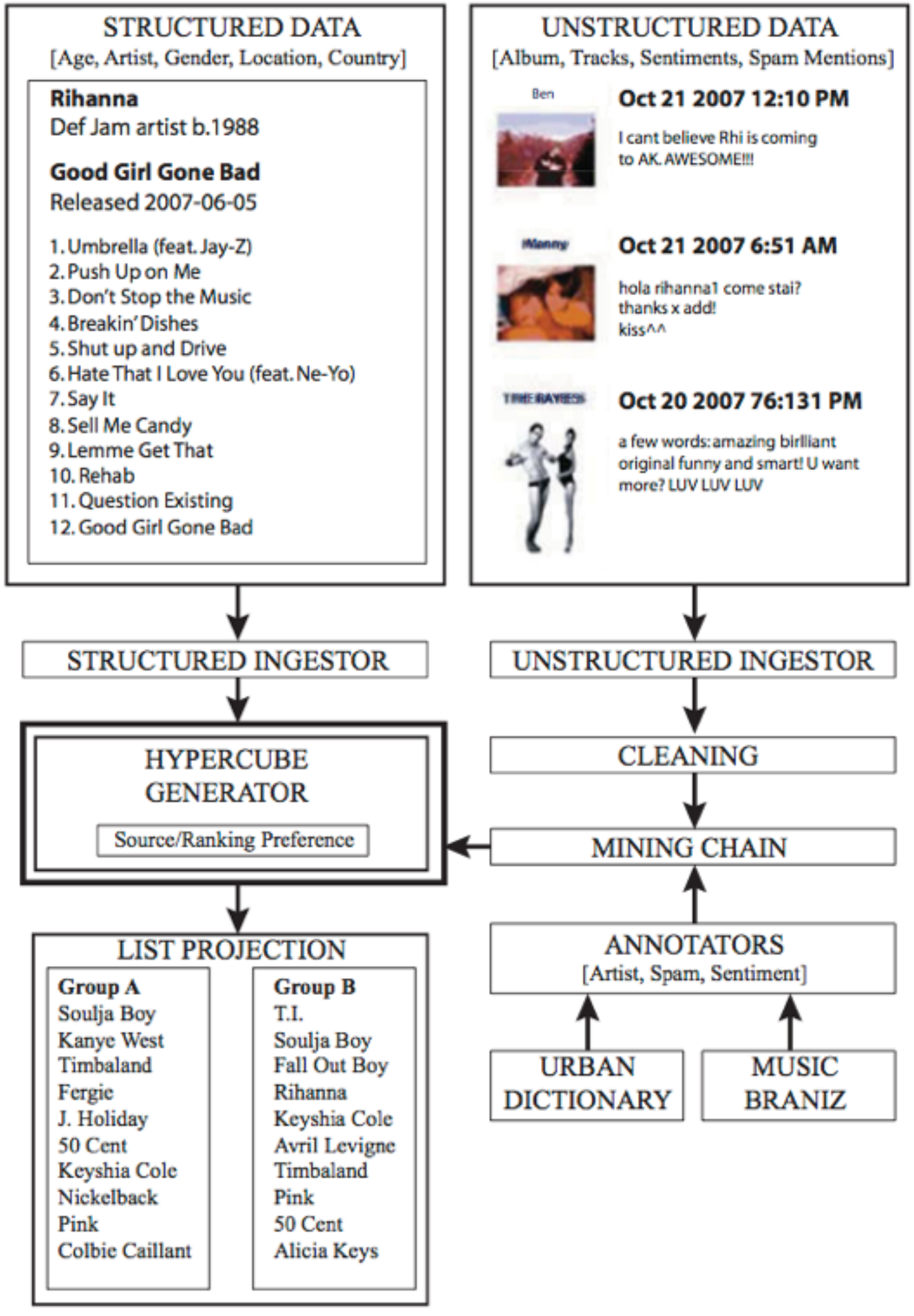
Analyzing user comments: SoundIndex

- Popularity of musical artists
- Sales not necessarily a good indication
- Popularity can now be determined by monitoring on-line public discussions
  - measure music popularity by mining music enthusiasts' comments on artist pages on MySpace

BBC SoundIndex

- Enables real-time analytics of music popularity using data from a variety of Social Networks

- Crawling and Ingesting (MySpace/Twitter, iTunes/Amazon, YouTube, LastFM)
- Annotating
- Hypercube construction
- Projection to a list


STRUCTURED DATA
[Age, Artist, Gender, Location, Country]
Rihanna
Def Jam artist b.1988
Good Girl Gone Bad
Released 2007-06-05
1. Umbrella (feat. Jay-Z)
2. Push Up on Me
3. Don't Stop the Music
4. Breakin' Dishes
5. Shut up and Drive
6. Hate That I Love You (feat. Ne-Yo)
7. Say It
8. Sell Me Candy
9. Lemme Get That
10. Rehab
11. Question Existing
12. Good Girl Gone Bad
UNSTRUCTURED DATA
[Album, Tracks, Sentiments, Spam Mentions]
Ben
Oct 21 2007 12:10 PM
I cant believe Rhi is coming to AK. AWESOME!!!
Manny
Oct 21 2007 6:51 AM
hola rihanna1 come stai? thanks x add! kiss^^
Oct 20 2007 76:131 PM
a few words: amazing birlliant original funny and smart! U want more? LUV LUV LUV
STRUCTURED INGESTOR
UNSTRUCTURED INGESTOR
CLEANING
MINING CHAIN
HYPERCUBE GENERATOR
Source/Ranking Preference
ANNOTATORS
[Artist, Spam, Sentiment]
URBAN DICTIONARY
MUSIC BRANIZ
LIST PROJECTION
Group A
Soulja Boy
Kanye West
Timbaland
Fergie
J. Holiday
50 Cent
Keyshia Cole
Nickelback
Pink
Colbie Caillant
Group B
T.I.
Soulja Boy
Fall Out Boy
Rihanna
Keyshia Cole
Avril Levigne
Timbaland
Pink
50 Cent
Alicia Keys

Crawling and Ingesting

- extracts plain text data and stores in common data store for further processing
- There are nearly 50,000 artists in an initial set
  - Need strategies for collecting entries
- Each comment is uniquely identified by a combination of user name, data source, time-stamp, etc.

Annotating

- Artist and Music Annotator: Spotting artist, album, track, and other music related (e.g. labels, tours, shows, concerts) mentions.
- Sentiment Annotator: Spotting and transliterating sentiments in comments.
- Spam Annotator: Identifying comments that are spam or do not directly contribute to artist/music popularity figures (e.g. comments about an artist's DUI charge).

Artist annotator

- Window of words + Jaccard distance of a dictionary entity and entity spotted in text (MusicBrainz)
- A shallow NL parse of the comment to verify the spotted entity's part of speech tag
- Look up the spotted entity's corpus-wide statistics
- If the combined score of the three steps is greater than a threshold (e.g., 0.9 for artists and 0.8 for tracks), record the annotation with the dictionary value of the spotted entity. e.g, 'Aiimmy' in the comment is annotated as 'Amy Winehouse' to facilitate aggregation of number of artist mentions.

Sentiment annotator (1/2)

- very large number of ways users express sentiment
- annotator translates a variety of slang expressions to a finite set of known bad and good sentiments using a popular slang dictionary – UrbanDictionary.com

Your music is really bangin!
You're a genius! Keep droppin bombs!
u doin it up 4 real. i really love the album.
keep doin wat u do best. u r so bad!

hey just hittin you up showin love to one of chi-town's own. MADD LOVE.

"U R SO BAD!" and other compliments. . .

- coarse assignments of positive and negative comments on an artist's page
- youth slang dictionary used to identify word use
  - bathroom stalemate
  - going batman
  - ass-tag convention

Sentiment annotator (2/2)

- a seed of 60 positive and 45 negative sentiments is created manually to assist in this transliteration
- UD provides a set of related tags and user-defined and voted definitions for a slang term
- A shallow NL parse of a sentence to identify adjectives or verbs to suggest the presence of a sentiment.
- compute the corpus-wide statistic of a related tag and pick the one that occurs most frequently to be a transliteration for the slang term
    - tight = awesome 456, sweet 136, hot 429, sick 23, dope 182....

Spam detection

- online public sources can be infested with bot-generated spam content
- content-based identification of spam
  - testing content on patterns or regular expressions
  - learning Bayesian models over spam and non- spam content
- ineffective on our corpus because
  - comments are rather short
  - share similar buzz words with non-spam content
  - poorly formed and contain frequent variations of word/ slang

Spam annotator

- spots possible spam phrases and their variations in text using the mined spam patterns
  - Classifying a comment as spam or non-spam is done using a set of rules over the results of all the three annotators
  - An example of such a rule would be that if a spam phrase, artist and music entities, and a positive sentiment were spotted; the comment was probably not spam.

Turning into a list

- Uses standard data warehousing analystics
- Structured + unstructured data
  - M : (Age, Gender, Location, Time, Artist, Sentiment, Spam analysis, …)
- “What is hot in New York City for 19 year old males?”
- “Who are the most popular artists from San Francisco?”

Results

38%	of total comments were spam
61%	of total comments had positive sentiments
4%	of total comments had negative sentiments
35%	of total comments had no identifiable sentiments

Table 7: Annotation Statistics

Groups and Age Range	No. of male respondents	No. of female respondents
Group 1 (8-15)	8	9
Group 2 (17-22)	21	26
Group 3 (17-22)	7	3

Table 8: Survey Group Statistics

Billboard.com	MySpace Analysis
Soulja Boy	T.I.
Kanye West	Soulja Boy
Timbaland	Fall Out Boy
Fergie	Rihanna
J. Holiday	Keyshia Cole
50 Cent	Avril Lavigne
Keyshia Cole	Timbaland
Nickelback	Pink
Pink	50 Cent
Colbie Caillat	Alicia Keys

Table 9: Billboard's Top Artists vs our generated list

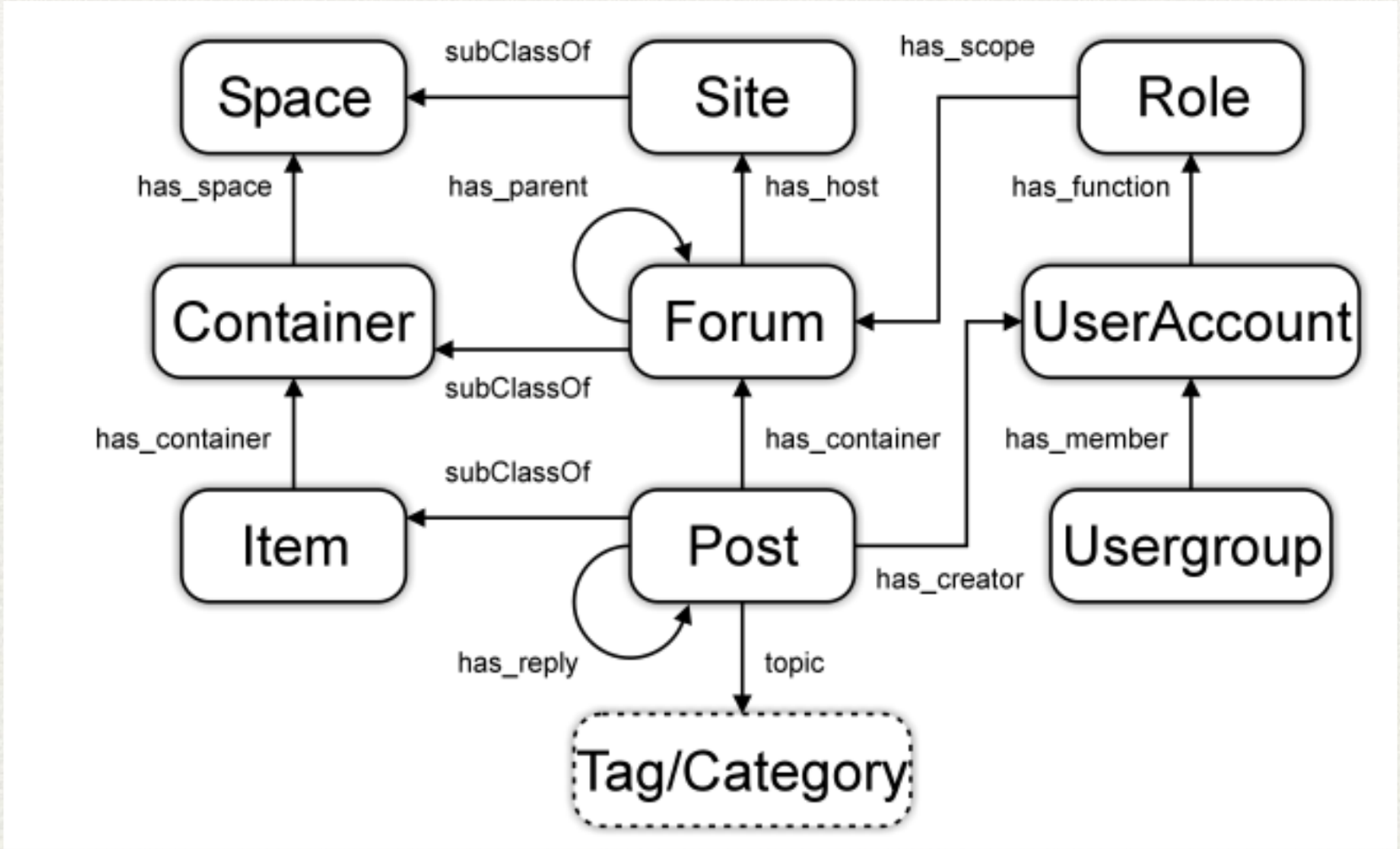
	Group 1	Group 2	Group 3
MySpace-Generated List	15	30	6
Billboard List	2	17	4

Table 10: Experiment Results: number of people who prefered each list

Improving SSocW

- FOAF (Friend of a Friend): An ontology for describing people and their relationships. (http://foaf-project.org/).
- SIOC (Semantically-Interlinked Online Communities): To fully describe content and structure of social websites, and facilitate creation and integration of online communities. (http:// sioc-project.org/).
- Semantic MediaWiki: An extension of MediaWiki, allowing users to add structured information to pages. (http://semantic-mediawiki.org/).
- WikiData https://www.wikidata.org/wiki/Wikidata:Main_Page

SIOC (1/2)


Space
subClassOf
Site
has_scope
Role
has_space
has_parent
has_host
has_function
Container
Forum
UserAccount
subClassOf
has_container
has_container
has_member
subClassOf
Item
Post
Usergroup
has_creator
has_reply
topic
Tag/Category

SIOC (2/2)

```
<sioc:Post rdf:about="http://johnbreslin.com/blog/2006/09/07/creating-connections-between-discussion-clouds-with-sioc/">
    <dcterms:title>Creating connections between discussion clouds with SIOC</dcterms:title>
    <dcterms:created>2006-09-07T09:33:30Z</dcterms:created>
    <sioc:has_container rdf:resource="http://johnbreslin.com/blog/index.php?sioc_type=site#weblog"/>
    <sioc:has_creator>
        <sioc:UserAccount rdf:about="http://johnbreslin.com/blog/author/cloud/" rdfs:label="Cloud">
            <rdfs:seeAlso rdf:resource="http://johnbreslin.com/blog/index.php?sioc_type=user&amp;sioc_id=1"/>
        </sioc:UserAccount>
    </sioc:has_creator>
    <sioc:content>SIOC provides a unified vocabulary for content and interaction description: a semantic layer that can co-ex
    <sioc:topic rdfs:label="Semantic Web" rdf:resource="http://johnbreslin.com/blog/category/semantic-web/"/>
    <sioc:topic rdfs:label="Blogs" rdf:resource="http://johnbreslin.com/blog/category/blogs/"/>
    <sioc:has_reply>
        <sioc:Post rdf:about="http://johnbreslin.com/blog/2006/09/07/creating-connections-between-discussion-clouds-with-sioc
            <rdfs:seeAlso rdf:resource="http://johnbreslin.com/blog/index.php?sioc_type=comment&amp;sioc_id=123928"/>
        </sioc:Post>
    </sioc:has_reply>
</sioc:Post>
```

Semantic Media-Wiki

- Semantic MediaWiki (SMW) is an extension of MediaWiki – the wiki application best known for powering Wikipedia – that helps to search, organise, tag, browse, evaluate, and share the wiki's content
- http://artwiki.org/ArtWiki
- http://directory.fsf.org/wiki/Main_Page