Entity Reconstruction: Putting the pieces of the puzzle back together

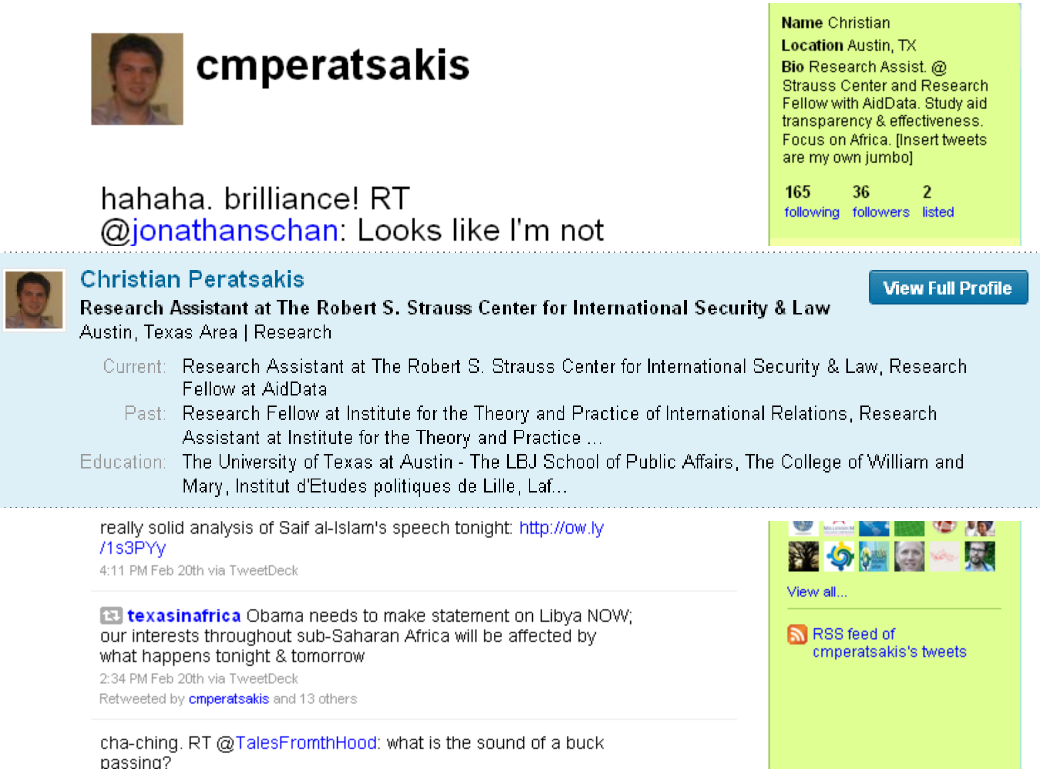
Georgia Koutrika - HP Labs, Palo Alto, USA

The Findability Challenge

Once upon a time, it was hard to find information about a person

Who is this person?

Now, the Web and people's online activities offer a breath of information!



cmperatsakis

hahaha. brilliance! RT @jonathanschan: Looks like I'm not

Name Christian
Location Austin, TX
Bio Research Assist. @ Strauss Center and Research Fellow with AidData. Study aid transparency & effectiveness. Focus on Africa. [Insert tweets are my own jumbo]

165	36	2
following	followers	listed

Christian Peratsakis
Research Assistant at The Robert S. Strauss Center for International Security & Law
Austin, Texas Area | Research

View Full Profile

Current:	Research Assistant at The Robert S. Strauss Center for International Security & Law, Research Fellow at AidData
Past:	Research Fellow at Institute for the Theory and Practice of International Relations, Research Assistant at Institute for the Theory and Practice ...
Education:	The University of Texas at Austin - The LBJ School of Public Affairs, The College of William and Mary, Institut d'Etudes politiques de Lille, Laf...

really solid analysis of Saif al-Islam's speech tonight: http://ow.ly/1s3PYy
4:11 PM Feb 20th via TweetDeck

texasinafrica Obama needs to make statement on Libya NOW; our interests throughout sub-Saharan Africa will be affected by what happens tonight & tomorrow
2:34 PM Feb 20th via TweetDeck
Retweeted by cmperatsakis and 13 others

cha-ching. RT @TalesFromthHood: what is the sound of a buck passing?

User Online Trails

Textual data

Web pages (personal, news, …)

User histories (search, browsing, purchases, …)

Posts (Blogs, Twitter, Facebook)

Comments (Yelp, Netflix, …)

Media

Movies viewed (Netflix, Hulu, …)

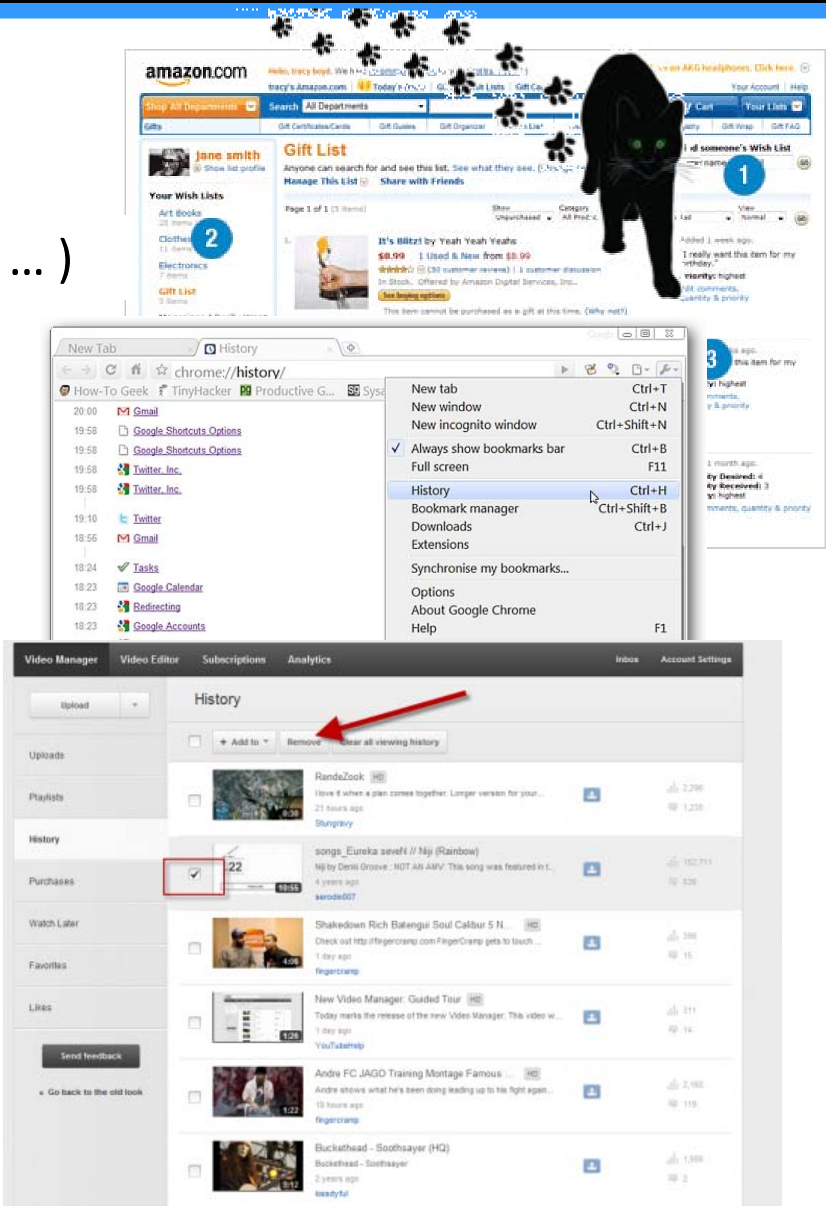
Images shared (Facebook, Google+, …)

Videos watched/shared (Youtube, …)

Social Networks

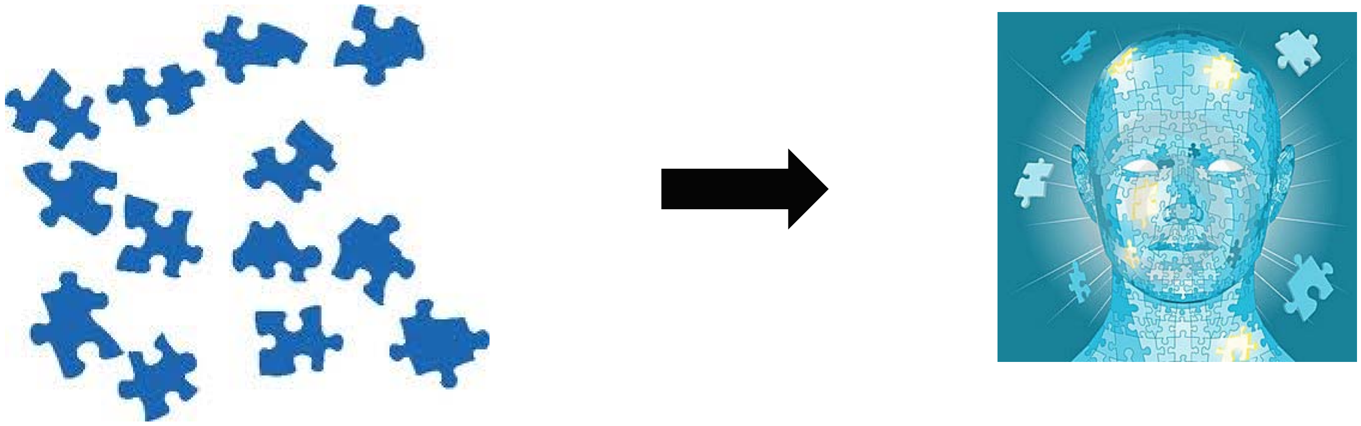
Connections (friends, family, …)

Social activity



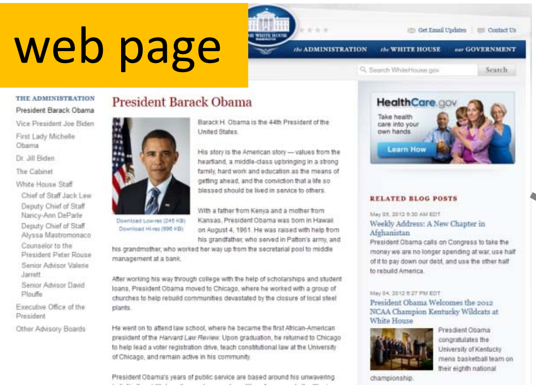
From User Online Trails to …. You!

Analyzing and combining these pieces of information together can lead to **valuable insights** about users and opens up the door to **tremendous opportunities** in sectors including education, health, marketing, law enforcement



Example Applications: Education

Personal web page



THE ADMINISTRATION
President Barack Obama
Vice President Joe Biden
First Lady Michelle Obama
Dr. Jill Biden
The Cabinet
White House Staff
Chief of Staff Jack Lew
Deputy Chief of Staff Nancy-Ann DeParle
Deputy Chief of Staff Alyssa Mastromonaco
Counselor to the President Peter Rouse
Senior Advisor Valerie Jarrett
Senior Advisor David Plouffe
Executive Office of the President
Other Advisory Boards

President Barack Obama

Download Low-res (245 KB)
Download Hi-res (996 KB)

Barack H. Obama is the 44th President of the United States.

His story is the American story — values from the heartland, a middle-class upbringing in a strong family, hard work and education as the means of getting ahead, and the conviction that a life so blessed should be lived in service to others.

With a father from Kenya and a mother from Kansas, President Obama was born in Hawaii on August 4, 1961. He was raised with help from his grandfather, who served in Patton's army, and his grandmother, who worked her way up from the secretarial pool to middle management at a bank.

After working his way through college with the help of scholarships and student loans, President Obama moved to Chicago, where he worked with a group of churches to help rebuild communities devastated by the closure of local steel plants.

He went on to attend law school, where he became the first African-American president of the *Harvard Law Review*. Upon graduation, he returned to Chicago to help lead a voter registration drive, teach constitutional law at the University of Chicago, and remain active in his community.

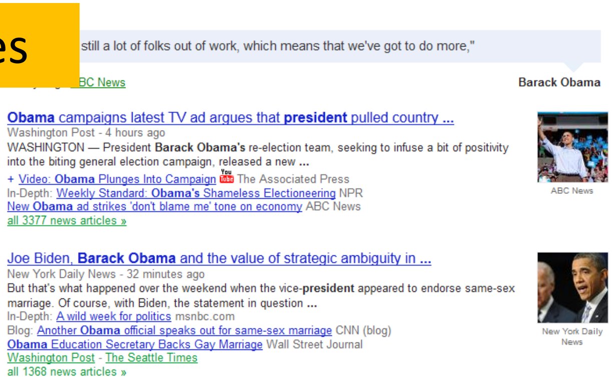
President Obama's years of public service are based around his unwavering

RELATED BLOG POSTS

May 05, 2012 5:30 AM EDT
Weekly Address: A New Chapter in Afghanistan
President Obama calls on Congress to take the money we are no longer spending at war, use half of it to pay down our debt, and use the other half to rebuild America.

May 04, 2012 6:27 PM EDT
President Obama Welcomes the 2012 NCAA Champion Kentucky Wildcats at White House
President Obama congratulates the University of Kentucky mens basketball team on their eighth national championship.

News pages



still a lot of folks out of work, which means that we've got to do more,"

BC News

Barack Obama

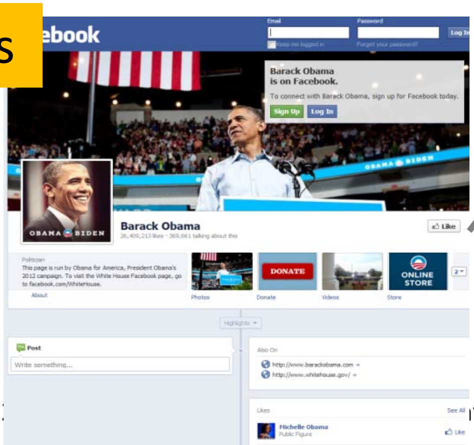
Obama campaigns latest TV ad argues that **president** pulled country **...**
Washington Post - 4 hours ago
WASHINGTON — President **Barack Obama's** re-election team, seeking to infuse a bit of positivity into the biting general election campaign, released a new **...**
+ Video: **Obama** Plunges Into Campaign The Associated Press
In-Depth: Weekly Standard: **Obama's** Shameless Electioneering NPR
New **Obama** ad strikes 'don't blame me' tone on economy ABC News
all 3377 news articles »

ABC News

Joe Biden, **Barack Obama** and the value of strategic ambiguity in **...**
New York Daily News - 32 minutes ago
But that's what happened over the weekend when the vice-**president** appeared to endorse same-sex marriage. Of course, with Biden, the statement in question **...**
In-Depth: A wild week for politics msnbc.com
Blog: Another **Obama** official speaks out for same-sex marriage CNN (blog)
Obama Education Secretary Backs Gay Marriage Wall Street Journal
Washington Post - The Seattle Times
all 1368 news articles »

New York Daily News

Social networks



Barack Obama is on Facebook.

To connect with Barack Obama, sign up for Facebook today.

Sign Up Log In

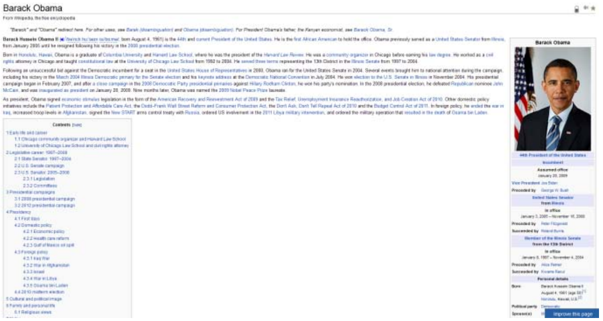
Barack Obama
26,409,213 likes · 369,661 talking about this

Politician
This page is run by Obama for America, President Obama's 2012 campaign. To visit the White House Facebook page, go to facebook.com/WhiteHouse.

Also On
http://www.barackobama.com
http://www.whitehouse.gov/

Likes
See All
Michelle Obama
Public Figure

Automatically Updated Dynamic Wikipedia Page

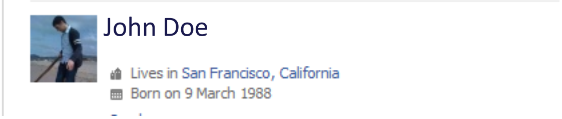


Barack Obama

From Wikipedia, the free encyclopedia

Example Applications: Health

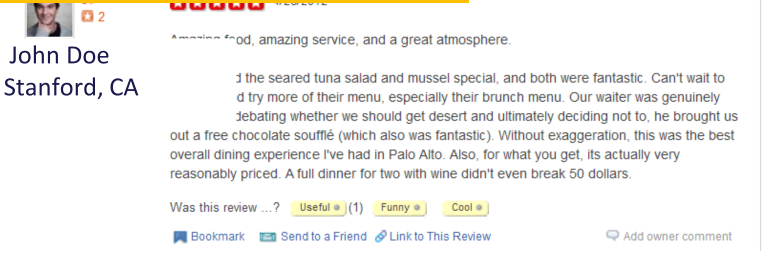
Social site presence



John Doe

Lives in San Francisco, California
Born on 9 March 1988

Posts on social sites



2

John Doe
Stanford, CA

Amazing food, amazing service, and a great atmosphere.

d the seared tuna salad and mussel special, and both were fantastic. Can't wait to d try more of their menu, especially their brunch menu. Our waiter was genuinely debating whether we should get desert and ultimately deciding not to, he brought us out a free chocolate soufflé (which also was fantastic). Without exaggeration, this was the best overall dining experience I've had in Palo Alto. Also, for what you get, its actually very reasonably priced. A full dinner for two with wine didn't even break 50 dollars.

Health Record

Health Record for John Doe

My Medicare History

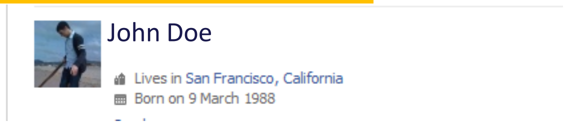
Medications

Browsing History

http://www.lovefood.com/guide/recipes/11389/claudia-rodens-orange-and-almond-cake
http://www.yelp.com/c/palo-alto-ca/breakfast_brunch
…

Example Applications: Marketing

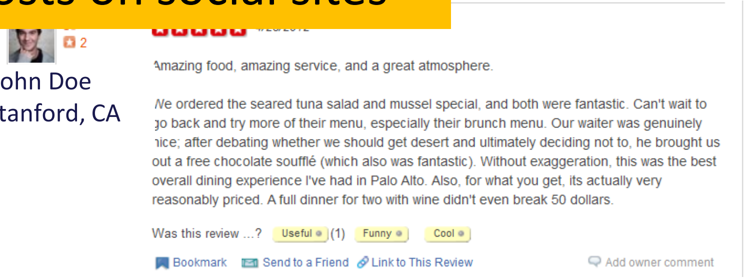
Social site presence



John Doe

Lives in San Francisco, California
Born on 9 March 1988

Posts on social sites



John Doe
Stanford, CA

Amazing food, amazing service, and a great atmosphere.

We ordered the seared tuna salad and mussel special, and both were fantastic. Can't wait to go back and try more of their menu, especially their brunch menu. Our waiter was genuinely nice; after debating whether we should get desert and ultimately deciding not to, he brought us out a free chocolate soufflé (which also was fantastic). Without exaggeration, this was the best overall dining experience I've had in Palo Alto. Also, for what you get, its actually very reasonably priced. A full dinner for two with wine didn't even break 50 dollars.

Browsing History

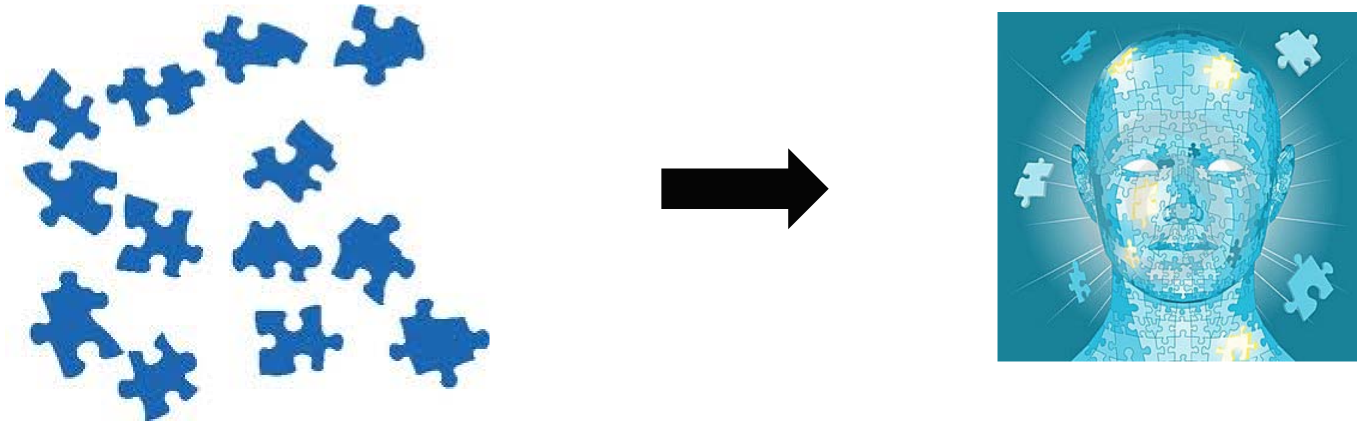
http://www.lovefood.com/guide/recipes/11389/claudia-rodens-orange-and-almond-cake
http://www.yelp.com/c/palo-alto-ca/breakfast_brunch
…

Consumer Profile

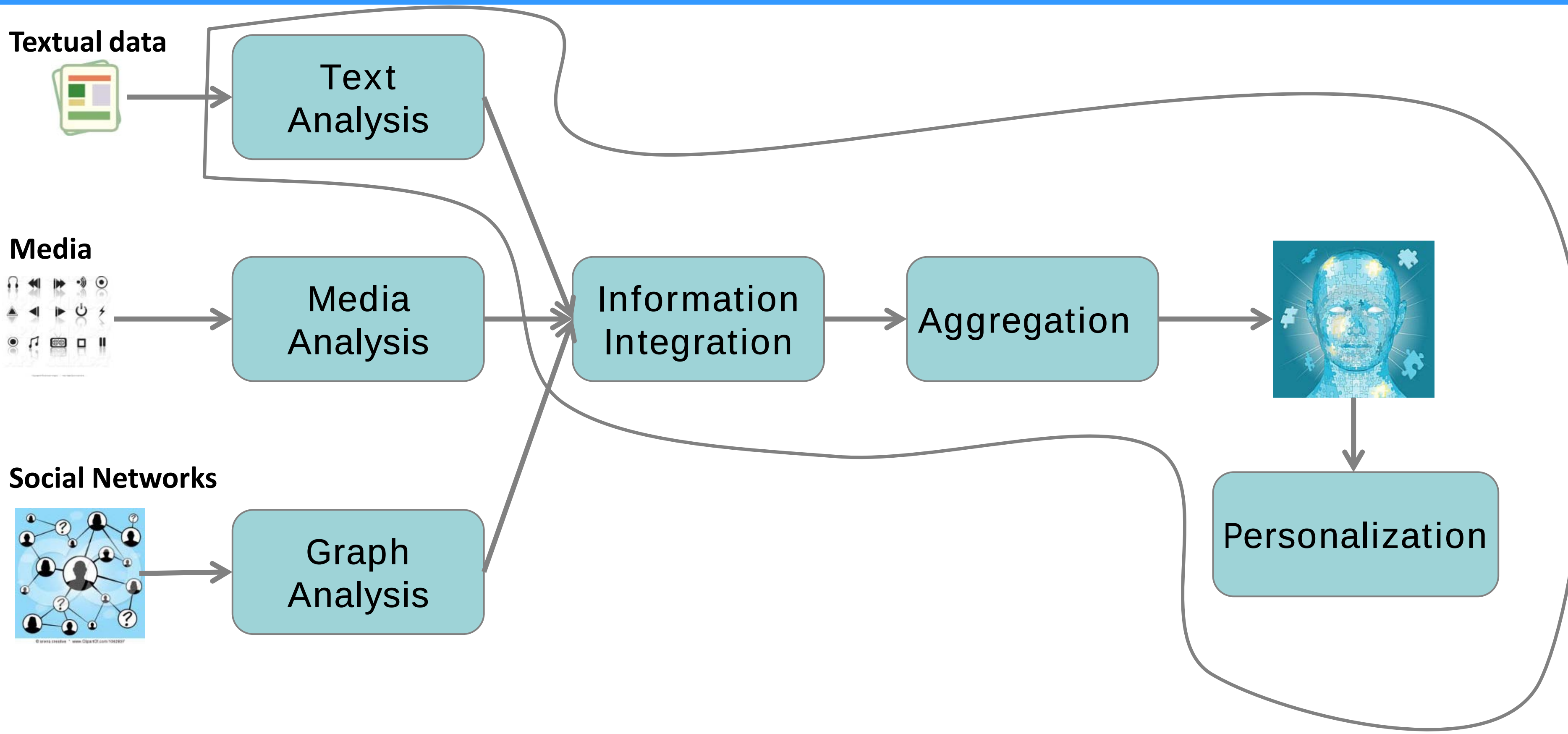
Socio-demography	**Consumer's habits**
age	
sex	buying habits
status	general consumption
education	of consumer goods
income	monthly consumption
	buying factors
Values, lifestyle, …	loyalty of tradesmen
Time spending	**Nutrition**
time for activites	habits
frequency	adittudes
Media Consumption	**Cars**
TV	driver
radio	No. of cars
print	type of car
BTL	age of car
internet	choice
mobilna telefonija	gas station - which
Household	and freq of visit
	goods consumption
	on gas station
Banks and insurance	
most used bank	
all used bank	**bank services** use
bank experinece	bank services use
future used banks	in the future
changing banks	payment of goods
bad experience	credit cards, which,
which you would	**credit**, if, which,
never choose	by which bank

From User Online Trails to …. You!

Analyzing and combining these pieces of information together can lead to **valuable insights** about users and opens up the door to **tremendous opportunities** in sectors including education, health, marketing, law enforcement



The Entity Reconstruction Workflow



Textual data

Text Analysis

Media

Media Analysis

Social Networks

Graph Analysis

Information Integration

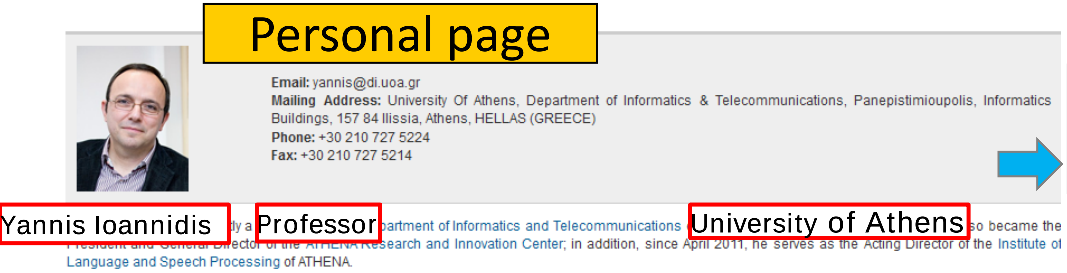
Aggregation

Personalization

Text Analysis

The purpose of this step is to model and structure the information content of textual sources

Personal page



Email: yannis@di.uoa.gr
Mailing Address: University Of Athens, Department of Informatics & Telecommunications, Panepistimioupolis, Informatics Buildings, 157 84 Ilissia, Athens, HELLAS (GREECE)
Phone: +30 210 727 5224
Fax: +30 210 727 5214

Yannis Ioannidis ly a Professor partment of Informatics and Telecommunications University of Athens so became the President and General Director of the ATHENA Research and Innovation Center; in addition, since April 2011, he serves as the Acting Director of the Institute of Language and Speech Processing of ATHENA.

Person entity

Name	Title	Organization
Yannis Ioannidis	Professor	University of Athens

tweet

I need a new digital camera, any recommendations around 300?

Intent to buy

Product	Category	Price
Digital camera	Electronics	~300

review

★★★★★ 5/5/2012

One of best restaurants I have ever been to and I would highly recommend dining here. Lamb off of the rotisserie was unbelievable! The service was outstanding and the atmosphere was perfect.

Sentiment

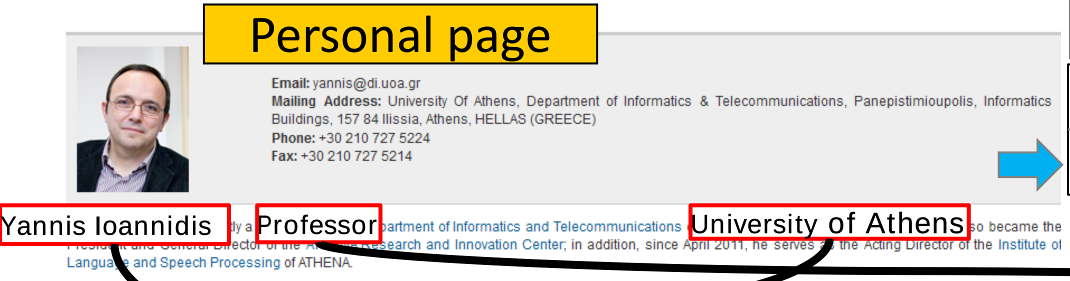
Category	Sentiment	Polarity
Restaurant	best	positive

Aspect	Sentiment	Polarity
Rotisserie	unbelievable	positive
Service	outstanding	positive

Text Analysis Tasks: Information Extraction

Information extraction

the task of automatically extracting **structured information** from unstructured data



Personal page

Email: yannis@di.uoa.gr
Mailing Address: University Of Athens, Department of Informatics & Telecommunications, Panepistimioupolis, Informatics Buildings, 157 84 Ilissia, Athens, HELLAS (GREECE)
Phone: +30 210 727 5224
Fax: +30 210 727 5214

Yannis Ioannidis ly a Professor partment of Informatics and Telecommunications University of Athens so became the
ector of the A esearch and Innovation Center; in addition, since April 2011, he serves the Acting Director of the Institute of
Langua and Speech Processing of ATHENA.

Person entity

Name	Title	Organization
Yannis Ioannidis	Professor	University of Athens

Named entity detection:

recognition of (known) entity names (e.g., people and organizations), places, temporal expressions (e.g., dates)

Relationship extraction:

identification of relations between entities, such as

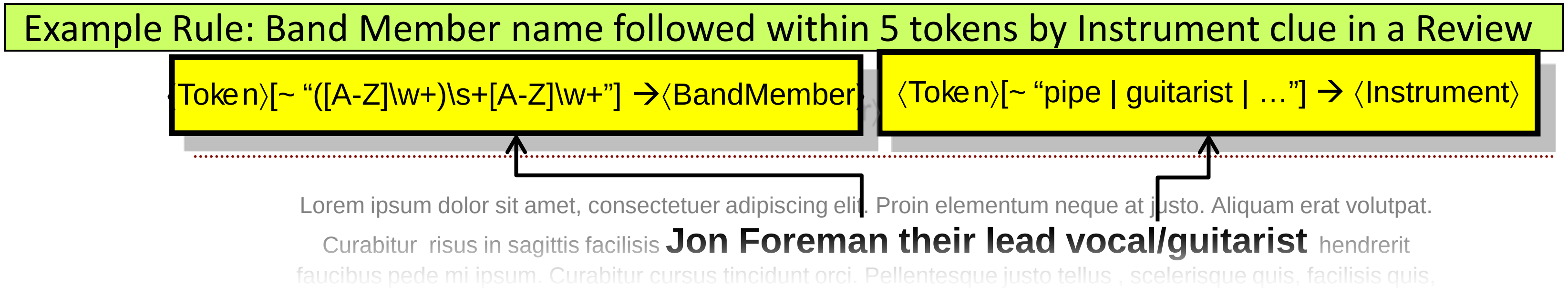
PERSON <works as> COMPANY

Text Analysis Tasks: Information Extraction

Information extraction approaches

Rule-based approaches

e.g., Autoslog, Circus (see [1]) , ANNIE (GATE framework)



Example Rule: Band Member name followed within 5 tokens by Instrument clue in a Review

⟨Token⟩[~ “([A-Z]\w+)\s+[A-Z]\w+”] →⟨BandMember⟩

⟨Token⟩[~ “pipe | guitarist | …”] → ⟨Instrument⟩

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Proin elementum neque at justo. Aliquam erat volutpat. Curabitur risus in sagittis facilisis **Jon Foreman their lead vocal/guitarist** hendrerit faucibus pede mi ipsum. Curabitur cursus tincidunt orci. Pellentesque justo tellus , scelerisque quis, facilisis quis,

[1] Mena B. Habib, Maurice van Keulen Information Extraction, Data Integration, and Uncertain Data Management: The State of The Art. Technical Report

Text Analysis Tasks: Information Extraction

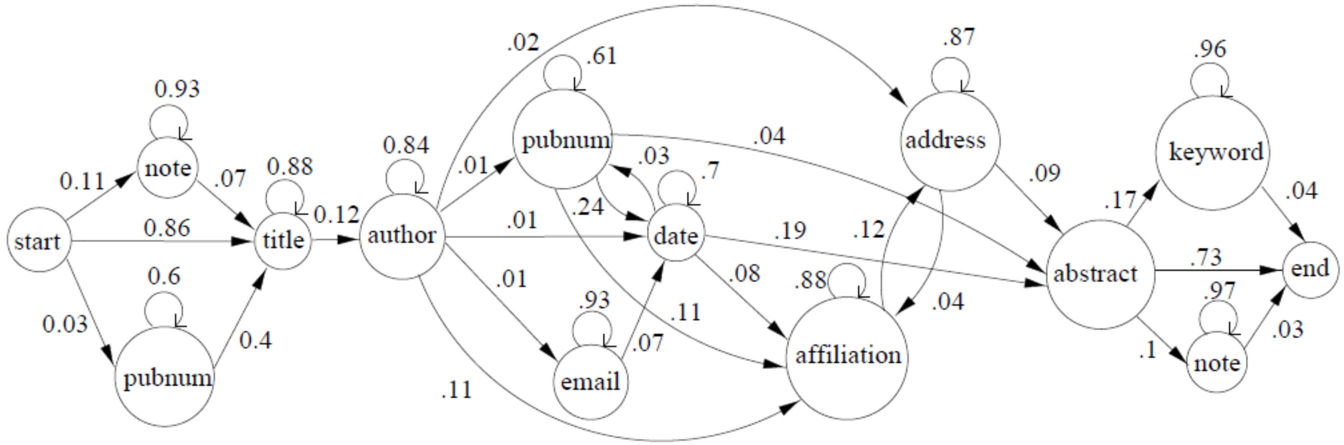
Information extraction approaches

Rule-based approaches

e.g., Autoslog, Circus (see [1]), ANNIE (GATE framework)

Machine Learning approaches

e.g., Rapier [2], SNoW [3], WHISK [4]



[2] Cali, M. E.: Relational learning techniques for natural language information extraction. PhD thesis, University of Texas at Austin, 1998,
[3] Roth, D., Yih, W. T.: Relational learning via propositional algorithms: an information extraction case study. IJCAI, 2001.
[4] Soderland, S.: Learning information extraction rules for semi-structured and free text. Machine Learning, 34 (1999) .

Text Analysis Tasks: Information Extraction

Information extraction approaches

Rule-based approaches

e.g., Autoslog, Circus (see [1]), ANNIE (GATE framework)

Machine Learning approaches

e.g., Rapier [2], SNoW [3], WHISK [4]

Declarative approaches

AQL/SystemT, PSOX, SQoUT, xLog, and RAD

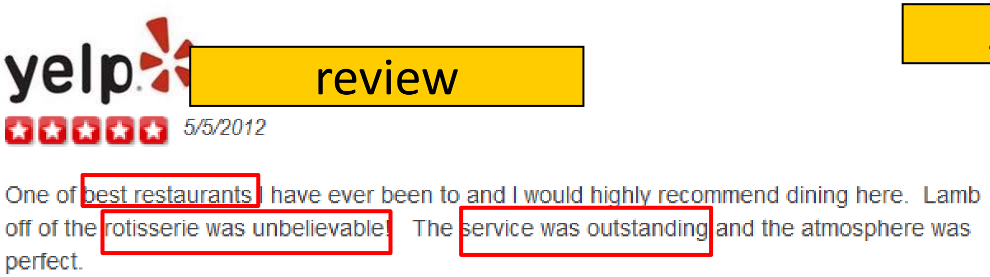
(see SIGMOD Record 37(4), 2010)

Rule-based/Declarative approaches can obtain better precision, but at the cost of lower recall and more work

Text Analysis Tasks: Sentiment Analysis

Sentiment analysis

the task of **determining the attitude** of a speaker or a writer
with respect to some topic or the **overall contextual polarity** of a document.



review

5/5/2012

One of best restaurants I have ever been to and I would highly recommend dining here. Lamb off of the rotisserie was unbelievable! The service was outstanding and the atmosphere was perfect.

Sentiment

Category	Sentiment	Polarity
Restaurant	best	positive

Aspect	Sentiment	Polarity
Rotisserie	unbelievable	positive
Service	outstanding	positive

It is cast as a classification or extraction problem

However, compared to topic/information, sentiment can often be expressed in a more subtle manner, making it difficult to be identified by any of a sentence or document’s terms when considered in isolation.

Text Analysis Tasks: Sentiment Analysis

Examples

"If you are reading this because it is your darling fragrance, please wear it at home exclusively, and tape the windows shut."
(review by Luca Turin and Tania Sanchez of the Givenchy perfume Amarige, in *Perfumes: The Guide, Viking* 2008.)

No ostensibly negative words occur

"This film should be brilliant. It sounds like a great plot, the actors are first grade, and the supporting cast is good as well, and Stallone is attempting to deliver a good performance. However, it can't hold up.

Wishful thinking

Why Text Analysis is hard

You can find only what you are looking for

Fred Flintstone was named CTO of Time Bank Inc. in 2010. The next year he got married and became CEO of Dinosaur Savings.

person	company	position	year in/out
Fred Flintstone	Time Bank Inc.	CTO	2010 in
Fred Flintstone	Time Bank Inc.	CTO	2011 out
Fred Flintstone	Dinosaur Savings	CEO	2011 in

information about his marriage was not captured; extraction seeks to cover only a predefined set of predications.

Why Text Analysis is hard

Variations and Ambiguity

tweet

College: Off to Stanford for my MBA! Bbye chicago!

Luvvvvv my iphn

I got a new iphone. And then I woke up

Typos, abbreviations, short text, sarcasm are just a few of the many issues that make text analysis hard

Why Text Analysis is hard

Scalability (Data)

Twitter: 140 million active users as of 2012, generating over 340 millions tweets daily

Facebook: 300 million photos are uploaded to the site each day.
3.2 billion Likes and Comments are posted daily. [1]

3.146 billion email accounts worldwide. [2]

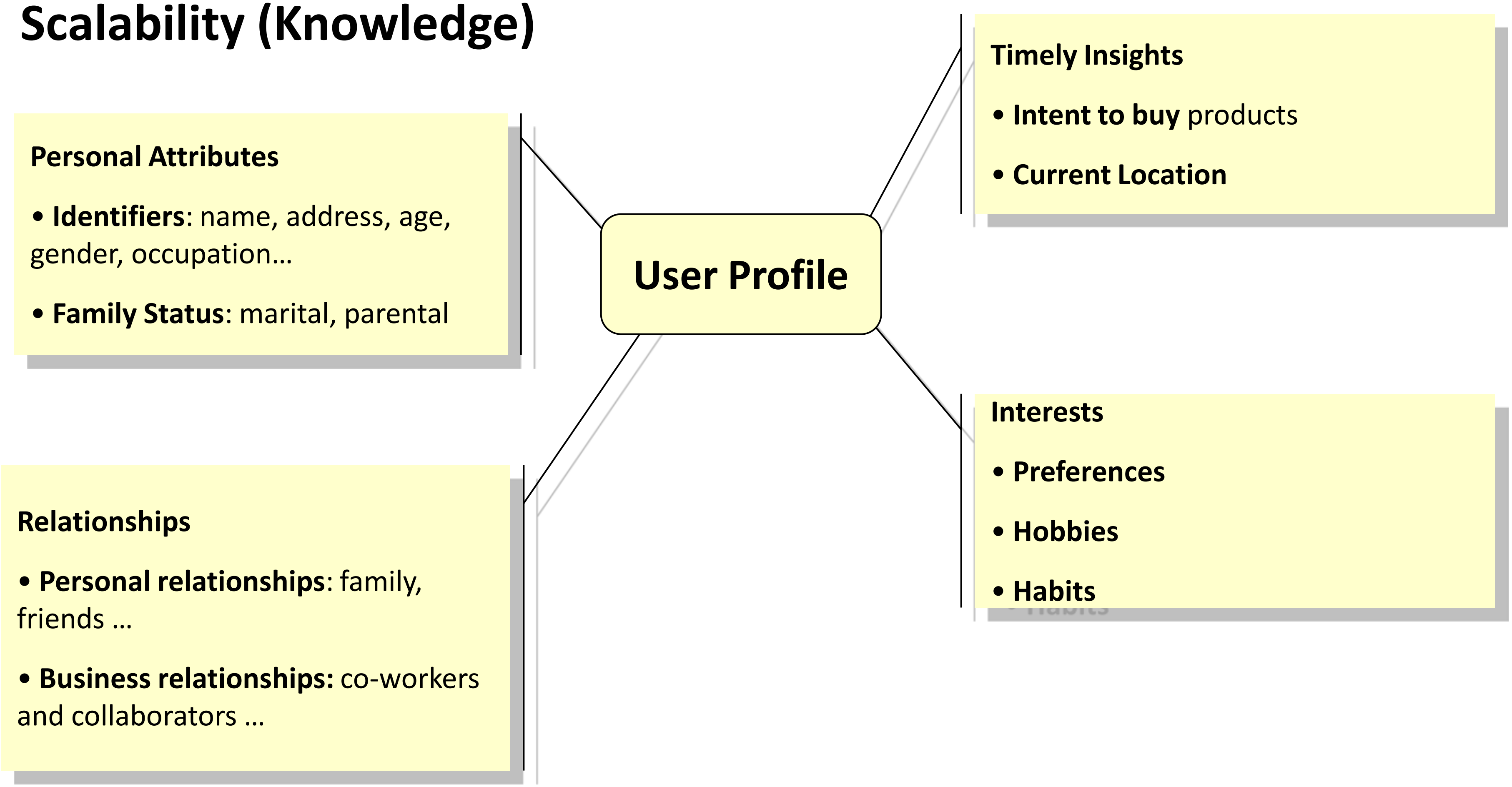
Keeping up with the amount of input data is a challenge

[1] http://www.huffingtonpost.com/2012/04/23/facebook-s-1-amendment_n_1446853.html

[2] http://royal.pingdom.com/2012/01/17/internet-2011-in-numbers/

Why Text Analysis is hard

Scalability (Knowledge)



User Profile

Personal Attributes

- **Identifiers**: name, address, age, gender, occupation…
- **Family Status**: marital, parental

Relationships

- **Personal relationships**: family, friends …
- **Business relationships:** co-workers and collaborators …

Timely Insights

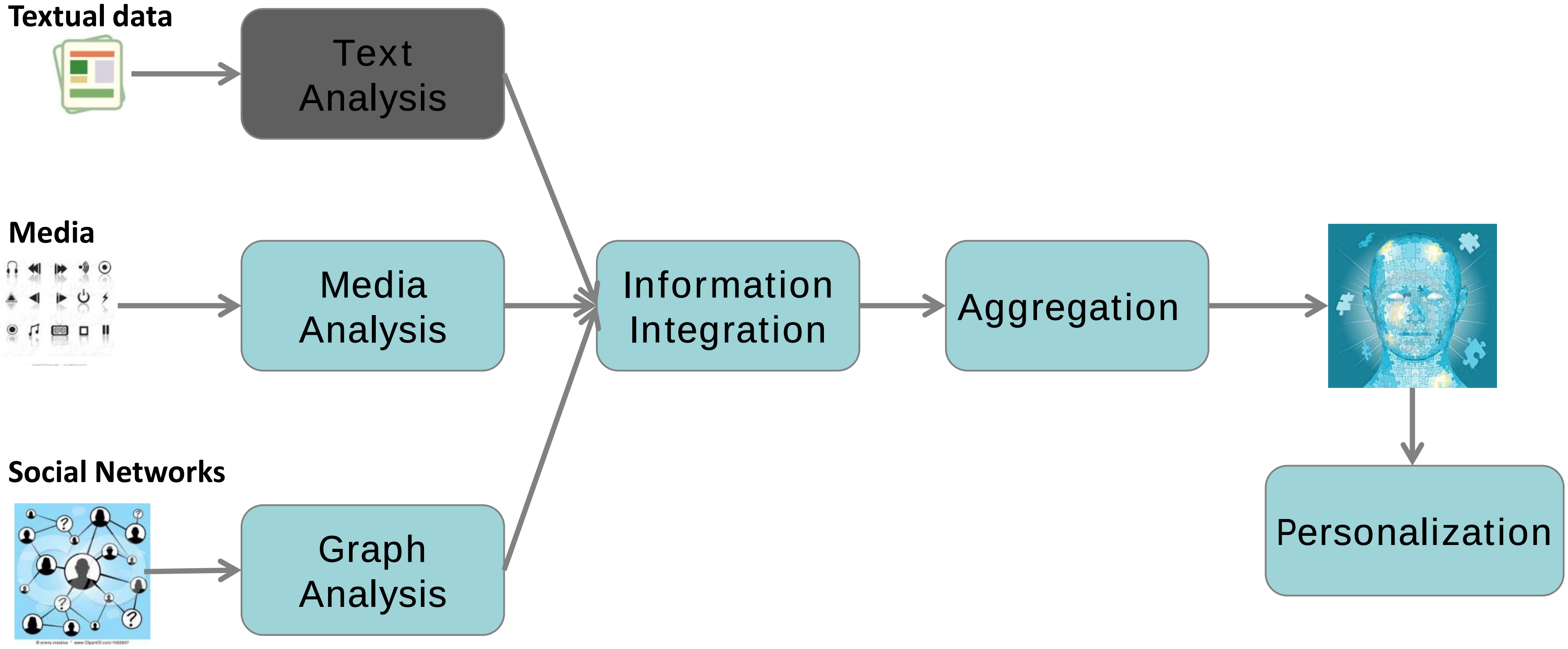
- **Intent to buy** products
- **Current Location**

Interests

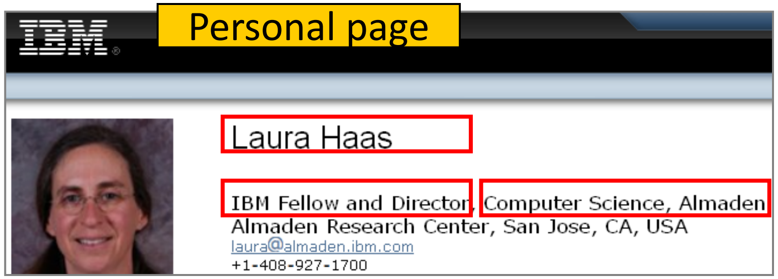
- **Preferences**
- **Hobbies**
- **Habits**

Identifying and keeping up with the types of user knowledge that may be of interest is a challenge

The Entity Reconstruction Workflow



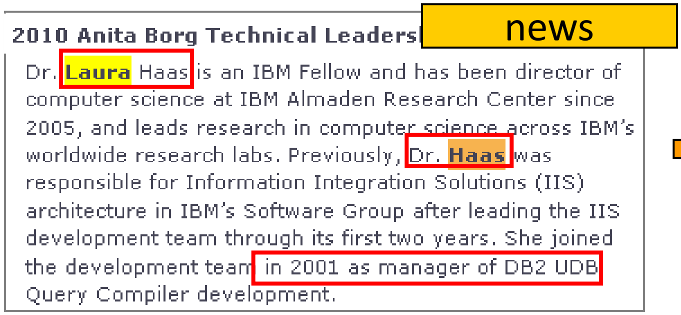
Information Integration



Personal page

Laura Haas

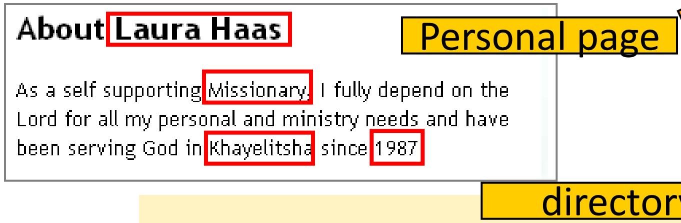
IBM Fellow and Director, Computer Science, Almaden
Almaden Research Center, San Jose, CA, USA
laura@almaden.ibm.com
+1-408-927-1700



news

2010 Anita Borg Technical Leaders

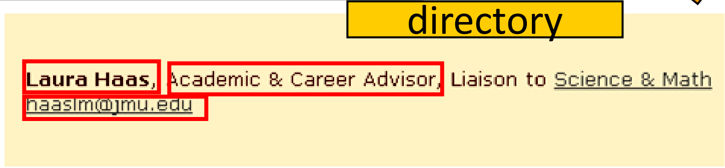
Dr. Laura Haas is an IBM Fellow and has been director of computer science at IBM Almaden Research Center since 2005, and leads research in computer science across IBM's worldwide research labs. Previously, Dr. Haas was responsible for Information Integration Solutions (IIS) architecture in IBM's Software Group after leading the IIS development team through its first two years. She joined the development team in 2001 as manager of DB2 UDB Query Compiler development.



About Laura Haas

Personal page

As a self supporting Missionary, I fully depend on the Lord for all my personal and ministry needs and have been serving God in Khayelitsha since 1987



directory

Laura Haas, Academic & Career Advisor, Liaison to Science & Math
haaslm@jmu.edu

The purpose of this step is to integrate disparate facts about a single entity

Information Integration

Laura Haas	IBM	Director (today-2005) DB2 UDB Manager (2005-2001)
Laura Haas	Khayelitsa	Missionary (today-1987)
Laura Haas	JMU	Academic & Career Advisor

Do these documents refer to the same person ?

- Variability in the person's name
- Lack of a key identifier
- Supporting attributes vary depending on the context
- Multiple (approximate) ways to resolve mentions

Information Integration Tasks: Entity Resolution

Entity Resolution

The problem of **linking facts** that refer to **the same entity** when **integrating** two or more disparate sources.

ER is a complex, trial-and-error process

- It requires domain-specific knowledge
- It is hard to achieve high precision and recall

Laura Haas	IBM Research
L. Haas	IBM Almaden
L. Haas	Computer Science, IBM Almaden

Information Integration Tasks: Entity Resolution

Entity Resolution Approaches

Algorithms and Metrics

e.g., Jaro, edit distance, multi-attribute similarity measures (e.g., [1,2])

Tailor, iFuice (see [3])

Declarative approaches

e.g., WHIRL, Dedupalog, LinQL (see [3])

[1] A. K. Elmagarmid, P. G. Ipeirotis, and V. S. Verykios, "Duplicate Record Detection: A Survey," IEEE TKDE, vol. 19, no. 1, pp. 1–16, 2007.

[2] I. P. Fellegi and A. B. Sunter, "A Theory for Record Linkage," J. Am.Statistical Assoc., vol. 64, no. 328, pp. 1183–1210, 2007.

[3] Hanna Köpcke, Erhard Rahm: Frameworks for entity matching: A comparison. Data Knowl. Eng. 69(2): 197-210 (2010)

Why Information Integration is hard

Information on the Web may be

-incomplete and in variations

e.g., EDBT 2012 web site:

"Adaptive Indexing in Modern Databases"
Stratos Idreos (CWI, The Netherlands); Stefan Manegold (CWI, The Netherlands);
Goetz Graefe (HP Labs, Palo Alto)

Intention Insider: Discovering People's Intentions in the Social Channel
Malu Castellanos, HP Labs, USA; ….

Session Chair: Ronald Fagin (IBM Research - Almaden)

Adaptive MapReduce using Situation-Aware Mappers:
Rares Vernica (HP Labs), Andrey Balmin (IBM Almaden),
Kevin Beyer (IBM Almaden Research Center), ….

Why Information Integration is hard

Information on the Web may be

- intentionally faked

  e.g., a small experiment in Twitter: almost half of the times, the combination name/city/state did not retrieve any person from peoplefinder.com

- bogus or ambiguous

  e.g., "user location in Twitter": "wish I were in California"

Little or untrustworthy evidence hinders information integration

Why Information Integration is hard

Handling Conflicts within and across sources

- Each attribute has specific semantics for integration

Name	Title	Organization
Yannis Ioannidis	Professor	University of Athens

Name	Title	Organization
Y. Ioannidis	General Director	ATHENA RC

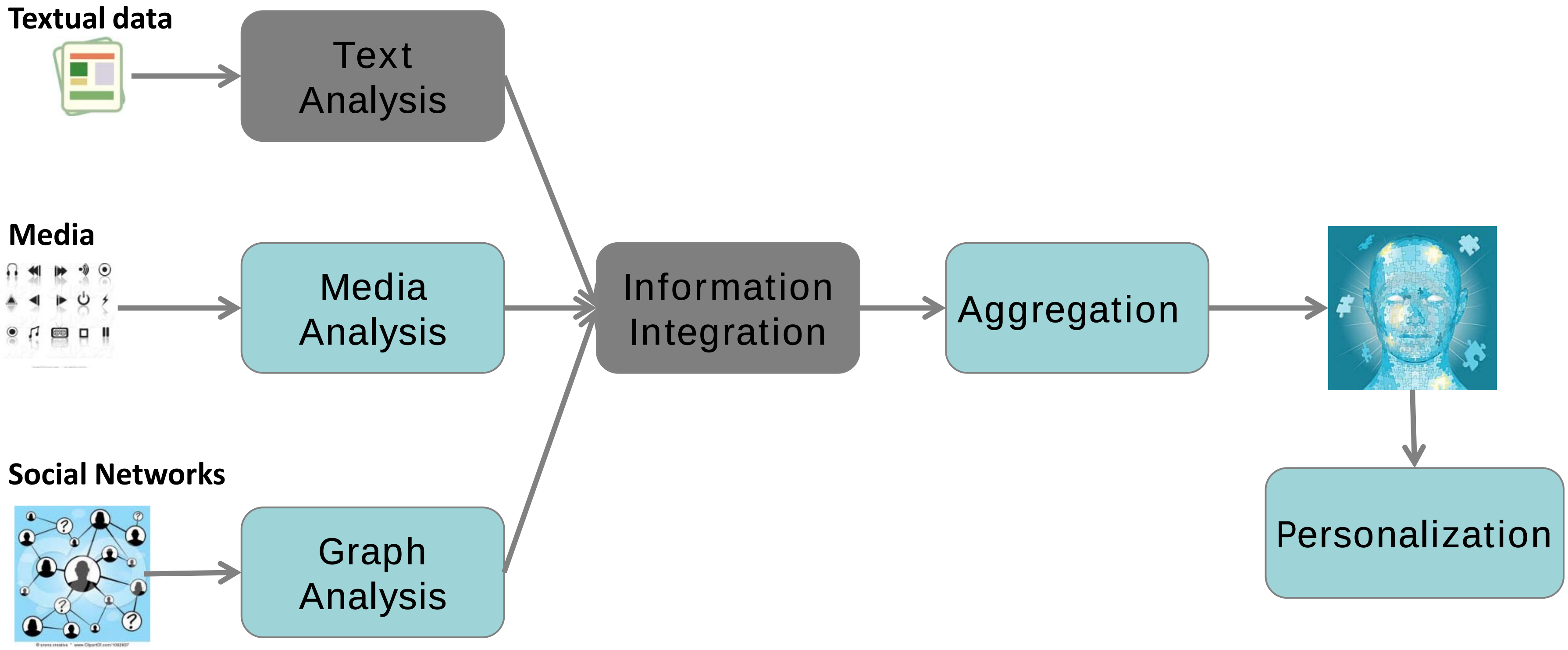
Is this a conflict ?

Name	genre	occupation
Cameron Black	female	CEO

Name	genre	occupation
Cameron Black	male	CEO

how to integrate conflicting gender?

The Entity Reconstruction Workflow



Textual data

Text Analysis

Media

Media Analysis

Social Networks

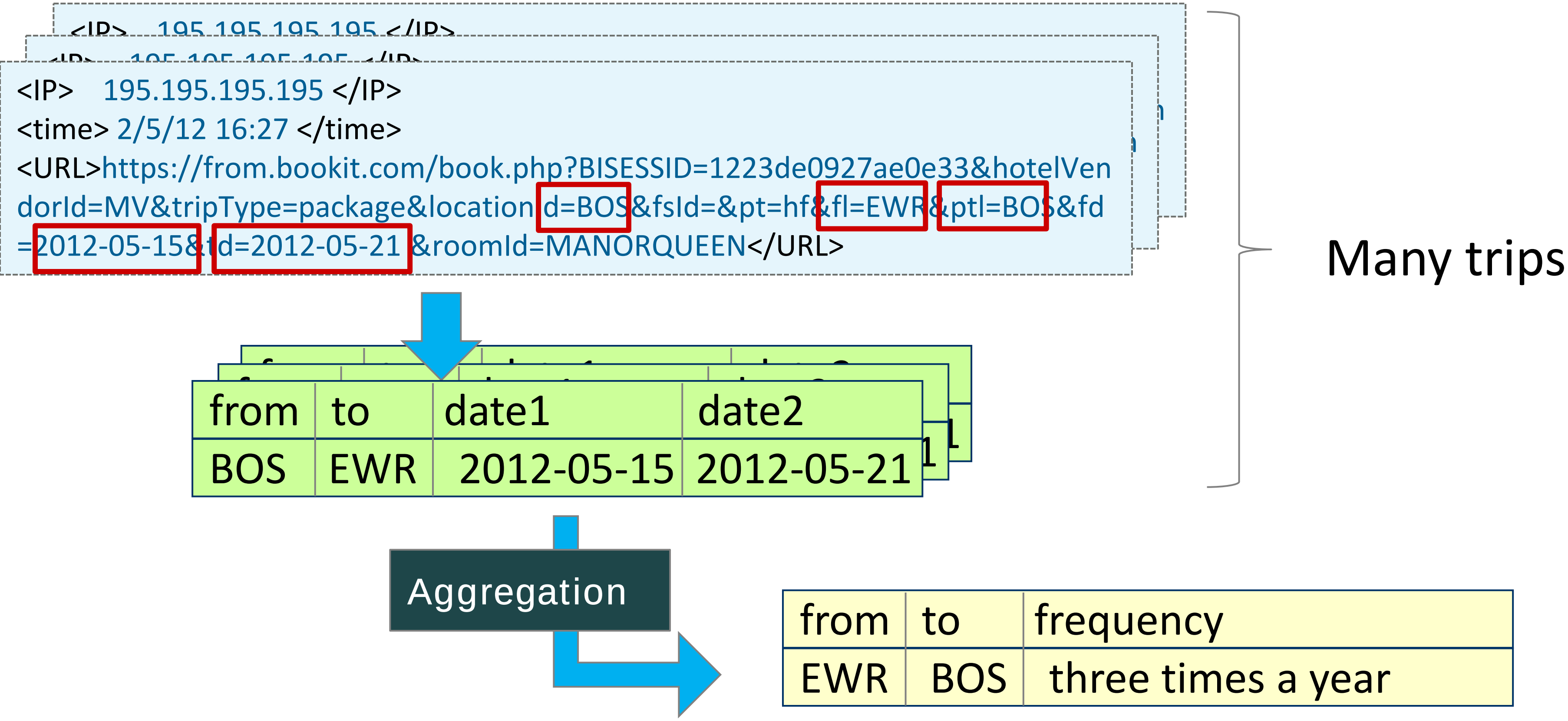
Graph Analysis

Information Integration

Aggregation

Personalization

Aggregation



<IP> 195.195.195.195 </IP>
<time> 2/5/12 16:27 </time>
<URL>https://from.bookit.com/book.php?BISESSID=1223de0927ae0e33&hotelVendorId=MV&tripType=package&locationId=BOS&fsId=&pt=hf&fl=EWR&ptl=BOS&fd=2012-05-15&td=2012-05-21&roomId=MANORQUEEN</URL>

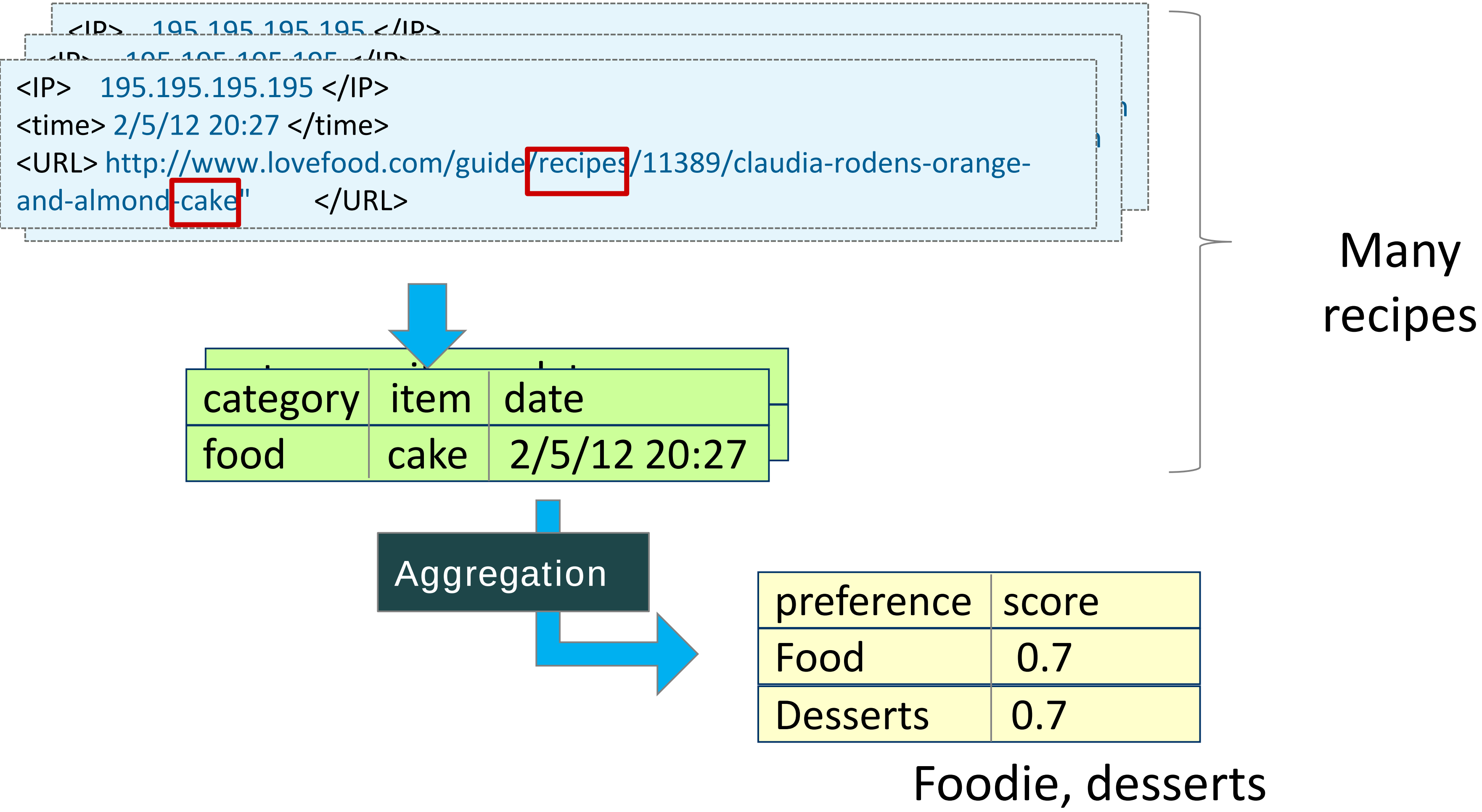
from	to	date1	date2
BOS	EWR	2012-05-15	2012-05-21

Many trips

from	to	frequency
EWR	BOS	three times a year

Frequent flyer to Boston

Aggregation



<IP> 195.195.195.195 </IP>
<time> 2/5/12 20:27 </time>
<URL> http://www.lovefood.com/guide/recipes/11389/claudia-rodens-orange-and-almond-cake" </URL>

category	item	date
food	cake	2/5/12 20:27

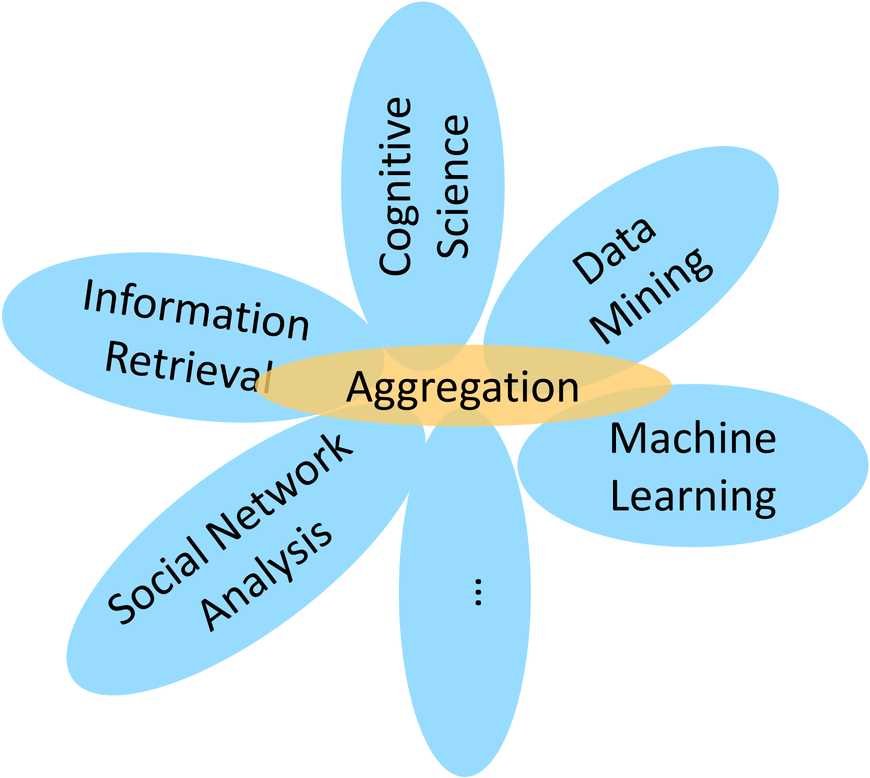
Aggregation

preference	score
Food	0.7
Desserts	0.7

Foodie, desserts

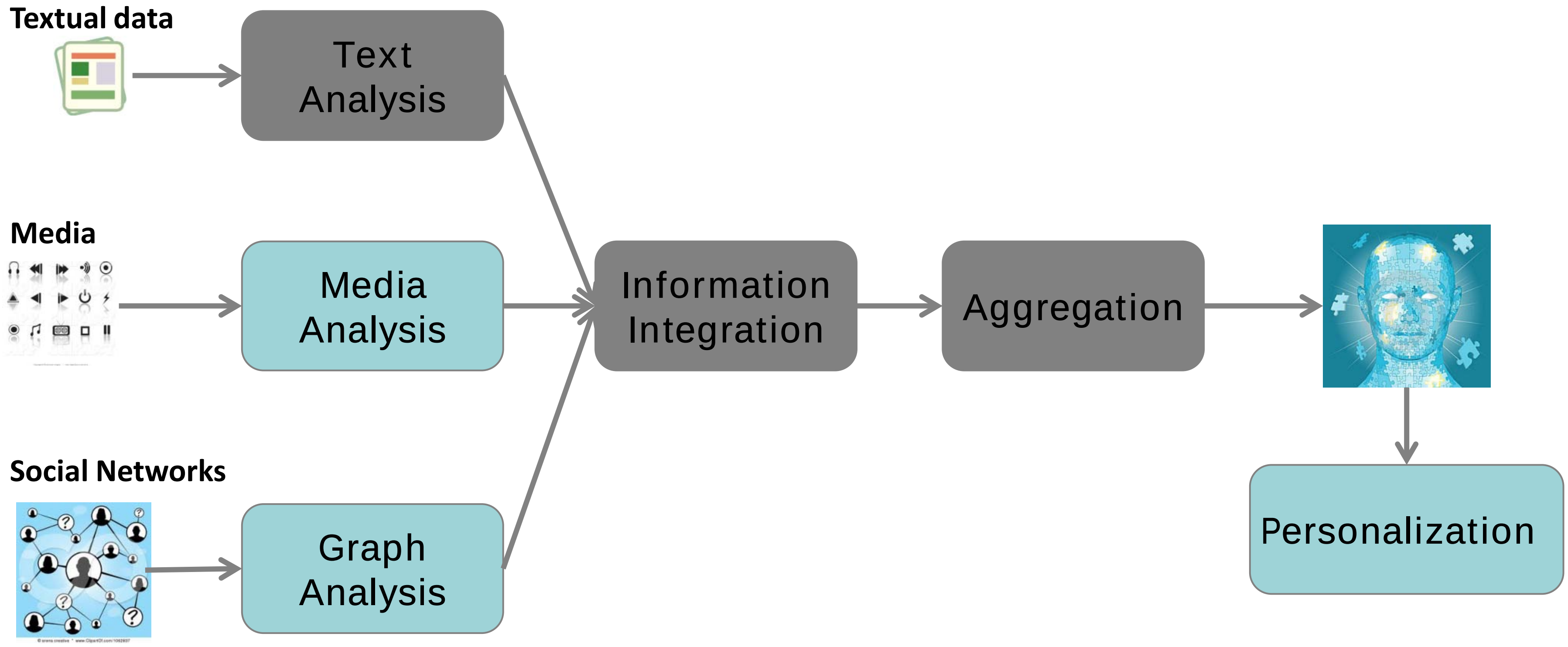
Many recipes

Aggregation Techniques



The more data collected about a person
the more things we could learn about this person!

The Entity Reconstruction Workflow



Textual data

Text Analysis

Media

Media Analysis

Social Networks

Graph Analysis

Information Integration

Aggregation

Personalization

Personalization

Product Recommendations

Content Delivery

Targeted Advertisements

Personalized Services

Personalization

http://www.youtube.com/watch?v=RNJl9EEcsoE

The Findability Challenge

- Heterogeneity
- Distributed Content
- Incompleteness
- Timeliness
- Privacy

Thank you!