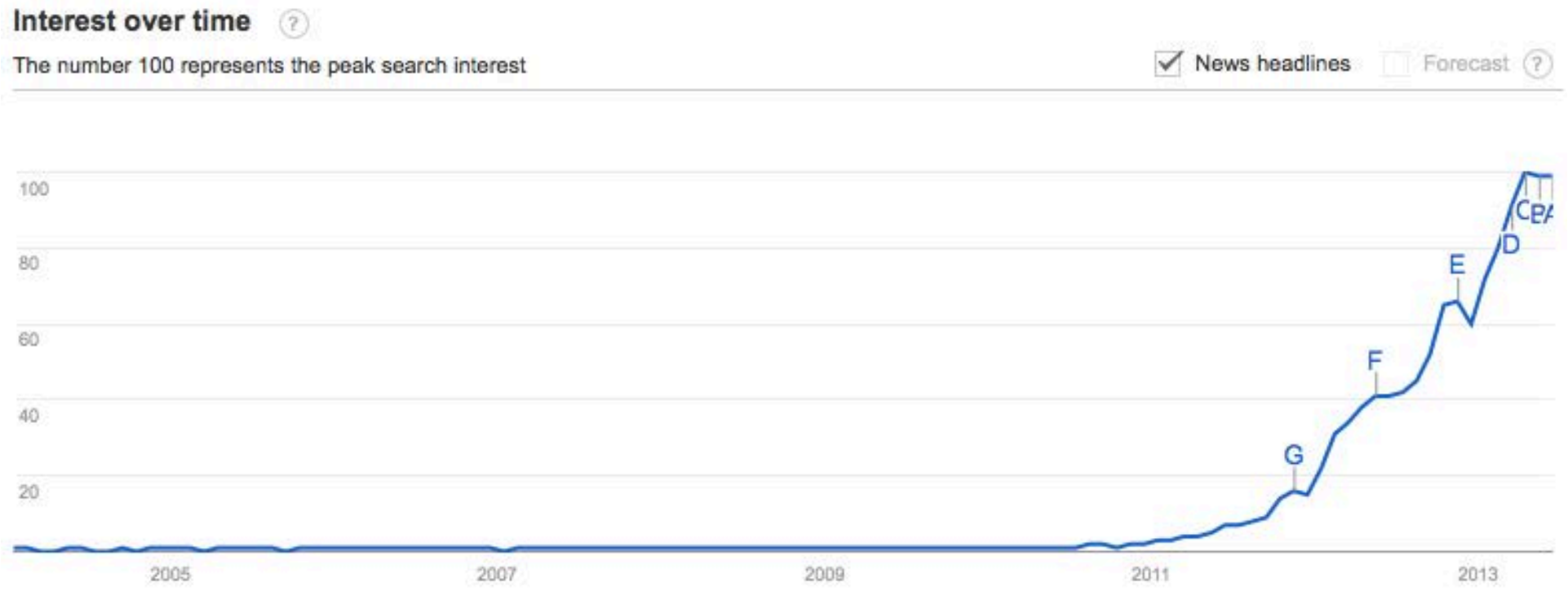
Big Data Analysis and Integration

Juliana Freire
juliana.freire@nyu.edu
Visualization and Data Analysis (ViDA) Center
http://bigdata.poly.edu
NYU Poly

NEW YORK UNIVERSITY

Big Data: What is the **Big** deal?



http://www.google.com/trends/explore#q=%22big%20data%22

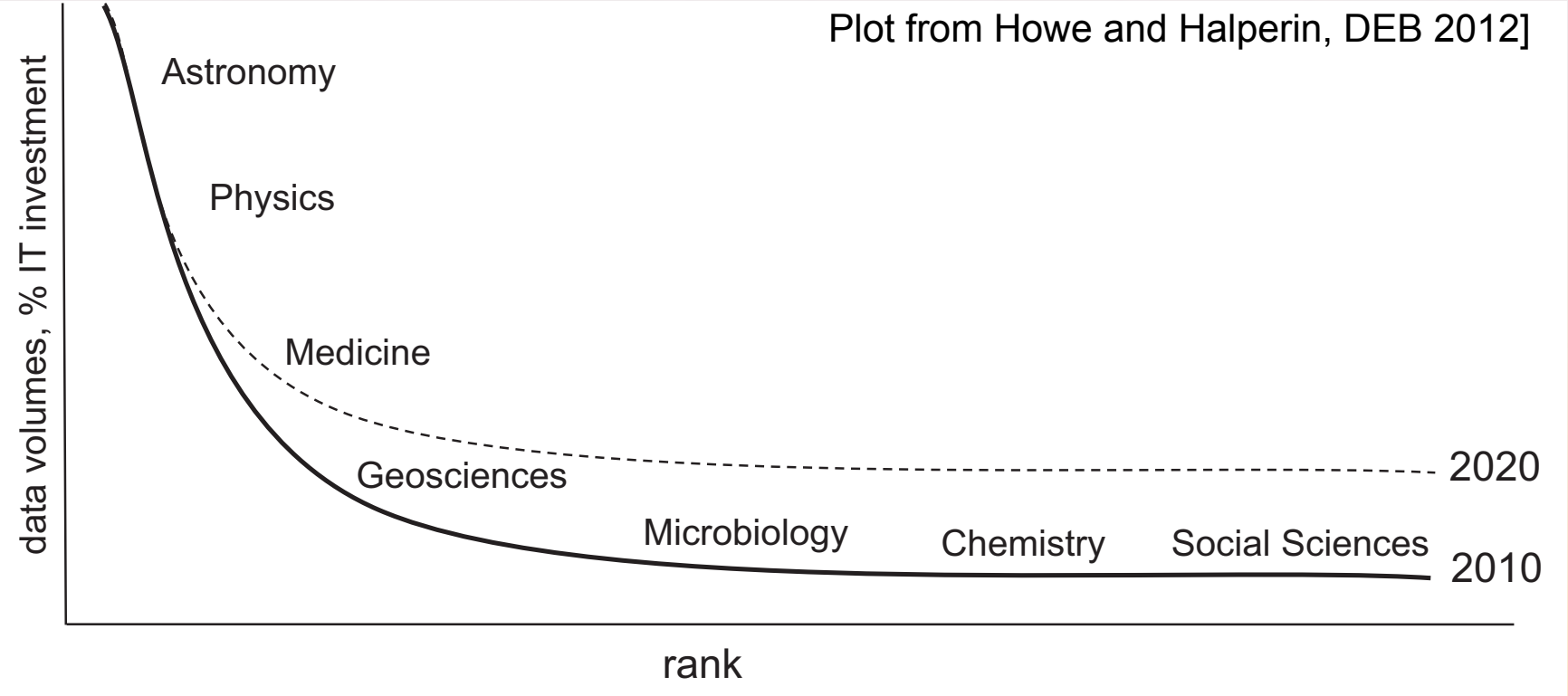
Big Data: What is the **Big** deal?

- Smart Cities: 50% of the world population lives in cities
  - Census, crime, emergency visits, taxis, public transportation, real estate, noise, energy, …
  - Make cities more efficient and sustainable, and improve the lives of their citizens http://cusp.nyu.edu/
  - Success stories: Mike Flowers and NYC inspections
- Enable scientific discoveries: science is now data rich
  - Petabytes of data generated each day, e.g., Australian radio telescopes, Large Hadron Collider, climate data, …
  - Social data, e.g., Facebook, Twitter (~~2,380,000~~ 3,180,000 and ~~2,880,000~~ 3,410,000 results in Google Scholar!)
- Data is currency: companies profit from Big Data
  - Better understand customers, targeted advertising, …

Big Data: What is the **Big** deal?

- Big data is not new: financial transactions, call detail records, astronomy, …
- What is new:

- Many more *data enthusiasts*



Big Data: What is the **Big** deal?

- Big data is not new: financial transactions, call detail records, astronomy, …
- What is new:

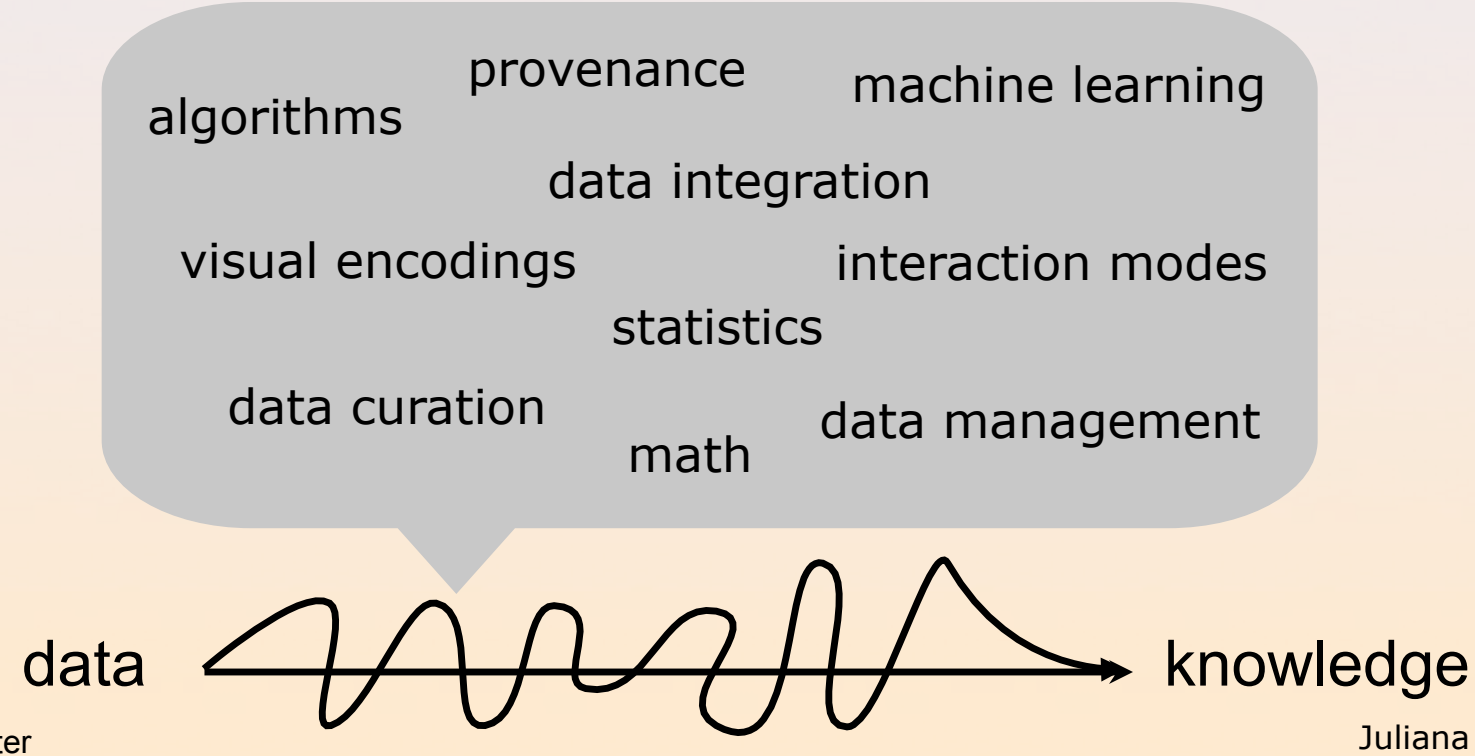
- Many more *data enthusiasts*

- More data are widely available, e.g., Web, data.gov, scientific data, social and urban data

- Computing is cheap and easy to access
  - Server with 64 cores, 512GB RAM ~$11k
  - Cluster with 1000 cores ~$150k
  - Pay as you go: Amazon EC2

Big Data: What is hard?

- Scalability for computations? NOT!
  - Lots of work on distributed systems, parallel databases, …
  - Elasticity: Add more nodes!
- Scalability for people: Data integration and exploration is hard regardless of whether data are big or small



(Big) Data Exploration: Desiderata

- Tools and techniques that aid people find, integrate, and explore data
- *Automate* as much as possible tedious tasks
- Enable data enthusiasts/experts analyze their data
- **Usability** is a **Big** issue
- Key ingredients (that we work on)
  - Data integration
  - Visualization and visual analytics
  - Data and provenance management

(Big) Data Analysis Pipeline

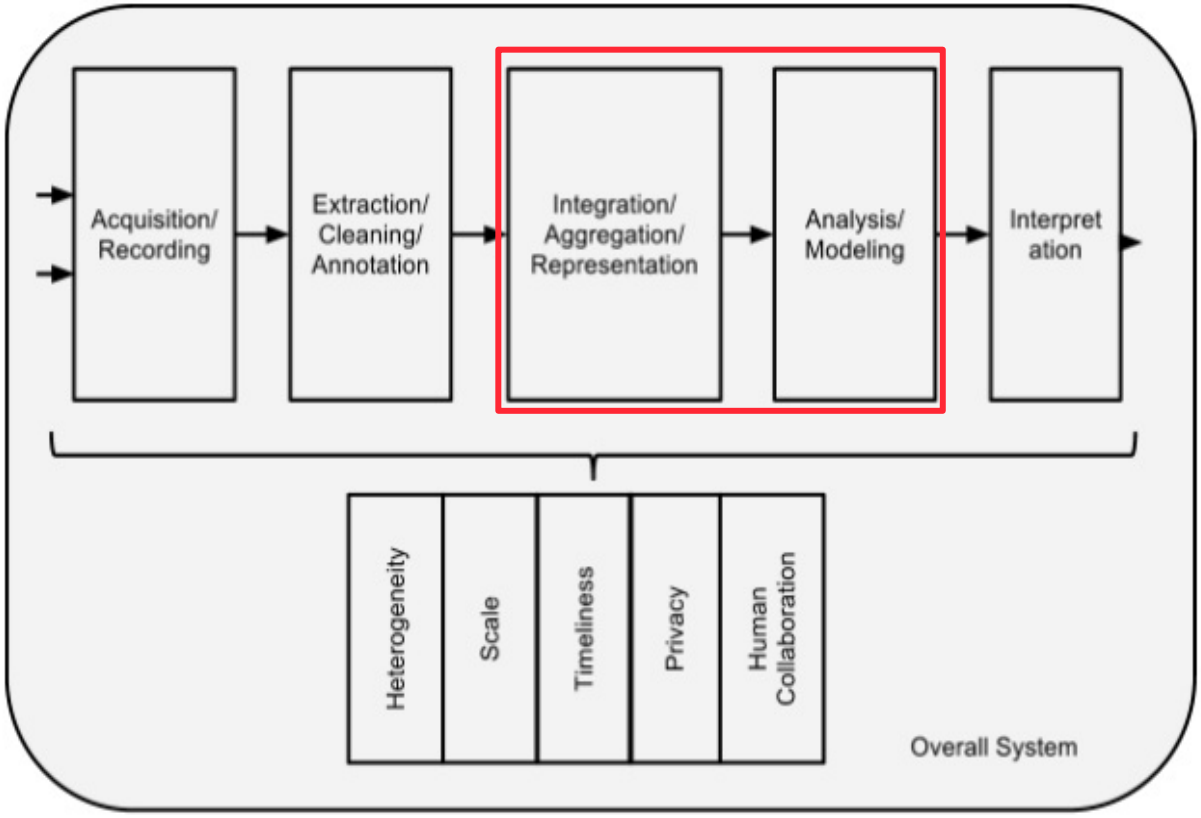
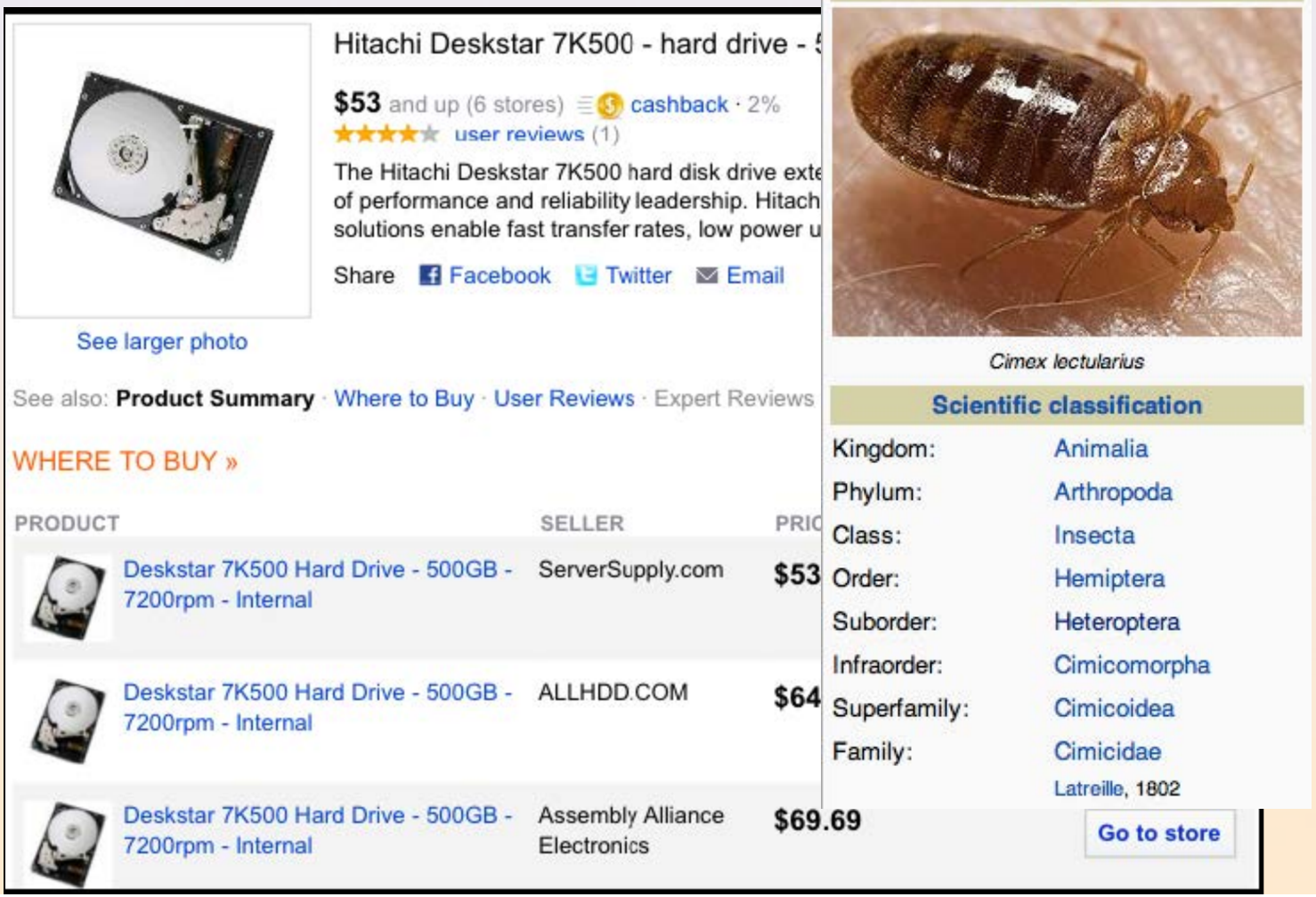


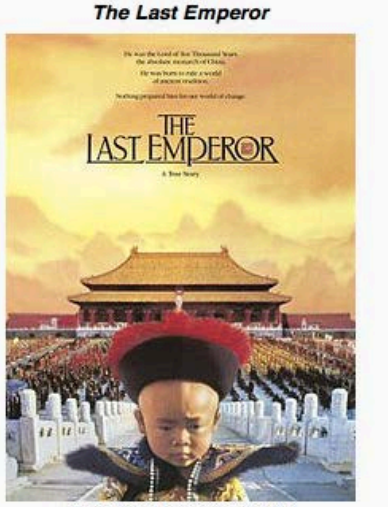
Figure 1: The Big Data Analysis Pipeline. Major steps in analysis of big data are shown in the flow at top. Below it are big data needs that make these tasks challenging.

http://cra.org/ccc/docs/init/bigdatawhitepaper.pdf

Structured Data Everywhere

- Millions of online databases [Madhavan, CIDR 2007]

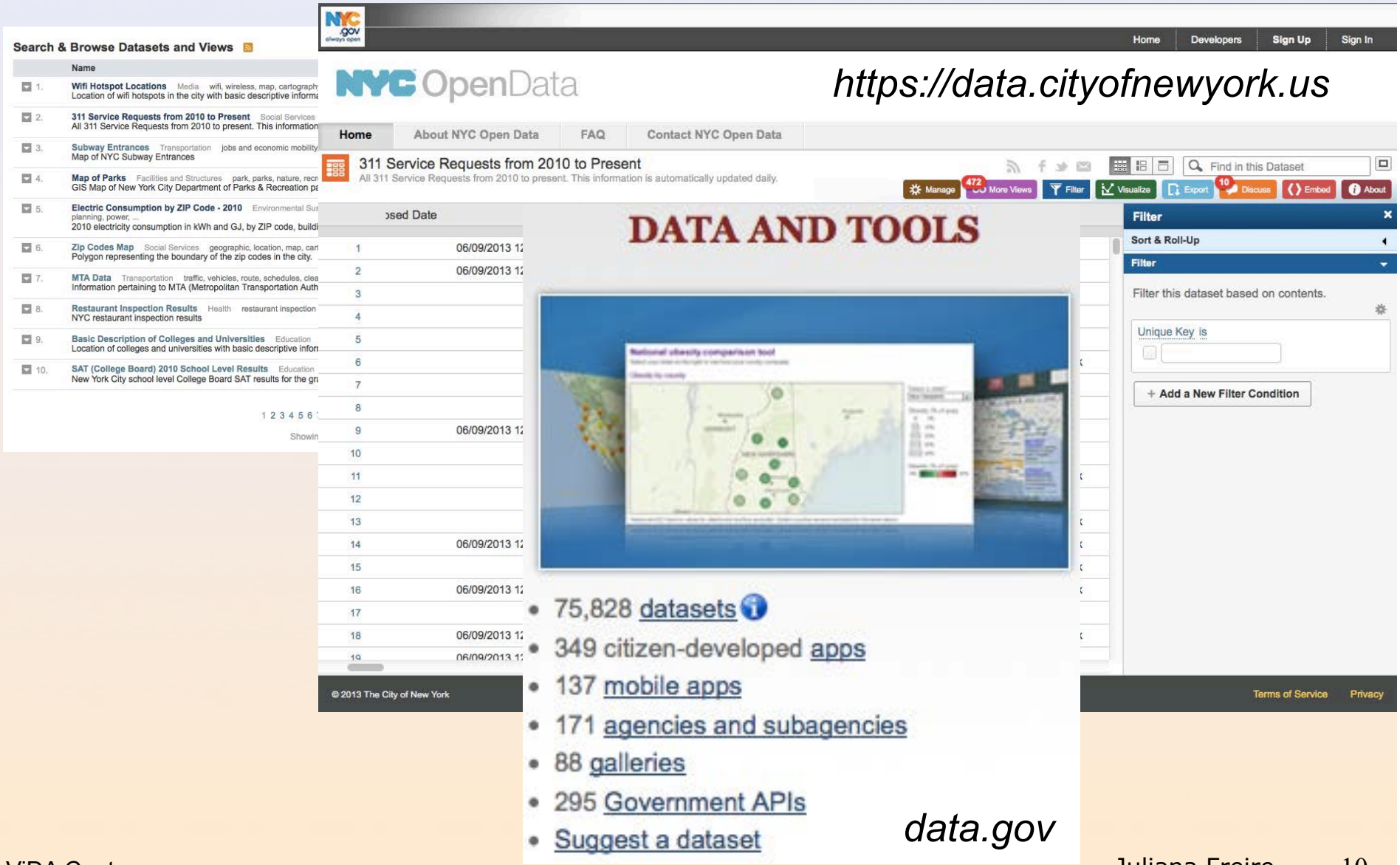




Promotional poster for the film

Directed by	Bernardo Bertolucci
Produced by	Jeremy Thomas
Written by	Mark Peploe Bernardo Bertolucci
Starring	John Lone Joan Chen Peter O'Toole Ruocheng Ying Victor Wong Dennis Dun Ryuichi Sakamoto Maggie Han Ric Young Vivian Wu Chen Kaige
Music by	Ryuichi Sakamoto David Byrne Cong Su
Cinematography	Vittorio Storaro
Editing by	Gabriella Cristiani
Studio	Recorded Picture Company
Distributed by	Columbia Pictures
Release date(s)	23 October 1987 (Italy) 18 November 1987 (New York City, New York Premire) 19 November 1987 (Los Angeles, California Premiere) 18 December 1987 (USA)
Running time	160 minutes
Country	China

Structured Data Everywhere



Information Integration: Challenges

- Information integration is **hard**, even at a small scale
- One notable example:

New York City gets 25,000 illegal-conversion complaints a year, but it has only 200 inspectors to handle them.

Flowers' group integrated information from 19 different agencies that provided indication of issues in buildings

Result: hit rate for inspections went from 13% to 70%

Integration took several months…

Todd Heisler/The New York Times

Michael Flowers, right, oversees a small group of tech-savvy and civic-minded statisticians working across from City Hall.

Enlarge This Image

Todd Heisler/The New York Times

"All we do," Mr. Flowers said, is "process massive amounts of information and use it to do things more effectively."

Information Integration: Challenges

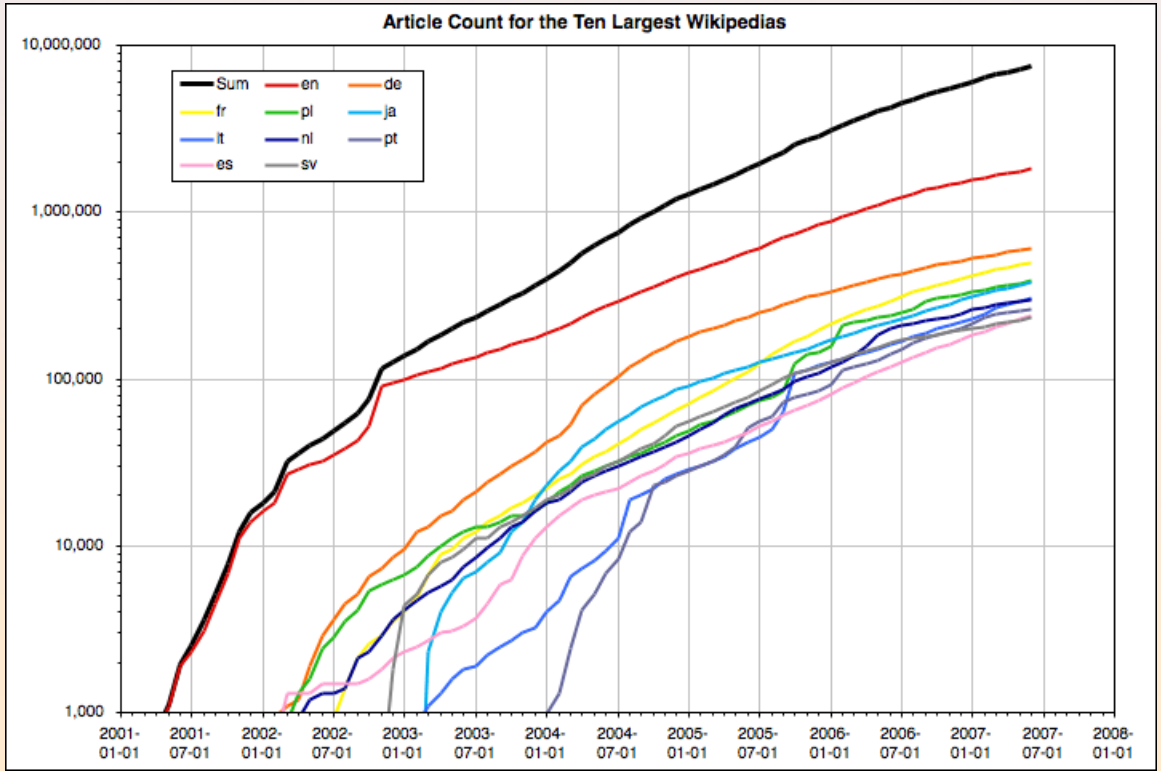
- Information integration is **hard**, even at a small scale
- 'Big data' is harder…
  - Large, heterogeneous and noisy data
  - Great variation in both the structure and how values are represented
- 'Big data' is easier…
  - Lots of examples
  - Many potential sources of similarity
- Need scalable and usable approaches

Big Data Integration Problems and Solutions

- Synthesizing products for online catalogs [Nguyen et al., VLDB 2011]
  - 800k offers, 1000 merchants, 400 product categories
- Integrating online databases [Nguyen et al., CIKM 2010]
  - 4,500 web forms, 33,000 form elements
- Matching multi-lingual Wikipedia infoboxes [Nguyen et al., VLDB 2012]
  - ~9,000 infoboxes
- Integrating NYC data
  - Still looking for a solution ☺

Wikipedia and Multilingualism

- There are articles in over 270 languages!
- A disproportionate number of Wikipedia documents are in English and out of reach for many people
  - 328M EN speakers, EN Wikipedia 20%
  - 178M PT speakers, PT Wikipedia 3.7%



Wikipedia and Multilingualism

- There are articles in over 270 languages!
- A disproportionate number of Wikipedia documents are in English and out of reach for many people
  - 328M EN speakers, EN Wikipedia 20%
  - 178M PT speakers, PT Wikipedia 3.7%
- Important to support multilingual queries – give users access to a larger segment of Wikipedia
- Enrich Wikipedia by integrating information in different languages

Querying Wikipedia in Multiple Languages

Find the ***genre*** and ***studio*** that produced the film "***The Last Emperor***"

O Último Imperador

O Último Imperador (PT/BR)

Reino Unido / Itália
França / China
1987 • cor • 165 min

Produção

Direção	Bernardo Bertolucci
Roteiro	Mark Peploe / Bernardo Bertolucci
Elenco original	John Lone Joan Chen Peter O'Toole Ryuichi Sakamoto
Género	drama biográfico / épico
Idioma original	inglês / mandarim / japonês

IMDb: (inglês) (português)

Projeto Cinema • Portal Cinema

Directed by	Bernardo Bertolucci
Produced by	Jeremy Thomas
Written by	Mark Peploe Bernardo Bertolucci
Starring	John Lone Joan Chen Peter O'Toole
Music by	Ryuichi Sakamoto
Cinematography	Vittorio Storaro
Editing by	Gabriella Cristiani
Studio	Hemdale Film
Distributed by	Columbia Pictures
Release date(s)	**United States:** November 18, 1987
Running time	160 minutes (theatrical) 218 minutes (television)
Country	China Italy United Kingdom France
Language	English Mandarin Chinese
Budget	$23.8 million[1]

Multilingual Wikipedia Integration: Challenges

- Goal: Identify correspondences between attributes
- Using dictionaries and translation is not sufficient:

  starring – *elenco original vs estrelando*

- WordNet is *incomplete for* many languages

O Último Imperador	
O Último Imperador (PT/BR)	
Reino Unido / Itália França / China	
1987 • cor • 165 min	
Produção	
Direção	Bernardo Bertolucci
Roteiro	Mark Peploe / Bernardo Bertolucci
Elenco original	John Lone Joan Chen Peter O'Toole Ryuichi Sakamoto
Género	drama biográfico / épico
Idioma original	inglês / mandarim / japonês
IMDb: (inglês) (português)	
Projeto Cinema • Portal Cinema	

Directed by	Bernardo Bertolucci
Produced by	Jeremy Thomas
Written by	Mark Peploe Bernardo Bertolucci
Starring	John Lone Joan Chen Peter O'Toole
Music by	Ryuichi Sakamoto
Cinematography	Vittorio Storaro
Editing by	Gabriella Cristiani
Studio	Hemdale Film
Distributed by	Columbia Pictures
Release date(s)	United States: November 18, 1987
Running time	160 minutes (theatrical) 218 minutes (television)
Country	China Italy United Kingdom France
Language	English Mandarin Chinese
Budget	$23.8 million[1]

- Infoboxes across languages are not comparable – overlap can be small
- Label similarity can be misleading: e.g., editor – editora
- Attribute values are heterogeneous and sometimes inconsistent, e.g., is the running time 160 or 165 minutes?

Related Work

- Cross-language infobox alignment:
  - [Adar et al., 2009]: train a classifier to identify cross-language infobox alignments (English, German, French and Spanish)

    Require training data – which may not be available for under-represented languages
  - Bouma et al., 2009: rely on identical values or on the existence of a cross-language path between values (English and Dutch)

    High precision, low recall
  - Effective only for to languages that are morphologically similar
- Cross-language ontology alignment
  - [Fu et al. and Santos et al.]: Machine translation + monolingual ontology matching algorithms
  - Well-defined and clean schema – Wikipedia infoboxes are heterogeneous and loosely defined
  - Do not take values into account

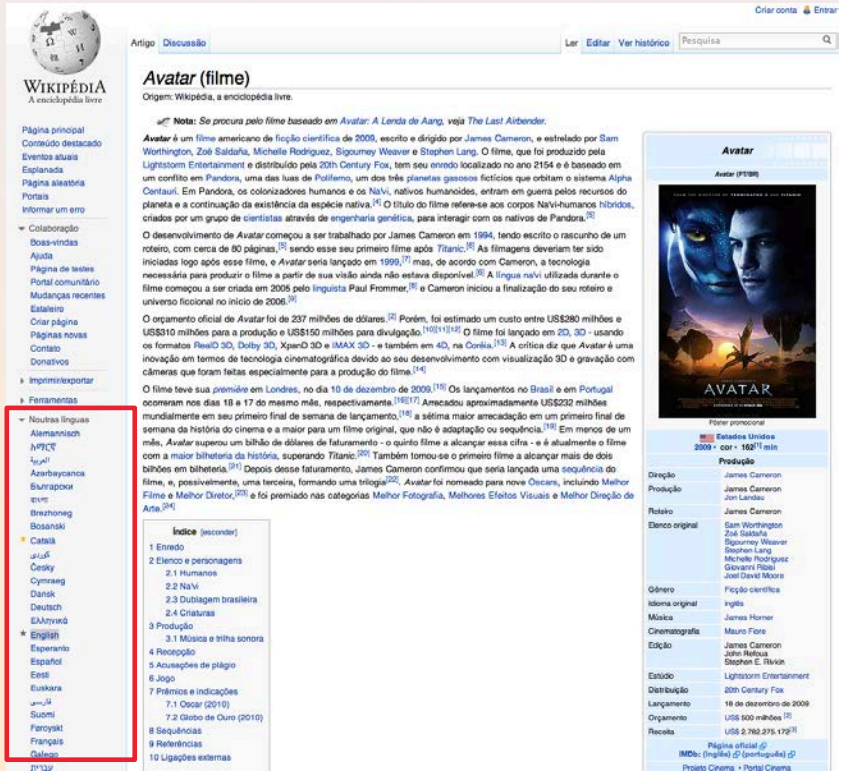
Our Approach: WikiMatch [Nguyen et al., VLDB 2012]

- Group infoboxes and attributes *
- Combine similarity information from multiple sources:
  - Attribute correlation *
  - Value similarity
  - Link structure
- Apply a multi-step approach to minimize error propagation and to increase recall *
  - Prioritize high-confidence correspondences
- Benefits:
  - *No need for external resources* such as bilingual dictionaries, thesauri, ontologies, or automatic translator
  - *No need for training* *

Big Data considerations

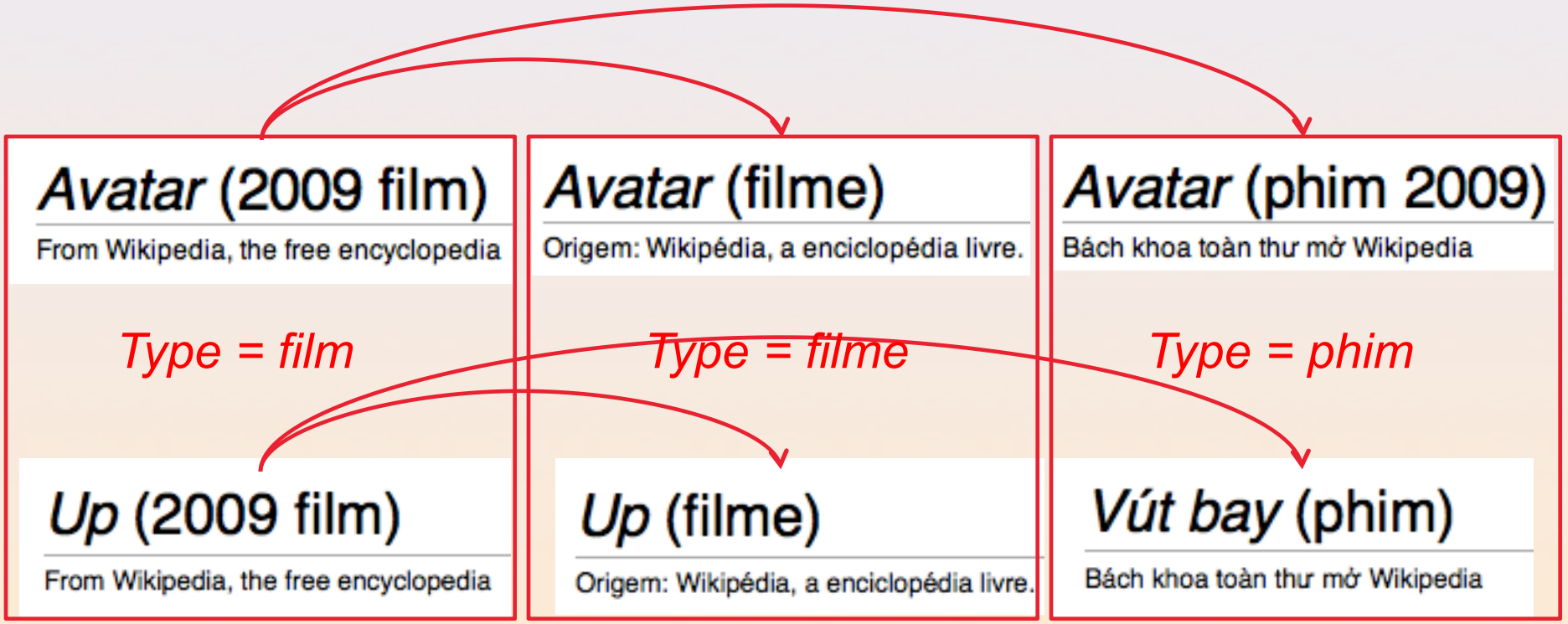
Matching *Entity Types* across Languages

- Group infoboxes based on their types *[Nguyen et al., CIKM2012]*
- Use cross-language links to cluster infoboxes across languages
- Intuition: If a set of infoboxes belonging to entity type T often link to infoboxes in a different language of type T', then it is likely that types T and T' are equivalent



Matching *Entity Types* across Languages

Type(film) = Type(filme) = Type(phim)



Computing Cross-Language Similarity

- Comparing pairs of infoboxes is not effective – too much heterogeneity
- Leverage the large number of infoboxes to build a *super-schema* for each type: *Given a type T, create schema S_T where each attribute a in S_T is associated with a set v of values that occur in infoboxes of type T for attribute a*

- **Problem:** Given two super-schemata S_T and S'_T for a type T, in languages L and L' respectively, our goal is to identify correspondences between attributes in these schemata
- *Our approach*: Combine similarity for different components of the schemata – *link structure, value, correlation*

Cross-Language ***Value*** Similarity

- Given attributes a_1 and a_2 in languages L and L' respectively: $vsim(a_1,a_2) = cos(v_1,v_2)$
- But values are represented differently in different languages, resulting in low value similarity

$v_{nascimento}$ ={1963:1, Irlanda:1, 18 de Dezembro 1950:1, Estados Unidos:2}

v_{born} ={1963:1, Ireland:1, June 4 1975:1, United States: 3}

- Automatically create a dictionary from language L to L' [Oh et al., 2008]

For each article A in L with a cross-language link to article A' in L', add an entry to the dictionary that translates the title of article A to the title of article A'

Automatically Create a Dictionary

DICTIONARY
Estados Unidos: United States
República da Irlanda: Republic of Ireland
Dezembro: December

Compute Similarity for *Translated* Values

- Given attributes a and a' , vsim(a,a') = cos(v^t_a,$v_{a'}$)

  $v_{nascimento}$ ={1963:1, Irlanda:1, 18 de Dezembro 1950:1, Estados Unidos:2}

  $v^t_{nascimento}$ ={1963:1, **Ireland:1**, **December 18 1950:1**, **United States:2**}

  v_{born} ={1963:1, Ireland:1, June 4 1975:1, United States: 3}

  vsim(nascimento, born) = cos($v^t_{nascimento}$,$v_{born'}$) = 0.62

Link Structure Similarity

O Último Imperador

O Último Imperador (PT/BR)

Reino Unido / Itália
França / China
1987 • cor • 165 min

Produção

Direção	Bernardo Bertolucci
Roteiro	Mark Peploe / Bernardo Bertolucci
Elenco original	John Lone Joan Chen Peter O'Toole Ryuichi Sakamoto
Género	drama biográfico / épico
Idioma original	inglês / mandarim / japonês

IMDb: (inglês) (português)

Projeto Cinema • Portal Cinema

Directed by	Bernardo Bertolucci
Produced by	Jeremy Thomas
Written by	Mark Peploe Bernardo Bertolucci
Starring	John Lone Joan Chen Peter O'Toole
Music by	Ryuichi Sakamoto
Cinematography	Vittorio Storaro
Editing by	Gabriella Cristiani
Studio	Hemdale Film
Distributed by	Columbia Pictures
Release date(s)	**United States:** November 18, 1987
Running time	160 minutes (theatrical) 218 minutes (television)
Country	China Italy United Kingdom France
Language	English Mandarin Chinese
Budget	$23.8 million[1]

Bernardo Bartolucci

Nascimento	16 de março de 1941 (71 anos) Parma Itália
Nacionalidade	Italiano
Ocupação	Diretor Roteirista
Cônjuge	Clare Peploe (1990 - atualmente)

IMDb: (inglês) (português)

Cross-language link

Bernardo Bertolucci

Born	16 March 1940 (age 72) Parma, Emilia-Romagna, Italy
Years active	1962–present
Spouse	Adriana Asti (div.) Clare Peploe (1990–)
Parents	Attilio Bertolucci (1911–2000) Ninetta Giovanardi

Link Structure Similarity

- The *link structure set* of an attribute in an entity type schema S is the set of outgoing links for all of its values
- Let ls(*a*) = {l_a|i = 1..n} and ls(*a'*) = {$l_{a'}$ |j = 1..m} be the link structure sets for attributes *a* and *a'*
- The link structure similarity between these attributes is measured as: linksim(*a,a'*) = cos(ls(*a*),ls(*a'*)).

  $ls_{nascimento}$ = {Irlanda:1, Estados Unidos:2}

  ls_{born} ={Ireland:1, United States:3}

  lsim(nascimento,born) = cos($ls_{nascimento}$, ls_{born}) = 0.99
- Link similarity can be misleading:

  $ls_{release\ date}$ ={1975:1, 1998:2, United States: 3}

  $ls_{quốc\ gia/country}$={Việt Nam:2, Hoa Kỳ:4}

  lsim(released date,quốc gia) = cos($ls_{released\ date}$, $ls_{quốc\ gia}$) = 0.72

Attribute Similarity: Correlation and ***LSI***

- LSI has been used to match terms across languages in *free text*
- Here, we use LSI as a correlation measure for *structured data*
- Create a set of dual-language infoboxes
  - E.g., actor-ator
- Build a co-occurrence matrix and apply SVD
- Cross-language synonyms are represented by similar vectors
- Intra-language synonyms are represented by distinct vectors

Attribute Correlation and ***LSI*** (cont.)

- Compute the cosine between vectors

$$LSI(a_p, a_q) = \begin{cases} cosine(\overrightarrow{a_p}, \overrightarrow{a_q}) & \text{if } a_p \text{ in } L \wedge a_q \text{ in } L' \\ 0 & \text{if } a_p,\ a_q \text{ in } I_L \text{ or } I_{L'} \\ 1 - cosine(\overrightarrow{a_p}, \overrightarrow{a_q}) & \text{if } a_p \wedge a_q \text{ in } L \text{ or } L' \end{cases}$$

LSI(a_p,a_q) = 1

→ intra-language synonyms, if same language

→ cross-language synonyms, if different languages

- Because cross-language infoboxes are not parallel, LSI by itself,ß is not sufficient
  - Need to combine LSI with other similarity measures

Combining Similarity Measures

- Group attributes with the same label, and for each group aggregate their values
- For each pair of attribute groups, compute similarities and sort by LSI, eliminating tuples whose LSI < T_{LSI}
- $<a_p,a_q>$ is a match if : $\max(vsim(a_p,a_q),lsim(a_p,a_q)) > T_{sim}$
- Grow match set carefully
- Revise uncertain matches

(see Nguyen et al., VLDB2012 for details)

LSI	vsim	lsim	Attribute Pair
0.99	0.45	0.73	born; nascimento
0.94	0.91	0.83	falecimento; morte
0.92	0.65	0.71	died; falecimento
0.73	0.73	0.26	spouse;cônjuge
0.39	0.60	0.38	died; nascimento
0.25	0.68	0.73	died;morte
0.20	0.47	0.00	other names; outros nomes
0.12	0.51	0.54	born; morte
0.00	0.95	0.58	nascimento; falecimento

M={died ~ falecimento}

- p1=<died, morte>
- p2=<died, nascimento>

LSI(nascimento,falecimento) = 0

- p_1 is integrated to M, but not p_2.
- M={died~ falecimento ~morte}

Experimental Evaluation

- Data: Wikipedia infoboxes related to movies in English (En), Vietnamese (Vn) and Portuguese (Pt)
  - Portuguese and English are morphologically similar, but Vietnamese is different from both; Vietnamese is under-represented
  - Construct dual-language infoboxes for Vn-En (659) and Pt-En (8,898)
- Ground truth: A bilingual expert labeled as correct or incorrect all the correspondences containing attributes from the two language pairs (Pt-En 315; Vn-En 160)
- Metrics: Weighted precision and recall to account for *important* attributes
- Baselines consisted of multiple configurations for
  - LSI
  - Coma++ (schema matching and translation)
  - Bouma (values and cross-language links)

Effectiveness: High Precision and Recall

Portuguese-English

Type	WikiMatch			Bouma			COMA++			LSI		
	P	R	F	P	R	F	P	R	F	P	R	F
film	0,97	0,95	0,96	0,79	**0,99**	0,88	**0,99**	0,95	**0,97**	0,01	0,20	0,02
show	**1,00**	**0,89**	**0,94**	0,82	0,68	0,75	0,98	0,52	0,68	0,07	0,05	0,06
actor	**1,00**	**0,52**	**0,68**	**1,00**	0,24	0,39	0,70	0,52	0,60	0,15	0,26	0,19
artist	**1,00**	**0,72**	**0,84**	**1,00**	0,55	0,71	**1,00**	0,34	0,51	0,75	0,50	0,60
channel	0,80	**0,69**	**0,74**	**1,00**	0,33	0,50	0,89	0,56	0,68	0,26	0,40	0,32
company	0,86	**0,87**	**0,87**	**1,00**	0,53	0,69	0,95	0,70	0,81	0,67	0,74	0,71
comics ch.	0,97	**0,87**	**0,92**	**0,99**	0,65	0,79	**0,99**	0,77	0,86	0,37	0,53	0,43
album	**1,00**	**0,93**	**0,96**	**1,00**	0,69	0,82	**1,00**	0,77	0,87	0,56	0,48	0,52
adult actor	0,84	**0,59**	**0,69**	**1,00**	0,26	0,41	0,73	0,43	0,54	0,22	0,19	0,20
book	0,80	**0,75**	**0,77**	0,75	0,58	0,66	0,75	0,66	0,70	0,15	0,36	0,21
episode	0,81	**0,90**	**0,85**	0,86	0,32	0,47	**1,00**	0,38	0,55	0,09	0,17	0,12
writer	**1,00**	**0,49**	**0,65**	**1,00**	0,22	0,36	**1,00**	0,27	0,43	0,60	**0,49**	0,54
comics	0,92	**0,65**	**0,76**	**1,00**	0,13	0,23	0,91	0,45	0,61	0,00	0,00	0,00
fictional ch.	**1,00**	0,69	**0,82**	**1,00**	0,06	0,11	0,81	**0,81**	0,81	0,36	0,37	0,36
Avg	0,93	**0,75**	**0,82**	**0,94**	0,45	0,55	0,91	0,58	0,69	0,30	0,34	0,31

Vietnamese-English

Type	WikiMatch			Bouma			COMA++			LSI		
	P	R	F	P	R	F	P	R	F	P	R	F
film	**1,00**	**0,99**	**0,99**	**1,00**	**0,99**	**0,99**	**1,00**	0,91	0,95	0,65	0,62	0,63
show	**1,00**	**0,88**	**0,93**	**1,00**	0,36	0,53	**1,00**	0,61	0,76	0,57	0,49	0,53
actor	**1,00**	**0,49**	**0,66**	**1,00**	0,28	0,44	**1,00**	0,39	0,56	0,49	0,35	0,41
artist	**1,00**	**0,65**	**0,79**	**1,00**	0,32	0,48	**1,00**	0,25	0,40	0,72	0,50	0,59
Avg	**1,00**	**0,75**	**0,84**	**1,00**	0,49	0,61	**1,00**	0,54	0,67	0,61	0,49	0,54

Effectiveness: High Precision and Recall

Portuguese-English

Type	WikiMatch			Bouma			COMA++			LSI		
	P	R	F	P	R	F	P	R	F	P	R	F
film	0,97	0,95	0,96	0,79	**0,99**	0,88	**0,99**	0,95	**0,97**	0,01	0,20	0,02
show	**1,00**	**0,89**	**0,94**	0,82	0,68	0,75	0,98	0,52	0,68	0,07	0,05	0,06
actor	**1,00**	**0,52**	**0,68**	**1,00**	0,24	0,39	0,70	0,52	0,60	0,15	0,26	0,19
artist	**1,00**	**0,72**	**0,84**	**1,00**	0,55	0,71	**1,00**	0,34	0,51	0,75	0,50	0,60
channel	0,80	**0,69**	**0,74**	**1,00**	0,33	0,50	0,89	0,56	0,68	0,26	0,40	0,32
company	0,86	**0,87**	**0,87**	**1,00**	0,53	0,69	0,95	0,70	0,81	0,67	0,74	0,71
comics ch.	0,97	**0,87**	**0,92**	**0,99**	0,65	0,79	**0,99**	0,77	0,86	0,37	0,53	0,43
album	**1,00**	**0,93**	**0,96**	**1,00**	0,69	0,82	**1,00**	0,77	0,87	0,56	0,48	0,52
adult actor	0,84	**0,59**	**0,69**	**1,00**	0,26	0,41	0,73	0,43	0,54	0,22	0,19	0,20
book	0,80	**0,75**	**0,77**	0,75	0,58	0,66	0,75	0,66	0,70	0,15	0,36	0,21
episode	0,81	**0,90**	**0,85**	0,86	0,32	0,47	**1,00**	0,38	0,55	0,09	0,17	0,12
writer	**1,00**	**0,49**	**0,65**	**1,00**	0,22	0,36	**1,00**	0,27	0,43	0,60	**0,49**	0,54
comics	0,92	**0,65**	**0,76**	**1,00**	0,13	0,23	0,91	0,45	0,61	0,00	0,00	0,00
fictional ch.	**1,00**	0,69	**0,82**	**1,00**	0,06	0,11	0,81	**0,81**	0,81	0,36	0,37	0,36
Avg	0,93	**0,75**	**0,82**	**0,94**	0,45	0,55	0,91	0,58	0,69	0,30	0,34	0,31

Vietnamese-English

Type	WikiMatch			Bouma			COMA++			LSI		
	P	R	F	P	R	F	P	R	F	P	R	F
film	**1,00**	**0,99**	**0,99**	**1,00**	**0,99**	**0,99**	**1,00**	0,91	0,95	0,65	0,62	0,63
show	**1,00**	**0,88**	**0,93**	**1,00**	0,36	0,53	**1,00**	0,61	0,76	0,57	0,49	0,53
actor	**1,00**	**0,49**	**0,66**	**1,00**	0,28	0,44	**1,00**	0,39	0,56	0,49	0,35	0,41
artist	**1,00**	**0,65**	**0,79**	**1,00**	0,32	0,48	**1,00**	0,25	0,40	0,72	0,50	0,59
Avg	**1,00**	**0,75**	**0,84**	**1,00**	0,49	0,61	**1,00**	0,54	0,67	0,61	0,49	0,54

Effectiveness: High Precision and Recall

Portuguese-English

Type	WikiMatch			Bouma			COMA++			LSI		
	P	R	F	P	R	F	P	R	F	P	R	F
film	0,97	0,95	0,96	0,79	**0,99**	0,88	**0,99**	0,95	**0,97**	0,01	0,20	0,02
show	**1,00**	**0,89**	**0,94**	0,82	0,68	0,75	0,98	0,52	0,68	0,07	0,05	0,06
actor	**1,00**	**0,52**	**0,68**	**1,00**	0,24	0,39	0,70	0,52	0,60	0,15	0,26	0,19
artist	**1,00**	**0,72**	**0,84**	**1,00**	0,55	0,71	**1,00**	0,34	0,51	0,75	0,50	0,60
channel	0,80	**0,69**	**0,74**	**1,00**	0,33	0,50	0,89	0,56	0,68	0,26	0,40	0,32
company	0,86	**0,87**	**0,87**	**1,00**	0,53	0,69	0,95	0,70	0,81	0,67	0,74	0,71
comics ch.	0,97	**0,87**	**0,92**	**0,99**	0,65	0,79	**0,99**	0,77	0,86	0,37	0,53	0,43
album	**1,00**	**0,93**	**0,96**	**1,00**	0,69	0,82	**1,00**	0,77	0,87	0,56	0,48	0,52
adult actor	0,84	**0,59**	**0,69**	**1,00**	0,26	0,41	0,73	0,43	0,54	0,22	0,19	0,20
book	0,80	**0,75**	**0,77**	0,75	0,58	0,66	0,75	0,66	0,70	0,15	0,36	0,21
episode	0,81	**0,90**	**0,85**	0,86	0,32	0,47	**1,00**	0,38	0,55	0,09	0,17	0,12
writer	**1,00**	**0,49**	**0,65**	**1,00**	0,22	0,36	**1,00**	0,27	0,43	0,60	**0,49**	0,54
comics	0,92	**0,65**	**0,76**	**1,00**	0,13	0,23	0,91	0,45	0,61	0,00	0,00	0,00
fictional ch.	**1,00**	0,69	**0,82**	**1,00**	0,06	0,11	0,81	**0,81**	0,81	0,36	0,37	0,36
Avg	0,93	**0,75**	**0,82**	**0,94**	0,45	0,55	0,91	0,58	0,69	0,30	0,34	0,31

Vietnamese-English

Type	WikiMatch			Bouma			COMA++			LSI		
	P	R	F	P	R	F	P	R	F	P	R	F
film	**1,00**	**0,99**	**0,99**	**1,00**	**0,99**	**0,99**	**1,00**	0,91	0,95	0,65	0,62	0,63
show	**1,00**	**0,88**	**0,93**	**1,00**	0,36	0,53	**1,00**	0,61	0,76	0,57	0,49	0,53
actor	**1,00**	**0,49**	**0,66**	**1,00**	0,28	0,44	**1,00**	0,39	0,56	0,49	0,35	0,41
artist	**1,00**	**0,65**	**0,79**	**1,00**	0,32	0,48	**1,00**	0,25	0,40	0,72	0,50	0,59
Avg	**1,00**	**0,75**	**0,84**	**1,00**	0,49	0,61	**1,00**	0,54	0,67	0,61	0,49	0,54

Effectiveness: High Precision and Recall

Portuguese-English

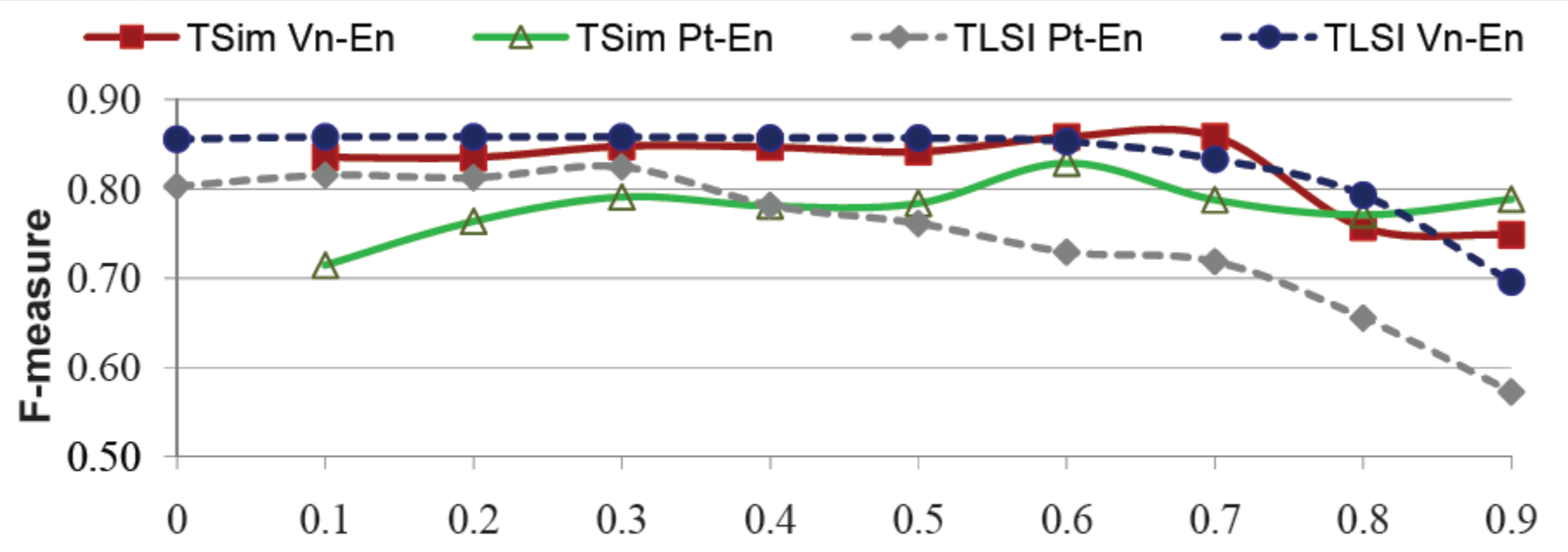
Type	WikiMatch			Bouma			COMA++			LSI		
	P	R	F	P	R	F	P	R	F	P	R	F
film	0,97	0,95	0,96	0,79	**0,99**	0,88	**0,99**	0,95	**0,97**	0,01	0,20	0,02
show	**1,00**	**0,89**	**0,94**	0,82	0,68	0,75	0,98	0,52	0,68	0,07	0,05	0,06
actor	**1,00**	**0,52**	**0,68**	**1,00**	0,24	0,39	0,70	0,52	0,60	0,15	0,26	0,19
artist	**1,00**	**0,72**	**0,84**	**1,00**	0,55	0,71	**1,00**	0,34	0,51	0,75	0,50	0,60
channel	0,80	**0,69**	**0,74**	**1,00**	0,33	0,50	0,89	0,56	0,68	0,26	0,40	0,32
company	0,86	**0,87**	**0,87**	**1,00**	0,53	0,69	0,95	0,70	0,81	0,67	0,74	0,71
comics ch.	0,97	**0,87**	**0,92**	**0,99**	0,65	0,79	**0,99**	0,77	0,86	0,37	0,53	0,43
album	**1,00**	**0,93**	**0,96**	**1,00**	0,69	0,82	**1,00**	0,77	0,87	0,56	0,48	0,52
adult actor	0,84	**0,59**	**0,69**	**1,00**	0,26	0,41	0,73	0,43	0,54	0,22	0,19	0,20
book	0,80	**0,75**	**0,77**	0,75	0,58	0,66	0,75	0,66	0,70	0,15	0,36	0,21
episode	0,81	**0,90**	**0,85**	0,86	0,32	0,47	**1,00**	0,38	0,55	0,09	0,17	0,12
writer	**1,00**	**0,49**	**0,65**	**1,00**	0,22	0,36	**1,00**	0,27	0,43	0,60	**0,49**	0,54
comics	0,92	**0,65**	**0,76**	**1,00**	0,13	0,23	0,91	0,45	0,61	0,00	0,00	0,00
fictional ch.	**1,00**	0,69	**0,82**	**1,00**	0,06	0,11	0,81	**0,81**	0,81	0,36	0,37	0,36
Avg	0,93	**0,75**	**0,82**	**0,94**	0,45	0,55	0,91	0,58	0,69	0,30	0,34	0,31

Vietnamese-English

Type	WikiMatch			Bouma			COMA++			LSI		
	P	R	F	P	R	F	P	R	F	P	R	F
film	**1,00**	**0,99**	**0,99**	**1,00**	**0,99**	**0,99**	**1,00**	0,91	0,95	0,65	0,62	0,63
show	**1,00**	**0,88**	**0,93**	**1,00**	0,36	0,53	**1,00**	0,61	0,76	0,57	0,49	0,53
actor	**1,00**	**0,49**	**0,66**	**1,00**	0,28	0,44	**1,00**	0,39	0,56	0,49	0,35	0,41
artist	**1,00**	**0,65**	**0,79**	**1,00**	0,32	0,48	**1,00**	0,25	0,40	0,72	0,50	0,59
Avg	**1,00**	**0,75**	**0,84**	**1,00**	0,49	0,61	**1,00**	0,54	0,67	0,61	0,49	0,54

Results at Different Thresholds

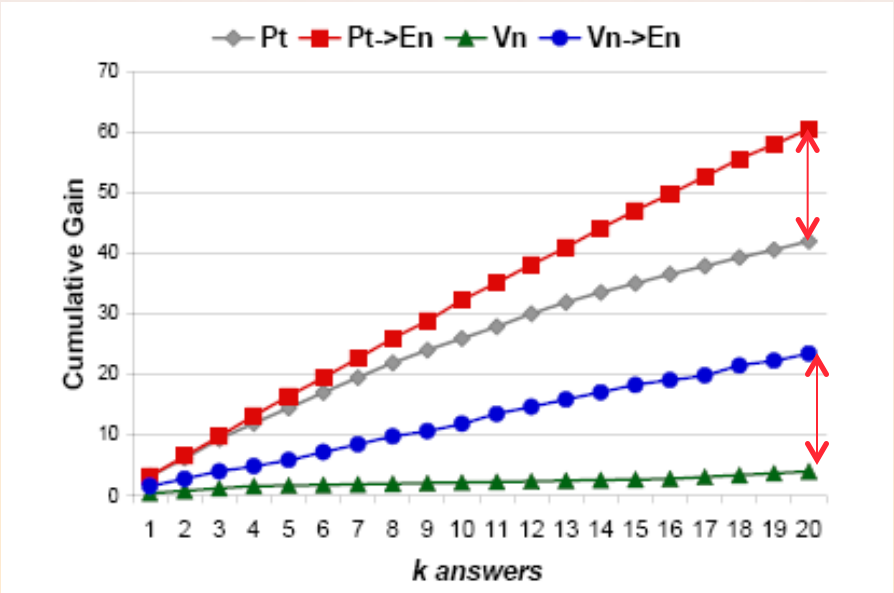
- TLSI should be low and TSim should be high



WikiMatch is robust to a wide variation of thresholds

Impact on Query Evaluation

- Run 10 queries in Pt and Vn
- Translate each query into En using our correspondences and run them
- Choose the top 20 answers for each run and give to an evaluator who rated each answer (scores from 1 to 5)
- Measure cumulative gain (CG)



Summary

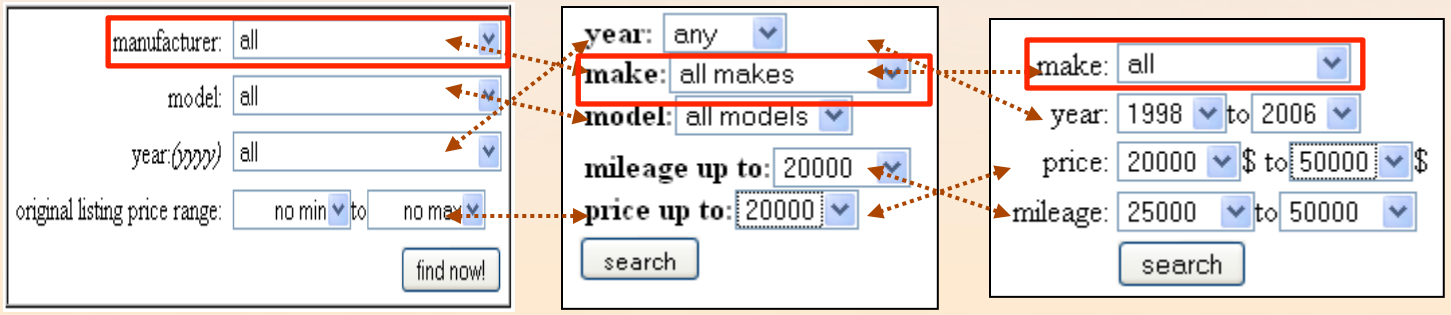
- WikiMatch provides a scalable approach to match infoboxes in different languages
  - Obtains *high* precision *and recall*
- No need for training
- Works for languages that are not syntactically similar and that are under-represented
- Future Work: Improve Wikipedia
  - Apply framework to more languages and entity types
  - Use results to identify inconsistencies and improve coverage for Wikipedia in multiple languages

Data Integration: Big Data Considerations

- Best effort invariably leads to errors: Automate with care!
- Lots of heterogeneity, but many examples – can use correlation!
  - Find multiple sources of similarity
  - Combine them prudently
- Rule of thumb: try to avoid error propagation – prioritize high-confidence matches
- Ideally, algorithms should allow tuning for recall or precision
- Evaluation is challenging
  - How to evaluate the other 267 language pairs?
  - How to check 800k offers?

Big Data Integration: Some Guidelines

Forms	Infoboxes
Group *forms* of the same type and attributes with the same label	Group *infoboxes* of the same type and attributes with the same label
Use multiple sources of similarity	Use multiple sources of similarity
Label, values, correlation	Link, values, correlation
Label and value similarity reinforce correlation	Link and value similarity reinforce correlation
Use high-confidence matches to find additional correspondences	Use high-confidence matches to find additional correspondences



[Nguyen et al., CIKM 2010]

(Big) Data Analysis Pipeline

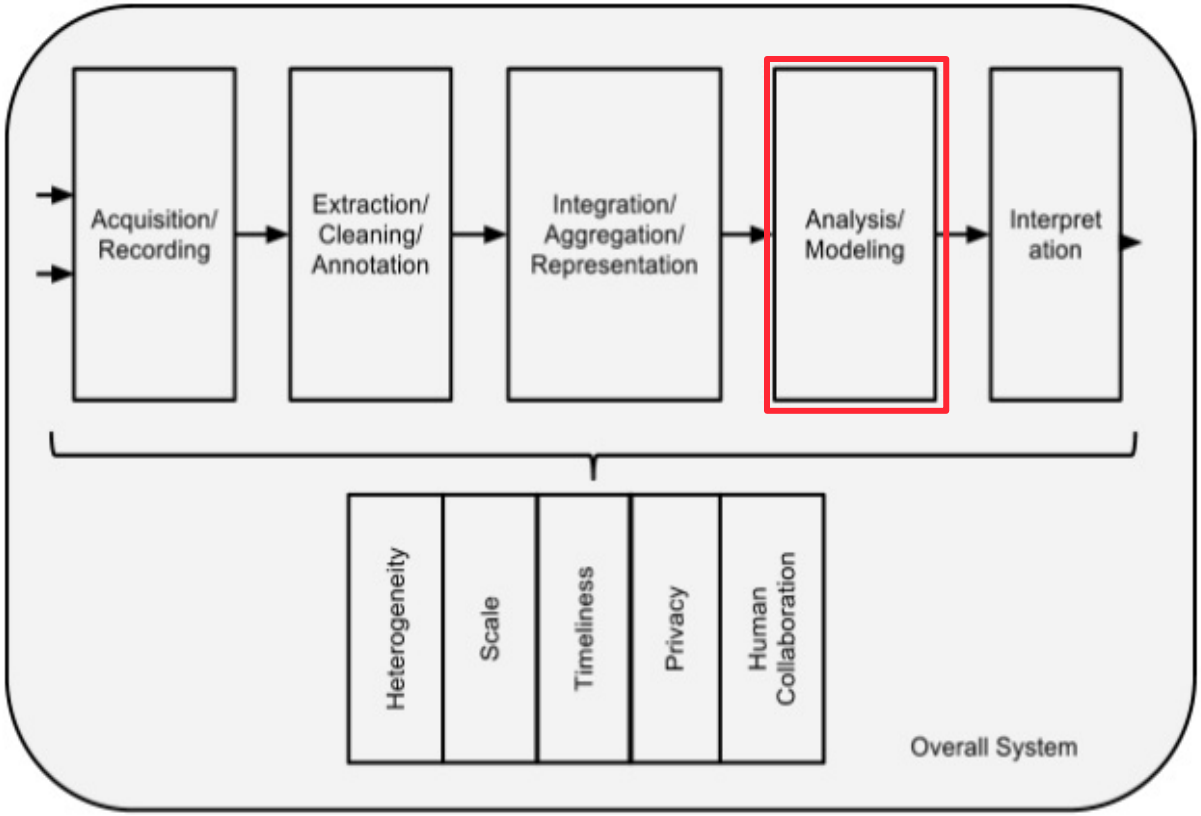


Figure 1: The Big Data Analysis Pipeline. Major steps in analysis of big data are shown in the flow at top. Below it are big data needs that make these tasks challenging.

http://cra.org/ccc/docs/init/bigdatawhitepaper.pdf

Data Analysis and Visualization

- Visualization is essential for exploring large volumes of data
  - "A picture is worth a thousand words"
- Pictures help us think *[Tamara Munzner]*
  - Substitute perception for cognition
  - Free up limited cognitive/memory resources for higher-level problems
- Active area of research
- Many open problems…

Visualization Research @NYU Poly

- Visualization Algorithms and Visual Representations
  - Large-data, streaming, parallel algorithms, etc.
  - "Smart" visualization algorithms (i.e., integration with machine learning)
  - Spatial-temporal data
- Visualization Systems
  - VisTrails, BirdVis, DEFOG, VisCareTrails, PedVis, UV-CDAT, TaxiVis, etc.
- Visualization Evaluation
  - Formal techniques for evaluating correctness and effectiveness of techniques (e.g., using EEG brain waves to measure "effort" for understanding plots)

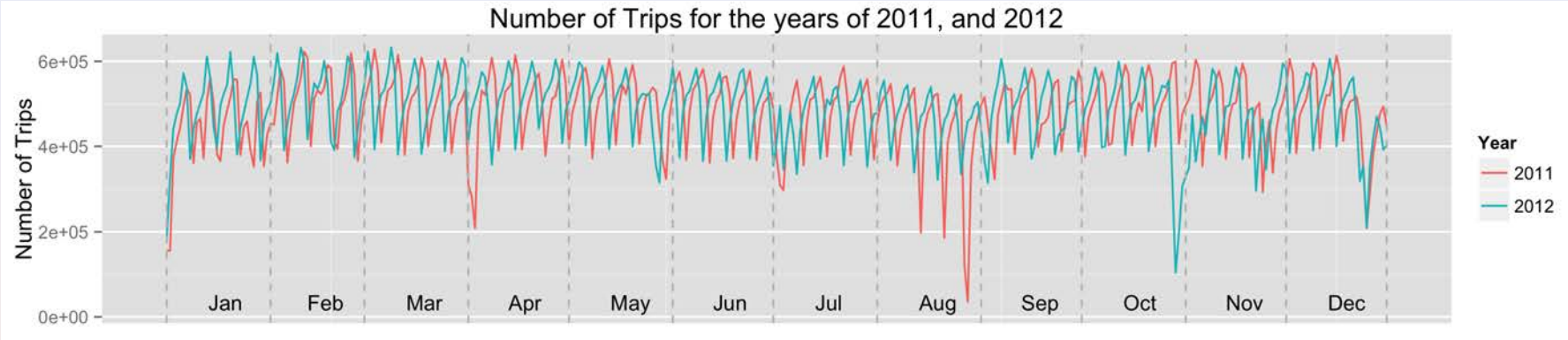
Exploring Big Urban Data

- More than half of the world's population lives in urban areas
- Through the large volumes of data are being collected and stored, it is possible to transform urban science
- Vision:

Enable researchers, decision makers, and citizens to perform complex analyses over an unprecedented collection of data sets never integrated before.

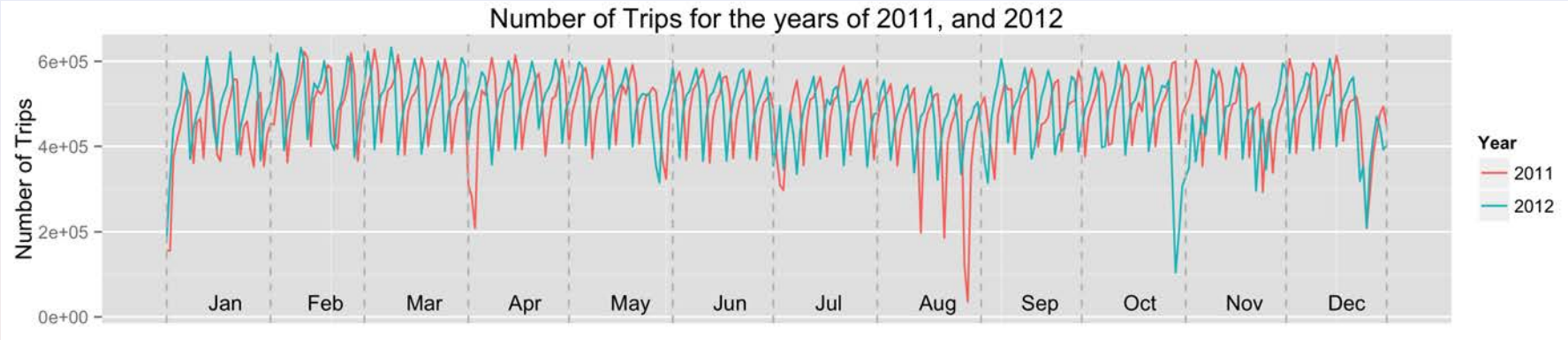
Enable cities to deliver services effectively, efficiently, and sustainably.

Exploring Urban Data: NYC Taxis



- Taxis as sensors for NYC: from economic activity and human behavior to mobility patterns
  - "What is the average trip time from Midtown to the airports during weekdays?"
  - "How the taxi fleet activity varies during weekdays?"
  - "How was the taxi activity in Midtown affected during a presidential visit?"
  - "How did the movement patterns change during Sandy?"
  - "Where are the popular night spots?"

Exploring Urban Data: NYC Taxis



- Data are big and complex
  - Multiple variables: *spatial temporal* + *trip attributes*
  - Large collection: 520 million trips -- ~500k trips/day
- Queries and analyses are hard to specify
- Domain scientists are unable to explore the *whole* data

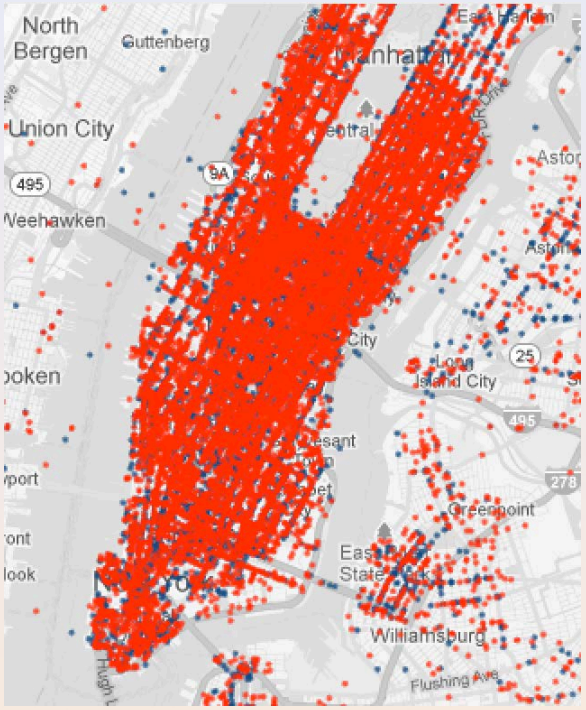
Managing Data

- Raw data:
  - 3 years: 2009, 2011, and 2012
  - 150 GB in 48 CSV files
  - 520M trips total
- After ETL:
  - 50GB in binary format
  - 12 fields with 2 temporal spatial attributes

	SQLite	Our solution
Storage Space	100 GB	30 GB
Building Indices (for 1 year of data)	52 hours	8 mins
Simple Queries	2s - 15s	0.2s
Complex Queries	1 min	2s

Visualizing Data

trips in an hour



trips in a day
too much information!

trips in a day
using level of detail and heat maps

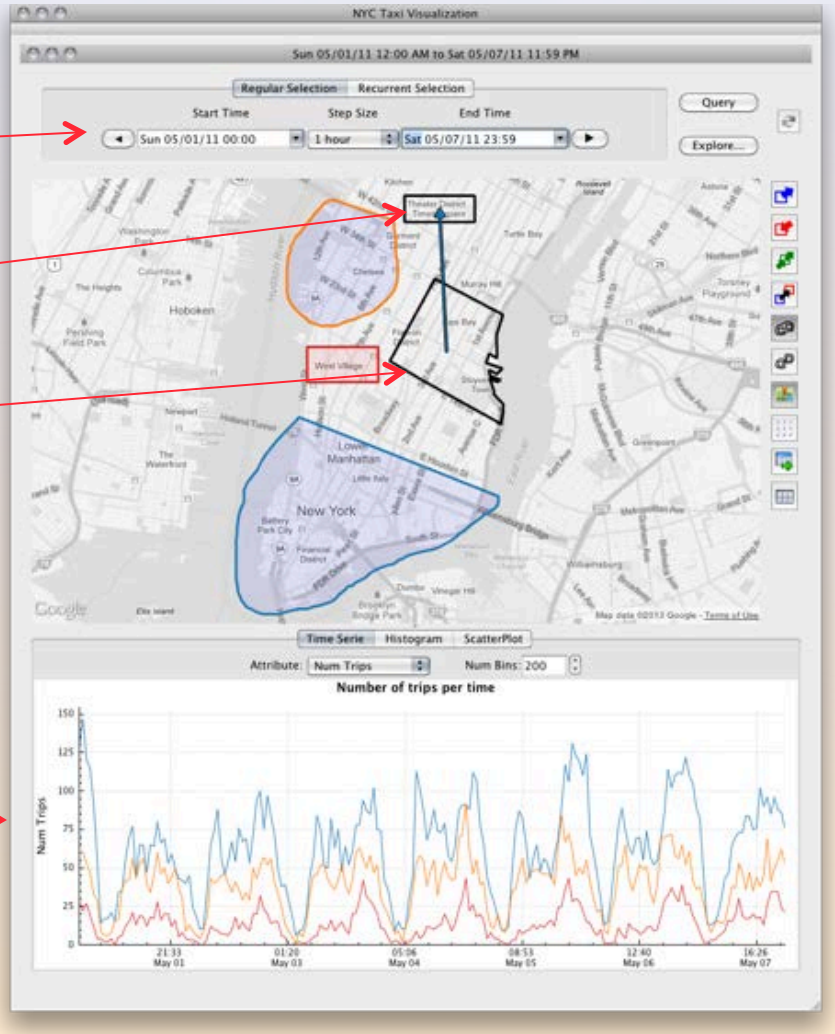
Data Exploration: A Two-Phase Process

- Data selection: Specify query constraints
- Visual analysis
  - Investigate selected data through visualization
  - Discover regions of interest
  - Define new data selections for further exploration

We unify the two through visual operations

Visual Data Selection

```
SELECT  *
FROM    trips
WHERE pickup_time in (5/1/11,5/7/11)
                AND
        dropoff_loc in “Times Square”
                AND
        pickup_loc     in “Gramercy”
```



Interactively explore data through the map view and plot widgets

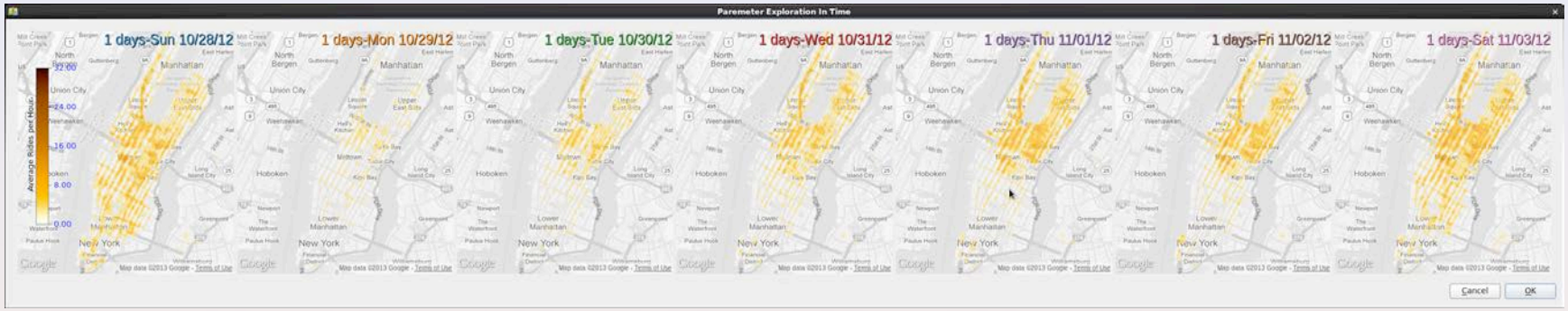
TaxiVis: Visually Exploring NYC Taxi Data

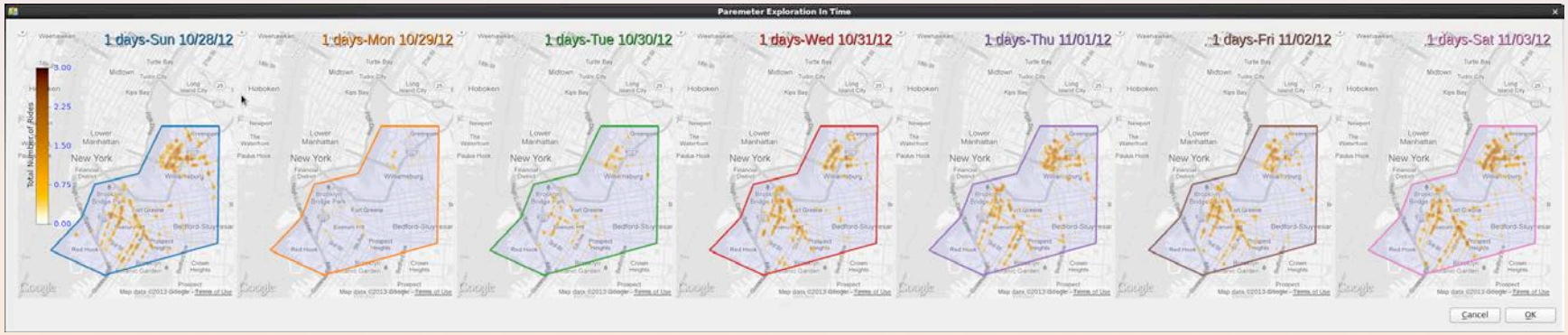
- New model that allows users to visually query taxi trips, easily select and compare different spatial-temporal slices
  - Data selection through visual manipulations
  - Use visualization to explore selected data
- Support for origin-destination queries that enable the study of mobility across the city
- Use multiple coordinated views to allow comparisons, and brushing to support query refinements
- Use of adaptive level-of-detail rendering and heat maps to generate clutter-free visualization for large results
- Scalable system that provides interactive response times for spatio-temporal queries over large data

Visual Query Model

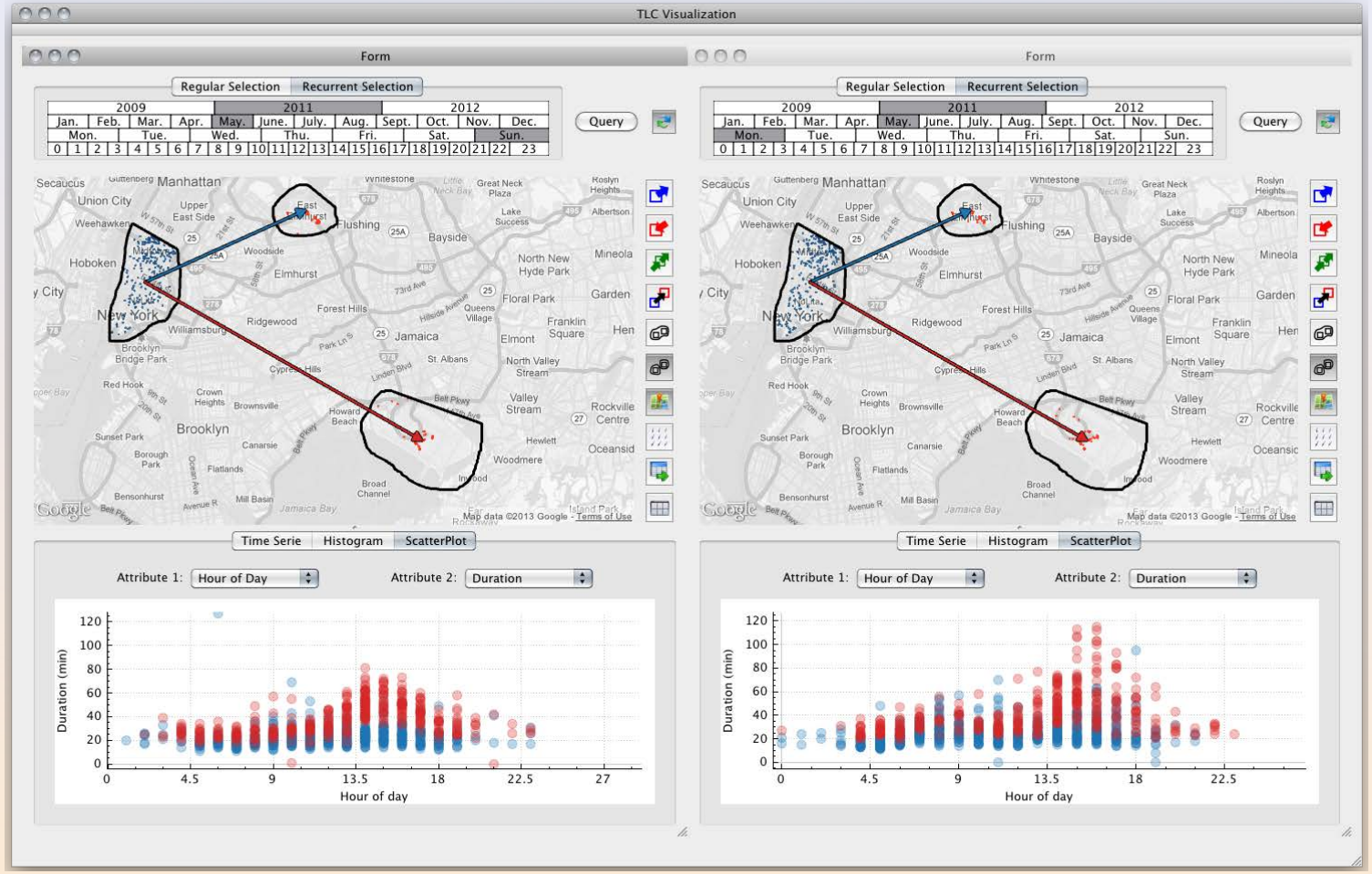
- Data selection by visual operations
- Each data selection can be assigned a different visual representation
  - Spatial context is maintained in the map view
- Query Expressiveness [Peuquet 1994]
  - when + where ➔ what
  - when + what ➔ where
  - where + what ➔ when

The Effects of Sandy: Temporal Comparison





Analyzing Movement

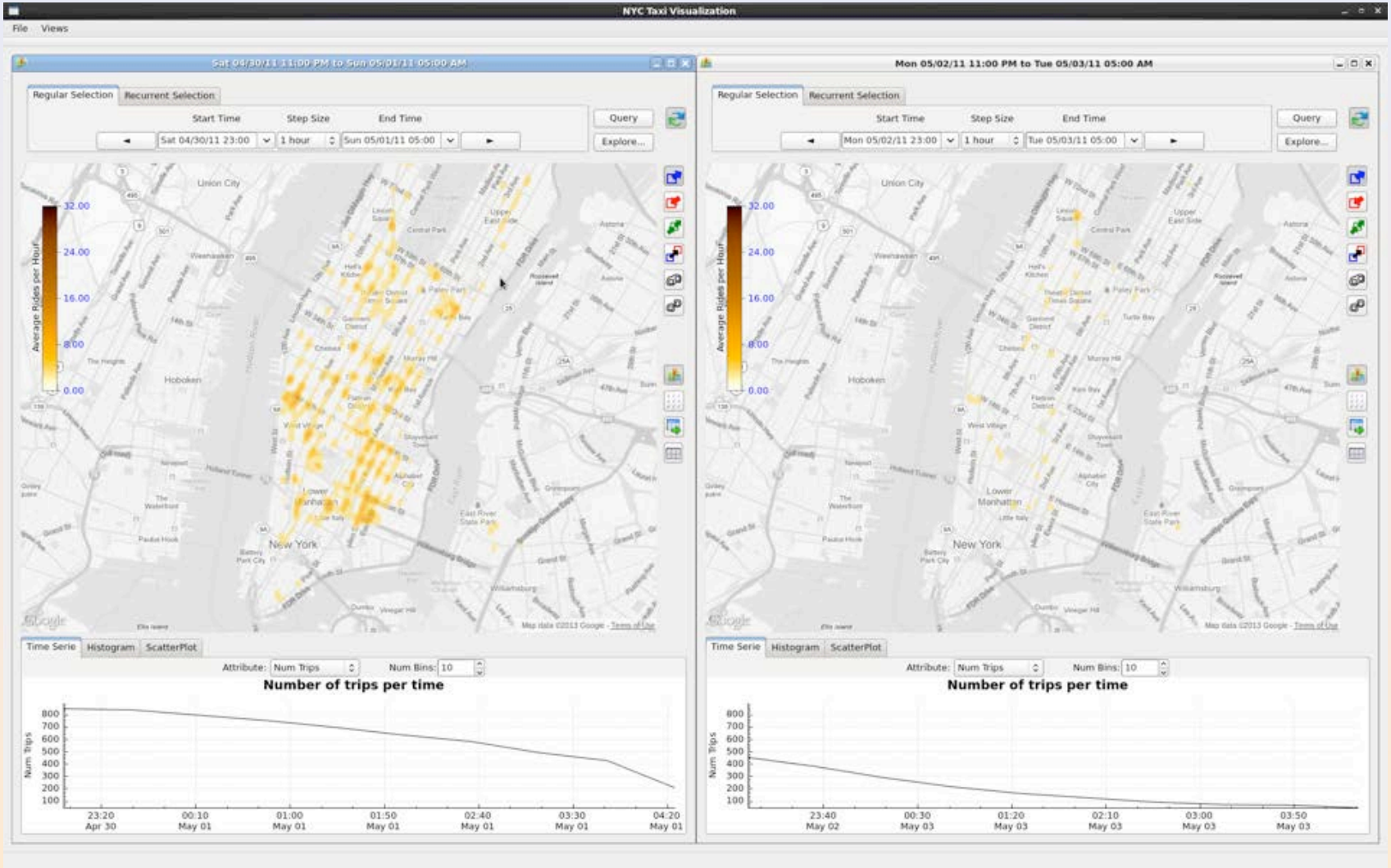


Detecting Events and Outliers

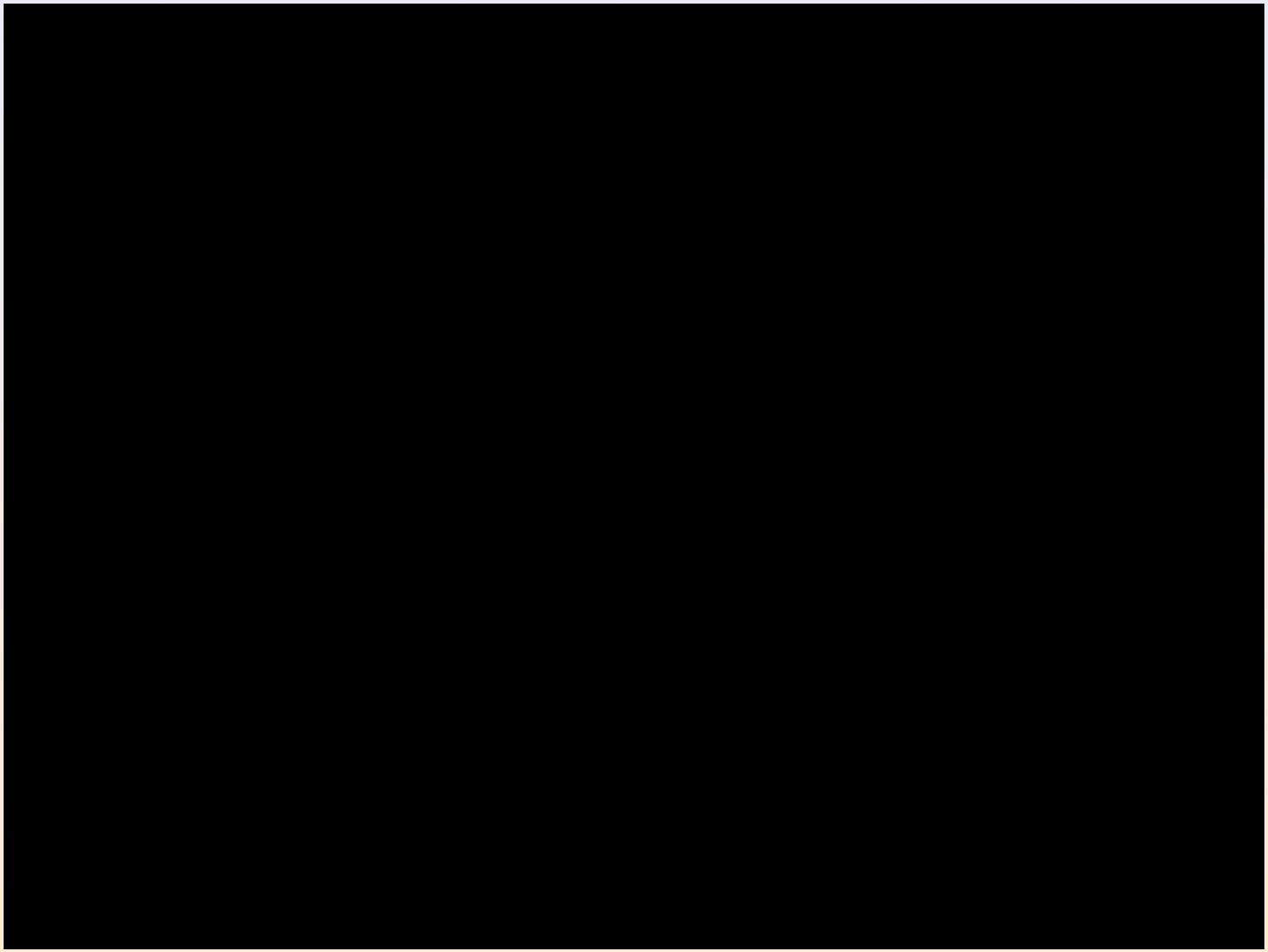
7-8am *8-9am* *9-10am* *10-11am*

Five Boro Bike Tour

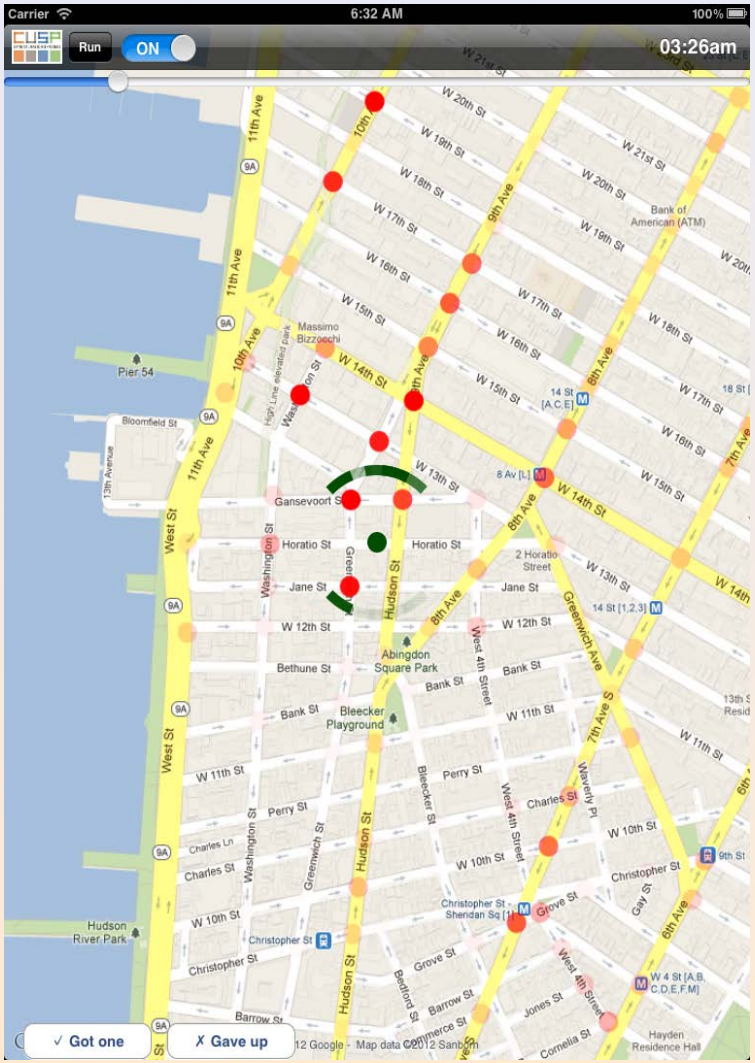
Night Life in NYC: Saturday vs. Monday



TaxiVis in Action (video)



CabFinder App

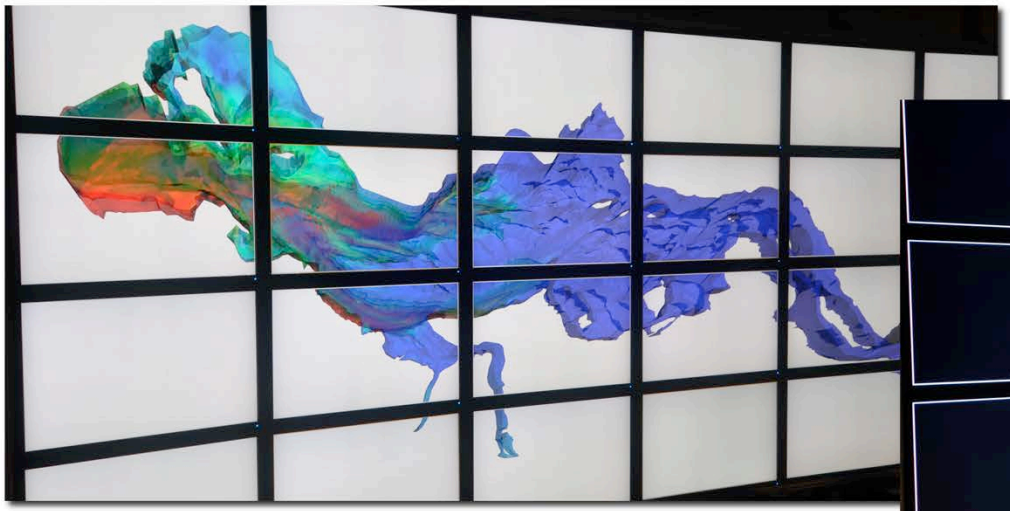


Summary

- Easy-to-use system to interactively explore large multivariate spatial-temporal data
- Future and ongoing work:
  - Apply to other urban mobility data, e.g., data from the NYC bike share program
  - Support additional data layers: weather, gas prices, news, tweets, etc.
  - Utilize parallel processing

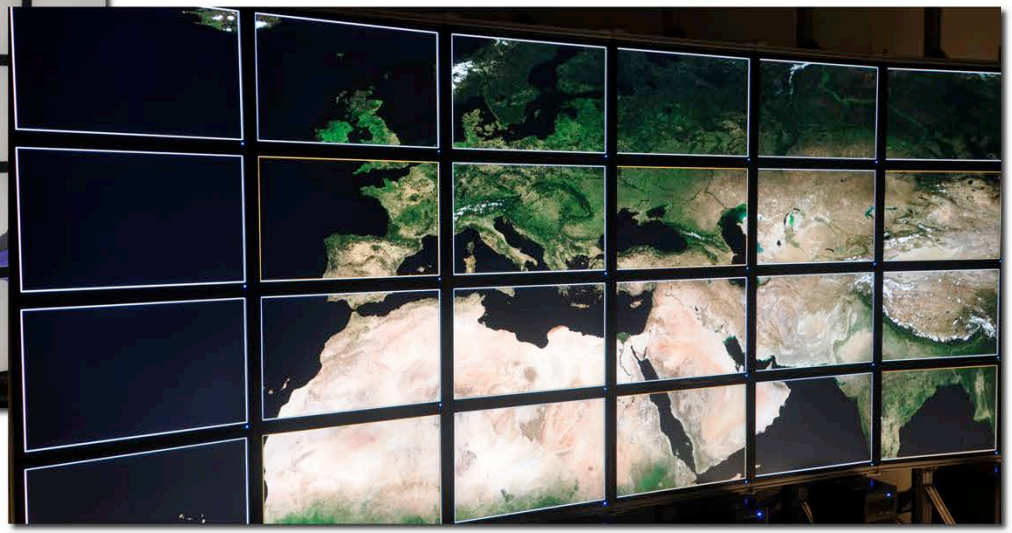
Visualization: Big Data Considerations

- There is a limit to what can fit in a screen, or that we can understand



River Estuary CFD Visualization

13GB Satellite Image Composite



Visualization: Big Data Considerations

- There is a limit to what can fit in a screen, or that we can understand
- Interactivity is key, but challenging for Big Data
  - Map Reduce has very high latency
  - RDBMS and even main-memory databases can be slow
- Need better integration between data management and visualization components [Fekete and Silva, DEB 2012]
  - Designed specialized index
- Need *usable tools designed for data enthusiasts* – both for data management and visualization

Conclusions and Future Work

- Data exploration is challenging for both small and big data – need tools that are easy to use
- Data integration at scale
  - Need automated methods that provide at least a starting point
  - Big data creates challenges but it is also an enabler: many samples, multiple sources of similarity
- Visualization is a powerful tool for data exploration
  - Its use is growing! [Halevy and McGregor, DEB 2012]
  - E.g., Google Fusion Tables
  - Need better integration with data management systems– ”visualization tools often implement from scratch their own main-memory databases” [Fekete and Silva, DEB 2012]
  - Challenging to design appropriate visual representations
- Analysis and visualization of large structured data opens up new opportunities and many challenges for computer science

Acknowledgments

- VisTrails group
- Thanh Nguyen, Viviane Moreira, Huy Vo, Lauro Lins, Nivan Ferreira, Jorge Poco, Fernando Chirigati, Claudio Silva
- This work is partially supported by the National Science Foundation, the Department of Energy, and IBM Faculty Awards.

Merci

Ευχαριστω

Thank you

Obrigada