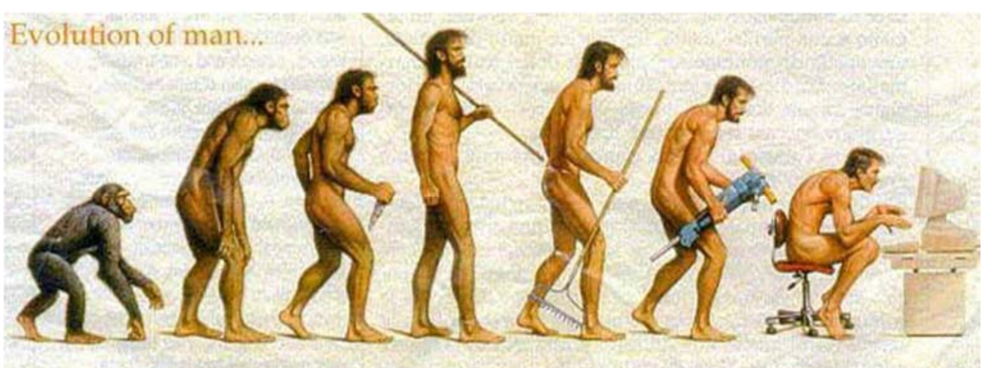
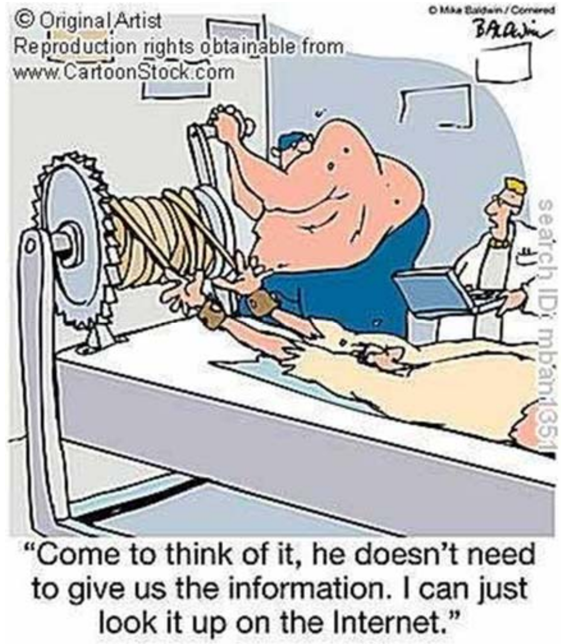
IN SEARCH OF TRUTH (ON THE DEEP WEB)

Divesh Srivastava
AT&T Labs-Research

The Web is Great



A Lot of Information on the Web



Information Can Be Erroneous

Telegraph.co.uk

Home | News | Sport | Finance | Comment | Travel | Lifestyle | Culture | Fas

UK | World | Politics | Celebrities | Obituaries | Weird | Earth | Science | Health News | Educatio

HOME > NEWS > NEWS TOPICS > HOW ABOUT THAT?

Steve Jobs obituary published by Bloomberg

An obituary of very-much-alive Apple founder Steve Jobs has been accidentally published by the respected Bloomberg business news wire.

By Matthew Moore
Last Updated: 7:05PM BST 28 Aug 2008

Text Size
Email this article
Print this article
Share this article
91 diggs digg it

How about that?
USA
News
The week in pictures
IN PICS
Pictures of the day

Steve Jobs was described as the man who 'refashioned the mobile phone' in the erroneous obituary. Photo: REUTERS

The story, marked "Hold for release – Do not use", was sent in error to the news service's thousands of corporate clients.

The story, marked "Hold for release – Do not use", was sent in error to the news service's thousands of corporate clients.

Information Can Be Erroneous

Maurice Jarre (1924-2009) French Conductor and Composer

"One could say my life itself has been one long soundtrack. Music was my life, music brought me to life, and music is how I will be remembered long after I leave this life. When I die there will be a final waltz playing in my head and that only I can hear."

2:29, 30 March 2009

Your continued donations keep Wikipedia running!

article | discussion | edit this page | history

Log in / cre

Maurice Jarre

From Wikipedia, the free encyclopedia

This is an old revision of this page, as edited by 86.42.227.123 (talk) at 02:29, 30 March 2009. It may differ significantly from the current revision.

(diff) ← Previous revision | Current revision (diff) | Newer revision → (diff)

Quotes

Nowadays, if a studio assumes that his film is bad, there is always an executive that gets more nervous than usual and thinks that if they change the music, the film will masterpiece.

One could say my life itself has been one long soundtrack. Music was my life, music brought me to life, and music is how I will be remembered long after I leave this life. will be a final waltz playing in my head and that only I can hear.

False Information Can Be Propagated

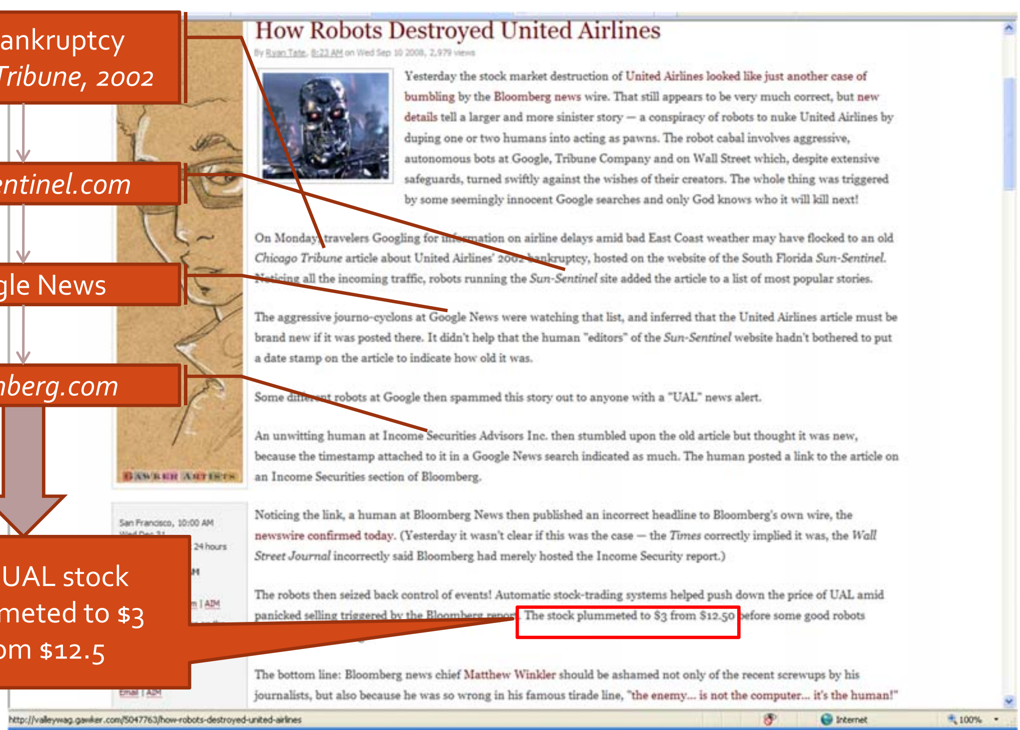
UA's bankruptcy
Chicago Tribune, 2002

Sun-Sentinel.com

Google News

Bloomberg.com

The UAL stock plummeted to $3 from $12.5



How Robots Destroyed United Airlines

By Ryan Tate, 8:23 AM on Wed Sep 10 2008, 2,979 views

Yesterday the stock market destruction of United Airlines looked like just another case of bumbling by the Bloomberg news wire. That still appears to be very much correct, but new details tell a larger and more sinister story — a conspiracy of robots to nuke United Airlines by duping one or two humans into acting as pawns. The robot cabal involves aggressive, autonomous bots at Google, Tribune Company and on Wall Street which, despite extensive safeguards, turned swiftly against the wishes of their creators. The whole thing was triggered by some seemingly innocent Google searches and only God knows who it will kill next!

On Monday, travelers Googling for information on airline delays amid bad East Coast weather may have flocked to an old *Chicago Tribune* article about United Airlines' 2002 bankruptcy, hosted on the website of the South Florida *Sun-Sentinel*. Noticing all the incoming traffic, robots running the *Sun-Sentinel* site added the article to a list of most popular stories.

The aggressive journo-cyclons at Google News were watching that list, and inferred that the United Airlines article must be brand new if it was posted there. It didn't help that the human "editors" of the *Sun-Sentinel* website hadn't bothered to put a date stamp on the article to indicate how old it was.

Some different robots at Google then spammed this story out to anyone with a "UAL" news alert.

An unwitting human at Income Securities Advisors Inc. then stumbled upon the old article but thought it was new, because the timestamp attached to it in a Google News search indicated as much. The human posted a link to the article on an Income Securities section of Bloomberg.

Noticing the link, a human at Bloomberg News then published an incorrect headline to Bloomberg's own wire, the newswire confirmed today. (Yesterday it wasn't clear if this was the case — the *Times* correctly implied it was, the *Wall Street Journal* incorrectly said Bloomberg had merely hosted the Income Security report.)

The robots then seized back control of events! Automatic stock-trading systems helped push down the price of UAL amid panicked selling triggered by the Bloomberg report. The stock plummeted to $3 from $12.50 before some good robots

The bottom line: Bloomberg news chief Matthew Winkler should be ashamed not only of the recent screwups by his journalists, but also because he was so wrong in his famous tirade line, "the enemy... is not the computer... it's the human!"

http://valleywag.gawker.com/5047763/how-robots-destroyed-united-airlines

IS DEEP-WEB DATA CONSISTENT & RELIABLE?

Study on Two Domains

	#Sources	Period	#Objects	#Local-attrs	#Global-attrs	Considered items
Stock	55	7/2011	1000*20	333	153	16000*20
Flight	38	12/2011	1200*31	43	15	7200*31

- Belief of clean data
- Poor data quality can have big impact

Study on Two Domains

	#Sources	Period	#Objects	#Local-attrs	#Global-attrs	Considered items
Stock	55	7/2011	1000*20	333	153	16000*20

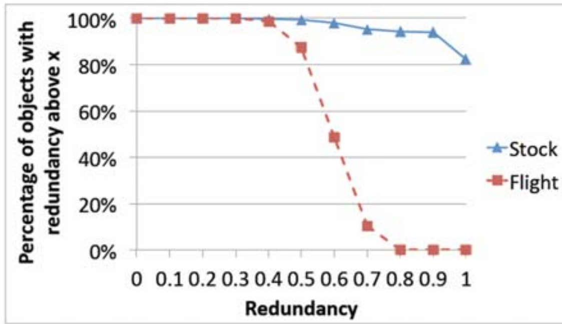
- Stock
  - Search "stock price quotes"
  - Sources: 200 (search results)→89 (deep web)→76 (GET method) →55 (no JavaScript)
  - 1000 "Objects": a stock with a particular symbol on a particular day
    - 30 from Dow Jones Index
    - 100 from NASDAQ100 (3 overlaps)
    - 873 from Russell 3000
  - Attributes: 333 (local) → 153 (global) → 21 (provided by > 1/3 sources) → 16 (no change after market close)

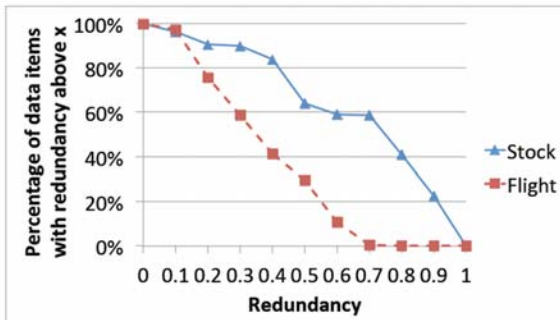
Study on Two Domains

	#Sources	Period	#Objects	#Local-attrs	#Global-attrs	Considered items
Flight	38	12/2011	1200*31	43	15	7200*31

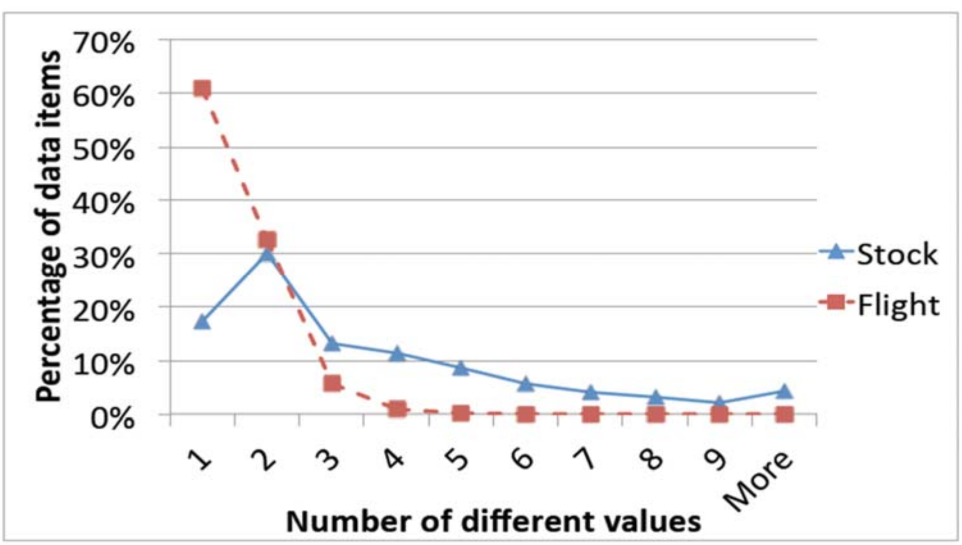
- Flight
  - Search "flight status"
  - Sources: 38
    - 3 airline websites (AA, UA, Continental)
    - 8 airport websites (SFO, DEN, etc.)
    - 27 third-party websites (Orbitz, Travelocity, etc.)
  - 1200 "Objects": a flight with a particular flight number on a particular day from a particular departure city
    - Departing or arriving at the hub airports of AA/UA/Continental
  - Attributes: 43 (local) → 15 (global) → 6 (provided by > 1/3 sources)
    - scheduled dept/arr time, actual dept/arr time, dept/arr gate

Q1. Is There a Lot of Redundant Data? ✓



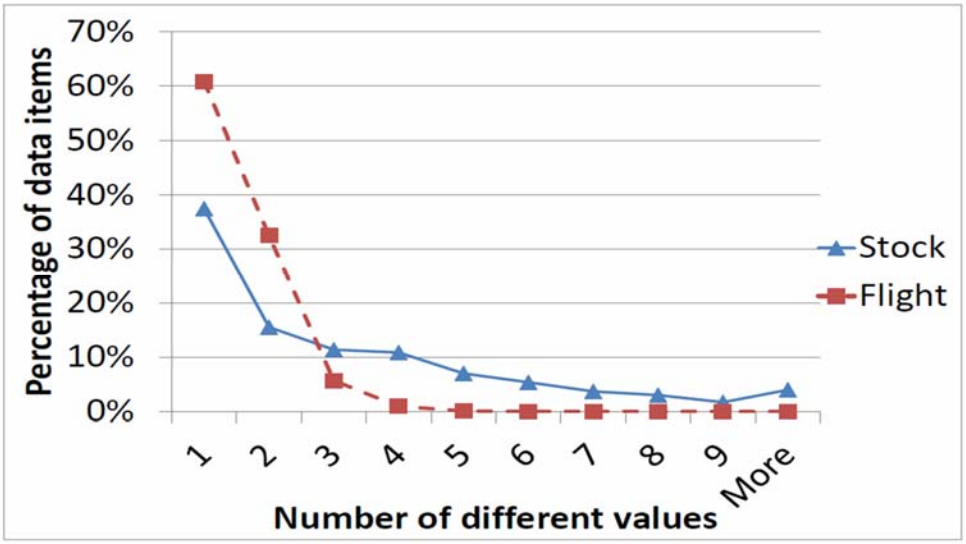


Q2. Is the Data Consistent? ✗



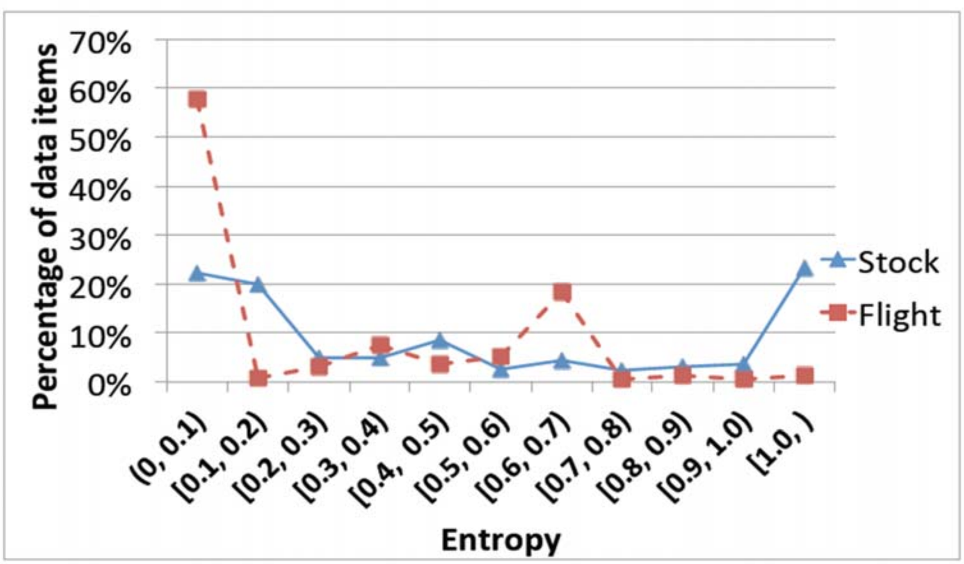
- Tolerance to 1% value difference

Q2. Is the Data Consistent?



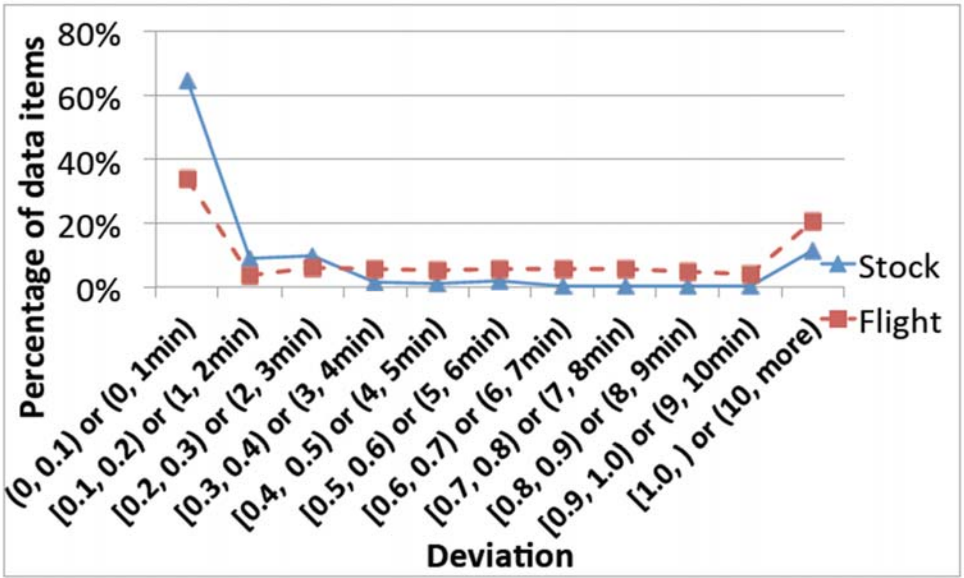
- Tolerance to 1% value difference
- Inconsistency on 50% items after removing *StockSmart*

Q2. Is the Data Consistent? (II)



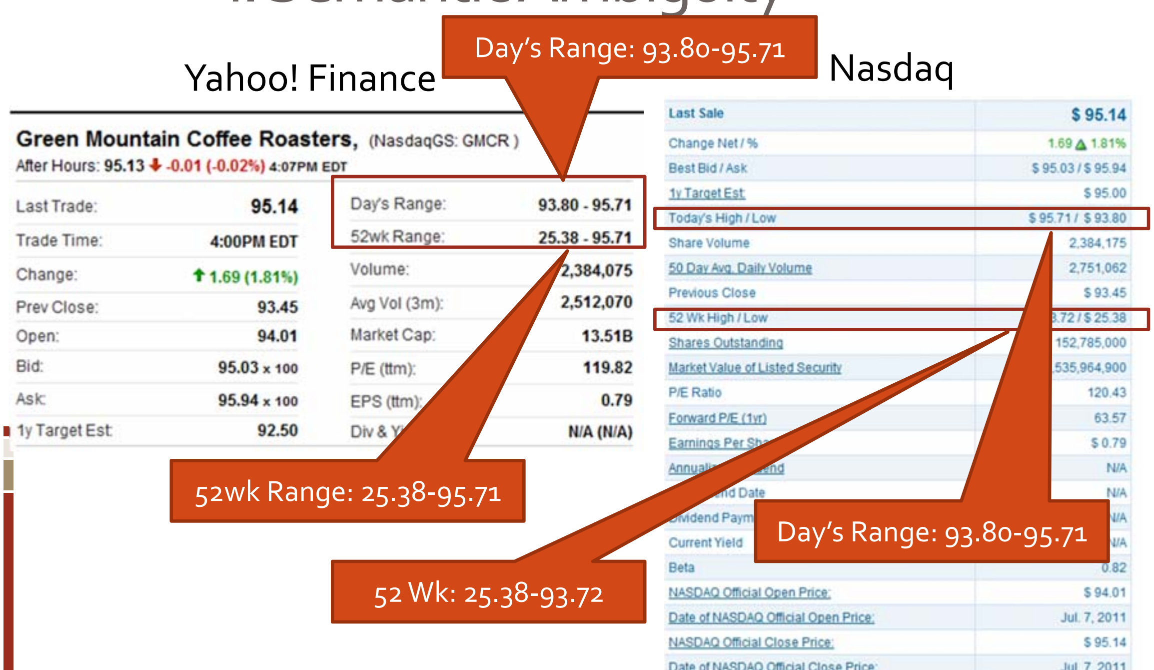
- Entropy measures distribution of different values
- Quite low entropy: one value provided more often than others

Q2. Is the Data Consistent? (III)

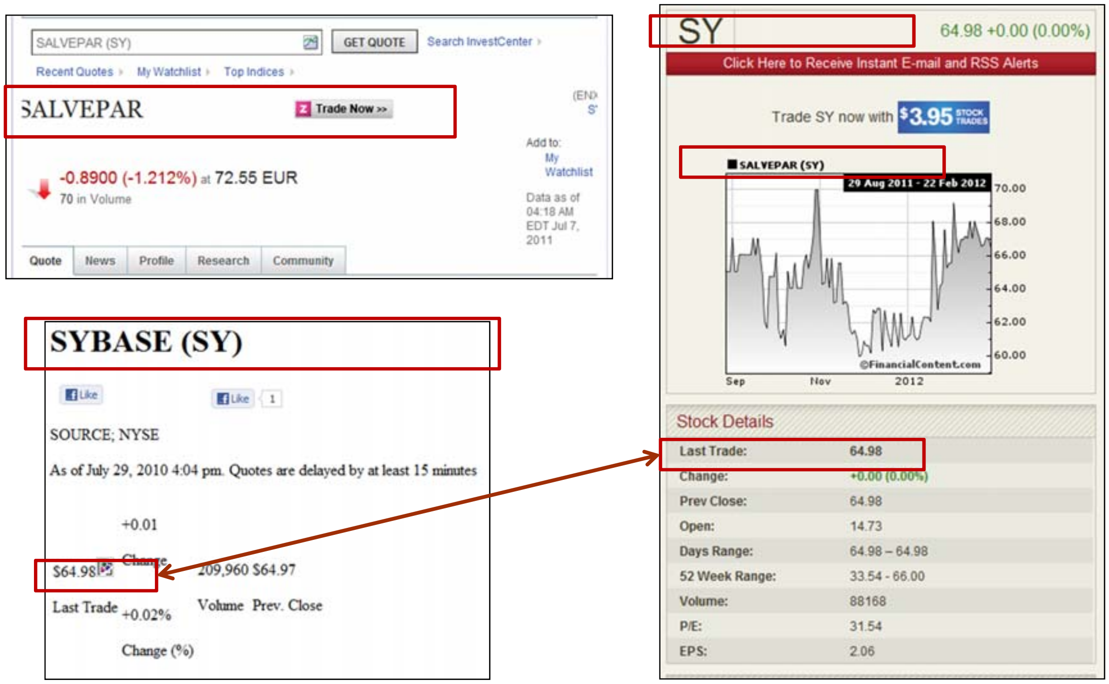


- ❑Deviation measures difference of numerical values
- ❑High deviation: 13.4 for Stock, 13.1 min for Flight

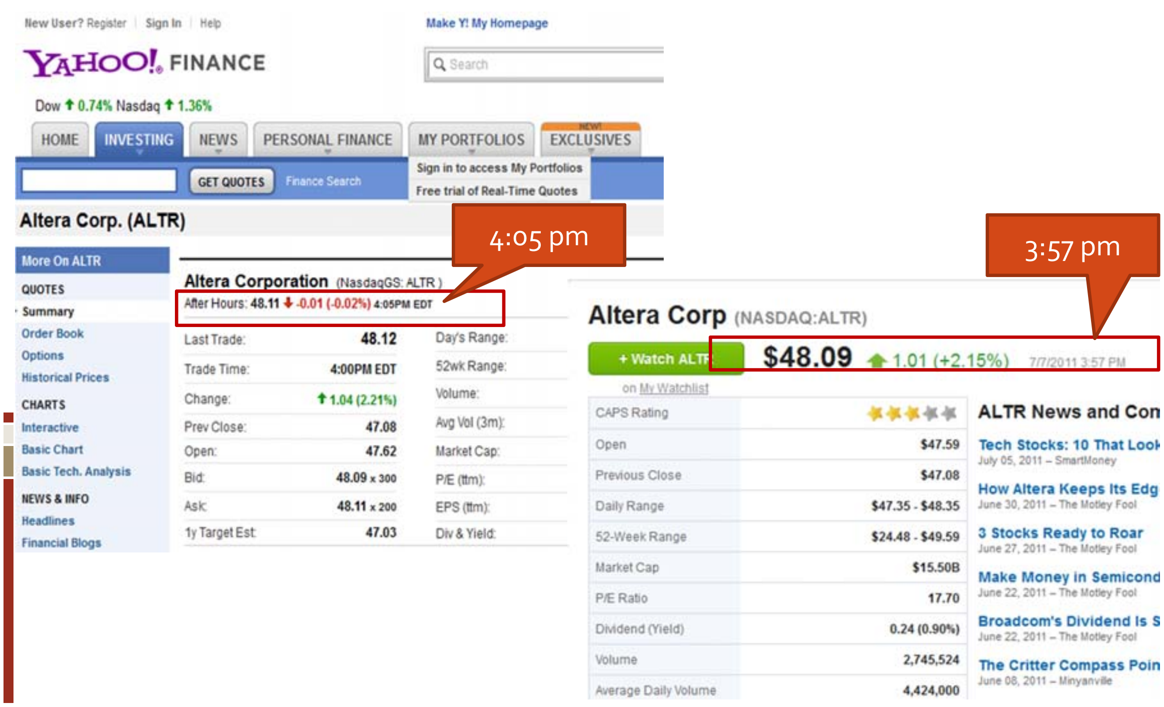
Why Such Inconsistency? — I. Semantic Ambiguity



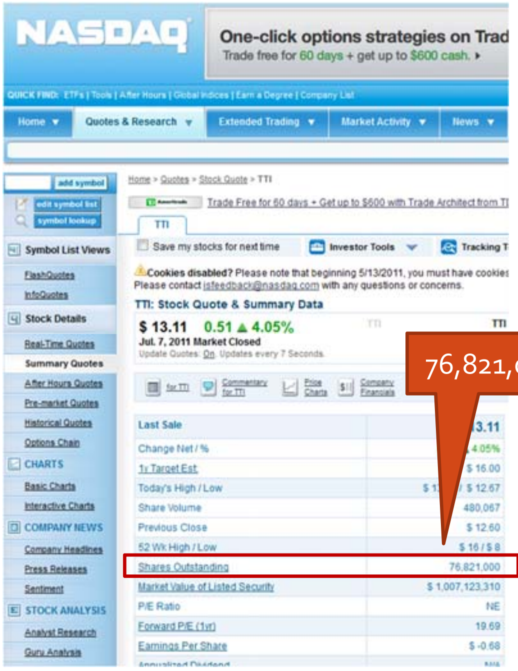
Why Such Inconsistency? — II. Instance Ambiguity

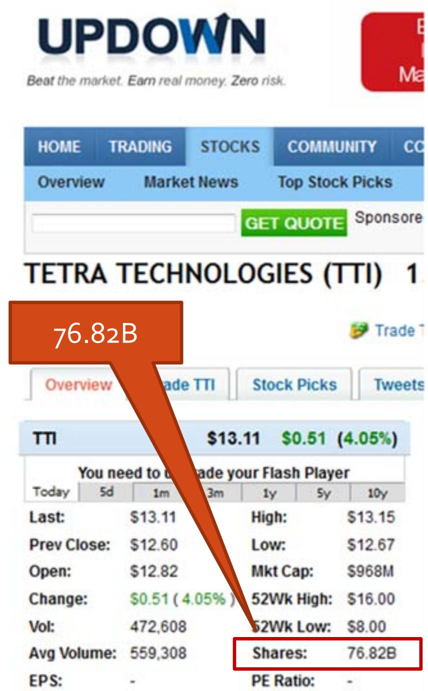


Why Such Inconsistency? — III. Out-of-Date Data



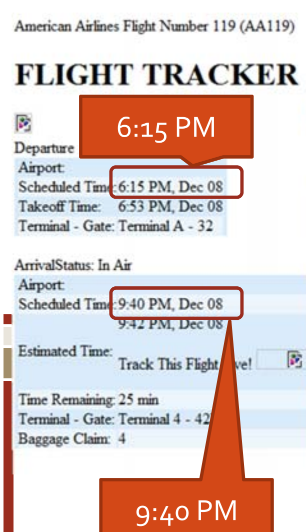
Why Such Inconsistency?
— IV. Unit Error



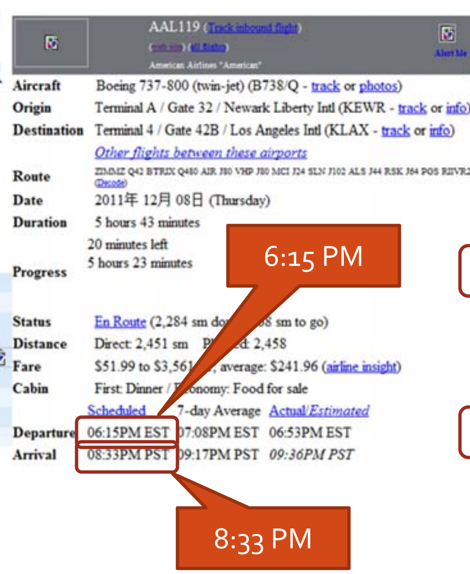


Why Such Inconsistency?
— V. Pure Error

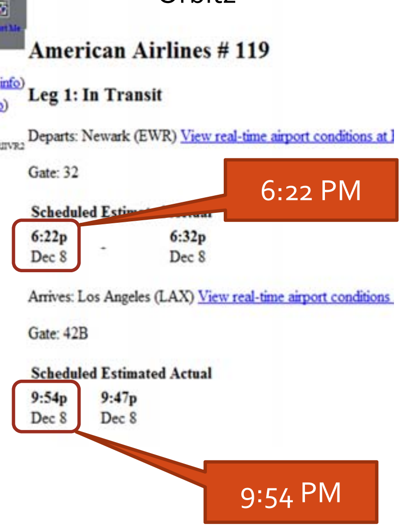
FlightView



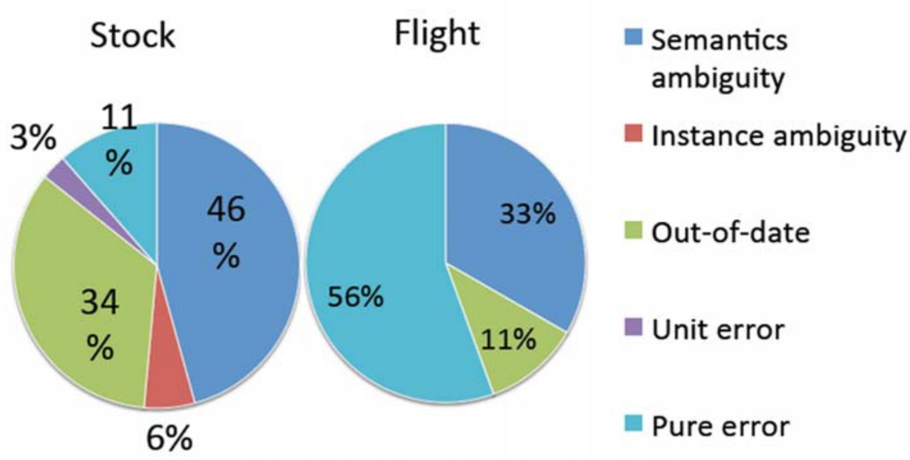
FlightAware



Orbitz

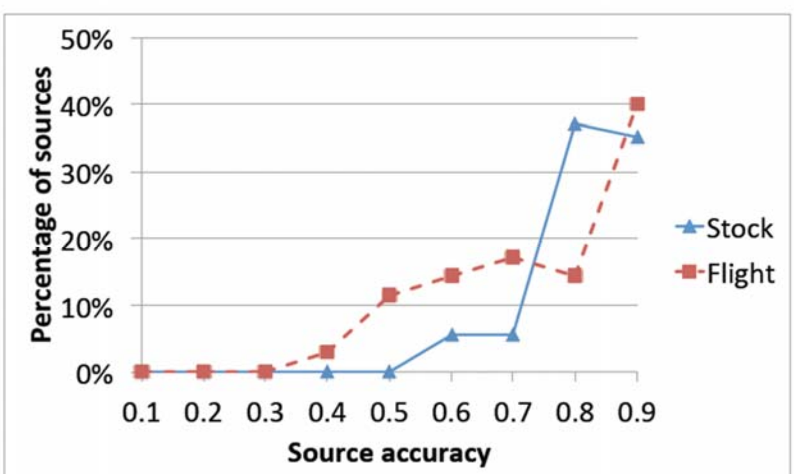


Why Such Inconsistency?



- Random sample of 20 data items and 5 items with the largest # of values in each domain

Q3. Do Sources Have High Accuracy? ✘



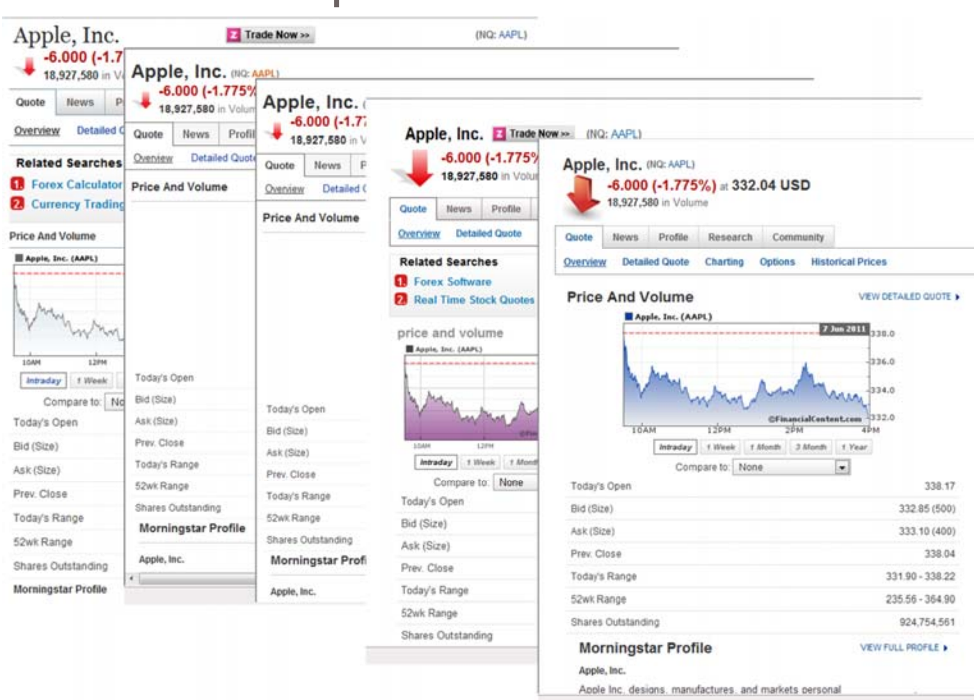
- Not high on average: .86 for Stock and .8 for Flight
- Gold standard
  - Stock: vote on data from *Google Finance, Yahoo! Finance, MSN Money, NASDAQ, Bloomberg*
  - Flight: from airline websites

Q3-2. What About Authoritative Sources? ✘

	Source	Accuracy	Coverage
Stock	*Google Finance*	.94	.82
	Yahoo! Finance	.93	.81
	NASDAQ	.92	.84
	MSN Money	.91	.89
	Bloomberg	.83	.81
Flight	*Orbitz*	.98	.87
	Travelocity	.95	.71
	Airport average	.94	.03

- ❑ Reasonable but not so high accuracy
- ❑ Medium coverage

Q4. Is There Copying or Data Sharing Between Deep-Web Sources? ✓



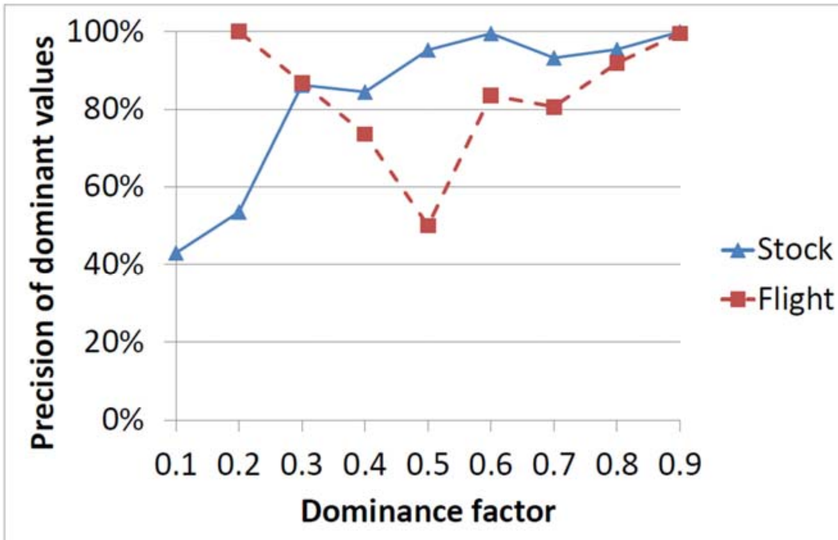
Q4-2. Is Copying or Data Sharing Mainly on Accurate Data?

✘

	Remarks	Size	Schema sim	Object sim	Value sim	Avg accu
Stock	Depen claimed	11	1	.99	.99	.92
	Depen claimed	2	1	1	.99	.75
Flight	Depen claimed	5	0.80	1	1	.71
	Query redirection	4	0.83	1	1	.53
	Dependence claimed	3	1	1	1	.92
	Embedded interface	2	1	1	1	.93
	Embedded interface	2	1	1	1	.61

HOW TO RESOLVE INCONSISTENCY (*DATA FUSION*)?

Basic Solution: Voting



- Only 70% correct values are provided by over half of the sources
  - .908 voting precision for Stock; i.e., wrong values for 1500 data items
  - .864 voting precision for Flight; i.e., wrong values for 1000 data items

Improvement I. Using Source Accuracy

	S1	S2	S3
Flight 1	7:02PM	6:40PM	7:02PM
Flight 2	5:43PM	5:43PM	5:50PM
Flight 3	9:20AM	9:20AM	9:20AM
Flight 4	9:40PM	9:52PM	8:33PM
Flight 5	6:15PM	6:15PM	6:22PM

Improvement I. Using Source Accuracy

	S1	S2	S3
Flight 1	7:02PM	6:40PM	7:02PM
Flight 2	5:43PM	5:43PM	5:50PM
Flight 3	9:20AM	9:20AM	9:20AM
Flight 4	9:40PM	9:52PM	8:33PM
Flight 5	6:15PM	6:15PM	6:22PM

Higher accuracy;
More trustable

Naïve voting obtains an accuracy of 80%

Improvement I. Using Source Accuracy

	S1	S2	S3
Flight 1	7:02PM	6:40PM	7:02PM
Flight 2	5:43PM	5:43PM	5:50PM
Flight 3	9:20AM	9:20AM	9:20AM
Flight 4	9:40PM	9:52PM	8:33PM
Flight 5	6:15PM	6:15PM	6:22PM

Higher accuracy;
More trustable

Challenges:

1. How to decide source accuracy?
2. How to leverage accuracy in voting?

Considering accuracy obtains an accuracy of 100%

Source Accuracy: Bayesian Analysis

- Goal: $Pr(v_i(D) \text{ true} \mid \Phi_D(\mathbf{S}))$, for each D, $v_i(D)$

- According to Bayes Rule, we need to know
  - $Pr(\Phi_D(\mathbf{S}) \mid v_i(D) \text{ true})$, $Pr(v_i(D) \text{ true})$, for each $v_i(D)$
- $Pr(\Phi_D(\mathbf{S}) \mid v_i(D) \text{ true})$ can be computed as:
  - $\prod_{S \in \mathbf{S}(v_i(D))}(A(S)) * \prod_{S \in \mathbf{S} \setminus \mathbf{S}(v_i(D))}((1 - A(S))/n)$
- $Pr(v_i(D) \text{ true} \mid \Phi_D(\mathbf{S})) = e^{Conf(v_i(D))}/(\sum_{v_0(D)} e^{Conf(v_0(D))})$
  - $Conf(v_i(D)) = \sum_{S \in \mathbf{S}(v_i(D))} \ln(nA(S)/(1 - A(S)))$

- $A(S) = Avg_{\, v_i(D) \in S}\, Pr(v_i(D) \text{ true} \mid \Phi_D(\mathbf{S}))$

Computing Source Accuracy

- Source accuracy A(S)

$$A(S) = Avg_{\, v_i(D) \in S}\, Pr(v_i(D) \text{ true} \mid \Phi)$$

  - $v_i(D) \in S$: S provides value v_i on data item D
  - Φ : observations on all data items by sources **S**
  - $Pr(v_i(D) \text{ true} \mid \Phi)$: probability of $v_i(D)$ being true

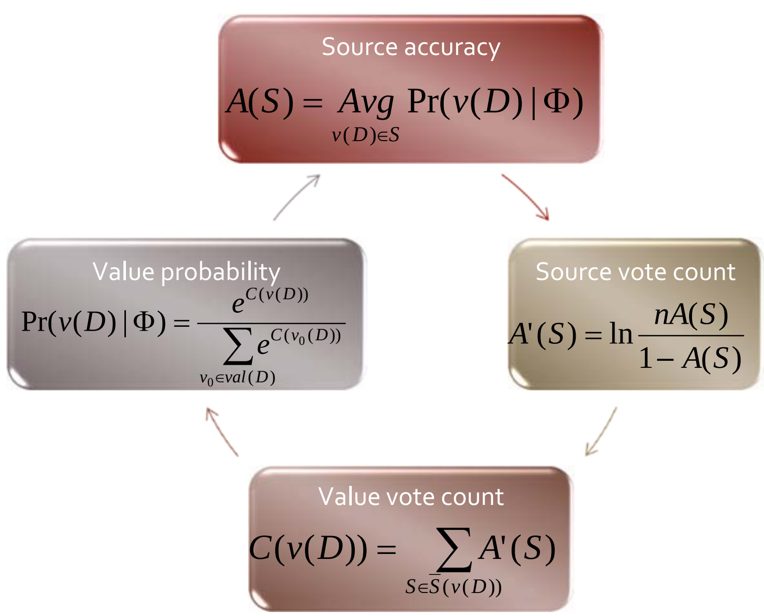
How to compute $Pr(v_i(D) \text{ true} \mid \Phi)$?

Using Source Accuracy in Data Fusion

- Input: data item D, val(D) = $\{v_0, v_1, \ldots, v_n\}$, Φ
- Output: Pr(v_i(D) true | Φ), for i=0,…, n (sum=1)
- Based on Bayes Rule, need Pr(Φ | v_i(D) true)
- Under independence, need Pr(Φ_D(S)|v_i(D) true)
  - If S provides v_i : Pr(Φ_D(S) |v_i(D) true) = A(S)
  - If S does not : Pr(Φ_D(S) |v_i(D) true) =(1-A(S))/n

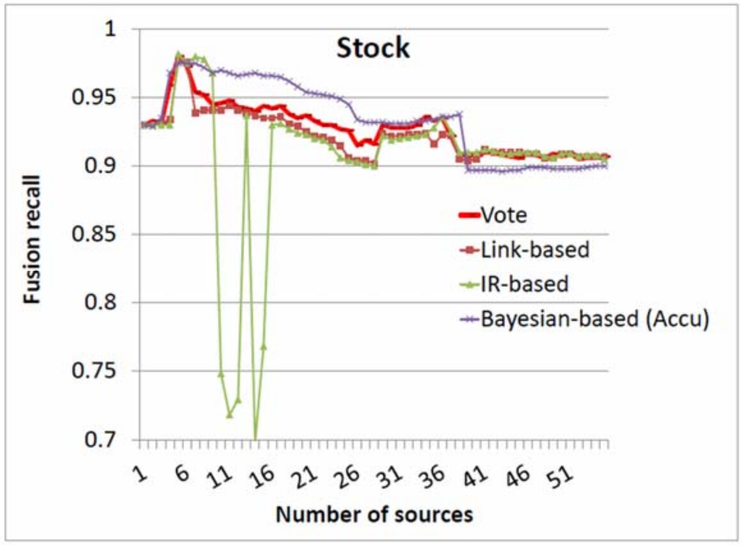
Challenge: How to handle inter-dependence between source accuracy and value probability?

Data Fusion Using Source Accuracy



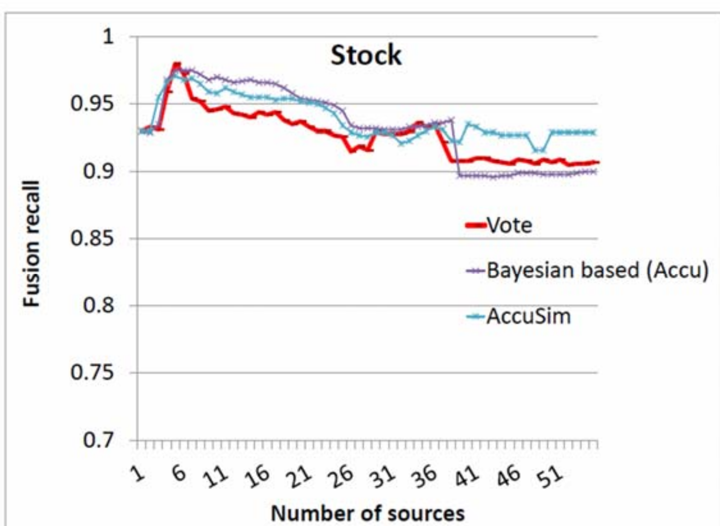
- Continue until source accuracy converges

Results on Stock Data (I)



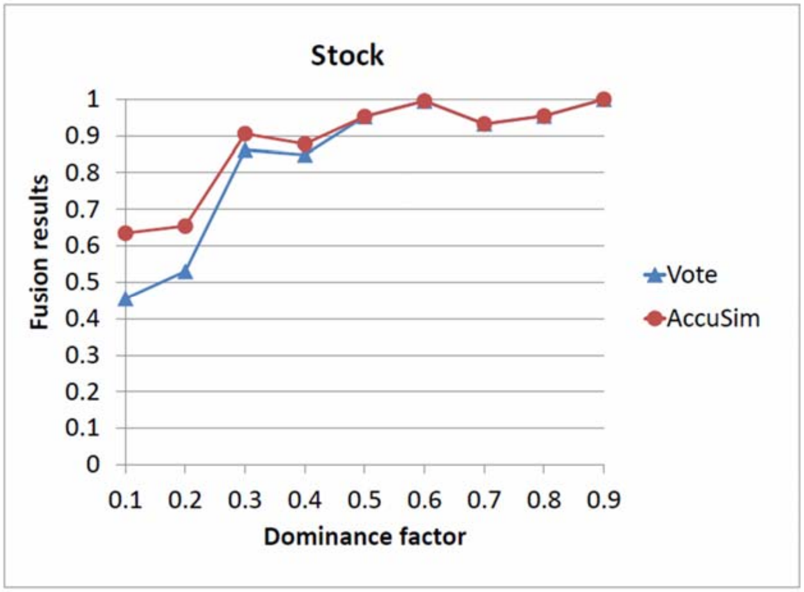
- ❑Sources ordered by recall (coverage * accuracy)
- ❑Among various methods, the Bayesian-based method (Accu) performs best at the beginning, but in the end obtains a final precision (=recall) of .900, worse than Vote (.908)

Results on Stock Data (II)

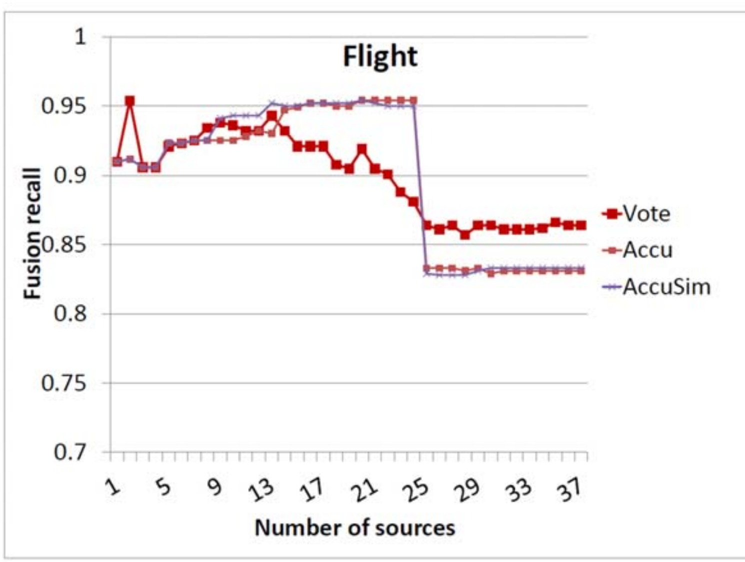


- ❑AccuSim obtains a final precision of .929, higher than Vote and any other method (around .908)
  - ❑ This translates to 350 more correct values

Results on Stock Data (III)



Results on Flight Data



- Accu/AccuSim obtain final precision of .831/.833, both lower than Vote (.857)
- WHY??? What is that magic source?

Copying on Erroneous Data

	Remarks	Size	Schema sim	Object sim	Value sim	Avg accu
Stock	Depen claimed	11	1	.99	.99	.92
	Depen claimed	2	1	1	.99	.75
Flight	Depen claimed	5	0.80	1	1	.71
	Query redirection	4	0.83	1	1	.53
	Dependence claimed	3	1	1	1	.92
	Embedded interface	2	1	1	1	.93
	Embedded interface	2	1	1	1	.61

Copying on Erroneous Data

	S1	S2	S3	S4	S5
Flight 1	7:02PM	6:40PM	7:02PM	7:02PM	8:02PM
Flight 2	5:43PM	5:43PM	5:50PM	5:50PM	5:50PM
Flight 3	9:20AM	9:20AM	9:20AM	9:20AM	9:20AM
Flight 4	9:40PM	9:52PM	8:33PM	8:33PM	8:33PM
Flight 5	6:15PM	6:15PM	6:22PM	6:22PM	6:22PM

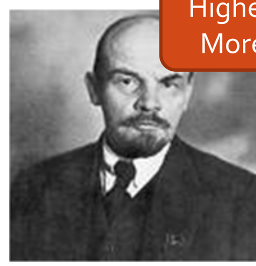
A lie told often enough becomes the truth.
—*Vladimir Lenin*

Copying on Erroneous Data

	S1	S2	S3	S4	S5
Flight 1	7:02PM	6:40PM	7:02PM	7:02PM	8:02PM
Flight 2	5:43PM	5:43PM	5:50PM	5:[illegible]PM	5:50PM
Flight 3	9:20AM	9:20AM	9:20AM	9:20[illegible]	9:20AM
Flight 4	9:40PM	9:52PM	8:33PM	8:33PM	8:33PM
Flight 5	6:15PM	6:15PM	6:22PM	6:22PM	[illegible]PM

Higher accuracy;
More trustable

A lie told often enough becomes the truth.
—*Vladimir Lenin*



Considering source accuracy can be worse when there is copying

Improvement II. Ignoring Copied Data

	S1	S2	S3	S4	S5
Flight 1	7:02PM	6:40PM	7:02PM	7:02PM	8:02PM
Flight 2	5:43PM	5:43PM	5:50PM	5:50PM	5:50PM
Flight 3	9:20AM	9:20AM	9:20AM	9:20AM	9:20AM
Flight 4	9:40PM	9:52PM	8:33PM	8:33PM	8:33PM
Flight 5	6:15PM	6:15PM	6:22PM	6:22PM	6:22PM

Challenges:

1. How to detect copying?
2. How to leverage copying in voting?

It is important to detect copying and ignore copied values in fusion

Copying?

Are Source 1 and Source 2 dependent? Not necessarily

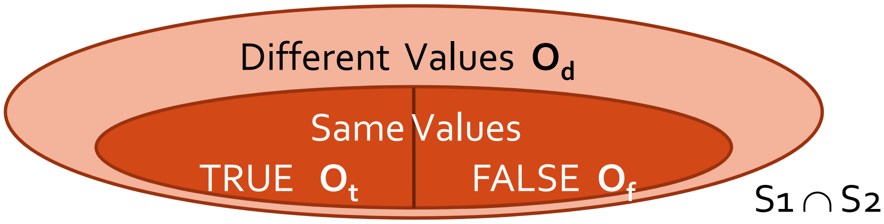
Source 1 on USA Presidents:	Source 2 on USA Presidents:	
1st : George Washington	1st : George Washington	✓
2nd : John Adams	2nd : John Adams	✓
3rd : Thomas Jefferson	3rd : Thomas Jefferson	✓
4th : James Madison	4th : James Madison	✓
…	…	
41st : George H.W. Bush	41st : George H.W. Bush	✓
42nd : William J. Clinton	42nd : William J. Clinton	✓
43rd : George W. Bush	43rd : George W. Bush	✓
44th: Barack Obama	44th: Barack Obama	✓

Copying? — Common Errors

Are Source 1 and Source 2 dependent? Very likely

Source 1 on USA Presidents:	Source 2 on USA Presidents:	
1st : George Washington	1st : George Washington	✓
2nd : Benjamin Franklin	2nd : Benjamin Franklin	✘
3rd : John F. Kennedy	3rd : John F. Kennedy	✘
4th : Abraham Lincoln	4th : Abraham Lincoln	✘
…	…	
41st : George W. Bush	41st : George W. Bush	✘
42nd : Hillary Clinton	42nd : Hillary Clinton	✘
43rd : Dick Cheney	43rd : Dick Cheney	✘
44th: Barack Obama	44th: John McCain	

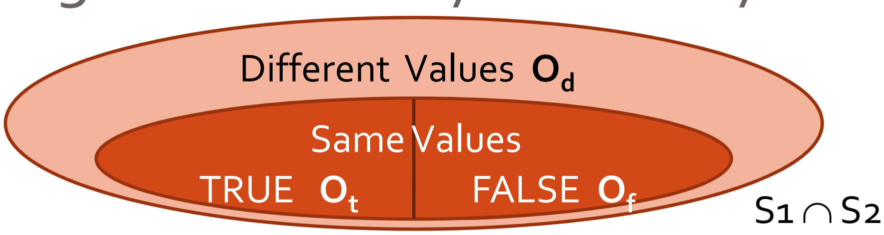
Copying Detection: Bayesian Analysis



- Goal: $\Pr(S1 \perp S2 \mid \Phi)$, $\Pr(S1 \sim S2 \mid \Phi)$ (sum = 1)

- According to Bayes Rule, we need to know
  - $\Pr(\Phi \mid S1 \perp S2)$, $\Pr(\Phi \mid S1 \sim S2)$
- Key: compute $\Pr(\Phi_D \mid S1 \perp S2)$, $\Pr(\Phi_D \mid S1 \sim S2)$
  - For each $D \in S1 \cap S2$

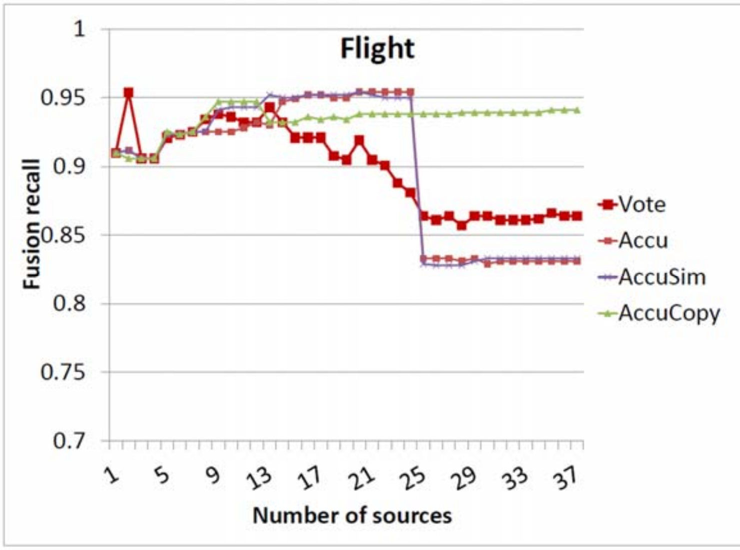
Copying Detection: Bayesian Analysis



Pr	Independence		Copying
O_t	A^2	$<$	$A \bullet c + A^2(1-c)$
O_f	$\frac{(1-A)^2}{n}$	$\ll$	$(1-A) \bullet c + \frac{(1-A)^2}{n}(1-c)$
O_d	$P_d = 1 - A^2 - \frac{(1-A)^2}{n}$	$>$	$P_d(1-c)$

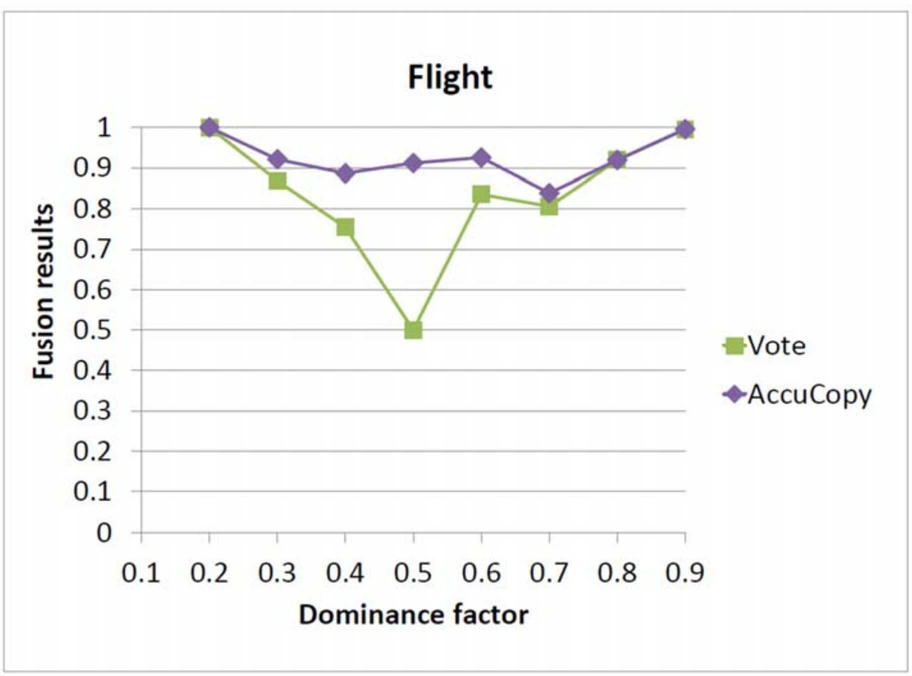
A-source accuracy; n-#wrong-values; c-copy rate

Results on Flight Data



- AccuCopy obtains a final precision of .943, much higher than Vote (.864)
  - This translates to 570 more correct values

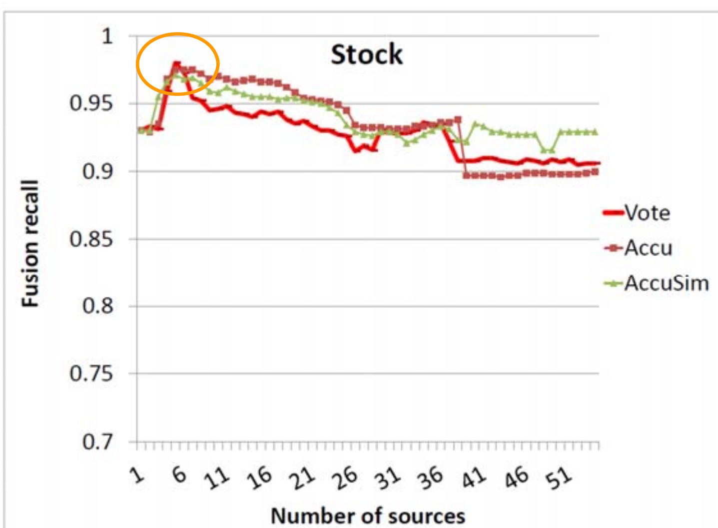
Results on Flight Data (II)



Take-Aways

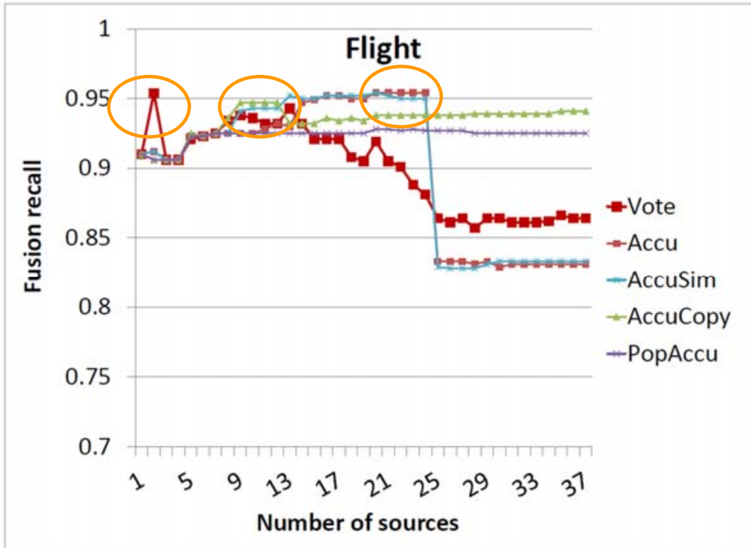
- Deep Web data is not fully trustable
  - Deep Web sources have different accuracies
  - Copying is common
- Truth finding on the Deep Web can leverage
  - source accuracy
  - copying relationships, and
  - value similarity

Important Direction: Source Selection



- Peaks happen before integrating all sources
- How to find the best set of sources while balancing quality gain and integration cost?

Important Direction: Source Selection



- Peaks happen before integrating all sources
- How to find the best set of sources while balancing quality gain and integration cost?

Acknowledgements

- Joint work with:
  - Xin Luna Dong, Yifan Hu, Ken Lyons (AT&T)
  - Laure Berti-Equille (IRD)
  - Xian Li, Weiyi Meng (SUNY-Binghamton)
- Selected research papers:
  - Truth Finding on the Deep Web: Is the Problem Solved? PVLDB 2013?
  - Global detection of complex copying relationships between sources. PVLDB 2010.
  - Integrating conflicting data: the role of source dependence. PVLDB 2009.


THANK YOU