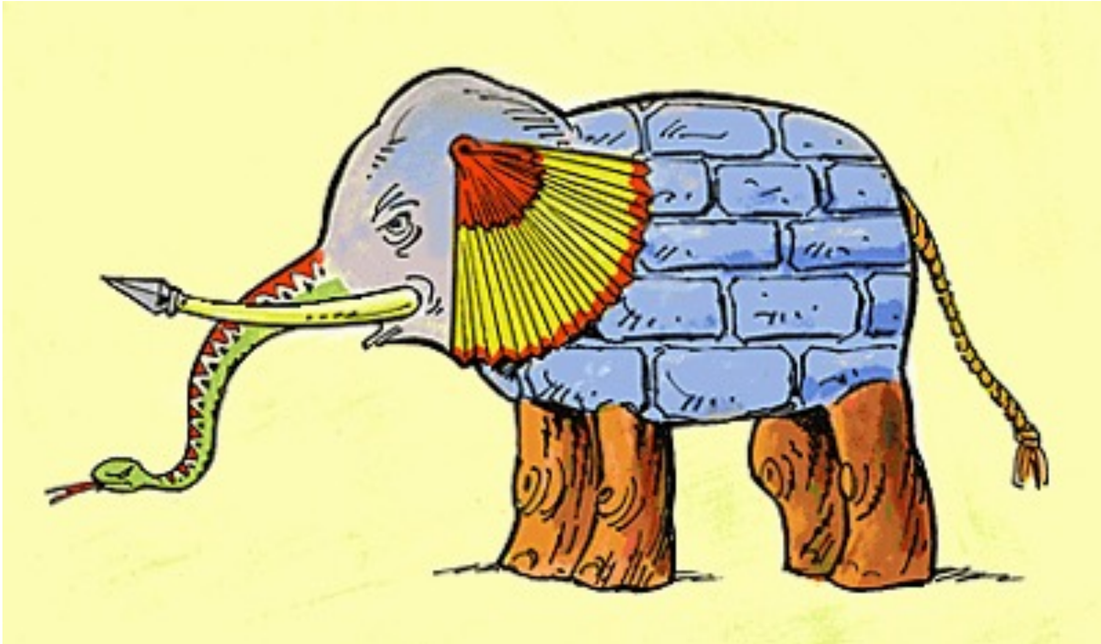
Towards High-Quality Big Data: A Focus on Time

Divesh Srivastava

Database Research, DSAIR, AT&T CDO

Big Data



- **Big Data** is different things to different people.
  - Volume, velocity, variety, variability, **value**, **veracity.**

Data Science

- Goal of **data science**: extract significant **value** from big data.

- Key stumbling block: data quality
  - Raw data is often of questionable **veracity.**
  - Data science using low quality data: **garbage in, garbage out.**

- Today's talk is on **data quality**, with a focus on **time.**

Outline

- Motivation.
  - Illustrative data quality examples.
  - “Small data” quality.
  - Towards big data quality.

- Obtaining high-quality long data.
  - Linking temporal records.
  - Discovering timestamp glitches.
  - The FIT family for real-time monitoring.

Data Quality: By the Numbers

- Impact of poor data quality.
  - In data science projects, data cleaning takes 30-80% of time/budget.
  - Erroneous data costs US businesses $600 billion/year [E02].
  - Data quality tools market is growing at 16% annually, way over 7% average for other IT segments [G07].
- How much data is erroneous.
  - Enterprise data error rates: average of 1-5%, some > 30% [R98].
- Next: **examples** to drive our intuitions, with a focus on **time** …

A Focus on Time

- Everything changes over time (abstracting Heraclitus).
  - Attributes of an entity evolve over time.

Divesh Srivastava (c.2000)

Divesh Srivastava (c.2020)

  - Different entities across time may have the same attributes.

Adam Smith (1723-1790)

Adam Smith (1965-)

Example: Changing Attributes Over Time

r1: Xin Dong
R. Polytechnic Institute

r2: Xin Dong
University of Washington

r3: Xin Dong
University of Washington

r4: Xin Luna Dong
University of Washington

r5: Xin Luna Dong
AT&T Labs-Research

r6: Xin Luna Dong
AT&T Labs-Research

1991 | 2004 | 2005 | 2006 | 2007 | 2008 | 2009 | 2010 | 2011

-Who is who?

r7: Dong Xin
University of Illinois

r8:Dong Xin
University of Illinois

r9: Dong Xin
Microsoft Research

r10: Dong Xin
University of Illinois

r11: Dong Xin
Microsoft Research

r12: Dong Xin
Microsoft Research

Example: Changing Attributes Over Time

dblp
computer science bibliography

search dblp

Xin Dong

> Home > Persons

by year | Dagstuhl

This is just a *disambiguation page*, and is not intended to be the bibliography of an actual person. The links to all actual bibliographies of persons of the same or a similar name can be found below. Any publication listed on this page has not been assigned to an actual author yet. If you know the true author of one of the publications listed below, you are welcome to contact us.

[–] Other persons with the same name

- Xin Dong 0001 (aka: Xin Luna Dong, Luna Dong) — Amazon (and 1 more)
- Xin Dong 0002 — Rensselaer Polytechnic Institute, Troy, USA
- Xin Dong 0003 — Zhejiang University, China
- Xin Dong 0004 — Northeastern University, Boston, USA
- Xin Dong 0005 — Central South University, Changsha, China
- Xin Dong 0006 — Communication University of China, Information Engineering School, Beijing, China
- Xin Dong 0007 — Shanghai Jiao Tong University
- Xin Dong 0008 — University of Nebraska-Lincoln
- Xin Dong 0009 — Harvard University, Cambridge, MA, USA (and 1 more)
- Xin Dong 0010 — Rutgers University, NJ, USA

[+] Other persons with a similar name

[–] 2020 – today

2020

[j13] Xin Dong, Yizhao Zhou, Lantian Wang, Jingfeng Peng, Yanbo Lou, Yiqun Fan:
Liver Cancer Detection Using Hybridized Fully Convolutional Neural Network Based on Deep Learning Framework. IEEE Access 8: 129889-129898 (2020)

[–] Refine list

showing all 33 records

refine by search term

Example: Changing Attributes Over Time

dblp.org/pid/35/7092.html

- Wei Wang 0266 — Iowa State University, Ames, IA, USA
- Wei Wang 0267 — Guangzhou Maritime University, GuanZhou, China
- Wei Wang 0268 — School of Resource and Environmental Sciences, Wuhan University, Wuhan, China
- Wei Wang 0269 — Beijing University of Chinese Medicine, Beijing, China
- Wei Wang 0270 — College of Information and Control Engineering, Nanjing University of Information Science And Technology, Nanjing, Jiangsu, China (and 1 more)
- Wei Wang 0271 — SER Group Ltd., Hong Kong (and 1 more)
- Wei Wang 0272 — Nanyang Technological University, Singapore
- Wei Wang 0273 — Beijing Institute of Technology, School of Information and Electronics, China

show less

[–] **Other persons with a similar name**

- Da-Wei Wang
- Liwei Wang (aka: Li-wei Wang, Li-Wei Wang) — *disambiguation page*
- Pengwei Wang (aka: PengWei Wang, Peng-Wei Wang) — *disambiguation page*
- Wei-Jen Wang
- Wei-Tsong Wang
- Wei-Yen Wang
- Jun-Wei Wang 0001 — University of Science and Technology Beijing, School of Automation and Electrical Engineering, China (and 1 more)
- Weifan Wang 0001 (aka: Wei-Fan Wang 0001) — Zhejiang Normal University, Department of Mathematics, Jinhua, China (and 2 more)
- Xingwei Wang 0001 (aka: Xing-Wei Wang 0001) — Northeastern University, College of Software, Shenyang, China
- Wang Wei — *disambiguation page*

show all similar names

[–] **2020 – today**

2021

[j560] Wei Wang, Lijuan Liu:
Complex L_p affine isoperimetric inequalities. Adv. Appl. Math. 122: 102108 (2021)

[j559] Wangli Hao, Ian Max Andolina, Wei Wang, Zhaoxiang Zhang:
Biologically inspired visual computing: the state of the art. Frontiers Comput. Sci. 15(1): 151304 (2021)

[j558] Junyang Chen, Zhiguo Gong, Wei Wang, Weiwen Liu:
HNS: Hierarchical negative sampling for network representation learning. Inf. Sci. 542: 343-356 (2021)

[–] **Refine list**

showing all 1326 records

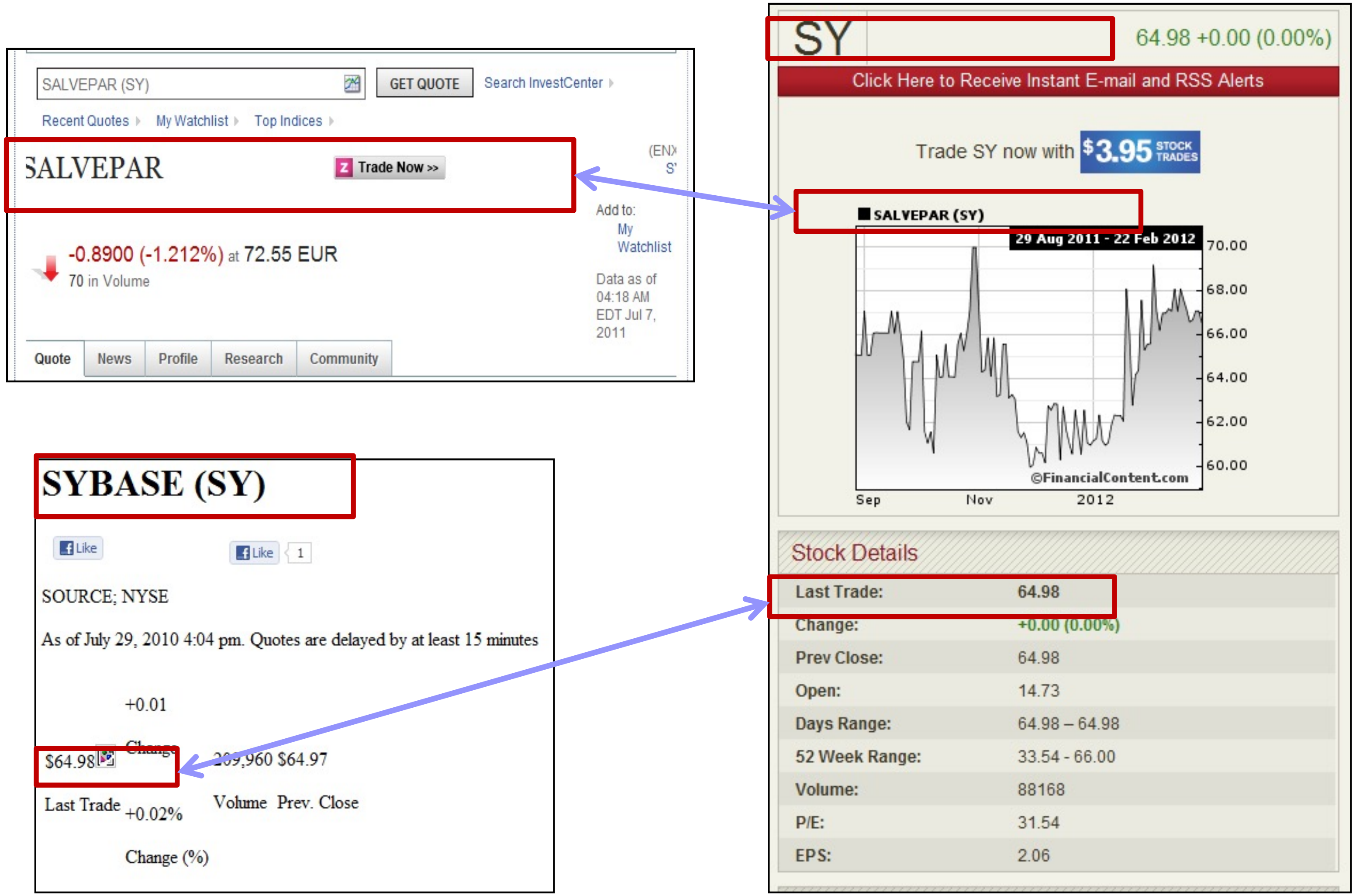
refine by search term

refine by type
- ☑ Journal Articles (only)
- ☑ Conference and Workshop Papers (only)
- ☑ Parts in Books or Collections (only)
- ☑ Editorship (only)
- ☑ Informal Publications (only)

select all | deselect all

refine by coauthor

Example: Instance Ambiguity Across Time



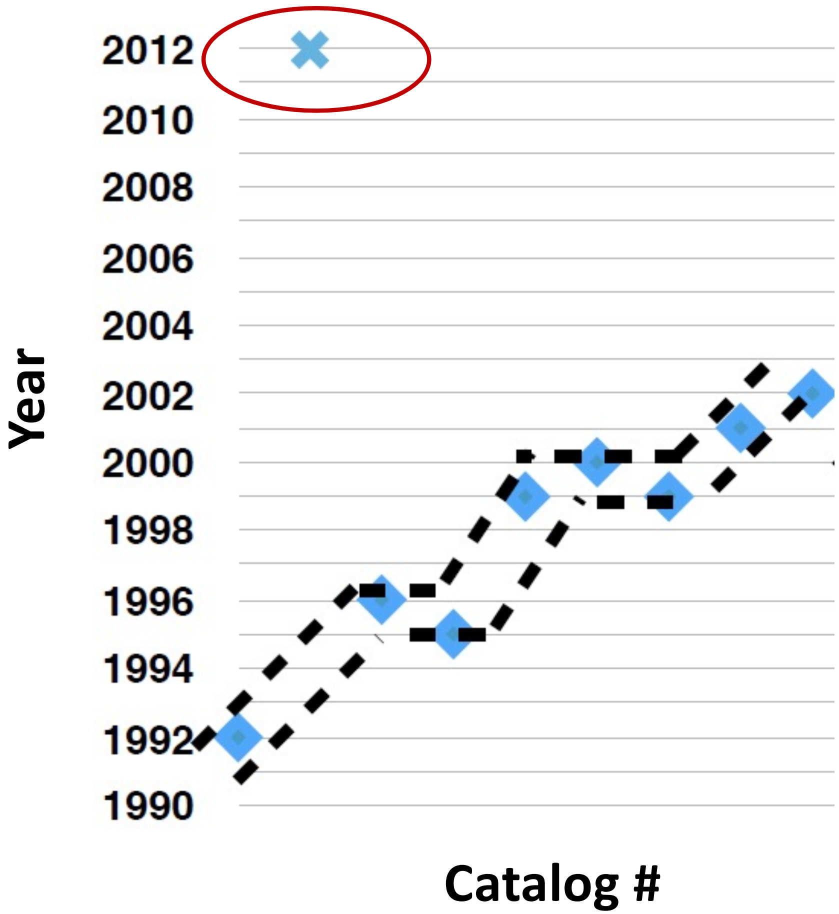
Example: Timestamps can be Erroneous

- Which record has an erroneous value of year?

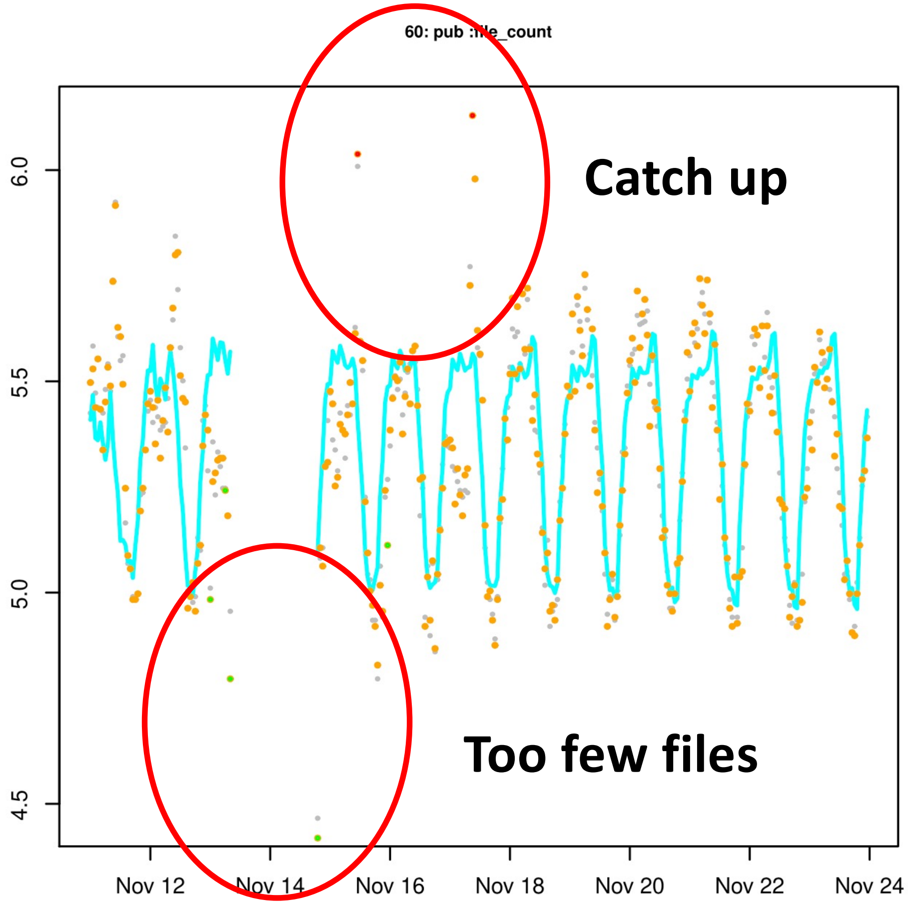
Tid	Release Title	Country	Year	Month	Catalog #
t1	Unplugged	Canada	1992	8	CDW45024
t2	Mirror Ball	Canada	2012	6	CDW45934
t3	Ether	Canada	1996	2	CDW46012
t4	Insomniac	Canada	1995	10	CDW46046
t5	Summerteeth	Canada	1999	3	CDW47282
t6	Sonic Jihad	Canada	2000	7	CDW47383
T7	Title of …	Canada	1999	7	CDW47388
t8	Reptile	Canada	2001	3	CDW47966
t9	Always …	Canada	2002	2	CDW48016

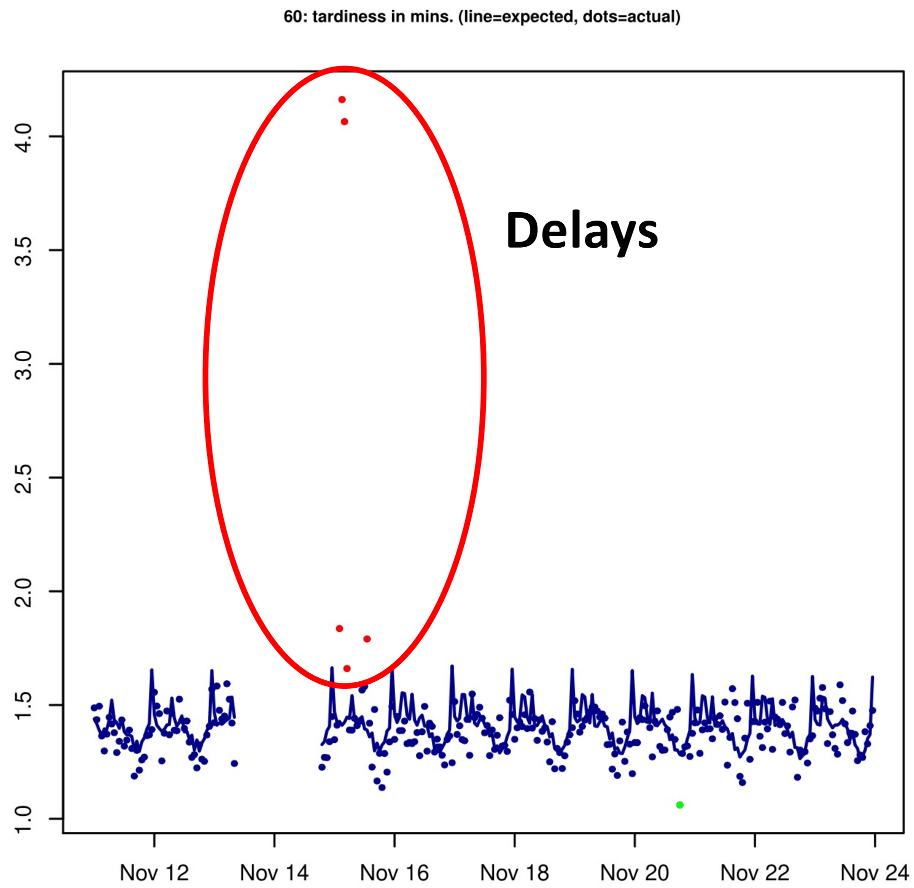
Example: Timestamps can be Erroneous

- Which record has an erroneous value of year?

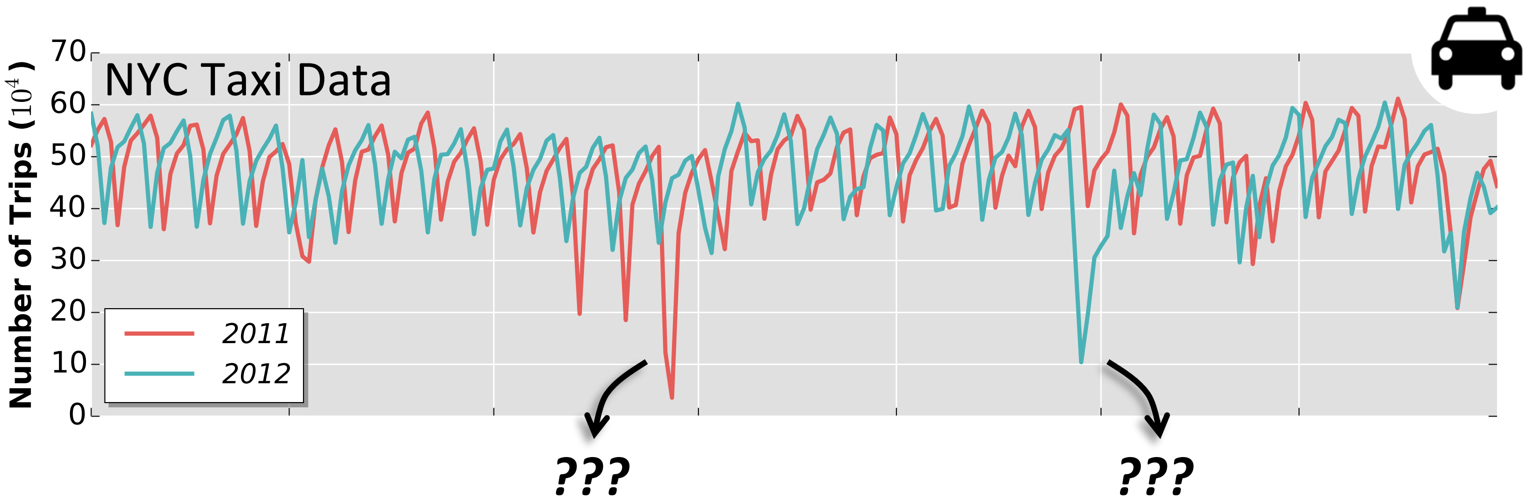


Example: Delayed Data Arrival Over Time

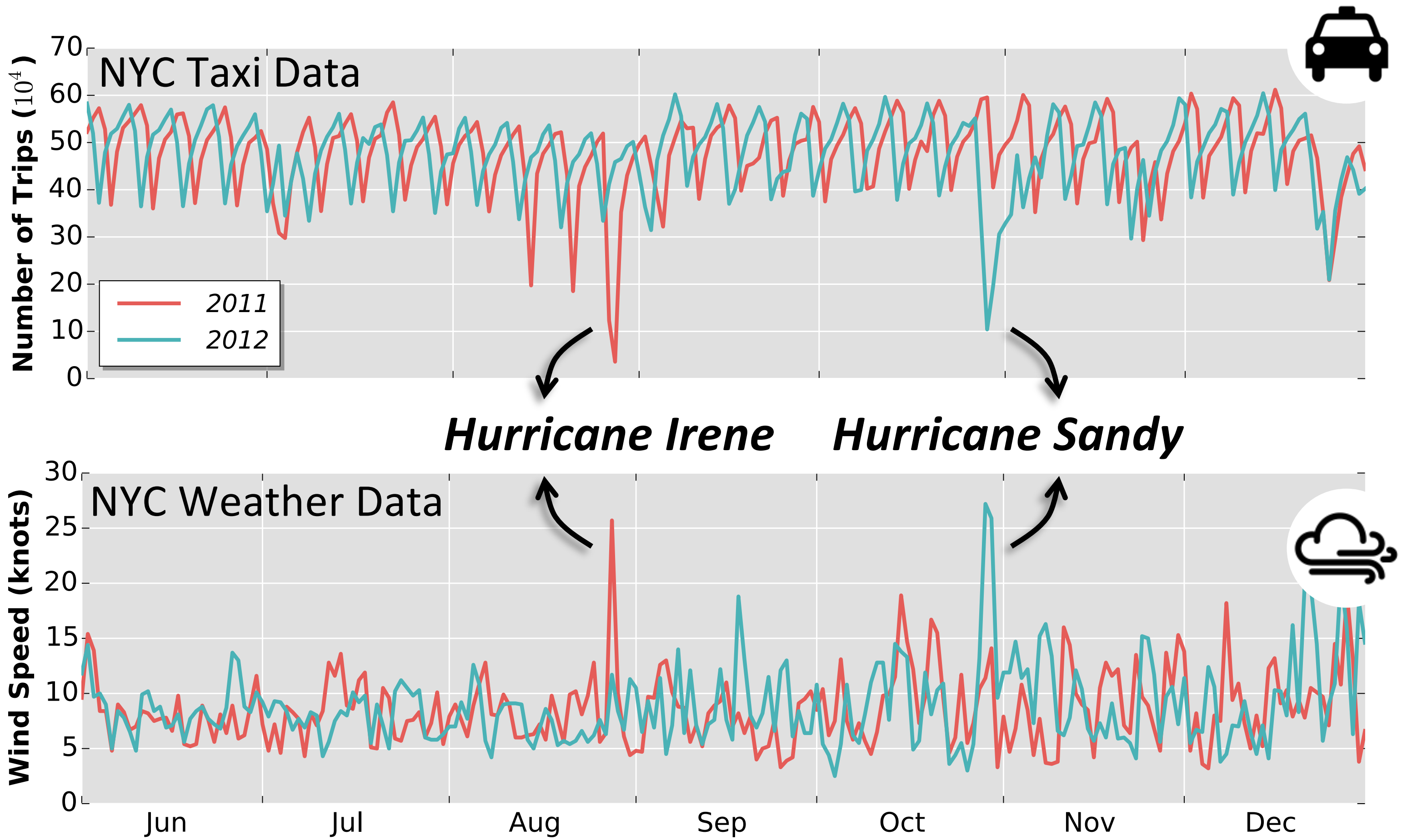




Example: Time Series Anomalies



Example: Correlated Time Series Anomalies



Examples: Lessons Learned

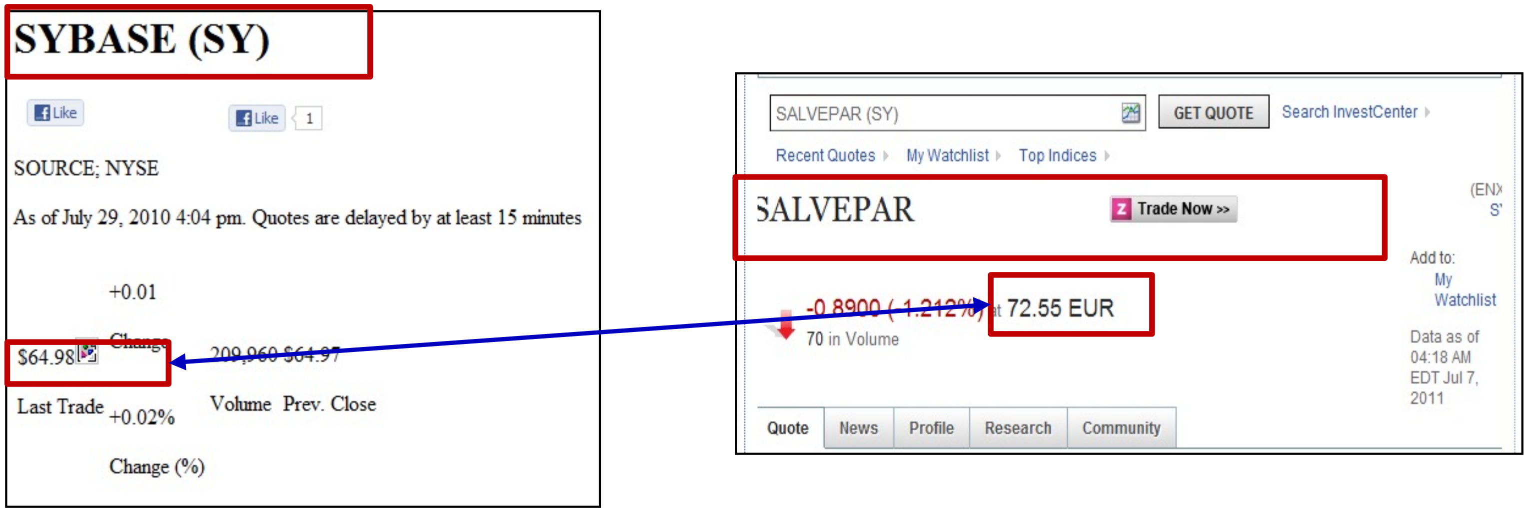
- Big data over time (i.e., long data) can have **veracity** issues.
  - Even in domains where poor-quality data can have big impact.
  - Diversity of data quality issues involving time.

- Obtaining high-quality long data is **challenging**!
  - How soon can missing, erroneous and biased data be identified?
  - Which data can be used and when can it be used by data science?

Small Data Quality: How Was It Achieved?

- Specify **all** domain knowledge as **integrity constraints** on data.

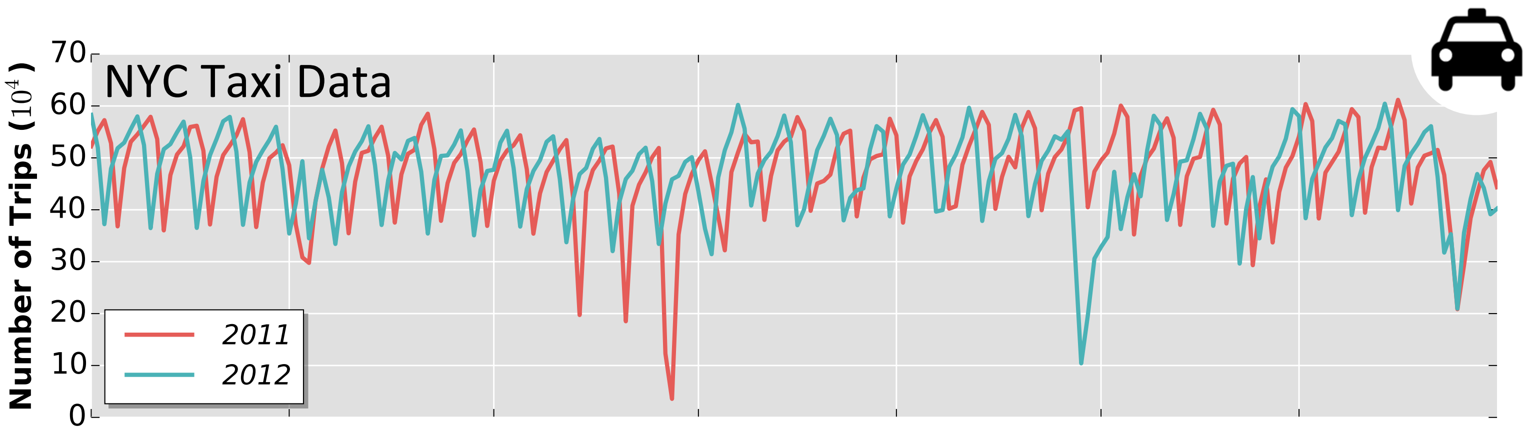
- Integrity constraint: formal specification that data must satisfy.
  - **Semantic** (SSN unique for person) vs **syntactic** (NNN-NN-NNNN).
  - **Qualitative** (FD on closing price) …



Small Data Quality: How Was It Achieved?

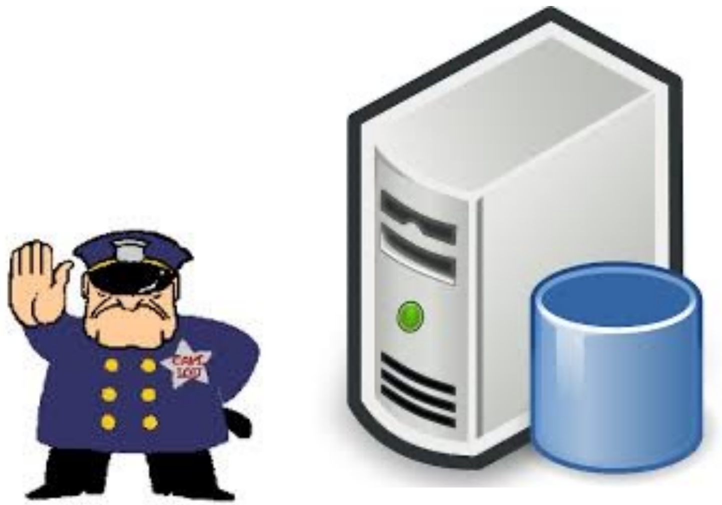
- Specify **all** domain knowledge as **integrity constraints** on data.

- Integrity constraint: formal specification that data must satisfy.
  - **Semantic** (SSN unique for person) vs **syntactic** (NNN-NN-NNNN).
  - **Qualitative** (FD on closing price) vs **quantitative** (# trips in 3σ of μ).



Small Data Quality: How Was It Achieved?

- Specify all domain knowledge as **integrity constraints** on data.
  - **Reject updates** that do not preserve integrity constraints.
  - Works well when the domain is **very well understood** and **static**.



Big Data Quality: A Different Approach?

- Big data: integrity constraints cannot be specified a priori.
  - Data **variety, volume** → complete domain knowledge is infeasible.
  - Data **velocity, variability** → domain knowledge becomes obsolete.
  - Too much rejected data → “small” data. ☺

Big Data Quality: A Different Approach?

- Big data: integrity constraints cannot be specified a priori.
  - Data **variety, volume** → complete domain knowledge is infeasible.
  - Data **velocity, variability** → domain knowledge becomes obsolete.
  - Too much rejected data → “small” data. ☺

- Solution: let the data speak for itself.
  - Learn **integrity constraints / models** (semantics) from the data.
  - Identify **data glitches** as violations of the learned models.
  - Repair **data glitches and models** in a timely manner.

Obtaining High-Quality Long Data

- What is special about time?
  - Time can be modeled as an ordered domain.
  - Everything happens in time; everything changes over time.

- A large variety of techniques to obtain high-quality long data.
  - **Linking temporal records** [LDM+11, LWT+12].
  - Data fusion over time [DBS09].
  - Discovering order dependencies [SGG+17, SGG+18], **band ODs** [LSB+20], **ABC ODs** [LSB+21].
  - **The FIT family** [DSS+15, DDS16, BDK+19, BDK+21].
  - Correlated time series anomalies [BDF+21].

Outline

- Motivation.
- Obtaining high-quality long data.
  - Linking temporal records.
  - Discovering timestamp glitches.
  - The FIT family for real-time monitoring.

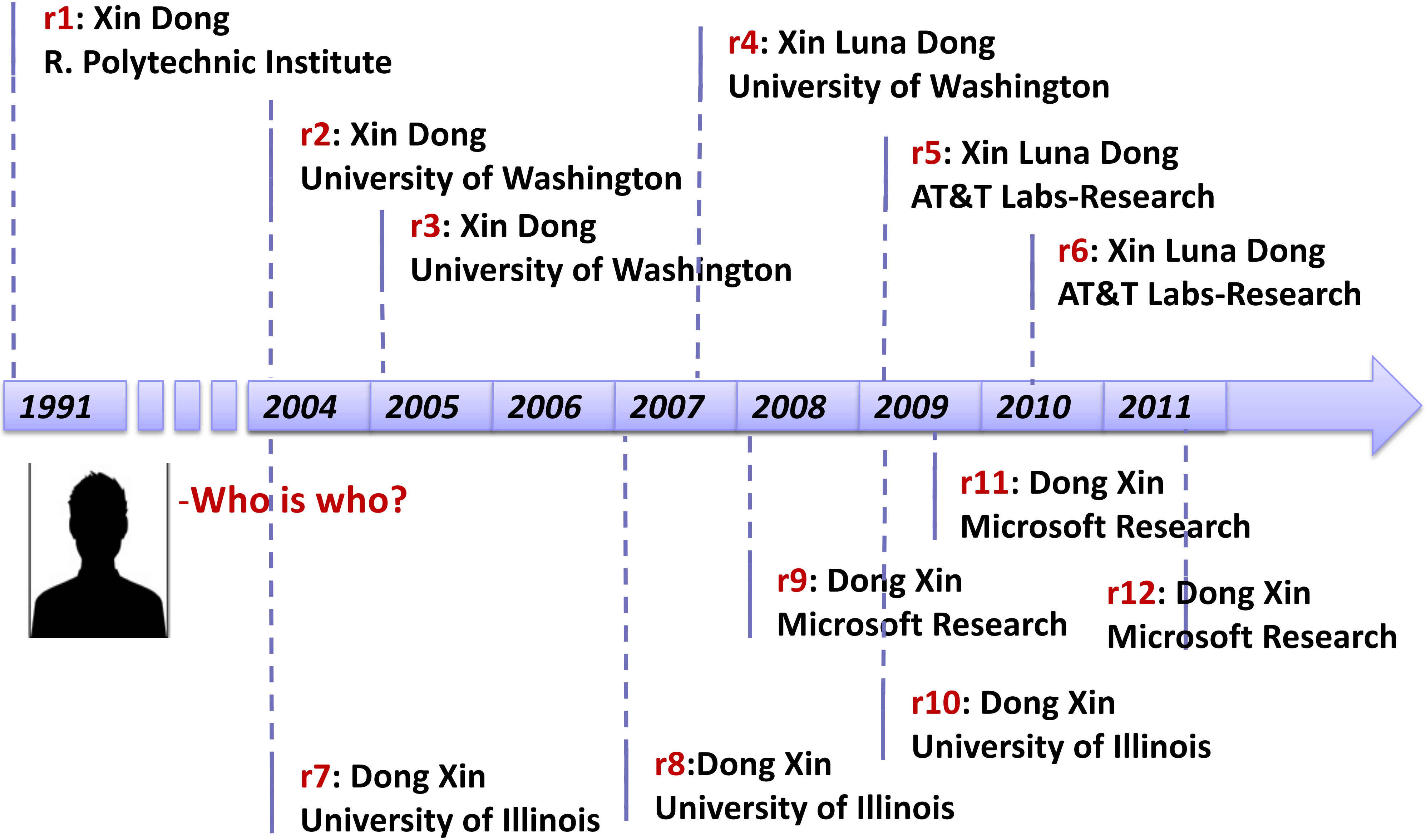
Linking Temporal Records

- Traditional record linkage.
  - Links records of an entity from multiple sources **at a point in time**.
  - Literature spanning 50+ years: statistical, rule-based, ML-based.

- Record linkage in long data [LDM+11, LWT+12]
  - Links records of an entity **over a long time period.**
  - Attributes of an entity evolve over time
  - Different entities across time may have the same attributes.

- Focus: insights that distinguish record linkage in long data.

Linking Temporal Records



Linking Temporal Records

r1: Xin Dong
R. Polytechnic Institute

r2: Xin Dong
University of Washington

r3: Xin Dong
University of Washington

r4: Xin Luna Dong
University of Washington

r5: Xin Luna Dong
AT&T Labs-Research

r6: Xin Luna Dong
AT&T Labs-Research

1991 | | | | *2004* | *2005* | *2006* | *2007* | *2008* | *2009* | *2010* | *2011*

-Traditional solution 1:
high value consistency

r11: Dong Xin
Microsoft Research

r9: Dong Xin
Microsoft Research

r12: Dong Xin
Microsoft Research

r10: Dong Xin
University of Illinois

r7: Dong Xin
University of Illinois

r8:Dong Xin
University of Illinois

Linking Temporal Records

r1: Xin Dong
R. Polytechnic Institute

r2: Xin Dong
University of Washington

r3: Xin Dong
University of Washington

r4: Xin Luna Dong
University of Washington

r5: Xin Luna Dong
AT&T Labs-Research

r6: Xin Luna Dong
AT&T Labs-Research

1991 | 2004 | 2005 | 2006 | 2007 | 2008 | 2009 | 2010 | 2011

-Traditional solution 2:
using similar names

r7: Dong Xin
University of Illinois

r8:Dong Xin
University of Illinois

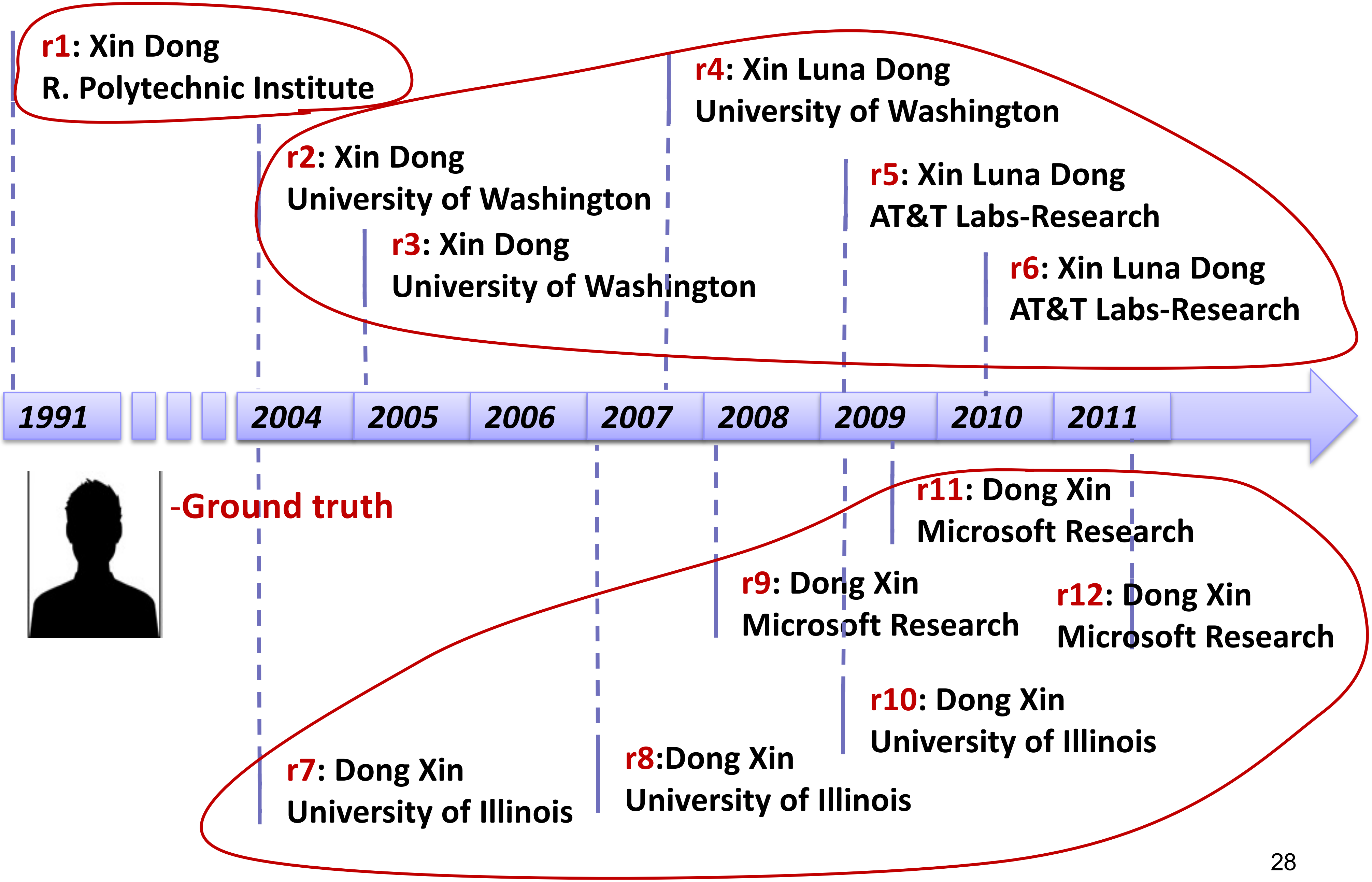
r9: Dong Xin
Microsoft Research

r10: Dong Xin
University of Illinois

r11: Dong Xin
Microsoft Research

r12: Dong Xin
Microsoft Research

Linking Temporal Records



Linking Temporal Records: Insights

- Smooth transition in one attribute, despite changes in another.

ID	Name	Affiliation	Co-authors	Year
r1	Xin Dong	R. Polytechnic Institute	Wozny	1991
r2	Xin Dong	University of Washington	Halevy, Tatarinov	2004
r7	Dong Xin	University of Illinois	Han, Wah	2004
r3	Xin Dong	University of Washington	Halevy	2005
r4	Xin Luna Dong	University of Washington	Halevy, Yu	2007
r8	Dong Xin	University of Illinois	Wah	2007
r9	Dong Xin	Microsoft Research	Wu, Han	2008
r10	Dong Xin	University of Illinois	Ling, He	2009
r11	Dong Xin	Microsoft Research	Chaudhuri, Ganti	2009
r5	Xin Luna Dong	AT&T Labs-Research	Das Sarma, Halevy	2009
r6	Xin Luna Dong	AT&T Labs-Research	Naumann	2010
r12	Dong Xin	Microsoft Research	He	2011

Linking Temporal Records: Insights

- Erratic changes in an attribute value are quite unlikely.

ID	Name	Affiliation	Co-authors	Year
r1	Xin Dong	R. Polytechnic Institute	Wozny	1991
r2	Xin Dong	University of Washington	Halevy, Tatarinov	2004
r7	Dong Xin	University of Illinois	Han, Wah	2004
r3	Xin Dong	University of Washington	Halevy	2005
r4	Xin Luna Dong	University of Washington	Halevy, Yu	2007
r8	Dong Xin	University of Illinois	Wah	2007
r9	Dong Xin	Microsoft Research	Wu, Han	2008
r10	Dong Xin	University of Illinois	Ling, He	2009
r11	Dong Xin	Microsoft Research	Chaudhuri, Ganti	2009
r5	Xin Luna Dong	AT&T Labs-Research	Das Sarma, Halevy	2009
r6	Xin Luna Dong	AT&T Labs-Research	Naumann	2010
r12	Dong Xin	Microsoft Research	He	2011

Linking Temporal Records: Insights

- Typically, there is continuity of history, i.e., no big gaps in time.

ID	Name	Affiliation	Co-authors	Year
r1	Xin Dong	R. Polytechnic Institute	Wozny	1991
r2	Xin Dong	University of Washington	Halevy, Tatarinov	2004
r7	Dong Xin	University of Illinois	Han, Wah	2004
r3	Xin Dong	University of Washington	Halevy	2005
r4	Xin Luna Dong	University of Washington	Halevy, Yu	2007
r8	Dong Xin	University of Illinois	Wah	2007
r9	Dong Xin	Microsoft Research	Wu, Han	2008
r10	Dong Xin	University of Illinois	Ling, He	2009
r11	Dong Xin	Microsoft Research	Chaudhuri, Ganti	2009
r5	Xin Luna Dong	AT&T Labs-Research	Das Sarma, Halevy	2009
r6	Xin Luna Dong	AT&T Labs-Research	Naumann	2010
r12	Dong Xin	Microsoft Research	He	2011

Linking Temporal Records: Solution Insights

- High penalty for value disagreement over a short time period.

ID	Name	Affiliation	Co-authors	Year
r1	Xin Dong	R. Polytechnic Institute	Wozny	1991
r2	Xin Dong	University of Washington	Halevy, Tatarinov	2004
r7	Dong Xin	University of Illinois	Han, Wah	2004
r3	Xin Dong	University of Washington	Halevy	2005
r4	Xin Luna Dong	University of Washington	Halevy, Yu	2007
r8	Dong Xin	University of Illinois	Wah	2007
r9	Dong Xin	Microsoft Research	Wu, Han	2008
r10	Dong Xin	University of Illinois	Ling, He	2009
r11	Dong Xin	Microsoft Research	Chaudhuri, Ganti	2009
r5	Xin Luna Dong	AT&T Labs-Research	Das Sarma, Halevy	2009
r6	Xin Luna Dong	AT&T Labs-Research	Naumann	2010
r12	Dong Xin	Microsoft Research	He	2011

Linking Temporal Records: Solution Insights

- Lower penalty for value disagreement over a long time period.

ID	Name	Affiliation	Co-authors	Year
r1	Xin Dong	R. Polytechnic Institute	Wozny	1991
r2	Xin Dong	University of Washington	Halevy, Tatarinov	2004
r7	Dong Xin	University of Illinois	Han, Wah	2004
r3	Xin Dong	University of Washington	Halevy	2005
r4	Xin Luna Dong	University of Washington	Halevy, Yu	2007
r8	Dong Xin	University of Illinois	Wah	2007
r9	Dong Xin	Microsoft Research	Wu, Han	2008
r10	Dong Xin	University of Illinois	Ling, He	2009
r11	Dong Xin	Microsoft Research	Chaudhuri, Ganti	2009
r5	Xin Luna Dong	AT&T Labs-Research	Das Sarma, Halevy	2009
r6	Xin Luna Dong	AT&T Labs-Research	Naumann	2010
r12	Dong Xin	Microsoft Research	He	2011

Linking Temporal Records: Solution

- High reward for value agreement across a small gap in time.

ID	Name	Affiliation	Co-authors	Year
r1	Xin Dong	R. Polytechnic Institute	Wozny	1991
r2	Xin Dong	University of Washington	Halevy, Tatarinov	2004
r7	Dong Xin	University of Illinois	Han, Wah	2004
r3	Xin Dong	University of Washington	Halevy	2005
r4	Xin Luna Dong	University of Washington	Halevy, Yu	2007
r8	Dong Xin	University of Illinois	Wah	2007
r9	Dong Xin	Microsoft Research	Wu, Han	2008
r10	Dong Xin	University of Illinois	Ling, He	2009
r11	Dong Xin	Microsoft Research	Chaudhuri, Ganti	2009
r5	Xin Luna Dong	AT&T Labs-Research	Das Sarma, Halevy	2009
r6	Xin Luna Dong	AT&T Labs-Research	Naumann	2010
r12	Dong Xin	Microsoft Research	He	2011

Linking Temporal Records: Solution Insights

- Lower reward for value agreement across a big gap in time.

ID	Name	Affiliation	Co-authors	Year
r1	Xin Dong	R. Polytechnic Institute	Wozny	1991
r2	Xin Dong	University of Washington	Halevy, Tatarinov	2004
r7	Dong Xin	University of Illinois	Han, Wah	2004
r3	Xin Dong	University of Washington	Halevy	2005
r4	Xin Luna Dong	University of Washington	Halevy, Yu	2007
r8	Dong Xin	University of Illinois	Wah	2007
r9	Dong Xin	Microsoft Research	Wu, Han	2008
r10	Dong Xin	University of Illinois	Ling, He	2009
r11	Dong Xin	Microsoft Research	Chaudhuri, Ganti	2009
r5	Xin Luna Dong	AT&T Labs-Research	Das Sarma, Halevy	2009
r6	Xin Luna Dong	AT&T Labs-Research	Naumann	2010
r12	Dong Xin	Microsoft Research	He	2011

Linking Temporal Records: Solution Insights

- Consider records in time order for linkage.

ID	Name	Affiliation	Co-authors	Year
r1	Xin Dong	R. Polytechnic Institute	Wozny	1991
r2	Xin Dong	University of Washington	Halevy, Tatarinov	2004
r7	Dong Xin	University of Illinois	Han, Wah	2004
r3	Xin Dong	University of Washington	Halevy	2005
r4	Xin Luna Dong	University of Washington	Halevy, Yu	2007
r8	Dong Xin	University of Illinois	Wah	2007
r9	Dong Xin	Microsoft Research	Wu, Han	2008
r10	Dong Xin	University of Illinois	Ling, He	2009
r11	Dong Xin	Microsoft Research	Chaudhuri, Ganti	2009
r5	Xin Luna Dong	AT&T Labs-Research	Das Sarma, Halevy	2009
r6	Xin Luna Dong	AT&T Labs-Research	Naumann	2010
r12	Dong Xin	Microsoft Research	He	2011

Linking Temporal Records: Results

- Quality experiments.
  - 2 real data sets (XD and WW from DBLP), 9 discovery algorithms.
  - F-1 of proposed approach > 0.9.

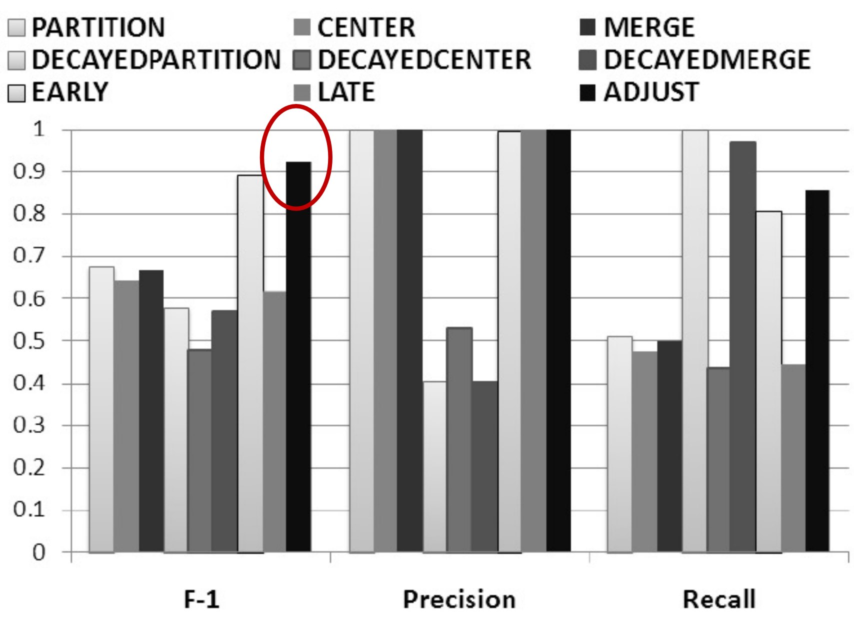


Figure 9: Results on *XD* set.

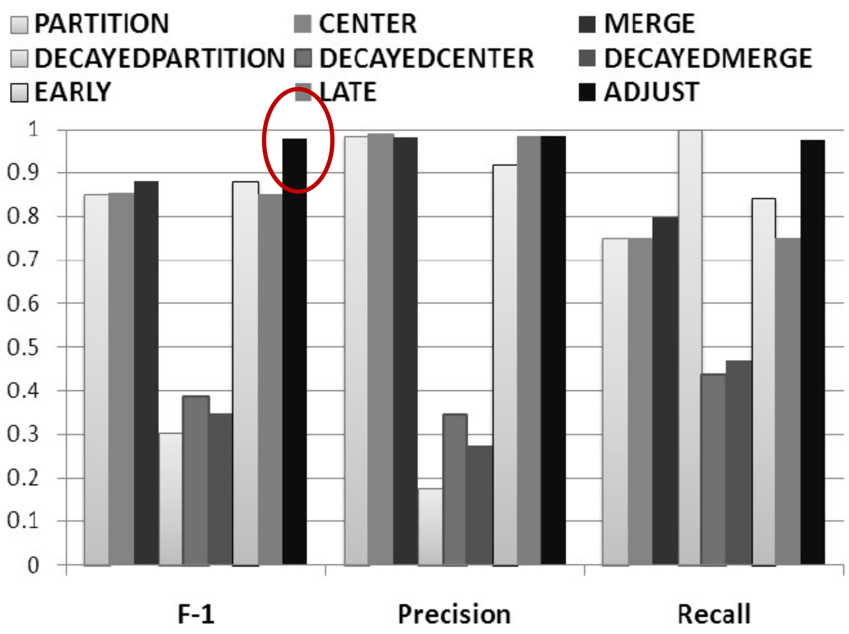


Figure 10: Results on *WW* set.

Outline

- Motivation.
- Obtaining high-quality long data.
  - Linking temporal records.
  - Discovering timestamp glitches.
  - The FIT family for real-time monitoring.

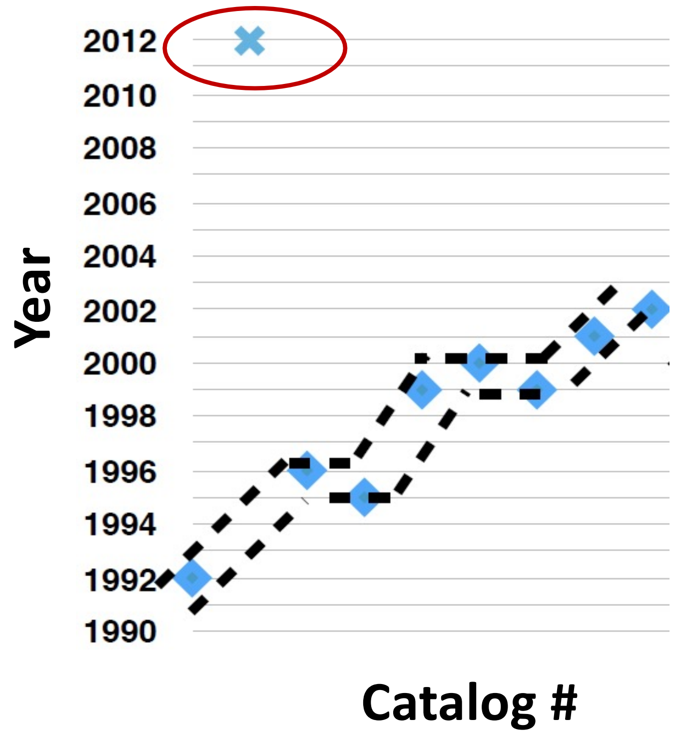
Discovering Timestamp Glitches

- Time plays a critical role in data science models.
  - Errors in timestamps can have serious consequences on models.

Tid	Release Title	Country	Year	Month	Catalog #
t1	Unplugged	Canada	1992	8	CDW45024
t2	Mirror Ball	Canada	2012	6	CDW45934
t3	Ether	Canada	1996	2	CDW46012
t4	Insomniac	Canada	1995	10	CDW46046
t5	Summerteeth	Canada	1999	3	CDW47282
t6	Sonic Jihad	Canada	2000	7	CDW47383
T7	Title of …	Canada	1999	7	CDW47388
t8	Reptile	Canada	2001	3	CDW47966
t9	Always …	Canada	2002	2	CDW48016

Discovering Timestamp Glitches: Challenges

- Idea: Use correlated ordered attributes in data to find **anomalies**.
- Semantic challenges.
  - Not all orderings are meaningful (e.g., lexicographic order on Release Title).
  - Correlations may be **non-strict** (e.g, Catalog # vs. Year).
- Efficiency challenges.
  - Large space of candidate attributes.
  - Data are big / long.



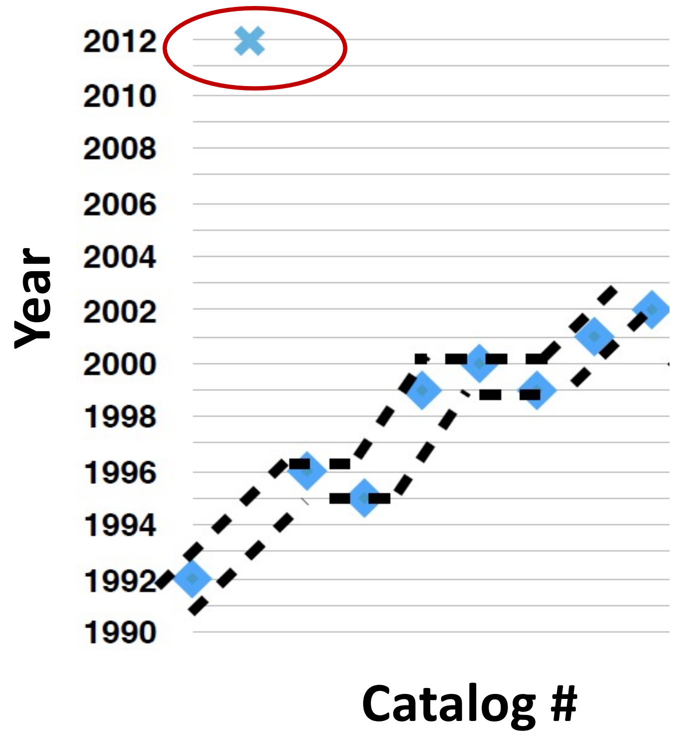
Discovering Timestamp Glitches: Solution I

- Idea: Use **non-strict** correlated ordered attributes in data to find **anomalies**.

- Step 1: Efficiently identify approx. OD between Catalog # and Year [SGG+17].

- Step 2: Explore candidate **longest monotonic bands (LMBs)** to determine optimal band width [LSB+20].

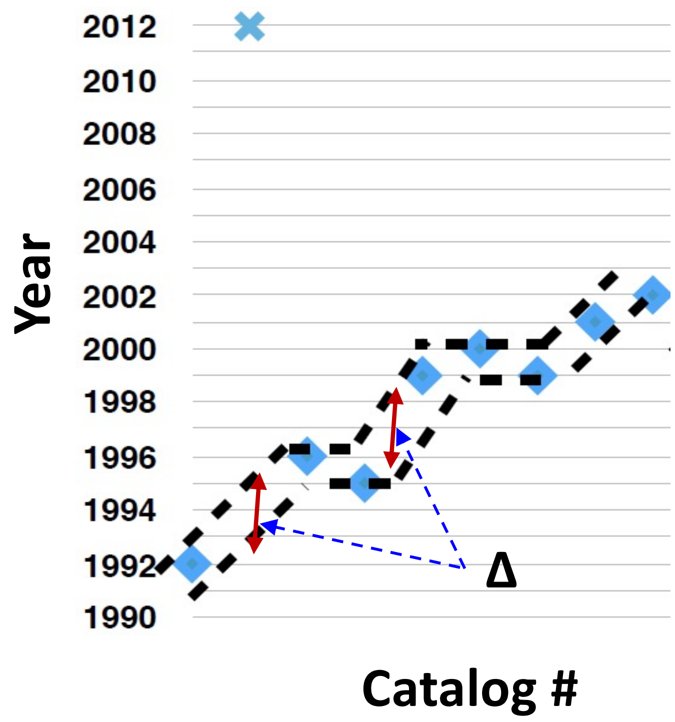
- Step 3: Use optimal band width to learn **AB OD model + glitches** [LSB+20].



Longest Monotonic Band

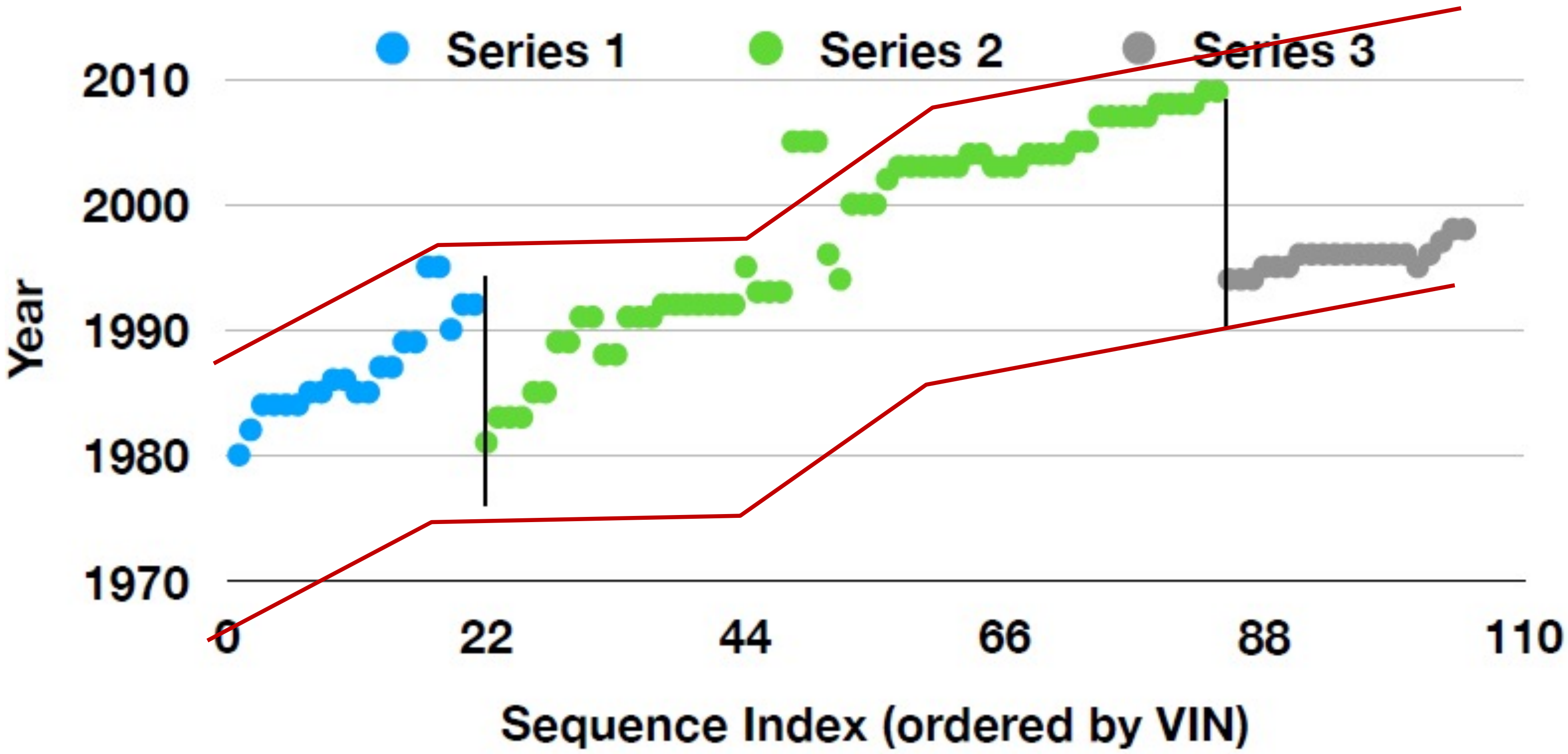
- Intuition: Longest subsequence of data inside a band of given width Δ whose lower and upper bounds are monotonic.

- Computing LMBs.
  - Generalizes the LIS problem.
  - $O(n^2)$ DP algorithm [LSB+20].
  - O(n*log(n)) DP algorithm [LSB+21].



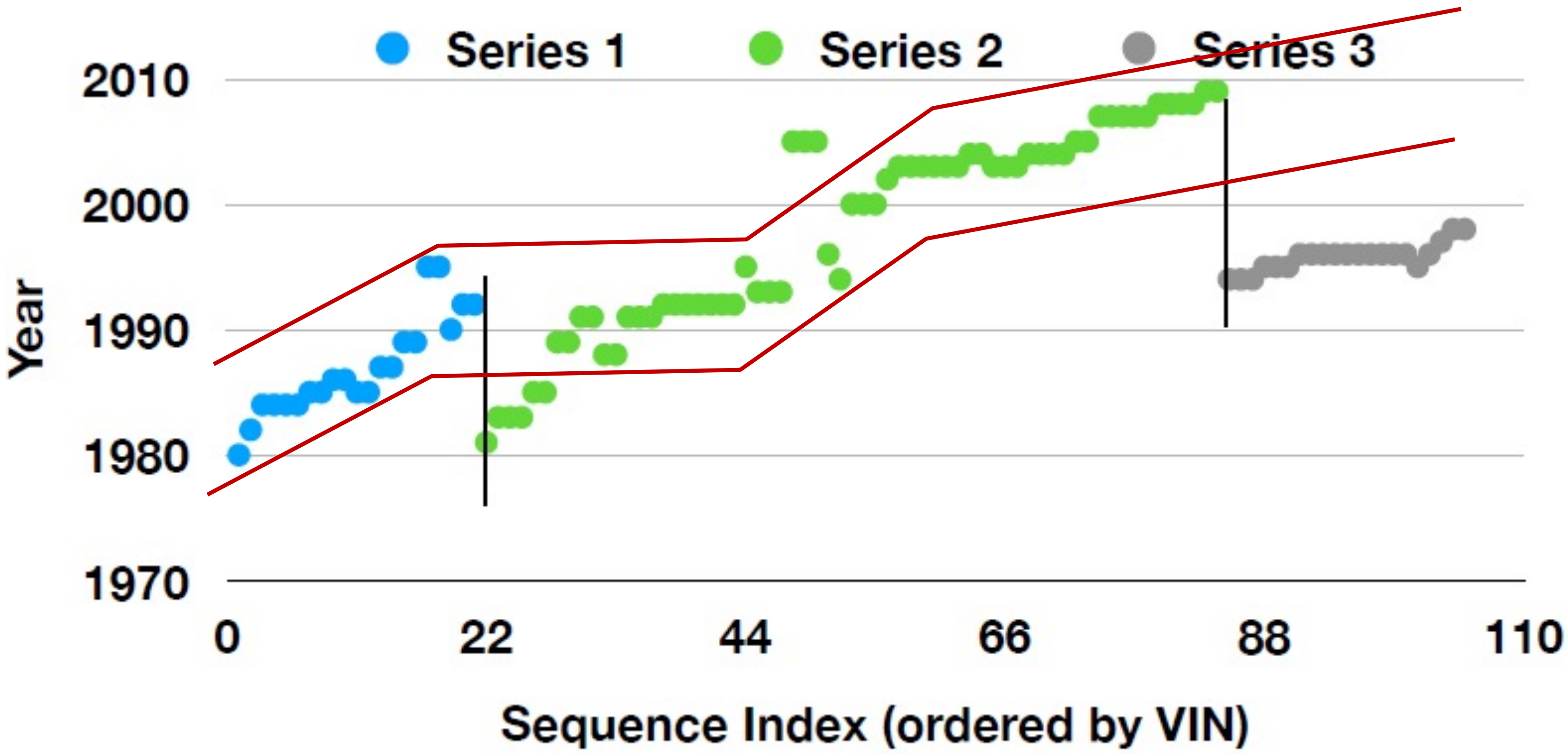
Discovering Timestamp Glitches: Challenges

- Using one approx. band OD to fit data may not always be ideal.
  - The band width is too large.



Discovering Timestamp Glitches: Challenges

- Using one approx. band OD to fit data may not always be ideal.
  - The band width is too large, or
  - There are too many anomalies.

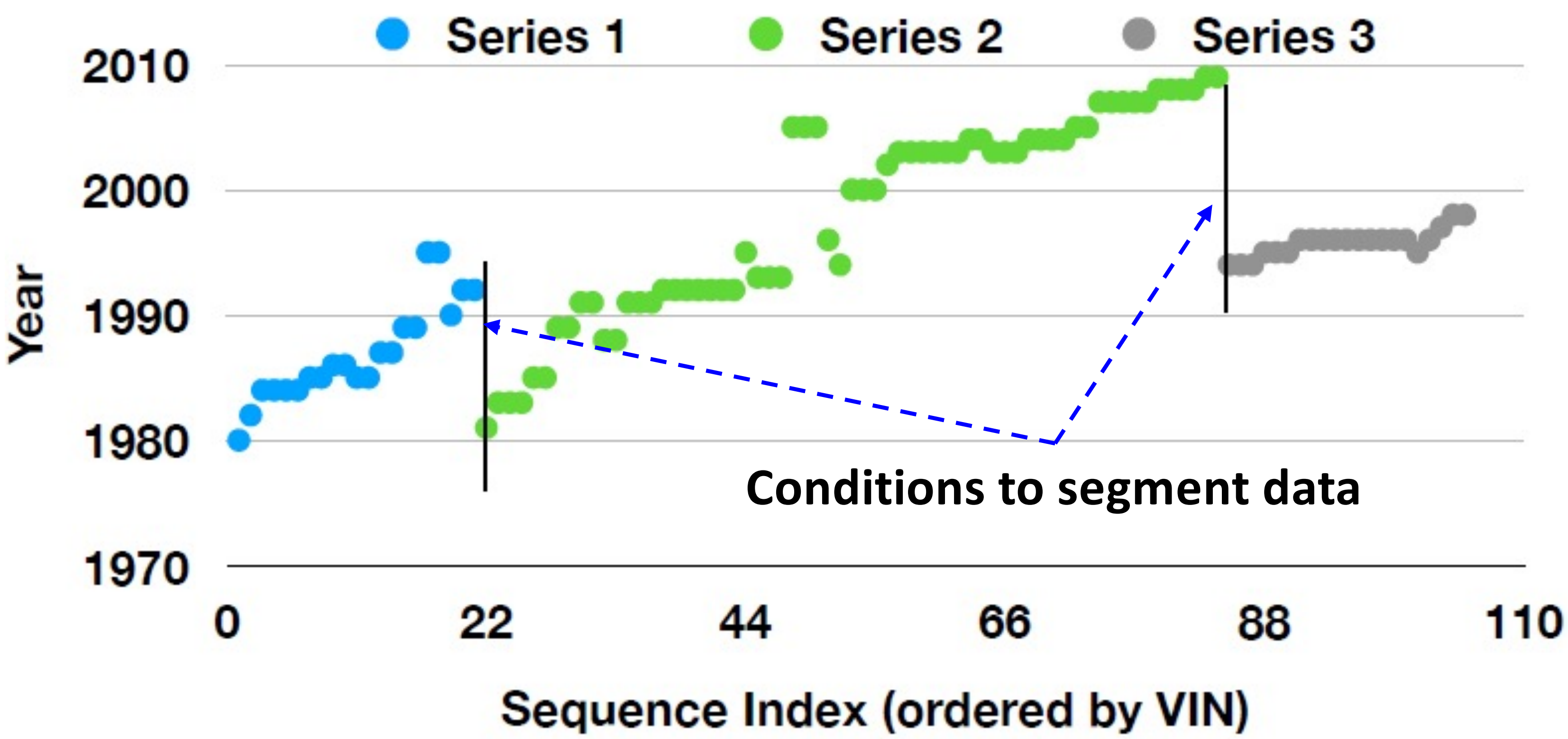


Impact of Big Data

- **Variety, variability** of data: one size does not fit all.
  - Learn **conditional** models (contextual semantics).

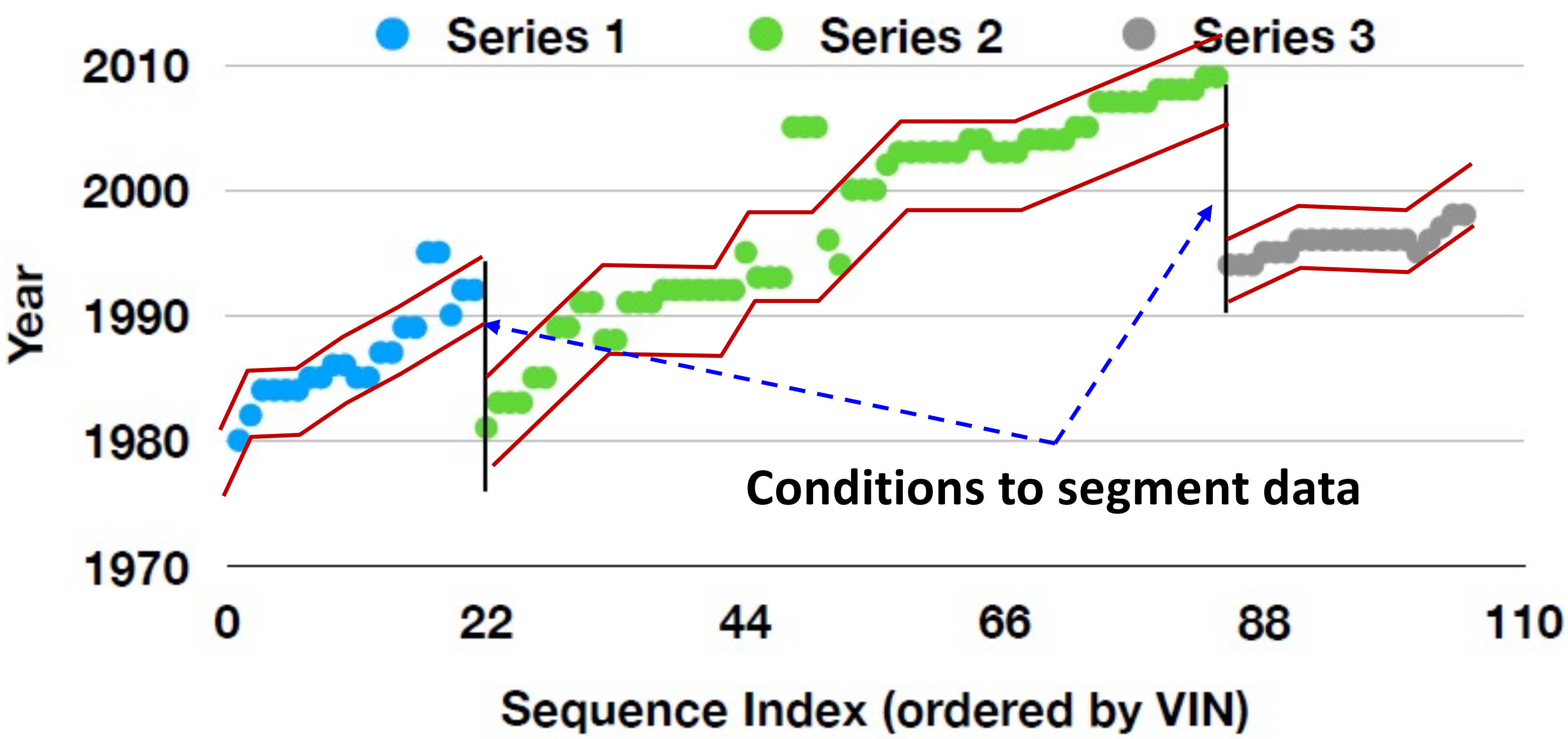
Discovering Timestamp Glitches: Solution 2

- Learn **ABC** (approximate band conditional) **OD** instead.
  - Need to learn **conditions** to partition/segment data.



Discovering Timestamp Glitches: Solution 2

- Learn **ABC** (approximate band conditional) **OD** instead.
  - Need to learn **conditions** to partition/segment data, and jointly
  - Determine LMBs of optimal band width **within each data segment**.



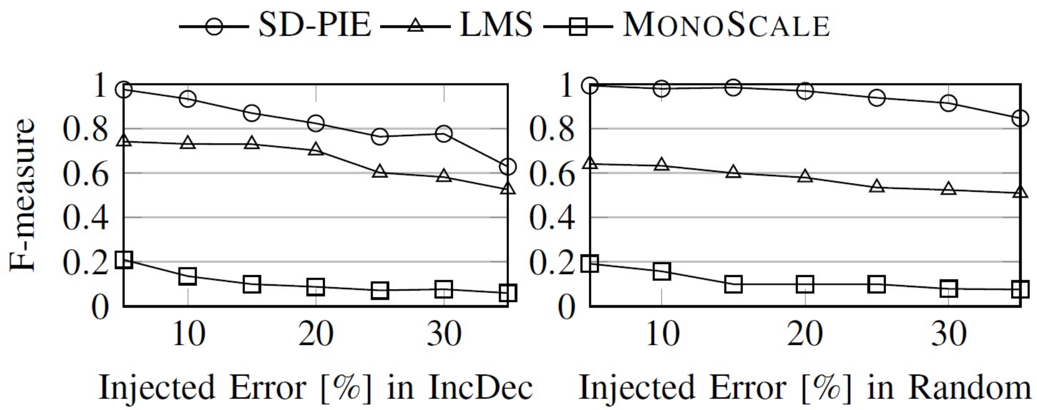
Discovering Timestamp Glitches: Results

- Quality experiments.
  - 2 real data sets (Music: ~0.9M, Car: ~350), 5 discovery algorithms.

DISCOVERY QUALITY ON *Music* AND *Car* DATASETS.

	GAP			MonoScale			A-MonoScale			LMS			SD-PIE		
	F-1	**Pre**	**Rec**	**F-1**	**Pre**	**Rec**	**F-1**	**Pre**	**Rec**	**F-1**	**Pre**	**Rec**	**F-1**	**Pre**	**Rec**
Music-Simple	0.97	1	0.95	0.29	1	0.17	**1**	1	1	**1**	1	1	**1**	1	1
Music-Inc	0.86	0.79	0.95	0.33	0.94	0.20	0.79	0.97	0.67	**0.99**	0.99	0.99	**0.99**	0.99	1
Music-IncDec	0.77	0.63	0.98	0.46	0.83	0.32	0.80	0.91	0.72	0.78	0.98	0.65	**0.95**	0.94	0.95
Music-Random	0.73	0.58	0.99	0.59	0.81	0.47	0.86	0.90	0.82	0.81	0.97	0.69	**0.93**	0.94	0.93
Car	0.53	0.73	0.41	0.35	0.91	0.22				0.96	0.98	0.94	**0.97**	0.98	0.97

  - Controlled errors injection for stress test experiments.



Outline

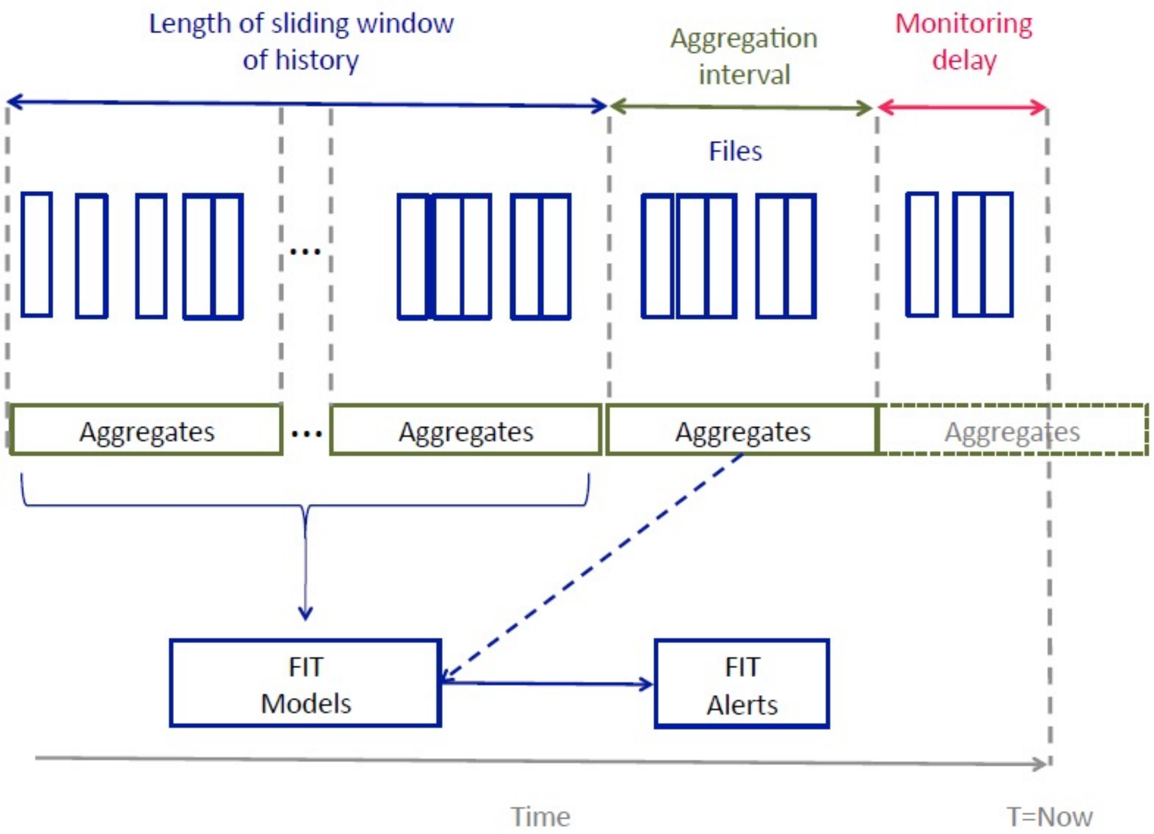
- Motivation.
- Obtaining high-quality long data.
  - Linking temporal records.
  - Discovering timestamp glitches.
  - The FIT family for real-time monitoring.

The FIT Family for DQ Monitoring

- Adaptive, data-driven **statistical** models/algos used at AT&T.
  - Continuous DQ monitoring on variety of **evolving data** streams.

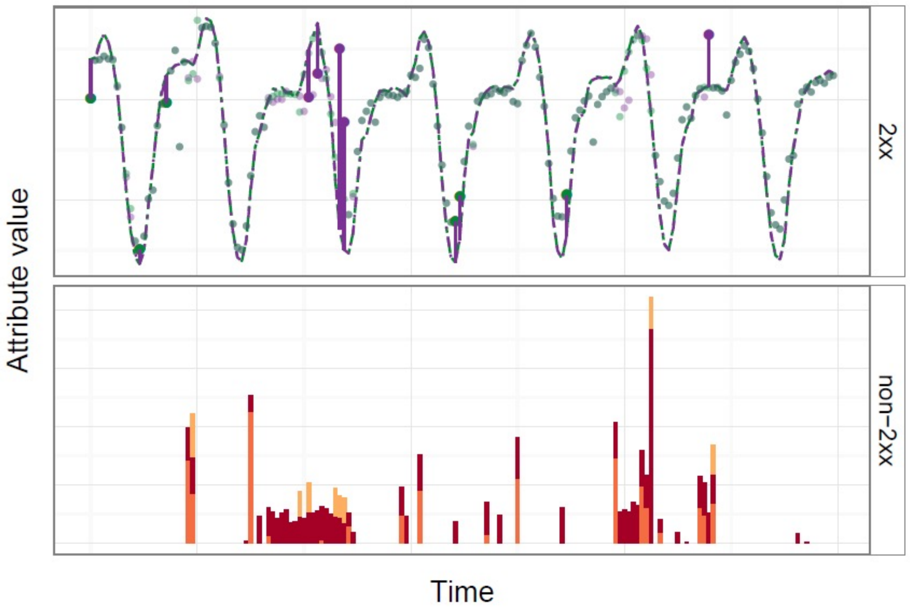
- FIT family members.
  - **ClassicFIT**: discovers data glitches in asynchronous data movement.
  - ContentFIT: discovers glitches in distributions of feed content.
  - SpaceFIT: discovers glitches in content of spatiotemporal feeds.
  - TimeFIT: learns models of delayed data arrivals over time.
  - ProcessFIT: learns process models based on multiple timestamps.
  - **SuperFIT**: discovers alert hotspots.

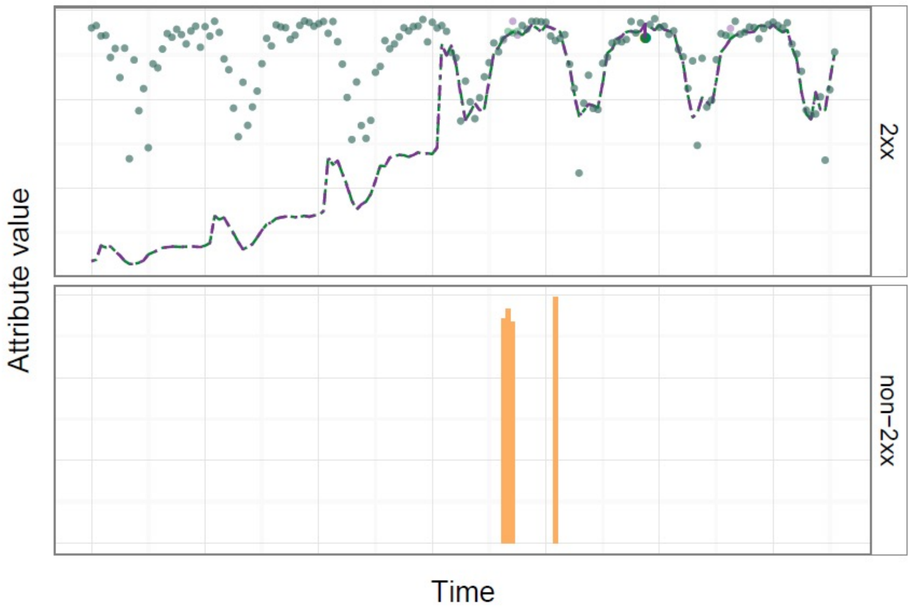
ClassicFIT [DSS+15]



- Monitors DQ in asynchronous data movement by analyzing logs.
  - Builds adaptive, data-driven statistical models in near real-time.
  - Alerts on missing, partial, duplicated & delayed data glitches.

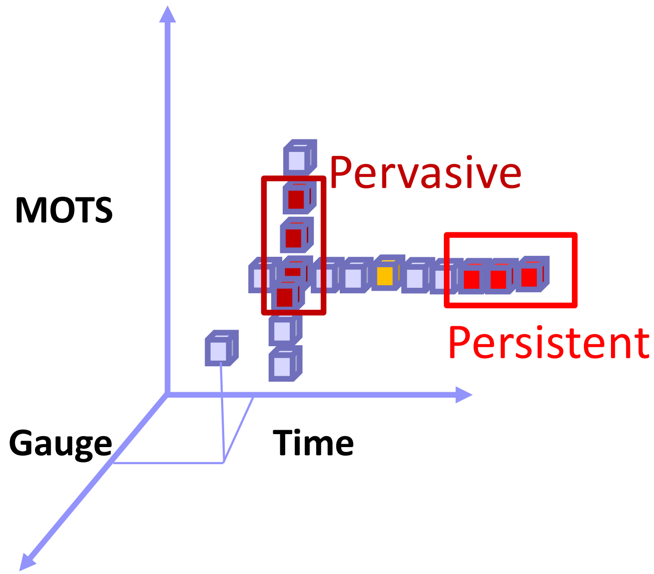
ClassicFIT [DSS+15]





- AT&T deployment.
  - Monitors > 3500 feeds and 57 million daily data router log records.
  - Generates alerts for abnormal file counts, file sizes and delays.

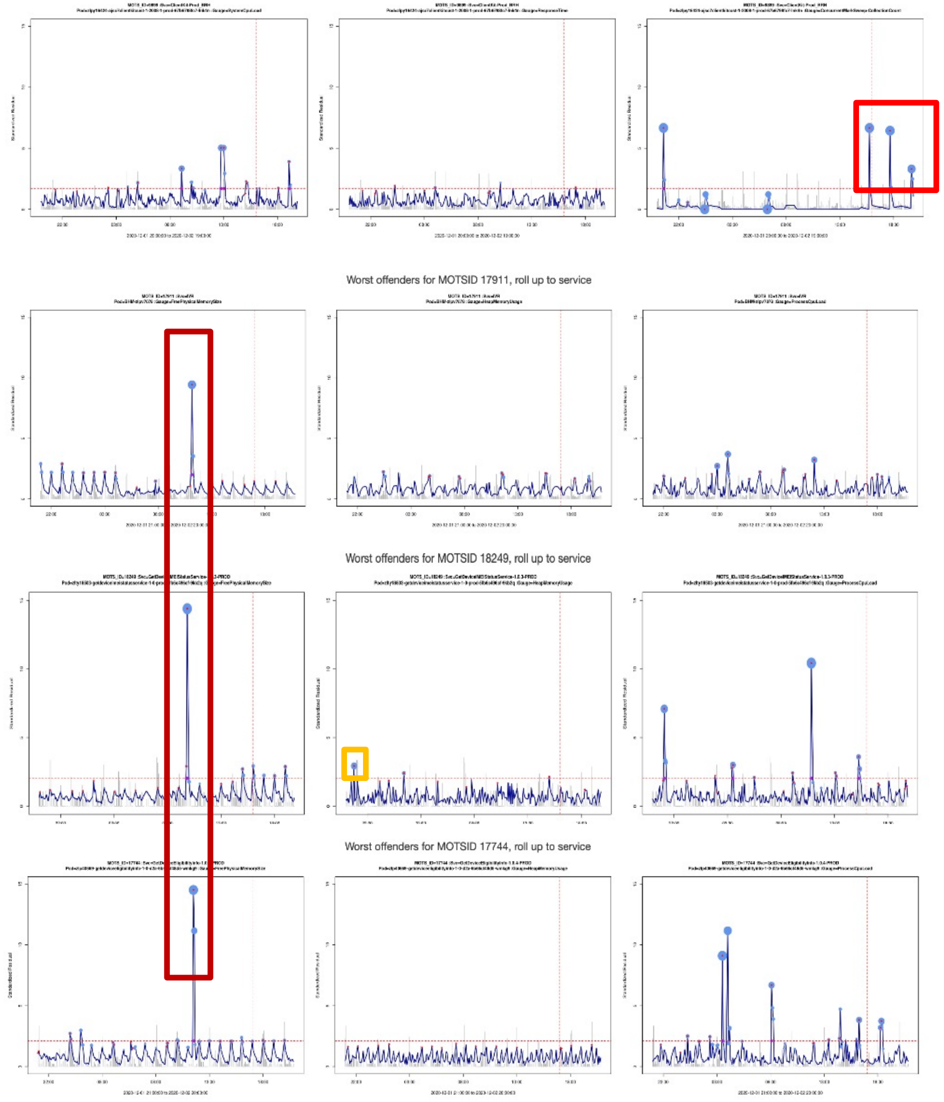
SuperFIT [BDK+19]



- Monitors alerts from inter-related high-volume data streams.
  - **Too many raw alerts**, not all equally critical, overwhelms agents.
  - SuperFIT discovers **alert hotspots** (super alerts, extreme alerts).
  - Based on **persistence** in time, **pervasiveness** in attribute space, and **priority** in terms of density of alerts and likelihood of occurrence.

SuperFIT [BDK+19]

- AT&T deployment.
  - Monitors infrastructure KPIs of critical AT&T cloud apps.
  - ~110 applications, 24M KPIs, 1M baseline alerts each day.
  - Generates ~50 **actionable extreme alerts** each day/app.



Conclusions

- Low quality long data is impediment to data science.
  - To achieve high quality data, **let the data speak for itself.**
  - Challenges due to volume, velocity, variety, variability of long data.
- Much interesting work has been done in this area.
  - **Learn** approximate, conditional models (semantics) from long data.
  - **Identify** data glitches as violations of the learned models.
  - **Repair** data glitches and models in a timely manner.
  - Real deployments at scale.
- A lot more research needs to be done!

Questions? Suggestions? Criticisms?

"I have a note from my doctor. I'm allergic to criticism!"