Data Glitches = Constraint Violations – Empirical Explanations

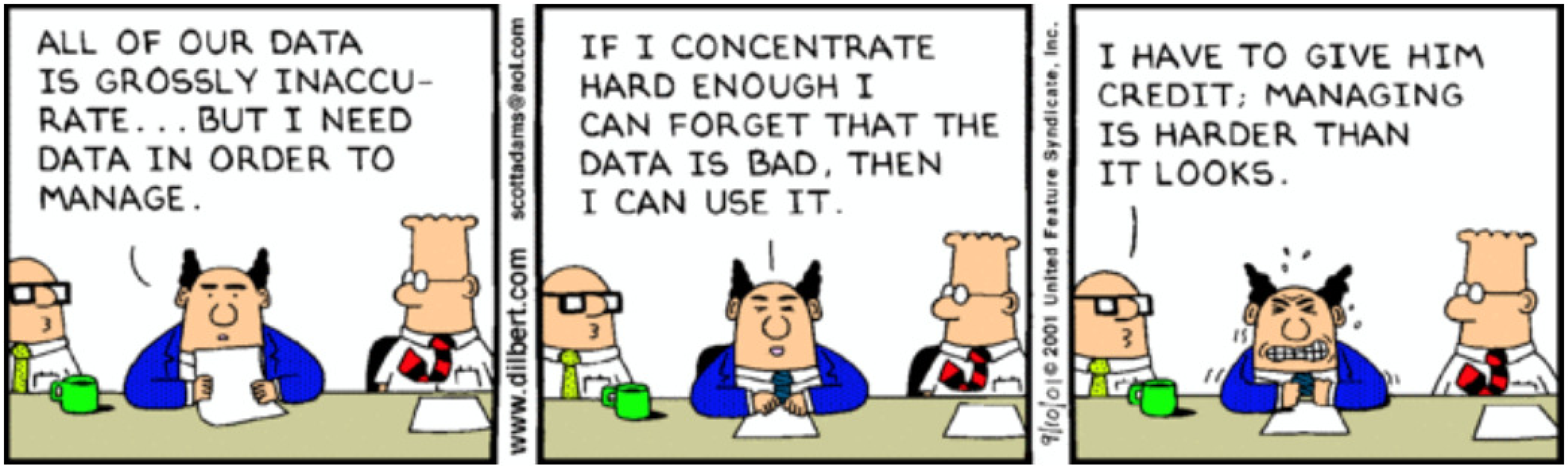
Divesh Srivastava

AT&T Labs-Research

What is a Glitch?

- A spaceman's word for irritating disturbances [Time, 23 Jul 1965].
  - "Something's gone wrong and you can't figure out what it is" [Daly].

What is a Data Glitch?



- A data scientist's phrase for irritating data quality problems.
  - Data that has gone wrong and **can't be used as desired**.
  - Unusual data that **does not conform** to data quality expectations.

What is an Integrity Constraint Violation?

- Integrity constraint: formal specification that data must satisfy.
  - **Semantic** (SSN unique for person) vs **syntactic** (NNN-NN-NNNN).
  - **Logical** (FD on 52wk low-high) vs **statistical** (# files within 3σ of μ).

- Violation: data that does not satisfy specified integrity constraint.

Yahoo! Finance

Green Mountain Coffee Roasters, (NasdaqGS: GMCR)
After Hours: 95.13 -0.01 (-0.02%) 4:07PM EDT

Last Trade:	95.14	Day's Range:	93.80 - 95.71
Trade Time:	4:00PM EDT	52wk Range:	25.38 - 95.71
Change:	1.69 (1.81%)	Volume:	2,384,075
Prev Close:	93.45	Avg Vol (3m):	2,512,070
Open:	94.01	Market Cap:	13.51B
Bid:	95.03 x 100	P/E (ttm):	119.82
Ask:	95.94 x 100	EPS (ttm):	0.79
1y Target Est:	92.50	Div & Yi	N/A (N/A)

52wk Range: 25.38-95.71

Nasdaq

Last Sale	$ 95.14
Change Net / %	1.69 ▲ 1.81%
Best Bid / Ask	$ 95.03 / $ 95.94
1y Target Est	$ 95.00
Today's High / Low	$ 95.71 / $ 93.80
Share Volume	2,384,175
50 Day Avg. Daily Volume	2,751,062
Previous Close	$ 93.45
52 Wk High / Low	$ 93.72 / $ 25.38
Shares Outstanding	152,785,000
Market Value of Listed Security	14,535,964,900
P/E Ratio	120.43
Forward P/E (1y	63.57
Earnings Per Sh	$ 0.79
Annualized Divi	N/A

52 Wk: 25.38-93.72

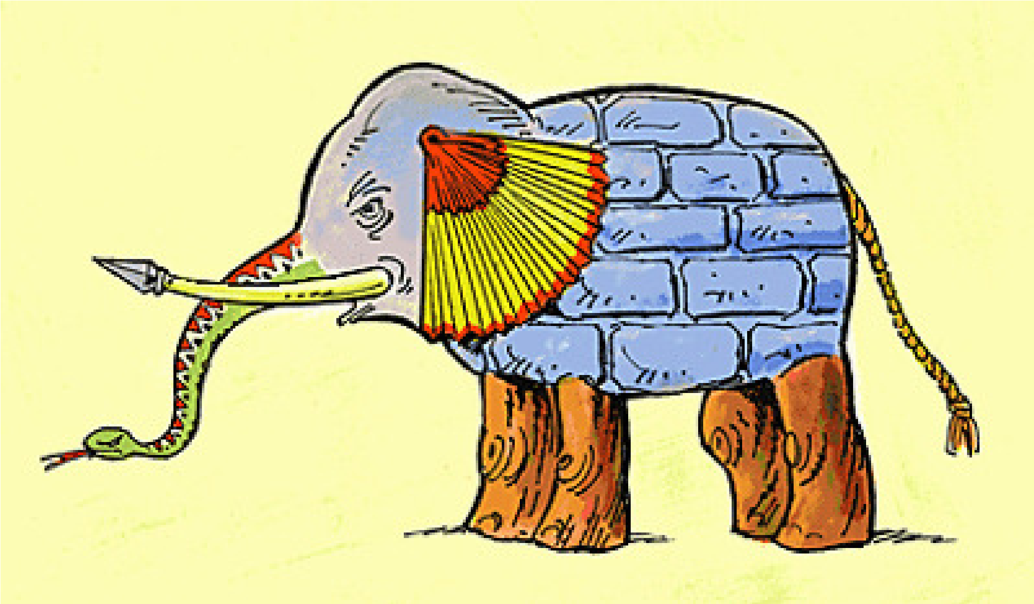
“Small Data” Quality: How Was It Achieved?

- 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.

Data Quality: Impact of Big Data

- **Variety, variability** of data: one size does not fit all.

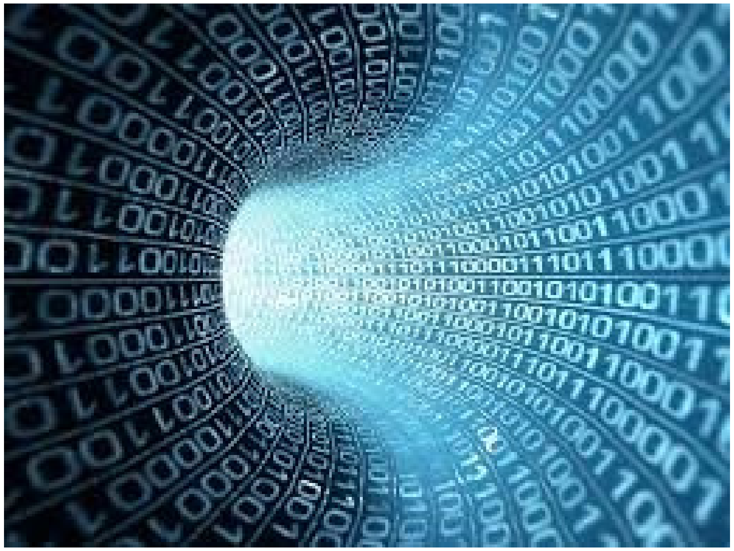
Big Data



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

Big Data Quality: A Different Approach?

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



Big Data Quality: A Different Approach?

- Big data: integrity constraints cannot be always specified a priori.
  - Data **variety** → complete domain knowledge is infeasible.
  - Data **variability** → domain knowledge becomes obsolete.

- Solution: let the data speak for itself.
  - Learn (simple) **integrity constraints / models** from the data.
  - Identify **violations** of the learned constraints.
  - Learn (complex) **empirical explanations** of the identified violations.
  - Declare **glitches** = constraint violations – empirical explanations.

In This Talk

- Big data: integrity constraints cannot be always specified a priori.
  - Data **variety** → complete domain knowledge is infeasible.
  - Data **variability** → domain knowledge becomes obsolete.

- Solution: let the data speak for itself.
  - Learn (simple) **integrity constraints / models** from the data.
  - Identify **violations** of the learned constraints.
  - Learn (complex) **empirical explanations** of the identified violations.
  - Declare **glitches** = constraint violations – empirical explanations.

Outline

- Introduction.
- What is an empirical explanation?
- Unsupervised learning of empirical explanations.

What is an Empirical Explanation?

ID	Status	Phone	Dept.	Rm.	Super_ID
ID_5	Active	1AAA3608776	D2300	A115	ID_9
ID_7	**New Hire**	1AAA3608776	D2300	D284	**ID_5**
ID_8	**New Hire**	1AAA3608776	D2300	B106	**ID_5**

- Data does not conform to expectation of “phone # uniqueness”.
  - Explanation = “new hires can have same phone # as supervisor”.
  - Explanation **can be learned from the data** → empirical explanation.

What is an Empirical Explanation? ✓

ID	Status	Phone	Dept.	Rm.	Super_ID
ID_10	Active	1AAA3605519	D8000	**A132**	ID_13
ID_11	Active	1AAA3605519	D8000	**A132**	ID_13
ID_12	Active	1AAA3605519	D8000	**A132**	ID_13

- Data does not conform to expectation of “phone # uniqueness”.
  - Explanation = “employees in same room can have same phone #”.
  - Is this an empirical explanation?

What is an Empirical Explanation?

ID	Status	Phone	Dept.	Rm.	Super_ID
ID_1	Active	1AAA3600000	D4000	-------	ID_4
ID_2	---------	1AAA3600000	-------	-------	------
ID_3	Active	1AAA3600000	D2200	E260	ID_6

- Data does not conform to expectation of "phone # uniqueness".
  - No empirical explanation is discernible.
  - May be a **data glitch**. ☺

What is an Empirical Explanation?

✓

Yahoo! Finance

Green Mountain Coffee Roasters, (NasdaqGS: GMCR)
After Hours: 95.13 ↓ -0.01 (-0.02%) 4:07PM EDT

Last Trade:	95.14	Day's Range:	93.80 - 95.71
Trade Time:	4:00PM EDT	52wk Range:	25.38 - 95.71
Change:	↑ 1.69 (1.81%)	Volume:	2,384,075
Prev Close:	93.45	Avg Vol (3m):	2,512,070
Open:	94.01	Market Cap:	13.51B
Bid:	95.03 x 100	P/E (ttm):	119.82
Ask:	95.94 x 100	EPS (ttm):	0.79
1y Target Est:	92.50	Div & Yi	N/A (N/A)

52wk Range: 25.38-95.71

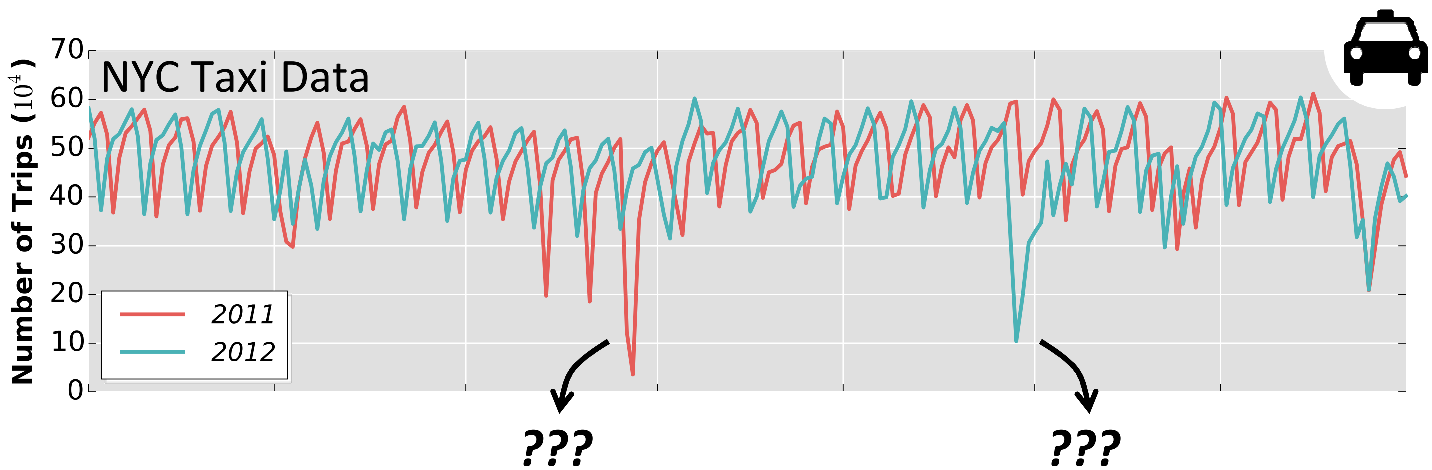
Nasdaq

Last Sale	$ 95.14
Change Net / %	1.69 ▲ 1.81%
Best Bid / Ask	$ 95.03 / $ 95.94
1y Target Est:	$ 95.00
Today's High / Low	$ 95.71 / $ 93.80
Share Volume	2,384,175
50 Day Avg. Daily Volume	2,751,062
Previous Close	$ 93.45
52 Wk High / Low	$ 93.72 / $ 25.38
Shares Outstanding	152,785,000
Market Value of Listed Security	14,535,964,900
P/E Ratio	120.43
Forward P/E (1y	63.57
Earnings Per Sh	$ 0.79
Annualized Divi	N/A

52 Wk: 25.38-93.72

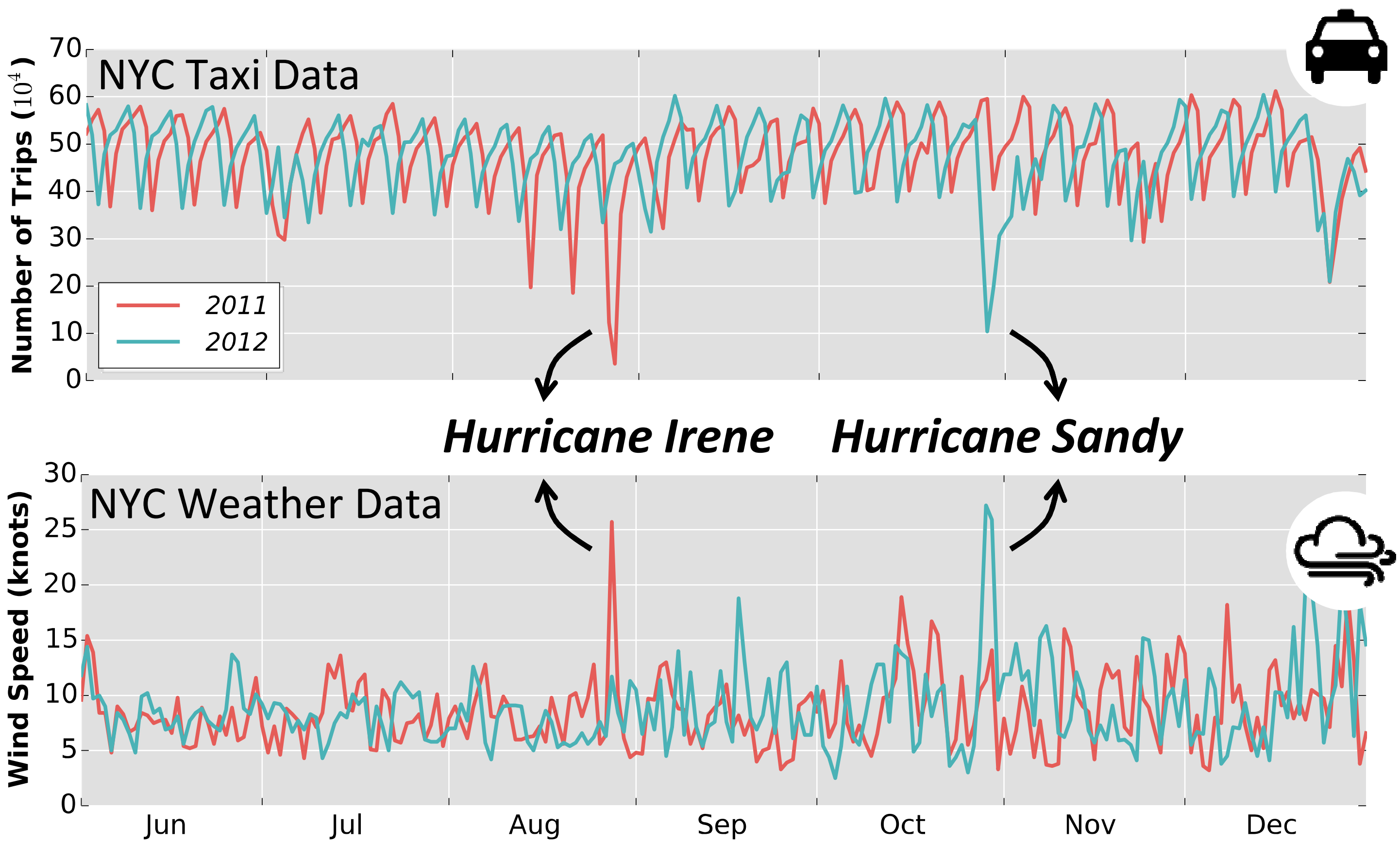
- Data does not conform to expectation of "FD on 52wk low-high".
  - Explanation = "52 wk low-high definitions differ between sources".
  - Is this an empirical explanation?

What is an Empirical Explanation?



- Data does not conform to (statistical) expectation of "≤ 3σ of μ".
  - No empirical explanation is discernible; could it be a data glitch?

What is an Empirical Explanation?

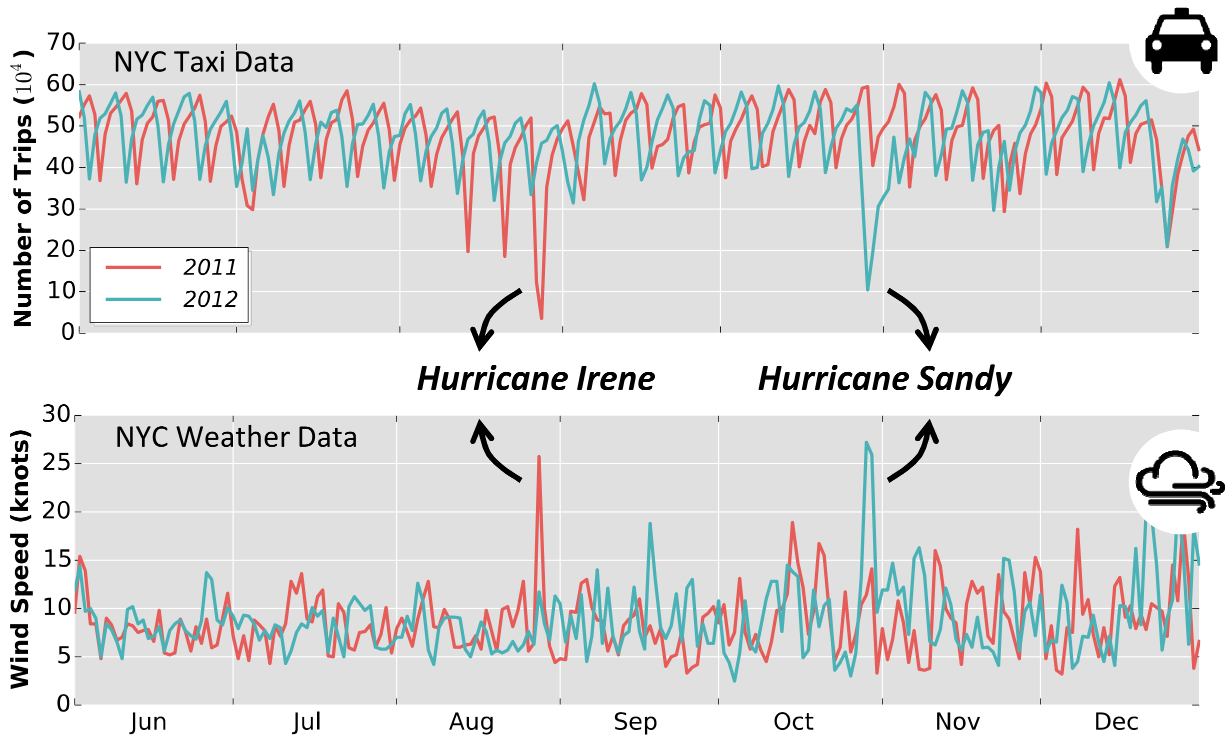


- Data does not conform to (statistical) expectation of "≤ 3σ of μ".
  - Empirical explanation = "Fewer taxi trips during high wind speeds".
  - An empirical explanation may **involve multiple data sets**.

Outline

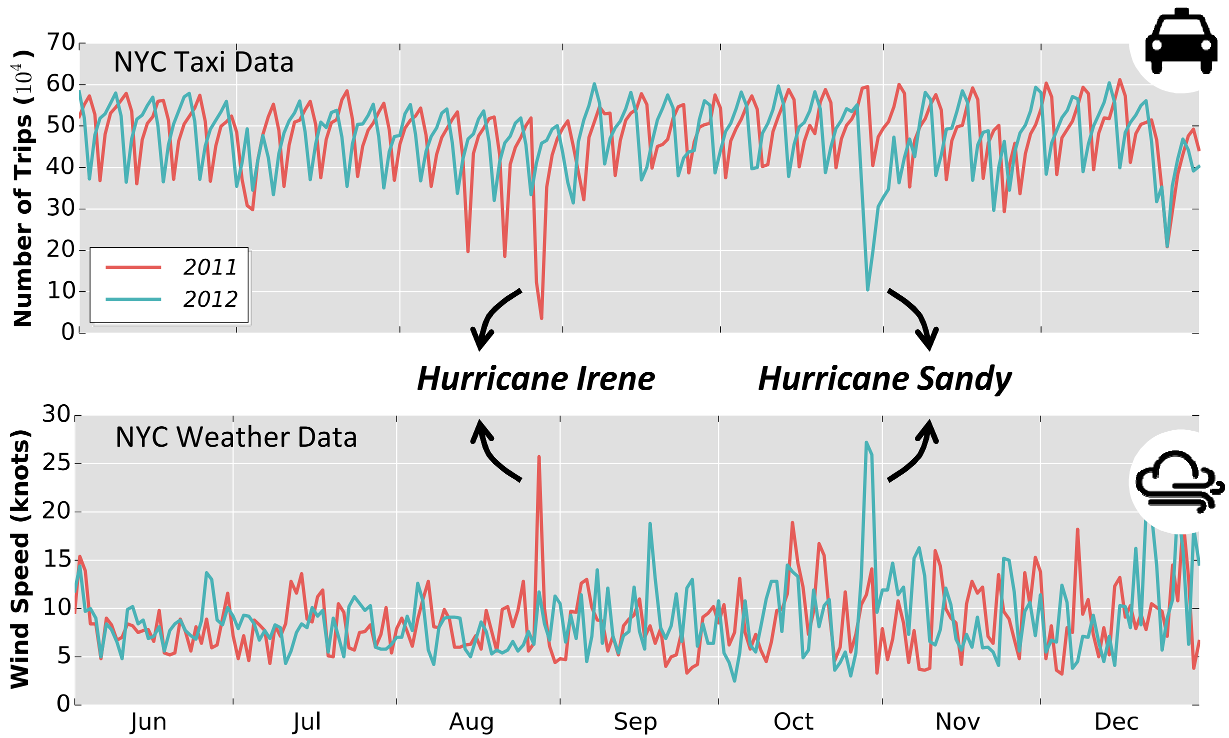
- Introduction.
- What is an empirical explanation?
- Unsupervised learning of empirical explanations.
  - Using spatio-temporal topological features [CD+16].
  - Using statistical signatures [DLS14].

Using Spatio-Temporal Features: Problem



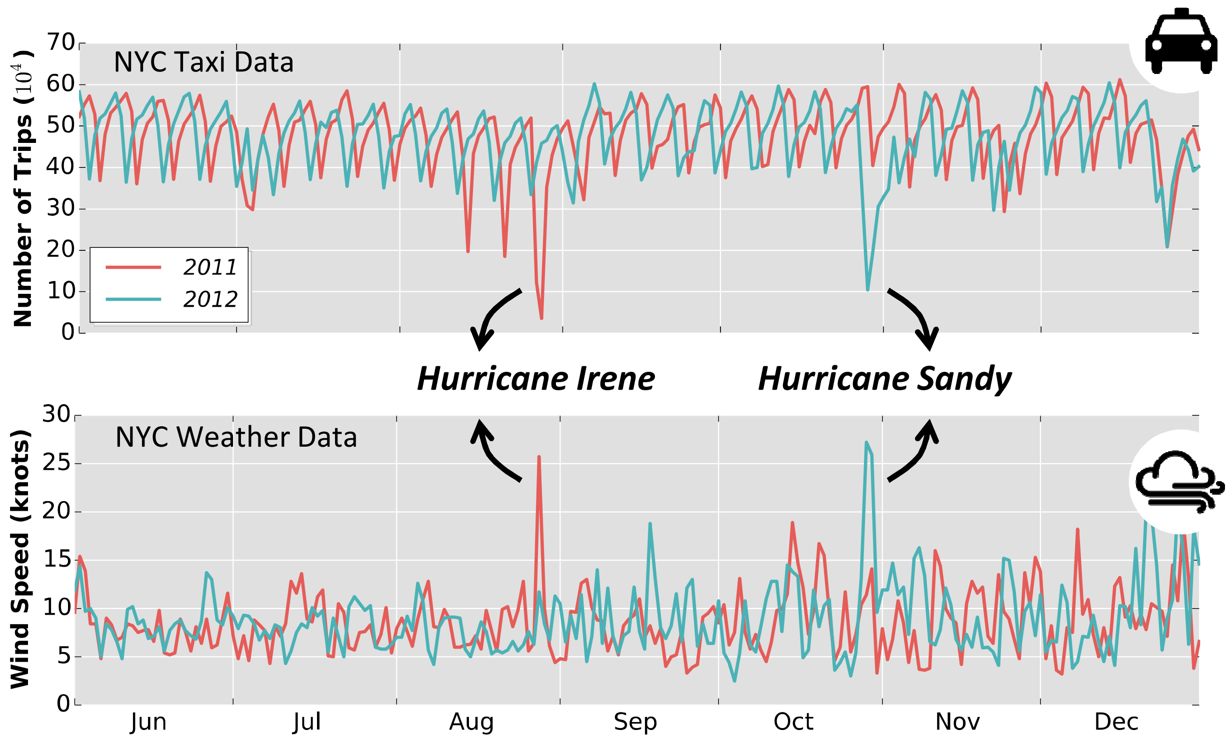
- Problem: Find data sets with **correlated spatio-temporal outliers**.

Using Spatio-Temporal Features: Alternatives



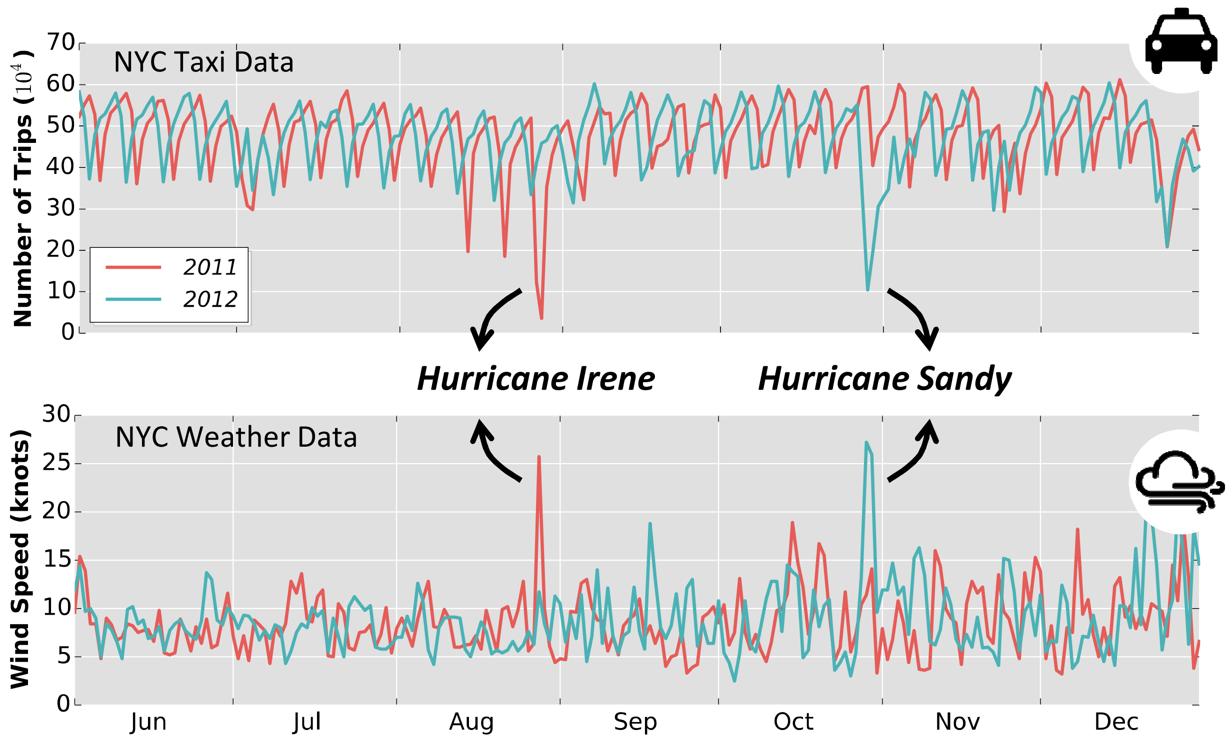
- Traditional approaches: Pearson's correlation, DTW, etc.
  - Miss **relationships that occur only at certain times / locations**, e.g., most of the time, # of taxi trips and wind speed are not related.

Using Spatio-Temporal Features: Challenges



- Finding correlated spatio-temporal outliers is challenging.
  - Big data sets, at different spatio-temporal resolutions.
  - Combinatorial # of possible correlations to evaluate.

Using Spatio-Temporal Features: Solution



- Solution: the **Data Polygamy** framework [CD+16].
  - Constraint violations = topological features (e.g., peaks, valleys).
  - Empirical explanations = significant (not a coincidence) correlations.

Interesting Relationships Discovered

- Data sets: NYC urban, NYC open data.

- Weather and vehicle collisions.
  - Strong correlation between heavy rainfall and motorist fatalities.
  - No significant relationship between rainfall and vehicle collisions.

- Weather and taxi availability.
  - Strong correlation between heavy rainfall and number of taxis.

Daily Intelligencer

Why You Can't Get a Taxi When It's Raining

By Annie Lowrey Follow @AnnieLowrey

Good luck, lady. Photo: Jacobs Stock Photography/Getty Images

It's pouring rain. You're running late. You desperately want to take a cab to the office. But, of course, there are none to be found. Happens all the time, right? Right, says science — or, to be specific, a new and exhaustive economic analysis of New York City taxi rides and Central Park meteorological data.

Outline

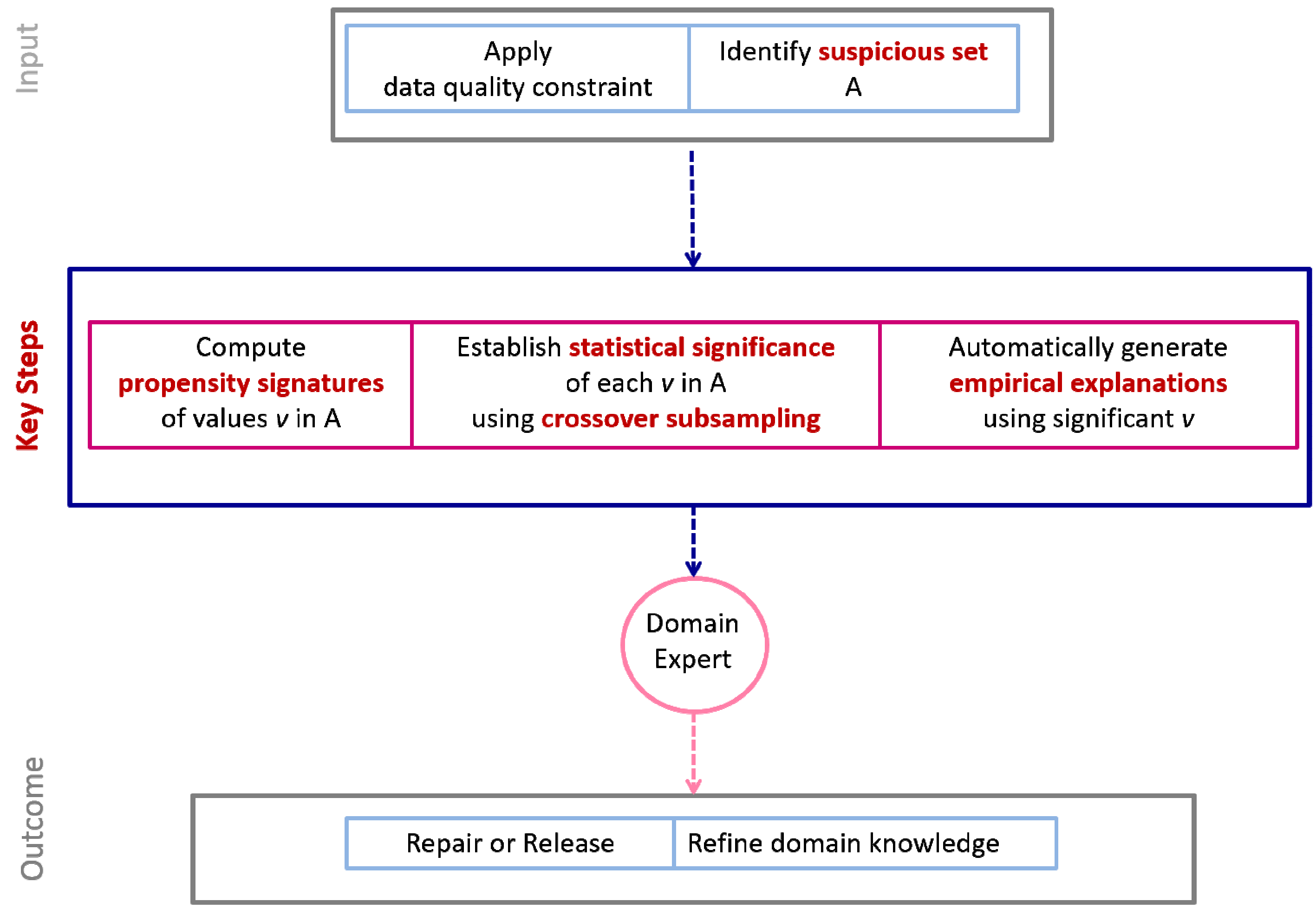
- Introduction.
- What is an empirical explanation?
- Unsupervised learning of empirical explanations.
  - Using spatio-temporal topological features [CD+16].
  - Using statistical signatures [DLS14].

Using Statistical Signatures: Problem

ID	Status	Phone	Dept	Rm.	Super
ID_5	Active	1AAA3608776	D2300	A115	ID_9
ID_7	**New Hire**	1AAA3608776	D2300	D284	**ID_5**
ID_8	**New Hire**	1AAA3608776	D2300	B106	**ID_5**

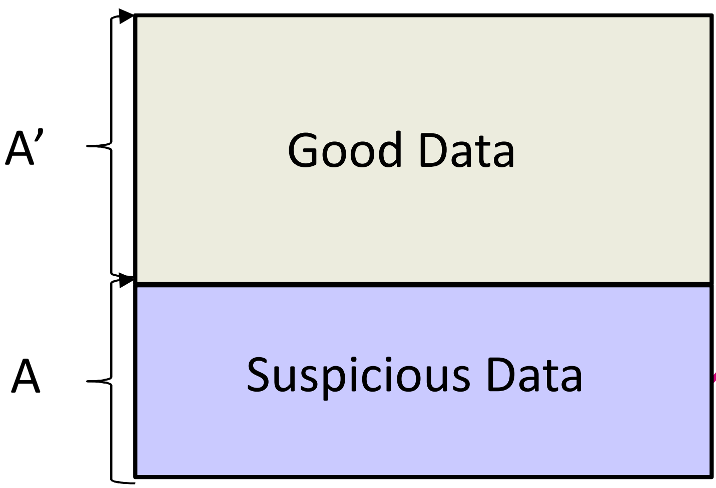
- Problem: Find **statistically significant explanations** of violations.
  - Needed because of incomplete, obsolete domain knowledge.

Using Statistical Signatures: Overview



Using Statistical Signatures: Step 1

Data D



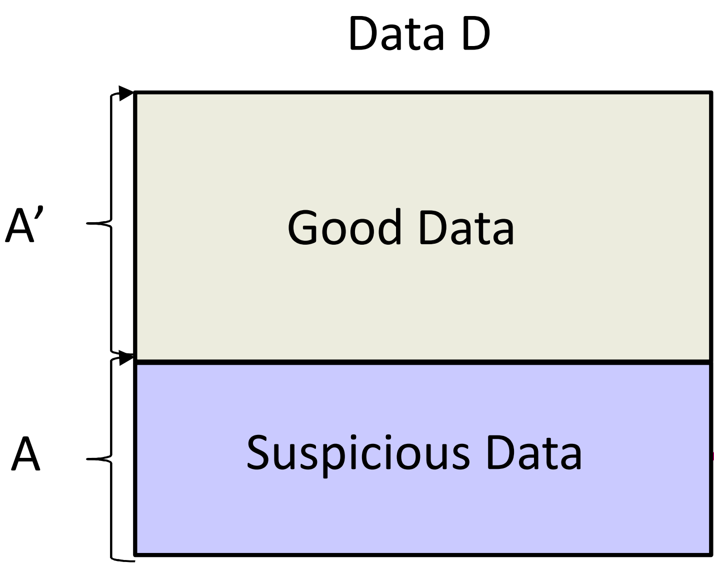
Suspicious set A: duplicate phone number

ID	Status	Phone	Dept	Rm.	Super
ID_5	Active	1AAA3608776	D2300	A115	ID_9
ID_7	**New Hire**	1AAA3608776	D2300	D284	**ID_5**
ID_8	**New Hire**	1AAA3608776	D2300	B106	**ID_5**

- Apply constraint on D, identify violations (suspicious set) A.

- For each value v in A, compute **propensity signatures** in A and A’.
  - s_A(New Hire) = {0.67, 0.0, 0.0, 0.0, 0.0, 0.0}
  - $s_{A'}$(New Hire) = {0.05, 0.0, 0.0, 0.0, 0.0, 0.0}

Using Statistical Signatures: Step 1

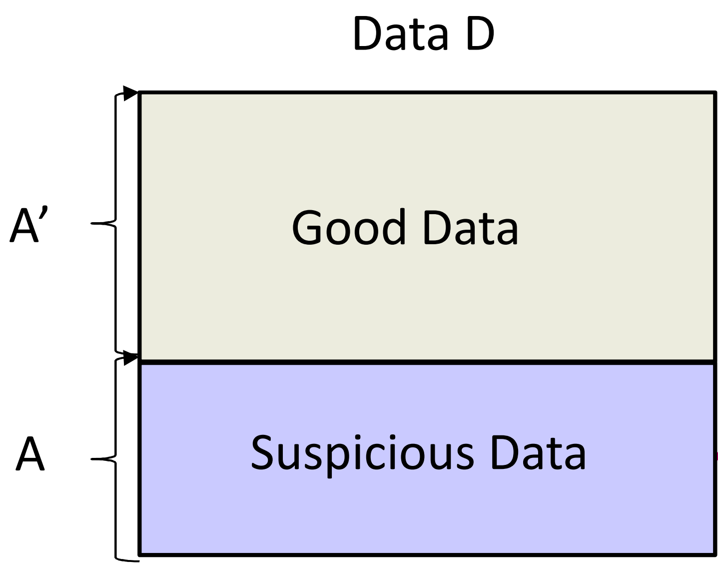


Suspicious set A: duplicate phone number

ID	Status	Phone	Dept	Rm.	Super
ID_5	Active	1AAA3608776	D2300	A115	ID_9
ID_7	**New Hire**	1AAA3608776	D2300	D284	**ID_5**
ID_8	**New Hire**	1AAA3608776	D2300	B106	**ID_5**

- Apply constraint on D, identify violations (suspicious set) A.
- For each value v in A, compute **propensity signatures** in A and A'.
  - $s_A(ID\_5) = \{0.33, 0.0, 0.0, 0.0, 0.0, 0.67\}$
  - $s_{A'}(ID\_5) = \{0.02, 0.0, 0.0, 0.0, 0.0, 0.05\}$

Using Statistical Signatures: Step 1



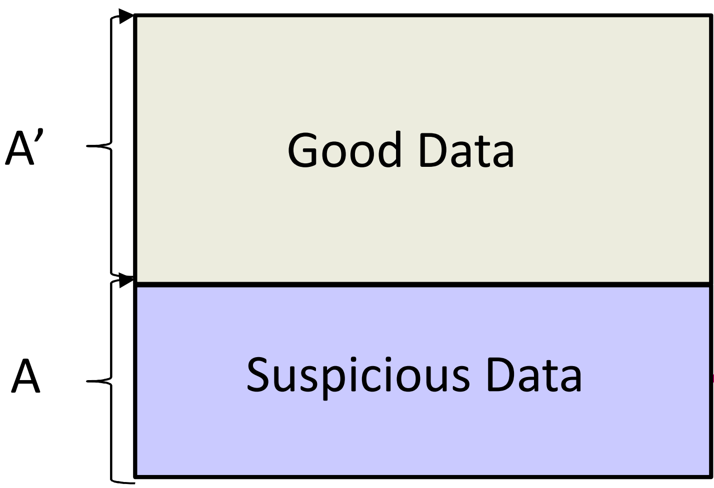
Suspicious set A: duplicate phone number

ID	Status	Phone	Dept	Rm.	Super
ID_5	Active	1AAA3608776	D2300	A115	ID_9
ID_7	**New Hire**	1AAA3608776	D2300	D284	**ID_5**
ID_8	**New Hire**	1AAA3608776	D2300	B106	**ID_5**

- Apply constraint on D, identify violations (suspicious set) A.
- For each value v in A, compute **propensity signatures** in A and A’.
  - Does value v have a **“sufficiently different” signature** in A vs A’?

Using Statistical Signatures: Step 1

Data D



Suspicious set A: duplicate phone number

ID	Status	Phone	Dept	Rm.	Super
ID_5	Active	1AAA3608776	D2300	A115	ID_9
ID_7	**New Hire**	1AAA3608776	D2300	D284	**ID_5**
ID_8	**New Hire**	1AAA3608776	D2300	B106	**ID_5**

- Apply constraint on D, identify violations (suspicious set) A.

- For each value v in A, compute **propensity signatures** in A and A'.
  - s_A(New Hire) = {0.67, 0.0, 0.0, 0.0, 0.0, 0.0}
  - $s_{A'}$(New Hire) = {0.05, 0.0, 0.0, 0.0, 0.0, 0.0}

Using Statistical Signatures: Step 2

A'

Good Data

A

ID	Status	Phone	Dept	Rm.	Super
ID_5	Active	1AAA3608776	D2300	A115	ID_9
ID_7	**New Hire**	1AAA3608776	D2300	D284	**ID_5**
ID_8	**New Hire**	1AAA3608776	D2300	B106	**ID_5**

Crossover Subsample

Blocks from A'

Block from A

- Goal: informative values that distinguish A from A'.
  - Establish statistical significance using **crossover subsampling**.
  - For an A block, sample A' blocks R times to create distribution.

Using Statistical Signatures: Step 3

ID	Status	Phone	Dept	Rm.	Super
ID_5	Active	1AAA3608776	D2300	A115	ID_9
ID_7	**New Hire**	1AAA3608776	D2300	D284	**ID_5**
ID_8	**New Hire**	1AAA3608776	D2300	B106	**ID_5**

- **Empirical explanation**: collection of all informative values for A.
  - Learned in an **unsupervised manner**, e.g., {ID_5, New Hire}.
  - Experts check empirical explanations, and decide on actions taken.

Summary

- Big data quality: let the data speak for itself.
  - Learn simple constraints from the data sets, identify violations.
  - Learn complex **empirical explanations** within and across data sets.
  - Data glitches = constraint violations – empirical explanations.

- Benefits: statistically robust, computationally efficient cleaning.
  - Reduces statistical distortion due to unnecessary cleaning.
  - Addresses challenges due to **variety, variability** in big data.

- Just the beginning, a lot of interesting work remains to be done …

Future Work

- **Improving efficiency** of learning empirical explanations.
  - Techniques presented are embarrassingly parallel.

- Use **supervised learning** for empirical explanations.
  - Current techniques use unsupervised techniques.

- **Combined learning** of constraints and empirical explanations.
  - Constraints used for data quality tend to be relatively simple.
  - Empirical explanations can be more complex.