

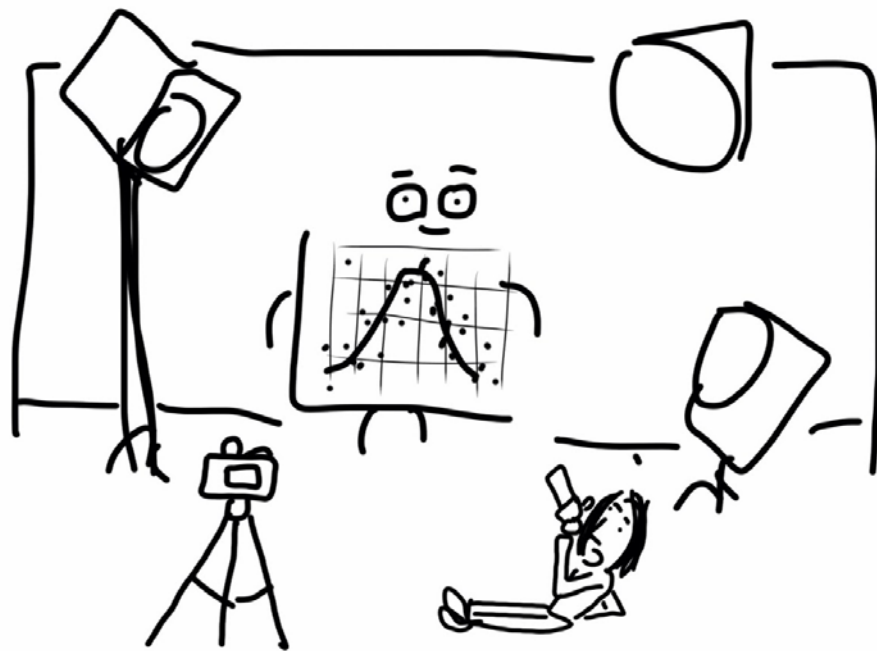


Practical Data Cleaning with Python

Tips and Tools for
Data Janitors



Sexy Data



Just keep working it, show off that bell curve

Real Data: 311 Calls, NYC

[illegible]

31872472,10/29/2015 11:08:21 PM,10/30/2015 01:36:18 AM,NYPD,New York City Police Department,Noise - Residential,Loud Music/Party,Residen
OF THE AMERICAS,7 AVENUE,,,ADDRESS,NEW YORK,,Precinct,Closed,10/30/2015 07:08:21 AM,The Police Department responded to the complaint and
MANHATTAN,MANHATTAN,985605,209402,Unspecified,MANHATTAN,Unspecified,Unspecified,Unspecified,Unspecified,Unspecified,Unspecified,Unspecif
-73.99511021832053)"

31872473,10/29/2015 11:44:42 AM,11/02/2015 11:23:02 AM,DOF,Correspondence Unit,DOF Property - Owner Issue,Remove Mortgage,Property Address
Department of Finance updated its records with the information provided.,11/02/2015 11:23:02 AM,08

[illegible]

31872490,10/29/2015 09:26:47 PM,10/30/2015 12:27:16 AM,NYPD,New York City Police Department,Noise - Residential,Banging/Pounding,Residen
AVENUE,FAILE STREET,BRYANT AVENUE,,,ADDRESS,BRONX,,Precinct,Closed,10/30/2015 05:26:47 AM,The Police Department responded to the complai
12:27:16 AM,02

[illegible]

Real Data: Unicef Survey in PDF

TABLE 9 | CHILD PROTECTION

Countries and areas	Child labour (%) ¹ 2005–2012*			Child marriage (%) 2005–2012*		Birth registration (%) ¹ 2005–2012*	Female genital mutilation/cutting (%) ¹ 2002–2012*		
	total	male	female	married by 15	married by 18		prevalence		attitudes support for the practice ²
							women ³	girls ³	
Afghanistan	10	11	10	15	40	37	–	–	–
Albania	12	14	9	0	10	99	–	–	–
Algeria	5 y	6 y	4 y	0	2	99	–	–	–
Andorra	–	–	–	–	–	100 v	–	–	–
Angola	24 x	22 x	25 x	–	–	36 x	–	–	–
Antigua and Barbuda	–	–	–	–	–	–	–	–	–
Argentina	7 y	8 y	5 y	–	–	99 y	–	–	–
Armenia	4	5	3	0	7	100	–	–	–
Australia	–	–	–	–	–	100 v	–	–	–
Austria	–	–	–	–	–	100 v	–	–	–
Azerbaijan	7 y	8 y	5 y	1	12	94	–	–	–
Bahamas	–	–	–	–	–	–	–	–	–
Bahrain	5 x	6 x	3 x	–	–	–	–	–	–
Bangladesh	13	18	8	29	65	31	–	–	–
Barbados	–	–	–	–	–	–	–	–	–
Belarus	1	1	2	0	3	100 y	–	–	–
Belgium	–	–	–	–	–	100 v	–	–	–
Belize	6	7	5	3	26	95	–	–	–
Benin	46	47	45	8	34	80	13	2 y	1
Bhutan	3	3	3	6	26	100	–	–	–
Bolivia (Plurinational State of)	26 y	28 y	24 y	3	22	76 y	–	–	–
Bosnia and Herzegovina	5	7	4	0	4	100	–	–	–

Real Data: GTFS (Transit Feeds)

TRANSITFEEDS

Feeds

API

Updates

[Home](#) / [Feeds](#) / [Europe](#) / [DE](#) / [Berlin](#) / [Verkehrsverbund Berlin-Brandenburg](#) / [VBB GTFS](#)

VBB GTFS

<div><div>Download Latest 53.9 MB</div><div>Routes 1,253</div><div>Stops 40,357</div></div>			
Date	Size	Routes	Status
10 April 2017	53.9 MB	1,253	Warnings 4 <input type="button" value="View"/> <input type="button" value="Download"/>
15 March 2017	51.8 MB	1,247	Warnings 4 <input type="button" value="View"/> <input type="button" value="Download"/>
16 December 2015	28.1 MB	1,532	Warnings 3 <input type="button" value="View"/> <input type="button" value="Download"/>
28 May 2015	25.9 MB	1,451	Warnings 3 <input type="button" value="View"/> <input type="button" value="Download"/>
2 November 2013	22.5 MB	1,363	Warnings 2 <input type="button" value="View"/> <input type="button" value="Download"/>

5 versions available



Data Wrangling with Python



What is Data Wrangling?



Big Data Borat

@BigDataBorat



Follow

In Data Science, 80% of time spent prepare data, 20% of time spent complain about need for prepare data.

RETWEETS

495

LIKES

251



6:47 PM - 26 Feb 2013



Why bother? (My Story)

Data affects lives!!

Machine learning, data and algorithms can determine who gets loans, who we marry, what we buy, what news we read.

The datasets we use matter. More data wrangling tools and skills create more accessibility & choice. Data wranglers ++!



Katharine Jarmul (@kjam): founder, kjamistan; data scientist & data engineer

Your Story

Poll Question:

How much time do you spend per week
wrangling data (on average)?



Image: Daily News % Felipe Ramales

kjamistan

Today's Agenda

Data Cleaning 101: Jupyter Workbooks & Active Learning

Automating Data Cleaning with Graphs

Case Study: Cleaning Web Data

What's Happening in Academia?

Quick Poll

Are you able to join code exercises throughout this two-day training?

How we'll use tools

Slack and Group Chat

asking questions, sharing more ideas

Polls

Surveying experiences, sharing knowledge

Pulse Check

Signaling understanding or completion of exercises



Data Cleaning 101



Following Along

- <http://github.com/kjam/data-cleaning-101>
- Python 3
- Requirements:
 - With pip: `install_reqs.txt`
 - With anaconda: `conda install (above)`
- If you get lost, please ask in chat (no stupid questions!)

Quick Poll

What Python version do you develop in?

Lesson One: Deduplication

- DataMade's dedupe library
 - <https://github.com/dedupeio/dedupe>
- (pip|conda) install dedupe
- See also: <https://github.com/dedupeio/csvdedupe>

Lesson Two: String Matching

- FuzzyWuzzy
 - <https://github.com/seatgeek/fuzzywuzzy>
- (pip|conda) install fuzzywuzzy
- See also: <https://github.com/chartbeat-labs/textacy>

Quick Poll

What are your biggest data cleaning problems?

Lesson Three: Managing Nulls

- Pandas capabilities
 - `na_values`
 - `fillna`
 - `dropna`
- (pip|conda) install pandas
- See also: <https://github.com/NathanEpstein/Dora> and <https://github.com/harshnisar/badfish>

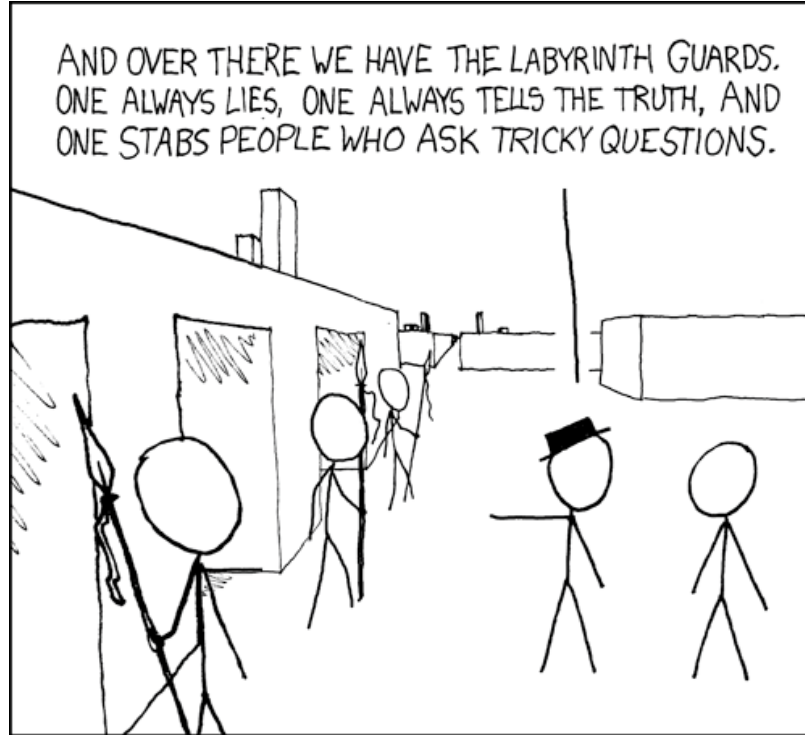
Lesson Four: Preprocessing

- Normalization, Impute missing values
 - <http://scikit-learn.org/stable/modules/preprocessing.html>
- (pip|conda) scikit-learn
- See also: <http://pandas.pydata.org/pandas-docs/stable/basics.html#descriptive-statistics>

Extra: Problem Specific Libs

- Privacy? <https://github.com/datascopeanalytics/scrubadub>
- Measurements? <http://pint.readthedocs.io/>
- Versioning ML Data?
<https://github.com/NathanEpstein/Dora>
- Dates? <http://arrow.readthedocs.io/en/latest/> or
<https://github.com/kennethreitz/maya>
- AutoClean? <https://github.com/rhiever/datacleaner>
- DIY Parser? <https://github.com/datamade/parserator>

Questions? (and short break)



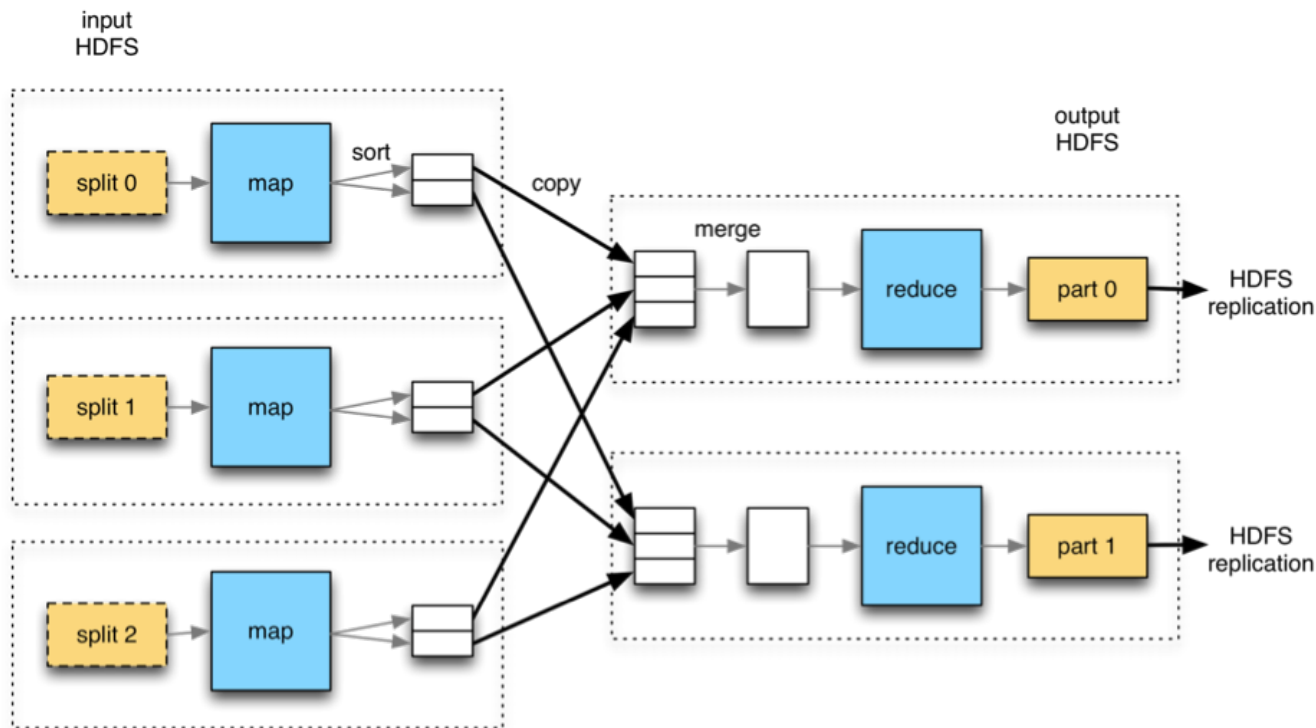
Source: <https://xkcd.com/246/>



Automating Data Cleaning with Graphs



Why Graphs?



DAG Benefits

- Parallelizable Workflow
- Easily Distributed
- “Replay” & Idempotency
- Failure Handling
- State Management

What if my Workflow is Unique?

- Is it actually 100% “unique”?
- Can you break it into smaller chunks which have more consistency and repetition?
- Are you doing it the hard way?

Quick Poll

What does your data extraction and cleaning look like?

Dask



- Parallel Out-of-Core Computations with Graphs
 - <https://github.com/dask/dask>
 - Implementations of: `pd.DataFrame`, `np.Array` and Bags (similar methods to Spark Text RDDs)
 - Delayed: Build your own Graph
 - Dask Distributed: Distributed Task Scheduling

Dask



Building a Data Pipeline with Dask...

TO TRACK STUFF IN SPACE !! 🚀

Questions? (and short break)

















Case Study: Preparing Web Data



Defining the Problem

Lobsters Recent Comments Search Filters Log

- ▲ **Remote security exploit in all 2008+ Intel platforms** hardware security [semiaccurate.com](#)
60  via albino 20 hours ago | cached | 35 comments
- ▲ **How Emotion Detection Technology Can Make Marketing More Effective** api [blog.paralldots.com](#)
2  authored by Gargi 22 hours ago | cached | 10 comments
- ▲ **neovim 0.2 released** release vim [github.com](#)
5  via geler 2 hours ago | cached | 1 comment
- ▲ **yieldfrom 1.0.0: A backport of the `yield from` semantic from Python 3.x to Python 2.7** python release [pypi.python.org](#)
10  authored by AmirRachum 18 hours ago | cached | 3 comments
- ▲ **Code Reviews Considered Harmful** practices [hackernoon.com](#)
2  via calvin 1 hour ago | cached | 1 comment
- ▲ **Six programming paradigms that will change how you think about coding** compsci programming [ybrikman.com](#)
30  via calvin 25 hours ago | cached | 5 comments
- ▲ **Readability Matters More Than Correctness** programming [xph.us](#)
26  via jcaudle 33 hours ago | cached | 31 comments
- ▲ **Open Sourcing our new Duckling** haskell [wit.ai](#)
12  via ehamberg 14 hours ago | cached | no comments
- ▲ **Searls-Briggs Type Indicator Survey (MBTI for programmers)** programming [testdouble.com](#)
16  via dgvs 31 hours ago | cached | 36 comments
- ▲ **Announcing kurly v1.0.0** show release go [davidjpeacock.ca](#)
11  authored by davidjpeacock 21 hours ago | cached | 3 comments
- ▲ **What are you working on this week?** ask culture
15  authored by calius 13 hours ago | 18 comments
- ▲ **terminal emulators' processing of escape sequences** security unix [openwall.com](#)
9  via jcs 18 hours ago | cached | 3 comments

Possible Solutions?

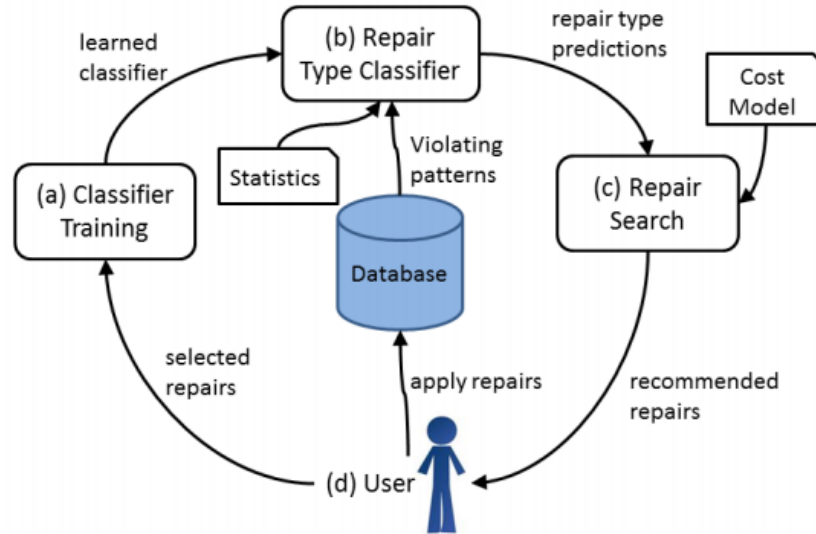
What are common fixes for handling external API
or web scraped data issues?



What's Happening in Academia?



Continuous Data Cleaning



M. Volkovs, F. Chiang, J. Szlichta, and R. J. Miller. ICDE, 2014

ActiveClean: ML + Crowd + Progressive Sampling

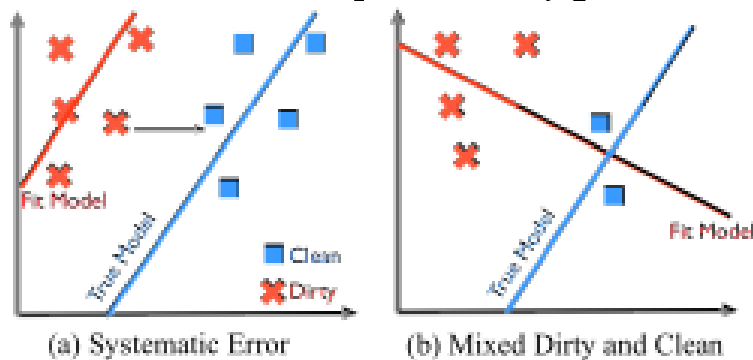


Figure 1: (a) Systematic corruption in one variable (axes) can lead to a shifted model (fitted lines). (b) Mixed dirty and clean data results in a less accurate model than no cleaning.

(Krishnan, Franklin, Goldberg, Wang, Wu, 2016)

ActiveClean

1. `Init(dirty_data, cleaned_data, dirty_model, batch, iter)`
2. `For each t in $\{1, \dots, T\}$`
- (a) `dirty_sample = Sampler(dirty_data, sample_prob, detector, batch)`
 - (b) `clean_sample = Cleaner(dirty_sample)`
 - (c) `current_model = Updater(current_model, sample_prob, clean_sample)`
 - (d) `cleaned_data = cleaned_data + clean_sample`
 - (e) `dirty_data = dirty_data - clean_sample`
 - (f) `sample_prob = Estimator(dirty_data, cleaned_data, detector)`
 - (g) `detector = Detector(detector, cleaned_data)`
3. `Output: current_model`

(Krishnan, Franklin, Goldberg, Wang, Wu, 2016)

Quick Poll

What data cleaning future excites you?

Whew! Day One Complete :)

- Tomorrow, same time: Data Validation ! Unit Tests ! Pipeline integration !
- Resource post: <https://blog.kjamistan.com/practical-data-cleaning-with-python-resources/>
- If you can, please take a minute to give me some feedback:
 - <http://bit.ly/practical-data-feedback>
- Reach out anytime:
 - @kjam on Twitter / Slack / GitHub
 - katharine@kjamistan.com

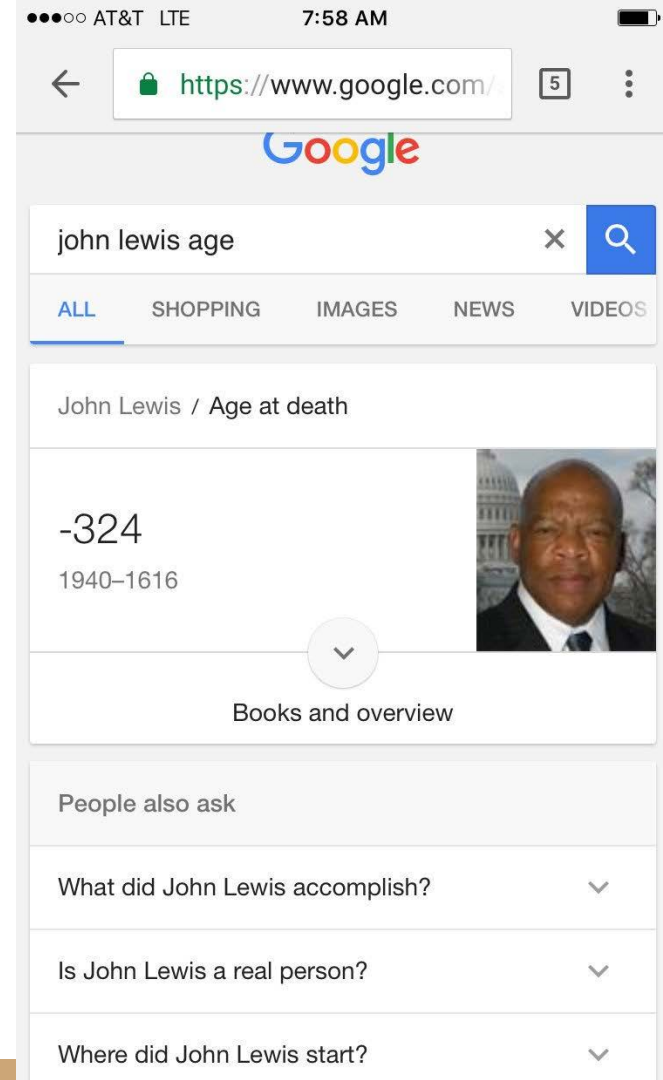


Practical Data Cleaning with Python

Tips and Tools for
Data Janitors



Data Validation in the Wild



Data Errors are Costly

..one consultant from a large database vendor noted that errors might be found well after some result is reported:

Most of these errors are subtle enough that the analysis will go through e.g., with standard null value semantics of SQL, but give an incorrect answer. Usually is only caught weeks later after someone notices something like... well the Wilmington branch cannot have 1M sales in a week.

Towards Reliable Interactive Data Cleaning: A User Survey and Recommendations
Sanjay Krishnan, Daniel Haas, Michael J. Franklin, 2016

Data Validation is Hard

“How do you determine whether the data is sufficiently clean to trust the analysis?”

Other than common sense we do not have a procedure to do this.

We usually do not do rigorous validation of data cleaning. We typically clean our data until the desired analytics works without error. This is not desirable but practical since in most cases data error is probably overshadowed by errors/inaccuracies in the models themselves.

Towards Reliable Interactive Data Cleaning: A User Survey and Recommendations
Sanjay Krishnan, Daniel Haas, Michael J. Franklin, 2016

Your Story

Poll Question: When I finish writing new reports / data science code or notebooks, I _____ add data validation.



Image: Basketball Wives via Tumblr

Today's Agenda

Data Validation 101: Definitions and Concepts

Data Validation and Testing with Python: Jupyter Workbooks & Active Learning

Data Unit Tests

Case Study: Data Validation with Machine Environment Data

What's Happening in Academia?

How we'll use tools

Slack and Group Chat

asking questions, sharing more ideas

Polls

Surveying experiences, sharing knowledge

Pulse Check

Signaling understanding or completion of exercises



Data Validation 101



Data Quality Evaluation

- ❑ Valid
- ❑ Accurate
- ❑ Complete
- ❑ Consistent
- ❑ Uniform
- ❑ Repeatable

Quick Poll

I find _____ most applicable to my current
data problems

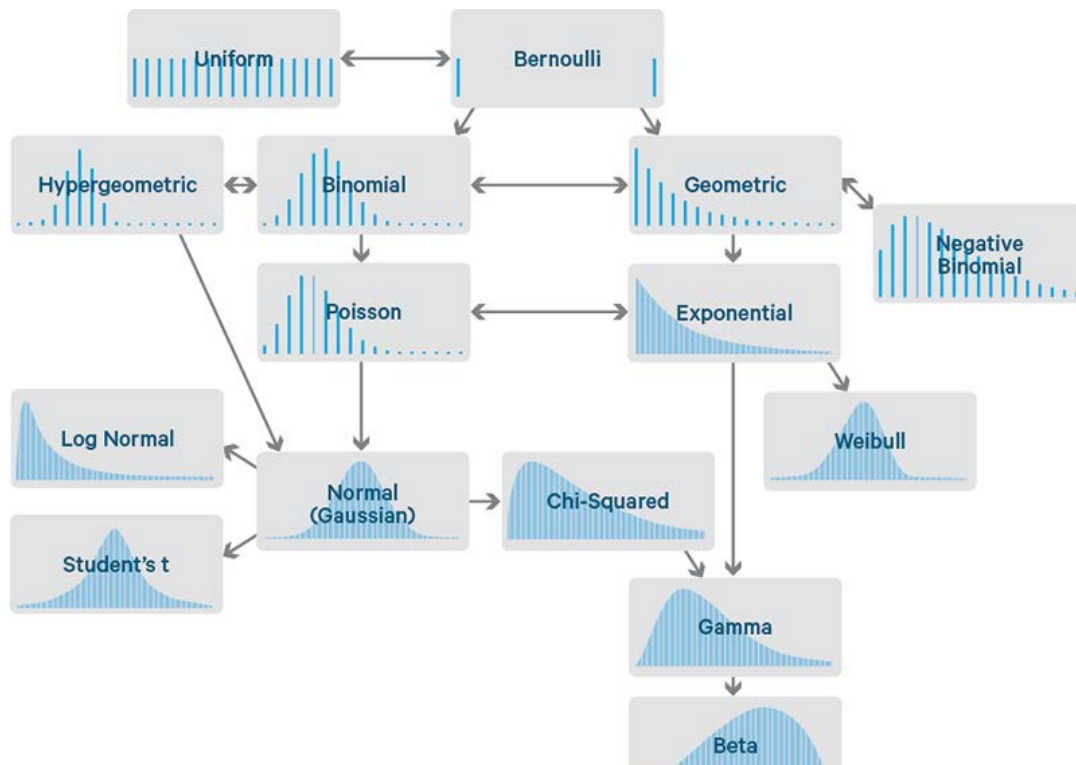
Data Reliability

- Correlation
 - More new users, more traffic and activity
- Temporal Stability
 - Traffic patterns
- Internal Consistency
 - More Clicks == More Page Views
- Determining good tests can be difficult, but useful

Data Validity

- Predictive Validity
 - Testing past models / Predictions
- Measurement or Metric Validity
 - What does it measure? How? Is it a valid measurement?
 - Fallacies of Metrics
- Trends or Noise?
 - Tracking patterns & outliers
 - Understanding biases
 - Handling Non-signals

Statistical Models & Measurements



Source: Cloudera (<http://blog.cloudera.com/blog/2015/12/common-probability-distributions-the-data-scientists-crib-sheet/>)

Quick Poll

Statistical measurements or distributions
can be used to verify my data ...

Outliers, Anomalies & Extreme Values

- Do normal outlier detection models work with your data?
- Can you easily predict normal values? (even if that means preprocessing and normalizing for day of week or seasonal trends)
- Do you throw out anomalies? How many? How come?
- Does anyone analyze average occurrence of anomalies or extreme values?

Quick Poll

Outliers are handled by...

Margin of Error (and Team MoE)

- How relevant are accuracy, error and confidence measurements for your data problems (or models)?
- Team “Margin of Error”
 - What’s the margin of error you need to meet for business purposes (or internal and external needs)?
 - Do you have the ability to test new ideas, models or exploratory analysis and make mistakes somewhere (A/B testing, subsets and samples)?

Quick Poll

Errors in data science modeling or reporting
are...



Data Validation and Testing with Python



Following Along

- <http://github.com/kjam/data-cleaning-101>
- Python 3
- Requirements:
 - `install_reqs.txt`
- If you get lost, please ask in chat (no stupid questions!)

Lesson One: Valid Values / Types

- Voluptuous
 - <https://github.com/alecthomas/voluptuous>
- (pip|conda) install voluptuous
- See also: <https://github.com/guyskk/validr>

Lesson Two: Dat aframe Validation

- Engarde
 - <https://github.com/TomAugspurger/engarde>
- (pip|conda) install engarde
- See also: <https://github.com/jnmclarty/validada>

Lesson Three: TDDA

- Test-Driven Data Analysis
 - <https://github.com/tdda/tdda>
- (pip|conda) install tdda
- See also: <https://docs.scipy.org/doc/scipy-0.19.0/reference/stats.html#statistical-functions>

Lesson Four: Property-Based Tests

- Hypothesis
 - <https://hypothesis.readthedocs.io/>
- (pip|conda) install hypothesis
- See also:
<https://hackage.haskell.org/package/QuickCheck>

Many More Libraries

- Schema Validation and Serialization:
 - <https://marshmallow.readthedocs.io/en/latest/>
 - For JVM / Apache: <https://avro.apache.org/>
- Model Validity
 - http://scikit-learn.org/stable/modules/cross_validation.html
- Testing ML features:
<https://github.com/machinalis/featureforge>
- Built-in Stats:
<https://docs.python.org/3/library/statistics.html>

Questions? (and a short break)



Source: <https://gradientproductions.wordpress.com/>

kjamistan



Data Unit Tests



Quick Poll

I write tests for my data science code...

What is Unit Testing?

- Test a small unit of code
 - Define inputs, outputs and behavior
- Not for outside software, APIs or integration testing
 - Usually internal code and tools only
- Can exist in larger suites
- Often used with automated testing before releases (or even just on merges)
- Code Coverage

Why Test ?

It is important to test your own code: don't assume that some testing organization or user will find things for you. But it's easy to delude yourself about how carefully you are testing, so try to ignore the code and think of the hard cases, not the easy ones. To quote Don Knuth describing how he creates tests for the TEX formatter, "I get into the meanest, nastiest frame of mind that I can manage, and I write the nastiest [testing] code I can think of; then I turn around and embed that in even nastier constructions that are almost obscene."

Why Data Unit Testing?

- Data Science uses Code!
 - data engineering, pipelines, extraction
- Testing small units of data
 - Within expected ranges
 - Display expected heuristics (correlation)
 - Show anomalies or erratic behavior
- More data science code, means more (generally untested) code. There should be tests! 🏹

BUT HOW?!?



Testing Basics

- Use libraries, don't reinvent the wheel
- Learn about mocking outside APIs
 - <https://docs.python.org/3/library/unittest.mock-examples.html>
- Fake the data!
 - <https://faker.readthedocs.io/en/master/>
 - <https://github.com/pereorga/csvfaker>
- Watch Ned Batchelder's testing talk:
<https://www.youtube.com/watch?v=FxSsnHeWQBY>

How to Implement Testing

- Use Version Control
- Use Automated Testing
 - Pytest library: <https://docs.pytest.org/en/latest/>
 - Continuous Integration Tests (Jenkins, Travis, TeamCity, etc)
- Regular code reviews and merge procedure
 - <http://www.bettercode.reviews/>

Quick Poll

I find code reviews...

Testing for Pipelines

- Does your framework have a built-in testing or validation toolset?
 - Testing Data Quality in Apache Spark:
<http://blog.cloudera.com/blog/2015/07/how-to-do-data-quality-checks-using-apache-spark-dataframes/>
 - <https://github.com/FRosner/drunken-data-quality>
- Testing with Apache Beam:
<https://beam.apache.org/documentation/pipelines/test-your-pipeline/>
- Tip: Check your framework's documentation first, then look for possible third-party fits

Testing Incoming (or existing) Data

- Tests:
 - Expected thresholds
 - Specific distributions or correlations
 - What to do when criteria not met?
- What determines an anomaly or outlier?
 - How are outliers handled?
- Automated “fixing”
 - Invalid data > Valid data via simple functions

Ok, I'm sold... But...



Testing & Validation Benefits

- Lost Revenue
 - Hours lost chasing bugs, outliers, bad data
 - Poor and costly decisions
 - Inaccurate predictions due to invalid data
- Gained Revenue
 - Happier employees and easier hires
 - Ability to automate (and TRUST) more workflows
 - Easier to find bugs
 - High priority to higher level thinking tasks

Questions? (and a short break)



Quick Poll

What pipeline do you use?



Case Study: Validating Router Environment Data



Defining the Problem

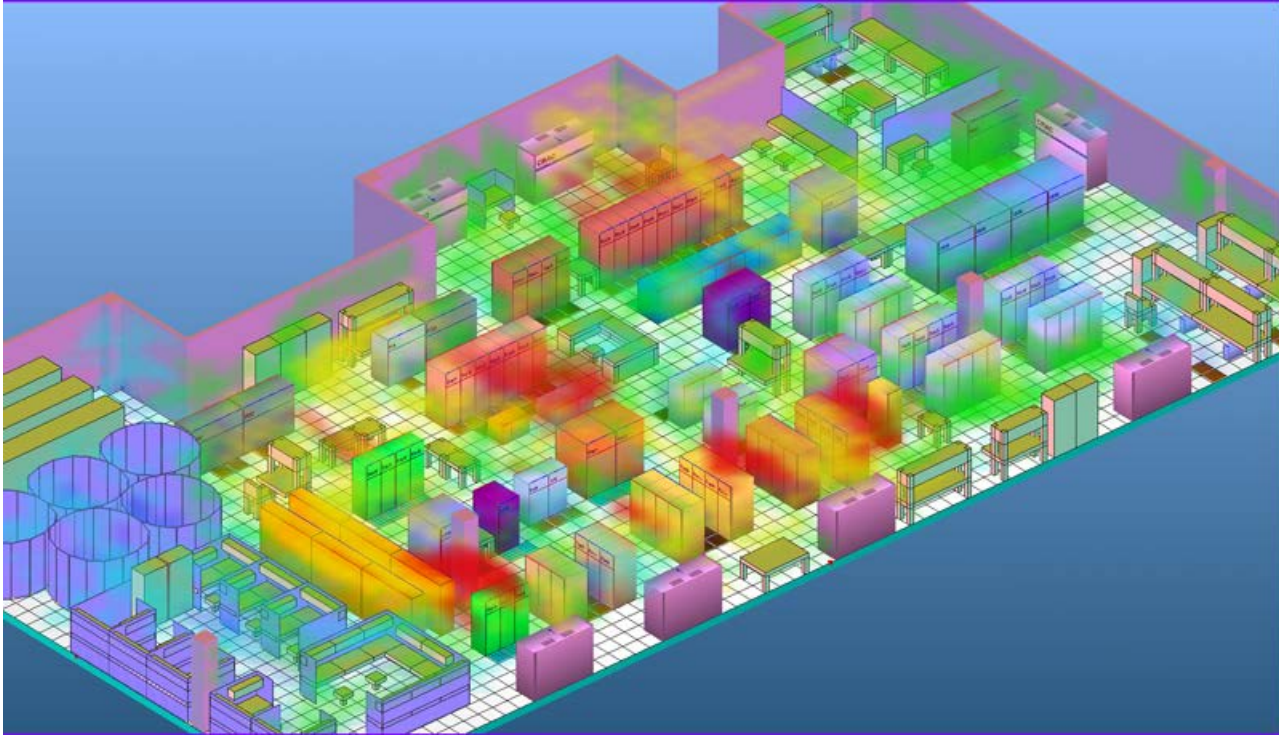


Image Source: <http://tileflow.com/>

Possible Solutions?

Poll Question:

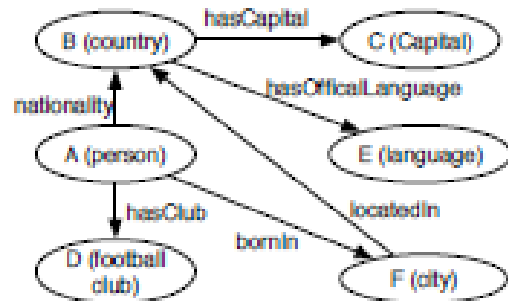
How might you validate incoming sensor data?



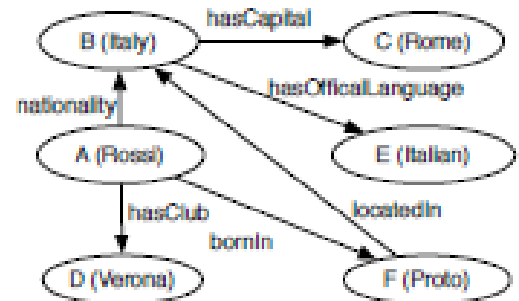
What's Happening in Academia?



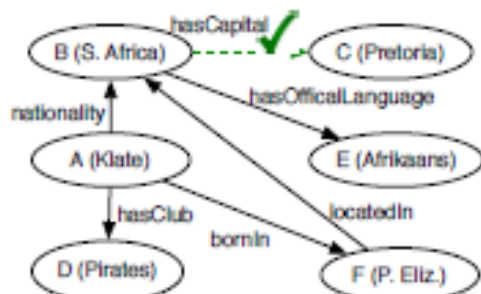
Katara: Crowd + IE



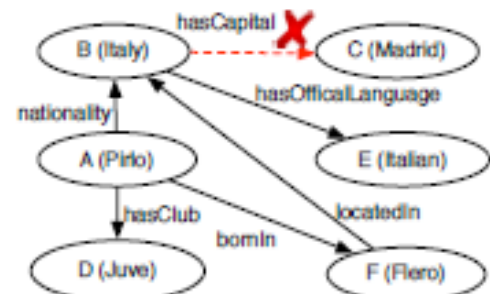
(a) A table pattern φ_s



(b) t_1 : KB validated



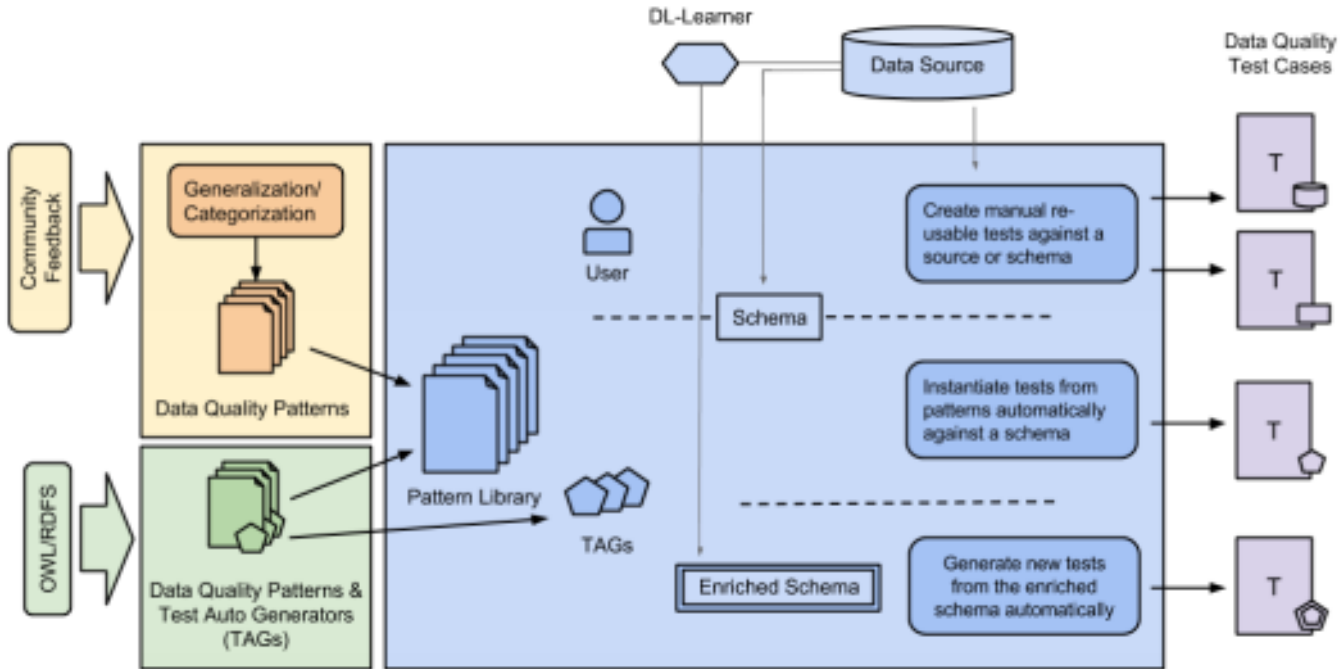
(c) t_2 : KB & crowd validated



(d) t_3 : erroneous tuple

X Chu, Morcos, Ilyas et al. 2015

Databugger



Unsupervised Anomaly Detection

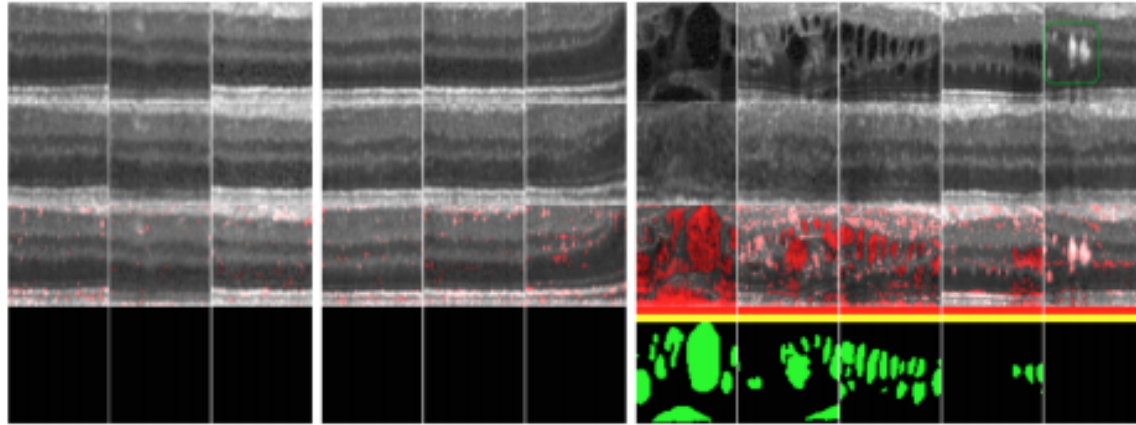


Fig. 3. Pixel-level identification of anomalies on exemplary images. First row: Real input images. Second row: Corresponding images generated by the model triggered by our proposed mapping approach. Third row: Residual overlay. Red bar: Anomaly identification by *residual score*. Yellow bar: Anomaly identification by *discrimination score*. Bottom row: Pixel-level annotations of retinal fluid. First block and second block: Normal images extracted from OCT volumes of healthy cases in the training set and test set, respectively. Third block: Images extracted from diseased cases in the test set. Last column: Hyperreflective foci (within green box). (Best viewed in color)

Quick Poll

My favorite part of this seminar was...

Congrats! 🎉

THANK YOU!

- Resource post: <https://blog.kjamistan.com/practical-data-cleaning-with-python-resources/>
- If you can, please take a minute to give me some feedback
 - <http://bit.ly/practical-data-feedback>
- Reach out anytime:
 - @kjam on Twitter / Slack / GitHub
 - katharine@kjamistan.com