Time Series Data

Clarifying Practical Approaches

Ananya Joshi

Warm-Up Until 9:35

A. Where are you in the project?
Write it on a note & post on board.

0: What project?
1: I found a dataset.
2: I explored the dataset.
3: I have a project question.
4: I have ideas for approaches.
5: I finished the project.

B. Open up the companion doc!

https://shorturl.at/vy089

C. Make groups of max 5!

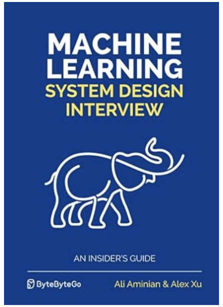
Agenda

By the end of class, you should be able to:

- Plan out your own applied project using time series data
- Identify and compare the different components of working with time series data
- Practically apply basic skills corresponding to each of these components

[Selected] Components of Time Series Data

1. **Curation:** What are properties of informative time series data?
2. **Task Selection:** What tasks can I complete with this data?
3. **Preparation:** How can I prepare my data before I feed it into a model?
4. **Evaluation:** How can I evaluate my approaches?



MACHINE LEARNING SYSTEM DESIGN INTERVIEW

AN INSIDER'S GUIDE

ByteByteGo Ali Aminian & Alex Xu

Intern Project Activity

We are starting a new job at a healthcare startup which is deploying a smart watch. Our only objective is to develop a useful analytics tool using data from this watch*

Basics of Time Series Data

Time series data has measurements occurring over time.

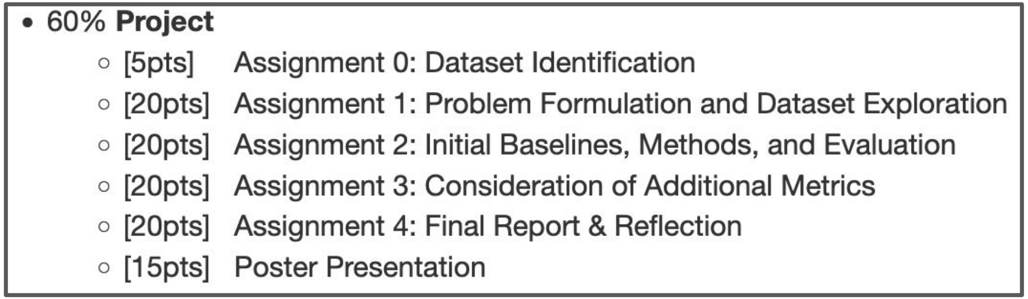
Examples:

Stock Prices

Hours of Sleep

G.P.A

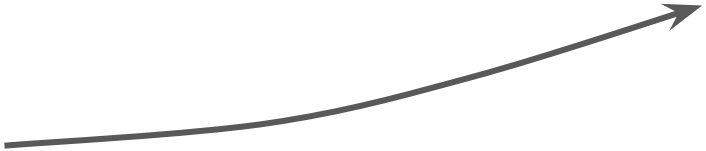
<u>**Think about your project!**</u>



- 60% **Project**
  - ○ [5pts] Assignment 0: Dataset Identification
  - ○ [20pts] Assignment 1: Problem Formulation and Dataset Exploration
  - ○ [20pts] Assignment 2: Initial Baselines, Methods, and Evaluation
  - ○ [20pts] Assignment 3: Consideration of Additional Metrics
  - ○ [20pts] Assignment 4: Final Report & Reflection
  - ○ [15pts] Poster Presentation

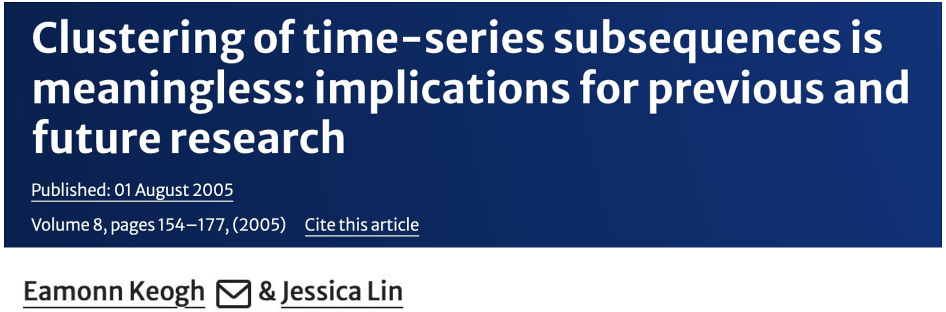
Why: Many practical projects rely on time series data because the world is changing, and projects need to keep up with these changes.

Examples of Pitfalls



Why might your time series data look like this? (COVID-19 Cases)

Data Curation



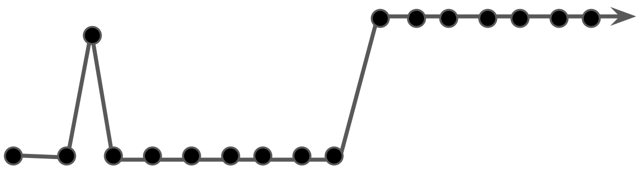
Clustering of time-series subsequences is meaningless: implications for previous and future research

Published: 01 August 2005

Volume 8, pages 154–177, (2005) Cite this article

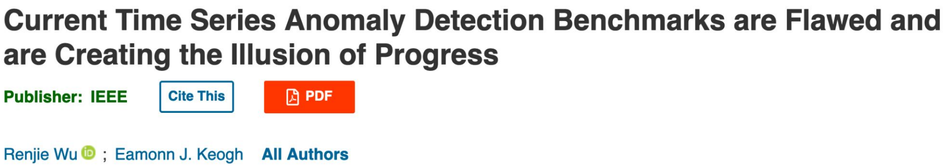
Eamonn Keogh & Jessica Lin

Task Selection



What would happen if you tried to cluster these values?

Data Preparation



Current Time Series Anomaly Detection Benchmarks are Flawed and are Creating the Illusion of Progress

Publisher: IEEE

Renjie Wu ; Eamonn J. Keogh All Authors

Evaluation

https://forms.gle/5Zwen3rYMxmb38oF7

Part 1: Data Curation

The smartwatch has the following sensors that provide data per second.

- GPS
- Gyroscope
- Heart Rate Monitor
- Blood Oxygen Saturation
- Skin Temperature
- Room Temperature
- IDs of Nearby Smartwatches

How do we identify data streams worth using?

Vote for up to 3 Streams Using Google Forms Link on Document!

Understanding Your Data

Guiding Questions [Together]

1. What phenomena is being measured?
   a. What values are not measurable?
   b. How does the quality of the measurement change?
   c. Are you ok with using this data? [fairness? bias?]
2. What aspects of the phenomena are not captured by the measurement?
3. What data standards do you need?
   a. How will you ensure that standard is met continuously?

<u>Heart Rate Sensor Specifications</u>

- Maximum 10 mV delta
- 16 bit resolution
- 30 second to 5 minute recording duration before necessary cool-down
- 300 samples per second sampling rate (max reported to watch per second)

Modified from Kardia Mobile Specifications

1/3 Intentional Data Aggregation/Fusion

Example Objectives: Surface behaviors while reducing noise, data overwhelm

Example: Taking the maximum over the samples in a time range (like the heart sensor)

Guiding Question: What assumptions am I making when (and how) I aggregate data?

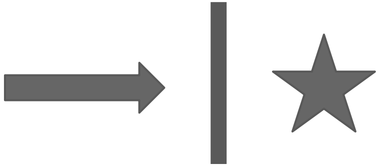
Over Time

Over Space

Over Component

Enrichment: Sensor takes 300 samples/second, but watch only reports per second. Compare using the max vs. average metrics with the heart monitor

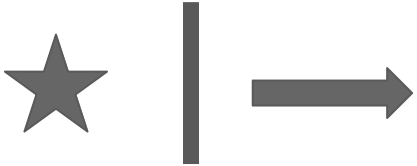
2/3 Censoring Data - Survival Analysis



Right Censoring

The data collection ended before the event of interest could occur.

E.g. Missing events because of the sensor’s cool down time



Left Censoring

The event occurred before the study began but the exact time is not known.

E.g. Testing satisfaction among subgroup of customers who have issued a complaint

Interval Censoring

The event occurred sometime in the interval but it is not clear when.

E.g. Aggregation over time

3/3 Exploratory Strategies and Validation Checks

Data Validation: *Without* Foundational Models

1. What is the form of the data that I expect?
   Think: ranges, data types, invariants

2. **What should I do with data that doesn’t fit my expectations?**

3. **How can I ensure that the data remains in the form I expect?**
   - Imputation decisions
   - Deleting data
   - Retrospective & Prospective

Committing to these design choices!

Slides by Ananya Joshi: aajoshi@andrew.cmu.edu

Activity 1: Data Curation (10 mins)

The smartwatch has the following sensors that provide data per second.

- GPS
- Gyroscope
- Heart Rate Monitor
- Blood Oxygen Saturation
- Skin Temperature
- Room Temperature
- IDs of Nearby Smartwatches
- Money Spent that Day

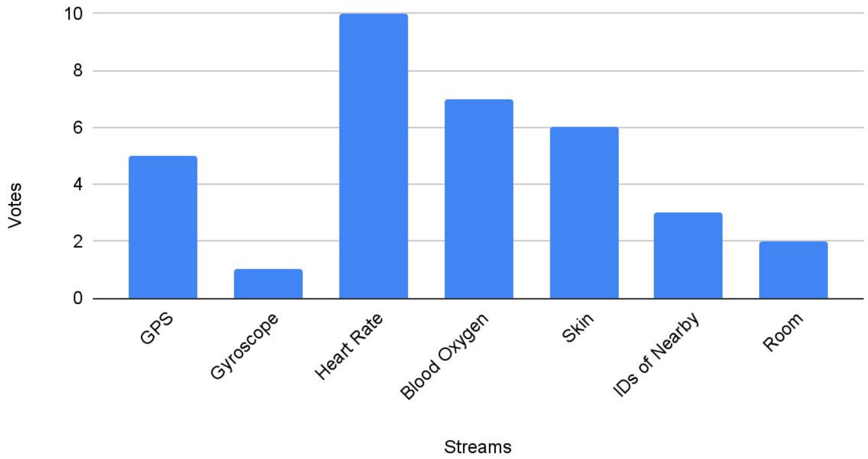
Question:

A. Discuss these streams considering any possible
   a. Aggregation strategies
   b. Censoring impacts
   c. Validation approaches
B. Pick 3 data streams you would use and **vote on the link!**

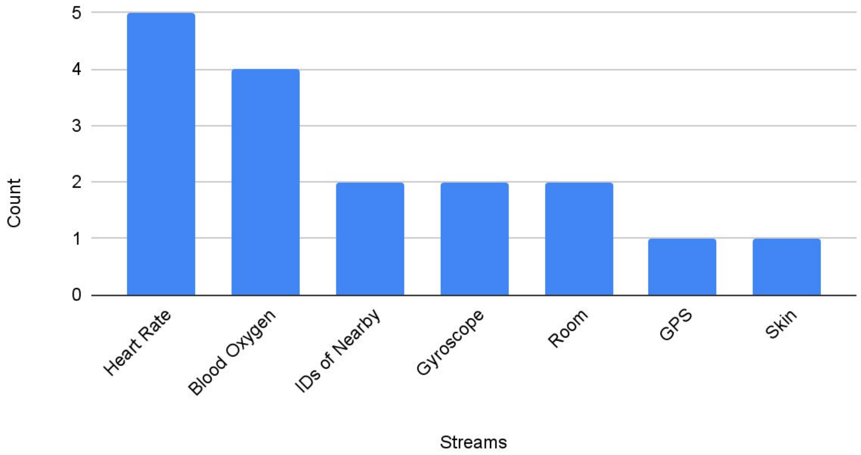
https://forms.gle/M514x63zvd9rX7BVA

Debrief 1:

Initially Selected Streams



Final Streams



How did your opinion change?

Part 2: Task Design

Data Sensors Using

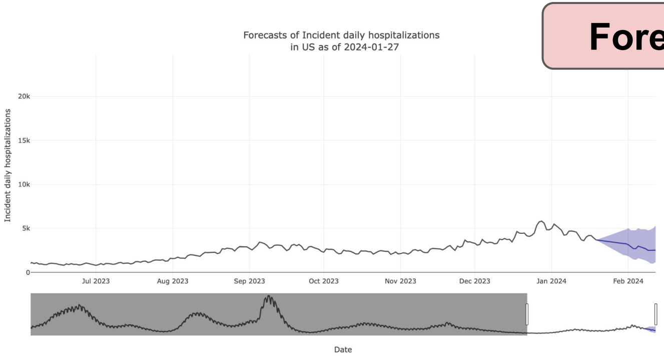
- GPS
- ~~Gyroscope~~
- ~~Heart Rate Monitor~~
- ~~Blood Oxygen Saturation~~
- ~~Skin Temperature~~
- ~~Money Spent that Day~~
- Room Temperature
- IDs of Nearby Smartwatches

What should we do with this data?

Data Science Project Scoping Guide

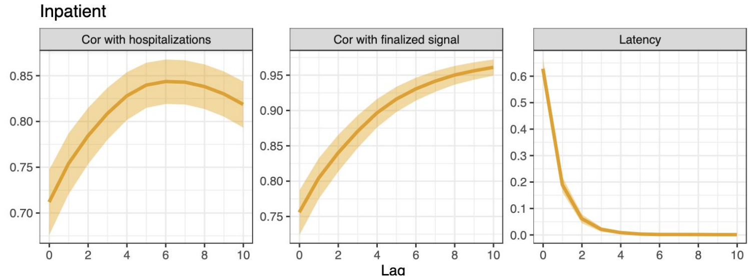
Categories of Deployable Tasks

ForecastHub



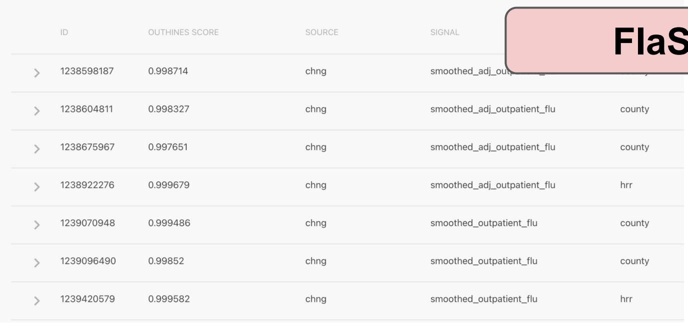
1. Forecasting Future Values

Predicting Hospitalizations



2. Correlation Analysis

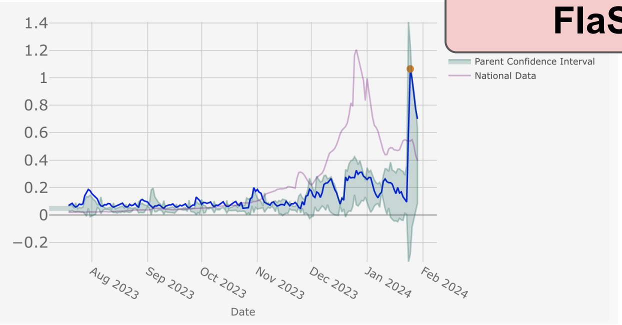
FlaSH



	ID	OUTHINES SCORE	SOURCE	SIGNAL	
>	1238598187	0.998714	chng	smoothed_adj_out	
>	1238604811	0.998327	chng	smoothed_adj_outpatient_flu	county
>	1238675967	0.997651	chng	smoothed_adj_outpatient_flu	county
>	1238922276	0.999679	chng	smoothed_adj_outpatient_flu	hrr
>	1239070948	0.999486	chng	smoothed_outpatient_flu	county
>	1239096490	0.99852	chng	smoothed_outpatient_flu	county
>	1239420579	0.999582	chng	smoothed_outpatient_flu	hrr

3. Ranking Data

FlaSH



4. Outlier Detection

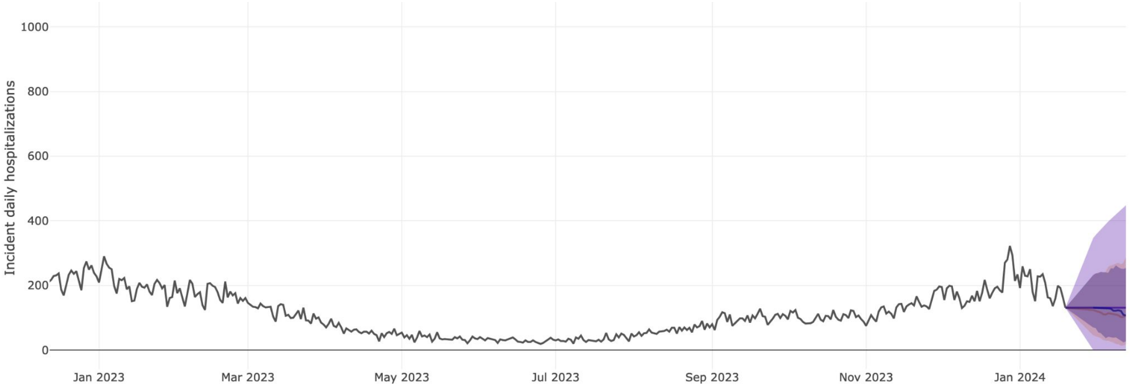
Curate | **Task** | Prep | Evaluate

Slides by Ananya Joshi: aajoshi@andrew.cmu.edu

Forecasting

Active Research Area!

Sample Question: What might this data look like in the next two weeks?



Ensemble

Baseline

Q1. What could go wrong with using rolling averages aggregation strategy for forecasting?

Q2. What techniques can you use to identify how much you weigh each estimator for the ensemble model?

Outlier Detection

Active Research Area!

Standard Approach follows a formula:

1. $\hat{X}_t$

Predict a Value

2. $\hat{X}_t - X_t$

Calculate Difference

3. $(\hat{X}_t - X_t)/\sigma$

Contextualize Difference

4. $\cancel{(\hat{X}_t - X_t)/\sigma \overset{?}{>} 3}$

~~**Send alerts**~~

Via Forecasting

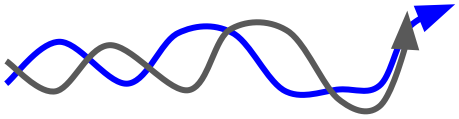
In a population-sensitive way

Compare to historical values to **rank data.**

Slides by Ananya Joshi: aajoshi@andrew.cmu.edu

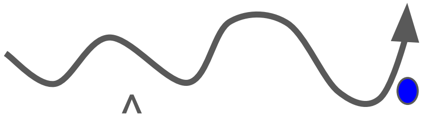
* Interesting Combined Strategy

1. **Correlate** Different Streams to Identify Relationships



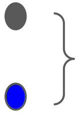
$A_{t+1} = B_t$

2. **Forecast** Current Values Based on those Relationships



$\hat{B}_t = A_{t-1}$

3. Use those forecasts to generate **Outlier** scores



> 99% of historical values

4. **Rank** those outliers

1. A_t
2. B_t
3. C_t

Part 3: Data Preparation

Data Sensors Using

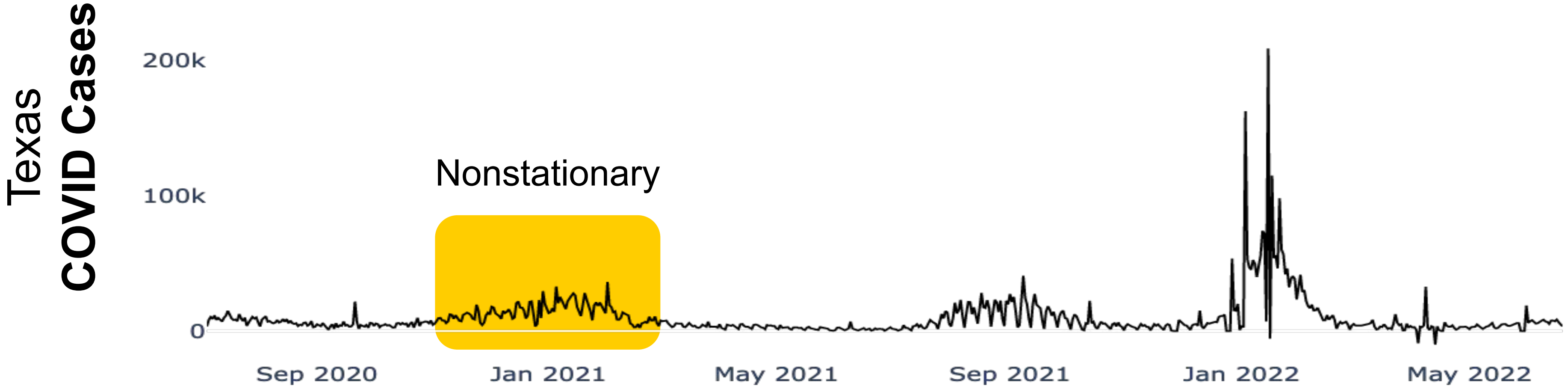
- GPS
- Room Temperature
- IDs of Nearby Smartwatches

+ Publicly available weather data.

Business Task: Provide city planners information on where to add more portable heaters for bus stops.

1. How might this relate to outlier detection?
2. What might be problems with the outlier detection approach here?
3. *How can we prepare the data?*

1/3 Addressing Data



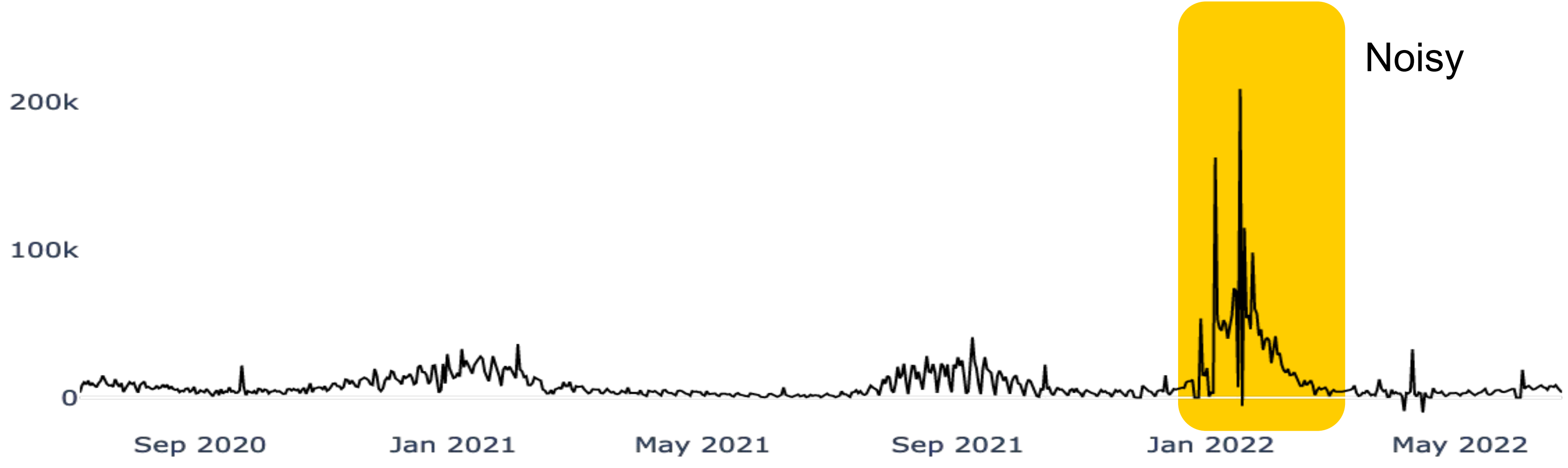
What do you notice?

- Changepoint detection approach (How might this fail?)

Slides by Ananya Joshi: aajoshi@andrew.cmu.edu

2/3 Addressing Data

Texas
COVID Cases



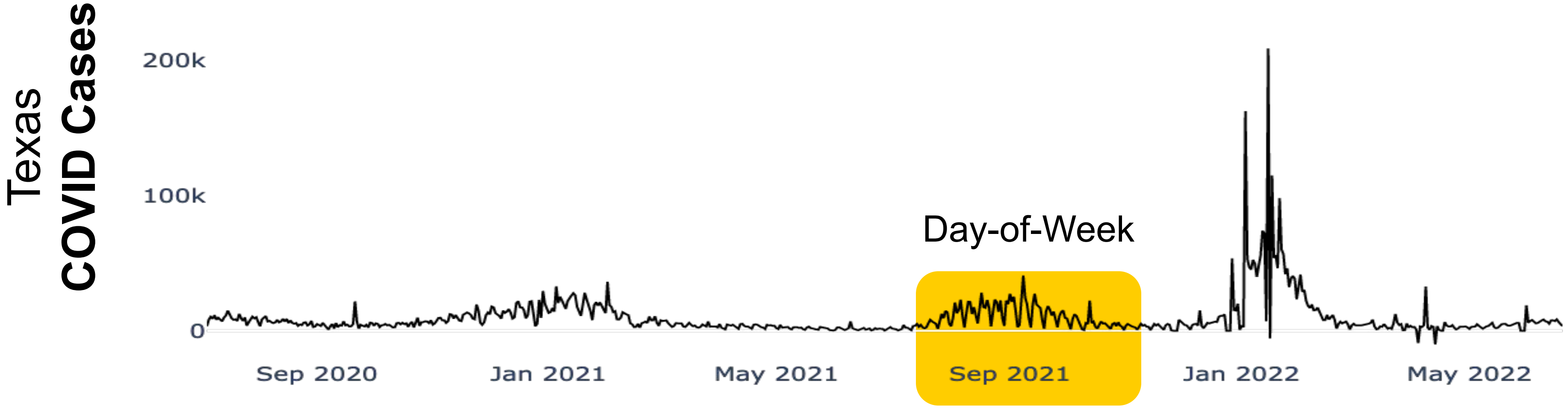
What do you notice?

- What denoising strategies can we apply (pros/cons?)

Curate | Task | **Prep** | Evaluate

Slides by Ananya Joshi: aajoshi@andrew.cmu.edu

3/3 Addressing Data



What do you notice?

- How can you identify seasonality in the data?
- What mechanisms can we use to reduce the impact of seasonality?

Curate | Task | **Prep** | Evaluate

Slides by Ananya Joshi: aajoshi@andrew.cmu.edu

Part 2 & 3 Activity (10 mins):

1. **Data Streams**

- GPS
- Room Temperature
- IDs of Nearby Smartwatches

+ Publicly available weather data.

2. Business Task: Provide city planners with data on where to add more portable heaters for bus stops.

3. Requested Deliverable: Flesh out a 5 step process, from input to output, for this task in your groups. Consider taking on different roles e.g.: curator, task definer, data preparer.

Hint (if needed) at the 2 minute mark!

Part 4: Evaluation & Monitoring Strategies

1. **Data Streams**
   - GPS
   - Room Temperature
   - IDs of Nearby Smartwatches

   + Public weather data

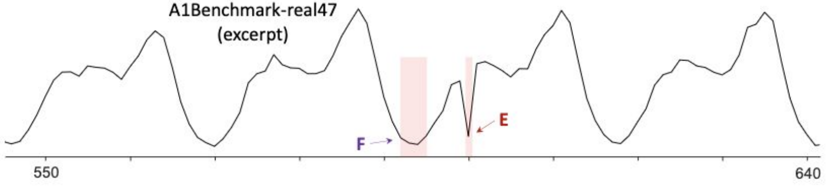
2. Business Task: Provide city planners with data on where to add more heated bus-stops.

3. Simple Approach: Every week, we estimate bus stop temperatures (with some noise estimate), detect regions which deviate from their historical average, and rank them by the magnitude of deviation.

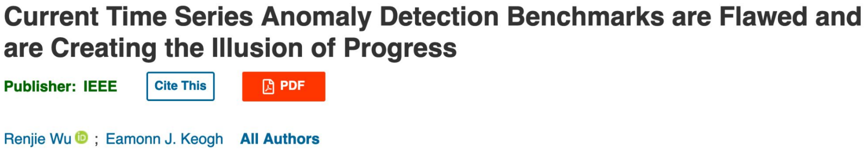
4. Does it actually work?

Designing an Evaluation when you have Labels

Do these labels even mean anything?



Wu and Keogh et. al.



Current Time Series Anomaly Detection Benchmarks are Flawed and are Creating the Illusion of Progress

Publisher: IEEE Cite This PDF

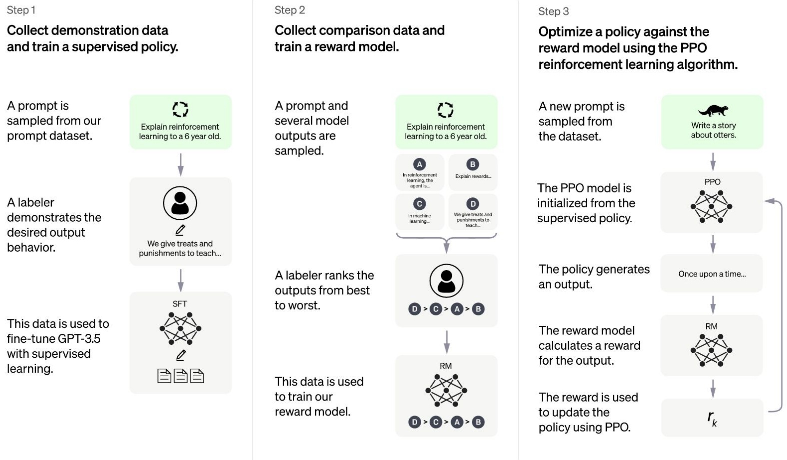
Renjie Wu ; Eamonn J. Keogh All Authors

Common Pitfalls:

1. Are only global outliers labeled? Could these be detected using 1 line of code?
2. Are there too many outliers (e.g. > 1%) ?
3. Are there any explanations available for the classification?
4. Left-censoring bias affecting anomaly detection sets.

What new challenges do we have when using synthetic labels?

Designing an Evaluation when you Need Labels



Step 1

Collect demonstration data and train a supervised policy.

A prompt is sampled from our prompt dataset.

A labeler demonstrates the desired output behavior.

This data is used to fine-tune GPT-3.5 with supervised learning.

Step 2

Collect comparison data and train a reward model.

A prompt and several model outputs are sampled.

A labeler ranks the outputs from best to worst.

This data is used to train our reward model.

Step 3

Optimize a policy against the reward model using the PPO reinforcement learning algorithm.

A new prompt is sampled from the dataset.

The PPO model is initialized from the supervised policy.

The policy generates an output.

The reward model calculates a reward for the output.

The reward is used to update the policy using PPO.

Explain reinforcement learning to a 6 year old.
We give treats and punishments to teach...
SFT
Explain reinforcement learning to a 6 year old.
A
In reinforcement learning, the agent is...
B
Explain rewards...
C
In machine learning...
D
We give treats and punishments to teach...
D > C > A > B
RM
D > C > A > B
Write a story about otters.
PPO
Once upon a time...
RM
r_k

OpenAI's Evaluation Strategy (openai.com)

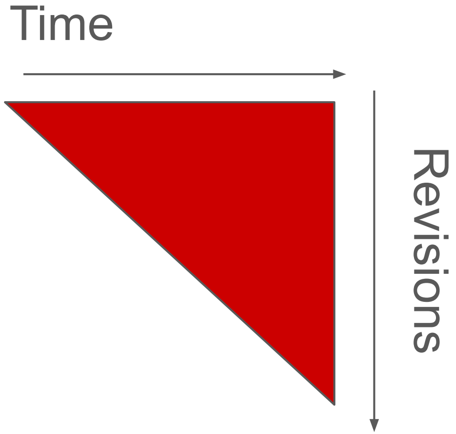
+ Synthetic Strategies

Keep In Mind:

1. Pre-registration before data collection begins is important.
2. People (even experts) don't agree. How will you address this?
3. You may need to design your own metrics.
4. Is there a way to get proxy evaluation data?

What are other ways to evaluate your time series experiments?

Putting it All Together: A FlaSH Demonstration



Rank outliers from the most recently received raw data at Delphi daily so that data reviewers can find notable events of interest from the data, quickly.

Data Curation

Task Identification

Preparation / Method

Evaluation Strategies

Quick Tour of Delphi’s FlaSH Project

Reflections & Takeaways

[Curating, Task Selection, Preparing, Evaluating]

Questions:

- How well did our initial imputation/validation strategies serve us?
- Which components can save you the most time in time series analysis?
- Which components need the most external feedback/information?
- What other aspects of a time series project falls outside these components?
- Which components is the most often ignored?

This is an iterative process that should be routinely revised for deployed tools!

Thank you!

By the end of class, you should be able to:

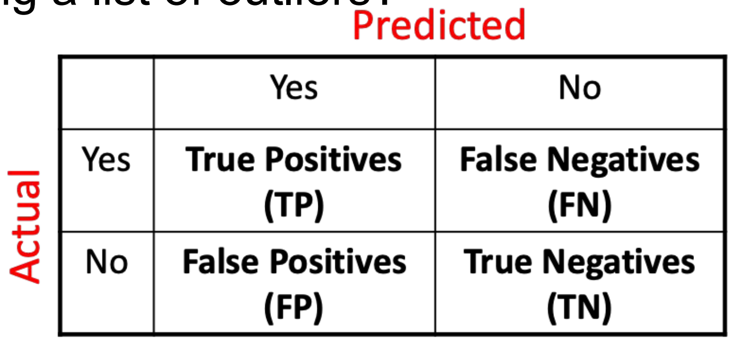
- Plan out your own applied project using time series data
- Identify and compare the different components of working with time series data
- Practically apply basic skills corresponding to each of these components

Quick Survey
https://forms.gle/HMLvNDA8GExTYG4a7

Email: aajoshi@andrew.cmu.edu

Metrics & Meanings Activity

1. What metric will help me better understand my model’s anomaly detection performance ?
   - A. Accuracy = (TP + TN) / (TP + TN + FP + FN)
   - B. Balanced Accuracy = TP / (TP + FN) + TN / (TN + FP)
2. My classifier has a threshold, and I want to see how the performance of this classifier varies over time. Which is better for anomaly or outlier detection?
   - A. ROC Curve
   - B. B. Pr-K Curve
3. Which gives me more information if I am making a list of outliers?
   - A. Precision : TP/(TP+FP)
   - B. Recall : TP/(TP+FN)



		Predicted	
		Yes	No
Actual	Yes	**True Positives (TP)**	**False Negatives (FN)**
	No	**False Positives (FP)**	**True Negatives (TN)**