Introduction to Anomaly Detection

MinneBOS 2019

Terran Melconian

terran@terran.us

413 376 5199

http://www.terran.us/talks/

Introduction

- **Outlier**: A data point which you suspect came from a different process than your other data.
- **Anomaly**: An outlier that you're happy you were bothered about.
- Can be caused either by:
  - changes in the world
  - changes in your data recording or processing systems
- You may care because:
  - you need to fix the situation
  - you consume the data
- Audience: people who have to deliver anomaly detection

Agenda

- Introduction and context
- Nature of the problem and objective
- Some specific techniques
- Meta-approach and framework
- Some more specific techniques

Speaker Background

Operations research & aeronautics

Software engineering & information retrieval

Consumer web operations & warehousing

Data science & analytics

Data science training & consulting

Bioinformatics & high dimensional ML

Why is this hard?

- High volumes of data at multiple levels of granularity, in time and entities
- Data points are not independent and statistical test assumptions are nowhere close to reasonable.
- Disagreement/uncertainty about what is or is not an anomaly
- Shortage of labeled training data

Realistic Expectations

- Expect to work at transforming your data prior to fitting your models
- Expect to build multiple models for different types of anomalies
- Expect to manually label detected outliers as desired and undesired on an ongoing basis
- Expect to have a substantial false positive rate
- Expect to manually diagnose the anomalies flagged by your models

What Ops is Like

- Operations runs on intuition and heuristics
  - Even if the person responding understands the theory, there usually isn't time to use it.
- What's it like being paged?
  - Christmas morning, about to leave the house to meet the family for a meal…
  - 5 AM, after having already been paged at 3 AM and 1 AM…
- Alerts need to be simple to diagnose and remediate.

Desirable Properties

- Interpretable with accessible intuition
- Easy to investigate further to diagnose problem
  - The data on which the alert was based should be stored somewhere convenient, before and after transformation
- Low false positive rate
  - Research says above 50% will surely be ignored
- Alert from only one place, as close to the root cause as possible.

Development Approach

- If you have known techniques, you can use them to evaluate unknown data.
- If you have known data, you can use it to try out new techniques.

- Unknown techniques *and* unknown data?

Data Sources

- Fully synthetic data
  - Captures arbitrarily complex scenarios
  - Time-consuming to write simulation at high level of fidelity
- Normal data with overlaid anomalies
  - Captures all system dynamics during normal operation
  - The “anomaly” is unambiguous and labelled
  - May not have correct system dynamics of the anomaly itself
- Real data with real anomalies
  - Not recommended starting point

Types of Anomalies

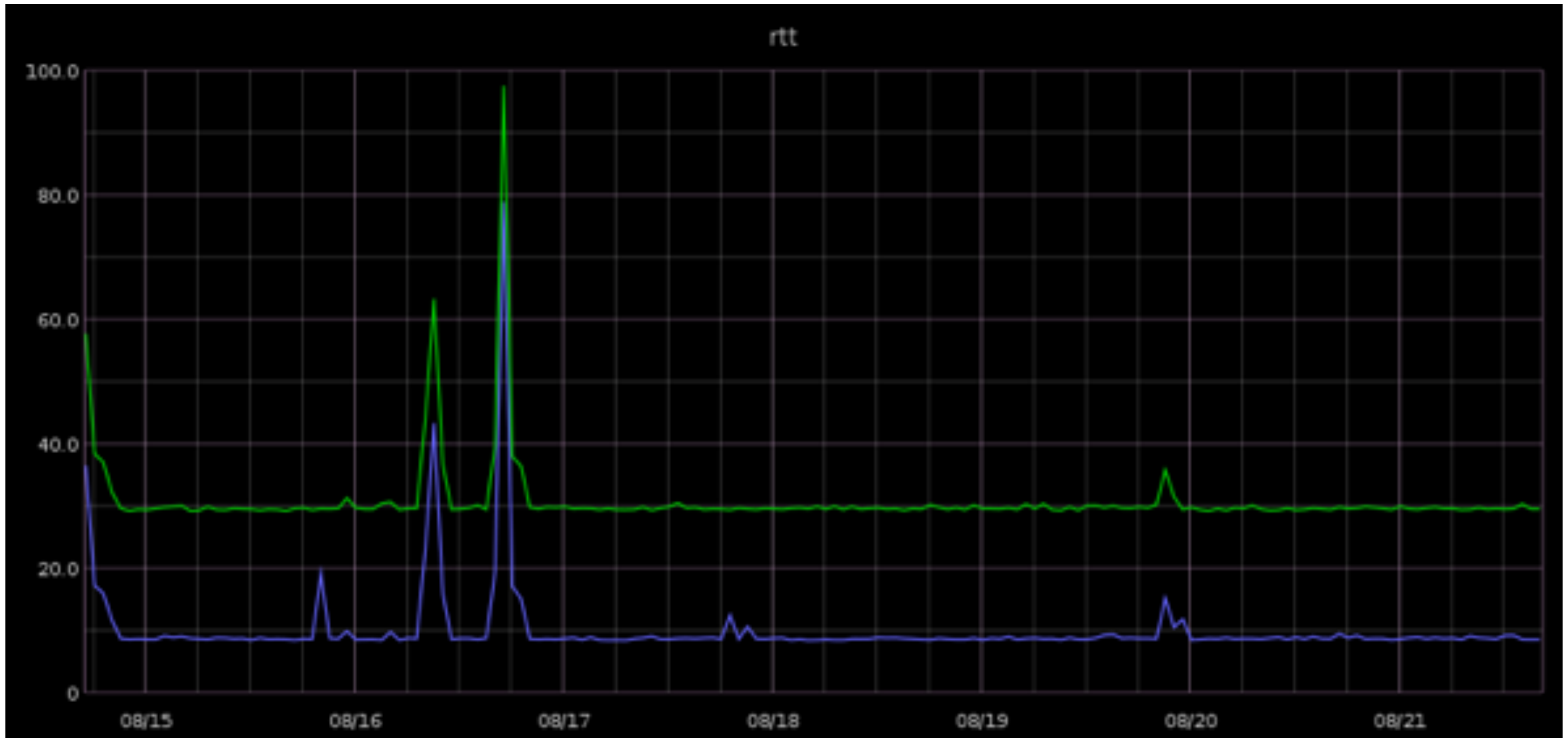
- Single unusual data point
- Unusual cluster of related data points
- Contextual anomaly relative to its neighbors
- Trend/trajectory in a new or problematic direction

- Types of data:
  - Quantitative or categorical data
  - Time series or independent multivariate points

  - All-discrete data must be time series or high dimensional

Fixed Thresholds

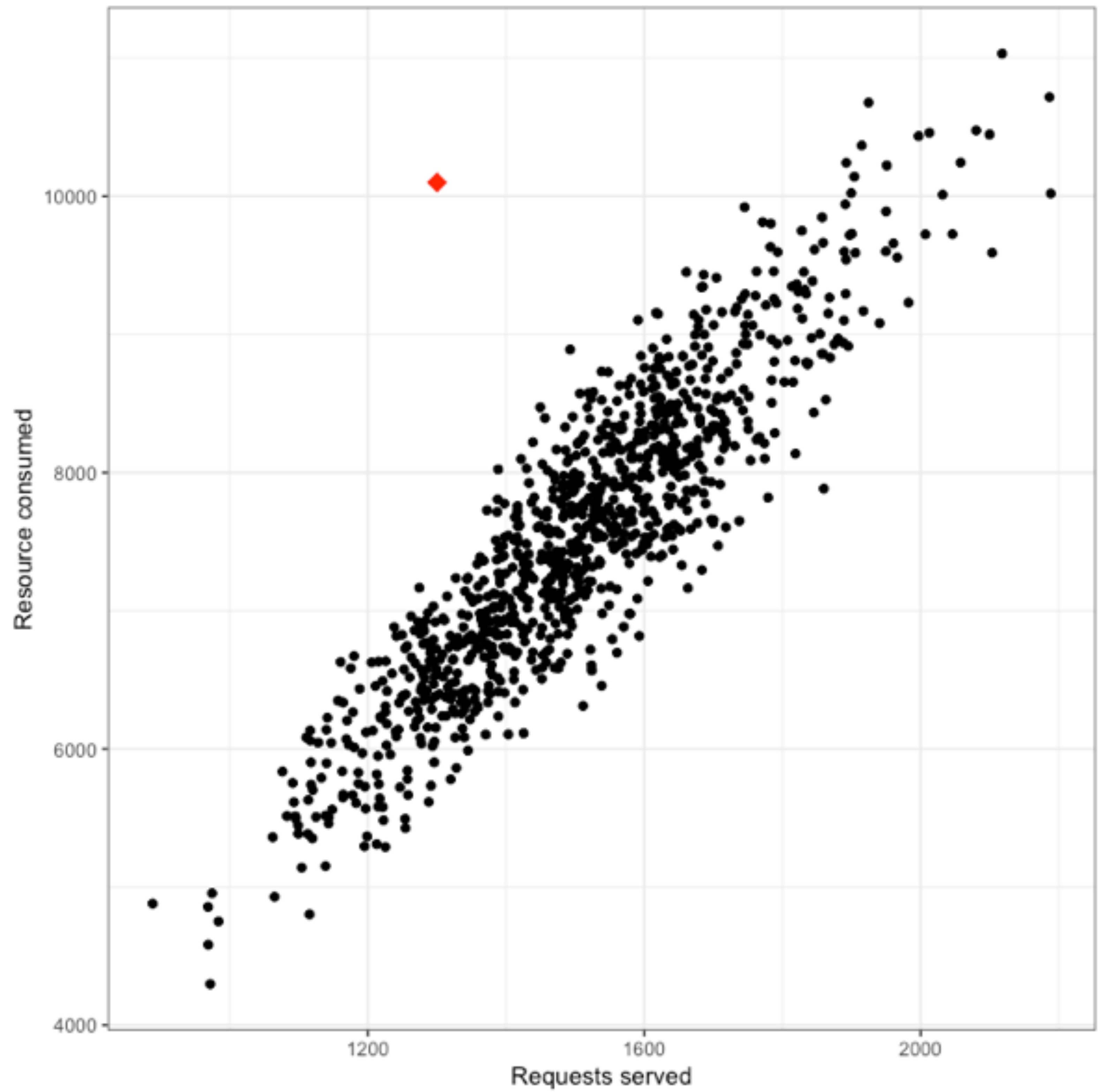
- Simplest and most commonly implemented method.
- Fixed min/max on metric values, tuned by hand. Alert when exceeded.
- Applied to one metric at a time, or to a simple aggregation such as a sum.
- Usually has a high false negative rate (missed true anomalies)



Fixed Thresholds

- Applies to single points or time series, quantitative metrics only.

- Can't find some pretty common cases

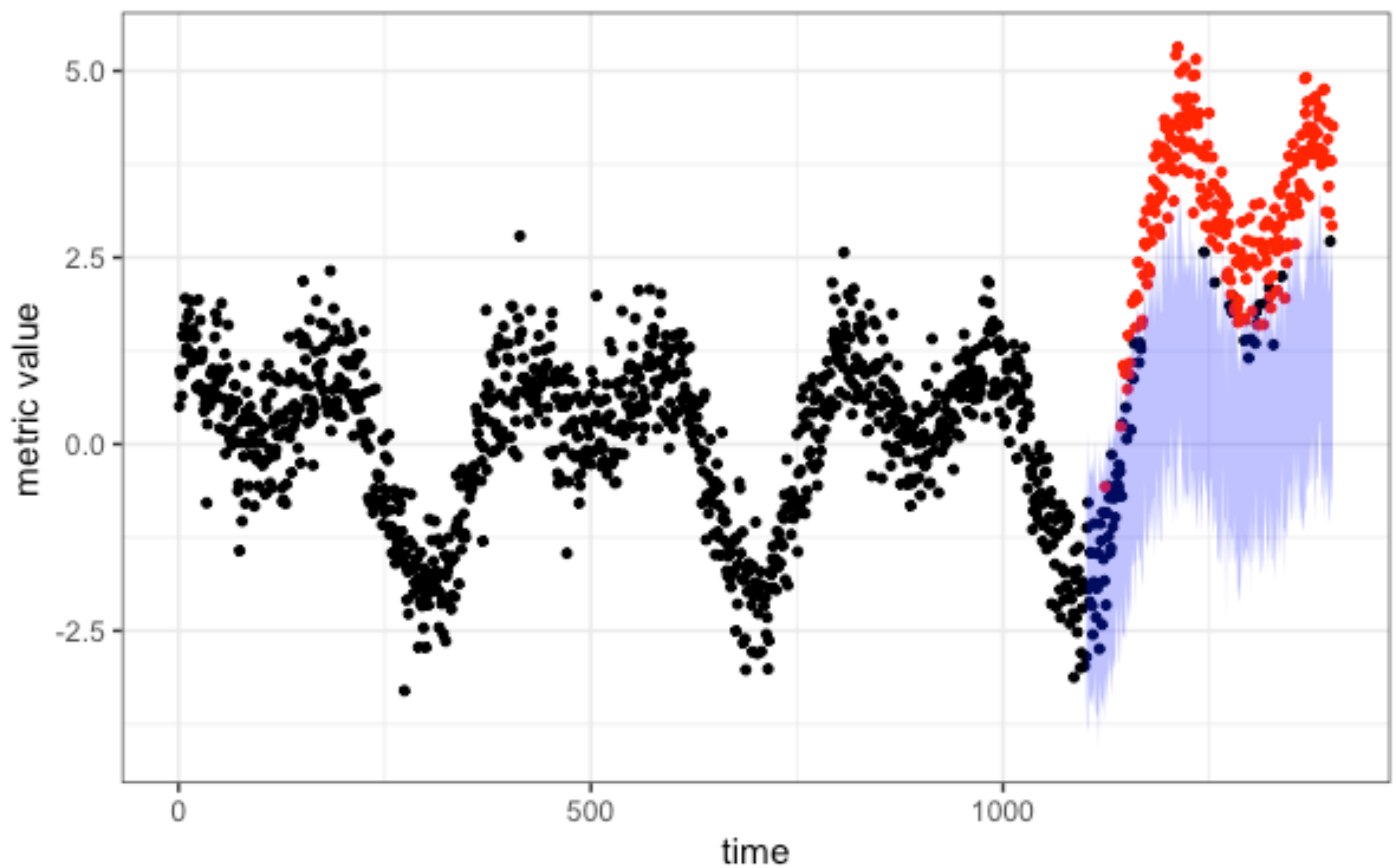


Seasonal Time Series

- “Seasonal” refers to any time period, including daily or weekly, not just annual.
- Time series techniques, e.g.
  - Holt-Winters
  - ARIMA models with external seasonal terms
- Commonly implemented in SaaS monitoring tools.
- Works well when the primary driver of variation is time of day, day of week, or season of year.
- Fails as soon as you have a holiday, deployment, or feature change on a different schedule.

Seasonal Time Series

- Works well when your system actually has consistent periodicity and you can alert on a single metric at a time.



Ratios

- Good place to start. Deserves respect.
- Decorrelates your metrics along one relationship.
- Reduces the number of places you alert for the same phenomenon.
- Try decomposing as **effect / cause** or **part / whole**
- Try to avoid ratios like **Part A / Part B** as this can be noisy and nonlinear when Part B is low.
  - If you must do this, log transform it.

Ratios

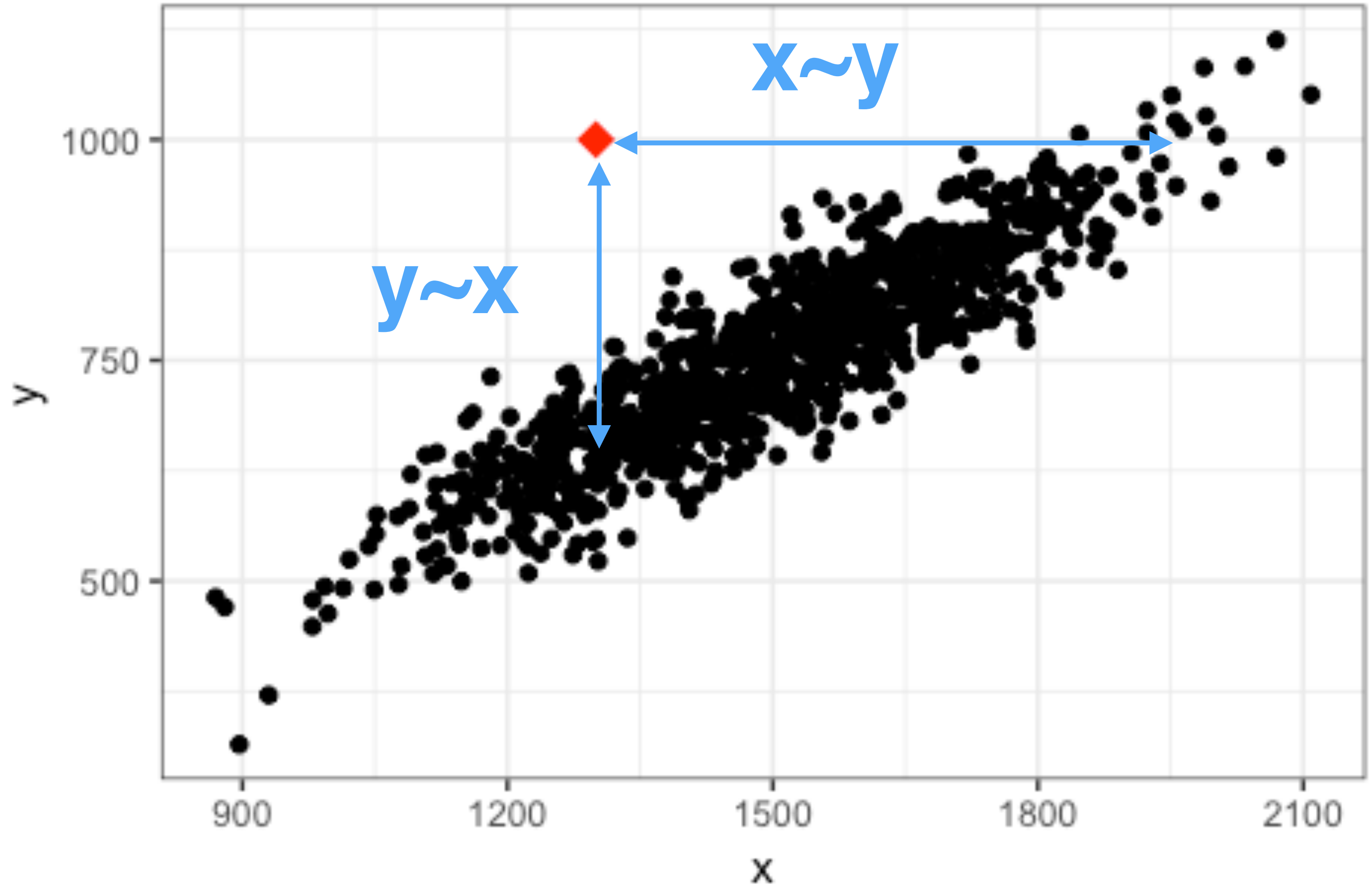
- Web ops example:
  - Request rate (apply your seasonal model here)
  - Database calls / request
  - Disk seeks / database call
  - milliseconds / disk seek
- another example:
  - dollars / request
  - North America dollars / total dollars
  - New England dollars / North America dollars
  - display ad dollars / total dollars

Linear Regression

- Most accessible multivariate technique.
- Choose one quantitative feature and build a model to predict it with your other features of any type.
- Predict the value of your feature with the model and get a z-score for how far the actual value is from the prediction.
- Remember your observations are probably not independent and the z-score cannot be turned into a p-value in the usual way.

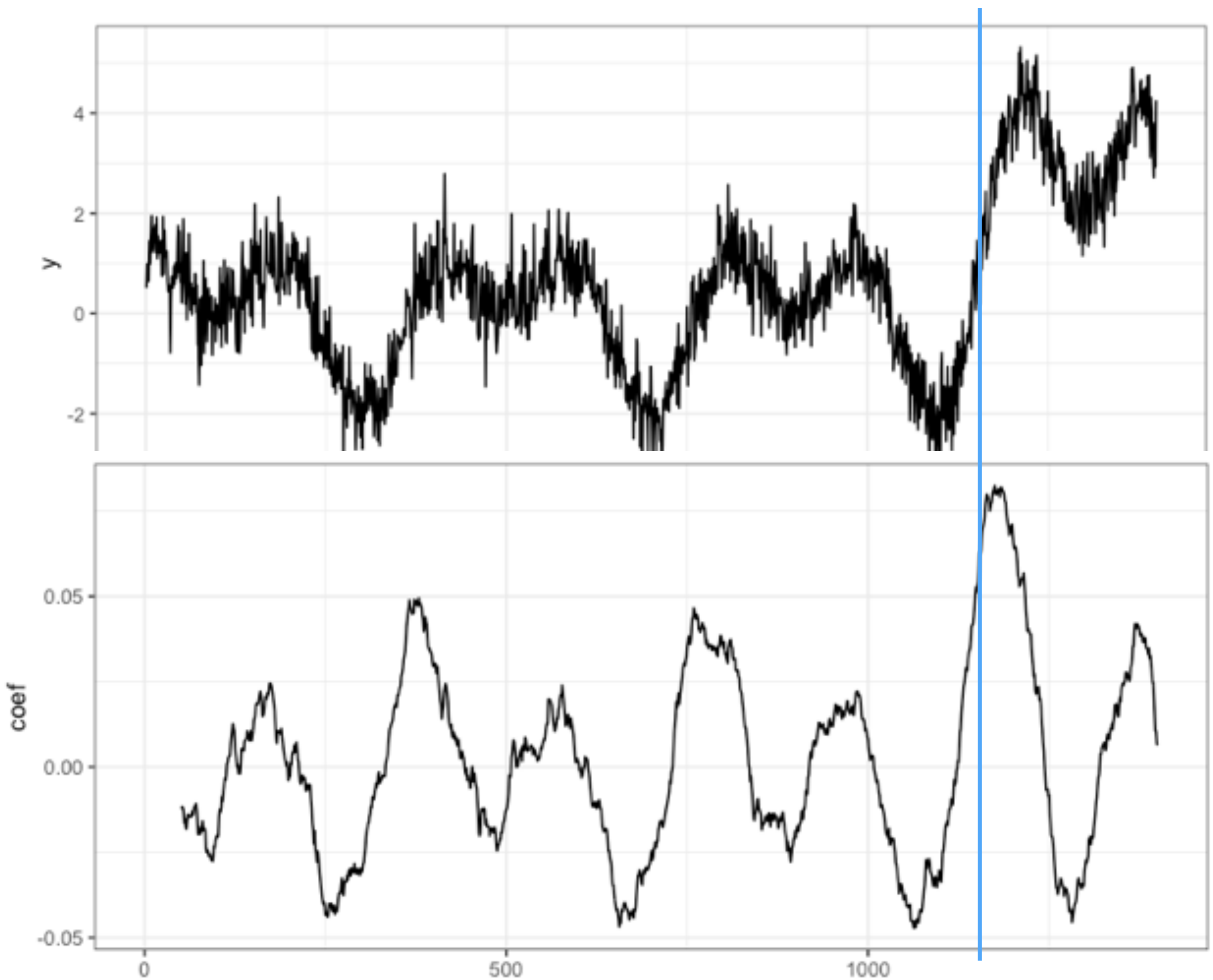
Linear Regression

- Works on data with at least one quantitative feature
- You may want to try multiple features as the dependent variable



Linear Regression

- Another use: fit trends to local regions in time series, then use the coefficients of the model as the feature.



Meta-Approach

- Transform original features into a few scalar metrics which indicate “unusualness”

- Combine these metrics and tune thresholds using your limited supply of labeled data

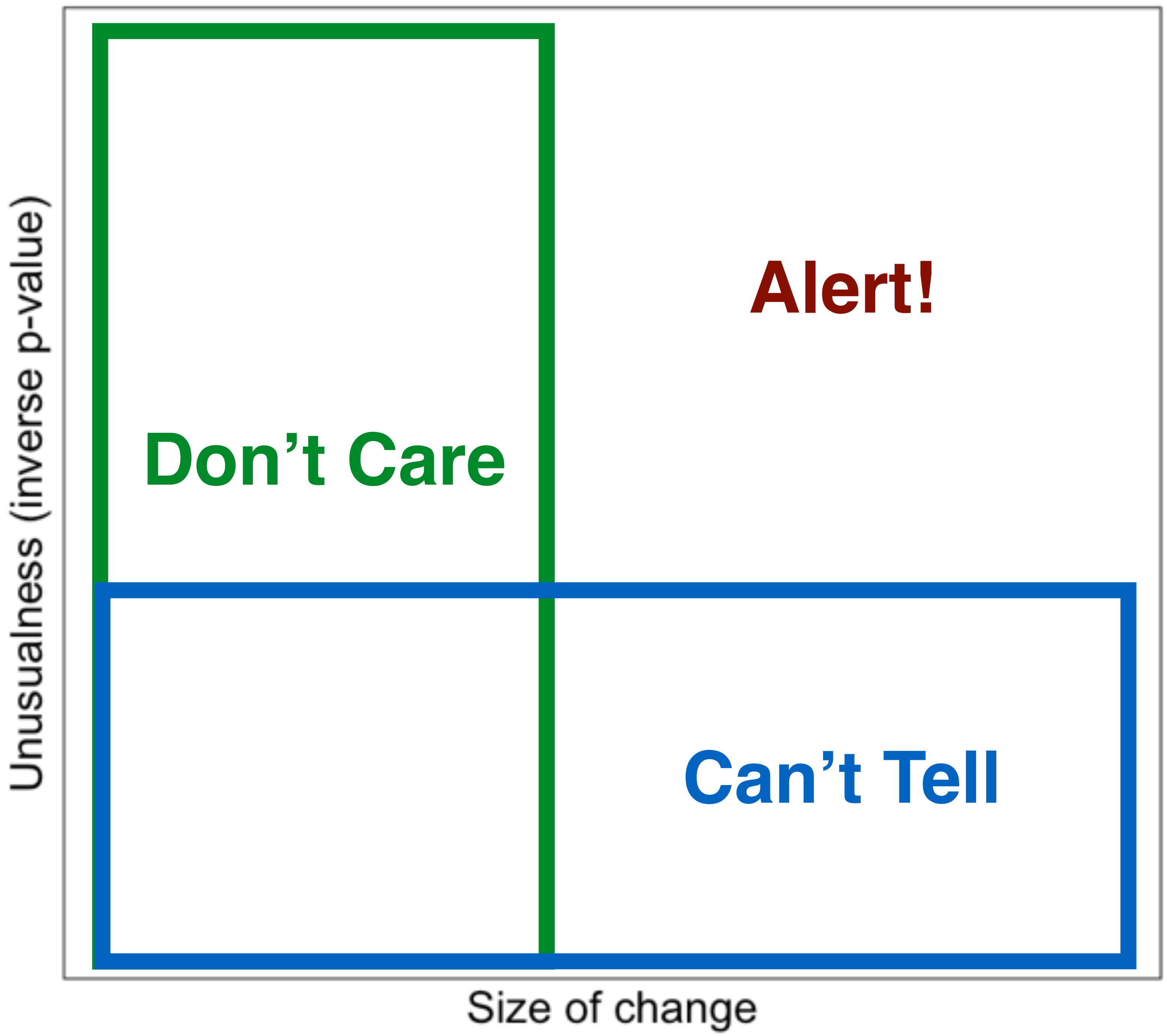
Meta-Approach

- Transform original features into a few scalar metrics which indicate "unusualness"
  1. Manual feature engineering, and/or
  2. Dimension reduction, then
  3. Model whose output is an unusualness metric
- Combine these metrics and tune thresholds using your limited supply of labeled data
  - Need both a "rareness" metric and an "importance" metric.
  - With large data, trivial changes can be statistically significant.

Data Transformations

- Time series data
  - Bucket by time interval and apply statistical tests
  - Difference/smooth/lag and convert to multidimensional
  - Fit trends over varying time windows, use model coefficients as the new feature
- Multidimensional data
  - Convert with ratios: part/whole, effect/cause
  - Reduce dimensions with matrix factorization techniques
- Think about global vs. local structure

Alerting Thresholds



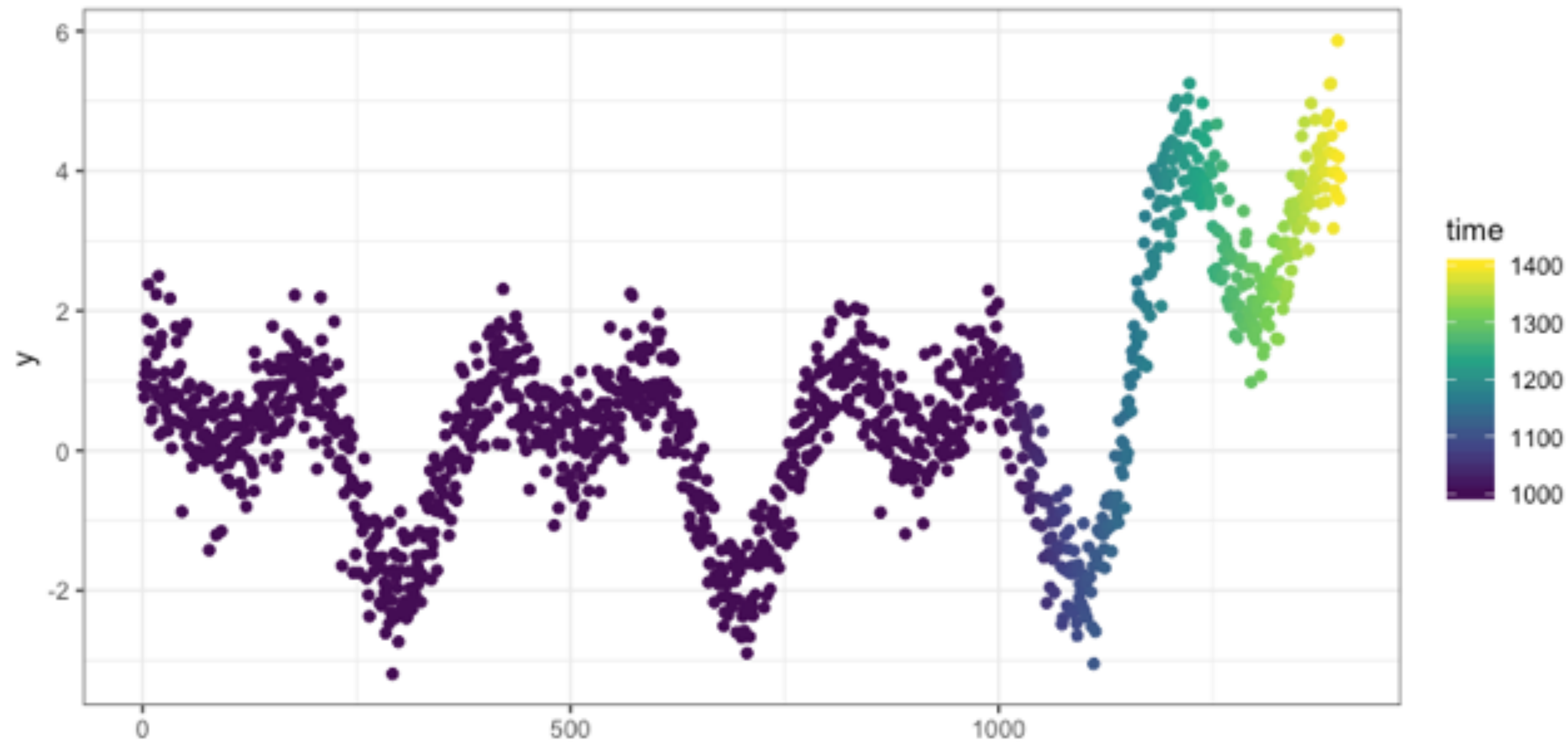
Threshold Fitting

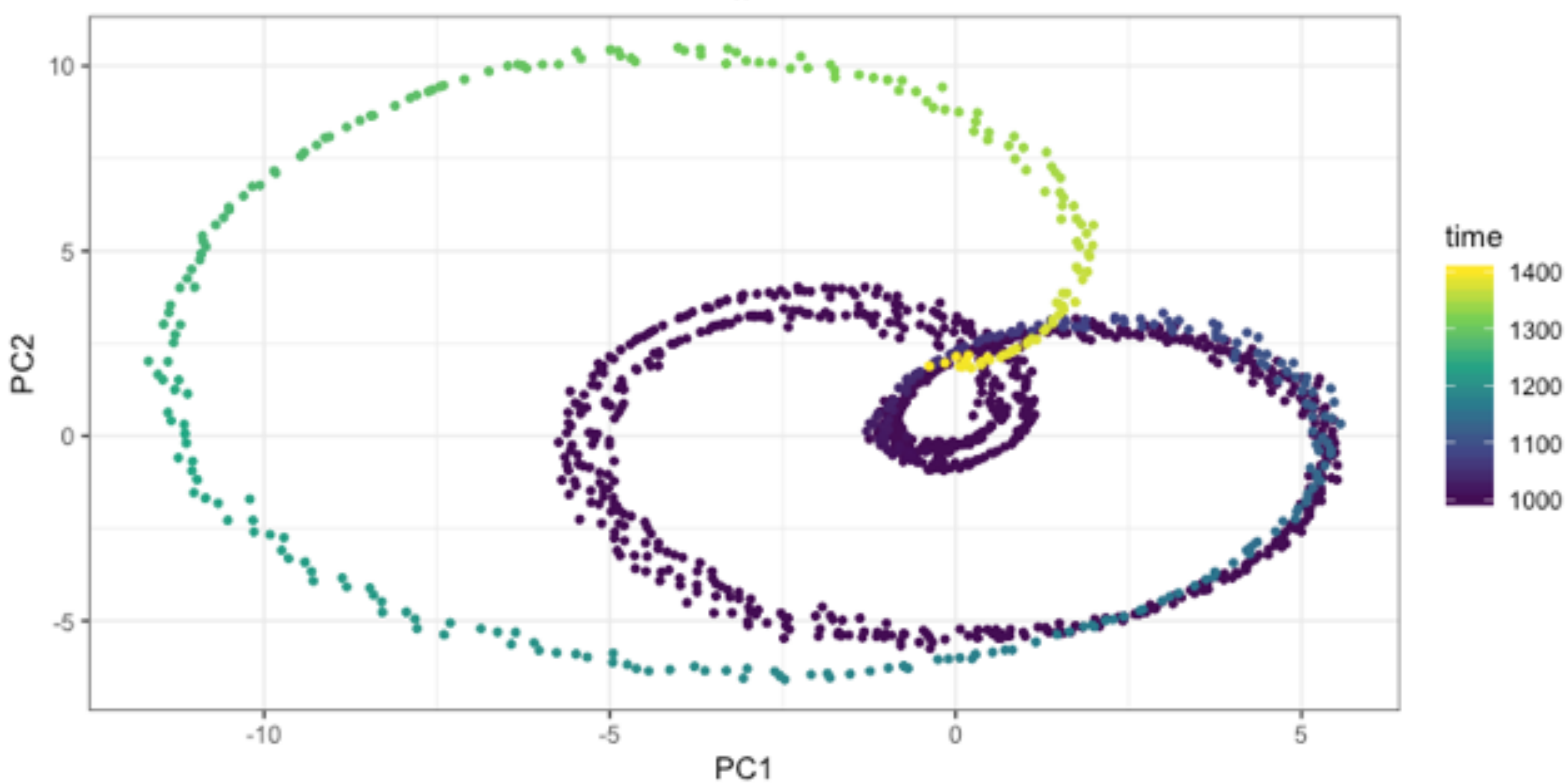
- “I’m tuning these thresholds to try to divide my outliers into two classes, ‘wanted’ and ‘unwanted’…”
- This reduces to fitting a classifier on a small amount of data, something you already know how to do.
- Complexity of the model you can use is driven by your willingness and ability to label data as wanted (“real”) and unwanted (“spurious”) anomalies.
- You can treat the input parameters to your data transformation steps or models as hyperparameters of the final classifier.
    - Don’t overfit your training data!

Principal Component Analysis

- Reduces data to a lower dimensional space while preserving as much of the variance as possible.
- Works well when several features are highly correlated. Computed with SVD.
- Two ways to use:
  - Transform into fewer dimensions and look for anomalies in the lower dimensional space
  - Compare the data reconstructed from lower dimensions with the original, and look for unusually high reconstruction error (outliers often represent poorly in the lower dimensional space)

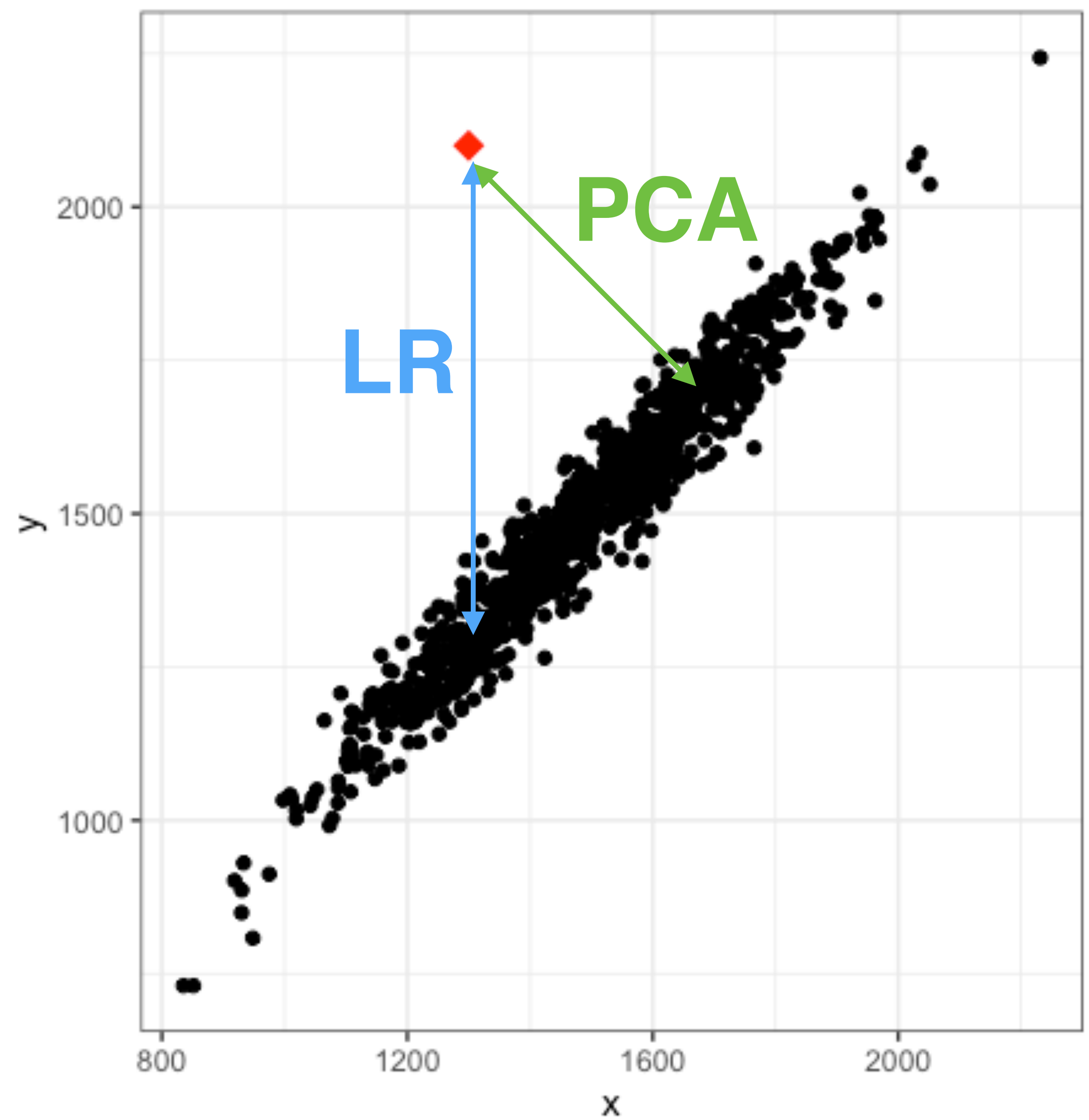
Principal Component Analysis





Principal Component Analysis

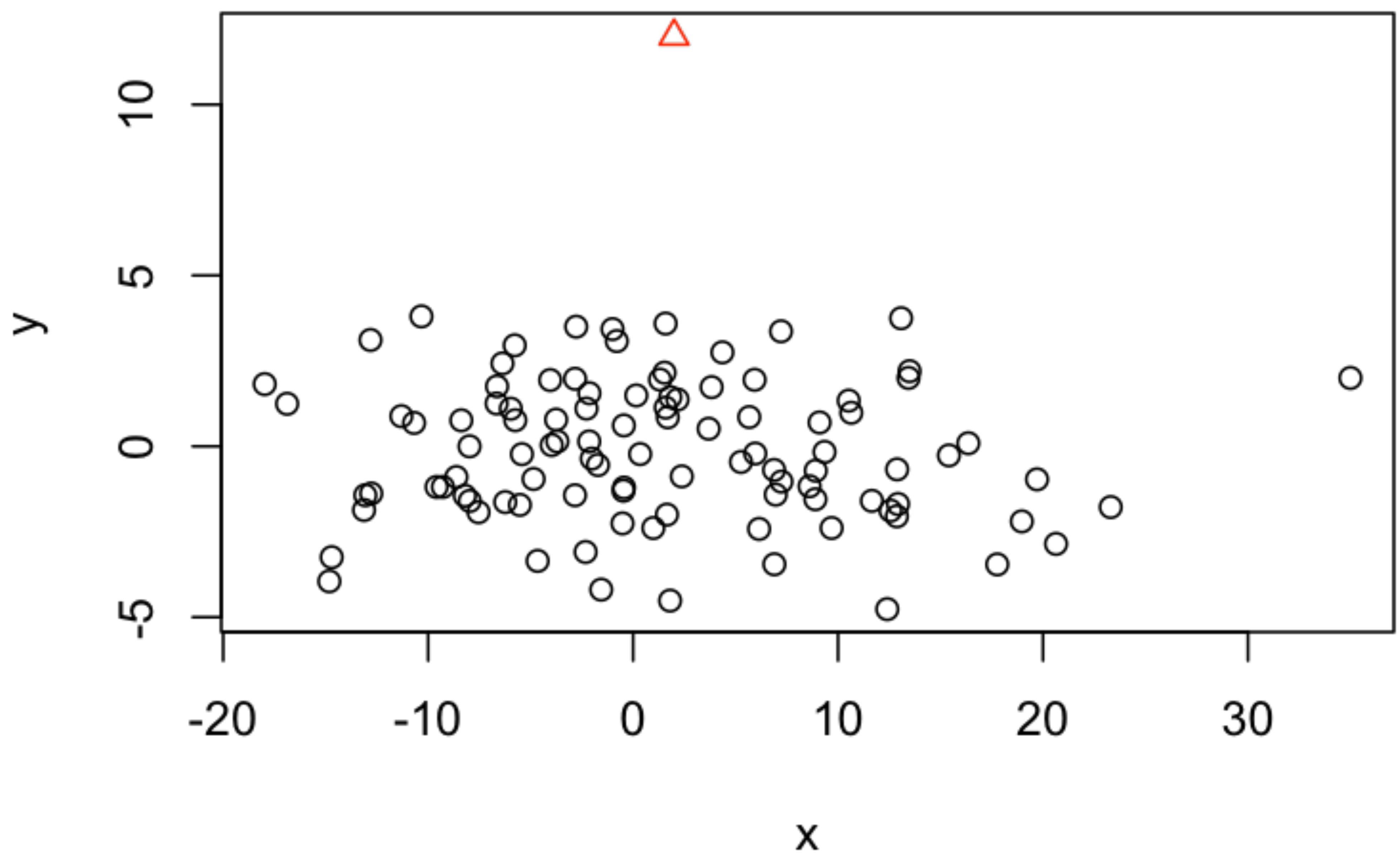
- Few advantages over linear regression in low-dimensional data.
- In high-dimensional data, more correlated features are an asset instead of a liability.



Clustering Models

- Run standard clustering algorithms and look for small clusters and/or points which fit their assigned cluster badly
  - Mahalanobis distance for point closeness to centroid
- Effective on individual outlier points or outlier clusters.
- Less effective on a continuous trajectory away from normal.
- Requires low dimensional data. In high dimensions, most points are about equally distant from each other and clustering doesn't work.

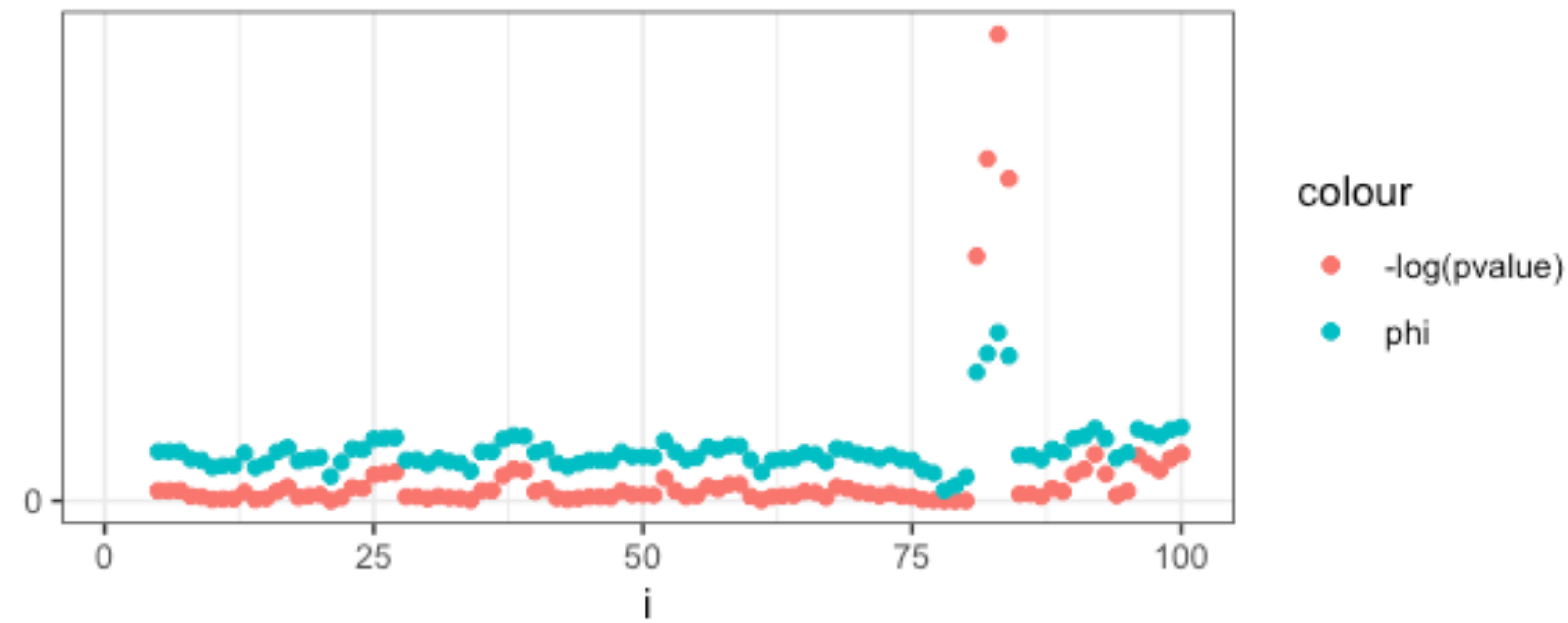
Clustering Models



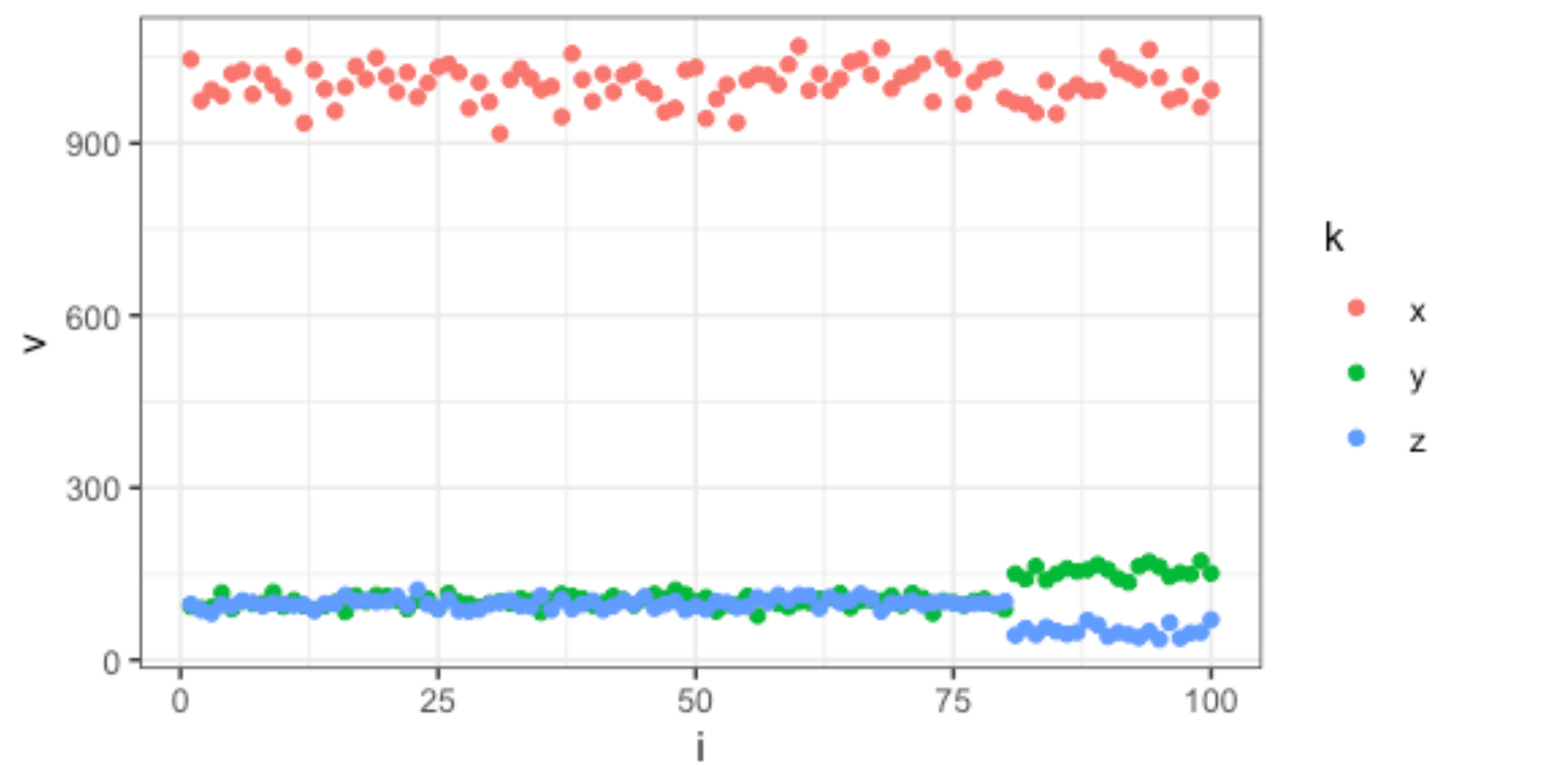
X^2 Test

- Useful for categorical data which can be bucketed: often time series into windows
- Run test over last N rolling time buckets (N is a hyperparameter to tune)
- Look at phi (measure of effect size) and p-value

X^2 Test



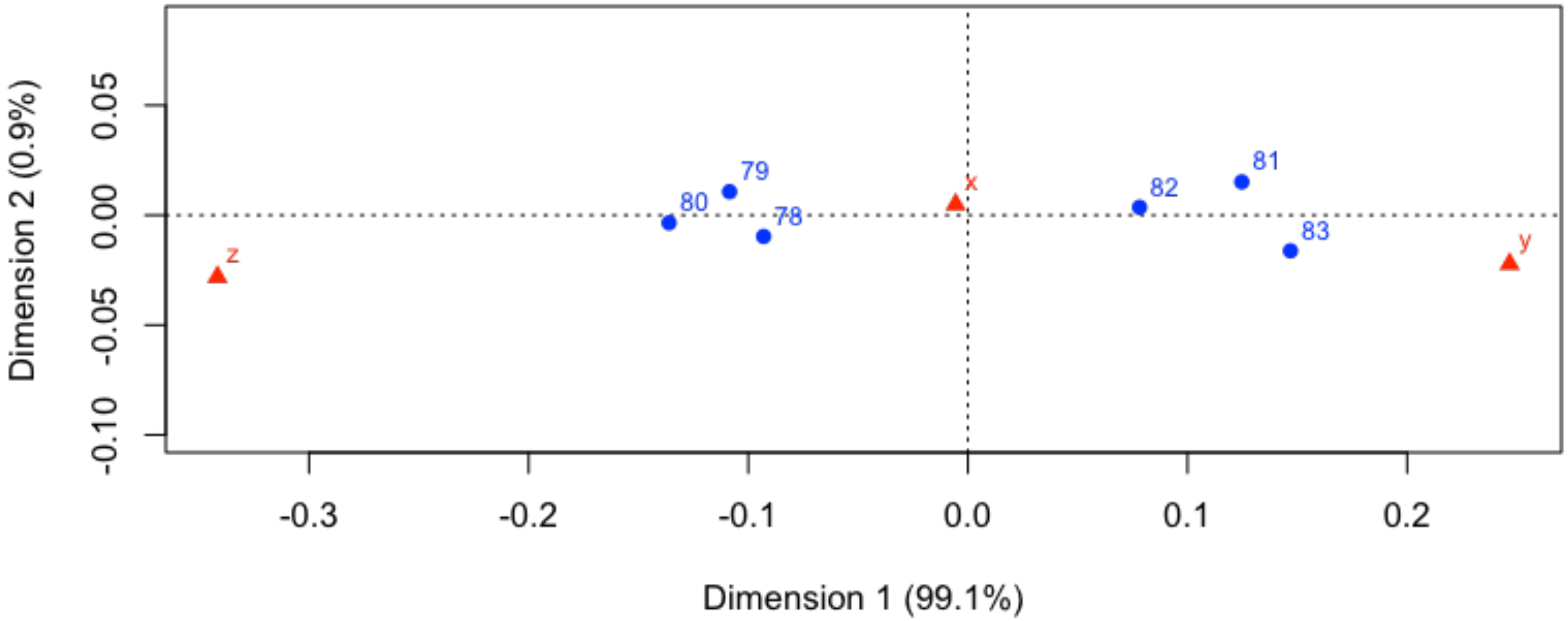
0
0
25
50
75
100
i
colour
-log(pvalue)
phi

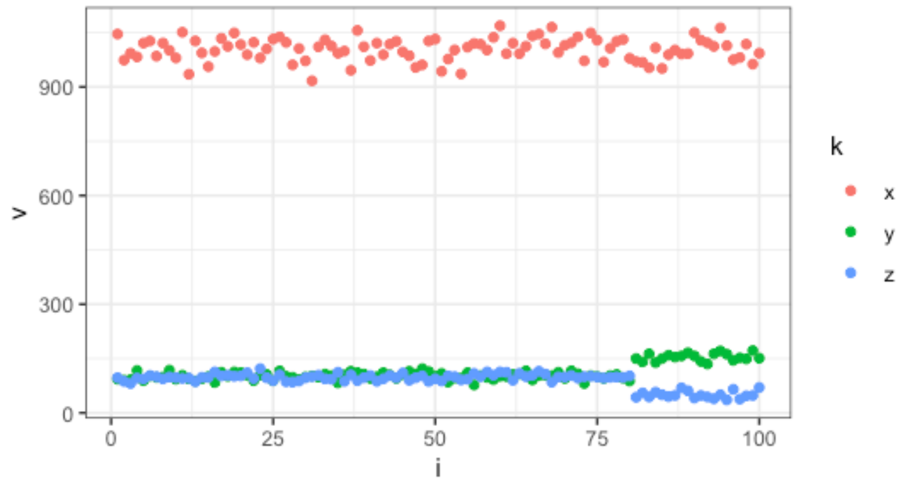


v
0
300
600
900
0
25
50
75
100
i
k
x
y
z

X^2 Test

- Correspondence Analysis is a way to visualize X^2





Specialized Models

- Markov models for discrete sequences
- Density-based or nearest neighbor methods
  - Local Outlier Factor, DBSCAN
- Deep neural network autoencoders
- One-class support vector machines
- Ensemble models
  - Over model type
  - Over hyperparameters

Specialized Transformations

- From time series
  - Wavelets to multidimensional points
  - Discretization of time series data to alphabet
    - using either shape or level within a window
- From graphs (networks)
  - MultiDimensional Scaling (preserves global dist)
  - Spectral methods (preserves local dist)
- From nonlinear multidimensional data
  - Kernel PCA

Questions?

Terran Melconian

terran@terran.us

413 376 5199

http://www.terran.us/talks/

Additional Reading

- Articles
  - Outlier Detection Overview, Sergio Santoyo
- Books
  - Outlier Analysis, Charu Aggarwal
    - Well-written, but fairly challenging algorithm book about all aspects of anomaly detection. Also consider his *Data Mining* book for general algorithms including data transformation techniques.
  - Correspondence Analysis, Michael Greenacre
    - Specifically about the CA technique I demonstrated for chi-squared test visualization.