Anomaly detection

A very brief introduction

Nate Monnig

Kount®
Protecting Your Digital Innovation

Quick Introduction

Kount® Protecting Your Digital Innovation

Nate Monnig

- Senior Data Scientist at Kount Inc. (Boise, ID)
- Research Scientist at Numerica Corp. (Fort Collins, CO)
- PhD in Applied Mathematics from University of Colorado Boulder
- Hydrogeologist at Golder Inc. (Lakewood, CO)
- MS in Hydrogeology from Colorado School of Mines
- BA in Physics from Dartmouth College

Anomaly Detection

What is Anomaly Detection?

- The identification of rare items, events or observations which raise suspicions by differing significantly from the majority of the data.
- **Outlier detection**
  - The training data contains outliers which are defined as observations that are far from the others. Outlier detection estimators thus try to fit the regions where the training data is the most concentrated, ignoring the deviant observations.
  - Unsupervised anomaly detection.
- **Novelty Detection**
  - The training data is not polluted by outliers and we are interested in detecting whether a new observation is an outlier. In this context an outlier is also called a novelty.
  - Semi-supervised anomaly detection.

Kount®

The good and the bad

- **Pros**
    - Potential to detect of anomalous events you hadn't anticipated
    - Threats you've never seen before (e.g. zero-day attacks)
    - Identify data quality / consistency issues (e.g. changes and/or problems with data collection pipelines)
- **Cons**
    - Can be difficult to detect the things you want to find
    - Anomalies != bad things
    - Difficult to tune thresholds (often find way too many anomalies or few to none)
    - Potentially manually intensive process of diagnosing root cause of anomaly
    - Alert fatigue

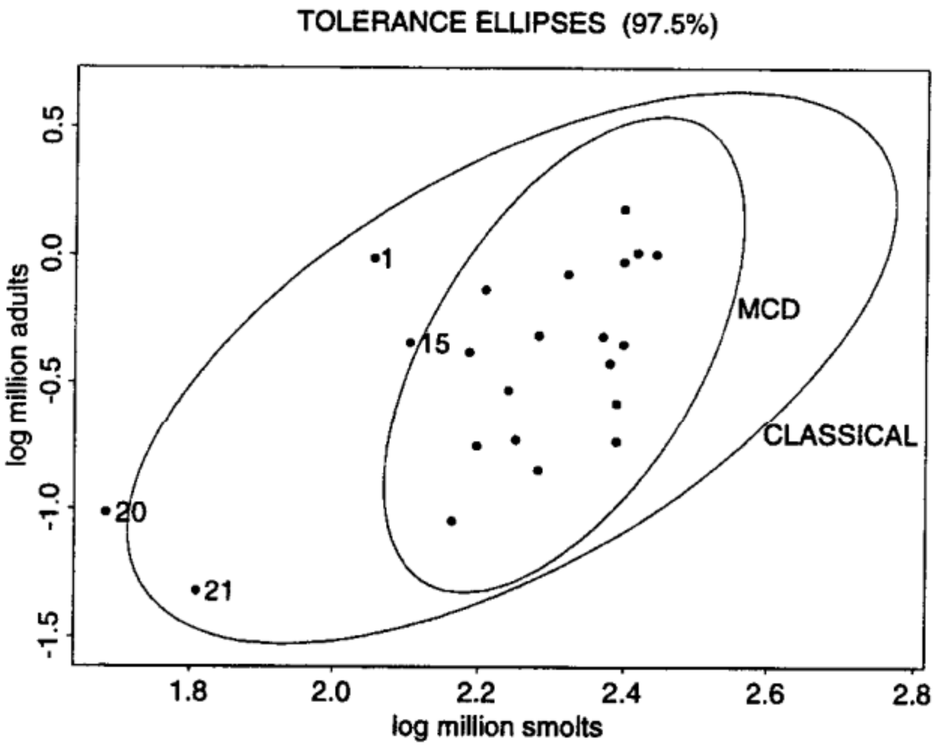
Kount®

Lots of algorithms for outlier detection

- **Robust Covariance (Minimum Covariance Determinant)**
- **Isolation Forest**
- One-class SVM
- Local Outlier Factor
- Robust PCA
- And many more…

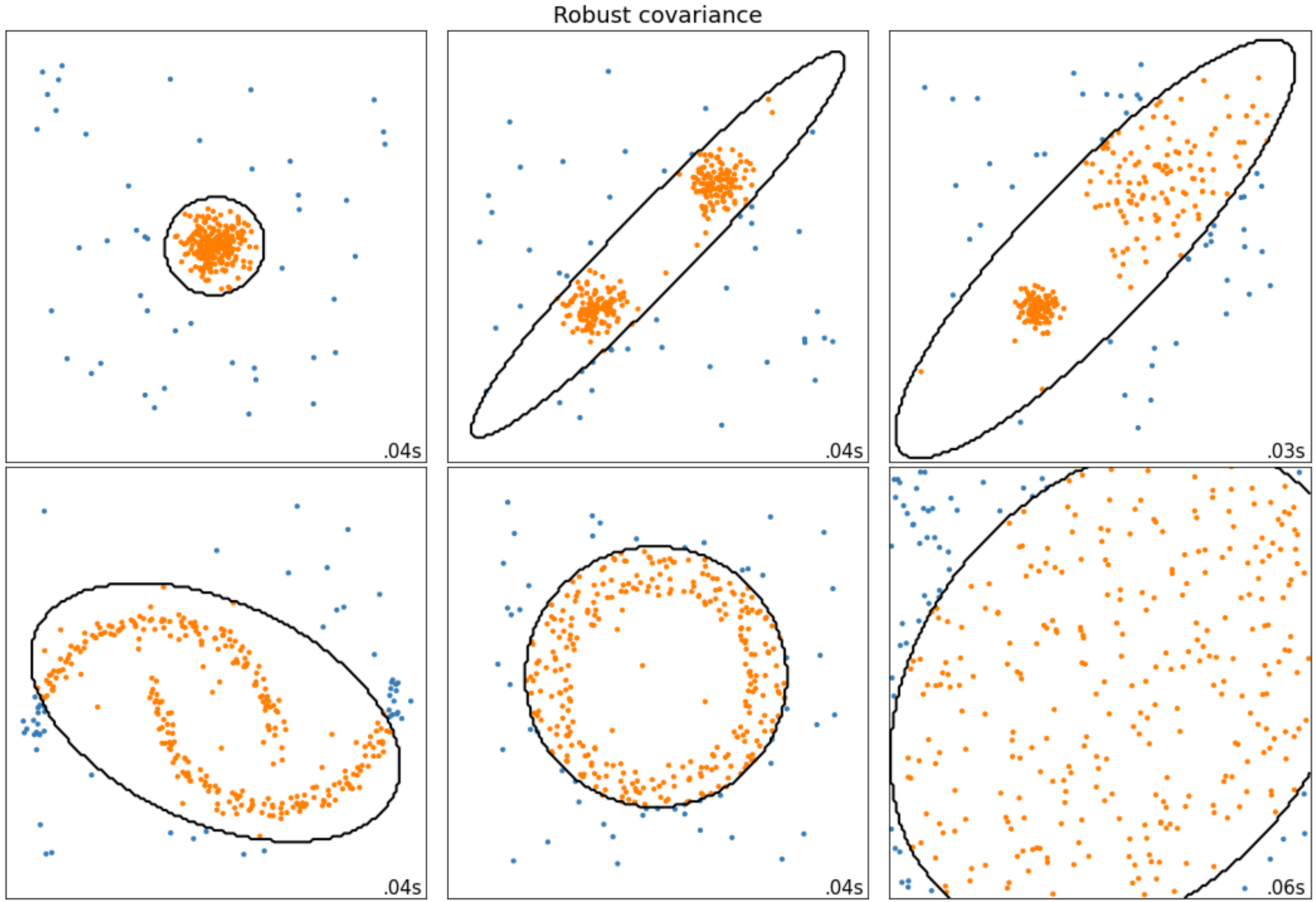
Robust Covariance (Minimum Covariance Determinant)

- Assumes inliers have a (multivariate) Gaussian distribution.
- **Conceptually**
  - Attempts to find the smallest ellipsoid which contains "most" of the data.
- **A bit more precisely**
  - Objective is to find h observations (out of n > h) whose covariance matrix has the smallest determinant.



Rousseeuw, P.J., Van Driessen, K. "A fast algorithm for the minimum covariance determinant estimator" Technometrics 41(3), 212 (1999)

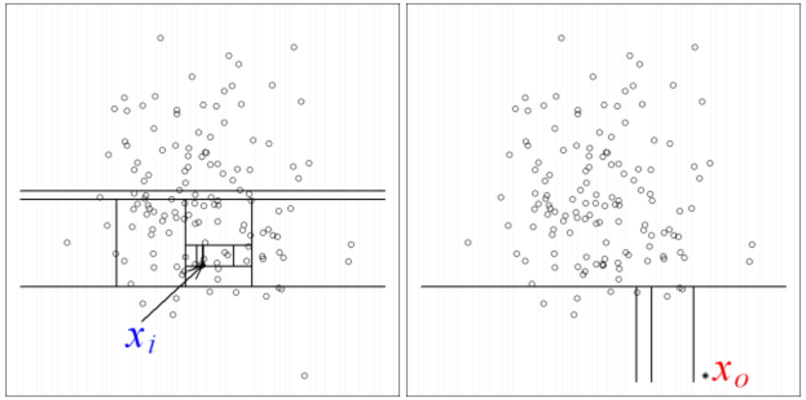
Robust Covariance (Minimum Covariance Determinant)



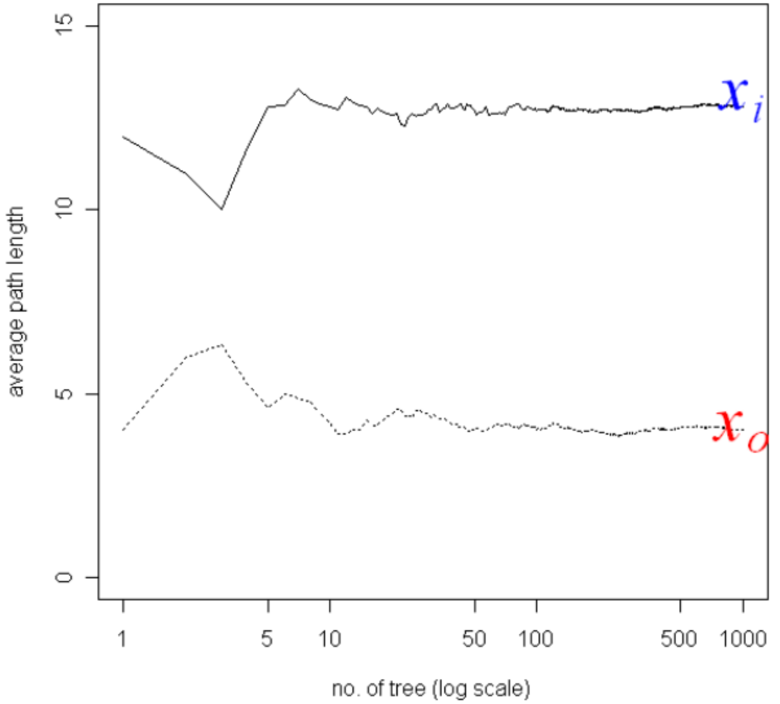
Isolation Forest

- Non-parametric method (no assumptions about distribution of inliers)
- Randomized method for detecting outliers
- Ensemble tree-based approach
    - Random selection of features
    - Random cutoffs on each feature between min and max values
    - Length of path to isolate a data point is an indicator of anomalous-ness
        - Short paths to isolation imply a likely outlier
        - Long paths to isolation imply a likely inlier

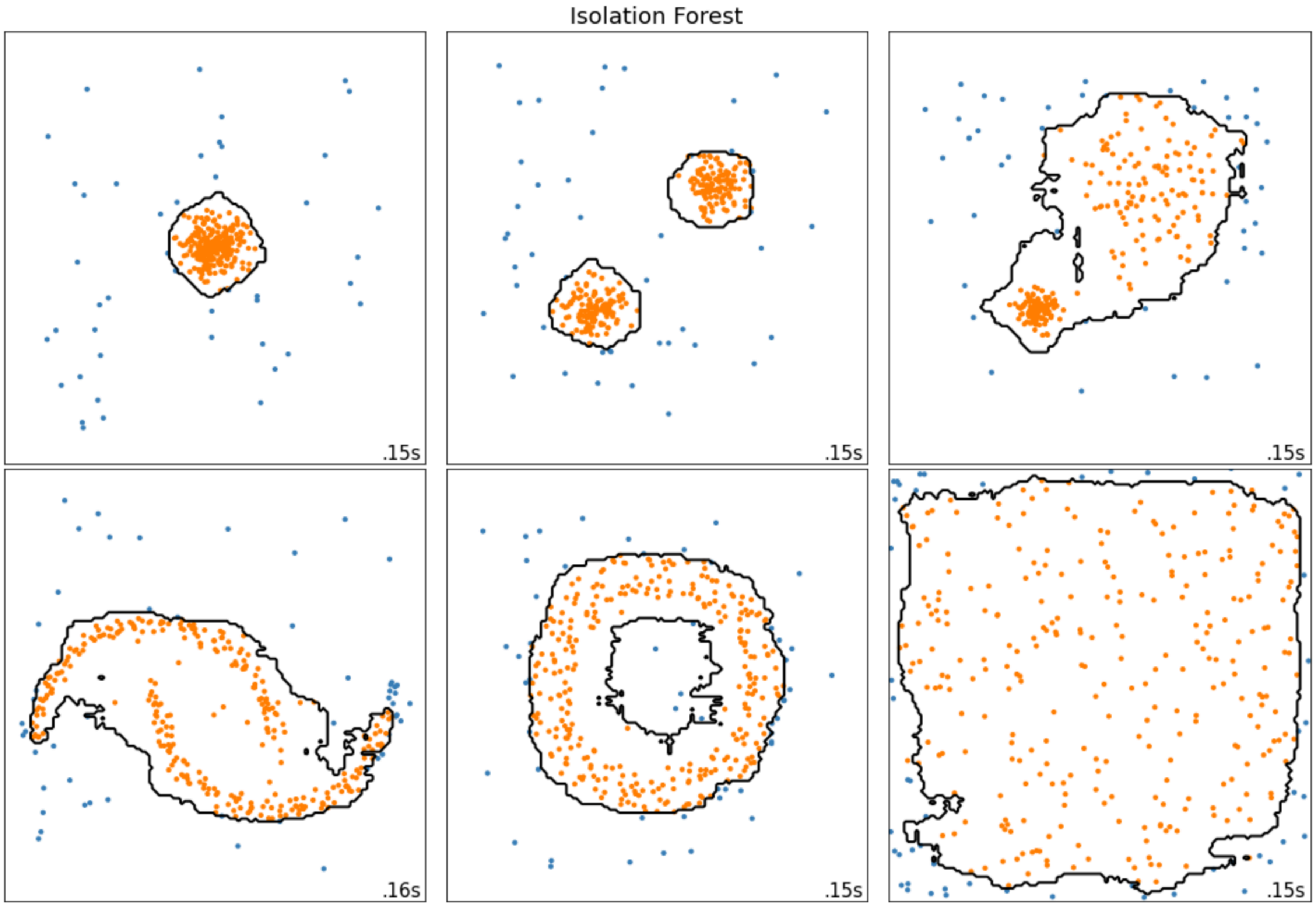
Liu, Fei Tony, Ting, Kai Ming and Zhou, Zhi-Hua. "Isolation forest." Data Mining, 2008. ICDM'08. Eighth IEEE International Conference on.



(a) Isolating x_i (b) Isolating x_o



Isolation Forest



To the codes!


YOU GET A CODE, YOU GET A CODE
2004
YOU ALL GET CODES
imgflip.com

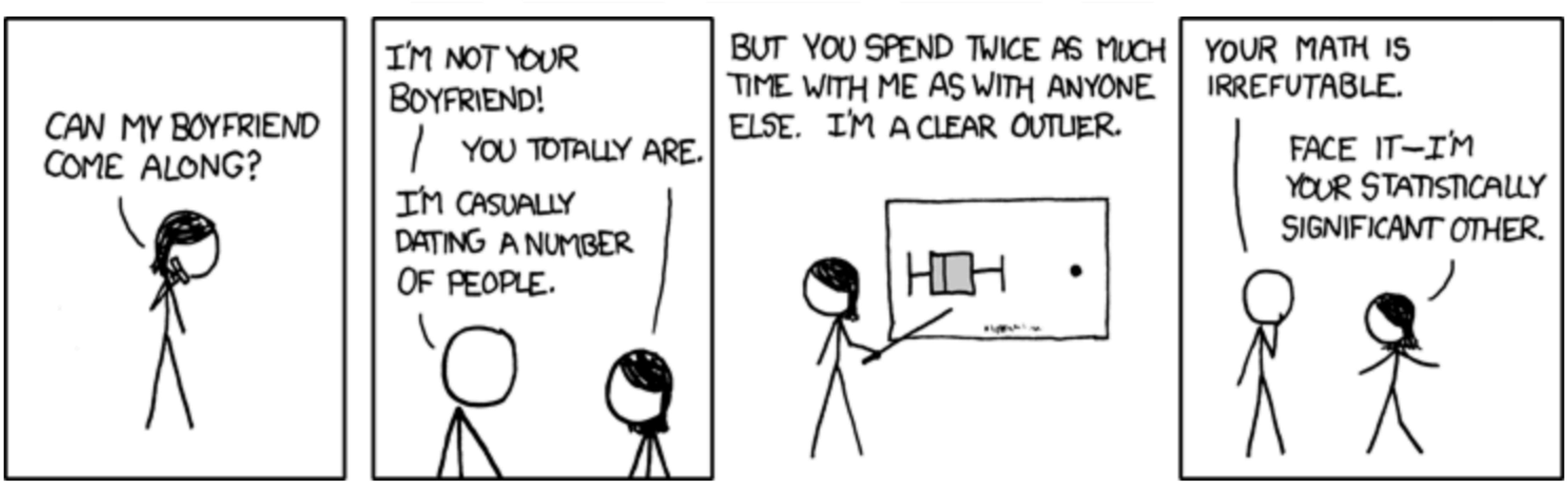
Next Steps

- Tune algorithm parameters
- Add (or synthesize) more features
  - Other device data
  - Shopping cart data
  - Customer history
- Get more data (more events)
  - May require a distributed compute platform like Spark
- Experiment with other anomaly detection algorithms
- Investigate correlation with variables or outcomes of interest (e.g. fraud)

The good and the bad

- **Pros**
  - Potential to detect of anomalous events you hadn't anticipated
  - Threats you've never seen before (e.g. zero-day attacks)
  - Identify data quality / consistency issues (e.g. changes and/or problems with data collection pipelines)
- **Cons**
  - Can be difficult to detect the things you want to find
  - Anomalies != bad things
  - Difficult to tune thresholds (often find way too many anomalies or few to none)
  - Potentially manually intensive process of diagnosing root cause of anomaly
  - Alert fatigue
- **Lessons**
  - Get super familiar with the data
  - Think carefully about feature set to ensure anomalies are more likely to be interesting
  - Align anomaly detection algorithm with the distribution of "normal data"
  - Tune thresholds carefully

Questions?


CAN MY BOYFRIEND COME ALONG?
I'M NOT YOUR BOYFRIEND!
YOU TOTALLY ARE.
I'M CASUALLY DATING A NUMBER OF PEOPLE.
BUT YOU SPEND TWICE AS MUCH TIME WITH ME AS WITH ANYONE ELSE. I'M A CLEAR OUTLIER.
YOUR MATH IS IRREFUTABLE.
FACE IT—I'M YOUR STATISTICALLY SIGNIFICANT OTHER.