Quantitative Cleaning

- Worldviews
  - Model-Free/Non-parametric
    - AKA "descriptive" statistics
    - e.g. histograms, SQL aggregates, distinct values, etc
      - Most of what Bellman proposes
    - Natural fit to set-theoretic/logic viewpoint
  - Model-Based/Parametric
    - data doesn't just appear: it samples or gives evidence of some other phenomenon
    - Ideally describes all attributes as random variables in a probability density function representing the joint distribution of the variables
- Outlier: something too far from typical
  - What's typical?
    - center (core)
  - What's too far?
    - spread (dispersion)
  - Intuition: the bell curve
  - Model-driven approach
    - Fit data to an appropriate probability model (parameterize)
    - Measure "residuals" -- distance from model
    - Look for outliers among residuals
      - Central Limit Theorem says that random ("iid") perturbations will give normally distributed errors!
  - Model-free approach?
    - order statistics
    - reduce quantitative values to ordinal values
- Robust Statistics
  - One important metric: the Breakdown Point of an estimator
    - What percentage "spam" before an adversary can choose the value?
    - Ideal: 50%
    - Worst case: 1/|S|
  - One important class -- L-estimators -- are easy to understand and implement
    - based on order statistics
  - Robust Core metrics
    - Median
    - Trimmed mean
    - Winsorized mean
    - Breakdown points?
    - How to choose?
  - Robust Spread
    - MAD
    - Subtopic
  - Hampel X84
    - In case you want a textbook method!
    - Outliers are 1.4826x MADS from the median
      - x is the # of standard deviations you would have chosen had you not been robust
  - DB Implementation
    - Median in SQL is a great quiz question
    - median,trimming,winsorizing all based on efficient order stats
    - Realistically, approximation algorithms via UDAs are the right way to go
      - especially to support simultaneous order stats on different columns
    - Open problem: one-pass, small-space MAD
- Multivariate Statistics
  - The general case!
  - Correlation and multivariate outliers
  - center: multivariate mean
  - dispersion: covariance
    - 1-d variance on the diagonal
    - correlations captured elsewhere
    - Mahalanobis depth
      - where stddevs measures sqrt(distance from ctr squared), this adjusts for the covariance matrix
  - Robust estimators
    - Componentwise
      - center: componentwise median
      - rescale data into componentwise dispersion metric (e.g. MADs)
      - threshhold distance from center
    - True Robust estimators
      - They do exist
      - They're a bit complex, and typically iterative computations
      - Open problem: scale these to big datasets!
    - Depth-based techniques
      - analogy to trimming via iterative deletion of max-mahalanobis points
        - breakdown point is 1/d
      - Analogy to order stats
      - Peeling depth
      - Tukey Half-plane depth
      - Simplicial depth
        - "# of triangles"
  - Distance-based outlier detection
    - Non-parametric approaches not based on order stats
    - E.g. how many points within distance D
    - Have a hard time with clusters of differet densities
    - NOT scale-invariant
- Timeseries outliers
  - Robust timeseries is a relatively young field
  - Additive outliers ("hiccups")
  - Innovation Outliers
  - Often handled by windowing or EWMA
  - Traditionally assume variance remains constant
    - new methods to adjust for changes
- Resampling: a cool thing to know
  - Bootstrap estimator
    - n items: n samples with replacement, followed by estimate
    - repeat thousands of times
    - Very handy when the input sample is small
  - Jacknife
    - leave one random item out
    - repeat dozens of times
  - Given the sampling distribution...
    - specific techniques to compute mean/ variance of estimator