QUANTITATIVE DATA CLEANING FOR LARGE DATABASES

JOSEPH M. HELLERSTEIN

Berkeley
UNIVERSITY OF CALIFORNIA

BACKGROUND

- a funny kind of keynote
  - a trip to the library
    - robust statistics, DB analytics
  - some open problems/directions
    - scaling robust stats, intelligent data entry forms

J. M. Hellerstein, “Quantitative Data Cleaning for Large Databases”, http://db.cs.berkeley.edu/jmh/papers/cleaning-unece.pdf

TODAY

- background
- outliers and robust statistics
- multivariate settings
- research directions

QDB ANGLES OF ATTACK

- data entry
  - data modeling, form design, interfaces
- organizational management
  - TDQM
- data auditing and cleaning
  - the bulk of our papers?
- exploratory data analysis
- the more integration, the better!

CULTURAL VALUES: WHAT IS A VALUE?

DB View: data	Stat View: evidence
descriptive statistics	*inductive (inferential)* statistics
model-free (nonparametric)	model the process producing the data (*parametric*)
+ works with any data + no model fitting magic	+ probabilistic interpretation - *likelihoods* on values - *imputation* of missing data - *forecasting* future data

TODAY

- background
- outliers and robust statistics
- multivariate settings
- research directions

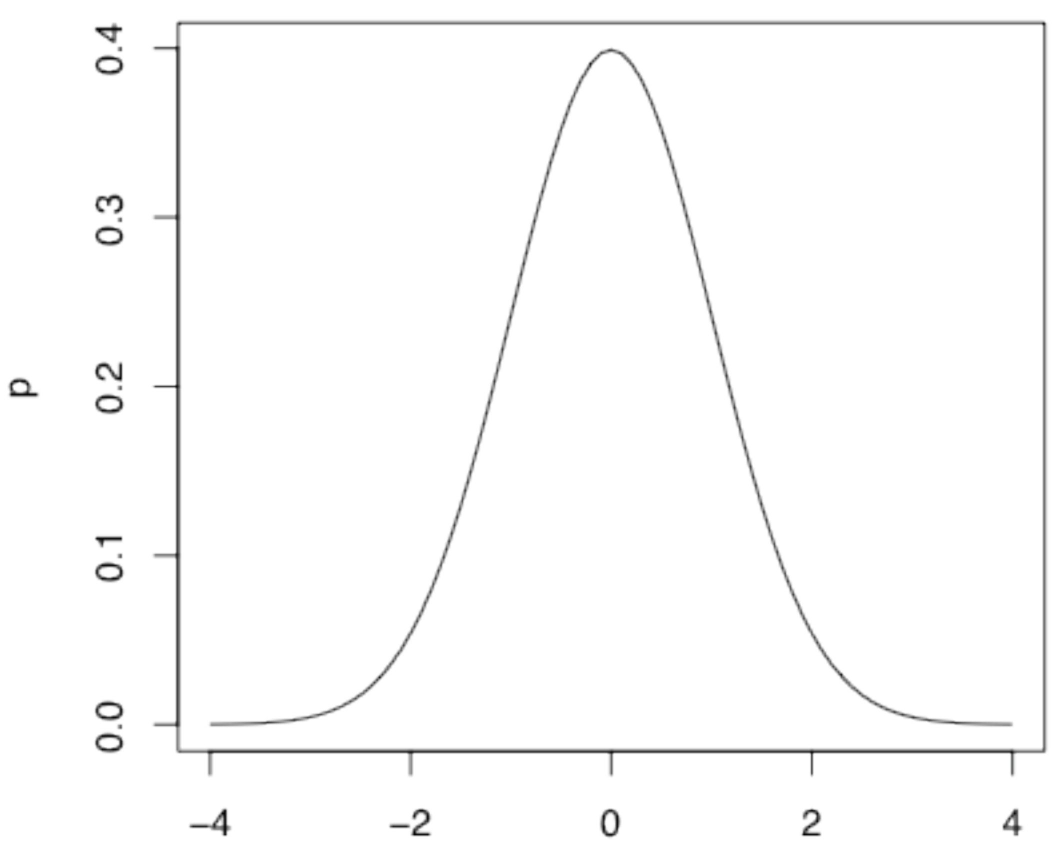
DAD, WHAT'S AN OUTLIER?

FAR FROM THE CENTER

- center
- dispersion

FAR FROM THE CENTER

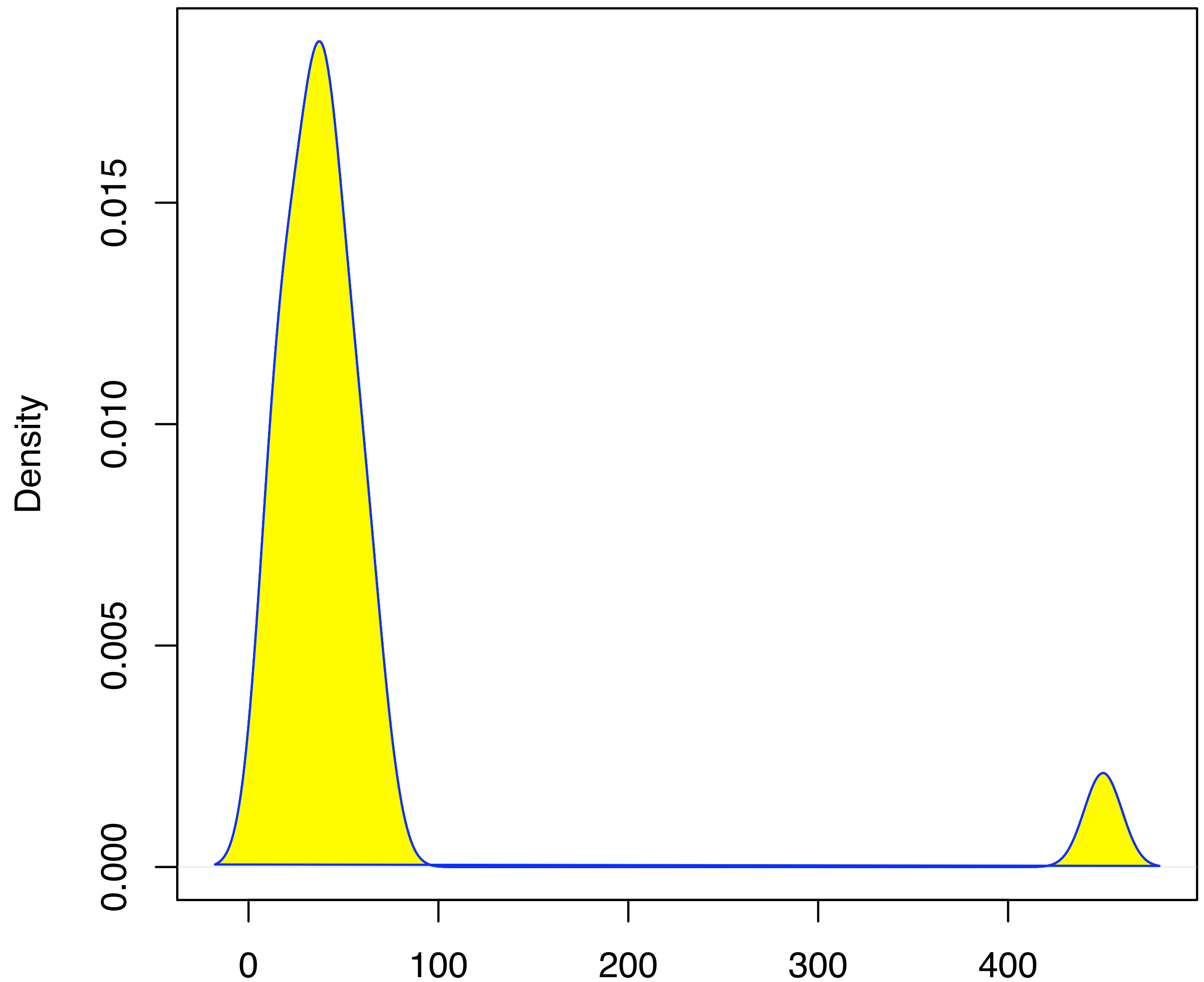
- center
- dispersion
- Normal distribution!
  - a.k.a Gaussian, bell curve
  - mean, variance



CENTER/DISPERSION (TRADITIONAL)

12	13	14	21	22	26	33	35	36	37	39	42	45	47	54	57	61	68	450

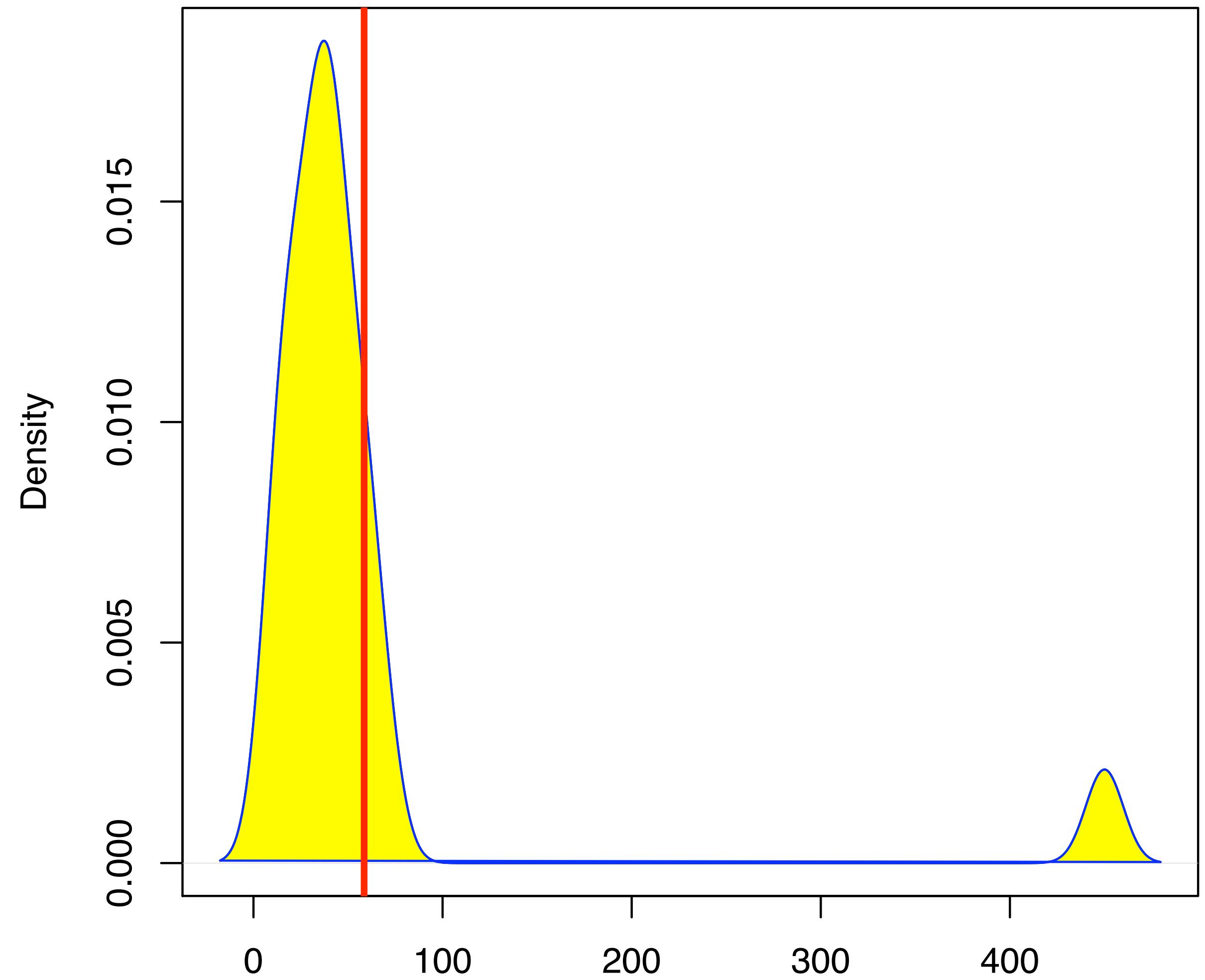
ages of employees (US)



CENTER/DISPERSION (TRADITIONAL)

12	13	14	21	22	26	33	35	36	37	39	42	45	47	54	57	61	68	450

ages of employees (US)

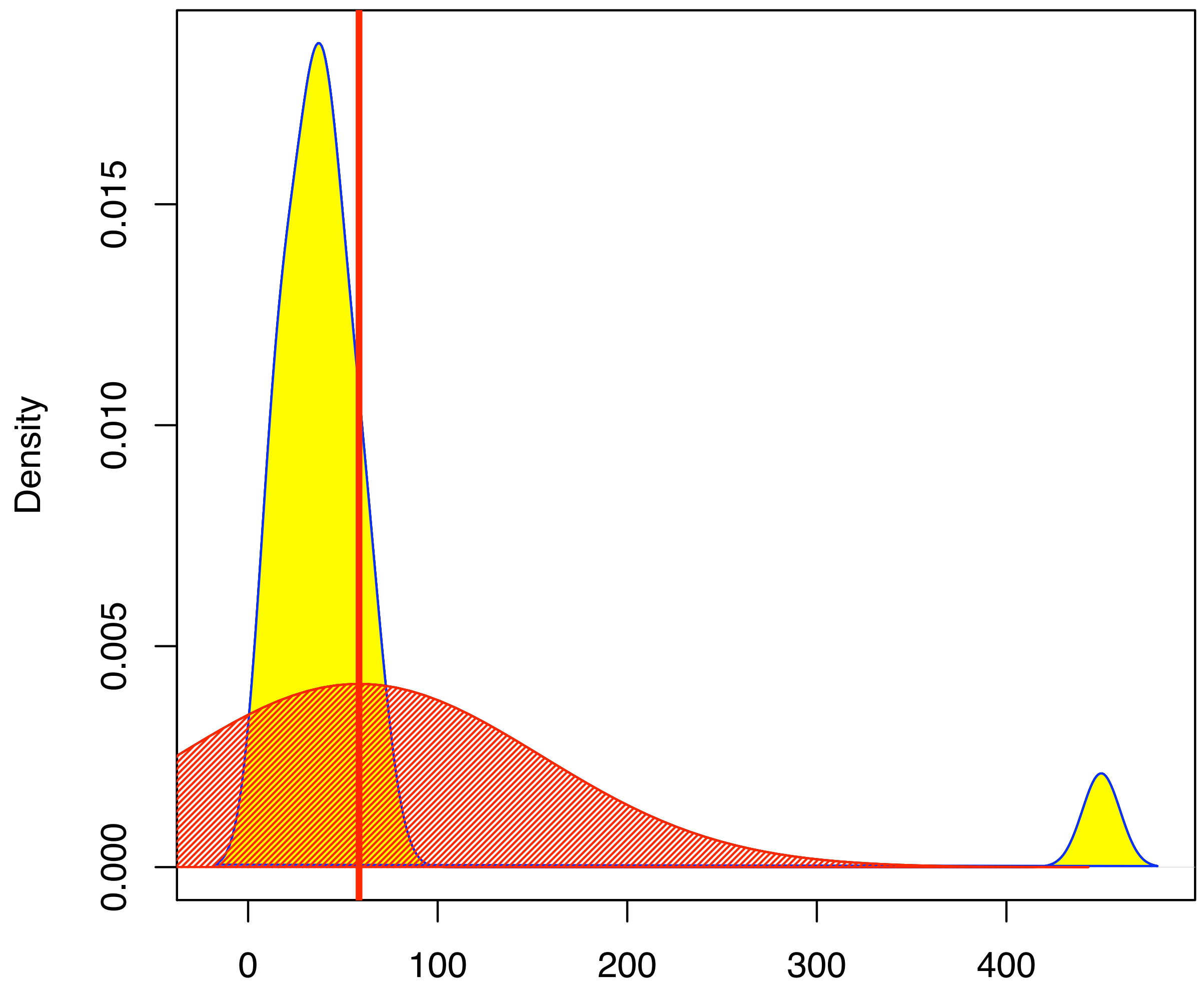


- mean 58.52632

CENTER/DISPERSION (TRADITIONAL)

12	13	14	21	22	26	33	35	36	37	39	42	45	47	54	57	61	68	450

ages of employees (US)

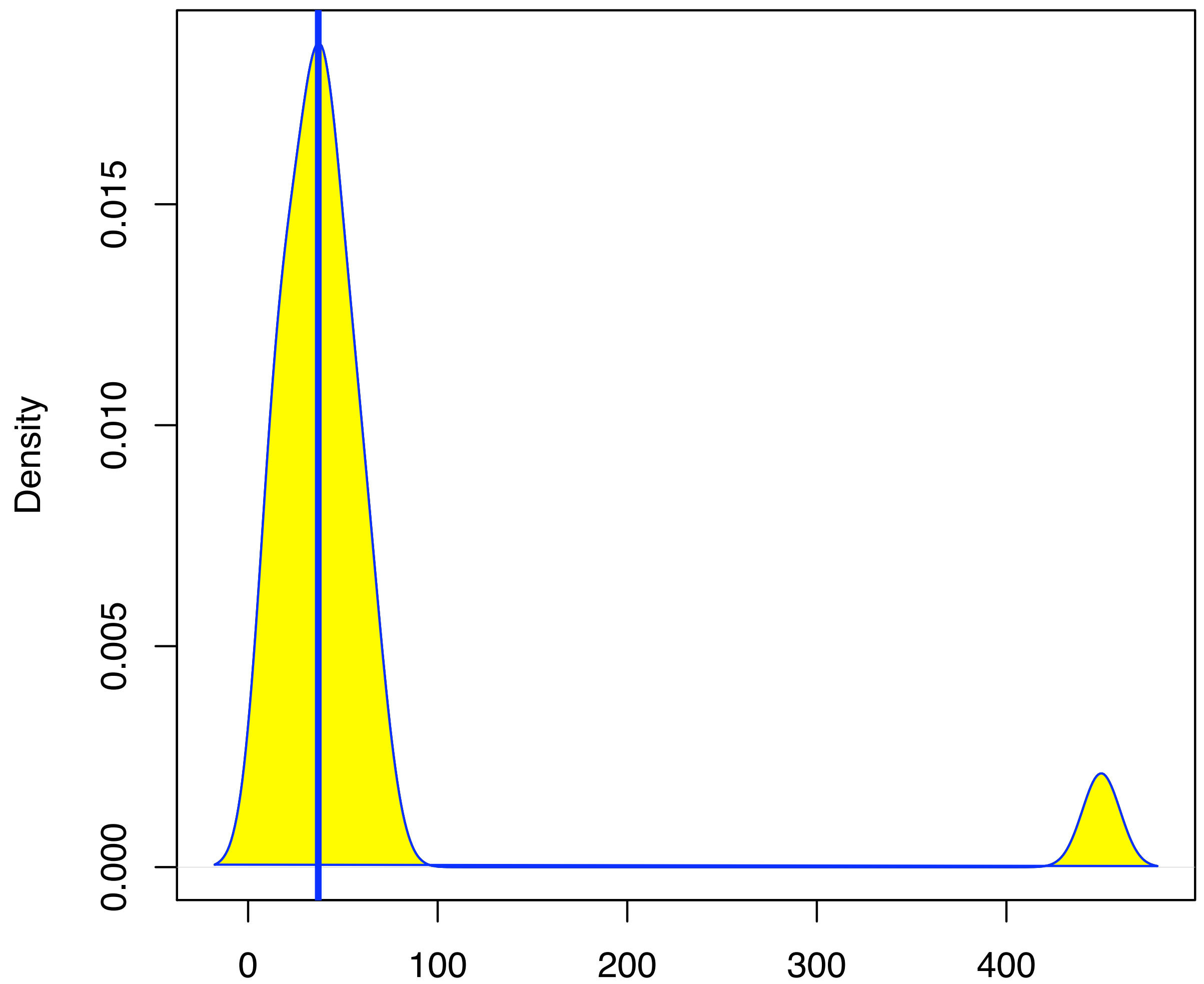


- mean 58.52632
- variance 9252.041

CENTER/DISPERSION (TRADITIONAL)

12	13	14	21	22	26	33	35	36	37	39	42	45	47	54	57	61	68	450

ages of employees (US)

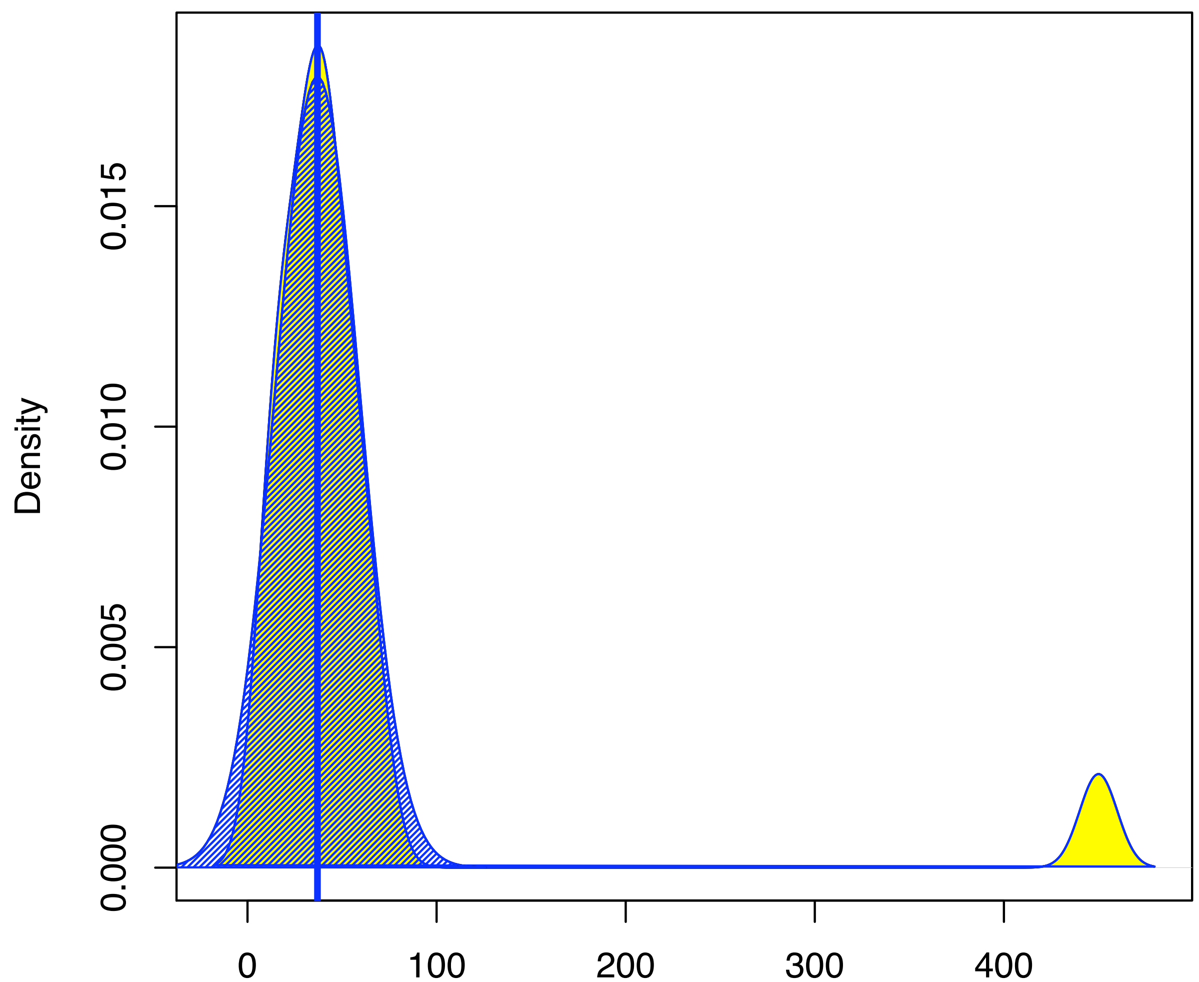


- median 37

CENTER/DISPERSION (ROBUST)

12	13	14	21	22	26	33	35	36	37	39	42	45	47	54	57	61	68	450

ages of employees (US)



- median 37
- MAD 22.239

SUBTLER PROBLEMS

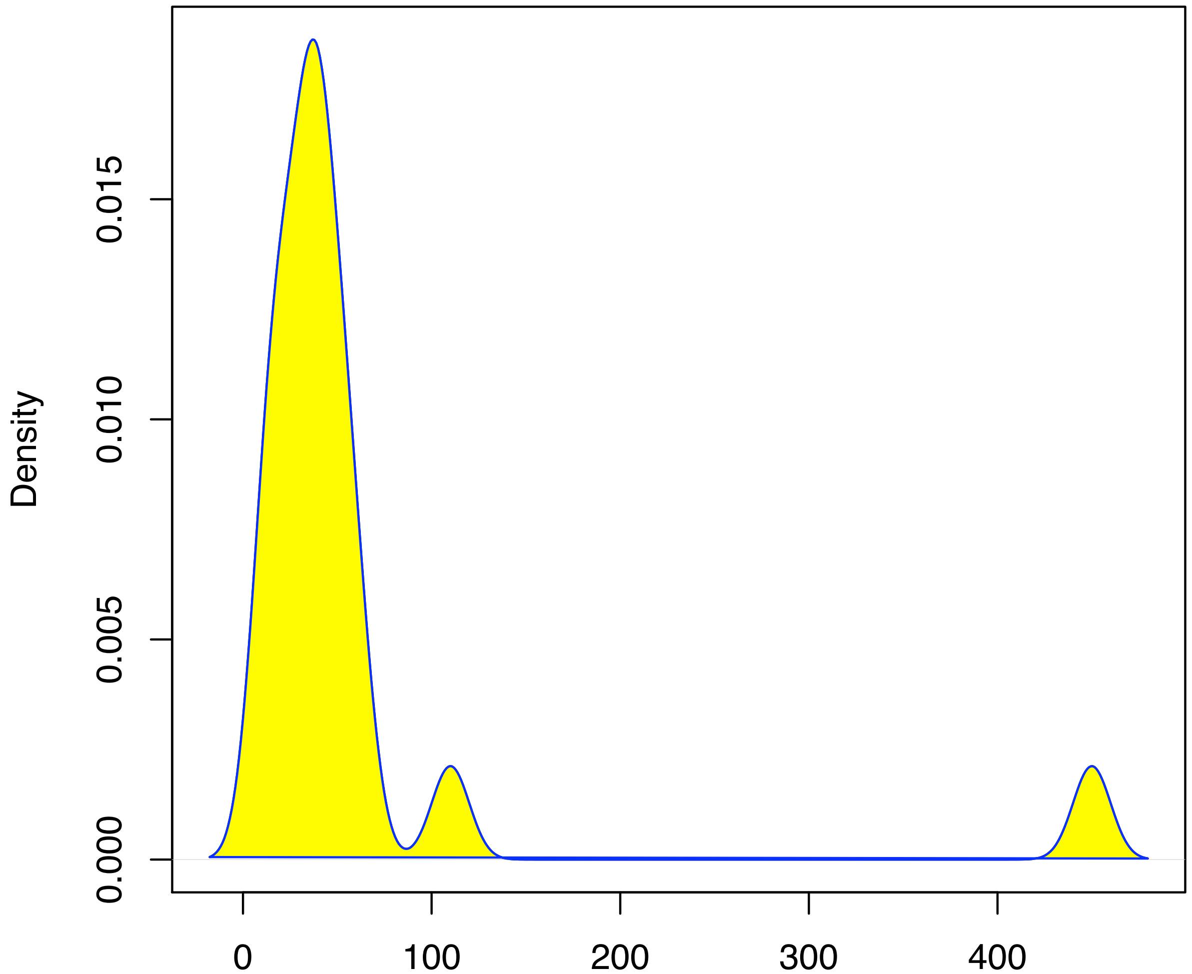
12	13	14	21	22	26	33	35	36	37	39	42	45	47	54	57	61	68	450

SUBTLER PROBLEMS

12	13	14	21	22	26	33	35	36	37	39	42	45	47	54	57	61	110	450

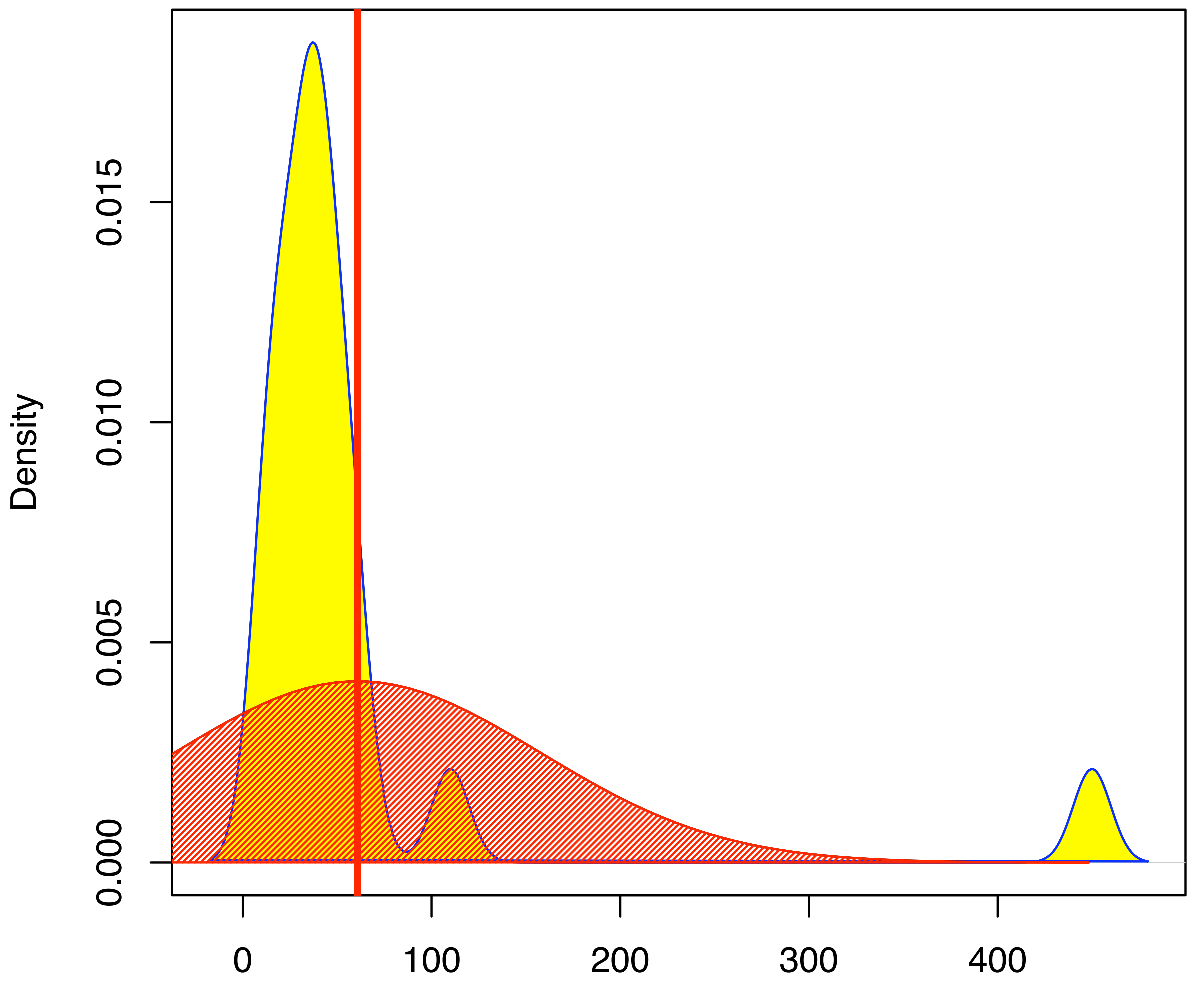
SUBTLER PROBLEMS

12	13	14	21	22	26	33	35	36	37	39	42	45	47	54	57	61	110	450



SUBTLER PROBLEMS

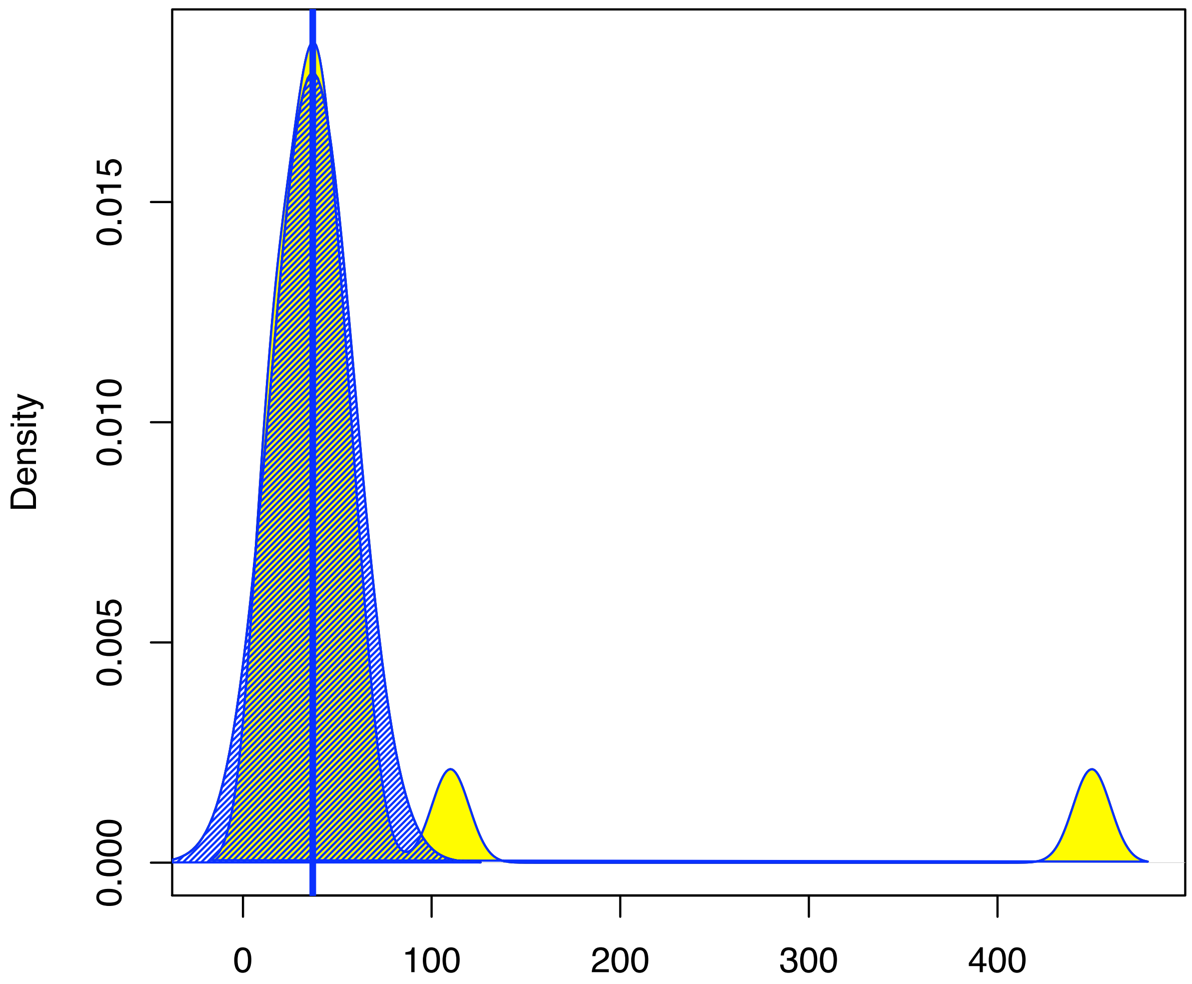
12	13	14	21	22	26	33	35	36	37	39	42	45	47	54	57	61	110	450



- *Masking*
  - magnitude of one outlier masks smaller outliers
  - makes manual removal of outliers tricky

SUBTLER PROBLEMS

12	13	14	21	22	26	33	35	36	37	39	42	45	47	54	57	61	110	450



- Robust stats:
  - handle multiple outliers
  - robust w.r.t. magnitude of outliers

ROBUSTNESS: INTUITION

- handle multiple outliers
- robust to magnitude of an outlier

HOW ROBUST IS ROBUST?

- *Breakdown Point*
  measures robustness of an estimator
  - *proportion of "dirty" data the estimator can handle before giving an arbitrarily erroneous result*
  - think adversarially
- best possible breakdown point: 50%
  - beyond 50% "noise", what's the "signal"?

SOME BREAKDOWN POINTS

- mean?
- mode?
- standard deviation?

SOME ROBUST CENTERS

12	13	14	21	22	26	33	35	36	37	39	42	45	47	54	57	61	110	450

- *median*
  - value that evenly splits set/distribution into higher and lower halves
- *k% trimmed* mean
  - remove lowest/highest *k%* values
  - compute mean on remainder
- *k% winsorized* mean
  - remove lowest/highest *k%* values
  - replace low removed with lowest remaining value
  - replace high removed with highest remaining value
  - compute mean on resulting set

SOME ROBUST CENTERS

12	13	14	21	22	26	33	35	36	37	39	42	45	47	54	57	61	110	450

- *median* (37)
  - value that evenly splits set/distribution into higher and lower halves
- *k% trimmed* mean
  - remove lowest/highest *k%* values
  - compute mean on remainder
- *k% winsorized* mean
  - remove lowest/highest *k%* values
  - replace low removed with lowest remaining value
  - replace high removed with highest remaining value
  - compute mean on resulting set

SOME ROBUST CENTERS

12	13	14	21	22	26	33	35	36	37	39	42	45	47	54	57	61	110	450

- *median* (37)
  - value that evenly splits set/distribution into higher and lower halves
- *k% trimmed* mean (37.933)
  - remove lowest/highest *k%* values
  - compute mean on remainder
- *k% winsorized* mean
  - remove lowest/highest *k%* values
  - replace low removed with lowest remaining value
  - replace high removed with highest remaining value
  - compute mean on resulting set

SOME ROBUST CENTERS

14	14	14	21	22	26	33	35	36	37	39	42	45	47	54	57	61	61	61

- *median* (37)
  - value that evenly splits set/distribution into higher and lower halves
- *k% trimmed* mean (37.933)
  - remove lowest/highest *k%* values
  - compute mean on remainder
- *k% winsorized* mean (37.842)
  - remove lowest/highest *k%* values
  - replace low removed with lowest remaining value
  - replace high removed with highest remaining value
  - compute mean on resulting set

ROBUST CENTER BREAKDOWN POINTS

- median?
- k% trimmed/winsorized mean?
- k ~= 50% ?

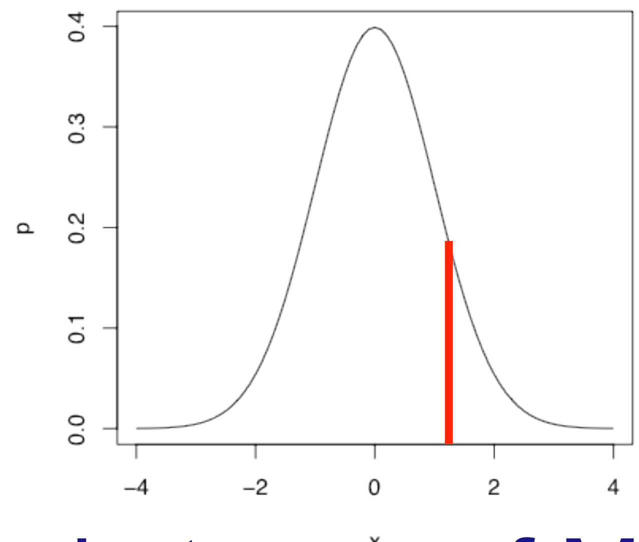
ROBUST DISPERSION (1D)

12	13	14	21	22	26	33	35	36	37	39	42	45	47	54	57	61	68	450

- interquartile range (IQR)
  - difference between 25% and 75% quartiles
- MAD: Median Absolute Deviation
  - $median(|Y_i - \tilde{Y}|) \quad \text{where } \tilde{Y} = median(Y)$
- breakdown points?
- note for symmetric distributions:
  - MAD is IQR/2 away from median

ROBUSTLY FIT A NORMAL

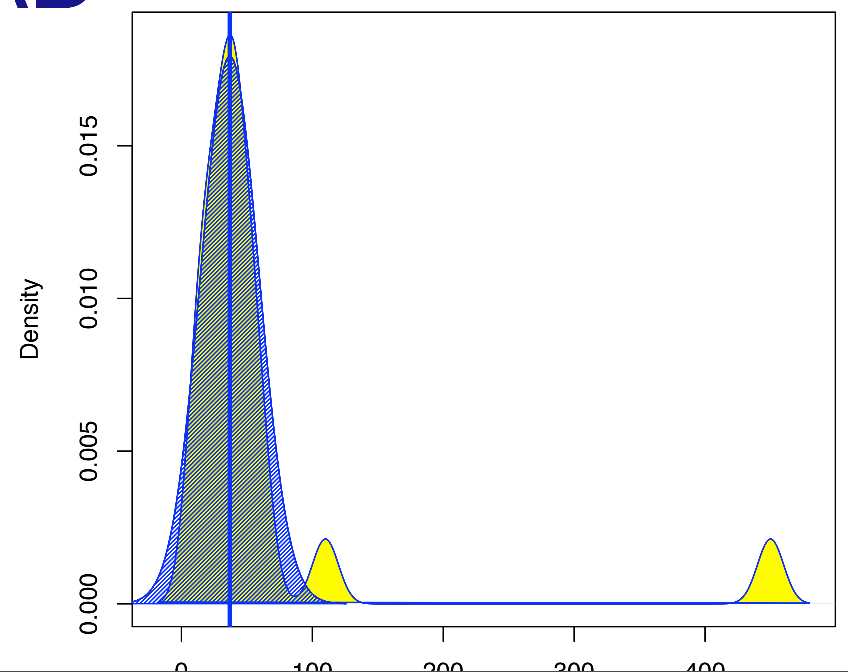
- base case: Standard Normal symmetric, center at 0
  - MAD: 75 %ile



- so estimate std dev in terms of MAD

$$\hat{\sigma} = 1.4826 \cdot \mathrm{MAD}$$

- center at median and off you go!



SCALABLE IMPLEMENTATION

- our metrics so far: *Order Statistics*
  - position in value order
- non-trivial to scale up to big data
  - but there are various tricks

SQL FOR MEDIAN?

SQL FOR MEDIAN?

```
-- A naive median query
SELECT c AS median
  FROM T
 WHERE (SELECT COUNT(*) from T AS T1 WHERE T1.c < T.c)
     = (SELECT COUNT(*) from T AS T2 WHERE T2.c > T.c)
```

SQL FOR MEDIAN?

[Rozenshtein, Abramovich, Birger 1997]

```
SELECT c as median
  FROM T x, T y
 GROUP BY x.c
HAVING SUM(CASE WHEN y.c <= x.c THEN 1 ELSE 0 END)
         >= (COUNT(*)+1)/2
   AND
       SUM(CASE WHEN y.c >= x.c THEN 1 ELSE 0 END)
         >= (COUNT(*)/2)+1
```

SORT-BASED SQL FOR MEDIAN

EFFICIENT APPROXIMATIONS

- one-pass, limited memory Median/Quantile
  - Manku, et al., SIGMOD 1998
  - Greenwald/Khanna, SIGMOD 2001
  - keep certain exemplars in memory (with weights)
    - bag of exemplars used to approximate median
- Hsiao, et al 2009: one-pass approximate MAD
  - based on Flajolet-Martin "COUNT DISTINCT" sketches
  - a *Proof Sketch*: distributed and verifiable!
- natural implementations
  - SQL: user-defined agg
  - Hadoop: Reduce function

SQL FOR APPROXIMATE MEDIAN

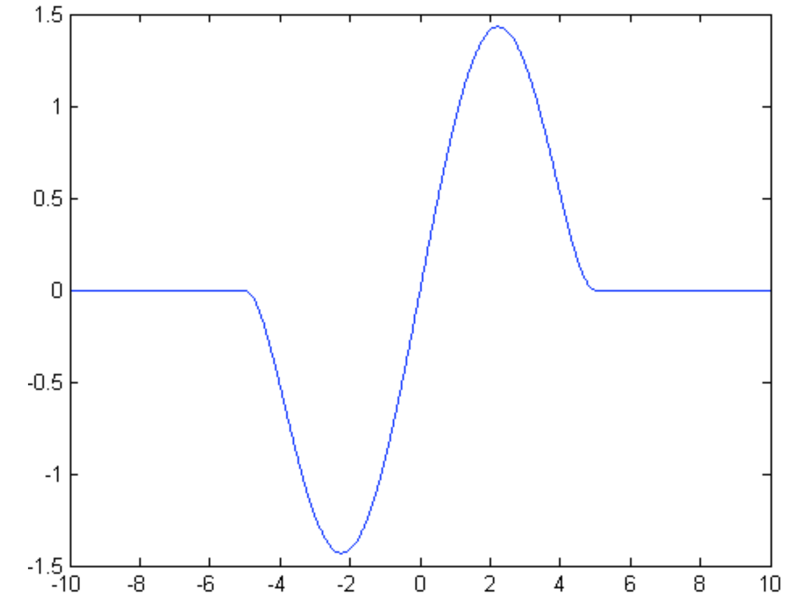
- given: UDF "approx_median"

ORDER STATISTICS

- methods so far: “L-estimators”
  - linear (hence “L”) combinations of order statistics
- simple, intuitive
- well-studied for big datasets
- but fancier stuff is popular in statistics
  - e.g. for multivariate dispersion, robust regression...

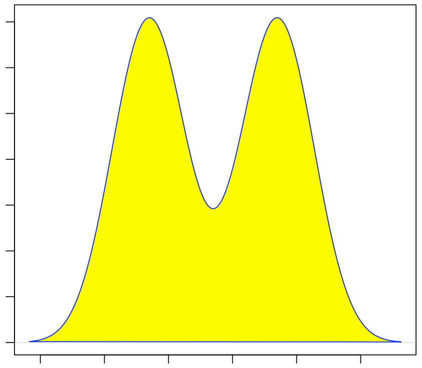
M-ESTIMATORS

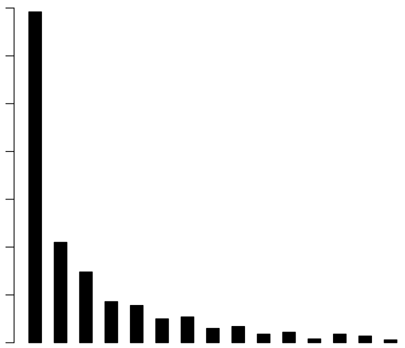
- widely used class
- based on Maximum Likelihood Estimators (MLEs)
  - MLE: maximize $\prod_{i=1}^{n} f(x_i)$ (minimize $\sum_{i=1}^{n} -\log f(x_i)$)
  - M-estimators generalize to minimize $\sum_{i=1}^{n} \rho(x_i)$
    - where *ρ* is chosen carefully
    - nice if d*ρ/dy* goes up near origin, decreasing to 0 far from origin
      - *redescending* M-estimators



STUFF IN THE PAPER

- No time today for outliers in:
  - indexes (e.g. inflation) and rates (e.g. car speed)
    - textbook stuff for non-robust case, robustification seems open
  - timeseries
    - a relatively recent topic in the stat and DB communities
  - non-normality
    - multimodal, power-series (zipf) distributions



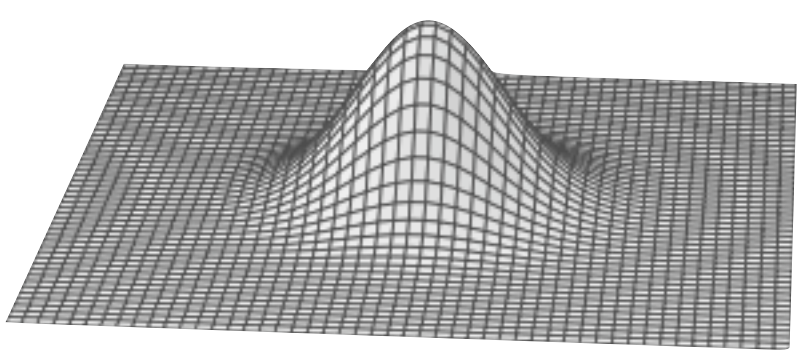


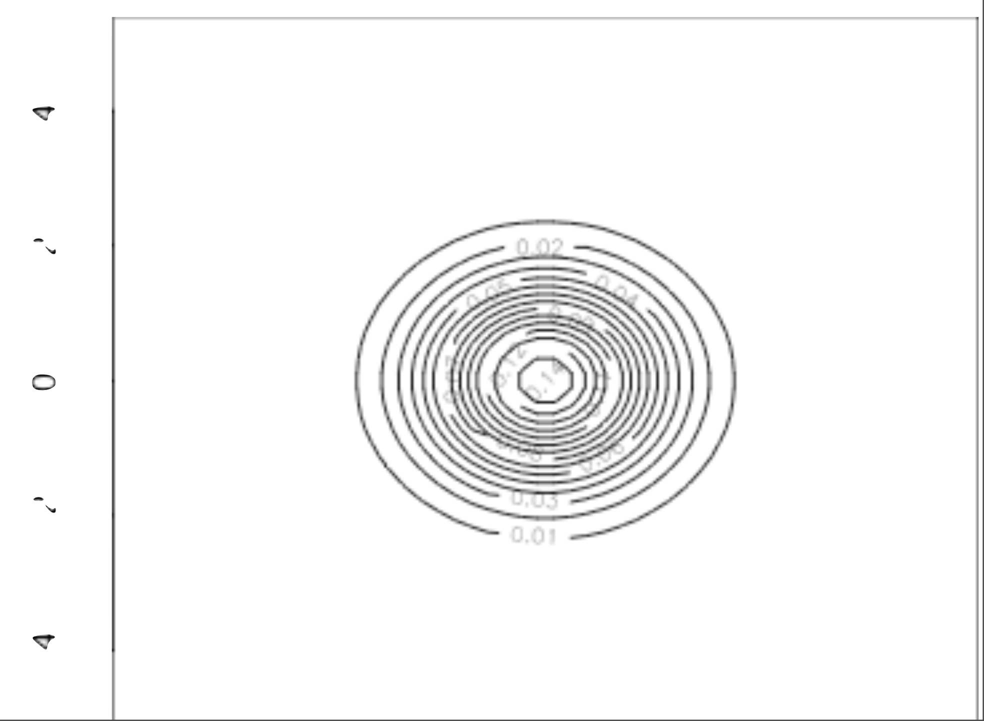
TODAY

- background
- outliers and robust statistics
- multivariate settings
- research directions

MOVING TO MULTIPLE DIMENSIONS

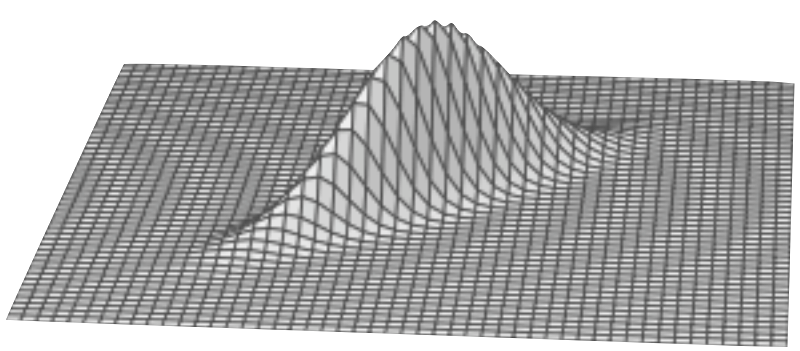
- intuition: multivariate normal
  - center: multidimensional mean
  - dispersion: ?

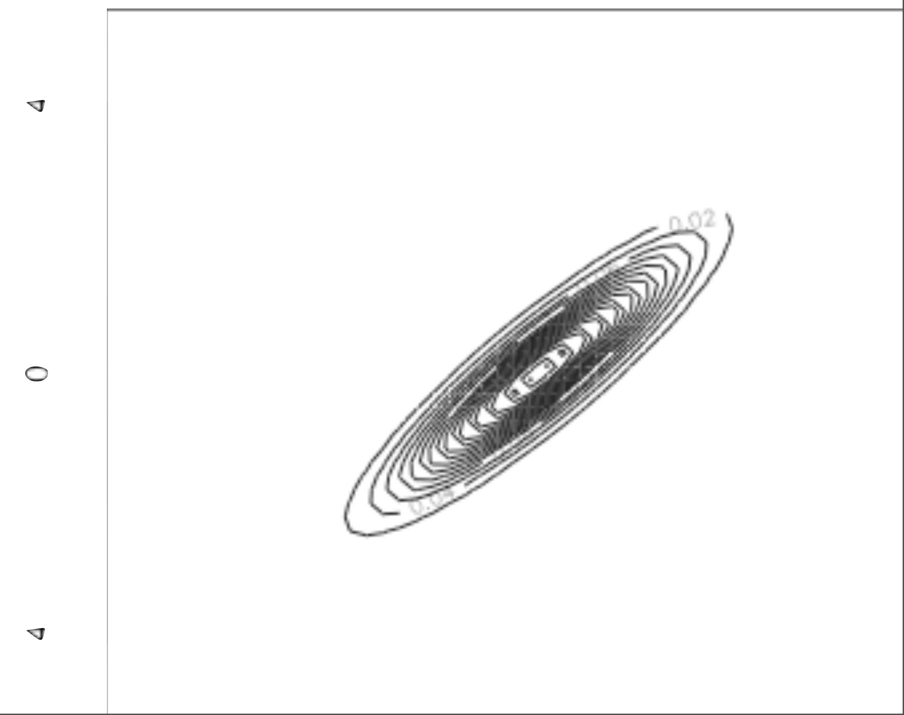




MOVING TO MULTIPLE DIMENSIONS

- intuition: multivariate normal
  - center: multidimensional mean
  - dispersion: ?





(SAMPLE) COVARIANCE

- *dxd* matrix for *N d*-dimensional points

$$q_{ij} = \frac{1}{N-1}\sum_{k=1}^{N}(x_{ik} - \bar{x}_i)(x_{kj} - \bar{x}_j)$$

- properties
  - symmetric
  - diagonal is independent variance per dimension
  - off-diagonal is (roughly) correlations

MULTIVARIATE DISPERSION

- *Mahalanobis* distance of vector x from mean μ:

$$\sqrt{(x-\mu)^T S^{-1} (x-\mu)}$$

  - where S is the covariance matrix
  - Not robust!
  - Simple SQL in 2d, much harder in >2d
    - requires matrix inversion!

ROBUST MULTIVARIATE OUTLIERS

- proposed Heuristics:
  - iteratively trim max-Mahalanobis point.
  - rescale units component-wise, then use Euclidean threshholds
- robust estimators for mean/covariance
  - this gets technical, e.g. Minimum Volume Ellipsoid (MVE)
  - scale-up of these methods typically open
- depth-based approaches
  - “stack of oranges”: Convex hull peeling depth
  - others...

TIME CHECK

- time for distance-based outlier detection?

DISTANCE-BASED OUTLIERS

- non-parametric
- various metrics:
  - *p* a *(k, D)*-outlier if at most *k* other points lie within *D* of *p* [Kollios, et al., TKDE 2003]
  - *p* an outlier if % of objects at large distance is high [Knorr/Ng, ICDE 1999]
  - top *n* elements in distance to their *k*th nearest neighor [Ramaswamy, et al. SIGMOD 2000]
- accounting for variations in cluster density
  - average density of the node' neigborhood w.r.t. density of nearest neighbors' neighborhoods [Breunig, et al, SIGMOD 2000]

ASSESSING DISTANCE-BASED METHODS

- descriptive statistics
  - no probability densities, so no expectations, predictions
- distance metrics not scale-invariant
  - complicates usage in settings where data or units not well understood

TODAY

- background
- outliers and robust statistics
- multivariate settings
- research directions

RESEARCH DIRECTIONS

- open problems in scaling
- new agenda: intelligent forms

SOME OPEN ISSUES

- scalable MAD
- robustly cleaning large, non-normal datasets
- scalable, robust multivariate dispersion
  - scalable matrix inversion for Mahalanobis (already done?)
  - Minimum-Volume Ellipsoid (MVE)?
- scale-invariant distance-based outliers?

OK, THAT WAS FUN

- now let's talk about filling out forms.

joint work ... with kuang chen, tapan parikh and others

http://www.flickr.com/photos/22646823@N08/3070394453/

DATA ENTRY

- repetitive, tedious, unglamorous
  - often contracted out to low-paid employees
  - often “in the way” of more valuable content
- the topic of surprisingly little CS research
  - compare, for example, to data visualization!

http://www.flickr.com/photos/zarajay/459002147/

DATA ENTRY!

- the first & best place to improve data quality
  - opportunity to fix the data at the source
- .. rich opportunity for new data cleaning research
  - with applications for robust (multidimensional) outlier detection!
  - synthesis of DB, HCI, survey design
- reform the form!

http://www.flickr.com/photos/48600101641@N01/316921200/

BEST PRACTICES (FROM OUTSIDE CS)

- *survey design* literature
  - question wording, ordering, grouping, encoding, constraints, cross-validation
- *double-entry*
  - followed by supervisor arbitration
- can these inform forms?
  - push these ideas back to point of data entry
  - computational methods to improve these practices

DATA COLLECTION IN LOW-RESOURCE SETTINGS

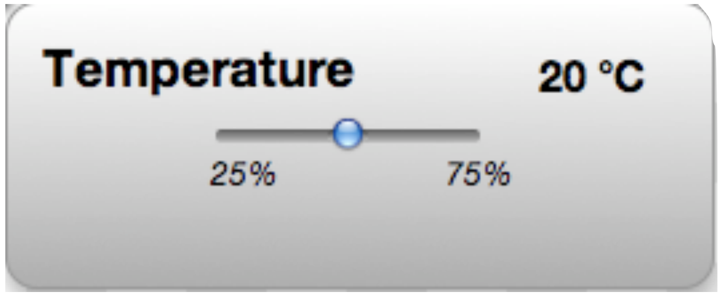
- lack of resources and expertise
- trend towards mobile data collection
  - opportunity for intelligent, *dynamic* forms

- though well-funded orgs often have bad forms too
  - deterministic and unforgiving
  - e.g. the spurious integrity problem
- time for automated and more statistical approach
  - informed by human factors

PROPOSED NEW DATA ENTRY RULES

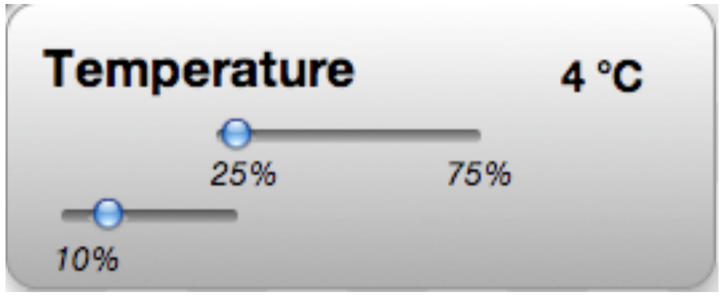
- feedback, not enforcement
  - interface *friction*
    - inversely proportional to *likelihood*
  - a role for data-driven *probabilities* during data entry
  - annotation should be easier than subversion
- friction merits explanation
  - role for *data visualization* during data entry
  - gather good evidence while you can!
- theme: *forms need the database*
  - and vice versa

FEEDBACK WIDGETS



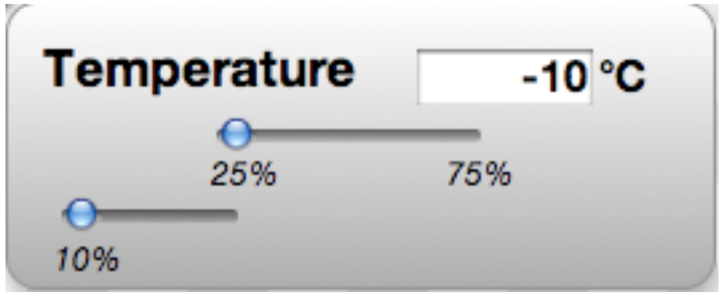
- a simple example
  - the point: these need not be exotic

FEEDBACK WIDGETS



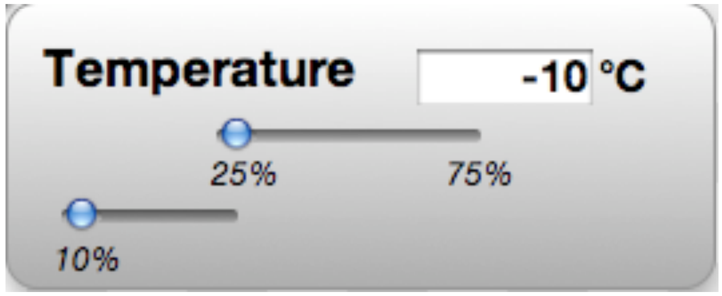
- a simple example
  - the point: these need not be exotic

FEEDBACK WIDGETS



- a simple example
  - the point: these need not be exotic

FEEDBACK WIDGETS



- a simple example
  - the point: these need not be exotic
  - a pure application of simple robust stats!

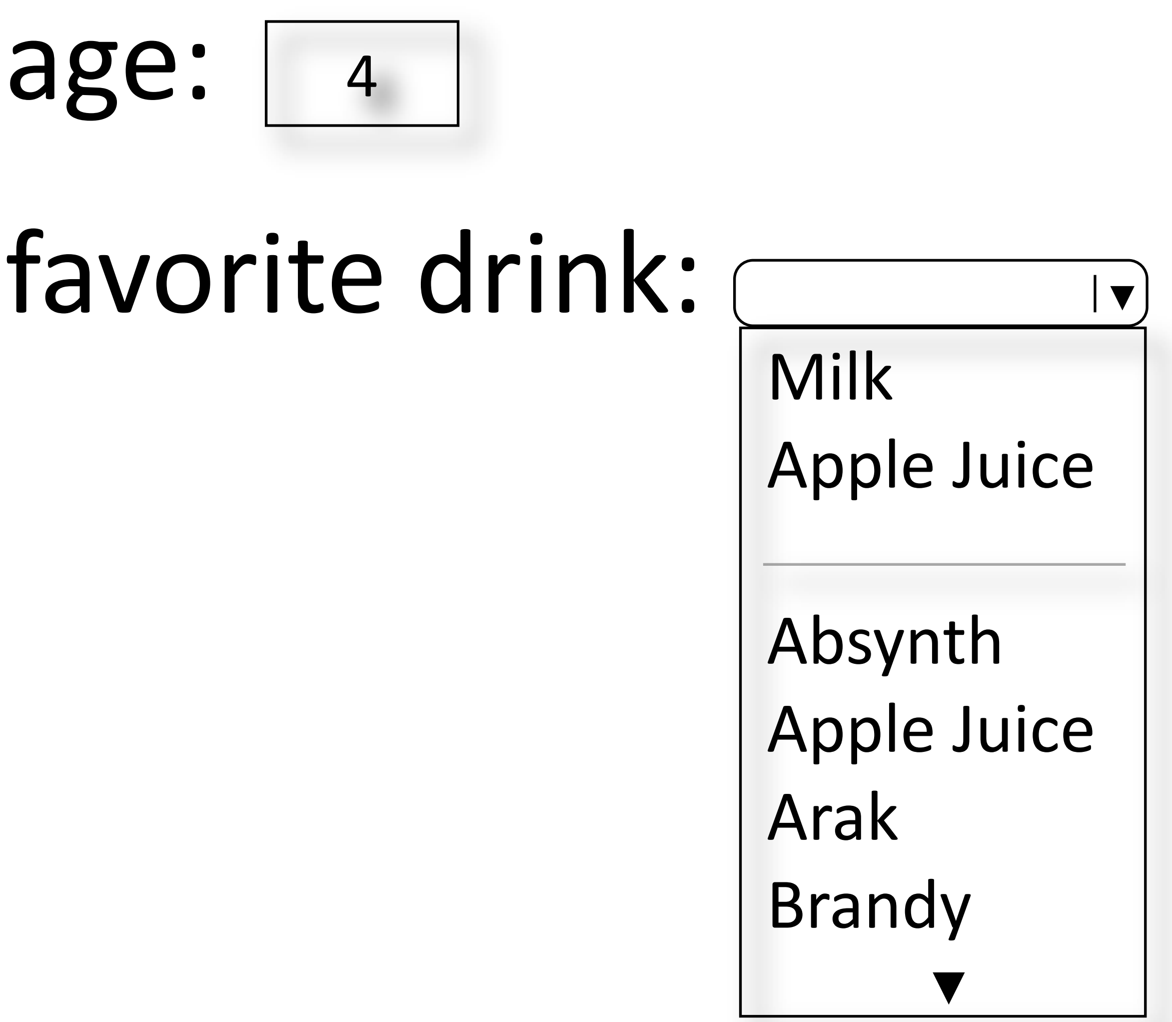
REQUIRES MULTIVARIATE MODELING

age:

favorite drink:

- this is harder to manage
  - computationally, and from HCI angle

REQUIRES MULTIVARIATE MODELING



- this is harder to manage
  - computationally, and from HCI angle

QUESTION ORDERING!

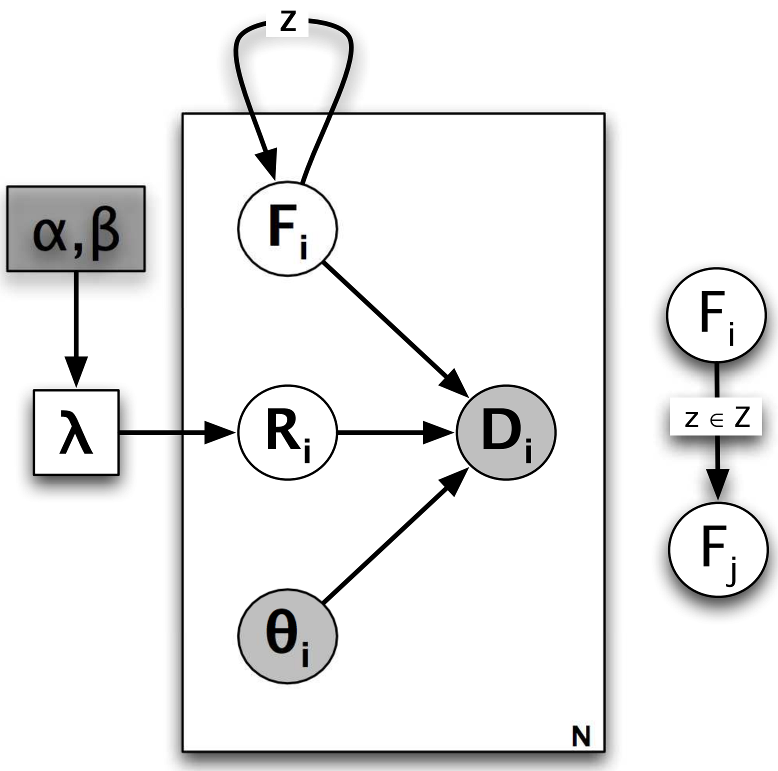
- *greedy information gain*
  - enables better form feedback
  - accounts for attention span
  - *curbstoning*

http://www.flickr.com/photos/25257946@N03/2686015697/

REASKING AND REFORMULATION

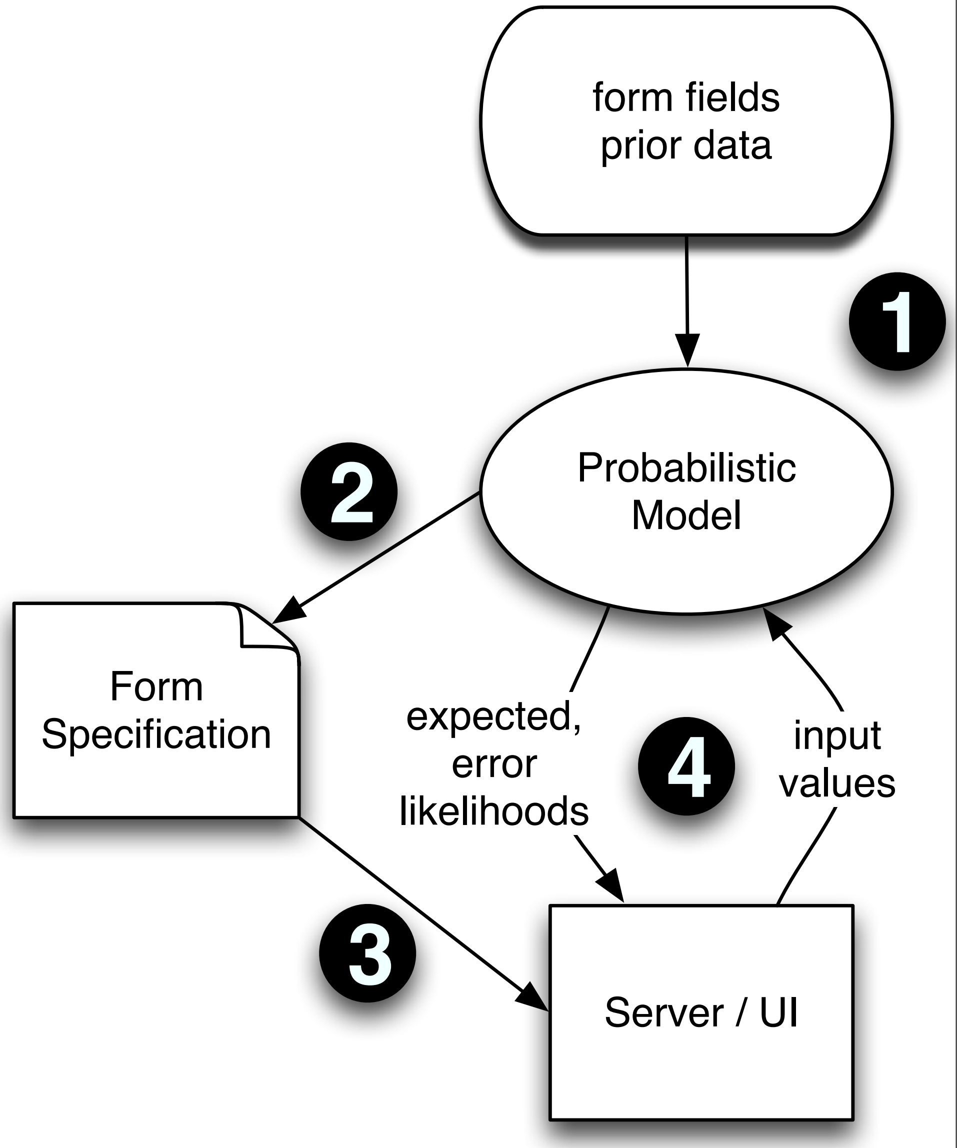
- need joint data model and *error model*
  - requires some ML sophistication
- error model depends on UI
  - will require some HCI sophistication
- reformulation can be automated:
  - e.g. quantization:

1. adult/child
2. age

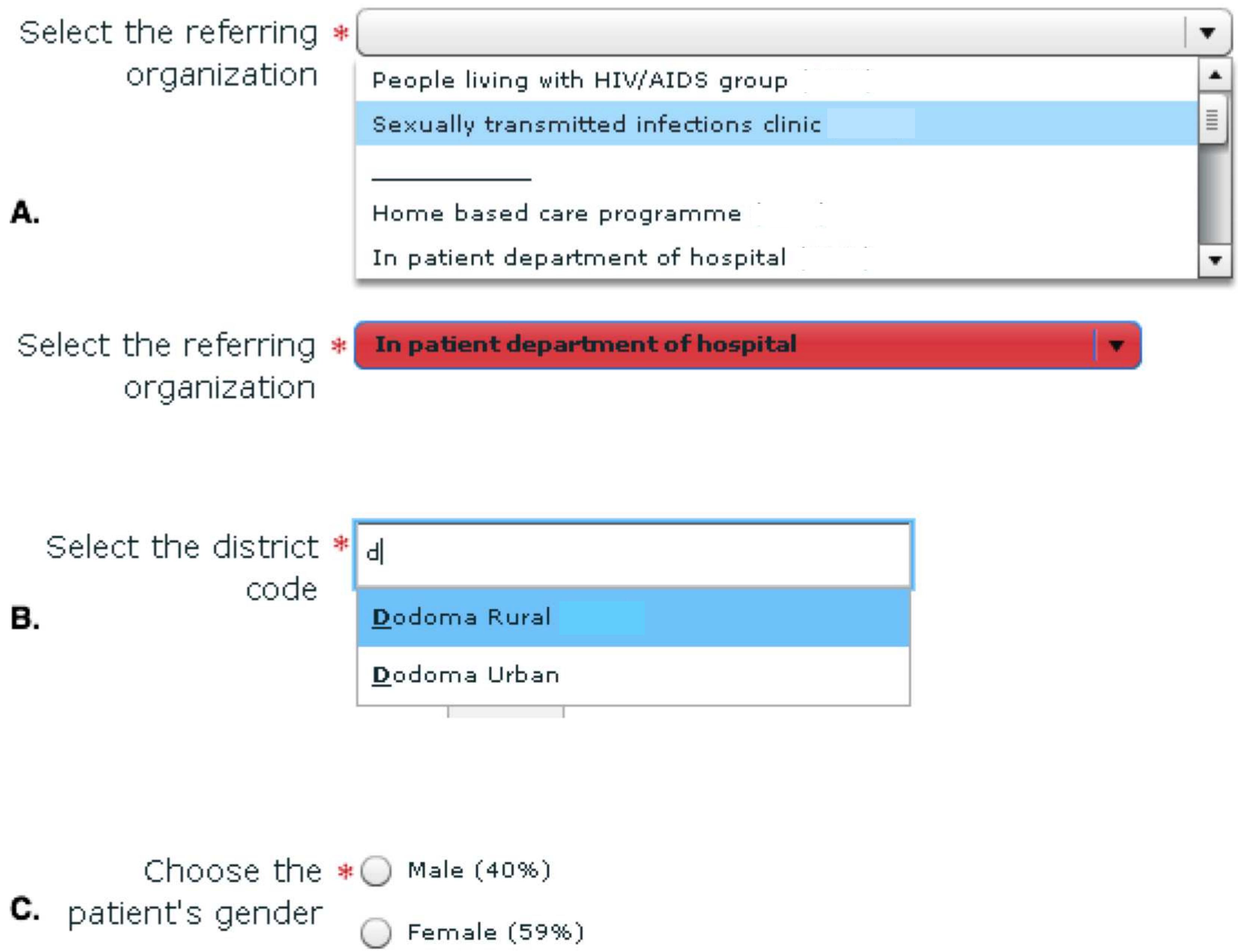


USHER

- learn a graphical model of all form variables, learn error model
  - structure learning & parameters
- optimize flexible aspects of form
  - *greedy information gain* principle for question ordering
    - subject to designer-provided constraints
- dynamically parameterize during form filling
  - decorate widgets
  - reorder, reask/reformulate questions



EXAMPLE WIDGETS



reduced friction, likelihood hints

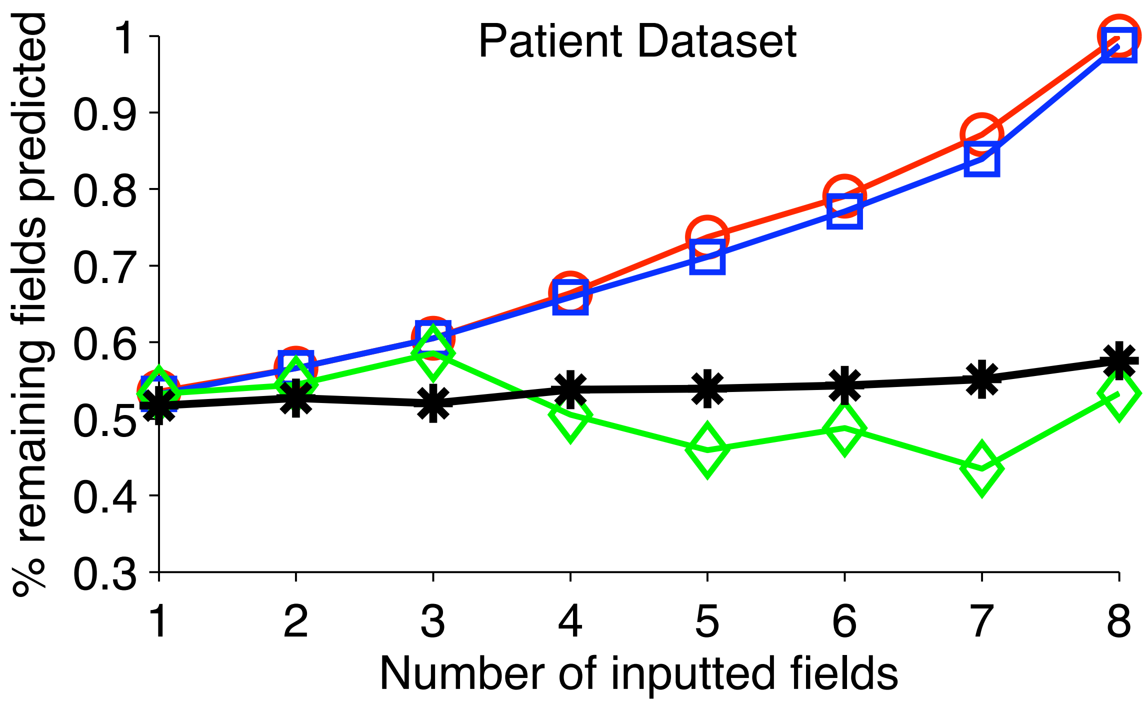
post-hoc assessment

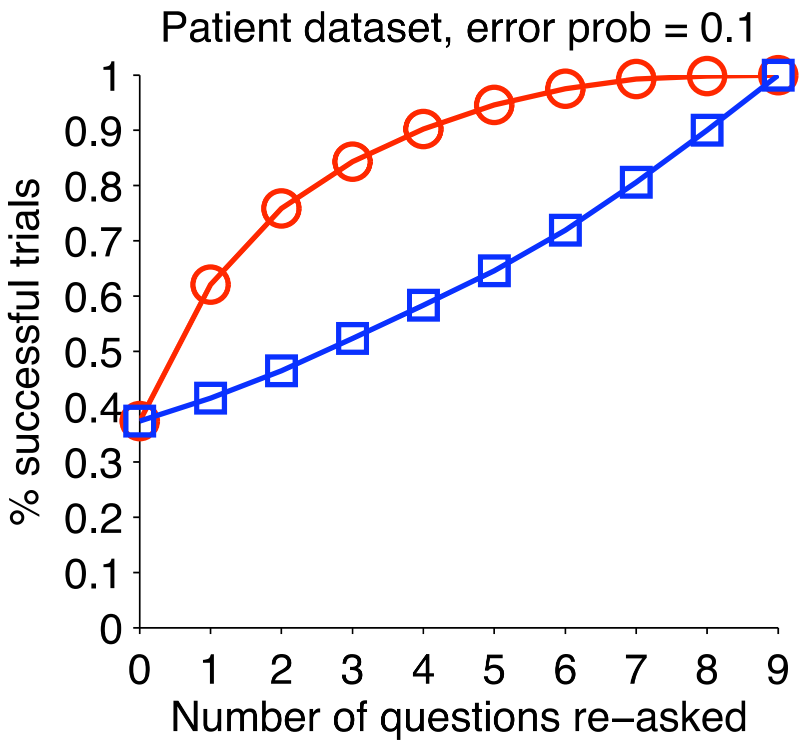
reduced friction

explicit probabilities

INITIAL ASSESSMENTS

- Tanzanian HIV/AIDS forms, US political survey
- Simulation shows significant benefits
  - both in reordering and reasking models
- User study in the works





CONCLUSIONS

- DB community has much to learn about quantitative data cleaning
  - e.g. robust statistics
- and much to offer
  - scalability, end-to-end view of data lifecycle
- note: everything is “quantitative”
  - we live in an era of big data and statistics!
- work across fields, build tools!
  - DB, stats, HCI, org mgmt, ...

ADDITIONAL READING

- *Exploratory Data Mining and Data Cleaning,* Tamraparni Dasu and Theodore Johnson, Wiley, 2003.
- *Robust Regression and Outlier Detection,* Peter J. Rousseeuw and Annick M. Leroy, Wiley 1987.
- "Data Streams: Algorithms and Applications"*.* S. Muthukrishnan. *Foundations and Trends in Theoretical Computer Science* 1(1), 2005.
- *Exploratory Data Analysis,* John Tukey, Addison-Wesley, 1977.
- *Visualizing Data.* William S. Cleveland. Hobart Press, 1993.

WITH THANKS TO...

- Steven Vale
  - UN Economic Council for Europe
- Sara Wood, PLOS
- the Usher team:
  - **Kuang Chen**, Tapan Parikh, UC Berkeley
  - Harr Chen, MIT

EXTRA GOODIES

RESAMPLING: BOOTSTRAP & JACKNIFE

- computational solution to small or noisy data
  - sample, compute estimator, repeat
  - at end, average the estimators over the samples
- recent work on scaling
  - see MAD Skills talk Thursday
- needs care: any bootstrap sample could have more outliers than breakdown point
- note: turns data into a sampling distribution!

ASIDE 1: INDEXES

- Rates of inflation over years
  - 1.03, 1.05, 1.01, 1.03, 1.06
  - $10 at start = $11.926 at end
  - want a center metric *μ* so 10**μ*^5 = $11.926
- *geometric mean:*

$$\sqrt[n]{\left(\prod_{i=1}^{n} k_i\right)}$$

  - sensitive to outliers near 0.
  - breakdown pt 0%

ASIDE 2: RATES

- Average speed on a car trip
  - 50km@10kph, 50km@50kph
  - travel 100km in 6 hours
  - "average" speed 100km/6hr = 16.67kph
- *harmonic mean*:

$$\frac{n}{\sum_{i=1}^{n} \frac{1}{k_i}}$$

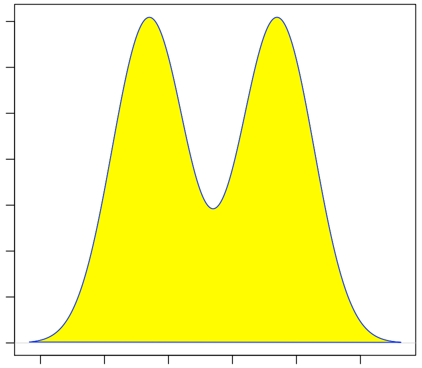
  - reciprocal of reciprocal of rates
  - sensitive to very large outliers
  - breakdown point: 0%

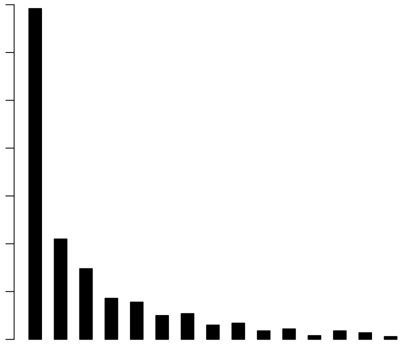
ROBUSTIFYING THESE

- Can always trim
- Winsorizing requires care
  - weight of "substitute" depends on its value
  - other proposals for indexes (geometric mean)
    - 100%
    - 1/2 the smallest measurable value
- Useful fact about means
  - harmonic <= geometric <= arithmetic
  - can compute (robust version of) all 3 to get a feel

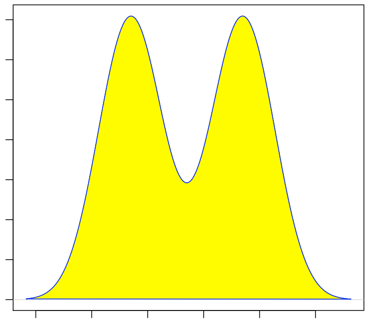
NON-NORMALITY

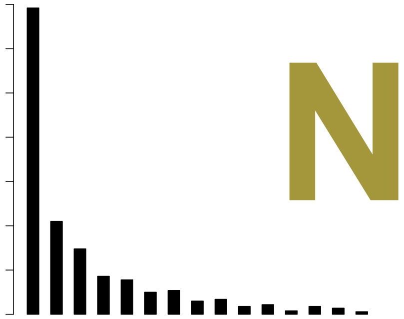
- Not everything is normal
  - Multimodal distributions
    - Cluster before looking for outliers
  - Power Laws (Zipfian)
    - Easy to confuse with normal data + a few frequent outliers
    - Nice blog post by Panos Ipeirotis





- Various normality tests
  - dip statistic is a robust test
  - Q-Q plots against normal good for intuition





NON-NORMAL. NOW WHAT?

- assume normality anyhow
  - consider likely false positives, negatives
- model data, look for outliers in residuals
  - often normally distributed if sources of noise are i.i.d.
- partition data, look in subsets
  - manual: data cubes, Johnson/Dasu's data spheres
  - automatic: clustering
- non-parametric outlier detection methods
  - a few slides from now...