2002 Mazda B-Series B2300
$5,850
122,617 miles
The Car Store
San Jose, CA (21-30 mi)

1999 Chevrolet Astro
$5,900
125,200 miles
Autometrics
El Cerrito, CA (21-30 mi)
Sketchy Approximations; Exactly Overrated

Eric Tschetter, Yahoo!
2002 Mazda B-Series B2300
$5,850
122,617 miles
The Car Store San Jose, CA (21-30 mi)

1999 Chevrolet Astro
$5,900
125,200 miles
Autometics El Cerrito, CA (21-30 mi)
Humans love approximations

- Two cars, $5,850, $5,900
  - Actually, both $6k
- \( \frac{6000 - 5850}{6000} = 2.5\% \) error
- \( \frac{6000 - 5900}{6000} = 1.66\% \) error
Blood Sugar Monitoring

- Diabetics monitor their blood sugar
- Lab test - 100% accurate, big blood draw
- OTC kits
  - FDA approved if 15% error, 95% of the time
- “Wrong”, but good enough
Approximations in Databases

- Should be acceptable
  - Unique users
    - Off by a smidge? Whatevs
  - Pageviews
    - This one too!
  - Ad impressions
    - Ok, that’s revenue data
SETTLE!?

WHY???????????
BECAUSE

ASSOCIATIVE
Sketches!

- Single-pass algorithms
- Current usage: Cardinality approximation
  a. Unique users
- Configurable error
- Distributable
  a. Compute local sketch
  b. Merge sketches
A Couple Options - HLL

- HyperLogLog (HLL)
  - Small memory (~1kb for ~2% error)
  - Fast Union
  - Cannot do intersection or disjunction
    - Inclusion-Exclusion = uncontrolled error
A Couple Options - KMV

- K-Minimum Values (KMV)
  - Medium memory (~256kb ~1.5% error @ 2 std. dev)
    - Constant memory better than O(n)
  - Optimizations available for union
  - Intersection/Disjunction allowed!