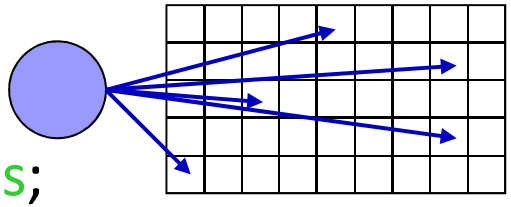
Sketches: Past, present and future

Graham Cormode
graham@research.att.com

A sketch by any other name…



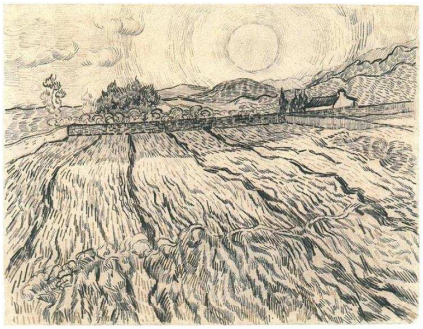
- Sketches, a.k.a: dimensionality reduction; summaries; synopses; sparse approximations; random projections; lossy compression; hash kernels; compressed sensing…
- Underlying idea: compact representation captures key features
  - Easy to compute and merge over large data
  - Provide accuracy guarantees as a function of size
- Different sketches have different properties:
  - Bloom Filter represents a set of items for membership queries
  - Count-Min sketch of a vector for point queries, inner products
  - Matrix random projections for matrix product, regression

Sketch Past (late 90's-)

- Large internet search engine analyzes logs using sketches
  - Instantly track popular queries, identify changes, spikes
- Large internet service provider monitors network using sketches
  - Real-time stats on (source, dest) pairs, round trip time distribution
- Large databases use sketches to optimize performance
  - Test (possible) presence of data before expensive look-up
- Primarily, fast real-time performance on streams of (vector) data
  - With or without slow access to exact data

Sketch Present

- Many opportunities for sketch ideas in this community
  - Wealth of techniques have been developed
  - Algorithms are simple, robust, implemented



- But sketches are no silver bullet for big data…
  - No general theory, new problems require custom solutions
  - No composability in general
    - Given matrices X, Y, can sketch for (XY) but not $(X^TY^{-1}X)^{-1}X^TY^{-1}$
  - Solve problems fast we could solve "exactly" given more resources

Sketch Future

- Bloom Filters have had huge impact in commercial 'big data'
  - Bloom Filter Principle: "Whenever a list or set is used, and space is an issue, a Bloom filter should be considered"
- Other sketches can have similar impact
  - Representing massive vectors, matrices, multisets
  - Sketch Principle: "Whenever a vector is used, and space is an issue, a sketch should be considered"?
- More methods for graphs, linear algebra is open problem
- More info:
  - Wiki: **sites.google.com/site/countminsketch/**
  - "Sketch Techniques for Approximate Query Processing" **dimacs.rutgers.edu/~graham/pubs/papers/sk.pdf**