Approximate Query Processing: Overview and Challenges

Peter J. Haas

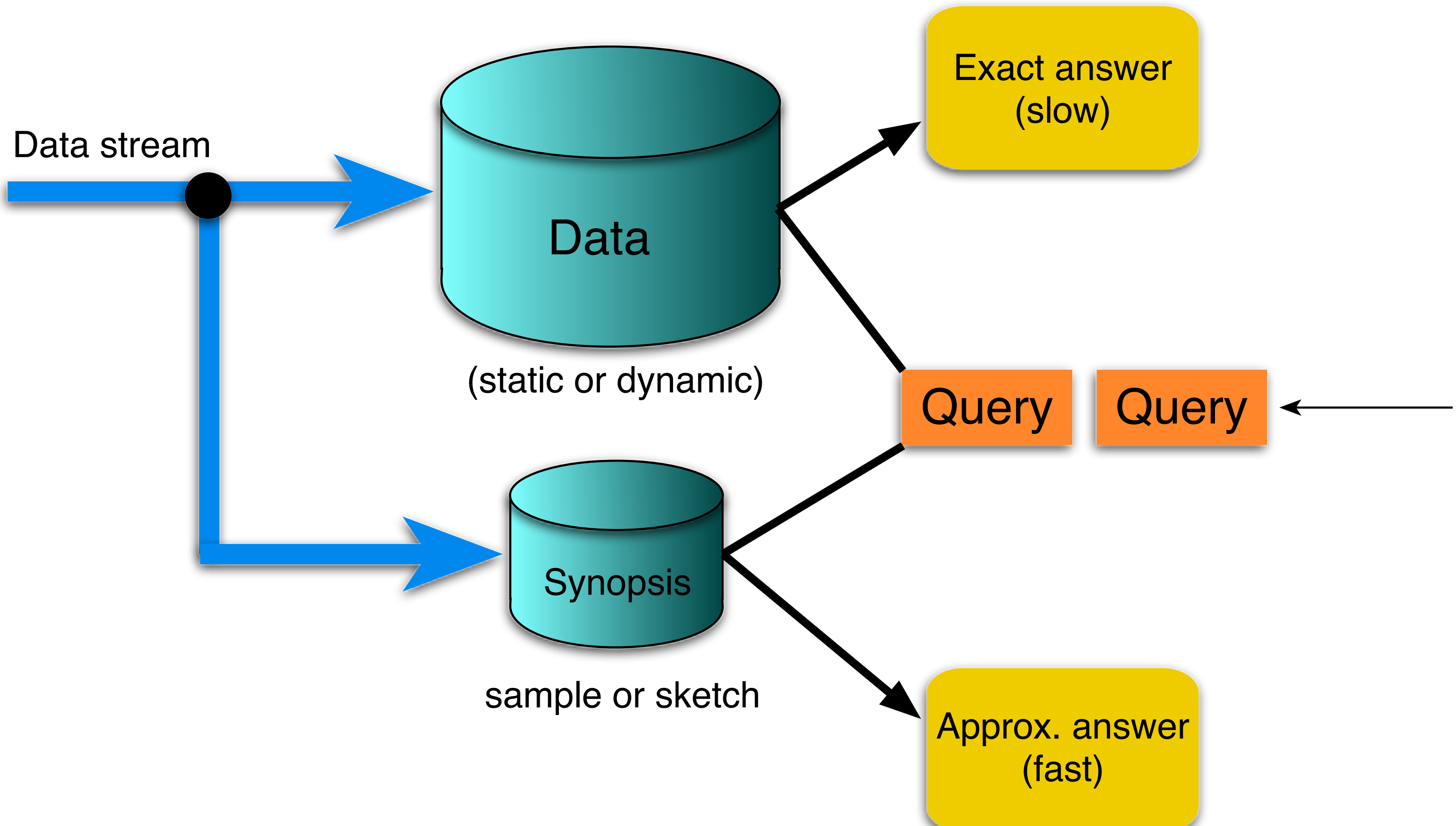
College of Information and Computer Sciences
University of Massachusetts Amherst

Thanks to:

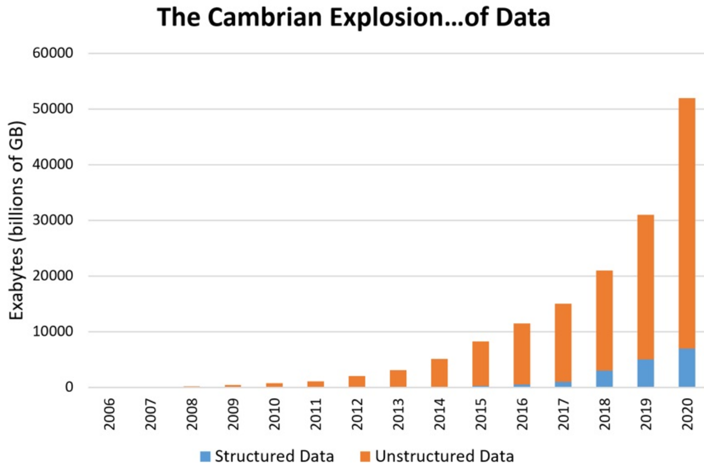
Andrew McGregor
Barzan Mozafari

EDBT 2018

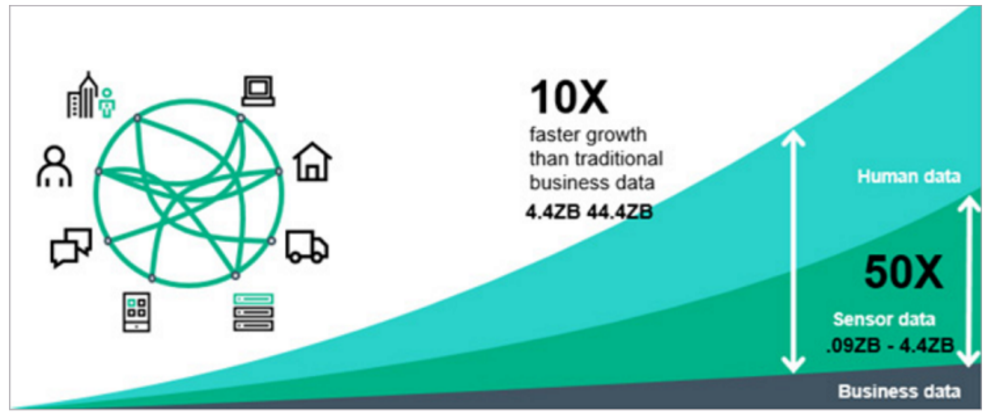
Approximate Query Processing (APQ)



AQP is More Important Than Ever



Source: Patrick Cheesman 2016



Source: InsideBIGDATA 2017

How to deal with data explosion?

→ Parallel/Distributed Computing
- Costly for ordinary people (EC2 = $)
- Not eco-friendly
- Some algorithms not embarrassingly parallel
- Concurrent queries degrade performance

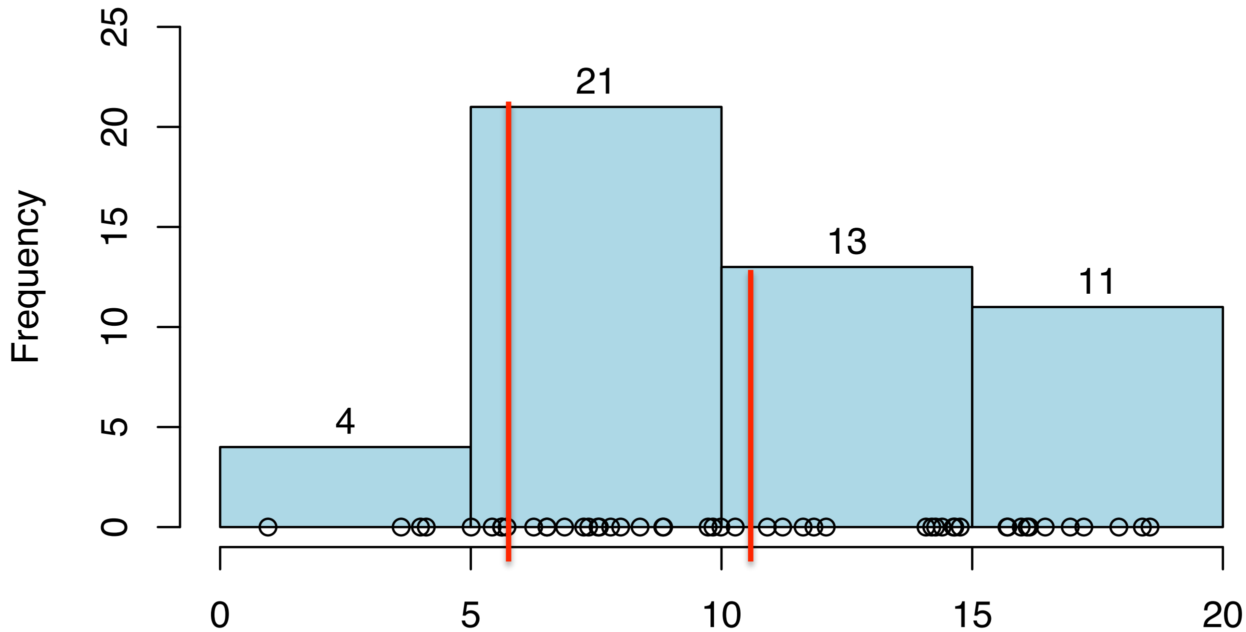
→ AQP
- Cheap
- Green

→ Can combine both

APQ Canonical Examples I

Histogram:

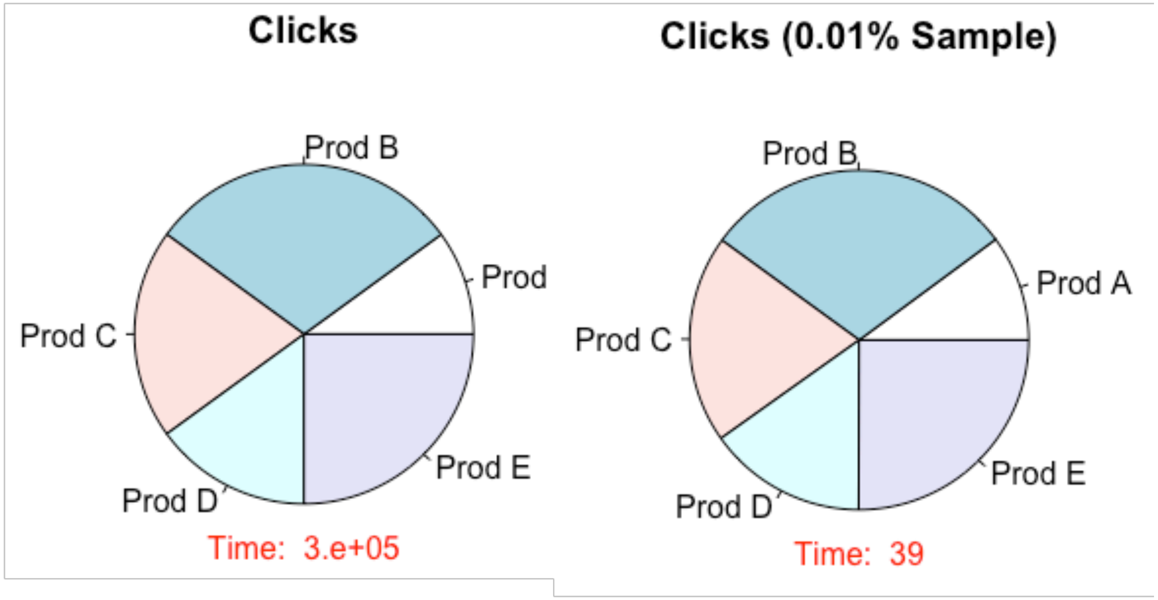
- SELECT COUNT(x) WHERE 5.1 < x < 10.3
- Exact answer: 21
- Approximate answer:
  (4.9/5) * 21 + (0.3/5) * 13 = 21.36



APQ Canonical Examples II

Sample:

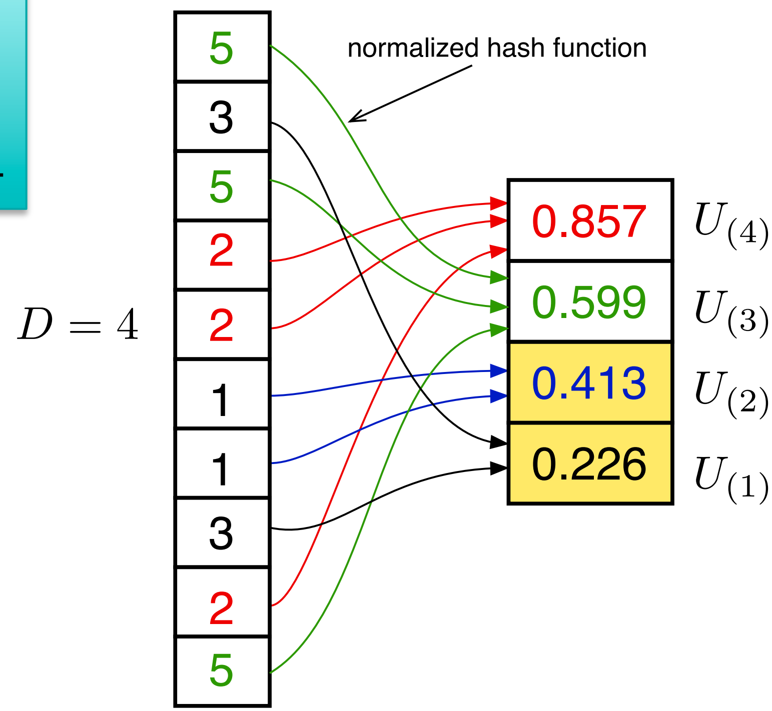
- SELECT SUM(prod) FROM clicks GROUP BY prod

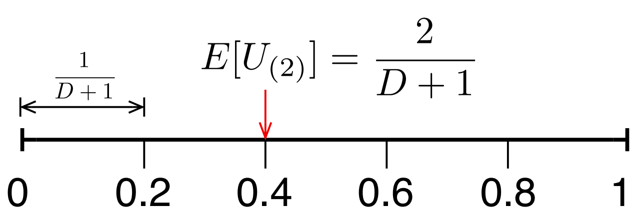


APQ Canonical Examples III

Sketch

- SELECT COUNT(DISTINCT x)
- Exact answer: 4
- Approximate answer: (2/0.413) – 1 = 3.84





$$D = \frac{2}{E[U_{(2)}]} - 1$$

$$\approx \frac{2}{U_{(2)}} - 1$$

A Taxonomy of APQ Problems

	Simple analytics		Complex analytics		Machine Learning	
Static queries	Heavy hitters, Max/min, Quantiles, Distinct values, Frequency moments	Sketches (FM, AMS, LSH, …) Random projections, Bayesian models …	Graph mining, Fixed analytic workflows	Spanner (distances) Sparsifer (cuts) SNAPE samples (vertex cover)	Clustering, Classification, Regression, Model mgmt, Data cleaning	CoreSets, Time-biased samples, Uniform/ stratified samples
Predict. queries and data	SPJ+agg queries, L_p distances Range sums, K-nearest neighbors, Subset sums	Stratified/VarOpt/ Measure-biased/ CR samples, Sample + index, Workload-based wavelets and histograms	SQL queries, Visual analytics Analytic workflows	Bayesian and maxEnt models	ML workflow	?
Ad hoc queries	SPJ+agg queries Visual analytics	Uniform samples, Multi-dim. histograms Bayesian models	SQL queries	Injected distinct samplers (Quickr)	Ad hoc ML	?

SPJ = Select, Project, Join

Challenge: Industrial Strength APQ Systems (Mozafari 2017)

OLAP Workloads	TPC-H	TPC-DS	Facebook	Conviva Inc.	Customer
System	ABM [1]	QuickR [2]	BlinkDB [3]	[1] + [3]	Verdict [5]
Unsupported Queries	See paper	Full outer joins	Joins of multiple fact tables	Joins of multiple fact tables	Multiple fact joins, nested, textual filters
Percentage of Supported Queries	68%	> 90%	> 96 %	91%	74%
Speedup	10x	2x	?	10-200x	2-20x

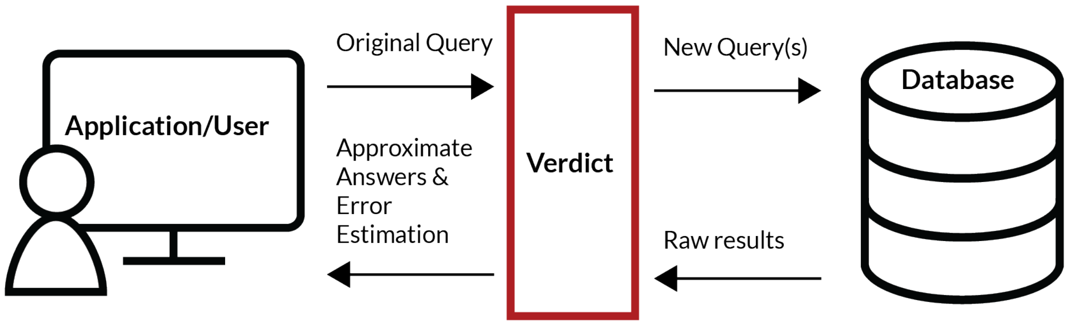
Source: Mozafari 2017

So far: relatively simple SQL queries

Challenge: Industrial Strength APQ Systems (Mozafari 2017)

Compatibility with existing engines: Middleware required

- Efficiency challenges)
- Automatic query rewrite needed



Verdict Architecture (http://verdictdb.org)

Source: Mozafari 2017

Dealing with existing interfaces

- Compatibility and user friendliness
- High-level accuracy contracts
  (at least p% accurate with p% prob and exist w. p% prob)

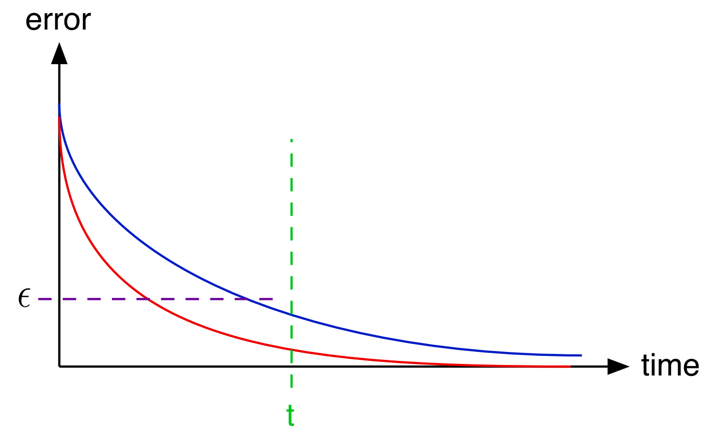
Challenge: Industrial Strength APQ Systems (Mozafari 2017)

Query planning

- Different query-plan criteria from traditional query optimization
  - Minimize time to acceptable error or error within time constraint
  - Error can be hard to predict and control
    - So far: Analytical formulas, Bayesian modeling, analytical/Poisson bootstrap
    - *A priori* error guarantees (*sample+seek* w. measure-biased sampling, indexes…)
  - Latency is *very* hard to predict (esp. in parallel/distributed setting)

- Automatically choosing the right synopsis
  - Run a competing set of synopses and combine answers
  - Theory? E.g, space complexity analysis [Kaushik et al. 2005]

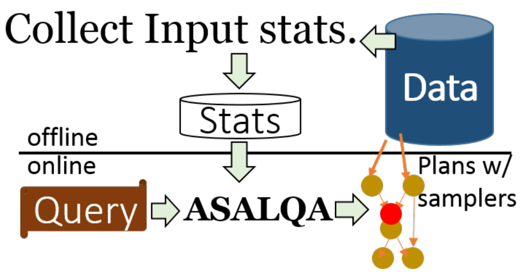
- Learning based on prior results + exploration (extend to dynamic data)



Challenge: Industrial Strength APQ Systems

Handling Complex analytics

- Arbitrary SQL aggregate queries
  - Subqueries: [Joshi and Jemaine 2009; Rusu et al. 2015]
  - Quickr [Kandula et al. 2016] inject distinct-samplers into query plan (multiple passes)



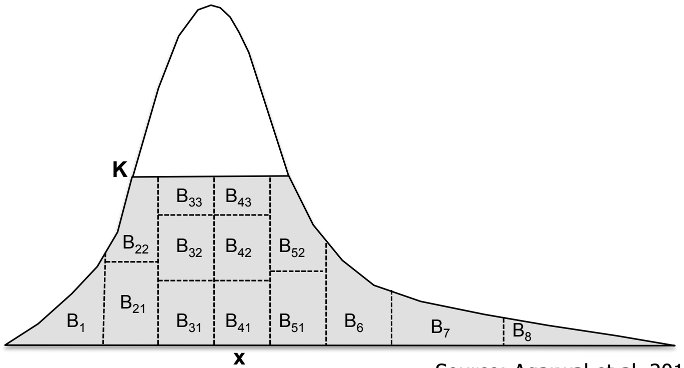
Source: Kandula et al. 2016

- Set-valued queries [Ioannidis and Poosala 1999]
- Modern queries
  - Graph queries
  - ML (coreSets, model management, sampleClean)
- Sequences of analytical operations: error propagation? [Ioannidis & Christodolakis 1991]
- Error estimation and guarantees
  - Even in "simple" SPJ+Agg setting with GROUP-BY and selection predicates

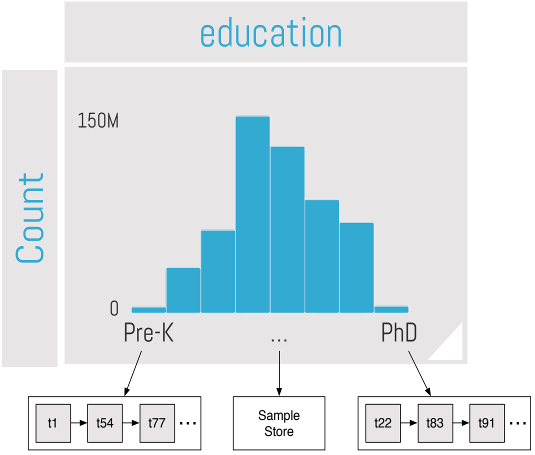
Challenge: APQ for Visual Analytics I

Achieving high interactivity

- Combine ad-hoc sampling with precomputed samples and indexes (e.g., AQUA, BlinkDB, IDEA, VisTrees)
- Reuse results between queries (IDEA, Verdict)
- Predict user behavior to fetch or precompute synopsis of interest (DICE, ForeCache)
- Use sketches for statistical guideposts (Foresight)



Source: Agarwal et al. 2013

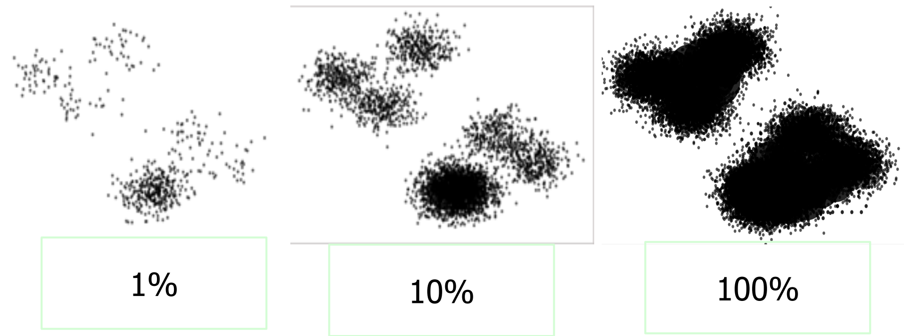


Source: Galakatos et al. 2017

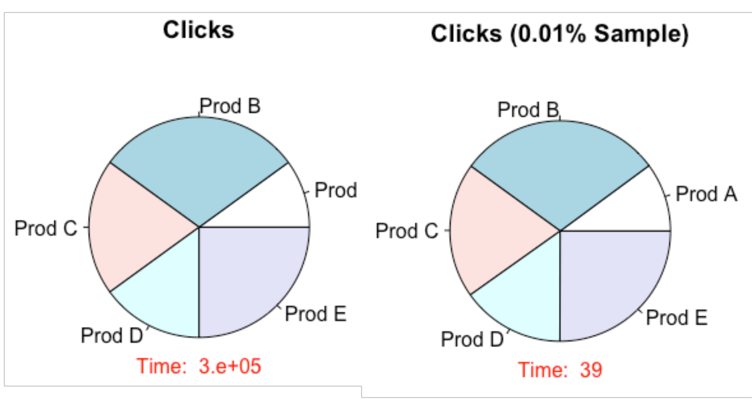
Challenge: APQ for Visual Analytics II

APQ and perception

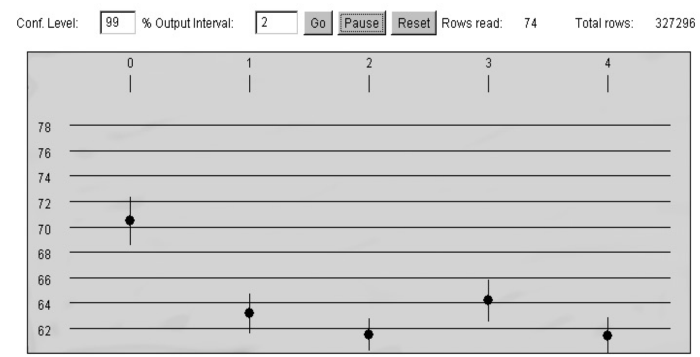
- Not well understood
- Need theory and user studies
- Need collaboration with HCI community



Sampling and cluster perception



A bad visualization [Few 2007]

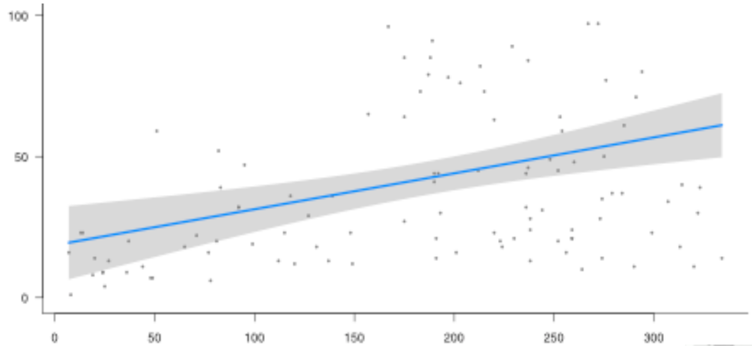


A bad interface [Fisher et al. 2012]

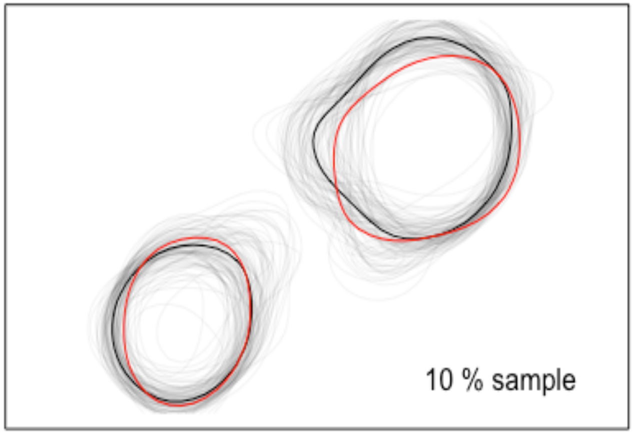
Challenge: APQ for Visual Analytics III

Visualizing uncertainty

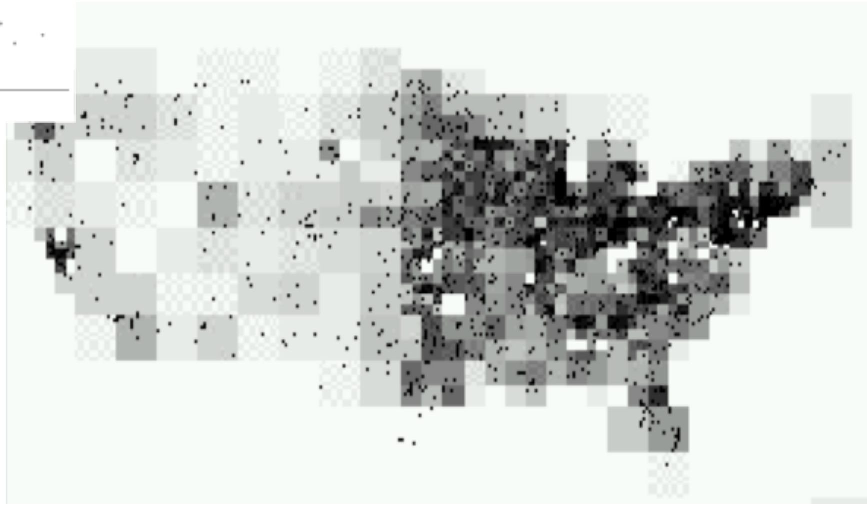
- Needed to engender trust, ensure proper inferences
- Don't need precision < screen resolution [Jugel, et al. 2014]



Finite-population confidence bands



Resampling [Kwon et al. 2017]

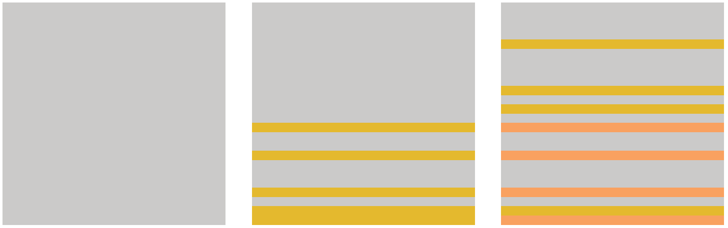


CLOUDS [Hellerstein et al. 1999]

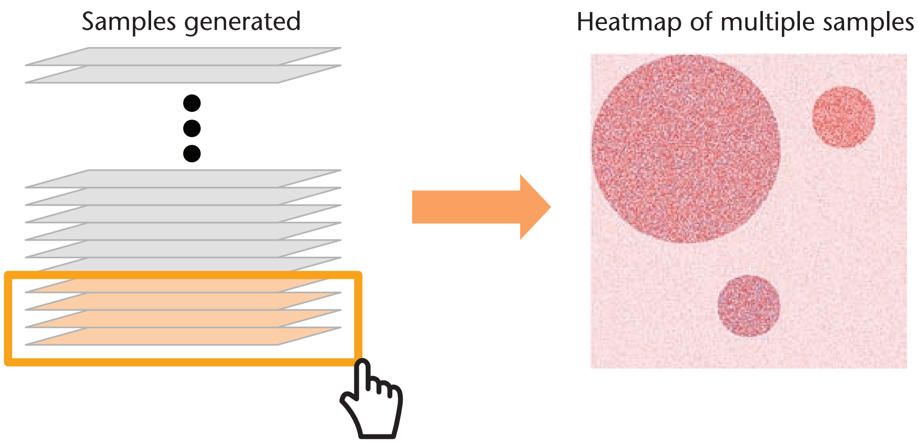
Challenge: APQ for Visual Analytics IV

Visualizing sample quality

- Helpful for building trust [Fisher et al. 2012]
- Interactive steering of sampling process [Kwon et al. 2017]

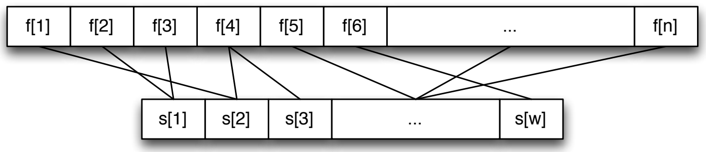


Visualizing sample quality (barrel plot)



Visualizing sample quality (dynamic layering)

Other Challenges



Combining synopses

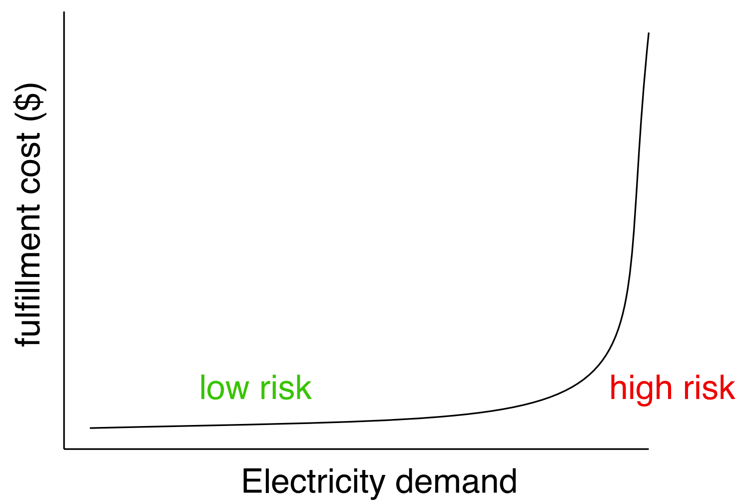
- Ex: count-min sketch ➔ l_2-sample ➔ estimate of F_2

l_2-sample: return (I, R), where

$$\Pr(I = i) = (1 \pm \varepsilon)\frac{f_i^2}{F_2} \text{ and } R = (1 \pm \varepsilon) f_i$$

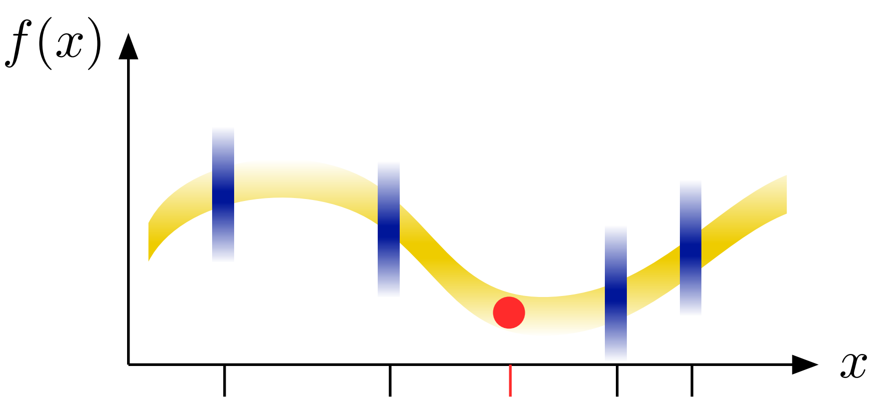
End-to-end incorporation of risk

- Data analysis for decision making under uncertainty
- Choose accuracy of approximation to control risk



Handling Multiple types of uncertainty

- Ex: AQP in probabilistic databases
- Ex: Gaussian random field interpolation



A Random Sample of References

APQ SYSTEMS

- BlinkDB: Queries with bounded errors and bounded response times on very large data. Agarwal et al., *Eurosys* 2015.
- A Handbook for Building an Approximate Query Engine. Mozafari and Niu, *IEEE Data Engrg. Bulletin* 38(3), 2015.
- Approximate query engines: Commercial challenges and research opportunities. Mozafari, *SIGMOD* 2017.
- Verdict: A system for stochastic query planning. Mozafari, *CIDR* 2015.
- Sample + Seek: Approximating Aggregates with Distribution Precision Guarantee. Ding et al., *SIGMOD* 2016.
- Quickr: Lazily approximating complex ad hoc queries in bigData clusters. Kandula et al., *SIGMOD* 2016.

SAMPLING

- Stream sampling for variance-optimal estimation of subset sums. Cohen et al., *SIAM J. Comput*. 40(5), 2011.
- A sampling algebra for aggregate estimation. Nirkhiwale et al., *PVLDB* 6(14), 2013.
- One sketch for all: Theory and application of conditional random sampling. Li et al., *NIPS* 2008.
- Temporally-biased sampling for online model management. Hentschel et al., *EDBT* 2018.
- The analytical bootstrap: A new method for fast error estimation in approximate query processing. Zeng et al., *SIGMOD* 2014.

MISCELLANEOUS

- Neighbor-sensitive hashing. Park et al., *VLDB* 2015.
- Histogram-based approximation of set-valued query-answers. Ioannidis and Poosala, *VLDB* 1999.
- Practical coreset constructions for machine learning. Bachem et al., arXiv:1703.06476 [stat.ML], *2017*.
- A sample-and-clean framework for fast and accurate query processing on dirty data. Wang et al., *SIGMOD* 2014.
- On the propagation of errors in the size of join results. Ioannidis and Christodoulakis, *SIGMOD* 1991.

References, Continued

APQ FOR VISUAL ANALYTICS

- Sampling for Scalable Visual Analytics. Kwon et al., *IEEE Comput. Graphics Appl*. 37(1), 2017.
- Visualization-aware sampling for very large databases. Park et al., *ICDE* 2016.
- Trust, but verify: Optimistic visualizations of approximate queries for exploring big data. Moritz et al., *CHI* 2017.
- VisTrees: Fast indexes for interactive data exploration. El-Hindi et al., *HILDA* 2016.
- Foresight: Rapid data exploration through guideposts. Demiralp et al., CoRR abs/1709.10513, 2017.
- Rolling the dice: Multidimensional visual exploration using scatterplot matrix navigation. Elmqvist et al.*, IEEE Trans. Visualization and Comput. Graphics* 14(6), 2008.
- Dynamic prefetching of data tiles for interactive visualization. Battle et al., *SIGMOD* 2016.

APQ SYNOPSES: SURVEYS AND COMPARISONS

- Graph stream algorithms: A survey. McGregor, *SIGMOD Record* 43(1), 2014.
- Synopses for massive data: Samples, histograms, wavelets, sketches. Cormode et al., Foundations and Trends in Databases 4(1-3), 2012.
- Approximate query processing: No silver bullet. Chaudhuri et al.*, SIGMOD* 2017.
- Synopses for query optimization: A space-complexity perspective. Kaushik et al., *ACM TODS* 30(4), 2005.

LEARNING AND BAYESIAN SYNOPSES

- Revisiting reuse for approximate query processing. Galakatos et al., *VLDB* 2017.
- Database learning: Toward a database that becomes smarter every time. Park et al., *SIGMOD* 2017.
- A Bayesian Method for Guessing the Extreme Values in a Data Set. Wu and Jermaine, *VLDB* 2007.
- Workload-Driven Antijoin Cardinality Estimation. Rusu et al., *ACM TODS* 40(3), 2015.
- Sampling-based estimators for subset-based queries. Joshi and Jermain *VLDB J.* 18(1), 2009.