From MUD to MIRE: Managing Inherent Risk in the Enterprise

Peter J. Haas

IBM Almaden Research Center
San Jose, CA

The Two Perpetual Questions

- “Where do the probabilities come from?”

- “Who is going to use this stuff in the real world?”

My background in probabilistic DB

RAQA: Resolution-Aware Query Answering for Business Intelligence
(Sismanis et al. 2009)

- OLAP querying
  (datacubes: roll-up, drill-down)
- Uncertainty due to entity resolution
- Bounds on query answers
- Implemented via SQL queries
- Conservative approach

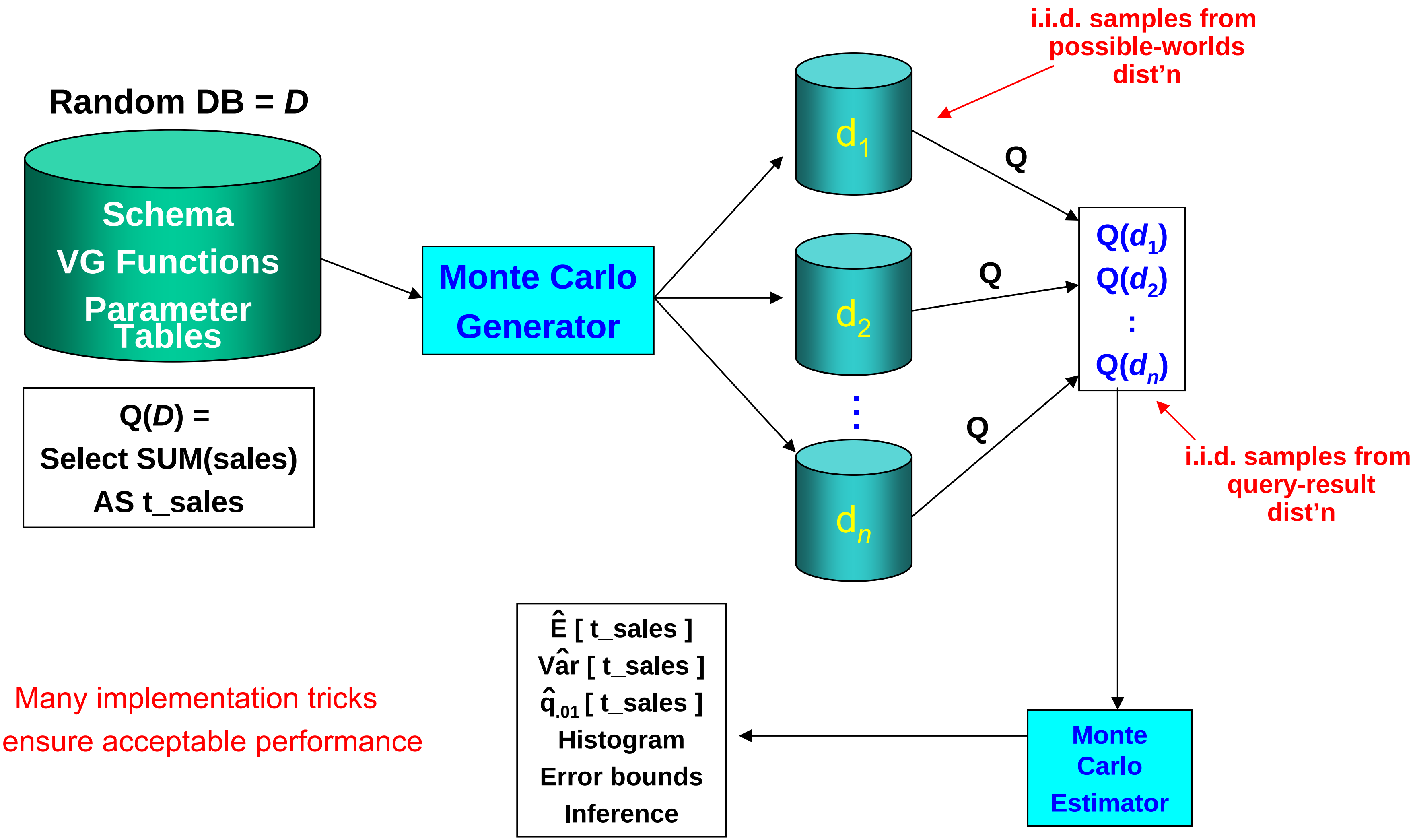
City	State	Strict range	Status
San Francisco	CA	[$30,$230]	guaranteed
San Jose	CA	[$70,$200]	non-guaranteed

Sum(Sales) group by City,State

State	Strict range	Status
CA	[$230,$230]	guaranteed

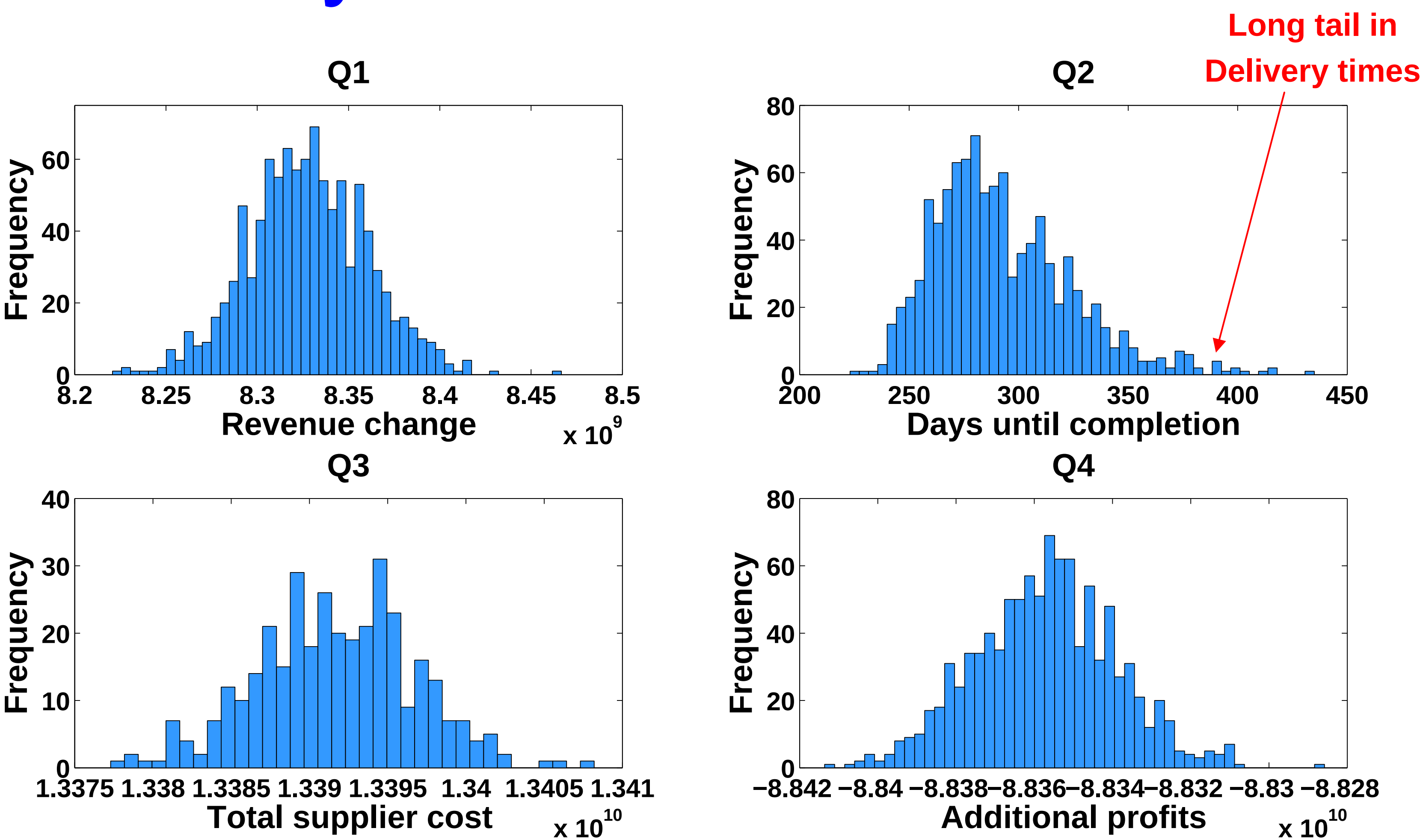
Sum(Sales) group by State

The MCDB System
(with Chris Jermaine & students)

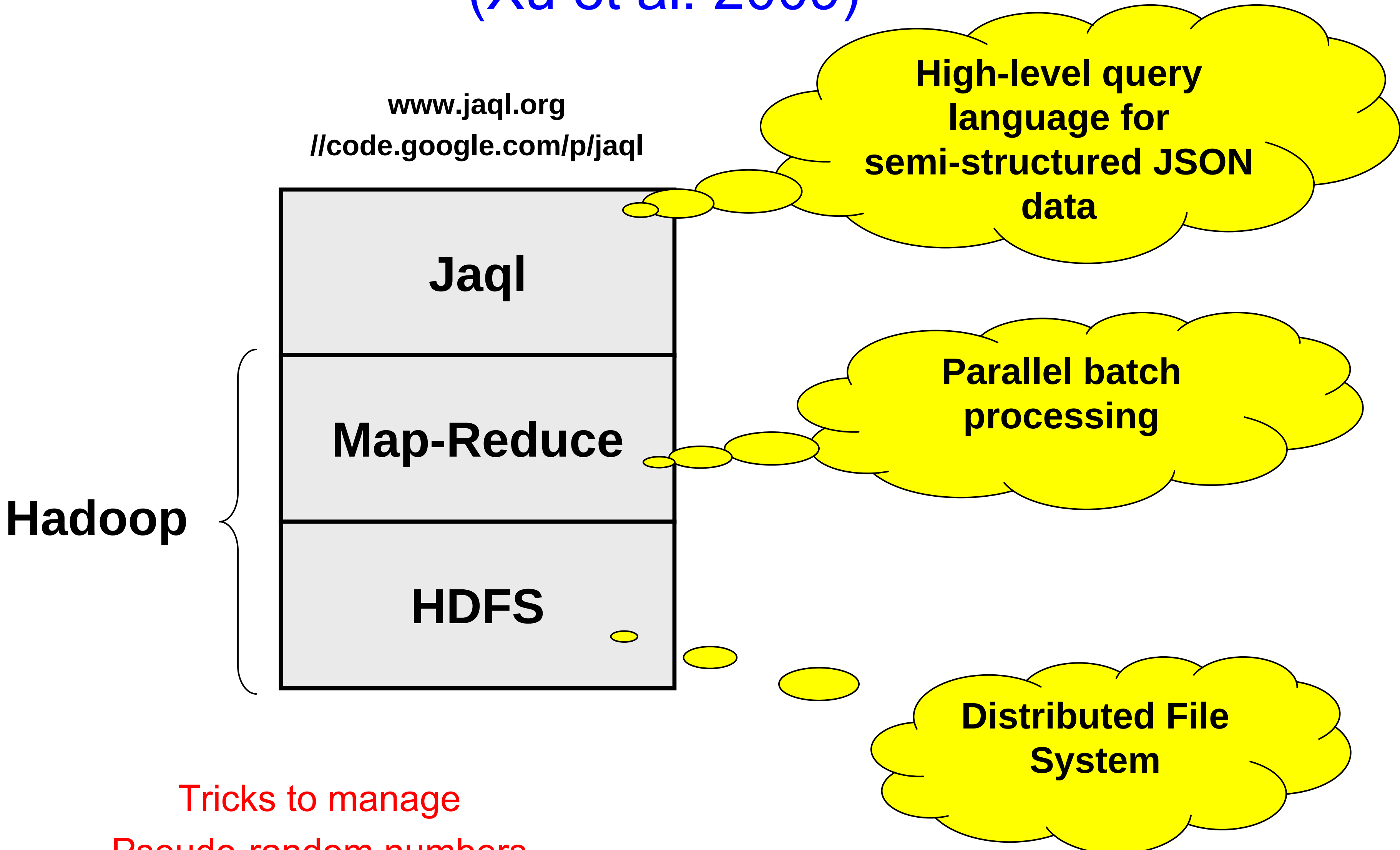


Many implementation tricks
to ensure acceptable performance

Query-Result Distributions



MC^3: MapReduce + MCDB
(Xu et al. 2009)



Tricks to manage
Pseudo-random numbers

Where do the probabilities come from?

Data-Warehouse Uncertainty

Data Integration

{John Smith, San Jose}
{John Smith, San Jose}
{John Smith, Los Angeles}

ETL

Name	City
John Smith	(SJ, 0.66), (LA, 0.33)

Name	City
John Smith	LA

Name	Sales
J. Smith	$50K

Similarity Join

City	Sales
LA	$50K

? (0.92)

Information extraction

A lovely thing to behold is Paris Hilton in the Springtime ...

System T Hotel Annotator

Hotels
NY Marriott
Paris Hilton

? (0.20)

(Michelakis et al., 2009)

09/09/2007
Re: system crash

This morning, my ORACLE system on LINUX exploded in a spectacular fireball ...

Text Miner

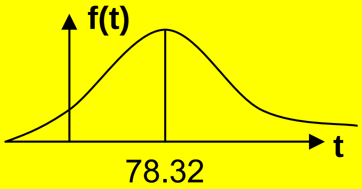
Source	Problem type
Cust0385	(DBMS, 0.8), (OS, 0.2)

Data-Warehouse Uncertainty – Cont'd

Measurement Uncertainty

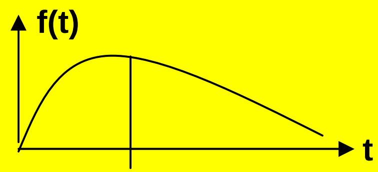
Sensor

Sensor_ID	Temp (F)
S23	78.32



System Monitor

Event	Time
Buffer overflow	10/17/2007:18:20:02



Real-World Challenges with Data-Warehouse Uncertainty

- People don't like to admit that it exists!
  - Retailers view uncertainty as failure of security, supply chain management
    - IBM research relationship manager for retail
  - Law enforcement
    - Photo ID in meth dealer trial
  - Scientists pretend data is perfect: uncertainty undermines results
    - Hans-Joachim Lenz
  - Database vendors
    - Data "cleaning" products

- Data warehouse may not even exist!
  - Ex: cancer data at medical center
  - Ex: tomato soup supply chain data

Stochastic Predictive Analytics on Big Data

- Uncertain data describes future or hypothetical events
  - Based on complex, fine-grained stochastic model over big data
  - Minimizes denial problem
- Intense recent interest in "business analytics" driven by
  - Need for low risk, quick payback projects (flexibility, low cost, fine data granularity)
  - Technical advances
    - Cloud computing
    - Software as a Service (SaaS)
    - Next generation tools, portals, visualization
- Often with a spreadsheet front end
  - $8 Billion of such tools [Gnatovich06]
  - IBM services pricing
- Lots of prototype activity
  - Fox/GreenPlum [Cohen09 MAD analytics paper]
  - VISA/IBM [Das10 SIGMOD paper]

Ex. 1: Portfolio Values

Customer

CustID	OptionID	NumShares	…
John Smith	23	50	…
…	…	…	…

EuroCallOptions

OptionID	InitVal	…	StrikeP	OVal
23	\$2.35	…	\$4.00	?
…	…	…		…

Option value
one month from now
(exercise date)

```
SELECT SUM (c.NumShares * o.Val)
FROM Customer c, EuroCallOptions o
WHERE  c.OptionID = o.OptionID
           AND c.CustType = 'Institutional'
```

Modified Black-Scholes model for European call option:

$$dV = rV\,dt + \left(a\sqrt{V}\right)V\,dW \qquad \mathrm{OVal} = \max\left(V(t_{\mathrm{final}}) - S, 0\right)$$

Simulation approximation (Euler approach):

$$V(t+\Delta t) = V(t) + rV(t)\Delta t + \left(a\sqrt{V(t)}\right)V(t)\sqrt{\Delta t}\,Z_j$$

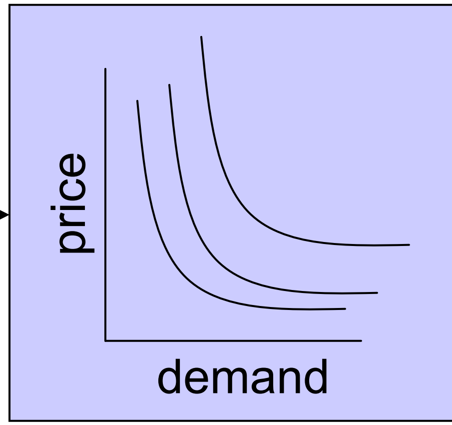
Sample from
Normal dist'n

Also CMOs, etc.

Ex. 2: Pricing Decisions

Data for all customers

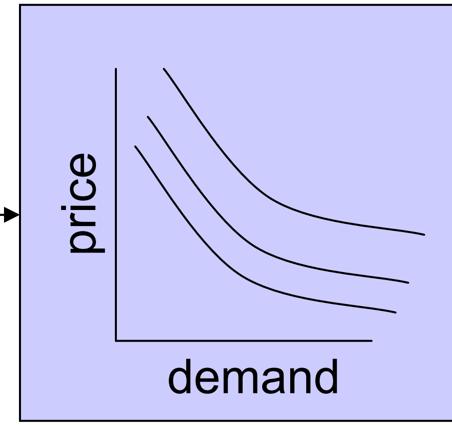
CustID	Unit Price	Order Amount
J. Smith	$10.20	500
…	…	…



Global demand distribution (prior)

Data for one customer

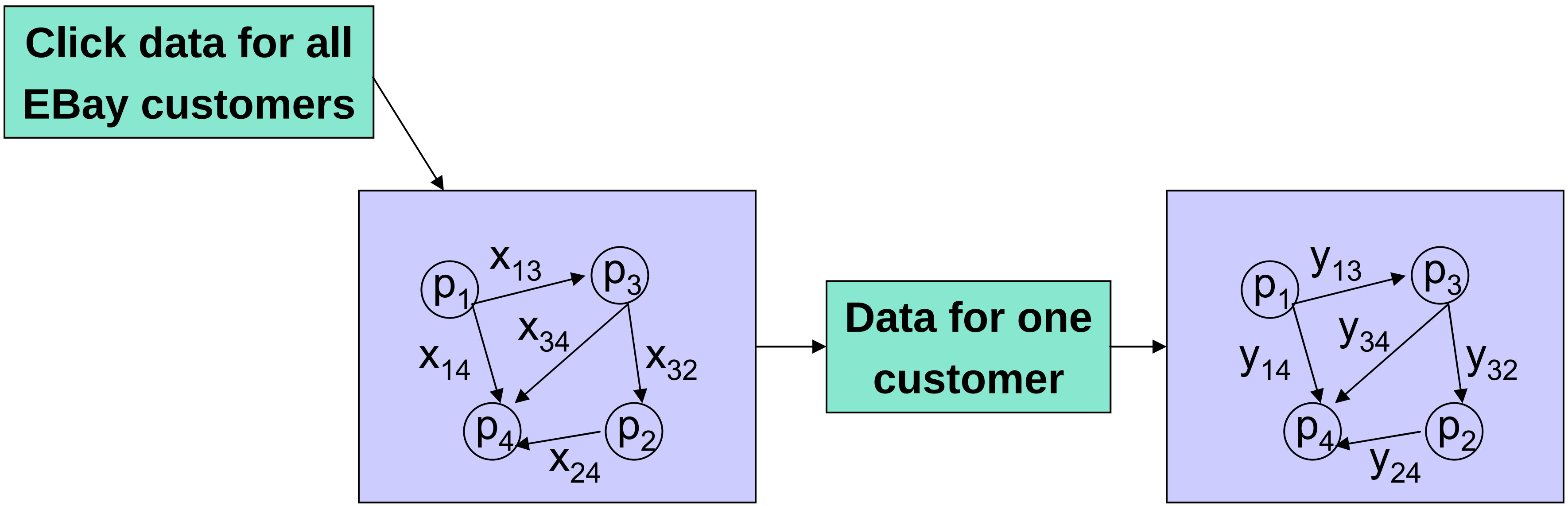
Bayes Theorem



Individual demand distribution (posterior)

- Can analyze arbitrary dynamically-defined customer segments when determining effect of price increase

Ex. 3: Individual Click Behavior (EBay)



Global Markov model distribution (Dirichelet prior)

Individual Markov model distribution (posterior)

- Can analyze arbitrary dynamic customer segments when determining effect of changing EBay pages

Ex. 4: Clinic-Capacity Risk

Medical data for all customers

Stochastic dosage model

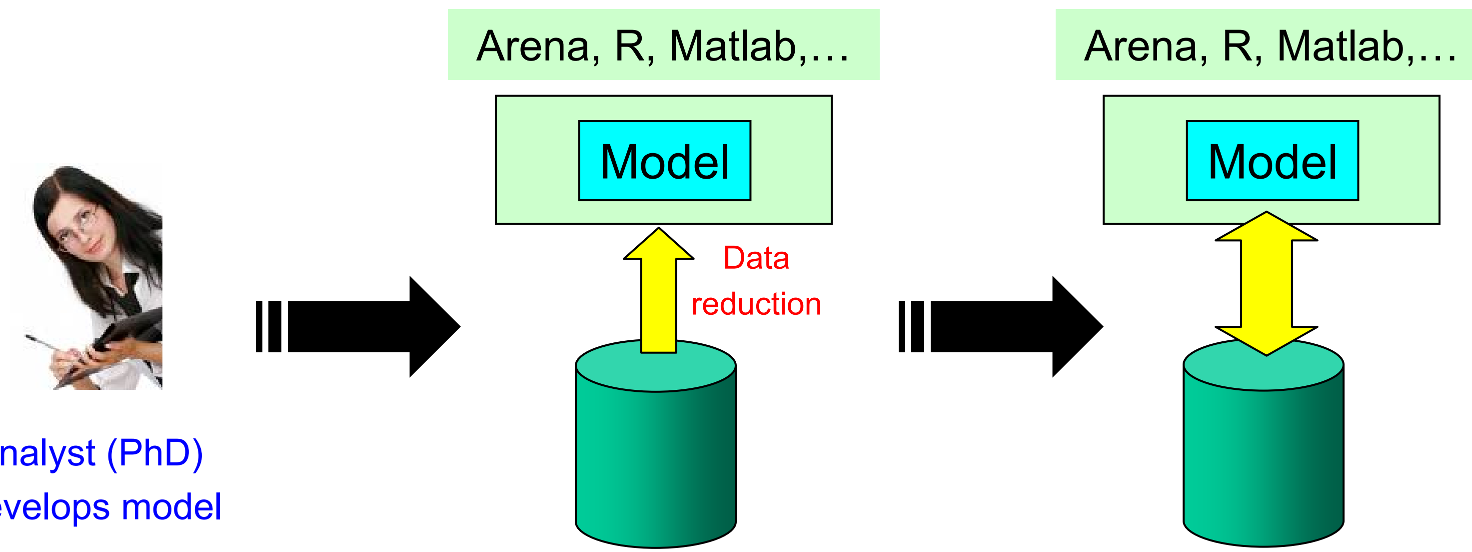
Pharmacy data for all customers

Cox hazard-rate disease model

Clinic-resource demand model

CustID	Time period	Resource needed
Jane Smith	June-Sept	?
…		…

MCDB: Improvement of Traditional Analytics Workflow



Analyst (PhD)
Develops model

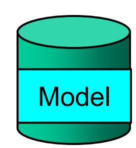
Model fitting

Model application
& querying

- Data extraction slow and bug-prone
- Only coarse-grained modeling
- No encapsulation for user

- Hard to re-link model results to DB
- Hard to deal with data updates
- Sensitivity, what-if analysis are hard

Goal: Integrate model with Database



Where do the probabilities come from?

From stochastic predictive models over big data

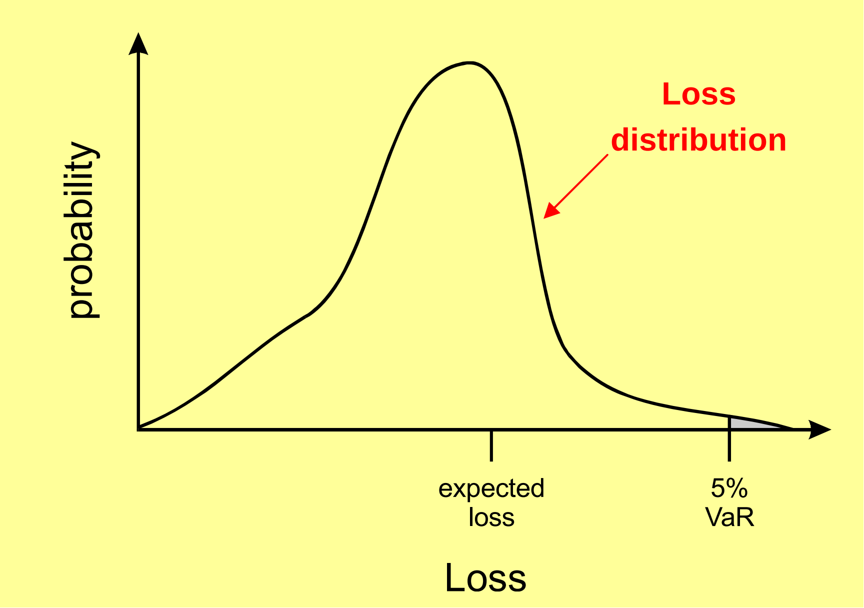
Who is going to use this stuff in the real world?

Key Driver: Risk Management

- Ex: Projected sales under micromarketing campaign
- Ex: ERP
  - # OS experts for help desk
  - Demand projected from historical text data (2x uncertainty)
  - Provide principled safety factor
- Regulatory pressure
  - Basel II, Solvency II
- Business pressure
  - Ex.: Energy Risk Professionals

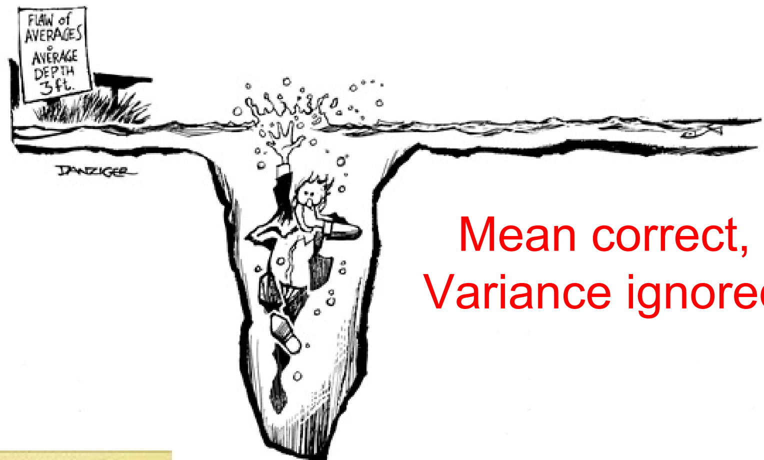
GARP Global Association of Risk Professionals

```
SELECT SUM (s.amount)
FROM SALES s, CUST c
WHERE s.ID = c.ID
   AND c.city = 'Los Angeles'
```



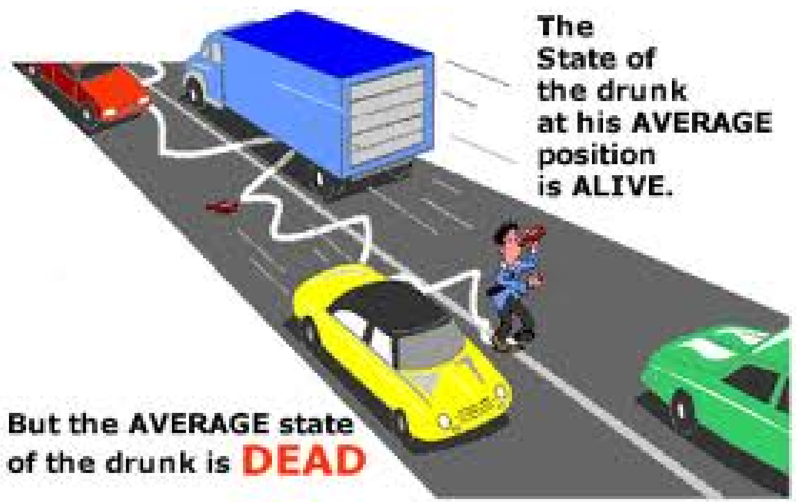
Challenge: Decision-makers' Poor Intuition About Risk

Flaw of averages (weak form):

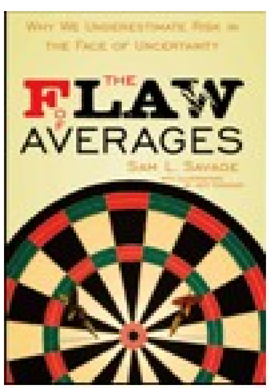


Mean correct,
Variance ignored

Flaw of averages (strong form):



Wrong value of mean:
$f(E[X]) \neq E[f(X)]$

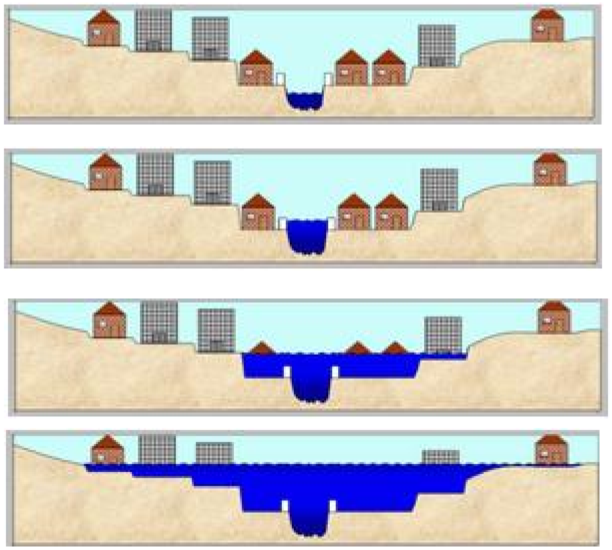


Sam Savage's book
(why we underestimate risk)

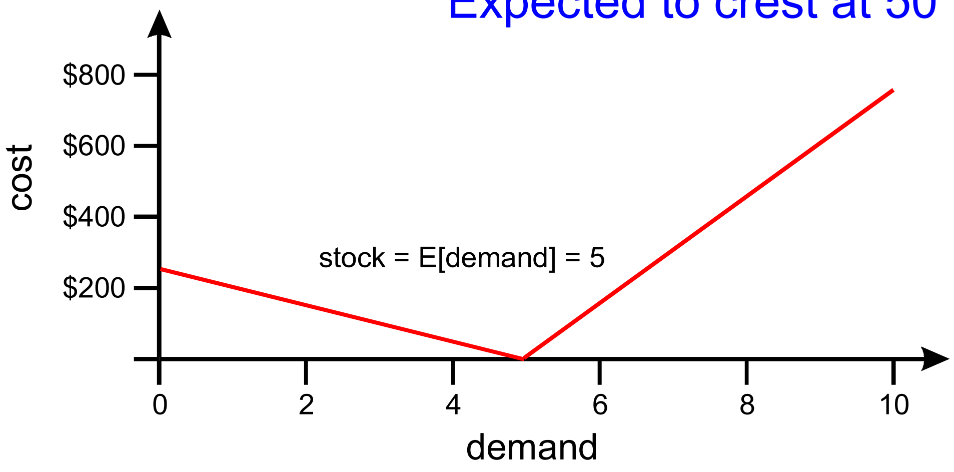
Examples

- Red River (ND) flooding
- Perishable Inventory (Red Lobster)
- U.S. accounting standards (FASB)
- Project completion time: 10 parallel tasks, $E[T_i]$ = 6 mo.
- Data cleansing
- Machine learning
- Trio agg. paper (MUD 2008)

- Basic probability

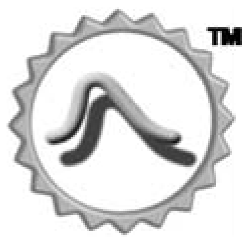


"Expected to crest at 50 feet"



Probability Management and Interactive Spreadsheets

- DIST 1.1 standard
  - DIST = distribution string
  - IID Monte Carlo (multivariate) samples
  - Compressed, with metadata
- Ensures correct, coherent risk computations throughout enterprise and beyond
  - E.g., Royal Dutch Shell
- "Electricity network" for probability
  - Royal Dutch Shell, Merck Pharmaceutical, Oracle, Wells Fargo Bank, Bessemer Trust, and IBM
- DISTs can be manipulated like numbers
  - Facilitates interactive spreadsheets (demo)



Audit seal of approval

Demo 1

BUSINESS PLAN: Requiring Capital Investment with Uncertain Demand

INVESTMENT	$1,600,000
CAPACITY	2000000
DEMAND	2507670
PRICE	$1
SALES	2000000
REVENUE	$2,000,000
PROFIT	$400,000
Average	$200,496

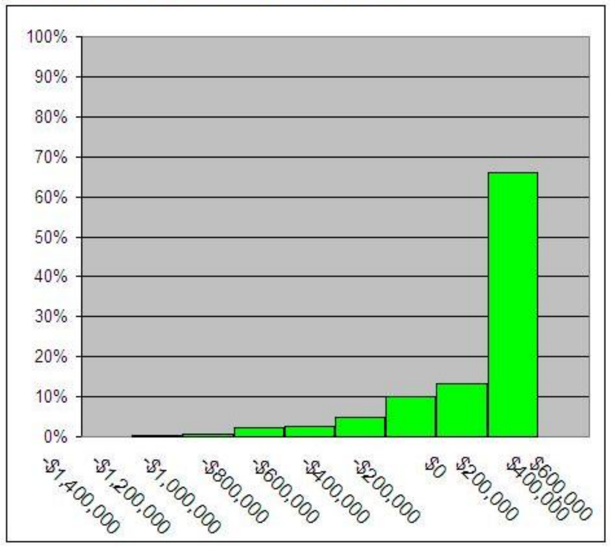
Inputs

- Decisions
- Certainties
- DIST String
- Dst(DIST)

Outputs

- Formulas

1



NOTE: You must activate the model with the Simulate Activate command or the button the the tool bar.

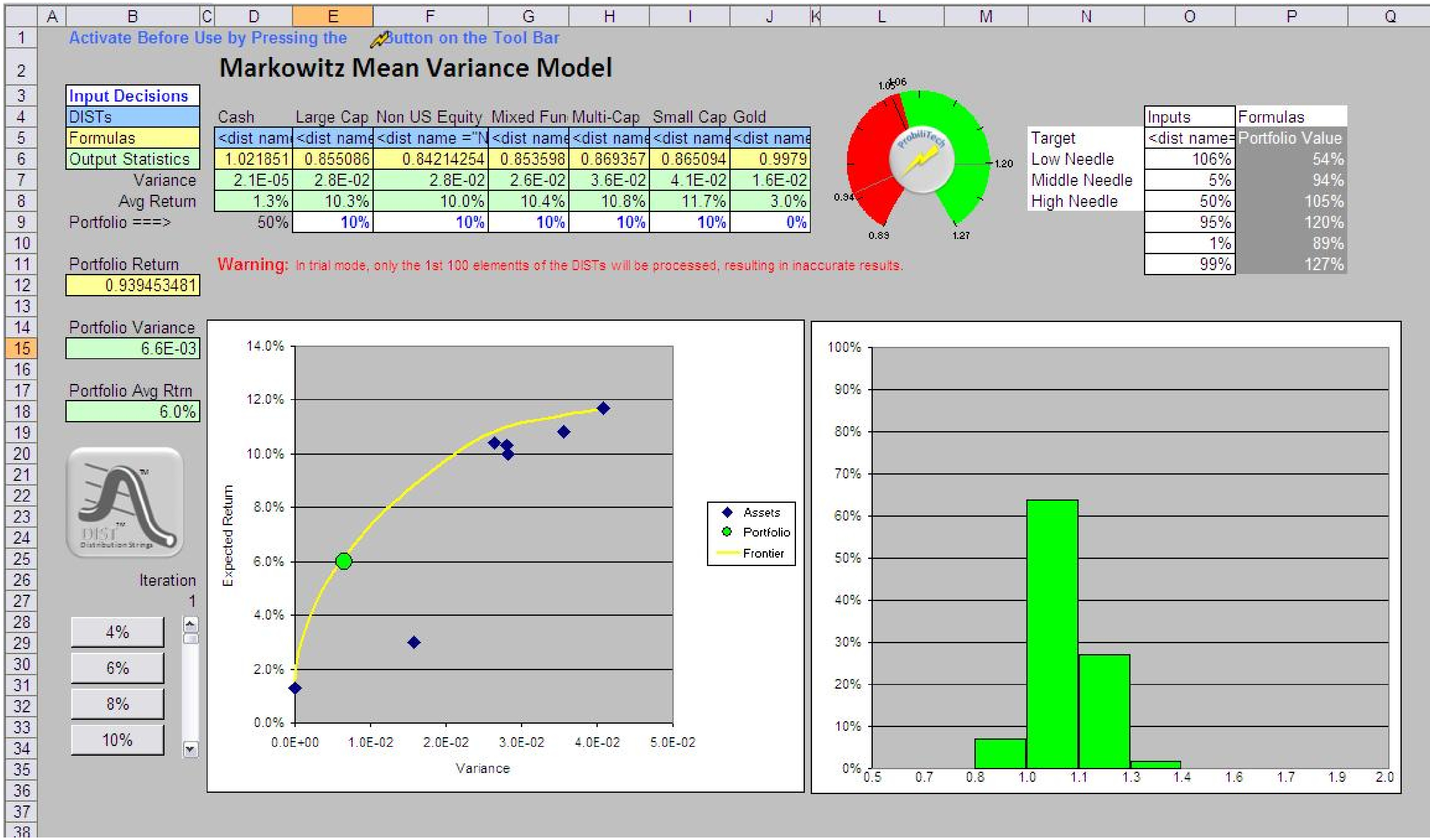
Investment in $10
16

Bins	Frequency
<dist name="	Unnamed---
-$1,400,000	0
-$1,200,000	0.003
-$1,000,000	0.007
-$800,000	0.022
-$600,000	0.027
-$400,000	0.05
-$200,000	0.1
$0	0.132
$200,000	0.659
$400,000	0
$600,000	

DEMAND DIST

<dist name="Demand DIST" avg="2.00000000E+006" min="5.03593565E+005" max="3.44361231E+006" count="1000" type="Single" origin="Normal(2m,0.5m)">rm24rMeuC3qlcW49UKSLmbV4sGamj3vbkv+8iJNrYqqOUmmDlzyEeKCQe5mvVYJZc

Demo 2


Activate Before Use by Pressing the Button on the Tool Bar
Markowitz Mean Variance Model
Input Decisions
DISTs
Formulas
Output Statistics
Cash
Large Cap
Non US Equity
Mixed Fun
Multi-Cap
Small Cap
Gold
<dist name
1.021851
0.855086
0.84214254
0.853598
0.869357
0.865094
0.9979
Variance
2.1E-05
2.8E-02
2.8E-02
2.6E-02
3.6E-02
4.1E-02
1.6E-02
Avg Return
1.3%
10.3%
10.0%
10.4%
10.8%
11.7%
3.0%
Portfolio ===>
50%
10%
10%
10%
10%
10%
0%
Portfolio Return
0.939453481
Warning: In trial mode, only the 1st 100 elementts of the DISTs will be processed, resulting in inaccurate results.
Portfolio Variance
6.6E-03
Portfolio Avg Rtrn
6.0%
Iteration
1
4%
6%
8%
10%
Inputs
Formulas
Target
<dist name=
Portfolio Value
Low Needle
106%
54%
Middle Needle
5%
94%
High Needle
50%
105%
95%
120%
1%
89%
99%
127%
Expected Return
Variance
Assets
Portfolio
Frontier

Probability Management and Probabilistic Databases

- ProbDBs can be a source of DISTs
  - Directly from MCDB
  - Can sample from
    - exact distributions
    - approximate empirical distributions
    - Fitted distributions (e.g., compute mean, var)

  for aggregation query or loss function
- Greater impact on decision-makers

Who is going to use this stuff in the real world?

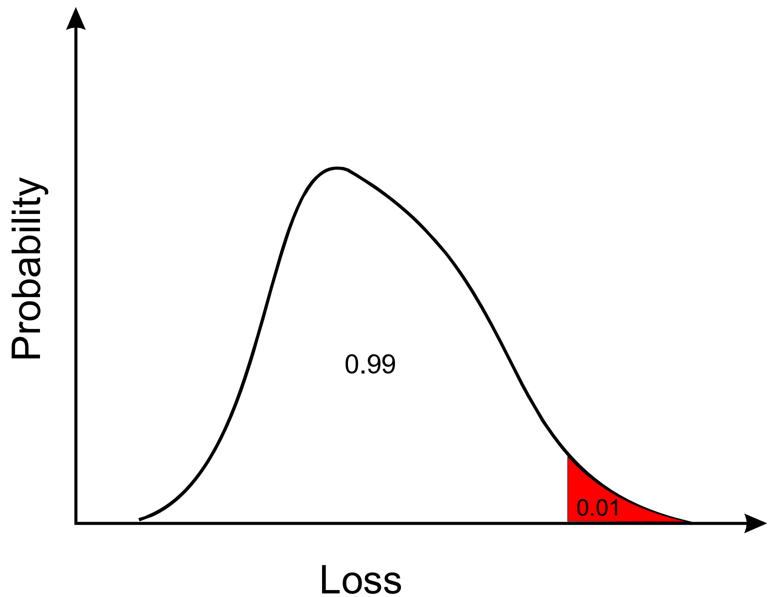
Decision-makers who care about risk
(Probability Management framework)

Risk-Orientation Leads to Interesting Research

- Ex 1: MCDB-R
- Ex 2: Risk in top-K queries

Ex. 1: MCDB-R

- Goals
  - Determine tail of query-result dist’n (e.g., 0.99-quantile = $VaR_{0.01}$)
  - Generate samples from tail*
- Challenge for naïve MCDB
  - Huge # of replications needed

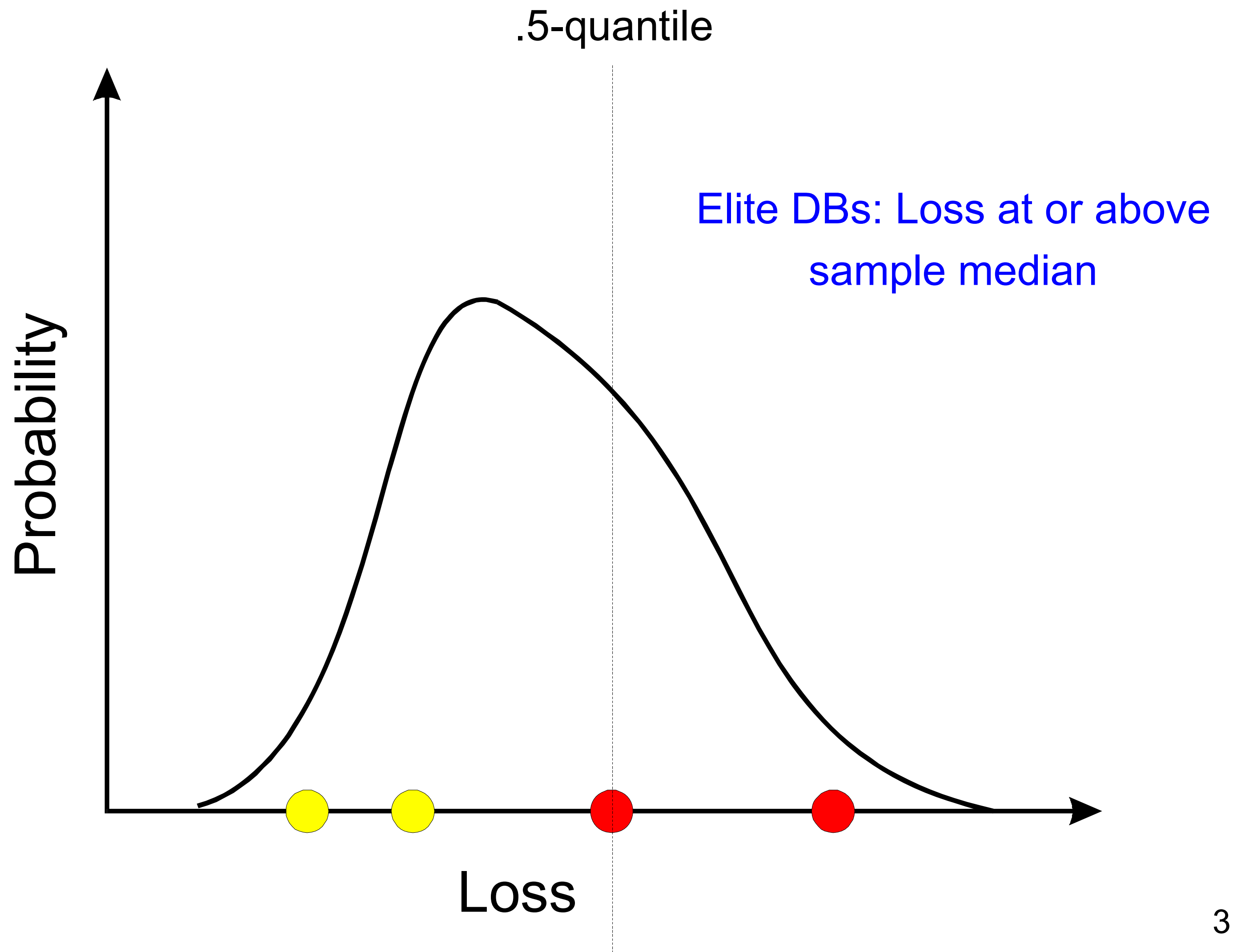


Normal($10M,$1M) loss:
On average, 3.5×10^6 reps before even one $15M loss is observed!

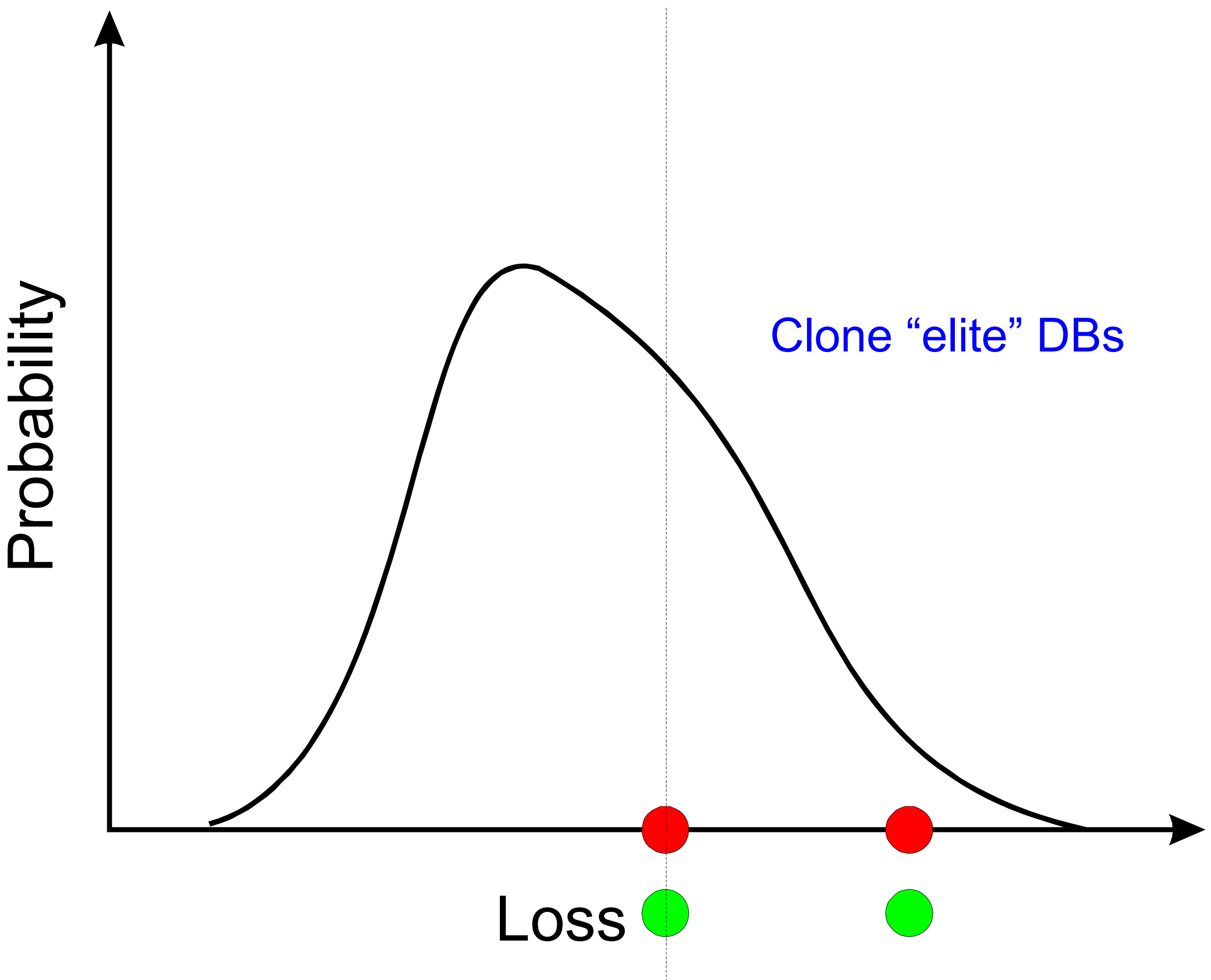
*Degen, M., Lambrigger, D.D., Segers, J.: Risk Concentration and Diversification - Second-Order Properties. *Insurance: Mathematics and Economics* 46(3), 2010

Gibbs-Cloner Approach

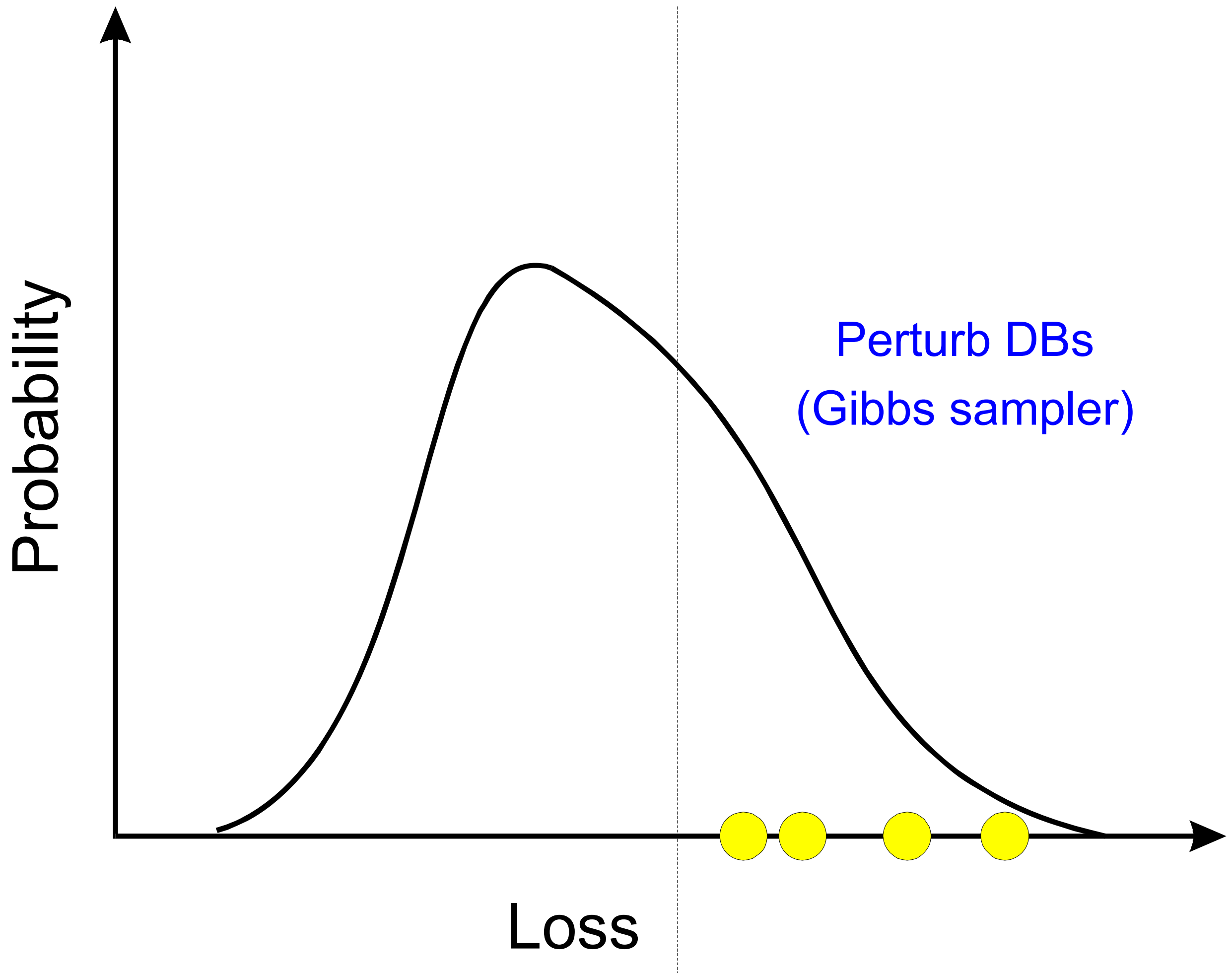
Gibbs-Cloner Approach



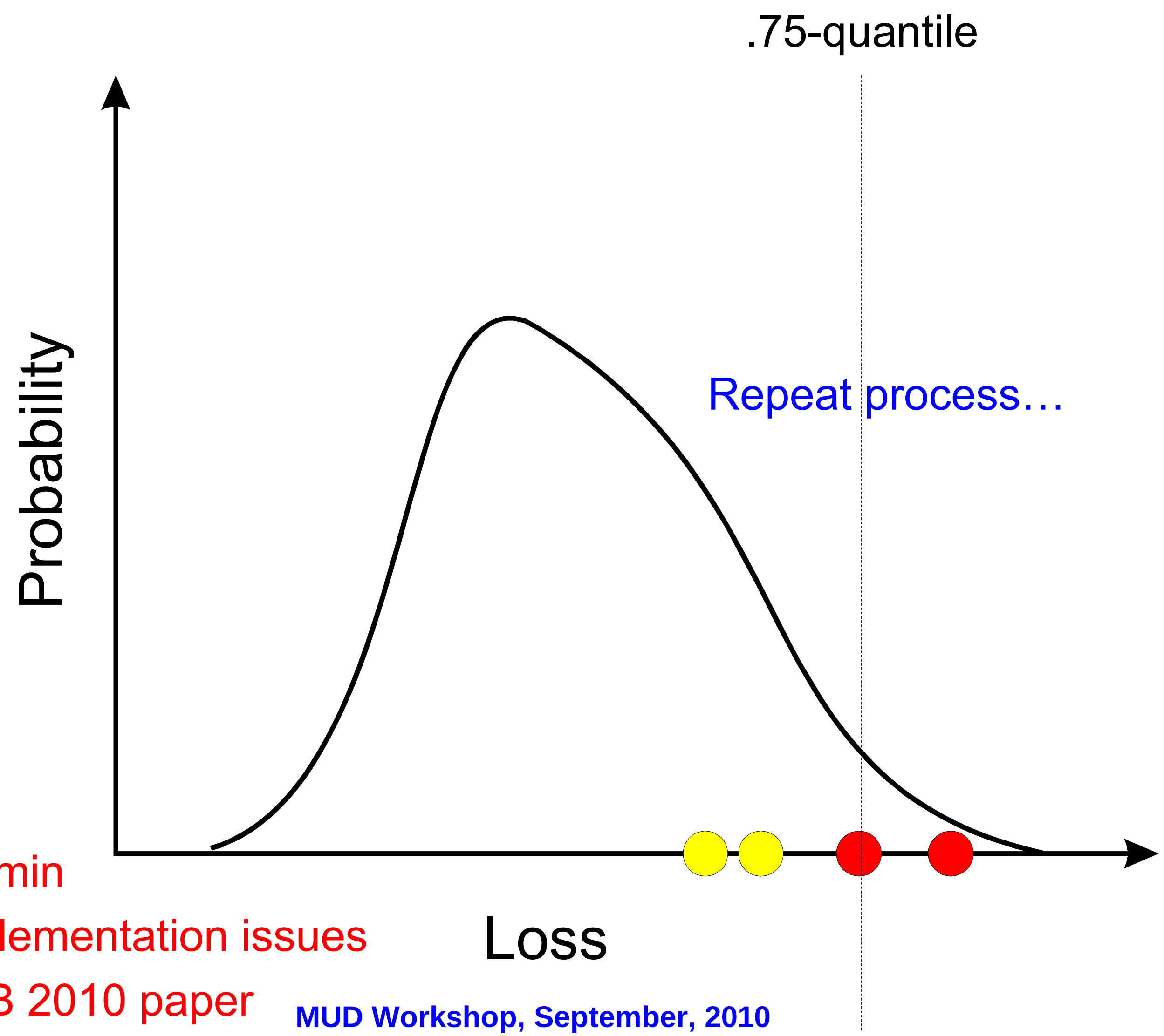
Gibbs-Cloner Approach



Gibbs-Cloner Approach



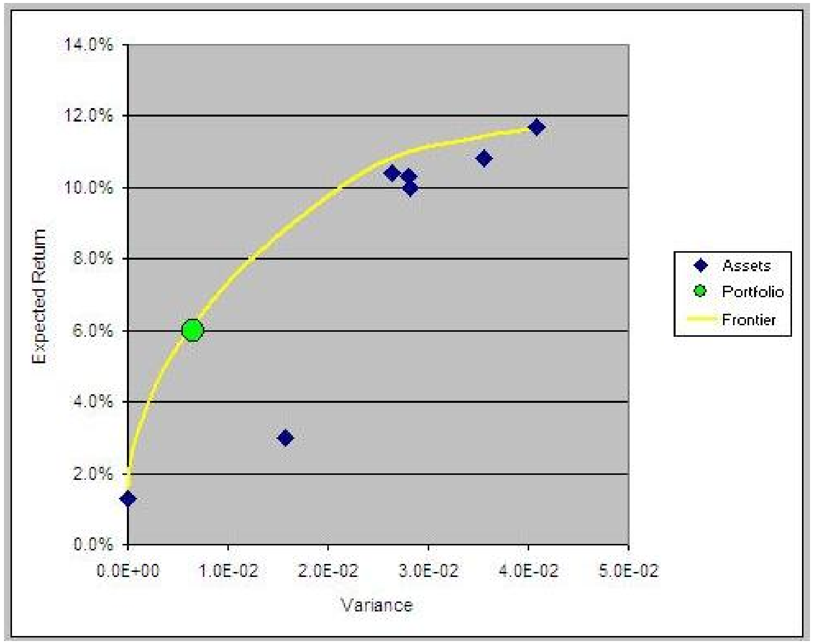
Gibbs-Cloner Approach



- 18 hrs ⇒ 11 min
- Complex implementation issues
- Details: VLDB 2010 paper

Ex. 2: Portfolio Theory of IR

- Wang and Zhu [SIGIR 2009]
  - Uncertain relevance (score)
  - Balance mean/variance of "overall relevance" of document group = $\text{sum}_i (R_i \times w_i)$
  - Diversification of documents
  - Q: Other loss functions?



Summary

- Easier to sell "stochastic predictive analytics over big data" than "data warehouse uncertainty" to real-world clients

- Risk management is a key driver in this setting but decision-makers are surprisingly clueless

- Probability-management ecosystem: a channel from ProbDBs to decision-makers?

- Risk-orientation leads to interesting research questions as well as potential impact

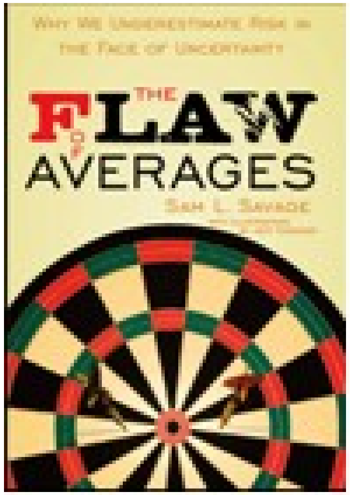
↓

Special Thanks

- Sam Savage
- Amol Deshpande
- Chris Jermaine and students
- Yannis Sismanis

Further Details:

- RAQA: ICDE 2009
- MCDB: SIGMOD 2008
- MC^3: SIGMOD 2009
- ProbIE: SIGMOD 2009
- MCDB-R: VLDB 2010



http://probabilitymanagement.org

www.almaden.ibm.com/cs/people/peterh

peterh@almaden.ibm.com

Thank You!