

Analytic DB Technology for the Data Enthusiast

**Pat Hanrahan
Stanford & Tableau**

SIGMOD Keynote 2012

My 1st Job: Analyzing Data

**University of Wisconsin
Experimental Particle Physics**

My 1st Job: Analyzing Data

Data Analysis is Spreading

Doctor:

**Why are my patients
returning to the hospital?**

Call Center Operator:

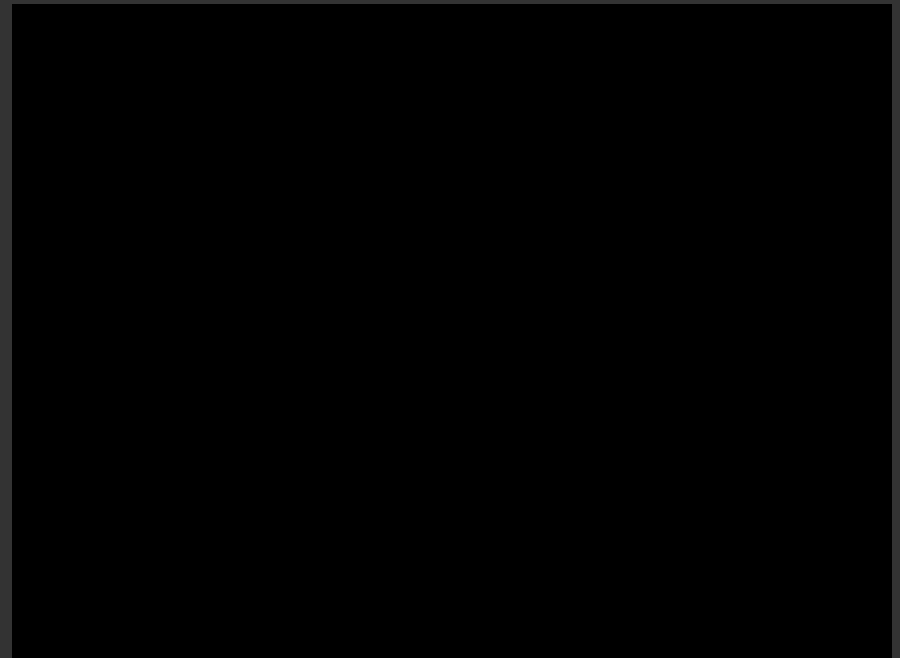
**Why does dispatching a tow
truck cost so much in ND?**

Doll Collector:

**What caused the price of
vintage Barbie dolls
to increase?**

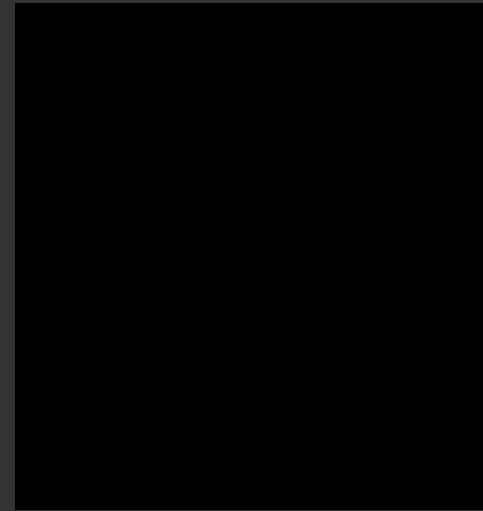
Game Producer:

**What causes players
to buy virtual goods?**



Why are databases so slow?

Data Scientist



Netflix Prize Team

Analytical Thinking?

Sherlock Holmes

"It is a capital mistake to theorize before one has data. Insensibly one begins to twist facts to suit theories, instead of theories to suit facts."

"Data, Data, Data! I can't make bricks without clay."

Definition: Analytical Thinking

**“A structured approach
to answering questions
and making decisions
based on facts and data”**

My Process

Pose the question

Find or collect the appropriate data

Check and verify

Clean and normalize

Contextualize the data by joining with other data

Explore relationships & patterns in the raw data

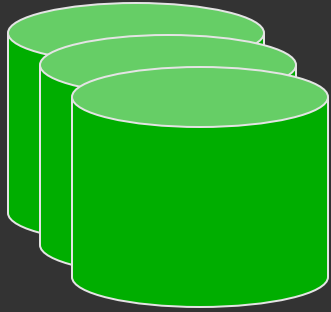
Generalize and summarize

Confirm hypotheses and analyze errors

Share findings with others

Decide and act

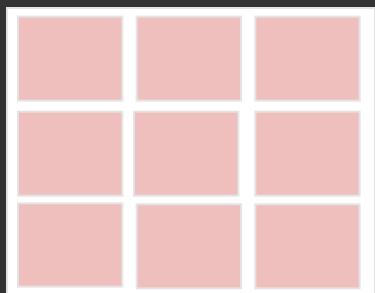
Question



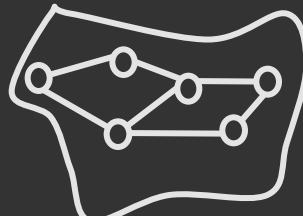
Forage
for data



Check and clean



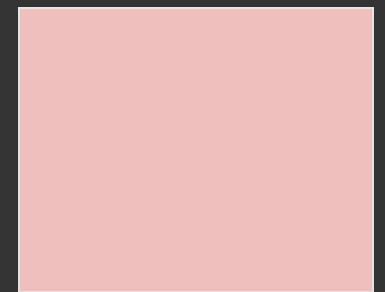
Show relationships and patterns
using visual representations

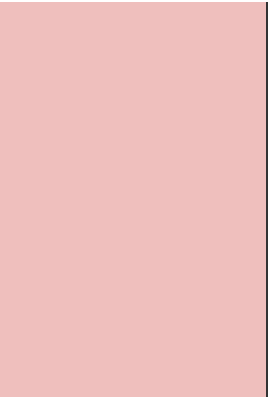


Decide and act



Test hypothesis,
analyze errors,
discover insight





**“Data Analysis is like
doing Experiments,” J. Tukey**

Experiments

- 1. Theorize and hypothesize**
- 2. Experiment**
- 3. Revise theory**
- 4. The craft occupies the experimenter allowing time to think**

Data Analysis

- 1. Theorize and hypothesize**
- 2. Find trends and relationships**
- 3. Find limitations of the model**
- 4. Provide insight to improve the model**

“State of the Art”

Spreadsheets

“State of the Art”

Crosstabs and Pivot Tables

Idea: Visual Analysis

**“Analytical Reasoning
Facilitated by
Interactive Visualization”**

Polaris / Tableau Demo

C. Stolte's PhD Thesis

1. Best Visualization Depends on the Question/Task

How much mint tea was sold in the west?

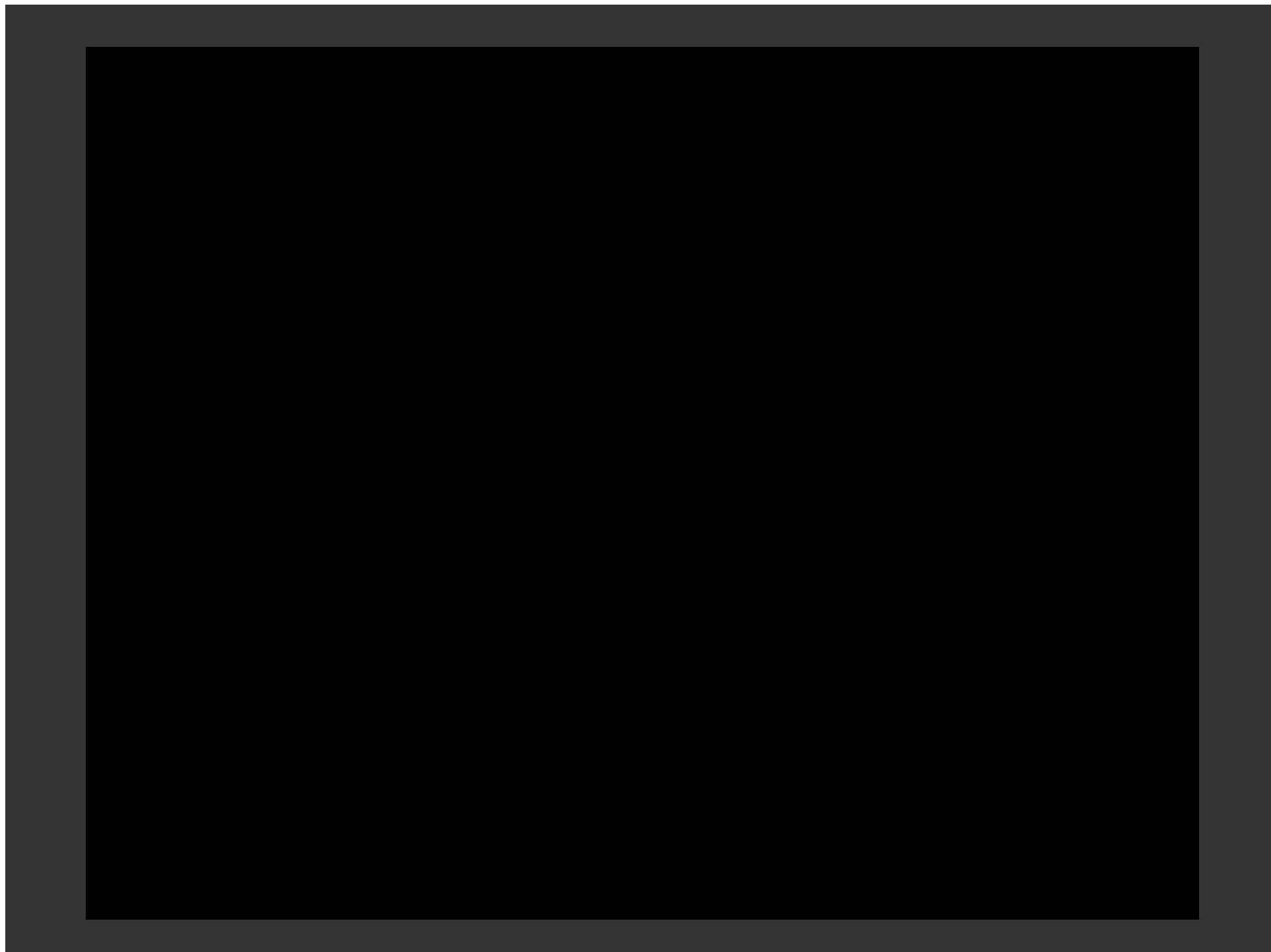
What product in what region sold the most?

What product in what region sold the most?

2. Formulate Any Query



Query-By-Example [Zloof, 1975]



SELECT AS CIRCLE

Candidate * Longitude ON X

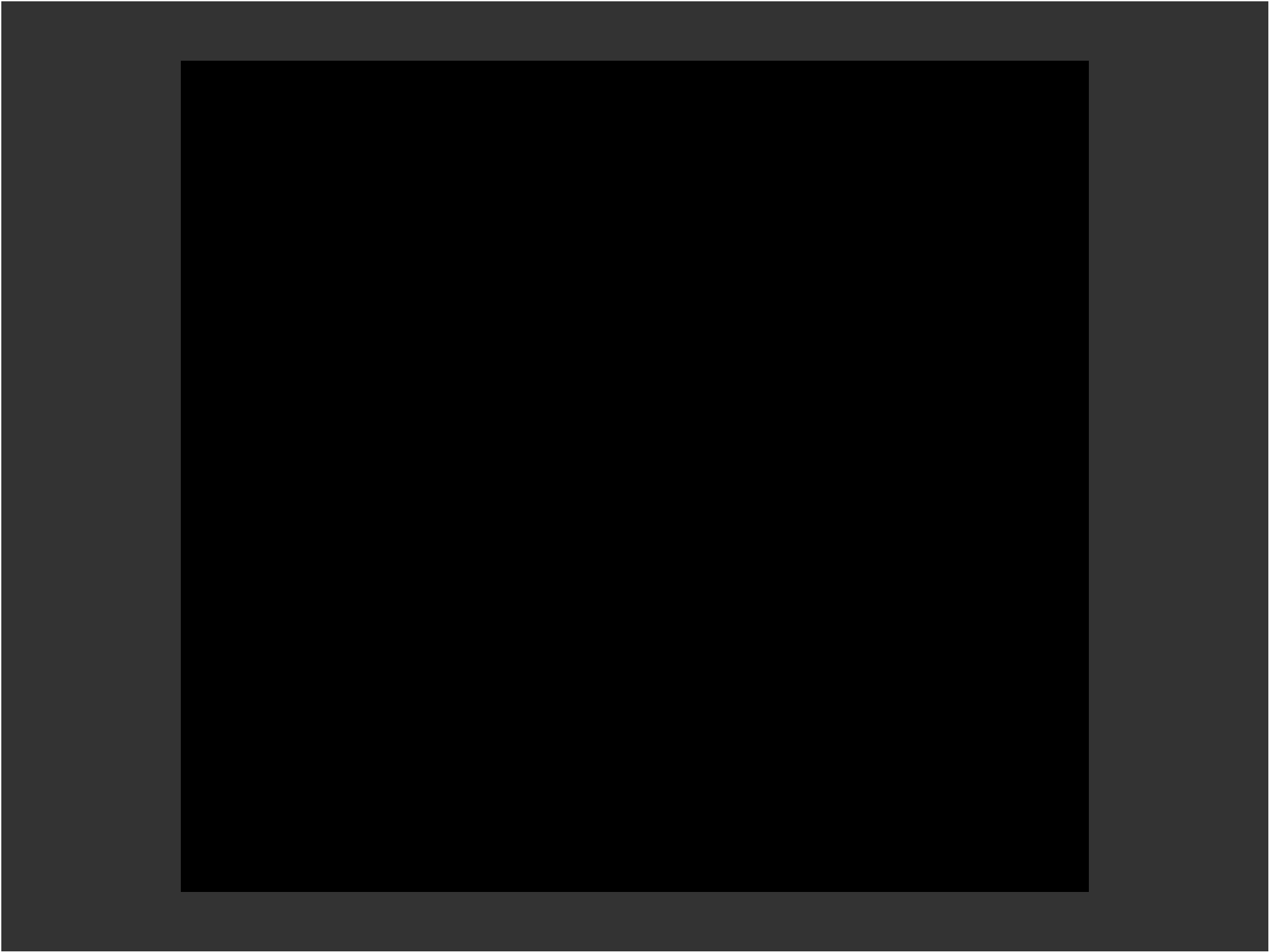
Latitude ON Y

Zipcode IN PANES

Party ON COLOR

Sum(Amount) ON SIZE

FROM ContributionsDatabase



SELECT AS SHAPE

Market * Sales ON COLS

Quarter * Profit ON ROWS

State * Product IN PANES

ProductType ON COLOR

Year ON SHAPE

FROM RetailDatabase

Four Main Ideas

- 1. Support cycle of analysis**
- 2. Answer a question by composing a picture**
- 3. Best visualization depends on question/task**
- 4. Must be able to generate any query**

+ Easy to use

Analysis at the Speed of Thought

Transactional Databases are Slow!!

TPC-H, 1 GB, Query 1*:

C Program	0.2 s
mysql	26.2 s
DBMS “X”	28.4 s

*Boncz et al., CIDR 2005

In-Memory Column Stores: 100x

Columns are efficient (MonetDB/X100, C-store, ...)

- A Only access needed columns**
- A Well-matched to processor+memory architecture**
- A Columns compress better than records**
- A Optimized for read/append**
- A Vector semantics instead of set semantics**

In-Memory reduces latency enabling interaction

- A Memory is cheap, memory hierarchy is expending**
- A Median business database fits in memory**

kx.com

2 B trades per day

STAC-M3

~10-20 msec latency

High Frequency Trading

Fully Utilize *All* Resources

Intel Ivy Bridge Processor (Core i7 3770K)

22 nm, 1.4B 3D transistors

4 3.5 Ghz cores

256-bit AVX vector instructions (16 FADD/FMUL)

Resource limits

Bandwidth limited: 2 DDR3 2133 = 34 GB/s

Compute limited: $4 * 3.5 \text{ Ghz} * 16 = 224 \text{ GFLOPS}$

Theoretical: 1B values can be summed in 125/5 msecs

2018 Laptop ~ CPU+GPU

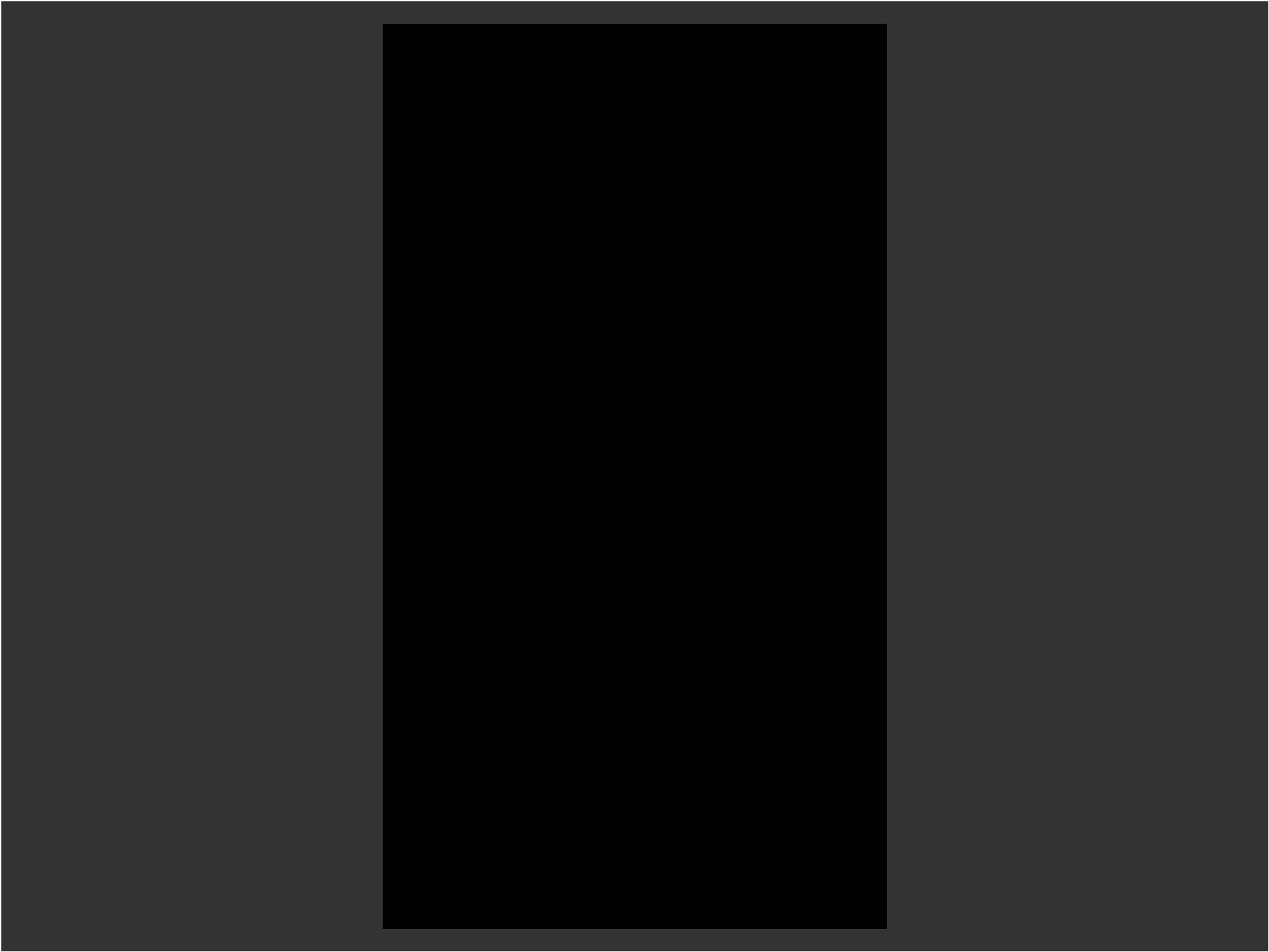
10 Teraflops

Supporting Data Enthusiasts

“Although we often hear that data speak for themselves, their voices can be soft and sly.

We need statistics to help them tell their story”

**Beginning Statistics with Data Analysis
Mosteller, Fienberg, Rourke**



Significant

Not Significant

Data Integration

Provides context for analysis

Semantic integration => people

Promising tools

- A Potters wheel**
- A Google fusion tables**
- A Data wrangler**
- A Data blending**

Dynamic Workload Driven Data Integration in Tableau

K. Morton, R. Bunker, J. Mackinlay, R. Morton, C. Stolte

Wrap Up

Summary

Large number of data enthusiasts

- ⌚ **Business users, with the questions, on a mission**
- ⌚ **Excellent analytical thinkers**
- ⌚ **Not DBAs, not programmers, not statisticians**

You can help them

- ⌚ **Current tools support only basic visual analysis**
- ⌚ **... not the entire process of analysis in the large**

Thank You