

**Analytic DB Technology  
for the  
Data Enthusiast**

**Pat Hanrahan  
Stanford & Tableau**

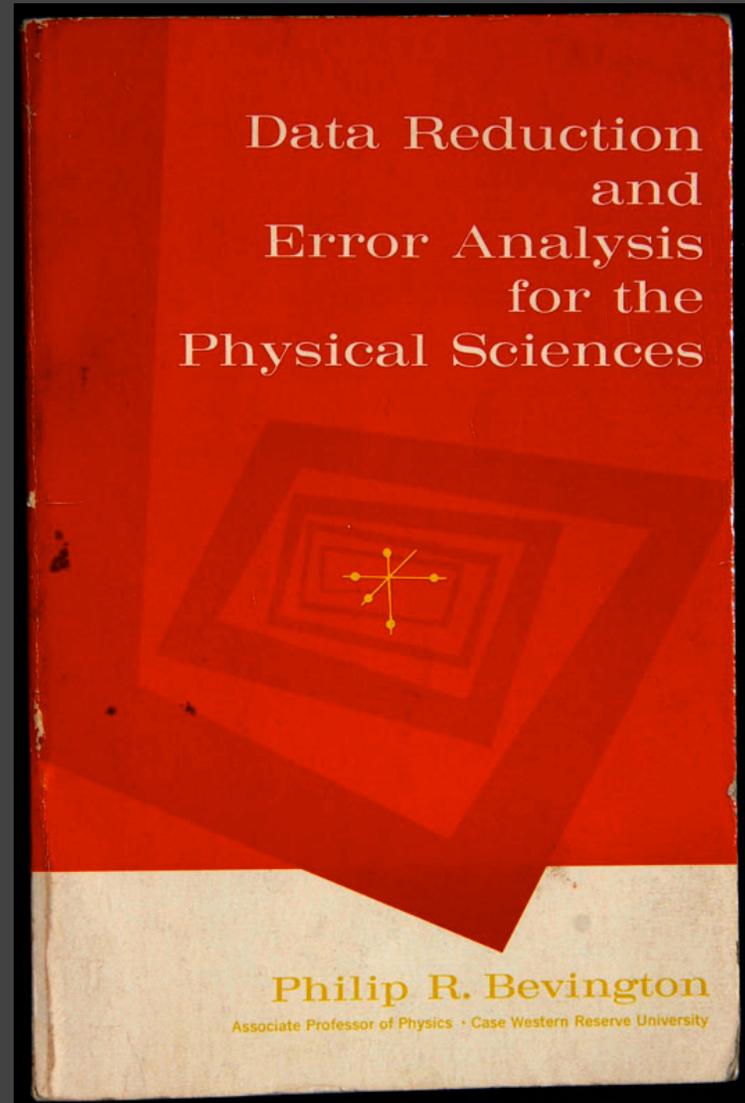
**SIGMOD Keynote 2012**

# My 1<sup>st</sup> Job: Analyzing Data



University of Wisconsin  
Experimental Particle Physics

# My 1<sup>st</sup> 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?**



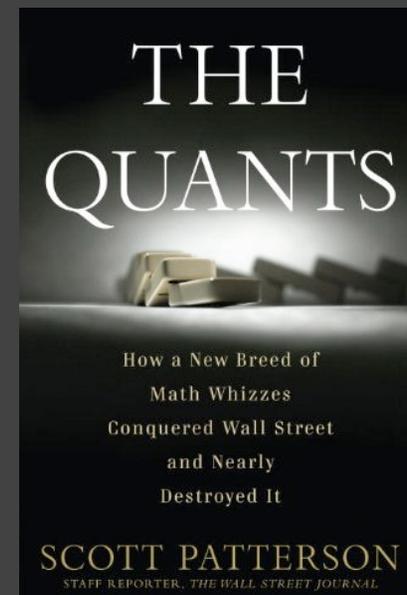
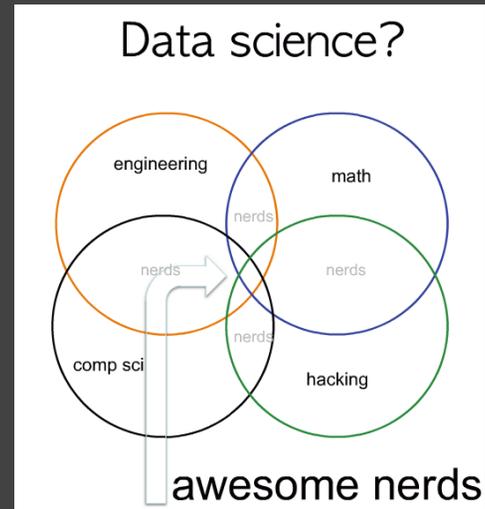


**Why are databases so slow?**

# Data Scientist



Netflix Prize Team



**Analytical Thinking?**



"HOLMES GAVE ME A SKETCH OF THE EVENTS."

# 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**

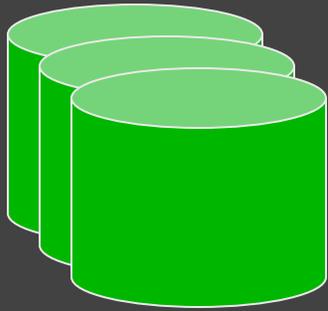
**Generalize and summarize**

**Confirm hypotheses and analyze errors**

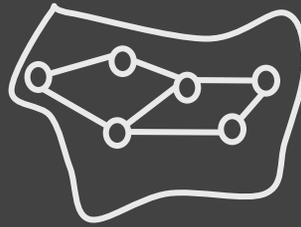
**Share findings with others**

**Decide and act**

# Question



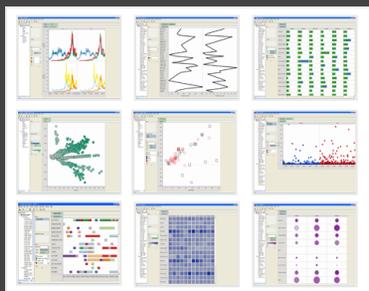
Forage  
for data



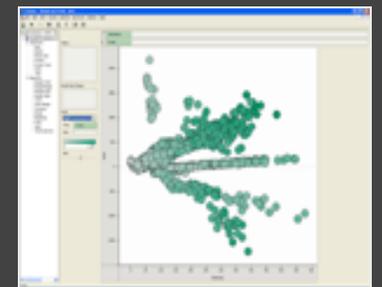
Decide and act

Check and clean

Test hypothesis,  
analyze errors,  
discover insight



Show relationships and patterns  
using visual representations





**“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”

The screenshot shows a Microsoft Excel window titled "Grades\_Access\_final.xlsx". The ribbon includes Home, Insert, Page Layout, Formulas, Data, Review, View, and Acrobat. The formula bar shows the formula `=AVERAGE(K6:K15)` in cell K16. The spreadsheet data is as follows:

Student No	Chap 1	Chap 2	Chap 3	Midterm	Chap 4	Chap 5	Final	Project	Activity	Total	Pe
Alyson Smith	40	40	20	186	40	20	259	55	85	745	
Jason Rost	40	40	40	264	35	40	253	60	90	862	
Jessica Kevin	40	20	36	158	40	35	238	50	90	707	
Joe Reboot	40	35	40	236	35	40	232	5	85	748	
Micah Khan	40	40	40	232	40	20	223	65	90	790	
Paulee Manson	40	40	0	196	40	20	301	5	85	727	
Scott Nelson	40	40	40	270	40	40	265	50	95	880	
Stephen Diaz	40	40	40	256	40	40	232	70	95	853	
Tamey Schaffer	40	40	32	225	35	35	285	65	90	847	
Average	40	37.222	32	224.77778	38.333	32.222	254.22222	47.222	89.444	795.444444	0

Row 18 contains the text "This is not the final version". The status bar at the bottom shows "Ready" and "100%".

## Spreadsheets

# “State of the Art”



## Title

Server Hostname: itm64vm13.tivlab.raleigh.ibm.com

AVG_Used_CPU_MHz		Feb 14, 2011	Feb 15, 2011	Feb 16, 2011	Feb 17, 2011	Feb 18, 2011	Feb 19, 2011	Feb 20, 2011	Feb 21, 2011
vsvdash1	itm64vm13.tivlab.raleigh.ibm.com	182.73	169.17	166.52666667	163.02	161.25	154.7	152.34	
	<b>vsvdash1</b>	<b>182.73</b>	<b>169.17</b>	<b>166.52666667</b>	<b>163.02</b>	<b>161.25</b>	<b>154.7</b>	<b>152.34</b>	
vsvdash2	itm64vm13.tivlab.raleigh.ibm.com	124.45	109.04333333	155.74333333	117.35	114.705	110.3375	109.87666667	
	<b>vsvdash2</b>	<b>124.45</b>	<b>109.04333333</b>	<b>155.74333333</b>	<b>117.35</b>	<b>114.705</b>	<b>110.3375</b>	<b>109.87666667</b>	
vsvdash3	itm64vm13.tivlab.raleigh.ibm.com	122.98	111.82	226.14333333	120.425	112.77	107.215	110.56666667	
	<b>vsvdash3</b>	<b>122.98</b>	<b>111.82</b>	<b>226.14333333</b>	<b>120.425</b>	<b>112.77</b>	<b>107.215</b>	<b>110.56666667</b>	
vsytaddm	itm64vm13.tivlab.raleigh.ibm.com	123.69	123.195	122.2	121.255	121.965	125.1475	123.55	
	<b>vsytaddm</b>	<b>123.69</b>	<b>123.195</b>	<b>122.2</b>	<b>121.255</b>	<b>121.965</b>	<b>125.1475</b>	<b>123.55</b>	
win2008vm1	itm64vm13.tivlab.raleigh.ibm.com	2,654.675	2,653.92666667	2,655	2,653.35	2,654.66	2,654.465	2,655.15666667	2,655.0
	<b>win2008vm1</b>	<b>2,654.675</b>	<b>2,653.92666667</b>	<b>2,655</b>	<b>2,653.35</b>	<b>2,654.66</b>	<b>2,654.465</b>	<b>2,655.15666667</b>	<b>2,655.0</b>
<b>Summary</b>		<b>641.705</b>	<b>669.87642857</b>	<b>703.90285714</b>	<b>635.08</b>	<b>633.07</b>	<b>630.373</b>	<b>666.49428571</b>	<b>671.5</b>

## Crosstabs and Pivot Tables

**Idea: Visual Analysis**

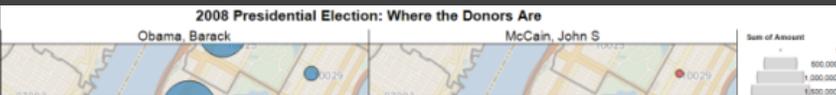
**“Analytical Reasoning**

**Facilitated by**

**Interactive Visualization”**

**Polaris / Tableau Demo**

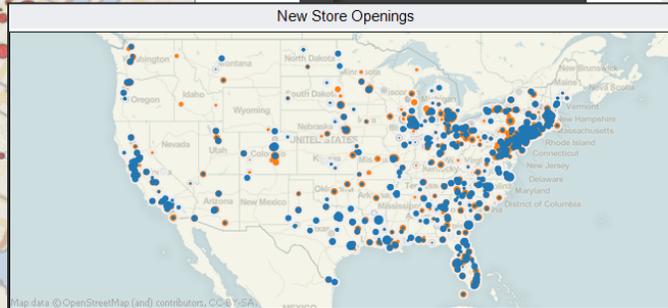
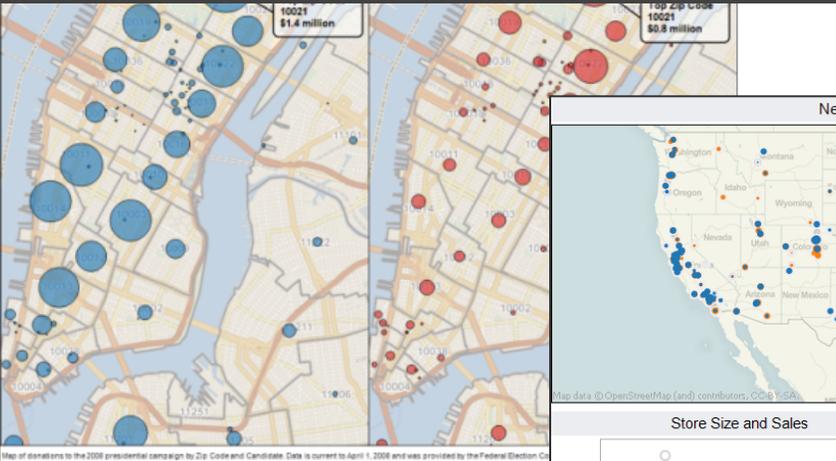
**C. Stolte's PhD Thesis**



### Spending by State

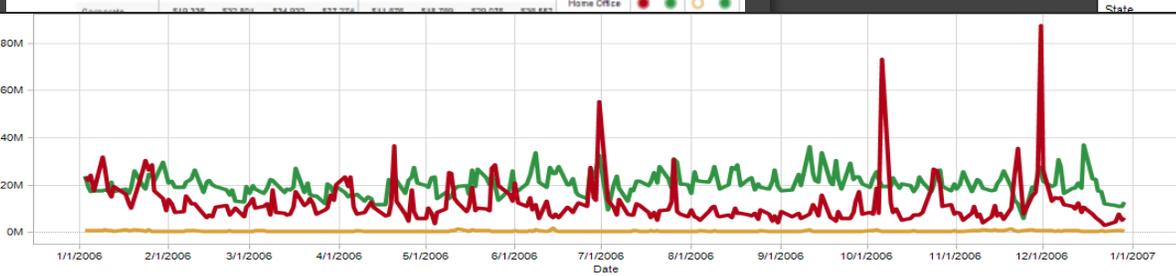
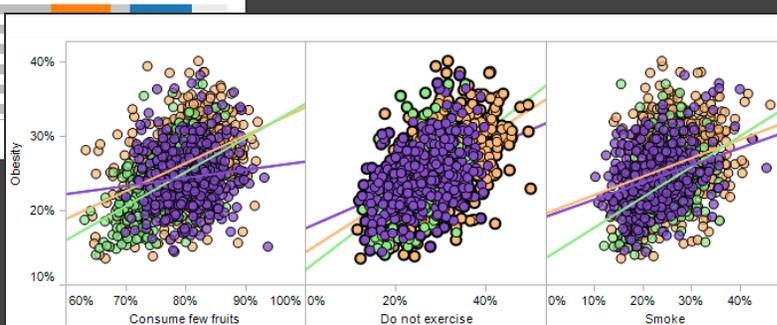
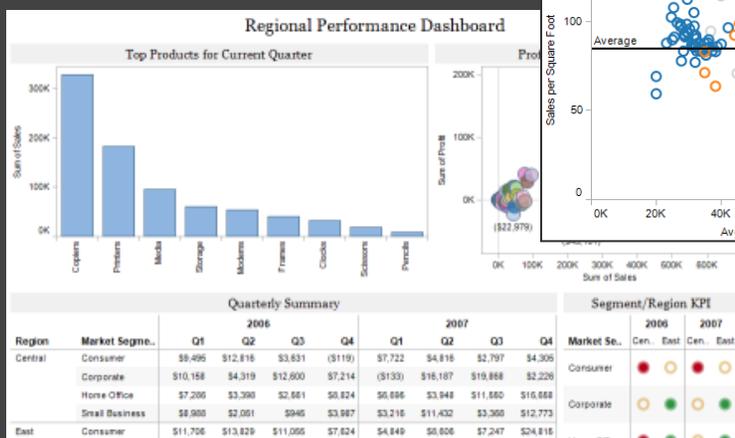
State	Department	1	2	3	4	Filter Spending
Florida	Health	\$253,821M	\$278,887M	\$303,060M	\$319,788M	\$816M \$934,426M

# 1. Best Visualization Depends on the Question/Task



Category	2006	2007	2008	2009
Veterans	\$35,461M	\$37,377M	\$37,748M	\$36,896M
Education	\$30,109M	\$24,557M	\$31,205M	\$29,727M
Transport	\$30,958M	\$33,563M	\$35,132M	\$35,144M
All Other	\$76,885M	\$100,937M	\$82,518M	\$91,623M
Apparel Special	\$3,389M	\$1,461,763M	\$1,513,753M	\$1,560,486M
Category Special	\$5,885M	\$465,326M	\$504,922M	\$542,982M
Department Special	\$1,257M	\$442,010M	\$461,400M	\$481,200M
Department Store	\$1,185M	\$331,845M	\$388,686M	\$437,034M
Home & Hardware	\$3,268M	\$371,187M	\$368,256M	\$375,844M
Miscellaneous	\$9,707M	\$64,686M	\$69,563M	\$56,687M
Supermarket	\$9,071M	\$68,622M	\$72,737M	\$71,560M
Value market	\$5,012M	\$50,868M	\$56,921M	\$59,554M
	\$5,323M	\$46,373M	\$57,145M	\$62,780M
	\$9,231M	\$56,252M	\$50,764M	\$54,879M
	\$9,767M	\$113,738M	\$129,512M	\$150,333M
	\$2,906M	\$2,010,907M	\$2,159,906M	\$2,292,853M
	\$2,295M	\$3,472,670M	\$3,675,659M	\$3,853,339M

Source: Office of Management and Budget



State	County	Obesity	Consume few fruits or veg.	Persons who do not exercise	Smokers
Texas	Dallas	37%	86%	32%	20%
	Escambia	34%	87%	34%	26%
	Talladega	33%	79%	34%	24%
	Chilton	31%	83%	30%	30%
	Walker	31%	79%	35%	34%
	Colbert	30%	77%	30%	22%
	Houston	30%	78%	33%	25%
	Jackson	29%	82%	34%	37%
	St. Clair	28%	80%	28%	29%
	Marshall	28%	82%	35%	32%

### Summary of Financial Performance

		Central		East		South		West	
		Profit	Sales	Profit	Sales	Profit	Sales	Profit	Sales
Coffee	Amaretto	\$5,104	\$14,012	\$1,010	\$2,994			-\$1,224	\$9,263
	Columbian	\$8,525	\$28,911	\$27,256	\$47,385	\$8,767	\$21,663	\$11,256	\$30,352
	Decaf Irish Cream	\$9,635	\$26,157	\$2,726	\$6,262	\$2,935	\$11,596	-\$1,307	\$18,233
Espresso	Caffe Latte					\$3,873	\$15,443	\$7,502	\$20,456
	Caffe Mocha	\$14,642	\$35,218	-\$6,232	\$16,646	\$5,202	\$14,166	\$4,066	\$18,874
	Decaf Espresso	\$8,859	\$24,483	\$2,411	\$7,720	\$5,930	\$15,381	\$12,302	\$30,578
	Regular Espresso			\$10,065	\$24,031				
Herbal Tea	Chamomile	\$14,435	\$36,571	\$764	\$2,193	\$3,178	\$11,183	\$8,854	\$25,631
	Lemon	\$6,253	\$21,982	\$7,902	\$27,177	\$2,593	\$14,494	\$13,121	\$32,273
	Mint	\$4,069	\$9,335	-\$2,243	\$11,991			\$4,328	\$14,384
Tea	Darjeeling	\$10,769	\$30,284	\$6,500	\$14,094			\$11,784	\$28,773
	Earl Grey	\$10,334	\$32,883	\$3,404	\$6,507			\$10,426	\$27,382
	Green Tea	\$1,227	\$5,209	\$5,654	\$11,576			-\$7,112	\$16,065

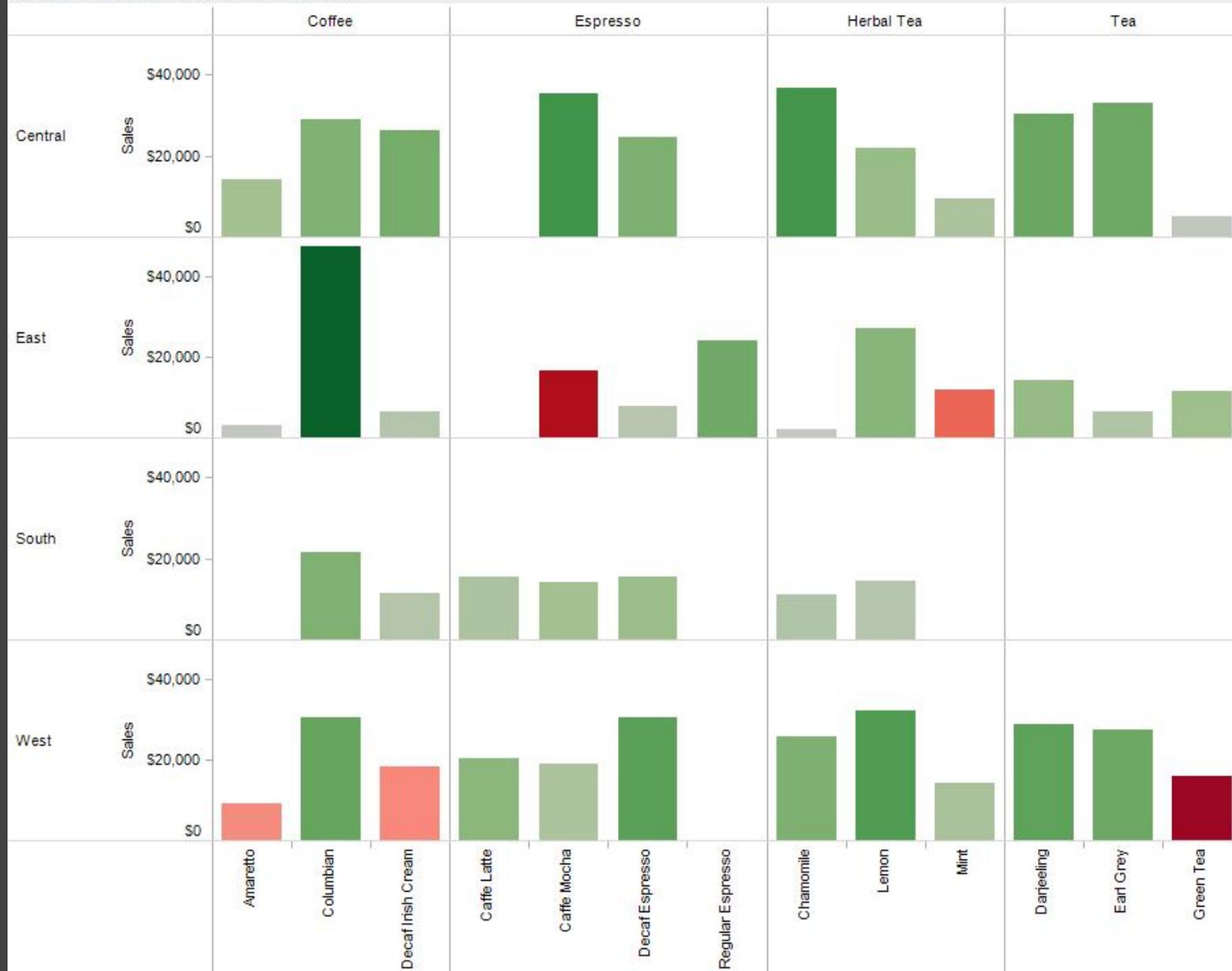
**How much mint tea was sold in the west?**

### Summary of Financial Performance

		Central		East		South		West	
		Profit	Sales	Profit	Sales	Profit	Sales	Profit	Sales
Coffee	Amaretto	\$5,104	\$14,012	\$1,010	\$2,994			-\$1,224	\$9,263
	Columbian	\$8,525	\$28,911	\$27,256	\$47,385	\$8,767	\$21,663	\$11,256	\$30,352
	Decaf Irish Cream	\$9,635	\$26,157	\$2,726	\$6,262	\$2,935	\$11,596	-\$1,307	\$18,233
Espresso	Caffe Latte					\$3,873	\$15,443	\$7,502	\$20,456
	Caffe Mocha	\$14,642	\$35,218	-\$6,232	\$16,646	\$5,202	\$14,166	\$4,066	\$18,874
	Decaf Espresso	\$8,859	\$24,483	\$2,411	\$7,720	\$5,930	\$15,381	\$12,302	\$30,578
	Regular Espresso			\$10,065	\$24,031				
Herbal Tea	Chamomile	\$14,435	\$36,571	\$764	\$2,193	\$3,178	\$11,183	\$8,854	\$25,631
	Lemon	\$6,253	\$21,982	\$7,902	\$27,177	\$2,593	\$14,494	\$13,121	\$32,273
	Mint	\$4,069	\$9,335	-\$2,243	\$11,991			\$4,328	\$14,384
Tea	Darjeeling	\$10,769	\$30,284	\$6,500	\$14,094			\$11,784	\$28,773
	Earl Grey	\$10,334	\$32,883	\$3,404	\$6,507			\$10,426	\$27,382
	Green Tea	\$1,227	\$5,209	\$5,654	\$11,576			-\$7,112	\$16,065

**What product in what region sold the most?**

## Summary of Financial Performance



**What product in what region sold the most?**

## 2. Formulate Any Query

Q2. Find the department(s) that sells an item(s) supplied by the supplier Parker.

Here the user fills in both the SALES and the SUPPLY Tables as follows.

SALES	DEPT	ITEM
	<u>P.TOY</u>	<u>PEN</u>

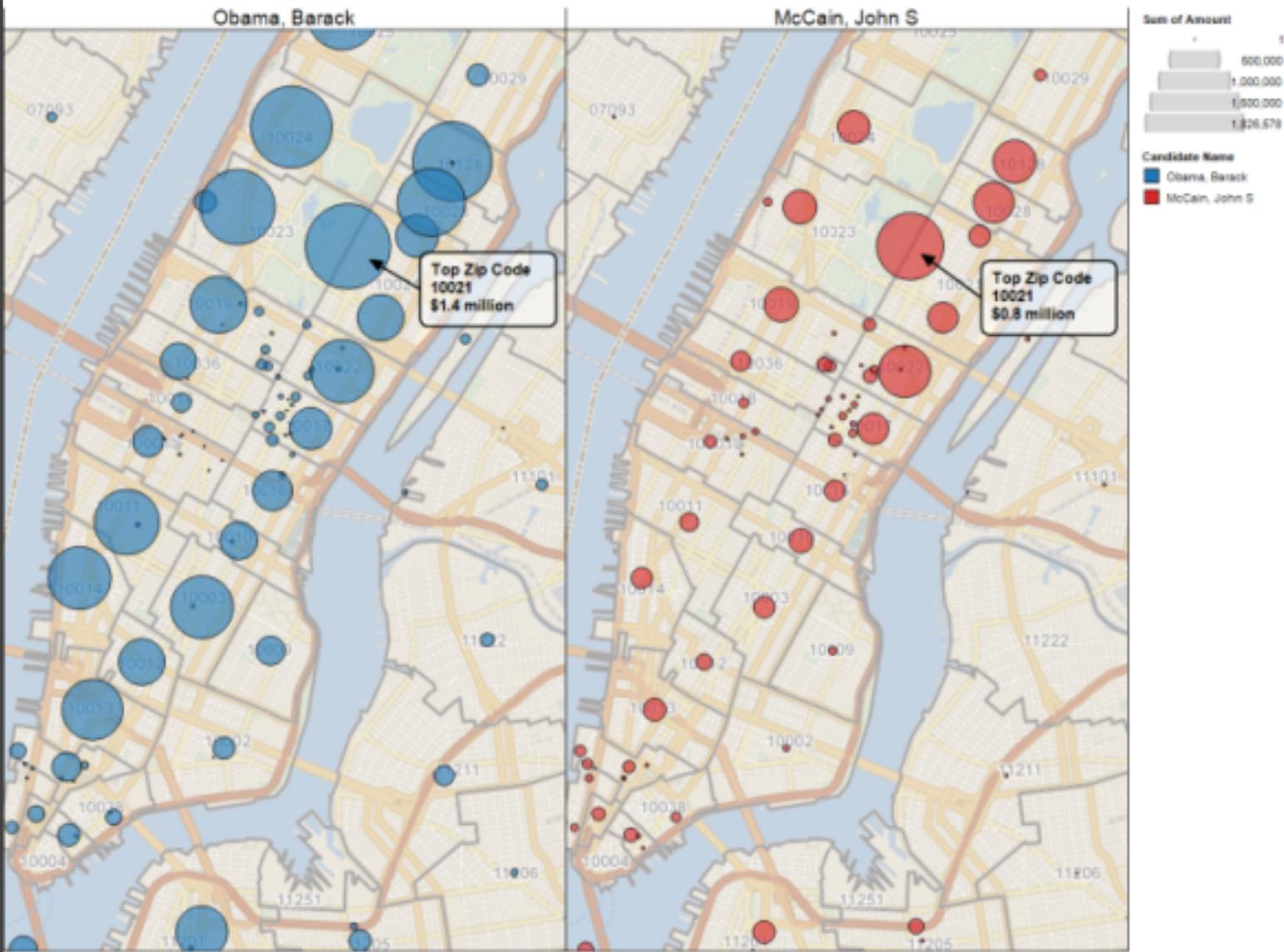
SUPPLY	ITEM	SUPPLIER
	<u>PEN</u>	PARKER

ANS :

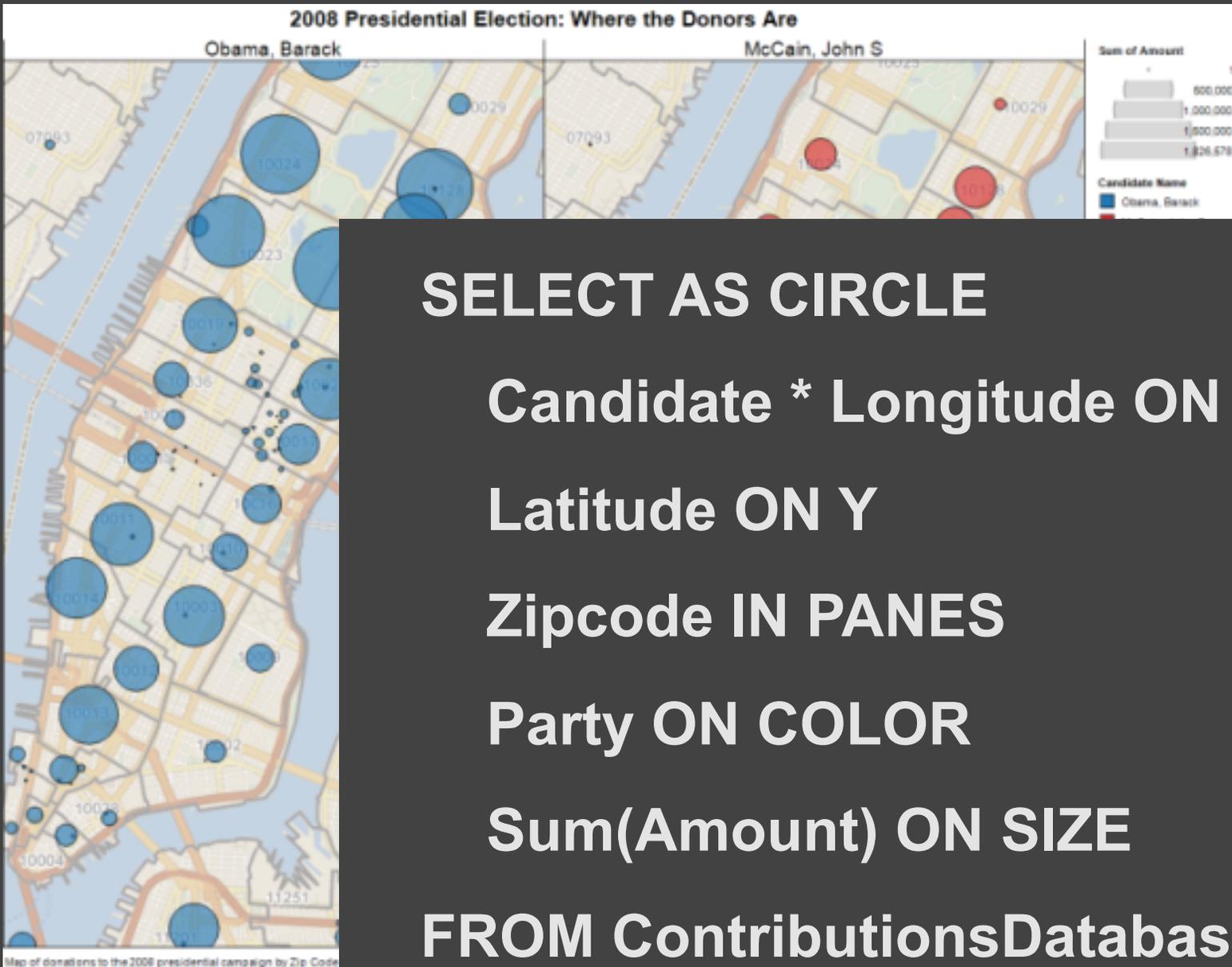
DEPT
HOUSEHOLD
TOY
STATIONARY
HARDWARE

Query-By-Example [Zloof, 1975]

### 2008 Presidential Election: Where the Donors Are



Map of donors to the 2008 presidential campaign by Zip Code and Candidate. Data is current to April 1, 2008 and was provided by the Federal Election Committee via their public website.



**SELECT AS CIRCLE**

**Candidate \* Longitude ON X**

**Latitude ON Y**

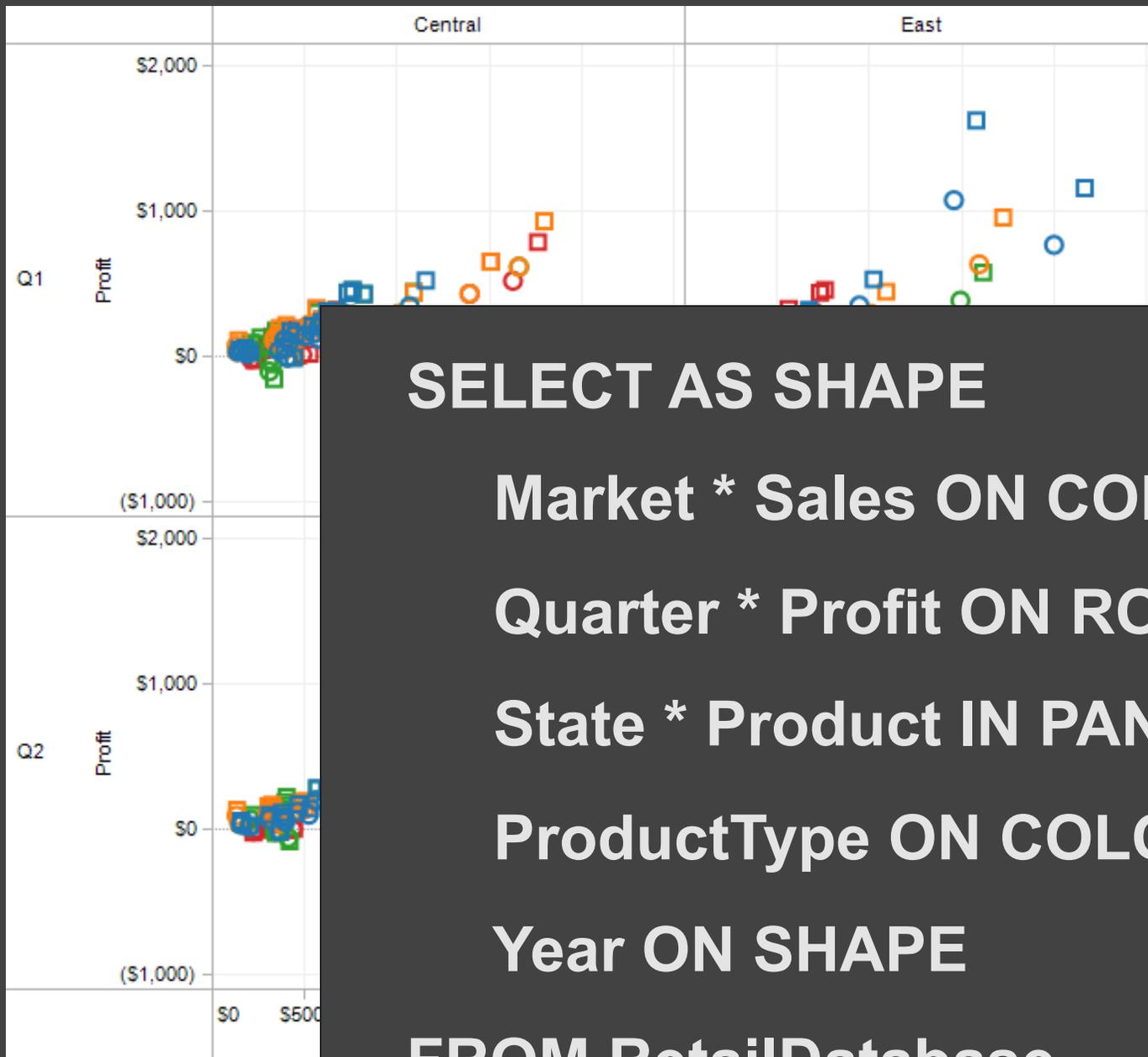
**Zipcode IN PANES**

**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\*:

<b>C Program</b>	<b>0.2 s</b>
<b>mysql</b>	<b>26.2 s</b>
<b>DBMS "X"</b>	<b>28.4 s</b>

\*Boncz et al., CIDR 2005

# In-Memory Column Stores: 100x

---

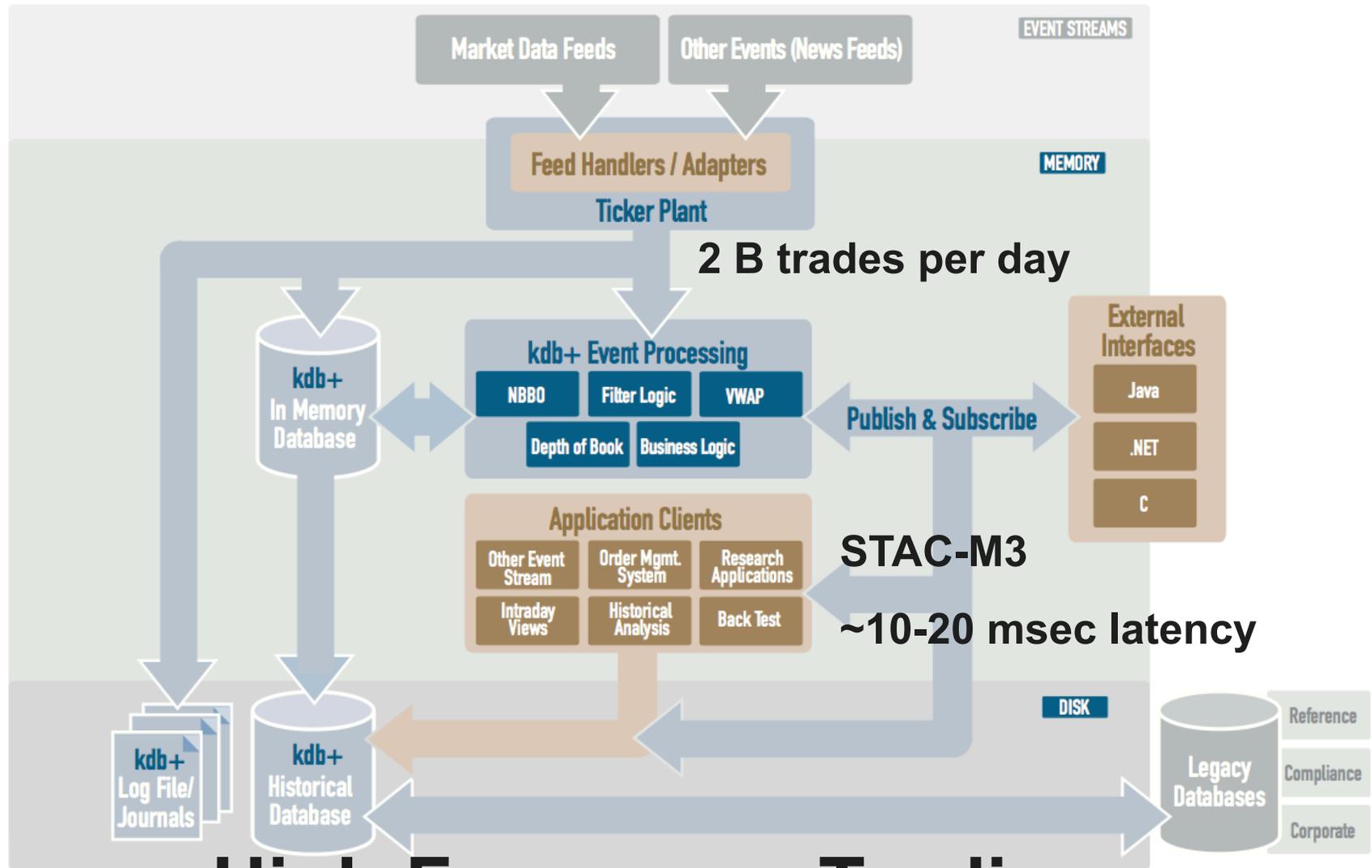
Columns are efficient (MonetDB/X100, C-store, ...)

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

In-Memory reduces latency enabling interaction

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

# kx.com



## 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

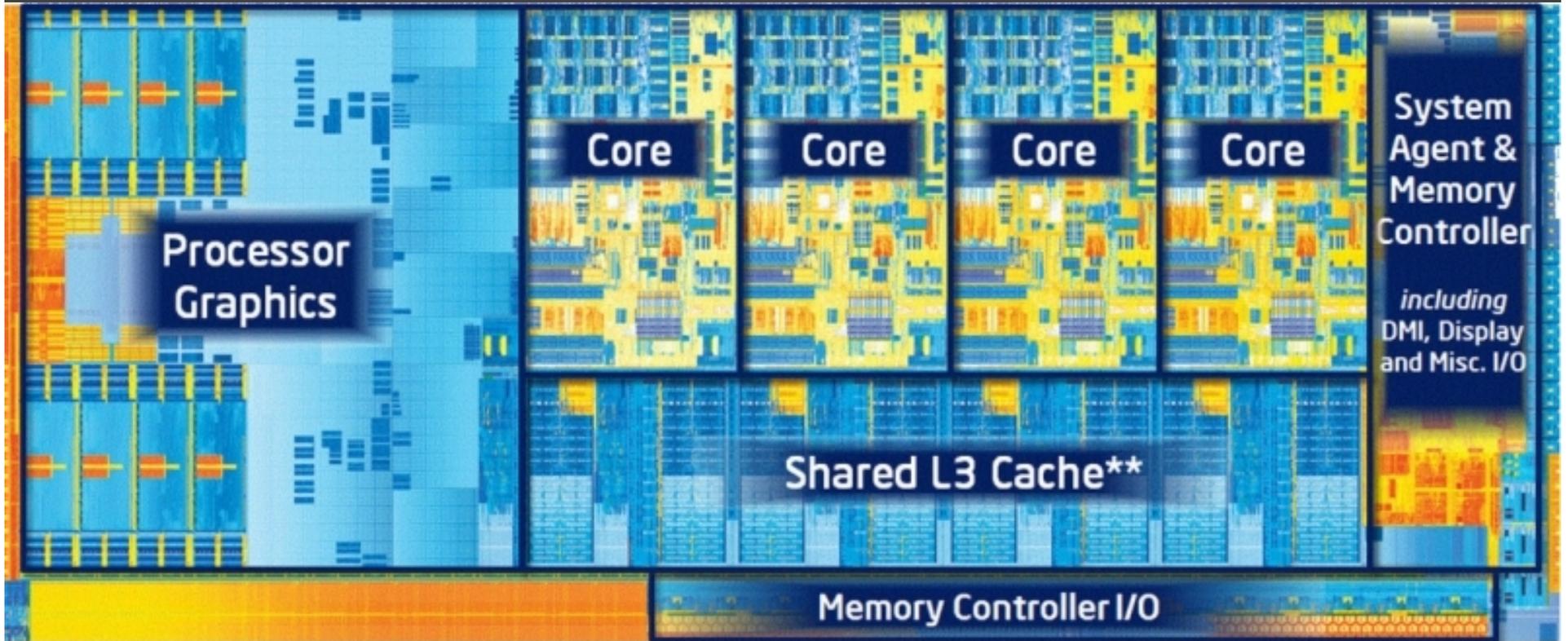
256-bit AVX vector instructions (16 FADD/FMUL)

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 msec



**2018 Laptop ~ CPU+GPU**

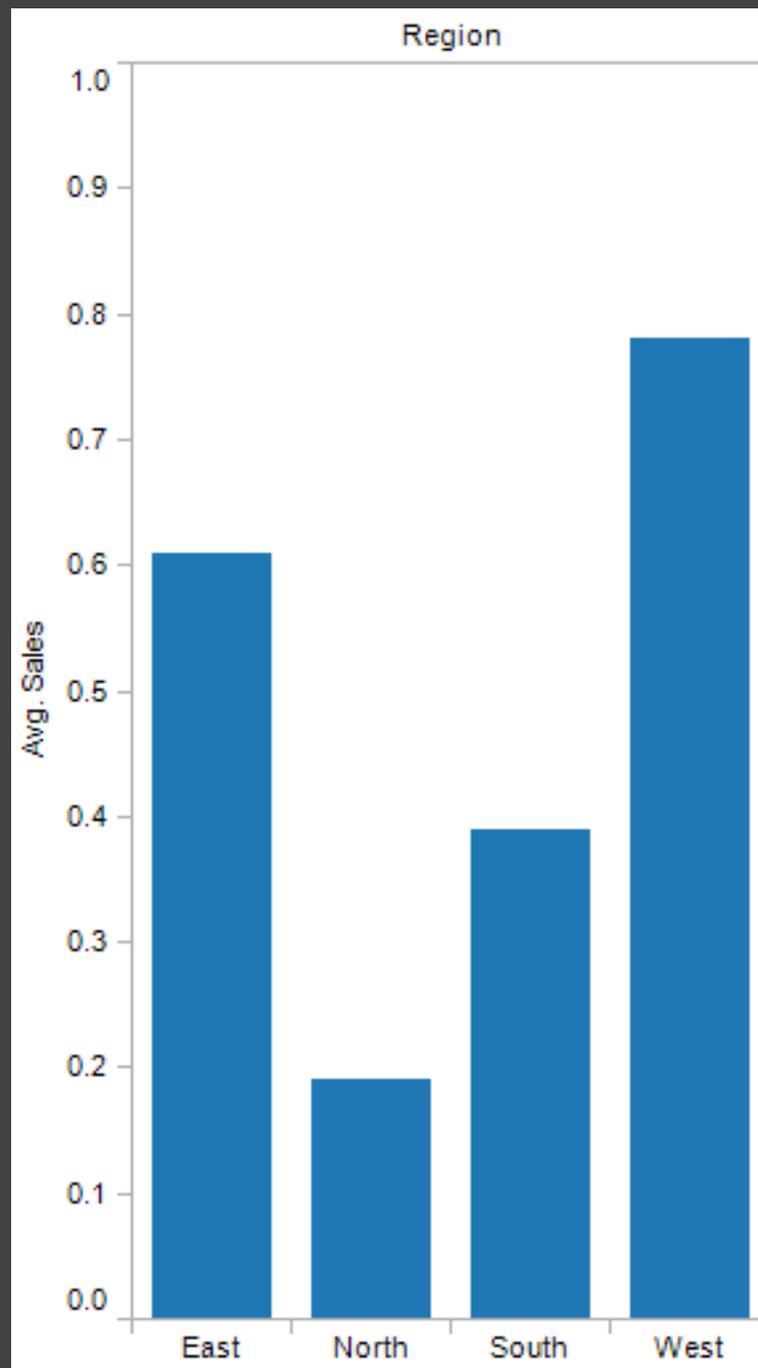
**10 Teraflops**

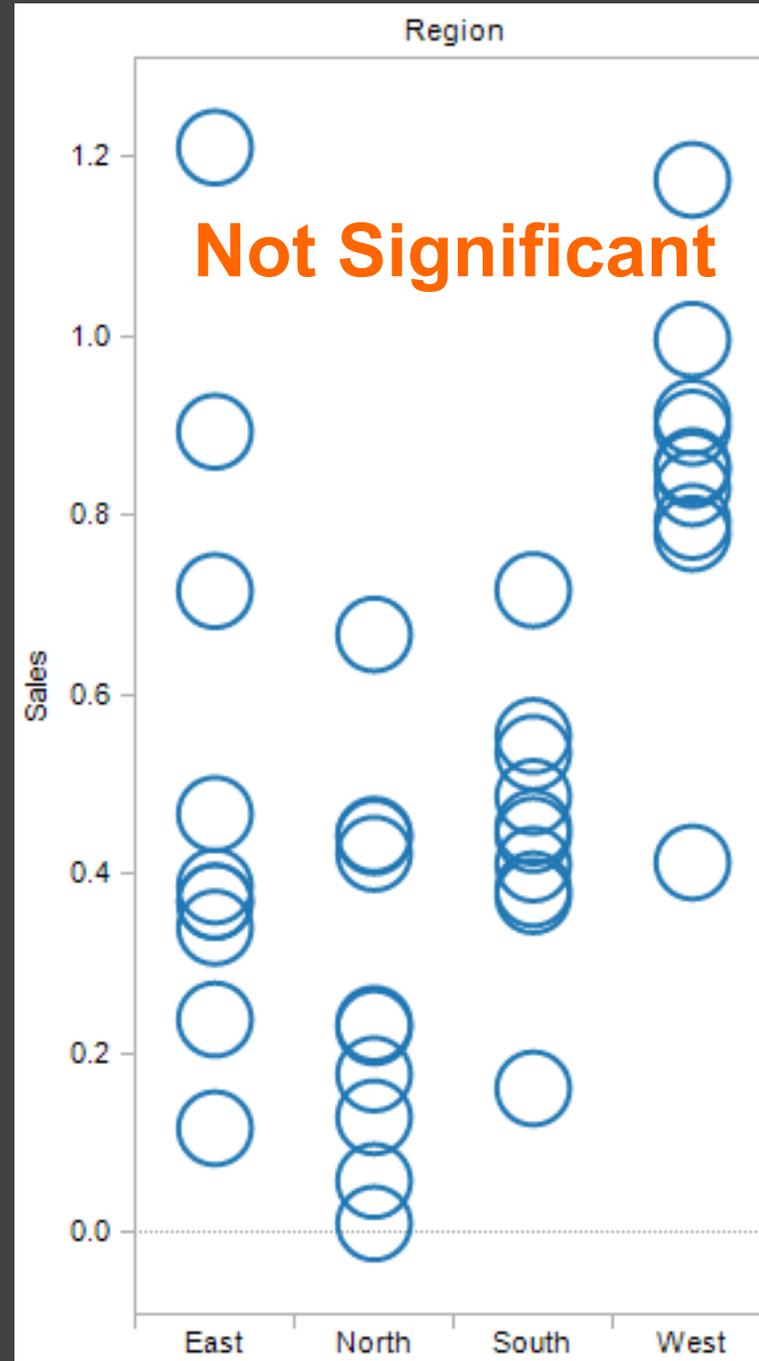
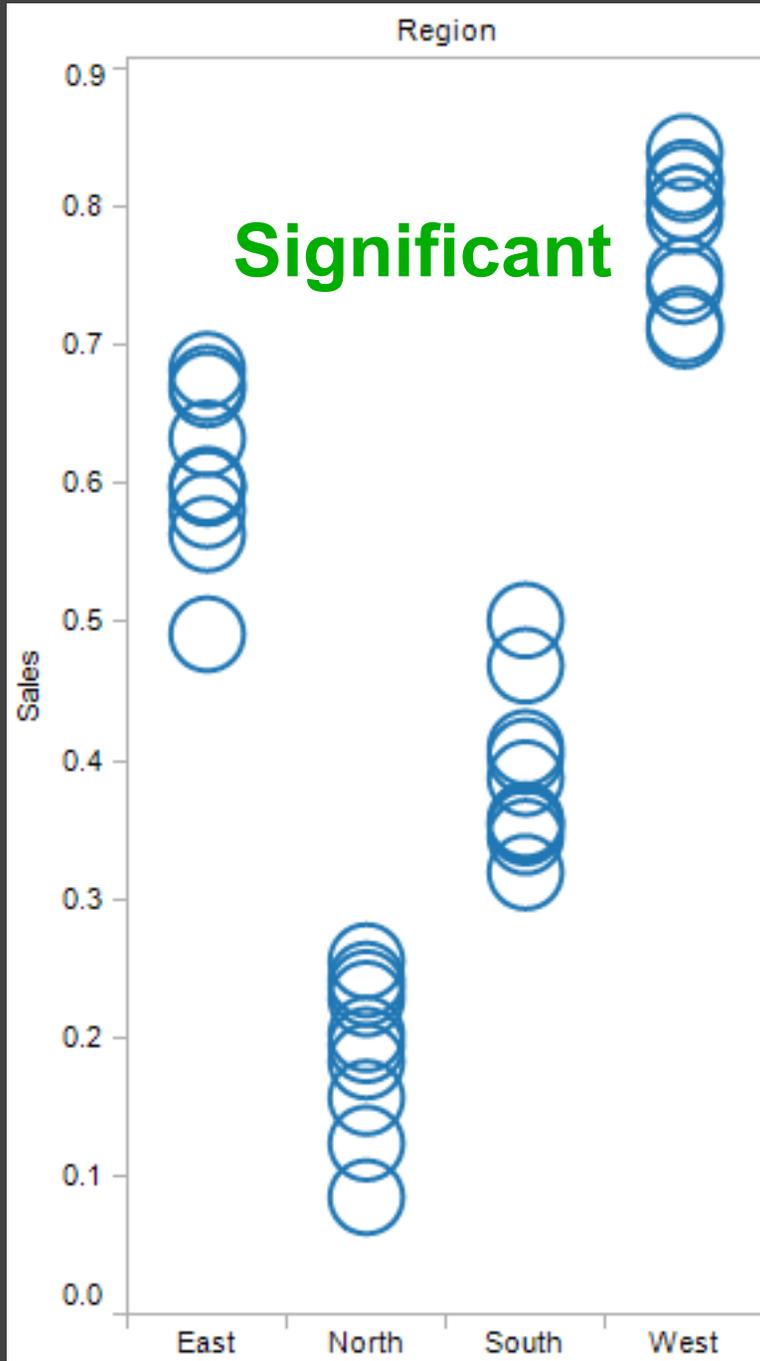
# Supporting Data Enthusiasts

**“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**





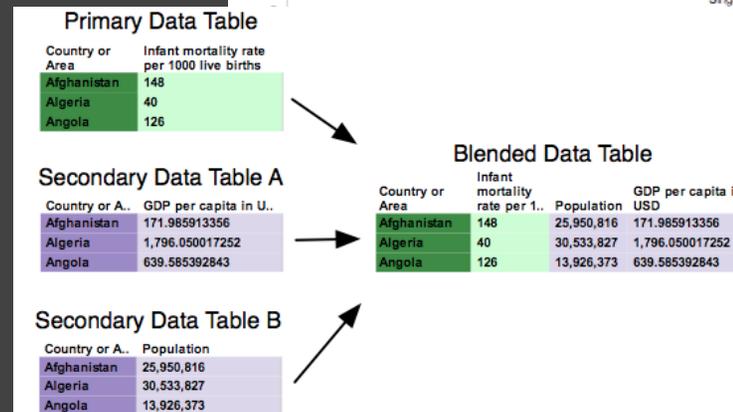
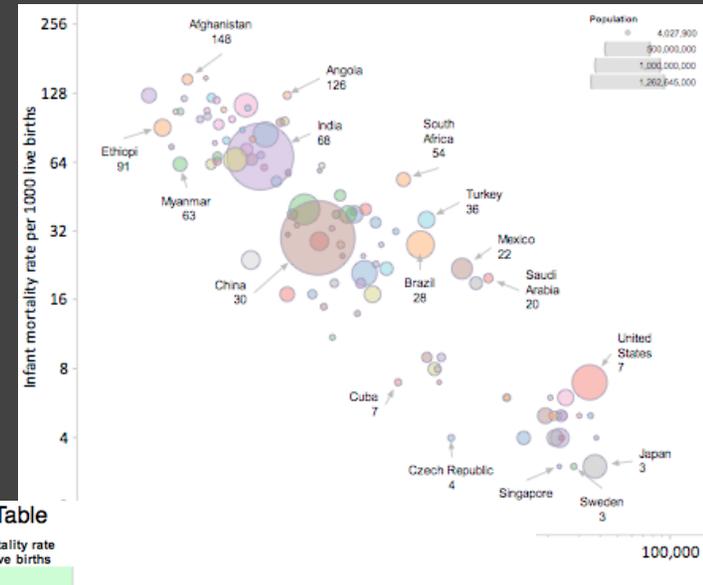
# Data Integration

Provides context for analysis

Semantic integration => people

Promising tools

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