

10 Best Practices for Faster Analytics using Super Graphics

Thursday Nov 8 2011

Montreal

Andrew Cardno
Managing Partner
American Kiwi LLC
andrewcardno@yahoo.com

About the Instructor: Andrew Cardno

Academic

- Started out as formally trained cartographer who professional practiced hand drawn cartography as part of an engineering/town planning practice.
- Bachelor of Surveying from Otago University, New Zealand and a Diploma of Computer Science from Victoria University, New Zealand.
- 16+ as a software engineer designing and/or hands on building software systems to deliver visualization software. 13+ years hands-on experience in data warehouse construction.
- Lead privately funded Phd and Masters research/development teams in visualization and math methods for 12+ years.
- Lead teams to win two Smithsonian laureates for heroism in information technology in data visualization (Telecommunications and Gaming).
- Lead teams to win 5 technology/innovation awards in data visualization in gaming

Publications

- Published over four dozen articles in industry journals
- Co-Author of The Math that Gaming Made (Volume 1) with Prof Singh (UNLV), pub 2011
- Named inventor on over 60 patent applications including high performance databases and data visualization and cartography.

Career

- CTO of BIS²
- Co-founder and CEO of Compudigm International.
- CTO of TableMax - Assisted with merger into pink sheet NASDAQ 2008
- Delivered visualization to : Seminole Gaming, Harrah's Entertainment Corporation, Penn National Gaming, Inc. (NASDAQ: PENN), Trump Entertainment Resorts, Inc. (NASDAQ: TRMP), Wal-mart Stores, Inc. (NYSE: WMT), Best Buy Co., Inc. (NYSE: BBY), Mayo Clinic, Sabre Systems, Continental(United) Airlines

Acknowledgements

Significant contributions to the content in this workshop come from working with the following individuals:

- Stephen Brobst, Sampo Technologies & Systems
- Richard Hackathorn, Bolder Technology
- Mark Madsen, Third Nature
- Sylvain Tassé President BI2U

The instructor gratefully acknowledges the contributions from these and many other individuals with whom I have worked over the past 16 years.

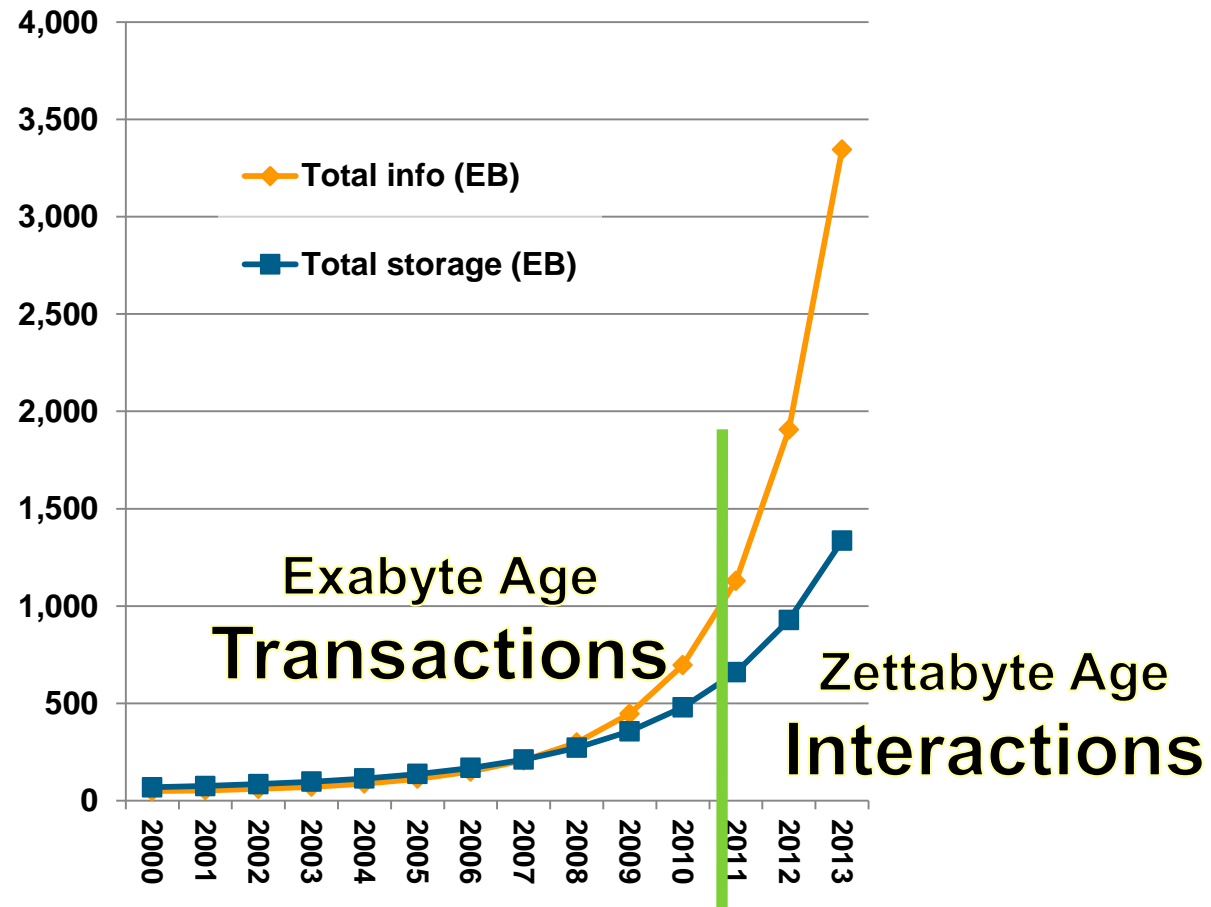
Agenda

- Driving the need
- 10 Best Practices for Faster Analytics
- Story 1: EDW + Spatial
- Story 2: EDW + Statistical + Visual
- Story 3: Electricity Control Room
- Story 4: Medical Diagnosis

**Driving The
Need**

The Data Explosion

- **More data has been created in the last three years than in the past 40,000 years.**
- **Almost all of this data has a location.**
- **Business and government decision-makers must have a strategy for dealing with location based data.**



One Zettabyte (ZB) = 1,000,000,000,000,000,000 bytes = 10^{21} bytes.
Based on IDC data growth estimates.

The Data Explosion

- **The worlds information is now doubling every 2 years***
- **1.8 Zetabites will be created in 2011**
 - > **In terms of sheer data 1.8 Zetabites is equal to:**
 - **Every person sending 3 tweets per minute = 4320 tweets per person per day FOR 26,976 years**
 - **200 Billion HD movies of 120 minutes**
 - **Storage of 1.8 Zetabites on a 32 gb Ipad would require 57.5 Billion Ipad**

*source: Extracting value from Chaos IDC digital universe study

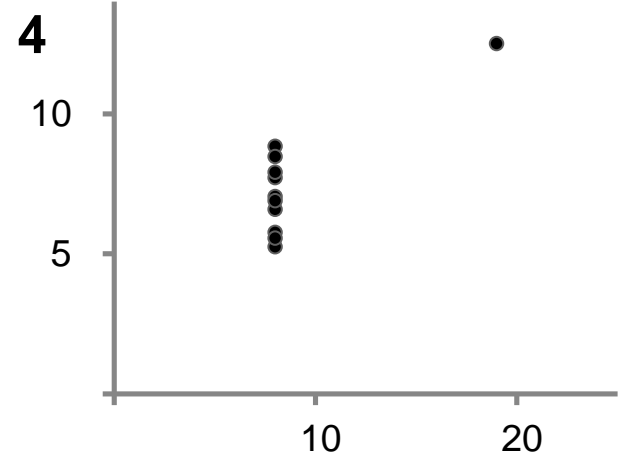
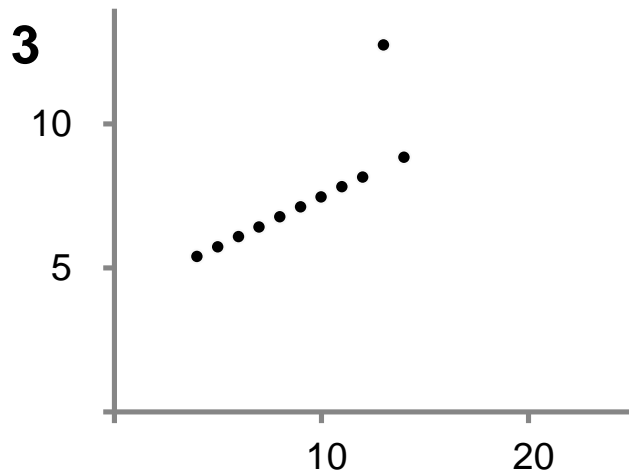
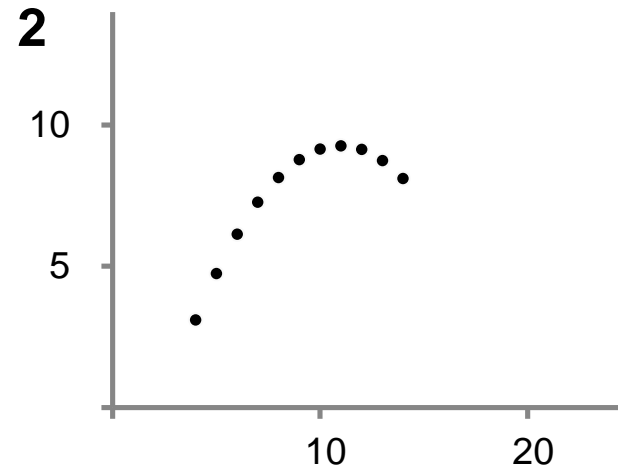
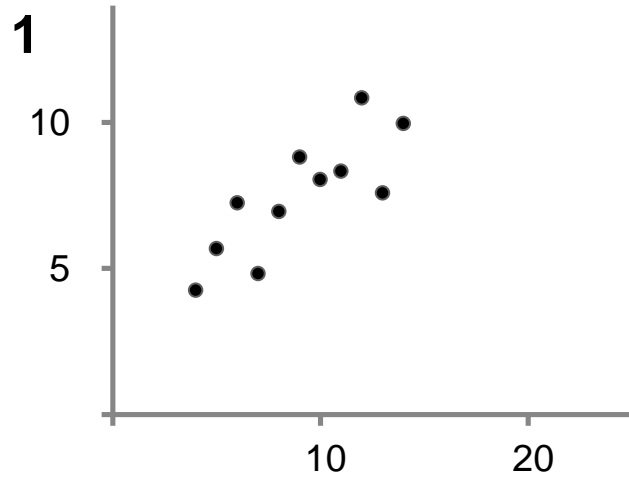
Computationally Centric Analysis

1		2		3		4	
X	Y	X	Y	X	Y	X	Y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89



N = 11
 mean of X's = 9.0
 mean of Y's = 7.5
 equation of regression line : $Y = 3 + 0.5X$
 standard error of estimate of slope = 0.118
 $t = 4.24$
 sum of squares $X - X = 110.0$
 regression sum of squares = 27.50
 residual sum of squares of Y = 13.75
 correlation coefficient = .82
 $r^1 = .67$

Visually Centric Analysis



F.J. Anscombe, "Graphics in Statistical Analysis", American Statistician, 37 (February 1973), 17-21

The Power of Visual Perception



70%



30%

Total sense receptors in humans...

Perception is sometimes serial and slow... ...and sometimes parallel and immediate.

How many fives are in this list of numbers?

987349702756479021947286240924060370804702890727
803208029007305901270238008374082078720272008083
247802602703793715709701379706674620970941027806
927979709123097230919592750927309272197873497260

9873497027**5**6479021947286240924060370804702890727
80320802900730**5**901270238008374082078720272008083
24780260270379371**5**709701379706674620970941027806
927979709123097230919**5**927**5**0927309272197873497260

10 Best Practices for Faster Analytics

10 Best Practices: Overview + Detail

Outcome Practices

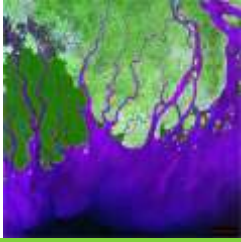
- Build on data infrastructure
- Give overview (context)
- Show detail (focus)
- Tell a story
- Generate excitement

Representation Practices

- Manage the ink
- Illuminate with detail
- Explore structural aspects
- Provide Overview and Detail
- Consider the art

4 Stories

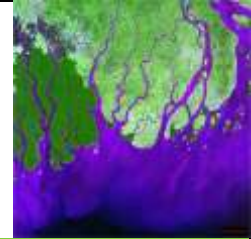
Story 1: InDatabase Locational



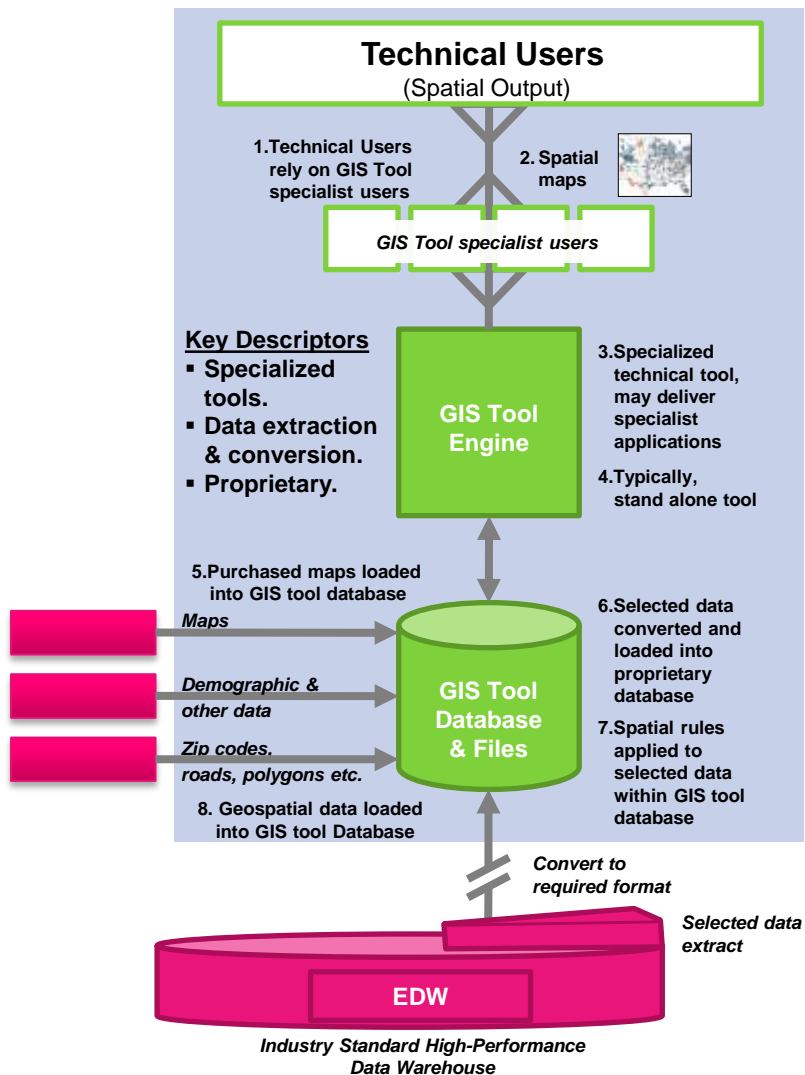
<http://landsat.gsfc.nasa.gov/earthstar/>

1: EDW+ Spatial

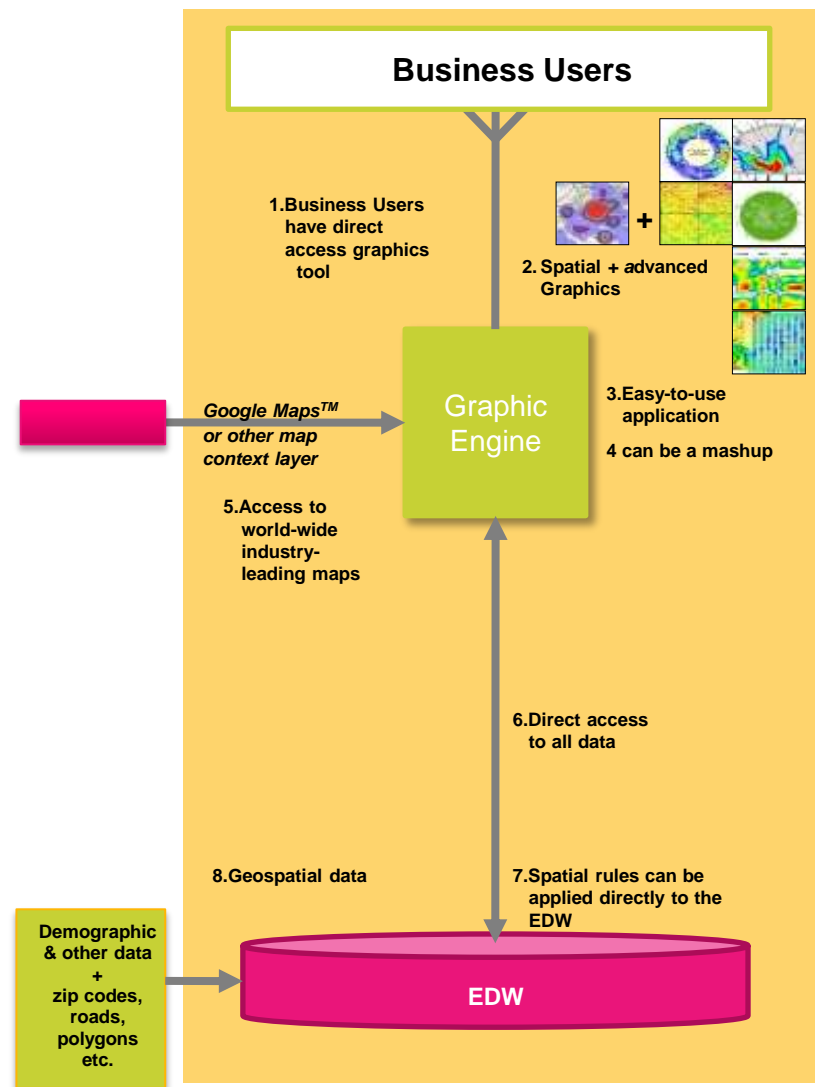
1: EDW + Spatial: Locational Data Infrastructure



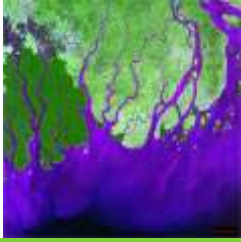
TYPICAL GIS TOOL APPROACH



IN-EDW APPROACH



1: EDW + Spatial : Locational Data Infrastructure



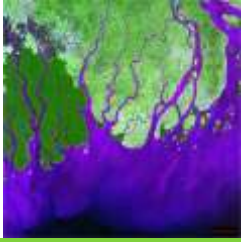
<http://landsat.gsfc.nasa.gov/earthstar/>

- In database spatial data
- Calculate spatial buffers
- Calculate join spatially

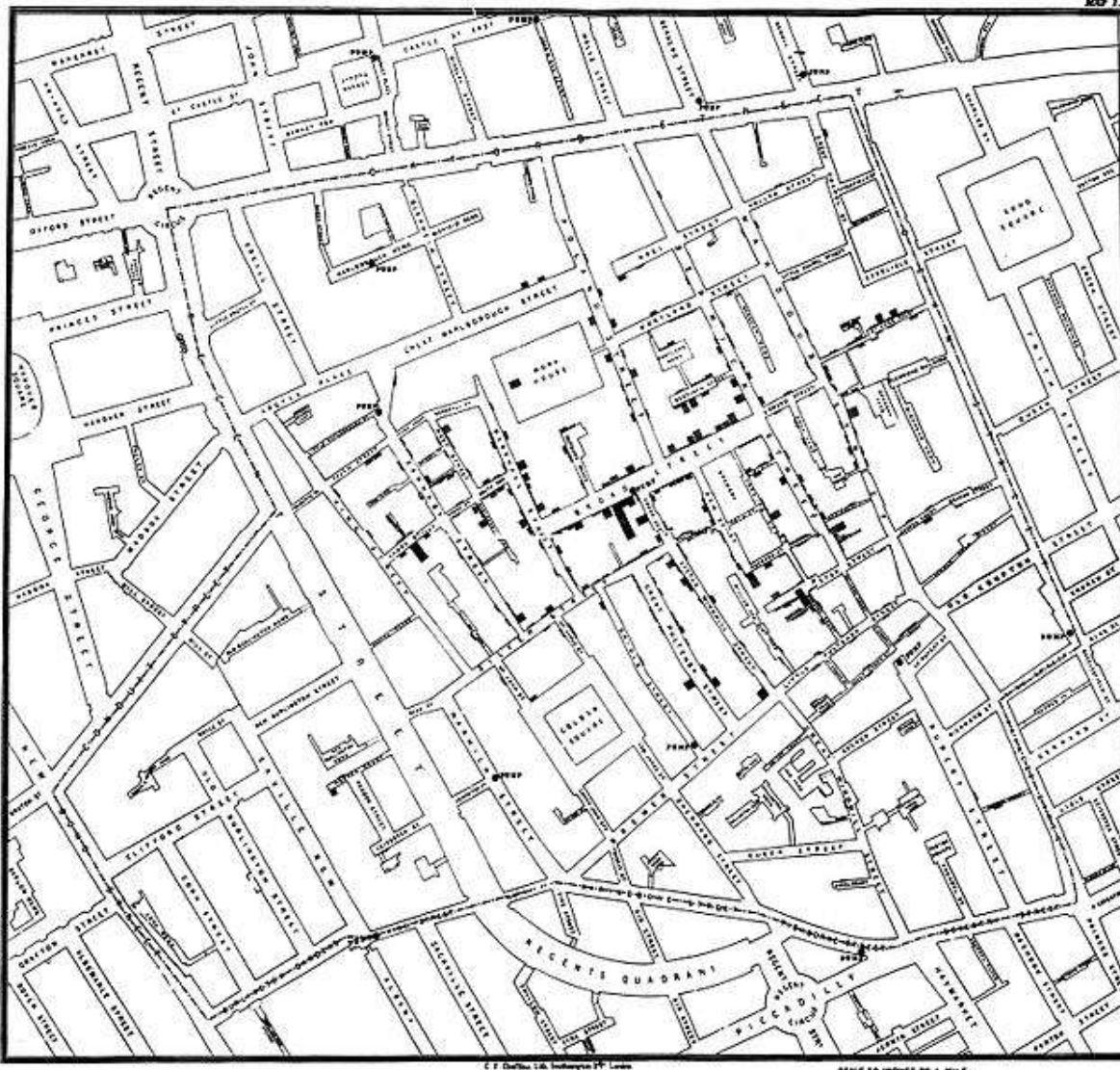
```
SELECT extract(year from cast (GamingDemo.StoreTransaction.DateTime as date)) AS
VBB2__NORM_FIELD,GamingDemo.geometry_county_CA.name AS
CountyName,GamingDemo.StoreLocation.StoreId AS
StoreIdDimProxy,GamingDemo.StoreLocation.Location AS
StoreLocation,SUM(GamingDemo.StoreTransaction.Revenue) AS Revenue2008
FROM GamingDemo.StoreTransaction INNER JOIN GamingDemo.StoreLocation ON
GamingDemo.StoreTransaction.StoreId = GamingDemo.StoreLocation.StoreId INNER JOIN
GamingDemo.geometry_county_CA ON
GamingDemo.StoreLocation.Location.ST_WITHIN(GamingDemo.geometry_county_CA.polygon) =
1
WHERE ((GamingDemo.StoreTransaction.DateTime >= (Timestamp '2008-01-01 00:00:00')) AND
(GamingDemo.StoreTransaction.DateTime < (Timestamp '2009-01-01 00:00:00')))
GROUP BY extract(year from cast (GamingDemo.StoreTransaction.DateTime as
date)),GamingDemo.geometry_county_CA.name,GamingDemo.StoreLocation.StoreId,GamingDe
mo.StoreLocation.Location
```



1: EDW + Spatial : Overview + Detail (1854)

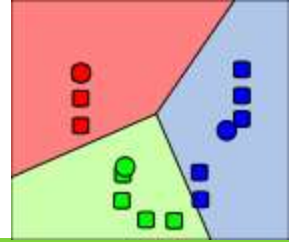


<http://landsat.gsfc.nasa.gov/earthart/>



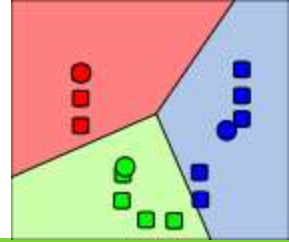
- Manage the ink
- Illuminate with detail
- Explore structural aspects
- Provide Overview and Detail
- Consider the art



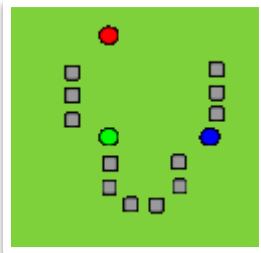


2: EDW + Statistical + Visual

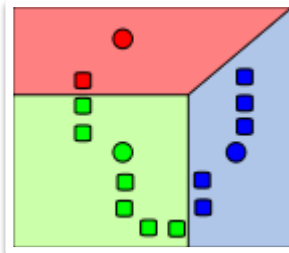
2: EDW + Statistical + Visual



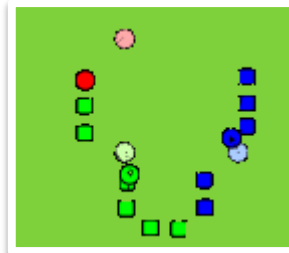
- Example: In database customer segmentation algorithm (Kmeans)



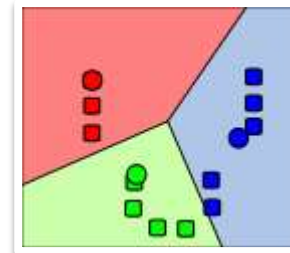
1) k initial "means" (in this case $k=3$) are randomly selected from the data set (shown in color).



2) k clusters are created by associating every observation with the nearest mean. The partitions here represent the Voronoi diagram generated by the means.



3) The centroid of each of the k clusters becomes the new means.

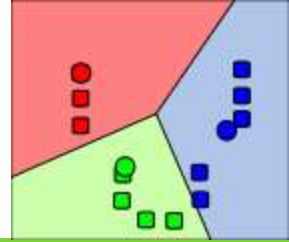


4) Steps 2 and 3 are repeated until convergence has been reached.

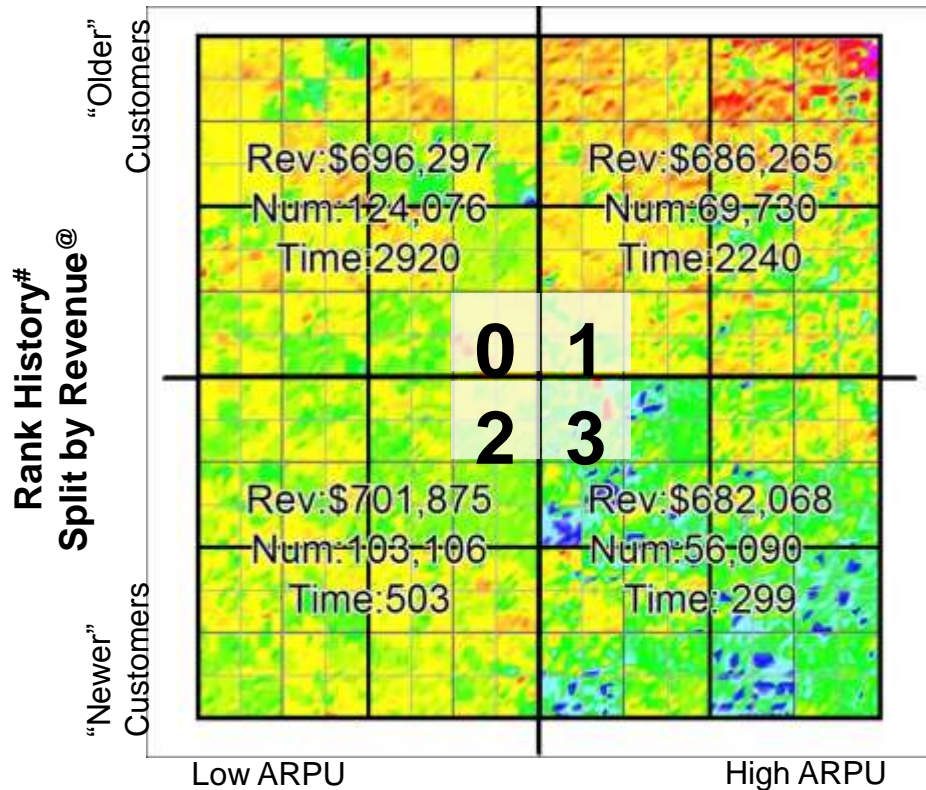
http://en.wikipedia.org/wiki/K-means_clustering

Produces a number per customer

2: EDW + Statistical + Visual



Identifying Telco customer segment opportunities:



Rank Average Revenue Per User (ARPU) Per Year Split by Revenue

0. Top left:

(Older customers and low ARPU)

Customers who have been with us a long time and have a low revenue per user.

1. Top right:

(Older customers and high ARPU)

Customers who have been with us a long time and have a high revenue per user.

2. Bottom left:

(Newer customers and low ARPU)

Customers who have been with us a short time and have a low revenue per user.

3. Bottom right:

(Newer customers and high ARPU)

Customers who have been with us a short time and have a high revenue per user.

History means the length of time someone has been a customer. An "older" customer means that they have been a customer for a relatively longer period, than a new customer. Each quadrant is split by an equal amount of revenue.

@ Alternatively, could be split by ARPU.

2: EDW + Statistical + Visual: Overview + Detail

ILLUSTRATIVE EXAMPLES ONLY

CHURNERS – Cluster 1

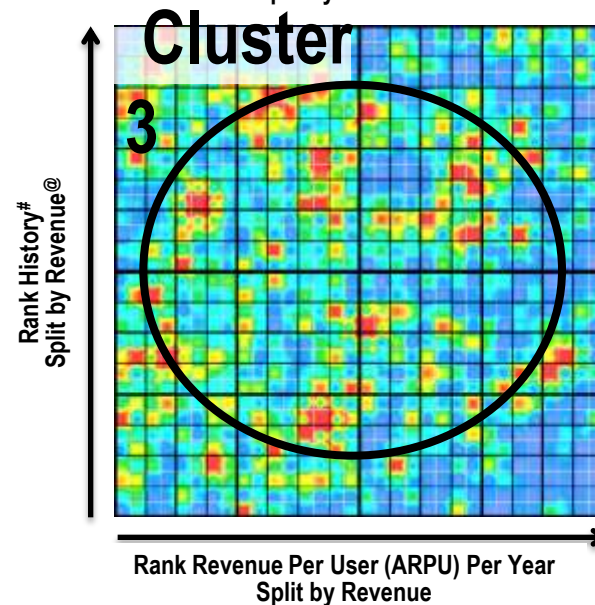
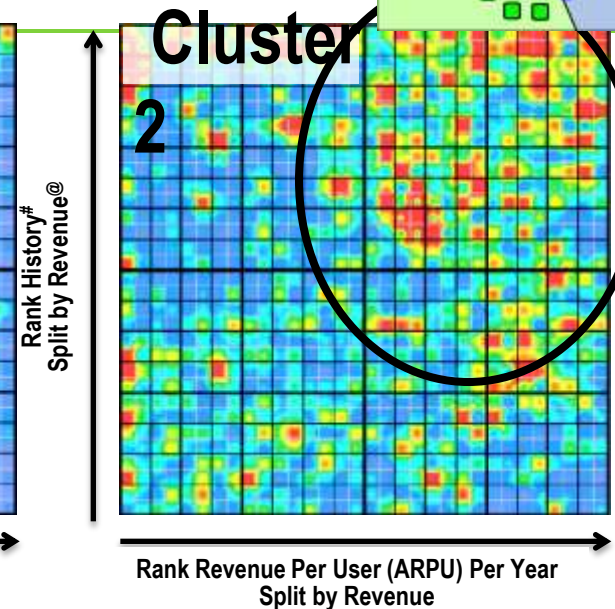
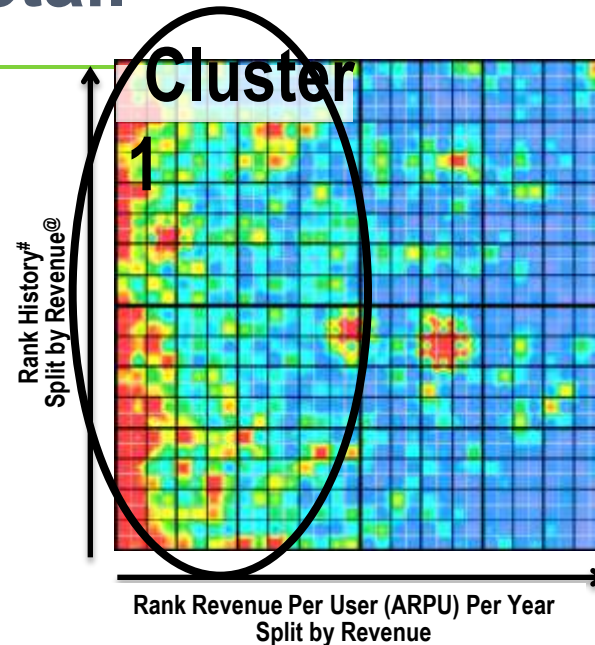
This cluster is low revenue customers, both old and new.

OLD GOLDIES' – Cluster 2

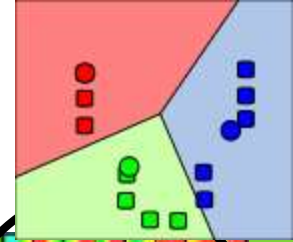
This cluster represents customers who have been with us a long time and high spend.

CENTRISTS – Cluster 3

This cluster represents average spend customers and neither old nor new



- Manage the ink
- Illuminate with detail
- Explore structural aspects
- Provide Overview and Detail
- Consider the art



Story 3: Active Decision Making



3: Electricity Control Room

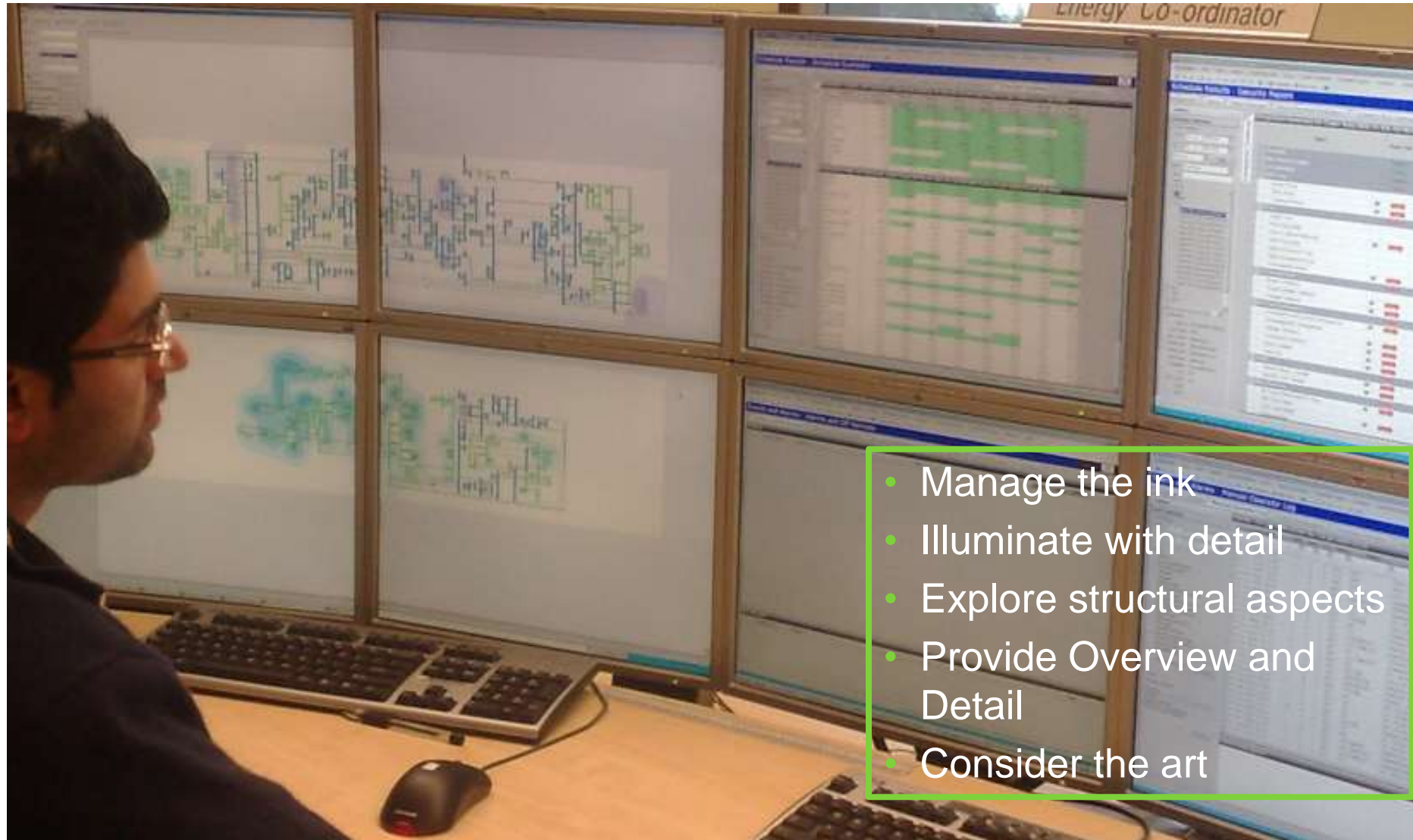
3: Active Decision Making



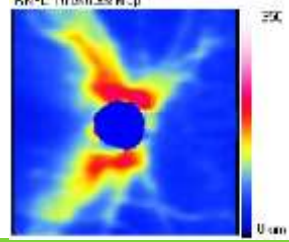
- Real time data management
- Active predictive models of power price, demand
- Humans do anomaly detection



3: Active Decision Making: Overview + Details

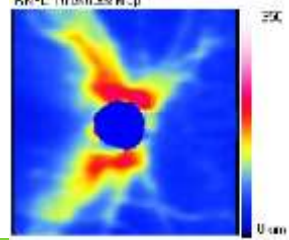


Story 4: Medical Diagnosis

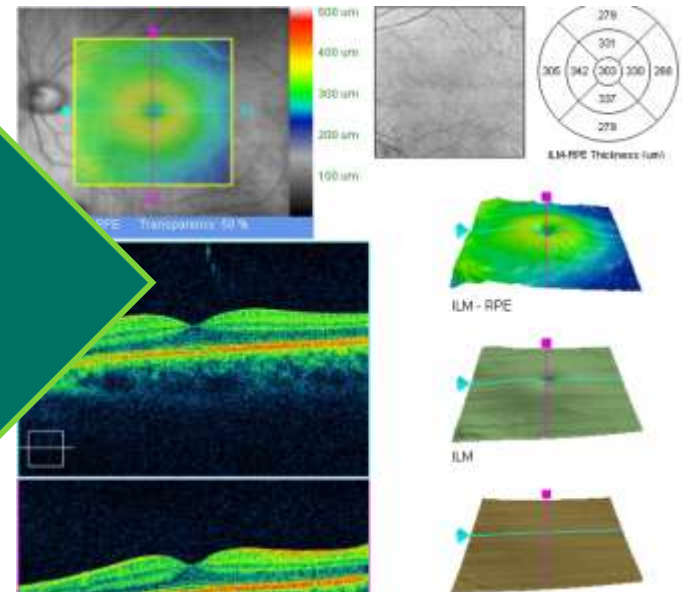


4: Medical Diagnosis

4: Medical Diagnosis

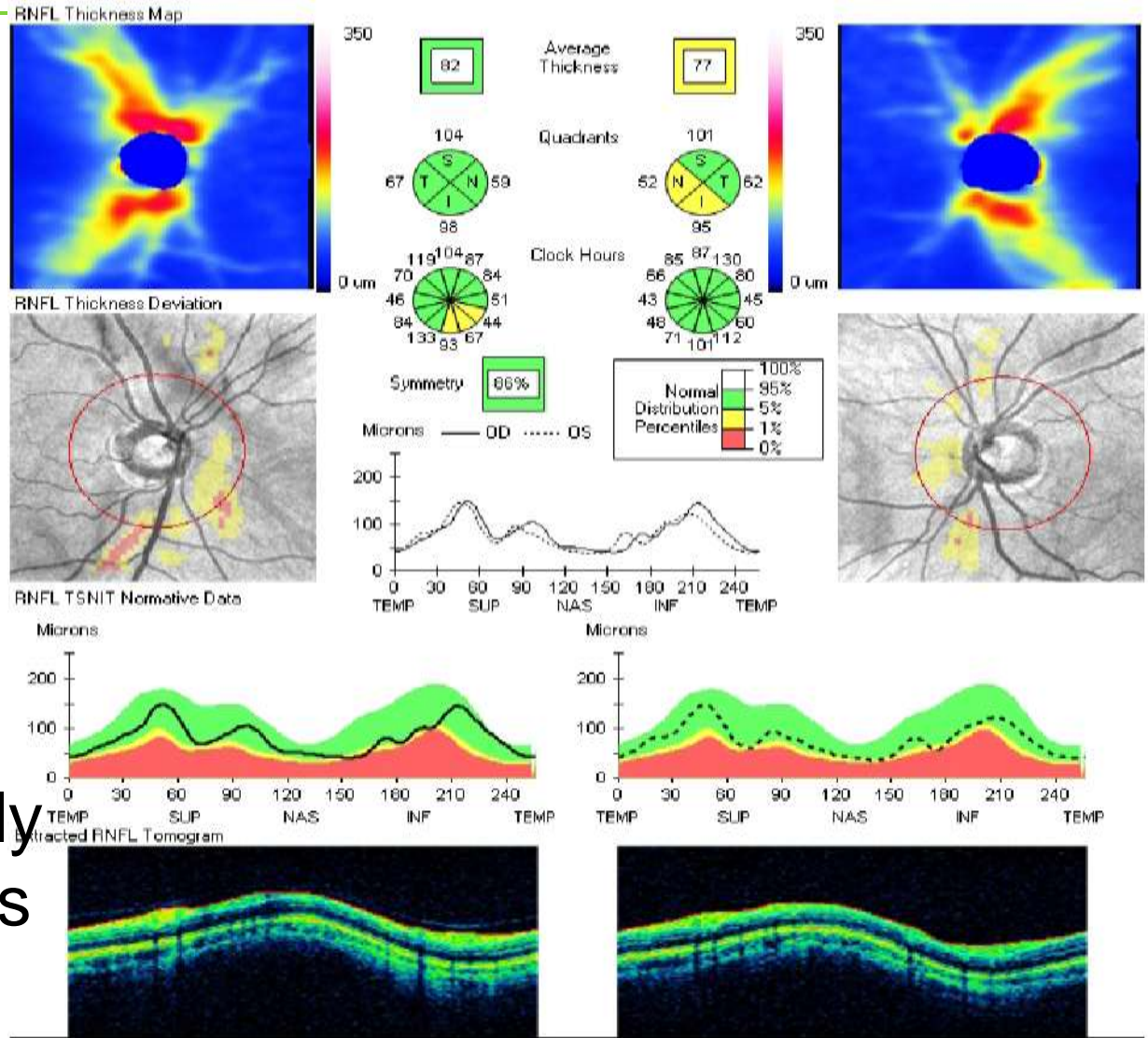


- High density measurement of the eye
- Models generate numbers showing potential issues
- Humans do anomaly detection



4: Medical Diagnosis: Overview + Detail

- Manage the ink
- Illuminate with detail
- Explore structural aspects
- Provide Overview and Detail
- Consider the art



Diagnosis is visually centric and involves skilled judgment

Summary

Overview
+
Detail
=
Speed

Questions?

4 Stories to Illustrate?

Outcome Practices

- Build on data infrastructure
- Give overview (context)
- Show outliers (focus)
- Tell a story
- Generate excitement

Representation Practices

- Manage the ink
- Illuminate with detail
- Explore structural aspects
- Provide Overview and Detail
- Consider the art

Recommended Reading

Edward Tufte, Yale University.

