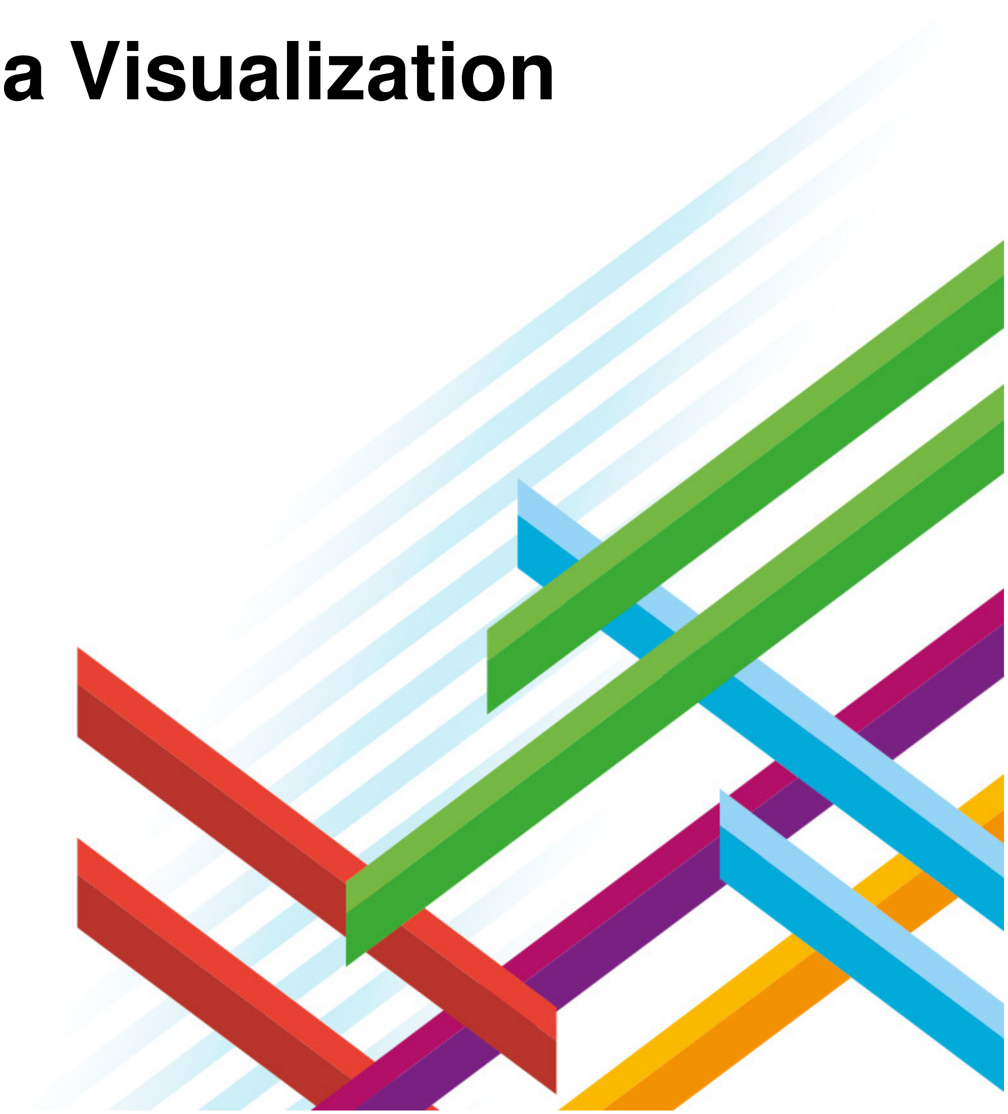
IBM BusinessConnect
Vernetzter, intelligenter und informierter denn je

Data Analytics and Data Visualization

Two sides of the same coin

Frank van Ham, PhD.

frankvanham@nl.ibm.com



Data has become a core resource in the 21st century

- Companies are very adept at managing different kinds of resources (Human, Financial, Physical)
- Most companies by now realize they need to manage their data resources as well.
  - Better data acquisition
  - Better/larger/cheaper data storage
  - Better and more consistent data models
  - Better security
  - …

- Big Data involves new engineering challenges in acquisition, storage and querying.

- However, the biggest challenge is in interpreting all of this data.

“So my 2TB Hadoop database is in the Cloud, now what?”

- Data is a **resource** which needs to be refined to be useable.
- This concept of refinement is typically expressed by the classic information hierarchy.

Storage	Data	**Raw facts**. No context. “male, 37, Dutch”
Classic BI	Information	**Actionable data**. “What … ?”, “How many …?”
Analytics	Knowledge	**Doing things right**. “Why is this happening?”, “How do I do this?”
Humans (for now)	Wisdom	**Doing the right things**. Strategy, vision.

Analytic algorithms can do the right things for us!

- We are moving into the era of analytics and smart algorithms.
- We can use powerful data mining and optimization algorithms to take decisions for us.
- Complex decision making tasks can be automated away, if we just deploy the appropriate analytic algorithm that will give us the answer we need.
- Any complex task can then be made faster by scaling computing hardware.

Will the start of my future working day look like this?

If what I told you were true, it will look more like this.

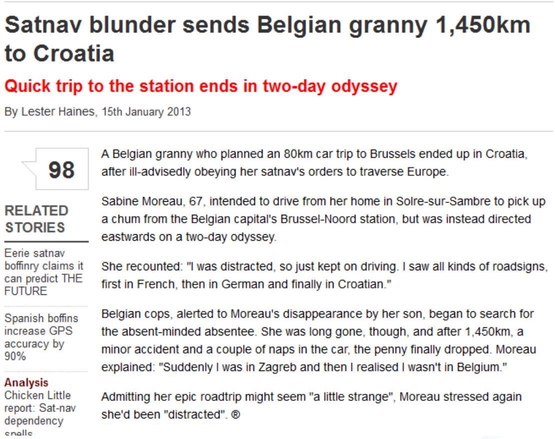
Actually, analytics are decision making **tools.**

- For many business problems there is no such thing as ‘the best answer’.
  - There might be multiple best answers, each with their own tradeoff.
  - The second best answer is better in practice.
  - The best answer might be completely impractical.
  - You may not even be able to specify the question clearly enough.

- For the foreseeable future, no combination of analytics can completely substitute for human contextual knowledge, intuition and creative problem solving.

- Instead, analytics are <u>tools</u> that can help us do our jobs better (more effective, faster, cheaper)

- No one in their right mind would apply tools blindly. Just like with any physical tool, we have to ‘see’ what we are doing when applying analytics to data.



Satnav blunder sends Belgian granny 1,450km to Croatia

Quick trip to the station ends in two-day odyssey

By Lester Haines, 15th January 2013

98

RELATED STORIES

Eerie satnav boffinry claims it can predict THE FUTURE

Spanish boffins increase GPS accuracy by 90%

Analysis Chicken Little report: Sat-nav dependency spells

A Belgian granny who planned an 80km car trip to Brussels ended up in Croatia, after ill-advisedly obeying her satnav's orders to traverse Europe.

Sabine Moreau, 67, intended to drive from her home in Solre-sur-Sambre to pick up a chum from the Belgian capital's Brussel-Noord station, but was instead directed eastwards on a two-day odyssey.

She recounted: "I was distracted, so just kept on driving. I saw all kinds of roadsigns, first in French, then in German and finally in Croatian."

Belgian cops, alerted to Moreau's disappearance by her son, began to search for the absent-minded absentee. She was long gone, though, and after 1,450km, a minor accident and a couple of naps in the car, the penny finally dropped. Moreau explained: "Suddenly I was in Zagreb and then I realised I wasn't in Belgium."

Admitting her epic roadtrip might seem "a little strange", Moreau stressed again she'd been "distracted". ®

Humans not really made for digital data input

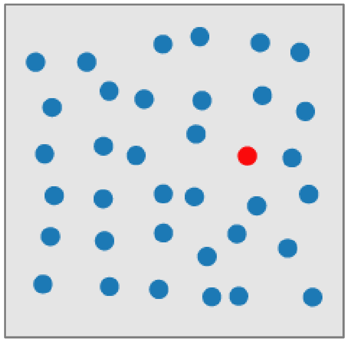
- Humans get most of their external input optically and our brains are highly tuned for visual processing.
- In the absence of direct digital input, visual input is the next best alternative channel we have.
- By using this channel smartly we can use ‘brain hardware’ instead of ‘brain software’ to transfer information from a computer to a brain.

Humans not really made for digital data input

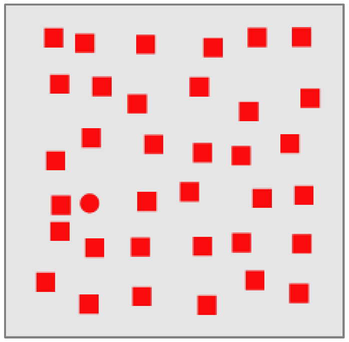
- Humans get most of their external input optically and our brains are highly tuned for visual processing.
- In the absence of direct digital input, visual input is the next best alternative channel we have.
- By using this channel smartly we can use ‘brain hardware’ instead of ‘brain software’ to transfer information from a computer to a brain.

Since we’re doing experiments, here’s another one

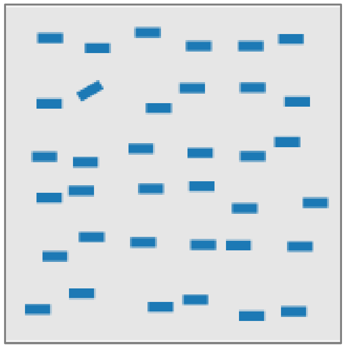
Find the outlier:



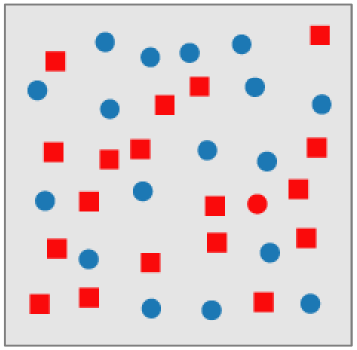
Color



Shape



Orientation



Color + Shape

Data Visualization encodes information visually.

- Data visualization is the process of encoding data as meaningful images which a human can digest more quickly.

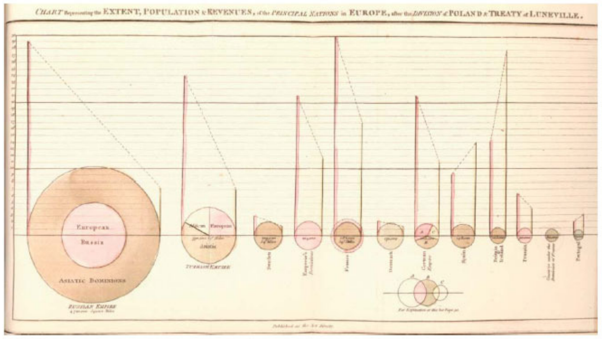
- Some encodings are more effective than others.

- Data visualization is just another medium, which happens to be well-suited to communicate data.

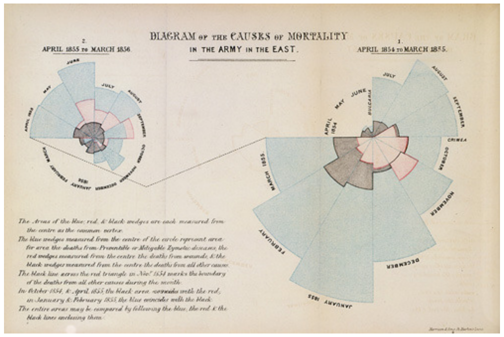
- Visualization in itself is not a product, but rather it's a 'force multiplier'. It will enhance the effectiveness of the product its embedded with.

Information Visualization is not a new idea in itself

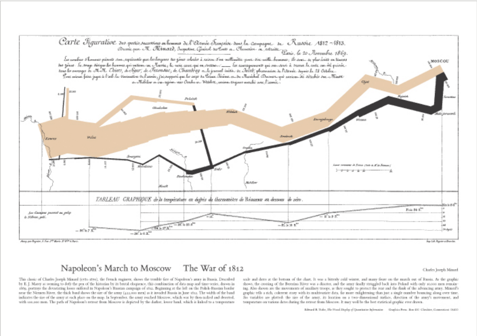
- Information Graphics have been around for a couple of centuries.
- Basic underlying mechanism is the same.
- However, modern computer and graphics technology make it possible to generate and share interactive versions quickly.



Playfair (1786)



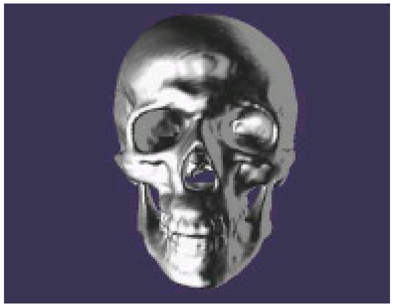
Nightingale (1858)



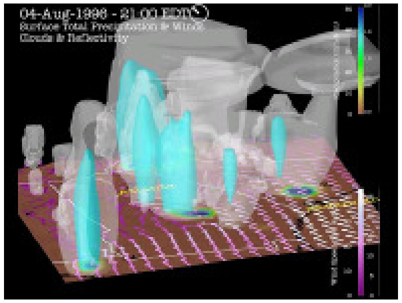
Minard (1869)

Scientific Visualization vs Information Visualization

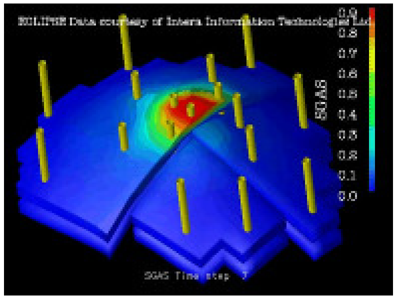
- **Scientific Visualization** creates imagery from data that have an inherently spatial component.



Medical



Weather forecasting

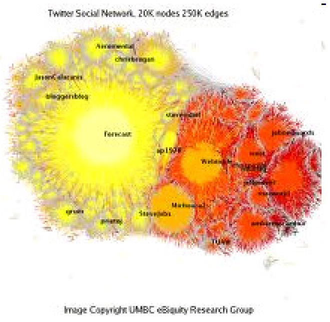


ECLIPSE Data courtesy of Intera Information Technologies Ltd
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
SGAS
SGAS Time step 7

Oil Drilling

Not a focus for BA, but check out IBM's OpenDX if you are interested in this area.

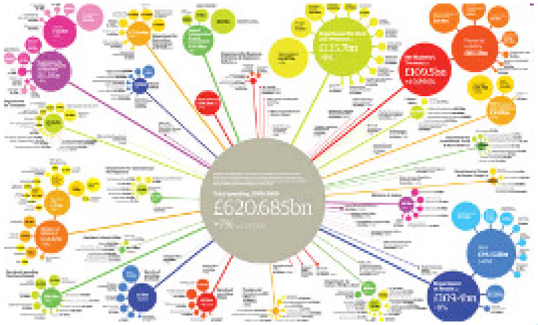
- **Information visualization** creates imagery from abstract non-spatial data.



Twitter Social Network, 20K nodes 250K edges
forecast
Steve Jobs
Image Copyright UMBC eBiquity Research Group

Social Networks

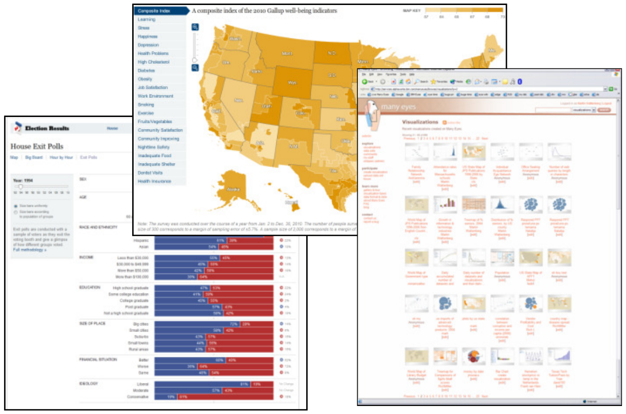
Projected Sales Data



Government Expenses by category

Two main usecases

- **Communication** : Communicate a predefined piece of information (infographics, informative charts, light interactive reports). Information **push**.



Data presentation, Data journalism

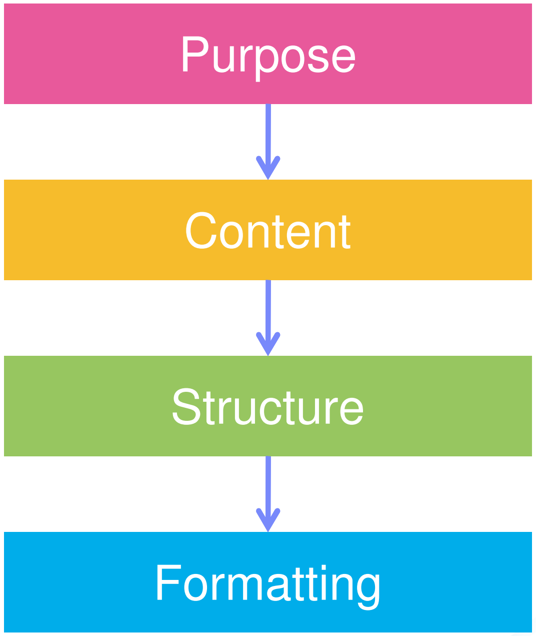
- **Analysis** : Help discover new information by allowing users to rapidly and interactively query their data in a visual way. Information **pull**.

Data Analysis, Data Analytics

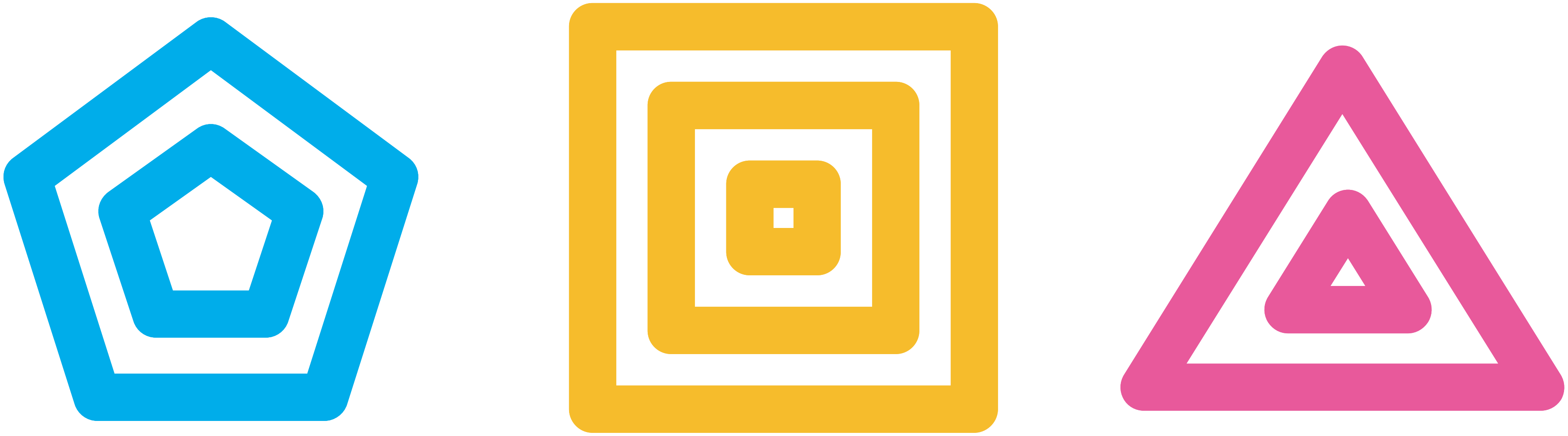
Communication : When telling a (data) story…

…keep in mind the following things

- **Purpose** : What am I trying to convey and for whom?
- **Content** : What data do I need to for my purpose?
- **Structure** : How do I structure my content such that it is clear?
- **Formatting** : How do I make my structure simple to consume?

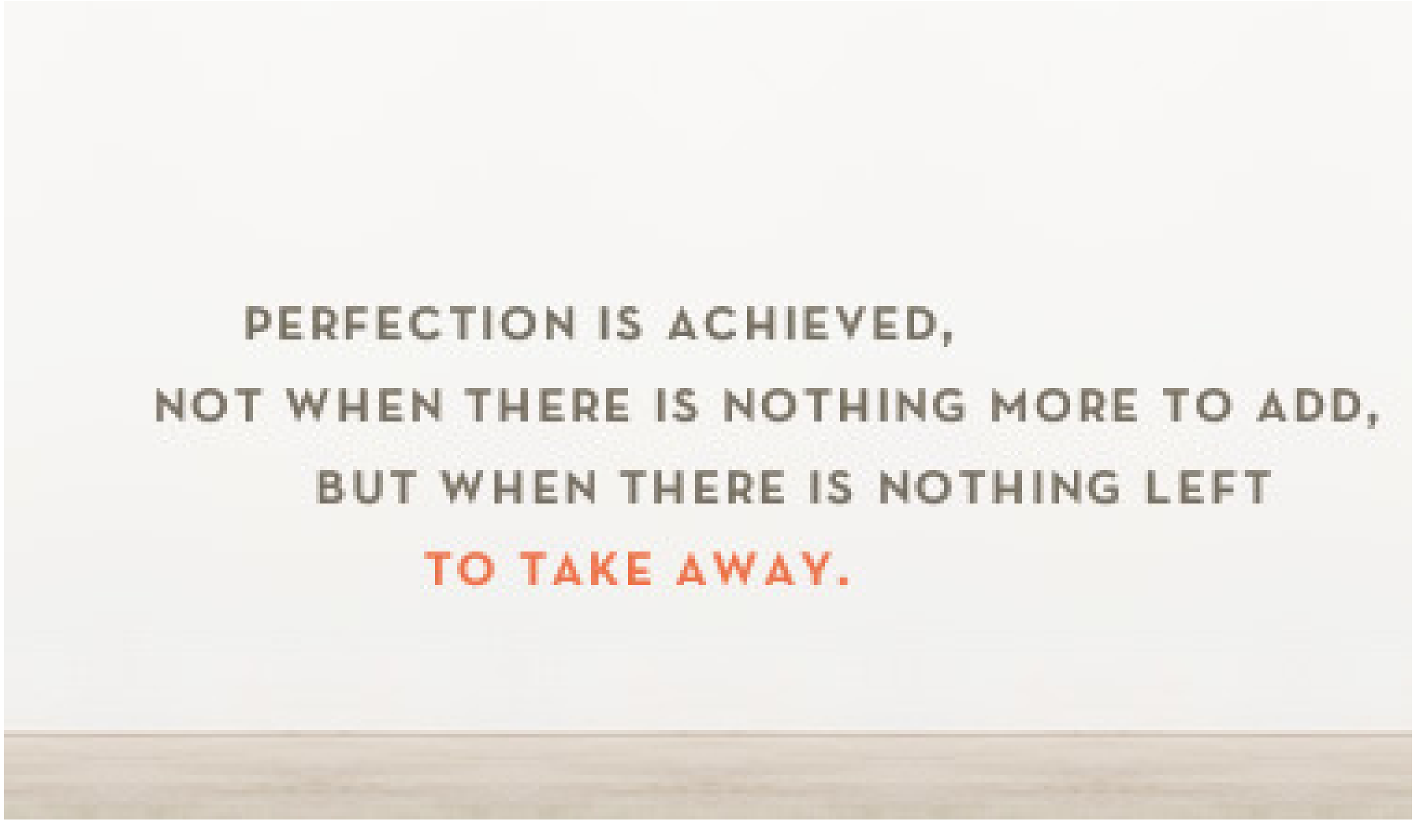


Purpose gives you a target



What the target is going to look like is still flexible.

Content

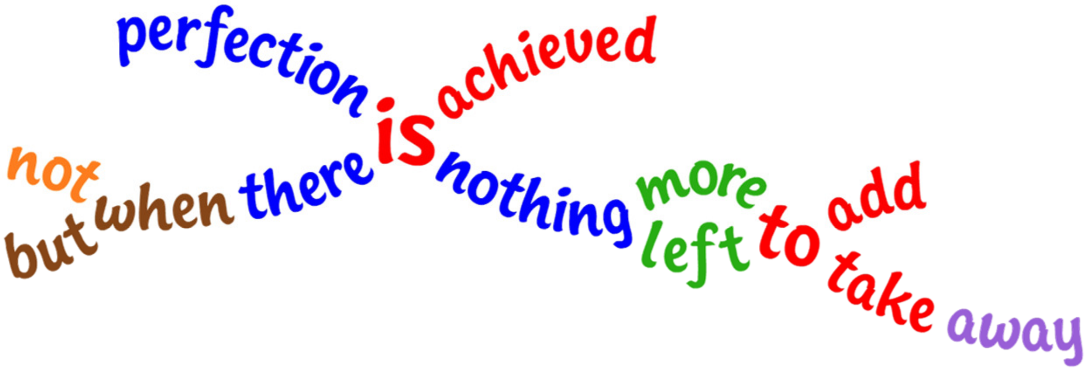


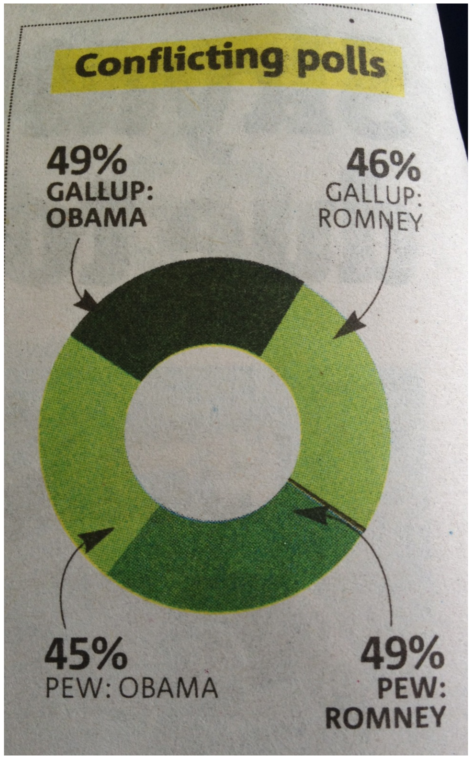
PERFECTION IS ACHIEVED,
NOT WHEN THERE IS NOTHING MORE TO ADD,
BUT WHEN THERE IS NOTHING LEFT
TO TAKE AWAY.

Only add the data needed for your purpose, no more.

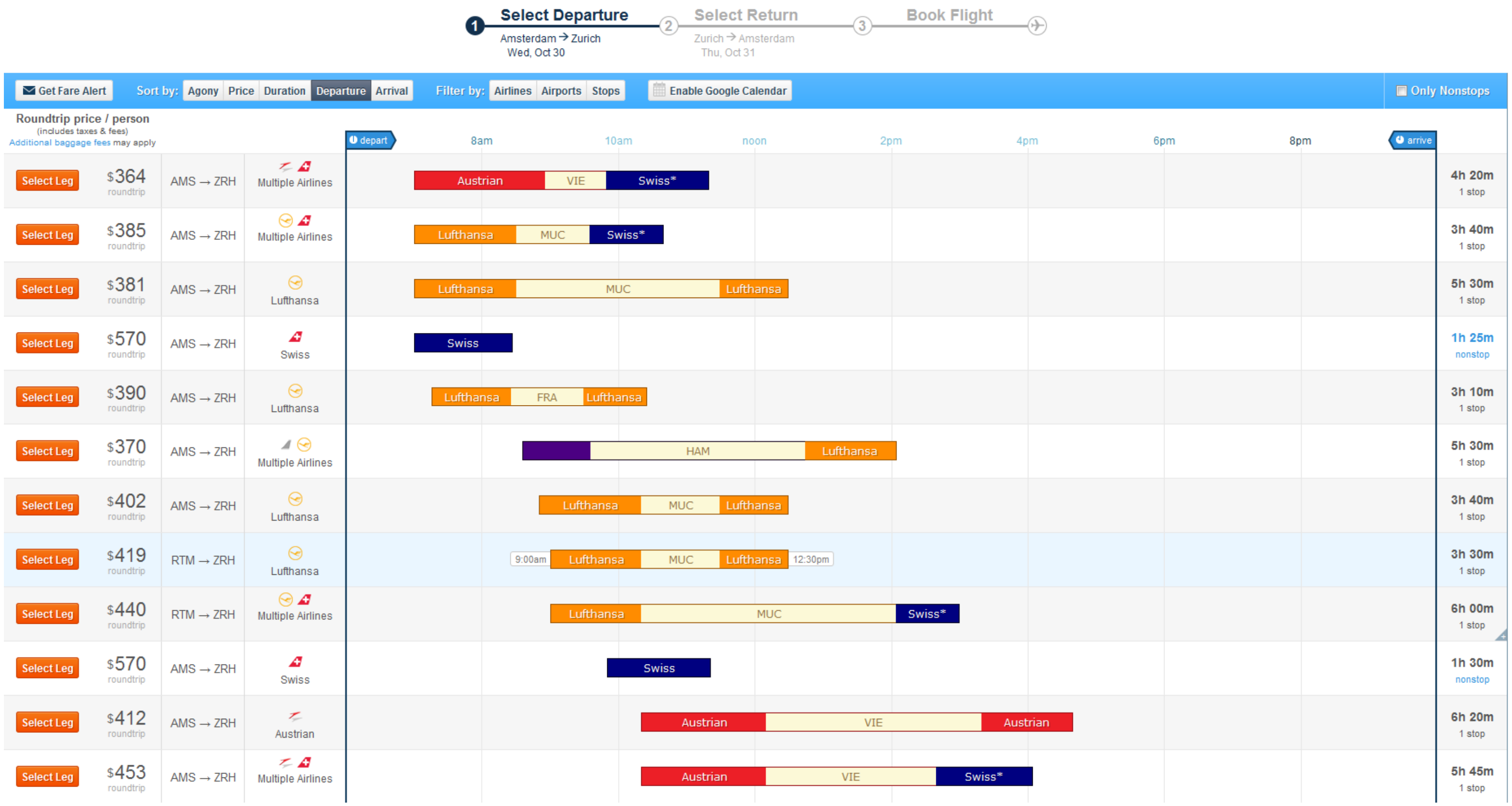
Try not to

Good content, but bad structure





Good structure



1 Select Departure — Amsterdam → Zurich, Wed, Oct 30
2 Select Return — Zurich → Amsterdam, Thu, Oct 31
3 Book Flight

Get Fare Alert | Sort by: Agony Price Duration **Departure** Arrival | Filter by: Airlines Airports Stops | Enable Google Calendar | □ Only Nonstops

Roundtrip price / person
(includes taxes & fees)
Additional baggage fees may apply

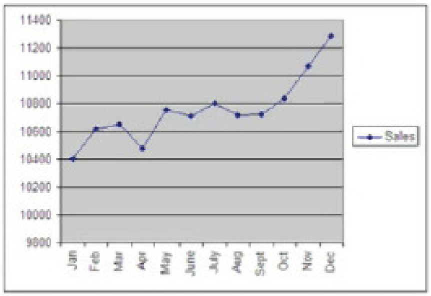
				depart · 8am · 10am · noon · 2pm · 4pm · 6pm · 8pm · arrive	
Select Leg	$364 roundtrip	AMS → ZRH	Multiple Airlines	Austrian · VIE · Swiss*	4h 20m 1 stop
Select Leg	$385 roundtrip	AMS → ZRH	Multiple Airlines	Lufthansa · MUC · Swiss*	3h 40m 1 stop
Select Leg	$381 roundtrip	AMS → ZRH	Lufthansa	Lufthansa · MUC · Lufthansa	5h 30m 1 stop
Select Leg	$570 roundtrip	AMS → ZRH	Swiss	Swiss	1h 25m nonstop
Select Leg	$390 roundtrip	AMS → ZRH	Lufthansa	Lufthansa · FRA · Lufthansa	3h 10m 1 stop
Select Leg	$370 roundtrip	AMS → ZRH	Multiple Airlines	HAM · Lufthansa	5h 30m 1 stop
Select Leg	$402 roundtrip	AMS → ZRH	Lufthansa	Lufthansa · MUC · Lufthansa	3h 40m 1 stop
Select Leg	$419 roundtrip	RTM → ZRH	Lufthansa	9:00am · Lufthansa · MUC · Lufthansa · 12:30pm	3h 30m 1 stop
Select Leg	$440 roundtrip	RTM → ZRH	Multiple Airlines	Lufthansa · MUC · Swiss*	6h 00m 1 stop
Select Leg	$570 roundtrip	AMS → ZRH	Swiss	Swiss	1h 30m nonstop
Select Leg	$412 roundtrip	AMS → ZRH	Austrian	Austrian · VIE · Austrian	6h 20m 1 stop
Select Leg	$453 roundtrip	AMS → ZRH	Multiple Airlines	Austrian · VIE · Swiss*	5h 45m 1 stop

Tips for good structure:
- Use position smartly
- Don’t use 3D (unless doing SciVis)
- Stick to defaults your audience already knows

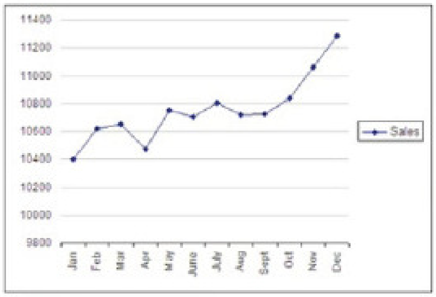
Formatting

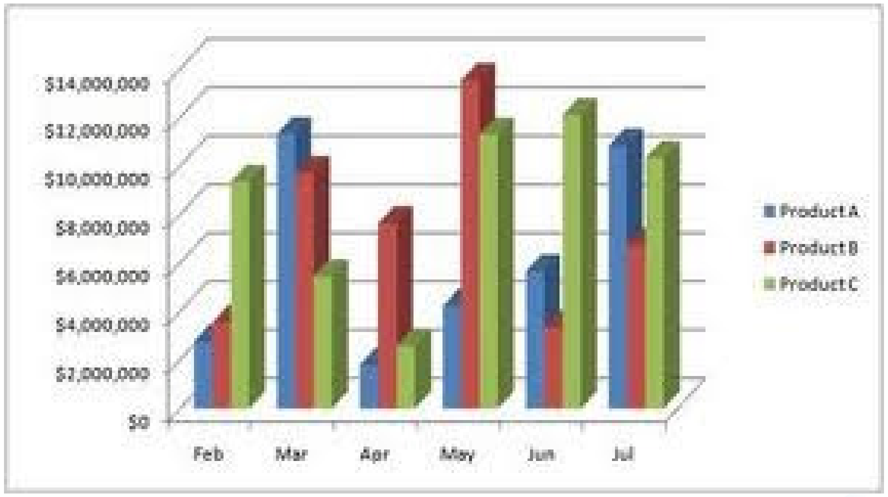
- The look and feel of your chart.
- Make sure important elements pop out visually.
- Reduce clutter to a minimum, less is more.
- Focus on legibility not on chart bling.
- Just because you could, doesn't mean you should.
- 3D and fake 3D don't help.

Bad

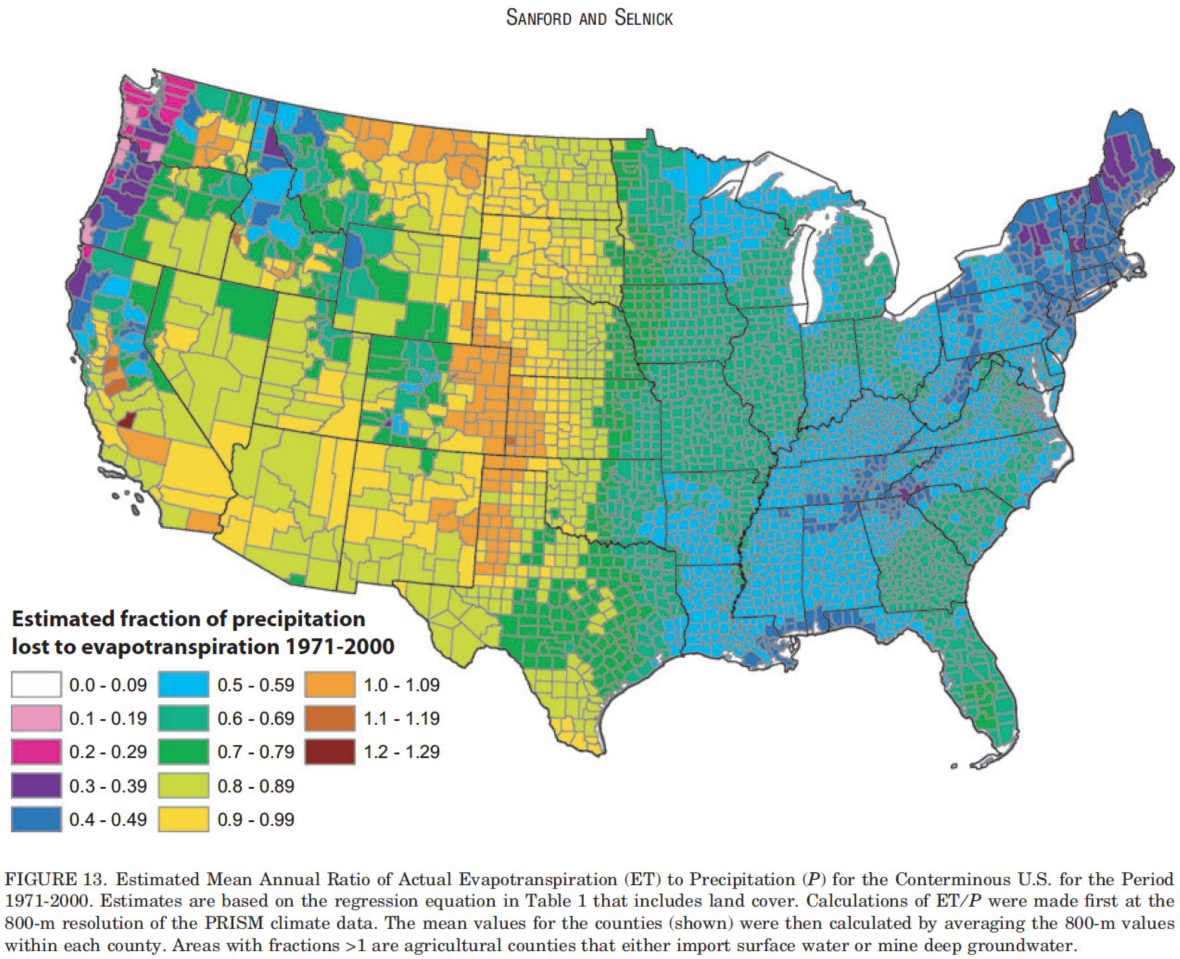


Good





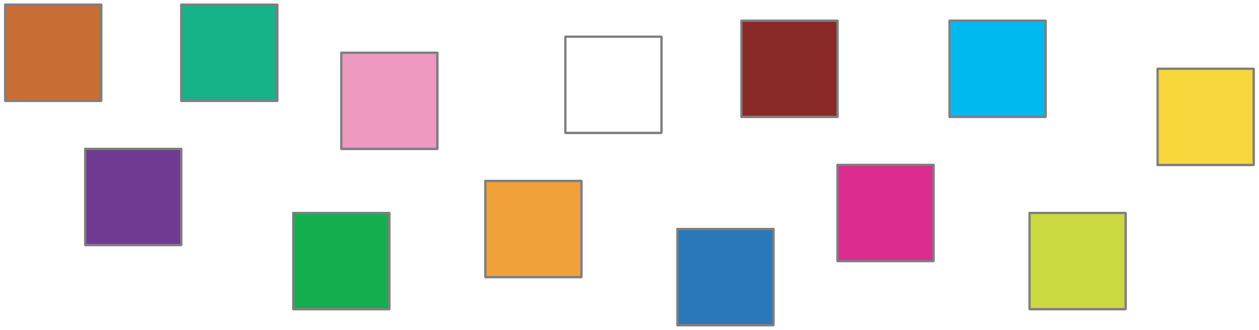
Good visual design but still bad formatting

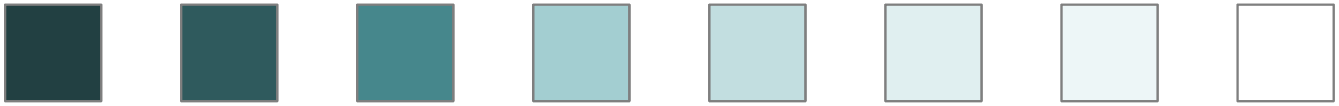


SANFORD AND SELNICK

FIGURE 13. Estimated Mean Annual Ratio of Actual Evapotranspiration (ET) to Precipitation (*P*) for the Conterminous U.S. for the Period 1971-2000. Estimates are based on the regression equation in Table 1 that includes land cover. Calculations of ET/*P* were made first at the 800-m resolution of the PRISM climate data. The mean values for the counties (shown) were then calculated by averaging the 800-m values within each county. Areas with fractions >1 are agricultural counties that either import surface water or mine deep groundwater.

Don’t use discrete colors for continuous values



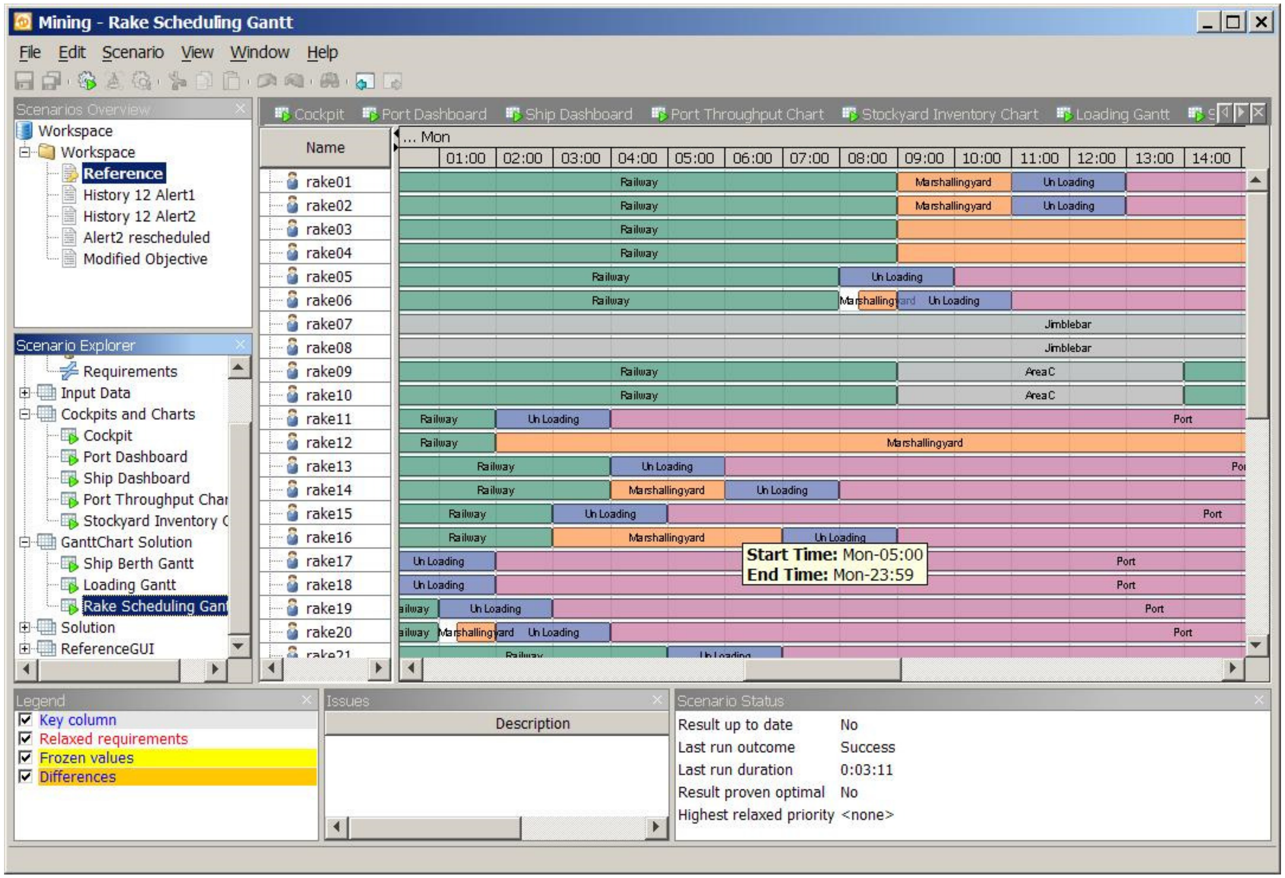


Use variations of the same color

Analysis usecase

- Main goal is to ‘understand’ data in order to use it to answer a hard question.
- Exactly how to achieve this answer is unknown at the start of analysis, but evolves throughout the process.
- Concrete samples
  - An financial investigator going over a large amount of transactions to find the cause of suspected fraud.
  - A sales professional charged with finding out why sales for a particular product are decreasing.
  - A medical insurance company looking for patterns in healthcare usage to optimize delivery.

Sample : optimization + visualization



Start Time: Mon-05:00
End Time: Mon-23:59

Legend:

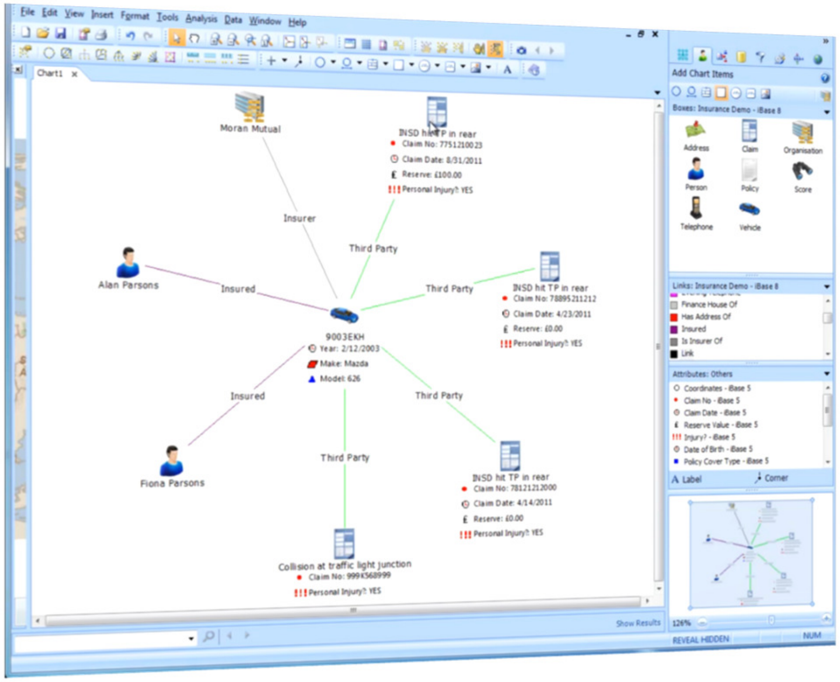
- ☑ Key column
- ☑ Relaxed requirements
- ☑ Frozen values
- ☑ Differences

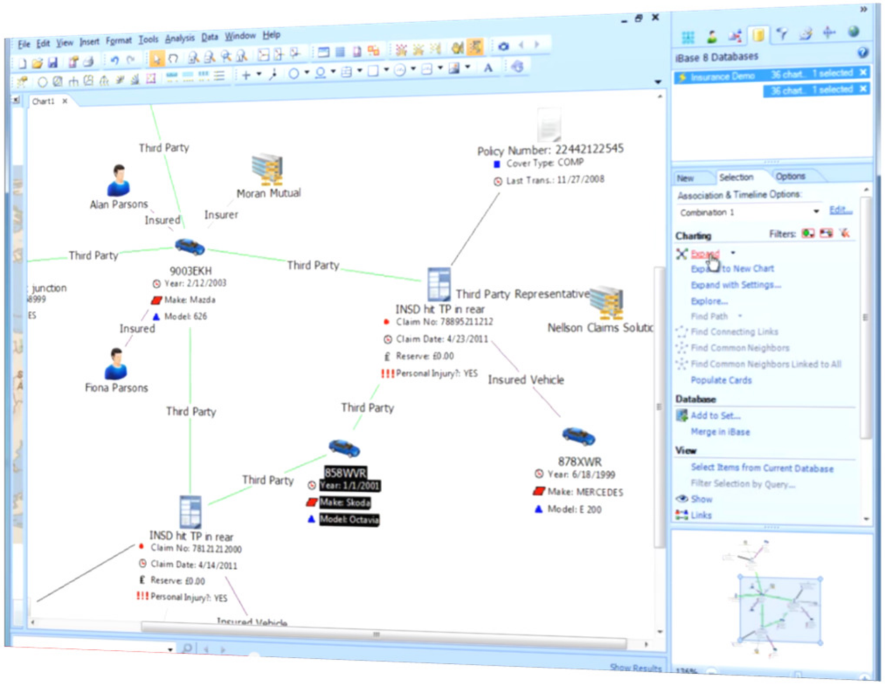
Scenario Status:

Result up to date	No
Last run outcome	Success
Last run duration	0:03:11
Result proven optimal	No
Highest relaxed priority	<none>

- Optimization algorithm computes an initial solution for a scenario.
- Users can interactively ‘tweak’ the optimization by changing constraints.
- Result of the optimization is displayed graphically. Users operate on the graphic.

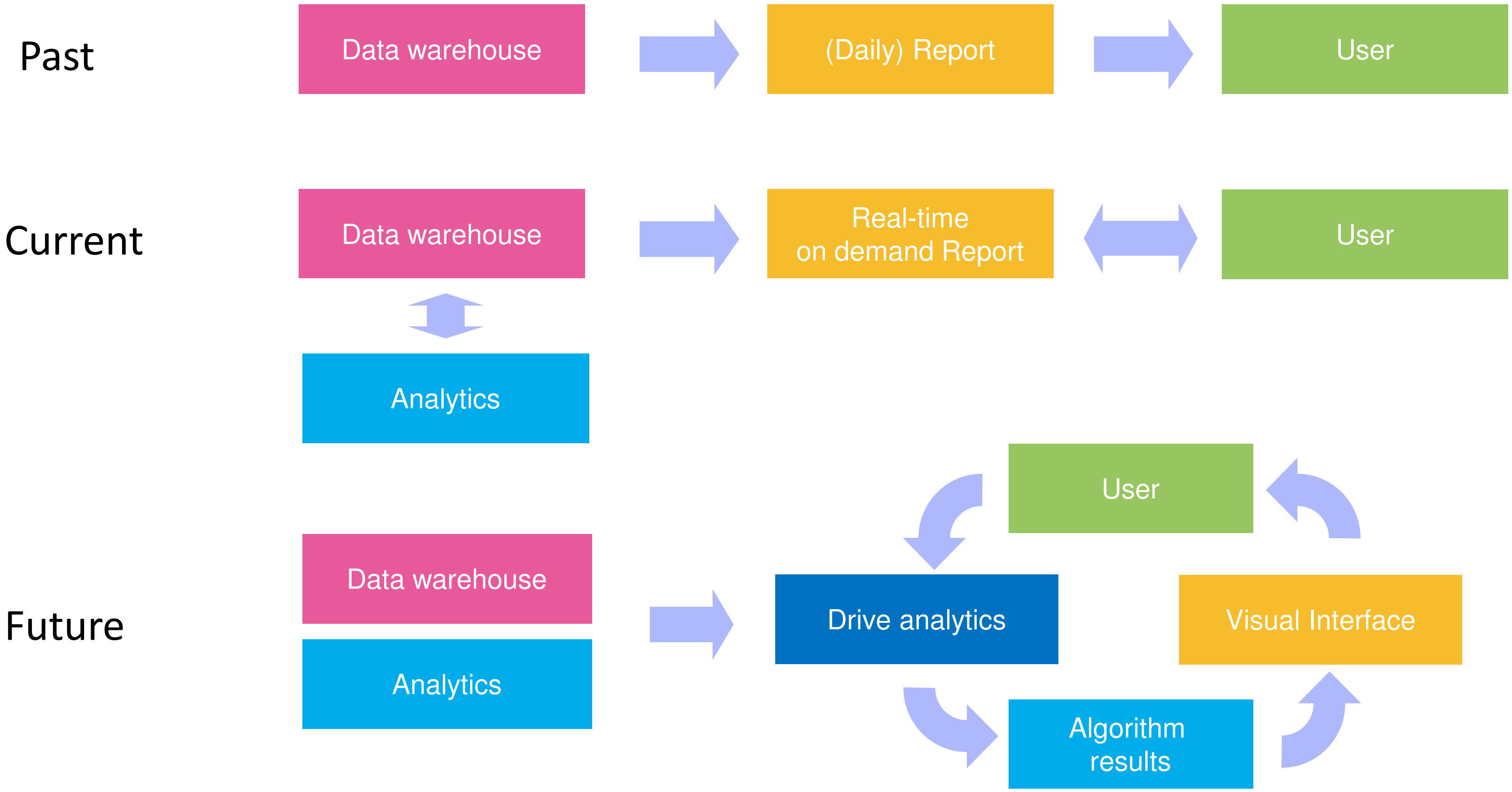
Sample : Insurance investigation





- Analytics can recommend potential fraudulent cases.
- Human investigator uses visualization to get a mental image of a particular case.
- Finds a common piece of information between two of the cases.

The visual analytics cycle



Past

Data warehouse

(Daily) Report

User

Current

Data warehouse

Real-time
on demand Report

User

Analytics

Future

Data warehouse

Analytics

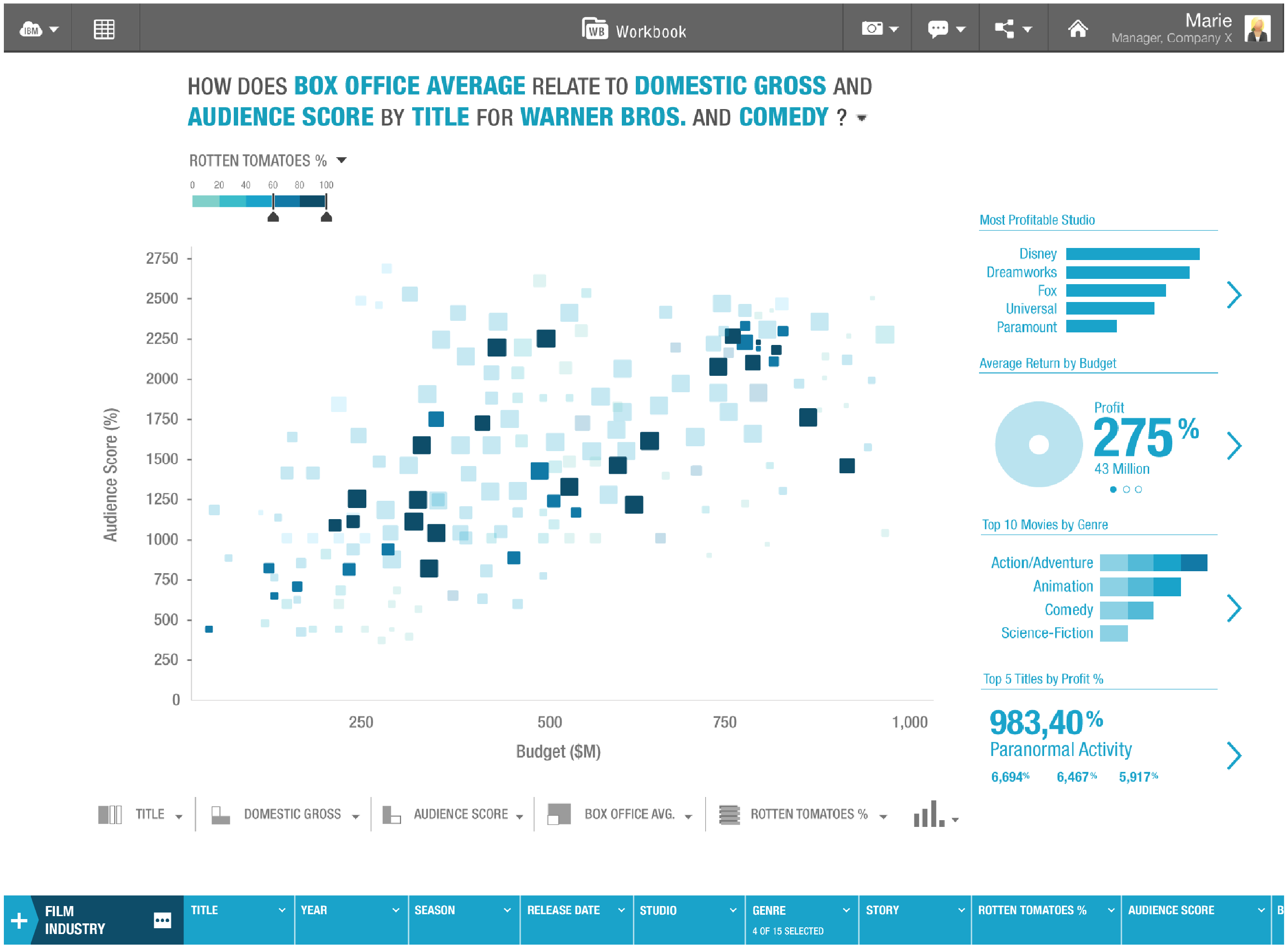
User

Drive analytics

Visual Interface

Algorithm
results

BA is pushing into the Visual Analytics space as well

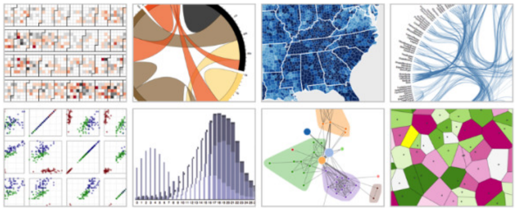


HOW DOES BOX OFFICE AVERAGE RELATE TO DOMESTIC GROSS AND AUDIENCE SCORE BY TITLE FOR WARNER BROS. AND COMEDY ?

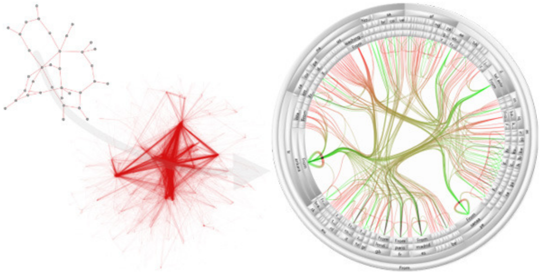
More Info @ Information on Demand – Las Vegas 3/11 – 7/11

Where is this going in the next years

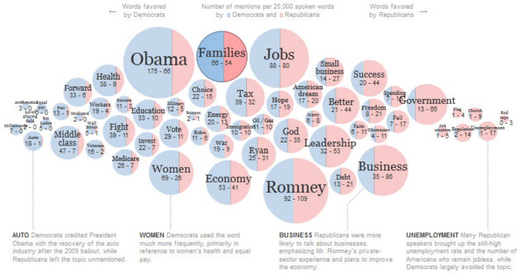
- **Tooling improvements** : Domain specific languages (D3, IBM RAVE) and tools to bridge the gap between design and implementation.



- **Network / relationship data** : Showing how multiple entities relate to each other and make that work for big data.

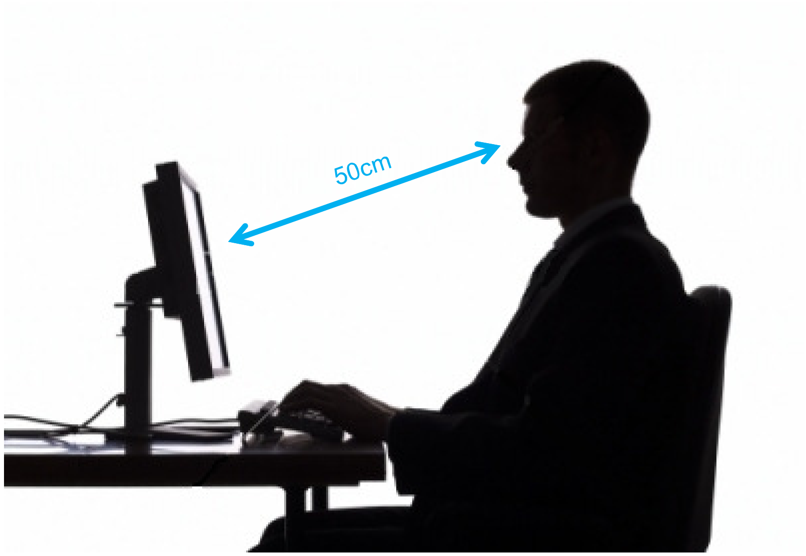


- **Text / unstructured data** : The majority of the world's information is not stored in structured databases.



- **Dealing with uncertainty**: How to convey the veracity of information visually?

Visualization : a medium for the last 50 cm.



References in this document to IBM products or services do not imply that IBM intends to make them available in every country.

Thank You!