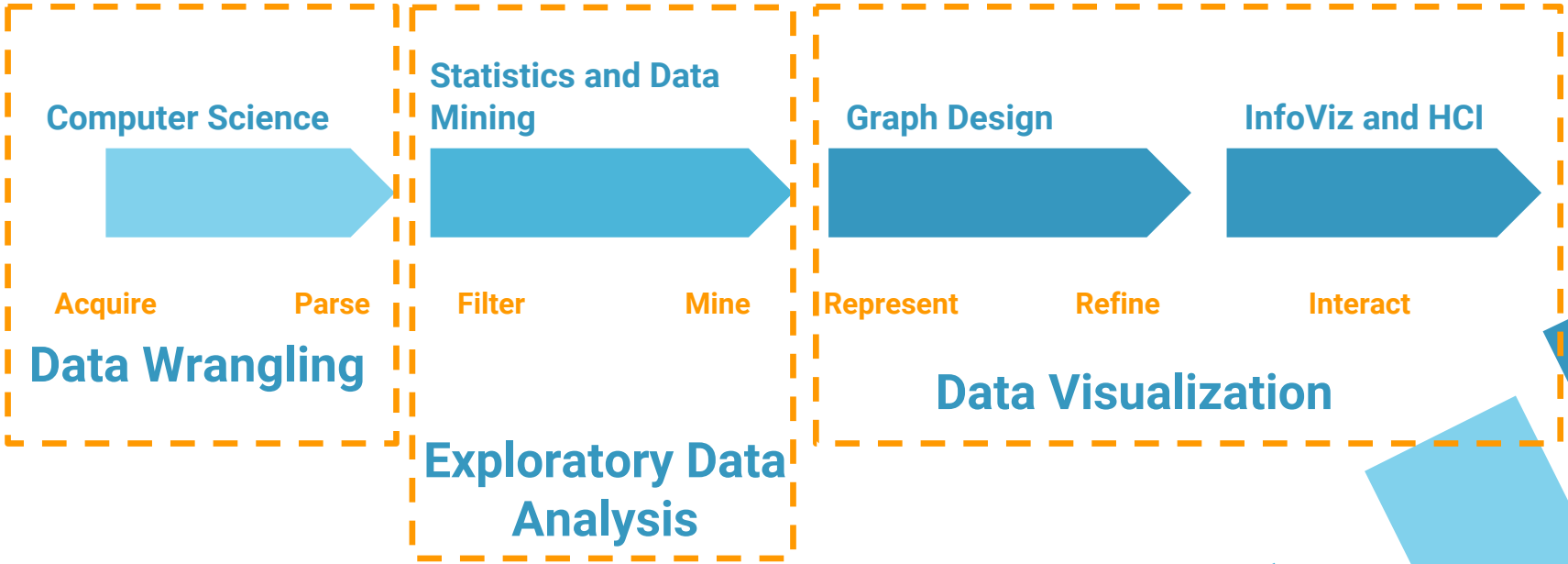
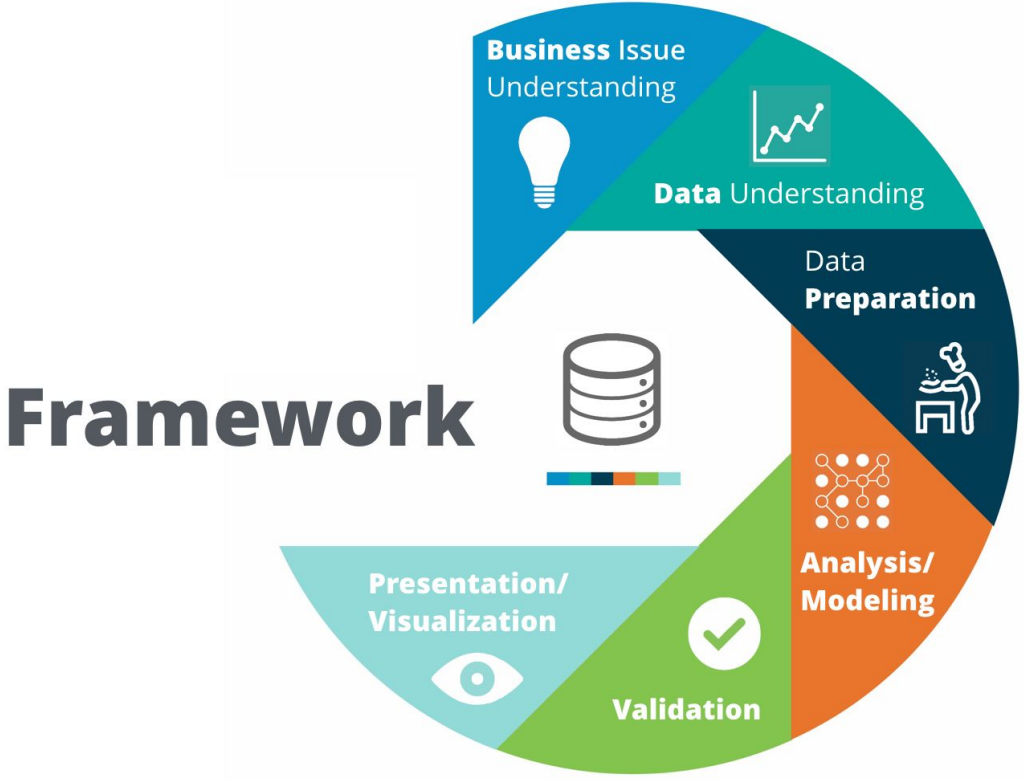
Let’s talk Data Science

Data Science Process



Data Science Process



Modern Data Scientist

A **data scientist** is a professional responsible for collecting, analyzing and interpreting large amounts of data to identify ways to help a business improve operations and gain a competitive edge over rivals.

Data Scientist is the sexiest job of the 21th century.

MODERN DATA SCIENTIST

Data Scientist, the sexiest job of 21th century requires a mixture of multidisciplinary skills ranging from an intersection of mathematics, statistics, computer science, communication and business. Finding a data scientist is hard. Finding people who understand who a data scientist is, is equally hard. So here is a little cheat sheet on who the modern data scientist really is.

MATH & STATISTICS

- Machine learning
- Statistical modeling
- Experiment design
- Bayesian inference
- Supervised learning: decision trees, random forests, logistic regression
- Unsupervised learning: clustering, dimensionality reduction
- Optimization: gradient descent and variants

PROGRAMMING & DATABASE

- Computer science fundamentals
- Scripting language e.g. Python
- Statistical computing package e.g. R
- Databases SQL and NoSQL
- Relational algebra
- Parallel databases and parallel query processing
- MapReduce concepts
- Hadoop and Hive/Pig
- Custom reducers
- Experience with xaaS like AWS

DOMAIN KNOWLEDGE & SOFT SKILLS

- Passionate about the business
- Curious about data
- Influence without authority
- Hacker mindset
- Problem solver
- Strategic, proactive, creative, innovative and collaborative

COMMUNICATION & VISUALIZATION

- Able to engage with senior management
- Story telling skills
- Translate data-driven insights into decisions and actions
- Visual art design
- R packages like ggplot or lattice
- Knowledge of any of visualization tools e.g. Flare, D3.js, Tableau

« Explatory »
Visualization

All visualizations are ‘explatory’

Explanatory

» Try to explain something to the viewer
» Tell a story to an audience
» Something specific you want to communicate

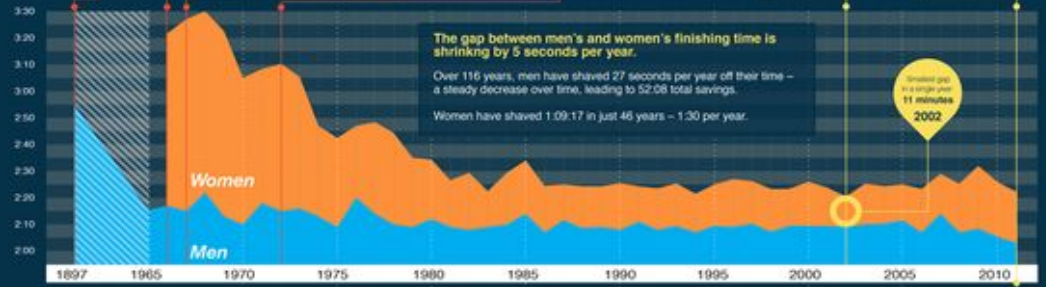
» Exploratory

» Analysis you do to get familiar with your data
» Trying to validate or not an hypothesis or a question

« Exploratory is the process of turning over 100 rocks to find perhaps 1 or 2 precious gemstones »

THE BOSTON MARATHON:
WOMEN ARE CLOSING THE GAP

Women got a late start to the Boston Marathon, but they've been making up time ever since – literally. While the men have had two course records in two years – shaving several minutes off the previous record times, women have overall done a much faster job of improving speed and race performance. If the trend continues, women will catch up to the men by about 2035.

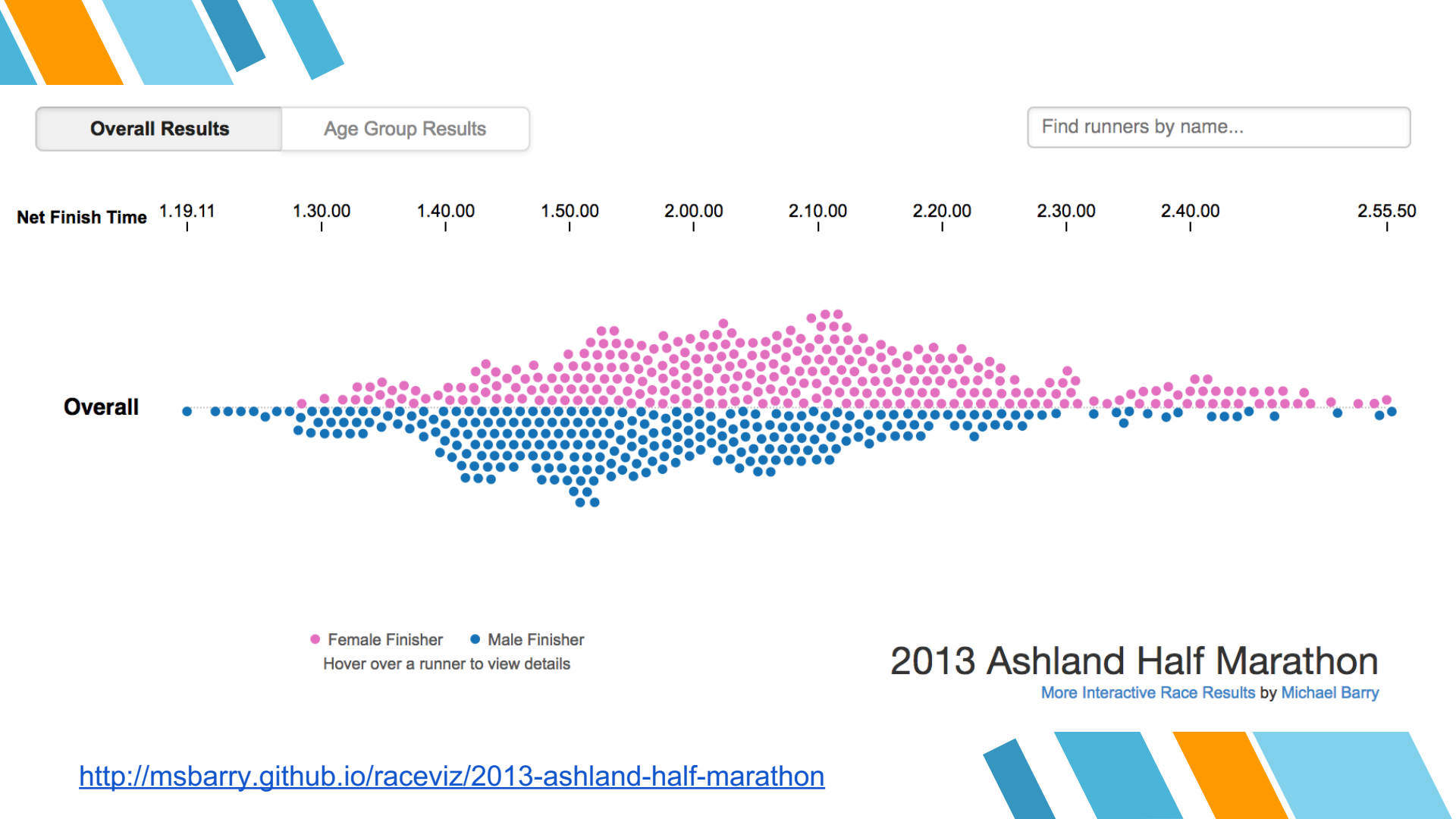


Despite It Being "Too Strenuous For Women" They Find a Way to Finish the Race

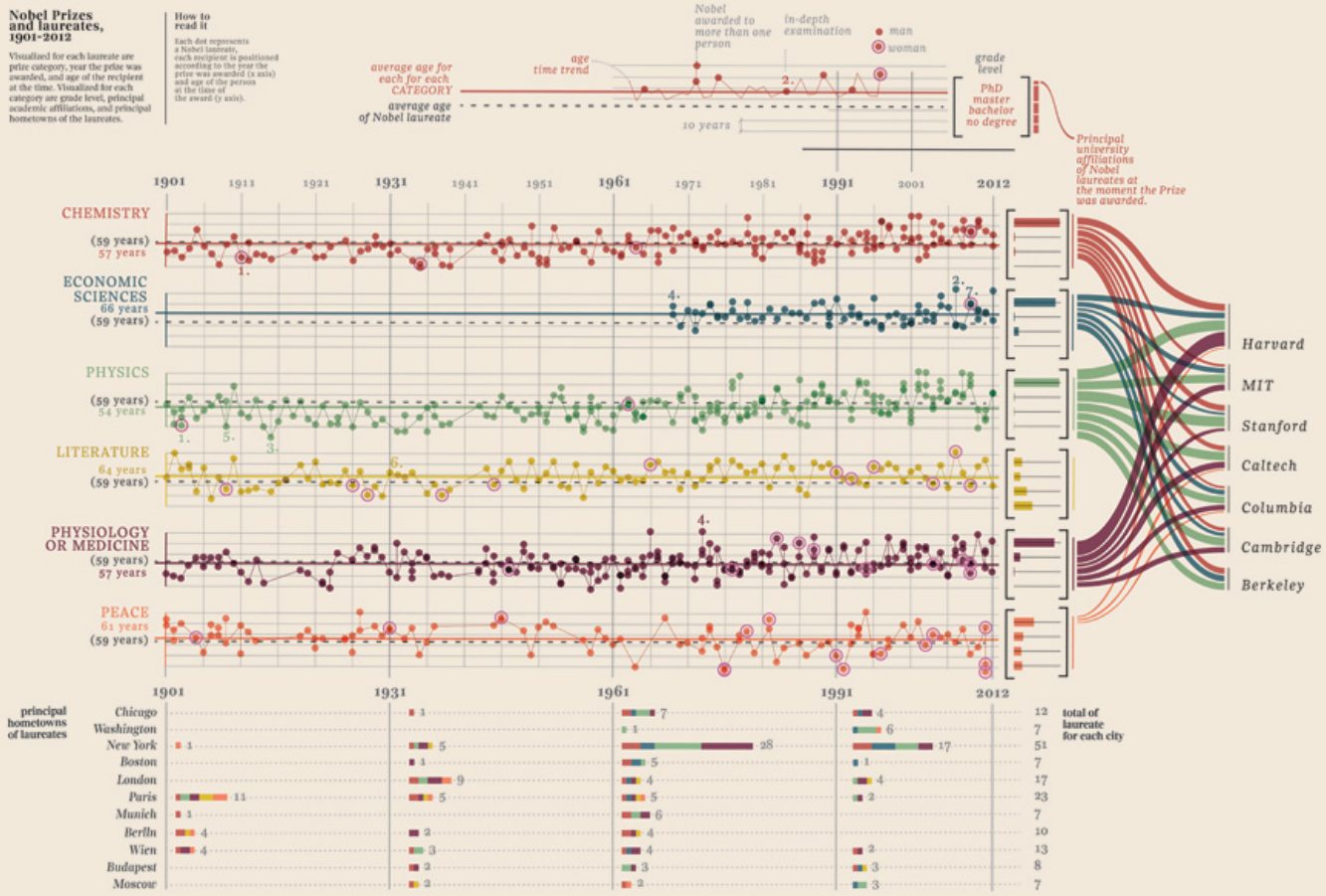
Long distance running was long denied to women. One reason was that it was felt to be "too strenuous" for women. One might expect finish rates to reflect this – that more women would quit the race than men. While this was somewhat true in the early years, the data shows that finish rates are only slightly higher for men – and that gap is narrowing every year.

BeehiveMedia

Most of this data is for the elite runners – the top of the pack. It is very possible that the AVERAGE performance of men and women is either more skewed or more equal than that provided here. We didn't have access to that data so couldn't compare.



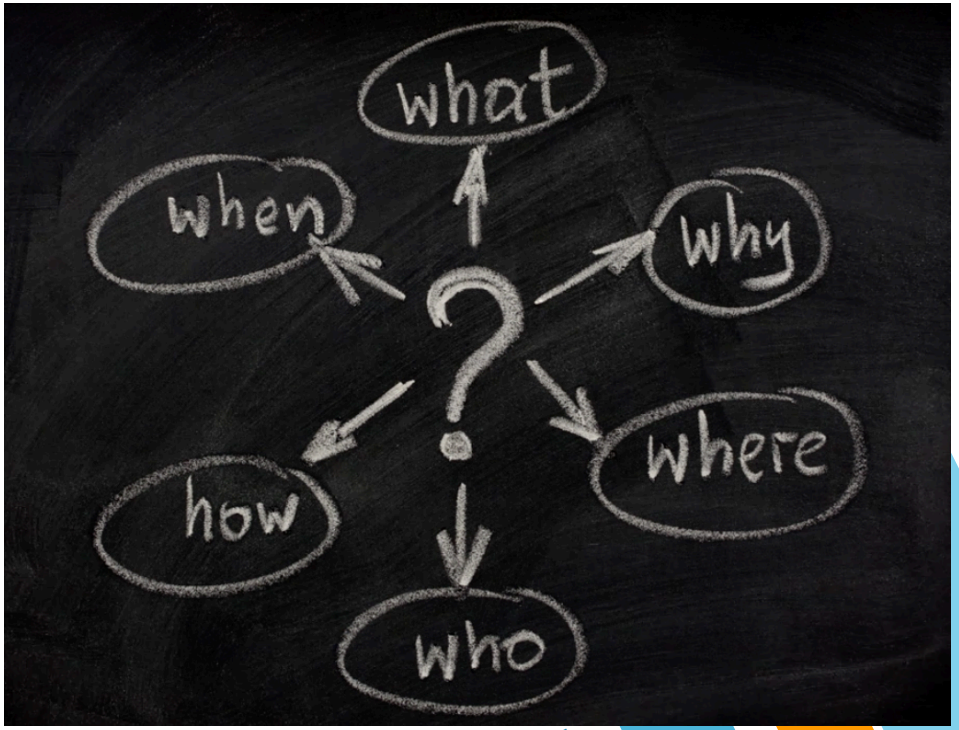
http://msbarry.github.io/raceviz/2013-ashland-half-marathon


Nobel Prizes and laureates, 1901-2012
Visualized for each laureate are prize category, year the prize was awarded, and age of the recipient at the time. Visualized for each category are grade level, principal academic affiliations, and principal hometowns of the laureates.
How to read it
Each dot represents a Nobel laureate, each recipient is positioned according to the year the prize was awarded (x axis) and age of the person at the time of the award (y axis).
average age for each for each CATEGORY
average age of Nobel laureate
age time trend
Nobel awarded to more than one person
in-depth examination
man
woman
grade level
PhD
master
bachelor
no degree
10 years
Principal university affiliations of Nobel laureates at the moment the Prize was awarded.
1901
1911
1921
1931
1941
1951
1961
1971
1981
1991
2001
2012
CHEMISTRY
(59 years)
57 years
ECONOMIC SCIENCES
66 years
(59 years)
PHYSICS
(59 years)
54 years
LITERATURE
64 years
(59 years)
PHYSIOLOGY OR MEDICINE
(59 years)
57 years
PEACE
61 years
(59 years)
Harvard
MIT
Stanford
Caltech
Columbia
Cambridge
Berkeley
principal hometowns of laureates
Chicago
Washington
New York
Boston
London
Paris
Munich
Berlin
Wien
Budapest
Moscow
total of laureate for each city

Unveiling Information Hierarchy

The 6 Ws

» Great way to organize and think about any story.



The 6 Ws – Hospital Use Case

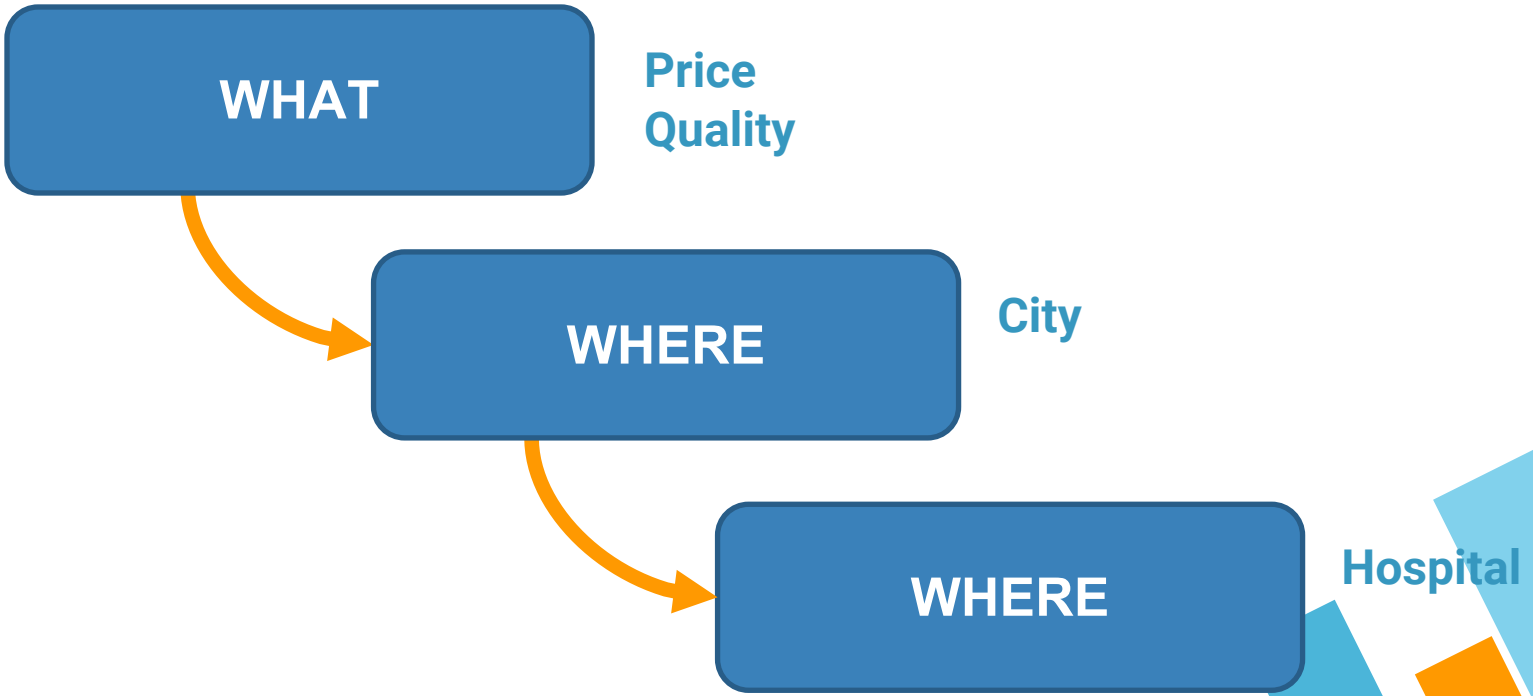
Hospital Pricing Data

» Where can I go for a specific treatment at a decent price AND good quality?

Applying the 6 Ws

W	Status
Who	Granular Answer
What	How I'll judge
When	
Where	Answer
Why	
How	

The 6 Ws – Hospital Use Case



Three more Ws

What's Wanting? (Missing)

» Keep calm and power on
» Go back to the source
» Generate more data

What in the World?

» More data from other sources
» Provide context
» Complete the picture

What's Wild?

» Unique visual approach
» Unexpected contextual data
» Out of the box interactive experience

Challenge

Obesity vs. Education

« Obesity is on average inversely proportional to the average education of the population »

Can you show the evidence for this assertion?

- Use the Tableau Public dashboard or
- The online version: http://www.vizwiz.com/2013/01/alberto-cairo-three-steps-to-become.html

Challenge

Obesity vs. Education

- What is the average percentage of obese people in all states? *(leave off the percent sign)*

  ____ %

- Which state has the largest percentage of its population that obtained a BA (Bachelor's degree or higher)? *(enter the state's two letter abbreviation)*

  ____

- A majority of states have greater percentages of *(choose one)*

  - ○ *obese people*
  - ○ *educated people*

THANKS!

Any questions?

You can find me at

- @manuGenard
- genard@fr.ibm.com