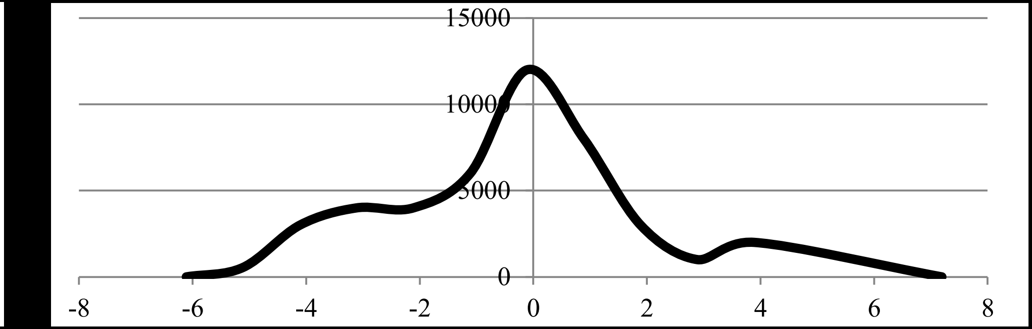


Exploring and Exploiting Interestingness in Data Science

Kirk Borne

@KirkDBorne

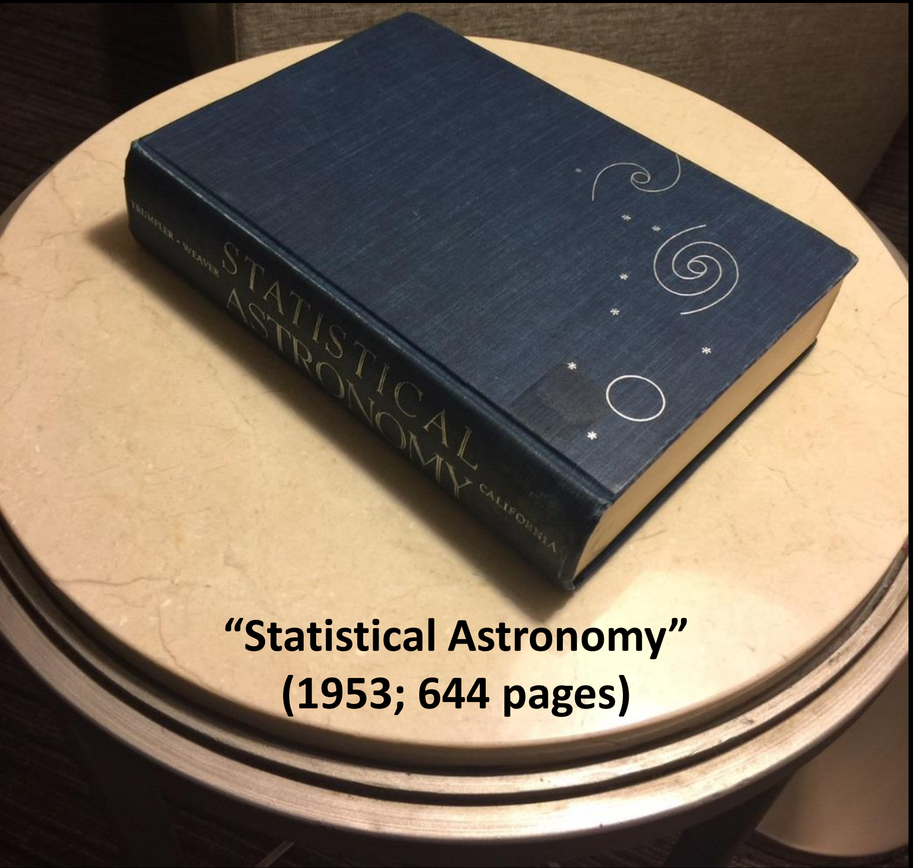
Principal Data Scientist

Booz Allen Hamilton

http://www.boozallen.com/datascience

Astronomy + Data + Statistics =

Long-time friends and acquaintances!

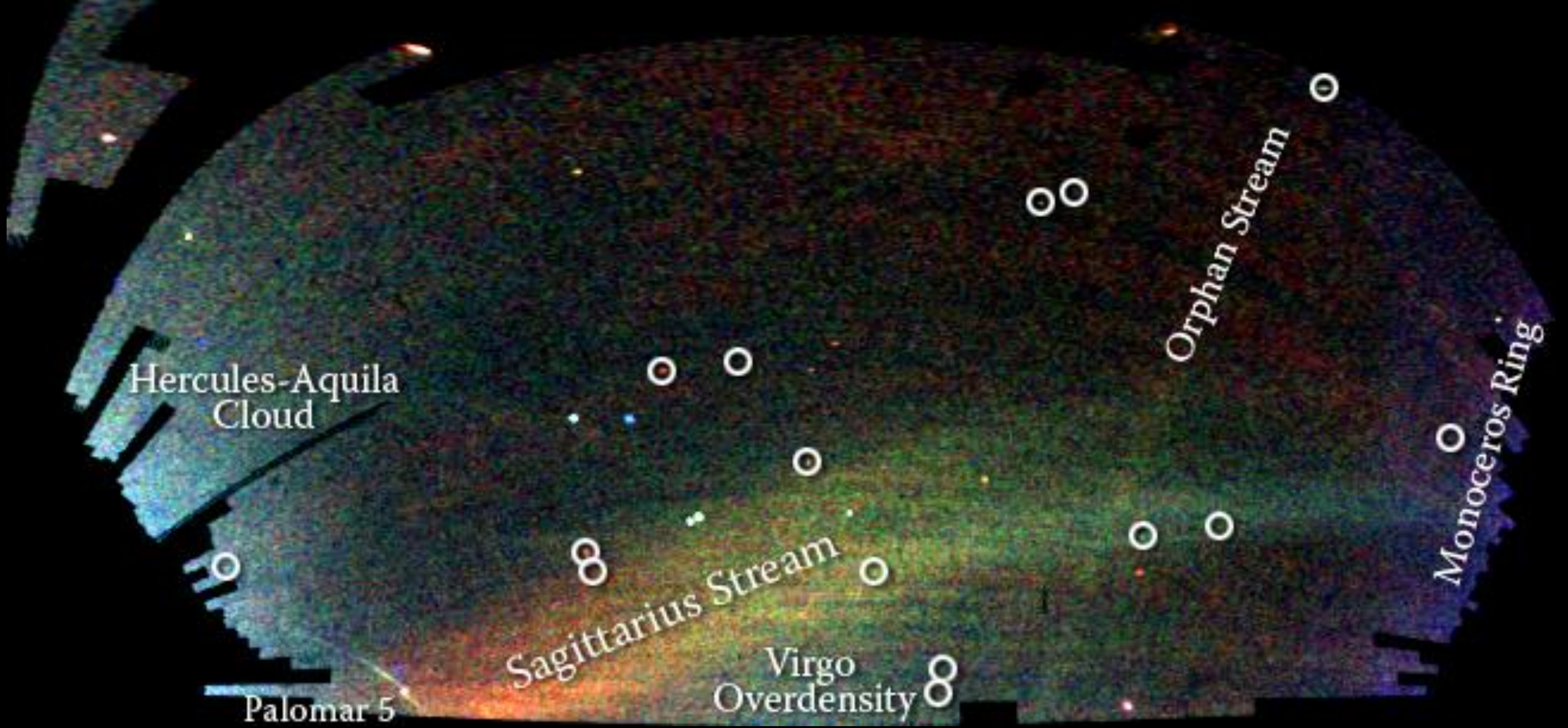


"Statistical Astronomy"
(1953; 644 pages)

Now there are new fields of research and education in Astronomy :
Astrostatistics and
Astroinformatics
http://asaip.psu.edu/

Sniffing out cold cases with DOGs:

Difference of Gaussians discovers Field of Streams around Milky Way galaxy



http://classic.sdss.org/includes/sideimages/fos_dr6_marked.html

Interestingness in Data:

Moving beyond Outlier Detection to Surprise Discovery!

- Outlier Detection 1.0 = Distance-based
- Outlier Detection 2.0 = Density-based
- Outlier Detection 3.0 = Pattern-based:
  - Finding the interesting, unexpected pattern (trend, correlation, change-point, segment, precursor signal, association) in your data
  - To facilitate more insightful ‘data-to-action’

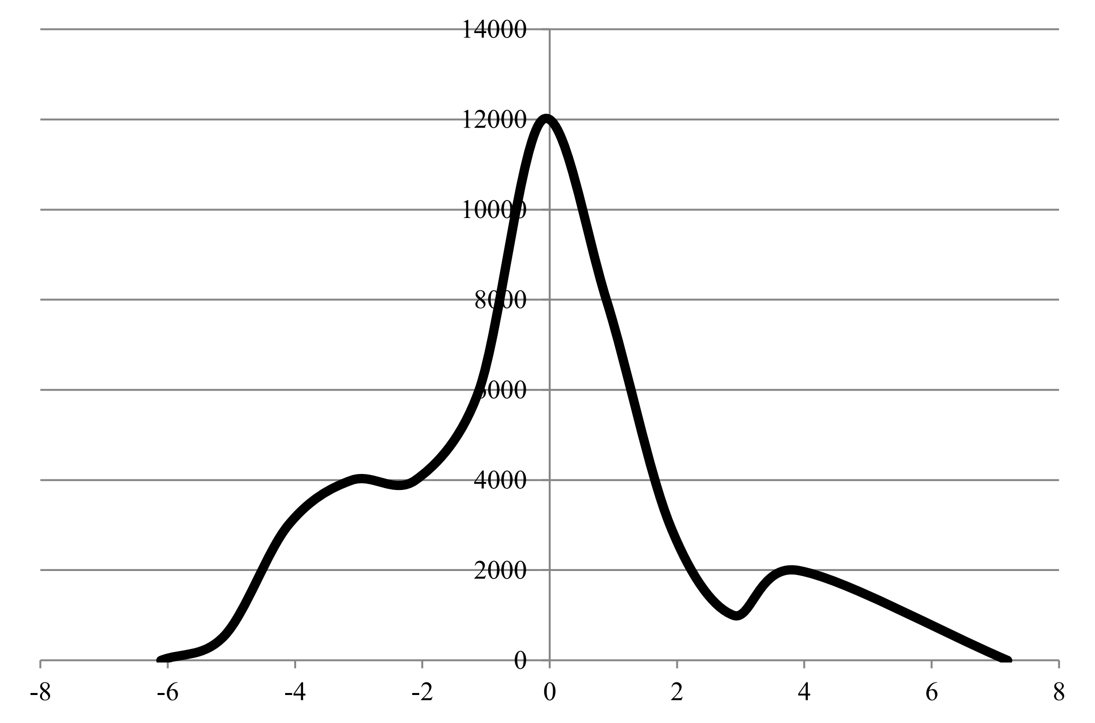
The Data Science Revolution =

Moving from data to insight to action!

All of the features in the data histogram convey valuable (actionable) information

(the long tail, outliers, multi-modal peaks, …)

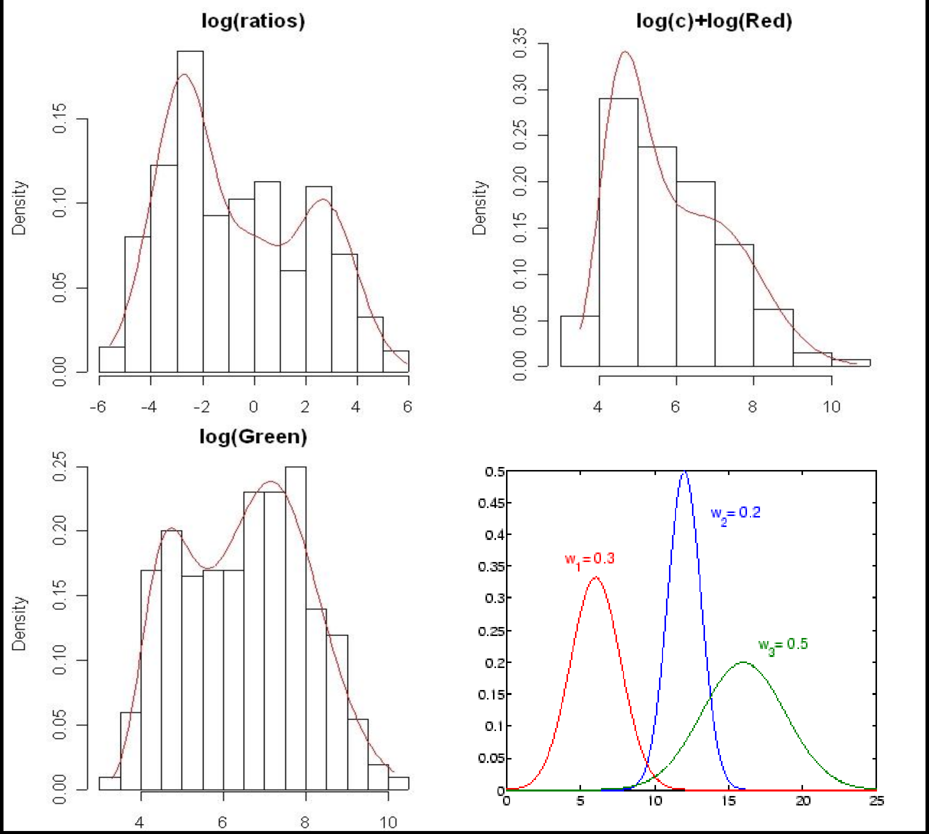
I ♥ DATA



Mixture Models = Statistical Clustering

- Each of these data histograms can be represented by the mixture (i.e., sum) of several Gaussian normal distributions, such as the 3 Gaussian distributions shown in the lower right.

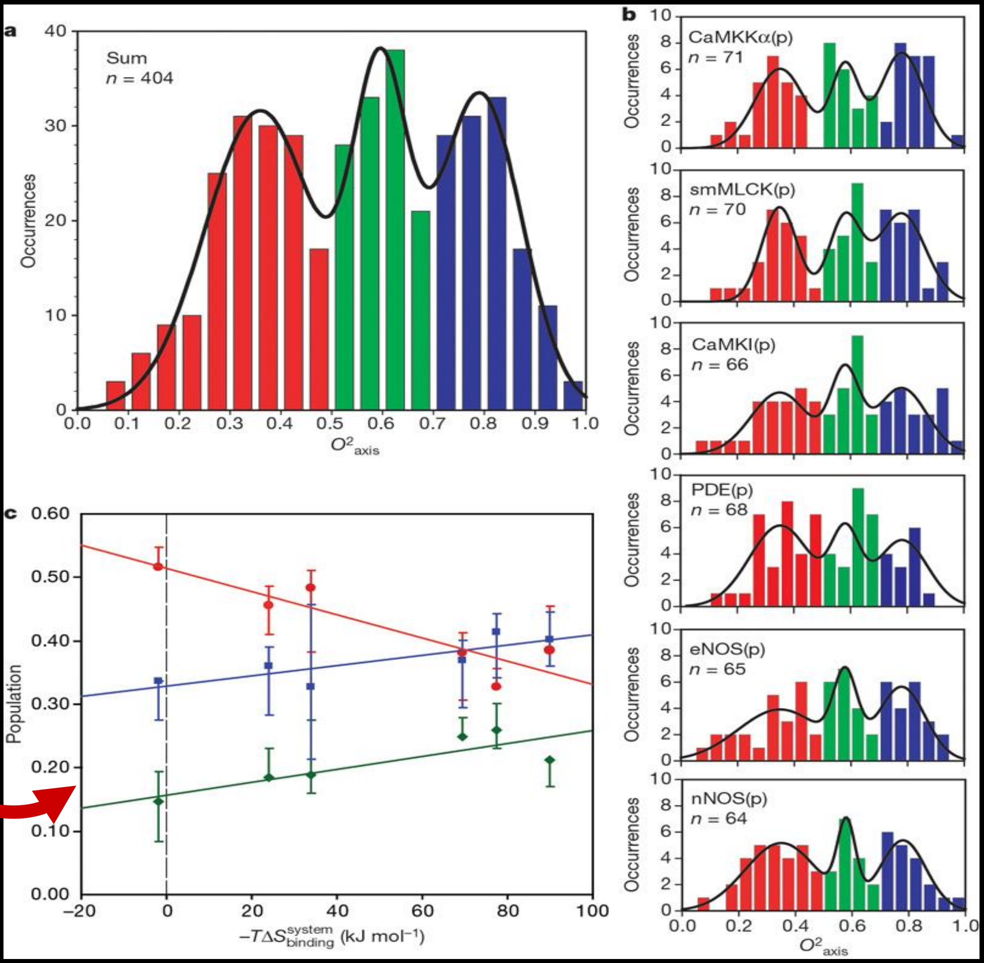
- Each Gaussian statistically represents (characterizes) one "cluster" of data values within the full set of data values.



Comprehensive web resource for Mixture Models for clustering and unsupervised learning in Data Mining:
http://www.csse.monash.edu.au/~dld/mixture.modelling.page.html

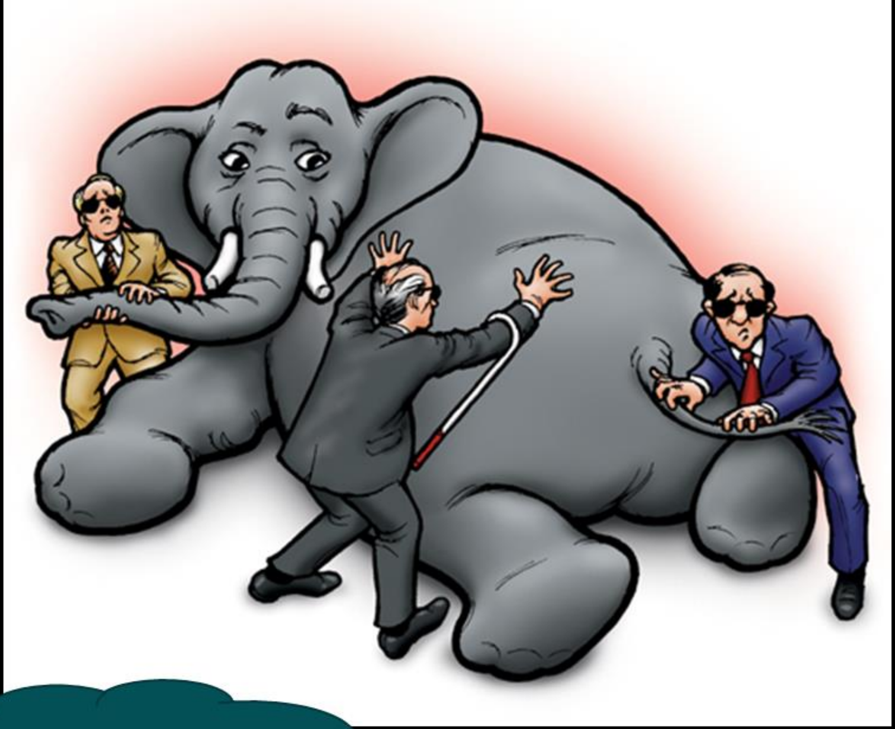
Statistical Clustering tags (characterizes) the data, enabling discovery: making the data “smart”!

- Each Gaussian in the mixture can be characterized by various parameters, such as the mean, variance (standard deviation), and amplitude (i.e., the strength of that particular Gaussian component within the mixture).
- These parameters can be plotted as a function of some independent (treatment) variable, to **discover trends and correlations** in the effects across the different segments of the population.



https://www.researchgate.net/publication/6200224_Conformational_entropy_in_molecular_recognition_by_proteins

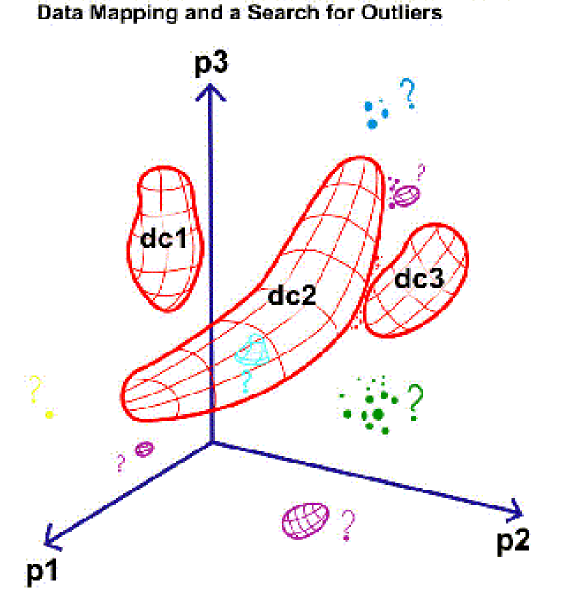
Massive data collections unlock deeper insights into hard problems and complex systems



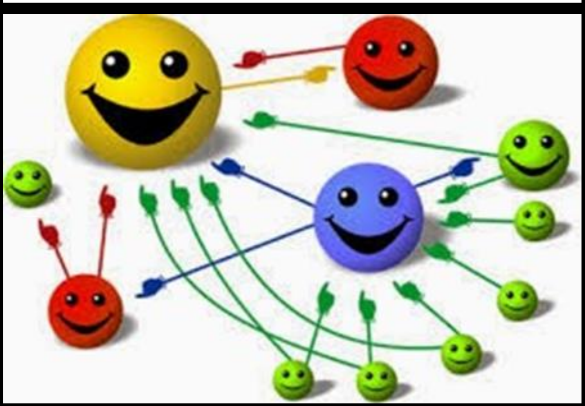
I wish we had more data…

…Be careful what you wish for!!!!

4 Types of Machine Learning Discovery from Data:



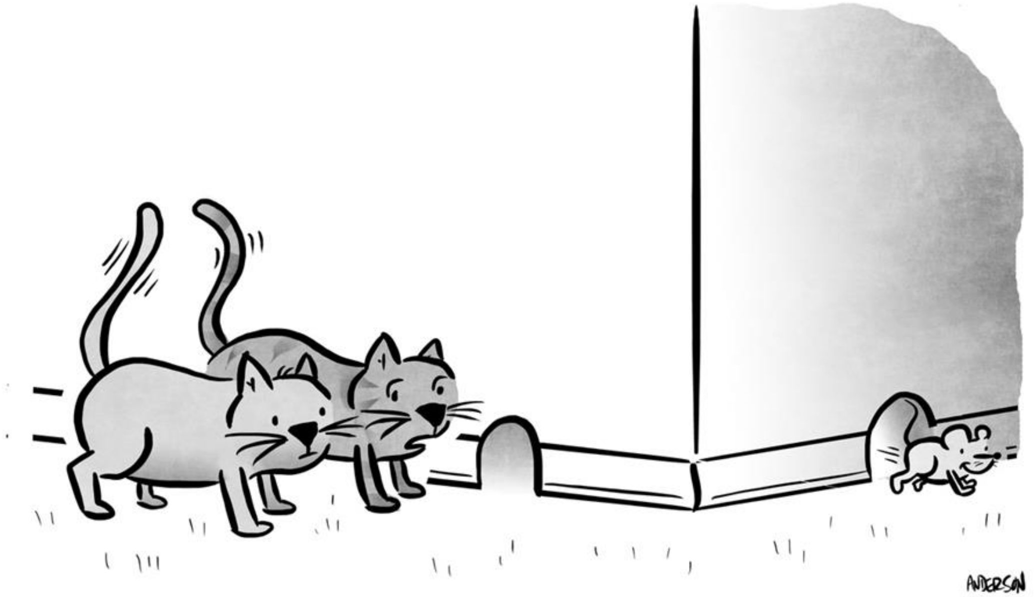
(Graphic by S. G. Djorgovski, Caltech)



1) **Class Discovery:** Find the categories of objects (population segments), events, and behaviors in your data. + Learn the rules that constrain the class boundaries (that uniquely distinguish them).
2) **Correlation (Predictive and Prescriptive Power) Discovery:** (insights discovery) – Find trends, patterns, dependencies in data that reveal the governing principles or behavioral patterns (the object's "DNA").
3) **Outlier / Anomaly / Novelty / Surprise Discovery:** Find the new, surprising, unexpected one-in-a-[million / billion / trillion] object, event, or behavior.
4) **Association (or Link) Discovery:** (Graph and Network Analytics) – Find both the typical (usual) and the atypical (unusual, interesting) data associations / links / connections in your domain.

Examples of Interestingness in Data

1) Outliers
2) Counting
3) Mapping
4) Associations
5) Linking
6) Clustering
7) Looking

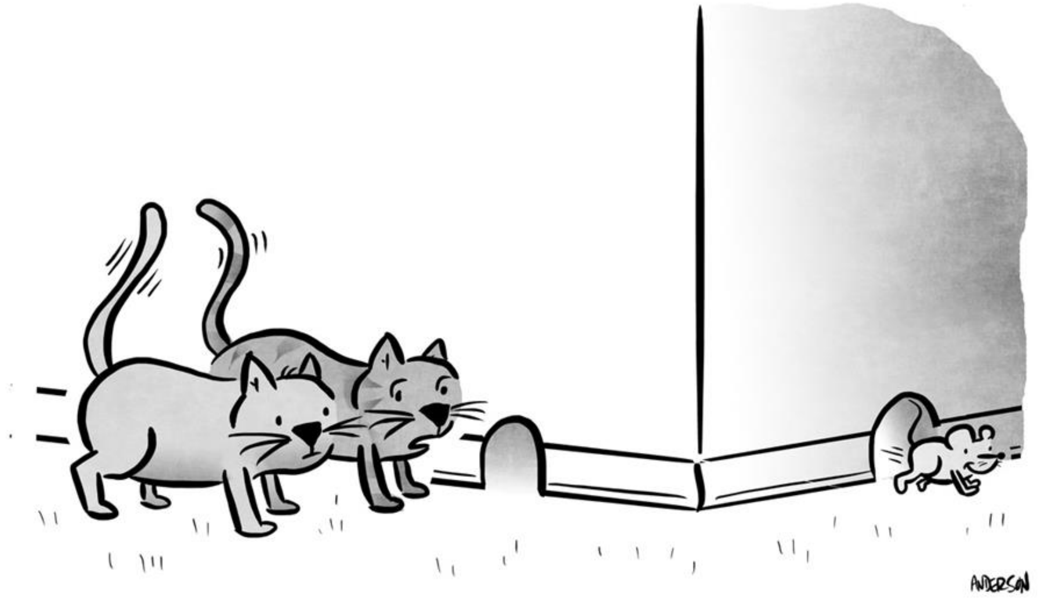


"According to our current predictive analytics solution, the mouse should be exiting from this hole in 3... 2... 1..." #betterdata

Source for image: https://www.hausmanmarketingletter.com/translating-analytics-to-action/

Examples of Interestingness in Data

1) **Outliers**
2) Counting
3) Mapping
4) Associations
5) Linking
6) Clustering
7) Looking

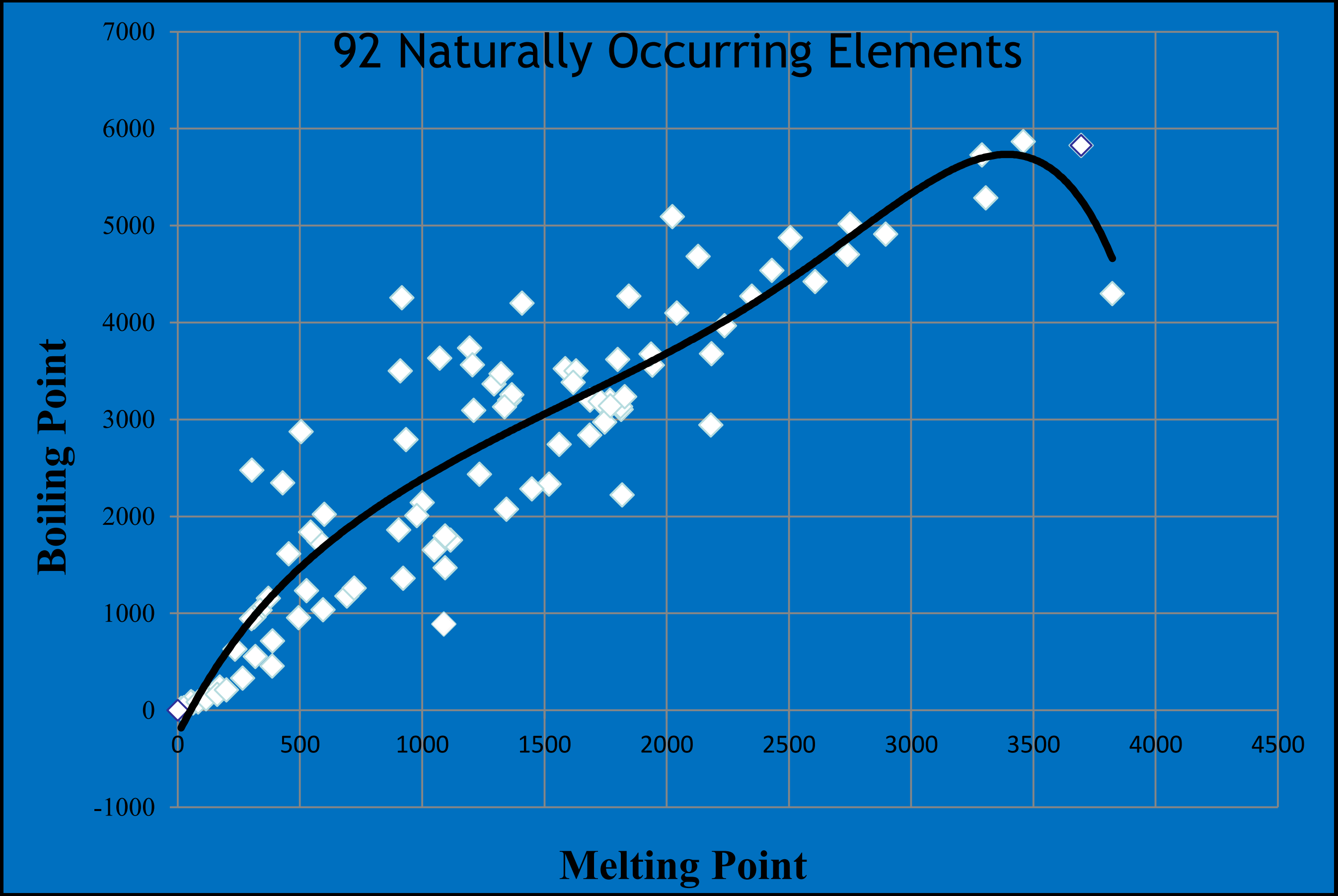


"According to our current predictive analytics solution, the mouse should be exiting from this hole in 3… 2… 1…" #betterdata

Source for image: https://www.hausmanmarketingletter.com/translating-analytics-to-action/

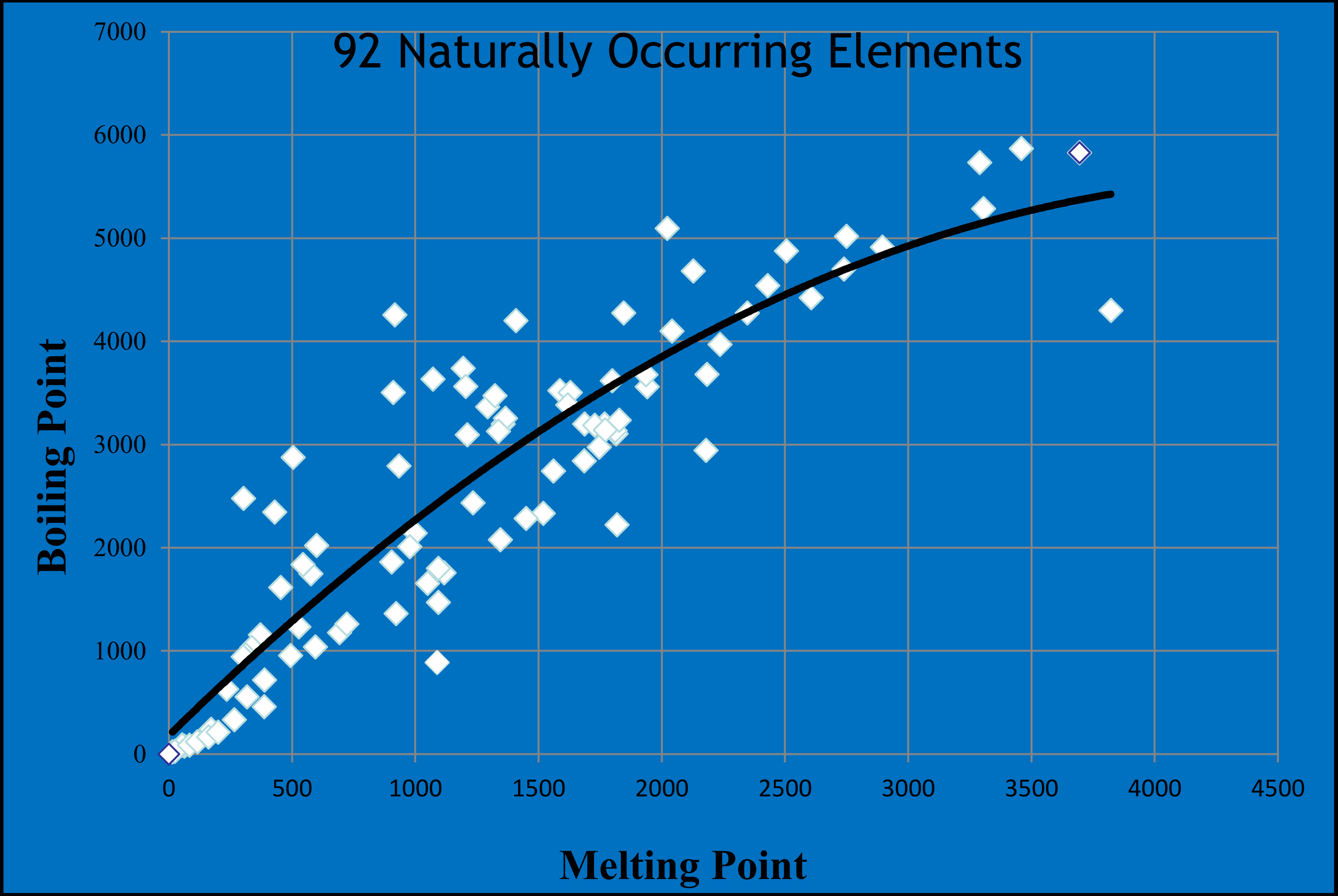
Trend Lines in big data sets: Descriptive Analytics!

It is tempting to over-fit every wiggle in the data.

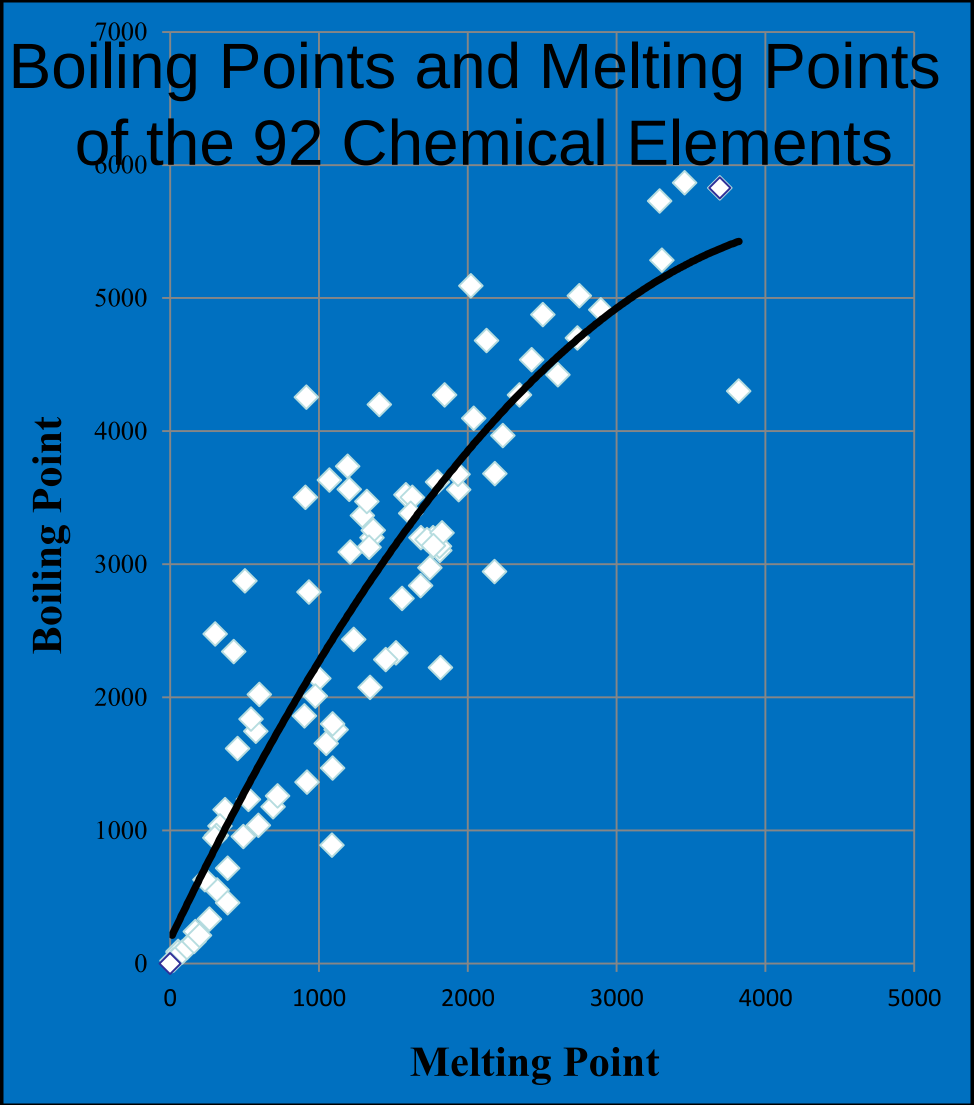


This is a better fit to the trend line…

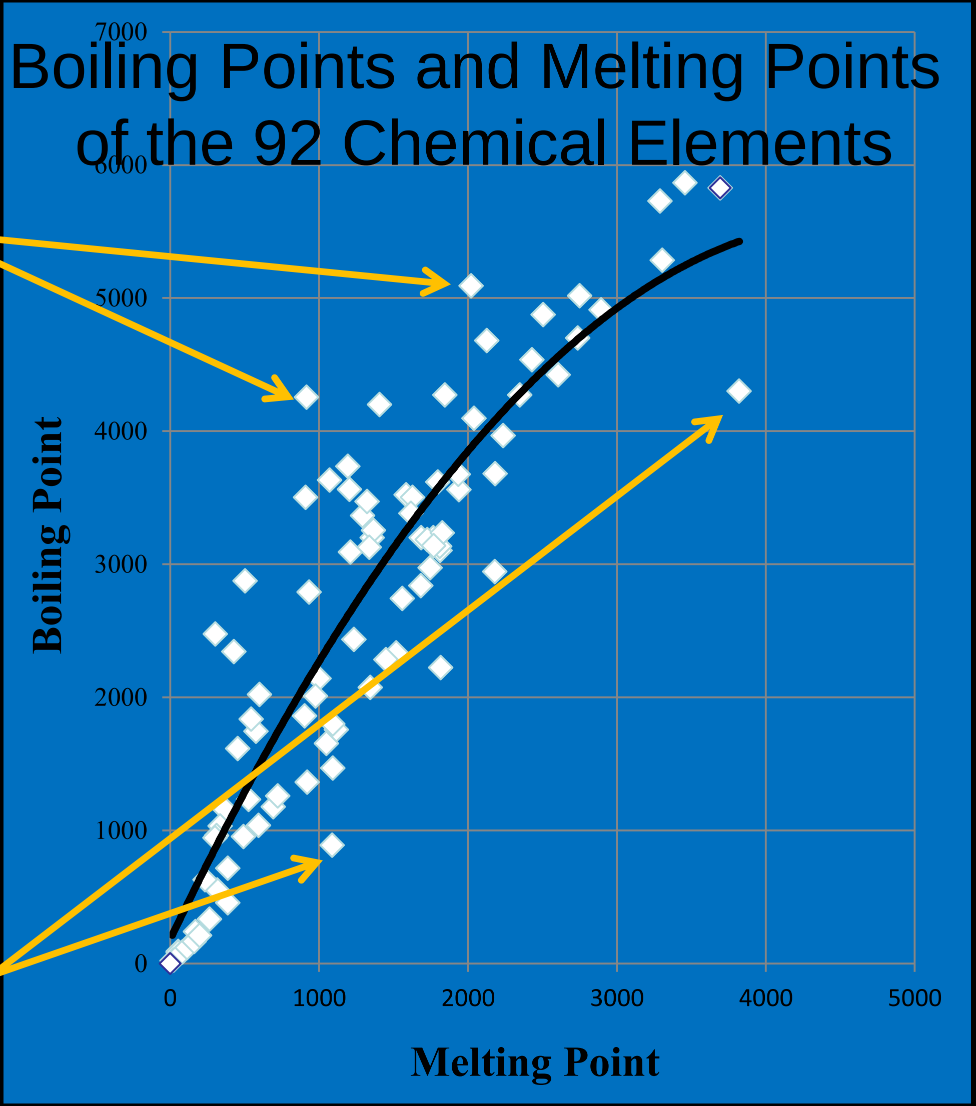
(generalization!) for use in Predictive Analytics!



Trend Line

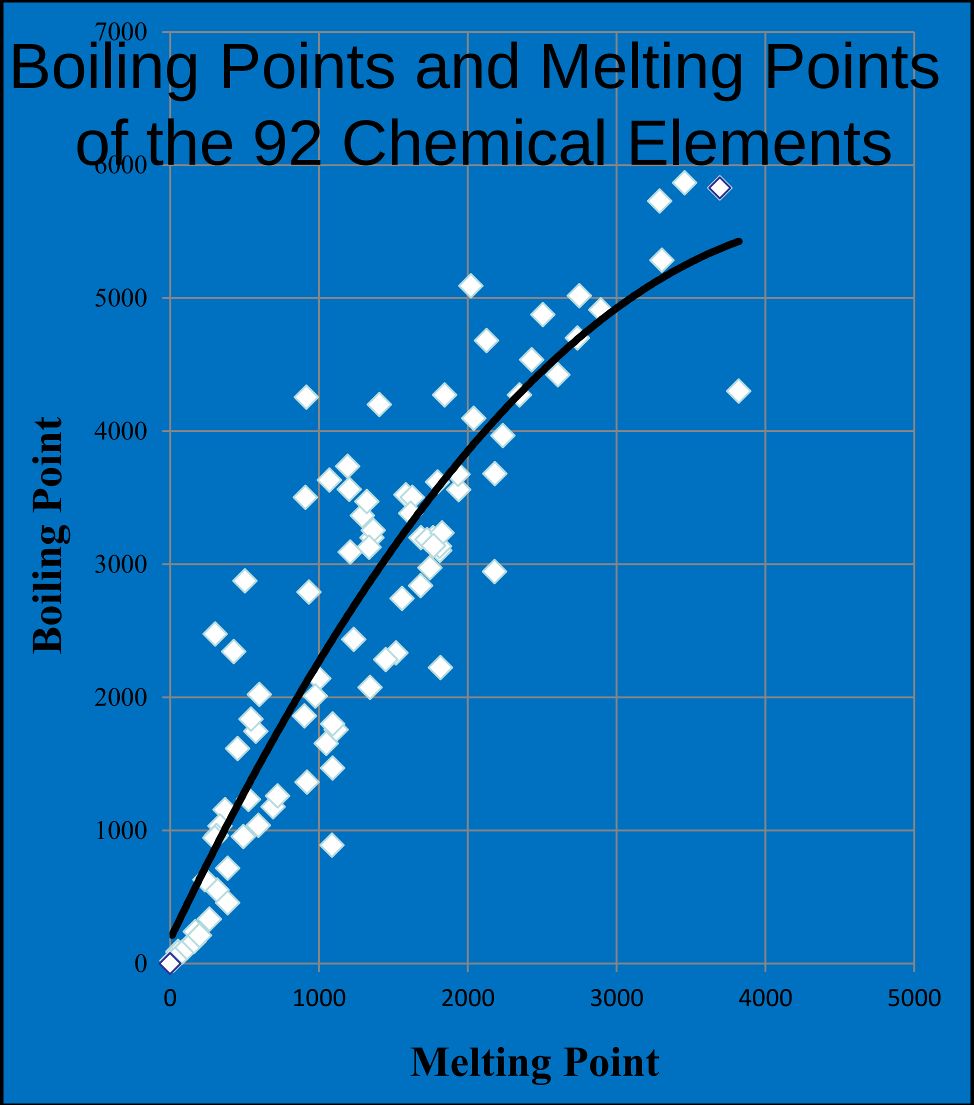

Boiling Points and Melting Points of the 92 Chemical Elements
Boiling Point
Melting Point
7000
6000
5000
4000
3000
2000
1000
0
0
1000
2000
3000
4000
5000

Trend Line and Outliers:



Sometimes we are tempted to think that outliers are just noise or natural variance.

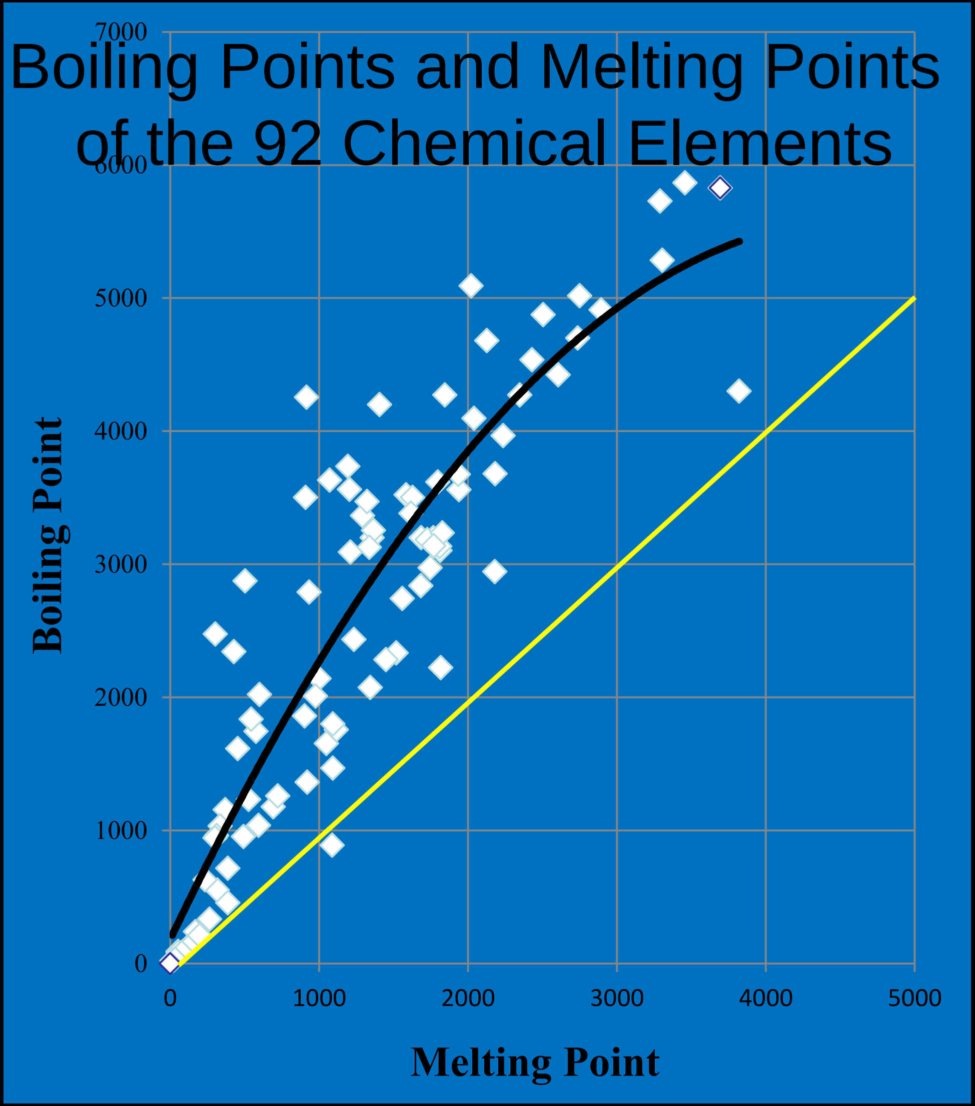
Trend Line and Outliers: where is the real discovery?


Boiling Points and Melting Points of the 92 Chemical Elements

Sometimes we are tempted to think that outliers are just noise or natural variance.

Trend Line and Outliers:

Add some context to the data!

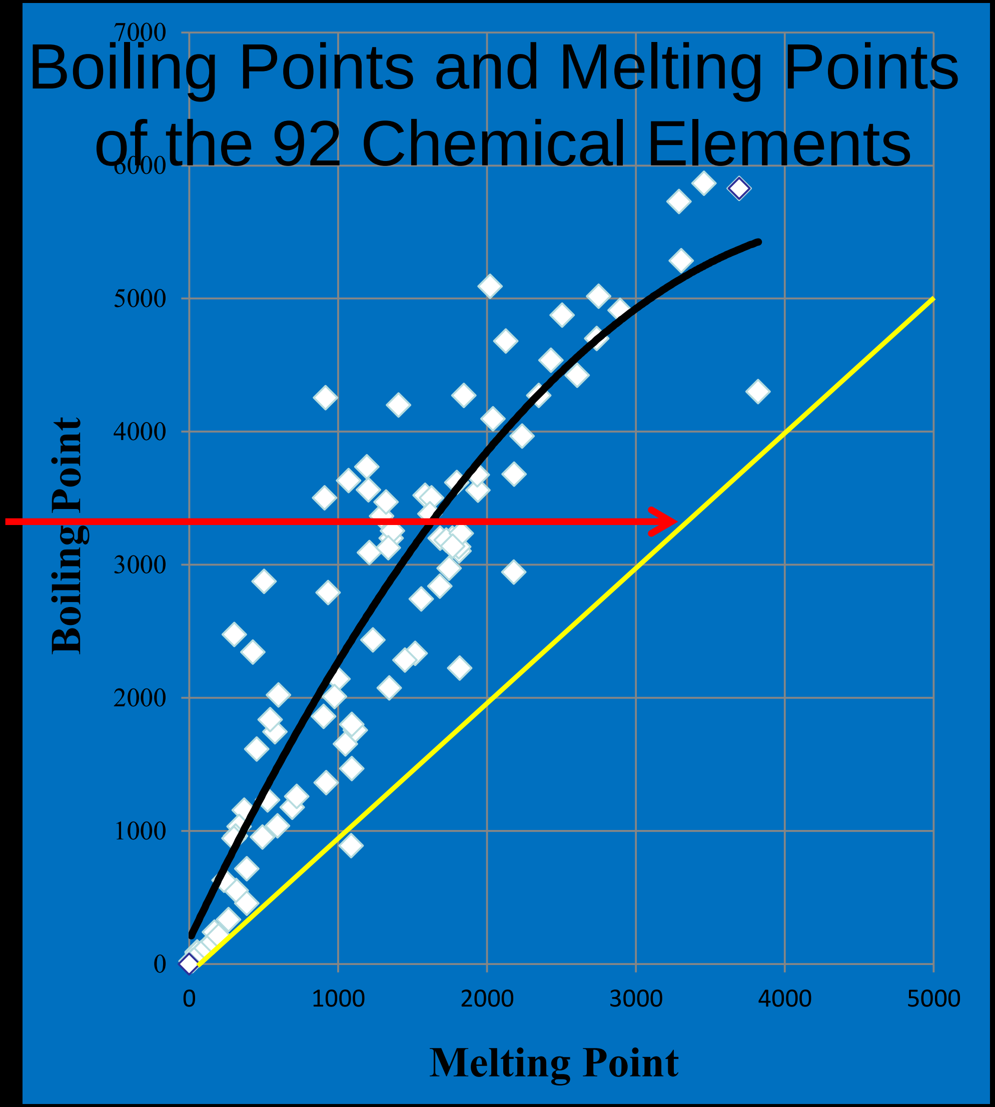


Trend Line and Outliers:

Add some context to the data!

…that diagonal line in the plot (where melting point = boiling point)

… this provides some context (related to your prior knowledge)!

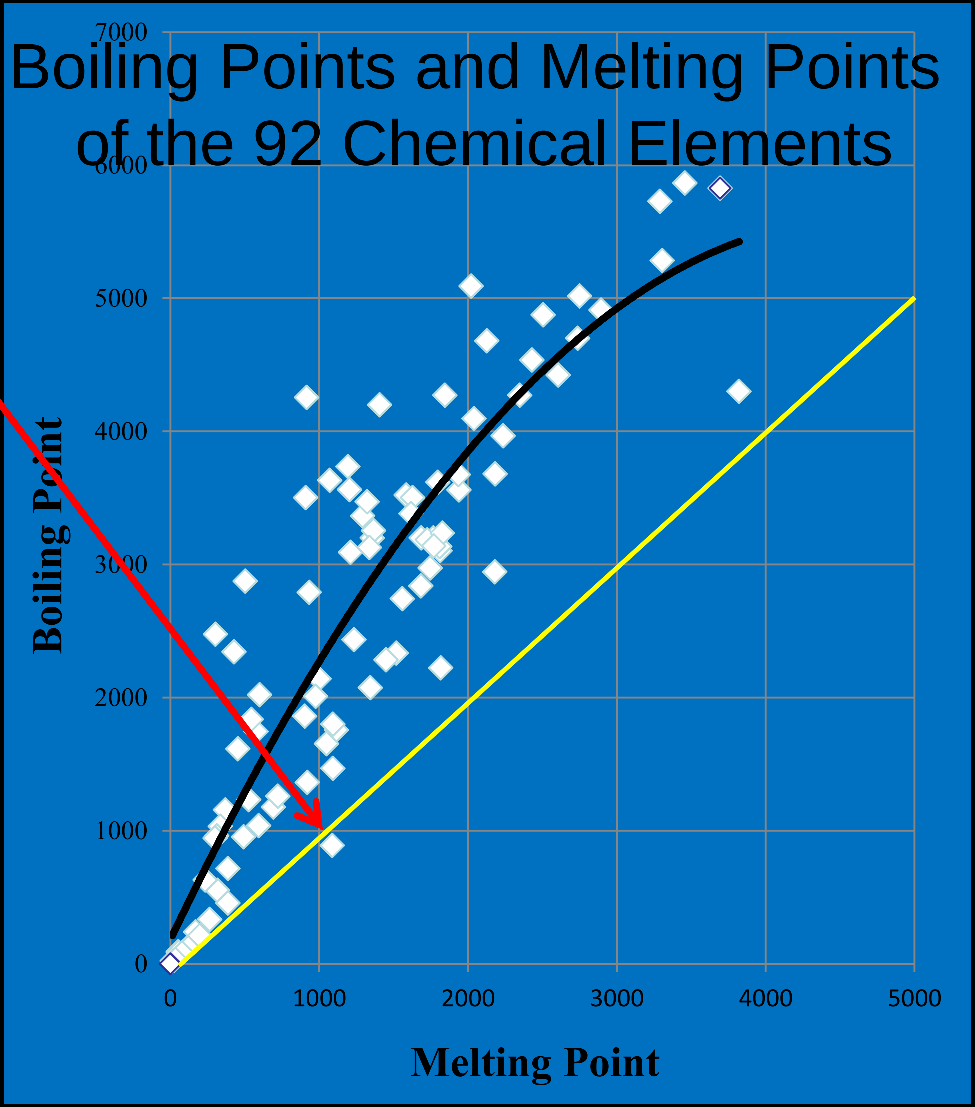


Trend Line and Outliers:

What is that point below the line?

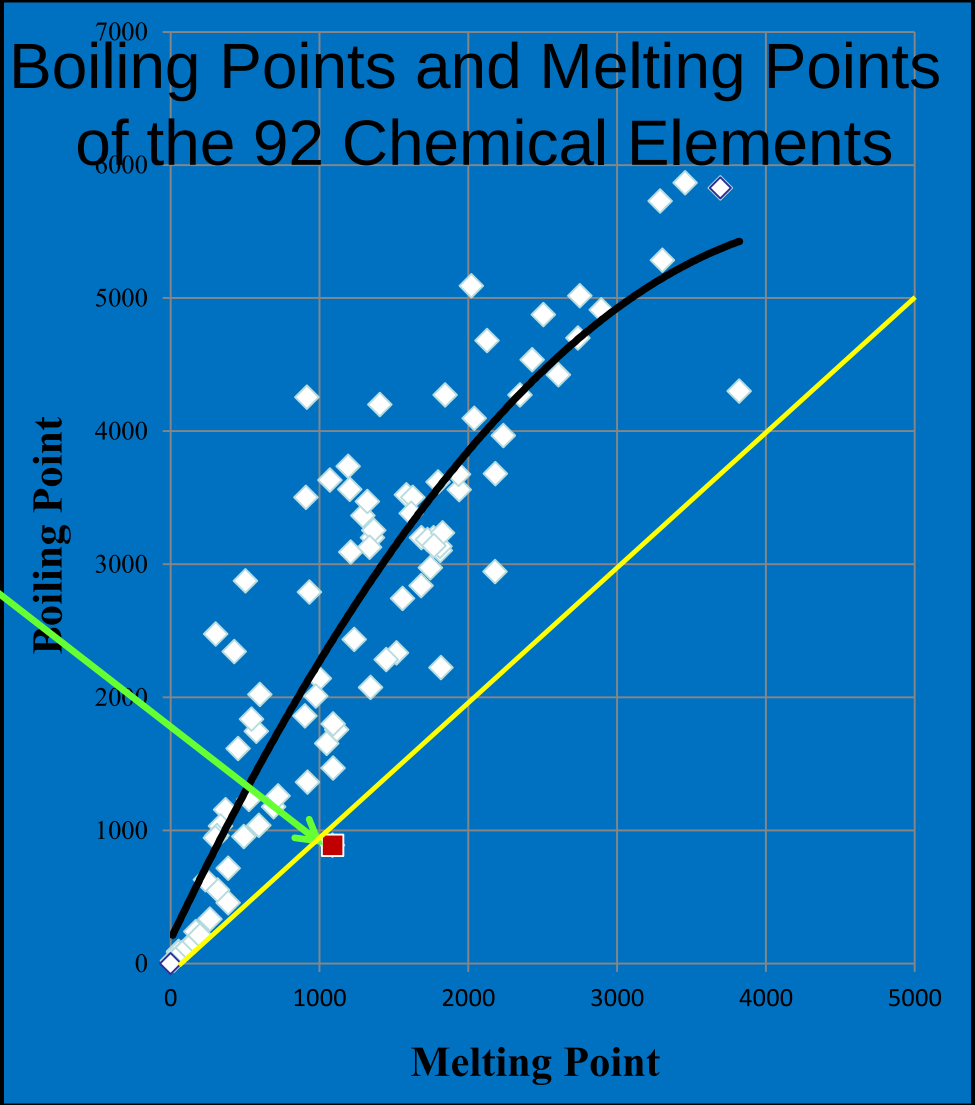
...that diagonal line in the plot (where melting point = boiling point)

... this provides some context (related to your prior knowledge)!

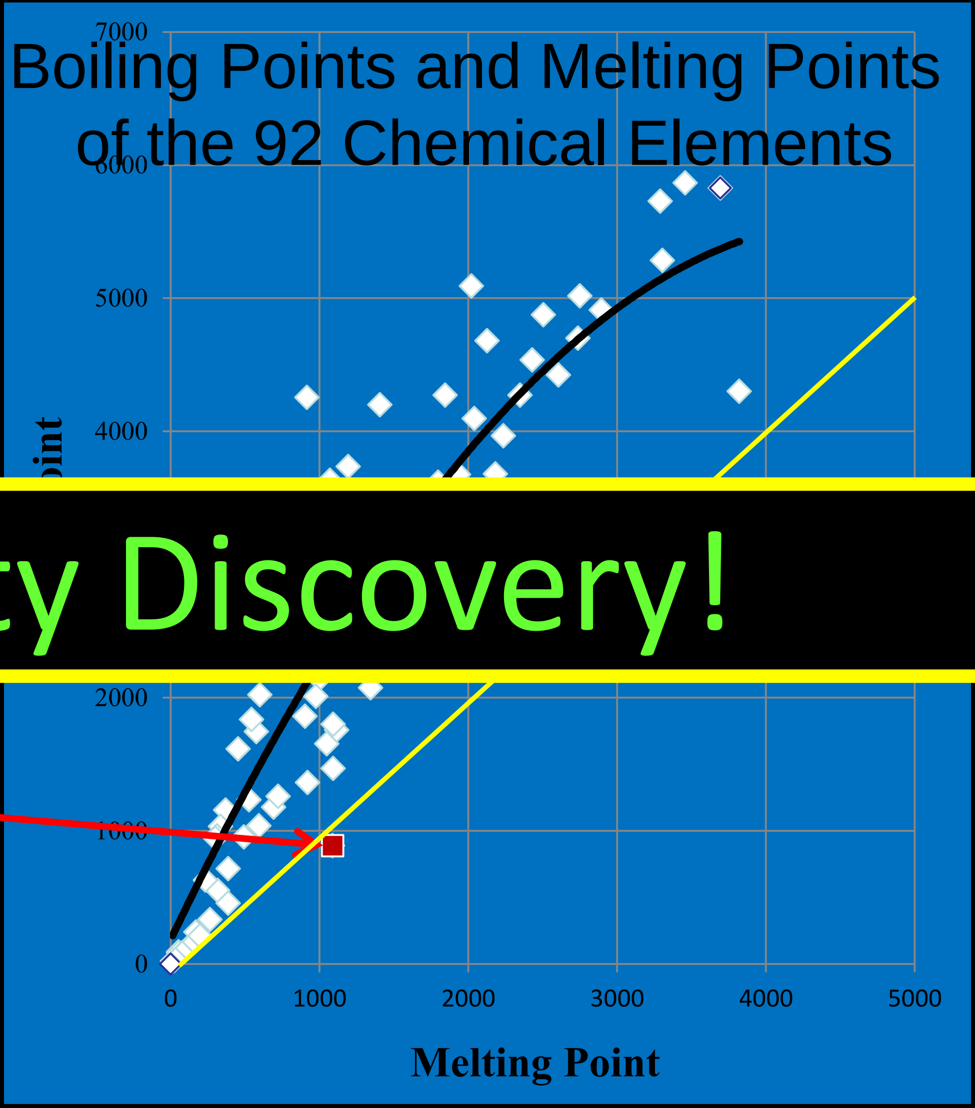


Boiling Points and Melting Points of the 92 Chemical Elements

Trend Line and Outliers: there’s the real discovery!



Trend Line and Outliers: there’s the real discovery!

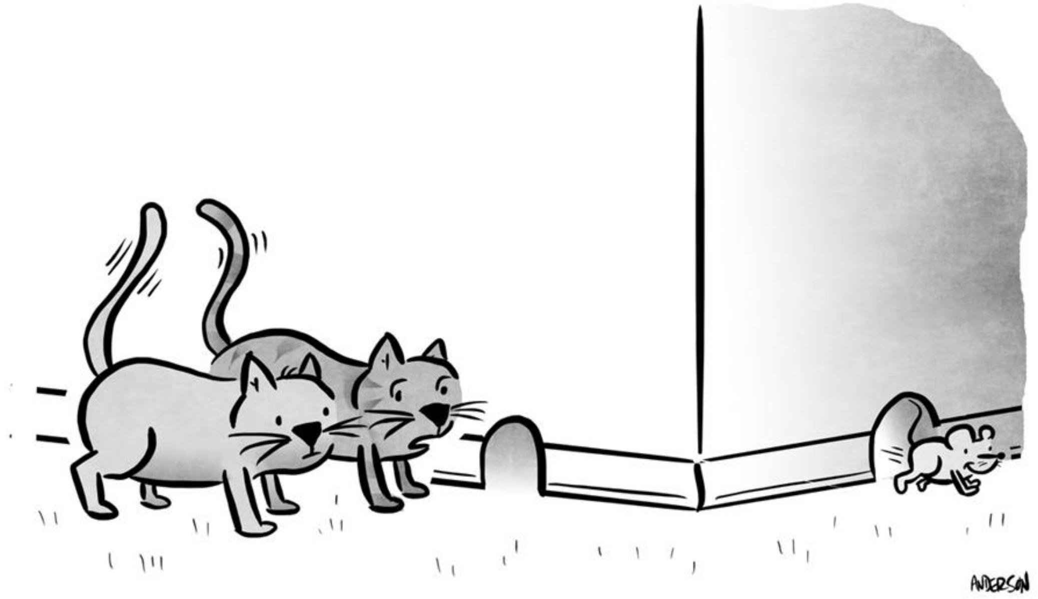


Novelty Discovery!

Melts @ 1089^{o}K
Boils @ 889^{o}K
Arsenic!

Examples of Interestingness in Data

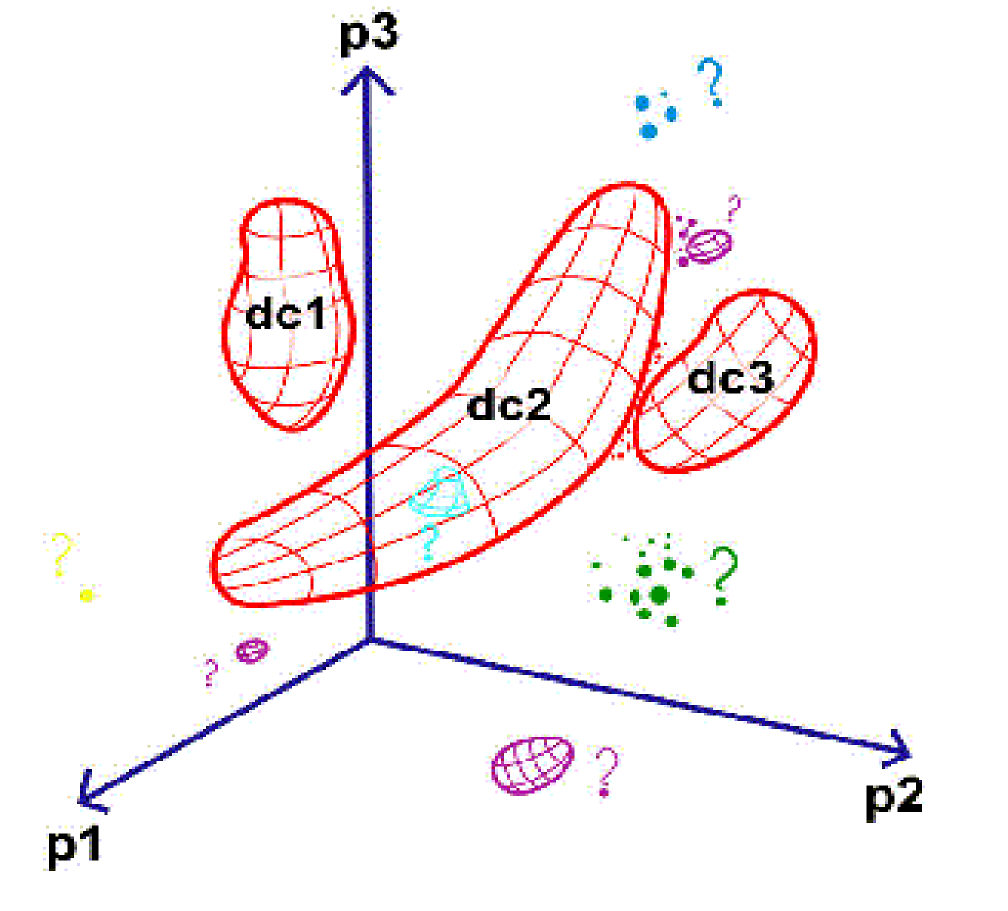
1) Outliers
2) **Counting**
3) Mapping
4) Associations
5) Linking
6) Clustering
7) Looking



"According to our current predictive analytics solution, the mouse should be exiting from this hole in 3… 2… 1…" #betterdata

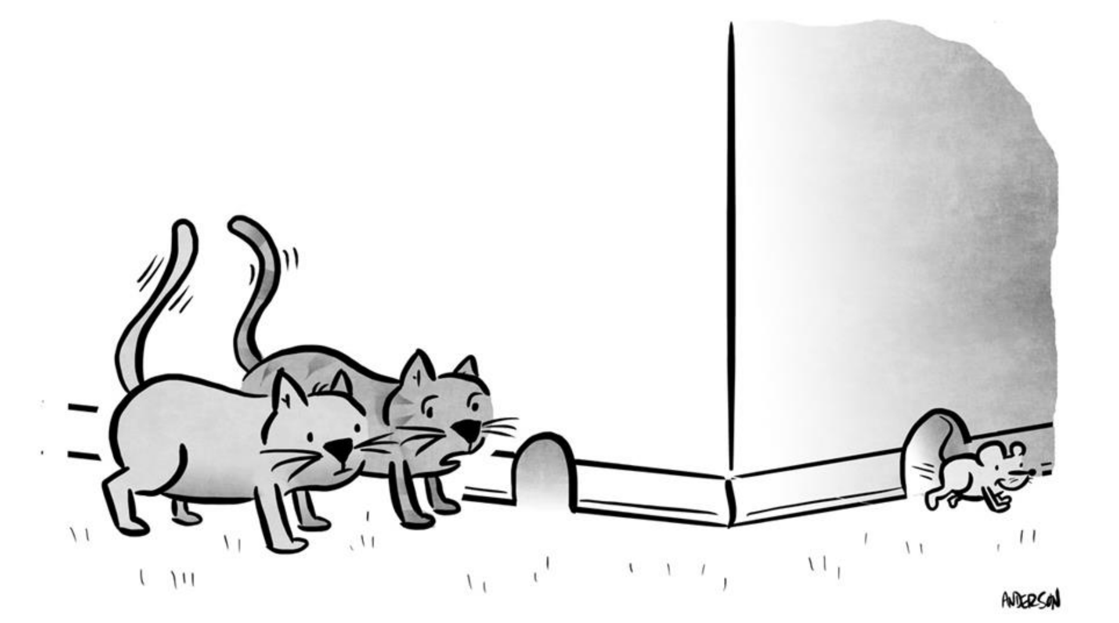
Source for image: https://www.hausmanmarketingletter.com/translating-analytics-to-action/

Correlation / Trend / Association Discovery =
= Predictive and Prescriptive Power Discovery!



Examples of Interestingness in Data

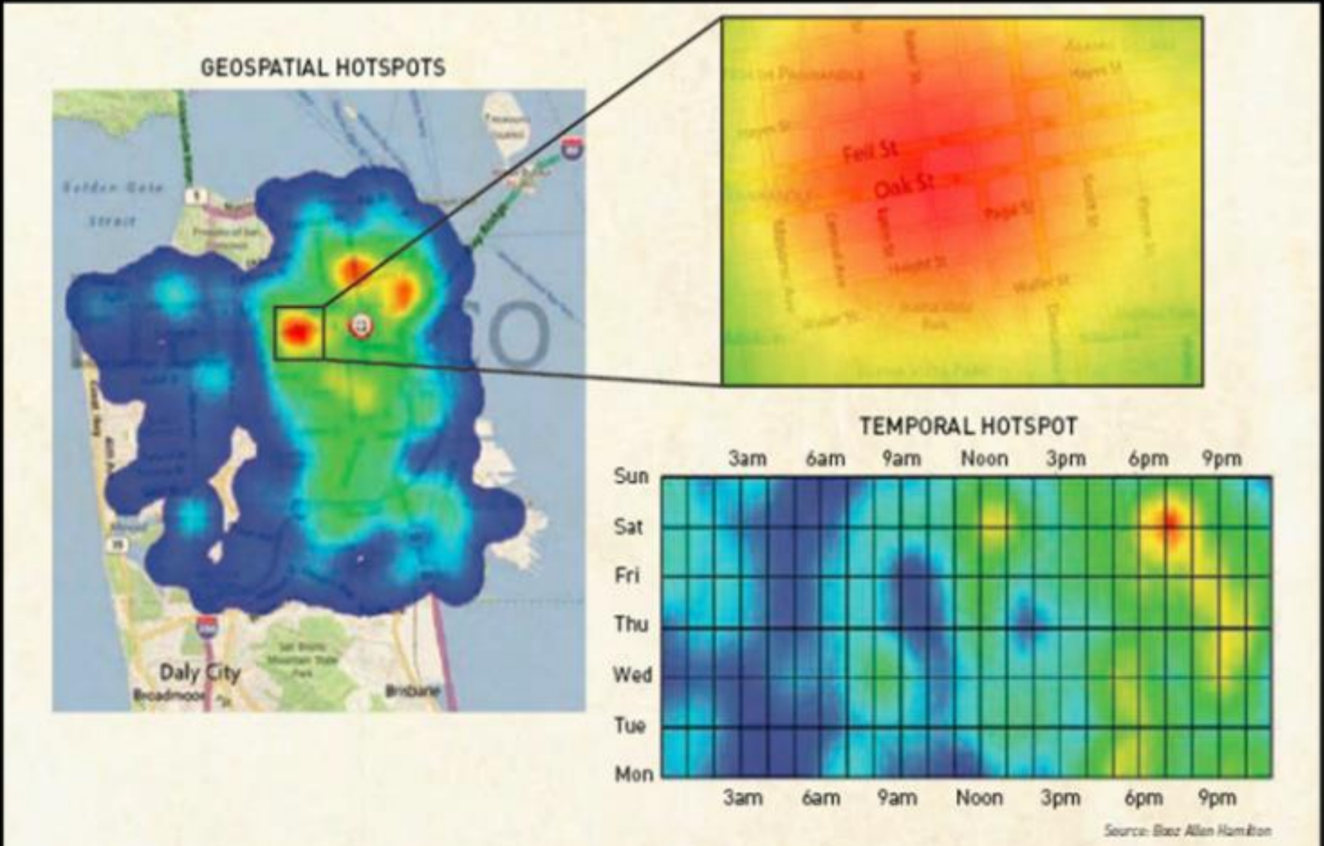
1) Outliers
2) Counting
3) **Mapping**
4) Associations
5) Linking
6) Clustering
7) Looking



“According to our current predictive analytics solution, the mouse should be exiting from this hole in 3… 2… 1…” #betterdata

Source for image: https://www.hausmanmarketingletter.com/translating-analytics-to-action/

“What is going on in that neighborhood on Saturday evenings between 6pm and 8pm?”



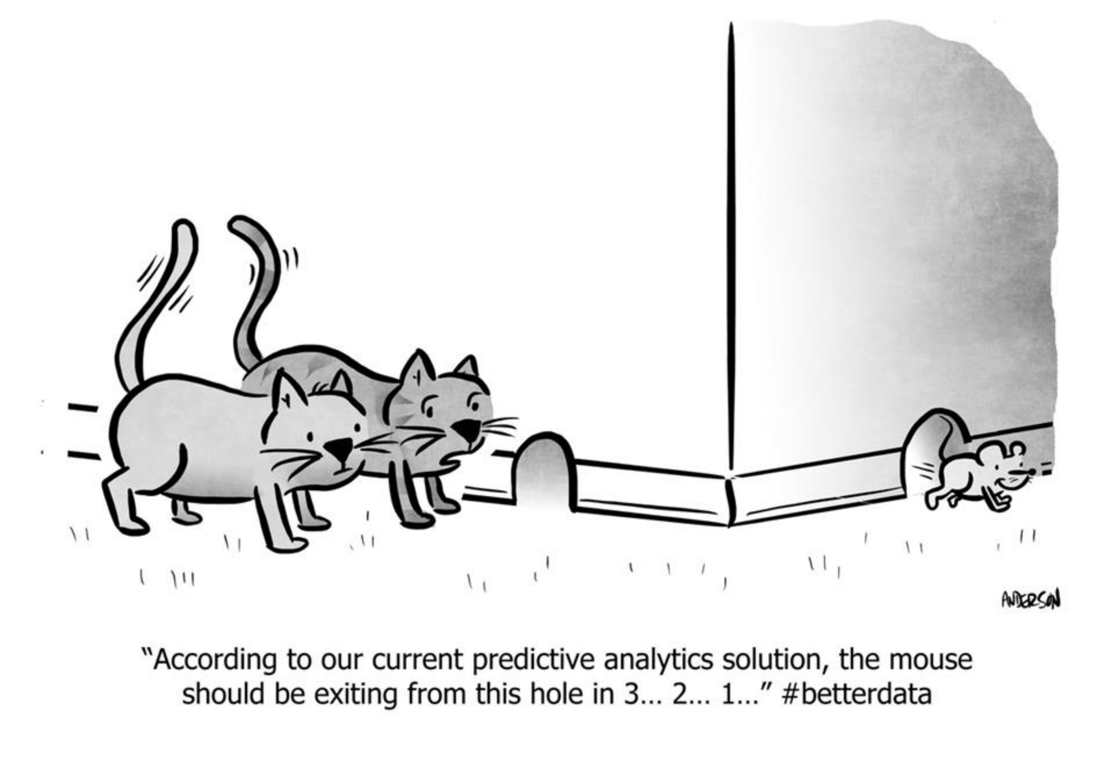
Source: Booz Allen Hamilton

Geospatial and Temporal Car Theft Hotspots

Source for graphic: https://www.boozallen.com/s/insight/publication/field-guide-to-data-science.html

Examples of Interestingness in Data

1) Outliers
2) Counting
3) Mapping
4) **Associations**
5) Linking
6) Clustering
7) Looking

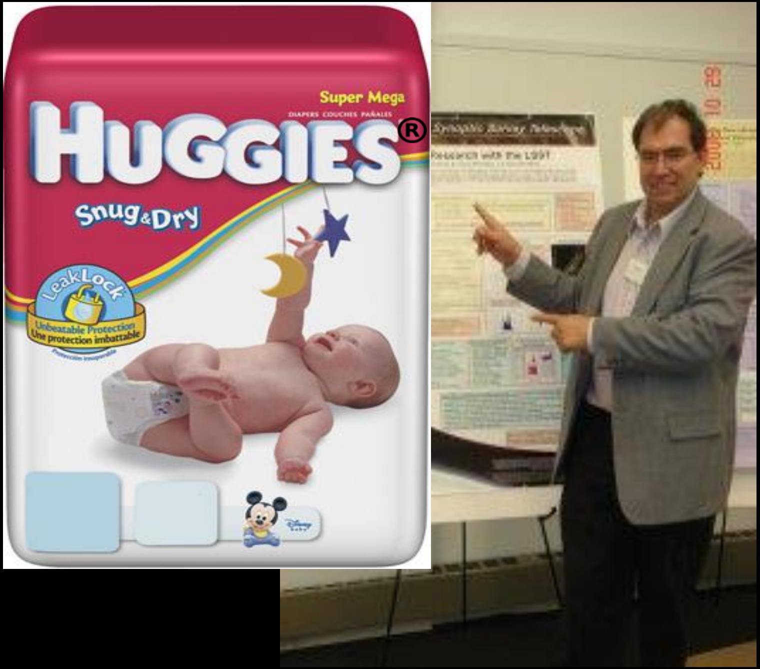


"According to our current predictive analytics solution, the mouse should be exiting from this hole in 3... 2... 1..." #betterdata

Source for image: https://www.hausmanmarketingletter.com/translating-analytics-to-action/

Association Discovery Example #1

- **Classic Textbook Example of Data Mining** (Legend?): Data mining of grocery store logs indicated that men who buy diapers also tend to buy beer at the same time.



Association Discovery Example #2

- **Wal-Mart** studied product sales in their Florida stores in 2004 when several hurricanes passed through Florida.

- Wal-Mart found that, before the hurricanes arrived, people purchased 7 times as many of ***{one particular product}*** compared to everything else.

Association Discovery Example #2

- **Wal-Mart** studied product sales in their Florida stores in 2004 when several hurricanes passed through Florida.

- Wal-Mart found that, before the hurricanes arrived, people purchased 7 times as many **strawberry pop tarts** compared to everything else.

Strawberry pop tarts???

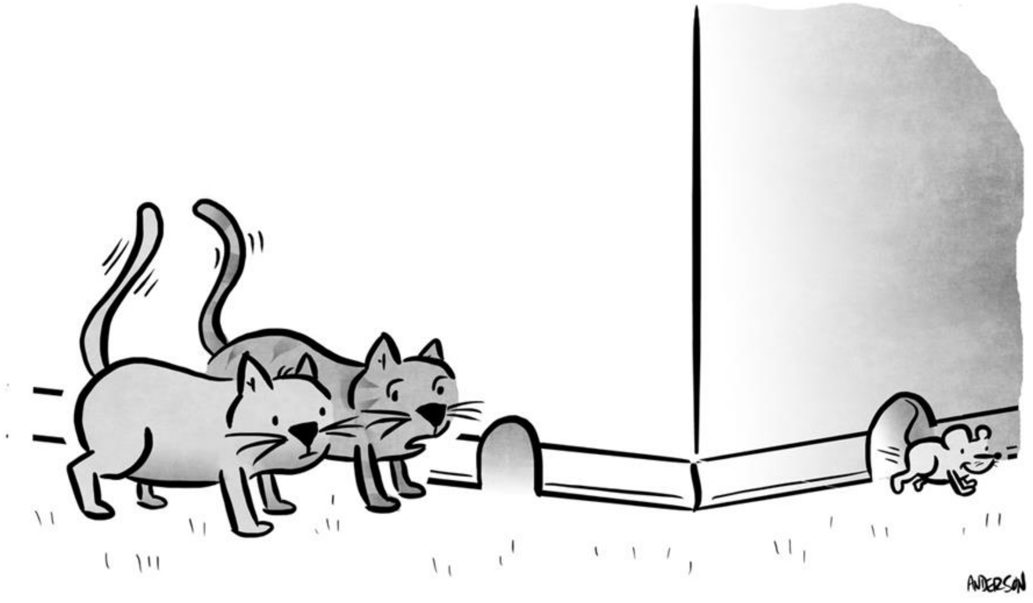
http://www.nytimes.com/2004/11/14/business/yourmoney/14wal.html
http://www.hurricaneville.com/pop_tarts.html
http://bit.ly/1gHZddA

Association Rule Discovery for Hurricane Intensification Forecasting

- Research by GMU geoscientists
- Predict the final strength of hurricane at landfall.
- Find co-occurrence of final hurricane strength with specific values of measured physical properties of the hurricane *while it is still over the ocean*.
- Result: the association rule discovery prediction is better than National Hurricane Center prediction!
- Research Paper by GMU scientists: https://ams.confex.com/ams/pdfpapers/84949.pdf

Examples of Interestingness in Data

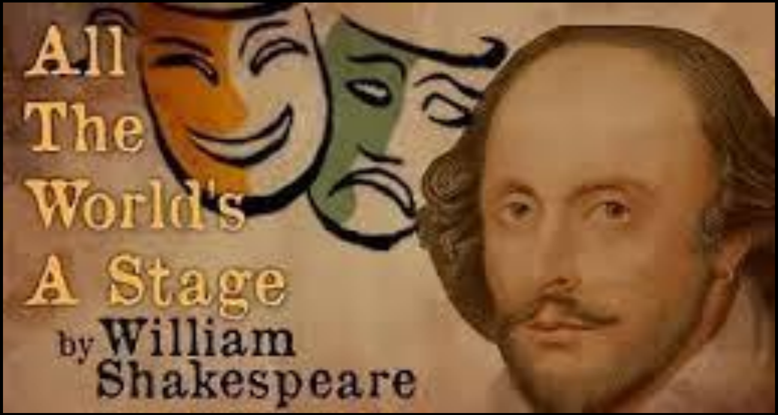
1) Outliers
2) Counting
3) Mapping
4) Associations
5) **Linking**
6) Clustering
7) Looking



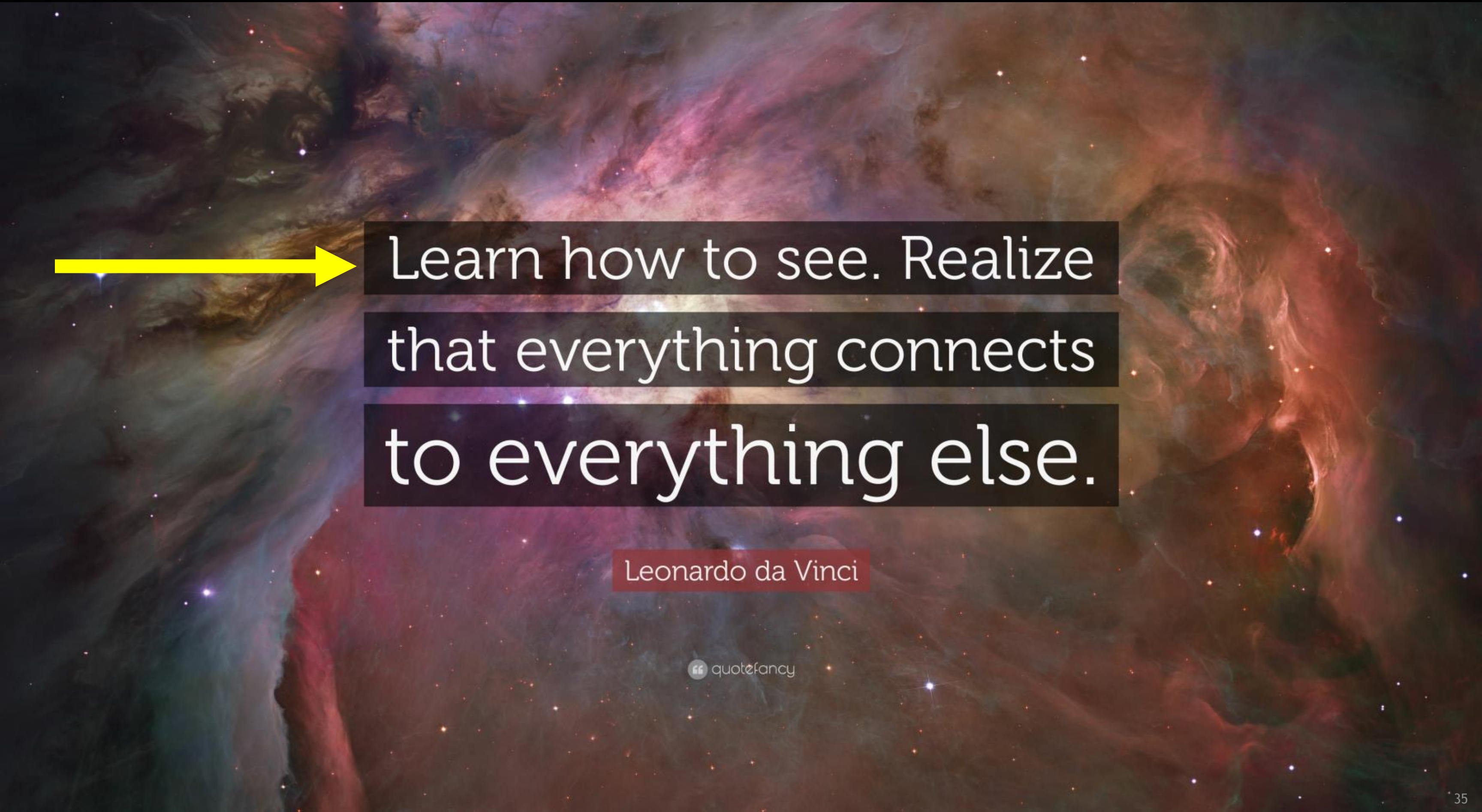
"According to our current predictive analytics solution, the mouse should be exiting from this hole in 3… 2… 1…" #betterdata

Source for image: https://www.hausmanmarketingletter.com/translating-analytics-to-action/

“All the World is a Graph” - Shakespeare?



“Everything connects to everything else”
- Leonardo da Vinci

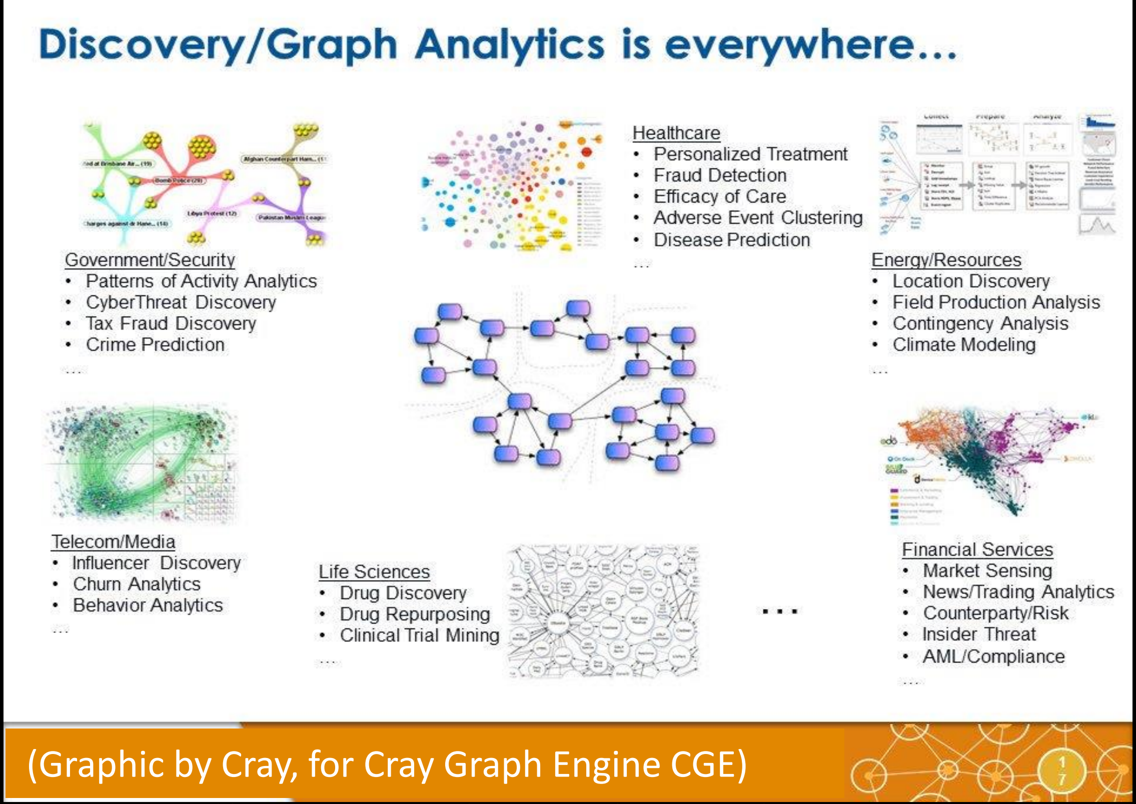


Learn how to see. Realize that everything connects to everything else.

Leonardo da Vinci

"All the World is a Graph" - Shakespeare?

The natural data structure of the world is not rows and columns, but a Graph!



Discovery/Graph Analytics is everywhere…

Government/Security
- Patterns of Activity Analytics
- CyberThreat Discovery
- Tax Fraud Discovery
- Crime Prediction

…

Healthcare
- Personalized Treatment
- Fraud Detection
- Efficacy of Care
- Adverse Event Clustering
- Disease Prediction

…

Energy/Resources
- Location Discovery
- Field Production Analysis
- Contingency Analysis
- Climate Modeling

…

Telecom/Media
- Influencer Discovery
- Churn Analytics
- Behavior Analytics

…

Life Sciences
- Drug Discovery
- Drug Repurposing
- Clinical Trial Mining

…

…

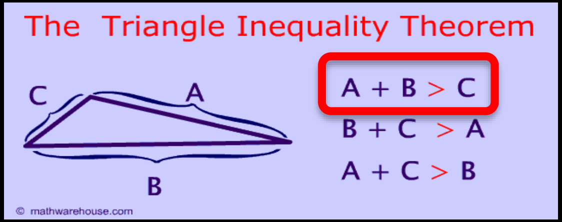
Financial Services
- Market Sensing
- News/Trading Analytics
- Counterparty/Risk
- Insider Threat
- AML/Compliance

…

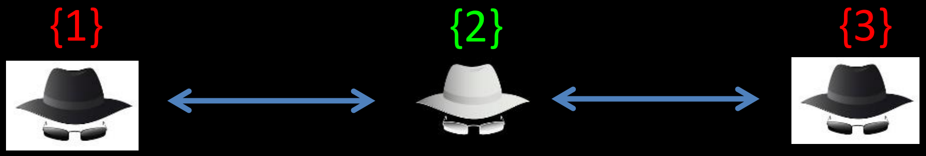
(Graphic by Cray, for Cray Graph Engine CGE)

http://www.cray.com/products/analytics/cray-graph-engine

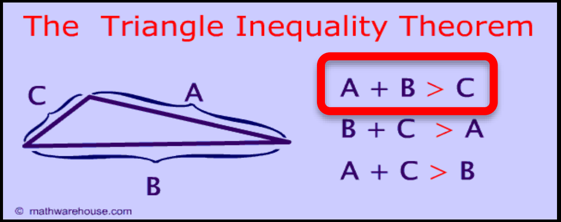
Simple Example of the Power of Graph: Semi-Metric Space



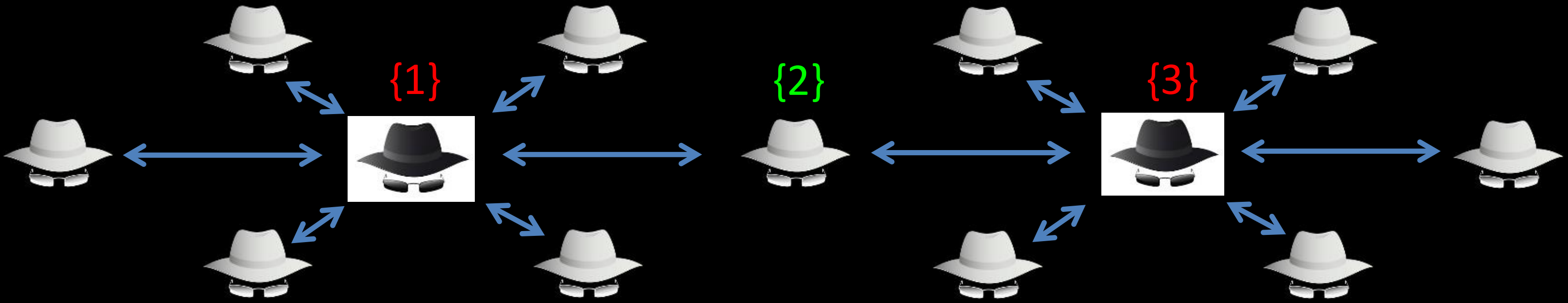
- Entity {1} is linked to Entity {2} (small distance A)
- Entity {2} is linked to Entity {3} (small distance B)
- Entity {1} is *not* linked directly to Entity {3} (Similarity Distance C = infinite)
- Similarity Distances between A, B, and C violate the triangle inequality!



Simple Example of the Power of Graph: Semi-Metric Space

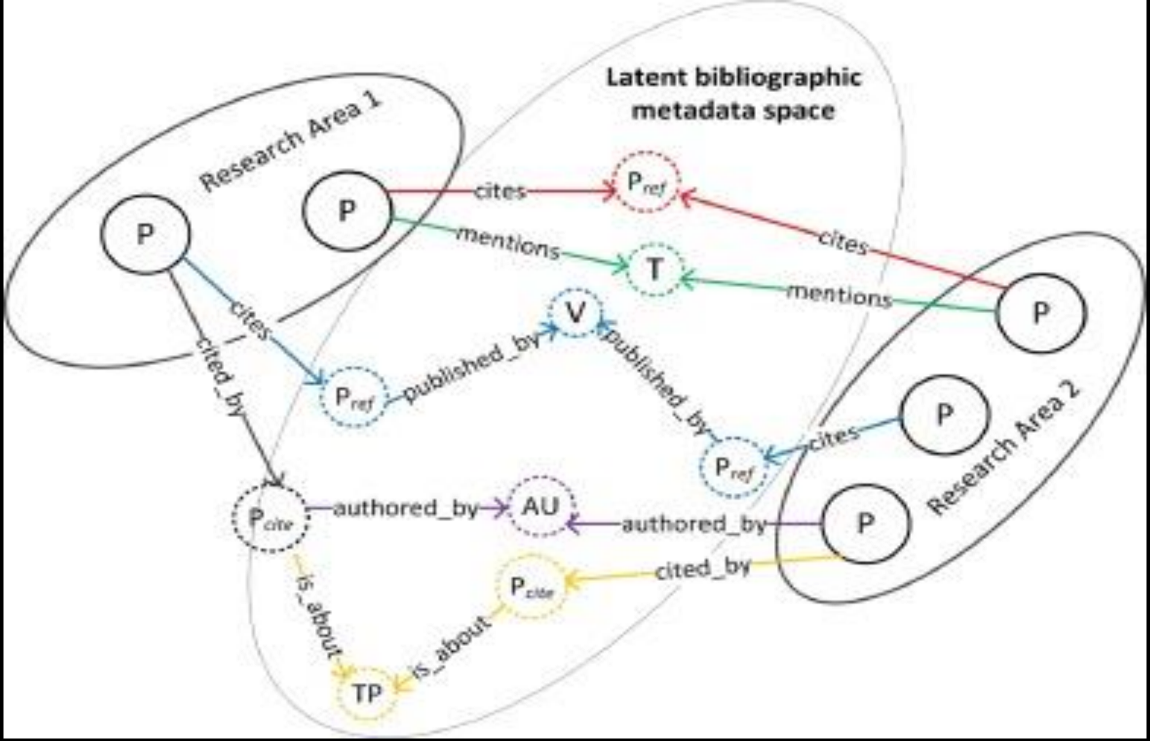


- Entity {1} is linked to Entity {2} (small distance A)
- Entity {2} is linked to Entity {3} (small distance B)
- Entity {1} is *not* linked directly to Entity {3} (Similarity Distance C = infinite)
- Similarity Distances between A, B, and C violate the triangle inequality!



- The connection between black hat entities {1} and {3} never appears explicitly within a transactional database.
- Examples: (a) Medical Research Discoveries across disconnected journals, through linked semantic assertions; (b) Customer Journey modeling; (c) Safety Incident Causal Factor Analysis; (d) Marketing Attribution Analysis; (e) Fraud networks, Illegal goods trafficking networks, Money-Laundering networks.

Research Example:
Literature-Based Discovery (LBD)



References:

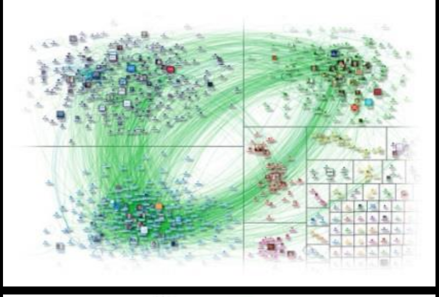
- https://www.sciencedirect.com/science/article/pii/S0950705116303860
- https://summerofhpc.prace-ri.eu/introducing-lbdream-and-literature-based-discovery/

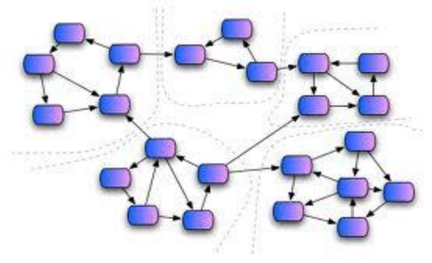
Research Example: Discovery in the NIH-NLM Semantic MEDLINE Database

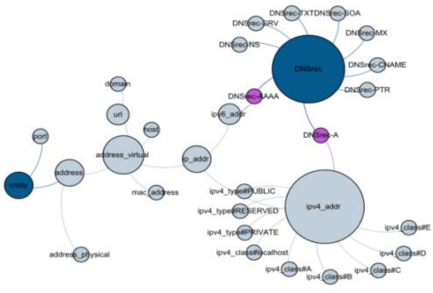
Project Description: Conduct semantic graph mining of the NIH-NLM metadata repository from ~26 million medical research articles.

Graph Database: ~90 million RDF triples (predications; semantic assertions).

Research Project: (PhD dissertation at GMU) Novel subgraph discovery; Context-based discovery; New concept emergence in medical research; Story discovery in linked graph network; and Hidden knowledge discovery through semi-metrics.



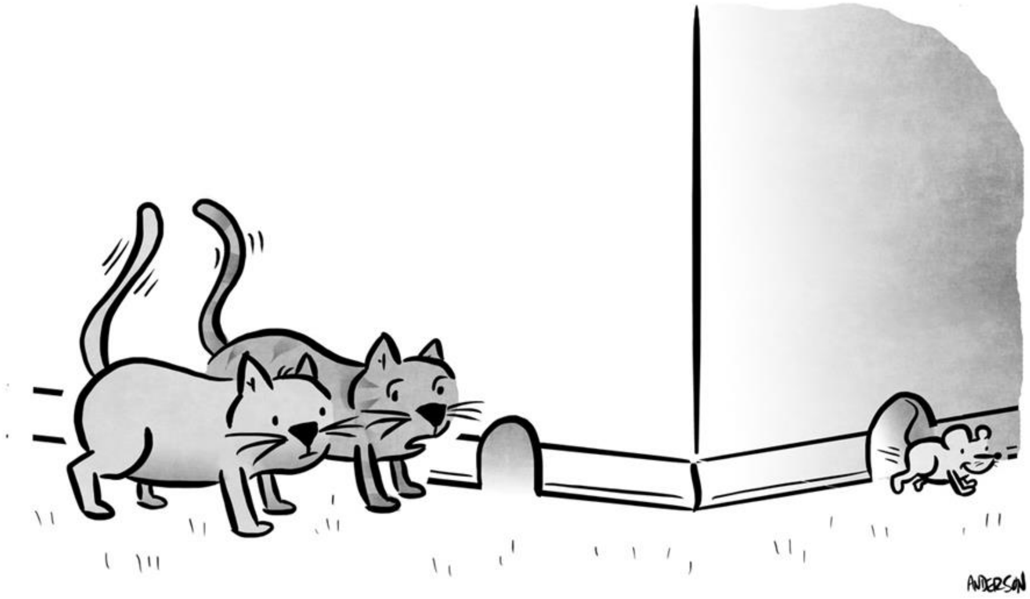




https://skr3.nlm.nih.gov/SemMedDB/

Examples of Interestingness in Data

1) Outliers
2) Counting
3) Mapping
4) Associations
5) Linking
6) **Clustering**
7) Looking



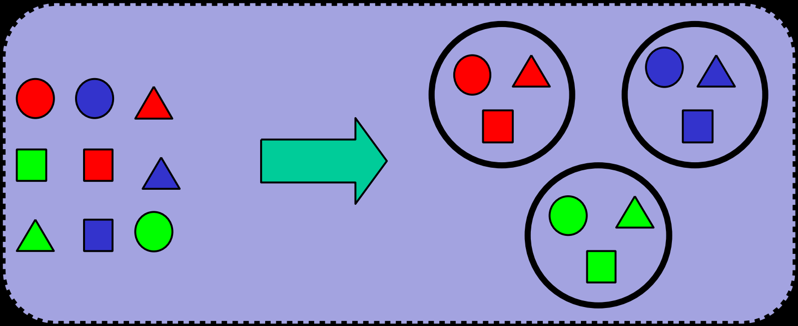
"According to our current predictive analytics solution, the mouse should be exiting from this hole in 3… 2… 1…" #betterdata

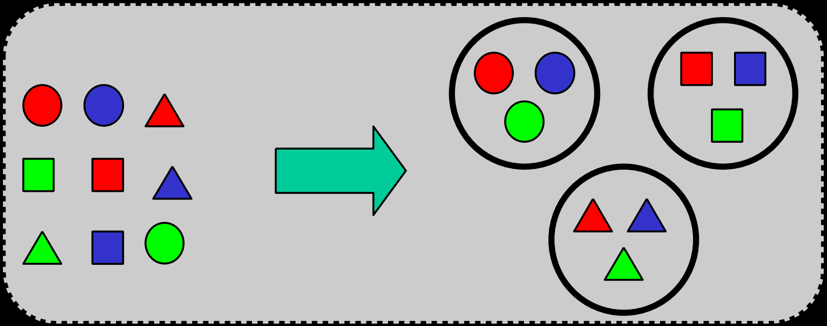
Source for image: https://www.hausmanmarketingletter.com/translating-analytics-to-action/

Clustering = *the process of partitioning a set of data into subsets (segments or clusters) such that a data element belonging to any chosen cluster is more similar to data elements belonging to that cluster than to data elements belonging to other clusters.*

= Group together similar items + separate the dissimilar items

= Identify similar characteristics, patterns, or behaviors among subsets of the data elements.





Challenge #1) No prior knowledge of the number of clusters.

#2) No prior knowledge of semantic meaning of the clusters.

#3) Different clusters are possible from the same data set!

#4) Different clusters are possible using different similarity metrics.

How to know if your clusters are good enough:

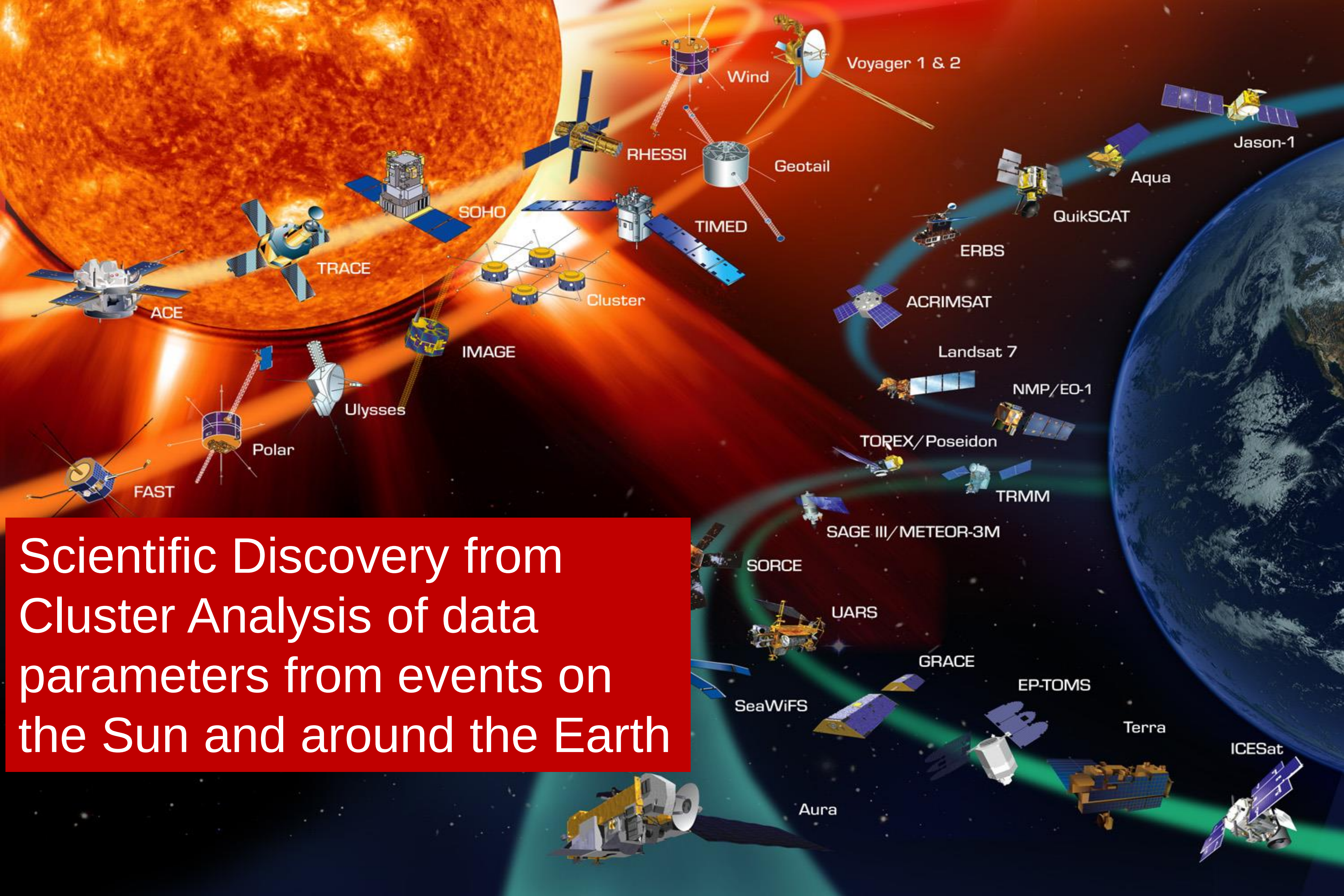
- You know the clusters are good …
  - … if the clusters are compact relative to their separation
  - … if the clusters are well separated from one another
  - … the “within cluster” errors are small (low variance within)
  - … if the number of clusters is small relative to the number of data points
- Various measures of cluster compactness exist, including the Dunn index , C-index, and the DBI (Davies-Bouldin Index)

Reference: http://www.biomedcentral.com/content/supplementary/1471-2105-9-90-S2.pdf

Application of Davies-Bouldin Index

- Assume K (the number of clusters) and assume other things (choice of clustering algorithm; the choice of clustering feature attributes; etc.)
- Measure DBI
- Test another set of values for the cluster input parameters (K, feature attributes, etc.)
- Measure DBI
- ... continue iterating like this until you find the set of cluster input parameters that yields the best (minimum) value for DBI.

Scientific Discovery from Cluster Analysis of data parameters from events on the Sun and around the Earth



Cluster Analysis:
Find the clusters, then **Evaluate** them

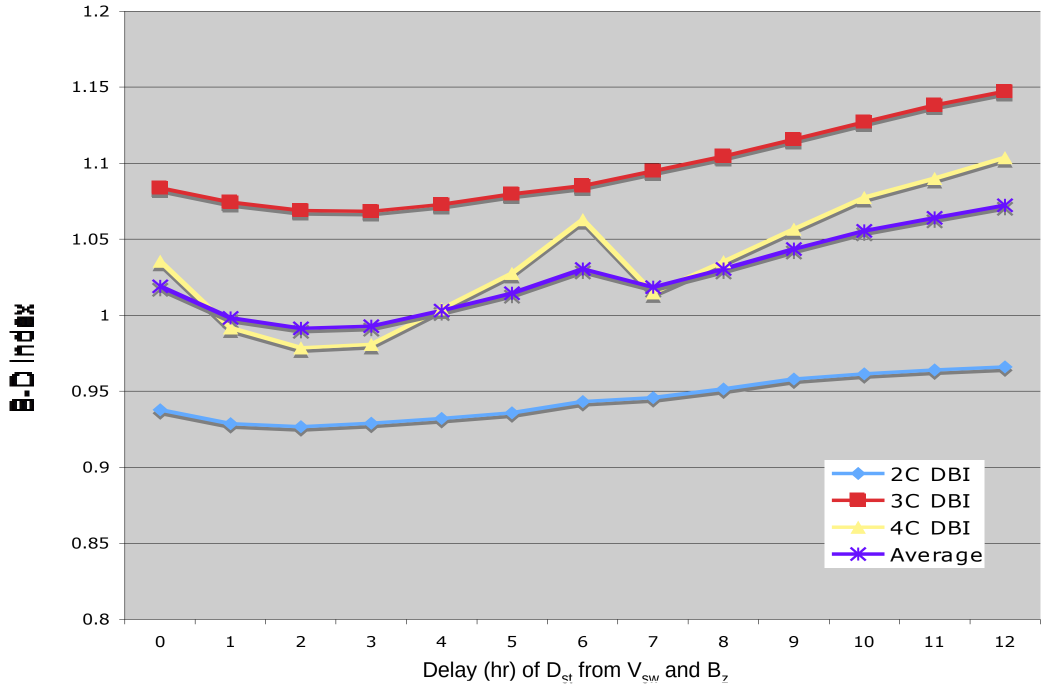


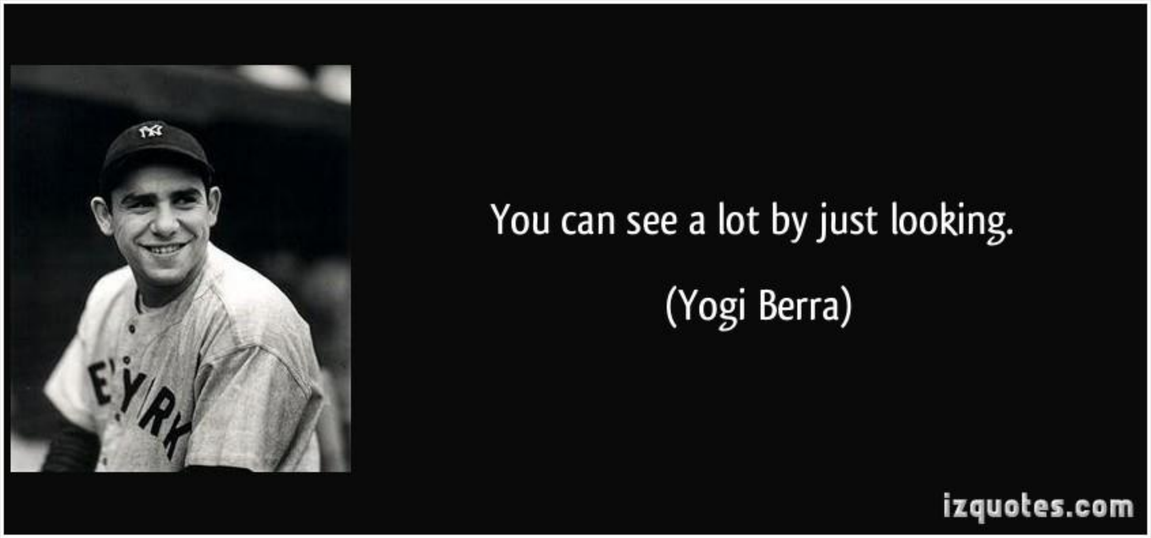
Figure 10. Davies-Bouldin index for various time delays of D_{st} from V_{sw} and B_z for cases of 2 (blue), 3 (red), 4 (yellow) clusters, and the overall average (purple), indicating an optimal delay of ~2-3 hours for Dst.

Good Clusters = Small Size relative to Cluster Separation.

DISCOVERY! …
Solar wind events have the strongest association (*i.e.*, the tightest clusters) with the space plasma events within the Earth's magnetosphere about 2-4 hours after a major plasma outburst occurs on the Sun.

Examples of Interestingness in Data

1) Outliers
2) Counting
3) Mapping
4) Associations
5) Linking
6) Clustering
7) **Looking**



You can see a lot by just looking.

(Yogi Berra)

"You can see a lot by just looking"
(and you can see around corners!)
Cognitive, Contextual, Insightful, Forecastful

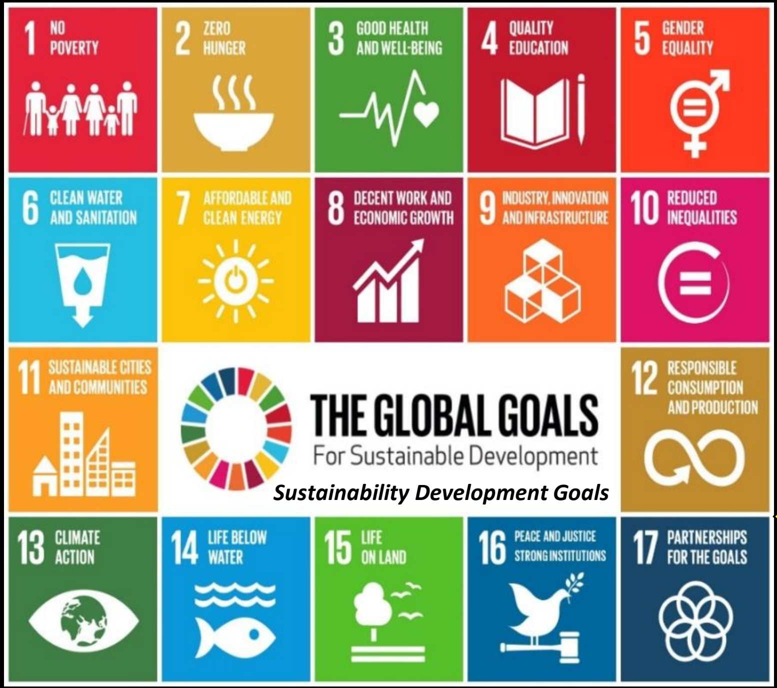
https://www.speedcafe.com/2017/07/12/f1-demo-take-place-london-streets/

Final Thoughts

Big Data + the IoT + Citizen Data Scientists =
= Partners in Sustainability

The Internet of Things (IoT):

Knowing the knowable via deep, wide, and fast data from ubiquitous sensors!


1 NO POVERTY
2 ZERO HUNGER
3 GOOD HEALTH AND WELL-BEING
4 QUALITY EDUCATION
5 GENDER EQUALITY
6 CLEAN WATER AND SANITATION
7 AFFORDABLE AND CLEAN ENERGY
8 DECENT WORK AND ECONOMIC GROWTH
9 INDUSTRY, INNOVATION AND INFRASTRUCTURE
10 REDUCED INEQUALITIES
11 SUSTAINABLE CITIES AND COMMUNITIES
THE GLOBAL GOALS
For Sustainable Development
12 RESPONSIBLE CONSUMPTION AND PRODUCTION
13 CLIMATE ACTION
14 LIFE BELOW WATER
15 LIFE ON LAND
16 PEACE AND JUSTICE STRONG INSTITUTIONS
17 PARTNERSHIPS FOR THE GOALS

Sustainability Development Goals

http://www.unglobalpulse.org

Big Data:

In the Big Data era, Everything is Quantified and Monitored :

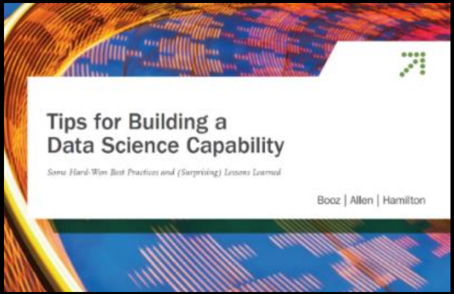
- Populations & Persons
- Smart Cities, Energy, Grids, Farms, Highways
- Environmental Sensors
- IoE = Internet of Everything!

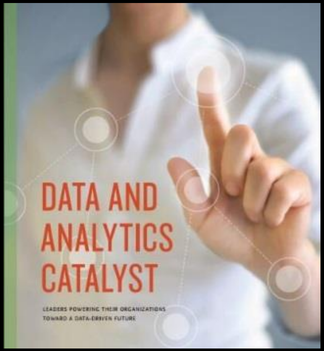
Discovery through Machine Learning and Data Science:

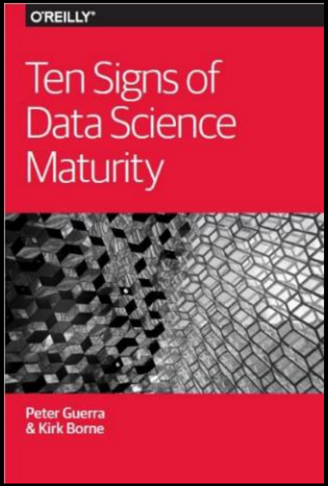
- Class Discovery, Correlation Discovery, Novelty Discovery, and
- Association Discovery: **Find interesting cases where condition X is associated with event Y with time shift Z.**

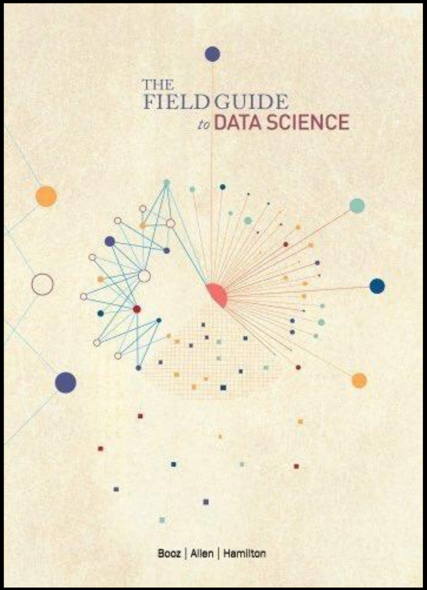
17 SDGs are KPIs for the World!
(currently, the SDGs have 229 Key Performance Indicators)
(SDG: Sustainability Development Goal)

NextGen Analytics and Data Science
Harnessing the Power of Data
Booz | Allen | Hamilton


Tips for Building a Data Science Capability
Some Hard-Won Best Practices and (Surprising) Lessons Learned
Booz | Allen | Hamilton


DATA AND ANALYTICS CATALYST
LEADERS POWERING THEIR ORGANIZATIONS TOWARD A DATA-DRIVEN FUTURE





Thank you!

Contact information, for further questions or inquiries:

Dr. Kirk Borne, Principal Data Scientist, Booz Allen Hamilton

Twitter: @KirkDBorne or Email: kirk.borne@gmail.com

Get slides here: http://www.kirkborne.net/ASA2018/