"Essentially, all models are wrong, …but some are useful."

- George Box

(One of the most influential statisticians of the 20th century and a pioneer in the areas of quality control, time series analysis, design of experiments and Bayesian inference.)

"No one ever made a decision because of a number. They need a story."

Daniel Kahneman, Quoted in Vanity Fair article "How Two Trailblazing Psychologists Turned the World of Decision Science Upside Down," November 2016

Data Literacy and Data Storytelling for Data Scientists

Get these slides here: http://www.kirkborne.net/INSAID2019/

Kirk Borne

@KirkDBorne

Principal Data Scientist

Booz Allen Hamilton

http://www.boozallen.com/datascience

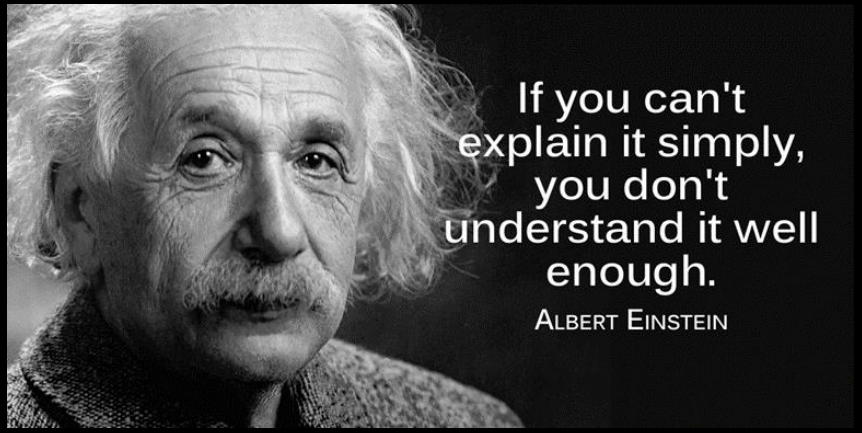
OUTLINE

- Talking the Walk
- Data Literacy
- Data Storytelling

Source: https://www.expertsystem.com/government-data-mining/

OUTLINE

- **Talking the Walk**
- Data Literacy
- Data Storytelling



We are *not* talking about this…

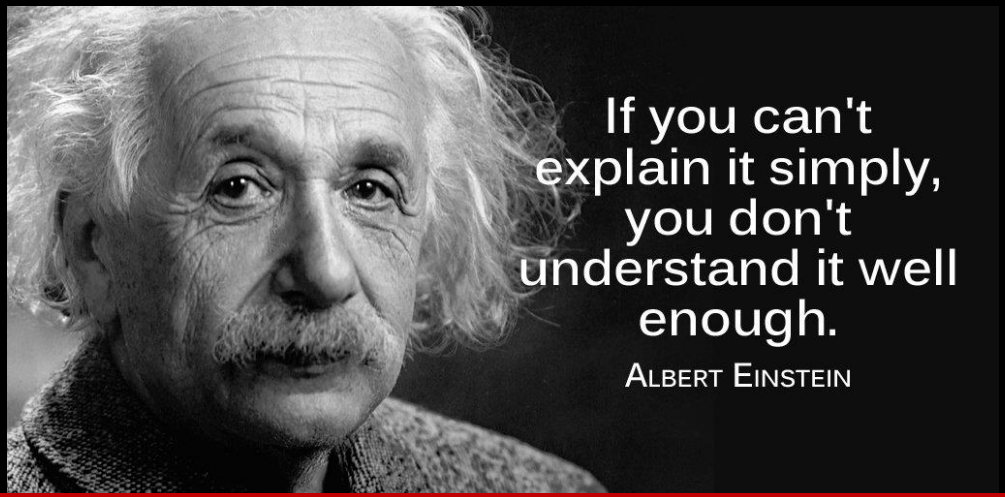
The unicorns of the new data world…

Which makes them hard to find…

https://www.figure-eight.com/figure-eight-2018-data-scientist-report/

http://www.marketingdistillery.com/2014/11/29/is-data-science-a-buzzword-modern-data-scientist-defined/

… but we are talking about this…



As data scientists, we must not only *Walk The Talk*,
but we must also must *Talk The Walk.*

There are 3 types of data folks:

1) **Talk the Talk**: those who use the buzzwords & talk the hype
2) **Walk the Talk**: those who can do the hard stuff
3) **Talk the Walk**: those who can explain the hard stuff

Data Literacy

"Data Literacy includes the ability to read, work with, analyze, and argue with data."

(Jordan Morrow, Qlik)

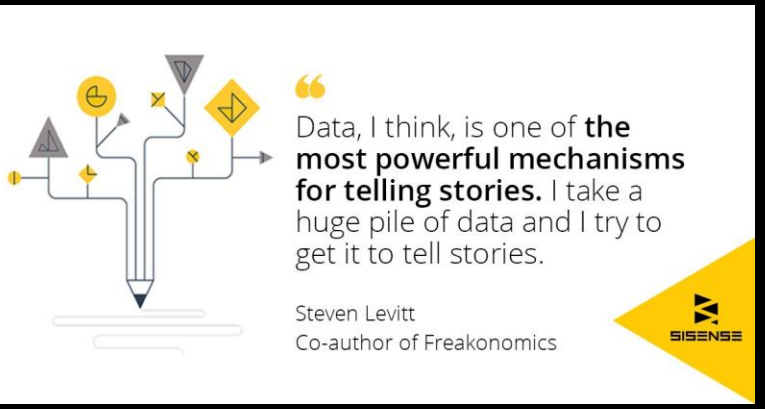
http://www.dataliteracynetwork.org/definitions.html

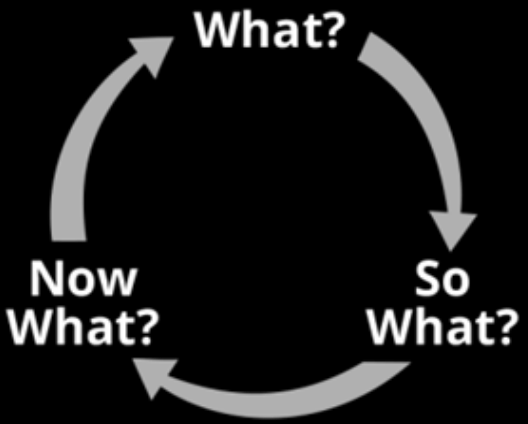
Source: http://bit.ly/2mEzJsr

Data Storytelling

"No one ever made a decision because of a number. They need a story."

Daniel Kahneman, Quoted in Vanity Fair article
"How Two Trailblazing Psychologists Turned the World of Decision Science Upside Down,"
November 2016





Focus on Creating Value from Data

Creating value at the "pull" of the customer!

"A pull strategy becomes more important than push because you want to create enough value so that the customer comes to you!"
https://www.marketing91.com/pull-strategy-in-marketing/

it's all about the
CUSTOMER

OUTLINE

- Talking the Walk
- **Data Literacy**
- Data Storytelling

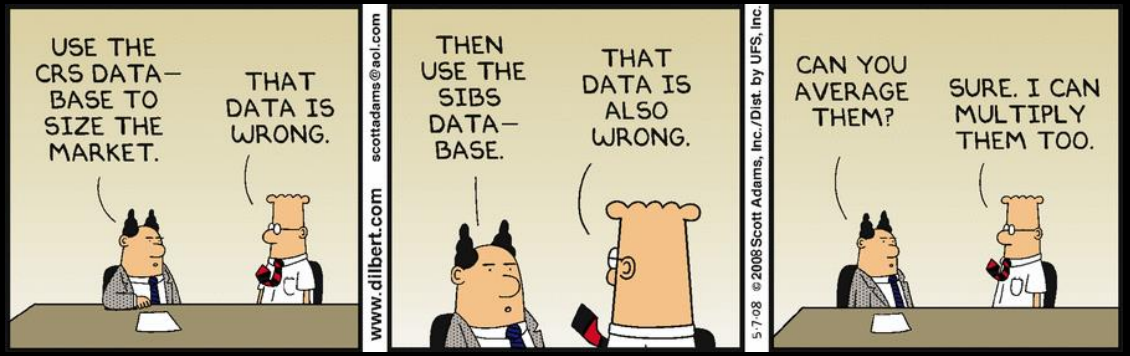
Source: https://en.wikipedia.org/wiki/New_Cuyama,_California

Data Literacy in 2 parts:

Data Science and Data Ethics

http://www.kirkborne.net/cds151/

1) How to use data correctly

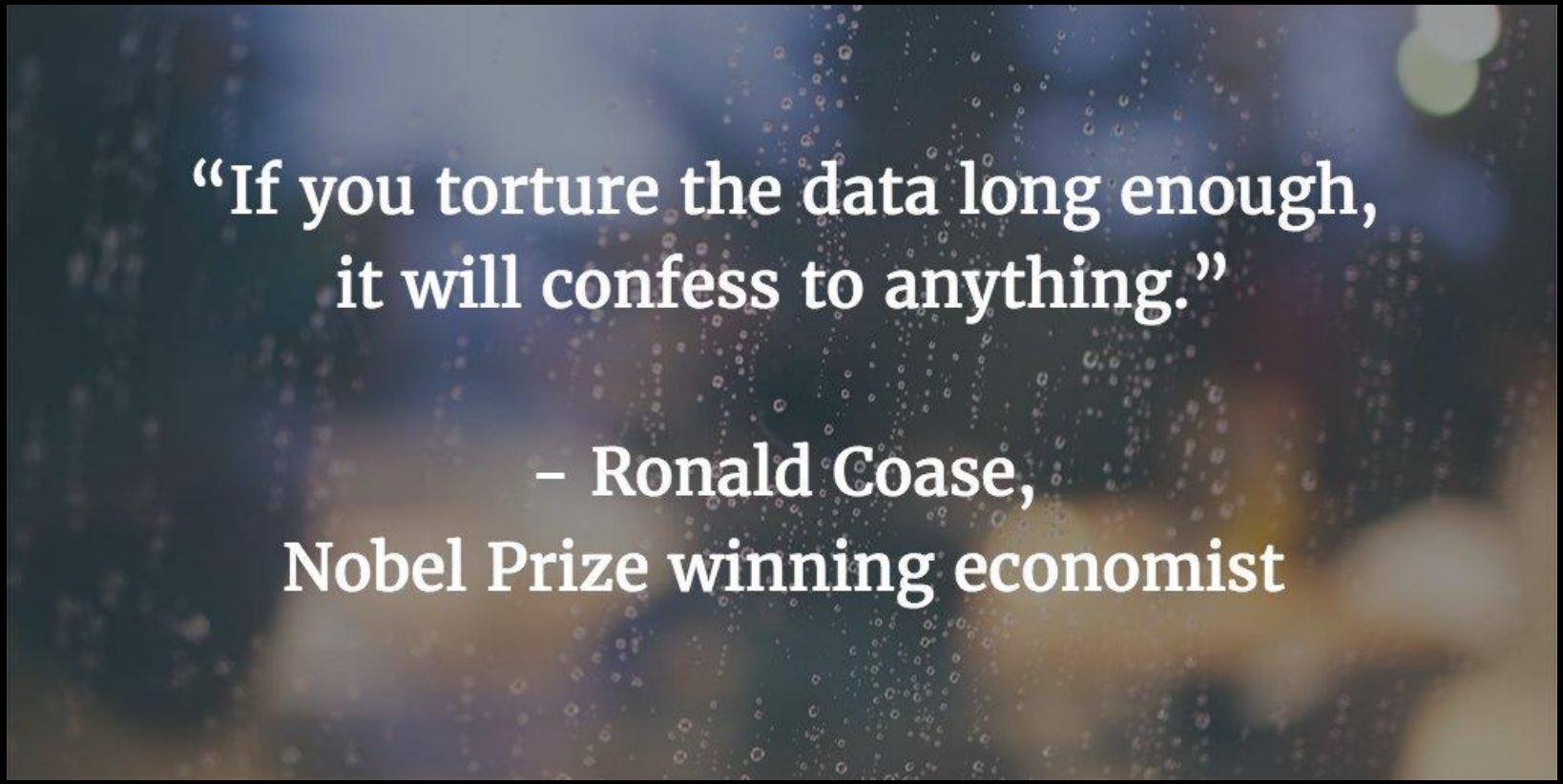


http://dilbert.com/strip/2008-05-07

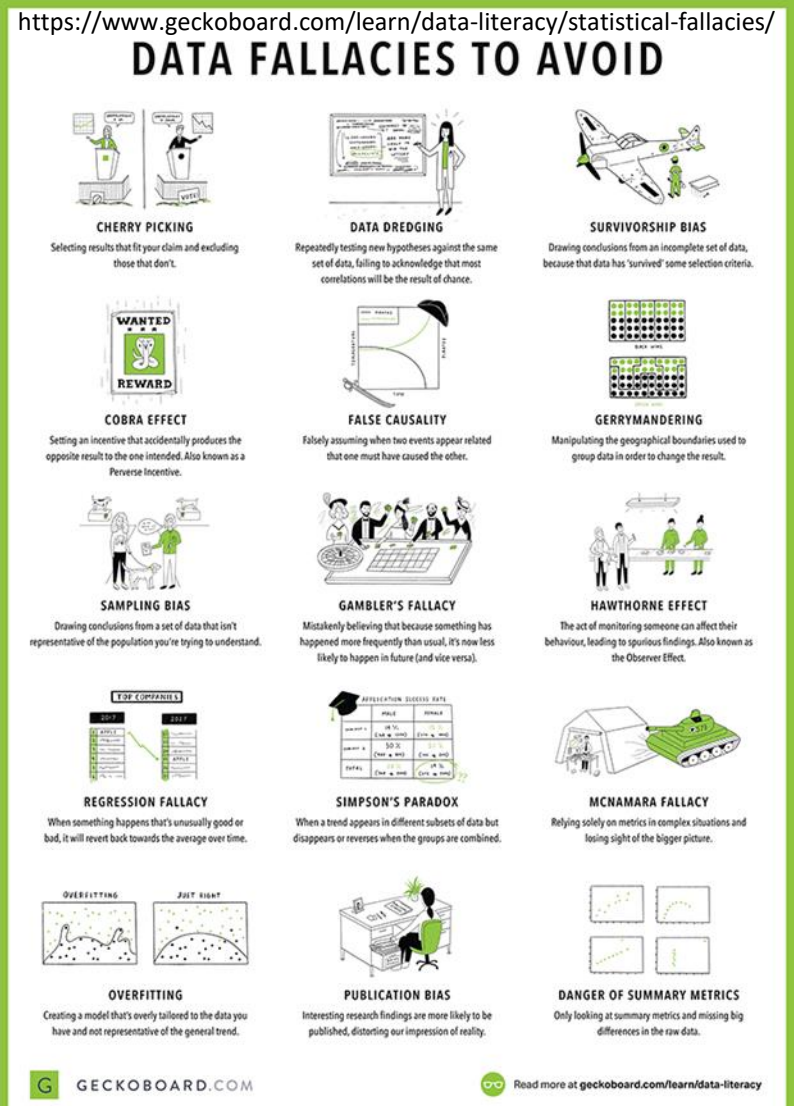
2) How to use data ethically

http://dilbert.com/strip/2000-11-13

In any collection of big data, you can always find correlations and patterns – but are these correlations random? or are they confessing some truth? or are they confirming some bias?



https://www.geckoboard.com/learn/data-literacy/statistical-fallacies/

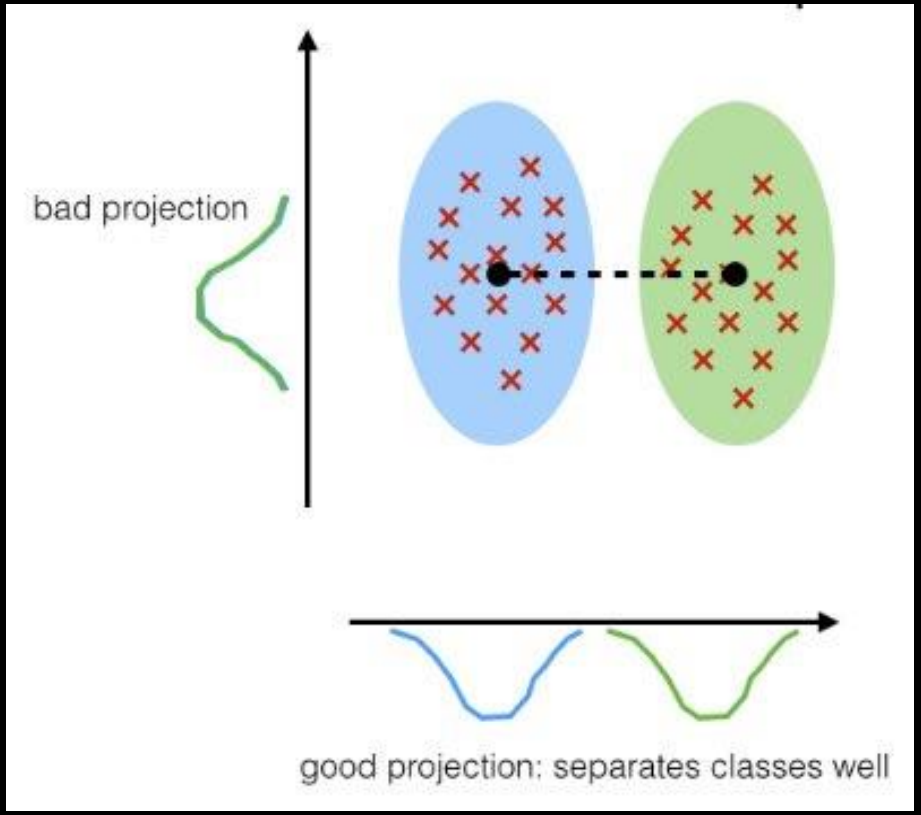


In our rush to build and to promote our models, we are often too quick to overlook our own cognitive biases and other data fallacies, such as:

"Correlation does not imply Causation!"

https://bit.ly/2pPnUSu

Feature Selection and Projection

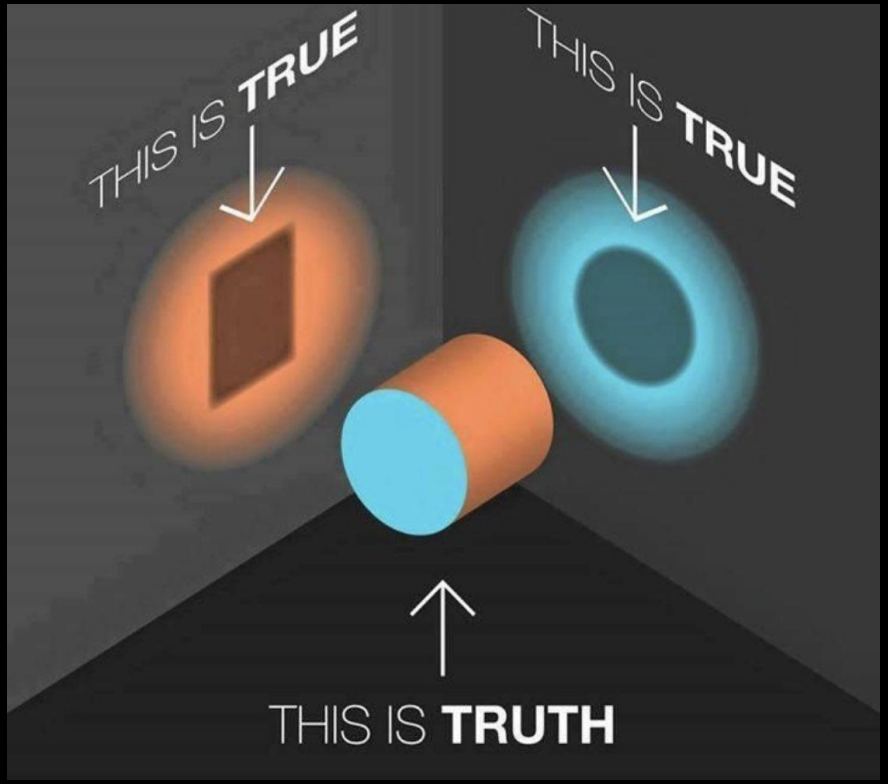


Source: https://www.quora.com/How-was-classification-as-a-learning-machine-developed

Feature Selection is important in order to disambiguate different classes.

More importantly, **Class Discovery** depends on choosing the right projection and selecting the right features!

High-Variety Data can be a Bias-Buster. Projection Matters!



Source: http://www.transformativeinsights.co.nz/blog/new-perspective-on-conflict

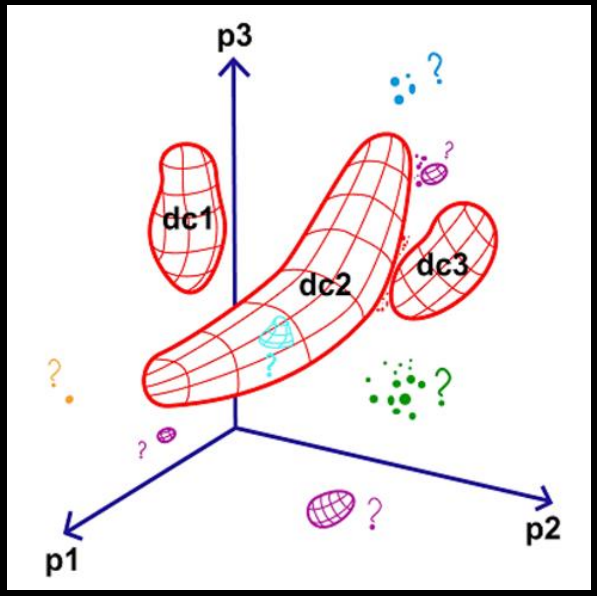
Your chosen data attributes represent a low-dimension projection of the full truth – the feature space (dimensions) in which you explore your data is a form of cognitive bias –

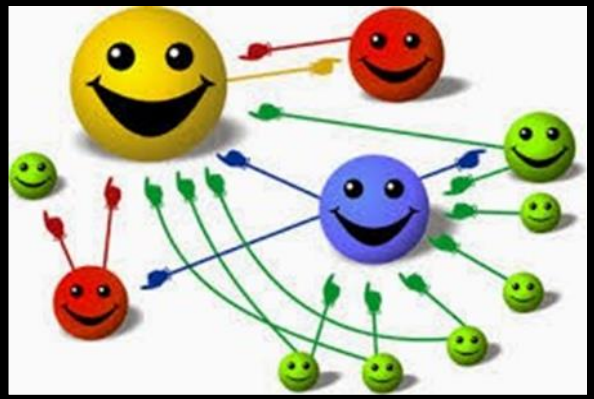
… **it matters!**

https://bit.ly/2CGHZjN

@KirkDBorne

4 Types of Insights Discovery from Data:





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 usual and the unusual (interesting) data associations / links / connections across the entities in your domain.

Levels of Analytics Maturity in Data-Driven Applications

1) Descriptive Analytics

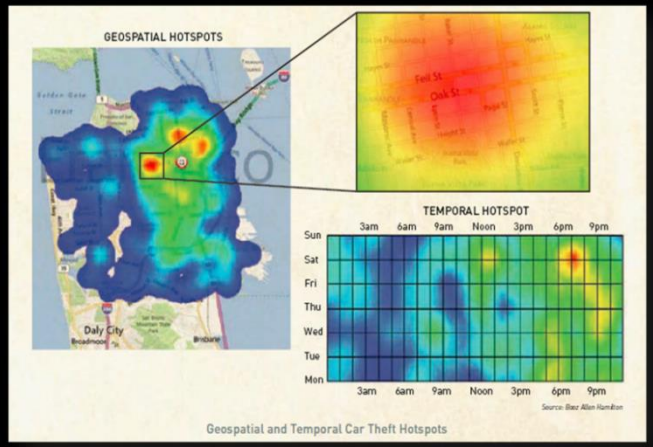
– **Hindsight** (What happened?)

2) Diagnostic Analytics

– **Oversight** (real-time / What is happening? Why did it happen?)

3) Predictive Analytics

– Foresight (What will happen?)



Geospatial and Temporal Car Theft Hotspots

5 Levels of Analytics Maturity in Data-Driven Applications

1) Descriptive Analytics

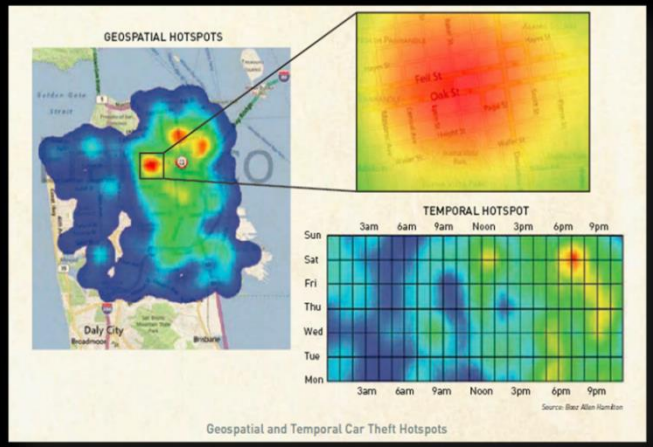
– **Hindsight** (What happened?)

2) Diagnostic Analytics

– **Oversight** (real-time / What is happening? Why did it happen?)

3) Predictive Analytics

– Foresight (What will happen?)



4) Prescriptive Analytics

– **Insight** (How can we optimize what happens?) (Follow the dots / connections in the graph!) **Insights Discovery**

5) Cognitive Analytics

– **Right Sight** (the 360 view , **what is the right question to ask for this set of data in this context** = Game of Jeopardy)

– Finds the right insight, the right action, the right decision,… right now!

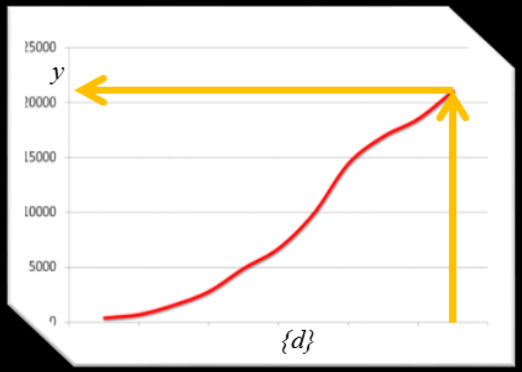
– Moves beyond simply providing answers, to generating new questions and hypotheses.

Predictive vs Prescriptive: What's the Difference?

PREDICTIVE *Analytics*

Find a function (i.e., the model) f(d,t) that predicts the value of some predictive variable y = f(d,t) at a future time t, given the set of conditions found in the training data {d}.

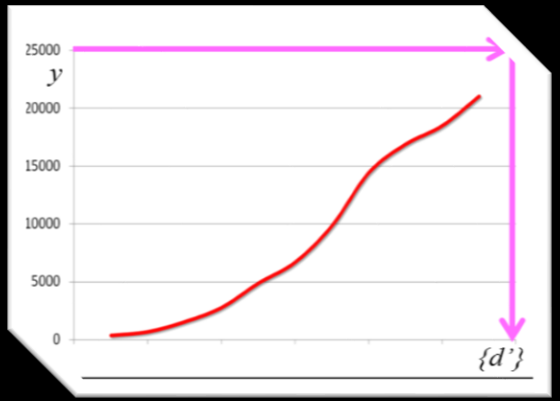
=> Given {d}, find y.



PRESCRIPTIVE *Analytics*

Find the conditions {d'} that will produce a prescribed (desired, optimum) value y at a future time t, using the previously learned conditional dependencies among the variables in the predictive function f(d,t).

=> Given y, find {d'}.

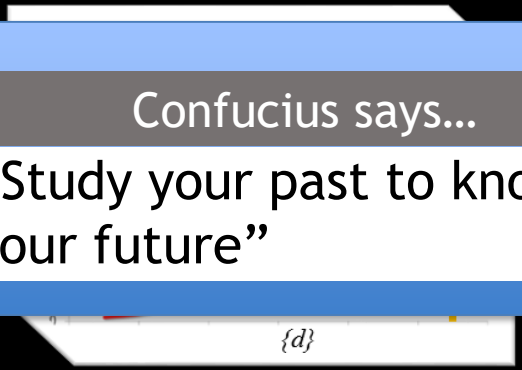


Predictive vs Prescriptive: What's the Difference?

PREDICTIVE *Analytics*

Find a function (i.e., the model) f(d,t) that predicts the value of some predictive variable y = f(d,t) at a future time t, given the set of conditions found in the training data {d}.

=> Given {d}, find y.

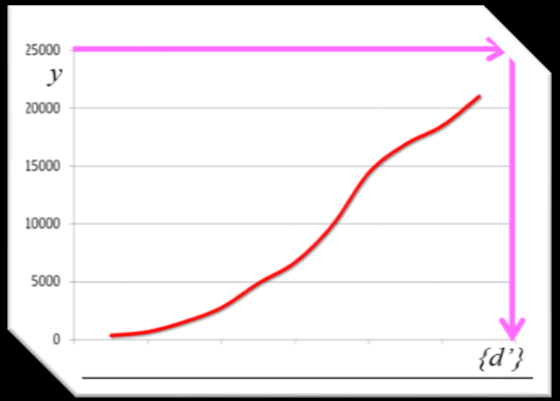


> Confucius says…
>
> "Study your past to know your future"

PRESCRIPTIVE *Analytics*

Find the conditions {d'} that will produce a prescribed (desired, optimum) value y at a future time t, using the previously learned conditional dependencies among the variables in the predictive function f(d,t).

=> Given y, find {d'}.

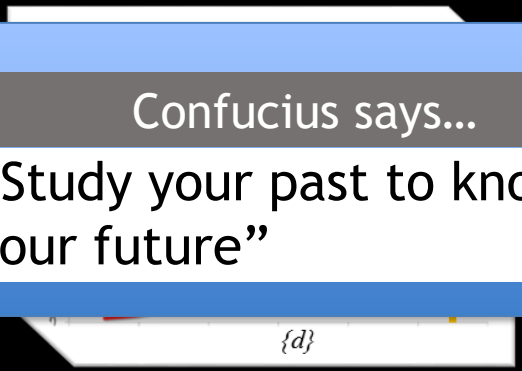


Predictive vs Prescriptive: What's the Difference?

PREDICTIVE *Analytics*

Find a function (i.e., the model) f(d,t) that predicts the value of some predictive variable y = f(d,t) at a future time t, given the set of conditions found in the training data {d}.

=> Given {d}, find y.



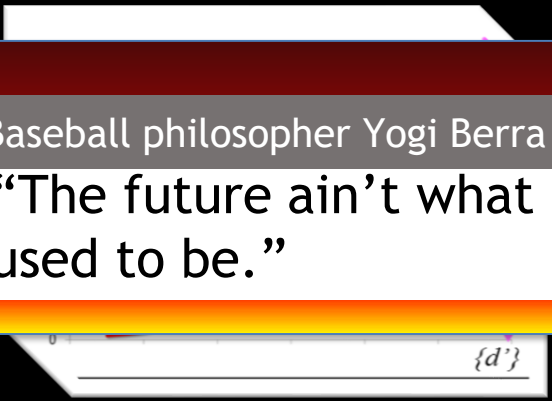
Confucius says…

"Study your past to know your future"

PRESCRIPTIVE *Analytics*

Find the conditions {d'} that will produce a prescribed (desired, optimum) value y at a future time t, using the previously learned conditional dependencies among the variables in the predictive function f(d,t).

=> Given y, find {d'}.

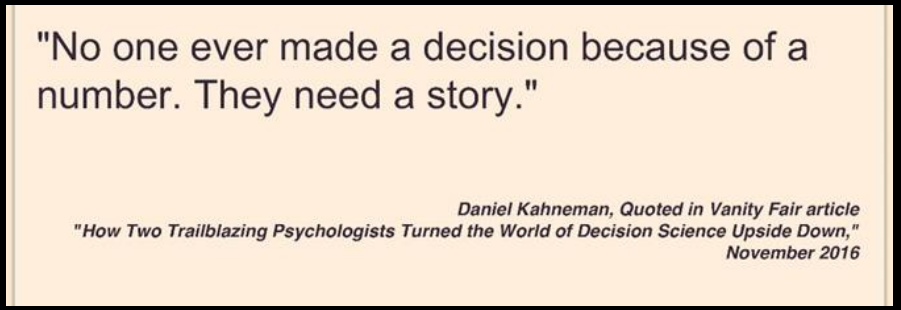


Baseball philosopher Yogi Berra says…

"The future ain't what it used to be."

OUTLINE

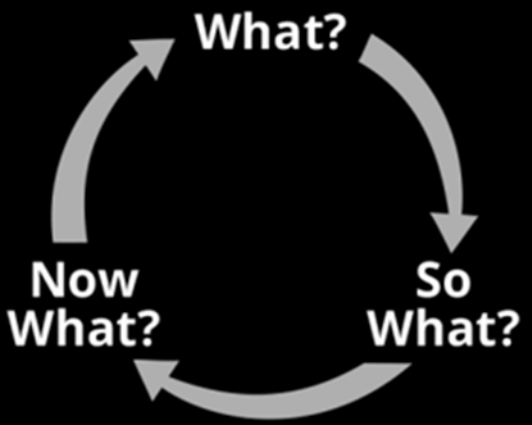
- Talking the Walk
- Data Literacy
- **Data Storytelling**



Source: https://www.vanityfair.com/news/2016/11/decision-science-daniel-kahneman-amos-tversky

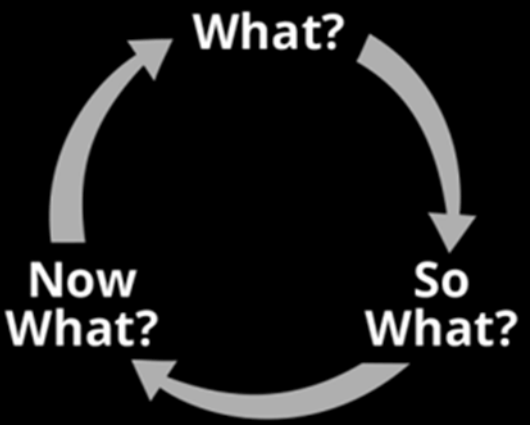
3 Short Data Stories

1) Counting
2) Associations
3) Graphs



3 Short Data Stories

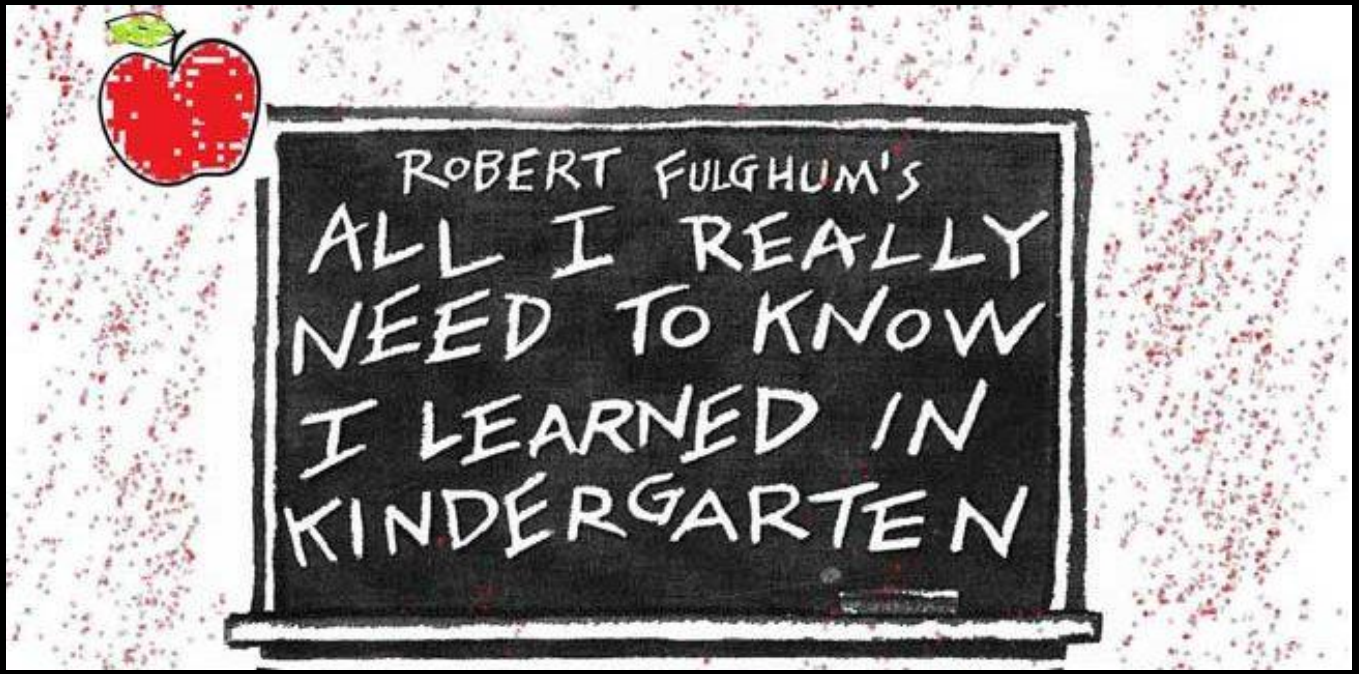
1) **Counting**
2) Associations
3) Graphs



Simple as possible, but very effective!

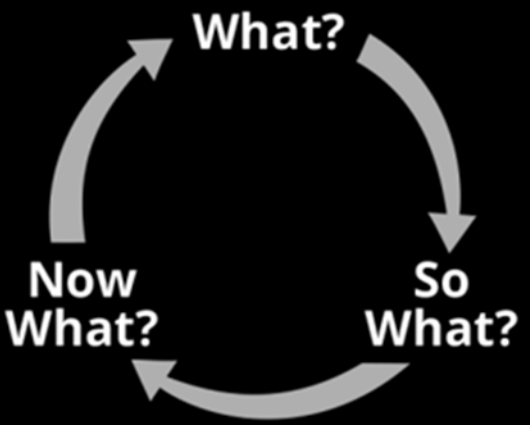
1) Counting!

Remember this…



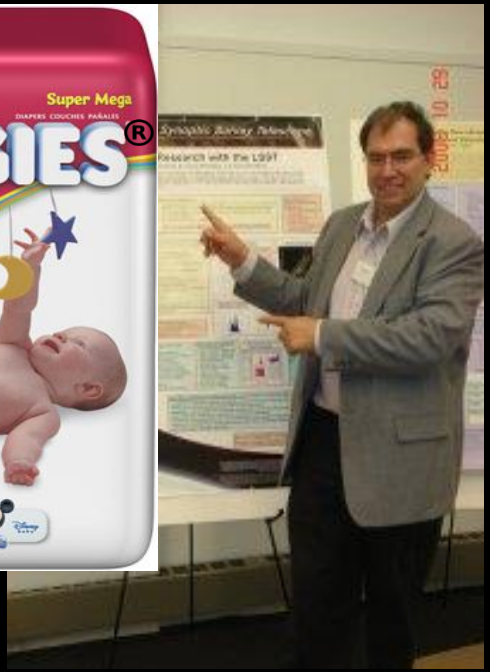
3 Short Data Stories

1) Counting
2) **Associations**
3) Graphs



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.

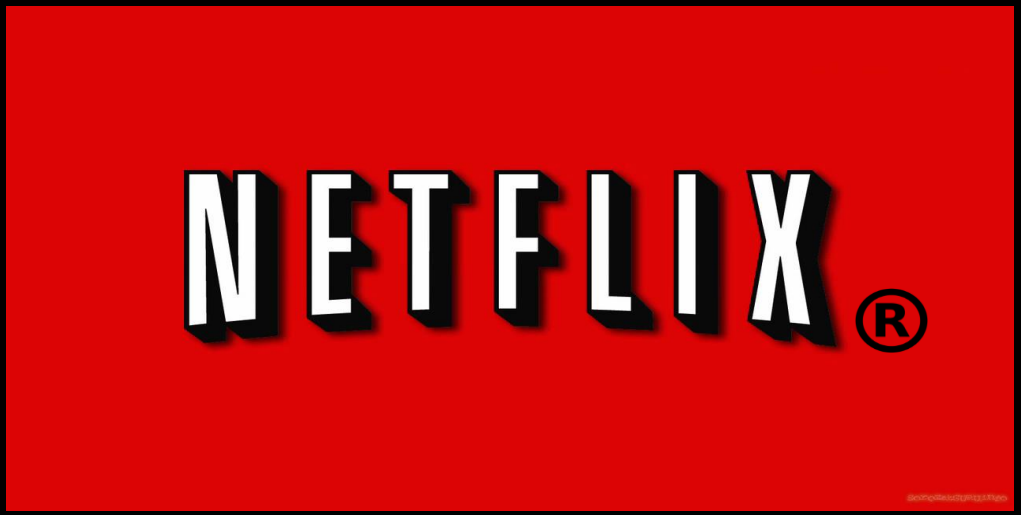


Association Discovery Example #2

- **Amazon.com** mines its customers' purchase logs to recommend books to you: *"People who bought this book also bought this other one."*

Association Discovery Example #3

- **Netflix** mines its video rental history database to recommend rentals to you based upon other customers who rented similar movies as you.



Association Discovery Example #4

- **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 #4

- **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???

May be explained with a composite scoring model:

https://en.wikipedia.org/wiki/Weighted_sum_model

Suggested readings:

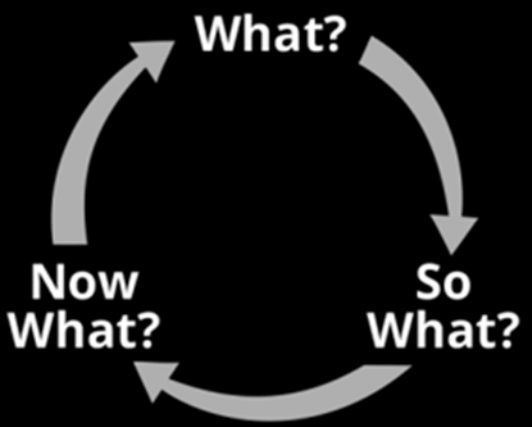
http://www.nytimes.com/2004/11/14/business/yourmoney/14wal.html

http://www.hurricaneville.com/pop_tarts.html

http://bit.ly/1gHZddA

3 Short Data Stories

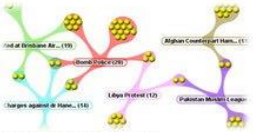
1) Counting
2) Associations
3) **Graphs**



"All the World is a Graph" - Shakespeare?

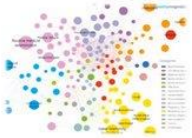
The natural data structure of the world is not rows and columns, but a Graph!

Discovery/Graph Analytics is everywhere…



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

…

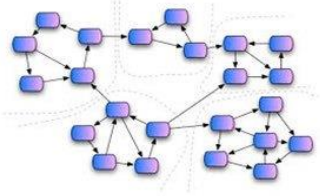


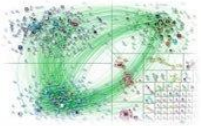
Healthcare
- Personalized Treatment
- Fraud Detection
- Efficacy of Care
- Adverse Event Clustering
- Disease Prediction

…

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

…

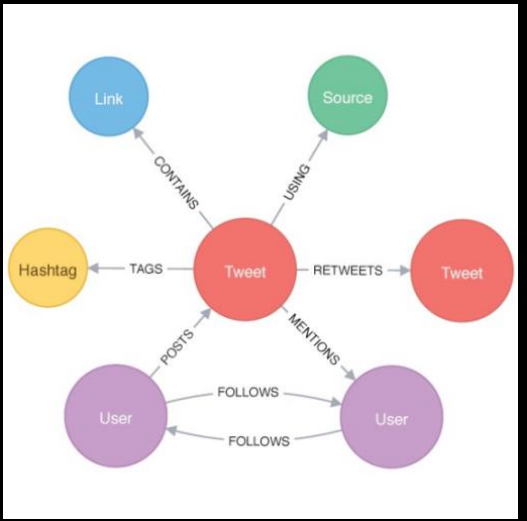
…

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

…

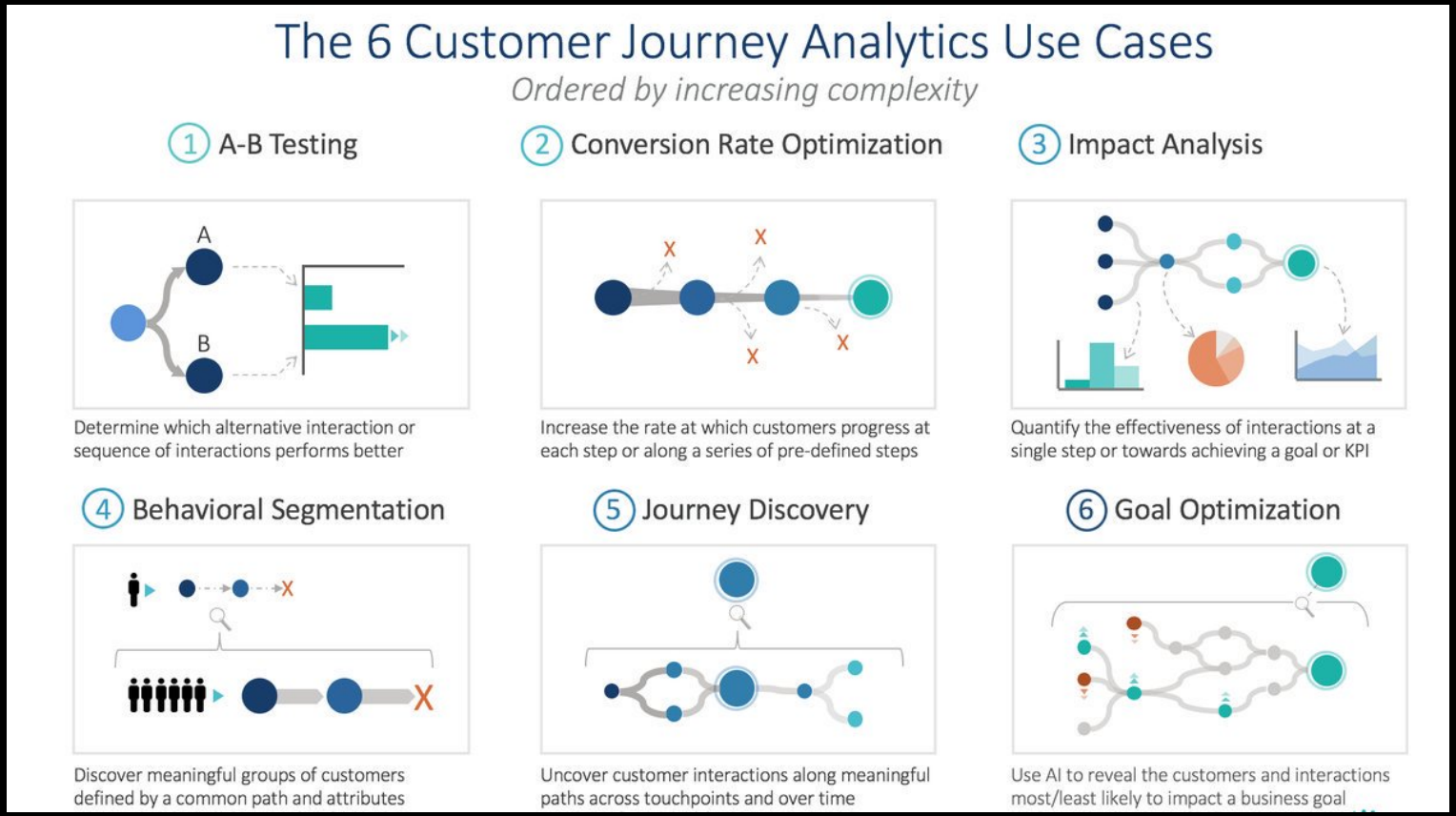
(Graphic by Cray, for Cray Graph Engine CGE)

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



Customer Journey Analytics by Pointillist.com –

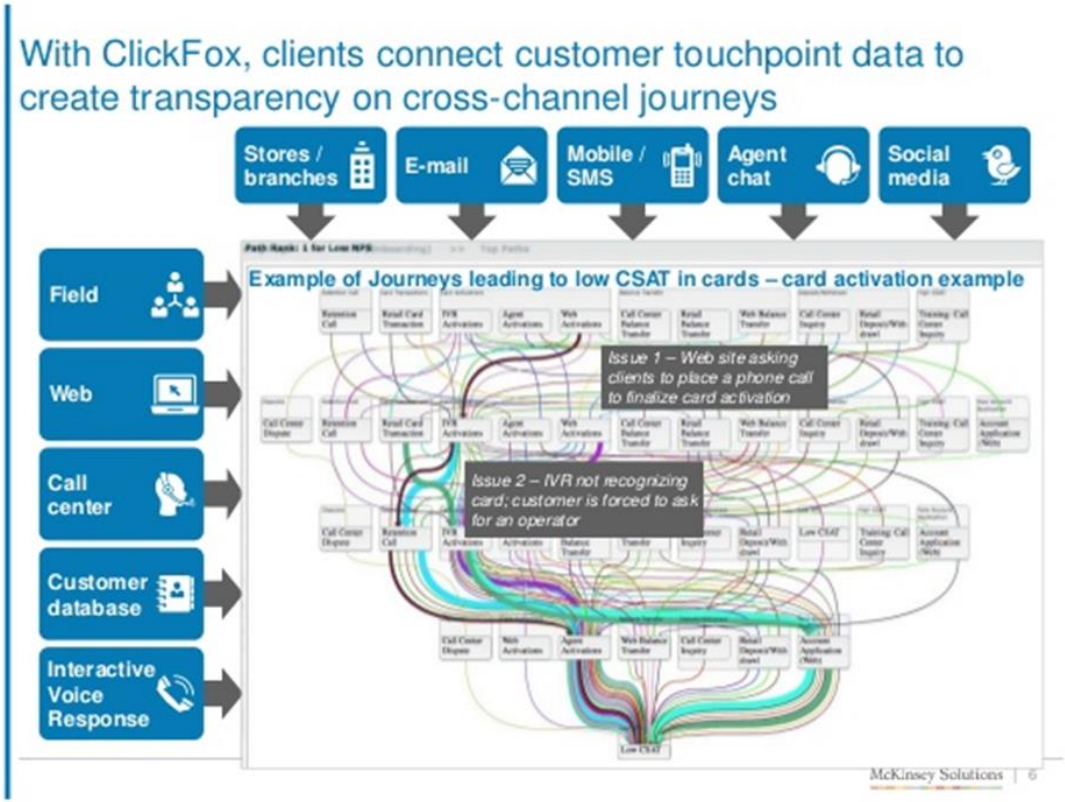
The Journey Graph tells the Customer's Story, and predicts Customer Outcomes with high accuracy!



https://www.pointillist.com/blog/how-to-successfully-implement-customer-journey-analytics-part-1/

Customer Journey Analytics by Clickfox.com –

The Journey Graph tells the Customer's Story, and predicts Customer Outcomes with high accuracy!



https://www.slideshare.net/Qualtrics/how-to-leverage-analytics-design-and-development-to-transform-customer-journeys

★ Data Journalism Handbook (2nd edition) – great resource for Data Literacy and Data Storytelling!

https://datajournalismhandbook.org/

The Data Journalism Handbook 2

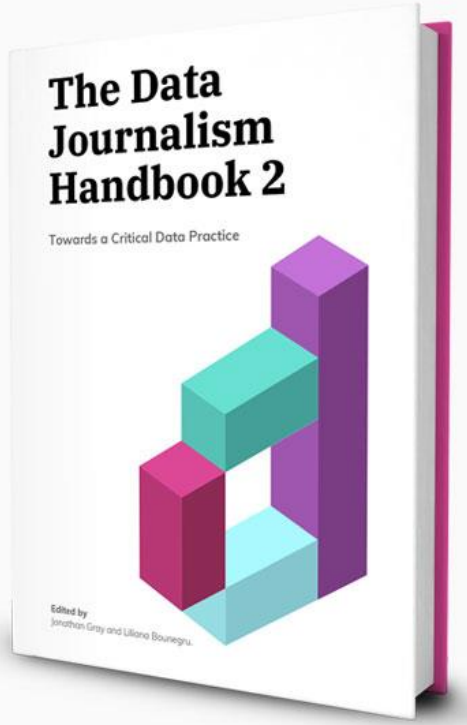
Towards a Critical Data Practice

BETA NOW AVAILABLE

Produced by

European Journalism Centre

Google News Initiative



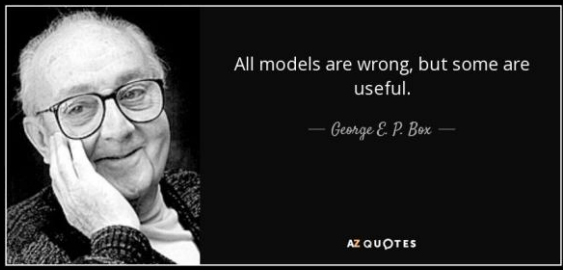
https://twitter.com/KirkDBorne/status/1088488281929588744

Final Reminders

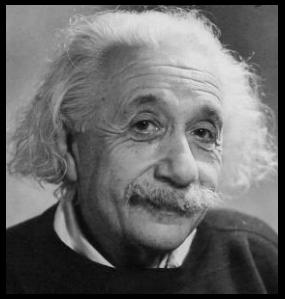
As data scientists, we must not only *Walk The Talk*, but we must also must *Talk The Walk.*

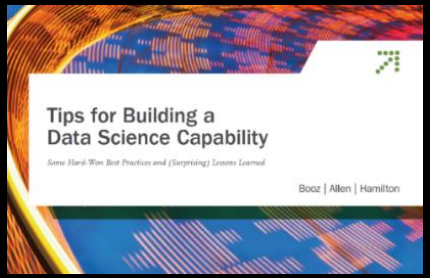
2 Reminders for Data Literacy and for Data Storytelling

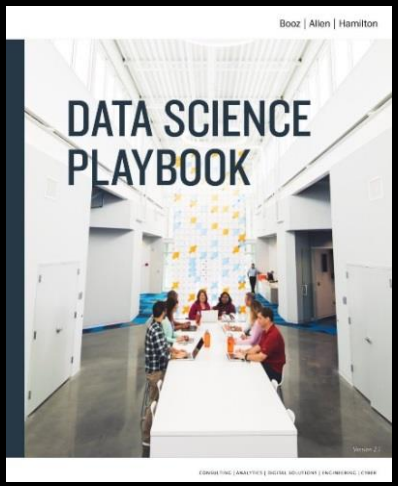
- ***"All models are wrong, but some are useful."*** – George Box

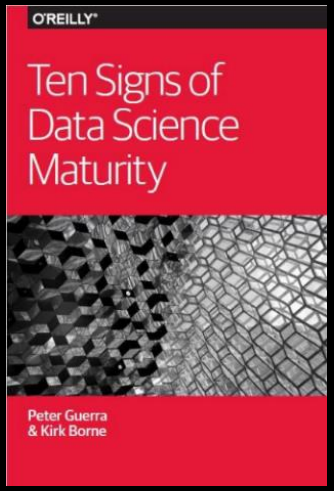


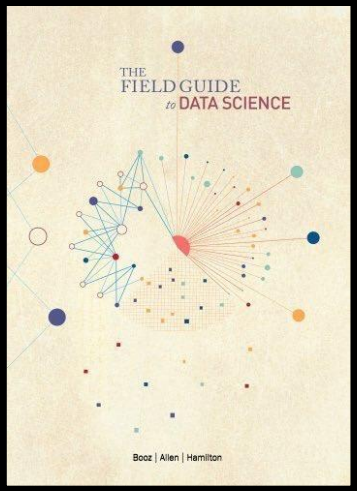
- ***"Everything should be made as simple as possible, but not simpler"*** – Albert Einstein











Come for the Data. Stay for the Science!

Thank you!

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

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

Get slides here: http://www.kirkborne.net/INSAID2019/

http://www.boozallen.com/datascience

@KirkDBorne