

**"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

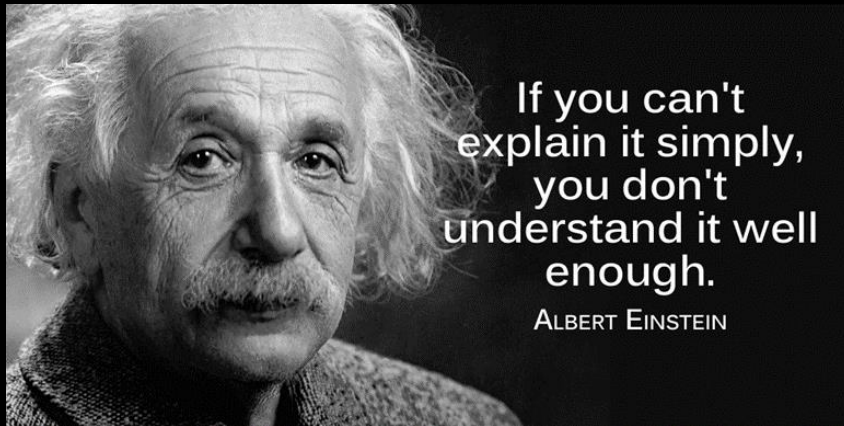
- Talking the Walk
- Data Literacy
- Data Storytelling



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...

Applied Science

- Statistics, applied math
- Machine Learning
- Tools: Python, R, SAS

Business Analysis

- Data Analysis, BI
- Business/domain expertise
- Tools: SQL, Excel, EDW

Data engineering

- Database technologies
- Computer science
- Tools: Java, Scala, Python, C++

Engineering

- Big data pipeline engineering
- Statistics and machine learning over large datasets
- Tools: Hadoop, PIG, HIVE, Cascading, SOLR, etc

figure eight

The Data Scientist Report 2018



EBOOK

MODERN DATA SCIENTIST

Data Scientist, the sexiest job of 21st 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

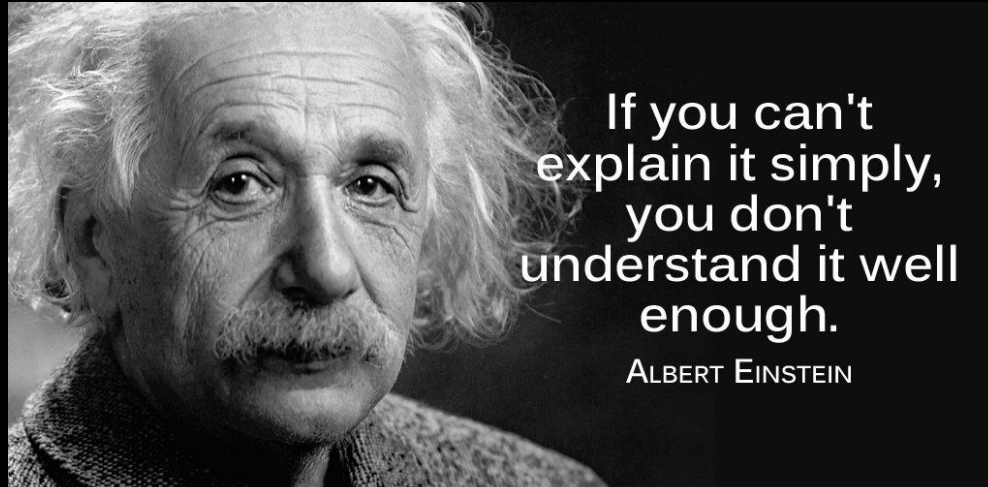
- ☆ Able to engage with senior management
- ☆ Story telling skills
- ☆ Translate data-driven insights into decisions and actions
- ☆ Visual art design
- ☆ R packages like ggplot or lattice
- ☆ Knowledge of any of visualization tools e.g. Flare, D3.js, Tableau



MarketingDistillery.com is a group of practitioners in the area of e-commerce marketing. Our fields of expertise include: marketing strategy and optimization; customer tracking and on-site analytics; predictive analytics and econometrics; data warehousing and big data systems; marketing channel insights in Paid Search, SEO, Social, CRM and brand.

Marketing
DISTILLERY

... 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>

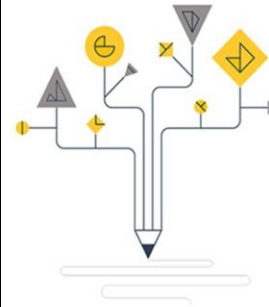


Source: <http://bit.ly/2mEzJsr>

Data Storytelling

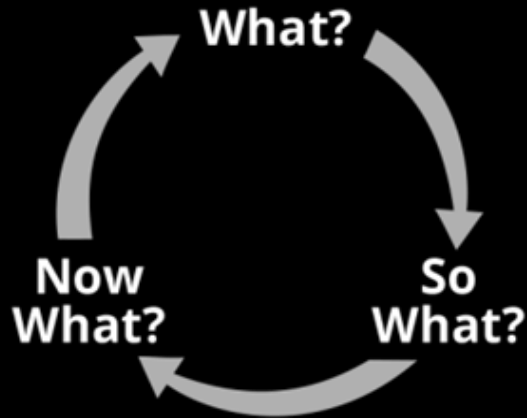
"No one ever made a decision because of a number. They need a story."

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



“Data, I think, is one of **the most powerful mechanisms for telling stories**. I take a huge pile of data and I try to get it to tell stories.

Steven Levitt
Co-author of *Freakonomics*



"People will forget what you said,
people will forget what you did,
but people will never forget
how you made them feel"

Maya Angelou

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



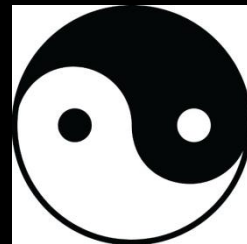
Source: https://en.wikipedia.org/wiki/New_Cuyama,_California



Data Literacy in 2 parts:

Data Science and Data Ethics

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

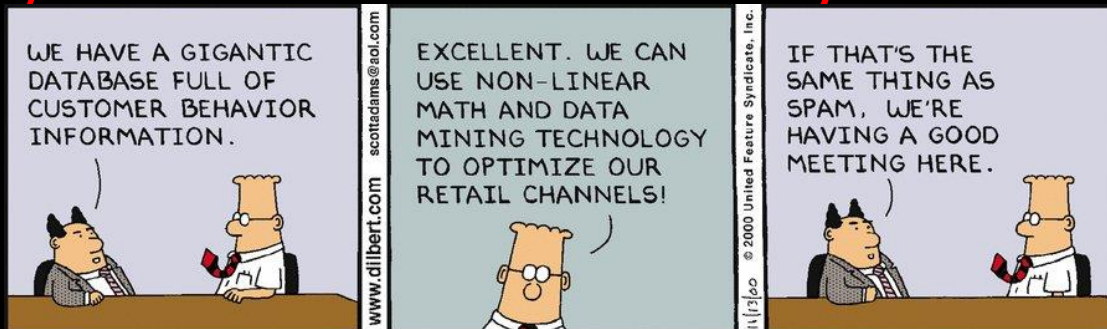


1) How to use data correctly



<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?

**“If you torture the data long enough,
it will confess to anything.”**

**– Ronald Coase,
Nobel Prize winning economist**

DATA FALLACIES TO AVOID



CHERRY PICKING

Selecting results that fit your claim and excluding those that don't.



DATA DREDGING

Repeatedly testing new hypotheses against the same set of data, failing to acknowledge that most correlations will be the result of chance.



SURVIVORSHIP BIAS

Drawing conclusions from an incomplete set of data, because that data has 'survived' some selection criteria.



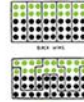
COBRA EFFECT

Setting an incentive that accidentally produces the opposite result to the one intended. Also known as a Perverse Incentive.



FALSE CAUSALITY

Falsely assuming when two events appear related that one must have caused the other.



GERRYMANDERING

Manipulating the geographical boundaries used to group data in order to change the result.



SAMPLING BIAS

Drawing conclusions from a set of data that isn't representative of the population you're trying to understand.



GAMBLER'S FALLACY

Mistakenly believing that because something has happened more frequently than usual, it's now less likely to happen in future (and vice versa).



HAWTHORNE EFFECT

The act of monitoring someone can affect their behaviour, leading to spurious findings. Also known as the Observer Effect.



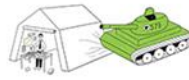
REGRESSION FALLACY

When something happens that's unusually good or bad, it will revert back towards the average over time.



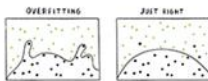
SIMPSON'S PARADOX

When a trend appears in different subsets of data but disappears or reverses when the groups are combined.



MCHAMARA FALLACY

Relying solely on metrics in complex situations and losing sight of the bigger picture.



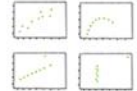
OVERFITTING

Creating a model that's overly tailored to the data you have and not representative of the general trend.



PUBLICATION BIAS

Interesting research findings are more likely to be published, distorting our impression of reality.



ANGER OF SUMMARY METRICS

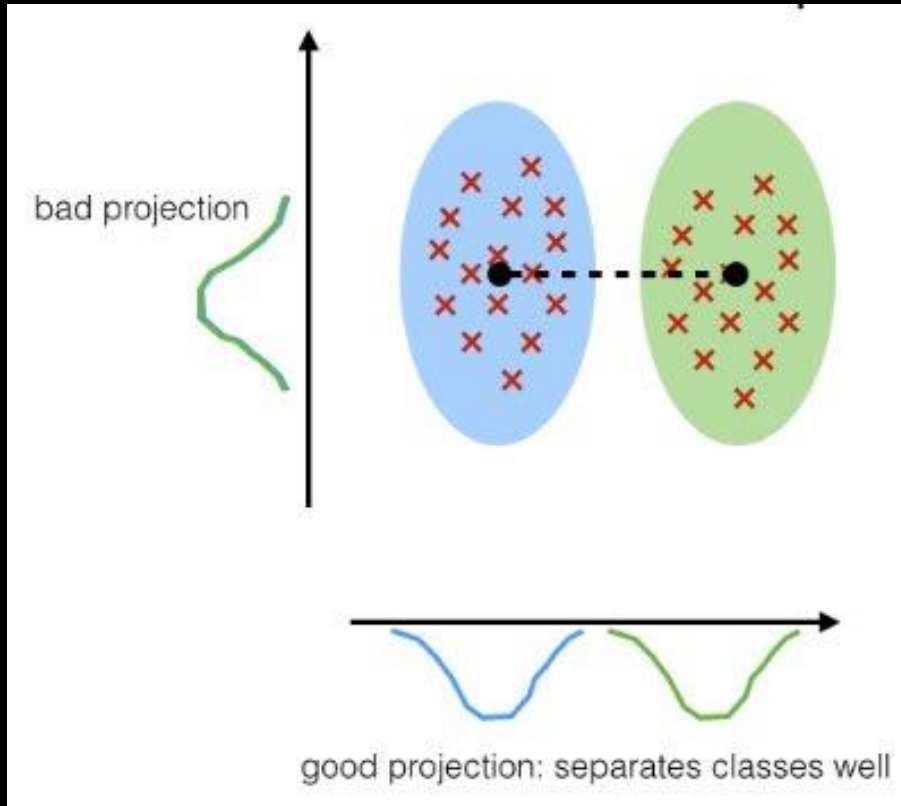
Only looking at summary metrics and missing big differences in the raw data.

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



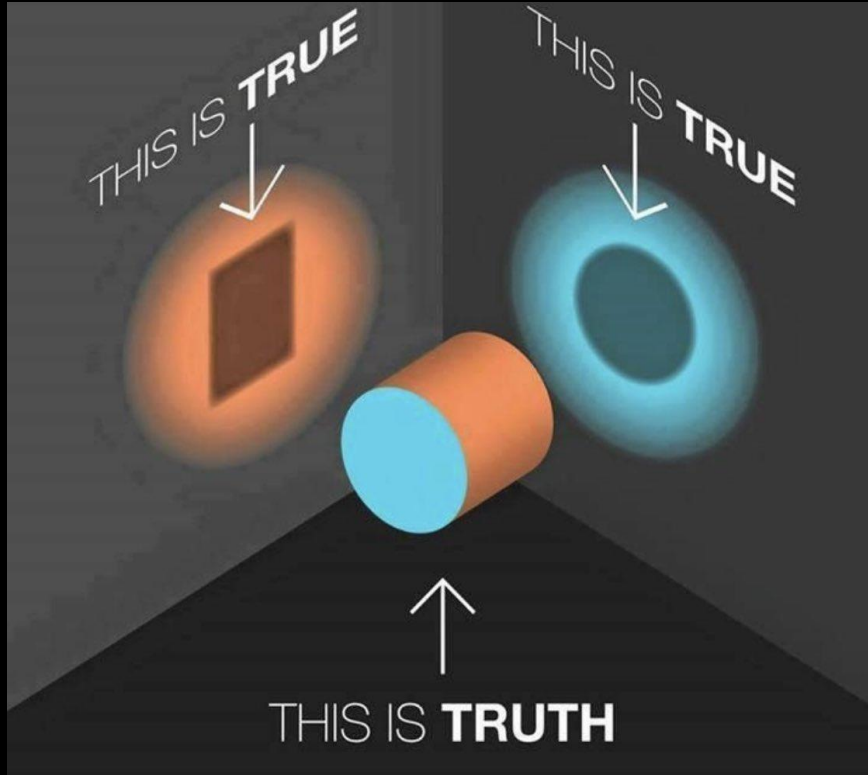
Feature Selection is important in order to disambiguate different classes.

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

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

High-Variety Data can be a Bias-Buster.

Projection Matters!



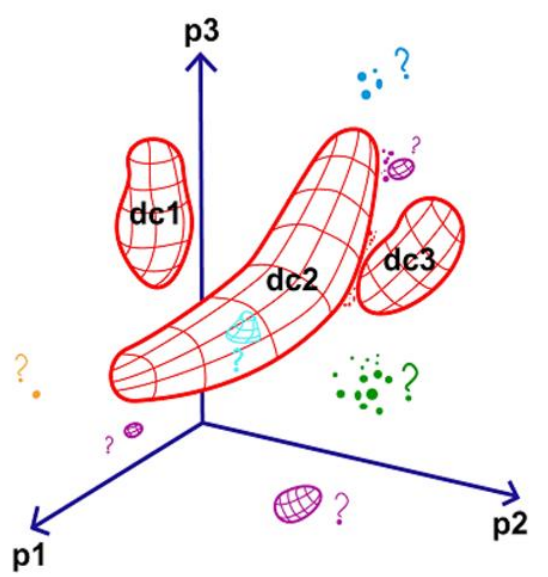
Source: <http://www.transformativeinsights.co.nz/blog/new-perspective-on-conflict>

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:

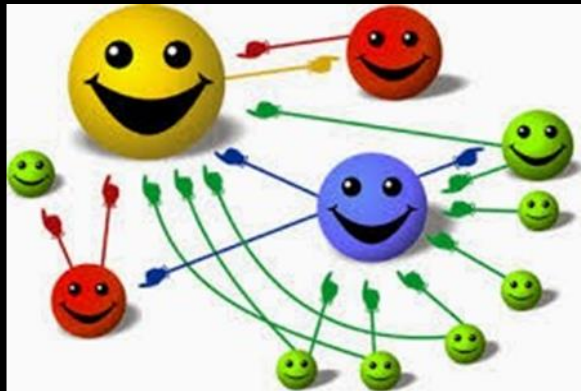


1) **Class Discovery:** Find the categories of objects (population segments), events, and behaviors in your data. + Learn the rules that constrain the class boundaries (that uniquely distinguish them).

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

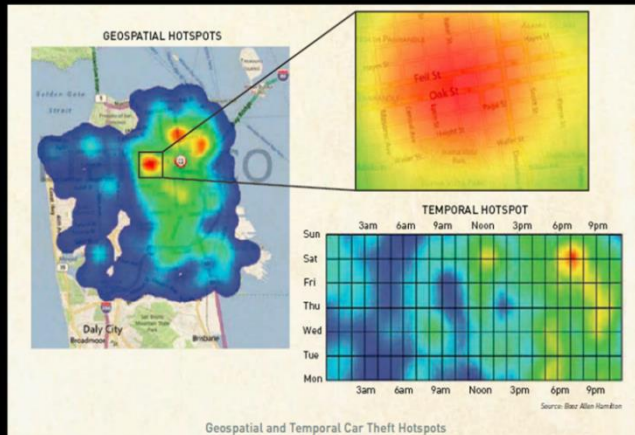
- **Hindsight** (What happened?)

2) Diagnostic Analytics

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

3) Predictive Analytics

- **Foresight** (What will happen?)



5 Levels of Analytics Maturity in Data-Driven Applications

1) Descriptive Analytics

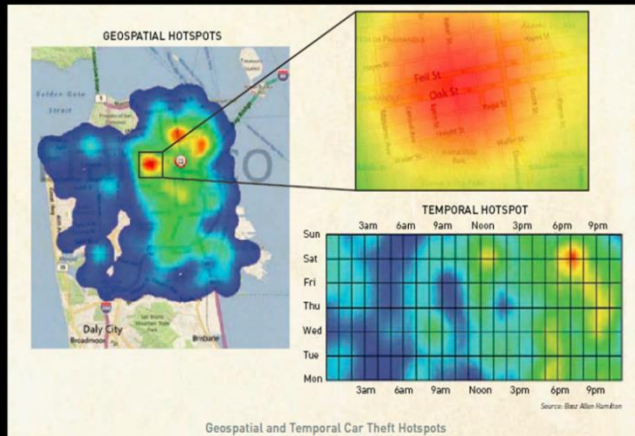
- **Hindsight** (What happened?)

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.**



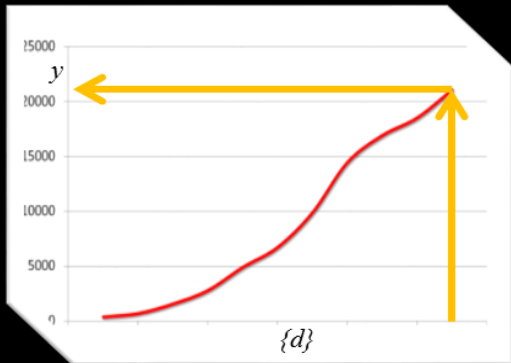
Predictive vs Prescriptive: What's the Difference?

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 .

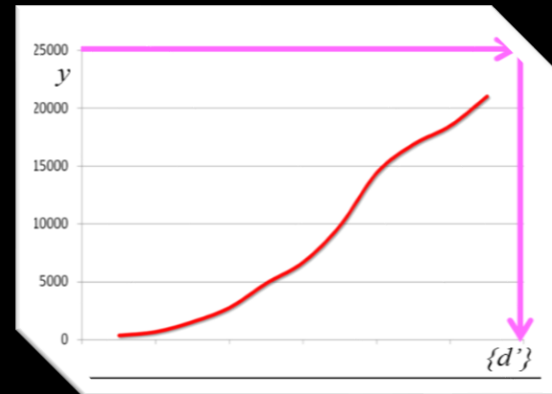


PRESCRIPTIVE

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”

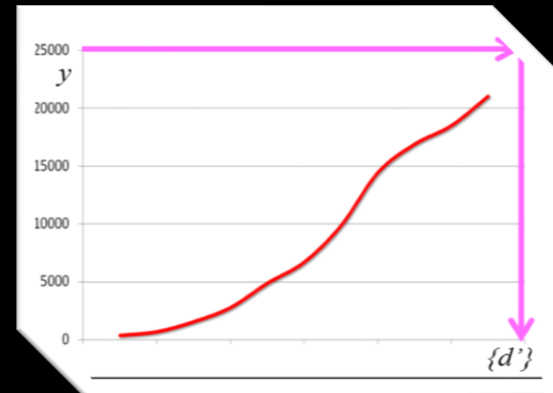
$\{d\}$

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”

$\{d\}$

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.”

$\{d'\}$

OUTLINE

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

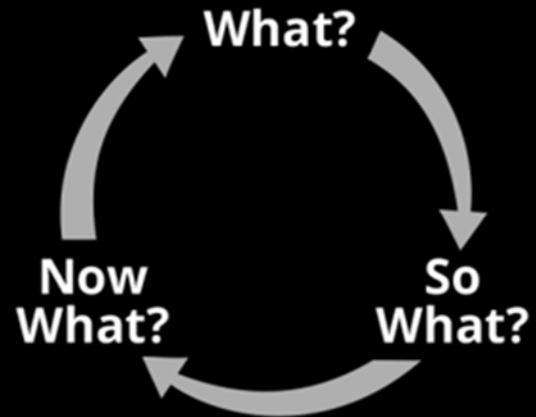
"No one ever made a decision because of a number. They need a story."

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

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

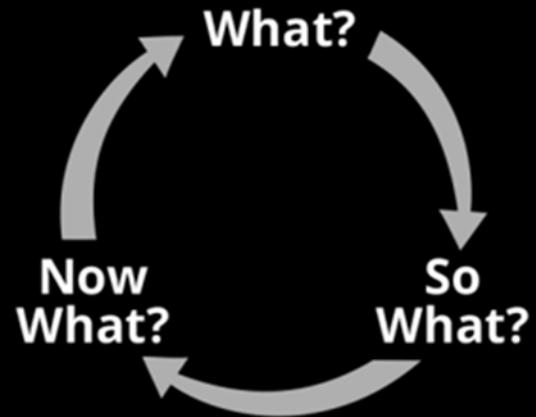
3 Short Data Stories

- 1) Counting
- 2) Associations
- 3) Graphs



3 Short Data Stories

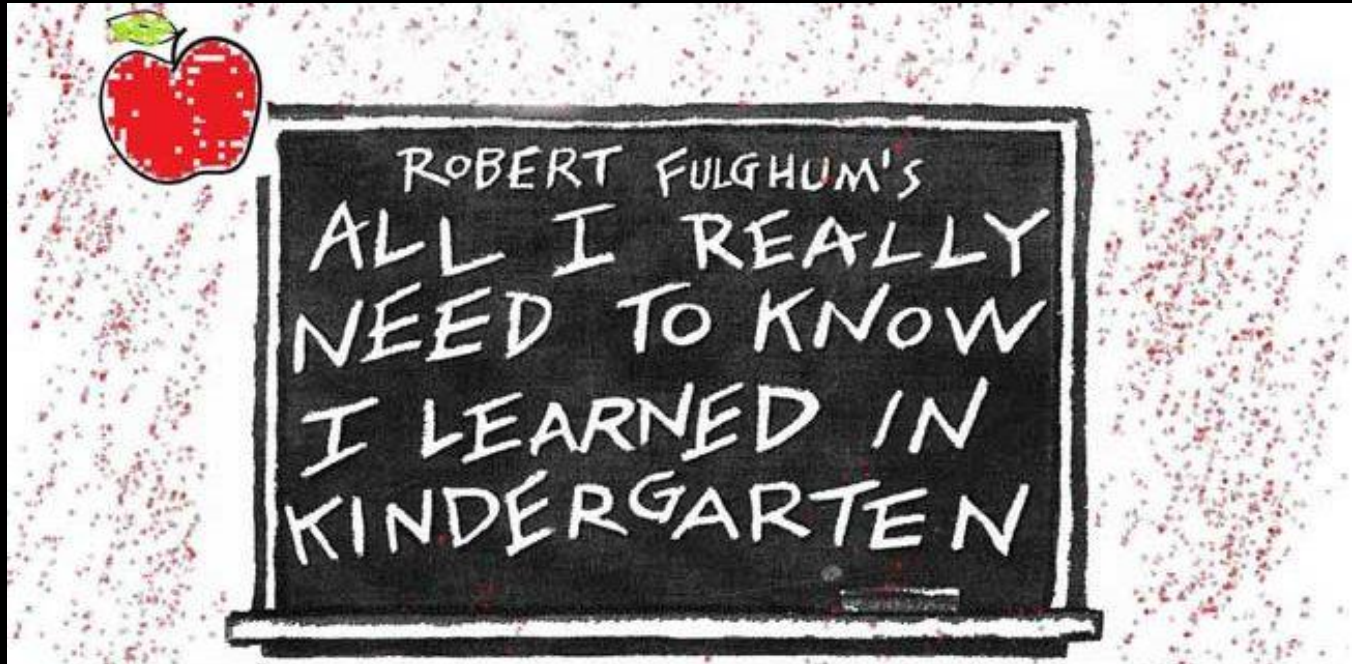
- 1) **Counting**
- 2) Associations
- 3) Graphs



Simple as possible, but very effective!

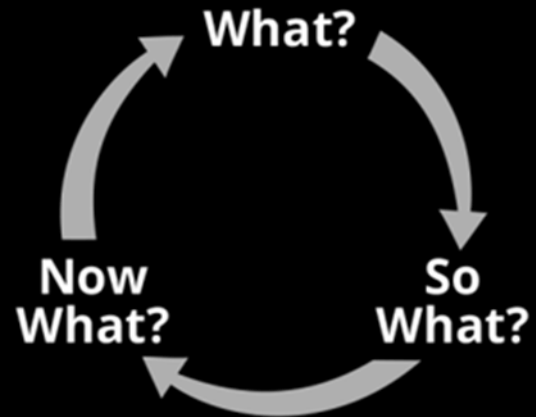
1) Counting!

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.**



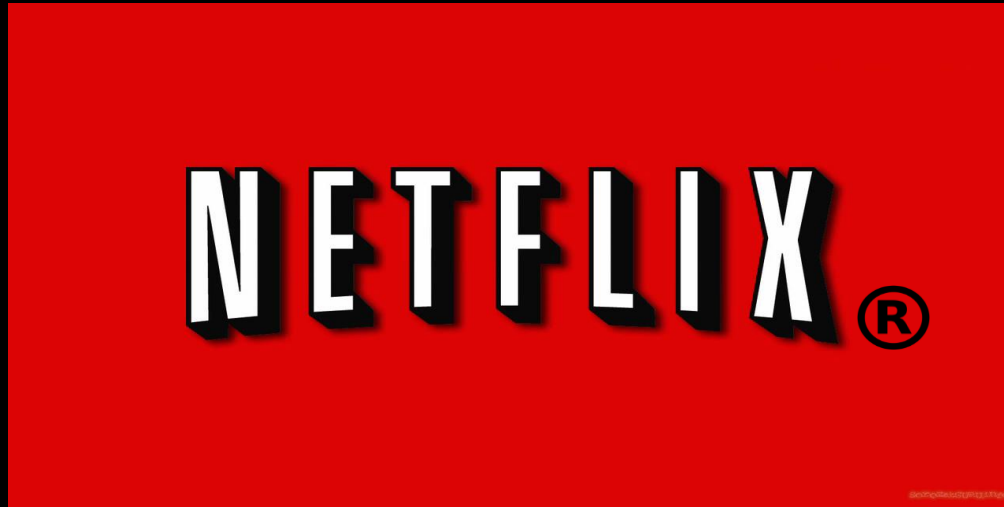
Association Discovery Example #2

- **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:

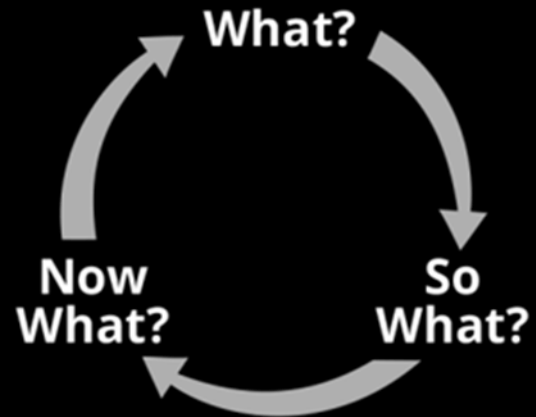
<http://www.nytimes.com/2004/11/14/business/yourmoney/14wal.html>

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

<http://bit.ly/1gHZddA>

3 Short Data Stories

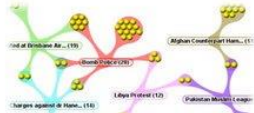
- 1) Counting
- 2) Associations
- 3) Graphs**



“All the World is a Graph” - Shakespeare?

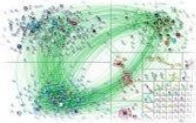
The natural data structure of the world is not rows and columns, but a Graph!

Discovery/Graph Analytics is everywhere...



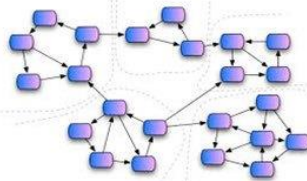
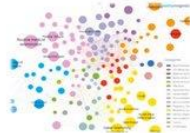
Government/Security

- Patterns of Activity Analytics
- CyberThreat Discovery
- Tax Fraud Discovery
- Crime Prediction



Telecom/Media

- Influencer Discovery
- Churn Analytics
- Behavior Analytics



Life Sciences

- Drug Discovery
- Drug Repurposing
- Clinical Trial Mining



Healthcare

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



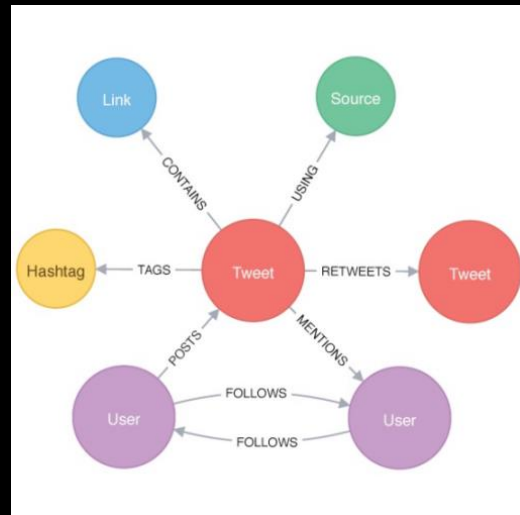
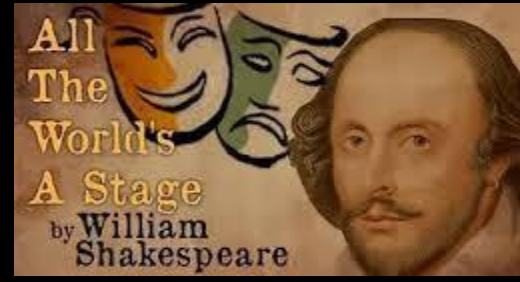
Energy/Resources

- Location Discovery
- Field Production Analysis
- Contingency Analysis
- Climate Modeling



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>

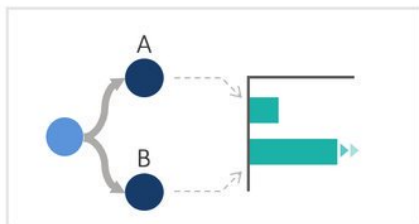
Customer Journey Analytics by Pointillist.com –

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

The 6 Customer Journey Analytics Use Cases

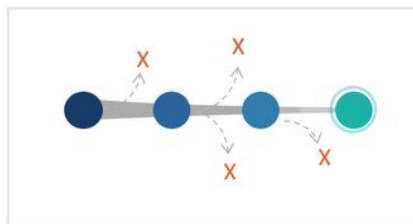
Ordered by increasing complexity

① A-B Testing



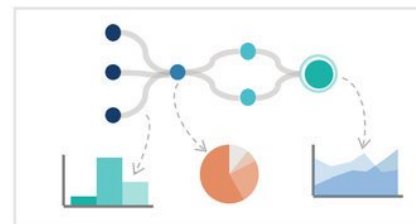
Determine which alternative interaction or sequence of interactions performs better

② Conversion Rate Optimization



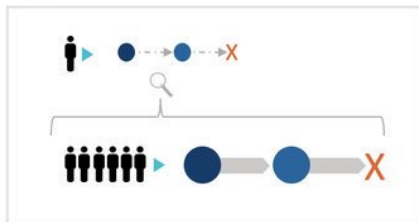
Increase the rate at which customers progress at each step or along a series of pre-defined steps

③ Impact Analysis



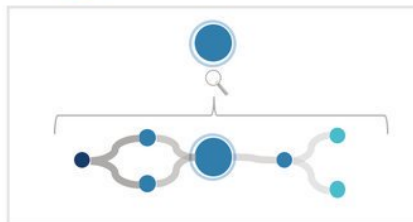
Quantify the effectiveness of interactions at a single step or towards achieving a goal or KPI

④ Behavioral Segmentation



Discover meaningful groups of customers defined by a common path and attributes

⑤ Journey Discovery



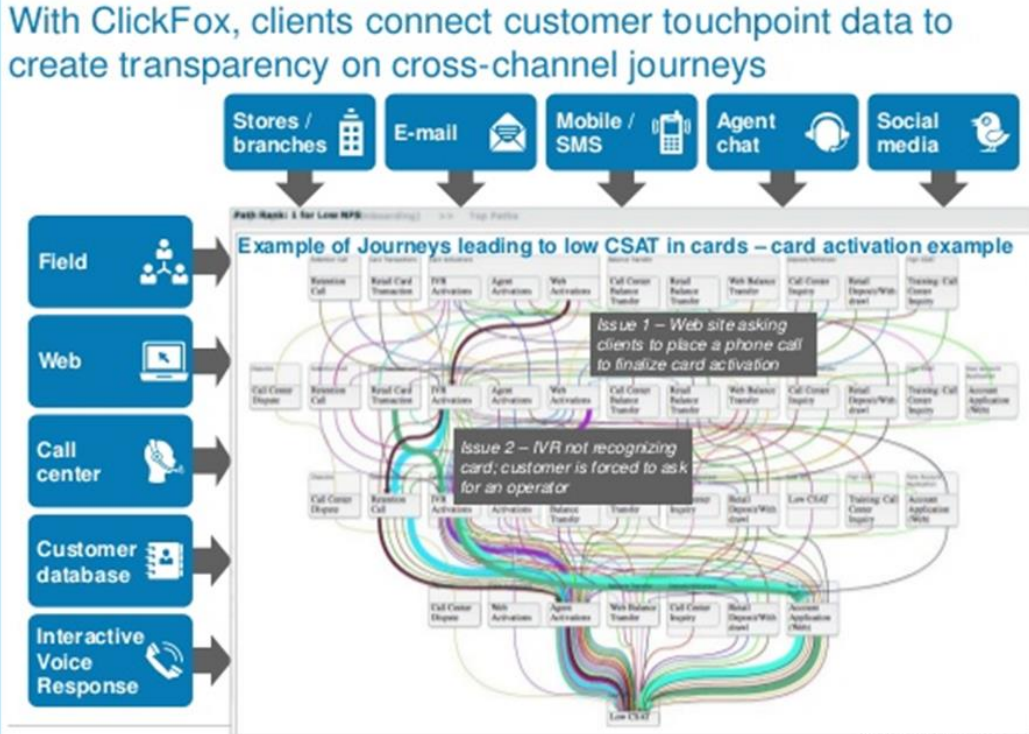
Uncover customer interactions along meaningful paths across touchpoints and over time

⑥ Goal Optimization



Use AI to reveal the customers and interactions most/least likely to impact a business goal

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



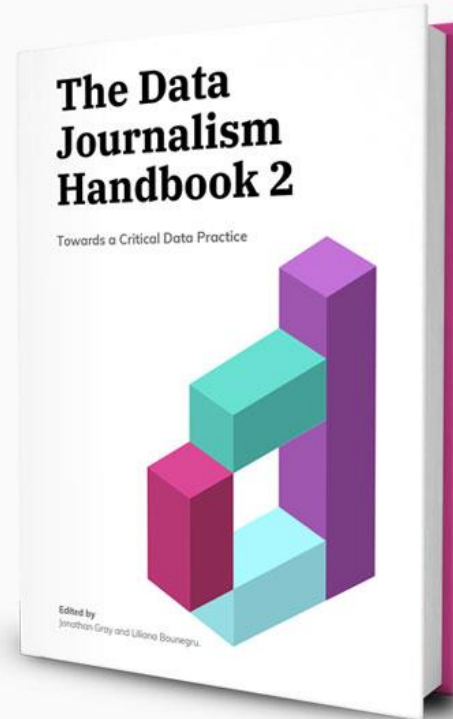
★ **Data Journalism Handbook (2nd edition) –
great resource for **Data Literacy** and **Data Storytelling!****
<https://datajournalismhandbook.org/>

The Data Journalism Handbook 2

Towards a Critical Data Practice

BETA NOW AVAILABLE

Produced by

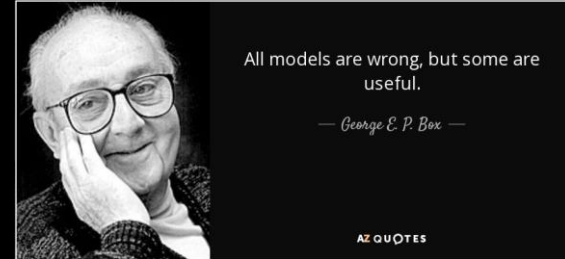


Final Reminders

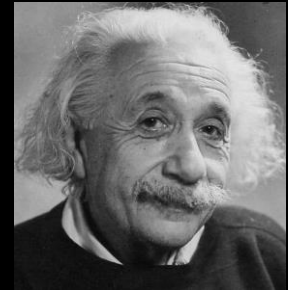
As data scientists, we must not only *Walk The Talk*, but we must also must *Talk The Walk*.

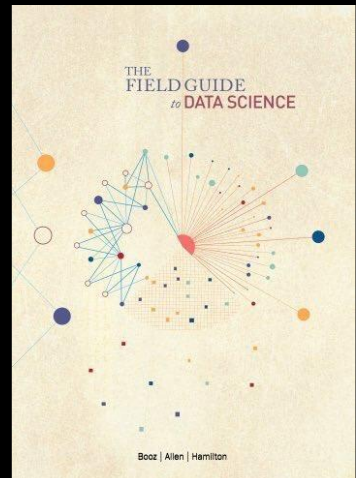
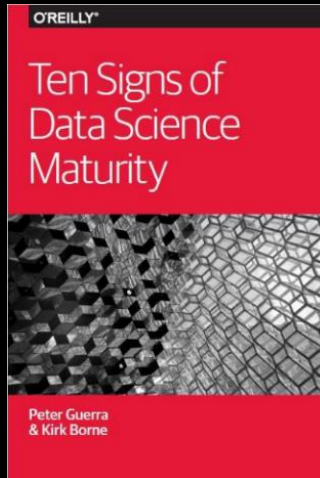
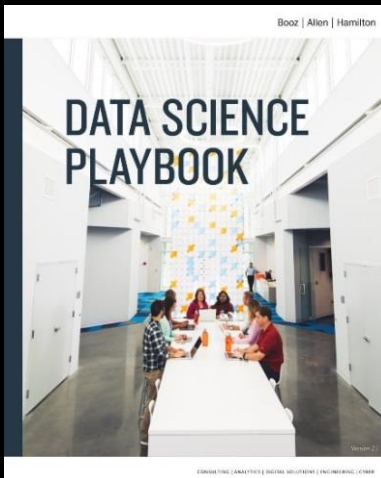
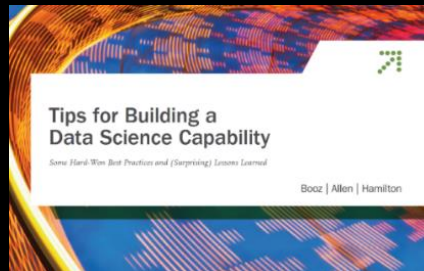
2 Reminders for Data Literacy and for Data Storytelling

- ***“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