



"Portrait of Madame Metzinger" - Jean Metzinger (1911)

Beer, diapers and correlation: A tale of ambiguity

Accelerate Boston, April 2017

Mark Madsen

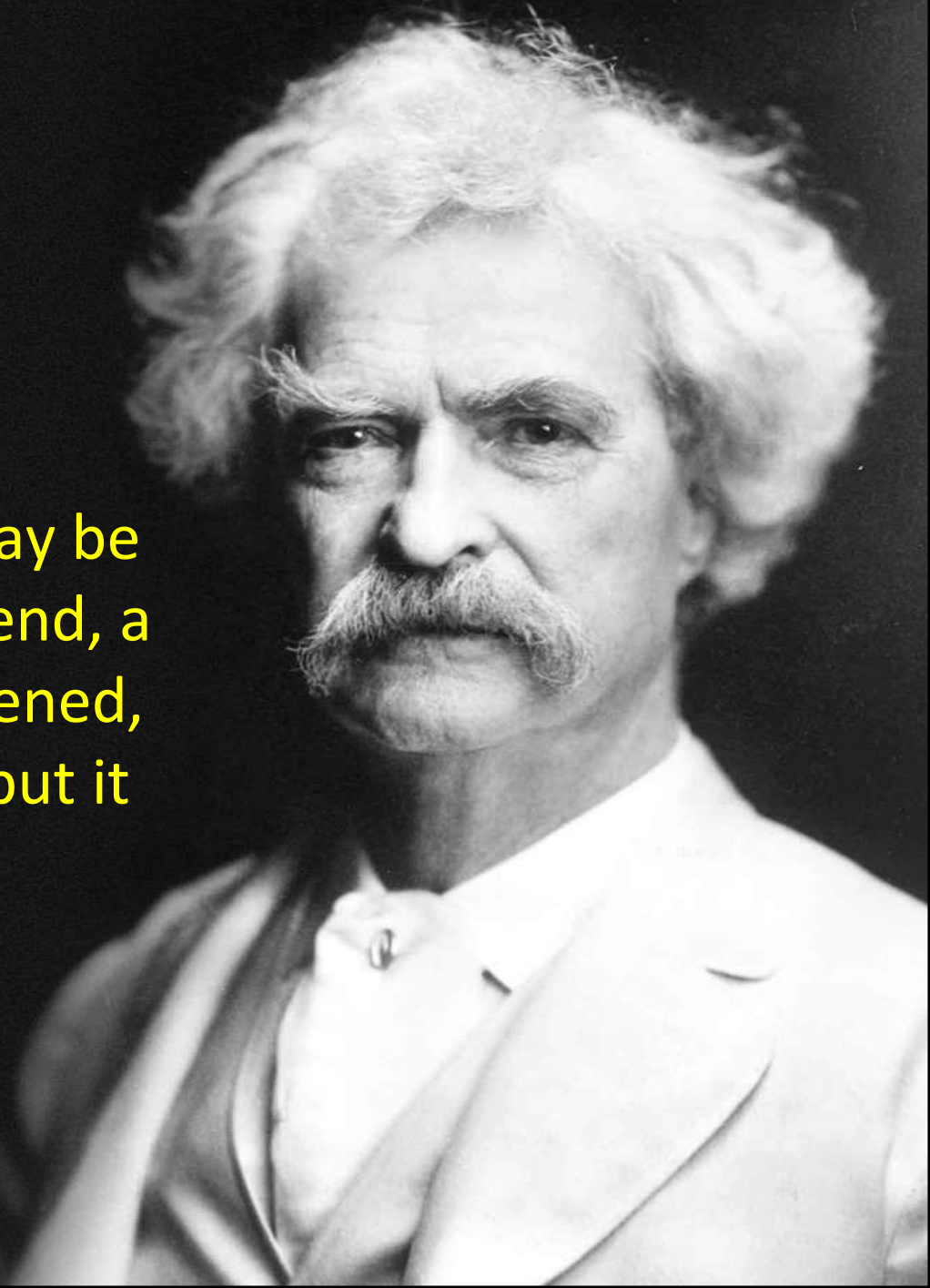
www.ThirdNature.net

@markmadsen



"I will set down a tale... it may be history, it may be only a legend, a tradition. It may have happened, it may not have happened: but it *could* have happened."

– Mark Twain



Analytics: half art, half science



"Portrait of Dora Maar" Pablo Picasso (1937)

Cubism paints a picture in a way that is similar to how we see and remember a subject.

Although cubism seems abstract, it tries to capture the truth – it may be a more realistic way to paint a portrait.



"Portrait de Ambroise Vollard" Pablo Picasso (1910)

Like art at the turn of the last century, analytics, particularly how people think they should be used, can get stuck in absolutist thinking.

Models abstract reality by capturing dominant features, the essence of a situation.

Like cubist paintings.

The beer and diapers story, as told today



Global



Why should a business do this sort of analysis?

Increase co-purchases of products bought together, e.g. office desk supplies and the forgotten item

Cross-promotion using directional relationships, e.g. chips and dip

Price optimization, particularly during promotions

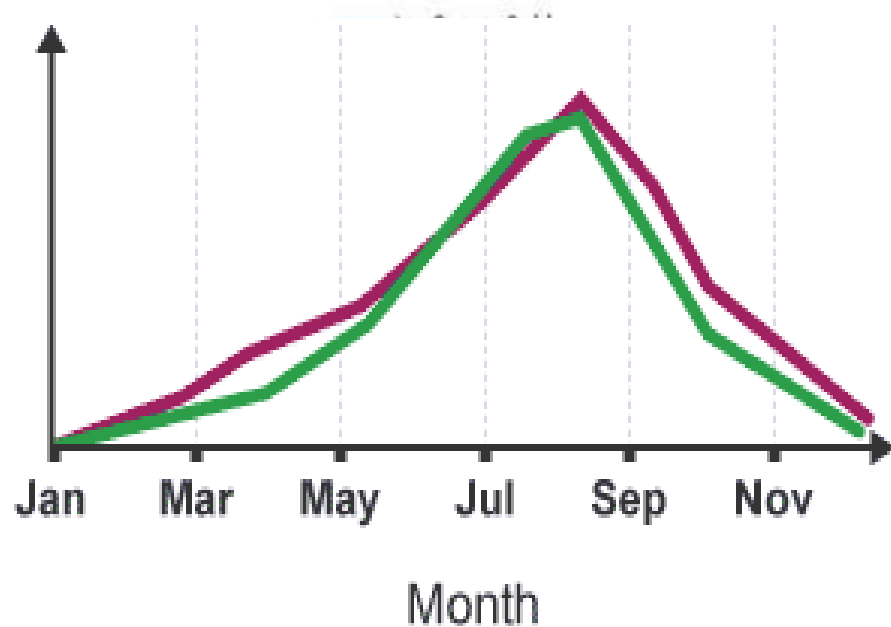
Inventory management, e.g. stocking the proper amount of the *dependent* product

Refine marketing, e.g. targeting segments based on their affinities



Correlation

cor·re·la·tion: a mutual relationship or connection between two or more things. *Statistics*: a quantity measuring the extent of interdependence between variable quantities.



 Ice cream sales  Shark attacks

How do you look for these patterns?

Many techniques you could use: Apriori , FP-Growth, Eclat, K-Apriori, SCOPE, etc.

For example, back in 1994:

- Chain stores of that size at that time stocked about 30,000 items
 - That's $30,000^2 = 900,000,000$ cells in a matrix
 - You need less than half of them since $A B = B A$,
 - So it's $30,000 * (29,999 / 2) = 449,985,000$
- Each cell is four sums, A, B, AB, 0 and this calc:

$$\varphi = \frac{ad - bc}{\sqrt{(a + b)(c + d)(a + c)(b + d)}}$$

I liked: Beyond Market Baskets Generalizing Association Rules to Correlations”, SIGMOD 97

I wonder...Is it true?

I read about the beer and diapers story in a Chain Store Age magazine in 1993-94. In 1994-ish I ran a test, in a grocery store operating in PA, OH, MD Specifically searched for this pattern.



What I discovered...

Original: 1992, **Correlation**

- **Stated cause: Men buying diapers**

My test: Grocery chain, 1994, **No correlation**

- **No causal relationship**

Is it true? No.

I'm glad that's settled.

Tried it again, just for fun: Is it true? Yes.

Original: Grocery, 1992,
Correlation

1. Grocery, ~1994, **No correlation**
2. Drug store, ~1995,
Correlation



I kept looking when I had the opportunity

Original: Grocery, 1992, Correlation

1. Grocery, ~1994, No correlation
2. Drug store, ~1995, Correlation
3. Drug store, 1997, Very high correlation
4. Drug store, 1997, No correlation
5. Grocery, 2000, Weak correlation
6. Multiple, 2013, Correlation

Is it true? Yes. No. Maybe. I don't know.

Is this for real?



OMG WTF

Where did this story come from? Here's a 1998 ad



Origin search: literature trails ends in ~1992

11,400 academic papers, 14,000 books, ~1,000,000 web pages

“The discount chain moved the beer and snacks such as peanuts and pretzels next to the disposable diapers and increased sales on peanuts and pretzels by more than 27%.”

“For example, one Midwest grocery chain used the data mining capacity of Oracle software to analyze local buying patterns. They discovered that when men bought diapers on Thursdays and Saturdays, they also tended to buy beer.”

“[UK] Dads with newborns cannot go out to pubs to socialize with friends (me including). So they buy more beers so that they can drink at home!”

“Some time ago, Wal-Mart decided to combine the data from its loyalty card system with that from its PoS...Once combined, the data was mined extensively ...On Friday afternoons, young American males who buy diapers (nappies) also have a predisposition to buy beer.”

“Shoppers who buy diapers for the first time at a Tesco store can expect to receive coupons by mail for baby wipes, toys -- and beer. Tesco's analysis showed that new fathers tend to buy more beer because they are home with the baby and can't go to the pub.”

“Sometimes the data can throw up surprises: mining of databases held by 7-Eleven stores in the US revealed a link between purchases of beer and nappies. When they were moved together, sales of both increased,”

Origin search: literature trails ends in ~1992

11,400 academic papers, 14,000 books, ~1,000,000 web pages

"The discount chain moved the beer and snacks such as peanuts and pretzels next to the disposable diapers and... sales on peanuts and pretzels by more than 27%."

"For example, one Michigan... used the data mining capacity of Oracle software to analyze local buying... that when men bought diapers on

Thursdays and Saturdays... (me including).

"[UK] Dads with new babies... So they buy more..."


"Some time ago, we... that from its PoS... Once... Friday afternoons, young American males who buy... position to buy beer."

"Shoppers who buy diapers for the first time... tend to receive coupons by mail for baby wipes, toys -- and beer. Tesco... showed that new fathers tend to buy more beer because they are home with the baby and can't go to the pub."

"Sometimes the data can throw up surprises: mining of databases held by 7-Eleven stores in the US revealed a link between purchases of beer and nappies. When they were moved together, sales of both increased,"

Lesson #1: in academic research on data science and machine learning, "PhD" really does mean "BS piled higher and deeper"

Doesn't the explanation seem trite?

A man with short dark hair, wearing a black t-shirt, is shown from the chest up, looking slightly to his right with a thoughtful expression. The background is a rural landscape featuring a vibrant rainbow arching across a cloudy sky. In the distance, there are green hills, a red cylindrical object, and a blue structure. A thought bubble is positioned above the man's head.

So you're
telling me...

Where did this story *really* come from?

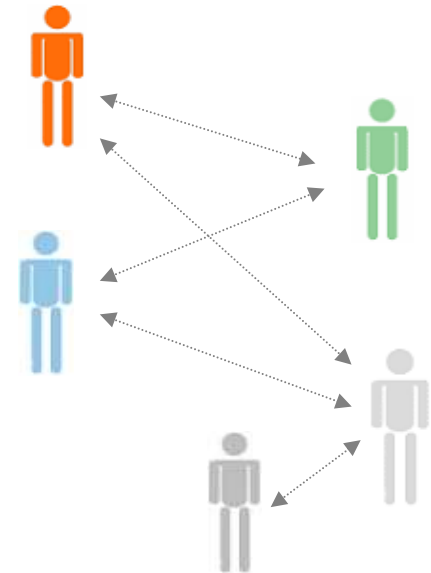
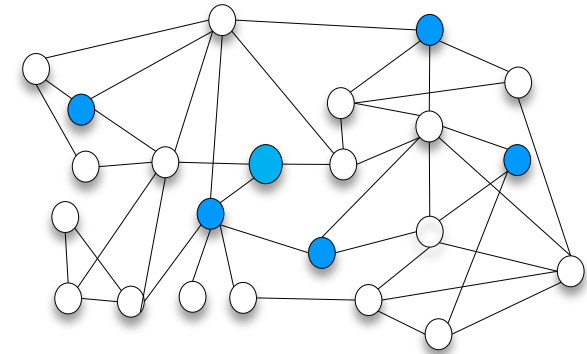
The trail is muddy but it's recent enough that people are still around. After searching, emailing and calling people, I found that...

It goes back to 1992, Osco:

“...90 days of point-of-sale data from Osco Drug stores - 1.2 million baskets...

...between 5pm and 7pm, customers tended to co-purchase beer and diapers.

...we have a correlation between beer, diapers and time...”



The source of the story: it's not (entirely) an urban myth



KAREN HEATH

**By the way, happy 25th
anniversary to beer & diapers**



KAREN HEATH

What did they find and how did they find it?

“...we looked for correlation with baby products because they were high margin...”

“We used SQL queries to find relationships.”

“...we have a correlation between beer, diapers and time, but no correlations with age, gender or day.”

“...does not appear to have exploited the information by moving the products around.”



The origin of the story

“Our 'fearless leader', Thom Blischok, when talking with prospects and the press, didn't distinguish between the actual affinities tested and our hypotheses. Our job was to sell the value of systems. Sometimes in selling, fact blurred with folklore.” – John Earle

Lesson #2: Don't let data get in the way of a good story.





Fragmentation,
ambiguity, what's
going on here?

There were mixed
answers though,
some correlations. Is
it real, or is it just a
story told by Tom?

Are you asking the
right question? Not
“Is it true” but “under
what circumstances is
it true?” and “Can I
use it?”

*Marianne Faithful's version of a
portrait from 'Destroy Rankin'*

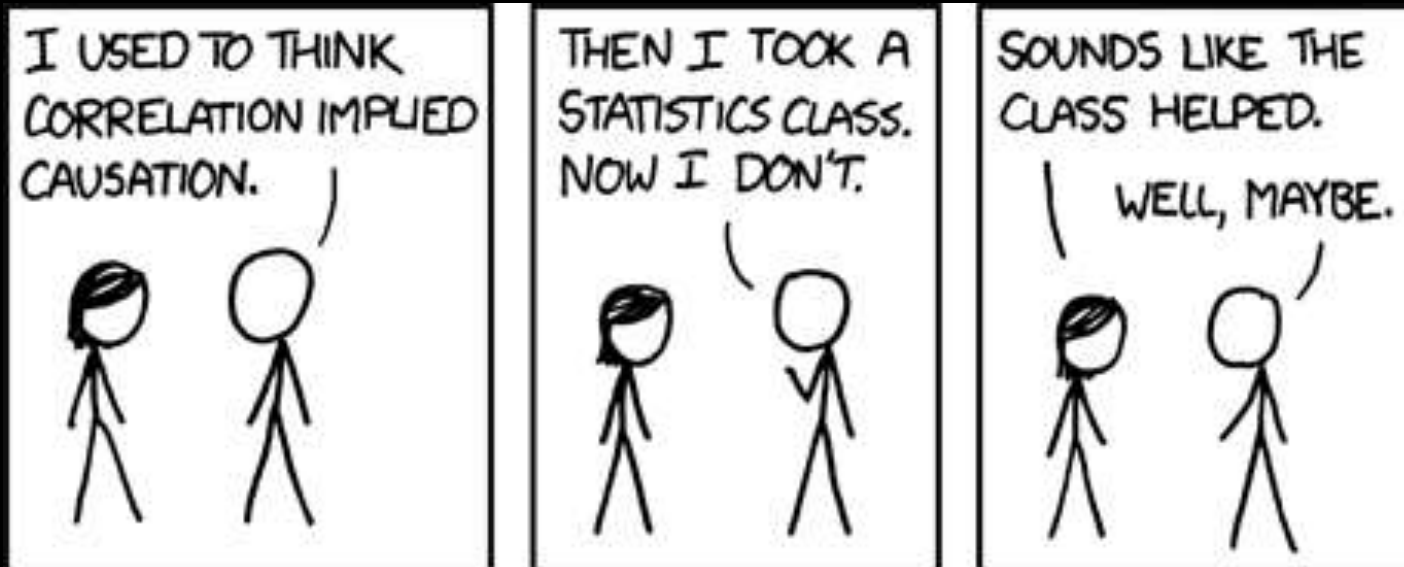
It's not the insight, but what you do with it, that matters
As a manager: what would you do in this situation?

Analytic insights that result in no action are expensive trivia.

Enterprise reality: causality > correlation

You wouldn't do anything without some idea of cause or context, correlation isn't enough .

sorry Economist, causation does matter, otherwise it will be ice cream & sharks



What is the explanation?

$$f(\text{REEFER, NO}) = \text{MONEY}$$

The image depicts a mathematical function f that takes two inputs: 'REEFER' (represented by three beer bottles) and 'NO' (represented by a cartoon of a child sitting in front of a sign that says 'NO'). The output of the function is 'MONEY' (represented by a stack of cash).

Forgetful husbands only remember the basics

“I send you to the store for a few necessities, and you come back with chips and beer instead of talcum powder and diapers.”

- *MacGyver*, 4/18/1988



Or harried dads rewarding themselves with impulse buys



Men can't go out with their friends any more



Maybe it's a nightcap so babies sleep well



Maybe it's mom, because in the late-1990s...



How Mother and Baby "Picked Up"

A case of Blatz Beer in your home means much to the young mother, and obviously baby participates in its benefits.

The malt in the beer supplies nourishing qualities that are essential at this time and the hops act as an appetizing, stimulating tonic.

Main 2400



BLATZ
MILWAUKEE

Always the same good old *Blatz*

Beer and Breastfeeding, Pubmed, <http://www.ncbi.nlm.nih.gov/pubmed/11065057>

Sometimes it's not "purchasing" exactly, but it still should still count as a correlation

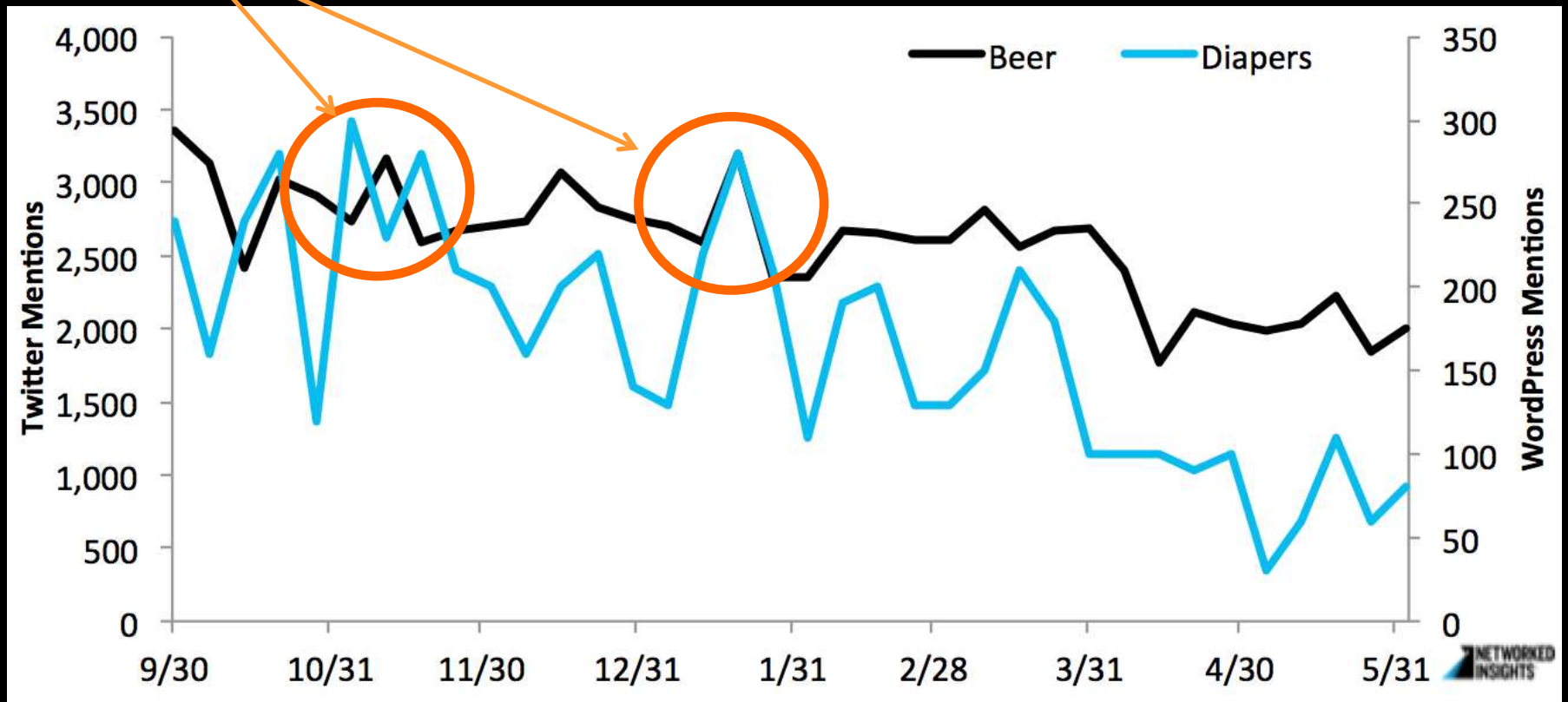


Nobody checked if it was diapers for *babies*...

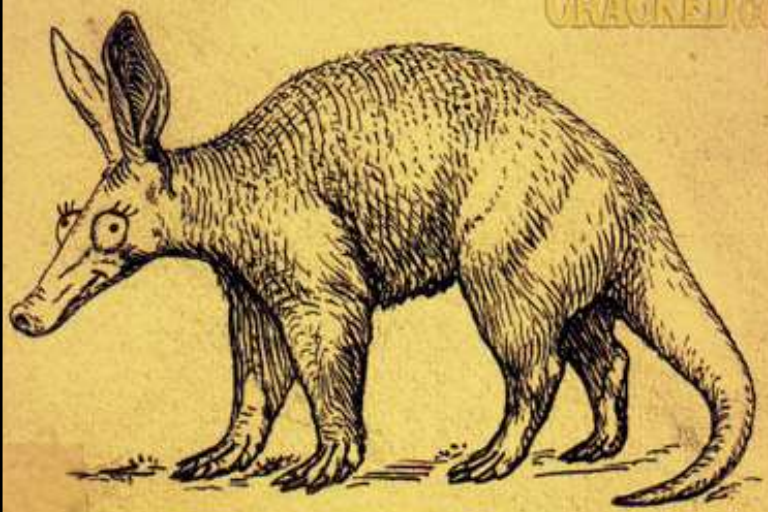


'Cause no presentation is complete without charts

Whazzat?



Men, shopping before Thanksgiving and end of January?
(in the USA)



Beausality: explanations that sound reasonable

A-2

AARDVARK Stupidus Faceus

The aardvark, aka, "God's Mistake," is the confused lovechild of a rabbit and pig. Aardvarks are insectivores, preferring to hunt by inviting ants to timeshare meetings, which are an elaborate ruse. Aaaaardvarks mate by passive aggressively alluding to how lonely they are. When encountering an aggressive aaaaaardvark in the wild, it is adviseable to loudly hum the "Cheers" theme song, which will frighten the aaaaaaaaaardvark away. Aaaaaaaaaardvarks, far from being endangered, are in the "not nearly threatened enough" category. Aaaaaaaaaardvarks are notorious for tampering with mail-slots, and are therefore not welcome in most USPS locations. Aaaaaaaaaardvarks hate mornings. Aaaaaaaaaaaaaaaaaardvarks smell like peanutbutter. Aaaaaaaaaaaaaaaaa

Beneficial to Young and Old

"GESUNDHEIT, GRANDPA"

Cultivate the **RAINIER BEER** habit

It brings the glow of health and gives a new lease on life... No medicine can equal it as a **TONIC**

SEATTLE BREWING & MALTING CO. Seattle, Wash.

THEY'RE HAPPY
Because they eat

LARD

Issued by the Lard Information Council

Do you want to know what the answers are?

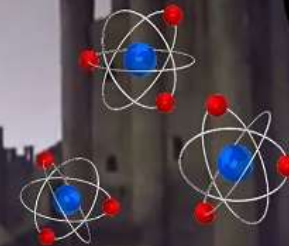
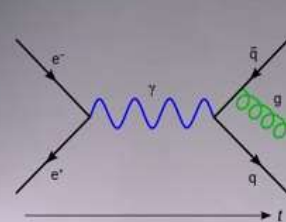
Who believes that...

Good news! You are all right. Bad news! You are all wrong.

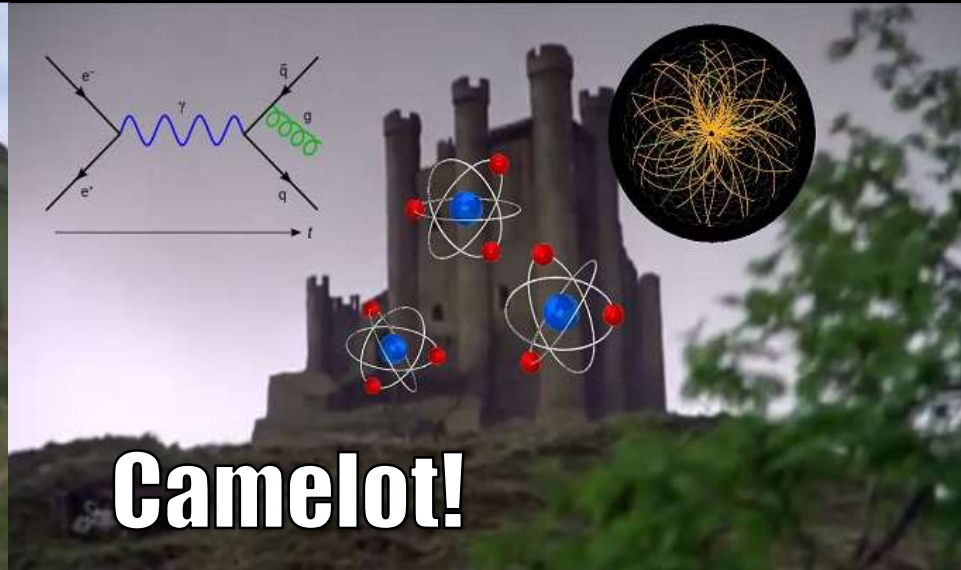
1. Grocery, 1992, Correlation – both, TH/SA, 5-7 PM, midwest
2. Grocery, ~1994, No correlation, mid-northeast
3. Drug store, ~1995, Correlation – male, weekends, east
4. Drug store, 1997, Very high correlation, both, always, US
5. Drug store, 1997, No correlation, US
6. Grocery, 2000, Weak correlation, female, weekdays, west/mid
7. Multiple, 2013, Correlation, male, seasonal dates, US
8. Grocery, 2006, Correlation, male, UK, (not 1st hand)

*The story is true, and false, and indeterminate.
It's ambiguous. So much for the absolute answer.*

LOOK, MY LIEGE!



Camelot!



IT'S ONLY A MODEL



SHHH!



Image credit: unknown

**Are you asking the right question?
Not "is it true" but "under what circumstances is it true?"**



Le Goûter / Tea Time / Femme à la Cuillère (1911)

“It is termed analytical cubism because of its *structured dissection of the subject, viewpoint-by-viewpoint, resulting in a fragmentary image of multiple viewpoints* and overlapping planes. Other distinguishing features of analytical cubism were a *simplified palette of colours*, so the viewer was not distracted from the structure of the form, and the density of the image at the centre of the canvas.”



What is happening here is that “the model” is seeing pieces of a whole because each of these models is independent, with its own observations and data, like a snapshot that covers only the focal point of your eye.



You are actually seeing the pieces individually, with little context to link them,



...not as independent, in their appropriate places within a larger context.





Joiner, David Hockney (1980s)

When they are lined up together the larger scene emerges, in much the same way that your brain puts together a scene by focusing on individual pieces and assembling the picture from them as your focus moves from place to place.



Except that the pieces you see in these models are not nicely bounded. They are assembled piecemeal, in overlapping bits, some including one element some another. Just like David Hockney's "joiners" from the 80s. This is a problem any time you generalize from someone else's model.

1970s - 1980s - 1990s - 2000s - 2010s - 2020s - 2030s - 2040s - 2050s - 2060s - 2070s - 2080s - 2090s - 2100s

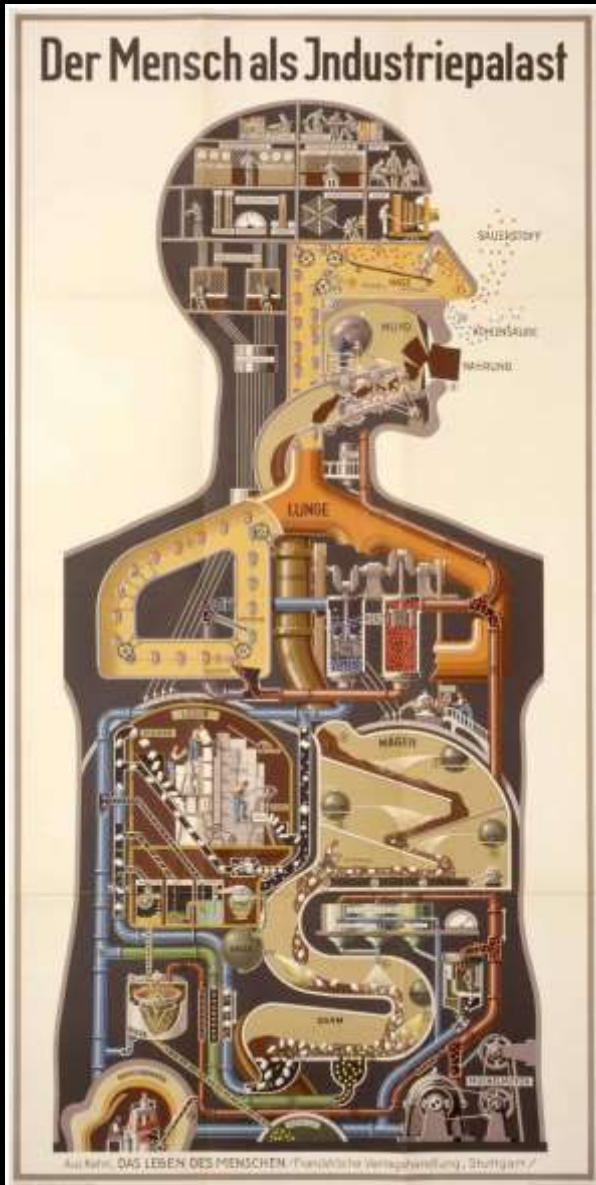
Joiner, David Hockney (1980s)

Therefore...

You would choose to merchandise, price or promote in accordance with the causality *you believe* is true.

The actions you take and the related outcomes will vary based on *your local context*.

Not one way to act but many, just as there are many attributes, not one.



There is no “single version of the truth”



Portrait #1 (self), Daniel Crooks (2007)

The artist is part of the art they create. They choose what goes into the art.

just like

The analyst is part of the model they create. They engineer the features.

Therefore

All models are relative.

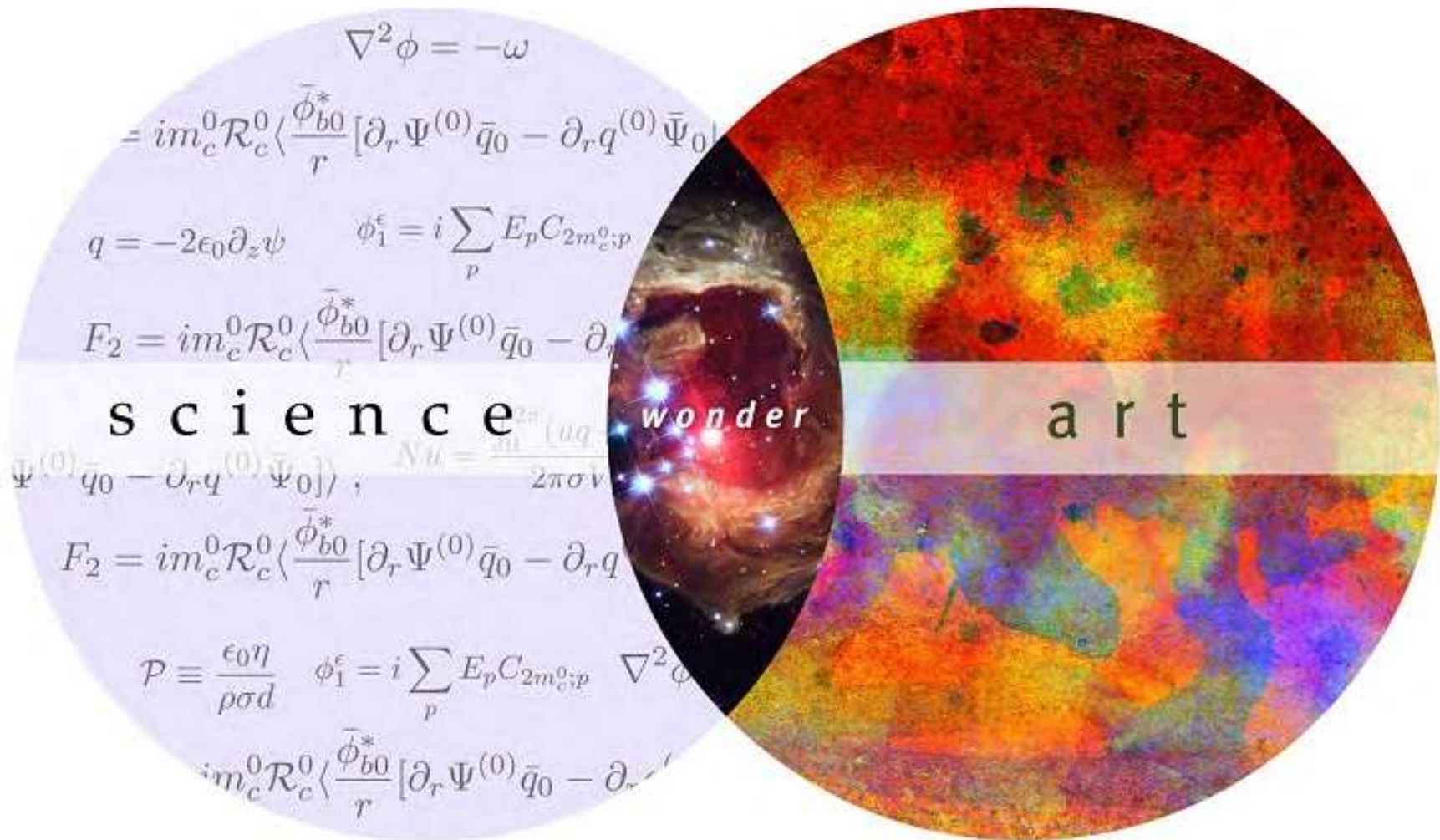
The data is inherent in the model created, just hidden from view.

People are the link that gives context. Without a person determining the applicability of analytical models or interpreting the results, little sense can be made of them. You need the entire scene.



'Pearblossom Highway, 11th to 18th April 1986 No.2', David Hockney

“Every science begins as philosophy and ends as art.” -- Will Durant





The power of myth

The origin story is no less true for having not happened. A myth:

- Reveals some mystery
- Shows the complexity of the world and explains what you encounter
- Validates a sociological system
- Shows that there's a role, a place for us

Lose Santa, gain a holiday, just like the Grinch.

(I still use this story 😊)

The final lesson learned at Osco in the 1990s:

If you put two products next to each other on a shelf then they are more likely to be sold together.

Thanks for coming to some deep learning!



Reflections

1. There is no “single version of the truth”, the word “version” is a clue. Truth is relative and contextual, not absolute. So don’t promote machine-learning as black and white answer-boxes.
2. “All models are wrong, some are useful” – George Box
3. Like a cubist, think of all the angles when you use analytics. “*Do what they did*” management strategy doesn’t work when it comes to contextual problems.
4. You need to have an idea about causation. Ignore the pundits.
5. Data science is one half of the solution, applying the results is the other half. “Science” because you go into the unknown when you act. Like Heisenberg , your actions cause effects, and the choice of action may preclude another action.
6. Focusing on the science misses the art.

Some reference material

Predictive Analytics and Data Mining. 2014 (chapter 6, association rules)

SCOPE: An Efficient One Pass Approach to find strongly Correlated Item Pairs

<http://ieeexplore.ieee.org/document/4731311/>

A Note on “Beyond Market Baskets Generalizing Association Rules to Correlations”

http://ftp10.us.freebsd.org/users/azhang/disc/disc01/cd1/out/websites/kdd_explorations_full/ahmed.pdf

Discovering temporal association rules: Algorithms, language and systems,

<http://www.computer.org/csdl/proceedings/icde/2000/0506/00/05060306.pdf>

Low Cost High Performance Uncertainty Quantification

<https://pdfs.semanticscholar.org/20ee/7a52a4a75762ddcb784b77286b4261e53723.pdf>

The delights of seeing: Cubism, Joiners and The Multiple Viewpoint,

<http://thedelightsofseeing.blogspot.com/2011/03/cubism-joiners-and-multiple-viewpoint.html>

Beer and Breastfeeding, Pubmed, <http://www.ncbi.nlm.nih.gov/pubmed/11065057>

Frames, Biases, and Rational Decision-Making in the Human Brain,

http://econ.as.nyu.edu/docs/IO/9877/De_Martino1.pdf



About the Presenter

Mark Madsen is president of Third Nature, a consulting and advisory firm focused on analytics, strategy and data management. Mark is an award-winning author, architect and CTO who has received awards for his work from the American Productivity & Quality Center, Smithsonian Institute and industry associations. He is an international speaker, a contributor to Forbes, co-chair of the Accelerate data science conference, and member of the O'Reilly Strata program committee. For more information or to contact Mark, follow @markmadsen on Twitter or visit <http://ThirdNature.net>



About Third Nature



Third Nature is a research and advisory firm focused on emerging technology and practices in analytics, information strategy, business intelligence and data management.

Our goal is to help organizations solve problems using data. We offer education, advisory and research services to support business and IT organizations. We also provide product-related consulting to software vendors in the industry.

We fill the gap between what the industry analyst firms cover and what IT needs. We specialize in strategy and architecture, so we look at the complex needs of organizations and evaluate how different technologies can be applied to solve problems.