Machine Learning safety reminders

Data Science PT #7

Rui Quintino
Data Research, DevScope
rui.quintino@devscope.net

Microsoft Partner
Gold Data Analytics
Gold Collaboration and Content
Gold Application Development
Gold Application Integration
Silver Customer Relationship Management
Silver Application Lifecycle Management
Silver Intelligent Systems
Silver Hosting

devscope

Where are we?

/datascienceportugal ~ 320

/groups/datascienceportugal ~ 278

/datascienceportugal ~ 110

/groups/8586496 ~ 134

/DataSciencePortugal

@datascience_pt

Join Us!

And a special thanks to...

http://shelf.ai/

https://www.uniplaces.com/

https://devscope.net/

About me

Data R&D @ DevScope

#PowerBI #SQLServer #Web

#Analytics #Azure #Microsoft

#MachineLearning #R #Linux

#Bots #Hadoop #Docker

#Python #Coaching #Learning

twitter.com/rquintino

rquintino.wordpress.com

rui.quintino@devscope.net

"jack of all trades (and master of none)"

1. a person who can do many different types of work but who is not (necessarily...) very competent at any of them...


DevScope

consulting

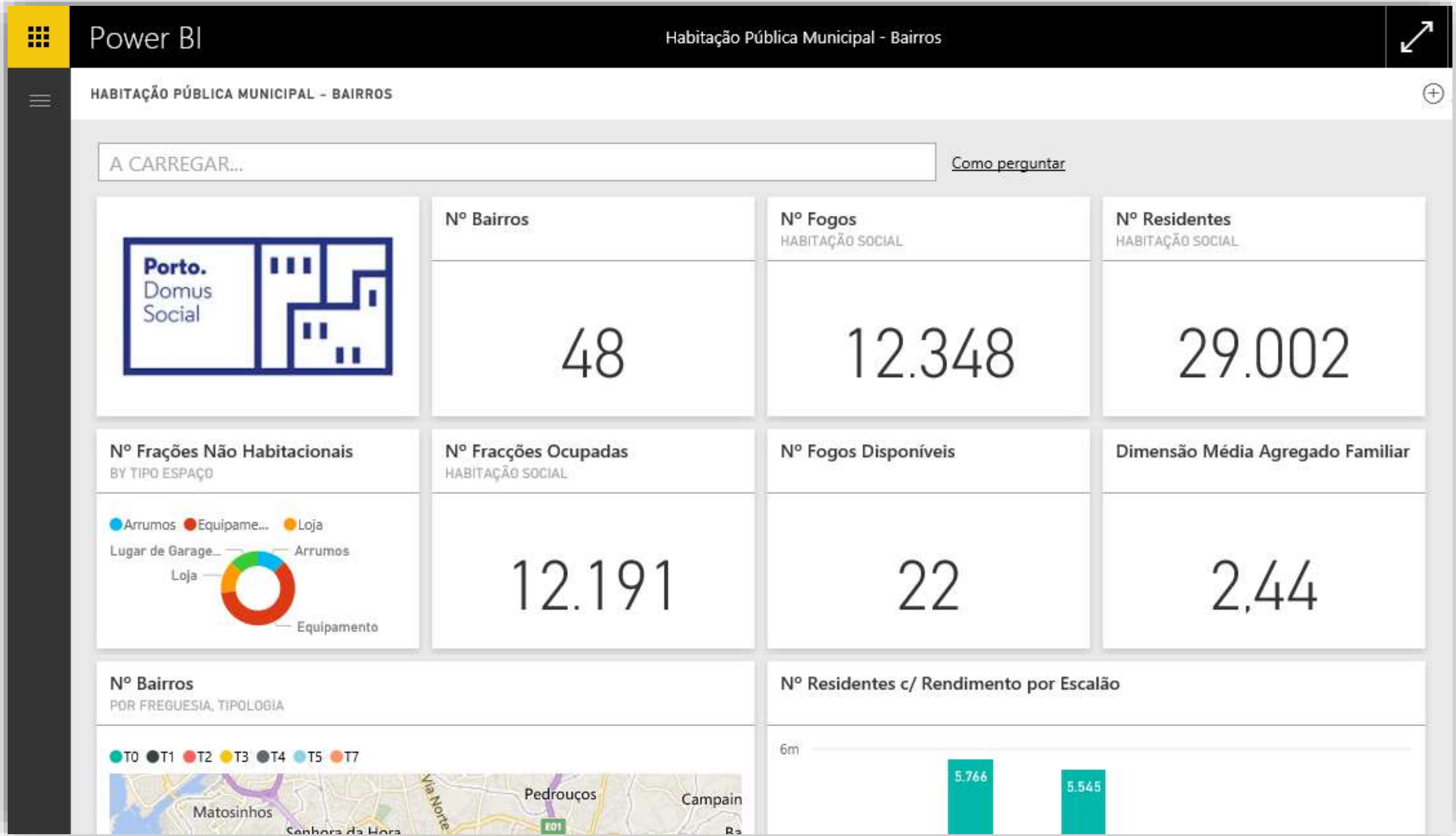
software

saas

devscope

http://travelbi.turismodeportugal.pt/

https://www.sns.gov.pt/profissional/recursos-humanos-da-saude/informacao-rh/analise-mensal-do-balanco-social/



http://www.domussocial.pt/domussocial/caracterizacao-sociodemografica

C.H. São João- Reagir a tempo

Resultados

Financeiros positivos, Topo dos rankings nacionais
MSHUG Innovation Awards 2014
IT Europa's BigData, BI & Analytics Solution of the Year 2014
Outstanding ICT Innovation Achievement HIMSS Europe 2016

Mais informação:

https://devscope.wordpress.com/2016/12/06/hvital-awarded-at-himss-europe-2016/

https://devscope.wordpress.com/2014/04/02/iteuropas-best-big-data-business-intelligence-and-analytics-solution-of-the-year/

https://devscope.wordpress.com/2014/02/24/hsjoao-devscope-winners-in-the-microsoft-health-users-group-innovation-awards-2014/

Centro Hospitalar de São João

www.hvital.com

TVI24 Eleições

https://news.microsoft.com/pt-pt/2015/09/29/legislativas-2015-com-informacao-em-tempo-real-numa-app-second-screen-criada-para-a-tvi24/

Media Capital - TVI

smartdocumentor

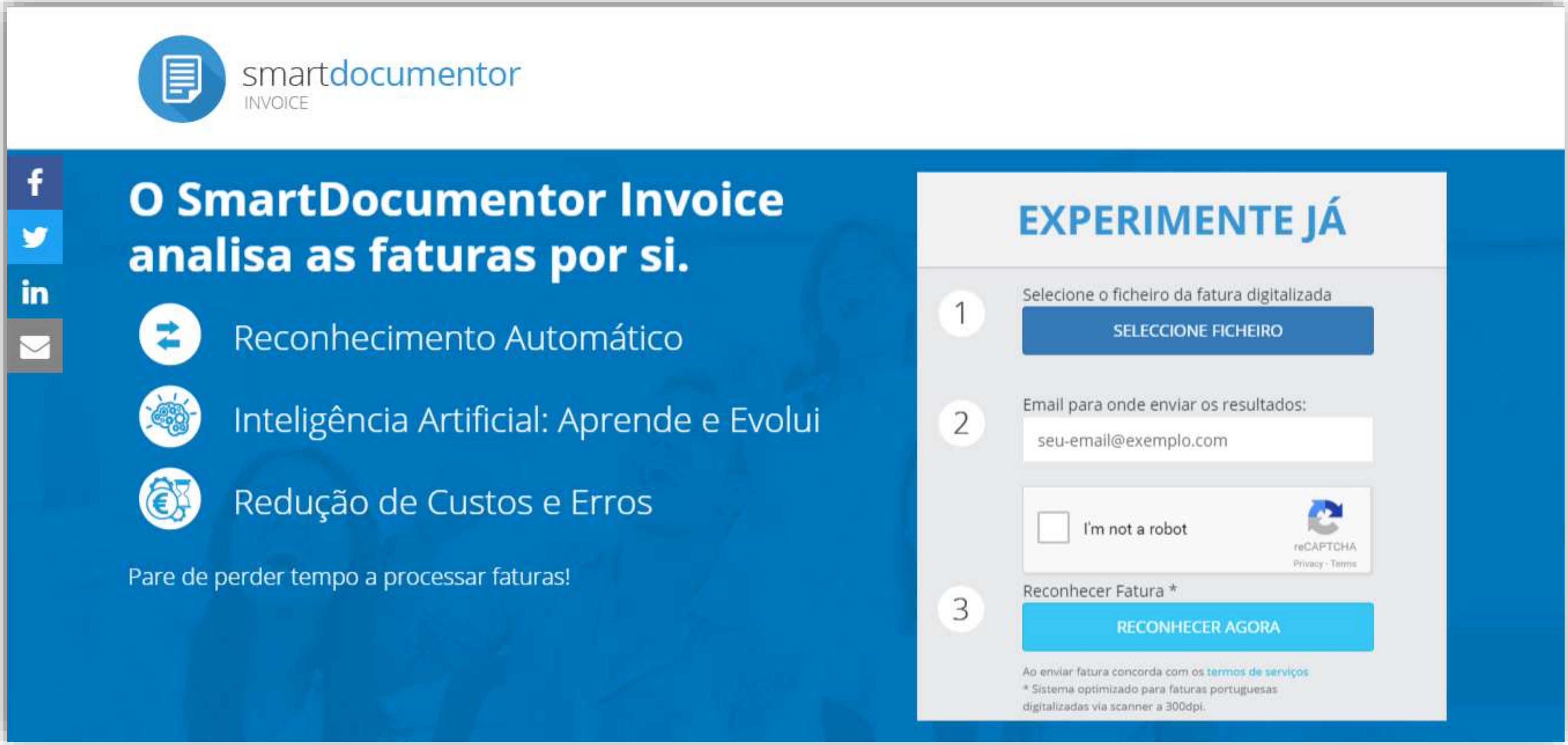
Power BI Scorecards

software

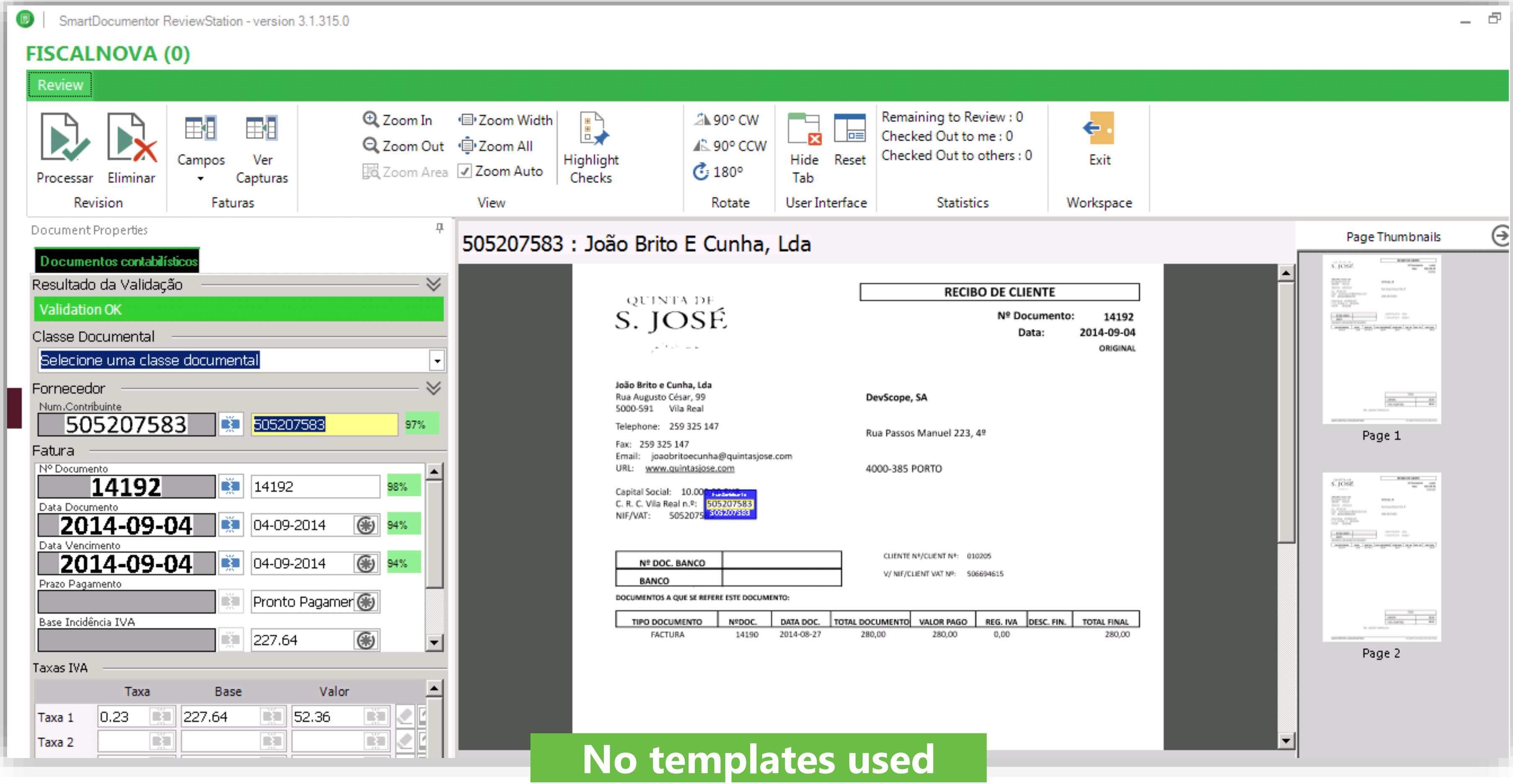
Power BI Tiles

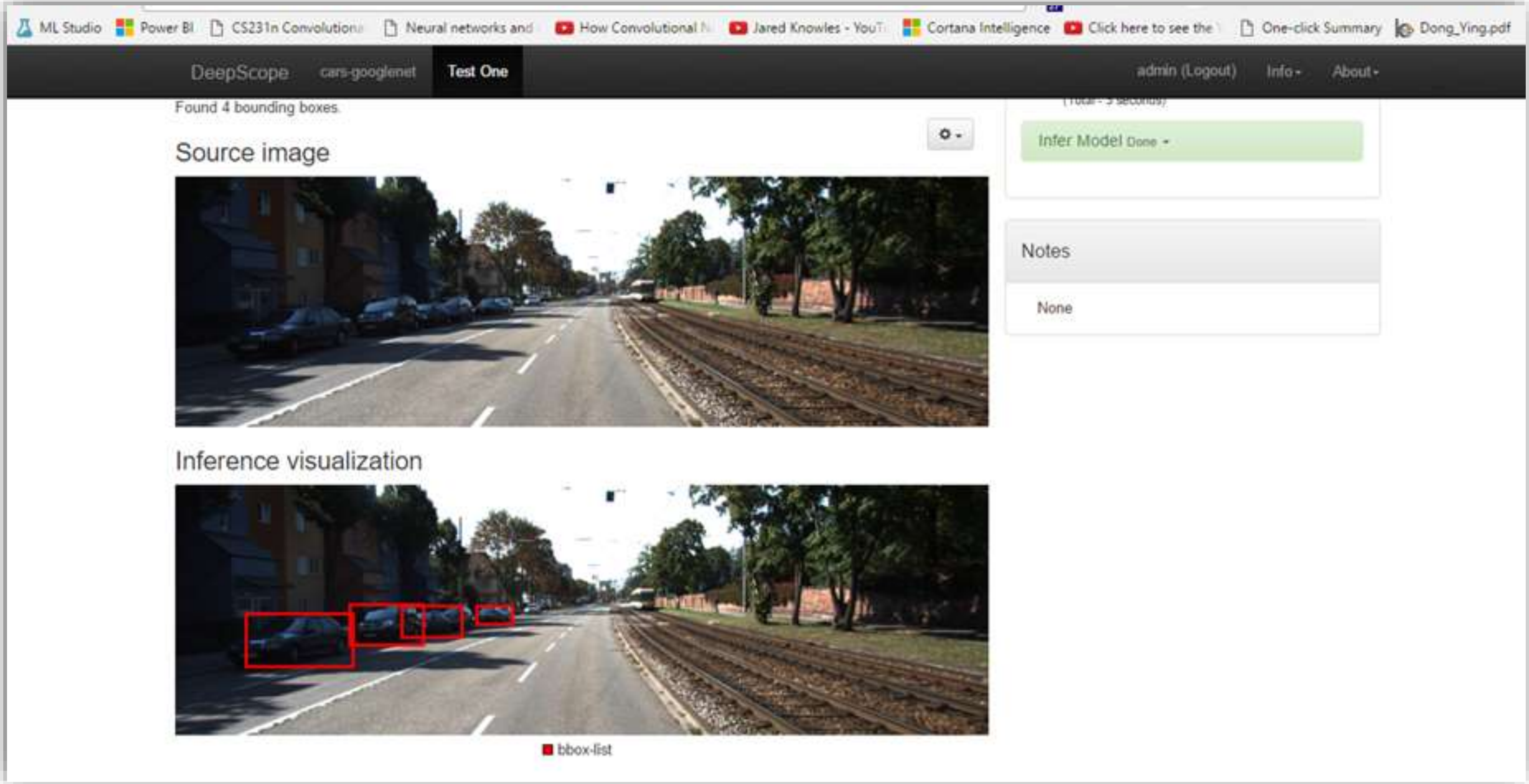
IMOMARKETING

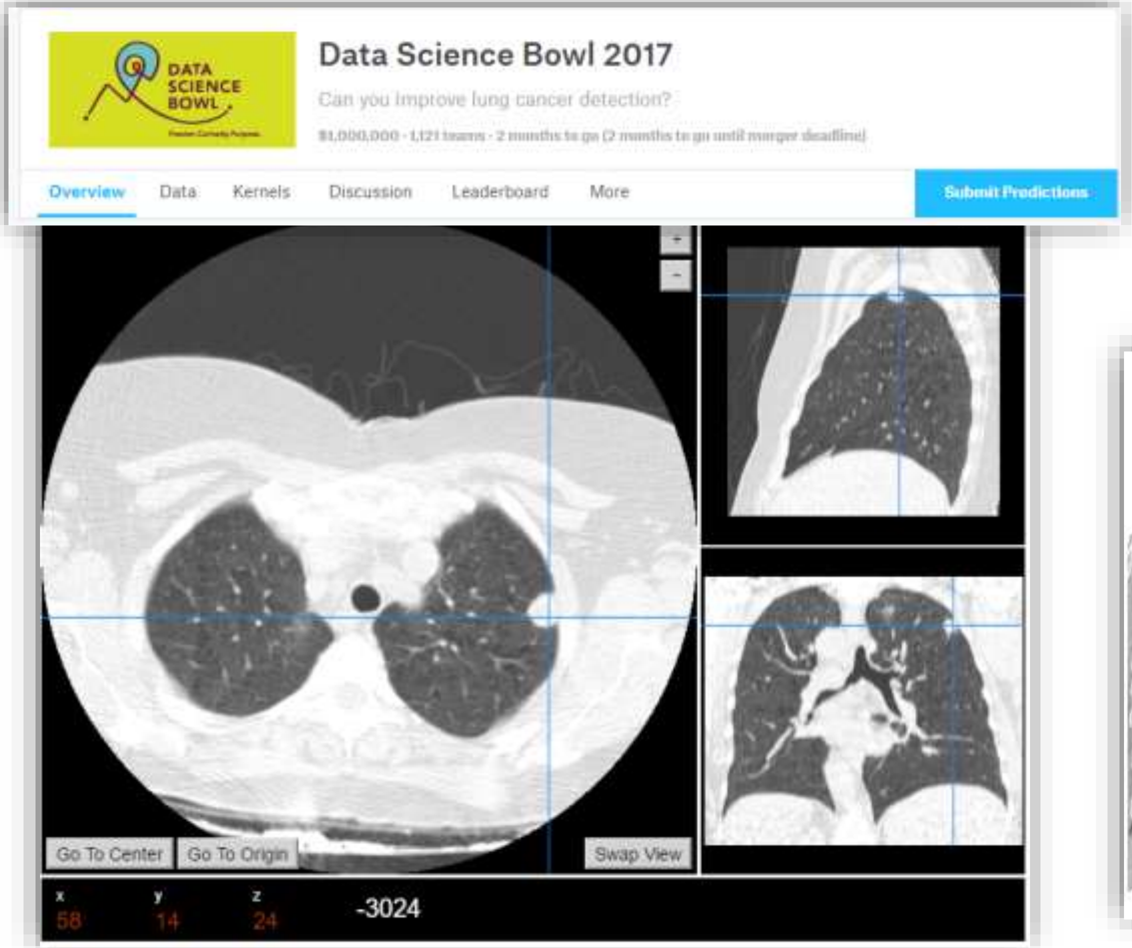
CRMARKETING

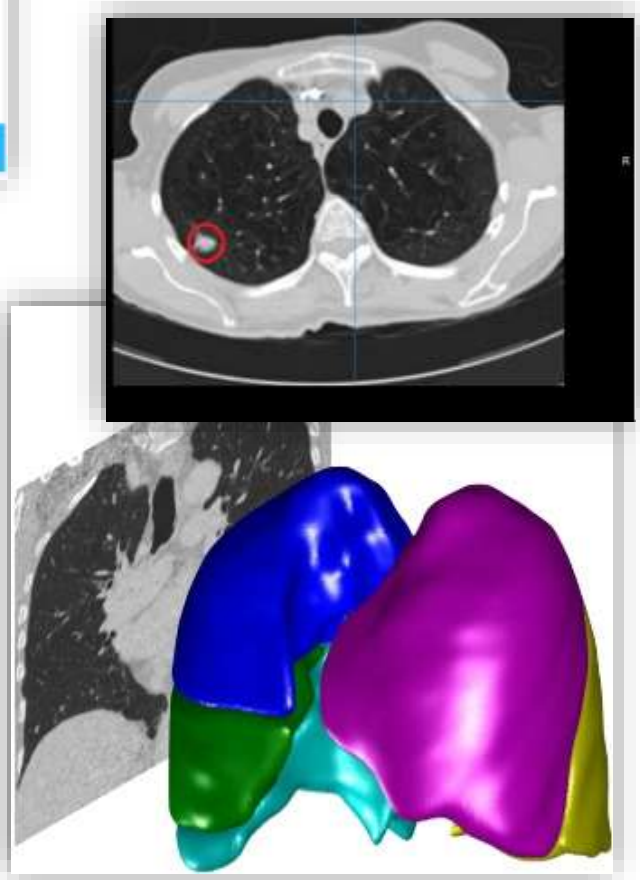


http://invoice.smartdocumentor.net

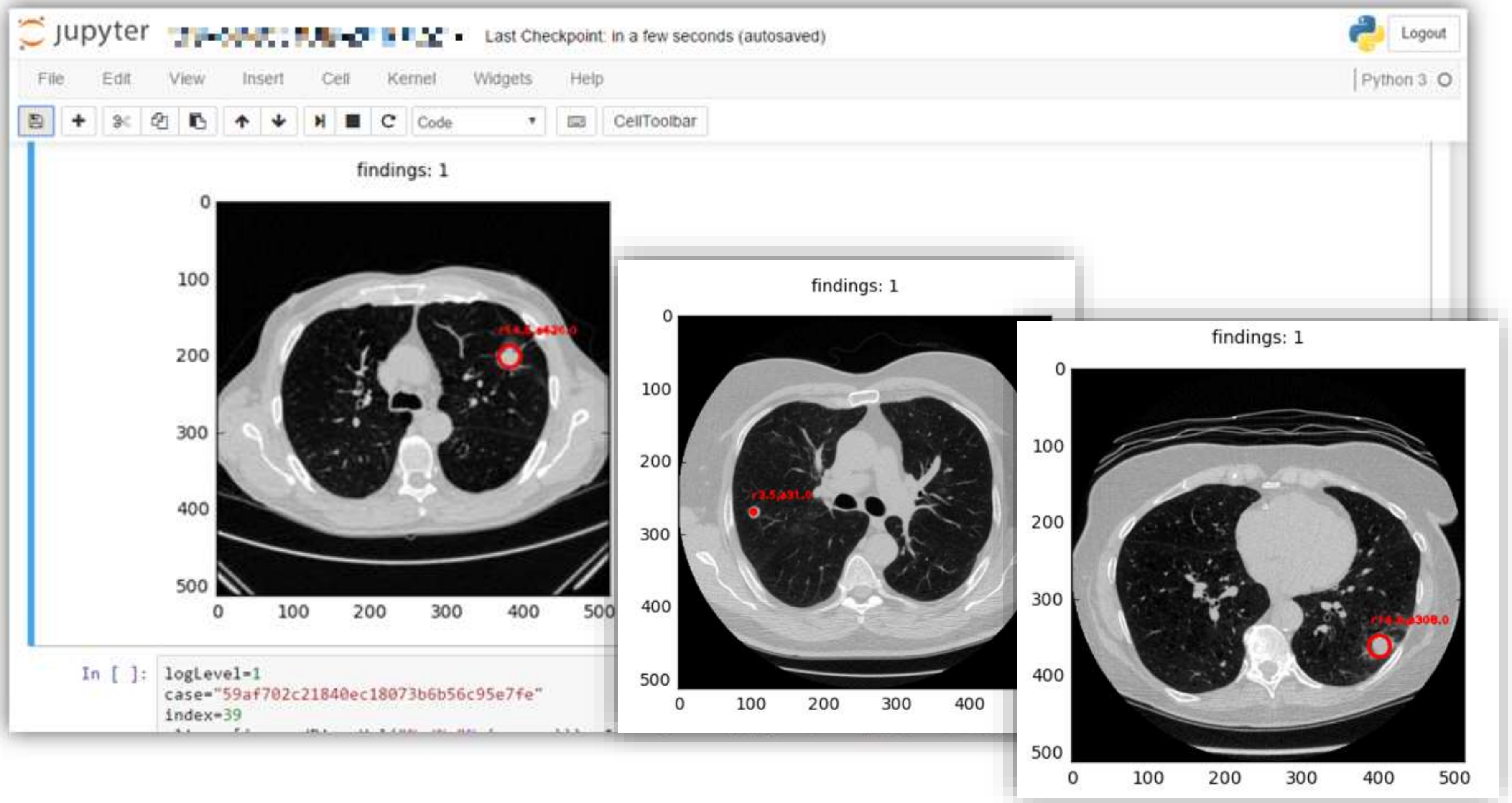

SmartDocumentor ReviewStation - version 3.1.315.0
FISCALNOVA (0)
Review
Processar
Eliminar
Revision
Campos
Ver Capturas
Faturas
Zoom In
Zoom Out
Zoom Area
Zoom Width
Zoom All
Zoom Auto
View
Highlight Checks
90º CW
90º CCW
180º
Rotate
Hide Tab
Reset
User Interface
Remaining to Review : 0
Checked Out to me : 0
Checked Out to others : 0
Statistics
Exit
Workspace
Document Properties
Documentos contabilísticos
Resultado da Validação
Validation OK
Classe Documental
Selecione uma classe documental
Fornecedor
Num.Contribuinte
505207583
505207583
97%
Fatura
Nº Documento
14192
14192
98%
Data Documento
2014-09-04
04-09-2014
94%
Data Vencimento
2014-09-04
04-09-2014
94%
Prazo Pagamento
Pronto Pagamer
Base Incidência IVA
227.64
Taxas IVA
Taxa
Base
Valor
Taxa 1
0.23
227.64
52.36
Taxa 2
505207583 : João Brito E Cunha, Lda
QUINTA DE S. JOSÉ
RECIBO DE CLIENTE
Nº Documento: 14192
Data: 2014-09-04
ORIGINAL
João Brito e Cunha, Lda
Rua Augusto César, 99
5000-591 Vila Real
Telephone: 259 325 147
Fax: 259 325 147
Email: joaobritoecunha@quintasjose.com
URL: www.quintasjose.com
DevScope, SA
Rua Passos Manuel 223, 4º
4000-385 PORTO
Nº DOC. BANCO
BANCO
CLIENTE Nº/CLIENT Nº: 010205
V/ NIF/CLIENT VAT Nº: 506694615
DOCUMENTOS A QUE SE REFERE ESTE DOCUMENTO:
TIPO DOCUMENTO
NºDOC.
DATA DOC.
TOTAL DOCUMENTO
VALOR PAGO
REG. IVA
DESC. FIN.
TOTAL FINAL
FACTURA
14190
2014-08-27
280,00
280,00
0,00
280,00
Page Thumbnails
Page 1
Page 2
No templates used


ML Studio
Power BI
CS231n Convolution
Neural networks and
How Convolutional N
Jared Knowles - YouT
Cortana Intelligence
Click here to see the
One-click Summary
Dong_Ying.pdf
DeepScope
cars-googlenet
Test One
admin (Logout)
Info
About
Found 4 bounding boxes.
Source image
Infer Model Done
Notes
None
Inference visualization
bbox-list





https://www.kaggle.com/c/data-science-bowl-2017



jupyter
Last Checkpoint: in a few seconds (autosaved)
Logout
File
Edit
View
Insert
Cell
Kernel
Widgets
Help
Python 3
Code
CellToolbar
findings: 1
findings: 1
findings: 1
In [ ]:
logLevel=1
case="59af702c21840ec18073b6b56c95e7fe"
index=39

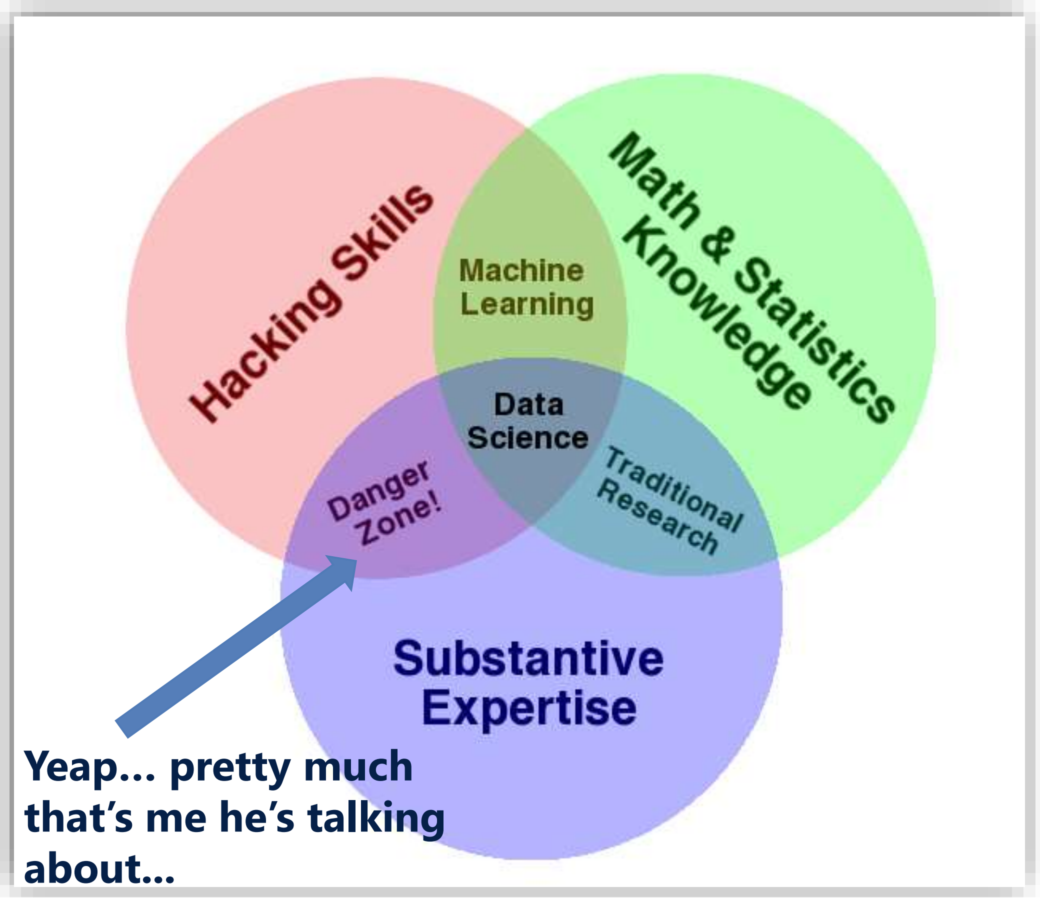
Machine Learning & Data Science Gotchas/pitfalls

#Business Intelligence #Analytics #DataViz #Dashboards #Reports …

- Easy ROI
- Mostly Observational Data
- Difficult but doable
- What is?
- Commodity these days
- Like our 5 senses (awareness, cognition)
- We can hardly live without it

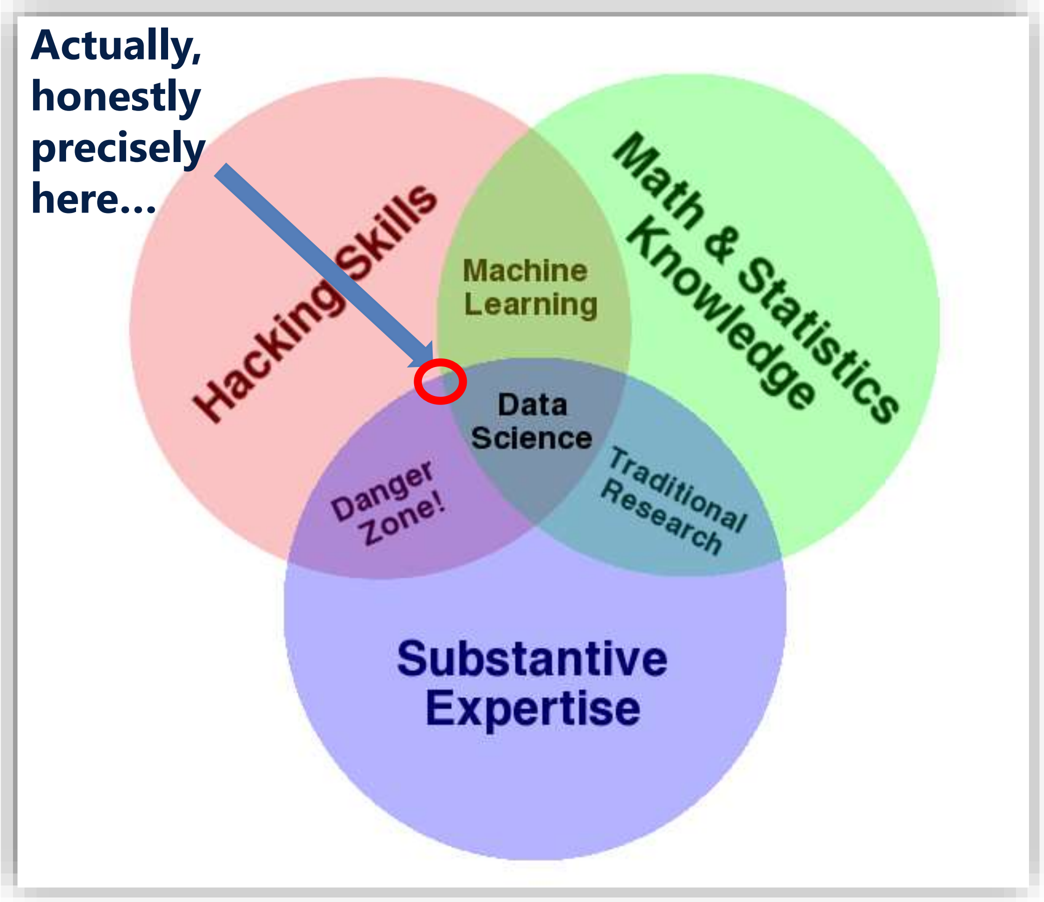
#Data Science #MachineLearning #Predictive Modelling #Statistics #Deep Learning …

- Risky, ROI uncertain
- Observational/Studies/Random Trials
- Very hard, complex
- Why? When (predict) ? (patterns)
- Comp. Advantage (& lots of hype too)
- Like our mind/ intelligence (intellect)
- Not a vital function



Yeap… pretty much that’s me he’s talking about…

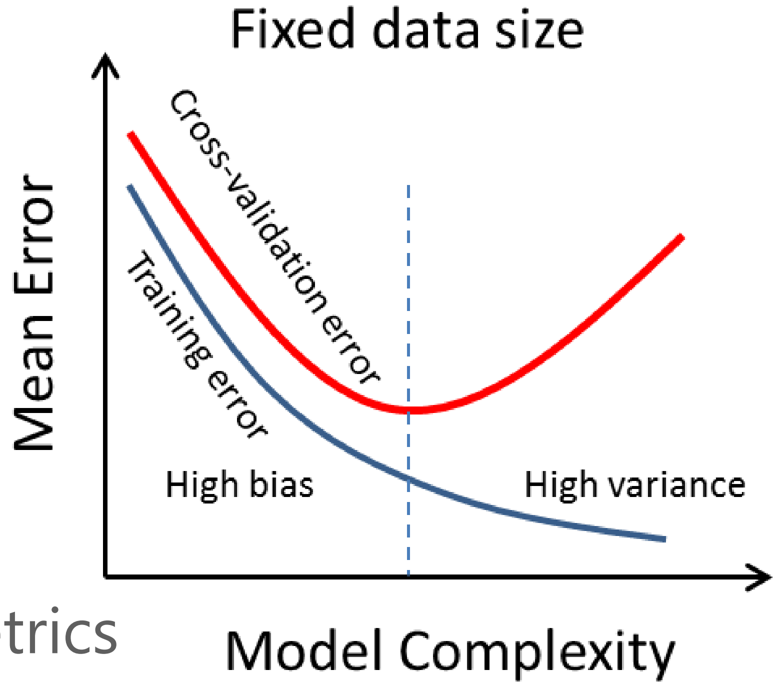
http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram



http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram

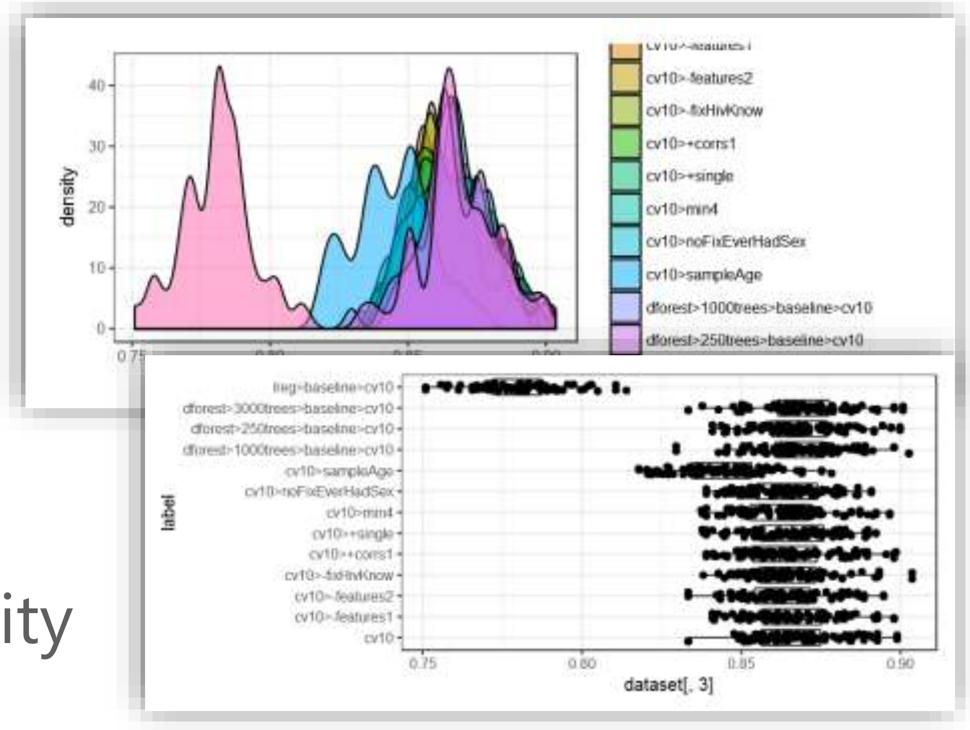
The basics

- Train vs Test vs Validation performance
- Overfitting & Under fitting
- Bias vs Variance
- Test splits, Cross validation
- Choose the right Evaluation metrics (loss weights, unbalanced datasets)



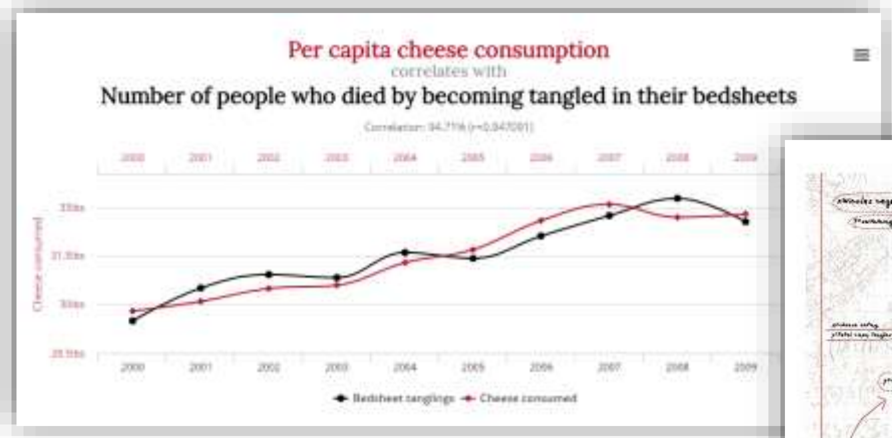
The basics

- Single number metrics…
- <u>Check your distributions!</u>
- Be aware/quantify <u>uncertainty</u>
- Do some baselines first (random guessing, majority class predictor)



Spurious correlations

- Correlation vs causation

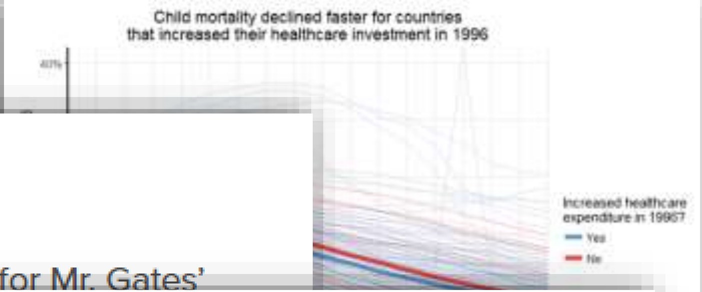


http://www.tylervigen.com/spurious-correlations

OCTOBER 18, 2016

Exploring the effects of healthcare investment on child mortality in R

@drsimon
mortality
interesting
informativ
ourworldi
peer-revie



Temporal precedence as an indicator of causality

The aim of this post is to provide some empirical support for Mr. Gates'
comment and investi
of causality: tempora
effect in time. Theref
healthcare expenditu
mortality rates.

A particular concern is whether temporal precendence, as evidenced here, is a solid enough indicator of a causal relationship. The truth is that it is not. Temporal precedence is a condition that is necessary, but not sufficient, to determine that a causal relationship exists. Thus, the evidence presented here might lend support to the notion of causality, but it is far from sufficient for being confident that it exists. As a scientist, I rely on randomized and controlled experiments to establish causality. But running such an experiment with healthcare will (hopefully) never happen. In my brief but

https://drsimonj.svbtle.com/exploring-a-causal-relation-between-healthcare-investment-and-child-mortality-in-r

Observational data vs Random Experiments

- Abundance & limits of observational data (aside from A/B testing,
- **pretty much of data we use these days is observational, limitations apply**
- What is? vs Why is?
- A/B Tests (Big Data Random Experiments)
- Correlation vs causation
- Con-founders

Experiments vs. Observational Studies

In an **experiment** investigators apply treatments to experimental units (people, animals, plots of land, etc.) and then proceed to observe the effect of the treatments on the experimental units.

In a **randomized experiment** investigators control the assignment of treatments to experimental units using a chance mechanism (like the flip of a coin or a computer's random number generator).

Experiments vs. Observational Studies (cont.)

In an **observational study** investigators observe subjects and measure variables of interest without assigning treatments to the subjects. The treatment that each subject receives is determined beyond the control of the investigator.

For example, suppose we want to study the effect of smoking on lung capacity in women.

Experiment

- Find 100 women age 20 who do not currently smoke.
- Randomly assign 50 of the 100 women to the smoking treatment and the other 50 to the no smoking treatment.
- Those in the smoking group smoke a pack a day for 10 years while those in the control group remain smoke free for 10 years.
- Measure lung capacity for each of the 100 women.
- Analyze, interpret, and draw conclusions from data.

Observational Study

- Find 100 women age 30 of which 50 have been smoking a pack a day for 10 years while the other 50 have been smoke free for 10 years.
- Measure lung capacity for each of the 100 women.
- Analyze, interpret, and draw conclusions from data.

http://www.public.iastate.edu/~dnett/S401/nexpvsobs.pdf

Fisher's Hypothesis

- Suppose there is a gene that causes smoking to appear to be a very pleasurable experience.
- Suppose the same gene also causes emphysema, lung cancer, throat cancer, etc.
- People who have the gene will be more likely to smoke than people who do not have the gene.
- People who have the gene will be more likely to get emphysema, lung cancer, throat cancer, etc.

5

Fisher's Hypothesis (cont.)

- So is it really smoking that causes health problems? Maybe it is just the gene?
- A **confounding** variable is related both to group membership and to the outcome of interest. Its presence makes it hard to establish the outcome as being a direct consequence of group membership.

6

Always Randomize if Possible

Consider a field experiment intended to compare the yield of two corn varieties (A and B).

Suppose the field is divided into 20 plots that run from one end of the field to the other.

Is there anything wrong with the following assignment of varieties to field plots?

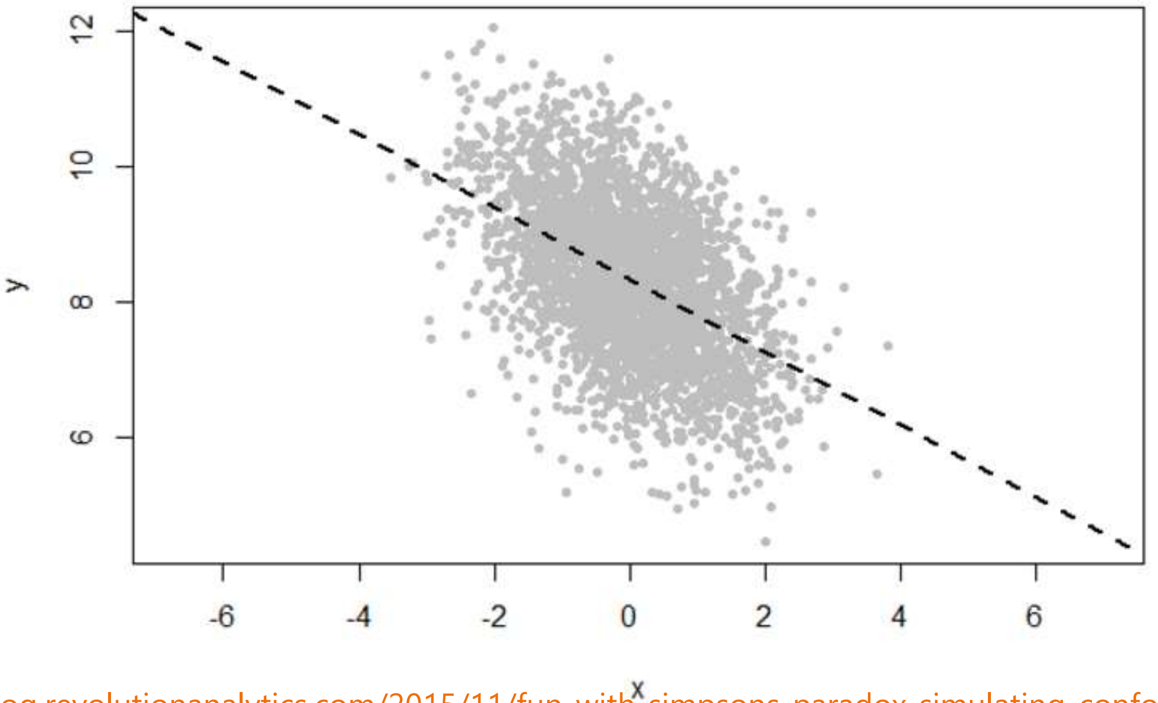
A B A B A B A B A B A B A B A B A B A B

http://www.public.iastate.edu/~dnett/S401/nexpvsobs.pdf

Observational data vs Random Experiments

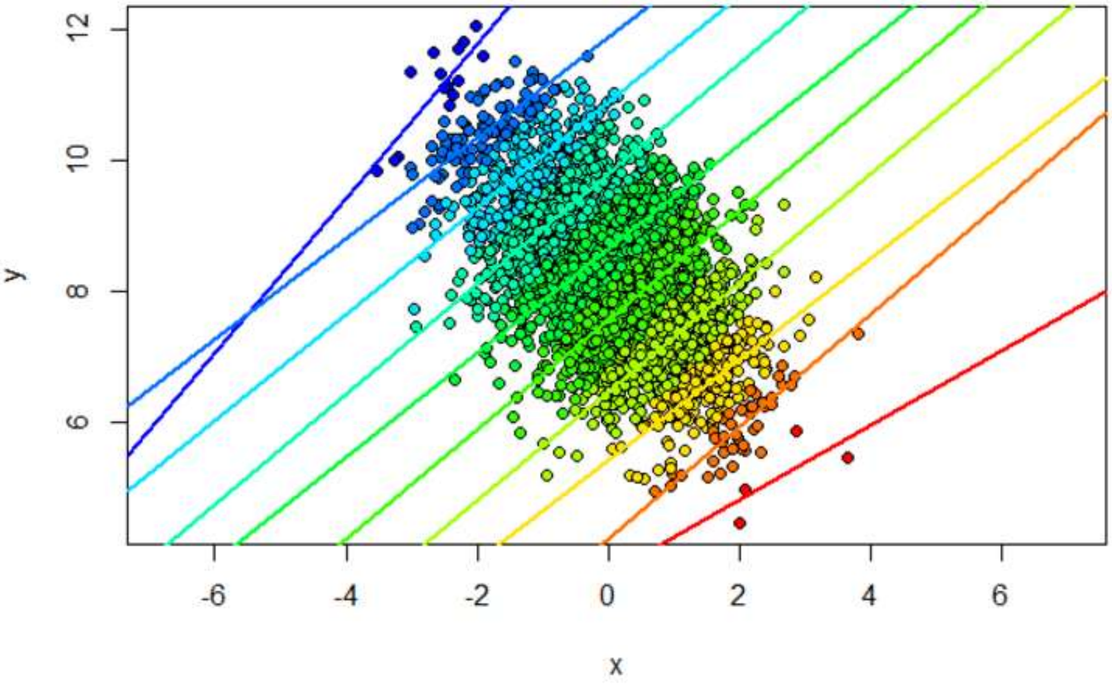
- Random Experiments try to approximate reality, controlling for every factor using chance/randomness (varying single variable between groups)
- Observational data, more real, but impossible to control for everything (ex: confounder/simpson's paradox)

Simpson's Paradox



http://blog.revolutionanalytics.com/2015/11/fun-with-simpsons-paradox-simulating-confounders.html

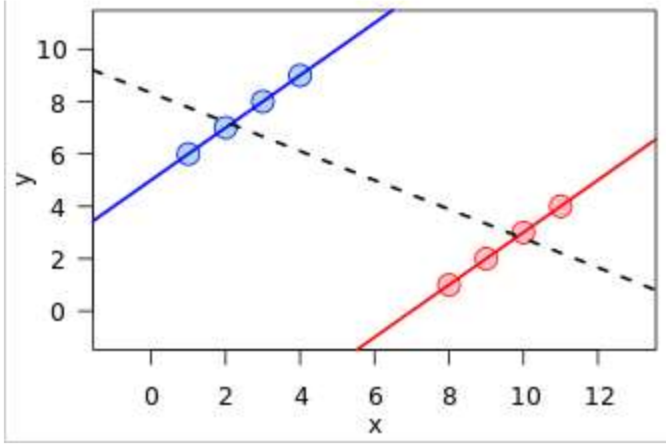
Simpson's Paradox



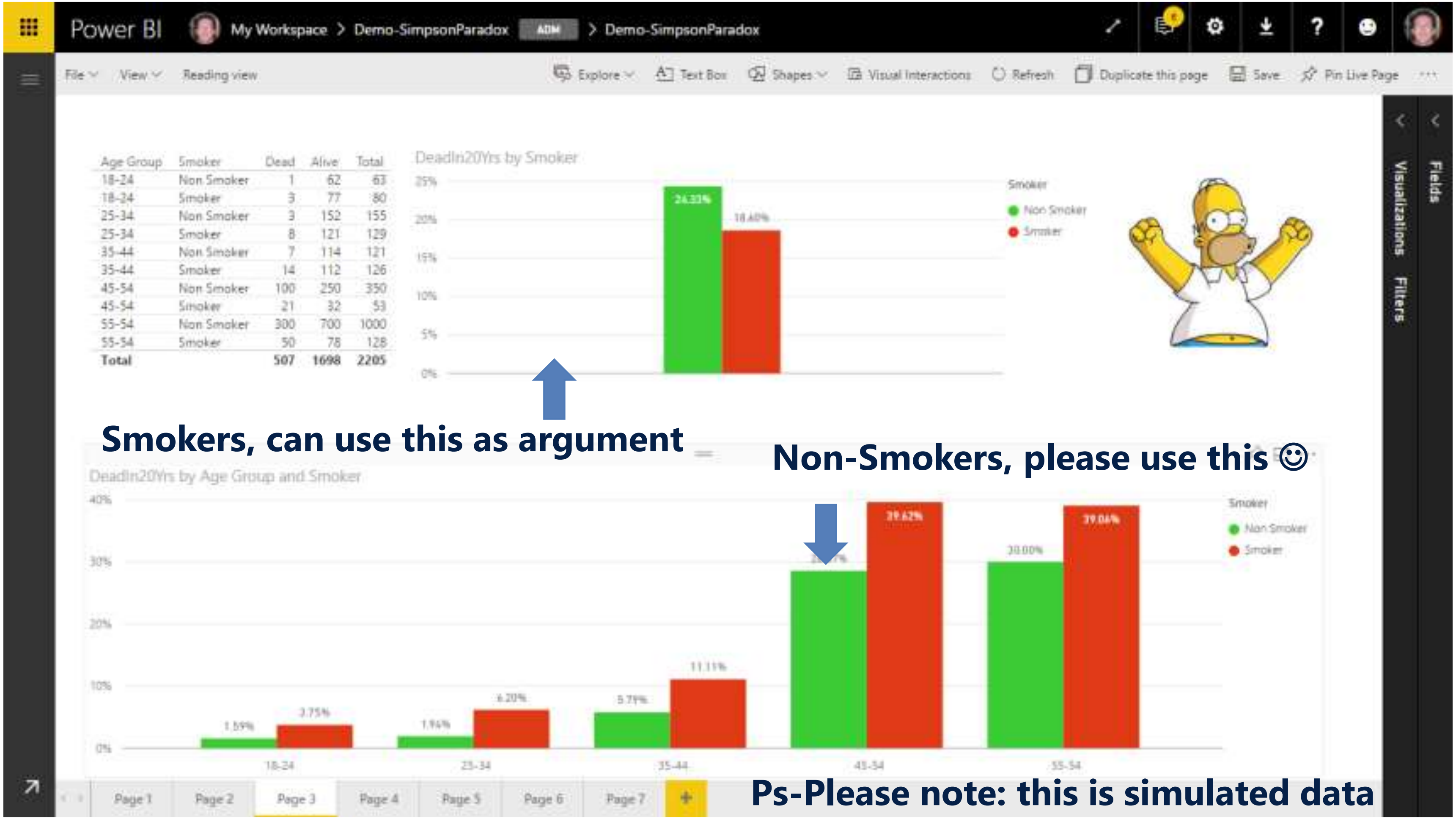
http://blog.revolutionanalytics.com/2015/11/fun-with-simpsons-paradox-simulating-confounders.html

Simpson’s Paradox

- Edward H. Simpson ,1951
- “single version of the truth”?
- Choose one!
- **“Any statistical relationship between two variables may be reversed by including additional factors in the analysis.” [Pearl2009]**

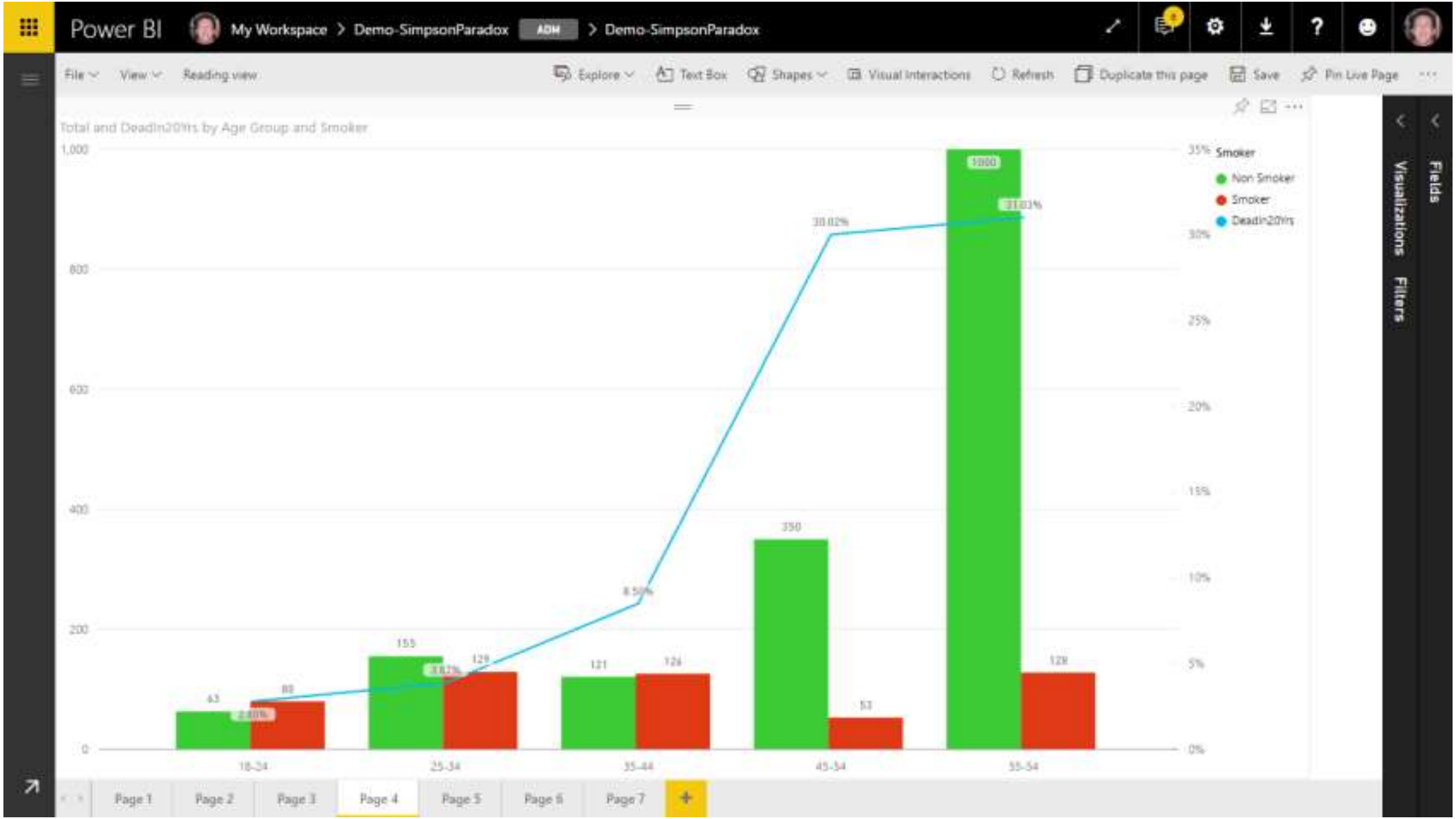


https://en.wikipedia.org/wiki/Simpson%27s_paradox

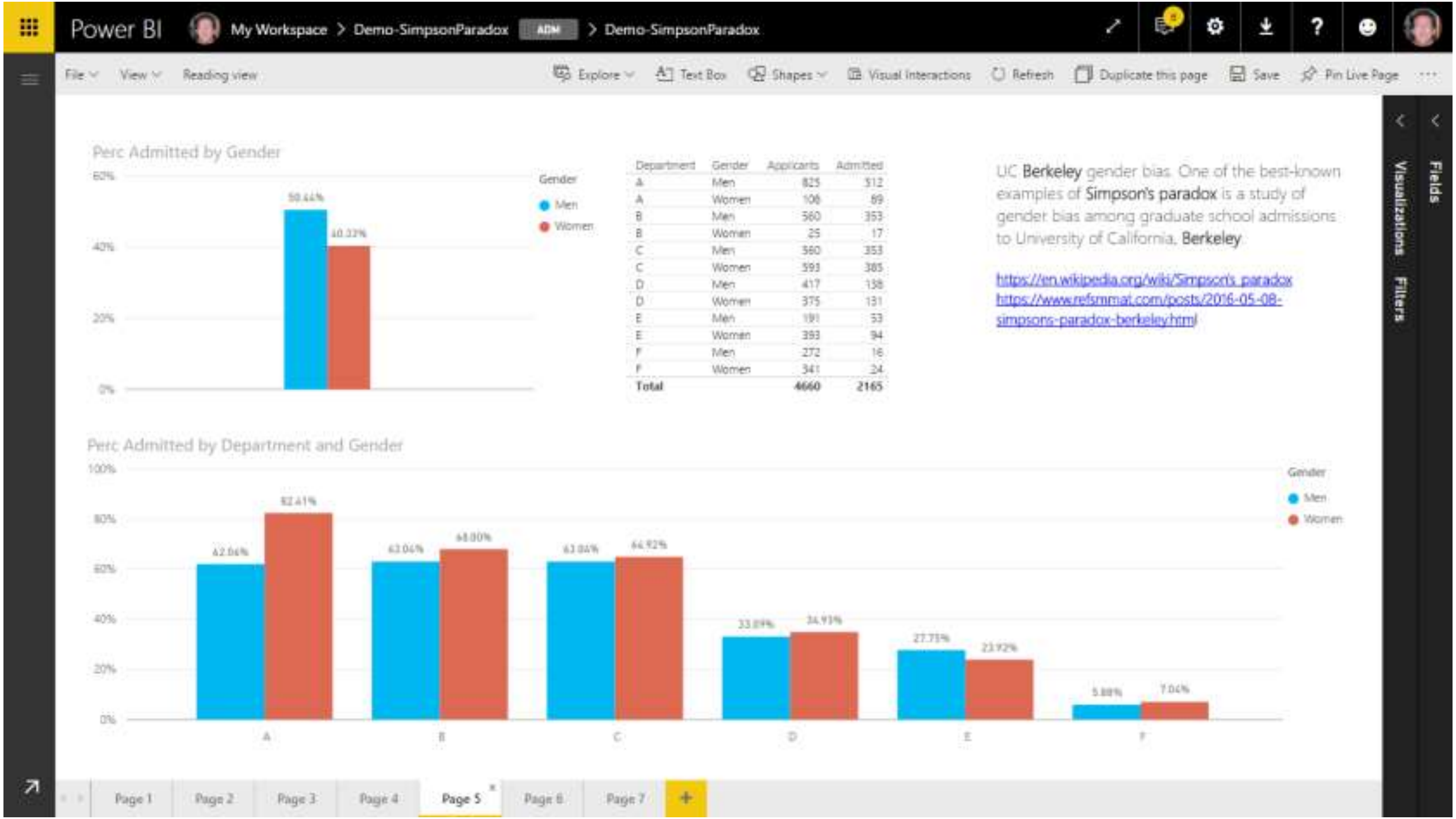


Age Group	Smoker	Dead	Alive	Total
18-24	Non Smoker	1	62	63
18-24	Smoker	3	77	80
25-34	Non Smoker	3	152	155
25-34	Smoker	8	121	129
35-44	Non Smoker	7	114	121
35-44	Smoker	14	112	126
45-54	Non Smoker	100	250	350
45-54	Smoker	21	32	53
55-54	Non Smoker	300	700	1000
55-54	Smoker	50	78	128
Total		507	1698	2205

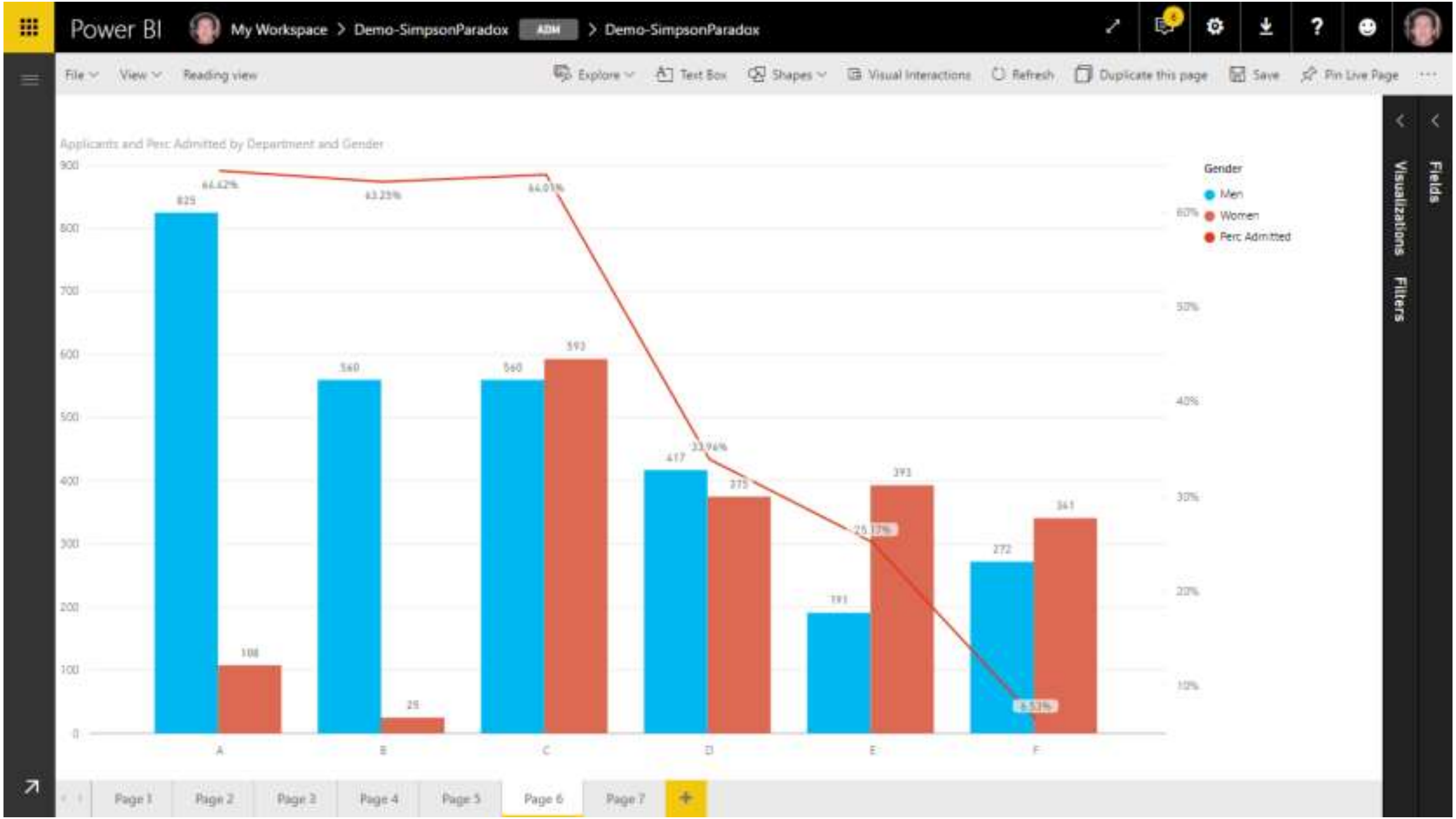
https://app.powerbi.com/view?r=eyJrIjoiMjk1N2JmNWMtZmZhNS00N2EzLWEzYTgtN2YxMzExNTYzZjlhIiwidCI6IjA5ZTI1MWRjLTVlODctNDhiZi1iNGQyLTcxYjAxYWRiOTg0YSIsImMiOjh9



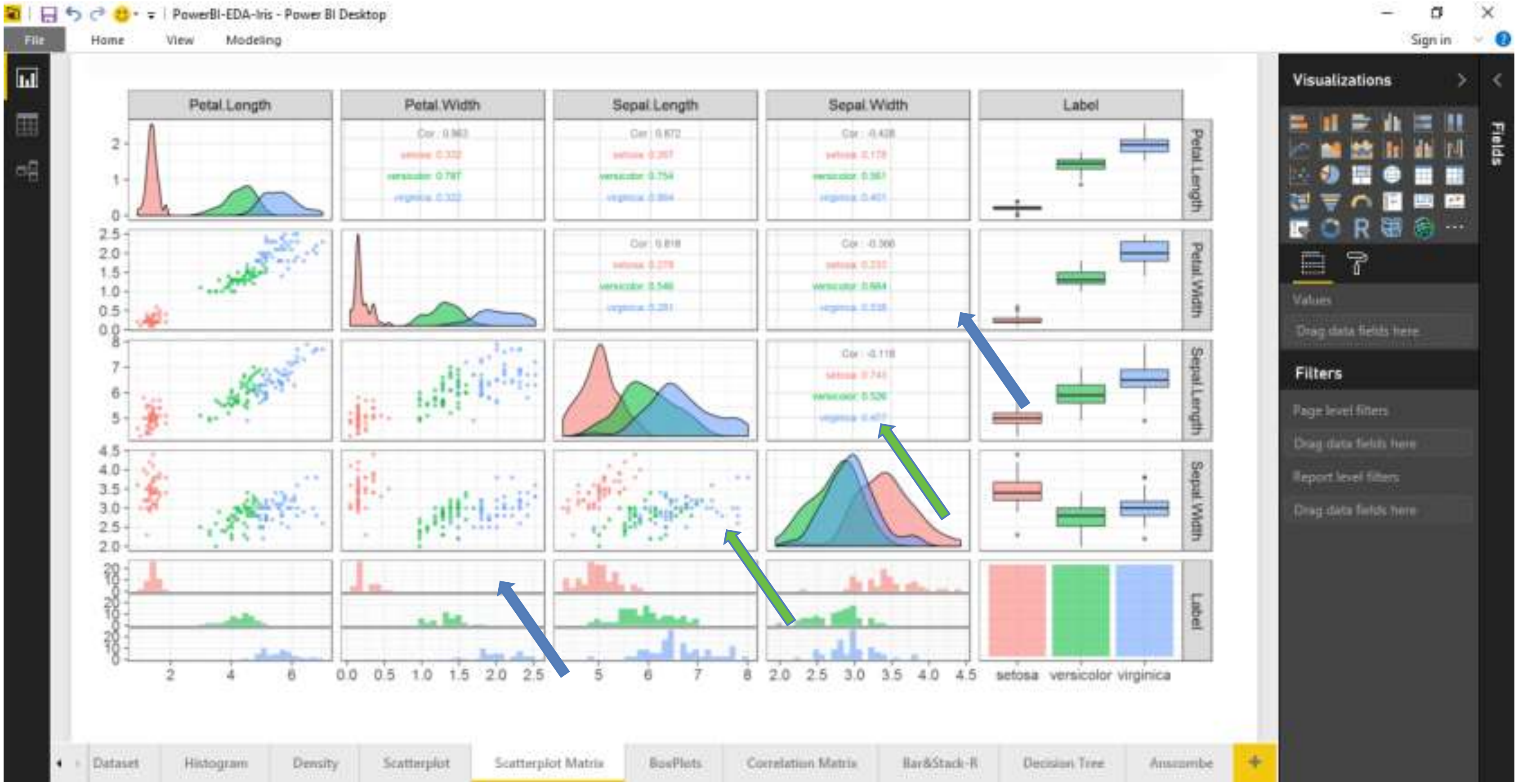
https://app.powerbi.com/view?r=eyJrIjoiMjk1N2JmNWMtZmZhNS00N2EzLWEzYTgtN2YxMzExNTYzZjlhIiwidCI6IjA5ZTI1MWRiLTVlODctNDhiZi1iNGQyLTcxYjAxYWRiOTg0YSIsImMiOjh9



https://www.refsmmat.com/posts/2016-05-08-simpsons-paradox-berkeley.html



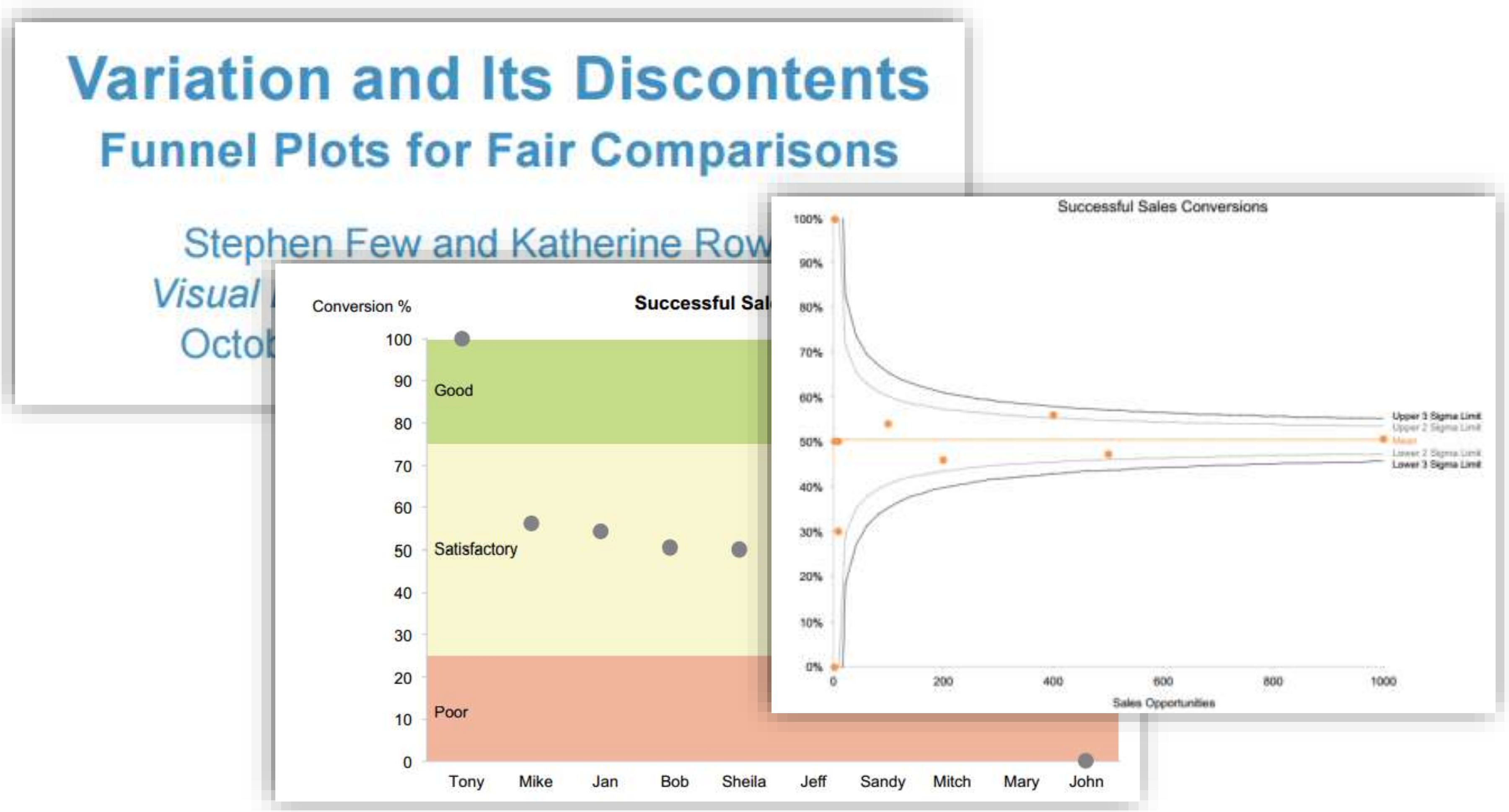
https://www.refsmmat.com/posts/2016-05-08-simpsons-paradox-berkeley.html



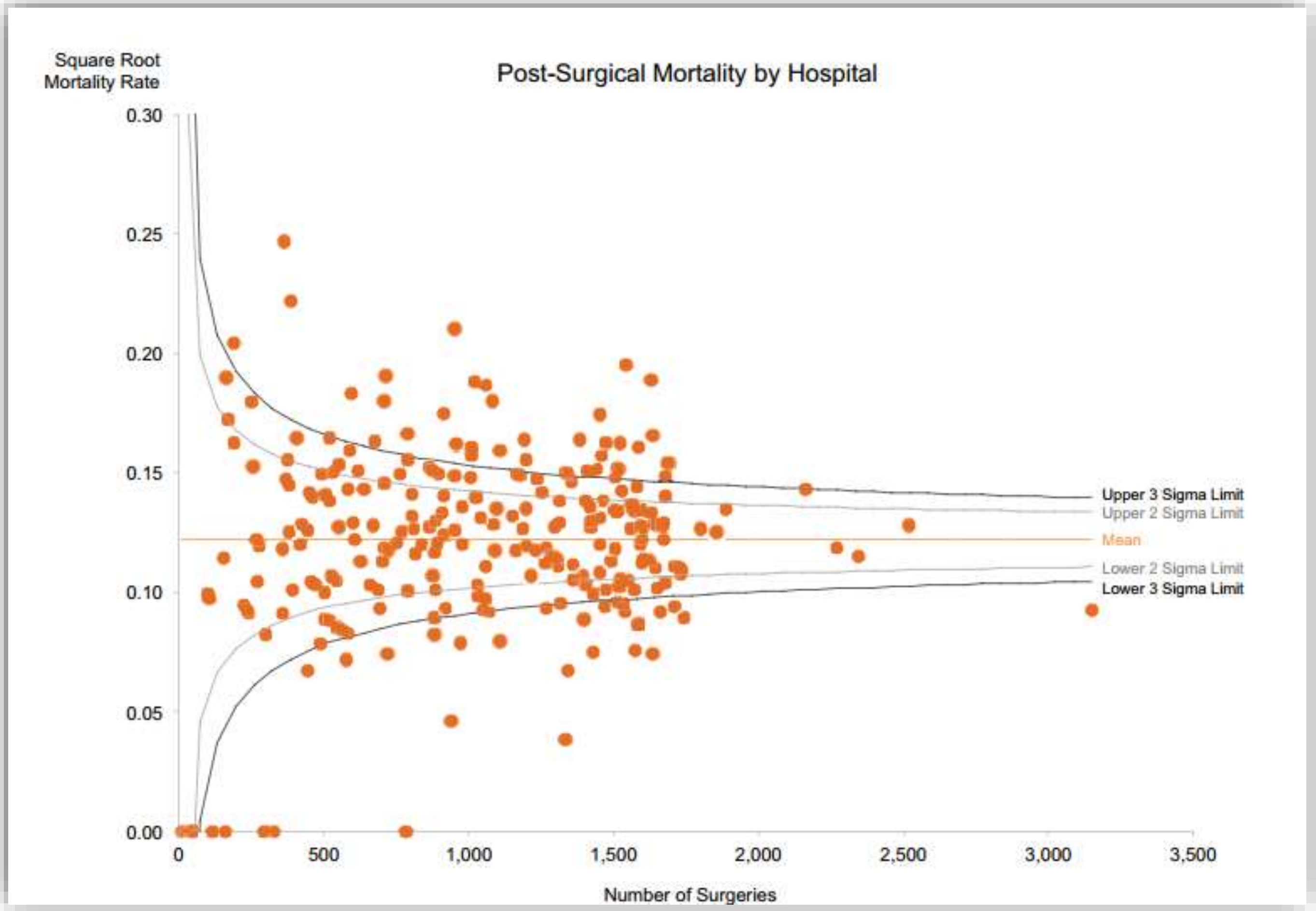
Ex: Simpson Paradox using Fisher's Iris Dataset

Statistically Significant result?

- a result (ex. a difference) that’s not likely attributed to chance
- P Values interpretation/validity



https://www.perceptualedge.com/articles/visual_business_intelligence/variation_and_its_discontents.pdf



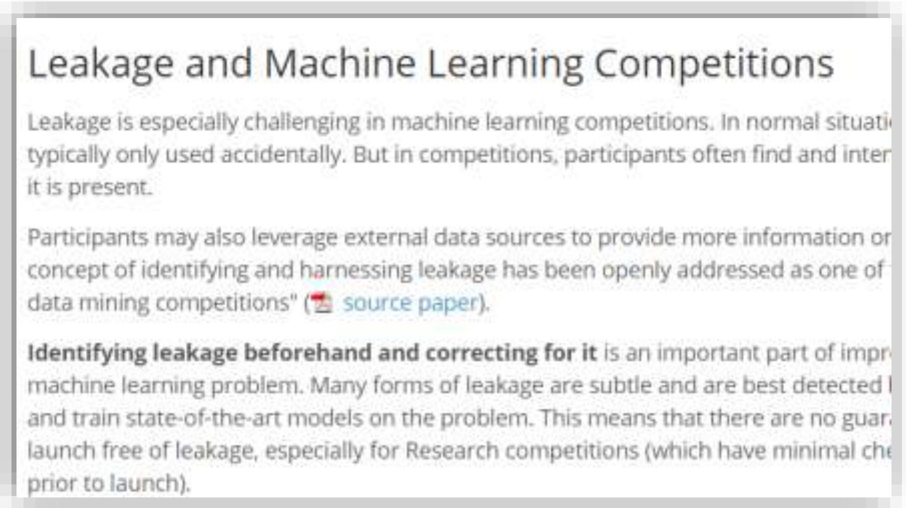
https://www.perceptualedge.com/articles/visual_business_intelligence/variation_and_its_discontents.pdf

Data Leakage

- “it’s sunny on sunny days”…
- Inflates generalization performance estimate
- Ex: label information leaks into features
- Ex: features capture events/data occurring after event of interest
- Can be very hard to detect
- Ex: label aware feature selection methods
- Touch/see data once (inner loop feature selection)
- (did you use data for EDA/feature selection/modelling decisions? -> don’t use it for evaluation)

Data Leakage

- Ps-extremely explored on ML competitions
- …creating models that can win competitions but be pretty much useless, unrealistic



Leakage and Machine Learning Competitions

Leakage is especially challenging in machine learning competitions. In normal situati typically only used accidentally. But in competitions, participants often find and inter it is present.

Participants may also leverage external data sources to provide more information or concept of identifying and harnessing leakage has been openly addressed as one of data mining competitions" (source paper).

Identifying leakage beforehand and correcting for it is an important part of impr machine learning problem. Many forms of leakage are subtle and are best detected and train state-of-the-art models on the problem. This means that there are no guar launch free of leakage, especially for Research competitions (which have minimal che prior to launch).

https://www.kaggle.com/wiki/Leakage

Machine Learning “Insights” – possible?

- Feature Importance is not causality
- Observational data remember?
- Side effects of “insights”

“Black Boxes”, “White Boxes”

NUMBERS | ARTIFICIAL INTELLIGENCE

Is Artificial Intelligence Permanently Inscrutable?

Despite new biology-like tools, some insist interpretation is impossible.

BY AARON M. BORNSTEIN
ILLUSTRATION BY EMMANUEL POLANCO
SEPTEMBER 1, 2016

ADD A COMMENT FACEBOOK TWITTER EMAIL SHARING

http://nautil.us/issue/40/learning/is-artificial-intelligence-permanently-inscrutable

“Black Boxes”, “White Boxes”

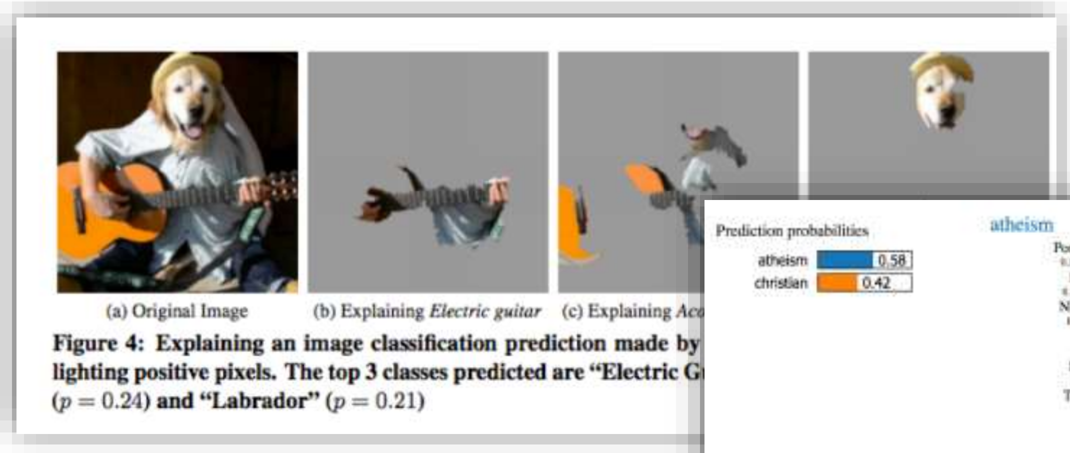
The neural networks were right more often than any of the other methods. But when the researchers and doctors took a look at the human-readable rules, they noticed something disturbing: One of the rules instructed doctors to send home pneumonia patients who already had asthma, despite the fact that asthma sufferers are known to be extremely vulnerable to complications.

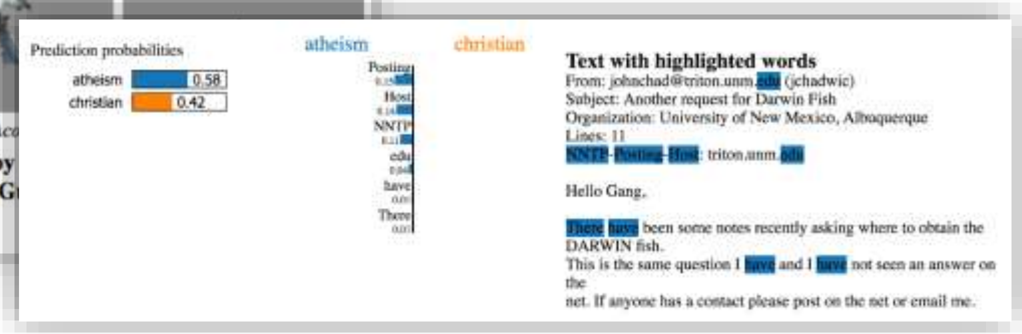
The model did what it was told to do: Discover a true pattern in the data. The poor advice it produced was the result of a quirk in that data. It was hospital policy to send asthma sufferers with pneumonia to intensive care, and this policy worked so well that asthma sufferers almost never developed severe complications. Without the extra care that had shaped the hospital’s patient records, outcomes could have been dramatically different.

http://nautil.us/issue/40/learning/is-artificial-intelligence-permanently-inscrutable

"Black Boxes", "White Boxes"

LIME - Local Interpretable Model-Agnostic Explanations

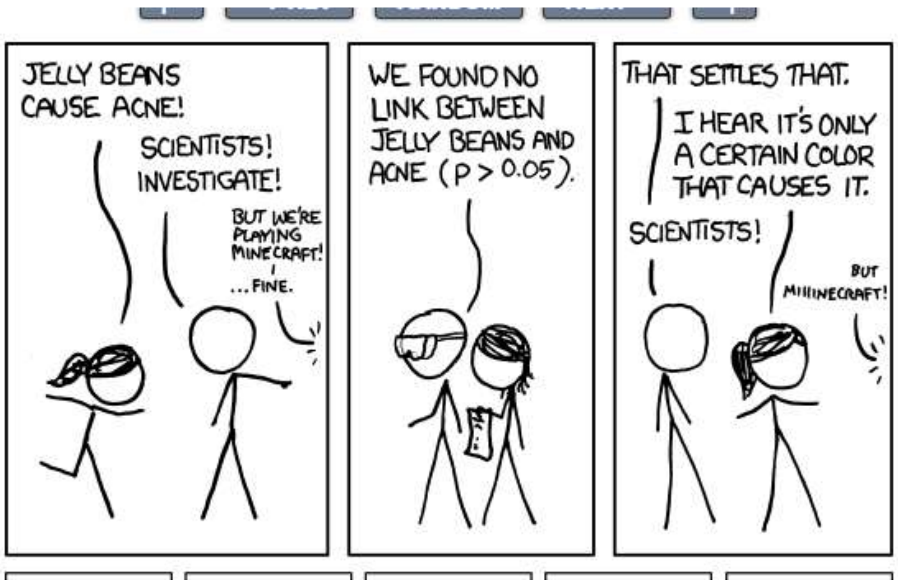




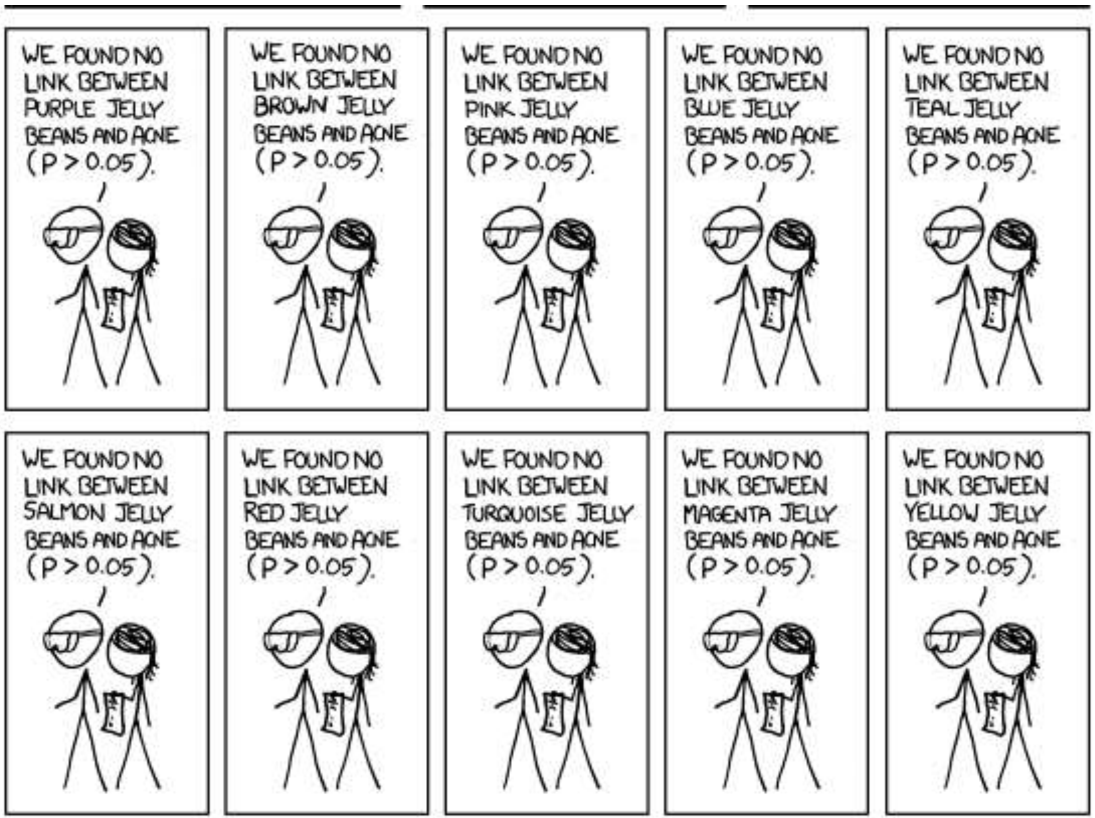
https://homes.cs.washington.edu/~marcotcr/blog/lime/

P Hacking

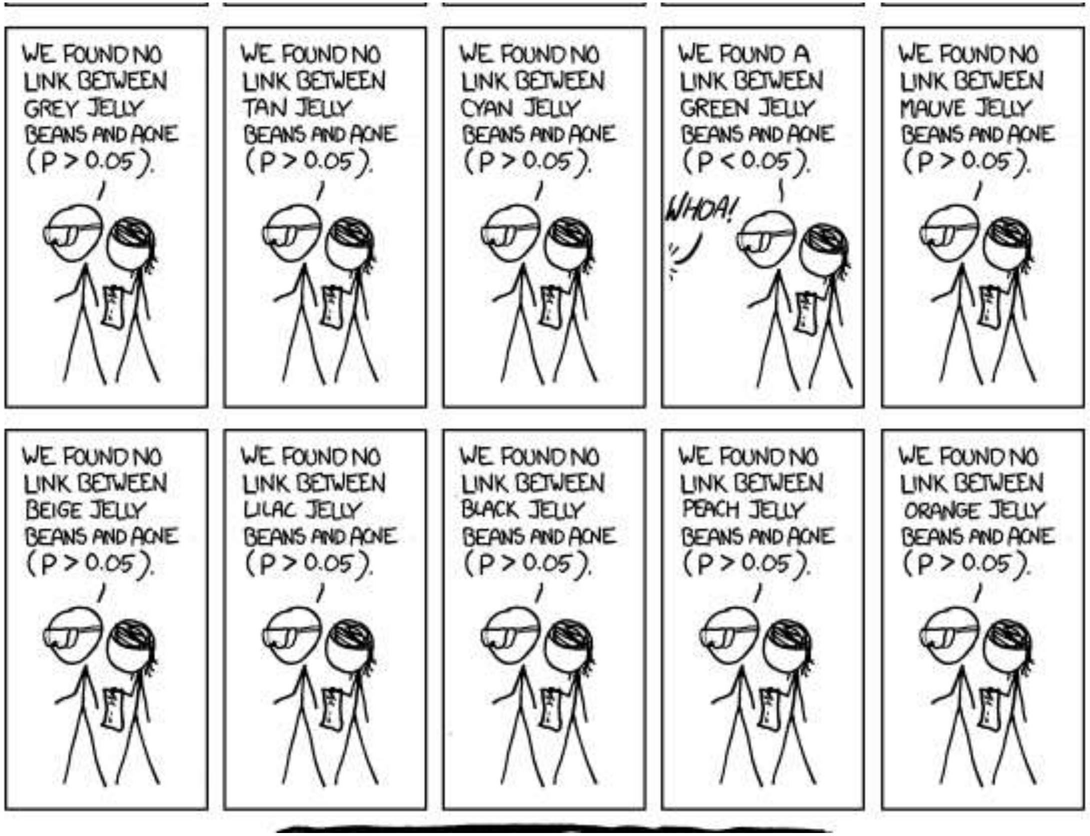
- Abundance of data
- Multiple statistical tests pitfalls



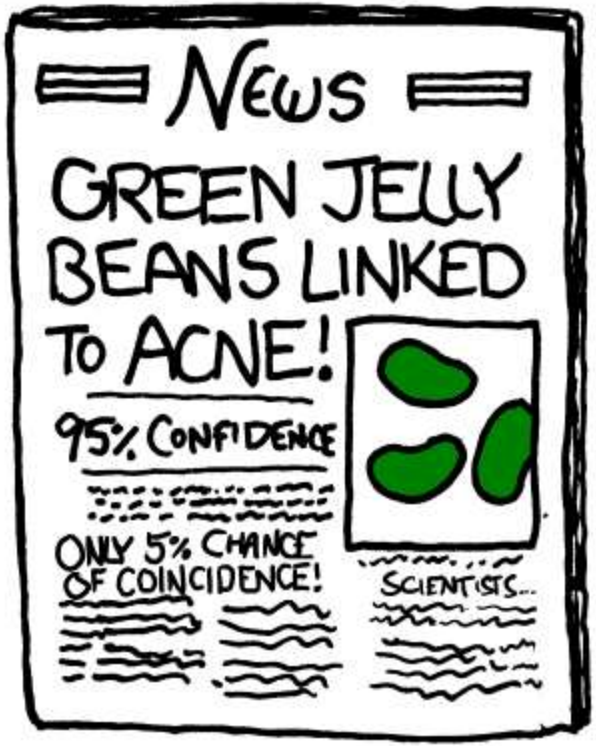
https://xkcd.com/882/



https://xkcd.com/882/

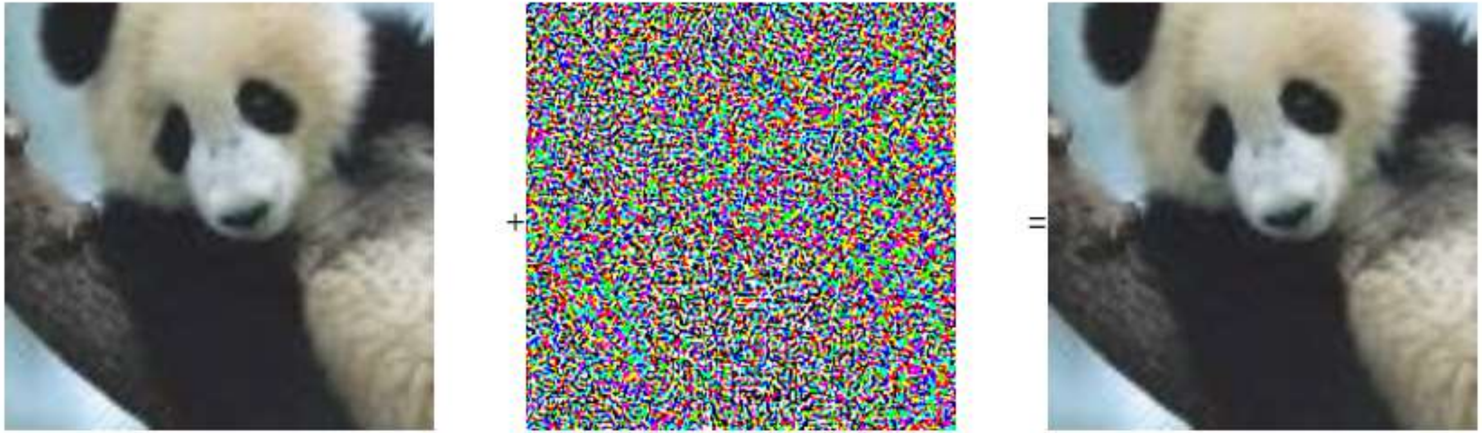


https://xkcd.com/882/



https://xkcd.com/882/

Adversarial Examples (Deep Learning/CNNs)



Original image classified as a panda with 60% confidence.

Tiny adversarial perturbation.

Imperceptibly modified image, classified as a gibbon with 99% confidence.

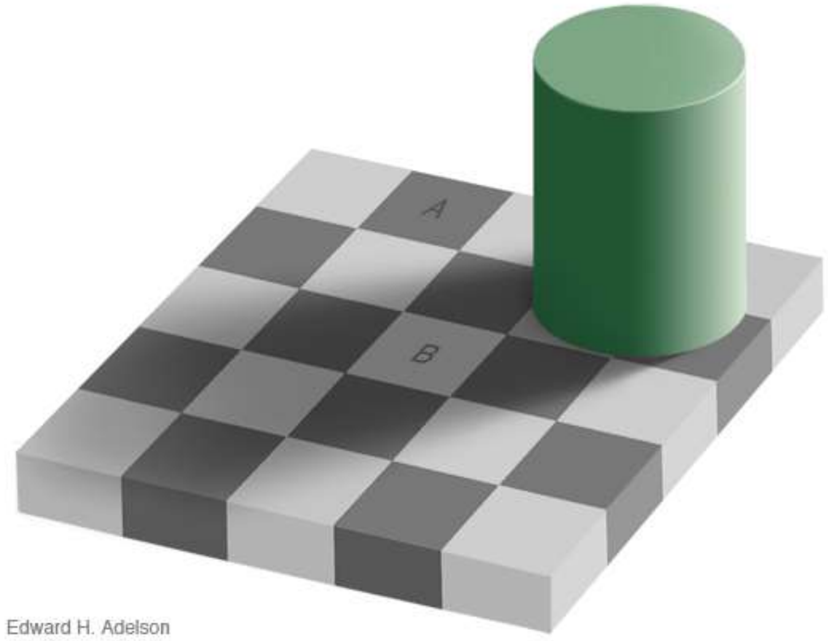
http://www.kdnuggets.com/2015/07/deep-learning-adversarial-examples-misconceptions.html

http://karpathy.github.io/2015/03/30/breaking-convnets/

Closing…

- Too good to be true? it probably isn't… (true!)
- Ego control, awareness of cognitive biases, sunk costs, confirmation bias & others
  - https://betterhumans.coach.me/cognitive-bias-cheat-sheet-55a472476b18
  - http://mentalfloss.com/article/68705/20-cognitive-biases-affect-your-decisions
- "Shitty hypothesis" ☺ - assuming I have this awesome results & "I really messed up" is true, what have I done wrong? (ask: what else could cause this?)
- Occam's Razor & Hickam's dictum
- A visual guide to Bayesian thinking
  - https://www.youtube.com/watch?v=BrK7X_XlGB8

fun: How sure are you?

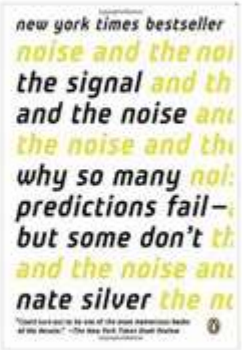


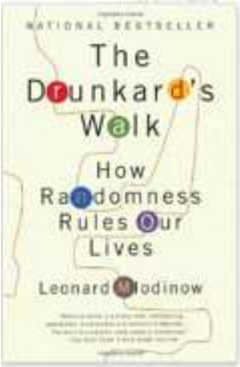
http://web.mit.edu/persci/people/adelson/checkershadow_illusion.html

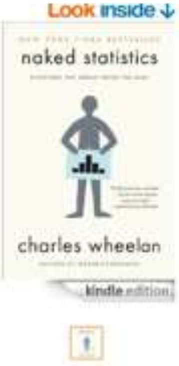
References

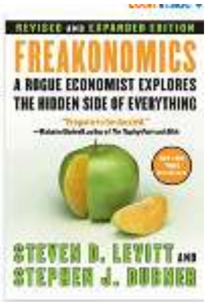
https://www.youtube.com/watch?v=tleeC-KlsKA

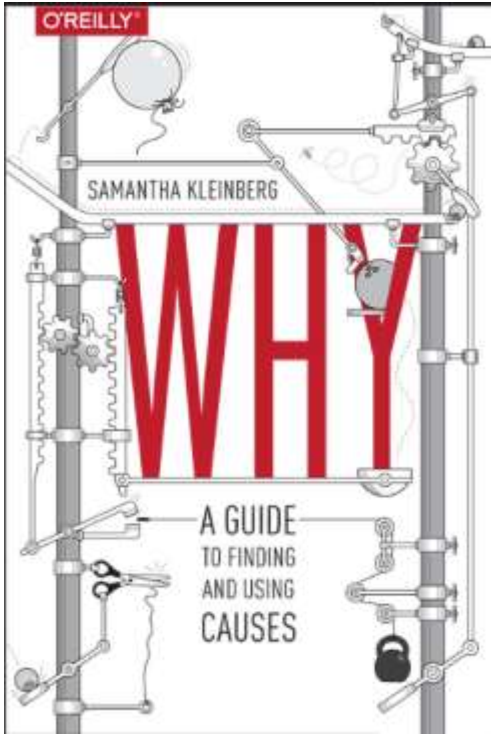
References

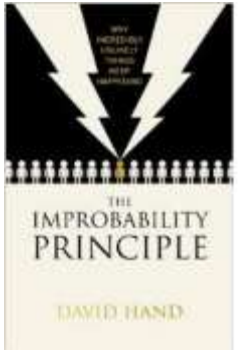

new york times bestseller
the signal and the noise
why so many predictions fail—but some don't
nate silver

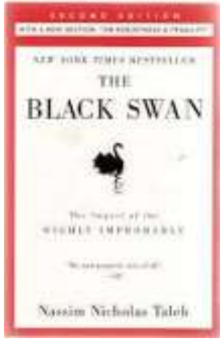

NATIONAL BESTSELLER
The Drunkard's Walk
How Randomness Rules Our Lives
Leonard Mlodinow

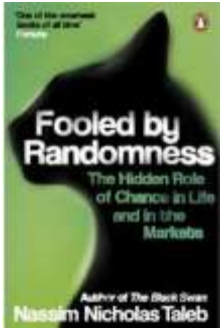

Look inside
naked statistics
charles wheelan
kindle edition

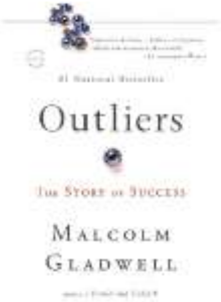

REVISED AND EXPANDED EDITION
FREAKONOMICS
A ROGUE ECONOMIST EXPLORES THE HIDDEN SIDE OF EVERYTHING
STEVEN D. LEVITT AND STEPHEN J. DUBNER


O'REILLY
SAMANTHA KLEINBERG
WHY
A GUIDE TO FINDING AND USING CAUSES


THE IMPROBABILITY PRINCIPLE
DAVID HAND


SECOND EDITION
NEW YORK TIMES BESTSELLER
THE BLACK SWAN
Nassim Nicholas Taleb


Fooled by Randomness
The Hidden Role of Chance in Life and in the Markets
Author of The Black Swan
Nassim Nicholas Taleb


Outliers
The Story of Success
Malcolm Gladwell

References

- Pedro Domingos-A Few Useful Things to Know about Machine Learning
- Claudia Perlich - Leakage in Data Mining: Formulation, Detection, and Avoidance
- Machine Learning Mastery- Common Pitfalls In Machine Learning Projects
- Daniel Nee- Common Pitfalls in Machine Learning

Coming soon…

- Out of sample
- Sample Bias
- suggestions?

  -> rui.quintino@devscope.net

devscope

Rua Passos Manuel Nº 223 – 4º Andar
4000-385 Porto

Av. Sidónio Pais, Nº 2 – 3º Andar
1050-214 5 Lisboa

T. +351 223 751 350/51
F. +351 223 751 352

info@devscope.net
www.devscope.net