2018 PSI Conference

Breaking Boundaries in Drug Development

Amsterdam

3-6 June 2018

Statistics and Data Science

Stephen J. Ruberg, PhD
President
Analytix Thinking, LLC

Motivation

"… the time may not be very remote when it will be understood that for complete initiation as an efficient citizen of one of the new great complex world-wide States that are now developing, it is as necessary to be able to compute, to think in averages and maxima and minima, as it is now to be able to read and write."

H. G. Wells
Mankind in the Making,

"Statistical thinking will one day be as necessary for efficient citizenship as the ability to read and write."

Samuel S. Wilks
Presidential Address to the American Statistical Association
December 28, 1950

Bringing data to life.

Example #1
Predicting Alzheimer’s Disease
Big Data - Biomarkers

Predicting Alzheimer's Disease

<u>Problem Statement</u>

There are no good treatments for Alzheimer's Disease

- By the time it is diagnosed, it may be too late.

Detecting it early may be a key to treatment or prevention

Current imaging approaches are expensive and invasive

Ideally, a blood test would be easy, cheap and very helpful

Predicting Alzheimer’s Disease

Researchers at …

- University of Rochester School of Medicine
- University of California, Irvine School of Medicine
- Georgetown University Medical Center

Study Outline

- Follow cognitively normal elderly patients over time
- Identify which patients “convert” to amnestic Mild Cognitive Impairment (aMCI) or Alzheimer’s Disease (AD)
- Examine baseline blood proteins from “converters” and “non-converters” for differences

Bringing data to life.

The (Statistical) Analytical Methods

"groups were defined primarily using a composite measure of memory performance"

"Metabolites defining the participant groups were selected using the least absolute shrinkage and selection operator (LASSO) penalty."

"… metabolomic data from the untargeted LASSO analysis to build separate linear classifier models …"

"… used receiver operating characteristic (ROC) analysis to assess the performance of the classifier models …"

"… employed internal cross-validation …"

> "The optimal value of the tuning parameter lambda, which was obtained by the cross-validation procedure, was then used to fit the model."

"… matched … participants on the basis of age, sex and education level."

"… used separate multivariate ANOVA (MANOVA) to examine discovery and validation group performance …"

"… used Tukey's honestly significant difference (HSD) procedure for post hoc comparisons."

"… quantitative profiling data was subjected to the nonparametric Kruskal-Wallis test … followed by Mann-Whitney U-tests for post hoc pairwise comparisons …. Significance was adjusted for multiple comparisons using Bonferroni's method ($P < 0.025$)."

Bringing data to life.

The Results

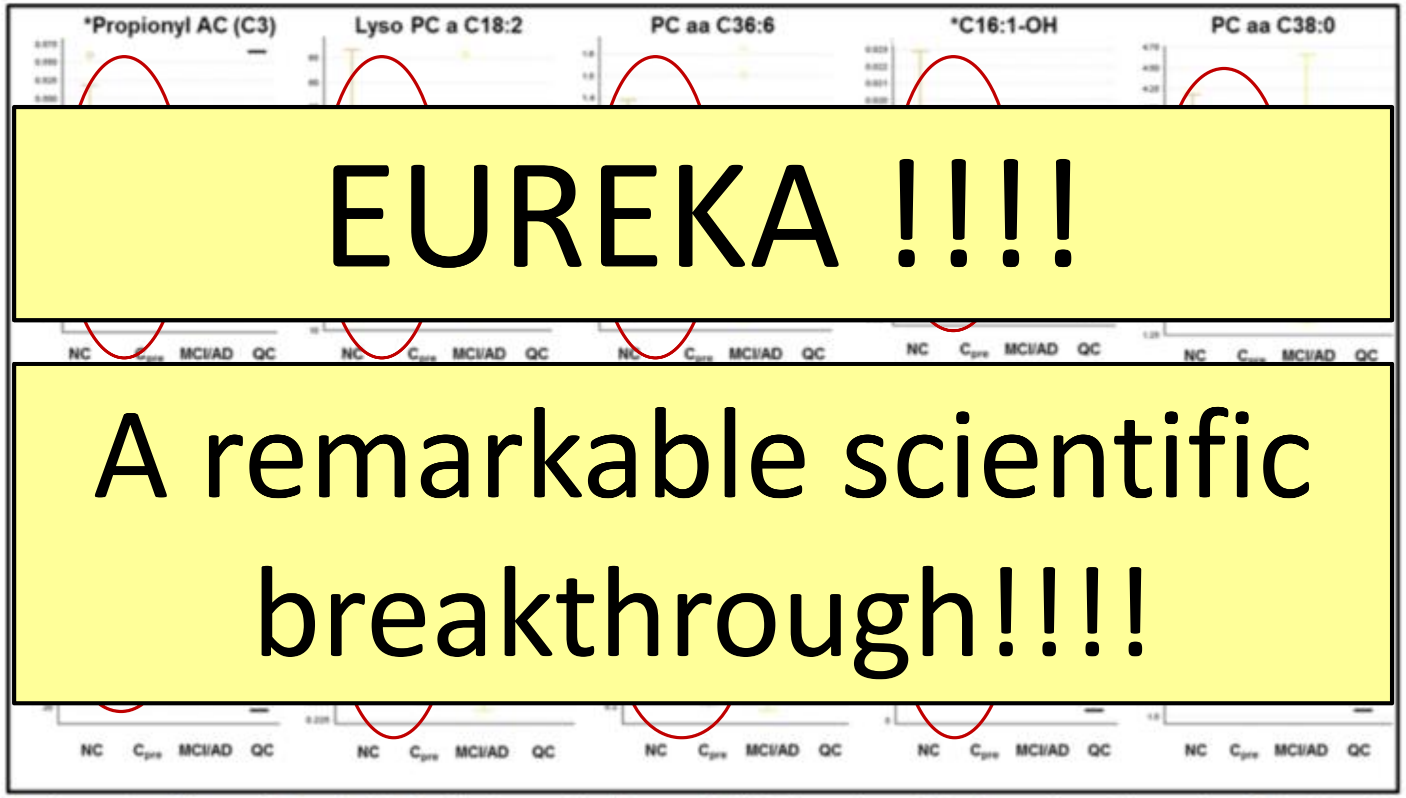
Mar 09, 2014

ABSTRACT

… Herein, we describe our lipidomic approach to detecting preclinical Alzheimer's disease in a group of cognitively normal older adults. **We discovered and validated a set of ten lipids from peripheral blood that predicted phenoconversion to either amnestic mild cognitive impairment or Alzheimer's disease within a 2–3 year timeframe with over 90% accuracy.** This biomarker panel, reflecting cell membrane integrity, may be sensitive to early neurodegeneration of preclinical Alzheimer's disease.

Bringing data to life.

The Results



EUREKA !!!!

A remarkable scientific breakthrough!!!!

Bringing data to life.

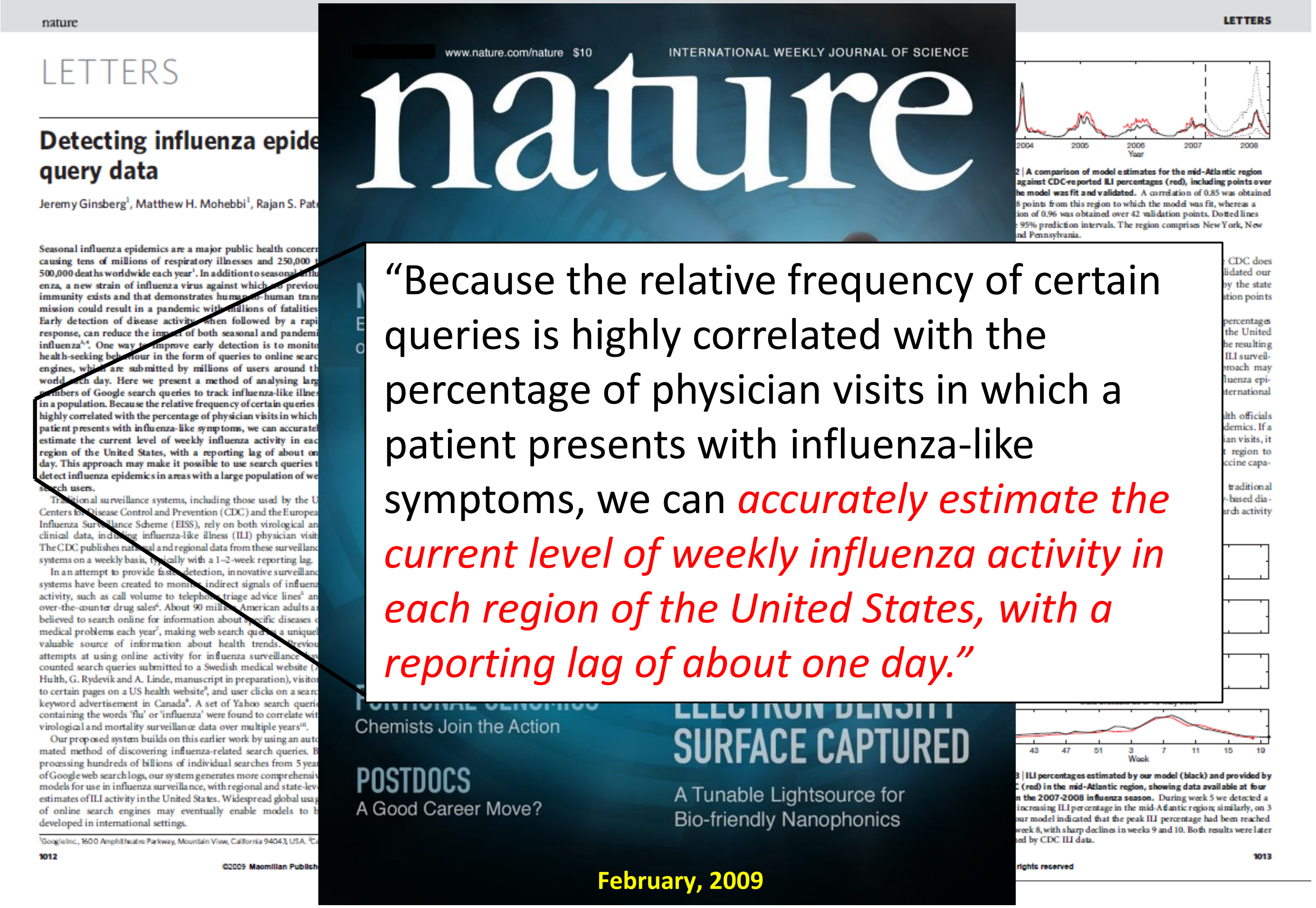
The Publicity

In a first-of-its-kind study, researchers have developed a **blood test for Alzheimer's disease that predicts with astonishing accuracy** whether a healthy person will develop the disease.

Bringing data to life.

Example #2
Public Health
Big Data – Social Media

Social Media in Action



"Because the relative frequency of certain queries is highly correlated with the percentage of physician visits in which a patient presents with influenza-like symptoms, we can *accurately estimate the current level of weekly influenza activity in each region of the United States, with a reporting lag of about one day.*"

Social Media in Action

nature International weekly journal of science

Detecting influenza epidemics using search engine query data

Seasonal influenza epidemics are a major public health concern, causing tens of millions of respiratory illnesses and 250,000 to 500,000 deaths worldwide each year[1]. In addition to seasonal influenza, a new strain of influenza virus against which no previous immunity exists and that demonstrates human-to-human transmission could result in a pandemic with millions of fatalities[2]. Early detection of disease activity, when followed by a rapid response, can reduce the impact of both seasonal and pandemic influenza[3,4]. One way to improve early detection is to monitor health-seeking behaviour in the form of queries to online search engines, which are submitted by millions of users around the world each day. Here we present a method of analysing large numbers of Google search queries to track influenza-like illness in a population. Because the relative frequency of certain queries is highly correlated with the percentage of physician visits in which a patient presents with influenza-like symptoms, we can accurately estimate the current level of weekly influenza activity in each region of the United States, with a reporting lag of about one day. This approach may make it possible to use search queries to detect influenza epidemics in areas with a large population of web search users.

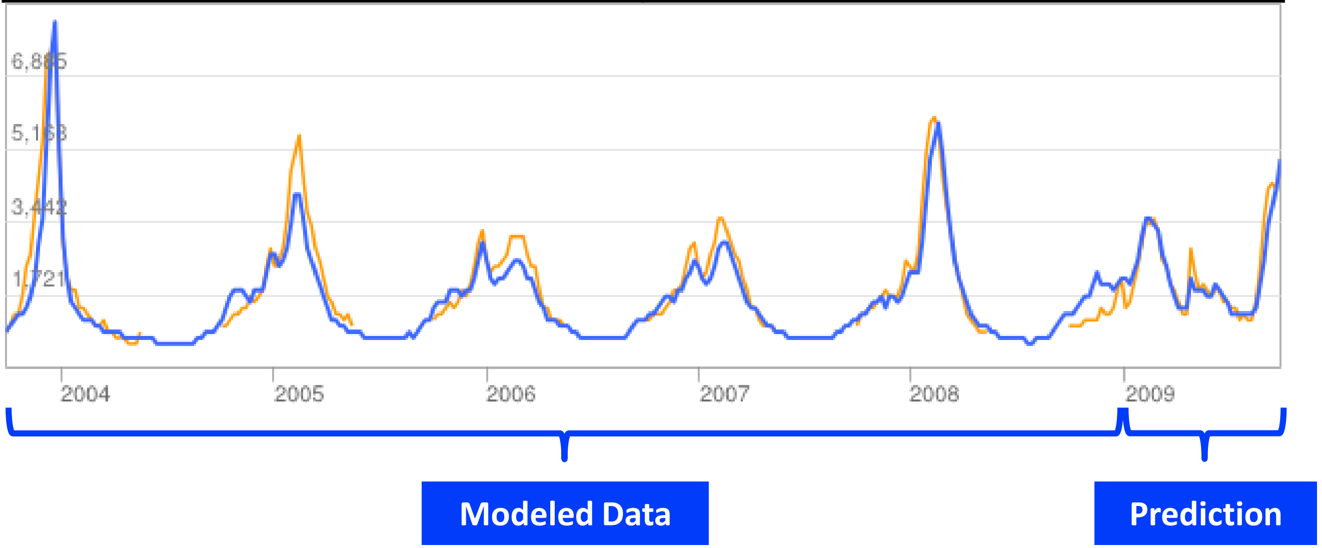
"… **analyzing large numbers of Google queries** to track influenza-like illness in a population. Because the relative **frequency of certain queries is highly correlated with the percentage of physician visits** in which a patient presents with flu-like symptoms, we can accurately estimate the current level of weekly influenza activity …"

Bringing data to life.

Social Media in Action

United States Flu Activity

Influenza estimate	**Google Flu Trends Estimate** **United States Data**



Bringing data to life.

Social Media in Action

“Google web search queries can be used to estimate ILI percentages accurately in each of the nine public health regions of the United States. Because search queries can be processed quickly, the resulting ILI estimates were consistently 1–2 weeks ahead of CDC ILI surveillance reports. The early detection provided by this approach may become an important line of defense against future influenza epidemics in the United States, and perhaps eventually in international settings.”

ILI = Influenza-like illness

Bringing data to life.

Triumph of Big Data

"… researchers have focused on the observation that search 'predicts the present,' meaning that search volume correlates with contemporaneous events."

S. Goel et al.
Proc. Natl. Acad. Sci. U.S.A., 107
Oct 12, 2010, p. 17486

"… simple models and big data trump more-elaborate analytics approaches."

A. McAfee, E. Brynjolfsson
Harvard Business Review, 90
Oct, 2012, p. 64

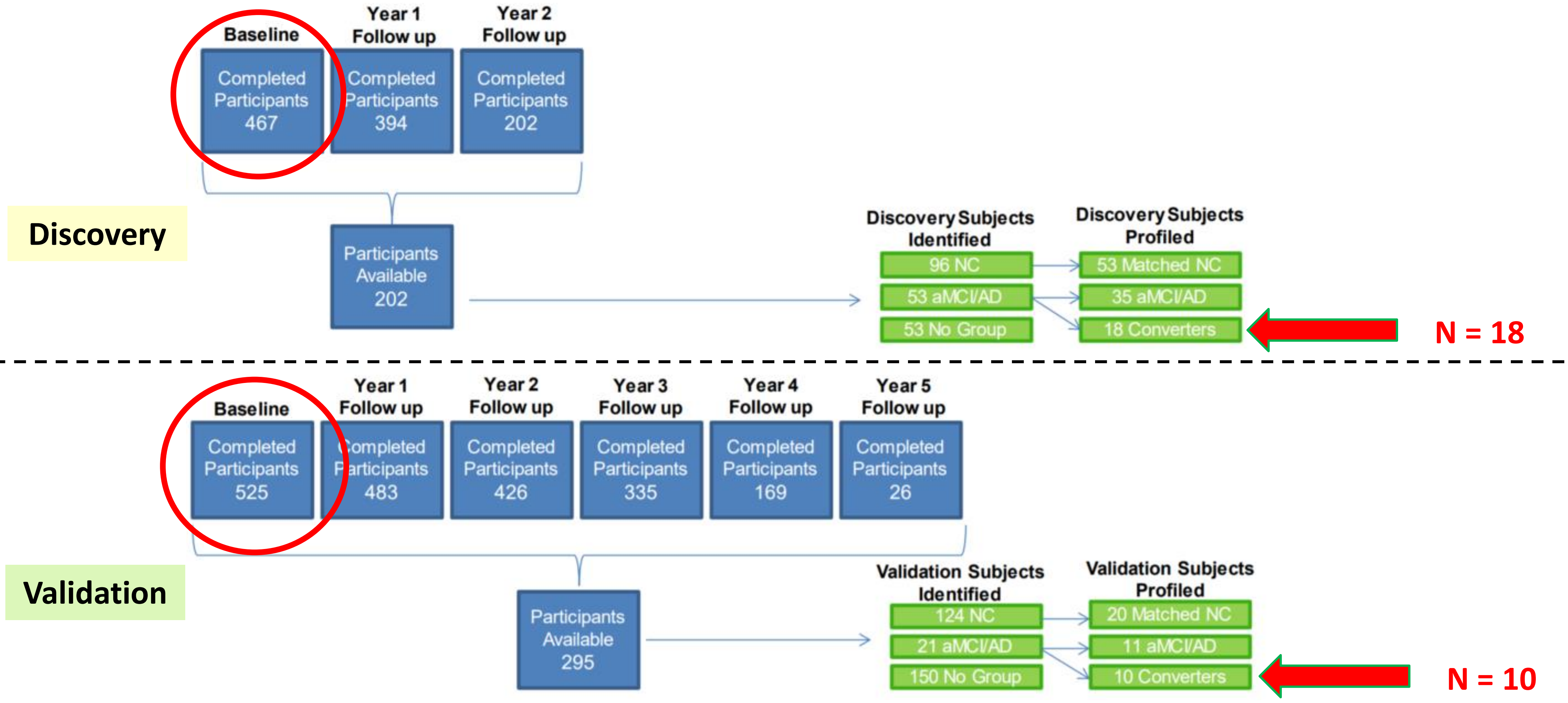
Bringing data to life.

The Rest of the Story
Part 1

Example #1
Predicting Alzheimer's Disease
Big Data - Biomarkers

Bringing data to life.

Patient Accounting



Bringing data to life.

Patient Accounting

467 volunteers

- Discovery Phase: 202 participants available
  - 149 met certain criteria for inclusion in the analysis
    - 53 with aMCI/AD
    - 96 normal
      - 18 converted from

Selection Bias ?!?

525 volunteers

- Validation Phase: 295 participants available
  - 144 met criteria for inclusion in the analysis
    - 21 with aMCI/AD
    - 124 normal
      - 10 converted from normal to aMCI/AD

Bringing data to life.

The Statistical Analytical Methods

Bringing data to life.

The Biological Analytical Methods

The actual data that was analyzed

- Sample storage and handling
- Sample storage time is confounded with groups

187 proteins analyzed

- Multiplicity !!!!

The Results – Part 2

July 2016

Blood metabolite markers of preclinical Alzheimer's disease in two longitudinally followed cohorts of older individuals

Ramon Casanova, Sudhir Varma, Brittany Simpson, Min Kim, Yang An, Santiago Saldana, Carlos Riveros, Pablo Moscato, Michael Griswold, Denise Sonntag, Judith Wahrheit, Kristaps Klavins, Palmi V. Jonsson, Gudny Eiriksdottir, Thor Aspelund, Lenore J. Launer, Vilmundur Gudnason, Cristina Legido Quigley, Madhav Thambisetty

RESULTS:
We failed to replicate these findings in a substantially larger study from two independent cohorts-the Baltimore Longitudinal Study of Aging ([BLSA], n = 93, AUC = 0.642, sensitivity/specificity of 51.6%/65.7%) and the Age, Gene/Environment Susceptibility-Reykjavik Study ([AGES-RS], n = 100, AUC = 0.395, sensitivity/specificity of 47.0%/36.0%). In analyses applying machine learning methods to all 187 metabolite concentrations assayed, we find a modest signal in the BLSA with distinct metabolites associated with the preclinical and symptomatic stages of AD, whereas the same methods gave poor classification accuracies in the AGES-RS samples.

Bringing data to life.

Example #2
Public Health
Big Data – Social Media

Social Media in Action

14 Mar 2014

BIG DATA

The Parable of Google Flu: Traps in Big Data Analysis

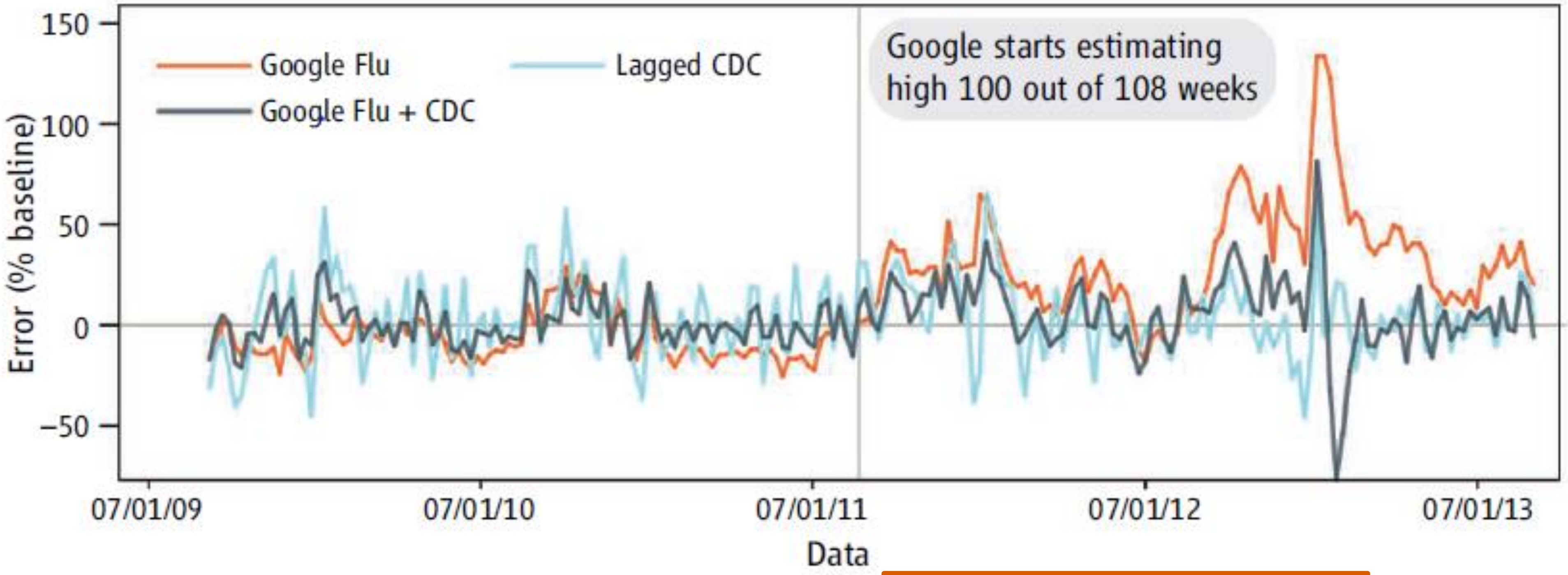
David Lazer,[1,2*] Ryan Kennedy,[1,3,4] Gary King,[3] Alessandro Vespignani[5,6,3]

Large errors in flu predictions were largely avoidable, which offers lessons for the use of big data.

Bringing data to life.

Social Media in Action

⟸ **Models built on data from 2003-2008.**



Predictions become worse over time. ⟹

What Is a Data Scientist?

uncertainty

probability

variability

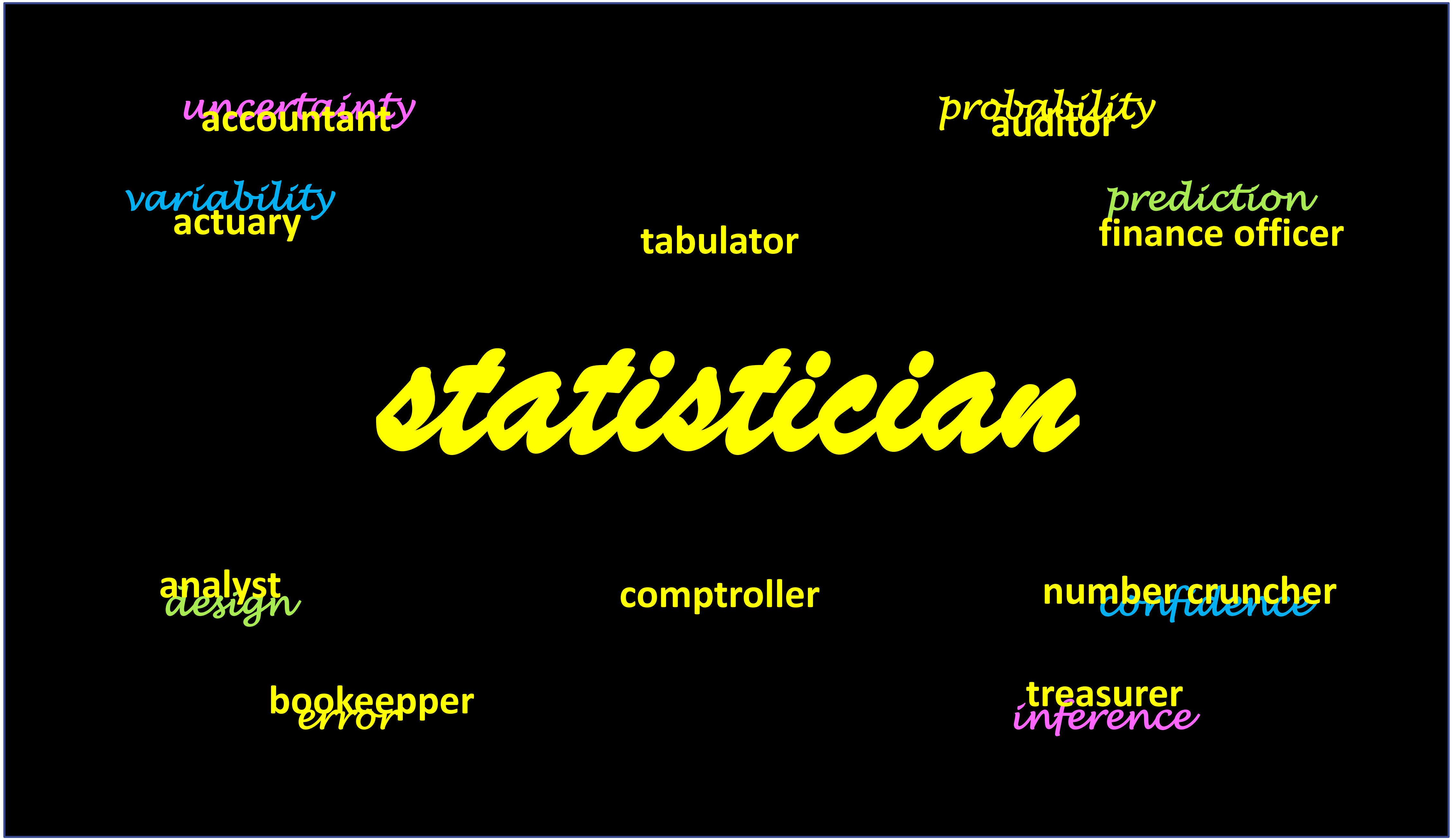
prediction

design

confidence

error

inference


uncertainty
accountant
probability
auditor
variability
actuary
tabulator
prediction
finance officer
statistician
analyst
design
comptroller
number cruncher
confidence
bookeepper
error
treasurer
inference

What Is a Data Scientist?

Thinking Like A Data Scientist - Part I[1]

The goal of the "thinking like a data scientist" process is to identify, brainstorm and/or uncover new variables that are better predictors of business performance.

Bill Schmarzo (1/5/2015)

Refined Thinking like

s[2]

!?!?

Data science is about identify

at might be better predictors of performance.

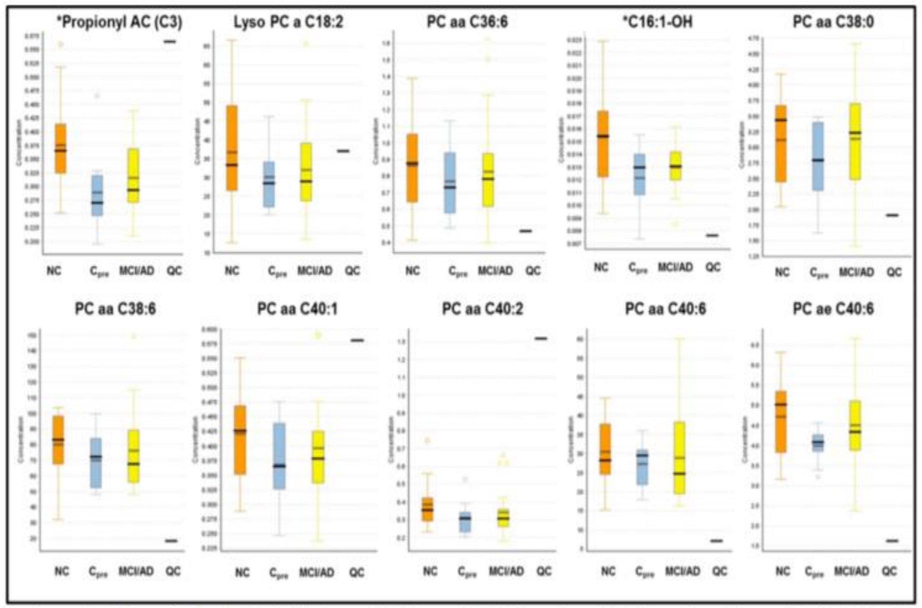
Bill Schmarzo 12/25/2017

[1] http://www.grroups.com/blog/thinking-like-a-data-scientist-part-i

[2] https://infocus.dellemc.com/william_schmarzo/refined-thinking-like-a-data-scientist-series/

Data Scientists and Statisticians

FOCUS ON THE ANSWER



FOCUS ON THE PROCESS / METHOD

Bringing data to life.

The Basics of Statistical Thinking
The Statistical Analysis Process

(that scientists can understand)

Bioanalytical Method Development

Sample Prep

Spiked Samples
Extraction
Same Matrix
etc.

Unknown Samples

LC-MS-MS

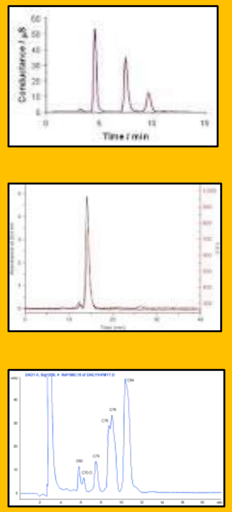
Lens Voltage

+/- Ion Mode

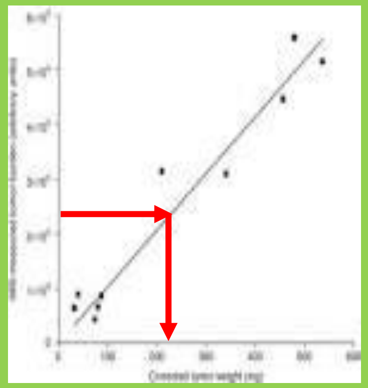
Collision Energy

Parameters for Tuning the Method

Outputs



Calibration



Goal

Understand the operating characteristics of the *whole system* so that results are trustworthy.

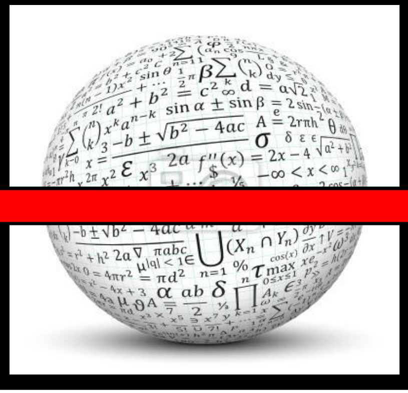
Statistical Method Development

Simulated Data

- ❖Known biomarkers
- ❖Effect size
- ❖Variability
- ❖Subgroup size

Data with Unknown Effects

Algorithms



Searching Algorithm

Variable Selection

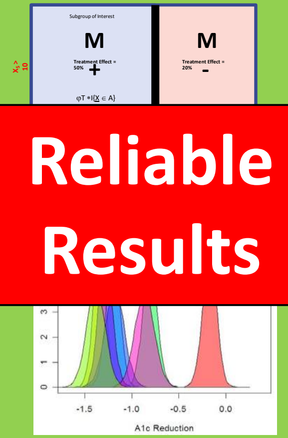
Model Complexity

Parameters for Tuning the Method

Outputs

- ❖ Biomarker defined subgroup
- ❖ Unadjusted statistical significance
- ❖ Effect size estimate

"Calibration"



Reliable Results

- ❖Compare to truth
- ❖Adjust significance
- ❖Correct for bias

Goal

Understand the operating characteristics of the *whole system* so that results are trustworthy.

Data Scientists and Statisticians

MULTIVARIATE **P-VALUES** **MULTIPLICITY** **BAYES**

Circulation: Cardiovascular Genetics

American Heart Association

Original Article 15 Jan 2015

Pharmacogenomic Determinants of the Cardiovascular Effects of Dalcetrapib

Jean-Claude Tardif[1*], Éric Rhéaume[1], Louis-Philippe Lemieux Perreault[2], Jean C. Grégoire[1], Yassamin Feroz Zada[2], Géraldine Asselin[2], Sylvie Provost[2], Amina Barhdadi[2], David Rhainds[3], Philippe L. L'Allier[1], Reda Ibrahim[1], Ruchi Upmanyu[4], Eric J. Niesor[4], Renée Benghozi[4], Gabriela Suchankova[4], Fouzia Laghrissi-Thode[5], Marie-Claude Guertin[6], Anders G. Olsson[7], Ian Mongrain[2], Gregory G. Schwartz[8] and Marie-Pierre Dubé[9]

Conclusions—The effects of dalcetrapib on atherosclerotic outcomes are determined by correlated polymorphisms in the ADCY9 gene.

100 potential biomarkers

- Prior probability of success (H_0 is false) = **0.20**
- Prior on H_0 is true (none are predictive) = 0.80
- Uniform prior per biomarker = **0.20**/**100** = **0.002**

Observed p-value = **0.0001** for one biomarker

- Bonferroni adjusted p-value $\leq$ **100** * **0.0001** = **0.01**

Bayesian posterior pr(H_0 is false) $\leq$ **0.44.**

Berger, JO, Wang X, Shen L (2014) A Bayesian Approach to Subgroup Identification, Journal of Biopharmaceutical Statistics, 24:1, 110-129.

Bringing data to life.

Data Scientists and Statisticians

<u>MULTIVARIATE</u>

"Give me a big enough data set, and I guarantee that I can find the patterns in it."

Prominent Data Science Researcher
Distinguished Professor
Major US University

<u>MULTIPLICITY</u>

"Torture the data long enough, and they will confess to anything."

Sir Ronald Coase
Nobel Prize in Economics
1991

Bringing data to life.

Data Scientists and Statisticians

BIG DATA

Google Flu Trends

Observation	Influenza Like Illness	Doctor Visit	Search Term 1	Search Term 2	Search Term 3	…	Search Term 50,000,000
1	Y	Y	Y	N	N		Y
2	N	N	Y	N	N		N
3	Y	N	N	N	Y		Y
…							
1152	N	N	Y	N	N		Y

Large p, small N problem

- As p gets large, more (useless?) correlations emerge
- An N gets large, everything becomes significant

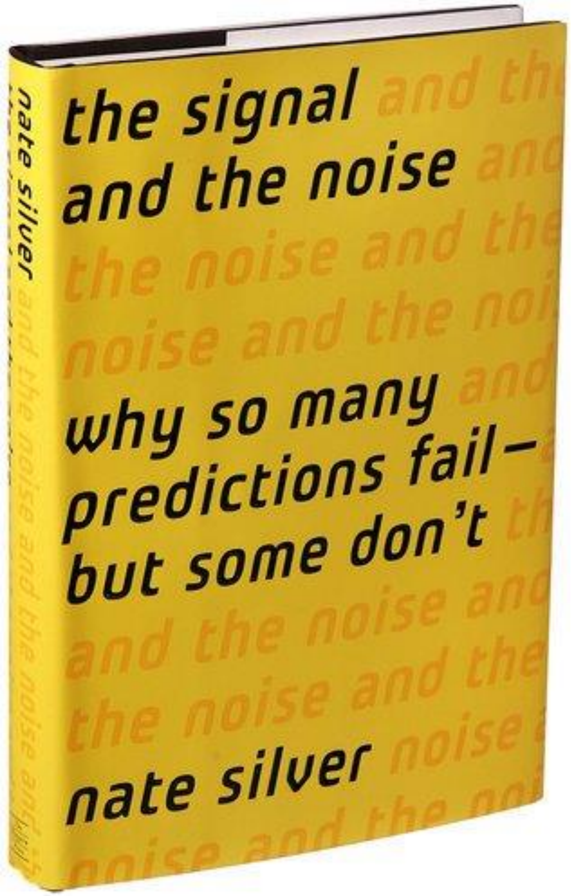
RIGHT DATA

"If you are trying to find a needle in a haystack, the worst thing you can do is add more hay."

Steve Ruberg
Your Run-of-the-Mill Statistician
July, 2015

Bringing data to life.

More Information, More Problems



"We face danger whenever information growth outpaces our understanding of how to process it. … it can still take a long time to translate information into useful knowledge, and that **if we are not careful, we may take a step back** in the meantime."

Data Scientists and Statisticians

Big Data is Dead

Alexander Thamm

November 23, 2017

"Big Data, Small Data, Little Data, Fast Data, and Smart Data are all 'Just Data.'"

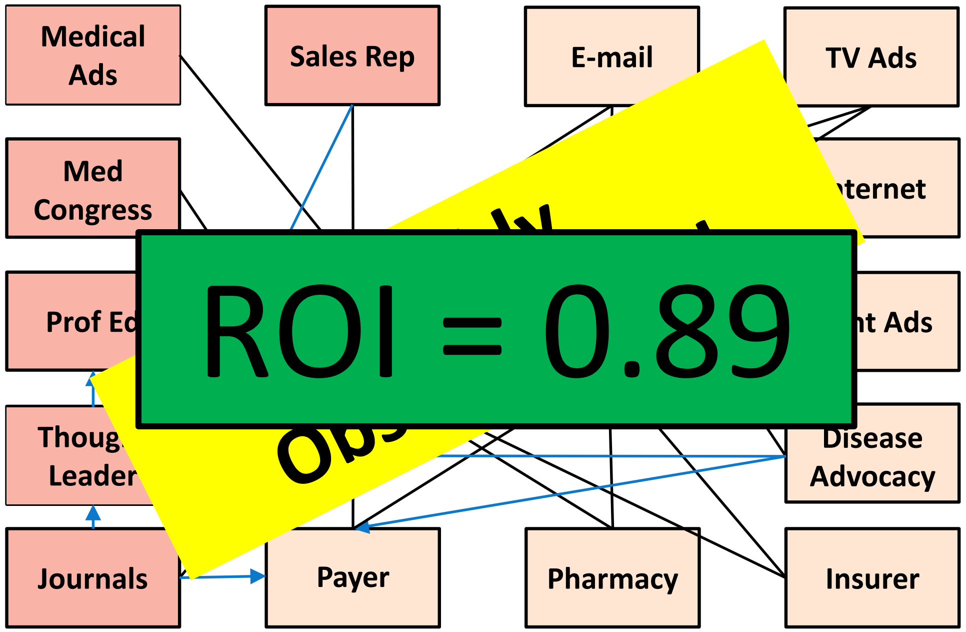
" … forecast quality does not depend on the volume of the data. Just Data means that, above all, the **right data** needs to be incorporated into analysis."

https://www.linkedin.com/pulse/big-data-dead-just-regardless-quantity-structure-speed-thamm

Bringing data to life.

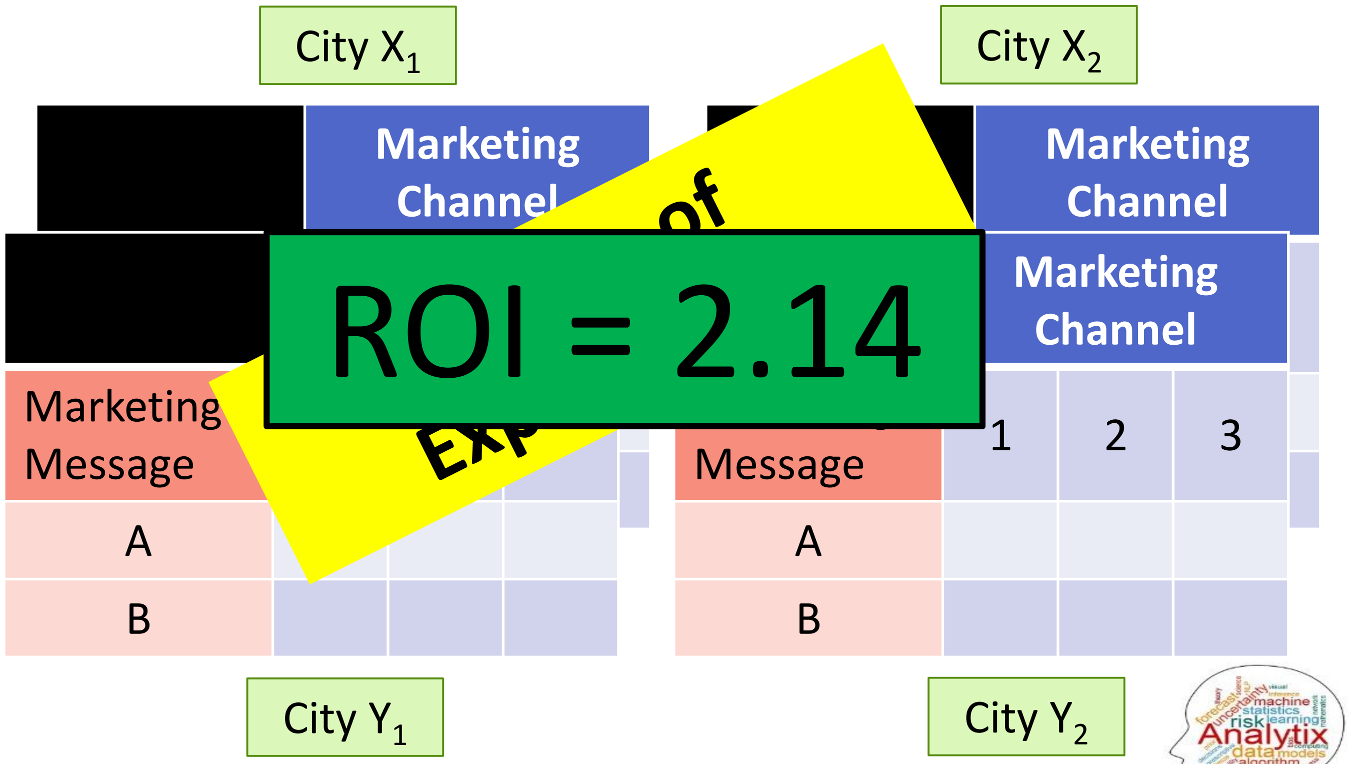
Data Scientists and Statisticians

BIG DATA



Prescription Ecosystem (US)

RIGHT DATA



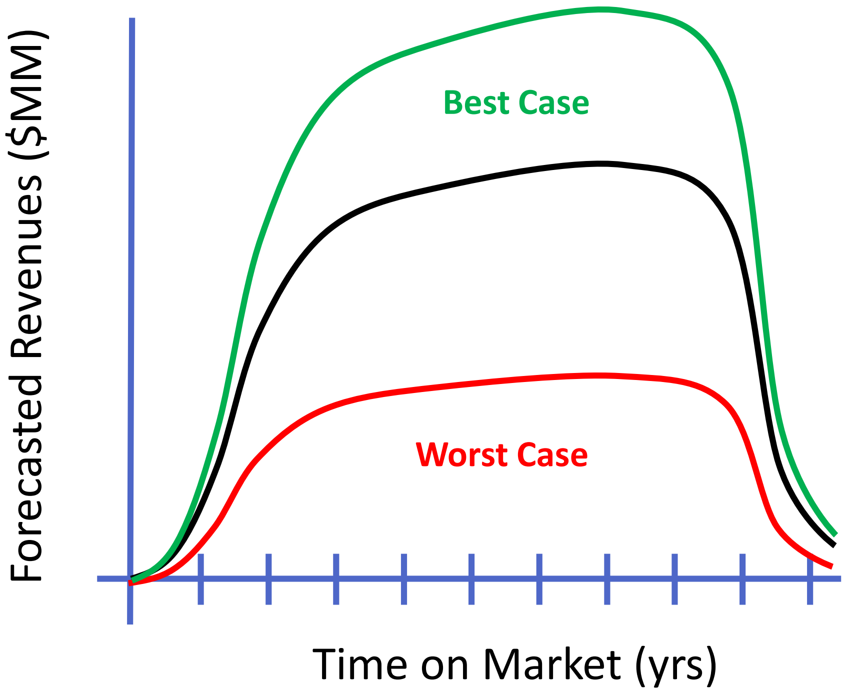
Bringing data to life.

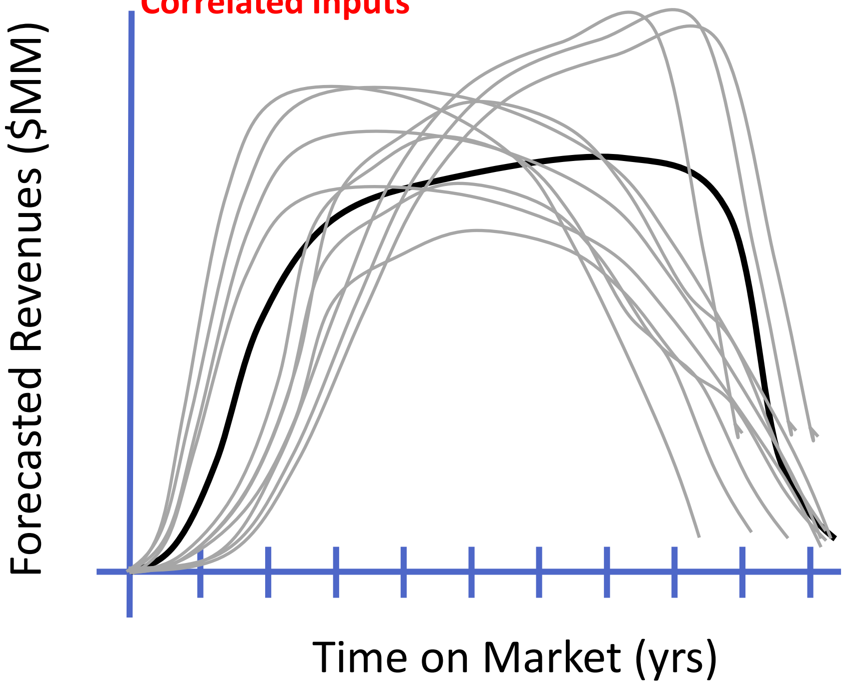
Data Scientists and Statisticians

Y = F(X) **MODEL**

Y = F(X) + ε **SIMULATE**

Uncertainty
Correlated Inputs

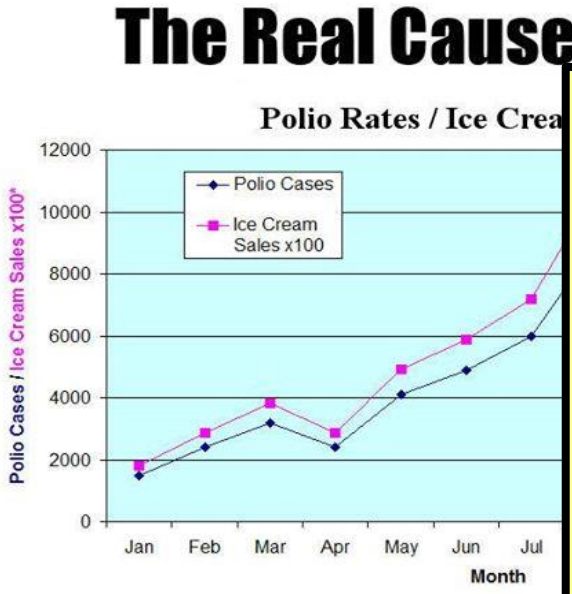




Bringing data to life.

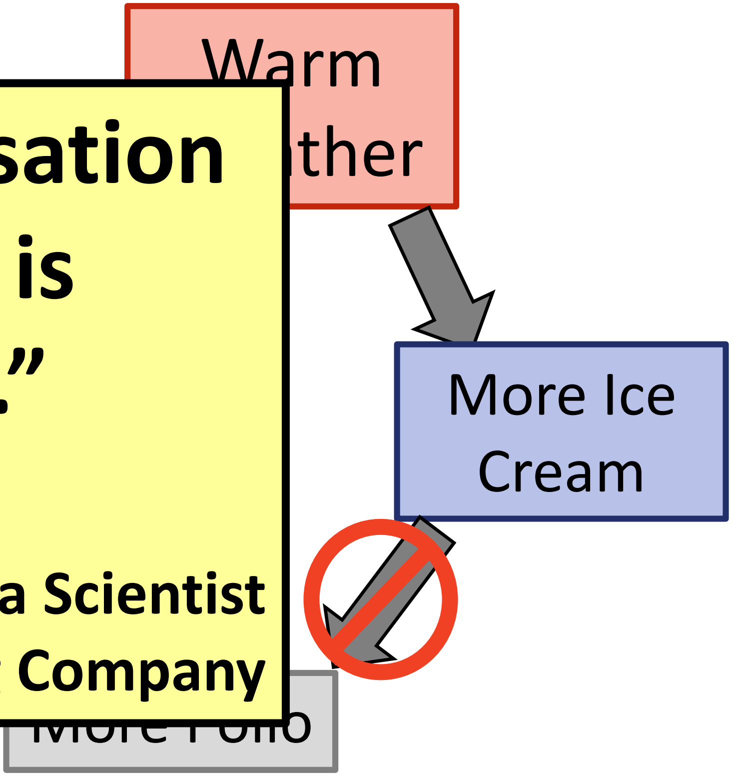
Data Scientists and Statisticians

CORRELATION



Eliminating ice cream was as part of an anti-polio diet!

CAUSATION



“We do not need causation anymore. Correlation is enough with big data.”

Partner and Data Scientist
Large Business Consulting Company

Bringing data to life.

Data Scientists and Statisticians

ASSOCIATION

ASSOCIATION OF DISRUPTED CIRCADIAN RHYTHMICITY WITH MOOD DISORDERS, SUBJECTIVE WELLBEING, AND COGNITIVE FUNCTION: A CROSS-SECTIONAL STUDY OF 91 105 PARTICIPANTS FROM THE UK BIOBANK

LAURA M LYALL[1], PHD, CATHY A WYSE[2], PHD, … DANIEL J SMITH[1], MD

INSTITUTE OF HEALTH AND WELLBEING, UNIVERSITY OF GLASGOW, GLASGOW, UK
INSTITUTE OF BIODIVERSITY, ANIMAL HEALTH AND COMPARATIVE MEDICINE, UNIVERSITY OF GLASGOW, GLASGOW, UK

The Lancet, Psychiatry
May 15, 2018

CAUSATION?

FINDINGS

"… 91 105 participants with accelerometry data … reduction … was associated with increased risk of lifetime major depressive disorder (odds ratio [OR] 1·06, 95% CI 1·04–1·08),

lifetime bipolar disorder (1·11, 1·03–1·20),

greater mood instability (1·02, 1·01–1·04),

higher neuroticism scores (incident rate ratio 1·01, 1·01–1·02),

… "

INTERPRETATION

"Circadian disruption is reliably associated with various adverse mental health and wellbeing outcomes, including major depressive disorder and bipolar disorder. Lower relative amplitude might be linked to increased susceptibility to mood disorders."

Bringing data to life.

Data Scientists and Statisticians

Evidation Health

Digital Track...
Sleep...
St...
ISPO...
Interna...
Shefali Kumar, Tom Quisel, Lu...

...ople with Type 2

Mark Heyman, Jessie Juusola

Intra-day Activity Better Predicts Chronic Conditions

NIPS 2016 - Neural Information Processing Systems 2016

Tom Quisel, David Kale, Luca Foschini

...nsistent
...ed He...

Behavioral Phenotyping of Digital Health Tracker Data

SDM 2016 - 16th SIAM International Conference on Data Mining

Thomas Quisel, Luca Foschini, Alessio Signorini

Weight Gain: A Within-Subject

...f the Society of Behavioral Medicine

...Unit Readmission Prediction

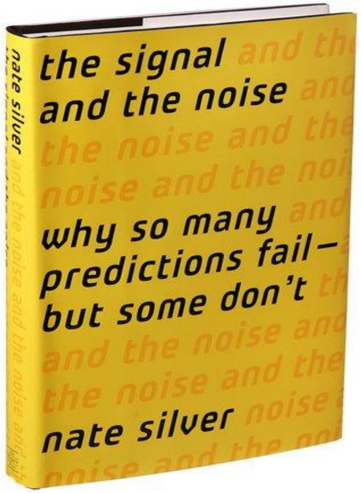
...CC 2017 - Journal of the American College of Cardiology, Vol. 69 No. 11

Tom Quisel, Richard Milani, Luca Foschini

Bringing data to life.

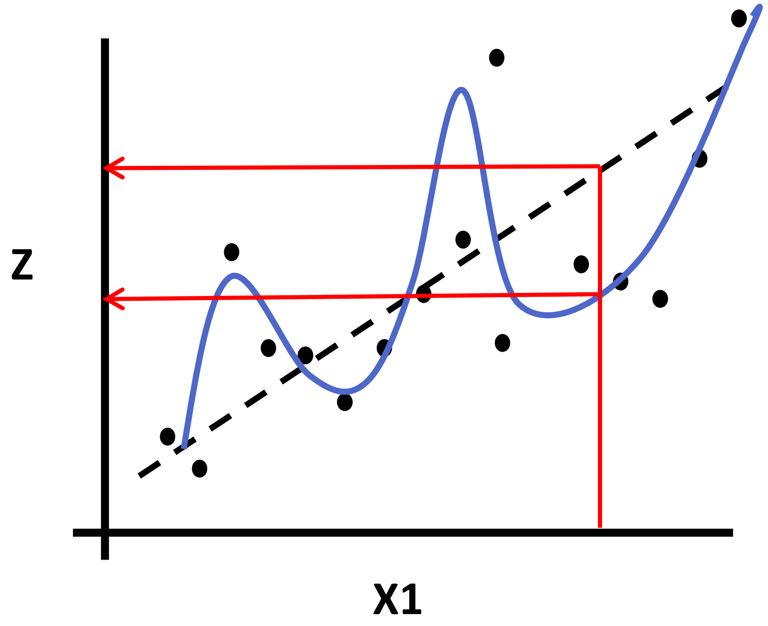
Data Scientists and Statisticians

OVER-FITTING



"The biggest problem you never heard of."

FITTING



"GFT overlooks considerable information that could be extracted by traditional statistical methods."

Science

Bringing data to life.

Data Scientists and Statisticians

DATA SCIENTISTS	STATISTICIANS
Applied – Good enough	Pure – Perfect
Cool problems	Cool math
Pictures – Holistic – Broad	Equations – Details – Caveats
Bold – Confident	Cautious – Skeptical
Communicators	Analyzers
Plentiful	Rare
Patterns – Associations (“might”)	**Causal effects**
Information	**Inference – Pr (Correct Decision)**
Big Data	**Right Data – Design of Experiments**

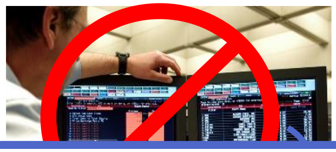
Bringing data to life.

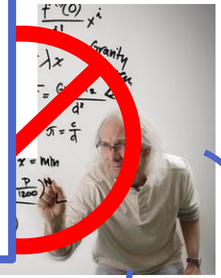
Data Scientists and Statisticians

DATA SCIENTISTS	**APPROACH**
Computer Scientists	“might be”
Epidemiologists	Associations
Pharmacokineticists	Y = f(X)
Outcomes Scientists	Associations
Bioinformaticists	Focus on WHAT not HOW
Business Analysts	Correlation
Econometrics	Y = f(X)

Bringing data to life.

Are Statisticians THE ANSWER?





"The future of statistics is bright. It is not as clear for statisticians."

Steve Ruberg and Bill Louv
JSM, 1991

S. J. Ruberg and W. C. Louv (1991) "The Statistician as Strategist." 1991 Proceedings of the Section on Statistical Education, American Statistical Association, 8-15.

Are Statisticians THE ANSWER?

Leading Disruptive/Complex Change

Vision →	Skills →	Incentives →	Resources →	Action Plan →	**CHANGE**
	Skills →	Incentives →	Resources →	Action Plan →	CONFUSION
Vision →		Incentives →	Resources →	Action Plan →	ANXIETY
Vision →	Skills →		Resources →	Action Plan →	GRADUAL CHANGE
Vision →	Skills →	Incentives →		Action Plan →	FRUSTRATION
Vision →	Skills →	Incentives →	Resources →	→	FALSE STARTS

Without adequate Communication of each of these elements, it is the same as not having the element present.

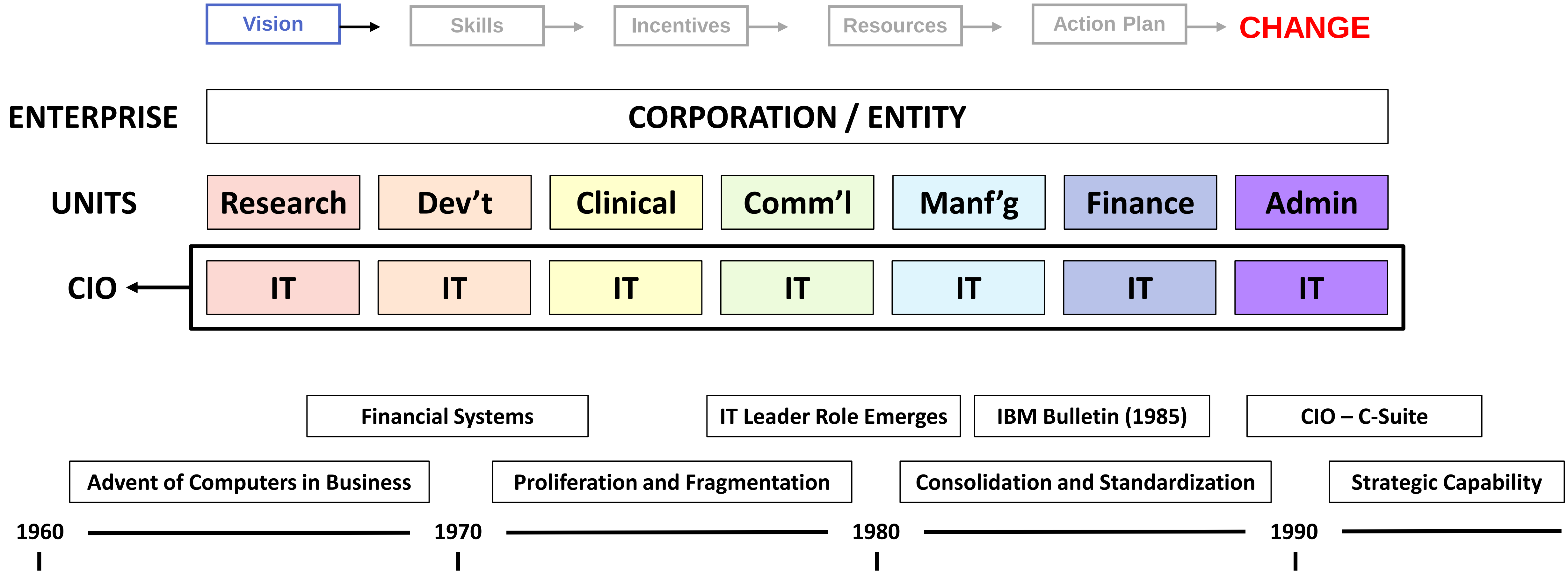
Leading Disruptive/Complex Change

Vision → Skills → Incentives → Resources → Action Plan → CHANGE

"A **vision** is a compelling image of an achievable future."

Laura Berman Fortgang

Leading Disruptive/Complex Change



Leading Disruptive/Complex Change

Vision → Skills → Incentives → Resources → Action Plan → **CHANGE**

ENTERPRISE | **CORPORATION / ENTITY**

UNITS | Research | Dev't | Clinical | Comm'l | Manf'g | Finance | Admin

CIO ← IT | IT | IT | IT | IT | IT | IT

R&D / Clinical Systems (Stats)

Analytics Role Emerges

Computer World (2012)*

CAO – C-Suite?

Advent of Analytics in Business

Proliferation and Fragmentation

Consolidation and Standardization

Strategic Capability

2000 — 2005 — 2015 — 2020

*https://www.computerworld.com/article/2492218/big-data/time-has-come-for-chief-analytics-officers.html

Leading Disruptive/Complex Change

Vision → Skills → Incentives → Resources → Action Plan → CHANGE

ENTERPRISE	CORPORATION / ENTITY						
UNITS	Research	Dev't	Clinical	Comm'l	Manf'g	Finance	Admin
CIO ←	IT	IT	IT	IT	IT	IT	IT
CAO ←	Stats+	Stats+	Stats+	Stats+	Stats+	Stats+	Stats+

Stats+ = Analytics = {Stats, OR, Applied Math, Econ, Comp Sci, Epidemiology, PK, Health Outcomes}

Communication: Focus and Alignment … + … Relentless pursuit and messaging

Leading Disruptive/Complex Change

Vision → **Skills** → Incentives → Resources → Action Plan → **CHANGE**

✓Technical Skills

Soft Skills

- **Leadership**
- Communication, negotiation, influence
- Practical, applied
- Subject matter acumen

How to get soft skills in school?

- Undergraduate training

Communication: Internal programs offered consistently

Leading Disruptive/Complex Change

Vision → Skills → **Incentives** → Resources → Action Plan → **CHANGE**

Show me the money!

- Reduce cost
- Speed cycle time
- Increase probability of success
- Change strategy, influence a decision

"I will give you 2 statisticians for 6 months …"

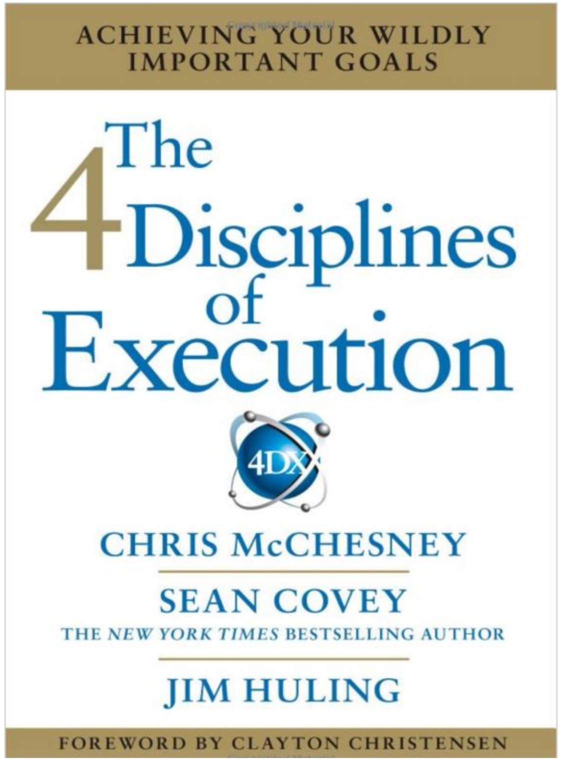
"Can you do that?"

Communication: Create metrics and tell the success stories → Management; internal publications; seminars …

Bringing data to life.

Leading Disruptive/Complex Change

Vision → Skills → Incentives → **Resources** → Action Plan → **CHANGE**



1. Focus on the Wildly Important

Initially … squeeze

- Prioritize, Synergize, Stop dabbling

Then re-assign people

- Experienced + newer

Link cross-functional resources

Communication: Workshops, Summits, Partnerships across functions … Explain what will NOT get done

Bringing data to life.

Leading Disruptive/Complex Change

Vision → Skills → Incentives → Resources → **Action Plan** → **CHANGE**

Need a Playbook

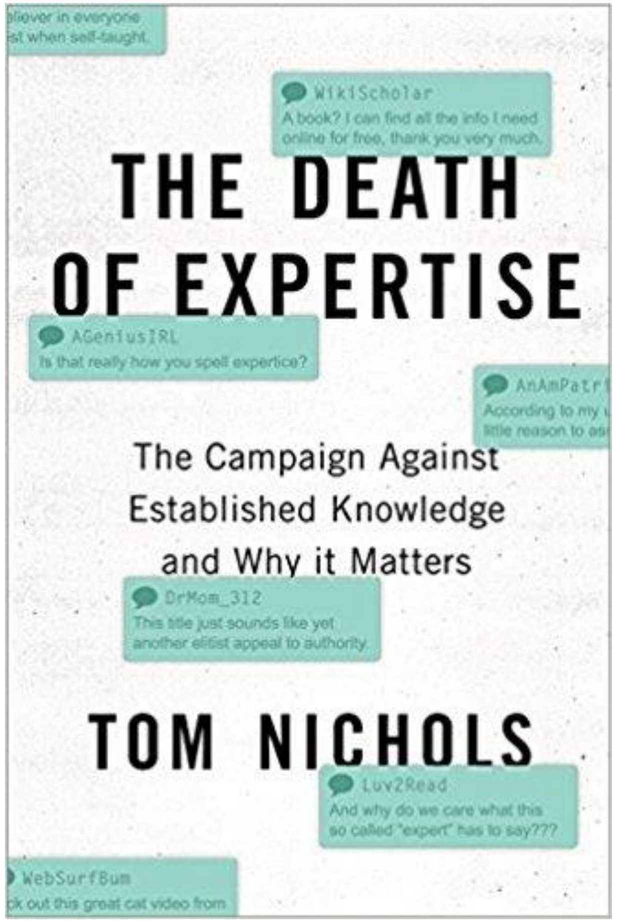
1. Define capabilities that align with business goals
2. **DEDICATE** resources (remember the squeeze)
3. Portfolio of projects / use cases / initiatives
4. Initially … quick wins
5. Metrics, story-telling
6. See #3 – rinse and repeat

Communication: Town Halls, Functional and Corporate forums

Bringing data to life.

DISCUSSION

DISCUSSION



Data Science Revolution

Fast answers

Superficial answers

Sans Uncertainty / Probability

Incorrect answers

Bias, Over-fitting, Multiplicity …

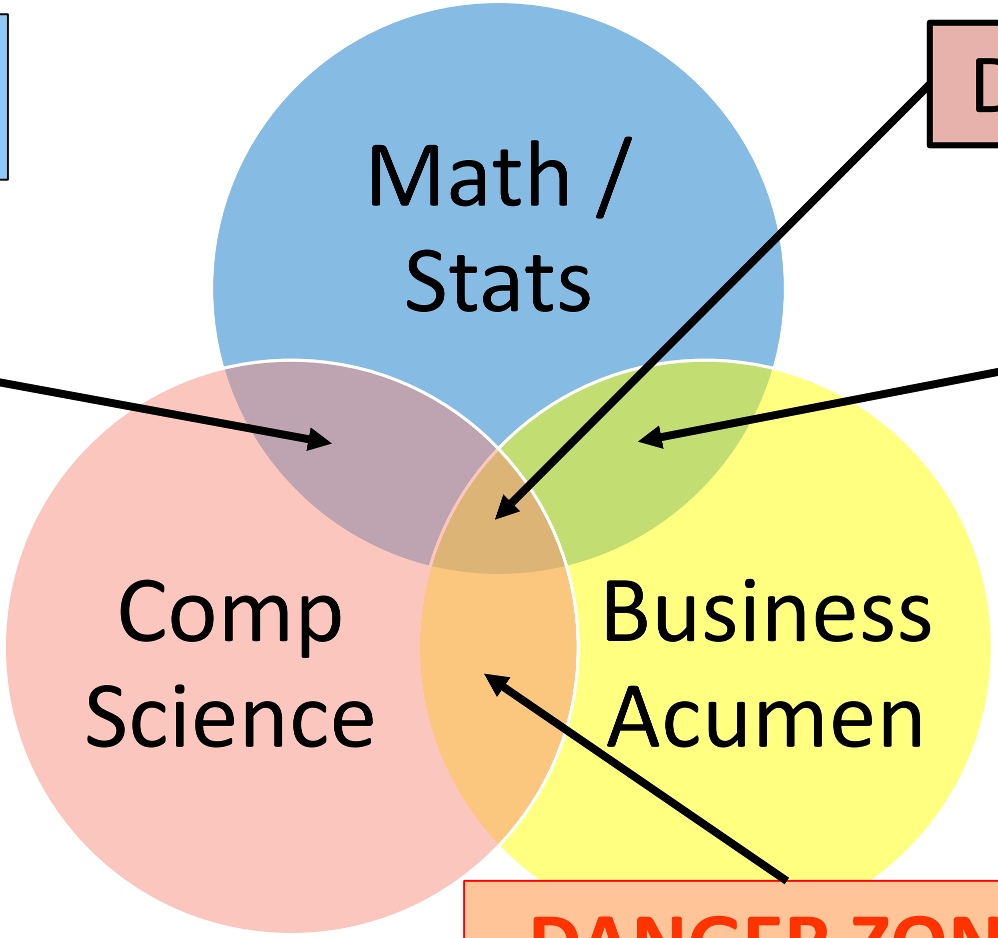
DISUCSSION

Math/Stats – quantitative knowledge and capabilities

Data Science !!!

Machine Learning

Traditional Research



Comp Sci – programming, data manipulations, algorithms, HPC

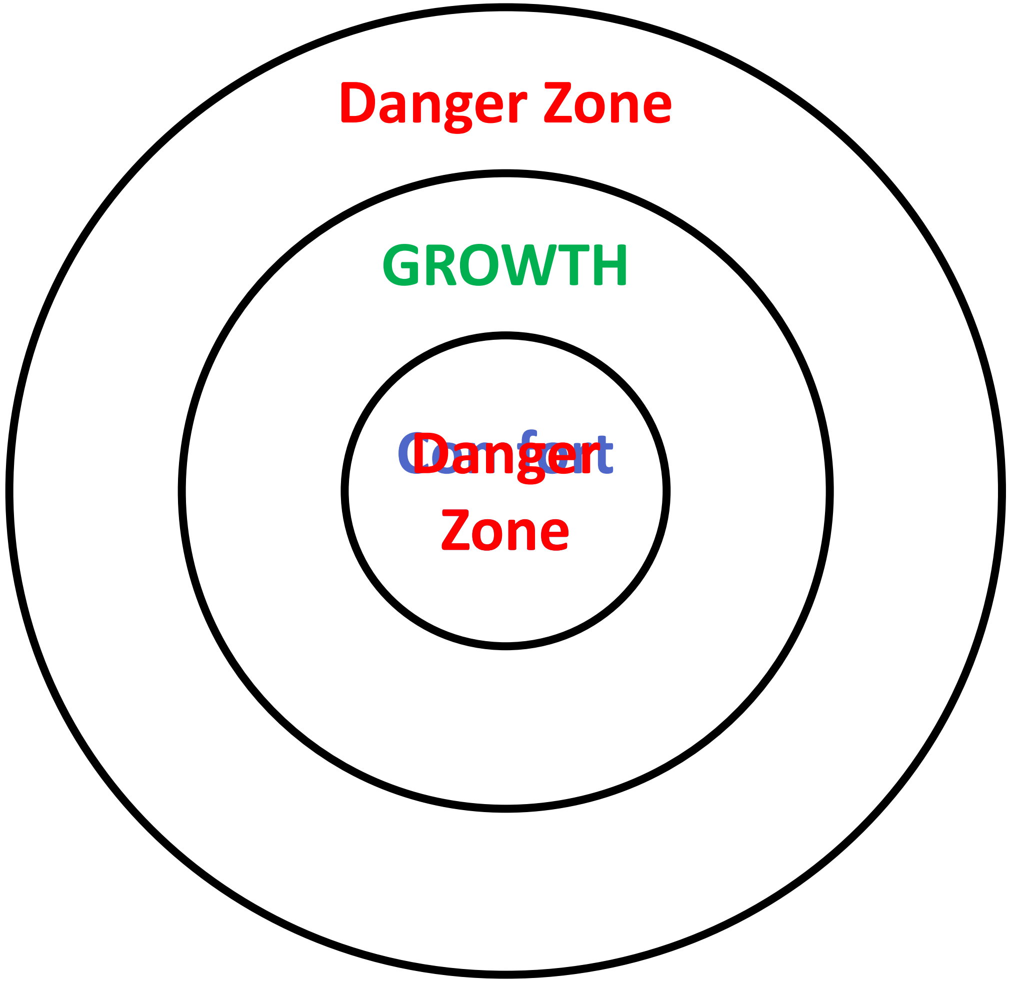
Bus Acumen – subject matter expertise

DANGER ZONE

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

Bringing data to life.

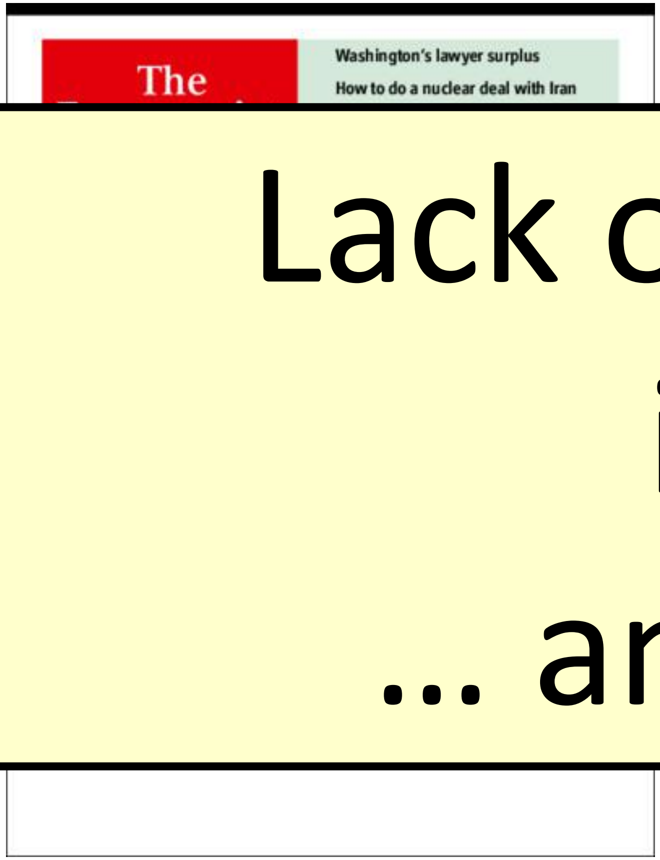
DISCUSSION



from Jessica Hagy, *Indexed*

Bringing data to life.

DISCUSSION



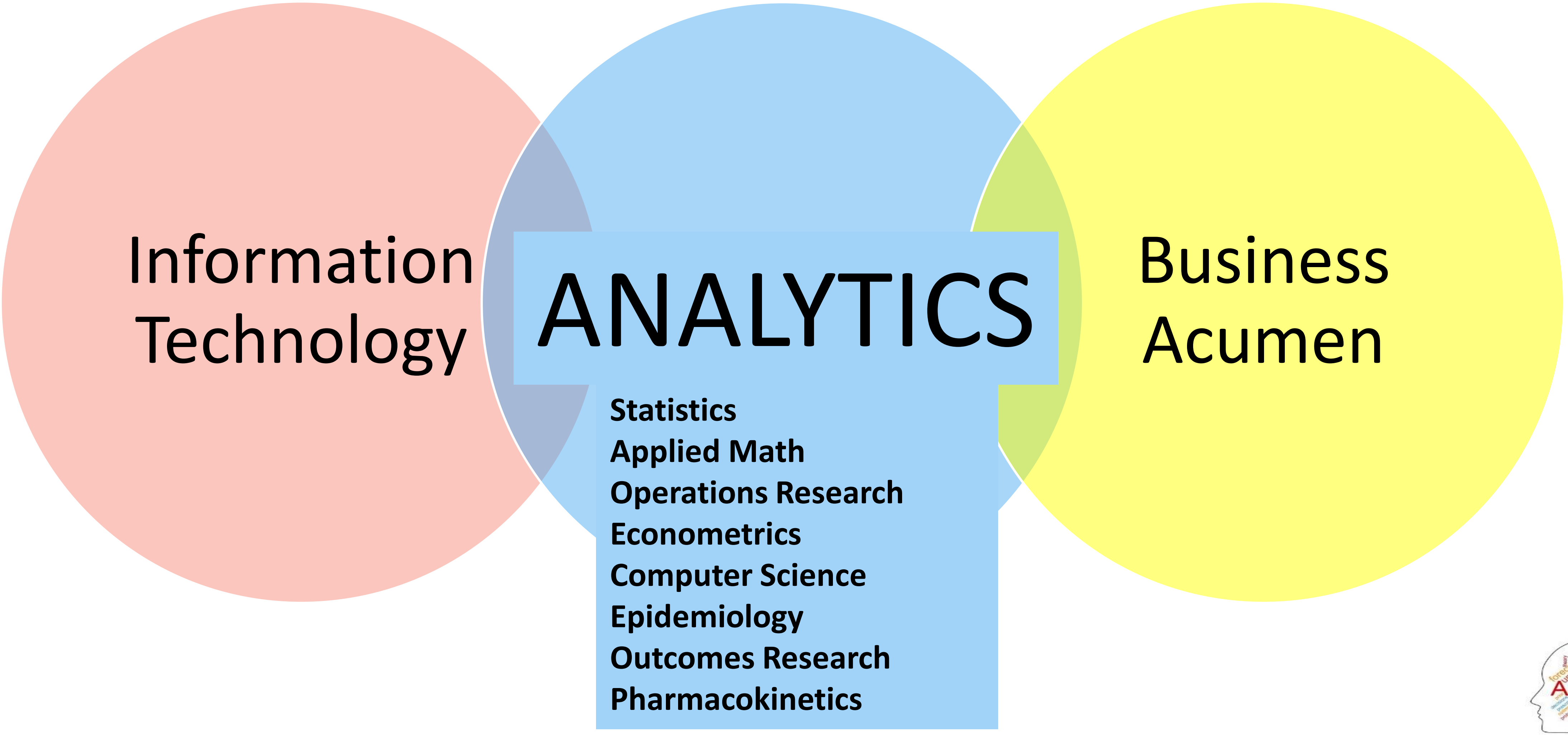
October 21, 2013

“Models which can be ‘tuned’ in many

Lack of reproducibility
in research
… and in business?

Bringing data to life.

DISCUSSION



Bringing data to life.

CONCLUSION

We (Stats) have we ... ore ...

Data Mining

(you nar

Informa

Machine Learning

...

... but not capitalize ... entum.

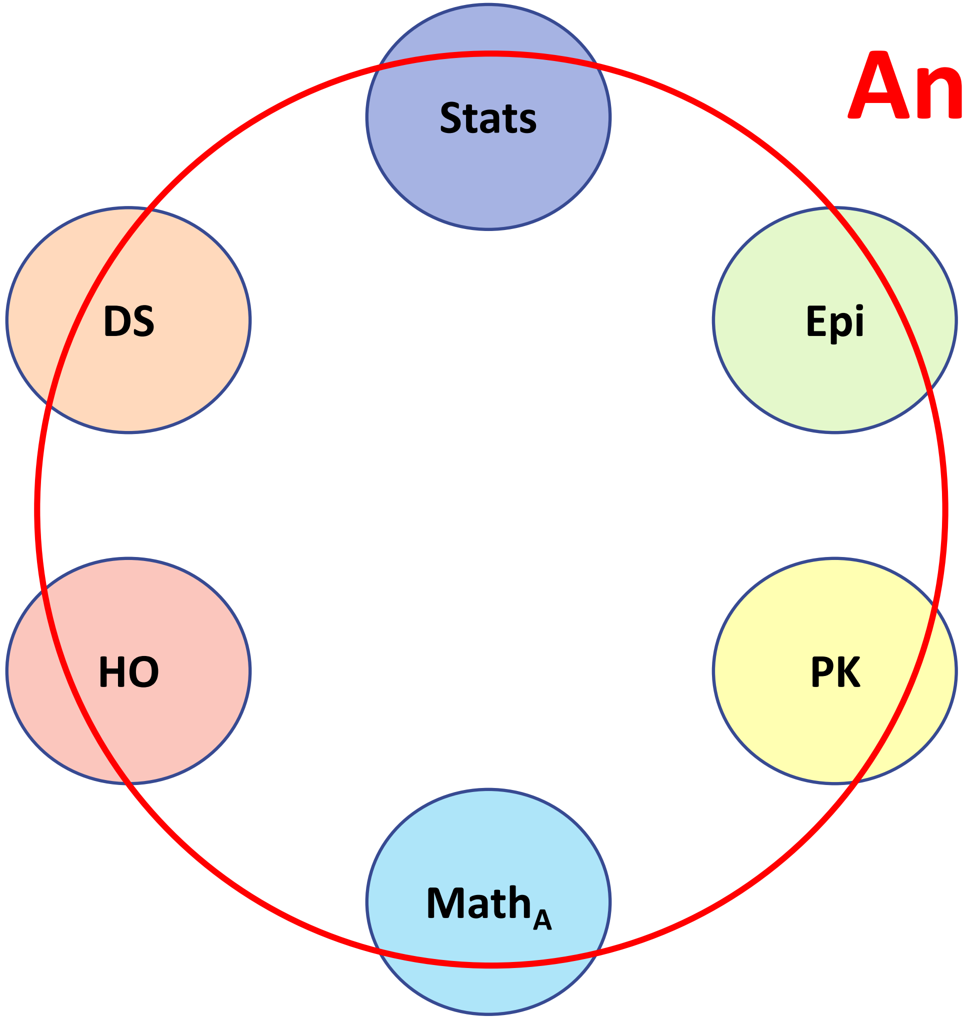
Bringing data to life.

CONCLUSION

Statistics	Analytics
American Statistical Association	International Association for Data and Analytics
Statisticians	Analyticians?

Bringing data to life.

CONCLUSION



Analytics Community

"There is no progress without conflict."

George Bernard Shaw

Bringing data to life.

CONCLUSION

Vision | Skills | Incentives | Resources | Action Plan

John Schaar

Bringing data to life.

Grazie!

ありがとうございました!

Gracias!

Спасибо!

धन्यवाद!

Thank you!

Dank je!

Danke!

Tack!

Je vous remercie!

Obrigado!

謝謝!