ASA Biopharmaceutical Section

Webinar

4 February 2020

Subgroup ~~Analysis~~ Identification

The Hardest Problem There Is

Stephen J. Ruberg, PhD
President
Analytix Thinking, LLC
AnalytixThinking@gmail.com
AnalytixThinking.blog

Bringing data to life.

Acknowledgement

My thinking on this problem has been significantly influenced by colleagues and collaborators, most notably

Lilly

Lei Shen

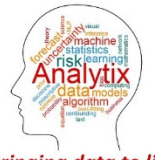
Rick Higgs

Ilya Lipkovich

Novartis

Bjoern Bornkamp

Mark Baillie



Bringing data to life.

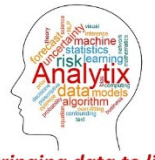
Introduction

Thinking about subgroup identification

- **WHAT** are we doing?
- **HOW** should we do it?

NOT the best subgroup ID method

Goal: Reliable, Credible, Actionable Inference



Bringing data to life.

Alice in Wonderland

Alice: "Would you tell me, please, which way I ought to go from here?"

Cheshire Cat: "That depends a good deal on where you want to get to."

Alice: "I don't much care where—"

Cheshire Cat: "Then it doesn't matter which way you go."

Alice: "—so long as I get SOMEWHERE."

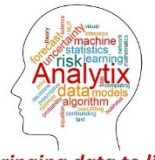
Cheshire Cat: "Oh, you're sure to do that, if you only walk long enough."

Lewis Carroll

Bringing data to life.

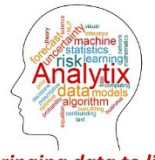
Outline

1. General Context
2. Prognostic biomarkers
    - A. Predicting Alzheimer's Disease
    - B. Predicting Acute Kidney Injury
3. Predictive Biomarkers
    - A. An Open Challenge
    - B. Disciplined Subgroup Search
    - C. An Oncology Example
        - Heterogeneity and Homogeneity of Response
    - D. Bayesian Thinking
4. Conclusion



Bringing data to life.

1. General Context



Bringing data to life.

General Context

Question: Is there a (sub)group of patients (M+) who can be identified by some measurable characteristics (i.e. biomarkers*) that have, on average, an exceptional response** compared to those patients in the complementary (sub)group (M-)?

*biomarkers can be phenotypic, genotypic, genomic, …

**exceptional response implies clinically meaningful
usually efficacy, but could be safety

Bringing data to life.

General Context

EMPIRICAL

Tailored Therapeutics

Discovering a subgroup

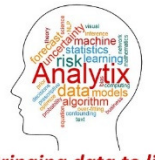
NOT MECHANISTIC

NOT Personalized medicine - CAR-T cell therapies

Kymriah®, Yescarta®, …

NOT gene therapy - Known genetic mechanism

Luxterna®, Zolgensma® …



Bringing data to life.

General Context

CLARIFYING GOALS

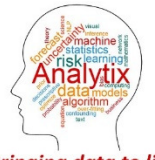
Do you want to find a subgroup or not?

YES – Heterogeneity is my friend!

- I want to find a targeted therapeutic!

NO – Heterogeneity is my enemy!

- I want the treatment effect to be homogeneous across subgroups.



Bringing data to life.

General Context

BIOMARKERS

A single biomarker

Sometimes plausible … sometimes not so much

A handful of biomarkers (i.e. biomarker signature)

Perhaps some combination of 2-3 biomarkers

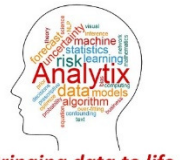
A (linear?) combination of many biomarkers

MammaPrint[1] (unsupervised learning)

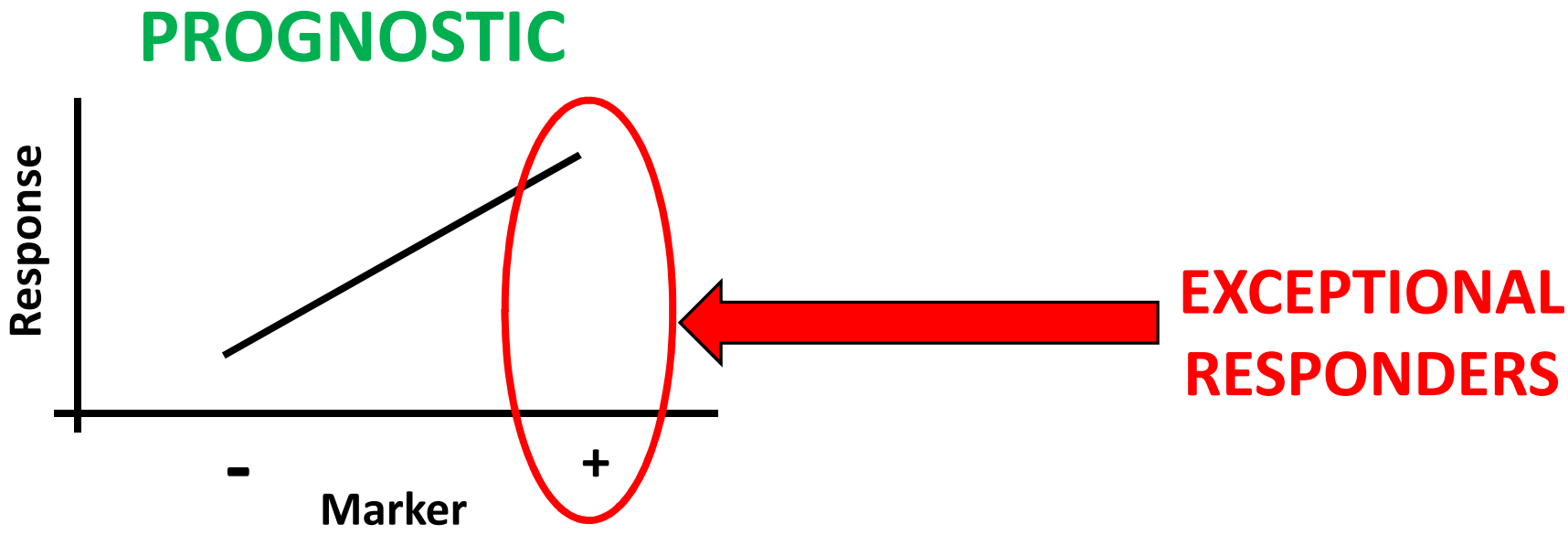
Enabled by “machine learning”

Plausibility? Overfitting?

[1]van 't Veer LJ, Dai H, van de Vijver MJ, et al. (2002). "Gene expression profiling predicts clinical outcome of breast cancer". *Nature*. **415** (6871): 530–6. doi:10.1038/415530a. hdl:1874/15552. PMID 11823860.



General Context



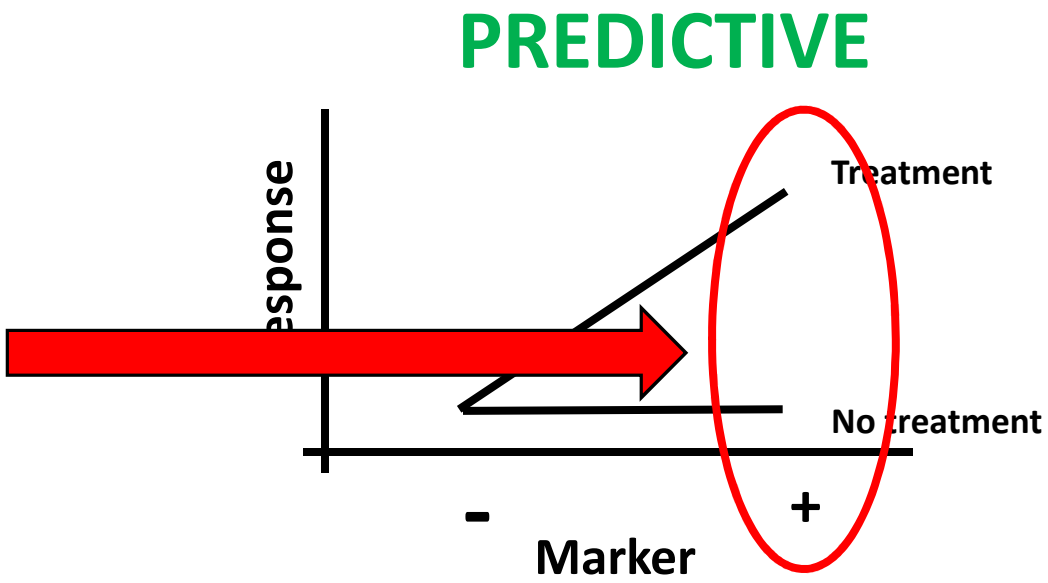
A biomarker or biomarker signature that identifies different groups of patients with respect to the risk of an outcome of interest ***in the absence of treatment***

Who is at risk?

- Diagnostic
- Who/when to intervene
- Enrollment in clinical trials

General Context

EXCEPTIONAL RESPONDERS



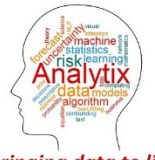
Who gets what treatment?

- Who to treat?
- Enrollment in clinical trials

A biomarker or biomarker signature that identifies different groups of patients with respect to the outcome of interest in ***response to a particular treatment***

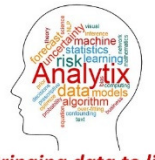
2. Prognostic Biomarker

Finding Heterogeneity



Bringing data to life.

2A. Predicting Alzheimer's Disease



Bringing data to life.

Predicting Alzheimer's Disease

Problem Statement

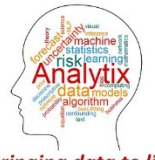
No good treatments for Alzheimer's Disease

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

Detecting it early - key to treatment or prevention

Current imaging approaches - expensive and invasive

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

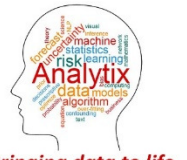


Bringing data to life.

Predicting Alzheimer's Disease

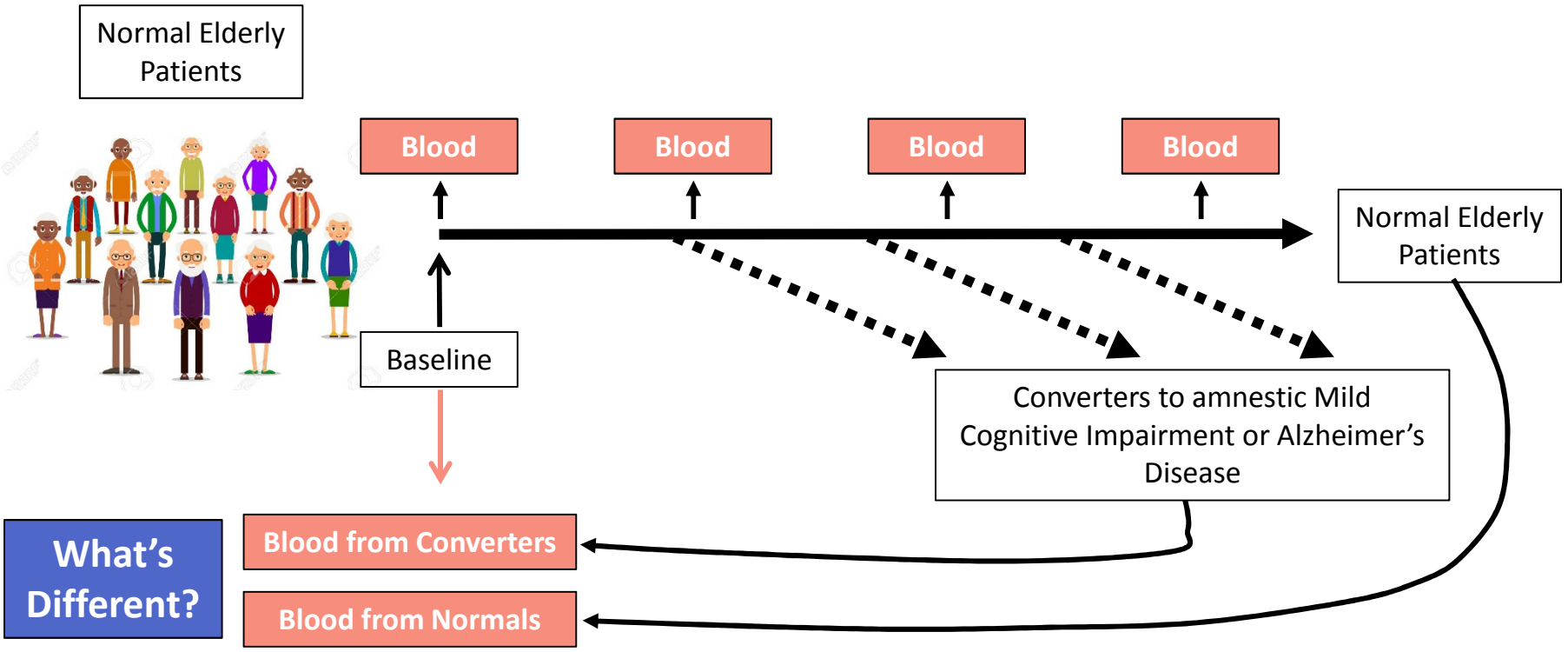
Study Outline

- Select cognitively normal elderly patients
- Collect blood samples at baseline and 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.

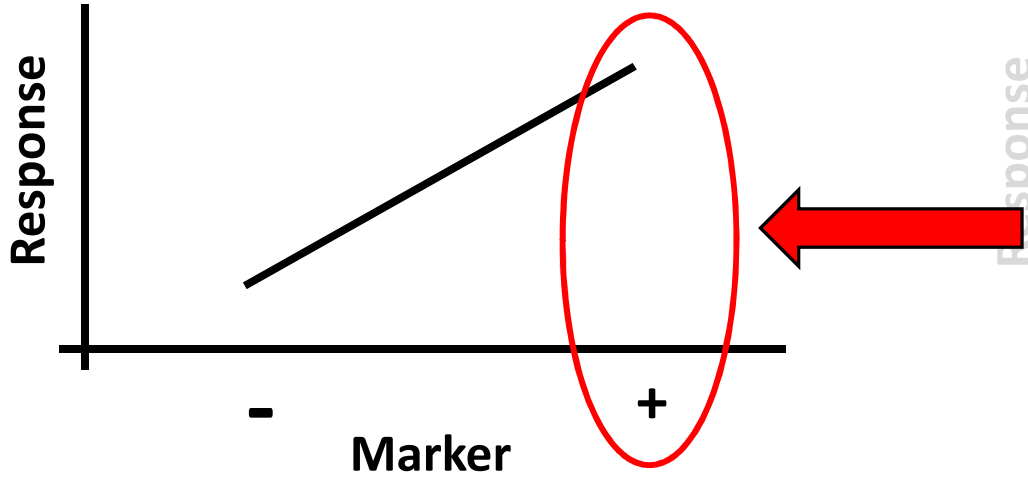
Predicting Alzheimer's Disease



Bringing data to life.

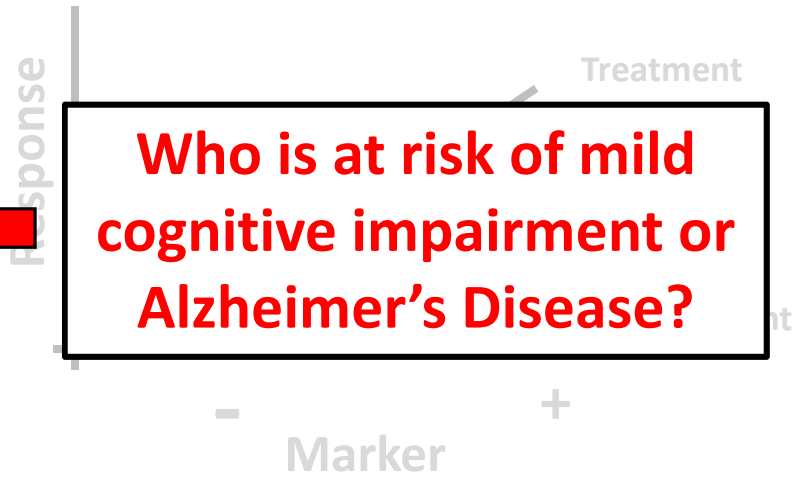
General Context

PROGNOSTIC



Single trait or signature of traits that identifies different groups of patients with respect to the risk of an outcome of interest ***in the absence of treatment***

PREDICTIVE



Who is at risk of mild cognitive impairment or Alzheimer's Disease?

Single trait or signature of traits that identifies different groups of patients with respect to the outcome of interest in **response to a particular treatment**

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 …"

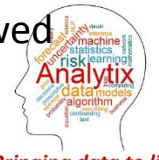
"The optimal value of the tuning parameter lambda, which was obtained by the cross-validation procedure, was then used to fit the model."

"… 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.

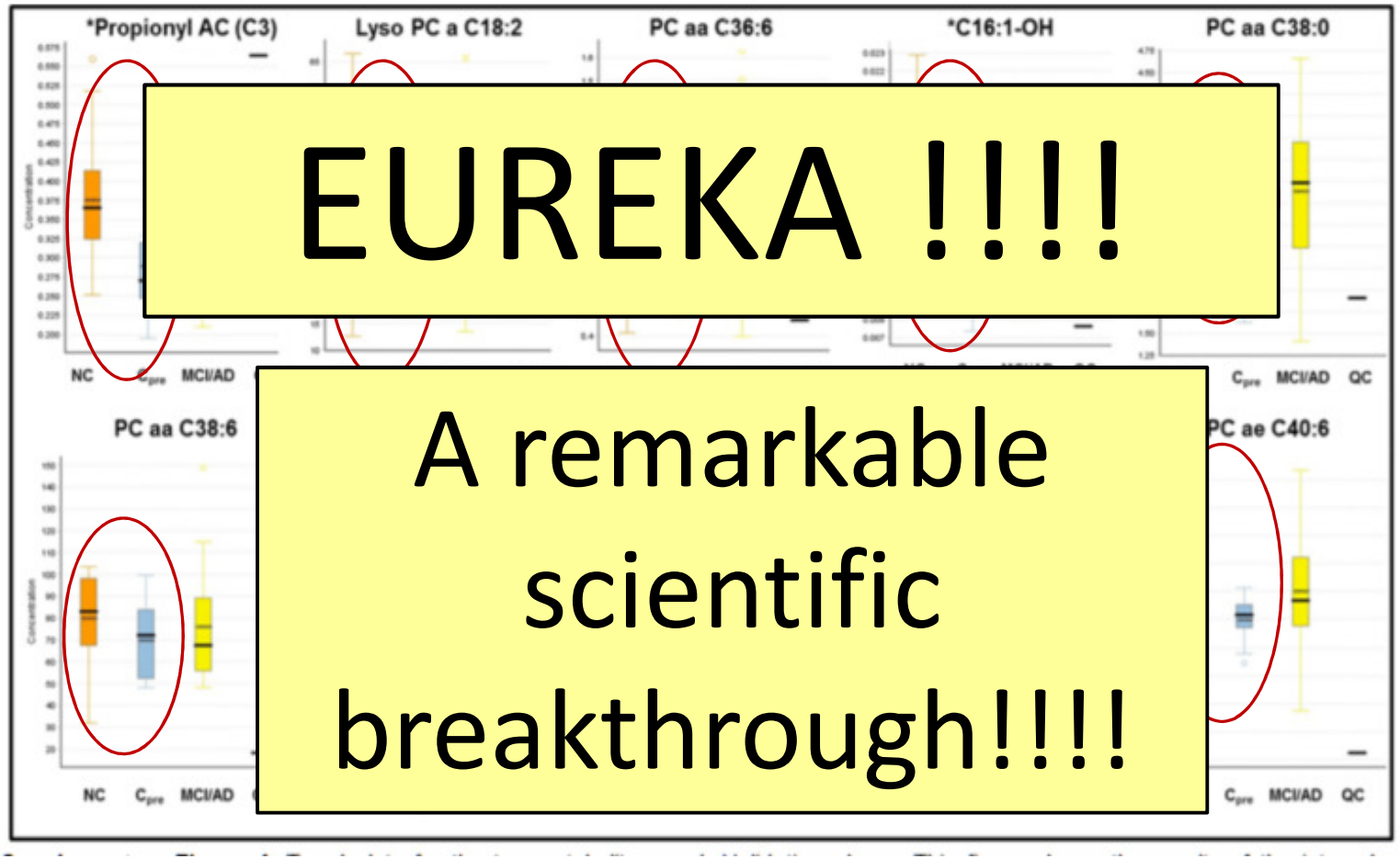
The Results

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



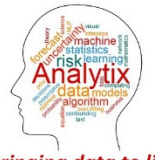
Bringing data to life.

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.

Predicting Alzheimer's Disease

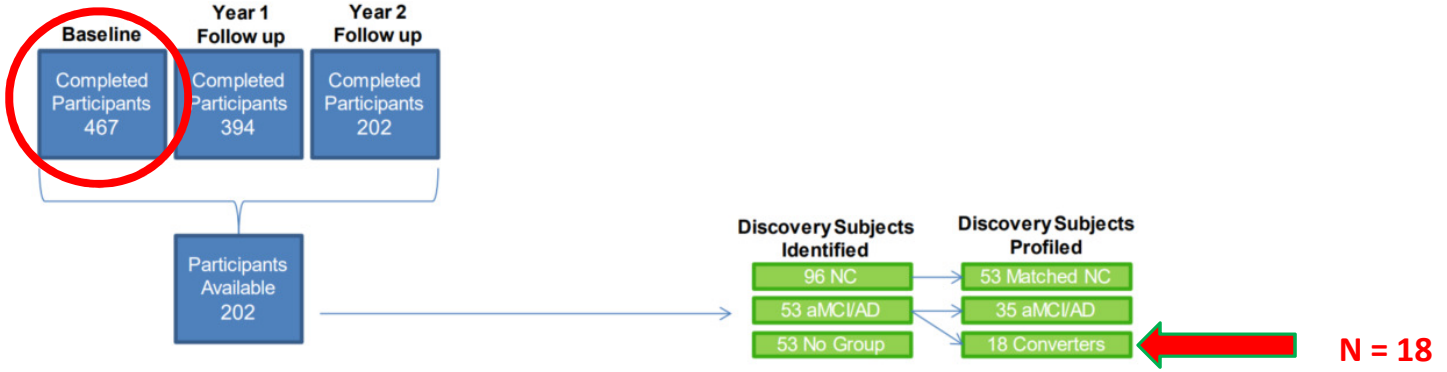
The Rest of the Story



Bringing data to life.

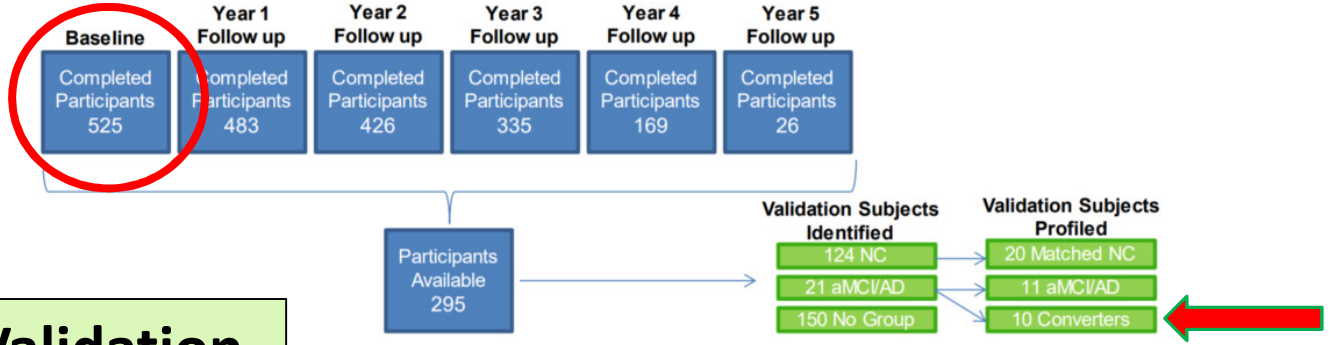
Patient Accounting

Discovery



N = 18

Validation



N = 10

Bringing data to life.

Patient Accounting

467 volunteers

- Discovery Phase: 202 (43%) participants available
  - 149 (74%) met certain criteria for inclusion in the analysis
    - 53 (36%) w
    - 96 (64%)n
      - 18 (19%) MCI/AD

Selection Bias ?!?

525 volunteers

- Validation Phase: 295 (56%) participants available
  - 145 (49%) met criteria for inclusion in the analysis
    - 21 (14%) with aMCI/AD
    - 124 (86%) normal
      - 10 (8%) converted from normal to aMCI/AD

Bringing data to life.

The Statistical Analytical Methods

celery

mustard

alfredo sauce

BBQ sauce

peanut butter

chicken

bleu cheese

onions

cilantro

lima beans

mango chutney

Bringing data to life.

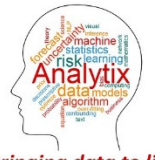
The Biological Analytical Methods

The actual data that was analyzed

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

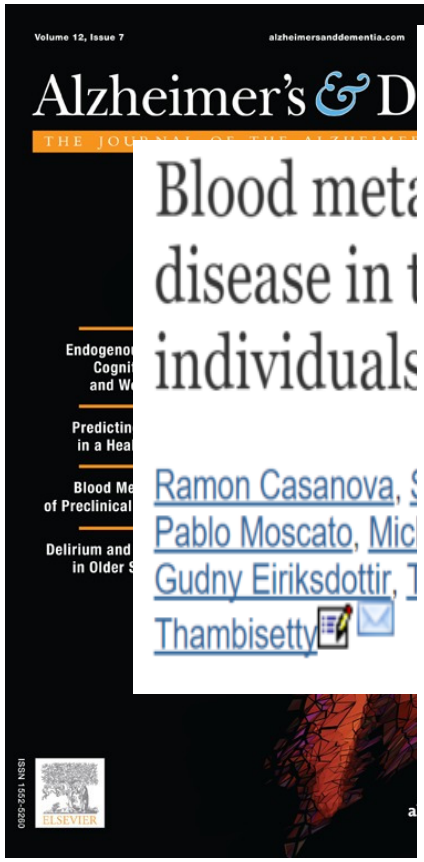
187 proteins analyzed

Multiplicity !!!!!



Bringing data to life.

The Rest of the Story



July 2016

RESULTS:

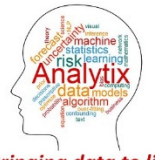
We failed to replicate these findings in a substantially larger study from two independent cohorts-the Baltimore Longitudinal Study of Aging ([BLSA], n = 93, AUC = 0.642, sensitivity/specificity of 51.6%/65.7%) and the Age, Gene/Environment Susceptibility-Reykjavik Study ([AGES-RS], n = 100, AUC = 0.395, 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.

The Publicity

Prognostic Biomarkers

An Alternate Approach



Bringing data to life.

A Somewhat Real Story

What's in this vial of fluid?

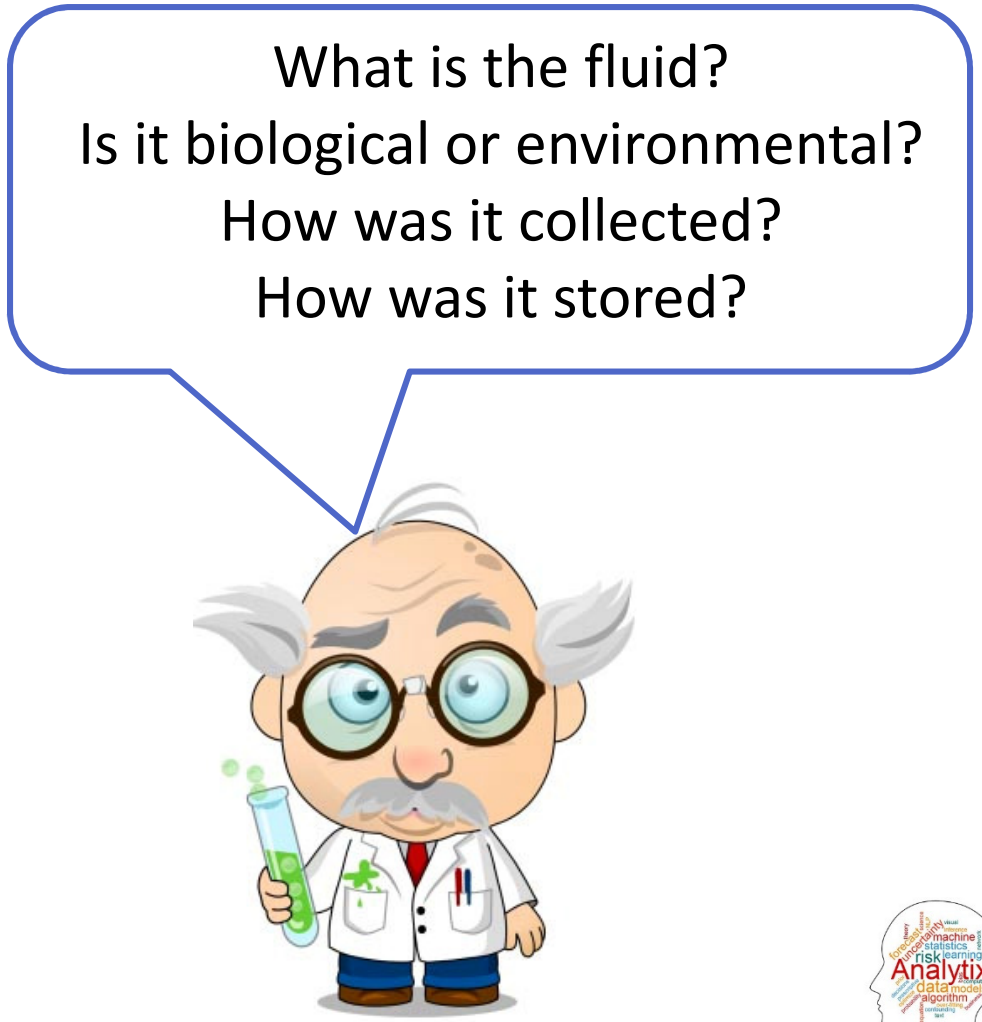
Does this guy have any idea what he is asking?

Bringing data to life.

A Somewhat Real Story

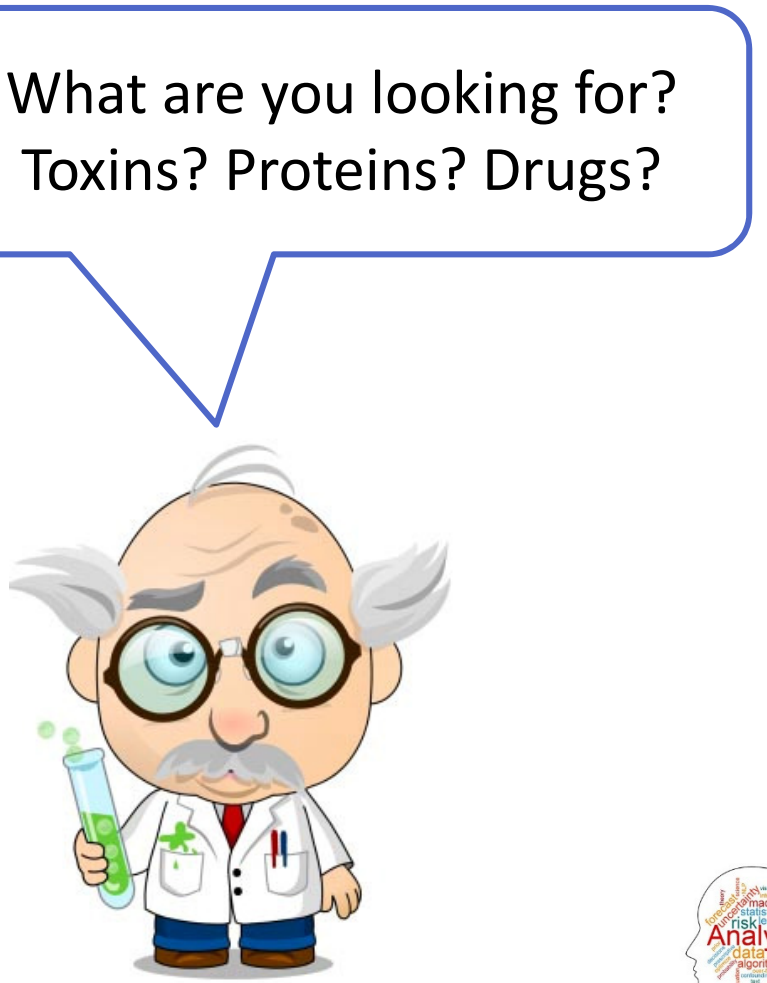
Bringing data to life.

A Somewhat Real Story



Bringing data to life.

A Somewhat Real Story



Bringing data to life.

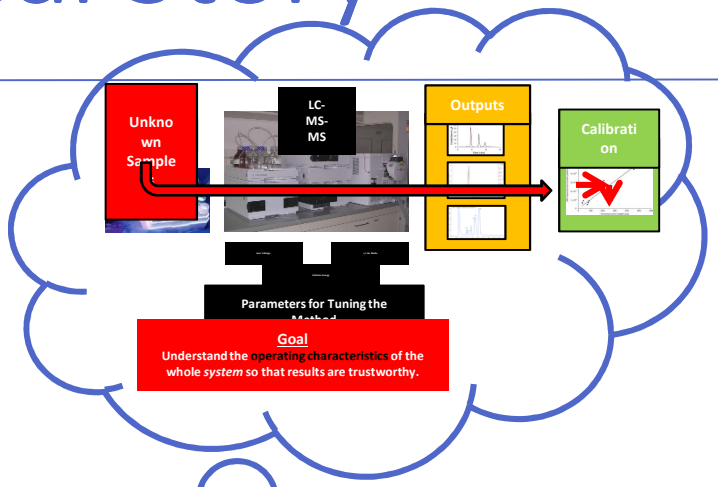
A Somewhat Real Story

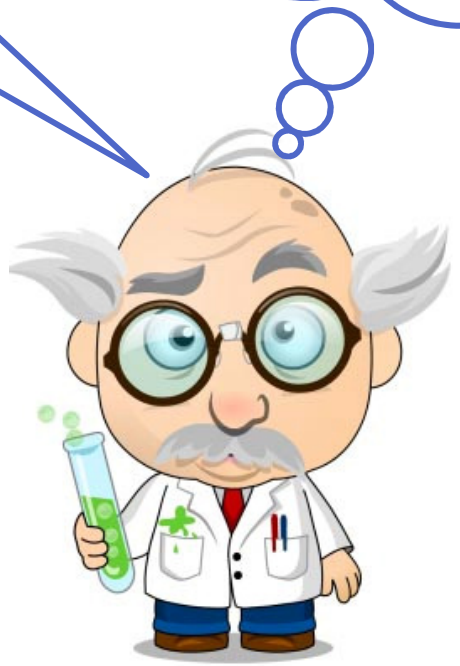
You don't understand. I just want you to explore this fluid and tell me everything you can learn from it.

Bringing data to life.

A Somewhat Real Story

Allow me to explain how this works.





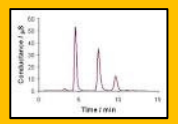
Bringing data to life.

Bio-Analytical Method Development

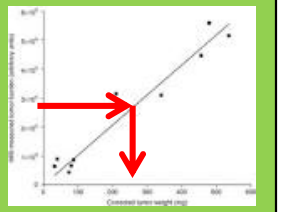
Samples with Unknown Contents

LC-MS-MS

Outputs



Calibration



+/- Ion Mode

ergy

rs for Method

Goal

U erating c he whole *system* so that resul y.

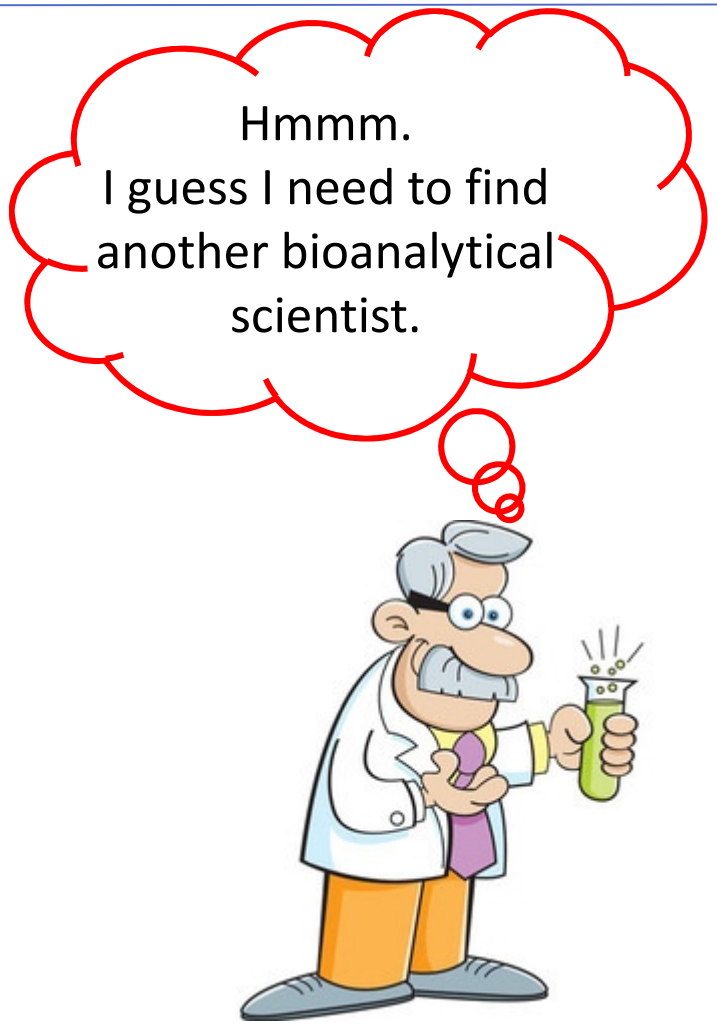
A Somewhat Real Story

Aha! Very interesting. But once you have your bioanalytical method or system validated, then you can measure whatever you want in my vial. Right?!?!

No! The method is specific to a particular matrix and the particular substances you want to find.

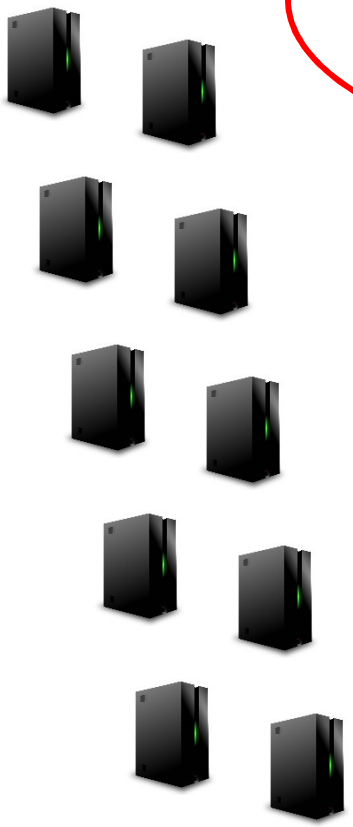
Bringing data to life.

A Somewhat Real Story



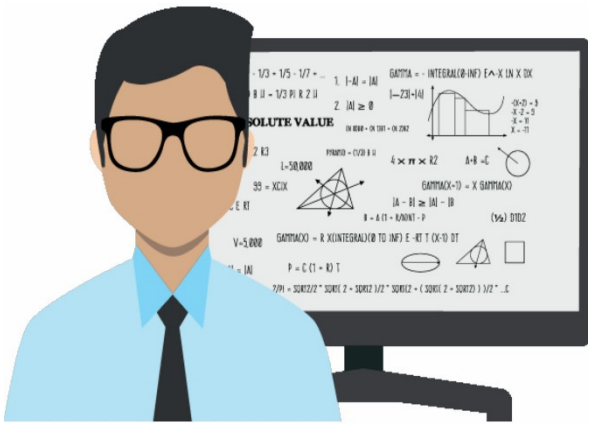
Bringing data to life.

A *Real* Story

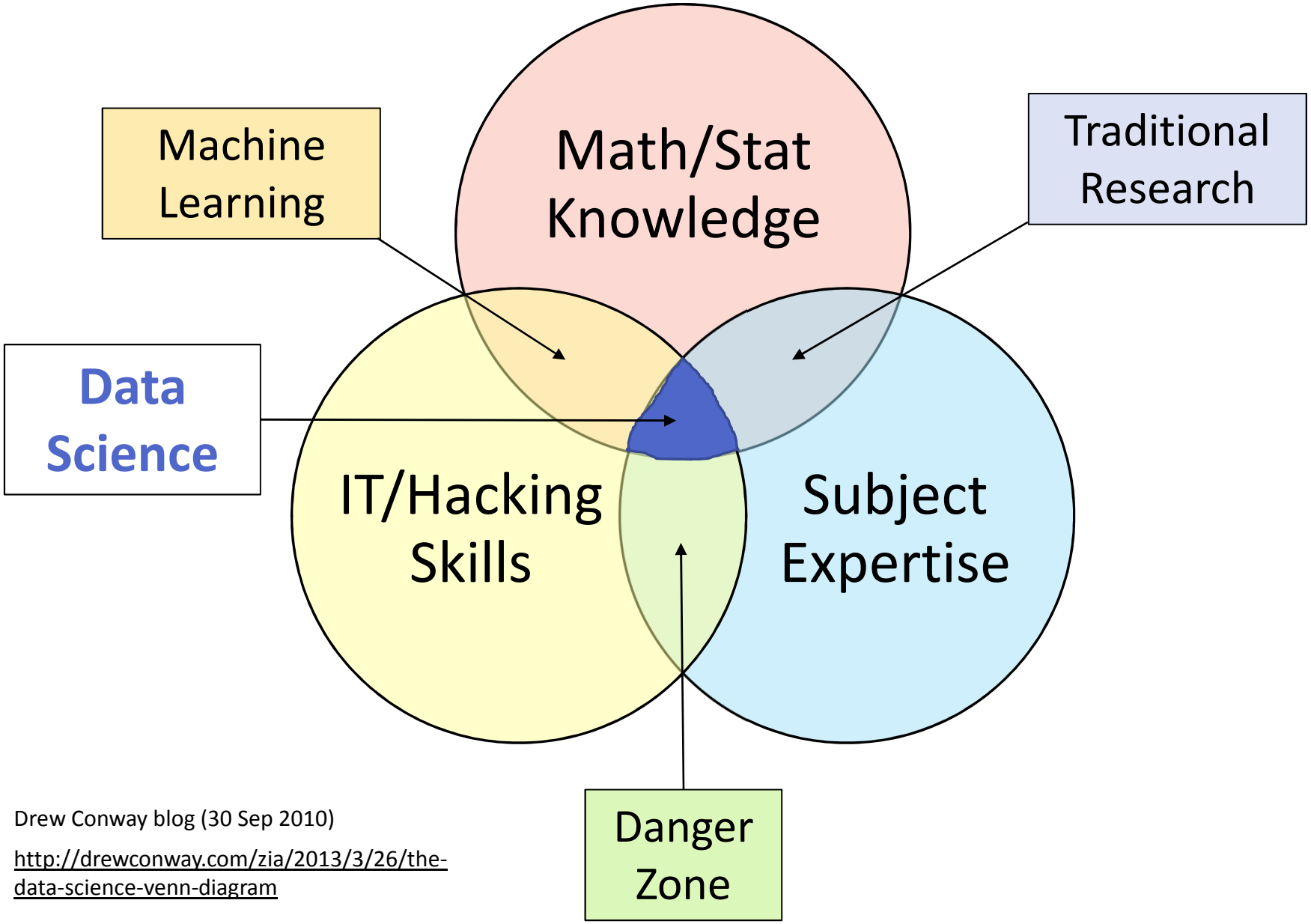


Can you analyze this data and tell me what I need to know about my business?

Absolutely!!!
When can I start?



Bringing data to life.



Drew Conway blog (30 Sep 2010)

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

Danger Zone

"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

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

Partner and Data Scientist
Large Business Consulting Company

"Here's what's in our data. It's not my job to talk about what it means, …"

Cassie Kozyrkov
Chief Decision Scientist at Google
HBR, Dec 4, 2018

"Models which can be 'tuned' in many different ways give researchers more scope to perceive a pattern where none exists. According to some estimates, **three-quarters of published scientific papers in the field of machine learning are bunk because of this 'overfitting',** says **Sandy Pentland, a computer scientist at the Massachusetts Institute of Technology**."

Bringing data to life.

Danger Zone

Geo data table

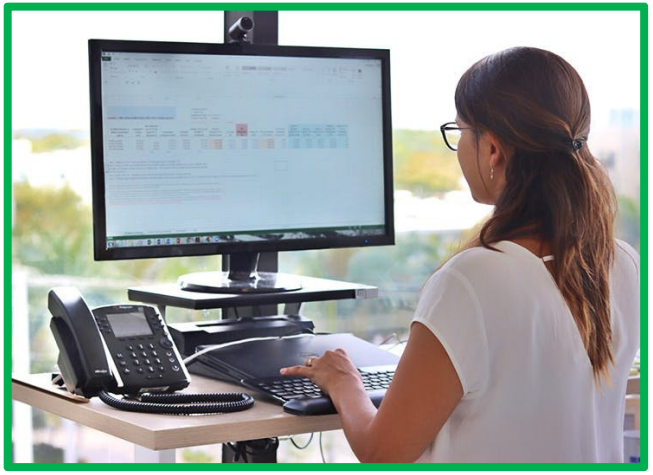
Edit row

Enter filter text

	Location	Gaging station	River	Level	Date	Average maxium water level	Average minimum water level	Water level	Hydrograph
1	Achleiten	Achleiten	Donau	400 cm	2016-06-08 11:30	-	-	Unknown	ganglinien390
2	Passau Ilzstadt	Passau Ilzstadt	Donau	614 cm	2016-06-08 09:15	827 cm	418 cm	Normal	ganglinien294
3	Passau Donau	Passau Donau	Donau	616 cm	2016-06-08 09:30	832 cm	403 cm	Normal	ganglinien170
4	Vilshofen	Vilshofen	Donau	413 cm	2016-06-08 10:15	535 cm	299 cm	Normal	ganglinien494
5	Hofkirchen	Hofkirchen	Donau	397 cm	2016-06-08 09:00	557 cm	196 cm	Normal	ganglinien576
6	Deggendorf	Deggendorf	Donau	414 cm	2016-06-08 11:30	615 cm	192 cm	Normal	ganglinien344
7	Pfelling	Pfelling	Donau	497 cm	2016-06-08 09:15	697 cm	268 cm	Normal	ganglinien492
8	Straubing	Straubing	Donau	371 cm	2016-06-08 09:00	577 cm	123 cm	Normal	ganglinien236
9	Pfatter	Pfatter	Donau	413 cm	2016-06-08 10:00	601 cm	307 cm	Normal	ganglinien266
10	Schwabelweis	Schwabelweis	Donau	358 cm	2016-06-08 11:45	520 cm	283 cm	Normal	ganglinien428
11	Eiserne Brücke	Eiserne Brücke	Donau	320 cm	2016-06-08 09:15	501 cm	195 cm	Normal	ganglinien500
12	Niederwinzer	Niederwinzer	Donau	504 cm	2016-06-06 04:00	-	-	Unknown	ganglinien122
13	Oberndorf	Oberndorf	Donau	308 cm	2016-06-08 11:45	518 cm	157 cm	Normal	ganglinien186
14	Kelheimwinzer	Kelheimwinzer	Donau	353 cm	2016-06-08 11:45	516 cm	257 cm	Normal	ganglinien488
15	Ingolstadt Luitpoldstrasse	Ingolstadt Luitpoldstrasse	Donau	302 cm	2016-06-08 07:15	-	-	Unknown	ganglinien288
16	Schöna	Schöna	Elbe	173 cm	2016-06-08 11:30	641 cm	91 cm	Normal	ganglinien568
17	Pirna	Pirna	Elbe	199 cm	2016-06-08 11:15	614 cm	110 cm	Normal	ganglinien358
18	Dresden	Dresden	Elbe	167 cm	2016-06-08 11:30	574 cm	78 cm	Normal	ganglinien280
19	Meissen	Meissen	Elbe	224 cm	2016-06-08 11:15	637 cm	126 cm	Normal	ganglinien130
20	Riesa	Riesa	Elbe	239 cm	2016-06-08 11:30	635 cm	148 cm	Normal	ganglinien404
21	Mühlberg	Mühlberg	Elbe	262 cm	2016-06-08 11:15	684 cm	177 cm	Normal	ganglinien198
22	Torgau	Torgau	Elbe	167 cm	2016-06-08 11:30	623 cm	70 cm	Normal	ganglinien160
23	Pretzsch-Mauken	Pretzsch-Mauken	Elbe	165 cm	2016-06-08 11:15	584 cm	71 cm	Normal	ganglinien162
24	Elster	Elster	Elbe	165 cm	2016-06-08 11:15	514 cm	60 cm	Normal	ganglinien140
25	Wittenberg	Wittenberg	Elbe	233 cm	2016-06-08 11:45	543 cm	114 cm	Normal	ganglinien436

Import data | Export data | Map legend

If you got data …



… and a computer …

… it's **EASY** to get an answer.

Danger Zone

false negative

uncertainty

Assessing the

variability

QUALITY

probability

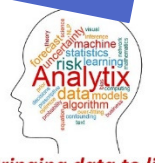
likelihood

of that Answer is

false positive

predictive value

***very difficult*.**

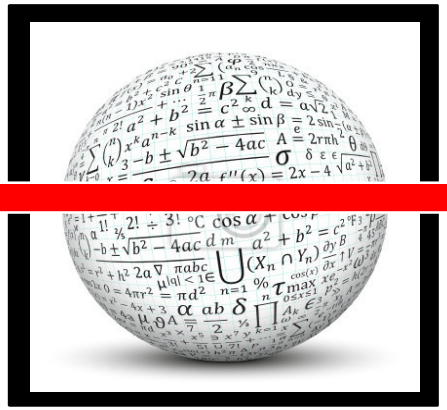


Bringing data to life.

Stat-Analytical Method Development

Data with Unknown Effects

Algorithms



Searching Algorithm

Variable Selection

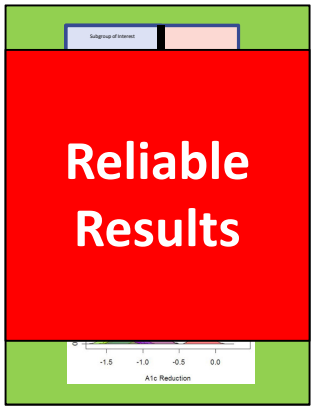
Model Complexity

Parameters for Tuning the Method

Outputs

- ❖Bias
- ❖Small sample performance

"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.

Analytical Methods Development

Predictive Medicine

Simulate
- Overall Sample Size
- Subgroup Sizes
- Variability
- Number of biomarkers
- Biomarker effects

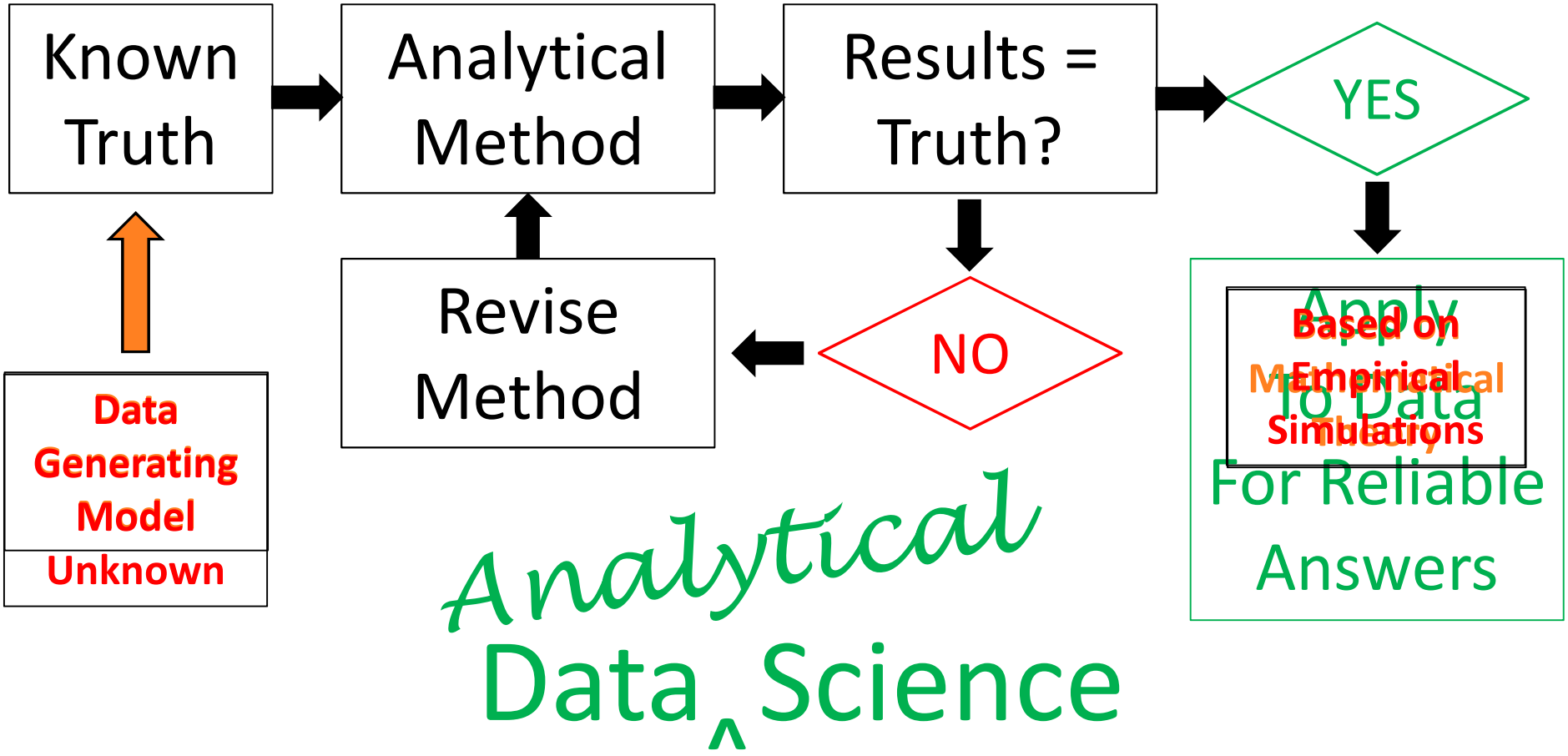
Is there a predictor (i.e. biomarker signature) of the subgroup of patients who progress to disease X?

For this Database, how often did the Analytical Method get the right answer?

Clinical Database → Analytical Methods

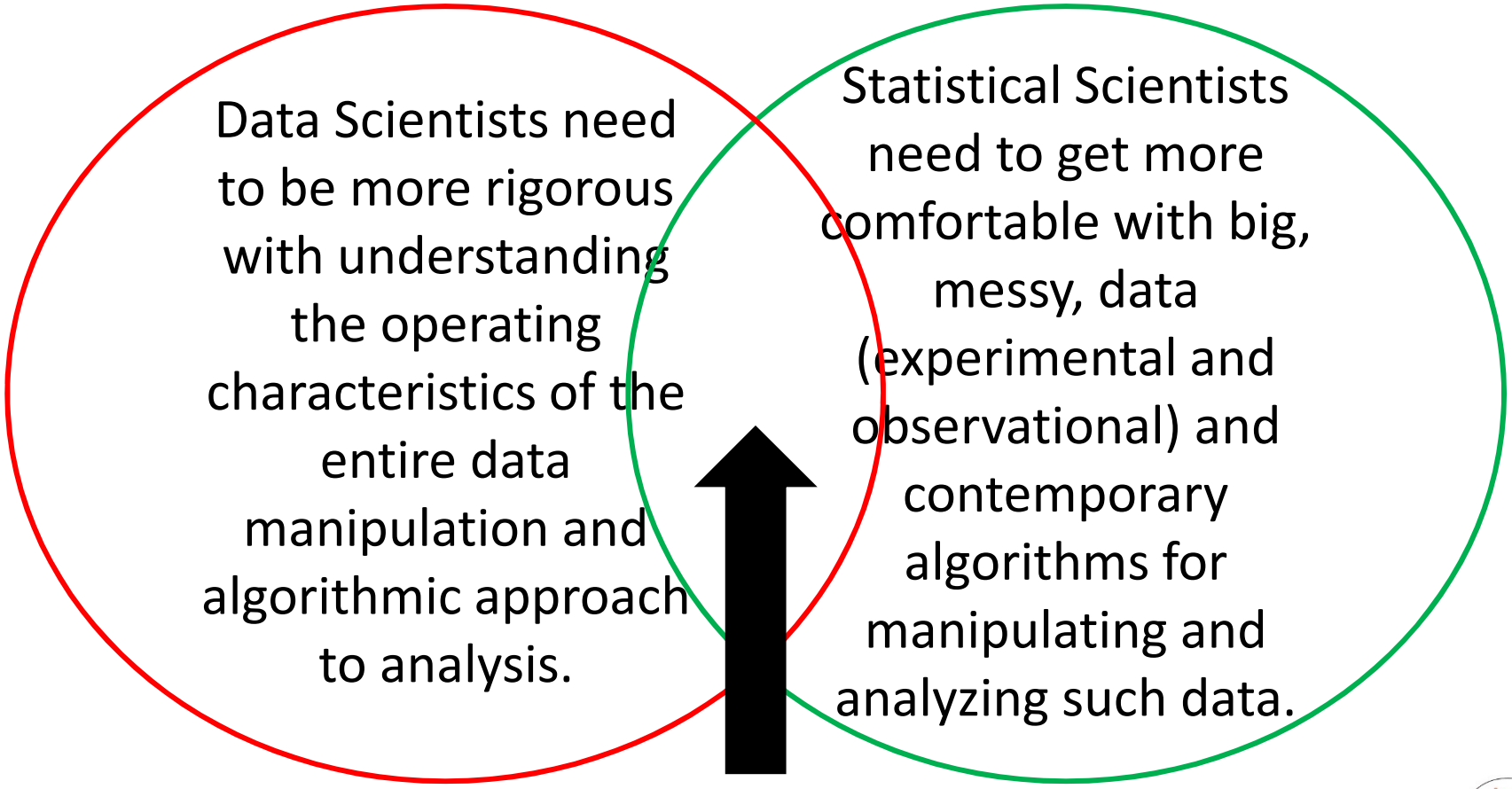
Bringing data to life.

Analytical Methods Development



Analytical Data Science

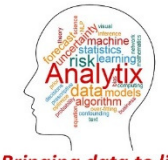
Statistical Science and Data Science



DATA ANALYTICAL SCIENTISTS

Bringing data to life.

2B. Predicting Acute Kidney Injury



Bringing data to life.

Acute Kidney Injury

LETTER

https://doi.org/10.1038/s41586-019-1390-1

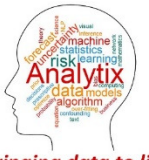
A clinically applicable approach to continuous prediction of future acute kidney injury

Nenad Tomašev[1]*, Xavier Glorot[1], Jack W. Rae[1,2], Michal Zielinski[1], Harry Askham[1], Andre Saraiva[1], Anne Mottram[1], Clemens Meyer[1], Suman Ravuri[1], Ivan Protsyuk[1], Alistair Connell[1], Cían O. Hughes[1], Alan Karthikesalingam[1], Julien Cornebise[1,12], Hugh Montgomery[3], Geraint Rees[4], Chris Laing[5], Clifton R. Baker[6], Kelly Peterson[7,8], Ruth Reeves[9], Demis Hassabis[1], Dominic King[1], Mustafa Suleyman[1], Trevor Back[1,13], Christopher Nielson[10,11,13], Joseph R. Ledsam[1,13]* & Shakir Mohamed[1,13]

The early prediction of deterioration could have an important role in supporting healthcare professionals, as an estimated 11% of deaths in hospital follow a failure to promptly recognize and treat deteriorating patients[1]. To achieve this goal requires predictions of patient risk that are continuously updated and accurate, and delivered at an individual level with sufficient context and enough time to act. Here we develop a deep learning approach for the

Promising recent work on modelling adverse events from electronic health records[2–17] suggests that the incorporation of machine learning may enable the early prediction of AKI. Existing examples of sequential AKI risk models have either not demonstrated a clinically applicable level of predictive performance[25] or have focused on predictions across a short time horizon that leaves little time for clinical assessment and intervention[26].

116 | NATURE | VOL 572 | 1 AUGUST 2019



Bringing data to life.

Acute Kidney Injury

AKI potentially life-threatening

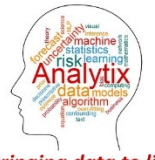
Predicting who will succumb to AKI 48 hours in advance allows for intervention

Data from US Department of Veteran Affairs

"The total number of independent entries in the dataset was approximately 6 billion ..."

- I think they mean "distinct", not "independent"
- 703,782 adult patients

620,000 features! (i.e. potential biomarkers)



Bringing data to life.

Acute Kidney Injury

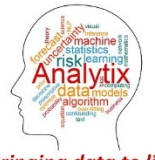
Model is recurrent neural network

Output is pr(AKI in next 48 hours) = p_{AKI}

When p_{AKI} > threshold, declare positive/alert

Retrospective model building

- 80% for training/model building
- 5% for validation
- 5% for calibration
- 10% for test



Bringing data to life.

Acute Kidney Injury

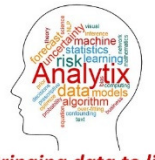
Results

Model predicts (with lead time of 48hrs)

- "55.8% of all inpatient episodes of acute kidney injury"
- "90.2% of all acute kidney injuries that required subsequent administration of dialysis"

"A ratio of 2 false alerts for every true alert."

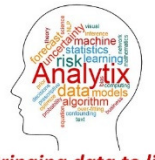
"Area under the receiver operating characteristic curve of 92.1%."



Bringing data to life.

Acute Kidney Injury

The Rest of the Story



Bringing data to life.

Controlled Experiment

npj | Digital Medicine

www.nature.com/npjdigitalmed

ARTICLE OPEN

Evaluation of a digitally-enabled care pathway for acute kidney injury management in hospital emergency admissions

Alistair Connell[1,2], Hugh Montgomery[1], Peter Martin[3], Claire Nightingale[3,4], Omid Sadeghi-Alavijeh[5], Dominic King[2], Alan Karthikesalingam[2], Cian Hughes[2], Trevor Back[2], Kareem Ayoub[2], Mustafa Suleyman[2], Gareth Jones[5], Jennifer Cross[5], Sarah Stanley[5], Mary Emerson[5], Charles Merrick[5], Geraint Rees[6], Chris Laing[5,7] and Rosalind Raine[3]

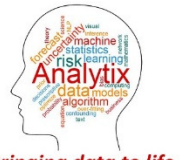
We developed a digitally enabled care pathway for acute kidney injury (AKI) management incorporating a mobile detection application, specialist clinical response team and care protocol. Clinical outcome data were collected from adults with AKI on emergency admission before (May 2016 to January 2017) and after (May to September 2017) deployment at the intervention site and another not receiving the intervention. Changes in primary outcome (serum creatinine recovery to ≤120% baseline at hospital discharge) and secondary outcomes (30-day survival, renal replacement therapy, renal or intensive care unit (ICU) admission, worsening AKI stage and length of stay) were measured using interrupted time-series regression. Processes of care data (time to AKI recognition, time to treatment) were extracted from casenotes, and compared over two 9-month periods before and after implementation (January to September 2016 and 2017, respectively) using pre–post analysis. There was no step change in renal recovery or any of the secondary outcomes. Trends for creatinine recovery rates (estimated odds ratio (OR) = 1.04, 95% confidence interval (95% CI): 1.00–1.08, $p = 0.038$) and renal or ICU admission (OR = 0.95, 95% CI: 0.90–1.00, $p = 0.044$) improved significantly at the intervention site. However, difference-in-difference analyses between sites for creatinine recovery (estimated OR = 0.95, 95% CI: 0.90–1.00, $p = 0.053$) and renal or ICU admission (OR = 1.06, 95% CI: 0.98–1.16, $p = 0.140$) were not significant. Among process measures, time to AKI recognition and treatment of nephrotoxicity improved significantly ($p < 0.001$ and 0.047 respectively).

npj Digital Medicine (2019)2:67 ; https://doi.org/10.1038/s41746-019-0100-6

Renal or ICU admission

OR = 1.06 (0.98, 1.16)
p = 0.140

Published online 31 Jul 2019



Bringing data to life.

The Rest of the Story

thebmj

BMJ 2019;366:l5011 doi: 10.1136/bmj.l5011 (Published 2 August 2019) Page 1 of 2

Check for updates

NEWS

App to help spot acute kidney injury had no clinical benefits, study finds

Nigel Hawkes

London

An alerting tool developed in cooperation with the Google company DeepMind to speed up the diagnosis of acute kidney injury has shown no clinical benefits when it was compared

recognition of AKI and treatment of nephrotoxicity improved significantly.

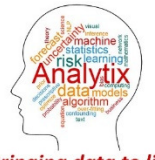
An earlier randomised trial published in the *Lancet* in 2015 had

BMJ 2019;366:l5011 doi: 10.1136/bmj.l5011 (Published 2 August 2019)

Bringing data to life.

Prognostic Biomarkers

An Alternate Approach



Bringing data to life.

Acute Kidney Injury

Machine learning models that act as diagnostic devices should follow the same principles and reporting as *in vitro* diagnostics

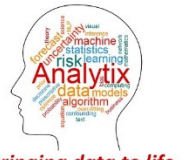
		TRUTH		
		Positive	Negative	
DIAGNOSTIC TEST RESULT	Positive	Sensitivity	False positive	PPV
	Negative	False negative	Specificity	NPV
		Prevalence		

Bringing data to life.

Acute Kidney Injury

Machine learning models that act as diagnostic devices should follow the same principles and reporting as *in vitro* diagnostics

		TRUTH		
		Positive	Negative	
DIAGNOSTIC TEST RESULT	Positive	1	2	PPV = 33%
	Negative	False negative	Specificity	NPV
		13.4%		



Bringing data to life.

Acute Kidney Injury

		TRUTH		
		Positive	Negative	
DIAGNOSTIC TEST RESULT	Positive	1	2	PPV = 33%
	Negative	False negative	Specificity	NPV
		13.4%		

PPV is decidedly a Bayesian notion

Bayesian approaches work best in conjunction with a utility function

Balance the cost of FP and FN and the value of TP and TN to optimize PPV

AnalytixThinking.Blog: Détente: The Peaceful Co-Existence of Significance Levels and Bayes

Acute Kidney Injury

Proposal

When used for making predictions (i.e. diagnosis),

all the hype about ML, digital medicine

(aka *in silico* diagnostics)

should be fit into existing analytical paradigms

(aka development of diagnostics tests)

in order to assess their validity and utility.

Statistical Science and Data Science

Data Analytical Scientists

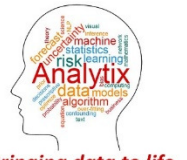
need to be more like

Bio-Analytical Scientists

regarding their approach to

validating their Analytical Methods

We are developing *in silico* assays.

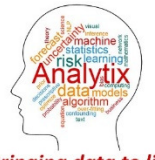


Bringing data to life.

3. Predictive Biomarker

Finding Heterogeneity

Randomized, Controlled Clinical Trials



Bringing data to life.

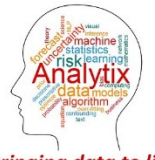
Subgroup Identification

"Always do subgroup *analysis*, but never believe them."

Attributed to Sir Richard Peto

Professor of Medical Statistics and Epidemiology

University of Oxford, England

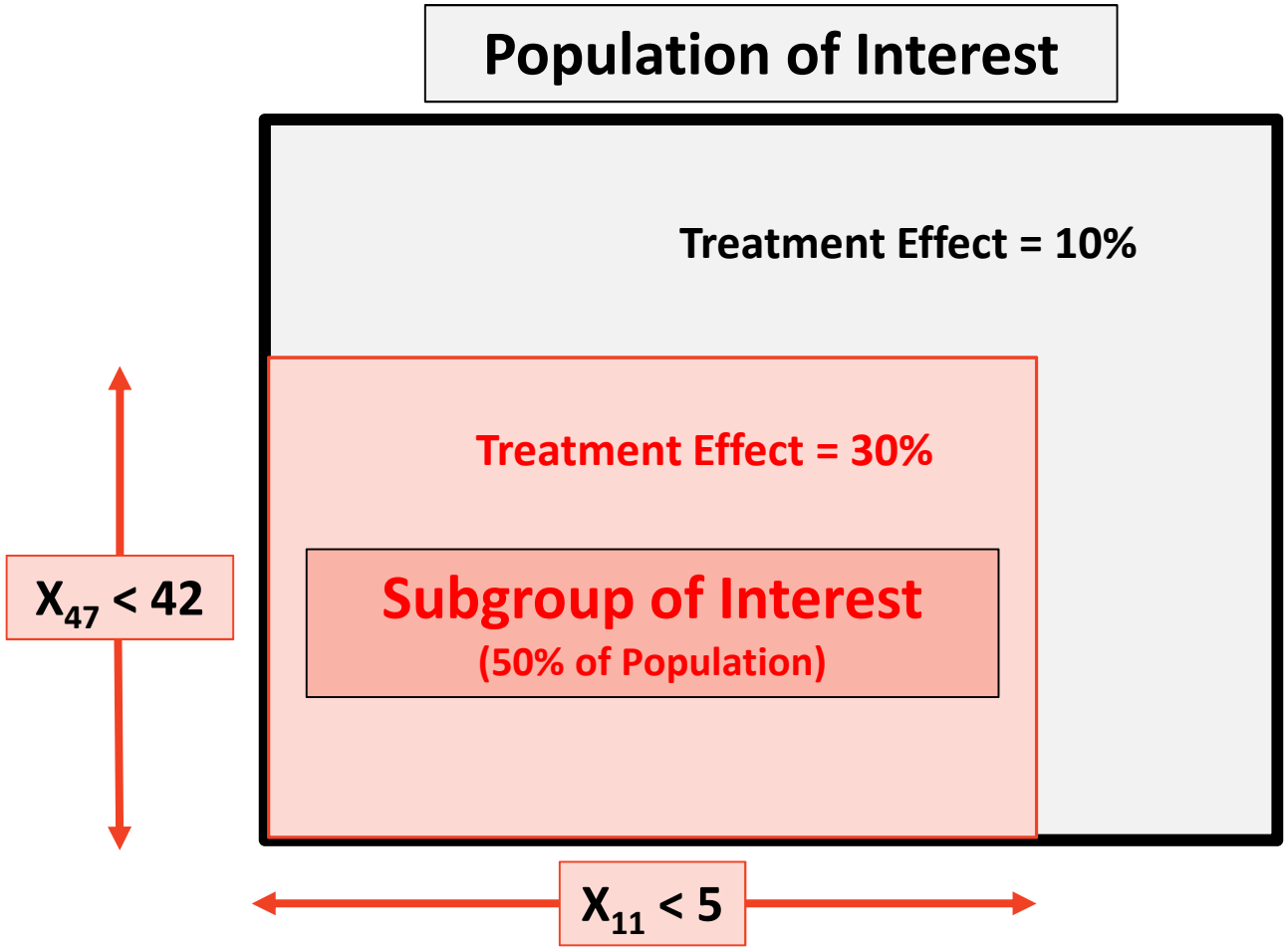


Bringing data to life.

3A. An Open Challenge

Clinical Trials

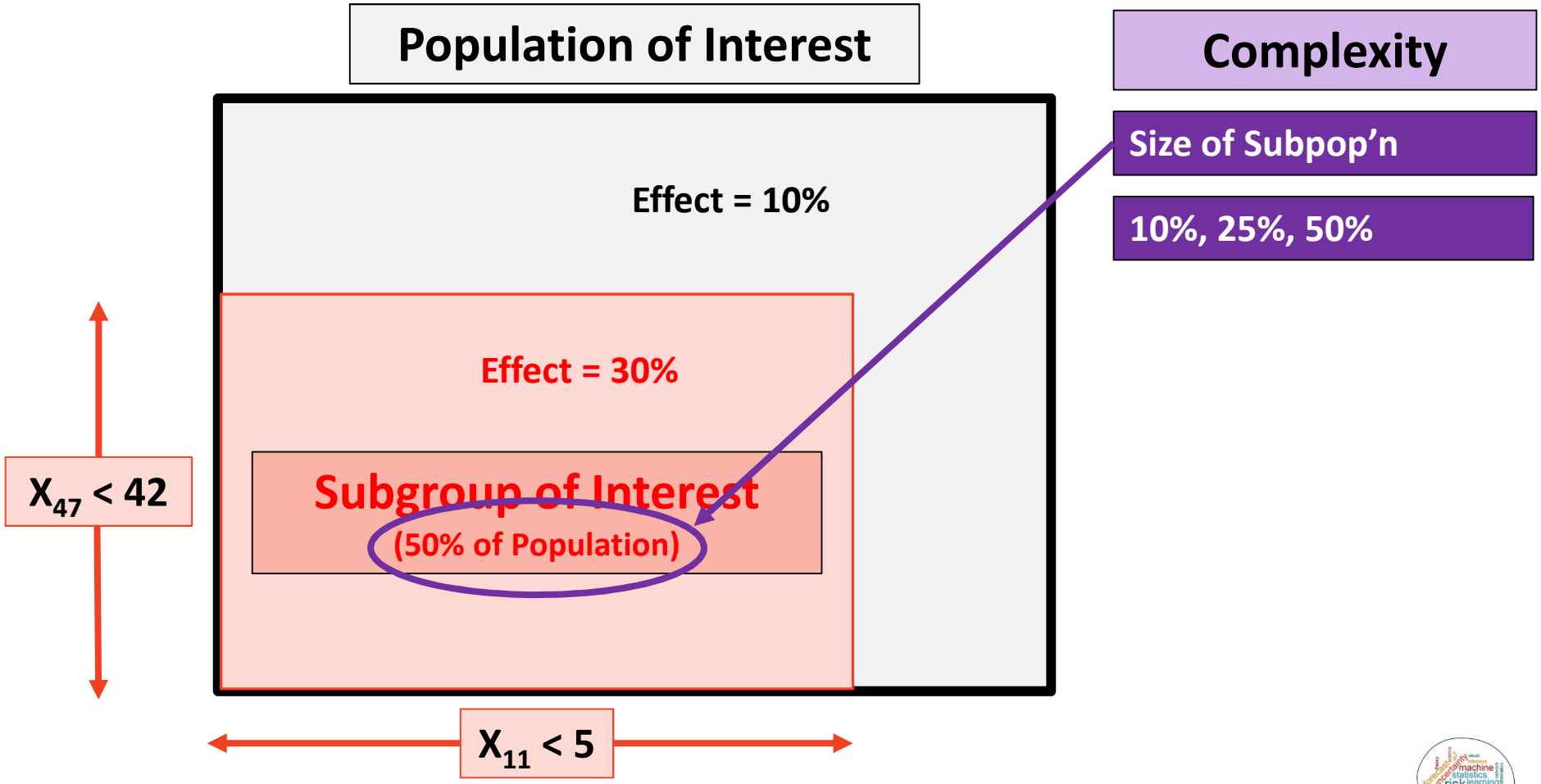
Treatment versus Control



Bringing data to life.

Clinical Trials

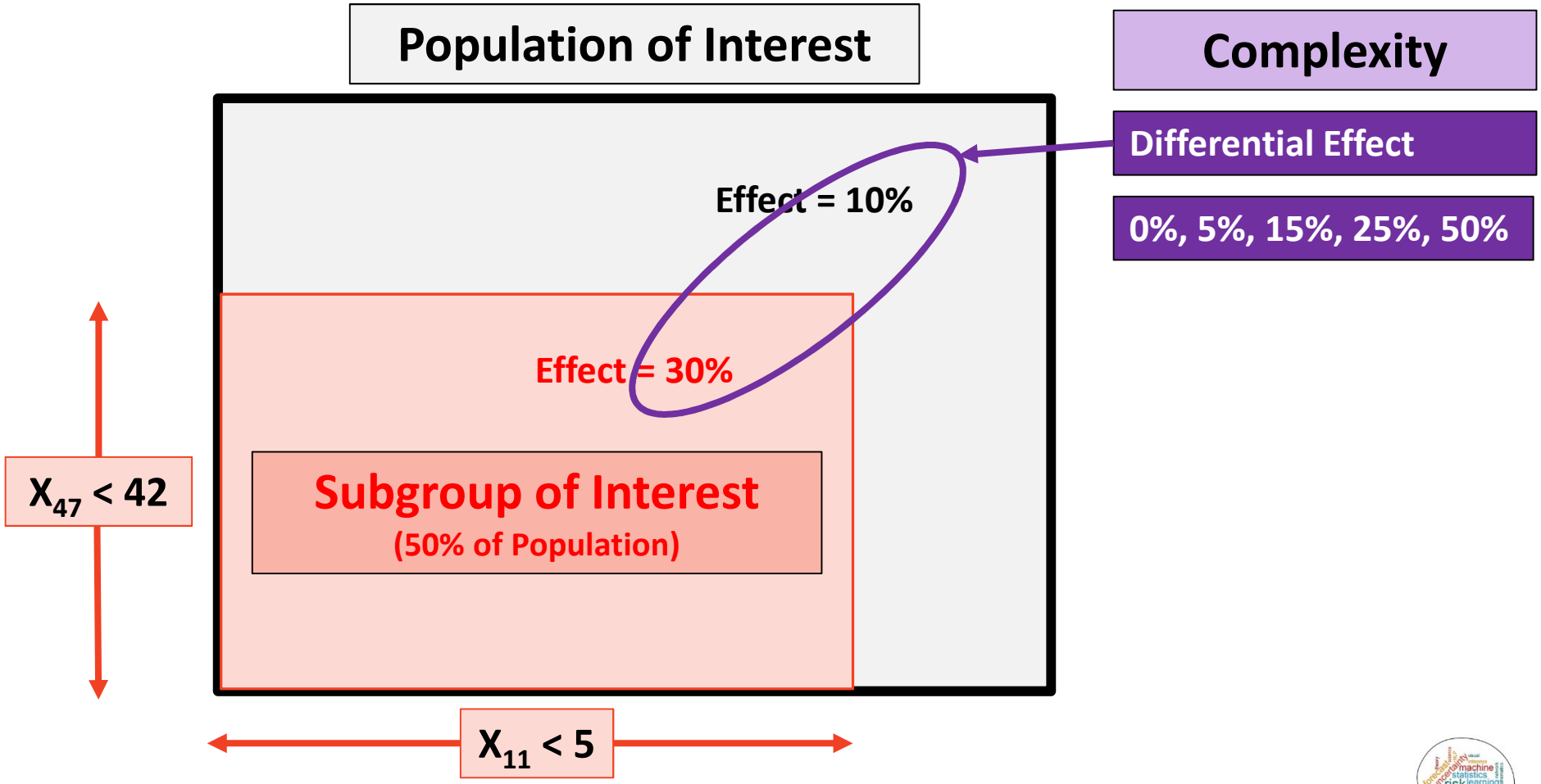
Treatment versus Control



Bringing data to life.

Clinical Trials

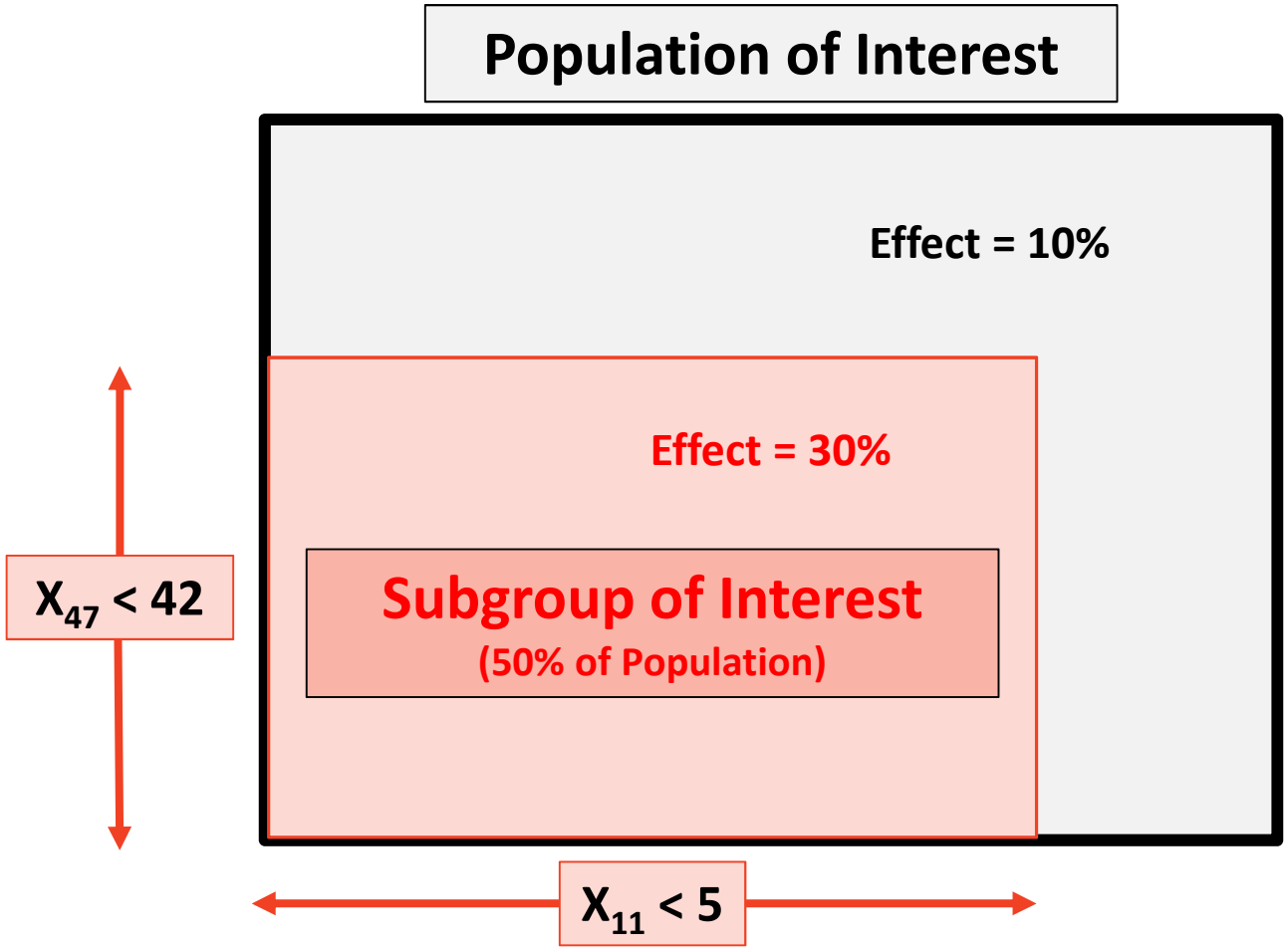
Treatment versus Control



Bringing data to life.

Clinical Trials

Treatment versus Control



Complexity

Number of biomarkers in the dataset

$X_1, X_2, X_3 \ldots X_{110}$

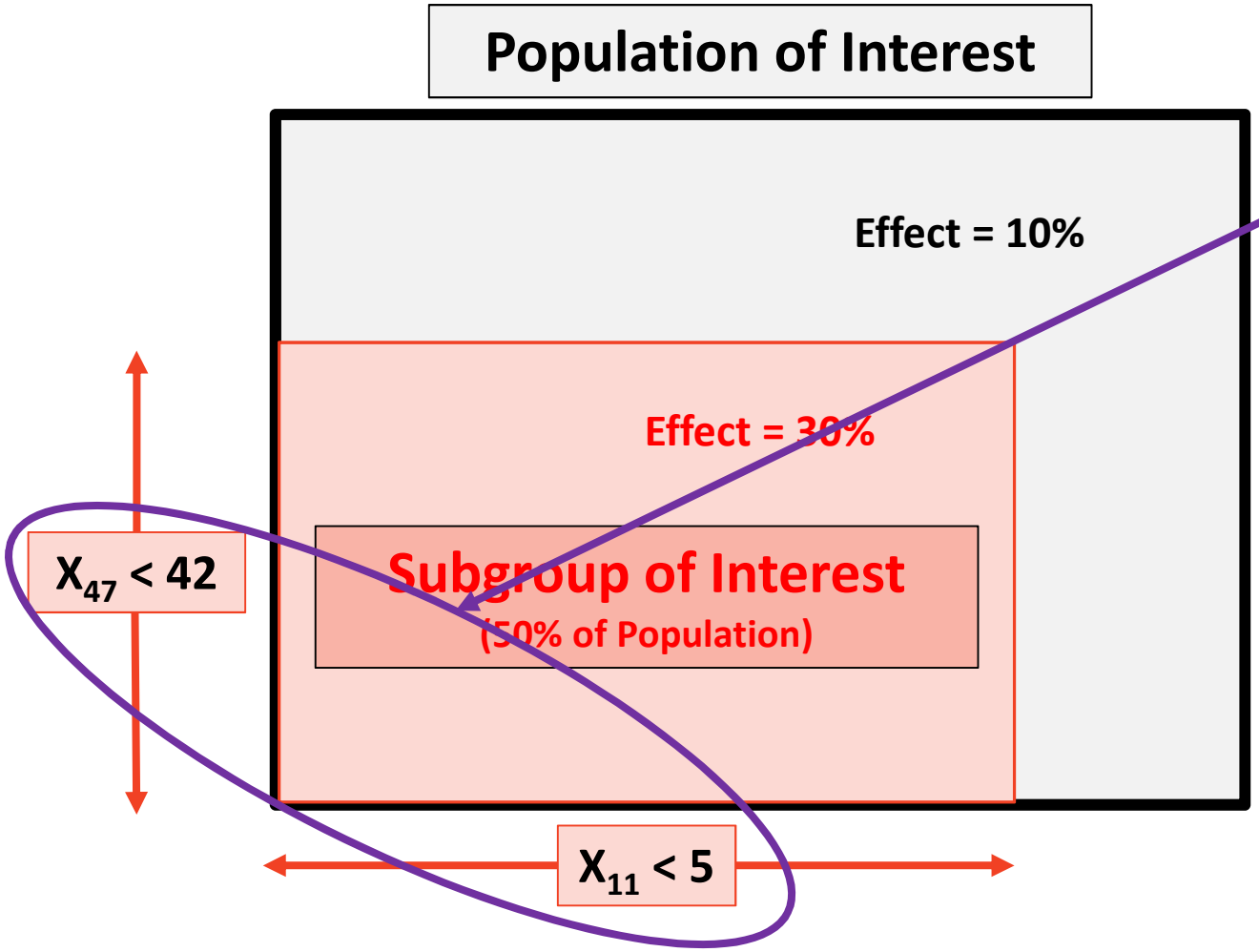
10, 25, 75, 250, 500

AnalytixThinking.Blog: Genetic Subgroups and CV Disease … 5.5×10^6 SNP biomarkers

Bringing data to life.

Clinical Trials

Treatment versus Control



Complexity

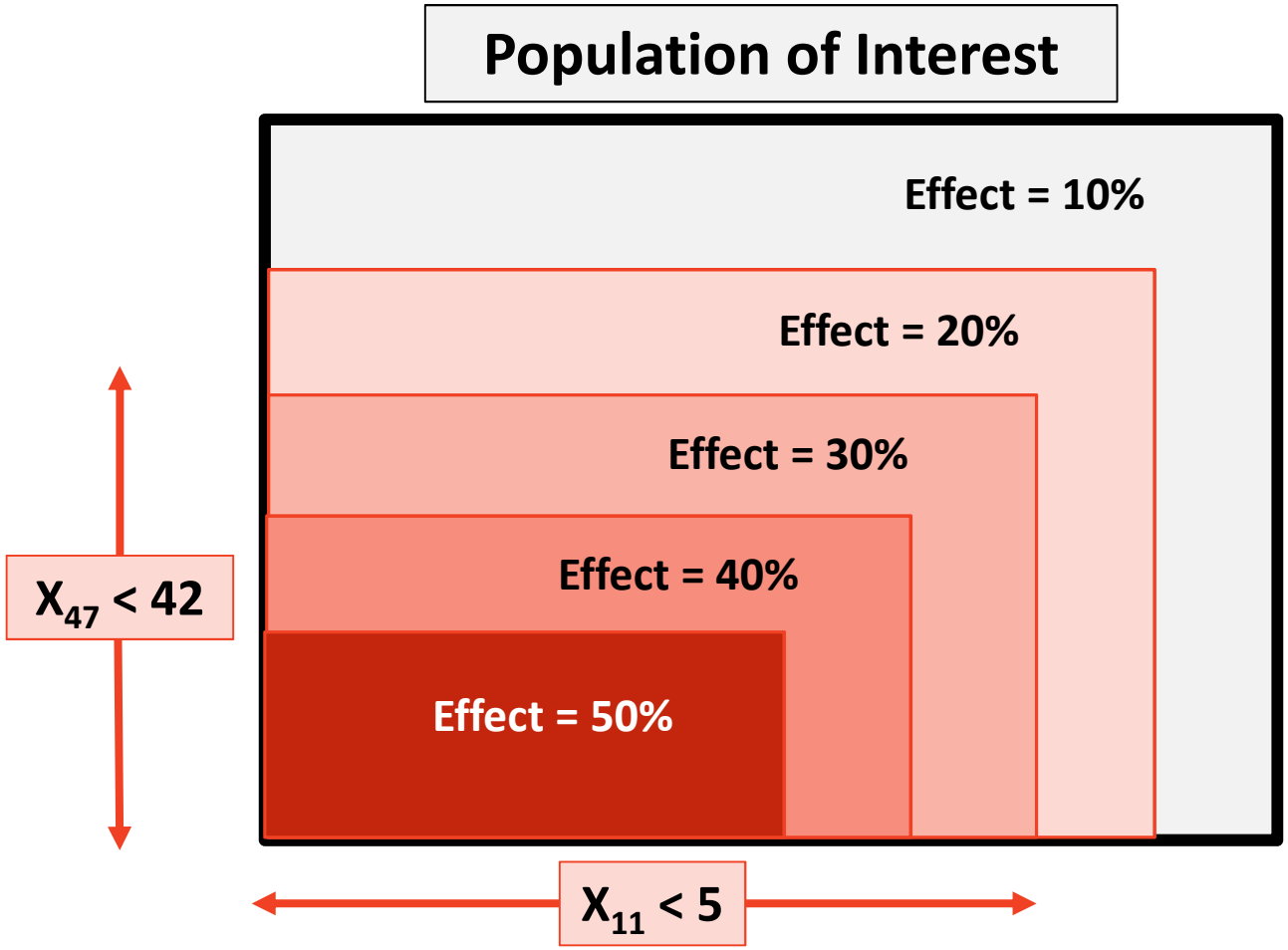
Number of biomarkers defining subpop'n

1, 2, …

Bringing data to life.

Clinical Trials

Treatment versus Control



Complexity

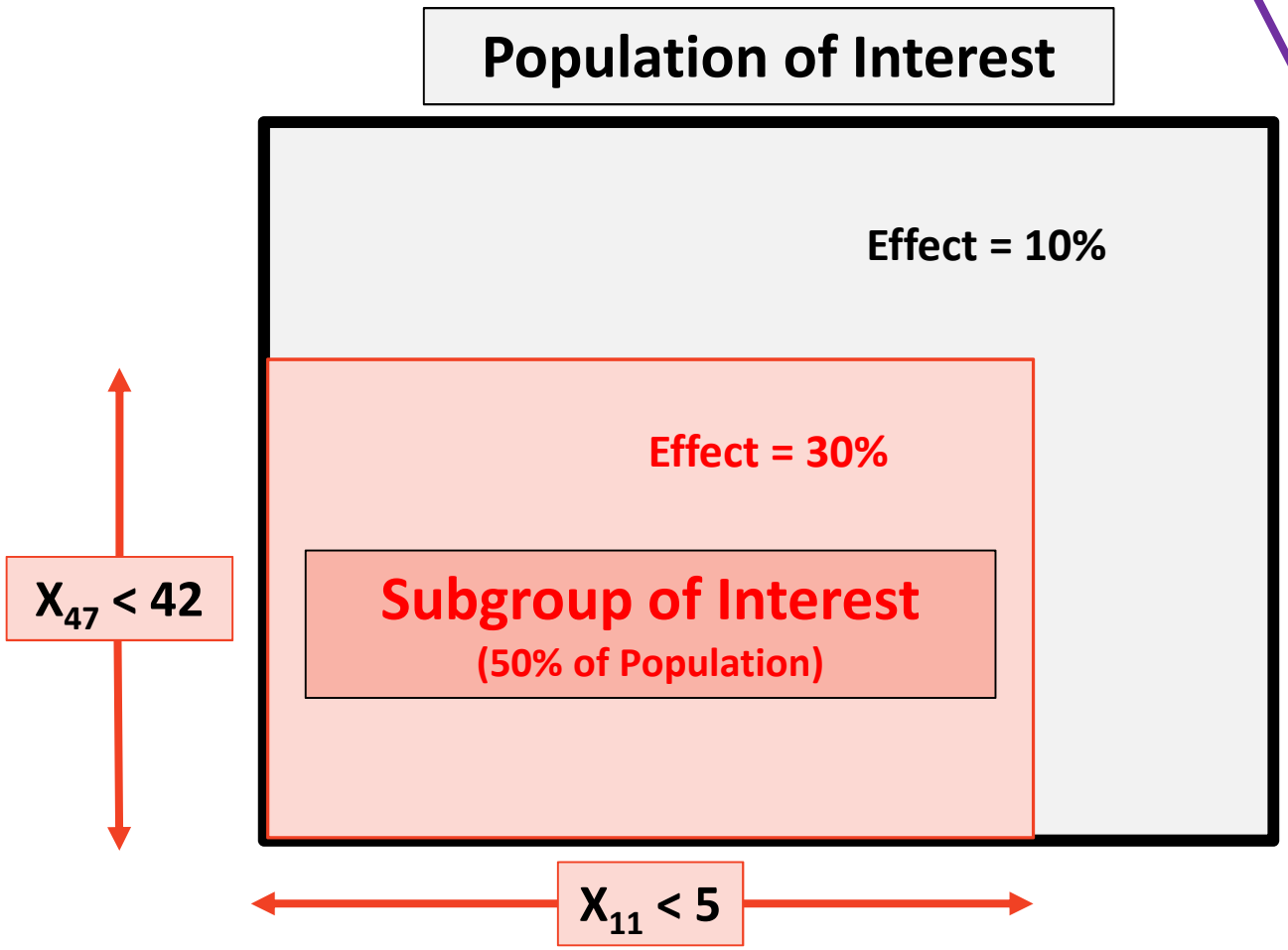
Nature of biomarkers effect

1, 2

Bringing data to life.

Clinical Trials

Treatment versus Control



Complexity

CT Sample Size

100, 250, 500, 1000

Bringing data to life.

Subgroup Identification Challenge

Scenarios (i.e. combinations of possibilities)

3 x 5 x 5 x 2 x 2 x 4 = 1200 !!!

Simulated 1200 datasets with these known parameters.

Posted on ***Innocentive*** and challenged the world …

FIND THE SUBGROUP (i.e. the X's and the cut-offs)

Created a scoring system to rank solutions (0, 100).

Participants could make 1 attempt per day over 3 months.

Subgroup Identification Challenge

Total of 748 entered the competition

- USA 279, India 69, UK 49, Canada 43, Germany 24, Australia 20, Russia 20, Italy 19, Spain 16,
- + 39 other countries (including Seychelles!)

62/120 (52%) did no better than flipping a coin !!!

Only 120 (16%) submitted a valid solution (that could be scored)

Bringing data to life.

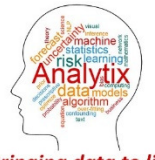
Subgroup Identification Challenge

Internal benchmark score = 62

- This problem is very hard !!

Only two submissions did marginally better with scores of 64 and 65.

- 118/120 (98%) did worse than the internal benchmark

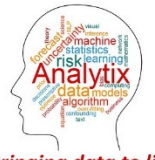


Bringing data to life.

Proceed with Caution

"In … clinical trials, we have a **medium data problem**. It's too big for a human to discern pattern recognition, but not big enough for most algorithms to be able to make sense of it. … **It's the perfect setup to make false discoveries."**

Dr. Donald Bergstrom
Relay Therapeutics
AI for drug development: What's possible and what's just hype
Oct 10, 2018



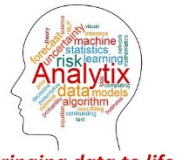
Bringing data to life.

Subgroup Identification Challenge

Clinical Trials

Treatment versus Control

Even with randomized, controlled trials/data, under normal circumstances (i.e. reasonable parameter values) and simple biomarkers relationships to response, **the subgroup is mis-identified (Type 1 error) or not identified (Type 2 error) a high percentage of the time.**



Bringing data to life.

Subgroup *Identification*

Important Distinction

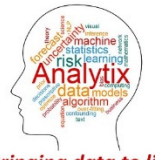
Subgroup *Analysis*

- Post hoc, little concern/control of Type 1 error
- Exploratory - go where the data leads you

Subgroup *Identification*

- Systematic approach
- **Disciplined Subgroup Search***

***Stephen J. Ruberg & Lei Shen (2015) Personalized Medicine: Four Perspectives of Tailored Medicine, Statistics in Biopharmaceutical Research, 7:3, 214-229.**

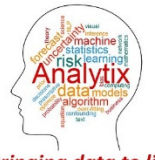


Bringing data to life.

Disciplined Subgroup Search

DSS characteristics

1. **Prespecification**: the algorithm/methodology to be used for identifying subgroups, the list of biomarkers that form the covariate space to be searched, complexity of subgroup definitions (i.e., how many covariates are allowed to define the subgroup), as well as any other options/decisions that can be made in the analysis process.
   - In short, this is no different than prespecification of any important analysis in a Phase 3 trial that adheres to the ICH-E9 Guideline.



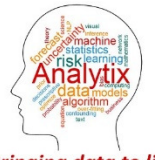
Bringing data to life.

Disciplined Subgroup Search

DSS characteristics

2. Adjusting for multiplicity: how statistical significance (i.e., *p*-values) of a subgroup finding will be adjusted for multiplicity. [Also consider Bayesian approaches.]

3. Bias correction: how estimates of treatment effect are corrected for bias due to the selection bias associated with searching multiple subgroups.



Bringing data to life.

Disciplined Subgroup Search

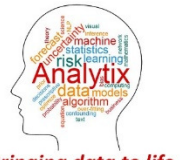
DSS characteristics

4. Biomarker effects: allows for separating prognostic biomarker effects from predictive biomarker effects.

5. Interactions: allows for multiple biomarkers to be included in the definition of a subgroup.

6. Partition: allows for identification of a cut-off value for a continuous biomarker that separates smaller treatment effects from larger treatment effects.

See also
Lipkovich I, Dmitrienko A, D'Agostino BR. (2017) Tutorial in biostatistics: data-driven subgroup identification and analysis in clinical trials. Statistics in Med 36:136-196.

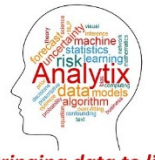


Subgroup Identification

Example

Ramucirumab for HCC

Note: This is a Lilly treatment. I am using only publicly available information in this presentation.



Ramucirumab in HCC

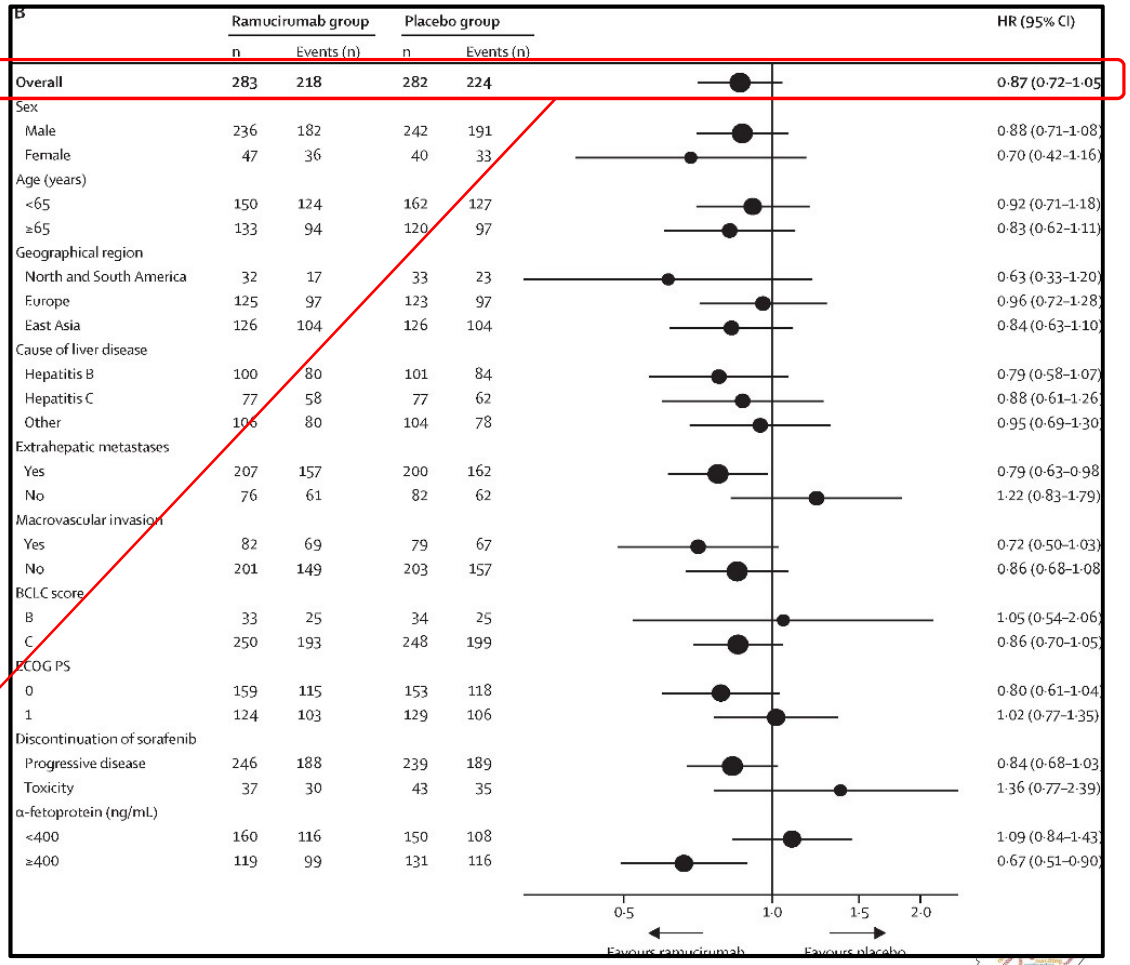
Ramucirumab vs Placebo in HCC (REACH)
Lancet Onc, 2015; 16, 859-870

Not a lot of good treatments for hepatocellular carcinoma

Double-blind, RCT (Phase 3)
N=565 (N_r=383; N_p=382)

Assess OS in ITT population

HR=0.87 (0.72, 1.05)
p=0.14



B	Ramucirumab group		Placebo group		HR (95% CI)
	n	Events (n)	n	Events (n)	
Overall	**283**	**218**	**282**	**224**	**0·87 (0·72–1·05)**
Sex					
Male	236	182	242	191	0·88 (0·71–1·08)
Female	47	36	40	33	0·70 (0·42–1·16)
Age (years)					
<65	150	124	162	127	0·92 (0·71–1·18)
≥65	133	94	120	97	0·83 (0·62–1·11)
Geographical region					
North and South America	32	17	33	23	0·63 (0·33–1·20)
Europe	125	97	123	97	0·96 (0·72–1·28)
East Asia	126	104	126	104	0·84 (0·63–1·10)
Cause of liver disease					
Hepatitis B	100	80	101	84	0·79 (0·58–1·07)
Hepatitis C	77	58	77	62	0·88 (0·61–1·26)
Other	106	80	104	78	0·95 (0·69–1·30)
Extrahepatic metastases					
Yes	207	157	200	162	0·79 (0·63–0·98)
No	76	61	82	62	1·22 (0·83–1·79)
Macrovascular invasion					
Yes	82	69	79	67	0·72 (0·50–1·03)
No	201	149	203	157	0·86 (0·68–1·08)
BCLC score					
B	33	25	34	25	1·05 (0·54–2·06)
C	250	193	248	199	0·86 (0·70–1·05)
ECOG PS					
0	159	115	153	118	0·80 (0·61–1·04)
1	124	103	129	106	1·02 (0·77–1·35)
Discontinuation of sorafenib					
Progressive disease	246	188	239	189	0·84 (0·68–1·03)
Toxicity	37	30	43	35	1·36 (0·77–2·39)
α-fetoprotein (ng/mL)					
<400	160	116	150	108	1·09 (0·84–1·43)
≥400	119	99	131	116	0·67 (0·51–0·90)

Bringing data to life.

Ramucirumab in HCC

Ramucirumab vs Placebo in HCC (REACH)
Lancet Onc, 2015; 16, 859-870

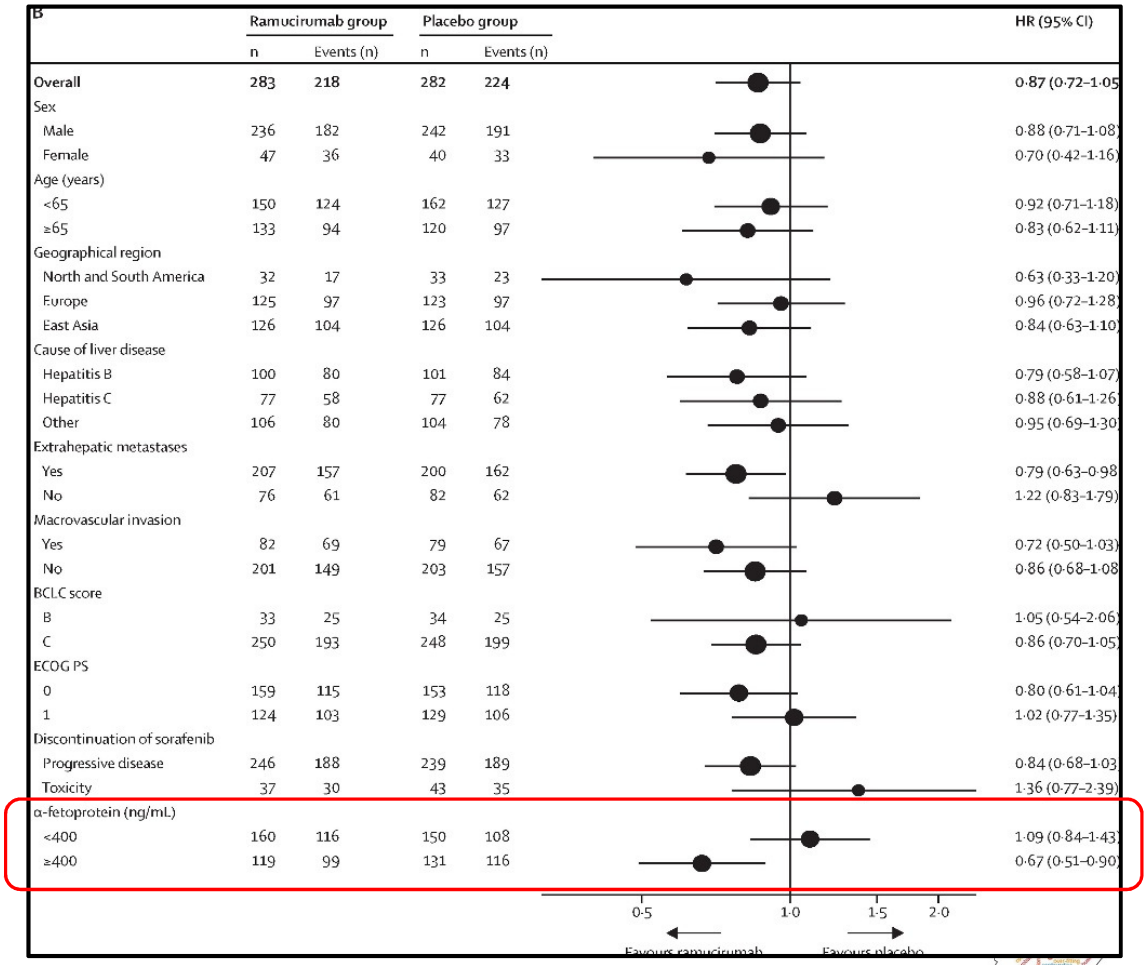
Not a lot of good treatments for hepatocellular carcinoma

Double-blind, RCT (Phase 3)
N=565 (N_r=383; N_p=382)

Assess OS in ITT population

10 subgroups pre-specified

Interaction: p=0.024
HR_{400+}=0.67 (0.51, 0.90)
p=0.006



Bringing data to life.

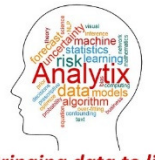
Ramucirumab in HCC

What are we to believe from these results?

Is the AFP finding spurious or real?

Could Subgroup Identification have been used to obtain confirmatory results?

If so, how?

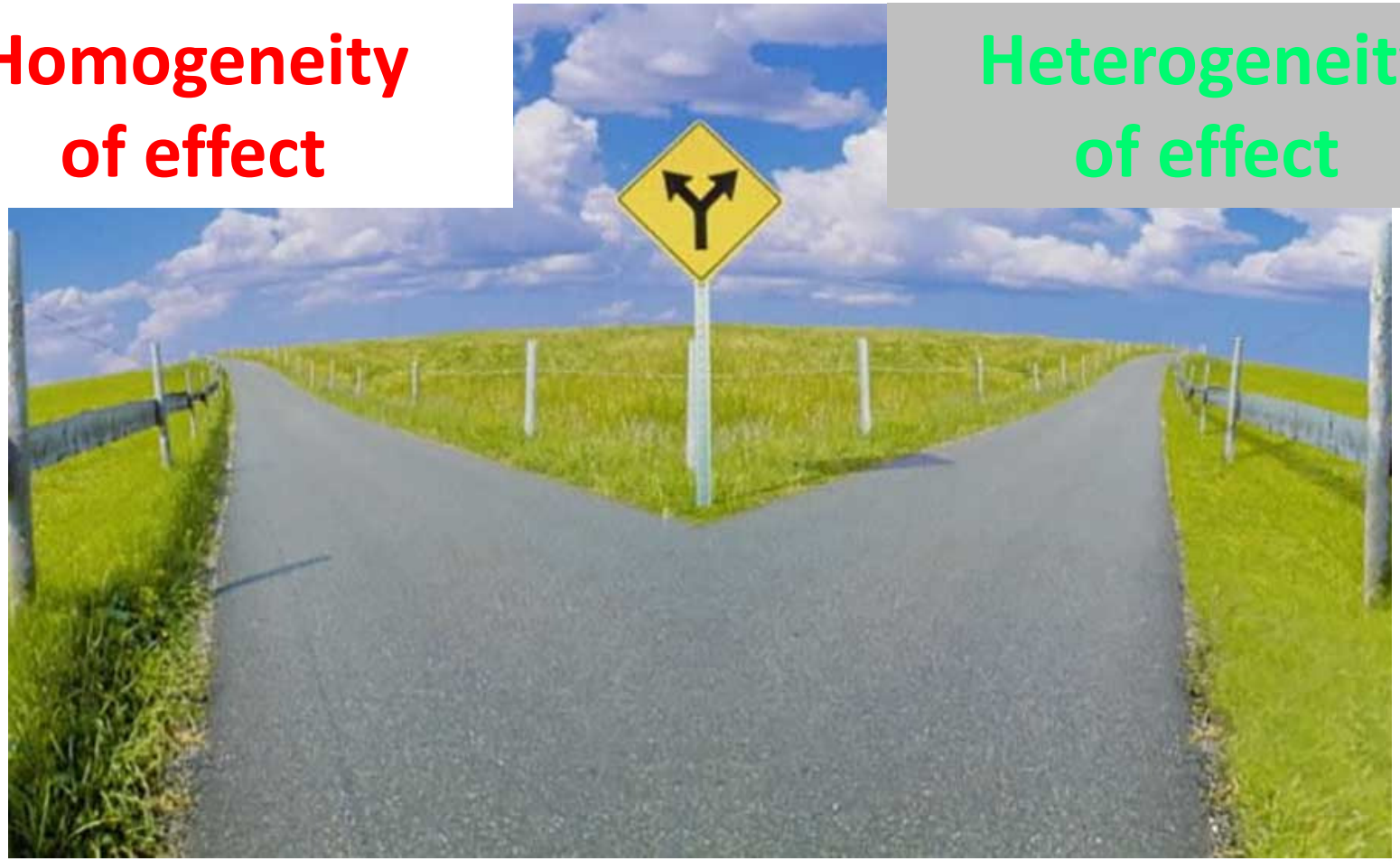


Bringing data to life.

Subgroup Identification

Homogeneity of effect

Heterogeneity of effect



Bringing data to life.

Ramucirumab in HCC

10 Subgroups defined and reported *a priori*

Sex
Age
Region

Homogeneity of effect

Cause of Liver Disease
Extrahepatic Metastases
Macrovascular invasions
BCLC Score
ECOG PS
Discontinuation of sorafenib
AFP

Heterogeneity of effect

Bringing data to life.

Homogeneity of Effect

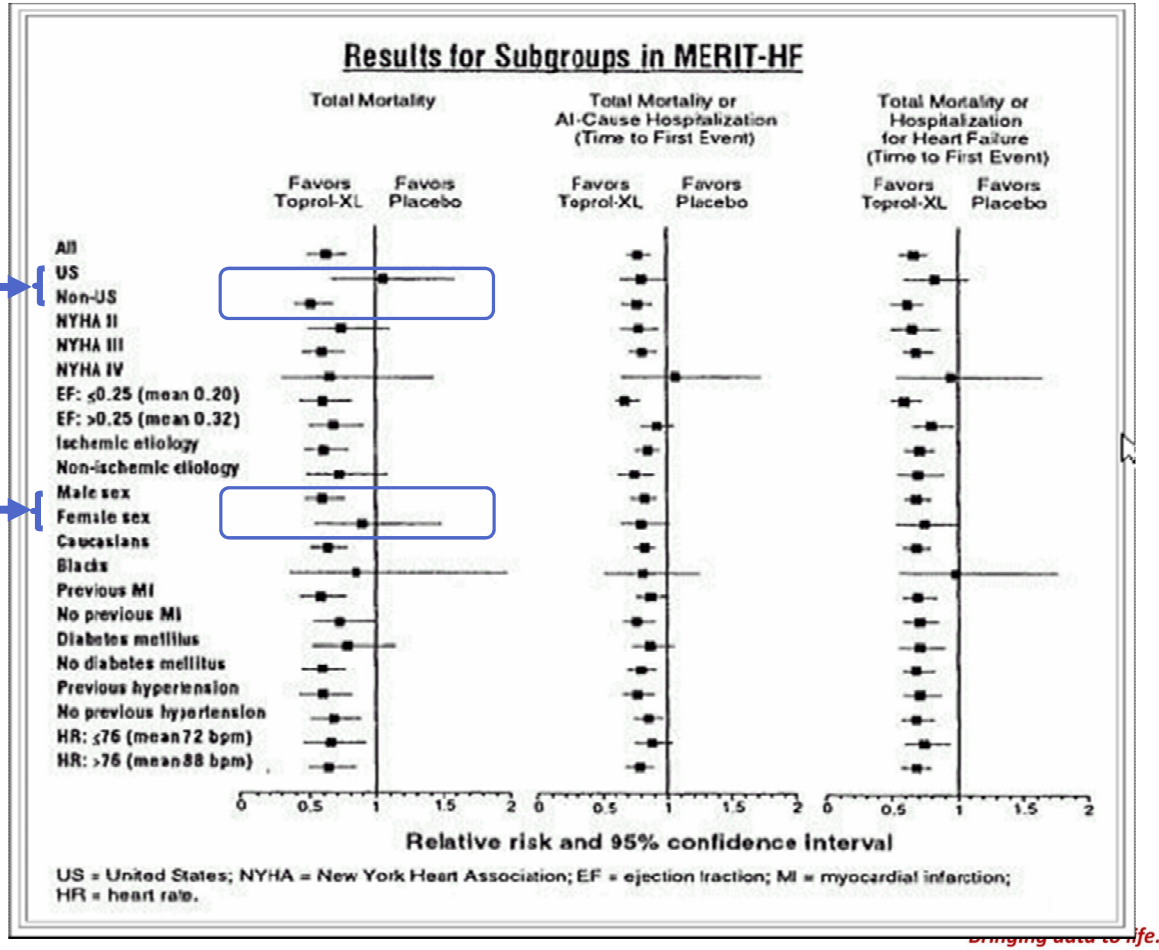
Homogeneity of effect

"Routine" baseline factors

Playing "defense"

Avoid trying to explain away (unusual?) findings

Toprol-XL Label



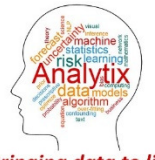
Homogeneity of Effect

From US Label for Toprol-XL

"The figure … illustrates principal results for a wide variety of subgroup comparisons, including US vs. non-US populations (the latter of which was not pre-specified).

… subgroup analyses can be difficult to interpret, and it is not known whether these represent true differences or chance effects."

Use Disciplined Subgroup Search!

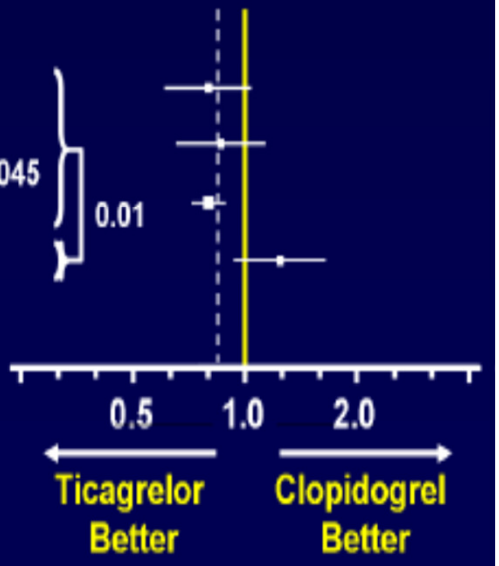


Bringing data to life.

Homogeneity of Effect

Ticagrelor Example

Characteristic	Total Patients	# of Events Tic	# of Events Clop	HR (95% CI)	Interaction p-value
Geographic Region					
Asia / Australia	1714	95	116	0.80 (0.61, 1.04)	
Cent / Sth America	1237	91	104	0.86 (0.65, 1.13)	
Euro / Md E / Afr	13859	576	712	0.80 (0.72, 0.90)	0.045
North America	1814	102	82	1.25 (0.93, 1.67)	0.01



Source: Sponsor presentation at CV and Renal Drugs Ad Comm Meeting July, 2010 CC-30

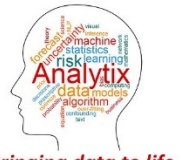
Bringing data to life.

Homogeneity of Effect

From US Label for Ticagrelor

"The individual results and nominal p-values, like all subset analyses, need cautious interpretation, and they could represent chance findings."

Use Disciplined Subgroup Search!



Bringing data to life.

Homogeneity of Effect

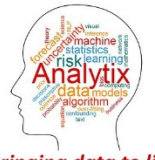
From US Label for Benlysta®

"Exploratory sub-group analyses of SRI response rate in patients of black race were performed.

… the SRI response rate in black patients … was less than that in the placebo group.

… Although no definitive conclusions can be drawn from these subgroup analyses, caution should be used when considering BENLYSTA treatment in black/African-American SLE patients."

Use Disciplined Subgroup Search!



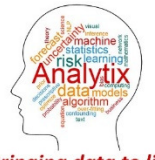
Bringing data to life.

Subgroup Identification

Always do subgroup identification !!!

What if DSS had been formally done?

- Often subgroups defined by baseline factors are described in the protocol (e.g. gender, race, baseline severity, etiology, etc.?
- What if the subgroup identification search methodology was pre-specified?
- What if *adjusted* p-values and effect estimates were calculated?
- *Would these "**surprising**" findings not be so confusing anymore?*

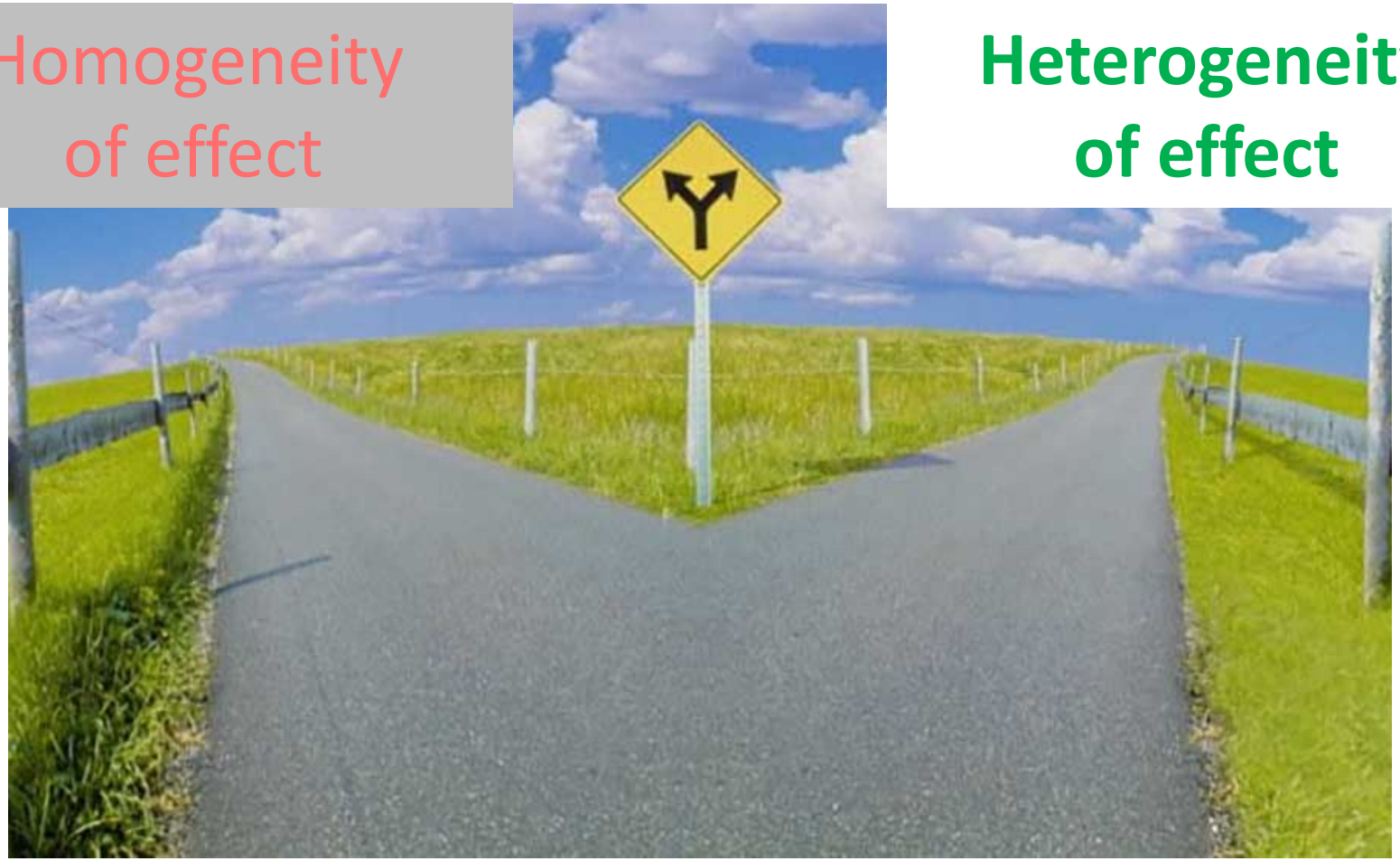


Bringing data to life.

Subgroup Identification

Homogeneity of effect

Heterogeneity of effect



Bringing data to life.

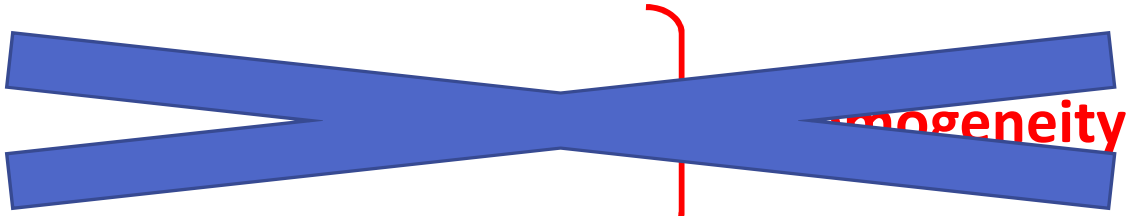
Ramucirumab in HCC

10 Subgroups defined and reported *a priori*

Sex

Age

Region



mogeneity of effect

Cause of Liver Disease

Extrahepatic Metastases

Macrovascular invasions

BCLC Score

ECOG PS

Discontinuation of sorafenib

AFP

Adjust for K=7
Baseline Factors

Heterogeneity of effect

Bringing data to life.

Ramucirumab in HCC

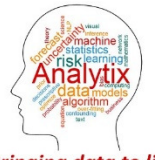
Multiple Comparisons Procedures

- AFP was NOT part of a formal multiplicity plan
- There was NO DSS procedure defined to examine all subgroups

HOWEVER, …

AFP was pre-specified and was a known strong *prognostic* biomarker for survival

- Could Type 1 Error (α) have been spent judiciously in order to have a statistically significant finding?



Bringing data to life.

Ramucirumab in HCC

Multiple Comparisons Approaches

Split α

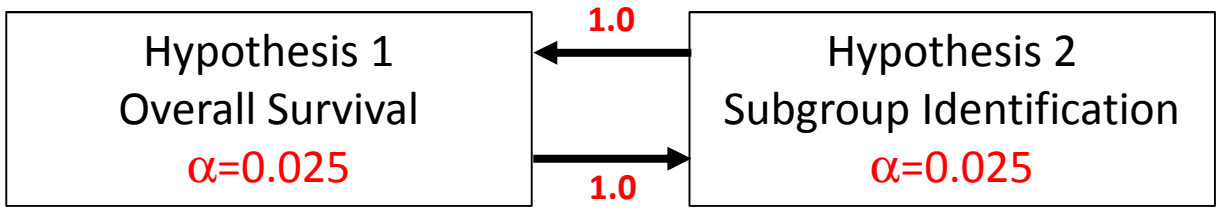
Hypothesis 1 Overall Survival α=0.04	Hypothesis 2 Subgroup Identification α=0.01
Subgroup definition non-descript	Nominal interaction p-value = 0.024 👎
Subgroup definition explicit – i.e. AFP 400+ will respond better	AFP 400+ nominal p-value = 0.006 Unlikely to "survive" multiplicity adjustment for all 7 subgroups investigated 👎

Bringing data to life.

Ramucirumab in HCC

Multiple Comparisons Approaches

Graphical



Subgroup definition non-descript

Nominal interaction p-value = 0.024
Unlikely to "survive" multiplicity adjustment for all 10 subgroups investigated 👎

Subgroup definition explicit – i.e. AFP 400+ will respond better

AFP 400+ nominal p-value = 0.006
"Survive" multiplicity adjustment for all 7 subgroups investigated ??? Maybe ???

Bringing data to life.

Ramucirumab in HCC

Interpretation

Regulatory

Not enough evidence for a regulatory approval

Company

How do we know what to believe?

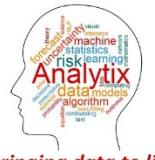
Should we proceed?

If we proceed, what is the likelihood of success?

Bringing data to life.

Ramucirumab in HCC

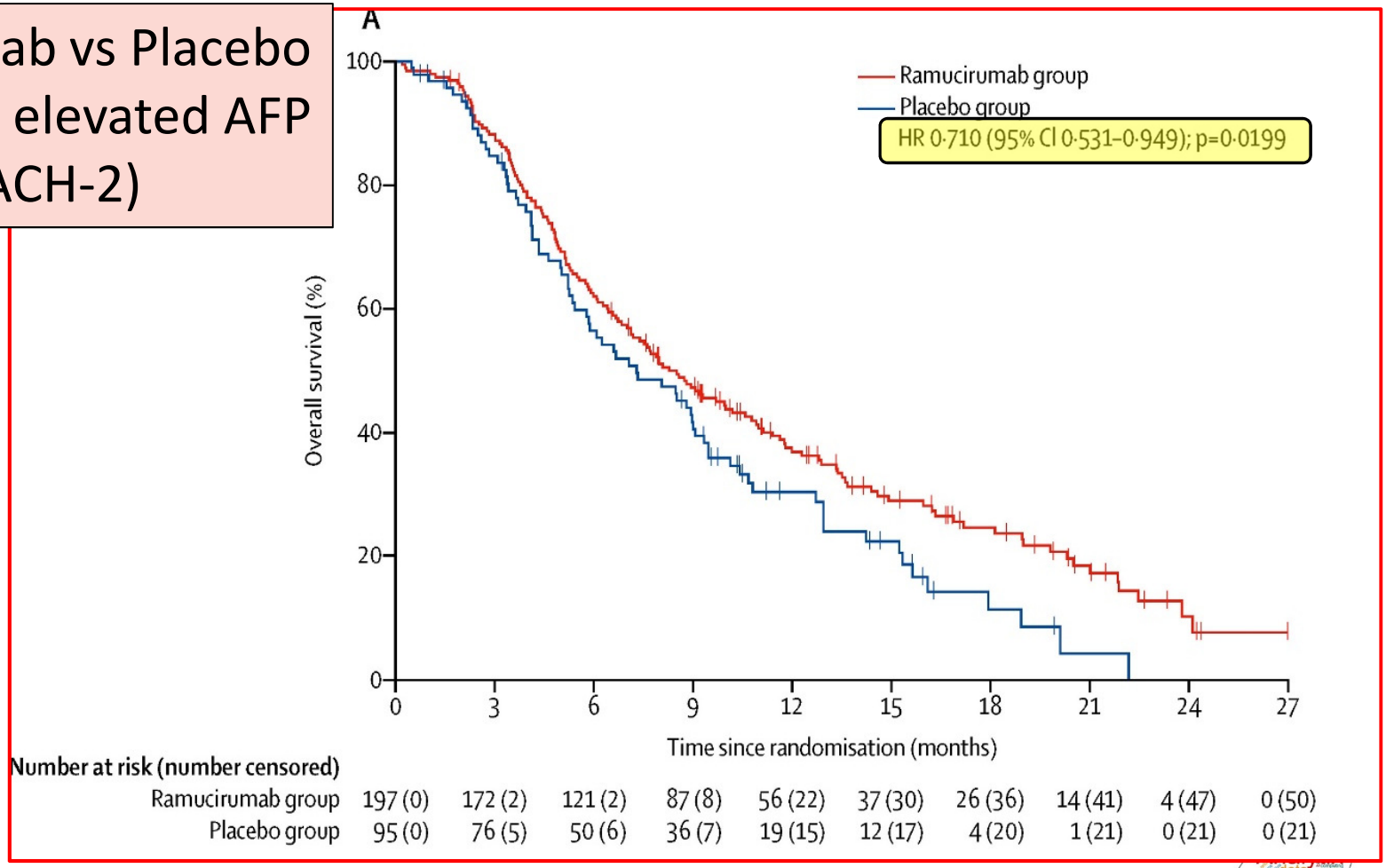
The Rest of the Story



Bringing data to life.

Ramucirumab in HCC

Ramucirumab vs Placebo in HCC with elevated AFP (REACH-2)



Lancet Onc, 2019; 20, 282-296

Bringing data to life.

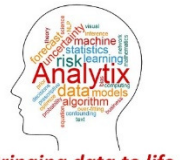
Ramucirumab in HCC

Always do subgroup identification !!!

What if DSS had been formally done in REACH?

- What if the AFP subgroup was pre-specified along with other subgroups?
- What if the subgroup identification search methodology was pre-specified?
- What if *adjusted* p-values and effect estimates were calculated?
- What if they were still significant and meaningful?

But … DSS is hard !!!!!



Bringing data to life.

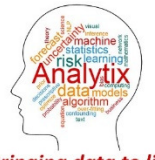
Ramucirumab in HCC

The $1,000,000,000 [read billion] question …

Could ramucirumab have been approved in the targeted subgroup based on REACH in 2015 instead of

- Spending 4 years, and
- Many, many millions of dollars, and
- Tens of thousands of patients not having access to an effective medication?

Is DSS a billion dollars hard?!?!?!?



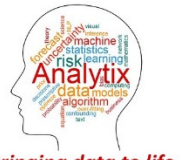
Bringing data to life.

Subgroup Identification

"Always do *subgroup identification using DSS* so the results are more interpretable."

Steve Ruberg

Your Run-of-the-Mill Statistician



Subgroup Identification

Bayesian Thinking

Frequentist Thinking

Bringing data to life.

Bayesian Thinking

Suppose there are **100** potential predictive biomarkers that could be important for a new treatment.

- **100** hypothesis tests of each biomarker

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

- Bonferroni adjusted p-value ≤ **100** * **0.0001** = **0.01**

EUREKA! We have discovered a novel biomarker-defined subgroup.

Bringing data to life.

Bayesian Thinking

ARE YOU SURE?

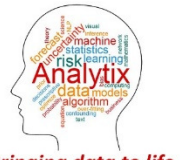
Suppose further

pr(success … finding a biomarker)

= pr(at least one 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**



Bringing data to life.

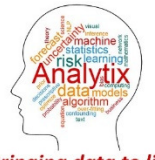
Bayesian Thinking

Let p_0 = prior probability that H_0 is false (e.g. the biomarker is predictive)

Let p = observed p-value for test statistics for H_0

Bayes factor* $\mathbf{[-e \times p \times ln(p)]^{-1}}$ can be used to give an upper bound on the posterior probability that H_0 is false

***Sellke et al (2001) Calibration of p Values for Testing Precise Null Hypotheses. The American Statistician, February 2001, Vol. 55, No. 1, pp 62-71.**



Bringing data to life.

Bayesian Thinking

Let p_0 = prior probability that H_0 is false (e.g. the biomarker is predictive)

Let p = observed p-value for test statistics for H_0

Bayes factor* $\mathbf{[-e \times p \times \ln(p)]^{-1}}$ can be used to give an upper bound on the posterior probability that H_0 is false

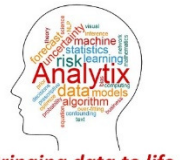
Posterior probability** for H_0 being false ($\mathbf{p_1}$) is (upper bound)

$$\mathbf{p_1} \leq \{1 + [(1-p_0)/p_0] \times [-e \times p \times \ln(p)]\}^{-1}$$

Posterior ← Prior ← New Data

*Sellke et al (2001) Calibration of p Values for Testing Precise Null Hypotheses. The American Statistician, February 2001, Vol. 55, No. 1, pp 62-71.

**If $p < 1/e = .368$



Bringing data to life.

Bayesian Thinking

ARE YOU SURE?

$p_0 = 0.002$ (uniform prior across 100 biomarkers)

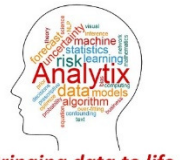
p = 0.0001 (from hypothesis test)

Recall Bonferroni adjusted p = 0.01

$$p_1 \leq \{1 + [(1-p_0)/p_0] \times [-e \times p \times \ln(p)] \}^{-1}$$

Bayesian posterior pr(H_0 is false) ≤ 0.44.

Berger J.O., Wang X., Shen L. (2014). A Bayesian approach to subgroup identification. *J Biopharm Stat,* 24(1), 110-29.



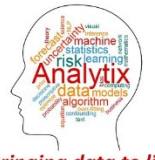
Bringing data to life.

Subgroup Identification

"Always use *Bayesian thinking* when doing subgroup identification so you can quantify how believable the results are."

Steve Ruberg

Your Run-of-the-Mill Bayesian Statistician



Bringing data to life.

Bayes and Ramucirumab

Suppose my prior is …

Pr(ramucirumab works in HCC) = 0.70

- Pr(ram works in all patients) = 0.50
- Pr(ram works in a subgroup) = 0.20
  - Pr(ram works in AFP400+) = 0.10
  - Pr(ram works in another subgroup) = 0.10

Recall: p=0.006 for AFP400+ subgroup

Posterior pr(ram works in AFP 400+) ≤ **0.57**

Bringing data to life.

Bayesian Thinking

Other Examples

Dalcetrapib – CV outcomes and genotypes

AnalytixThinking.Blog: Genetic Subgroups and CV Disease

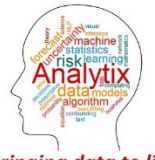
Solanezumab – Mild Alzheimer's patients

AnalytixThinking.Blog: Subgroups, Multiplicity and Bayes – A Case Study

AnalytixThinking.Blog

Bringing data to life.

4. Conclusion



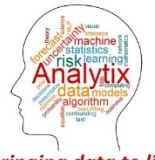
Bringing data to life.

Conclusion

Subgroup identification is the HOLY GRAIL.

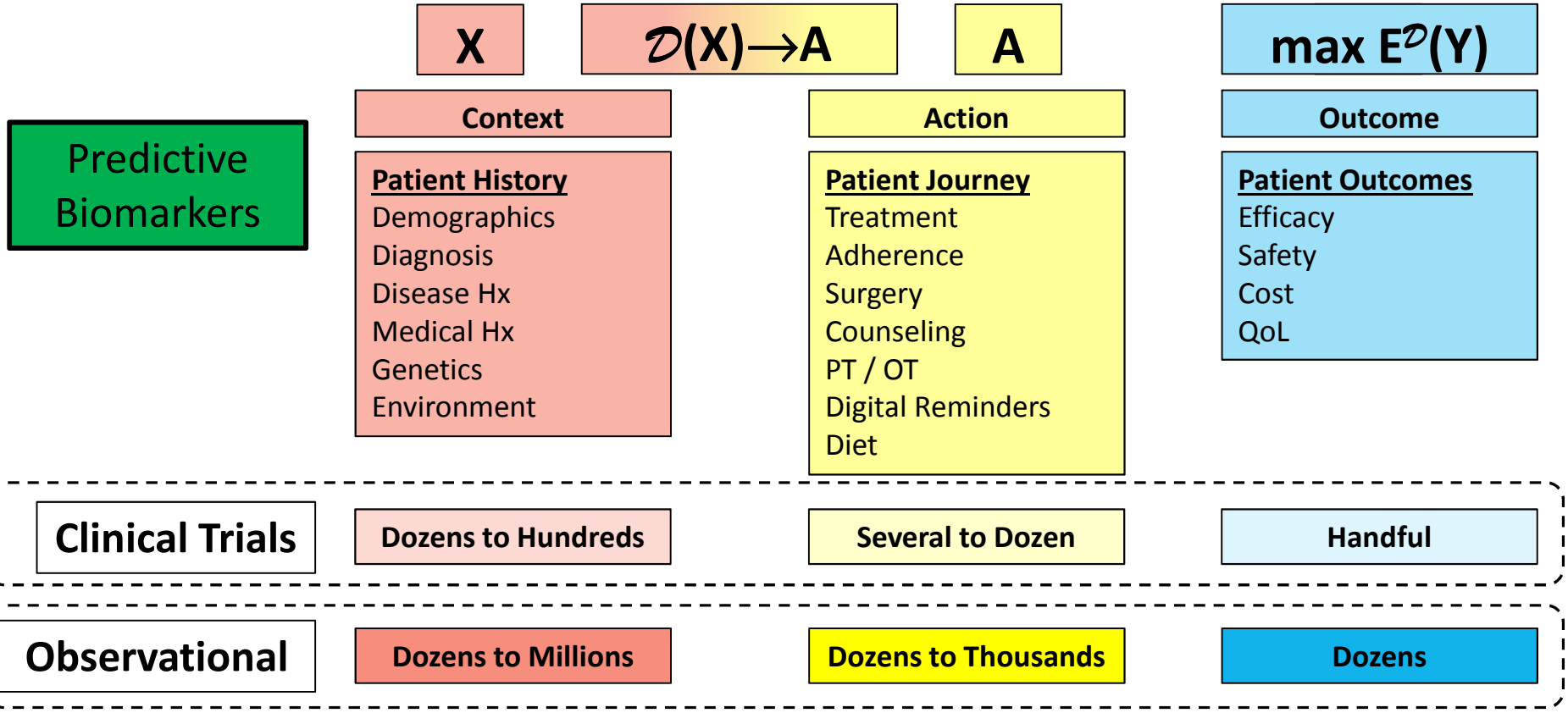
Not surprisingly, that makes it **the hardest problem there is**.

Dimensionality is enormous!



Bringing data to life.

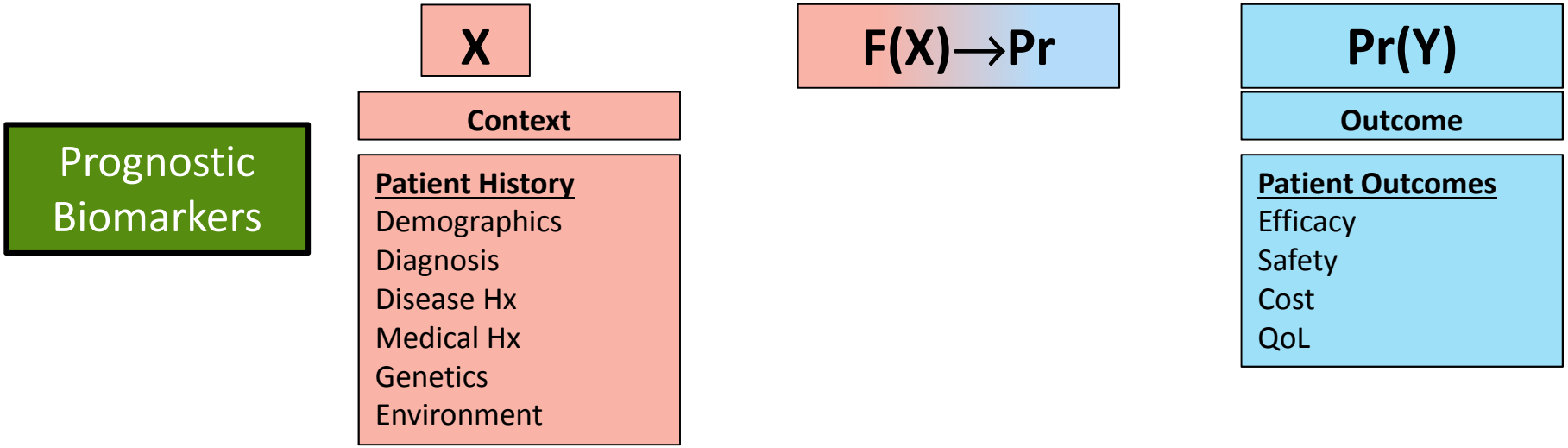
Conclusion



Individualized Treatment Regimes

Thanks to Haoda Fu

Conclusion


Prognostic Biomarkers
X
Context
Patient History
Demographics
Diagnosis
Disease Hx
Medical Hx
Genetics
Environment
F(X)→Pr
Pr(Y)
Outcome
Patient Outcomes
Efficacy
Safety
Cost
QoL
Observational
Dozens to Millions
Dozens

Prognostic Biomarkers

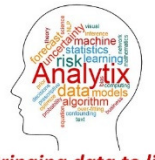
Data Analytical Scientists

need to think more like

Bio-Analytical Scientists

Assay Validation

Development of Diagnostic Tests



Bringing data to life.

Predictive Biomarkers

ALWAYS do Subgroup Identification!

(for trials of suitable size)

Do You Want To Find a Subgroup or Not?

NO ←

→ **YES**

Rule out spurious heterogeneity

Disciplined Subgroup Search

Confirm suspected heterogeneity

(replication and biological plausibility are very important)

(don't forget safety assessments as well)

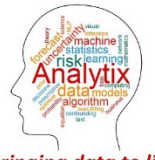
Bringing data to life.

Bayesian Thinking

Whenever I see a **significant finding**, I always ask …

"I wonder what their prior was?"

A Bayesian approach can help to quantify the likelihood of a finding being real.



Bringing data to life.

Conclusion

"If you don't know where you are going, any path will do."

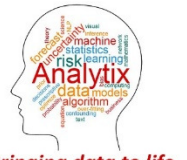
Lewis Carroll

Author of *Alice in Wonderland*

If you don't know what you are asking, any answer could be true.

Steve Ruberg

Author of …



Bringing data to life.

I hope that at least a SUBGROUP of you found this interesting, informative and possibly enlightening.

THANK YOU

AnalytixThinking.Blog

Bringing data to life.