Numbersense: Clearing the Fog of Big Data

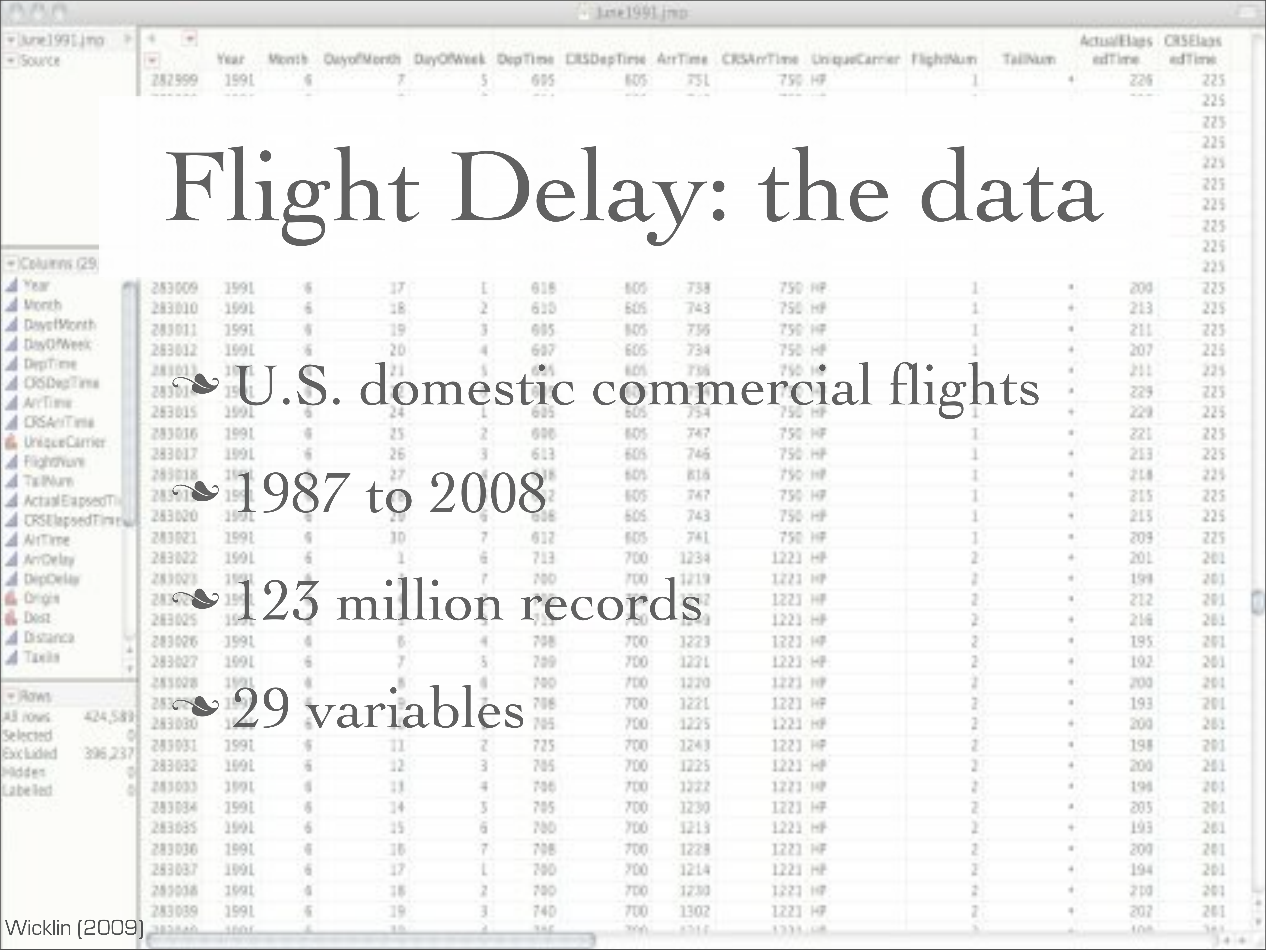
Kaiser Fung
INFORMS NYC Luncheon
9/18/2013

Big Data studies

- Observational data
- Co-opted
- Seemingly exhaustive N
- Fused data
- No controls

Flight Delay: the data

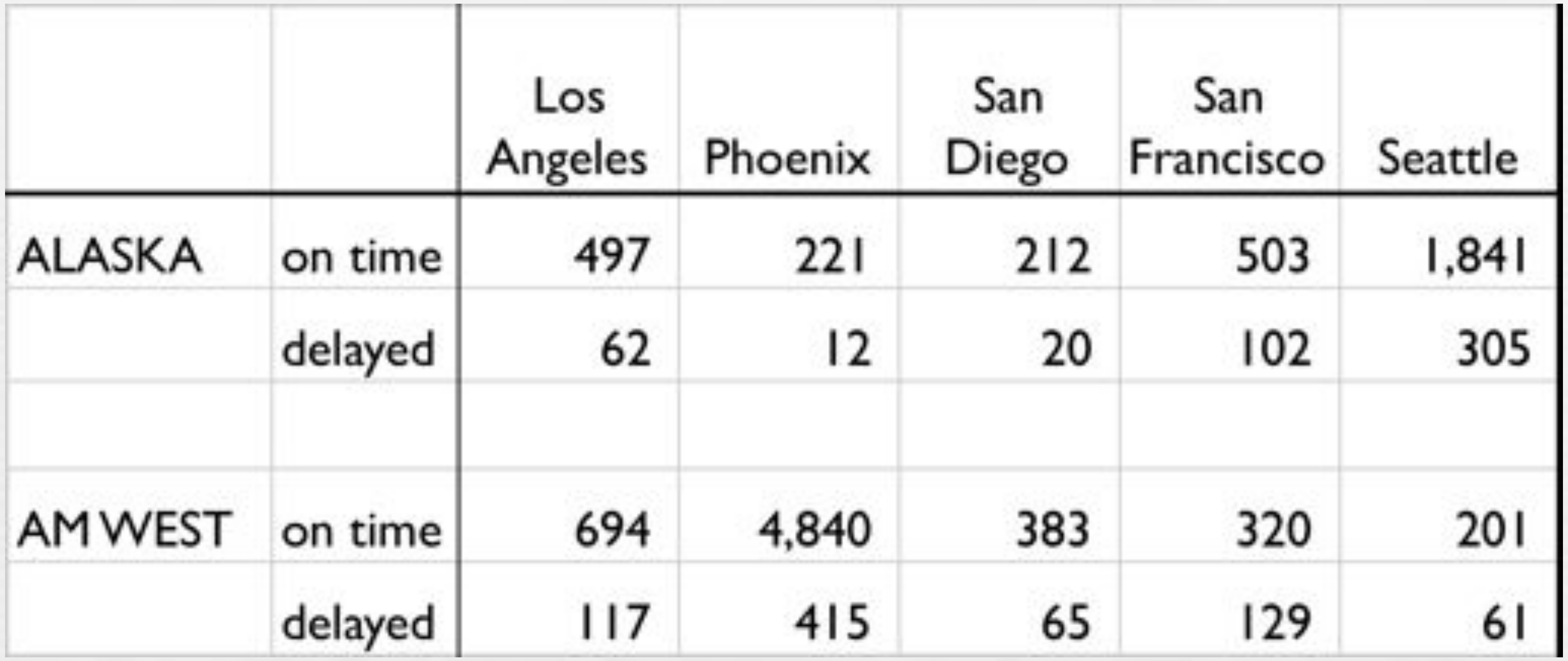
- U.S. domestic commercial flights
- 1987 to 2008
- 123 million records
- 29 variables



Wicklin (2009)

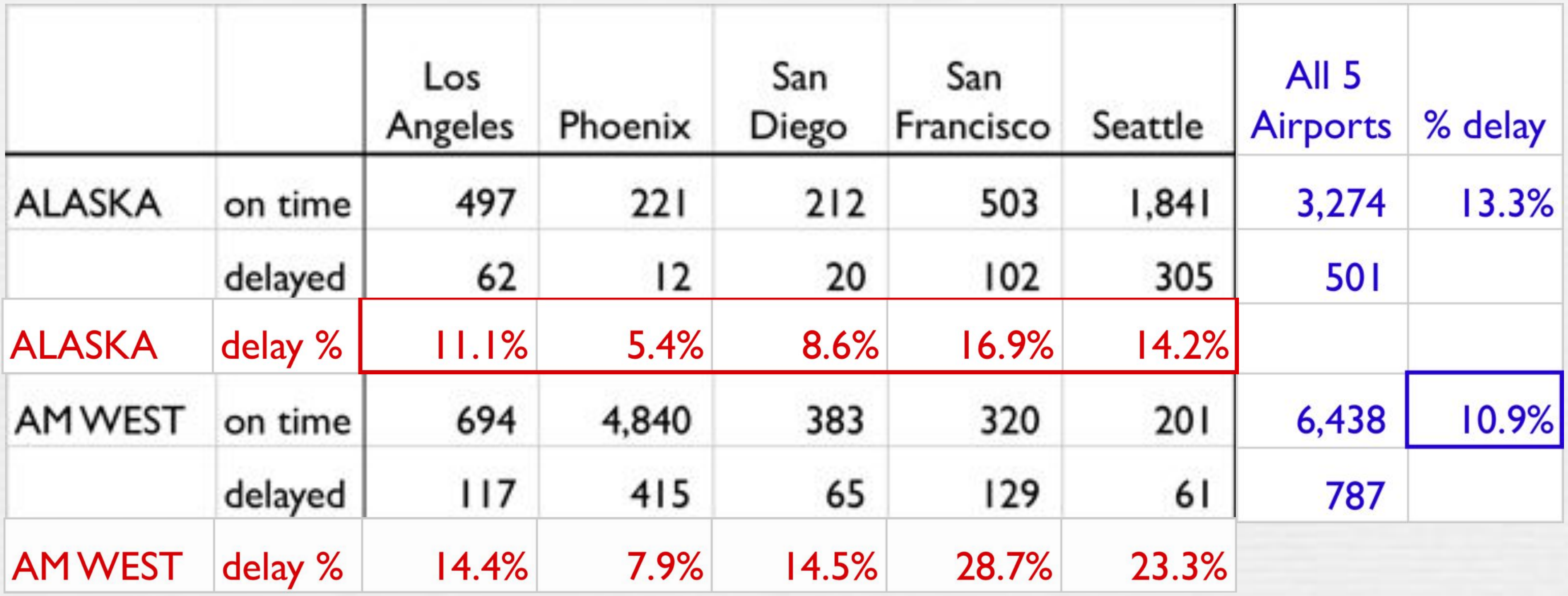
Monday, September 23, 2013

Which airline had a lower delay rate?



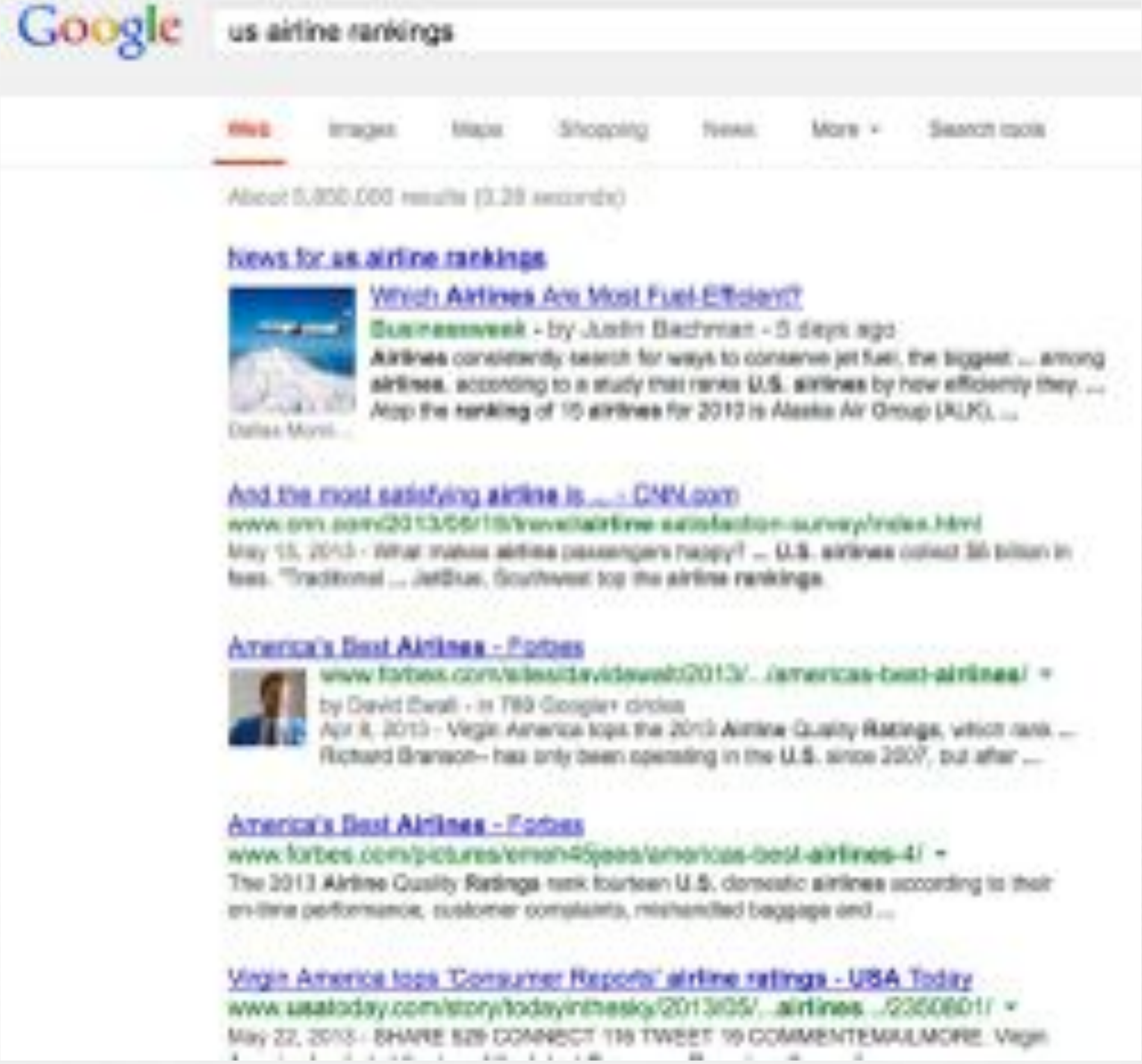
		Los Angeles	Phoenix	San Diego	San Francisco	Seattle
ALASKA	on time	497	221	212	503	1,841
	delayed	62	12	20	102	305
AM WEST	on time	694	4,840	383	320	201
	delayed	117	415	65	129	61

Which airline had a lower delay rate?



		Los Angeles	Phoenix	San Diego	San Francisco	Seattle	All 5 Airports	% delay
ALASKA	on time	497	221	212	503	1,841	3,274	13.3%
	delayed	62	12	20	102	305	501	
ALASKA	delay %	11.1%	5.4%	8.6%	16.9%	14.2%		
AM WEST	on time	694	4,840	383	320	201	6,438	10.9%
	delayed	117	415	65	129	61	787	
AM WEST	delay %	14.4%	7.9%	14.5%	28.7%	23.3%		

Ask the “right” question



us airline rankings

About 5,800,000 results (0.28 seconds)

News for us airline rankings

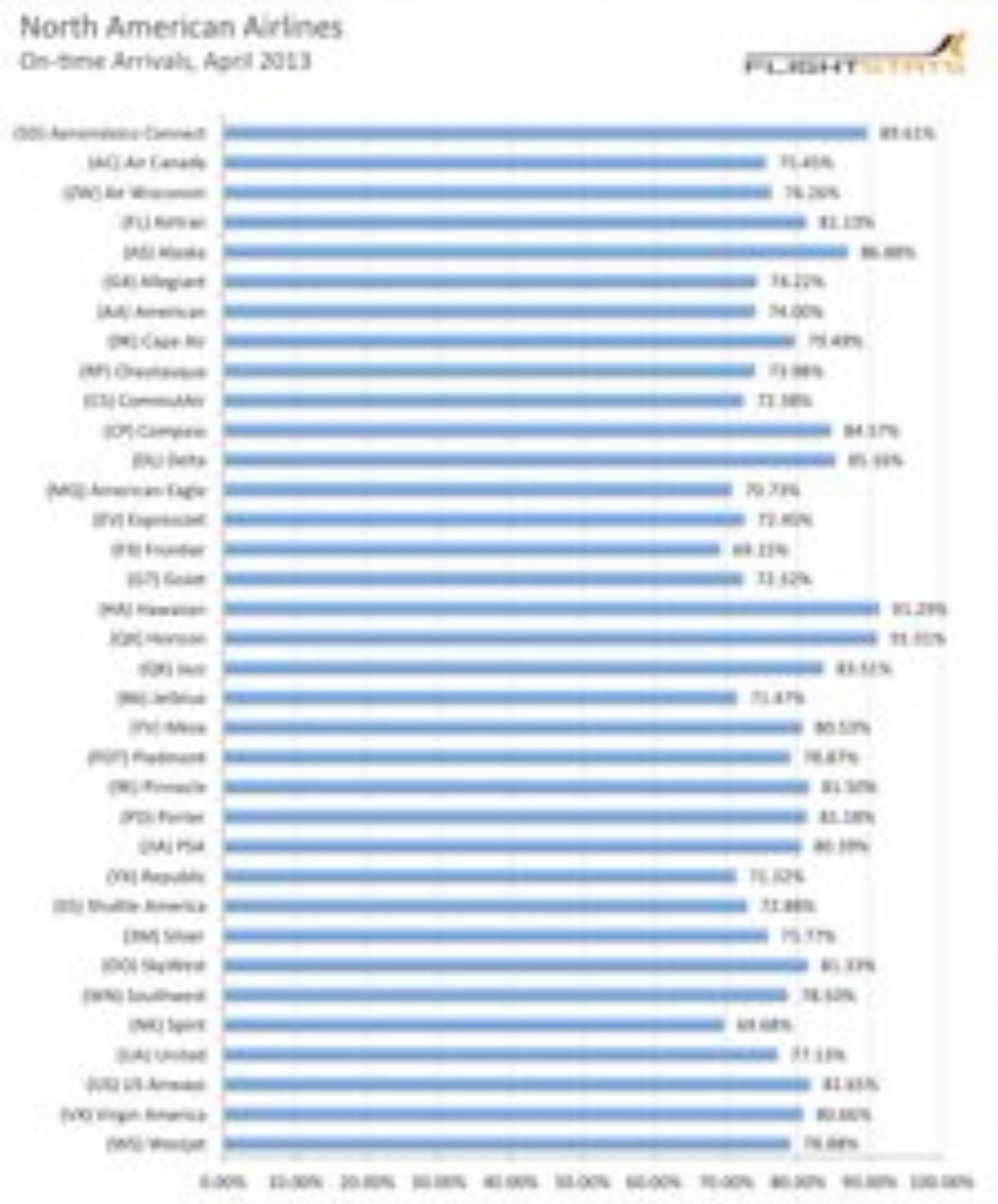
Which Airlines Are Most Fuel-Efficient?
Businessweek - by Justin Bachman - 5 days ago
Airlines consistently search for ways to conserve jet fuel, the biggest ... among airlines, according to a study that ranks U.S. airlines by how efficiently they. ... Atop the ranking of 15 airlines for 2013 is Alaska Air Group (ALK), ...
Dallas Morn...

And the most satisfying airline is ... - CNN.com
www.cnn.com/2013/05/15/travel/airline-satisfaction-survey/index.html
May 15, 2013 - What makes airline passengers happy? ... U.S. airlines collect $6 billion in fees. "Traditional ... JetBlue, Southwest top the airline rankings.

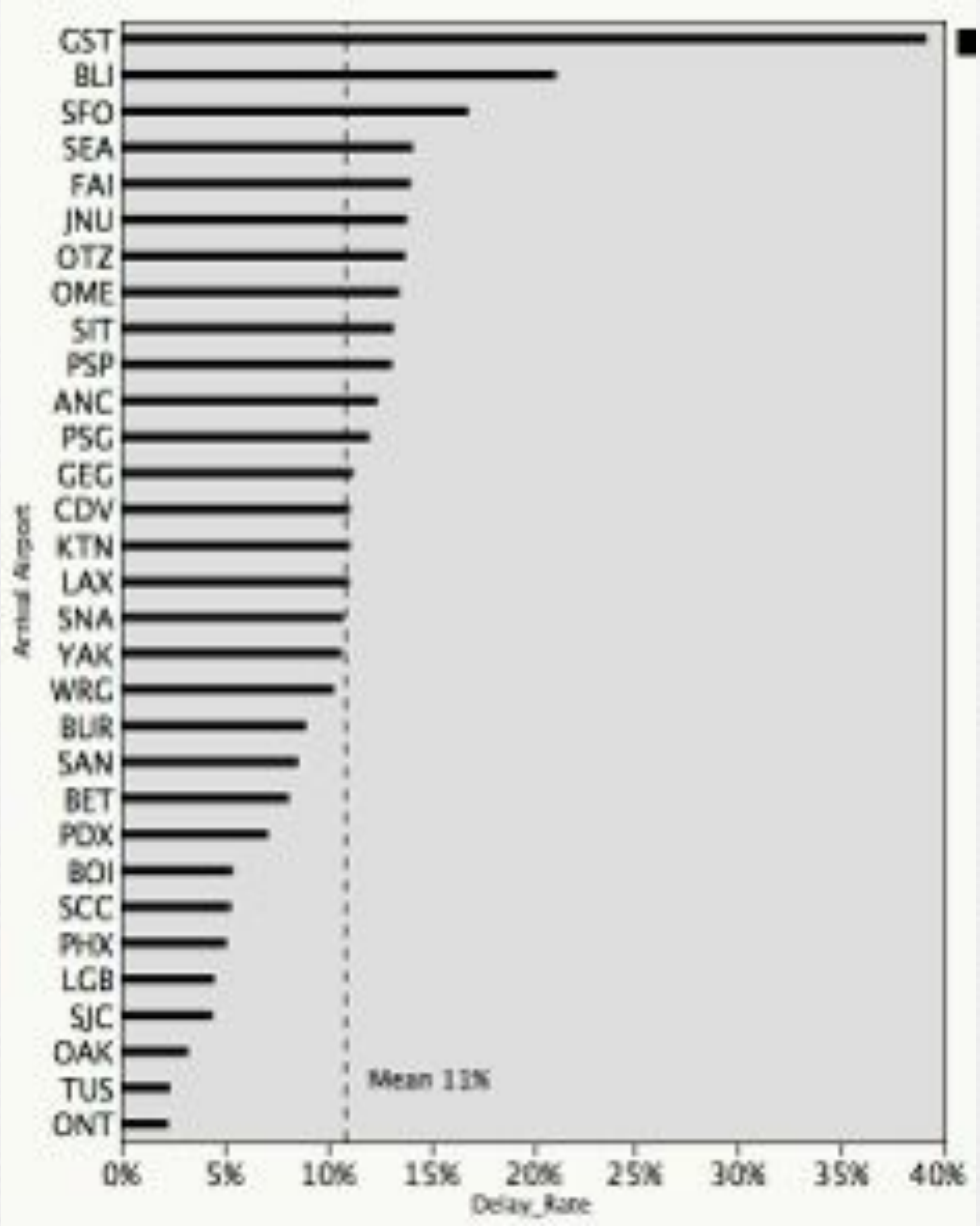
America's Best Airlines - Forbes
www.forbes.com/sites/davidewalt/2013/.../americas-best-airlines/ ▾
by David Ewalt - in 789 Google+ circles
Apr 8, 2013 - Virgin America tops the 2013 Airline Quality Ratings, which rank ... Richard Branson-- has only been operating in the U.S. since 2007, but after ...

America's Best Airlines - Forbes
www.forbes.com/pictures/emeh45eeis/americas-best-airlines-4/ ▾
The 2013 Airline Quality Ratings rank fourteen U.S. domestic airlines according to their on-time performance, customer complaints, mishandled baggage and ...

Virgin America tops 'Consumer Reports' airline ratings - USA Today
www.usatoday.com/story/todayinthesky/2013/05/...airlines.../2350801/ ▾
May 22, 2013 - SHARE 529 CONNECT 118 TWEET 19 COMMENTEMAILMORE. Virgin



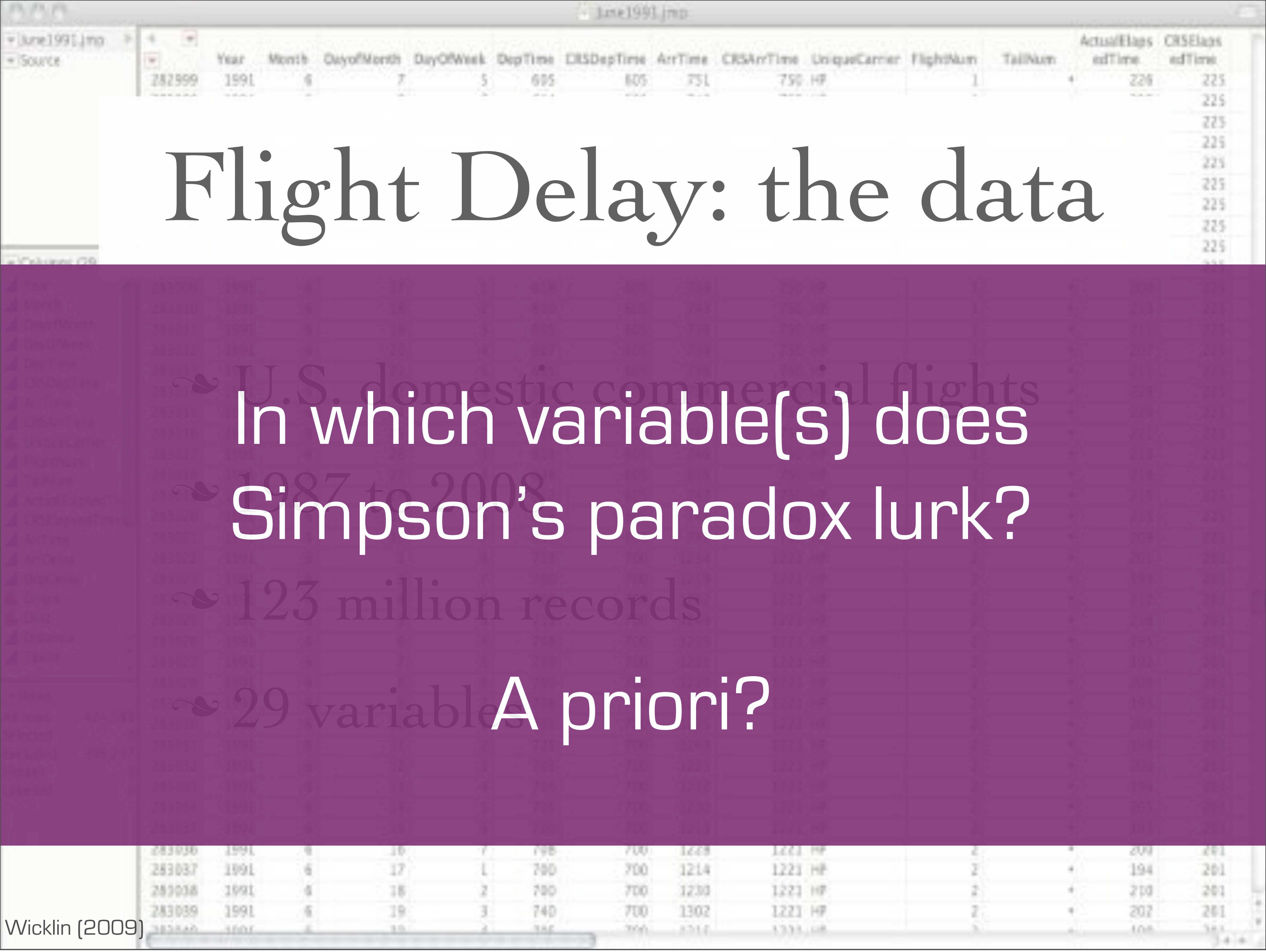
Alaska's On-time Performance



10.9%

Alaska is the industry leader in on-time flights

Flight Delay: the data



June1991.jmp

	Year	Month	DayofMonth	DayOfWeek	DepTime	CRSDepTime	ArrTime	CRSArrTime	UniqueCarrier	FlightNum	TailNum	ActualElapsedTime	CRSElapsedTime
282999	1991	6	7	5	605	605	751	750	HP	1	•	226	225
283036	1991	6	16	7	708	700	1228	1221	HP	2	•	209	201
283037	1991	6	17	1	700	700	1214	1221	HP	2	•	194	201
283038	1991	6	18	2	700	700	1230	1221	HP	2	•	210	201
283039	1991	6	19	3	740	700	1302	1221	HP	2	•	202	201

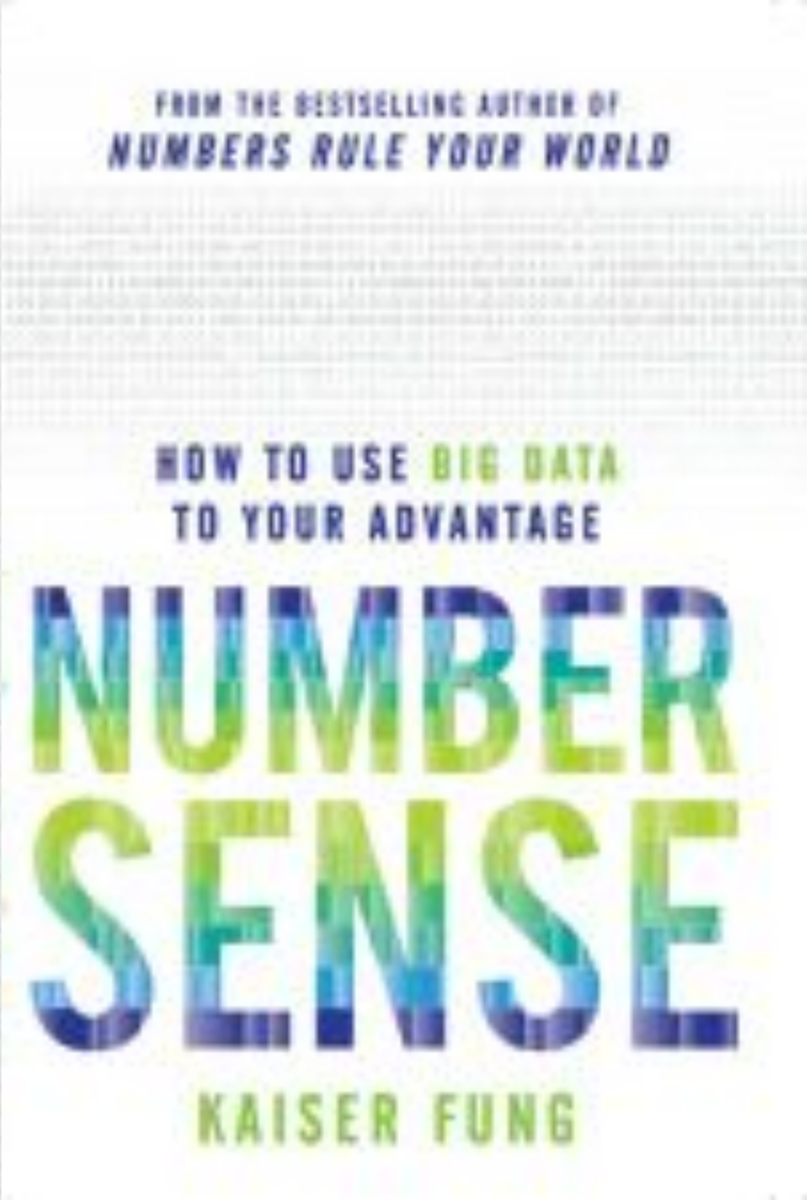
- U.S. domestic commercial flights
- 1987 to 2008
- 123 million records
- 29 variables

In which variable(s) does Simpson's paradox lurk?

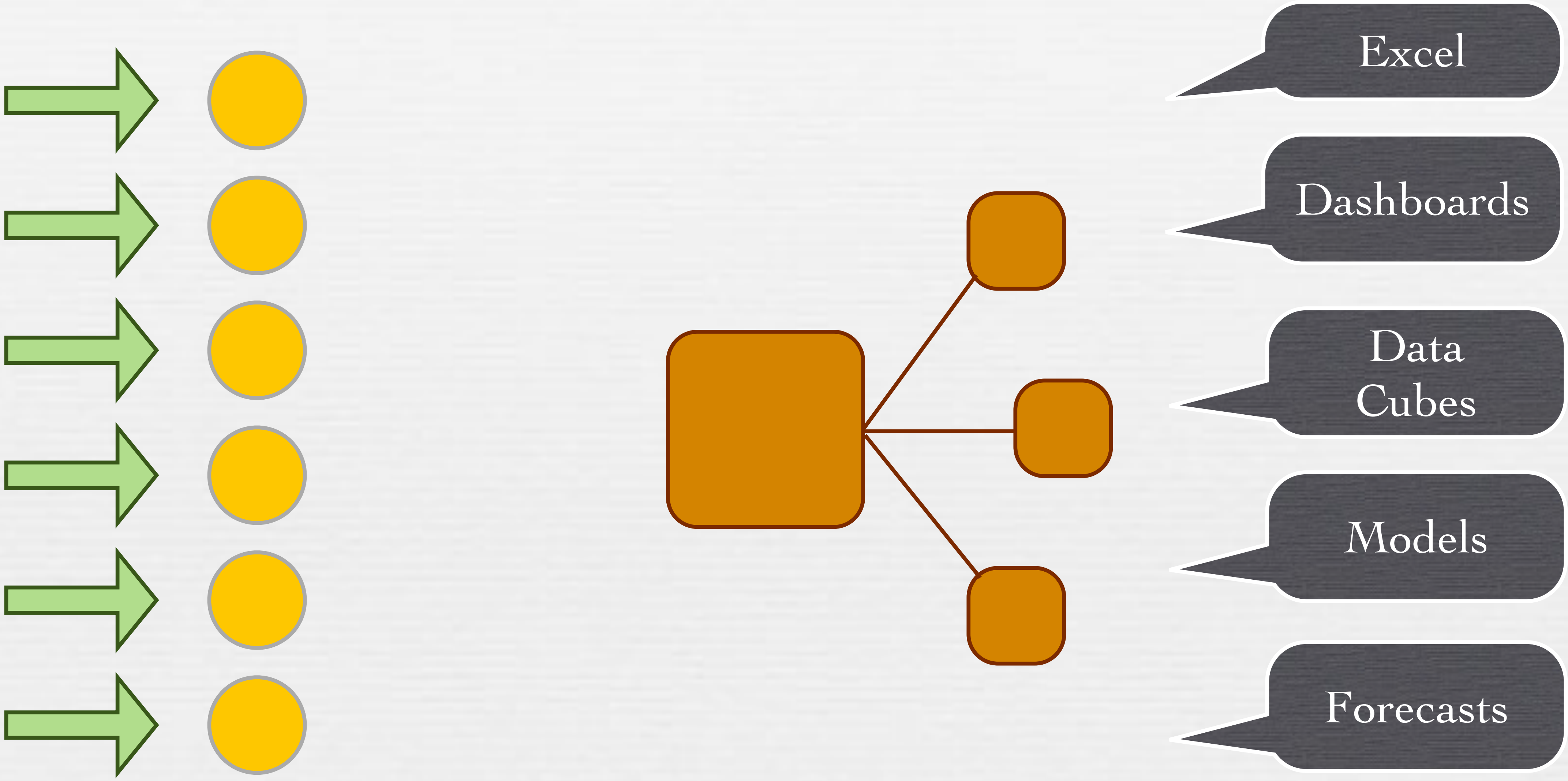
A priori?

Wicklin (2009)

When more people are performing more analyses more quickly, there are more theories, more points of view, more complexity, more conflicts and more confusion. There is less clarity, less consensus and less confidence.



Big Data: Producers



Excel

Dashboards

Data Cubes

Models

Forecasts

user interactions

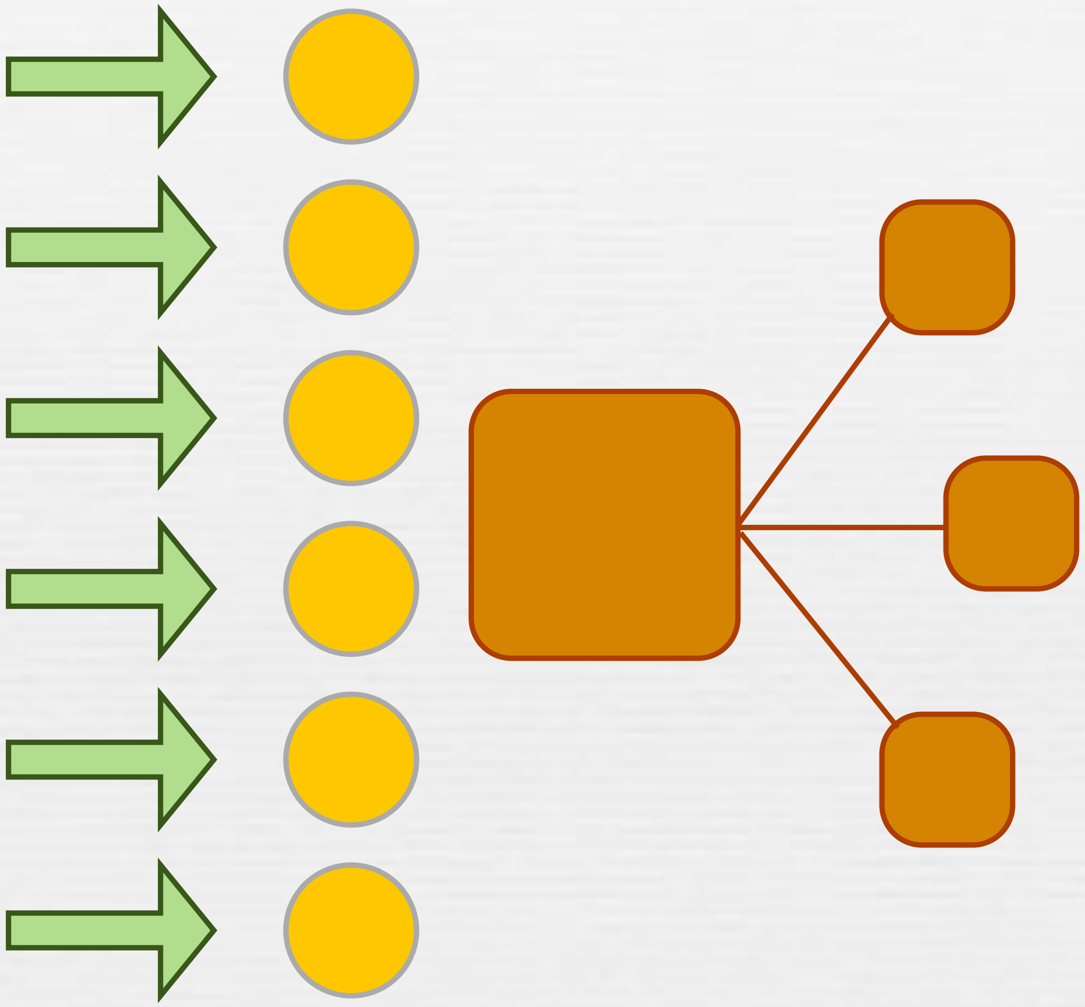
web logs: distributed servers

data ware house

data marts

displays

Big Data: Consumers



user interactions

web logs... datamarts

Excel

Dashboards

Data Cubes

Models

Forecasts

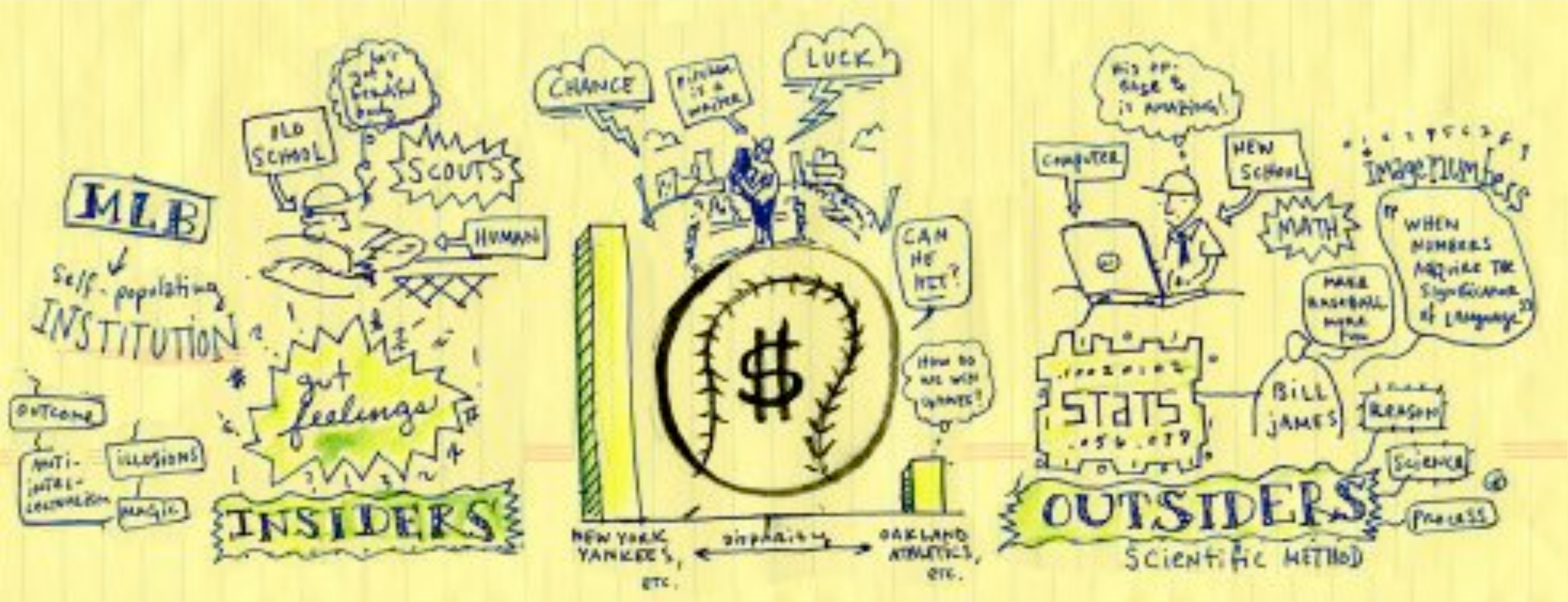
displays

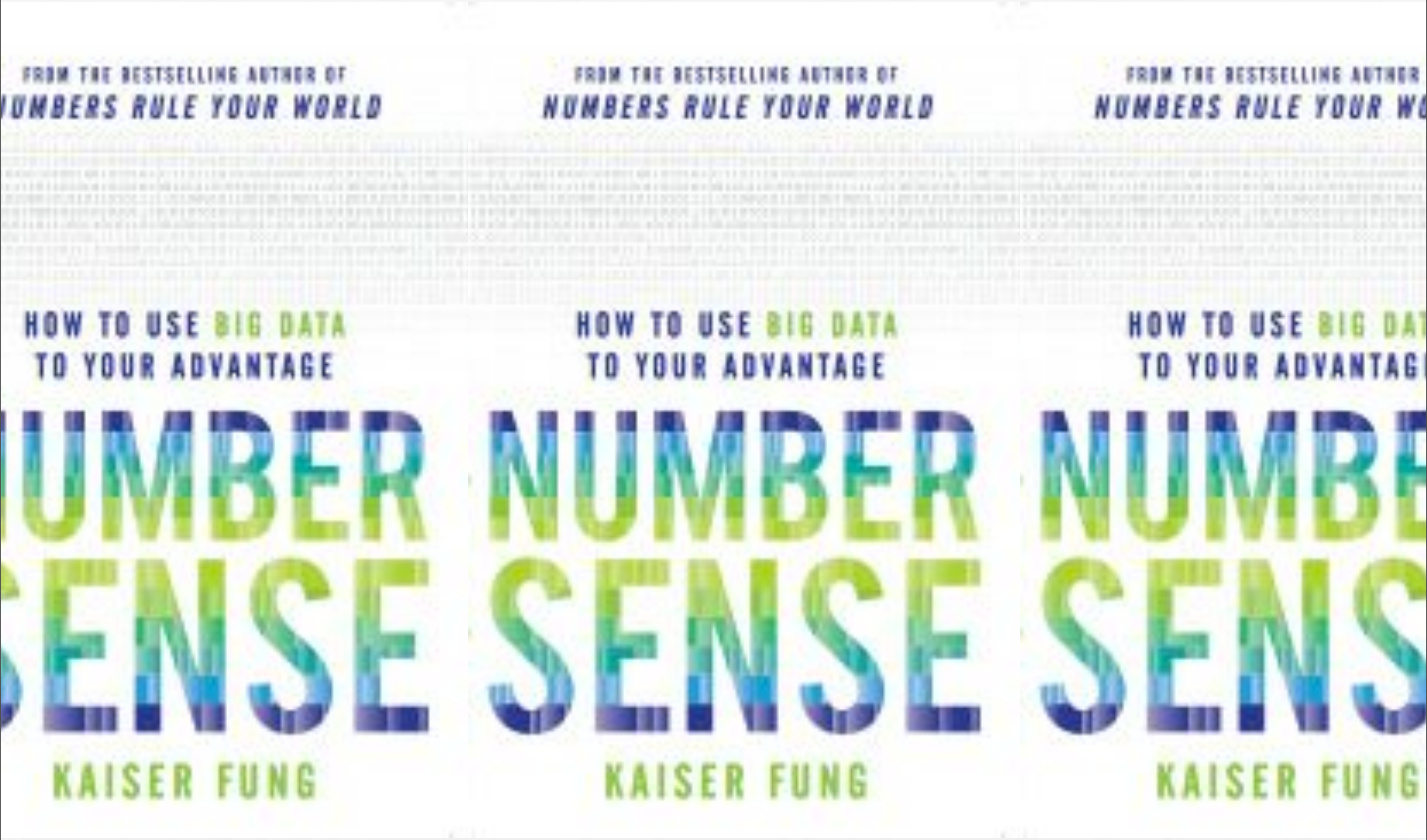
Strategies

Tactics

Plans

Moneyball


MLB
Self-populating
INSTITUTION
OUTCOME
ANTI-INTELLECTUALISM
ILLUSIONS
MAGIC
OLD SCHOOL
SCOUTS
HUMAN
gut feelings
INSIDERS
CHANCE
LUCK
CAN HE HIT?
$
NEW YORK YANKEES, ETC.
disparity
OAKLAND ATHLETICS, ETC.
COMPUTER
NEW SCHOOL
MATH
STATS
BILL JAMES
MAKE BASEBALL MORE FUN
"WHEN NUMBERS ACQUIRE THE SIGNIFICANCE OF LANGUAGE"
REASON
SCIENCE
PROCESS
OUTSIDERS
SCIENTIFIC METHOD



FROM THE BESTSELLING AUTHOR OF

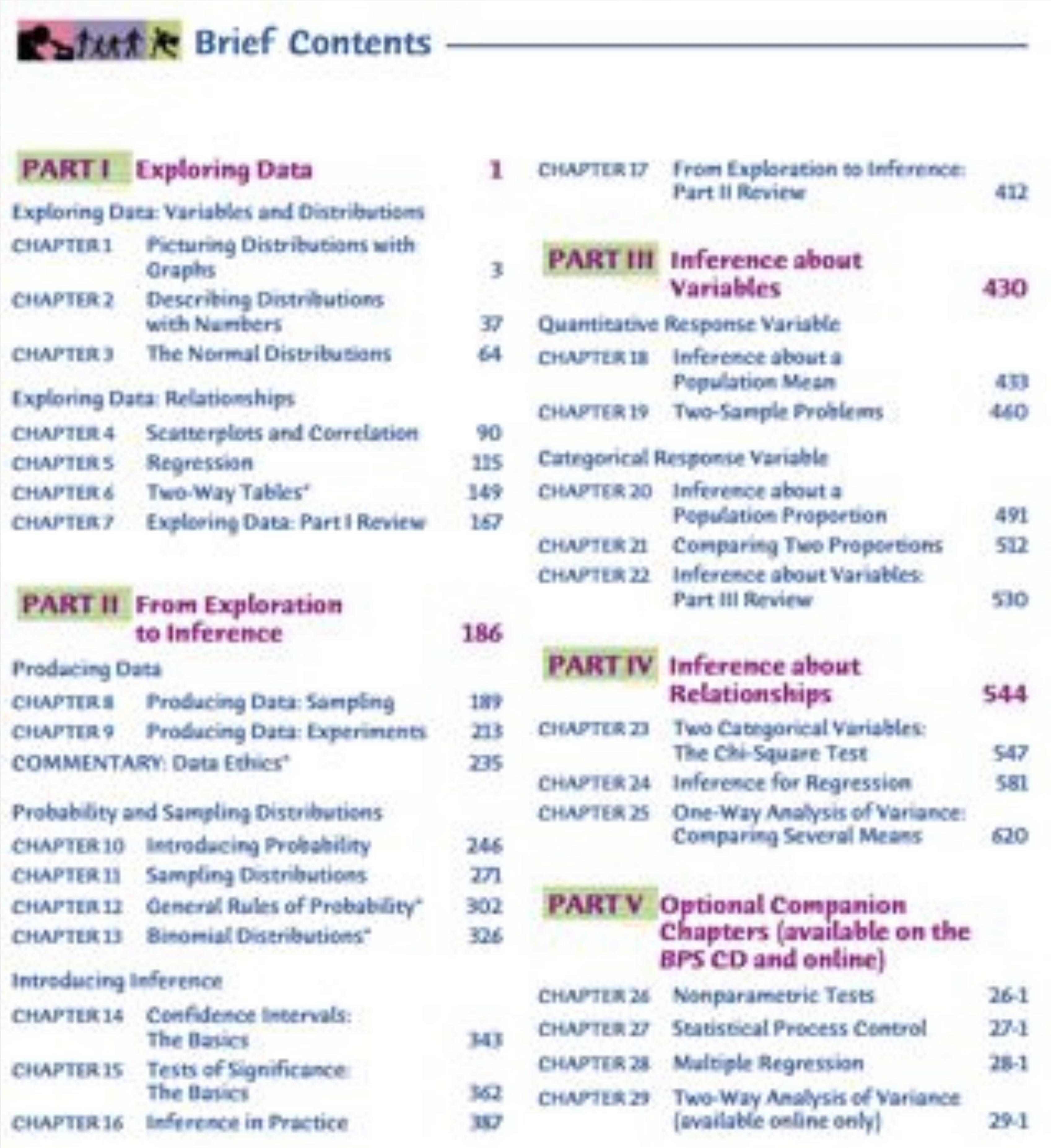
NUMBERS RULE YOUR WORLD

HOW TO USE BIG DATA
TO YOUR ADVANTAGE

NUMBER SENSE

KAISER FUNG

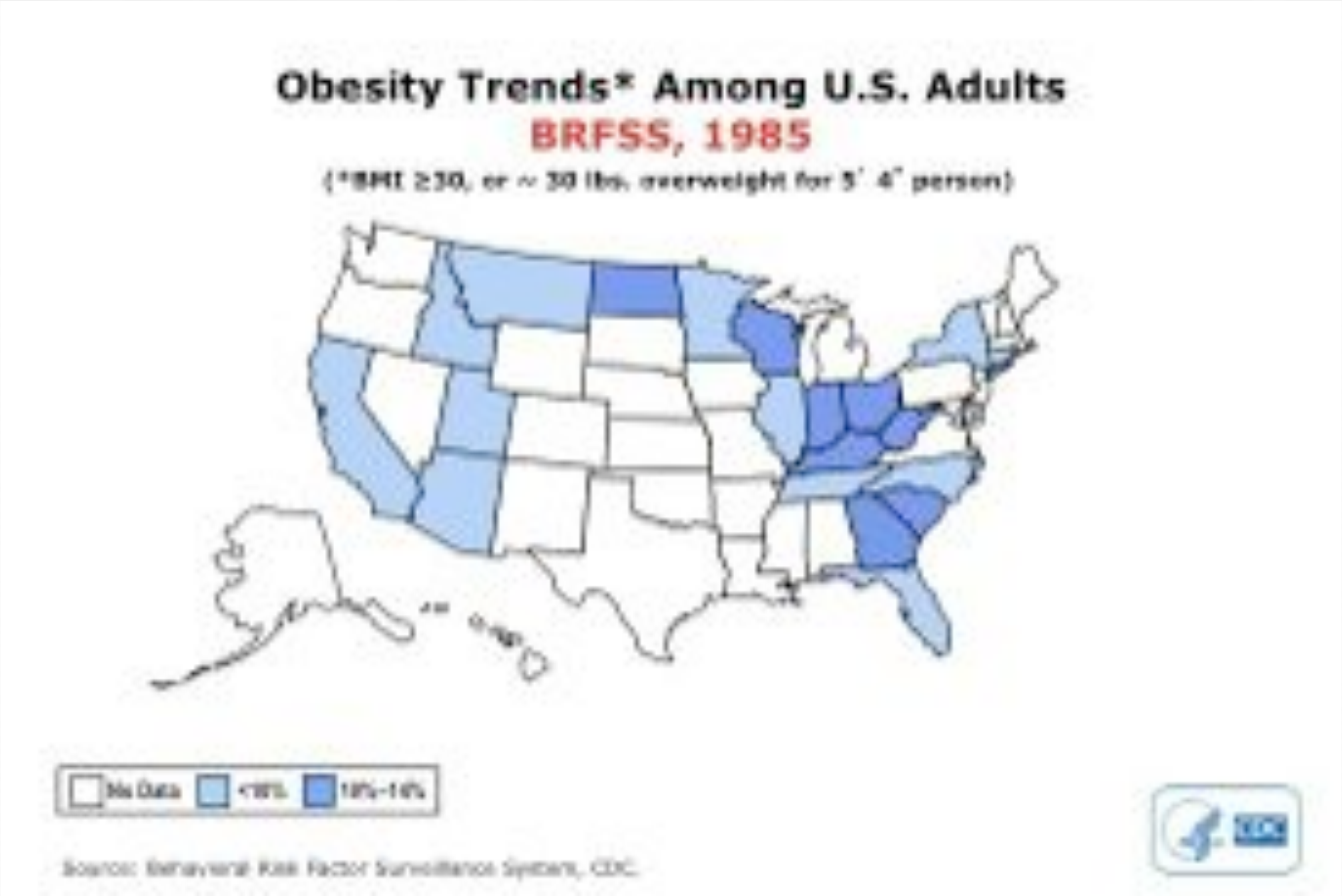
Statistics != Math



Brief Contents

David S. Moore, The Basic Practice of Statistics, ~2007

The Obesity Epidemic



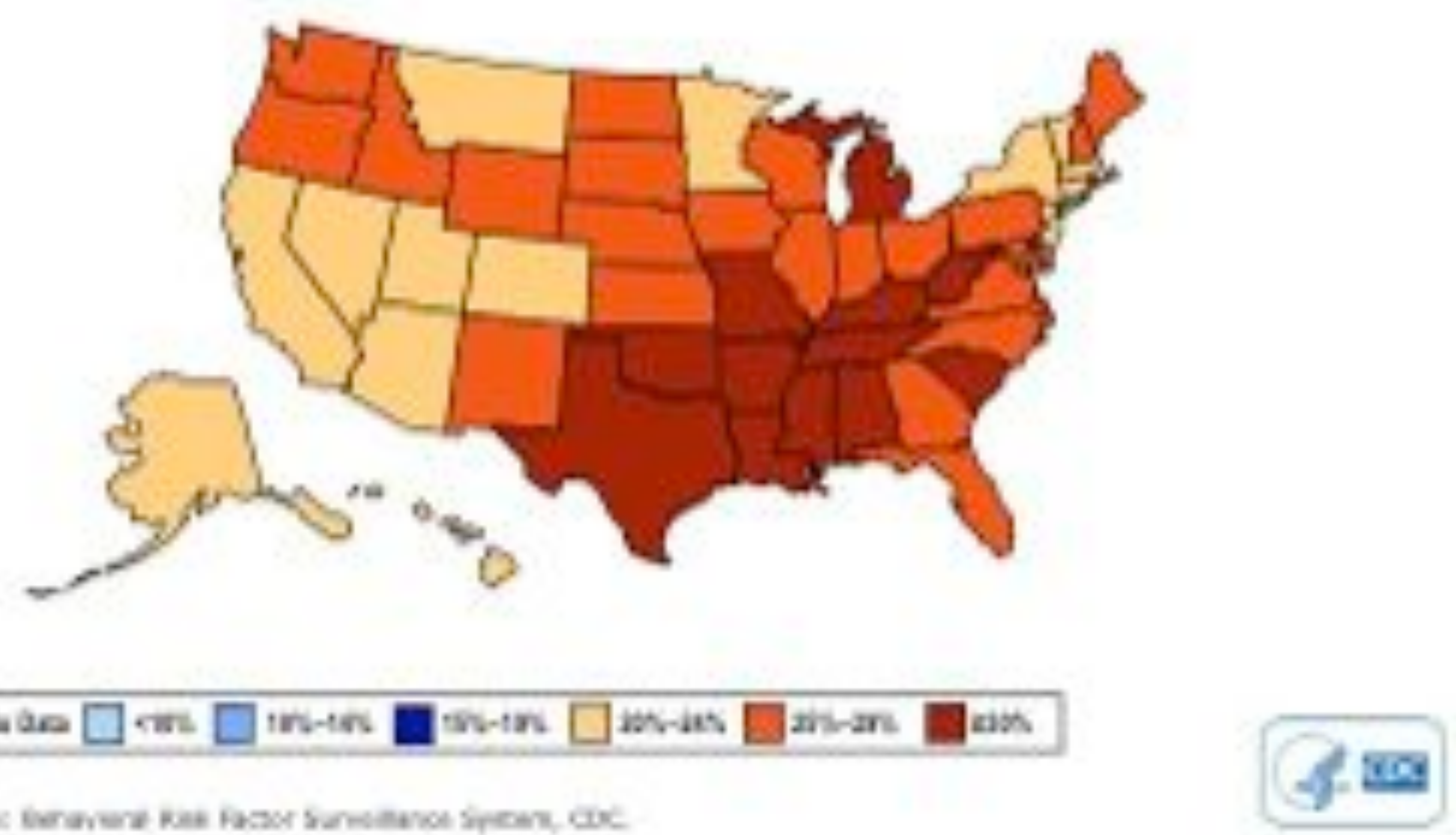
Obesity Trends* Among U.S. Adults

BRFSS, 1985

(*BMI ≥30, or ~ 30 lbs. overweight for 5' 4" person)

Source: Behavioral Risk Factor Surveillance System, CDC.

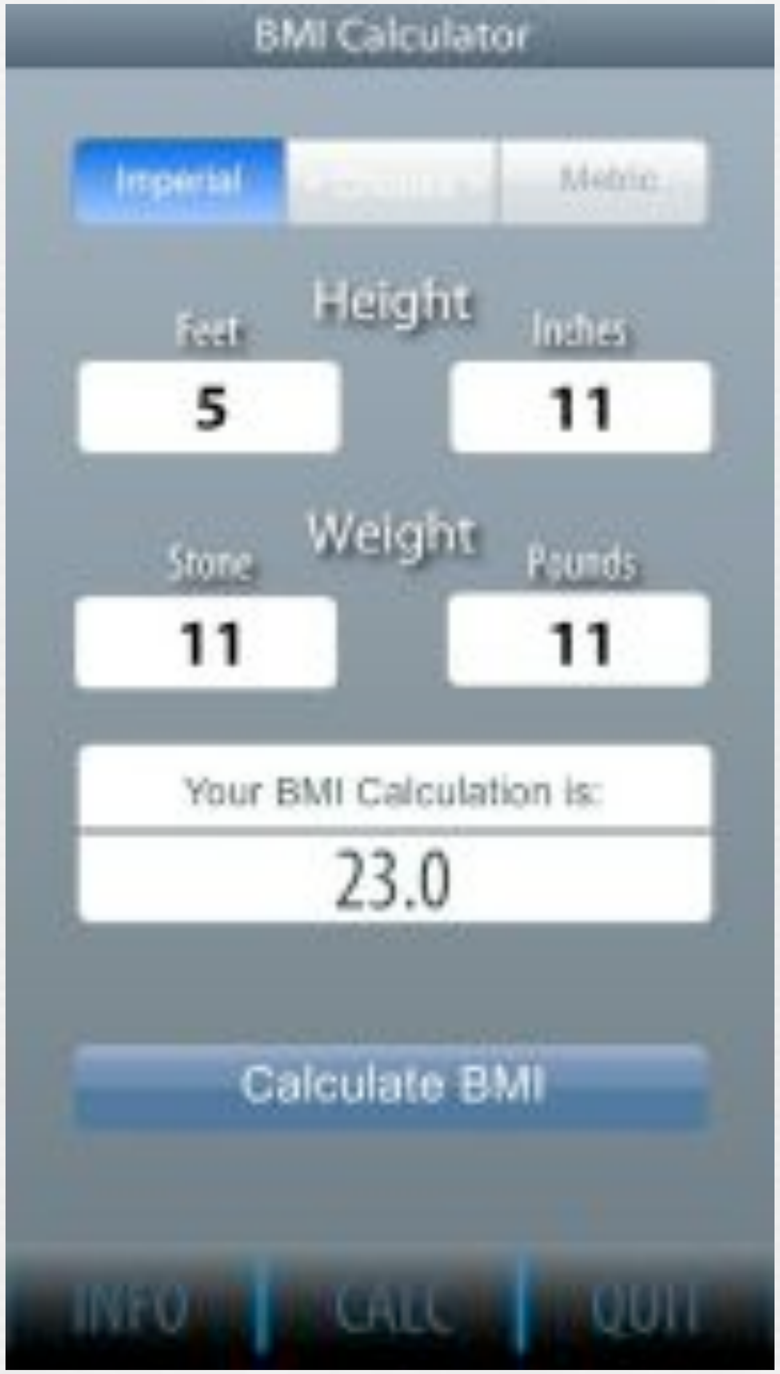
The Obesity Epidemic



Obesity Trends* Among U.S. Adults
BRFSS, 2010
(*BMI ≥30, or ~ 30 lbs. overweight for 5' 4" person)

Source: Behavioral Risk Factor Surveillance System, CDC.

Quetelet's Index (1830)



BMI Calculator

Height

Feet: 5

Inches: 11

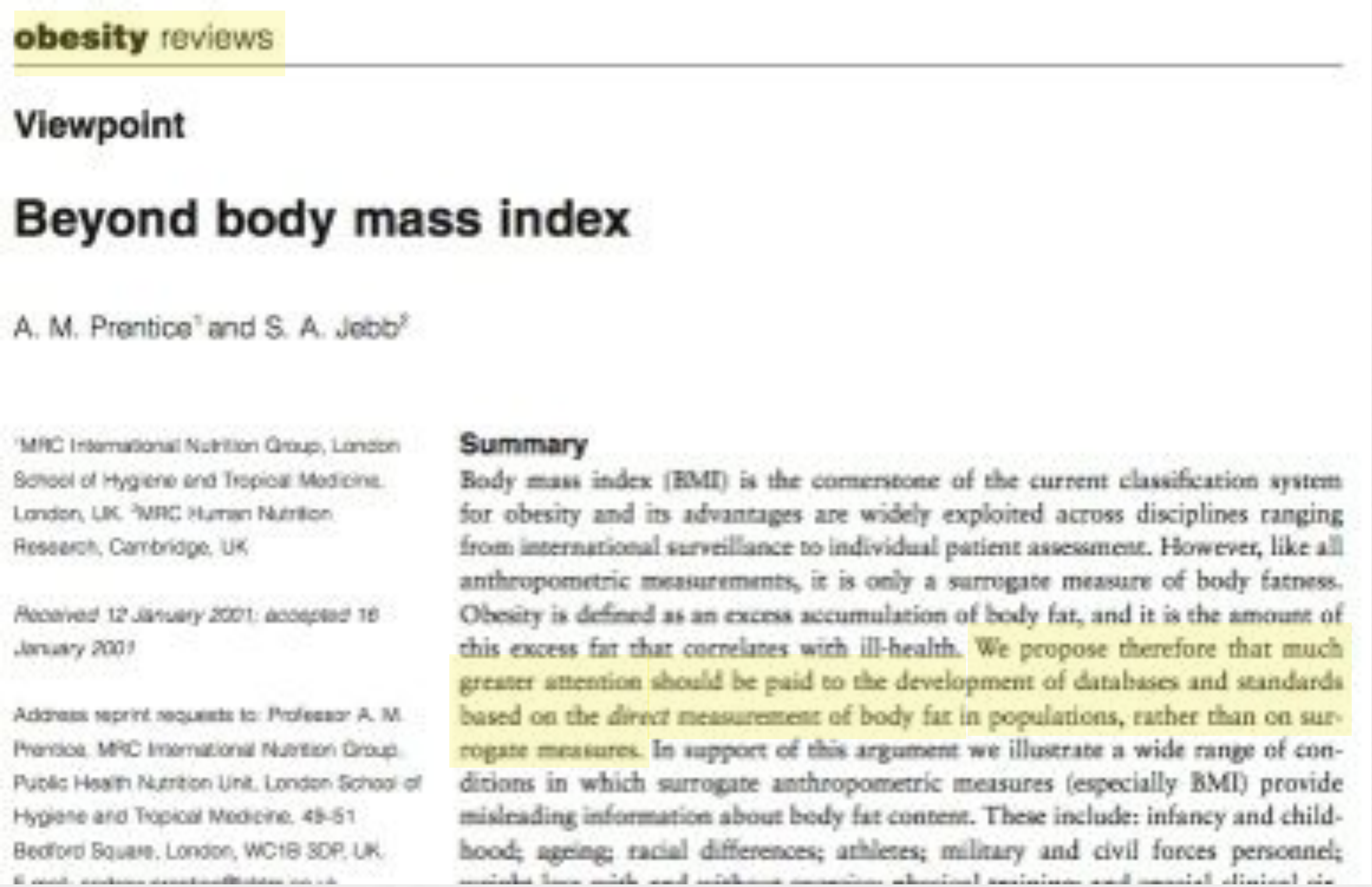
Weight

Stone: 11

Pounds: 11

Your BMI Calculation is: 23.0

BMI Critics (2000-)



obesity reviews

Viewpoint

Beyond body mass index

A. M. Prentice[1] and S. A. Jebb[2]

[1]MRC International Nutrition Group, London School of Hygiene and Tropical Medicine, London, UK. [2]MRC Human Nutrition Research, Cambridge, UK

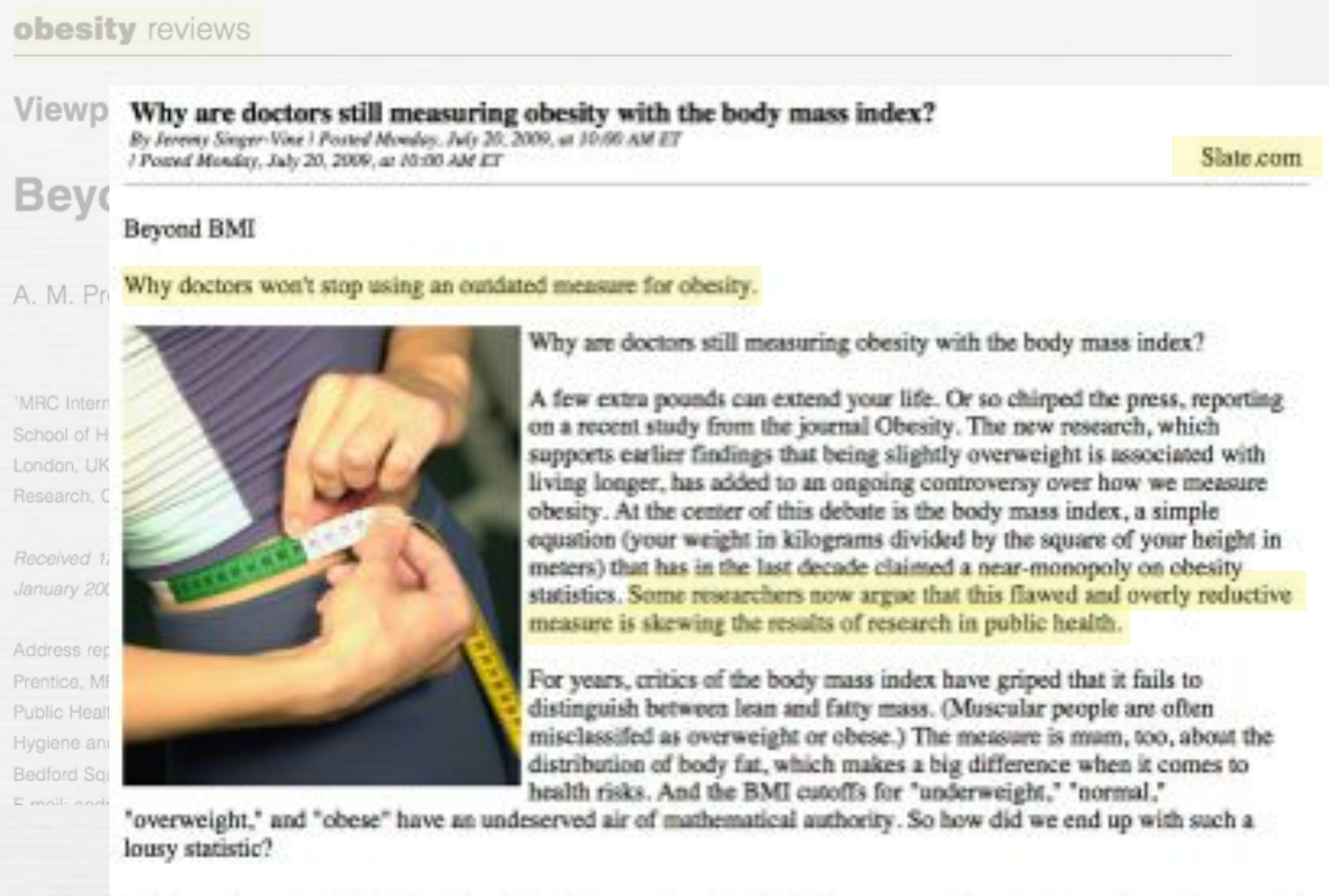
Received 12 January 2001; accepted 16 January 2001

Address reprint requests to: Professor A. M. Prentice, MRC International Nutrition Group, Public Health Nutrition Unit, London School of Hygiene and Tropical Medicine, 49-51 Bedford Square, London, WC1B 3DP, UK.

Summary

Body mass index (BMI) is the cornerstone of the current classification system for obesity and its advantages are widely exploited across disciplines ranging from international surveillance to individual patient assessment. However, like all anthropometric measurements, it is only a surrogate measure of body fatness. Obesity is defined as an excess accumulation of body fat, and it is the amount of this excess fat that correlates with ill-health. We propose therefore that much greater attention should be paid to the development of databases and standards based on the *direct* measurement of body fat in populations, rather than on surrogate measures. In support of this argument we illustrate a wide range of conditions in which surrogate anthropometric measures (especially BMI) provide misleading information about body fat content. These include: infancy and childhood; ageing; racial differences; athletes; military and civil forces personnel;

BMI Critics (2000-)



obesity reviews

Viewp

Beyo

A. M. Pr

¹MRC Intern
School of H
London, UK
Research, C

Received 1
January 20

Address rep
Prentice, MF
Public Healt
Hygiene an
Bedford Sq

Why are doctors still measuring obesity with the body mass index?

By Jeremy Singer-Vine | Posted Monday, July 20, 2009, at 10:00 AM ET
| Posted Monday, July 20, 2009, at 10:00 AM ET

Slate.com

Beyond BMI

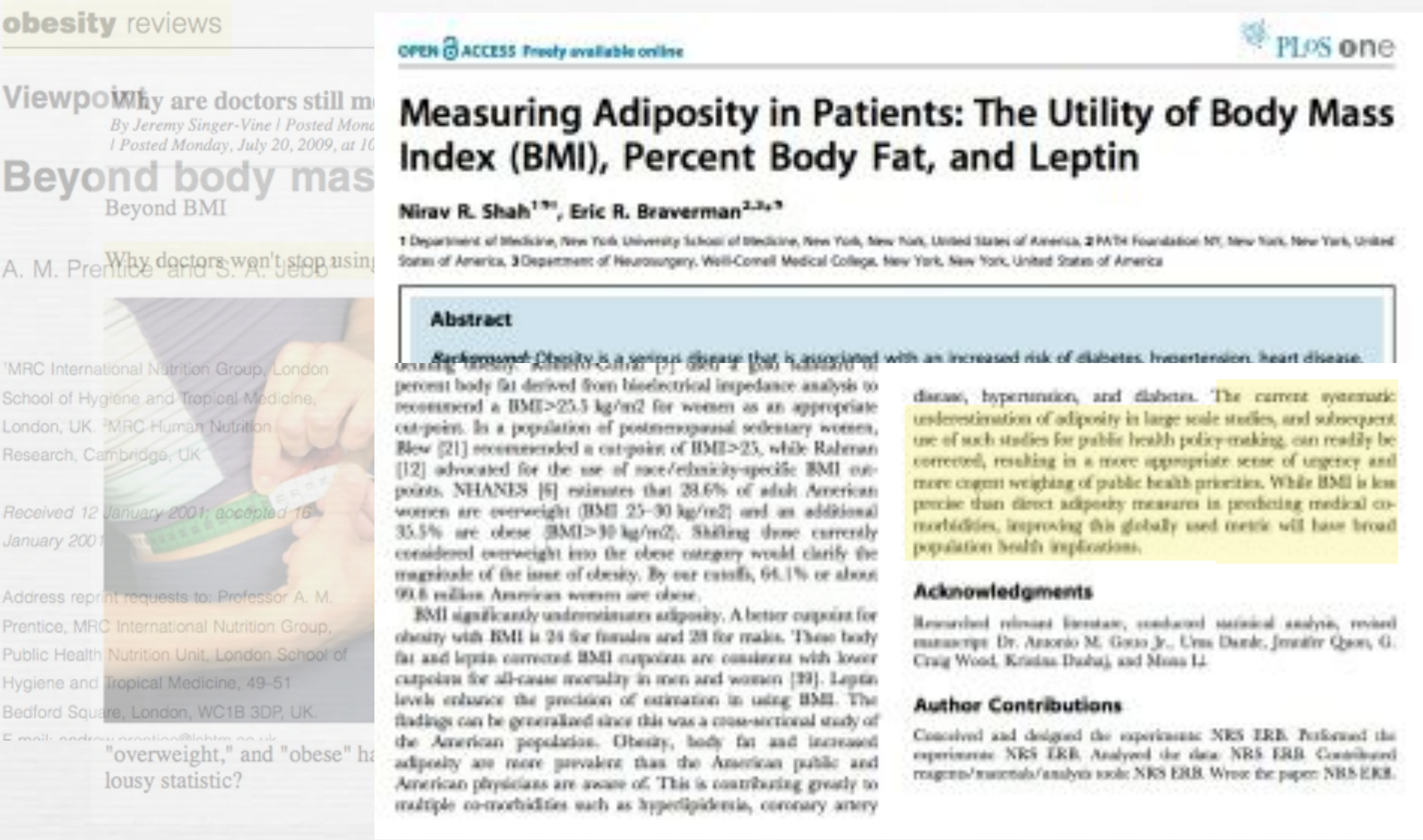
Why doctors won't stop using an outdated measure for obesity.

Why are doctors still measuring obesity with the body mass index?

A few extra pounds can extend your life. Or so chirped the press, reporting on a recent study from the journal Obesity. The new research, which supports earlier findings that being slightly overweight is associated with living longer, has added to an ongoing controversy over how we measure obesity. At the center of this debate is the body mass index, a simple equation (your weight in kilograms divided by the square of your height in meters) that has in the last decade claimed a near-monopoly on obesity statistics. Some researchers now argue that this flawed and overly reductive measure is skewing the results of research in public health.

For years, critics of the body mass index have griped that it fails to distinguish between lean and fatty mass. (Muscular people are often misclassifed as overweight or obese.) The measure is mum, too, about the distribution of body fat, which makes a big difference when it comes to health risks. And the BMI cutoffs for "underweight," "normal," "overweight," and "obese" have an undeserved air of mathematical authority. So how did we end up with such a lousy statistic?

BMI Critics (2000-)



obesity reviews

Viewpoint

Beyond body mas

A. M. Prentice and S. A. Jebb

¹MRC International Nutrition Group, London School of Hygiene and Tropical Medicine, London, UK. ²MRC Human Nutrition Research, Cambridge, UK

Received 12 January 2001; accepted 16 January 2001

Address reprint requests to: Professor A. M. Prentice, MRC International Nutrition Group, Public Health Nutrition Unit, London School of Hygiene and Tropical Medicine, 49-51 Bedford Square, London, WC1B 3DP, UK.

Why are doctors still m

By Jeremy Singer-Vine | Posted Mond

| Posted Monday, July 20, 2009, at 10

Beyond BMI

Why doctors won't stop using

"overweight," and "obese" ha

lousy statistic?

OPEN ACCESS Freely available online

PLoS one

Measuring Adiposity in Patients: The Utility of Body Mass Index (BMI), Percent Body Fat, and Leptin

Nirav R. Shah[1,*], Eric R. Braverman[2,3,*]

[1] Department of Medicine, New York University School of Medicine, New York, New York, United States of America, [2] PATH Foundation NY, New York, New York, United States of America, [3] Department of Neurosurgery, Weill-Cornell Medical College, New York, New York, United States of America

Abstract

Background: Obesity is a serious disease that is associated with an increased risk of diabetes, hypertension, heart disease,

defining obesity. Romero-Corral [7] used a gold standard of percent body fat derived from bioelectrical impedance analysis to recommend a BMI>25.5 kg/m2 for women as an appropriate cut-point. In a population of postmenopausal sedentary women, Blew [21] recommended a cut-point of BMI>25, while Rahman [12] advocated for the use of race/ethnicity-specific BMI cut-points. NHANES [6] estimates that 28.6% of adult American women are overweight (BMI 25–30 kg/m2) and an additional 35.5% are obese (BMI>30 kg/m2). Shifting those currently considered overweight into the obese category would clarify the magnitude of the issue of obesity. By our cutoffs, 64.1% or about 99.6 million American women are obese.

BMI significantly underestimates adiposity. A better cutpoint for obesity with BMI is 24 for females and 28 for males. These body fat and leptin corrected BMI cutpoints are consistent with lower cutpoints for all-cause mortality in men and women [39]. Leptin levels enhance the precision of estimation in using BMI. The findings can be generalized since this was a cross-sectional study of the American population. Obesity, body fat and increased adiposity are more prevalent than the American public and American physicians are aware of. This is contributing greatly to multiple co-morbidities such as hyperlipidemia, coronary artery

disease, hypertension, and diabetes. The current systematic underestimation of adiposity in large scale studies, and subsequent use of such studies for public health policy-making, can readily be corrected, resulting in a more appropriate sense of urgency and more cogent weighing of public health priorities. While BMI is less precise than direct adiposity measures in predicting medical co-morbidities, improving this globally used metric will have broad population health implications.

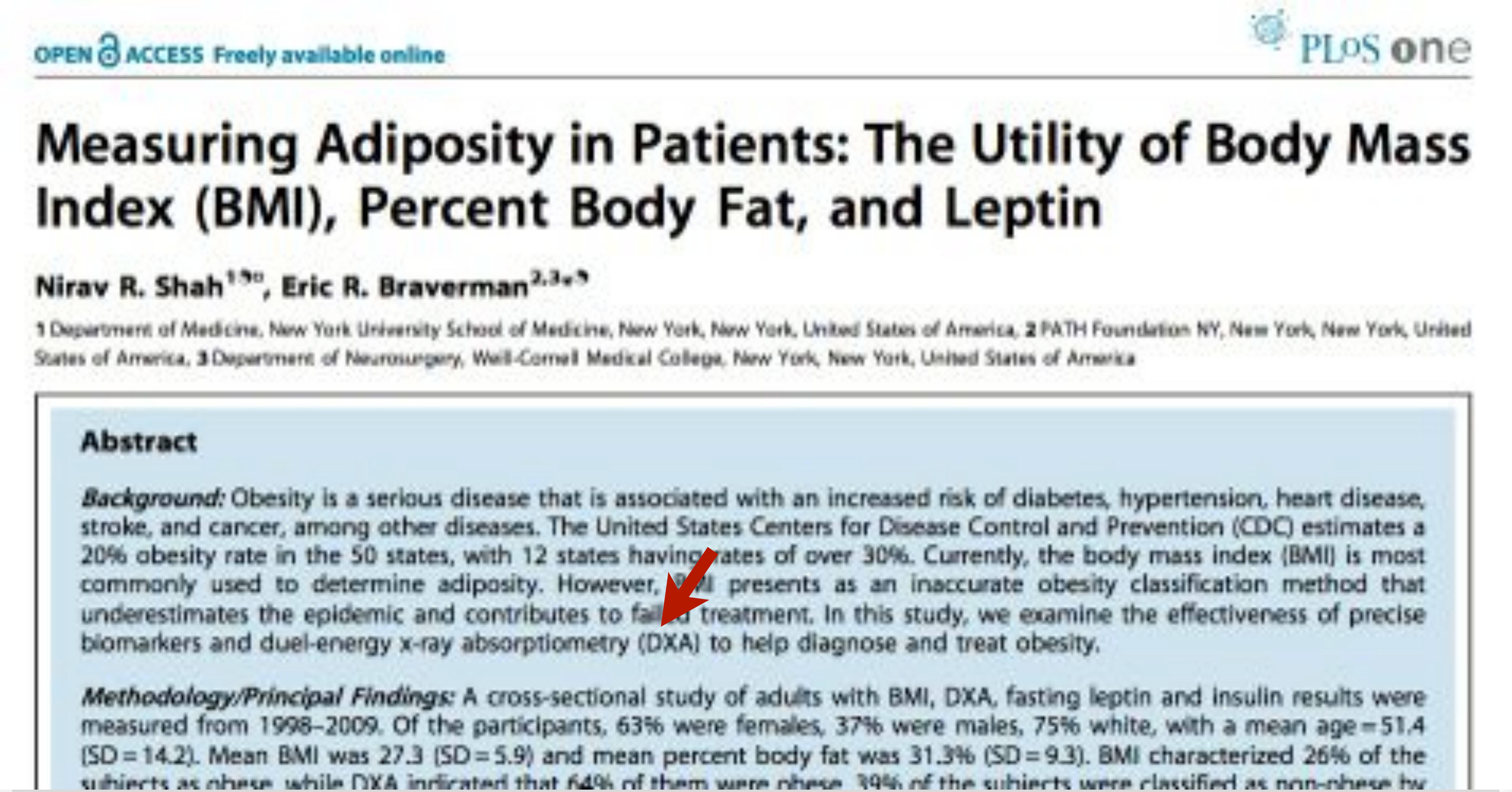
Acknowledgments

Researched relevant literature, conducted statistical analysis, revised manuscript: Dr. Antonio M. Gotto Jr., Uma Dande, Jennifer Quon, G. Craig Wood, Kristina Duhaj, and Mona Li.

Author Contributions

Conceived and designed the experiments: NRS ERB. Performed the experiments: NRS ERB. Analyzed the data: NRS ERB. Contributed reagents/materials/analysis tools: NRS ERB. Wrote the paper: NRS ERB.

Taking eyes off the ball



OPEN ACCESS Freely available online

PLoS one

Measuring Adiposity in Patients: The Utility of Body Mass Index (BMI), Percent Body Fat, and Leptin

Nirav R. Shah[1¶¤], Eric R. Braverman[2,3*¶]

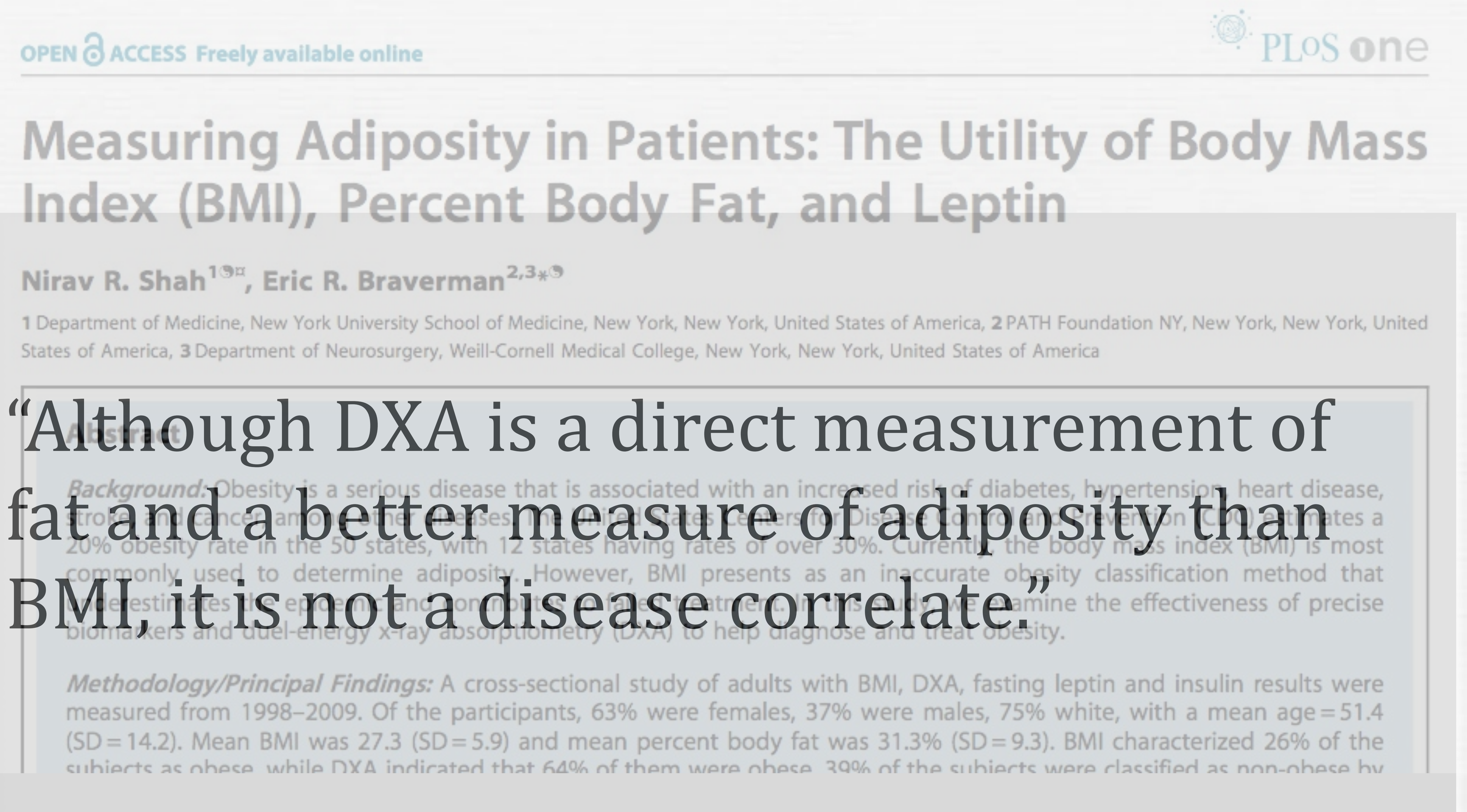
1 Department of Medicine, New York University School of Medicine, New York, New York, United States of America, 2 PATH Foundation NY, New York, New York, United States of America, 3 Department of Neurosurgery, Weill-Cornell Medical College, New York, New York, United States of America

Abstract

Background: Obesity is a serious disease that is associated with an increased risk of diabetes, hypertension, heart disease, stroke, and cancer, among other diseases. The United States Centers for Disease Control and Prevention (CDC) estimates a 20% obesity rate in the 50 states, with 12 states having rates of over 30%. Currently, the body mass index (BMI) is most commonly used to determine adiposity. However, BMI presents as an inaccurate obesity classification method that underestimates the epidemic and contributes to failed treatment. In this study, we examine the effectiveness of precise biomarkers and duel-energy x-ray absorptiometry (DXA) to help diagnose and treat obesity.

Methodology/Principal Findings: A cross-sectional study of adults with BMI, DXA, fasting leptin and insulin results were measured from 1998–2009. Of the participants, 63% were females, 37% were males, 75% white, with a mean age = 51.4 (SD = 14.2). Mean BMI was 27.3 (SD = 5.9) and mean percent body fat was 31.3% (SD = 9.3). BMI characterized 26% of the subjects as obese, while DXA indicated that 64% of them were obese. 39% of the subjects were classified as non-obese by

Taking eyes off the ball



OPEN ACCESS Freely available online

PLoS one

Measuring Adiposity in Patients: The Utility of Body Mass Index (BMI), Percent Body Fat, and Leptin

Nirav R. Shah[1], Eric R. Braverman[2,3]*

1 Department of Medicine, New York University School of Medicine, New York, New York, United States of America, 2 PATH Foundation NY, New York, New York, United States of America, 3 Department of Neurosurgery, Weill-Cornell Medical College, New York, New York, United States of America

Abstract

Background: Obesity is a serious disease that is associated with an increased risk of diabetes, hypertension, heart disease, stroke, and cancer, among other diseases. The United States Centers for Disease Control and Prevention (CDC) estimates a 20% obesity rate in the 50 states, with 12 states having rates of over 30%. Currently, the body mass index (BMI) is most commonly used to determine adiposity. However, BMI presents as an inaccurate obesity classification method that underestimates the epidemic and contributes to failed treatment. In this study, we examine the effectiveness of precise biomarkers and duel-energy x-ray absorptiometry (DXA) to help diagnose and treat obesity.

Methodology/Principal Findings: A cross-sectional study of adults with BMI, DXA, fasting leptin and insulin results were measured from 1998–2009. Of the participants, 63% were females, 37% were males, 75% white, with a mean age = 51.4 (SD = 14.2). Mean BMI was 27.3 (SD = 5.9) and mean percent body fat was 31.3% (SD = 9.3). BMI characterized 26% of the subjects as obese, while DXA indicated that 64% of them were obese. 39% of the subjects were classified as non-obese by

“Although DXA is a direct measurement of fat and a better measure of adiposity than BMI, it is not a disease correlate.”

The more things change

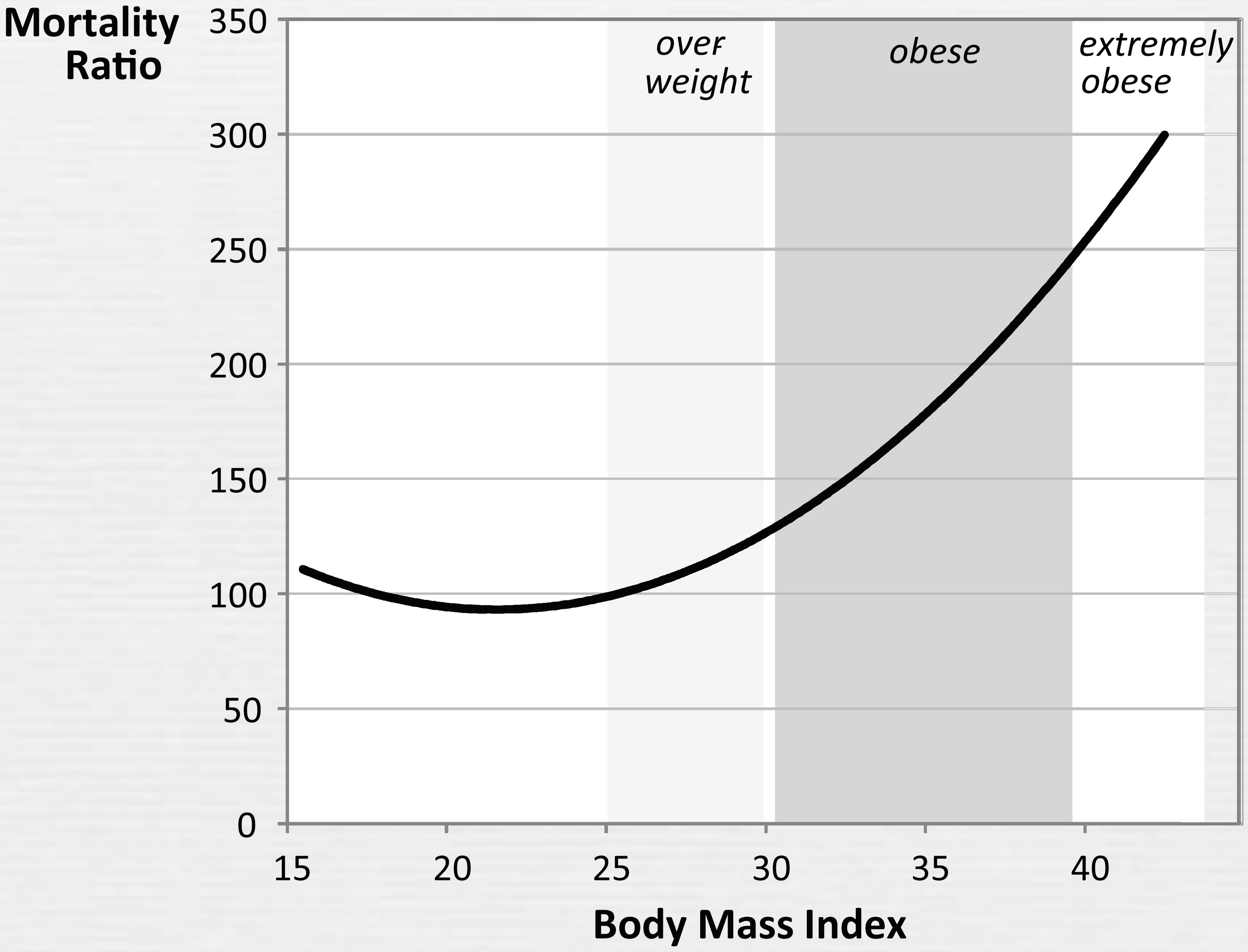
All Patients

		DXA		
		Not Obese	Obese	BMI Totals
BMI	Not Obese	35 %	39 %	74%
	Obese	1 %	25 %	26%
	DXA Totals	36 %	64 %	100%

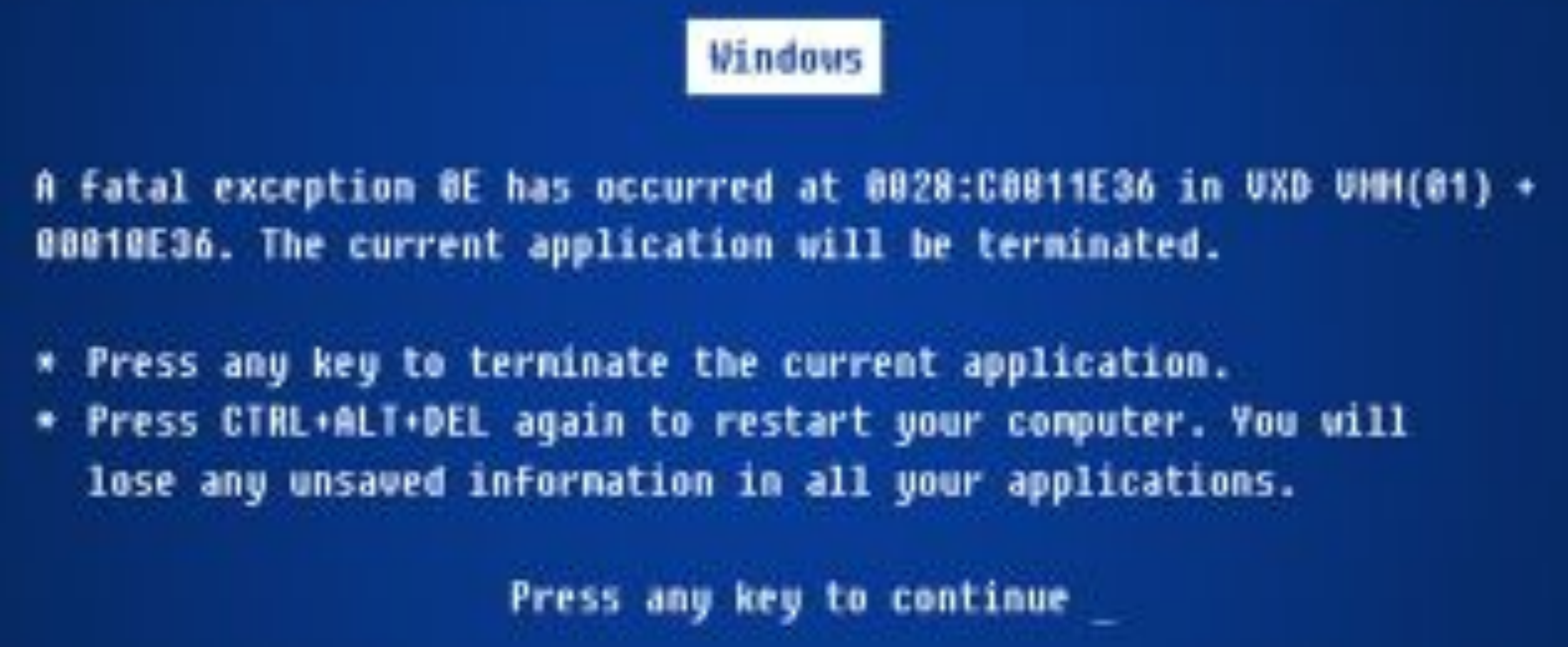
Female Patients

		DXA		
		Not Obese	Obese	BMI Totals
BMI	Not Obese	26 %	48 %	74%
	Obese	0 %	26 %	26%
	DXA Totals	26 %	74 %	100%

n-U-isance


Mortality Ratio
350
300
250
200
150
100
50
0
over weight
obese
extremely obese
15
20
25
30
35
40
Body Mass Index

Reinstall Windows



```
                          Windows

A fatal exception 0E has occurred at 0028:C0011E36 in VXD VMM(01) +
00010E36. The current application will be terminated.

*  Press any key to terminate the current application.
*  Press CTRL+ALT+DEL again to restart your computer. You will
   lose any unsaved information in all your applications.

                  Press any key to continue _
```

Trust, not Truth

Embarrassment of Riches

A team of psychologists performed personality tests on 100 professionals, of which 30 were engineers and 70 were lawyers.

Here is a brief description of one of the subjects:

> Jack is a 45-year-old man. He is married and has four children. He is generally conservative, careful, and ambitious. He shows no interest in political or social issues and spends most of his free time on his many hobbies, which include home carpentry, sailing, and mathematics.

Kahneman and Tversky (1974)

What is the probability that Jack is one of the 30 engineers?

A. 10 – 40 %

B. 41 – 60 %

C. 61 – 80 %

D. 81 – 100 %

Kahneman and Tversky (1974); Vanity Fair

The Law of Small Numbers is even more relevant in the era of Big Data

Target knows your daughter is pregnant

... before you do

Customer Acquisition

“Right around the birth of a child… parents are exhausted and overwhelmed and their shopping patterns and brand loyalties are up for grabs.”

Customer Acquisition

"We knew that if we could identify them in their second trimester, there's a good chance we could capture them for years."

Brochure Design

"We started mixing in all these ads for things we knew pregnant women would never buy, so the baby ads looked random."

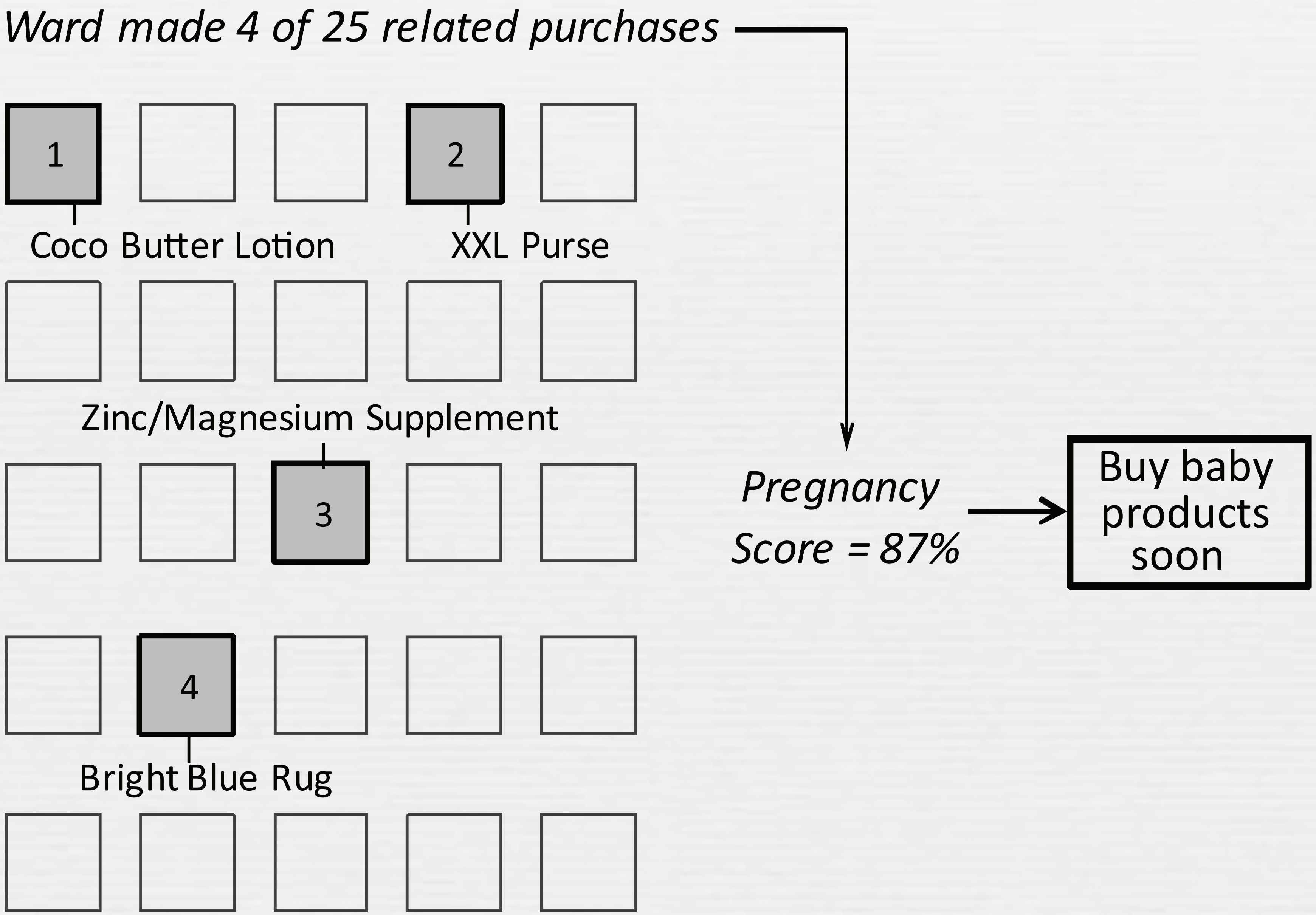
Brochure Design

“We’d put an ad for wineglasses next to infant clothes. That way, it looked like all the products were chosen by chance.”

Brochure Design

"As long as a pregnant woman thinks she hasn't been spied on, she'll use the coupons."

The Model



Ward made 4 of 25 related purchases

Coco Butter Lotion

XXL Purse

Zinc/Magnesium Supplement

Bright Blue Rug

Pregnancy Score = 87%

Buy baby products soon

Mad Dad

		Reality		
		Pregnant	Not	
Model Says	Pregnant	6%	14%	20
	Not	4%	76%	80
		10	90	100

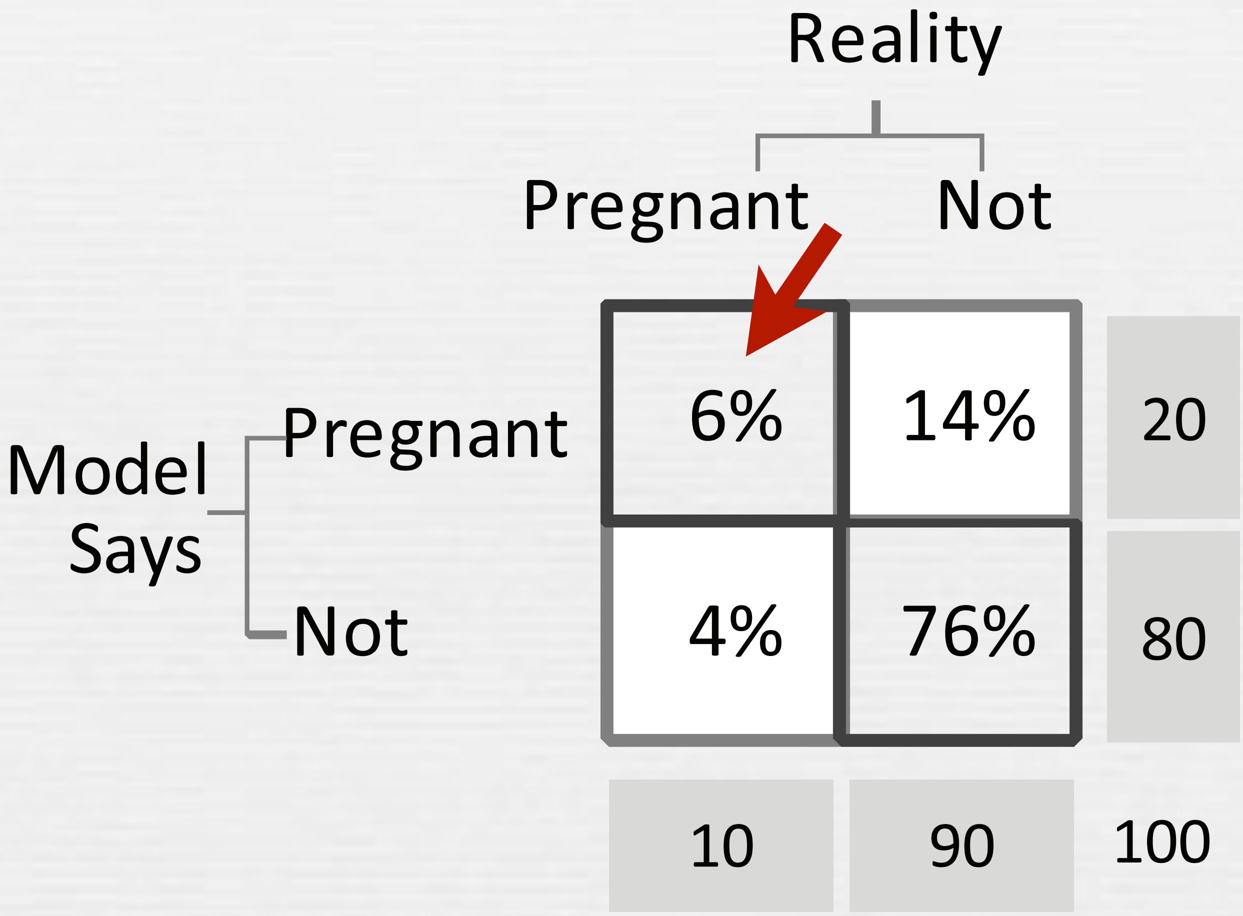
Incidence: $\frac{10}{100} = 10\%$

Positive predictive value: $\frac{6}{20} = 30\%$

False positive rate: $\frac{14}{90} = 16\%$

False negative rate: $\frac{4}{10} = 40\%$

Sending Mixed Messages



Model Says	Reality: Pregnant	Reality: Not	
Pregnant	6%	14%	20
Not	4%	76%	80
	10	90	100

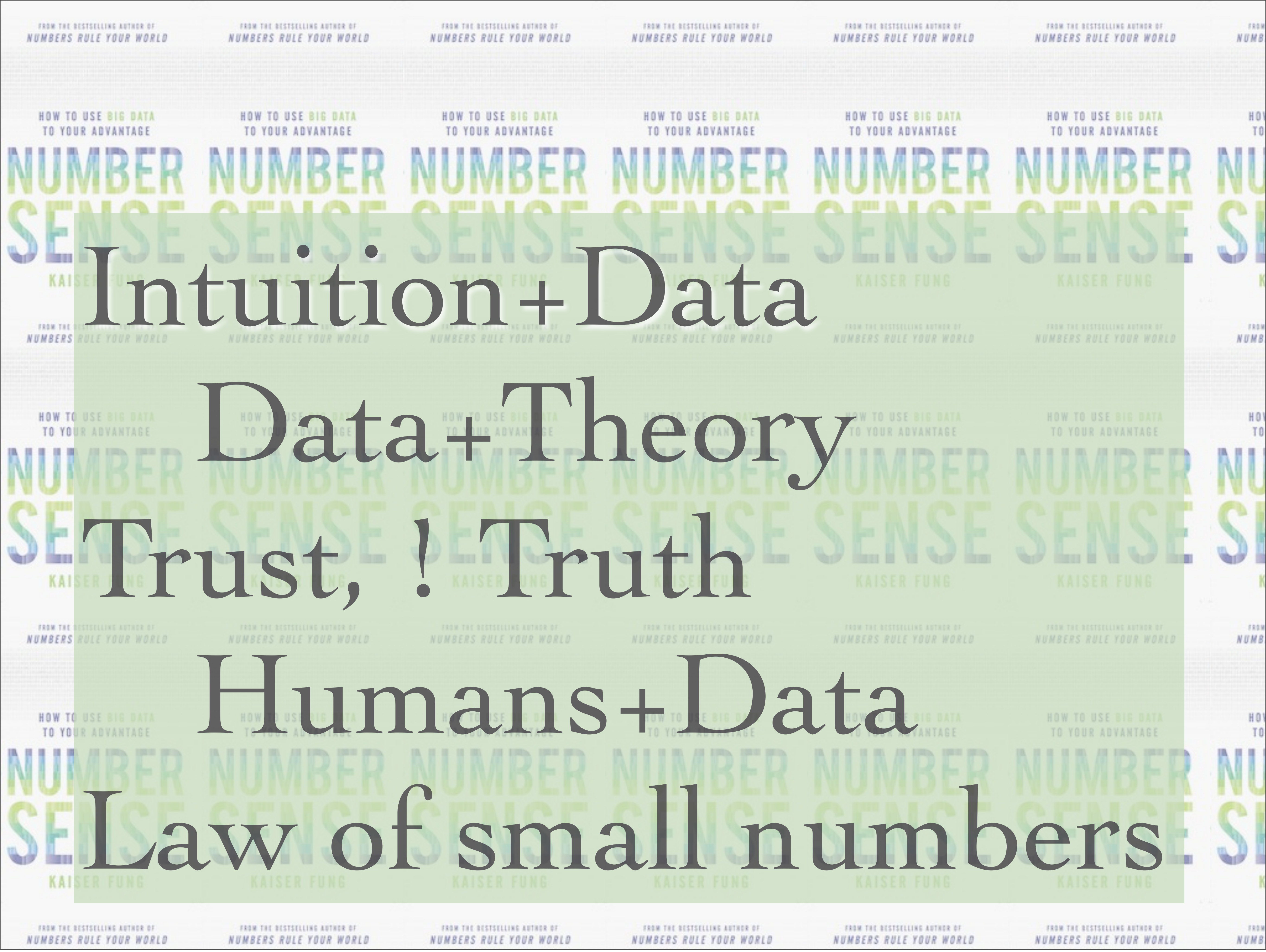
Incidence: $\frac{10}{100}$ = 10%

3x

Positive predictive value: $\frac{6}{20}$ = 30%

False positive rate: $\frac{14}{90}$ = 16%

False negative rate: $\frac{4}{10}$ = 40%



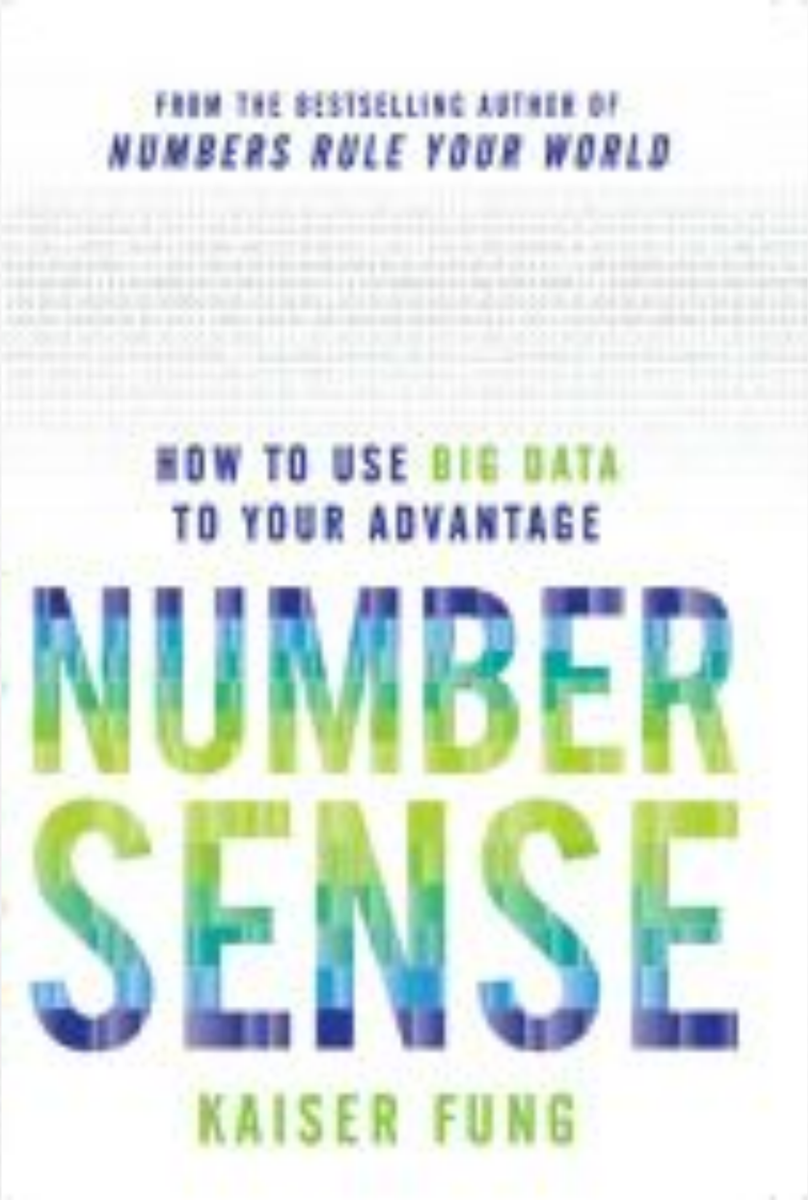
Intuition+Data

Data+Theory

Trust, ! Truth

Humans+Data

Law of small numbers



Thank you

Twitter: @junkcharts

Gmail: JunkCharts