Two examples

In both we emphasize ML or AI as a process that starts with some phenomenon in the world — it starts with the phenomenon you want to describe and perhaps some associated outcomes; then to the data you collect to describe each; and to the modeling that links some aspects of your phenomenon to your outcomes; and finally model interpretation and action in the world

Depending on your problem, these steps will take you more or less time to figure out and will push you back on your journalistic ethics and values — Robyn already talked about “proxy variables” and governing algorithms

Our examples — One creates algorithms to base reporting on (think Trevor Martin and Reddit) and the other was an algorithm being used in the wild that we had to report on

Two examples

Classically, we speak about two kinds of problems in learning — **supervised and unsupervised**

In the first case, you have input variables (also called “independent” variables, often denoted X) and outputs (also called a “dependent” variable, typically denoted Y) and you would like to predict Y from X.

> From a database of health outcomes, what is the chance that a person with your height, weight, age and blood pressure will have a heart attack in the next 5 years?
>
> Given browsing behaviors of other visitors to the New York Times, what is the chance that a person with your browsing history will click on an article about Trump and Putin?

In the second case, you want to infer (unexpected?) structure from data without the presence of a response Y. You simply want to understand what kind of clusters exist, what are the relationships between the variables? The clustering procedures we saw last time (k-means and hierarchical clustering) are two examples. We were just trying to see what was in the data, not trying to learn a prediction pattern.

MetaMind

Secure https://metamind.io

Apps ha Kernel youtube

RESEARCH CAREERS DOCS FREE SIGN UP

Smarter AI for developers.

MetaMind.io is now part of Salesforce Einstein. It helps power the world's smartest CRM by embedding state-of-the-art deep learning technology into the Salesforce Platform.

LEARN MORE

Algorithmia - Open Marketplac ×

Secure https://algorithmia.com

Apps ha Kernel youtube

ALGORITHMIA

Algorithms Docs Pricing Blog More Sign In

BUILD *INTELLIGENT APPS*

We're a common API for algorithms, functions, and models that run as scalable microservices.

FIND

Leverage a library of more than 3,000 microservices via an API.

TEST

Try any microservice in 5-lines of code or less.

DEPLOY

Secure and scalable infrastructure is ready when you are.

Username

Email

Password

Have a promo code?

SIGN UP FOR FREE

By signing up, you agree to our terms and conditions a policy.
We will occasionally send you account related emails.

Text Analysis Algorithms - Alg

Secure https://algorithmia.com/tags/text-analysis

Apps ha Kernel youtube

ALGORITHMIA Search algorithms... Algorithms Docs More Sign Up Sign In

Text Analysis

Browse algorithms related to text analysis:

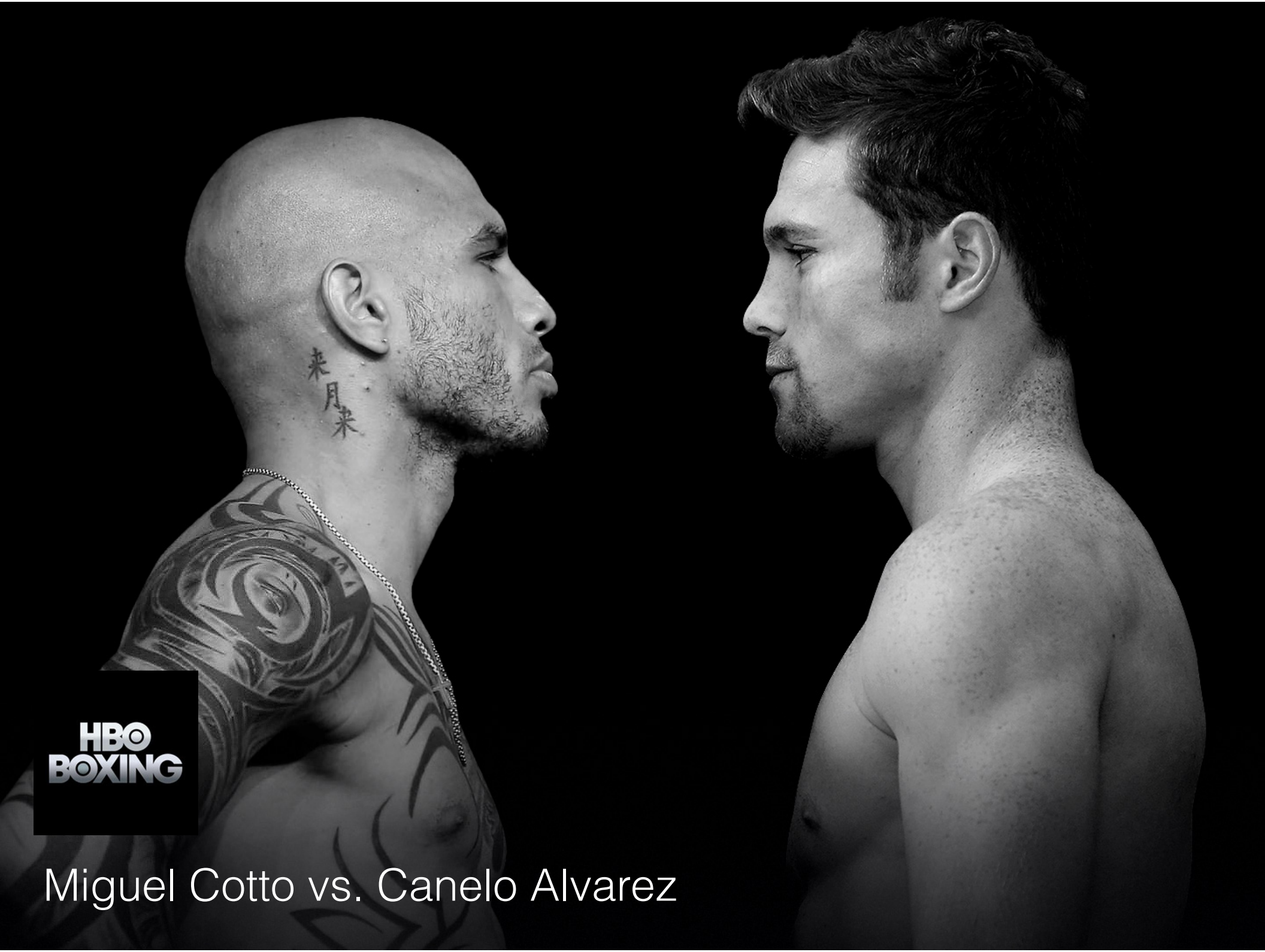
Top Rated | Most Called | Recently Added

Summarizer
Summarize english text
nlp

Sentiment Analysis
Determine positive or negative sentiment from text
nlp

Auto-Tag URL
Automatically generate keyword tags for a URL
tags

AutoTag
Automatically extract tags from text
nlp



Miguel Cotto vs. Canelo Alvarez

RIVAL
London 2012

Verizon 9:39 PM 67%

THROWDOWN SCORING

LIVE SCORING | LIVE STATS

HURD		HARRISON
152/424		118/361
35.8%	TOTALS	32.7%
48/219		44/165
21.9%	JABS	26.7%
104/205		74/196
50.7%	POWER	37.8%

TOTAL PUNCHES LANDED BY ROUND

	1	2	3	4	5	6	7	8	9
HURD	12	9	12	7	17	16	32	33	14
HARR...	9	16	20	8	22	14	13	6	10

VZW Wi-Fi 11:22 PM 19%

THROWDOWN SCORING

CABALLERO		RUIZ
184/657		179/668
28.0%	TOTALS	26.8%
36/268		44/218
13.4%	JABS	20.2%
148/389		135/450
38.0%	POWER	30.0%

TOTAL PUNCHES LANDED BY ROUND

	1	2	3	4	5	6	7	8	9	10
CABA...	15	10	20	18	19	23	17	17	20	25
RUIZ	19	25	27	14	23	9	14	16	13	19

TOTAL | JABS | POWER

HBO WORLD CHAMPIONSHIP BOXING
PUNCHZONE SM Beta
FOREMAN VS COTTO
154-POUND TITLE
FIGHT NIGHT BOUT SCHEDULE
MARTIROSYAN VS GREENE
FOREMAN VS COTTO
31
12
54
13
5
0
18
32
5
8
8
0
PUNCHES LANDED ON FOREMAN
PUNCHES LANDED ON COTTO
SEE BOUT RESULTS
Yuri Foreman
28-0, 8 KOs
PUNCHES LANDED ON FOREMAN
IN-ROUND FIGHT TOTAL
Head 0 97
Body 0 18
Total 0 115
CHART VIEW
Miguel Cotto
34-2, 27 KOs
PUNCHES LANDED ON COTTO
IN-ROUND FIGHT TOTAL
Head 0 55
Body 0 16
Total 0 71
CHART VIEW
ROUND 10
1 2 3 4 5 6 7 8 9 10 11 12
DISPLAY: NUMBERS PERCENTS
PUNCHES: THRU IN ROUND

CompuBox

compuboxonline.com

Apps ha Kernel youtube

COMPUBOX

30 YEARS AND COUNTING

SERVICES | ABOUT | CLIENTS | STORE | CONTACT

SERVICES

CompuBox can supply your website, app or promotion with various real-time stats and historical data. With 28 years of data, we can provide just about any and all boxing data that you may need.

- REAL-TIME STATS FOR YOUR EVENTS
- REAL-TIME DATA FEEDS
- DELAYED STATS FEED
- REPORTS

CONTACT US WITH ANY QUESTIONS ABOUT OUR SERVICES.

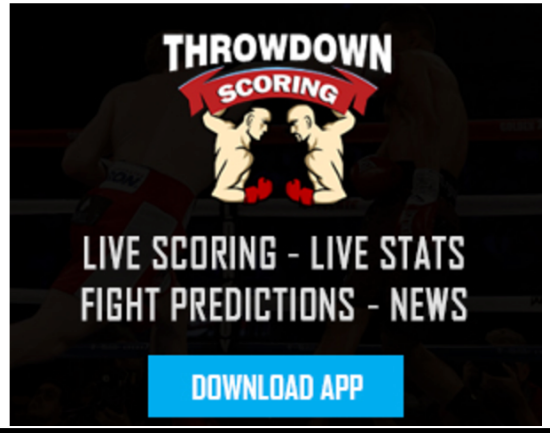
CompuBox Request Form

SUBSCRIBE to RSS Feed

3112 Followers

786 Fans

THROWDOWN SCORING



THE BOXING TRIBUNE

Compubox – Fact or Fiction?

January 20, 2011 by Paul Magno — Leave a Comment

By Rob Day

Quite a quiet, relaxing day full of nothingness and the usual humdrum and then I tuned into my favourite boxing show on the net: TalkinBoxing with Billy C, broadcast out of New York, in the good old US of A. Tuned in and listened to a fantastic debate about Compubox and how accurate it actually is.

CompuBox, Inc. started in 1985 and it really has gone from strength to strength in Boxing and more recently MMA, it's the market leader in this field and for me a welcome addition to a boxing broadcast. At each fight, two operators sit at ringside and each counts the punches thrown and landed by the fighters, one operator concentrates on Fighter A and the other operator on Fighter B. The punches are recorded by the operator tapping a pad that feeds the info in to the Laptop.

Is it accurate?

For me its not 100% accurate, as you have several different things coming into play. It sounds simple sitting watching a fighter and then tapping a pad each time he throws and then lands a punch but a similar system is used in amateur boxing where they have five judges counting the punches landing – if 3 out of the 5 judges see a punch land and tap their pad, then the punch gets counted as a point. Sounds good in theory but the judging at the recent Commonwealth Games in India was woeful. From my TV, I'd see the punch land but apparently all 5 judges missed it! Listening to former WBC Super

MOHEGAN SUN
SUN
Punch
Stat

It’s a hot potato….catch!

EMPIRE CITY
CASINO
DAILY NEWS
DAILY NEWS
REDIT UNION

AVERAGE SPEED COMPARISON (MPH)
THRU ROUND 2 LEAD HAND POWER HAND
MONAGHAN 16.9 28.2 44.1
CLAIBORNE 15.4 16.8 27.7

AVERAGE POWER COMPARISON (LBF)
THRU ROUND 1
LEAD HAND
POWER HAND
HARDEST
MONAGHAN 802 924 1142
CLAIBORNE 601 672 1184

09: 14:45
Corona
Corona

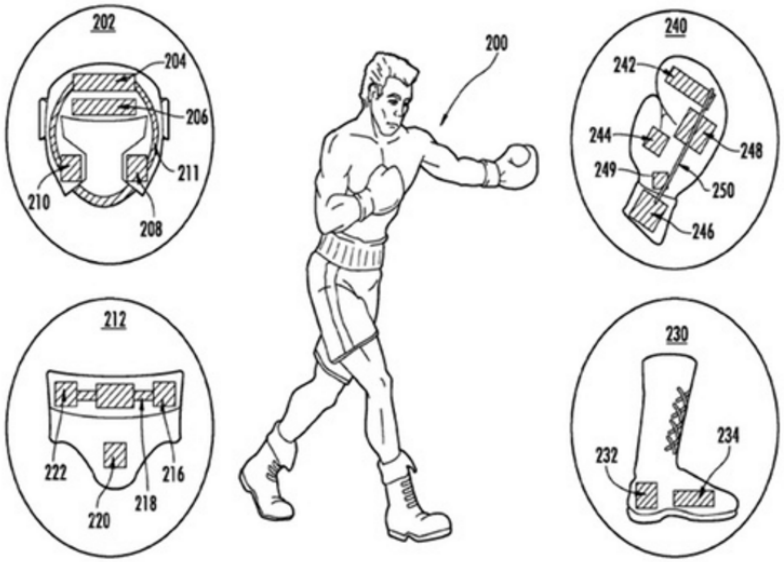
WORDS

HBO PATENTS TECHNOLOGY TO MEASURE PUNCHING POWER AND SPEED DURING FIGHTS

FIGHTLAND BLOG

By Nick Wong

Share 153 Like 153 Tweet 18 G+1 Submit



If you've ever wondered how hard a punch needs to land in order to knock someone out, that now may be a possibility. Late last week, media outlets announced HBO's recent patent on a new technology called PunchForce, a small 7.9-gram microchip that is inserted into the boxing glove to measure attributes like force of impact, punch velocity, and distance coverage. If successful, the added technology could change many aspects of the Sweet Science.

There have been plenty of gizmos in the past that measure the force of a punch, probably the most familiar is the arcade machine seen in gaming halls and **sports bars** that have questionable legitimacy behind their technological accuracy. But a variety of more serious attempts popped up in the past as well. **TenCount** made an attempt at designing a force-measuring punching bag back in 2013, **Puncho** launched a training glove that recorded all punch data through a smartphone, and Australian researchers once plotted out a **boxing vest** aimed at measuring the number and impact of landed blows to lessen scoring discrepancies seen in amateur boxing. None of these technologies have yet made an appearance onto the professional stage, however.

PunchForce does have a chance though. The Nevada State Athletic Commission and the Federal Communications Commission reportedly approved the device for usage back in 2012, and during that time a lot of media was drummed up about the technology, though for whatever reason never came to fruition on the big stage. Now given its new patent deal, news has begun to resurface. In the **patent**, the device is described as using a combined series of gyroscopes and accelerometers to measure the detailed physics of a punch. The intention is to arm analysts with more information for their commentary and add more color for

SPORTS

Say Goodbye to Boxing Judges

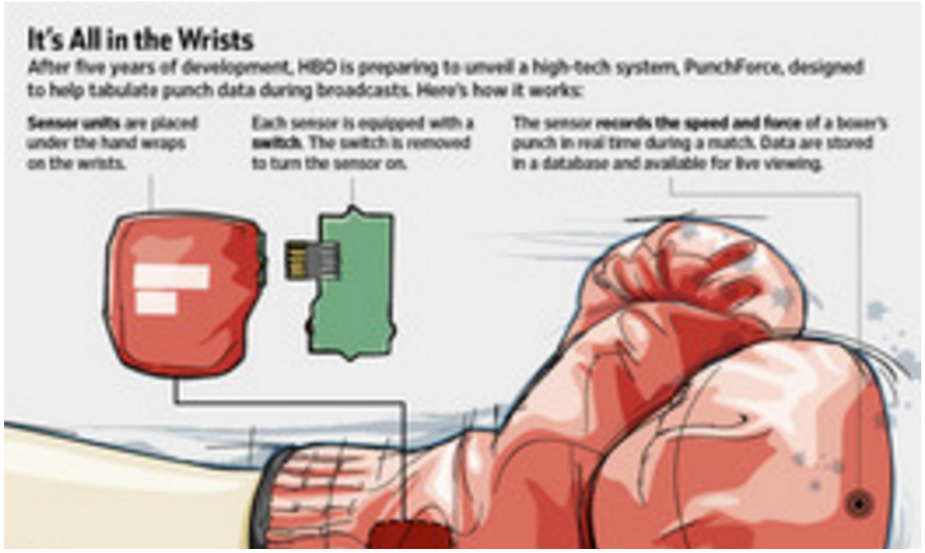
HBO's New Punch-Analysis Gizmo Earns Raves in Trials; Could It Prevent Blown Decisions?

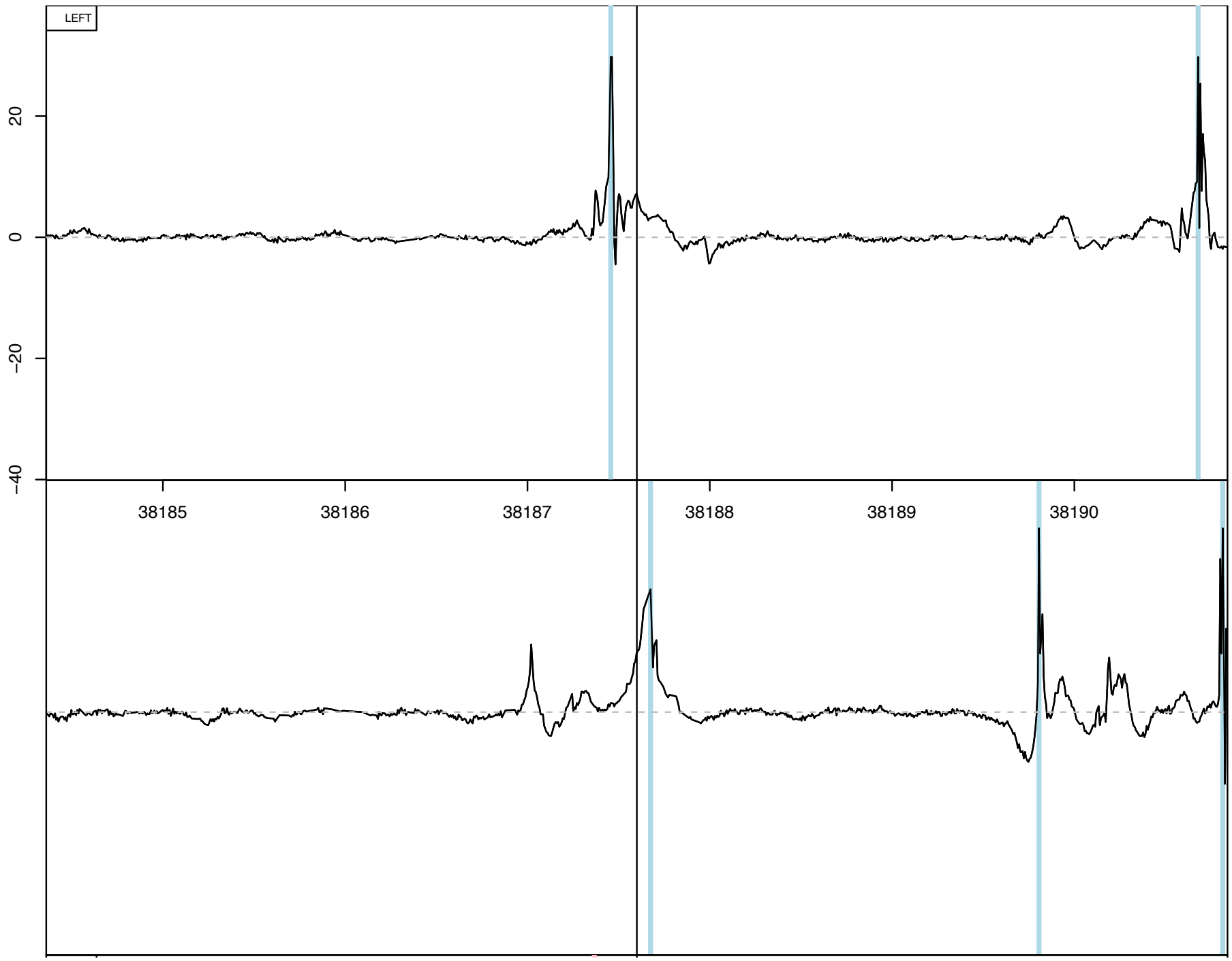
By **TONY OLIVERO**

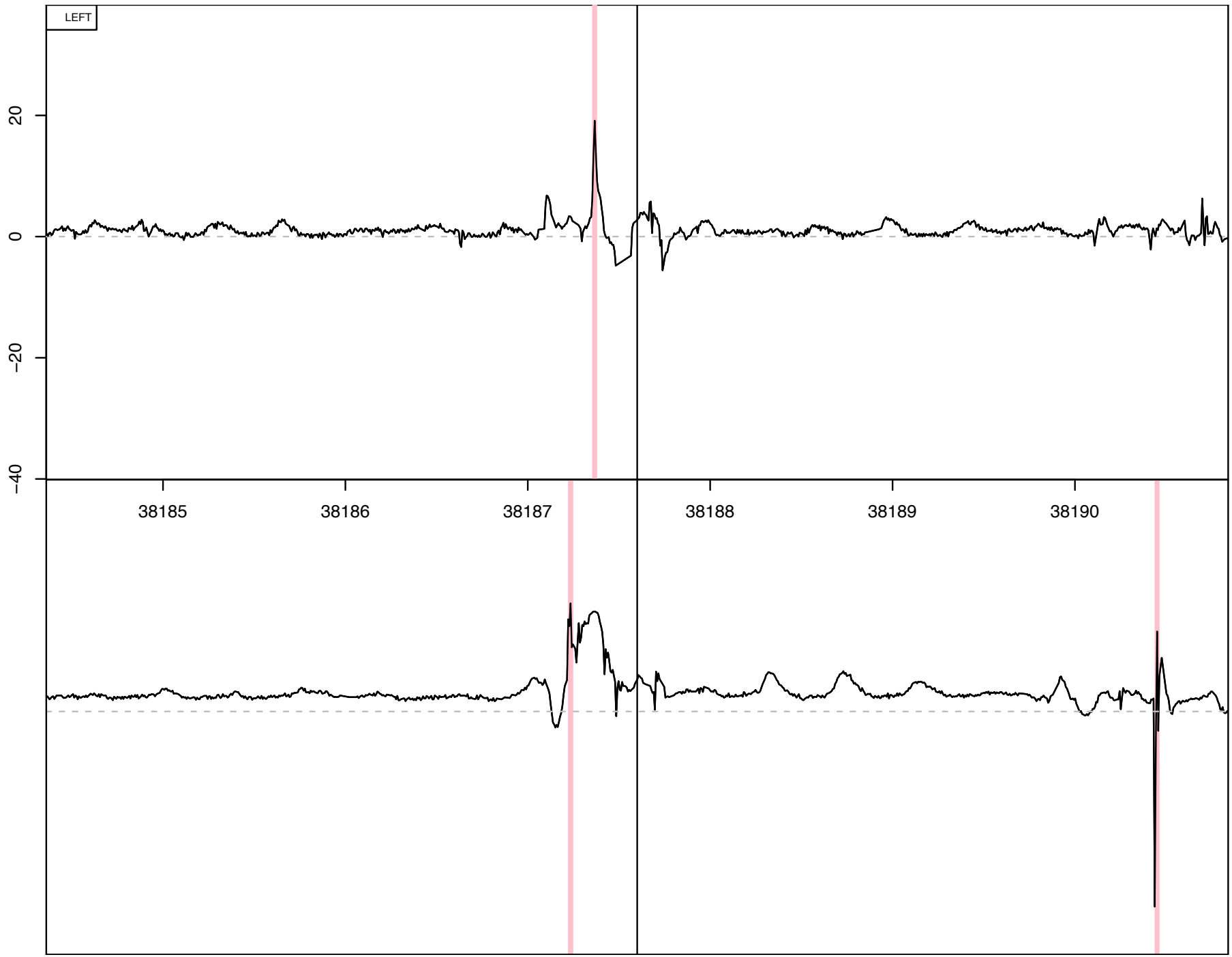
Updated June 25, 2012 8:28 p.m. ET

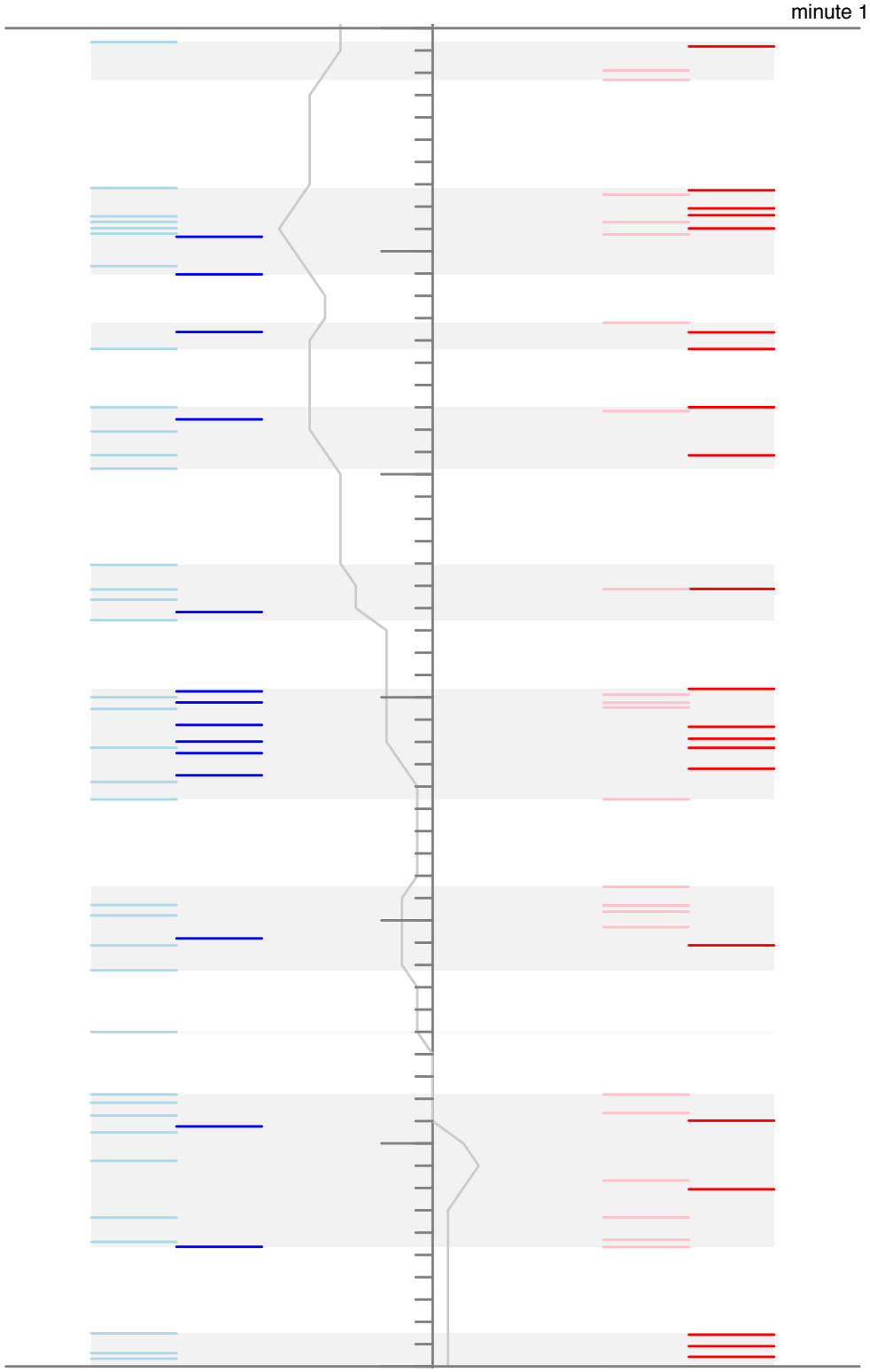
The device is shaped like a piece of Bazooka bubble gum. It weighs in at 7.9 grams and belongs to HBO, the heavyweight among boxing broadcasters.

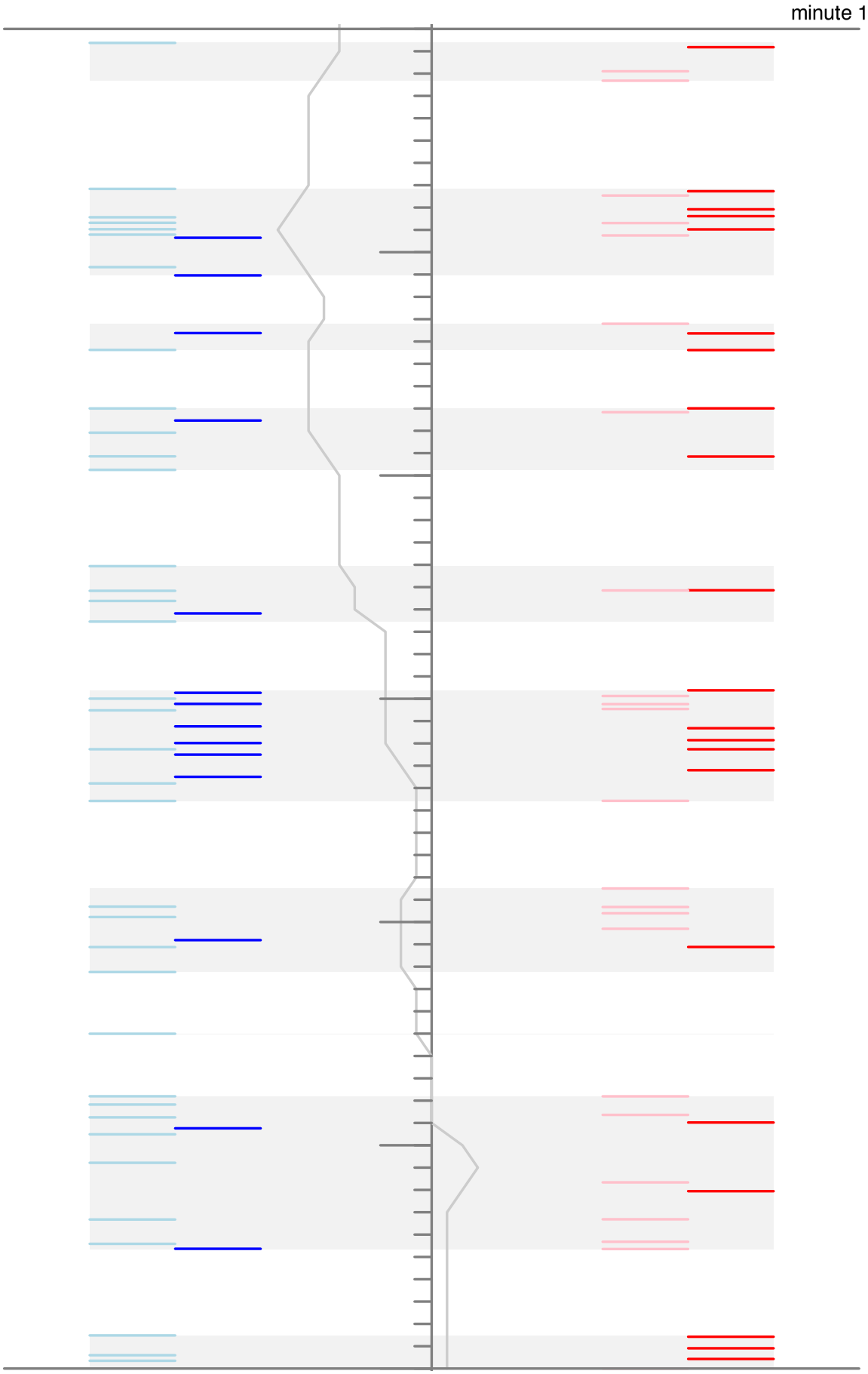
This technology, called PunchForce, is designed to measure the speed and force of a boxer's punches and transmit that information instantaneously to viewers of HBO broadcasts. But its real potential is far broader: If it works, it could help this struggling sport fix one of its most nagging flaws.

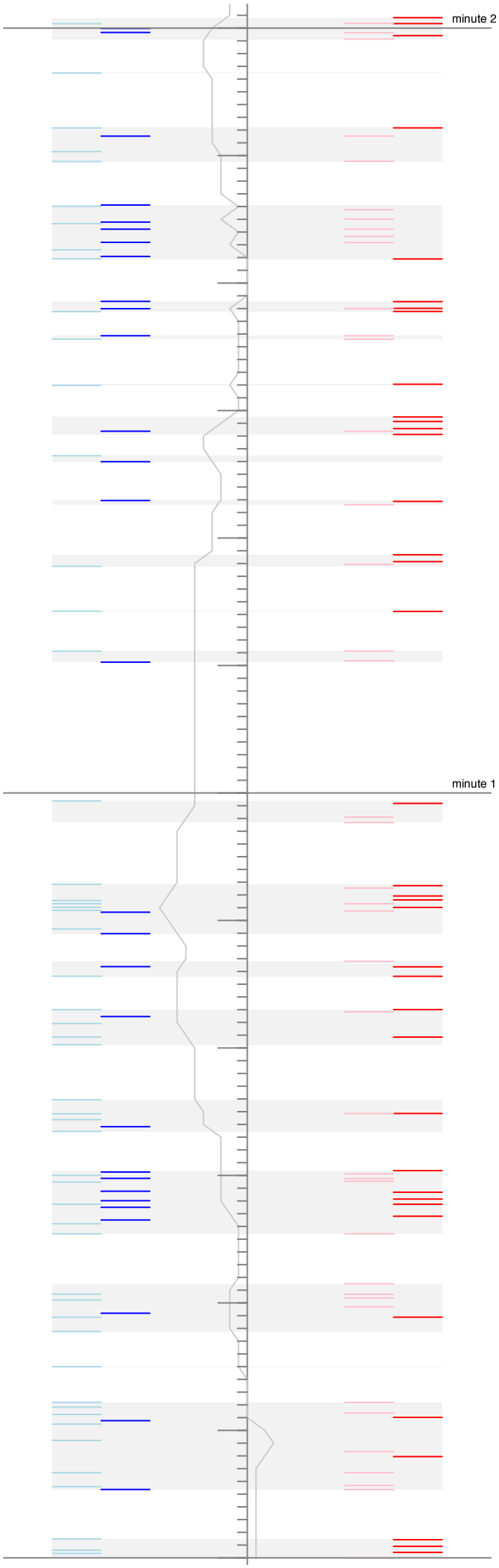


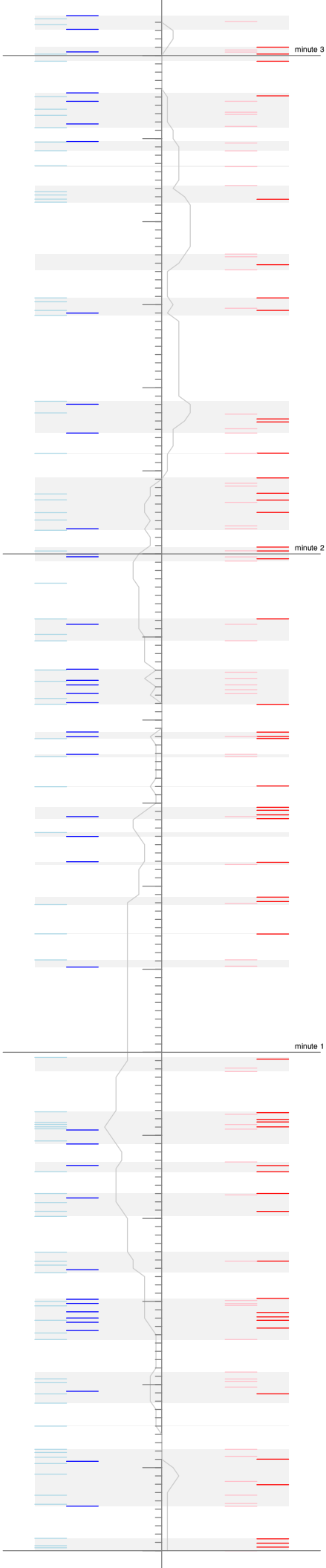

LEFT
20
0
−20
−40
38185
38186
38187
38188
38189
38190


LEFT
20
0
−20
−40
38185
38186
38187
38188
38189
38190


minute 1


minute 1


minute 2
minute 1


minute 3
minute 2
minute 1

EXIT
SUPER BAN
1970 RI

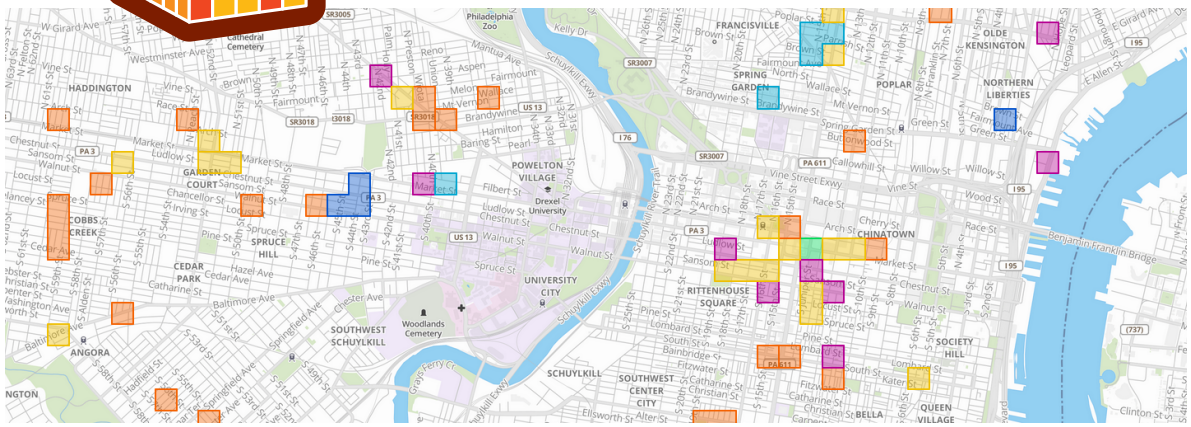
WBO CHAMPION
FEATHER WEIGHT
N GUZMAN
Leasch's

EVERLAST

EST.1910
IN ISLAM

Event Models
Aggravated Assault 41.88%
Gun Crimes 22.36%
Residential Burglary 20.67%
Larceny 6.44%
Simple Assault 4.44%
Trespassing 2.15%
Robbery 2.08%
Risk Percentile 99.38%
Risk Z Score 0.00
Leaflet | DC GIS, DDOT, DRES, OCT

HunchLab



HunchLab: Under the Hood

of academic researchers over the years and more recently re-branded under the "self-exciting point process" moniker.

We also aimed to include concepts such as the Risk Terrain Modeling research being published by Rutgers University, which describes crime locations through correlated geographic features such as bars, schools, and transit stops.

HunchLab 2.0 includes each crime theory by deriving individual sets of variables that represent the concepts within each theory. For instance, we may represent Risk Terrain Modeling by measuring the distance to the nearest bar and the density of bars at each raster cell. We may represent the near repeat pattern concept by measuring the amount of time since the most recent crime occurred in each raster cell. These sets of variables are then passed into the modeling process, which determines the useful theories for a given crime type. The system also determines how the theories interact. For instance, if the near repeat pattern effect is stronger in areas with lots of historic crimes, the system can use that information to enhance the forecast. Or, for example, if assaults frequently happen on Friday evenings near bars, the system can model that effect.

Analysts are able to select which data sets are used to model a particular type of crime. For instance, you could decide to give the assault modeling process access to all geographic layers, in which case the system will include the ones with which it finds correlations. Alternatively, you could pre-screen the geographic layers and exclude a particular layer because you do not feel there is a theoretical basis for its use with assaults.

What modeling methods does HunchLab use to generate forecasts of crime?

In some ways, the model building process in HunchLab mimics the thought process of an experienced analyst. For instance, consider asking an analyst to decide where to place patrol resources for a given upcoming time period. She may start by looking at where crimes have occurred in concentration previously and delineate hotspots of activity. Based on her past experience, she may know that during this particular time period, schools dismiss their students, which increases petty crimes around the schools in the neighborhood. She builds up many such layers of knowledge and balances these various concerns to form a plan. After the time period concludes, she may go back and look at where activities occurred to see if she can determine additional insights into the crime patterns to include in future plans. HunchLab incorporates machine learning concepts to help the software "think" like a crime analyst by imitating years of experience drawn from a police department's own data.

The concept of machine learning is to teach a computer to accomplish a particular task. In this case, we want to teach the computer to determine how likely a particular crime type is to occur at various locations for a given time period. We start this process in HunchLab by forming a set of training examples using the past several years of crime data. Each training example contains the theoretically derived variables we explained above, as well as the outcome (how many crime events occurred). For an entire municipality this training set will often include many millions of example observations. We can then start building the model.

The primary model HunchLab currently uses is a stochastic gradient boosting machine (GBM) comprised of decision trees trained using the AdaBoost loss function. This model is built to forecast whether a crime event will occur or not in a given space-time raster cell (a binary outcome). The general way this model works is as follows:

- Begin by selecting a random portion of the training examples.
- Build a decision tree that separates examples of where crimes occurred from ones that didn't based upon the variables.
  - For instance, the first decision within the tree might be interpreted as: "if no event happened in the last year in this location, it is very unlikely for a crime to occur today". The decision tree then splits the examples into two sets: (1) where a crime occurred during the past year and (2) where no crimes occurred during the past year.
  - Within each set, the process repeats. For example, the next decision for the set of locations with crimes in the past year might be interpreted as: "if an event happened in the last week, it is more likely for one to occur today". This set of examples would again be split based upon this decision rule.
  - This process continues to build out a decision tree that describes why crimes occur where they do.
- The decision tree is then used to make predictions of how likely crimes are for each observation in the entire training data set.
- This completes one training iteration within the boosting machine.
- The modeling process then begins again.
  - We start by selecting another random portion of the training examples. This random sampling process is why the model is stochastic.
  - In this next iteration, we build another decision tree (in the same manner as above). This time, however, we build the tree to predict the errors from applying the first decision tree model to this new sample of observations. In this way we are attempting to correct our mistakes. This concept is called boosting.
  - We then use these two trees to make predictions across the entire data set.
  - As we conduct this process, we can keep track of how many training iterations within the machine have made incorrect predictions for each training example. We increase the importance (via weights) of observations that we continue to get wrong and decrease the importance of observations that we continue to get correct. This process is called adaptive boosting (AdaBoost). When we build the next decision tree, we tell it to focus on the observations that we continue to get wrong via these weights.
- Training iterations continue several hundred times. The resulting model represents tens of thousands of decision rules of why crimes occur where they do.
- We conduct this entire process several times, each time holding back a portion of the example data. We can then use each of these models to make predictions for this held-out set of data to

Risk Terrain Modeling | Official

www.riskterrainmodeling.com

Apps ha Kernel youtube

HOME OVERVIEW GETTING STARTED RESOURCES TRAINING RTM IN PRACTICE CONTACT BLOG

RISK TERRAIN MODELING

DOCUMENTATION & DEMO SITE

GET STARTED

Intervention Planning - Risk Te ×

www.riskterrainmodeling.com/intervention-planning.html

Apps ha Kernel youtube

HOME OVERVIEW GETTING STARTED RESOURCES TRAINING RTM IN PRACTICE CONTACT BLOG

INTERVENTION PLANNING INTEL REPORTS

An Introduction to IPIRs

An Intervention Planning Intel Report (IPIR) has many components, but its main purpose is to communicate the ways in which the problem under study occurs in the study setting, and the factors that are important in elevating risks of it continuing. The IPIR is place-based. It prioritizes the risk factors to be addressed relative to their impact on the problem so that resources can be allocated and people can be tasked with what to do about the places when they get there. **Our analyses for police departments are routinely presented using the systematic framework of an IPIR.**

Problem: Violent Crime **Jurisdiction:** City, Urban

Scenario Title: "All In" **Key Risk Factors:** Gas stations, Bodegas, Vacant Buildings

Origin: Risk-based policing initiative, part of Project Safe Neighborhoods (PSN). IPIR presented for first time, discussed at task force meeting with variety of city and community stakeholders present.

>> Conversation quickly centered on "gas stations", one of the highest risk factors identified by RTM. Police officers in the room had a hunch that gas stations were "bad", particularly for robberies, carjackings, and assaults targeted at station attendants or customers fueling vehicles. A quick review of incident report narratives confirmed this *event context* in instances where 'gas stations' were mentioned.

>> A **community stakeholder** working with youth added to conversation. She explained: Many youth hangout after school around corner stores/bodegas where they can easily congregate and get food, drinks, and rolling paper for smoking at nearby vacant buildings. Bodegas close at 10pm, as required by city ordinance, but gas stations with food marts are exempted. These 24/7 gas stations provide space and "supplies" for youth to congregate late at night, creating a unique context for turf conflict, offending, or victimization. So, she explained, prioritize gas stations with food marts.

>> **Department of Public Works (DPW)** added to conversation, offering to prioritize DPW's existing efforts to board-up vacant buildings and clean lots near bodegas and gas stations w/ food marts.

>> **Mayor's Office** and **Department of Parks and Recreation** will enhance their advertising and recruitment campaigns for summer recreation activities and job opportunities at bodegas, gas stations and places nearby – to focus on where the youth are.

>> **Police (PD)** will deploy *directed patrols* and *meet-and-greets with managers* at risky facilities, at peak risky times; Patrol will make referrals to partnering community agencies/social workers when juveniles appear in need of outreach/support services. A new protocol for crime reporting will help with continued risk assessments: Officers will ask victims/witnesses about possible connections of crime incidents to known spatial risk factors (e.g. was victim on premise or coming from gas station before robbery?).

Table 3. Risk factor testing.

Risk Factor		Motor Vehicle Theft				Motor Vehicle Recovery			
	N	Op.	S.I.	Coef.	RRV	Op.	S.I.	Coef.	RRV
In the final RTM									
Disorder calls	16,325	Dens.	1104 ft	1.30	3.69	Dens.	552 ft	1.59	4.88
Foreclosures	899	Prox.	1656 ft	0.89	2.46	Prox.	1656 ft	0.75	2.11
Multi-family housing complexes	11,261	Prox.	1656 ft	0.90	2.46	Prox.	1656 ft	0.65	1.92
Hotels & motels	111	Prox.	552 ft	0.64	1.90	Prox.	552 ft	0.63	1.88
Sit-down restaurants	550	Prox.	1656 ft	0.61	1.85	Prox.	1656 ft	0.41	1.50
Parks	14,464	Prox.	1656 ft	0.44	1.55	Prox.	1656 ft	0.48	1.61
Commercial zoning	13,485	Prox.	1656 ft	0.31	1.37	Prox.	1656 ft	0.52	1.68
Convenience stores	77	–	–	–	–	Prox.	552 ft	0.62	1.85
Gas stations w/convenience stores	28	–	–	–	–	Prox.	1656 ft	0.50	1.68
Schools	106	–	–	–	–	Prox.	1656 ft	0.30	1.35
Intercept (rate)	–	–	–	−6.56	–	–	–	−6.48	–
Intercept (overdispersion)	–	–	–	−1.05		–	–	−0.82	–
Tested but not in the final RTM									
Bars	111	–	–	–	–	–	–	–	–
Bowling centres	8	–	–	–	–	–	–	–	–
Liquor stores	93	–	–	–	–	–	–	–	–
Malls	6	–	–	–	–	–	–	–	–
Night clubs	15	–	–	–	–	–	–	–	–
Parking stations & garages	6	–	–	–	–	–	–	–	–
Retail shops	57	–	–	–	–	–	–	–	–
Takeout restaurants	299	–	–	–	–	–	–	–	–
Variety stores	21	–	–	–	–	–	–	–	–

Notes: Abbreviations: Op., Operationalisation; Dens, Density; Prox., Proximity; S.I., Spatial influence; Coef., Coefficient; RRV, Relative risk value.
RTMDx only accepts point files as inputs. Multi-family housing complexes, parks and commercial zoning were provided as polygons and converted to polygons prior to the analysis. To conduct the conversion, researchers first converted the perimeter of each polygon to a series of points placed about one half block (i.e. 276 ft) from each other. To convert the interior of the polygon, the vector was first converted to a raster grid with cell size of 276 ft. The raster was then converted to a point file, with each raster cell centroid represented as a single point. The *N* values in Table 1 are the number of points input into RTMDx. The points were created from the following number of polygon features: 459 multi-family housing complexes, 26 commercial zoning areas and 530 parks.

Technology and story

In this example, we again see machine learning as a process that starts with the available data (and which variables are permissible and which are not) — actually, even before we collect a bit of data, there's the rationale for appealing to ML at all, which often involves some appeal to "objectivity"

Notice that the two estimation procedures we have here are quite different in terms of the "visibility" into the model — or, perhaps, it is the descriptive capability of the model that is used for RTM, whereas HunchLab is most interested in predictive accuracy

This divide is an old one — not all ML approaches are created equal. Some are opaque boxes that receive inputs and return outputs, while others have clear mathematical expressions for their computation, allowing you to isolate features of the model easily.

The understandability or "explainability" of a model is important in some cases, but perhaps less so in others — It does account for two modeling cultures that should be pointed out before we get too far into the subject.

Statistical Science
2001, Vol. 16, No. 3, 199–231

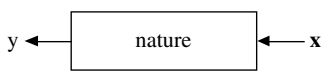
Statistical Modeling: The Two Cultures

Leo Breiman

Abstract. There are two cultures in the use of statistical modeling to reach conclusions from data. One assumes that the data are generated by a given stochastic data model. The other uses algorithmic models and treats the data mechanism as unknown. The statistical community has been committed to the almost exclusive use of data models. This commitment has led to irrelevant theory, questionable conclusions, and has kept statisticians from working on a large range of interesting current problems. Algorithmic modeling, both in theory and practice, has developed rapidly in fields outside statistics. It can be used both on large complex data sets and as a more accurate and informative alternative to data modeling on smaller data sets. If our goal as a field is to use data to solve problems, then we need to move away from exclusive dependence on data models and adopt a more diverse set of tools.

1. INTRODUCTION

Statistics starts with data. Think of the data as being generated by a black box in which a vector of input variables **x** (independent variables) go in one side, and on the other side the response variables **y** come out. Inside the black box, nature functions to associate the predictor variables with the response variables, so the picture is like this:



There are two goals in analyzing the data:

Prediction. To be able to predict what the responses are going to be to future input variables;
Information. To extract some information about how nature is associating the response variables to the input variables.

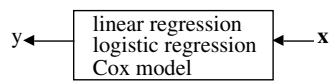
There are two different approaches toward these goals:

The Data Modeling Culture

The analysis in this culture starts with assuming a stochastic data model for the inside of the black box. For example, a common data model is that data are generated by independent draws from

$$\text{response variables} = f(\text{predictor variables, random noise, parameters})$$

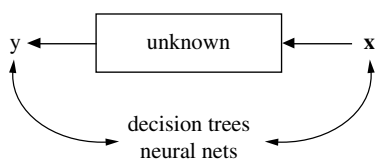
The values of the parameters are estimated from the data and the model then used for information and/or prediction. Thus the black box is filled in like this:



Model validation. Yes–no using goodness-of-fit tests and residual examination.
Estimated culture population. 98% of all statisticians.

The Algorithmic Modeling Culture

The analysis in this culture considers the inside of the box complex and unknown. Their approach is to find a function $f(\mathbf{x})$—an algorithm that operates on **x** to predict the responses **y**. Their black box looks like this:

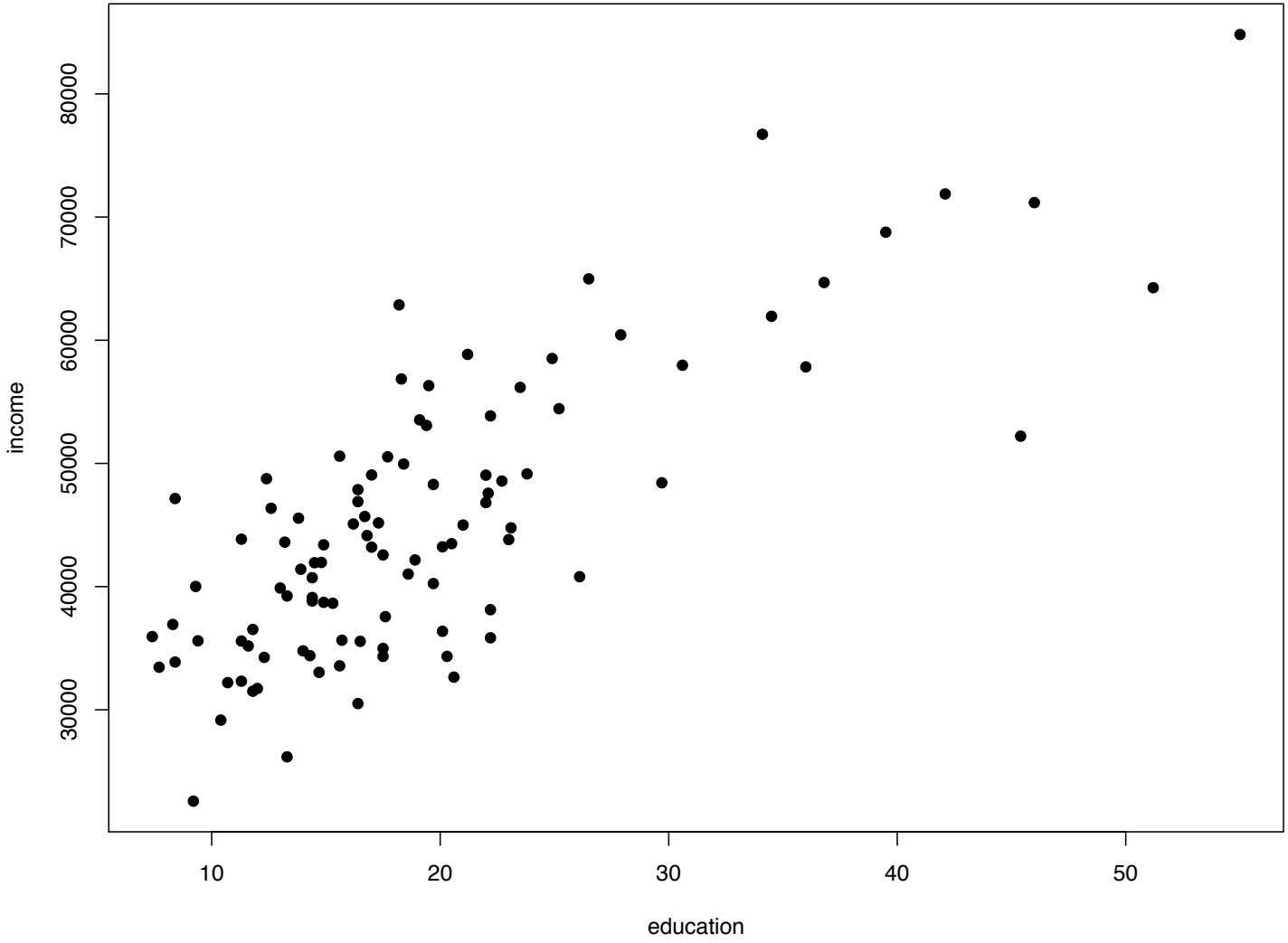


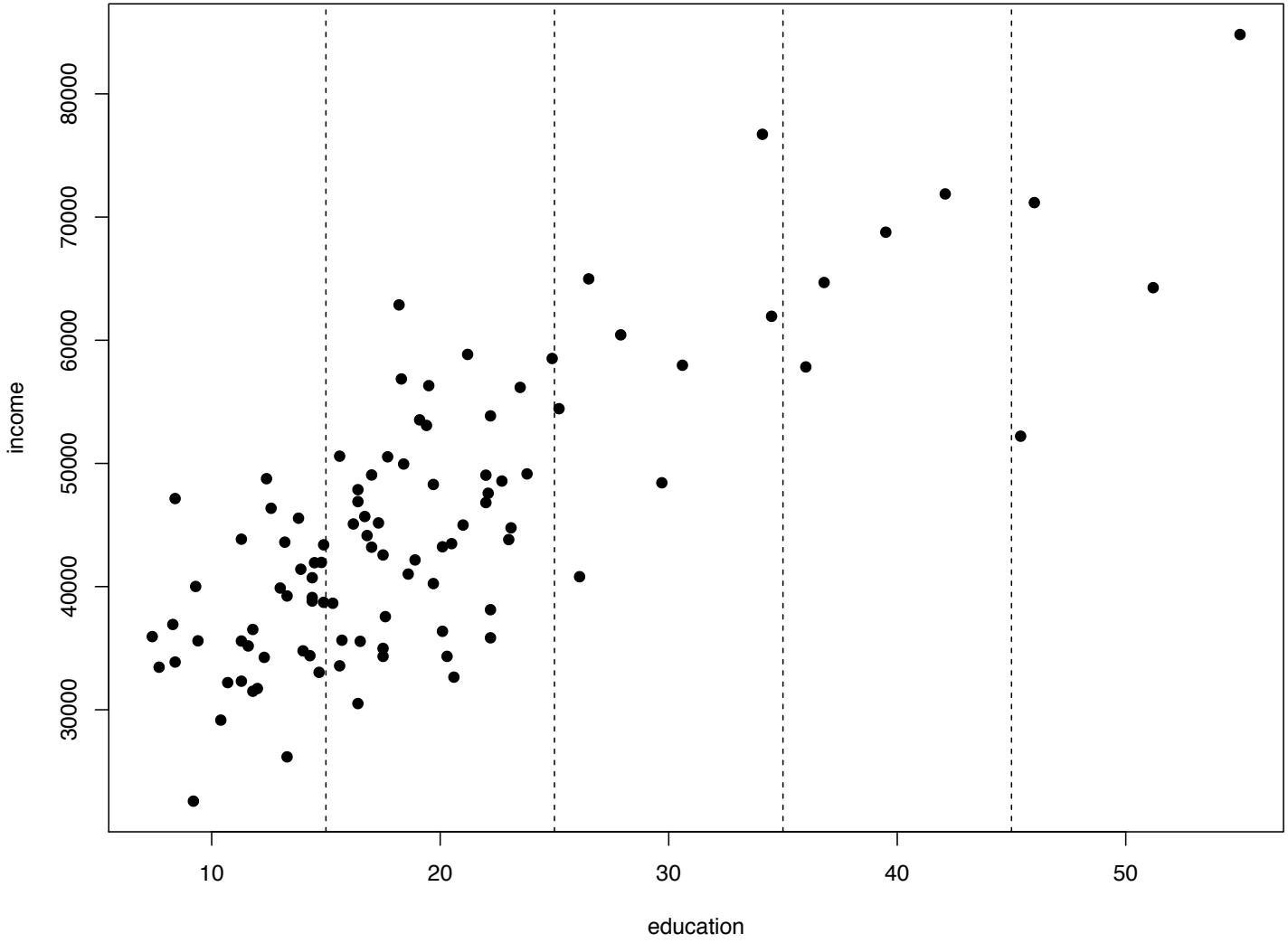
Model validation. Measured by predictive accuracy.
Estimated culture population. 2% of statisticians, many in other fields.

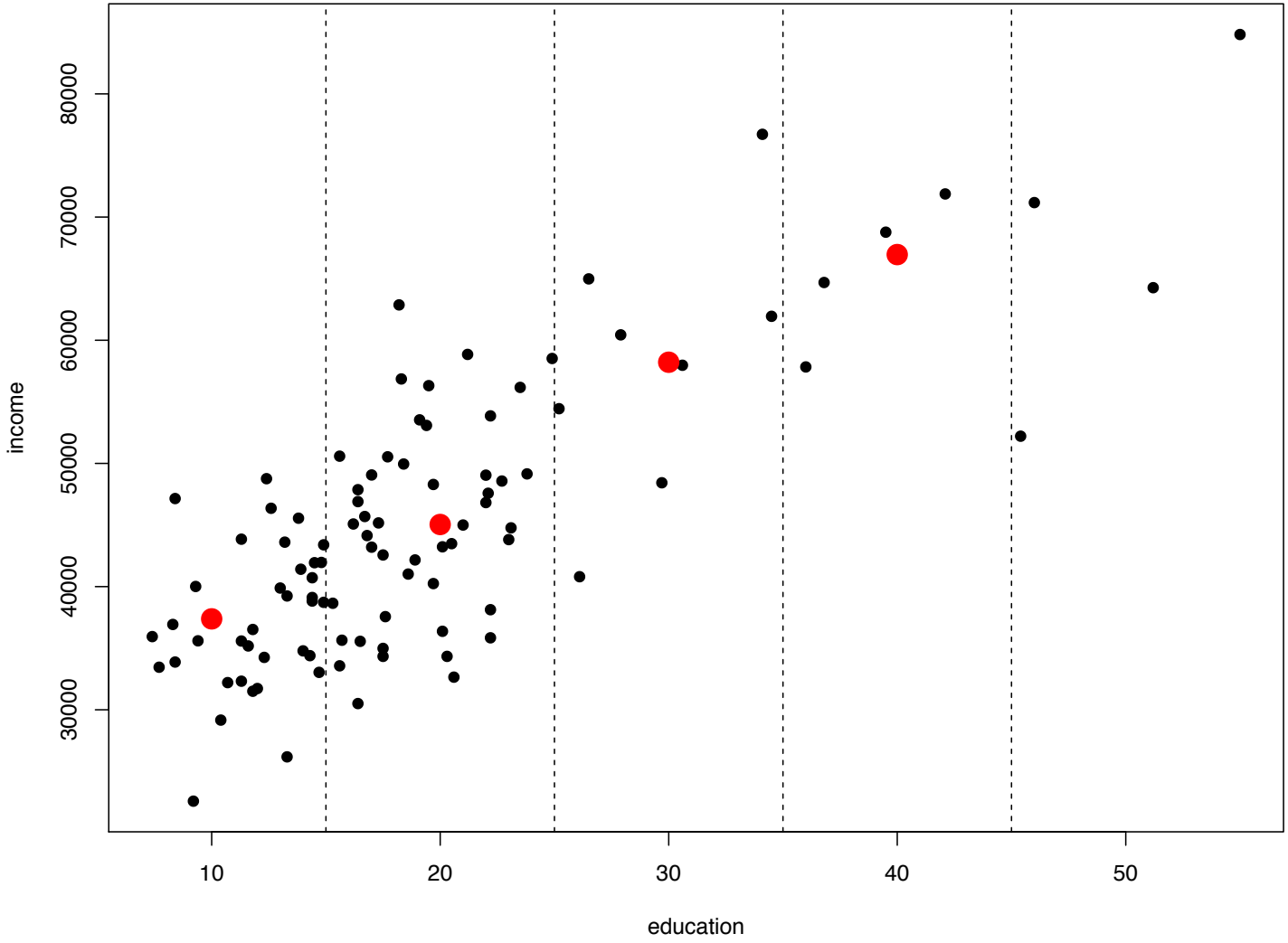
In this paper I will argue that the focus in the statistical community on data models has:

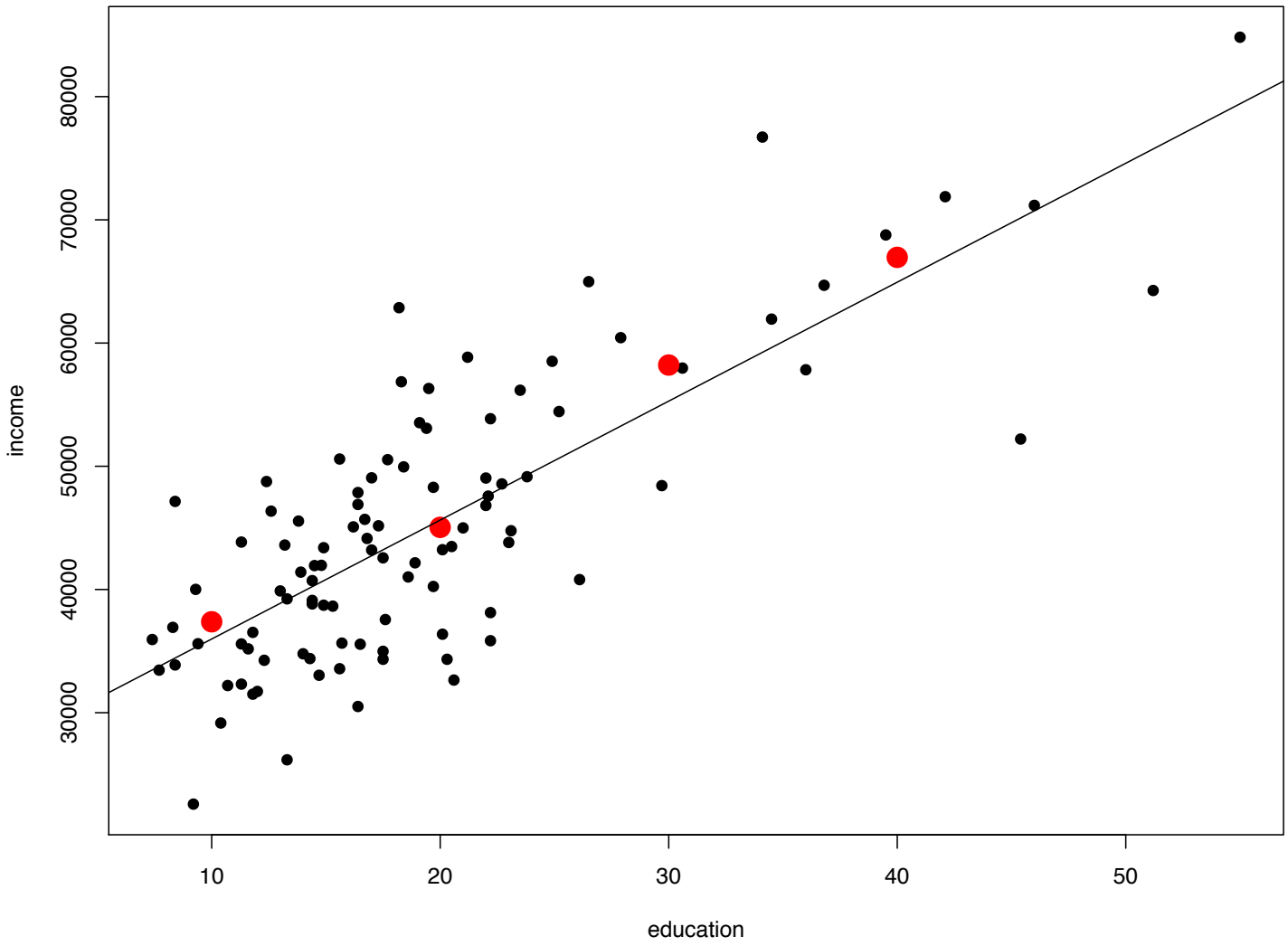
- Led to irrelevant theory and questionable scientific conclusions;

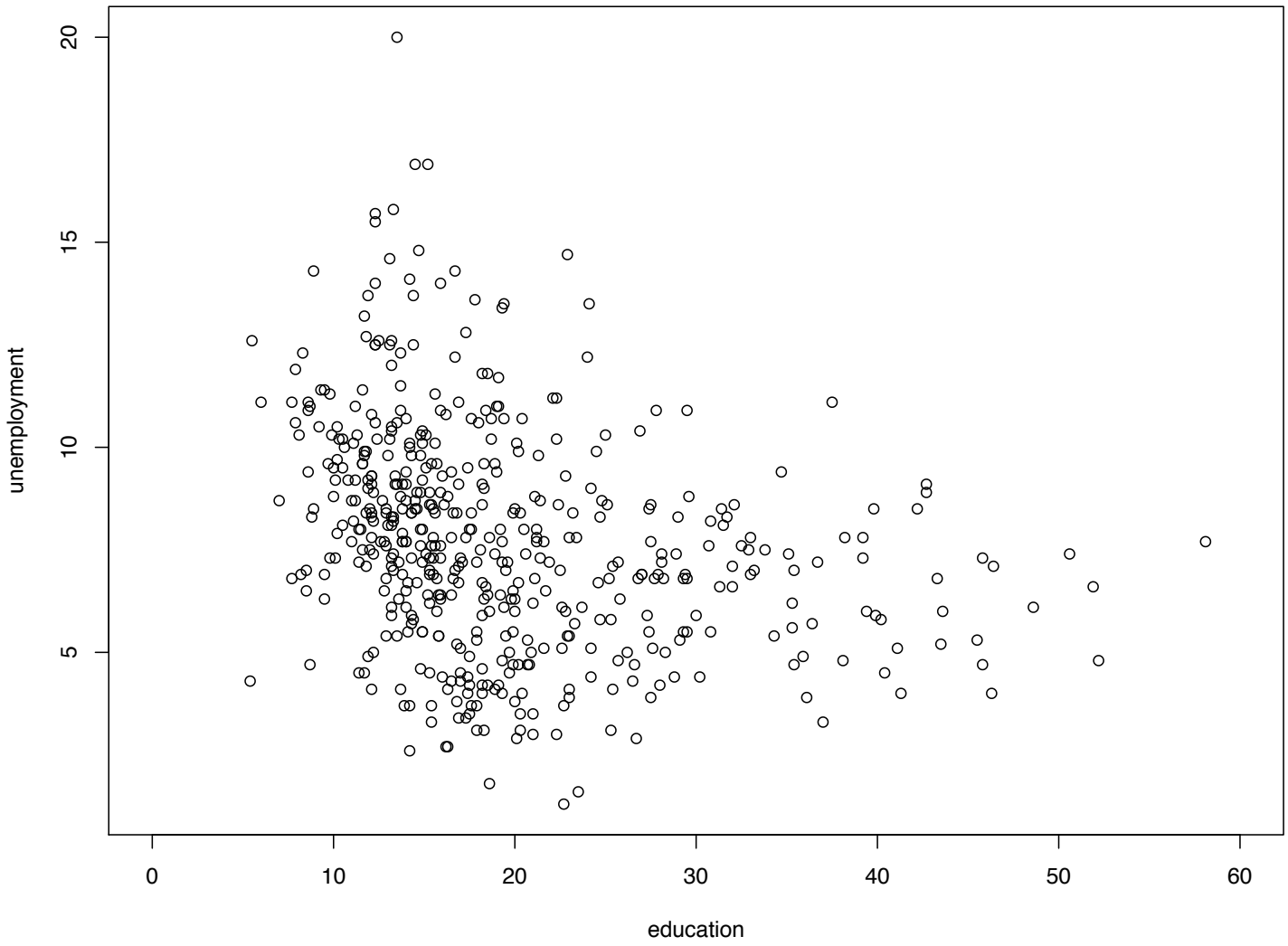
Leo Breiman is Professor, Department of Statistics, University of California, Berkeley, California 94720-4735 (e-mail: leo@stat.berkeley.edu).

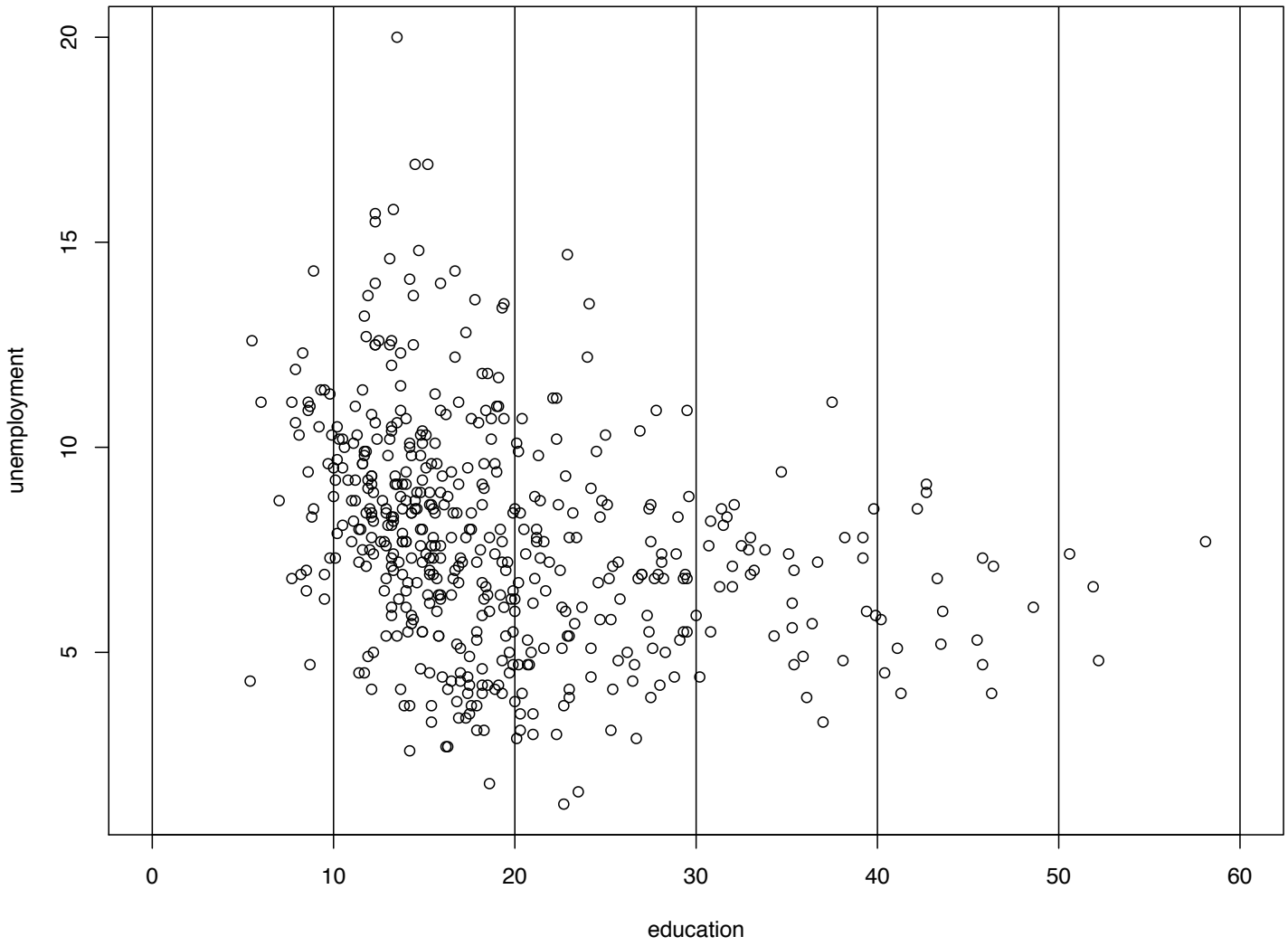

income
30000
40000
50000
60000
70000
80000
education
10
20
30
40
50

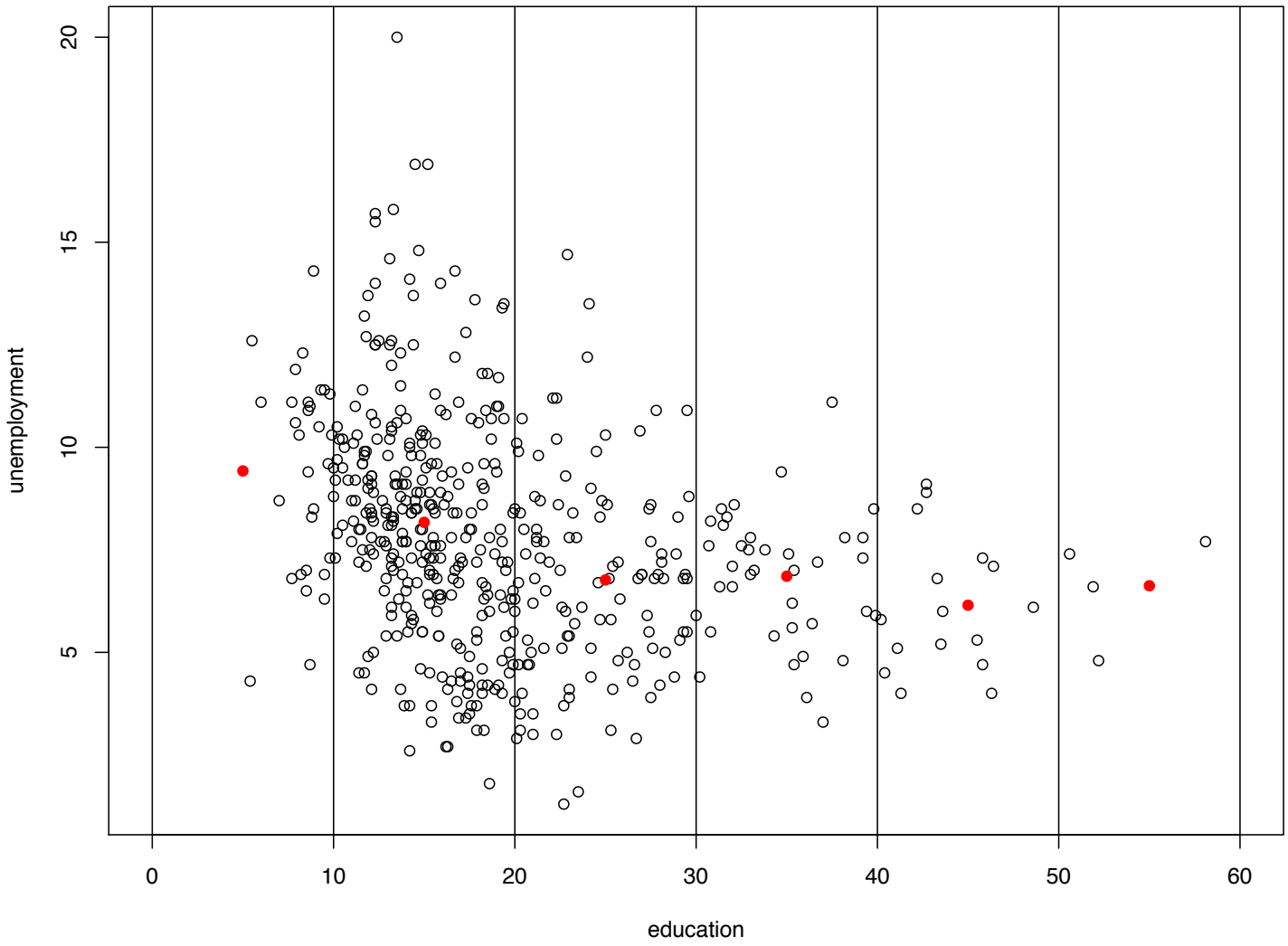

income
30000
40000
50000
60000
70000
80000
education
10
20
30
40
50

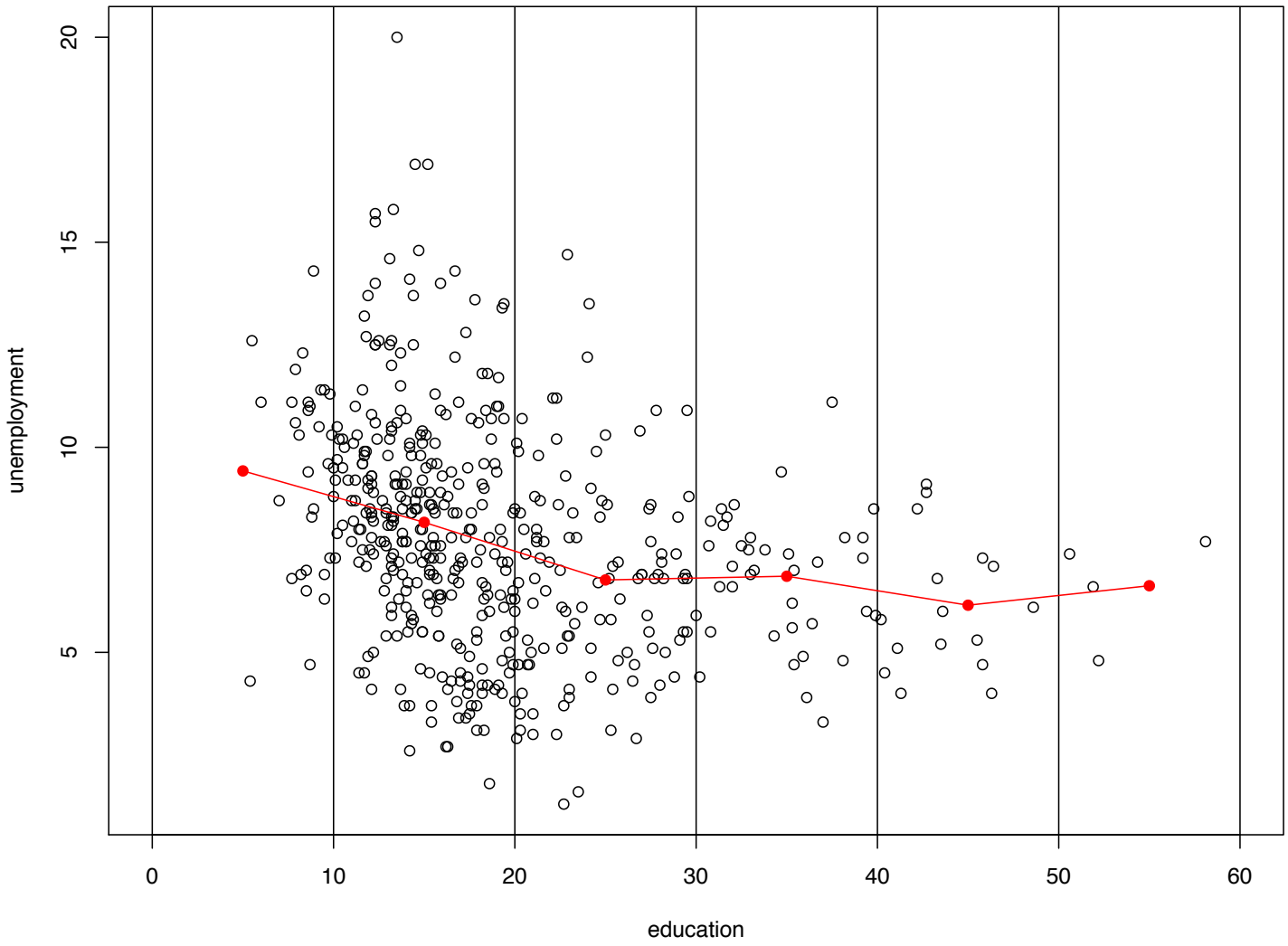

income
30000
40000
50000
60000
70000
80000
education
10
20
30
40
50


income
30000
40000
50000
60000
70000
80000
education
10
20
30
40
50


unemployment
5
10
15
20
education
0
10
20
30
40
50
60


unemployment
5
10
15
20
education
0
10
20
30
40
50
60


unemployment
5
10
15
20
education
0
10
20
30
40
50
60


unemployment
5
10
15
20
0
10
20
30
40
50
60
education