Diplomado de Modelado Predictivo y Machine Learning

Introduction

Hamdi Raissi
José Ruette

PONTIFICIA UNIVERSIDAD CATÓLICA DE VALPARAÍSO

Data Science context

PONTIFICIA UNIVERSIDAD CATÓLICA DE VALPARAÍSO

Data Science Initiatives

The Michigan Daily

NEWS SPORTS OPINION ARTS STATEMENT

TRENDING TOPICS JIM HARBAUGH FOOTBALL

University announces $100 million data science initiative

By TANYA MADHANI, Daily Staff Reporter
Published Tuesday, September 8, 2015 - 5:32pm

The University will invest $100 million in a new Data Science Initiative over the next five years with the aim of enhancing learning and research opportunities for students and faculty members.

To support the initiative, the University will hire 35 new faculty members over the next four years and launch the Michigan Institute for Data Science, which will lead educational and research opportunities related to big data. Massive sets of data can help researchers produce new insights into a broad spectrum of topics, from learning and medicine to transportation and social media.

"Big data can provide dramatic insights into the nature of disease, climate change, social behavior, business and economics, engineering, and the basic biological and physical sciences," University President Mark Schlissel wrote in a statement. "With our widely recognized strengths across all of these areas and our longstanding culture of collaboration across disciplines, U-M is in a unique position to leverage this investment in data science for the good of society."

More like this:

- Michael Smallegan: Change with MCubed
- University considers data collection policy changes
- Symposium helps cancer researchers share knowledge

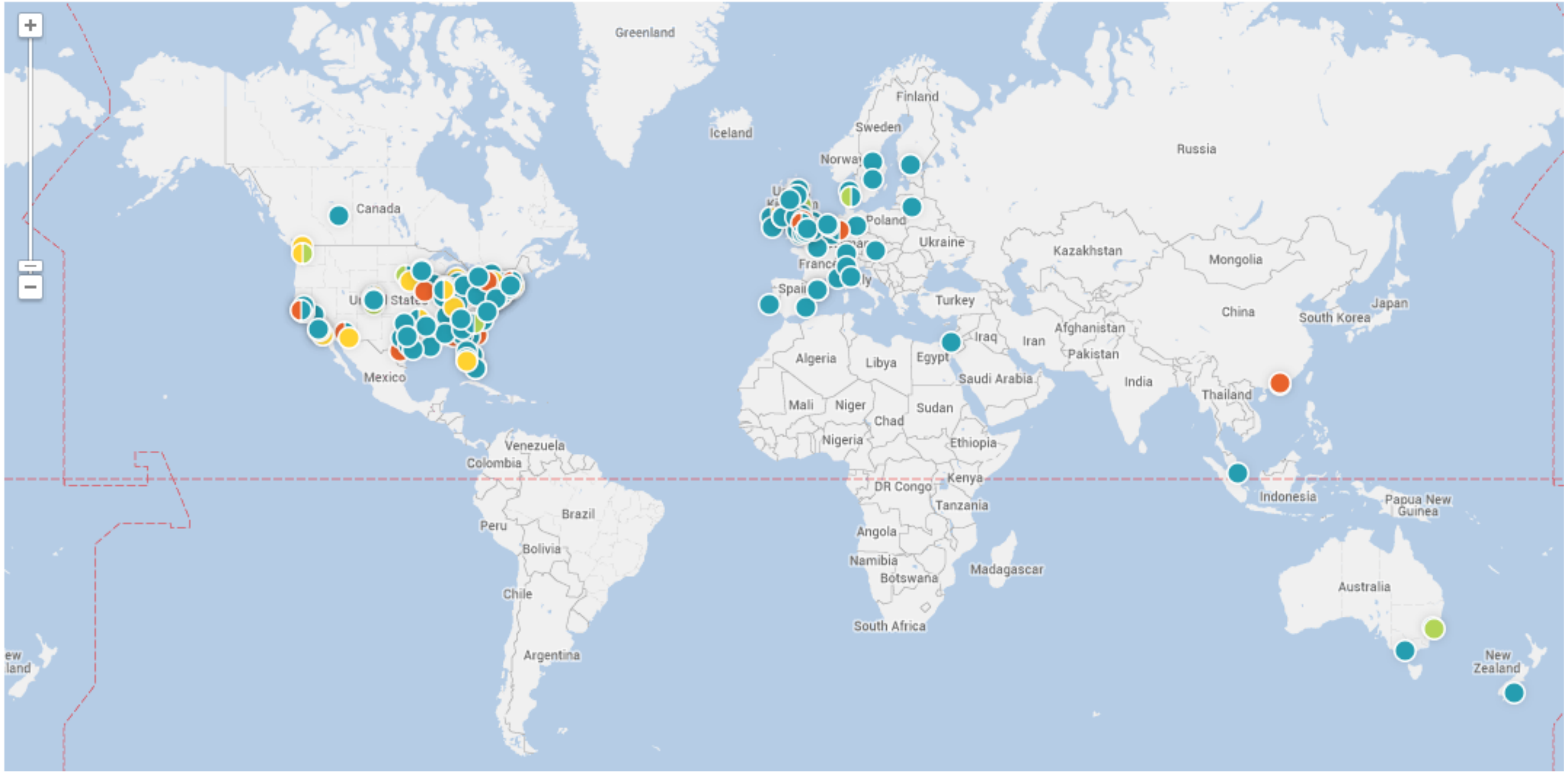
MOST READ

- WATCH LIVE: Schlissel to discuss diversity initiative at 11 a.m.
- CSG partners with University Dining to offer early game day breakfast
- Cartoon: Kim's new adventures
- City Council discusses proposed apartment complex
- At first meeting, CSG discusses fall priorities

Data Science Initiatives

A Map of Data Science Degree Programs Around the The World

Filter by: country state degree oncampus



Data Science Initiatives

UCDAVIS **News and Information**

9.9.2015

Events | Facts & Figures | Media Contacts | Adv

News home page

News by topic
- Agriculture
- Arts/Humanities
- Business
- Campus News
- Engineering
- Government/Law
- Health
- Science News
- Social Sciences
- Veterinary Medicine

Publications
- UC Davis Magazine
- Dateline UC Davis
- Dateline electronic newsletter
- Egghead blog
- Futurity.org
- California Aggie

Video
- iTunes U
- YouTube

Related news
- Chancellor Katehi
- UC Davis Athletics
- UC Davis Health System
- UC Newsroom
- UCTV

Computer scientist Koehl to lead Data Sciences Initiative

DATELI
News for Faculty and Staff

2.11.2014

By Dateline staff

Provost and Executive Vice Chancellor Ralph J. Hexter has appointed Professor Patrice Koehl, a comput biologist in the Department of Computer Science and an active researcher in the UC Davis Genome Ce founding director of the new Data Sciences Initiative, effective May 1.

The initiative emerged as a recommendation of the campus's Big Data Implementation Committee. He appointed the committee members, whom he charged, in part, "to propose the path for developing str potentially disruptive academic research and teaching programs that address the opportunities and ch presented by the Big Data revolution."

Koehl

As the founding director, Koehl will work with the Data Science Initiative Implementation (faculty and deans, to develop the initiative's mission and initial strategic plan, including a model and funding plan.

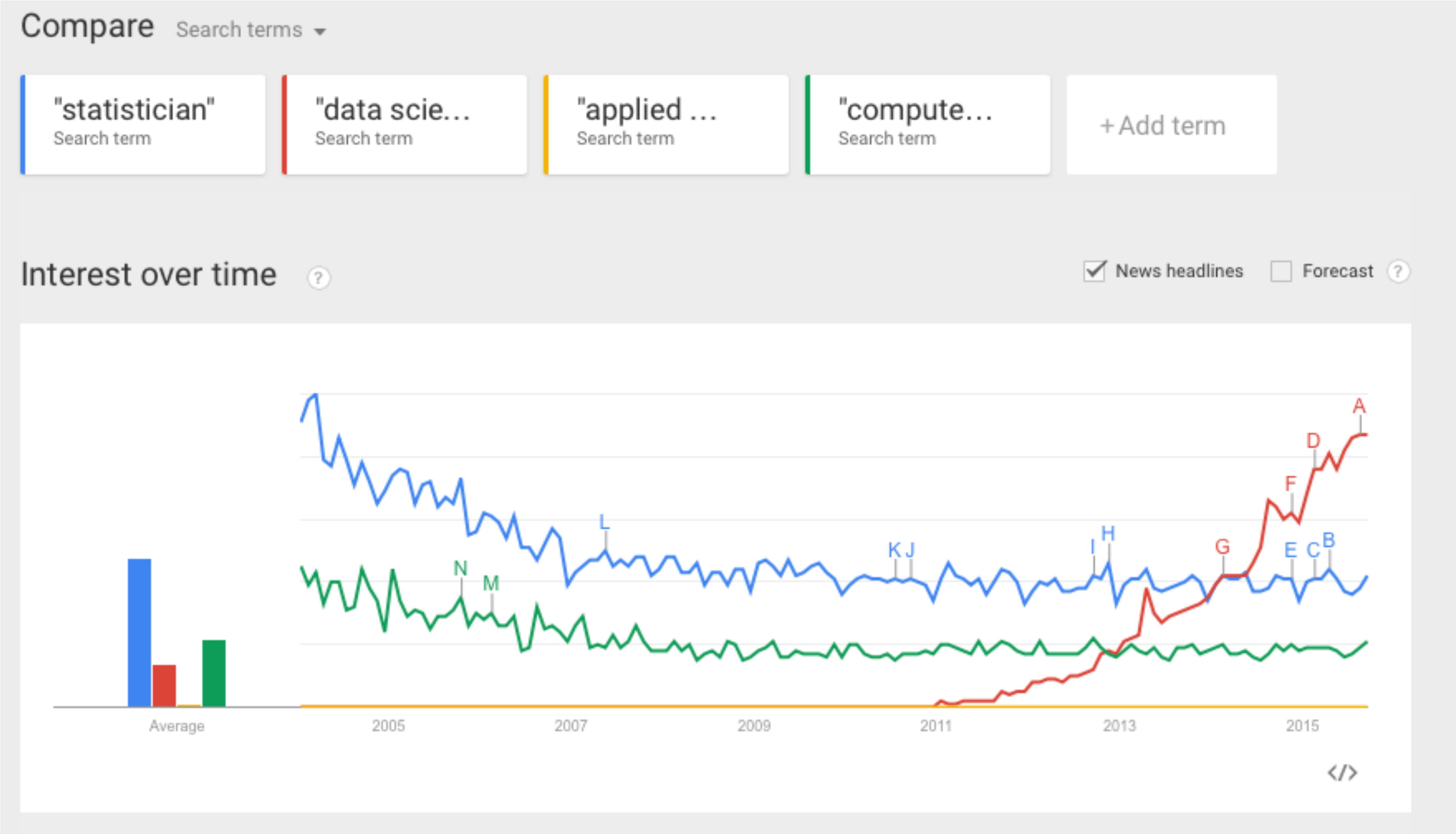
In partnership with the implementation committee, Koehl will further define the physical a infrastructure and services needed to support and coordinate various elements of Big Data UC Davis, where faculty are already accessing and analyzing data sets of unprecedented c and thereby opening up new vistas for inquiry and discovery.

In appointing Koehl, Hexter extolled his vision for the initiative and its significance for UC Davis, in adc combination of technical skills and administrative experience.

"Professor Koehl has the imagination, enthusiasm and practical savvy to help us maximize our potentia exciting new field," Hexter said.

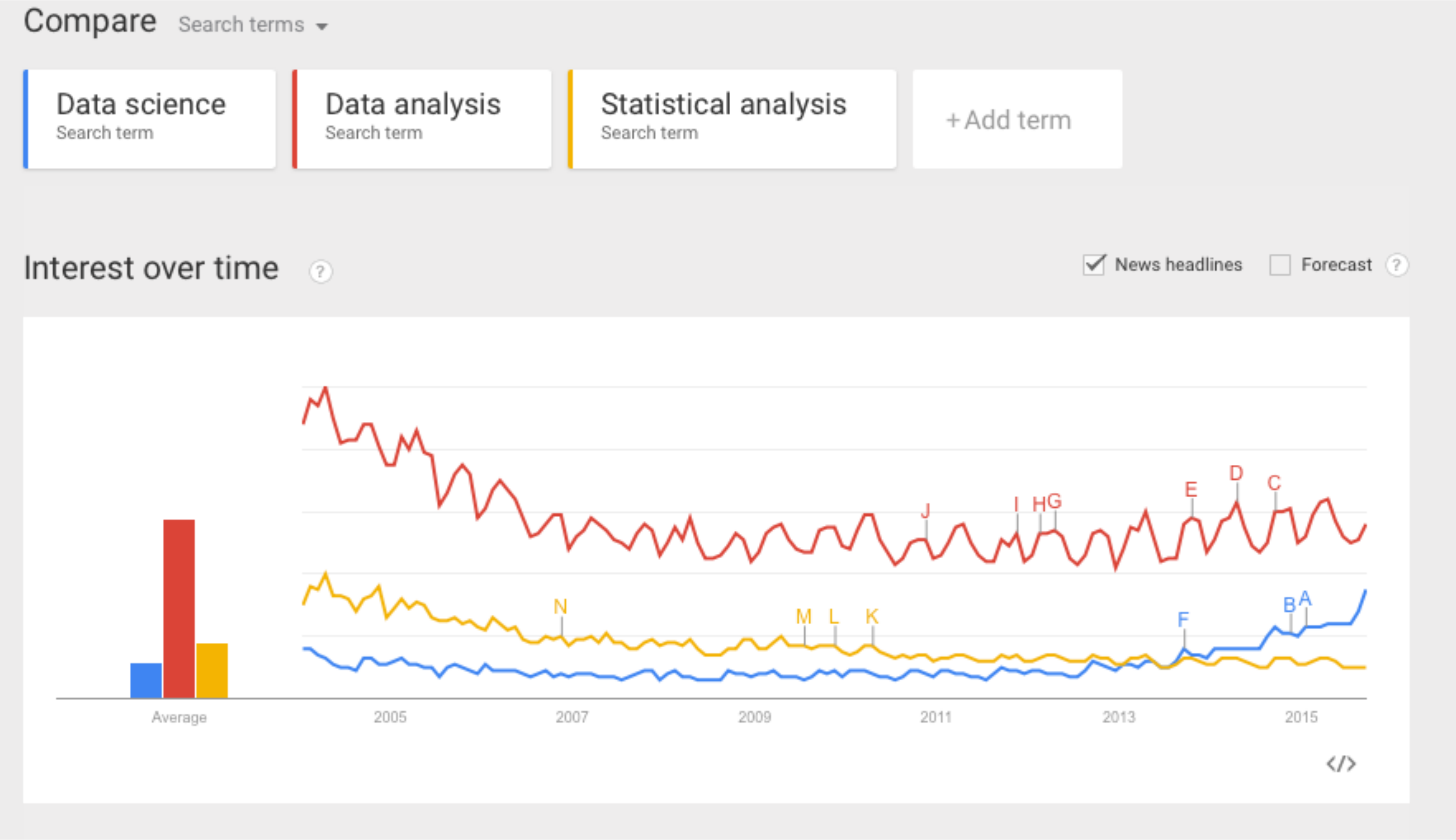
Data Science Initiatives

Search Metrics



PONTIFICIA UNIVERSIDAD CATÓLICA DE VALPARAÍSO

Search Metrics



Search Metrics

Defining data science

I really like the definition quoted above, of data science as *the intersection of software engineering and statistics*. Ofer Mendelevitch goes into more detail, drawing a continuum of professions that ranges from software engineer on the left to pure statistician (or machine learning researcher) on the right.

This continuum contains two additional roles, which are often confused with data scientists:

- *Data engineer:* a software engineer that deals with data plumbing (traditional database setup, Hadoop, Spark and all the rest)
- *Data analyst:* a person who digs into data to surface insights, but lacks the skills to do so at scale (e.g., they know how to use Excel, Tableau and SQL but can't build a web app from scratch)

Befuddment

THE WALL STREET JOURNAL. ≡ THE CIO REPORT

CIO Journal

Exclusive reporting and analysis for corporate-technology execut

CIO REPORT | CONSUMERIZATION | BIG DATA | CLOUD

11:46 am ET
May 2, 2014 BIG DATA

Why Do We Need Data Science When We've Had Statistics for Centuries?

ARTICLE | COMMENTS (13)

DATA SCIENCE DATA SCIENTIST

Email Print

By IRVING WLADAWSKY-BERGER

Data Science is emerging as one of the hottest new professions and academic disciplines in these early years of the 21st century. A number of articles have noted that the demand for data scientists is racing ahead of supply. People with the necessary skills are scarce, primarily because the discipline is so new. But, the situation is rapidly changing, as universities around the world have started to offer different kinds of graduate programs in data science. This year, for example, New York University is offering two new degrees–a general Master in Data Science, and a more domain-specific Master in Applied Urban Science and Informatics.

It's very exciting to contemplate the emergence of a major new discipline. It reminds me of the advent of computer science in the 1960s and 1970s. Like data science, computer science had its roots in a number of related areas, including math, engineering and management. In its early years, the field attracted people from a variety of other disciplines who started out using computers in their work or studies, and eventually switched to computer science from their original field.

IMS Bulletin online

HOME | LATEST ISSUE PDF | ARCHIVE (UNDER CONSTRUCTION) | ABOUT | ADVERTIS

HADLEY WICKHAM

Sep 4, 2014 Editor 18 Comments

Data science: how is it different to statistics?

Contributing Editor Hadley Wickham is Chief Scientist at RStudio and Adjunct Professor of Statistics at Rice University. He is interested in building better tools for data science. His work includes R packages for data analysis (ggplot2, plyr, reshape2); packages that make R less frustrating (lubridate for dates, stringr for strings, httr for accessing web APIs); and that make it easier to do good software development in R (roxygen2, testthat, devtools, lineprof, staticdocs). He is also a writer, educator, and frequent contributor to conferences promoting more accessible and more effective data analysis. He writes:

Befuddment

AMSTATNEWS
The Membership Magazine of the American Statistical Association

HOME | ABOUT | EDITORIAL CALENDAR | SUBMISSION INSTRUCTIONS | PDF ARCHIVES | ADVERTISE

Home » President's Corner

Aren't *We* Data Science?

1 JULY 2013   5,256 VIEWS   9 COMMENTS

Davidian

Last month, I shared this column with President-elect Nat Schenker and Past President Bob Rodriguez to announce an ASA strategic initiative to promote engagement of statisticians in Big Data. I'm following that announcement with an account of some of my recent experiences regarding data science, which inspire my enthusiasm for this effort. One in particular serves as a metaphor for the disconnect between statistics and data science we noted last month.

Around the time we were finalizing that column, Michelle Dunn, chair of the ASA Committee on Funded Research, forwarded an email to me. Michelle thought I would be interested in learning from the press release in the email that Eric Green would be speaking in Chapel Hill, North Carolina, 25 minutes from my office in Raleigh, on April 23. In January, the director of the National Institutes of Health (NIH), Francis Collins, announced the creation of a new NIH-wide position, the Associate Director for Data Science (ADDS), to "capitalize on the exponential growth of biomedical research data". Collins named Green, current director of the National Human Genome Research Institute, as acting ADDS. Green is also co-chair of the search committee charged with nominating the permanent ADDS.

Indeed, I was very interested! But what was even more interesting was the organization that had invited Green to speak. The press release announced "a new collaboration called the National Consortium for Data Science (NCDS) (aiming) to make North Carolina a national hub for data-intensive business and data science research." It went on to note that the NCDS had been launched at the Renaissance Computing Institute at The University of North Carolina at Chapel Hill (UNC-CH) and included among its founding members businesses, government organizations, and **major research universities**.

IMS Bulletin *online*

HOME | LATEST ISSUE PDF | ARCHIVE (UNDER CONSTRUCTION) | ABOUT | ADVERTISE

IMS NEWS | LECTURES AND ADDRESSES ,

Oct 1, 2014   Editor   3 Comments

IMS Presidential Address: Let us own Data Science

Each year the outgoing IMS President delivers an address at the IMS Annual Meeting, which, this year, was the Australian Statistical Conference in Sydney (July 9-14, 2014), a joint meeting of the Statistical Society of Australia Inc. (SSAI) and IMS. Bin Yu, Chancellor's Professor of Statistics and EECS, University of California at Berkeley, gave her Presidential Address, on which the following article is based:

Let us own data science

It is my honor and pleasure to deliver this IMS Presidential Address at the joint meeting of the Statistical Society of Australia Inc. (SSAI) and IMS.

Befuddment

50 years of Data Science

David Donoho

Sept. 18, 2015
Version 1.00

Abstract

More than 50 years ago, John Tukey called for a reformation of academic statistics. In 'The Future of Data Analysis', he pointed to the existence of an as-yet unrecognized *science*, whose subject of interest was learning from data, or 'data analysis'. Ten to twenty years ago, John Chambers, Bill Cleveland and Leo Breiman independently once again urged academic statistics to expand its boundaries beyond the classical domain of theoretical statistics; Chambers called for more emphasis on data preparation and presentation rather than statistical modeling; and Breiman called for emphasis on prediction rather than inference. Cleveland even suggested the catchy name "Data Science" for his envisioned field.

A recent and growing phenomenon is the emergence of "Data Science" programs at major universities, including UC Berkeley, NYU, MIT, and most recently the Univ. of Michigan, which on September 8, 2015 announced a $100M "Data Science Initiative" that will hire 35 new faculty. Teaching in these new programs has significant overlap in curricular subject matter with traditional statistics courses; in general, though, the new initiatives steer away from close involvement with academic statistics departments.

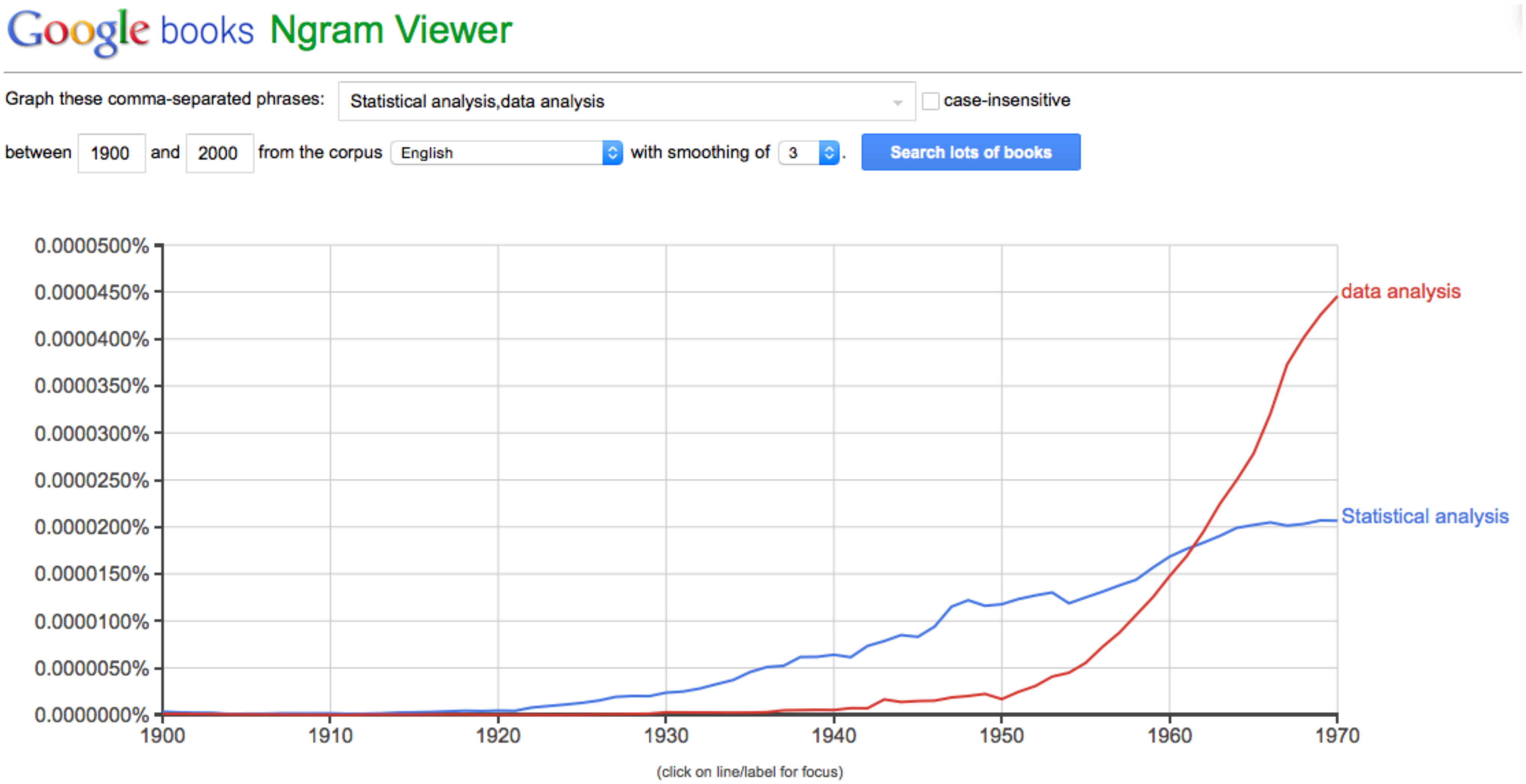
Tukey's Paper

THE FUTURE OF DATA ANALYSIS[1]

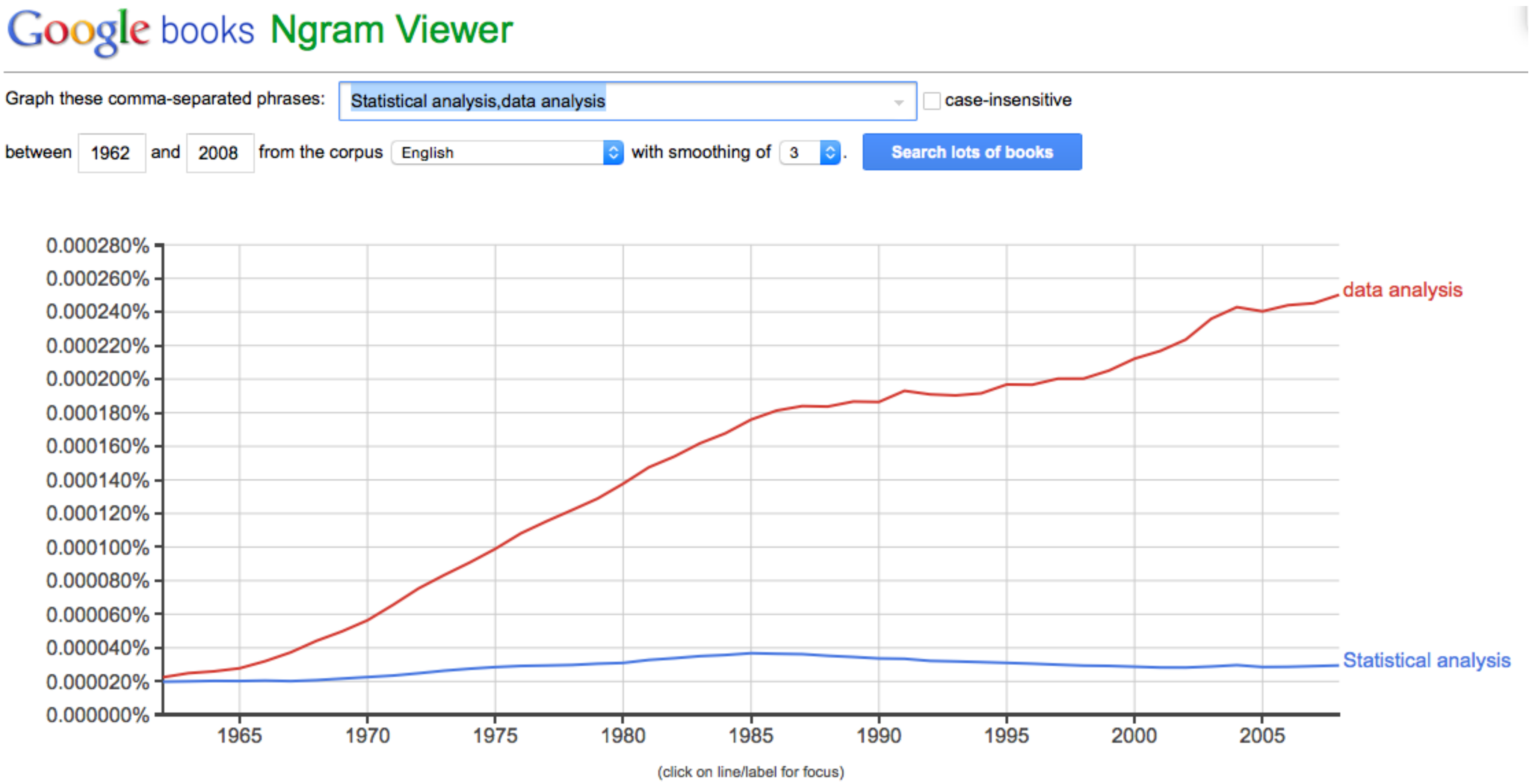
By John W. Tukey

Princeton University and Bell Telephone Laboratories

Tukey's Paper



Tukey's Paper



Chambers Paper

John Chambers 1992

Greater or Lesser Statistics: A Choice for Future Research

John M. Chambers
AT&T Bell Laboratories, Murray Hill, New Jersey

Abstract

The statistics profession faces a choice in its future research between continuing concentration on traditional topics, based largely on data analysis supported by mathematical statistics, and a broader viewpoint, based on an inclusive concept of learning from data. The latter course presents severe challenges as well as exciting opportunities. The former risks seeing statistics become increasingly marginal in activities to which it can make important contributions.

This paper is one of a set of short position papers on future research directions in statistics invited by the editor of *Statistics and Computation*.

Cleveland Paper

Bill Cleveland 2002

Data Science: An Action Plan for Expanding the Technical Areas of the Field of Statistics

William S. Cleveland
Statistics Research, Bell Labs
wsc@bell-labs.com

Abstract

An action plan to enlarge the technical areas of statistics focuses on the data analyst. The plan sets out six technical areas of work for a university department and advocates a specific allocation of resources devoted to research in each area and to courses in each area. The value of technical work is judged by the extent to which it benefits the data analyst, either directly or indirectly. The plan is also applicable to government research labs and corporate research organizations.

Emergence of Meta Analysis

Biostatistics (2014), **15**, 1, *pp.* 1–12
doi:10.1093/biostatistics/kxt007
Advance Access publication on September 25, 2013

R

An estimate of the science-wise false discovery rate and application to the top medical literature

LEAH R. JAGER

Department of Mathematics, United States Naval Academy, Annapolis, MD 21402, USA

JEFFREY T. LEEK*

Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD 21205, USA

jleek@jhsph.edu

Emergence of Meta Analysis

Meta Analysis
Systemic Failure
Causes
Solutions

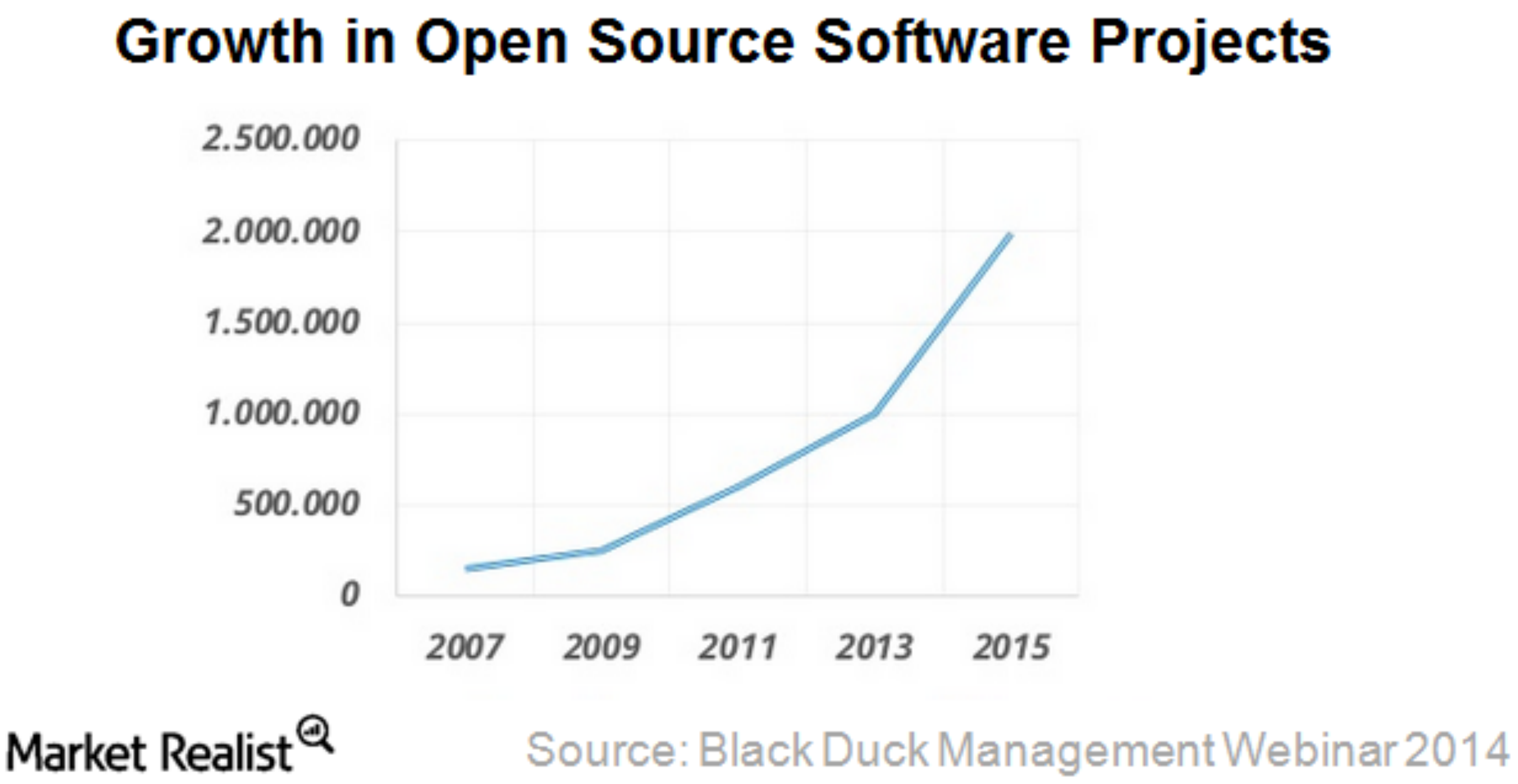
How Much of Our Published Research Can We Believe?

Systemic Failures, Their Causes, a Solution

David Donoho

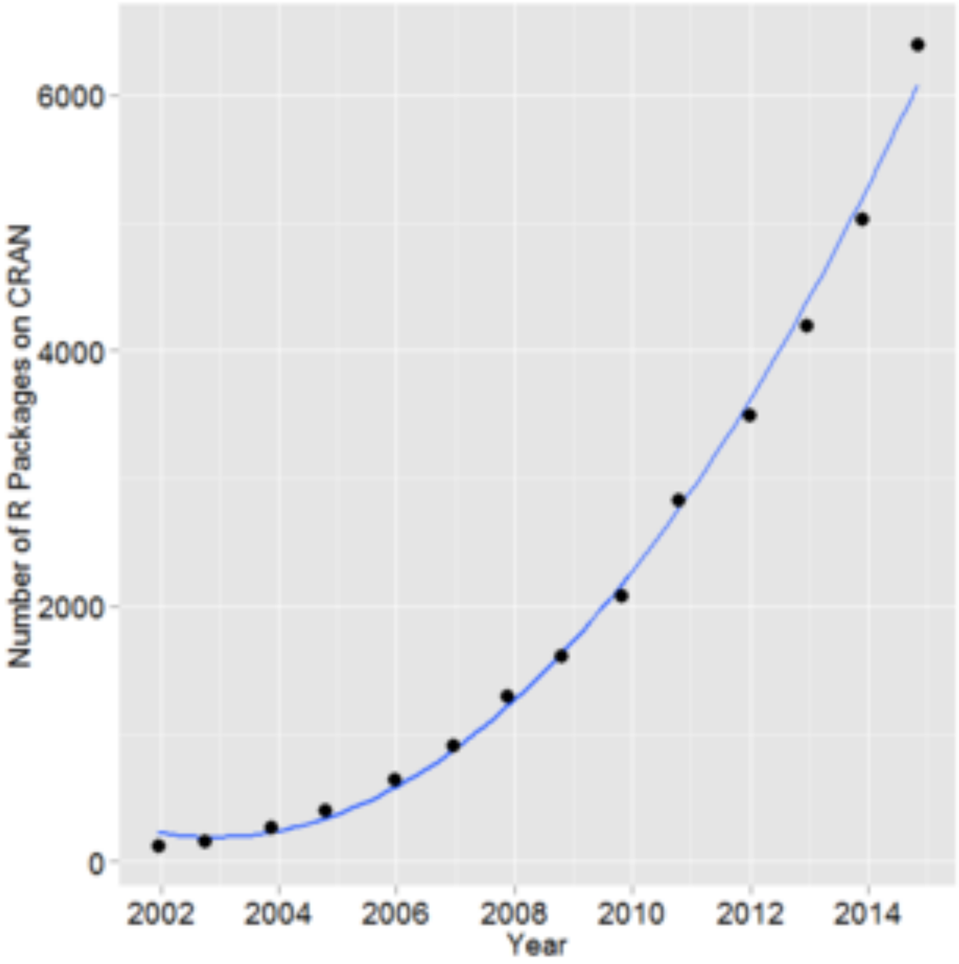
20140530

Emergence of Reproducible Data Analyses

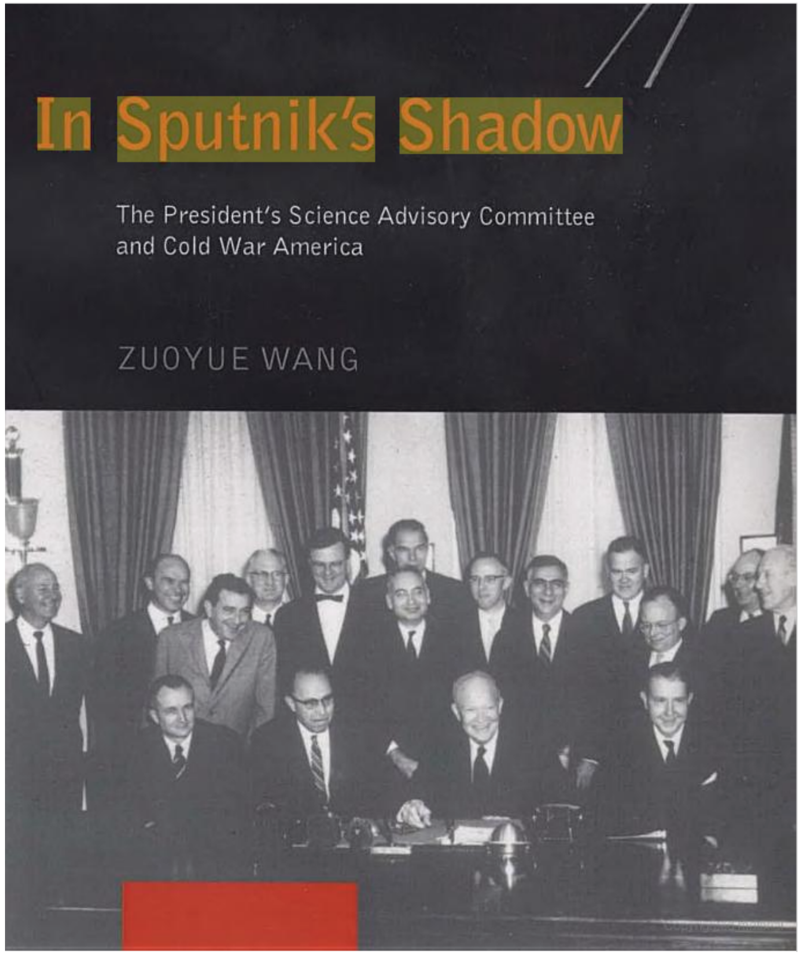


PONTIFICIA UNIVERSIDAD CATÓLICA DE VALPARAÍSO

Emergence of Reproducible Data Analyses



Crisis in Machine Translation mid 1960's



Crisis in Machine Translation mid 1960's

"We are safe in asserting that speech recognition is attractive to money. The attraction is perhaps similar to the attraction of schemes for turning water into gasoline, extracting gold from the sea, curing cancer, or going to the moon. One doesn't attract thoughtlessly given dollars by means of schemes for cutting the cost of soap by 10%. **To sell suckers, one uses deceit and offers glamor.**"

"It is clear that glamor and any deceit in the field of speech recognition blind the takers of funds as much as they blind the givers of funds. Thus, we may pity workers whom we cannot respect."

JR Pierce, 1969

Crisis in Machine Translation mid 1960's

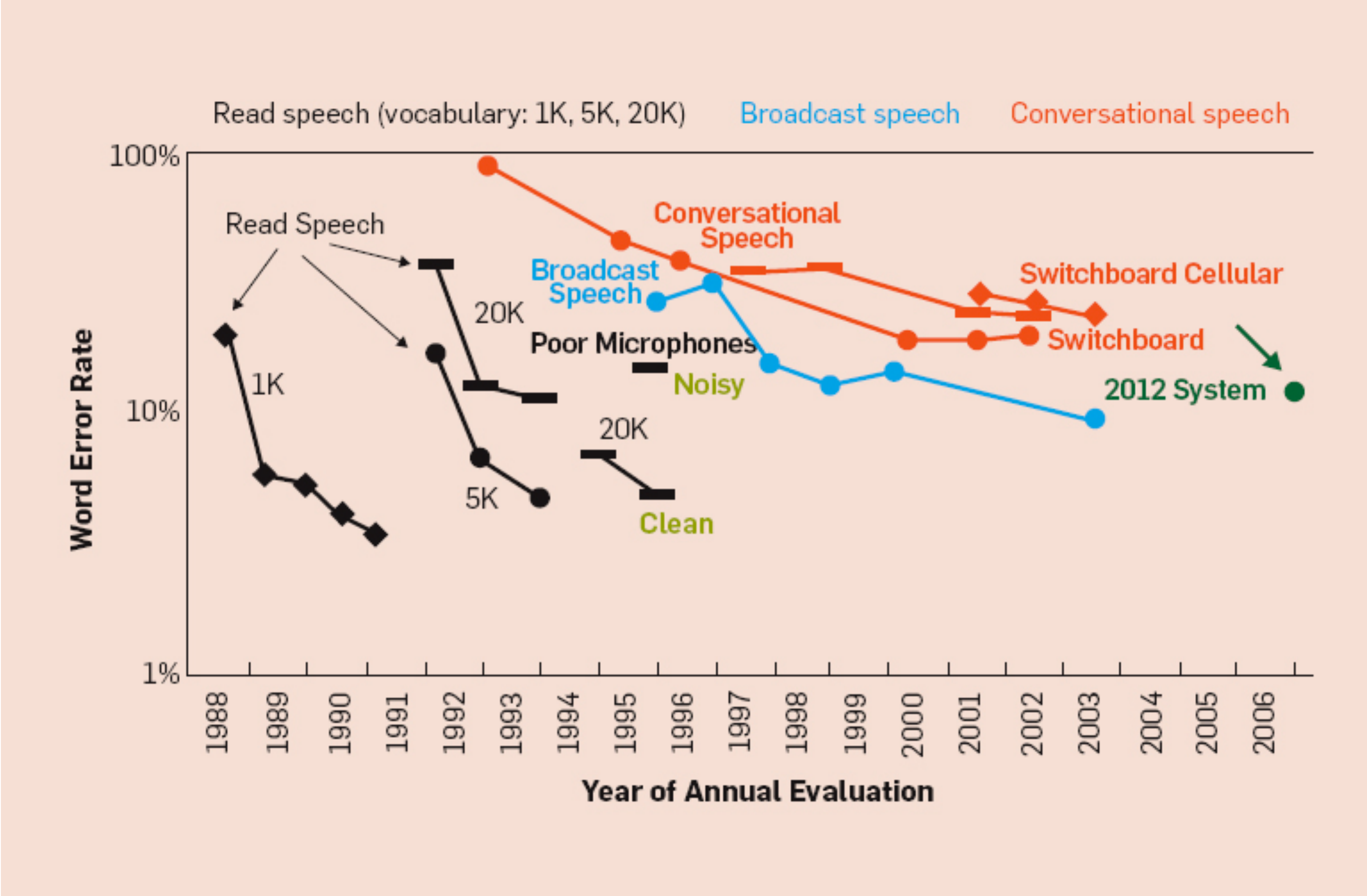
"Most recognizers behave, not like scientists, but like mad inventors or untrustworthy engineers. The typical recognizer gets it into his head that he can solve 'the problem.' The basis for this is either individual inspiration (the 'mad inventor' source of knowledge) or acceptance of untested rules, schemes, or information (the untrustworthy engineer approach). . . ."

"The typical recognizer ... builds or programs an elaborate system that either does very little or flops in an obscure way. A lot of money and time are spent. **No simple, clear, sure knowledge is gained.** The work has been an experience, not an experiment."

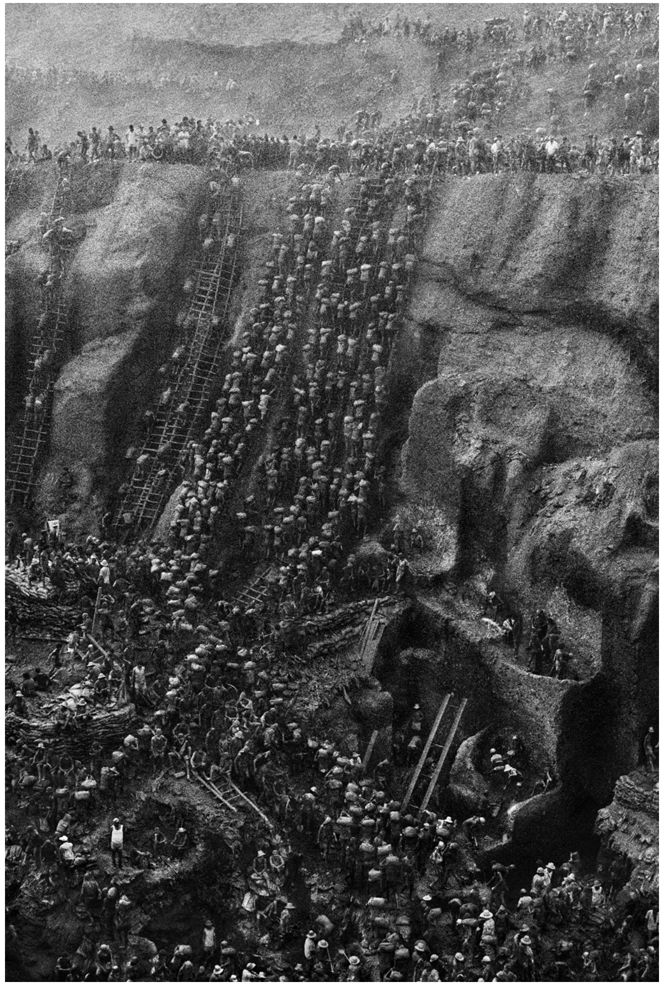
JR Pierce, 1969

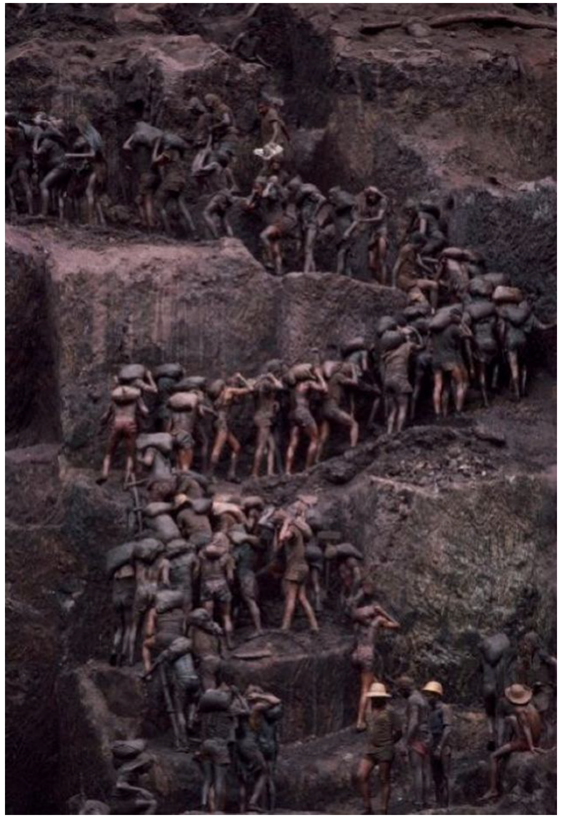
PONTIFICIA UNIVERSIDAD CATÓLICA DE VALPARAÍSO

Common Task Framework



Common Task Framework



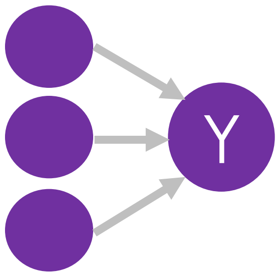


Sebastiao Salgado, Work

Predictive Modeling fundamentals

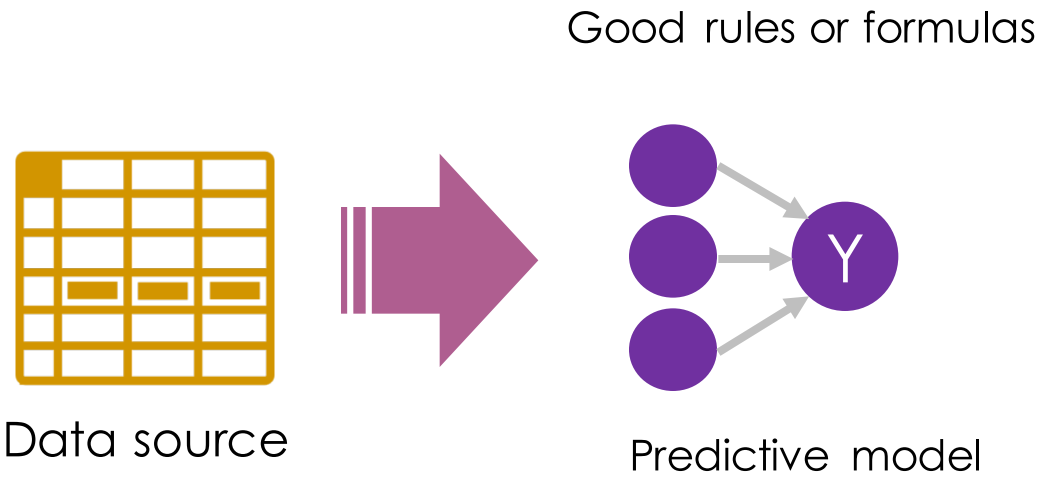
PONTIFICIA UNIVERSIDAD CATÓLICA DE VALPARAÍSO

Predictive modeling fundamentals

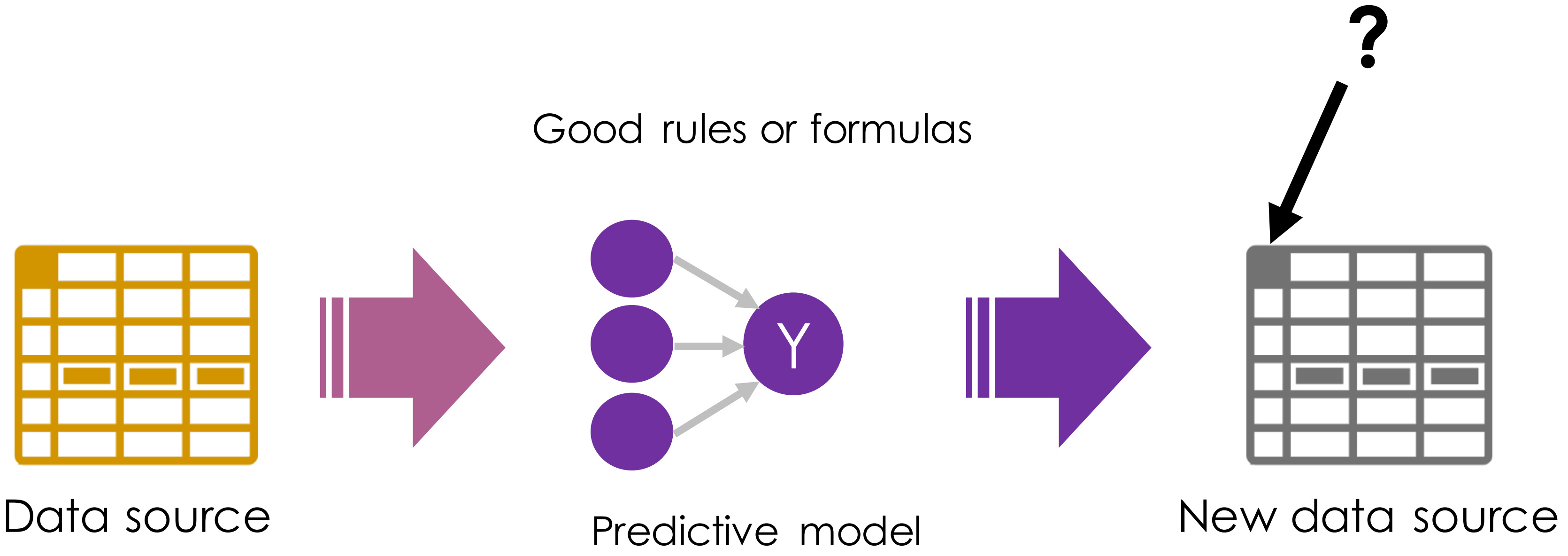


Predictive model

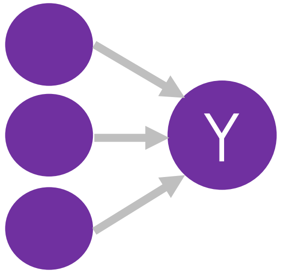
Predictive modeling fundamentals



Predictive modeling fundamentals



Predictive modeling fundamentals



Predictive model

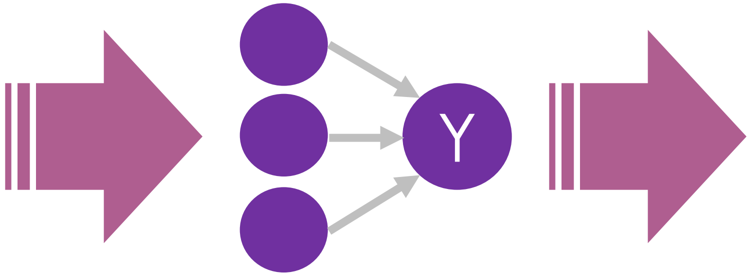
Objectives

- Fundamentals
- Business scenario data
- Modeling challenges and solutions

Predictive modeling fundamentals

Applications

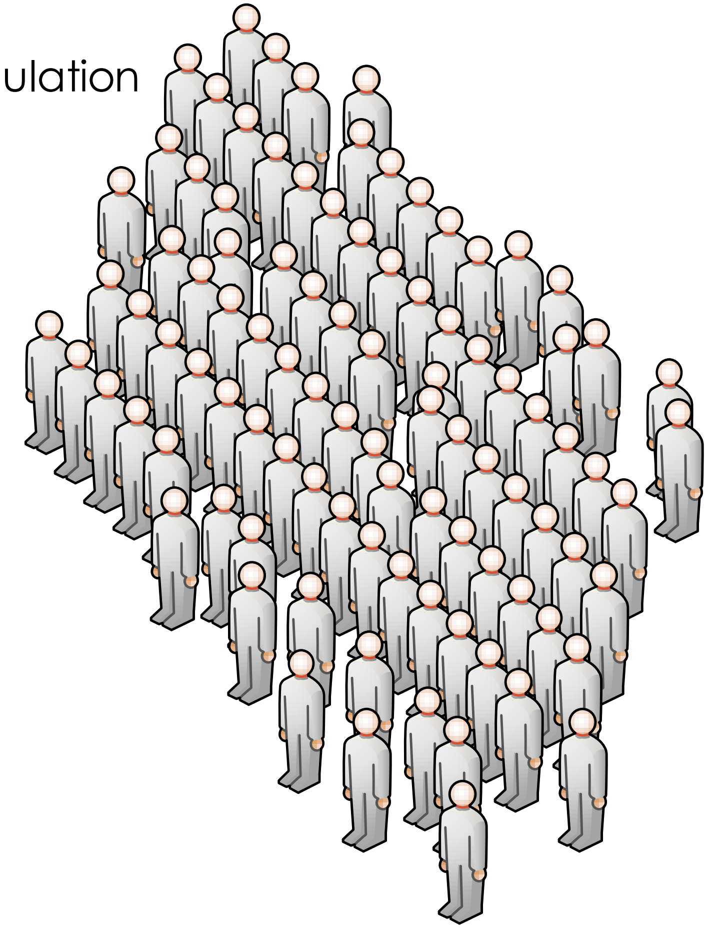
	y	X_1	X_2	$\cdots$	X_k
1	0	■	■	$\cdots$	■
2	1	■	■	$\cdots$	■
3	0	■	■	$\cdots$	■
$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$
n	1	■	■	$\cdots$	■



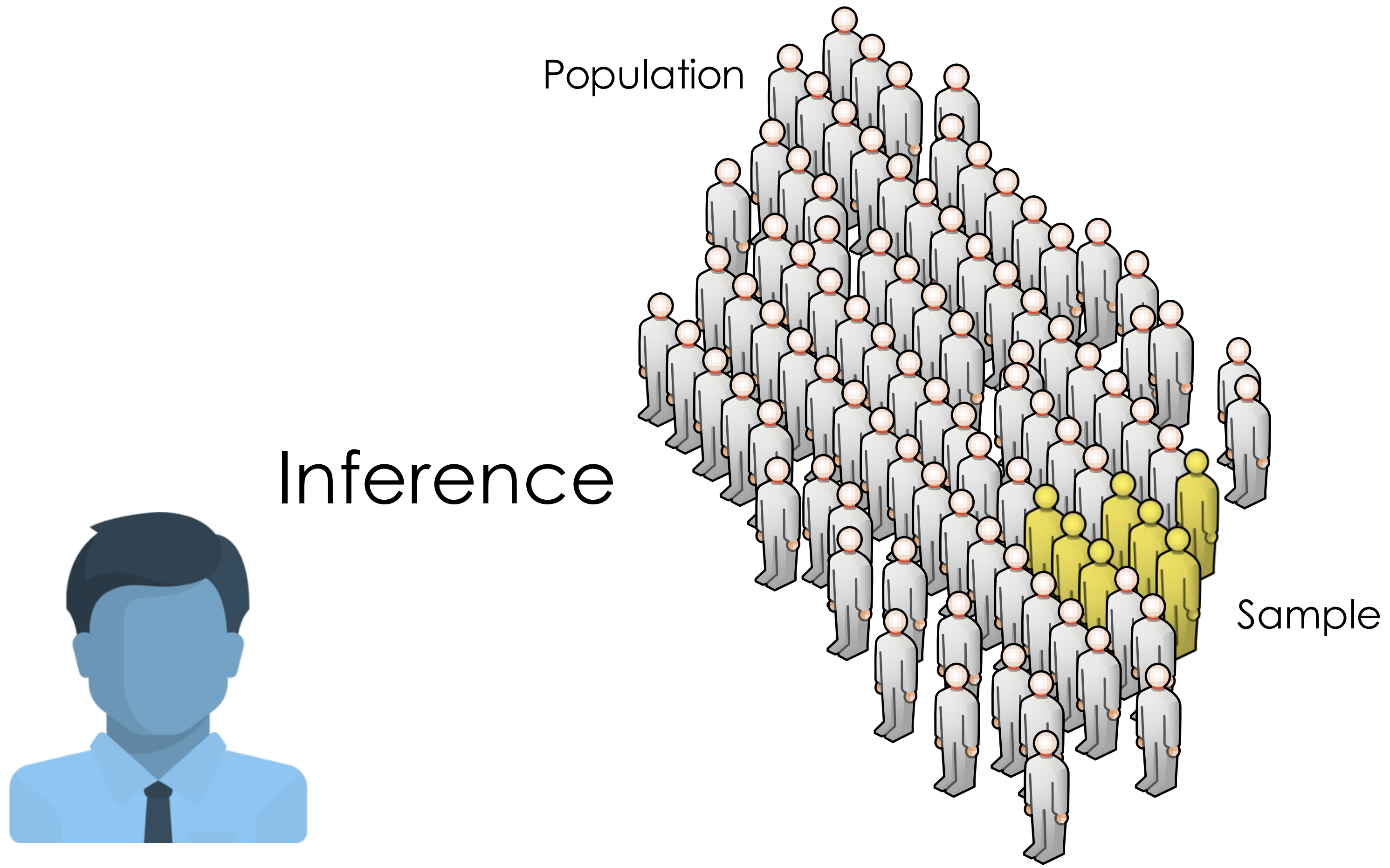
	y	X_1	X_2	$\cdots$	X_k
1	?	■	■	$\cdots$	■
2	?	■	■	$\cdots$	■
3	?	■	■	$\cdots$	■
$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$
n	?	■	■	$\cdots$	■

Terms and elements of predictive modeling

Population



Terms and elements of predictive modeling

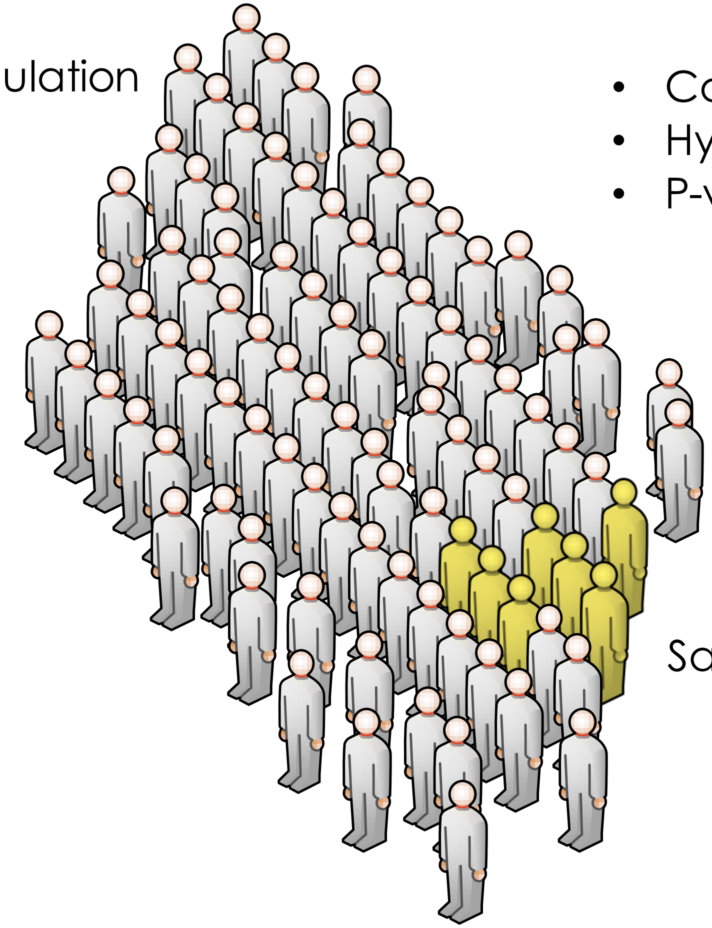


Terms and elements of predictive modeling

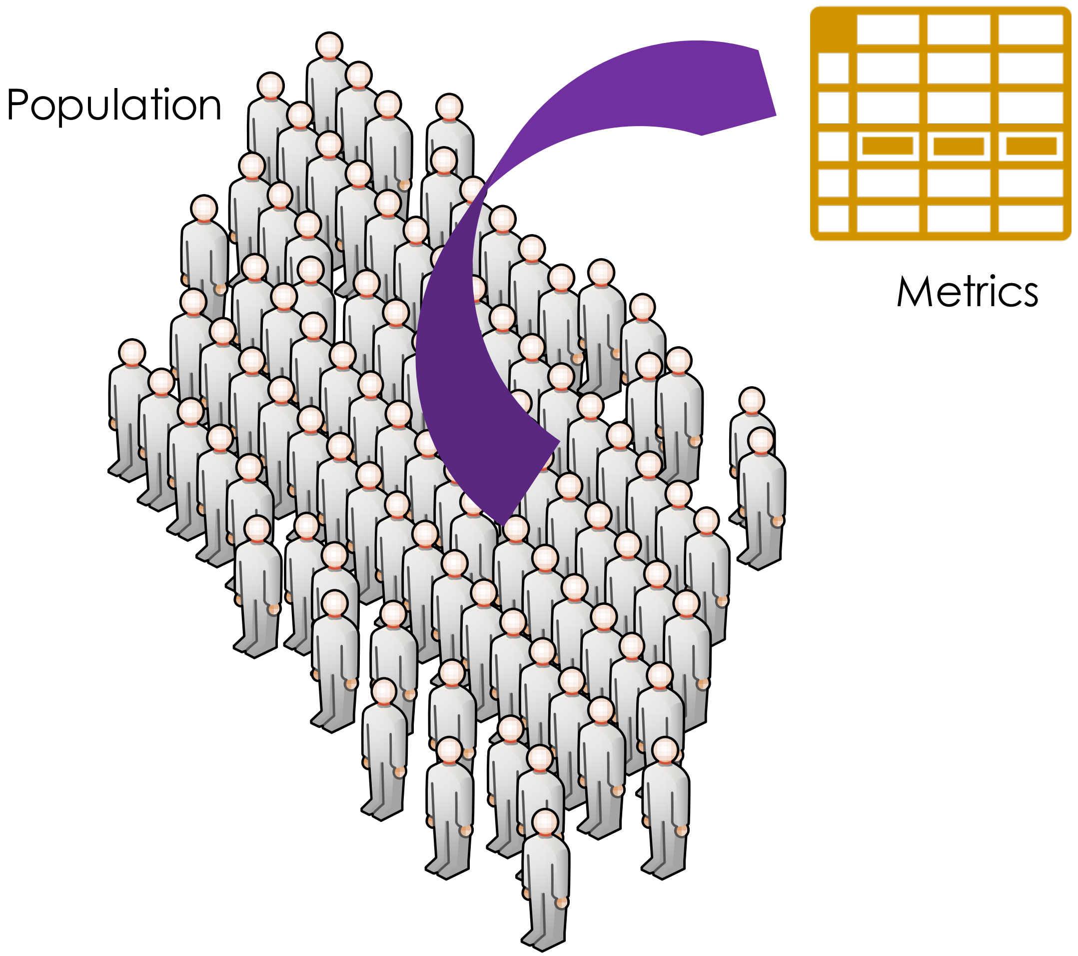
Common tools of inference

- Confidence intervals
- Hypothesis test
- P-values

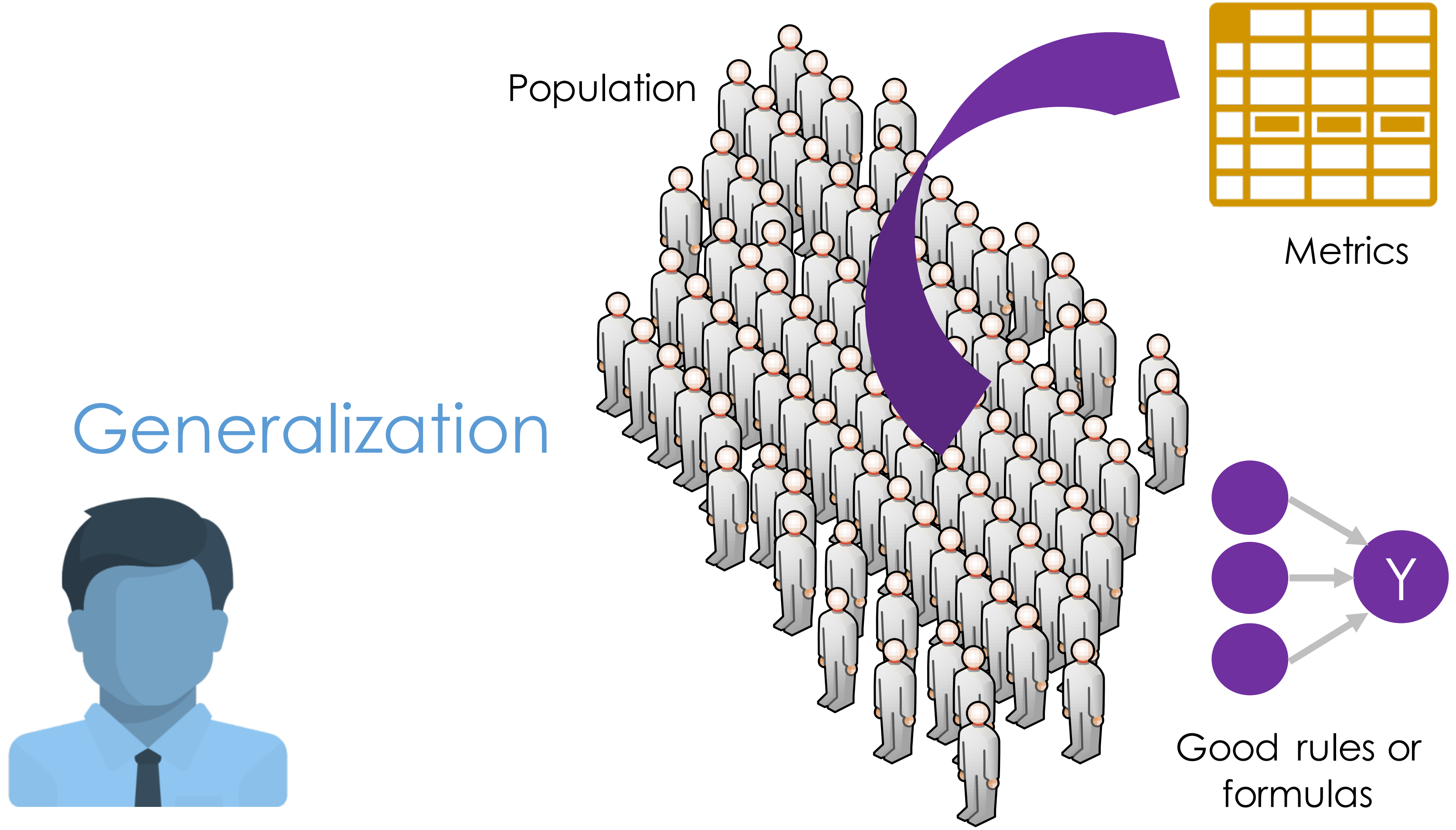
Inference



Terms and elements of predictive modeling

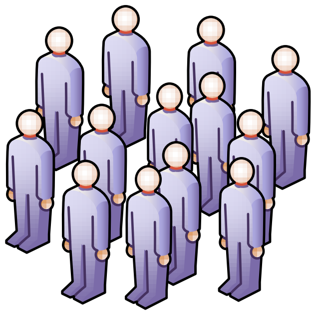


Terms and elements of predictive modeling



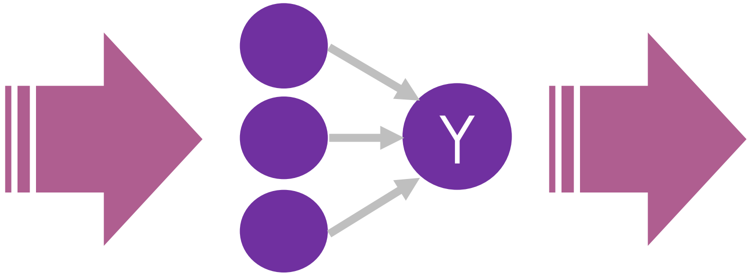
PONTIFICIA UNIVERSIDAD CATÓLICA DE VALPARAÍSO

Terms and elements of predictive modeling



New population

	y	X_1	X_2	$\cdots$	X_k
1	0	■	■	$\cdots$	■
2	1	■	■	$\cdots$	■
3	0	■	■	$\cdots$	■
$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$
n	1	■	■	$\cdots$	■



	y	X_1	X_2	$\cdots$	X_k
1	?	■	■	$\cdots$	■
2	?	■	■	$\cdots$	■
3	?	■	■	$\cdots$	■
$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$
n	?	■	■	$\cdots$	■

Emphasis

- Empirical quality of predictions
- Understanding relationships

Predictive modeling fundamentals

Target variable

Outcome
Response

Input variables

Predictors
Explanatory variables
Inputs
Features

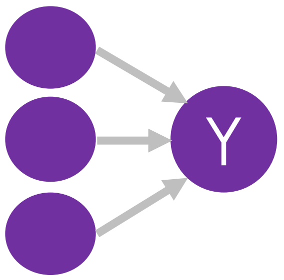
Cases

$$\begin{array}{cccccc} & y & X_1 & X_2 & \cdots & X_k \\ 1 & 0 & \blacksquare & \blacksquare & \cdots & \blacksquare \\ 2 & 1 & \blacksquare & \blacksquare & \cdots & \blacksquare \\ 3 & 0 & \blacksquare & \blacksquare & \cdots & \blacksquare \\ \vdots & \vdots & \vdots & \vdots & & \vdots \\ n & 1 & \blacksquare & \blacksquare & \cdots & \blacksquare \end{array}$$

Observation examples

Basic steps of predictive modeling

Build a model on historic data



$$
\begin{array}{cccccc}
 & y & X_1 & X_2 & \cdots & X_k \\
1 & 0 & \blacksquare & \blacksquare & \cdots & \blacksquare \\
2 & 1 & \blacksquare & \blacksquare & \cdots & \blacksquare \\
3 & 0 & \blacksquare & \blacksquare & \cdots & \blacksquare \\
\vdots & \vdots & \vdots & \vdots & & \vdots \\
n & 1 & \blacksquare & \blacksquare & \cdots & \blacksquare
\end{array}
$$

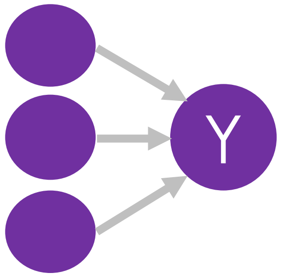
Basic steps of predictive modeling

Build a model on historic data

Supervised classification

The goal is to correctly classify

Known discrete variable



	y	X_1	X_2	$\cdots$	X_k
1	0	■	■	$\cdots$	■
2	1	■	■	$\cdots$	■
3	0	■	■	$\cdots$	■
$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$
n	1	■	■	$\cdots$	■

Basic steps of predictive modeling

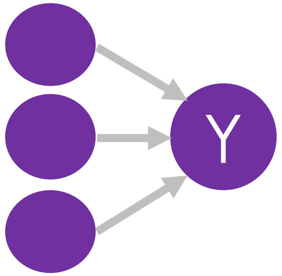
Build a model on historic data

Supervised classification

The goal is to correctly classify

Target variable Y
(binary response)

1 = Response
0 = No response



Known discrete variable

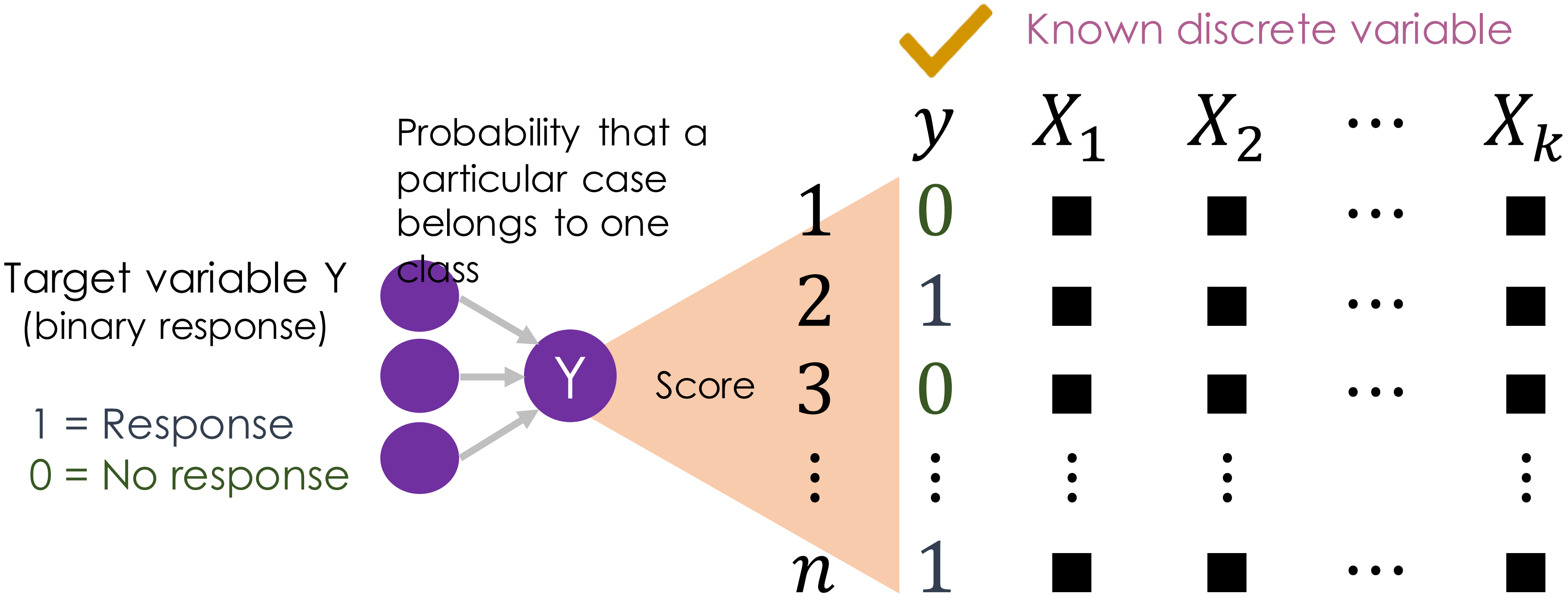
	y	X_1	X_2	$\cdots$	X_k
1	0	■	■	$\cdots$	■
2	1	■	■	$\cdots$	■
3	0	■	■	$\cdots$	■
$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$
n	1	■	■	$\cdots$	■

Basic steps of predictive modeling

Build a model on historic data

Supervised classification

The goal is to correctly classify



Basic steps of predictive modeling

Build a model on historic data

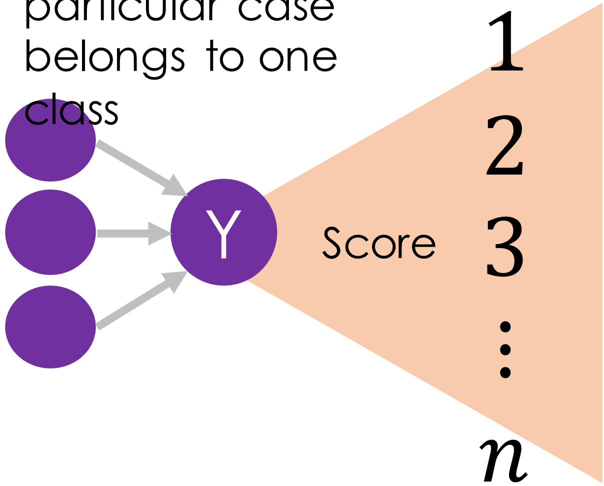
Supervised classification

The goal is to correctly classify

Known discrete variable

Why use a predictive model?

Probability that a particular case belongs to one class

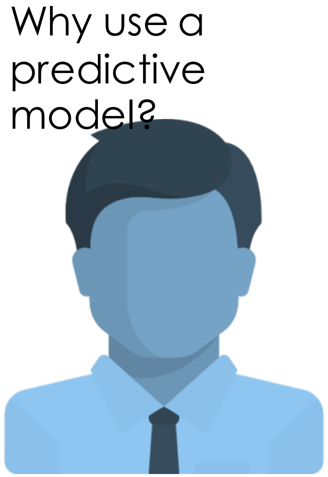


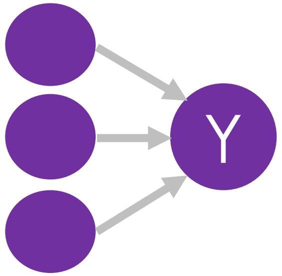
	y	X_1	X_2	$\cdots$	X_k
1	0	■	■	$\cdots$	■
2	1	■	■	$\cdots$	■
3	0	■	■	$\cdots$	■
$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$
n	1	■	■	$\cdots$	■

PONTIFICIA UNIVERSIDAD CATÓLICA DE VALPARAÍSO

Basic steps of predictive modeling

New data arrives





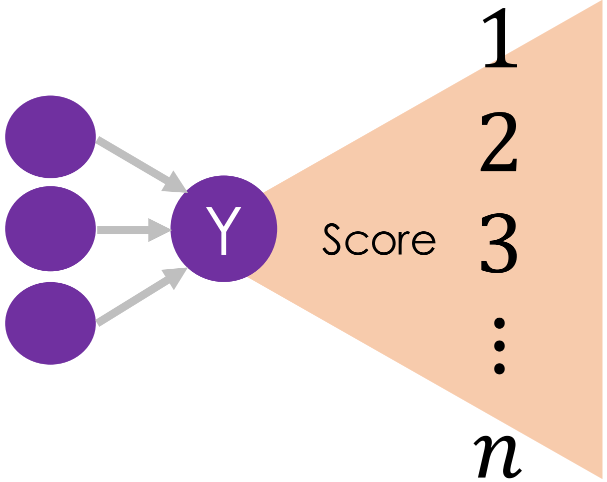
Unknown discrete variable

	y	X_1	X_2	$\cdots$	X_k
1	?	■	■	$\cdots$	■
2	?	■	■	$\cdots$	■
3	?	■	■	$\cdots$	■
$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$
n	?	■	■	$\cdots$	■

Basic steps of predictive modeling

New data arrives

Why use a predictive model?



Unknown discrete variable

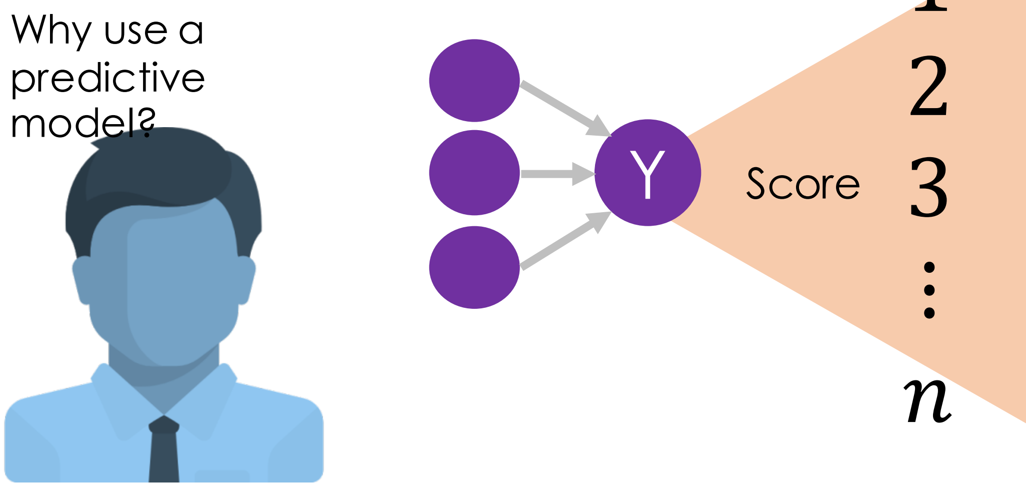
	y	X_1	X_2	$\cdots$	X_k
1	?	■	■	$\cdots$	■
2	?	■	■	$\cdots$	■
3	?	■	■	$\cdots$	■
$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$
n	?	■	■	$\cdots$	■

PONTIFICIA UNIVERSIDAD CATÓLICA DE VALPARAÍSO

Basic steps of predictive modeling

New data arrives

Generalize



Predicted discrete variable

	y	X_1	X_2	$\cdots$	X_k
1	1	■	■	$\cdots$	■
2	0	■	■	$\cdots$	■
3	1	■	■	$\cdots$	■
$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$
n	0	■	■	$\cdots$	■

Basic steps of predictive modeling

Input variables

$$
\begin{array}{cccccc}
 & y & X_1 & X_2 & \cdots & X_k \\
1 & 0 & \blacksquare & \blacksquare & \cdots & \blacksquare \\
2 & 1 & \blacksquare & \blacksquare & \cdots & \blacksquare \\
3 & 0 & \blacksquare & \blacksquare & \cdots & \blacksquare \\
\vdots & \vdots & \vdots & \vdots & & \vdots \\
n & 1 & \blacksquare & \blacksquare & \cdots & \blacksquare
\end{array}
$$

Basic steps

1, Supervised classification

- Prepare the inputs
- Select the most predictive inputs and fit models

2. Generalization

- Assess the models

Basic steps of predictive modeling

Most predictive Input variables

$$
\begin{array}{cccccc}
 & y & X_1 & X_2 & \cdots & X_k \\
1 & 0 & \blacksquare & \blacksquare & \cdots & \blacksquare \\
2 & 1 & \blacksquare & \blacksquare & \cdots & \blacksquare \\
3 & 0 & \blacksquare & \blacksquare & \cdots & \blacksquare \\
\vdots & \vdots & \vdots & \vdots & & \vdots \\
n & 1 & \blacksquare & \blacksquare & \cdots & \blacksquare
\end{array}
$$

Basic steps

1, Supervised classification

- Prepare the inputs
- Select the most predictive inputs and fit models

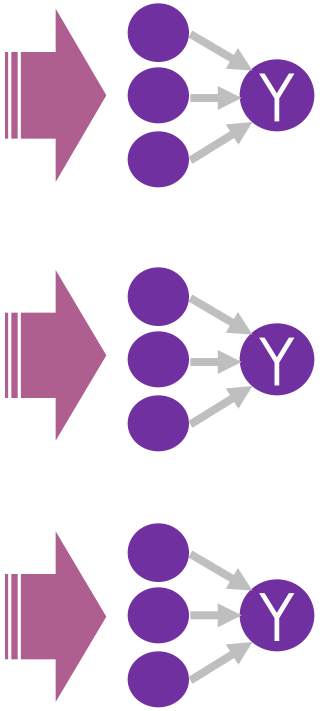
2. Generalization

- Assess the models

Basic steps of predictive modeling

Most predictive Input variables

	y	X_1	X_2	$\cdots$	X_k
1	0	■	■	$\cdots$	■
2	1	■	■	$\cdots$	■
3	0	■	■	$\cdots$	■
$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$
n	1	■	■	$\cdots$	■



Basic steps

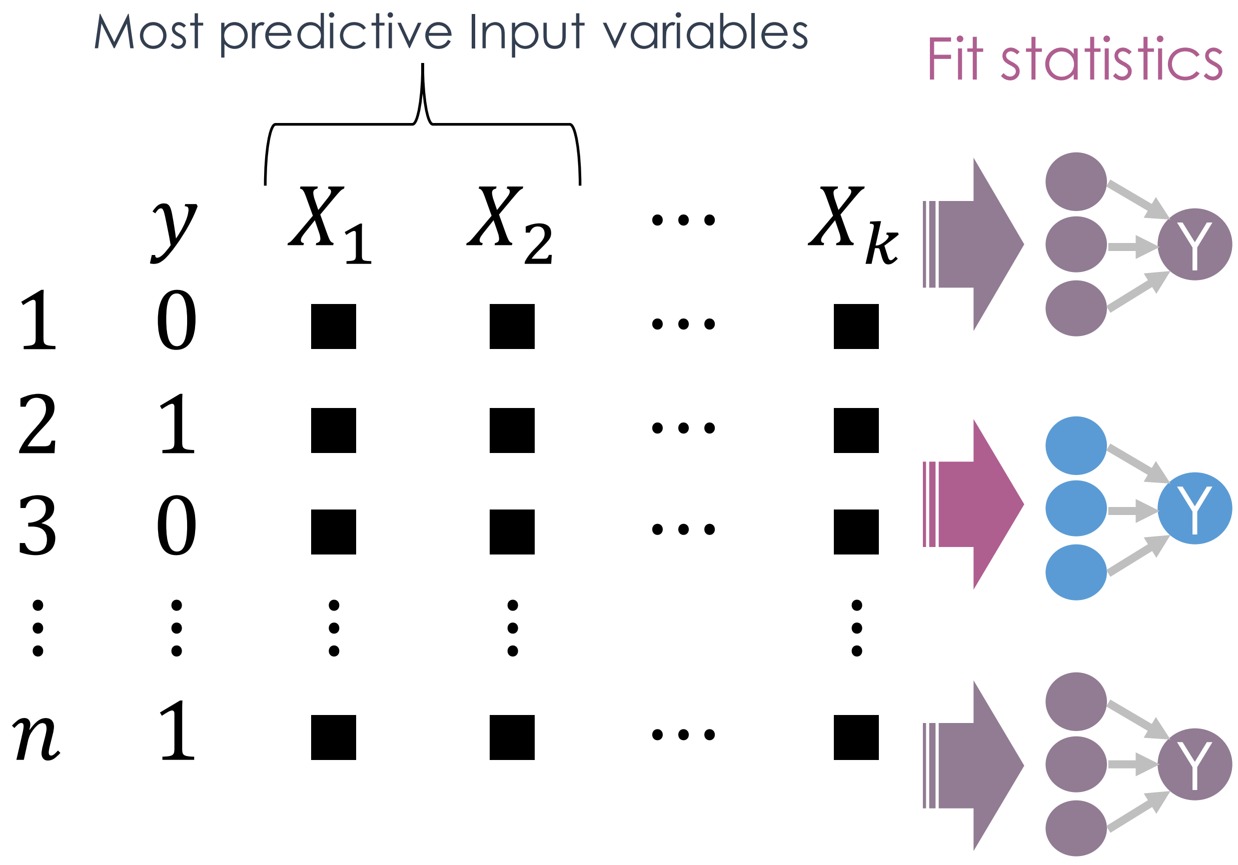
1, Supervised classification

- Prepare the inputs
- Select the most predictive inputs and fit models

2. Generalization

- Assess the models

Basic steps of predictive modeling



Basic steps

1, Supervised classification

- Prepare the inputs
- Select the most predictive inputs and fit models

2. Generalization

- Assess the models

PONTIFICIA UNIVERSIDAD CATÓLICA DE VALPARAÍSO

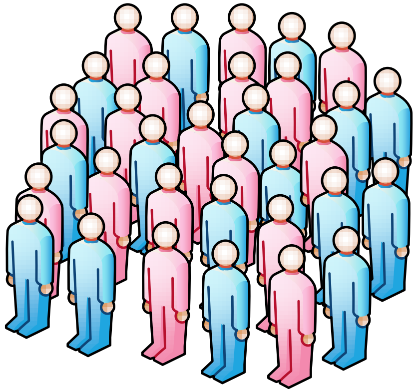
Applications of predictive modeling

Target marketing	Attrition predicción
Credit scoring	Fraud detection

Applications of predictive modeling

Target marketing

Customers

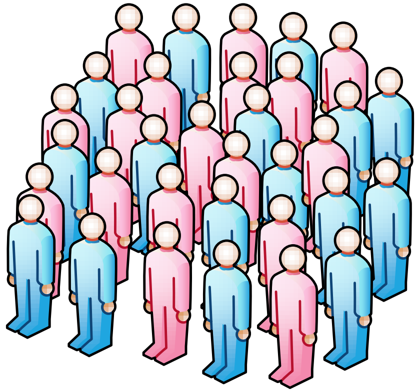


$$\text{Cases}\left\{\begin{array}{cccccc} & y & X_1 & X_2 & \cdots & X_k \\ 1 & 0 & \blacksquare & \blacksquare & \cdots & \blacksquare \\ 2 & 1 & \blacksquare & \blacksquare & \cdots & \blacksquare \\ 3 & 0 & \blacksquare & \blacksquare & \cdots & \blacksquare \\ \vdots & \vdots & \vdots & \vdots & & \vdots \\ n & 1 & \blacksquare & \blacksquare & \cdots & \blacksquare \end{array}\right.$$

Applications of predictive modeling

Target marketing

Customers



Input variables

Cases

	y	X_1	X_2	$\cdots$	X_k
1	0	$\blacksquare$	$\blacksquare$	$\cdots$	$\blacksquare$
2	1	$\blacksquare$	$\blacksquare$	$\cdots$	$\blacksquare$
3	0	$\blacksquare$	$\blacksquare$	$\cdots$	$\blacksquare$
$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$
n	1	$\blacksquare$	$\blacksquare$	$\cdots$	$\blacksquare$

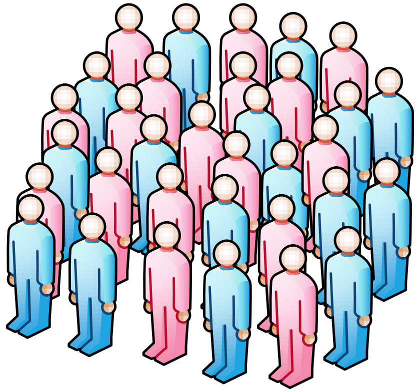
Applications of predictive modeling

Target marketing

Target variable

Input variables

Customers



Cases

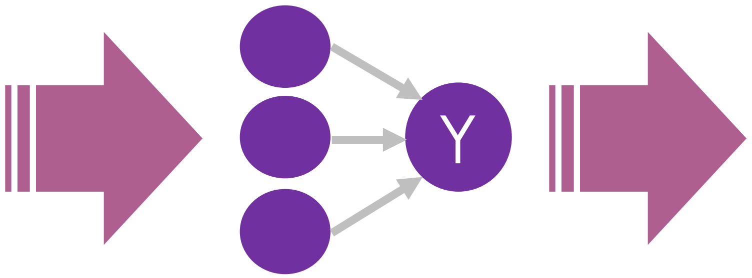
	y	X_1	X_2	$\cdots$	X_k
1	0	■	■	$\cdots$	■
2	1	■	■	$\cdots$	■
3	0	■	■	$\cdots$	■
$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$
n	1	■	■	$\cdots$	■

Target new potential customers

Applications of predictive modeling

Target marketing

	y	X_1	X_2	$\cdots$	X_k
1	0	■	■	$\cdots$	■
2	1	■	■	$\cdots$	■
3	0	■	■	$\cdots$	■
$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$
n	1	■	■	$\cdots$	■

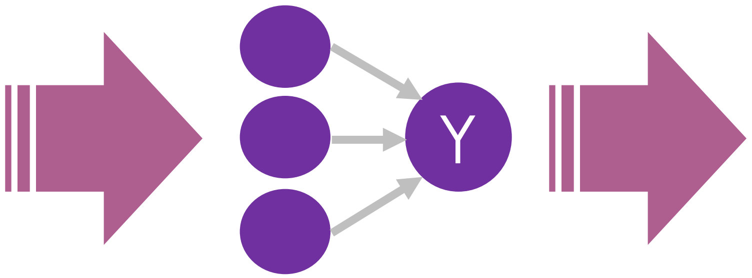


	y	X_1	X_2	$\cdots$	X_k
1	?	■	■	$\cdots$	■
2	?	■	■	$\cdots$	■
3	?	■	■	$\cdots$	■
$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$
n	?	■	■	$\cdots$	■

Applications of predictive modeling

Target marketing

	y	X_1	X_2	$\cdots$	X_k
1	0	■	■	$\cdots$	■
2	1	■	■	$\cdots$	■
3	0	■	■	$\cdots$	■
$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$
n	1	■	■	$\cdots$	■

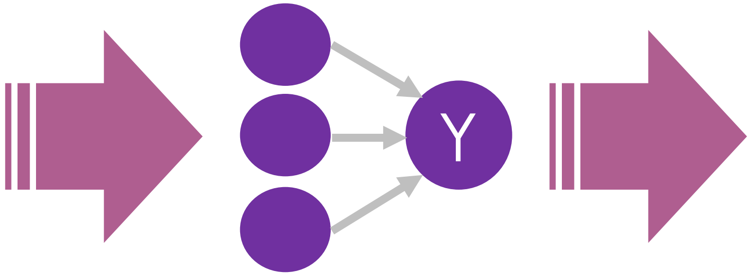


	y	X_1	X_2	$\cdots$	X_k
1	?	■	■	$\cdots$	■
2	?	■	■	$\cdots$	■
3	?	■	■	$\cdots$	■
$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$
n	?	■	■	$\cdots$	■

Applications of predictive modeling

Attrition
predicción

	y	X_1	X_2	$\cdots$	X_k
1	0	■	■	$\cdots$	■
2	1	■	■	$\cdots$	■
3	0	■	■	$\cdots$	■
$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$
n	1	■	■	$\cdots$	■

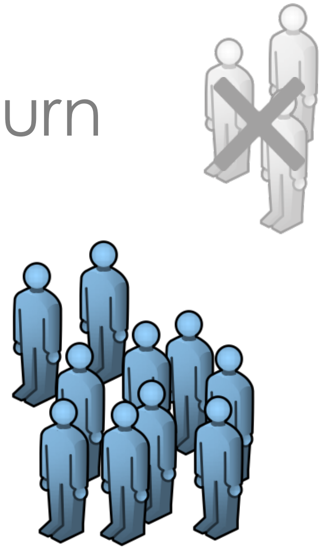


	y	X_1	X_2	$\cdots$	X_k
1	?	■	■	$\cdots$	■
2	?	■	■	$\cdots$	■
3	?	■	■	$\cdots$	■
$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$
n	?	■	■	$\cdots$	■

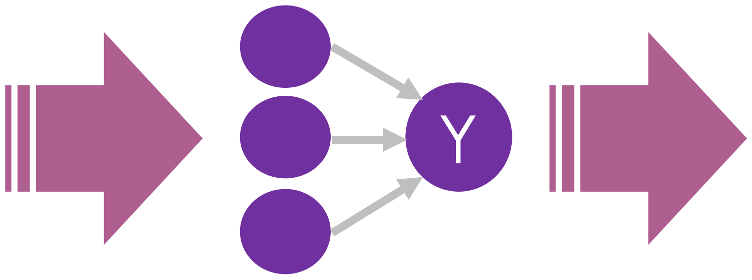
Applications of predictive modeling

Attrition predicción

Churn



	y	X_1	X_2	$\cdots$	X_k
1	0	■	■	$\cdots$	■
2	1	■	■	$\cdots$	■
3	0	■	■	$\cdots$	■
$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$
n	1	■	■	$\cdots$	■



	y	X_1	X_2	$\cdots$	X_k
1	?	■	■	$\cdots$	■
2	?	■	■	$\cdots$	■
3	?	■	■	$\cdots$	■
$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$
n	?	■	■	$\cdots$	■

Applications of predictive modeling

Credit scoring

Past applicants

Cases

	y	X_1	X_2	$\cdots$	X_k
1	0	■	■	$\cdots$	■
2	1	■	■	$\cdots$	■
3	0	■	■	$\cdots$	■
$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$
n	1	■	■	$\cdots$	■

Applications of predictive modeling

Credit scoring

Input variables

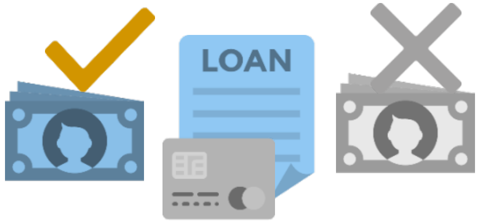
Past applicants

Cases

	y	X_1	X_2	$\cdots$	X_k
1	0	■	■	$\cdots$	■
2	1	■	■	$\cdots$	■
3	0	■	■	$\cdots$	■
$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$
n	1	■	■	$\cdots$	■

Applications of predictive modeling

Credit scoring



Target variable

Input variables

Past applicants

Cases

	y	X_1	X_2	$\cdots$	X_k
1	0	■	■	$\cdots$	■
2	1	■	■	$\cdots$	■
3	0	■	■	$\cdots$	■
$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$
n	1	■	■	$\cdots$	■

Reduce defaults and serious delinquencies

PONTIFICIA UNIVERSIDAD CATÓLICA DE VALPARAÍSO

Applications of predictive modeling

Fraud detection

Transactions

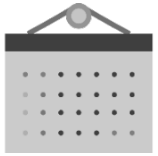
Cases

$$
\begin{array}{cccccc}
 & y & X_1 & X_2 & \cdots & X_k \\
1 & 0 & \blacksquare & \blacksquare & \cdots & \blacksquare \\
2 & 1 & \blacksquare & \blacksquare & \cdots & \blacksquare \\
3 & 0 & \blacksquare & \blacksquare & \cdots & \blacksquare \\
\vdots & \vdots & \vdots & \vdots & & \vdots \\
n & 1 & \blacksquare & \blacksquare & \cdots & \blacksquare
\end{array}
$$

Applications of predictive modeling

Fraud detection

Transactions



Input variables

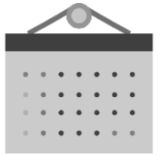
Cases

$$
\begin{array}{cccccc}
 & y & X_1 & X_2 & \cdots & X_k \\
1 & 0 & \blacksquare & \blacksquare & \cdots & \blacksquare \\
2 & 1 & \blacksquare & \blacksquare & \cdots & \blacksquare \\
3 & 0 & \blacksquare & \blacksquare & \cdots & \blacksquare \\
\vdots & \vdots & \vdots & \vdots & & \vdots \\
n & 1 & \blacksquare & \blacksquare & \cdots & \blacksquare
\end{array}
$$

Applications of predictive modeling

Fraud detection

Target variable



Input variables

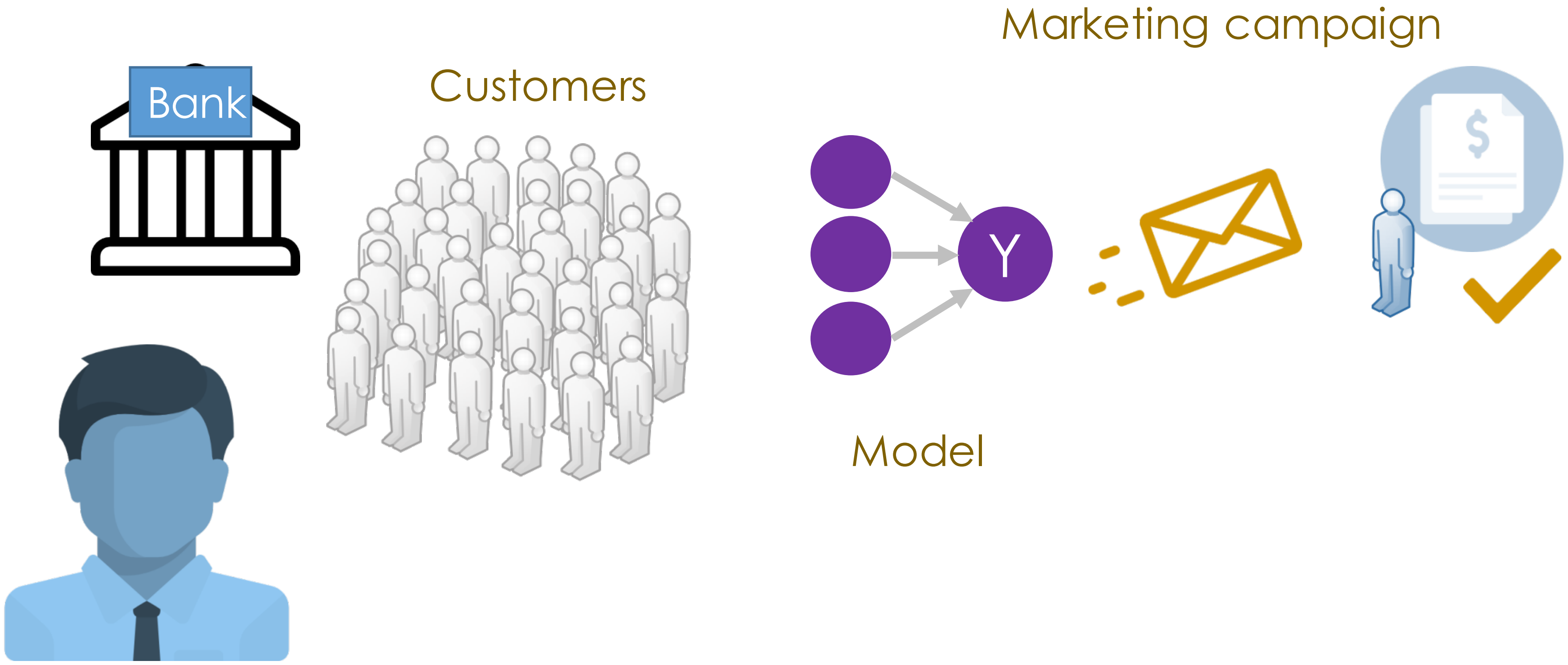
Transactions

Cases

	y	X_1	X_2	$\cdots$	X_k
1	0	■	■	$\cdots$	■
2	1	■	■	$\cdots$	■
3	0	■	■	$\cdots$	■
$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$
n	1	■	■	$\cdots$	■

Anticipate fraud or abuse

Demonstration Scenario: Target Marketing for a Bank



PONTIFICIA UNIVERSIDAD CATÓLICA DE VALPARAÍSO

Demonstration Scenario: Target Marketing for a Bank

```
%global inputs;
%let inputs=ACCTAGE DDA DDABAL DEP DEPAMT CASHBK
            CHECKS DIRDEP NSF NSFAMT PHONE TELLER
            SAV SAVBAL ATM ATMAMT POS POSAMT CD
            CDBAL IRA IRABAL LOC LOCBAL INV
            INVBAL ILS ILSBAL MM MMBAL MMCRED MTG
            MTGBAL CC CCBAL CCPURC SDB INCOME
            HMOWN LORES HMVAL AGE CRSCORE MOVED
            INAREA;

proc means data=work.develop n nmiss mean min max;
   var &inputs;
run;

proc freq data=work.develop;
   tables ins branch res;
run;
```

Demonstration Scenario: Target Marketing for a Bank

Questions

- What type of variable is Moved?

- How many variables have missing values?

- Look at the percentage of cases that have an Ins variable value of 1 versus those with a value of 0. What could you infer about the selection of cases for this data set?

- How many bank branches are represented in the data? Do you think this is a useful number of levels for the analysis?

- How many area classifications are represented in the data? Which area has the largest number of customers?

Demonstration Scenario: Target Marketing for a Bank
Questions

- What type of variable is Moved?

A: Moved is a binary variable.

- How many variables have missing values?

A: A total of 15 variables have missing values. Missing values are an issue. You learn how to handle missing values later in the course.

- Look at the percentage of cases that have an Ins variable value of 1 versus those with a value of 0. What could you infer about the selection of cases for this data set?

A: The results of PROC FREQ show that 34.6% of the customers in the develop data set purchased the insurance product. You might think that this percentage seems artificially high. In fact, the target event (buying the insurance product) is rare—only 2% of the population. To build the develop data set, the bank included all cases that have an Ins variable value of 1 and a representative sample of cases that have an Ins variable value of 0. This oversampling of the events increases the efficiency of the analysis because you are using a smaller sample and therefore have fewer cases to process. However, this oversampling also biases the results. You learn more about oversampling events, and how to adjust the model for it, later in the course.

- How many bank branches are represented in the data? Do you think this is a useful number of levels for the analysis?

A: The Branch of Bank table (the frequency table for Branch) indicates that the customers represented in the data do their banking in 19 different branches. When you determine that a categorical input variable has too many levels to be useful, you can collapse the levels. You learn to do this later in the course.

- How many area classifications are represented in the data? Which area has the largest number of customers?

A: The Area Classification table indicates that Res has three levels: R (rural), S (suburban), and U (urban). The largest number of customers live in urban areas, followed by suburban areas, and then rural areas.

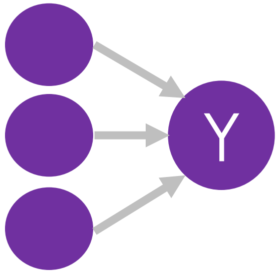
Predictive Modeling Challenges

PONTIFICIA UNIVERSIDAD CATÓLICA DE VALPARAÍSO

Predictive Modeling Challenges

Data challenges

Analytical challenges



Objectives

Describe challenges that predictive modelers commonly encounter	Identify solutions to some of these challenges
Define honest assessment	Split the data

Data challenges

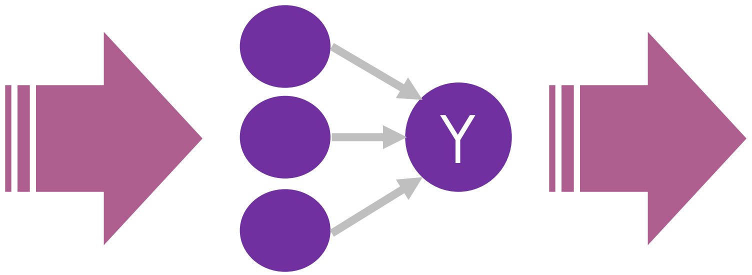
Observational data

Mixed measurement scales

High dimensionality

Rare target events

	y	X_1	X_2	$\cdots$	X_k
1	0	■	■	$\cdots$	■
2	1	■	■	$\cdots$	■
3	0	■	■	$\cdots$	■
$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$
n	1	■	■	$\cdots$	■



	y	X_1	X_2	$\cdots$	X_k
1	?	■	■	$\cdots$	■
2	?	■	■	$\cdots$	■
3	?	■	■	$\cdots$	■
$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$
n	?	■	■	$\cdots$	■

Data challenges

Observational data

Mixed measurement scales

High dimensionality

Rare target events

Redundant variab

Operationa

Opportunisti

	y	X_1	X_2	$\cdots$	X_k
1	0	■	■	$\cdots$	■
2	1	■	■	$\cdots$	■
3	0	■	■	$\cdots$	■
$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$
n	1	■	■	$\cdots$	■

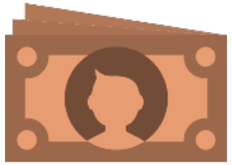
Missing value

Millions of case

Hundreds of variabl

Data challenges

Observational data	Mixed measurement scales	High dimensionality	Rare target events



Intervals 0, 0,85 , 2000, 50.15 , 10000,....

Discrete Male, Female

Count 0, 1, 2, 3,...

Nomina D, B, C, E, A

Data challenges

Observational data

Mixed measurement scales

High dimensionality

Rare target events

Account typ

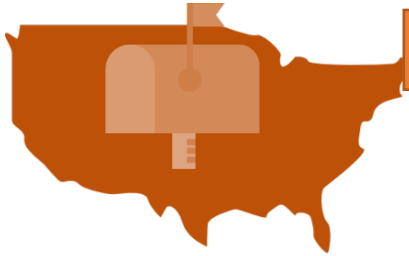
Dummy variabl

Value	Label	D1	D2
1	Checking	1	0
2	Savings	0	1
3	Other	0	0

PONTIFICIA UNIVERSIDAD CATÓLICA DE VALPARAÍSO

Data challenges

Observational data | **Mixed measurement scales** | High dimensionality | Rare target events



Zip code

Collapse labe

Zip Code	City	State
90620	Buena Park	California
90621	Buena Park	California
90622	Buena Park	California
90623	La Palma	California
90624	Buena Park	California
90630	Cypress	California
90631	La Habra	California
90632	La Habra	California
90633	La Habra	California
90680	Stanton	California
90720	Los Alamitos	California
90721	Los Alamitos	California
90740	Seal Beach	California
90742	Sunset Beach	California
90743	Surfside	California
92602	Irvine	California
92603	Irvine	California

Data challenges

Observational data

Mixed measurement scales

High dimensionality

Rare target events

Dimensions

$$\begin{array}{cccccc}
 & y & X_1 & X_2 & \cdots & X_k \\
1 & 0 & \blacksquare & \blacksquare & \cdots & \blacksquare \\
2 & 1 & \blacksquare & \blacksquare & \cdots & \blacksquare \\
3 & 0 & \blacksquare & \blacksquare & \cdots & \blacksquare \\
\vdots & \vdots & \vdots & \vdots & & \vdots \\
n & 1 & \blacksquare & \blacksquare & \cdots & \blacksquare
\end{array}$$

Data challenges

- Observational data
- Mixed measurement scales
- **High dimensionality**
- Rare target events

Dimensions

	y	X_1	X_2	$\cdots$	X_k	X_{k+1}	X_{k+2}	X_{k+3}	$\cdots$	X_{n+w}
1	0	$\blacksquare$	$\blacksquare$	$\cdots$	$\blacksquare$	$\blacksquare$	$\blacksquare$	$\blacksquare$	$\cdots$	$\blacksquare$
2	1	$\blacksquare$	$\blacksquare$	$\cdots$	$\blacksquare$	$\blacksquare$	$\blacksquare$	$\blacksquare$	$\cdots$	$\blacksquare$
3	0	$\blacksquare$	$\blacksquare$	$\cdots$	$\blacksquare$	$\blacksquare$	$\blacksquare$	$\blacksquare$	$\cdots$	$\blacksquare$
$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$
n	1	$\blacksquare$	$\blacksquare$	$\cdots$	$\blacksquare$	$\blacksquare$	$\blacksquare$	$\blacksquare$	$\cdots$	$\blacksquare$

Data challenges

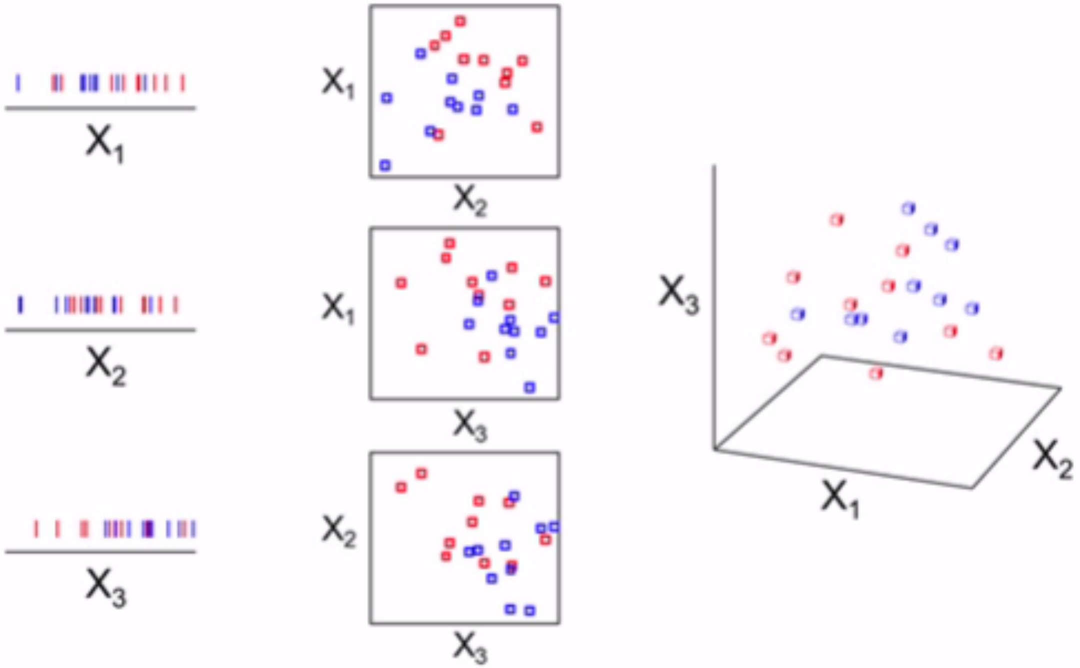
Observational data

Mixed measurement scales

High dimensionality

Rare target events

Really sparse dc



Data challenges

Observational data | Mixed measurement scales | **High dimensionality** | Rare target events

Curse of dimensiona

$$
\begin{array}{ccccccccccc}
 & y & \overbrace{X_1 \quad X_2 \quad \cdots \quad X_k}^{\text{Dimensions}} & X_{k+1} & X_{k+2} & X_{k+3} & \cdots & X_{n+w} \\
1 & 0 & \blacksquare \quad \blacksquare \quad \cdots \quad \blacksquare & \blacksquare & \blacksquare & \blacksquare & \cdots & \blacksquare \\
2 & 1 & \blacksquare \quad \blacksquare \quad \cdots \quad \blacksquare & \blacksquare & \blacksquare & \blacksquare & \cdots & \blacksquare \\
3 & 0 & \blacksquare \quad \blacksquare \quad \cdots \quad \blacksquare & \blacksquare & \blacksquare & \blacksquare & \cdots & \blacksquare \\
\vdots & \vdots & \vdots \quad \vdots \quad \quad \quad \vdots & \vdots & \vdots & \vdots & & \vdots \\
n & 1 & \blacksquare \quad \blacksquare \quad \cdots \quad \blacksquare & \blacksquare & \blacksquare & \blacksquare & \cdots & \blacksquare
\end{array}
$$

$$\boldsymbol{n < n + w}$$

Hard to asses variable relation

Data challenges

Observational data	Mixed measurement scales	High dimensionality	Rare target events

	y	X_1	X_2	X_3
1	0	■	■	■
2	1	■	■	■
3	0	■	■	■
⋮	⋮	⋮	⋮	⋮
n	1	■	■	■

Reduce dimension

Take the most important vari

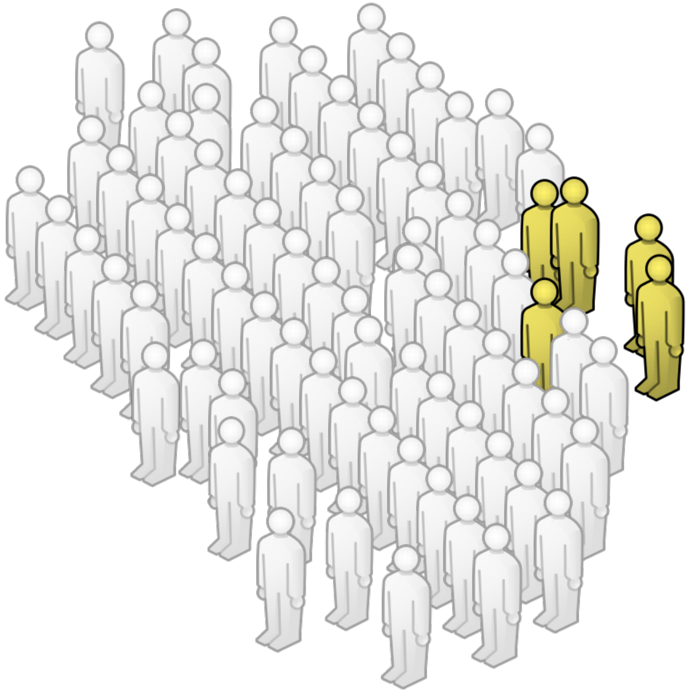
Data challenges

Observational data

Mixed measurement scales

High dimensionality

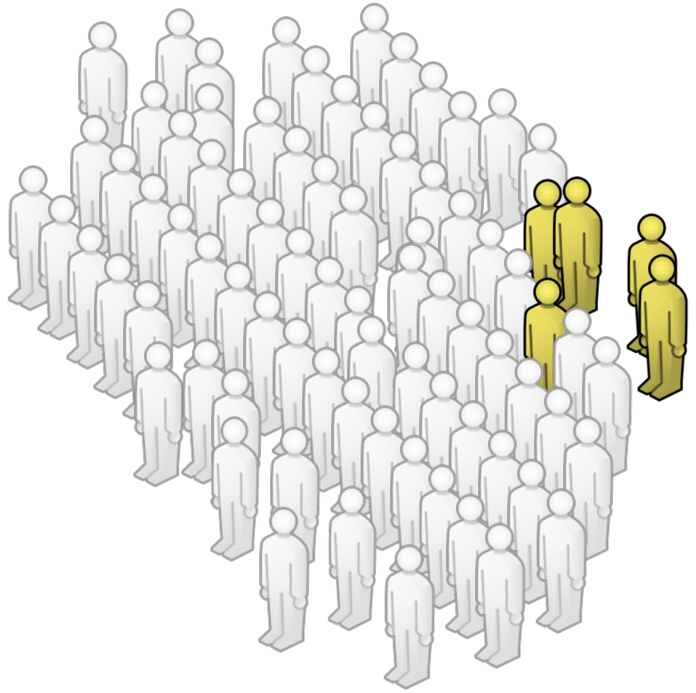
Rare target events



Data challenges

- Observational data
- Mixed measurement scales
- High dimensionality
- **Rare target events**

No - Even



Event

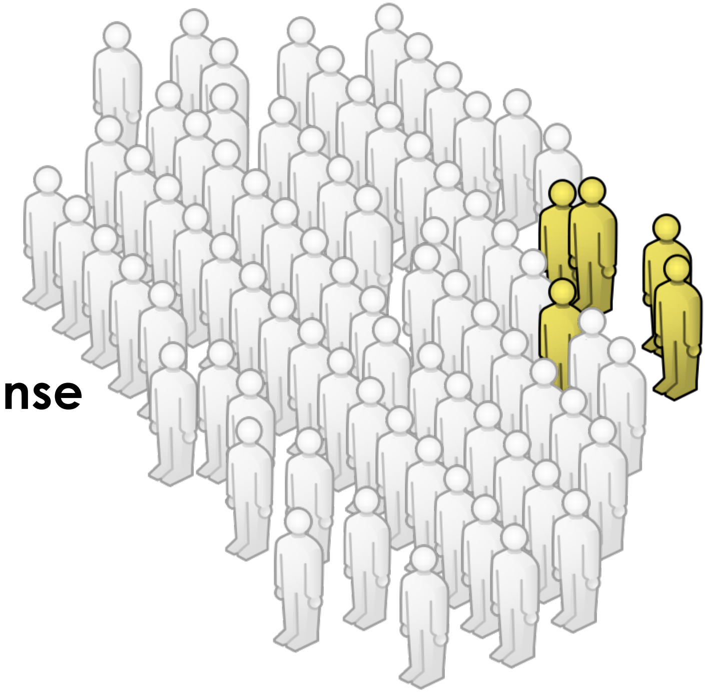
<1%

Response
Fraud
Churn
Default
Buy

Data challenges

Observational data | Mixed measurement scales | High dimensionality | Rare target events

No - Even



Event

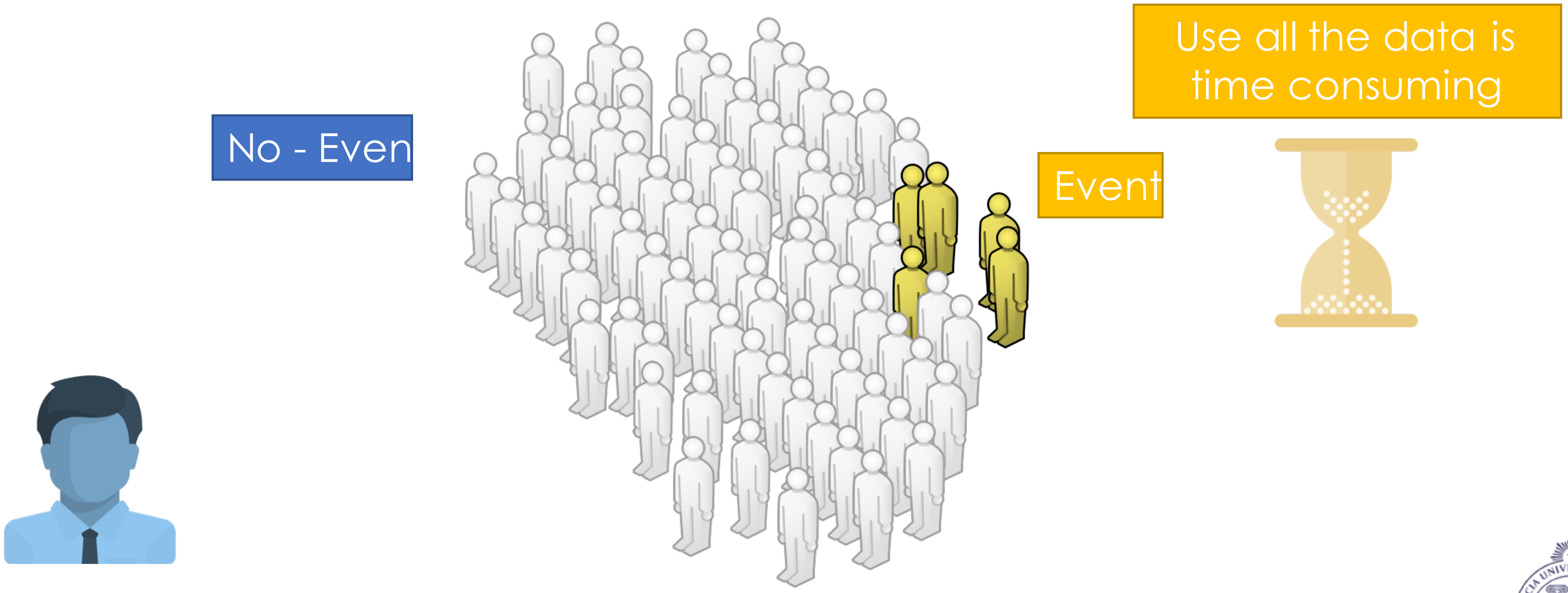
<1%

Response
Fraud
Churn
Default
Buy

No Response
Legitimate
Stay
Pay
Not Buy

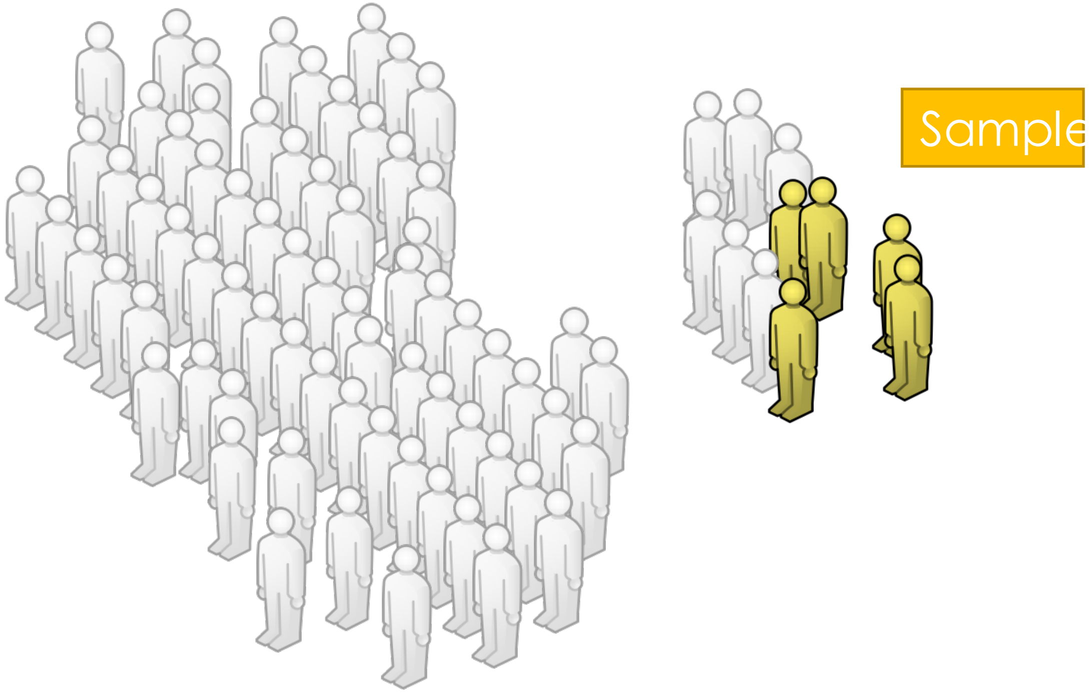
Data challenges

Observational data | Mixed measurement scales | High dimensionality | Rare target events



Data challenges

- Observational data
- Mixed measurement scales
- High dimensionality
- Rare target events



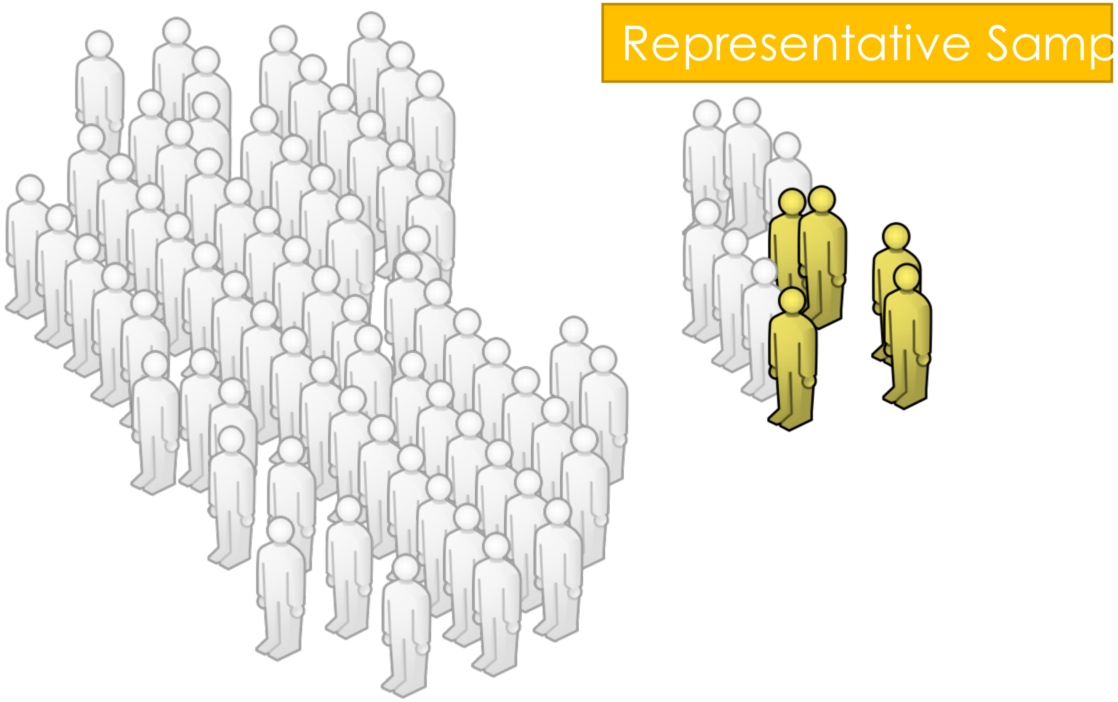
Data challenges

Observational data

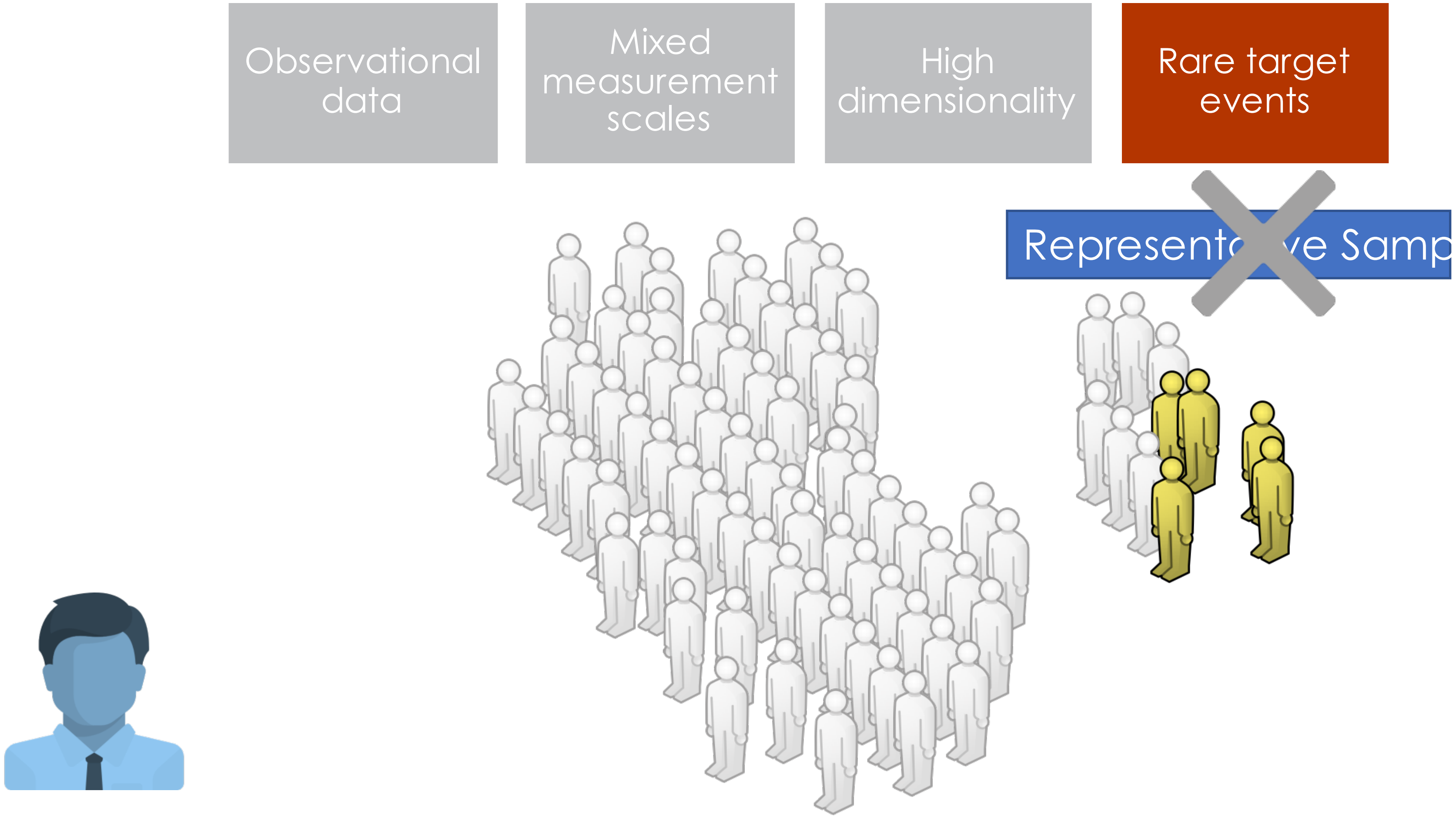
Mixed measurement scales

High dimensionality

Rare target events



Data challenges



Data challenges

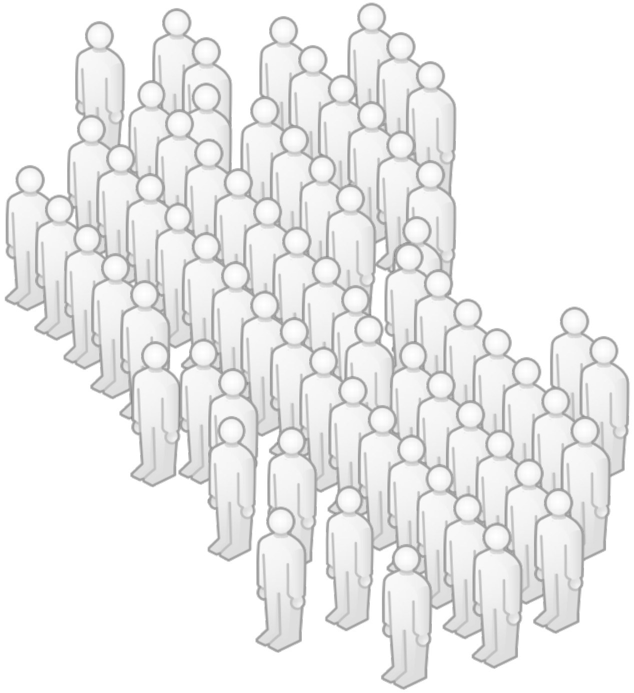
Observational data

Mixed measurement scales

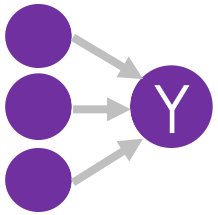
High dimensionality

Rare target events

x the number of event case



Oversampling



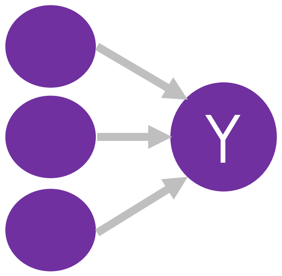
$\approx$ *Predictive power*

Analytical challenges

Non linearity and interactions

Model selection

	y	X_1	X_2	$\cdots$	X_k
1	0	■	■	$\cdots$	■
2	1	■	■	$\cdots$	■
3	0	■	■	$\cdots$	■
$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$
n	1	■	■	$\cdots$	■

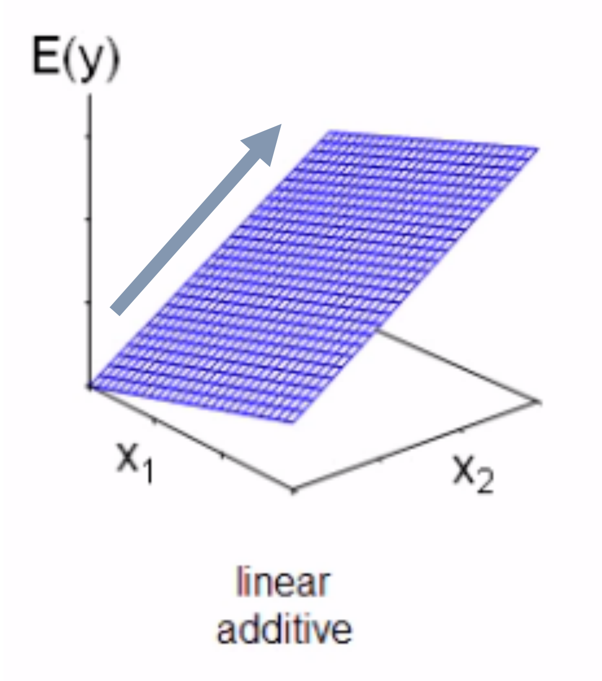


	y	X_1	X_2	$\cdots$	X_k
1	?	■	■	$\cdots$	■
2	?	■	■	$\cdots$	■
3	?	■	■	$\cdots$	■
$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$
n	?	■	■	$\cdots$	■

Analytical challenges

Non linearity and interactions

Model selection



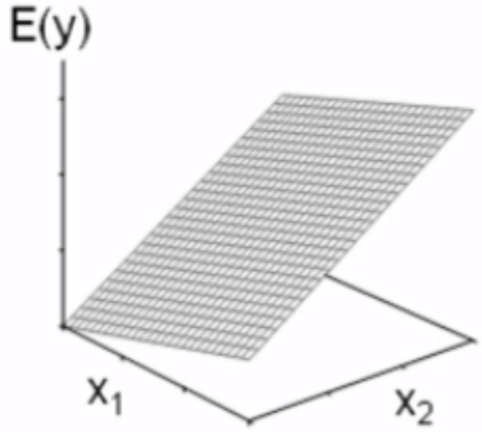
PONTIFICIA UNIVERSIDAD CATÓLICA DE VALPARAÍSO

Analytical challenges

Non linearity and interactions

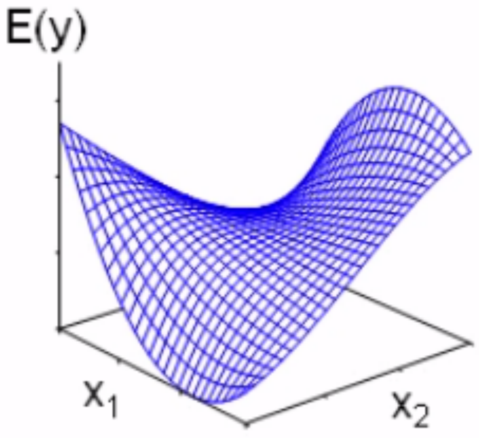
Model selection

Theory



linear
additive

Reality

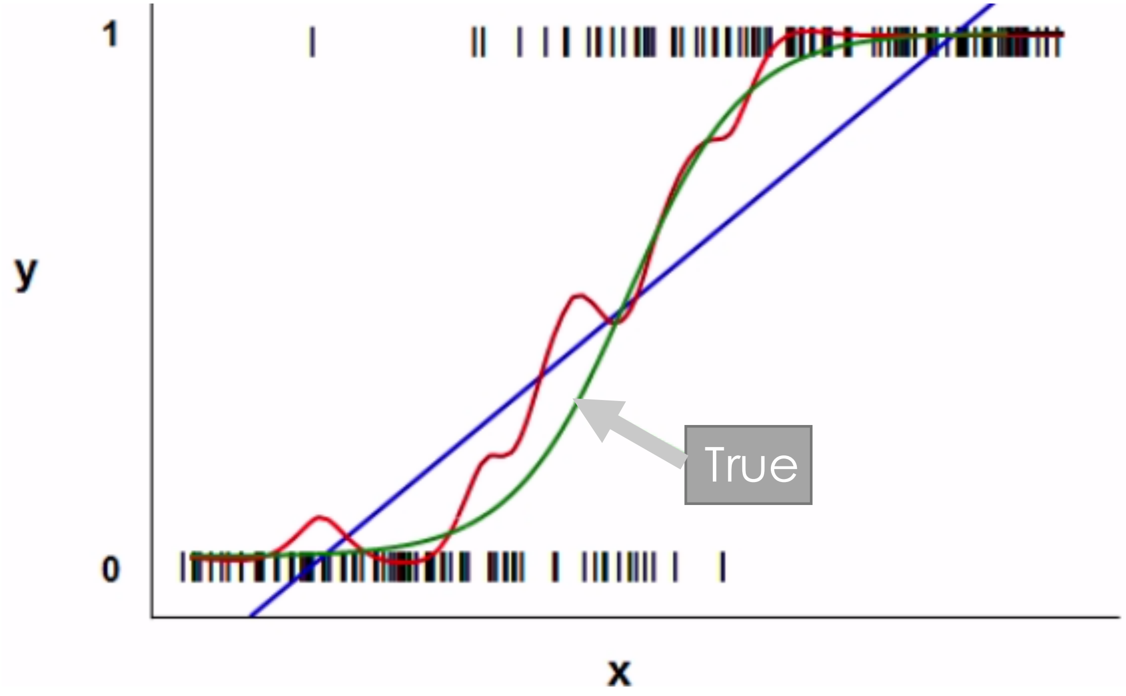


nonlinear
nonadditive

Analytical challenges

Non linearity and interactions

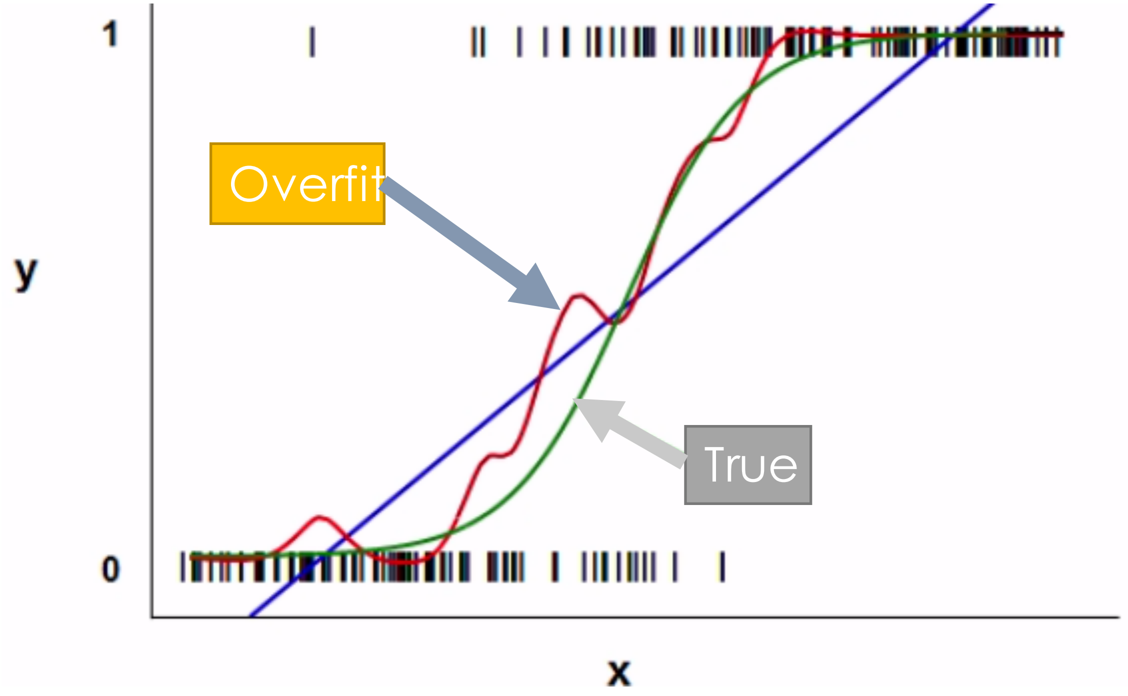
Model selection



Analytical challenges

Non linearity and interactions

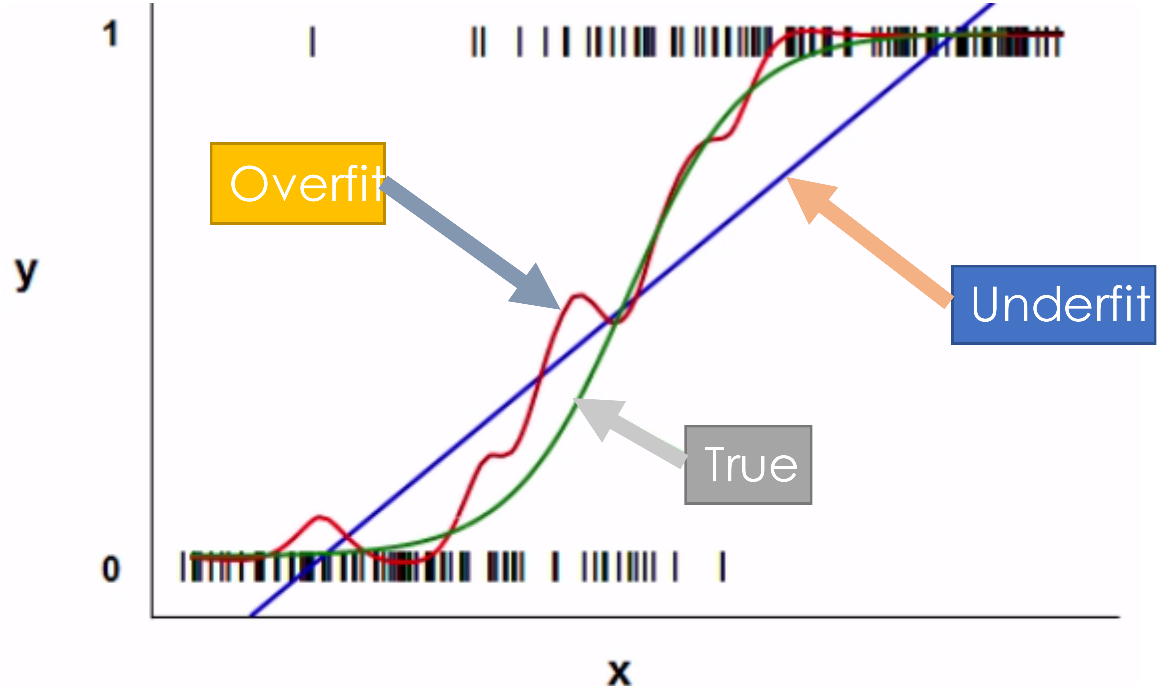
Model selection



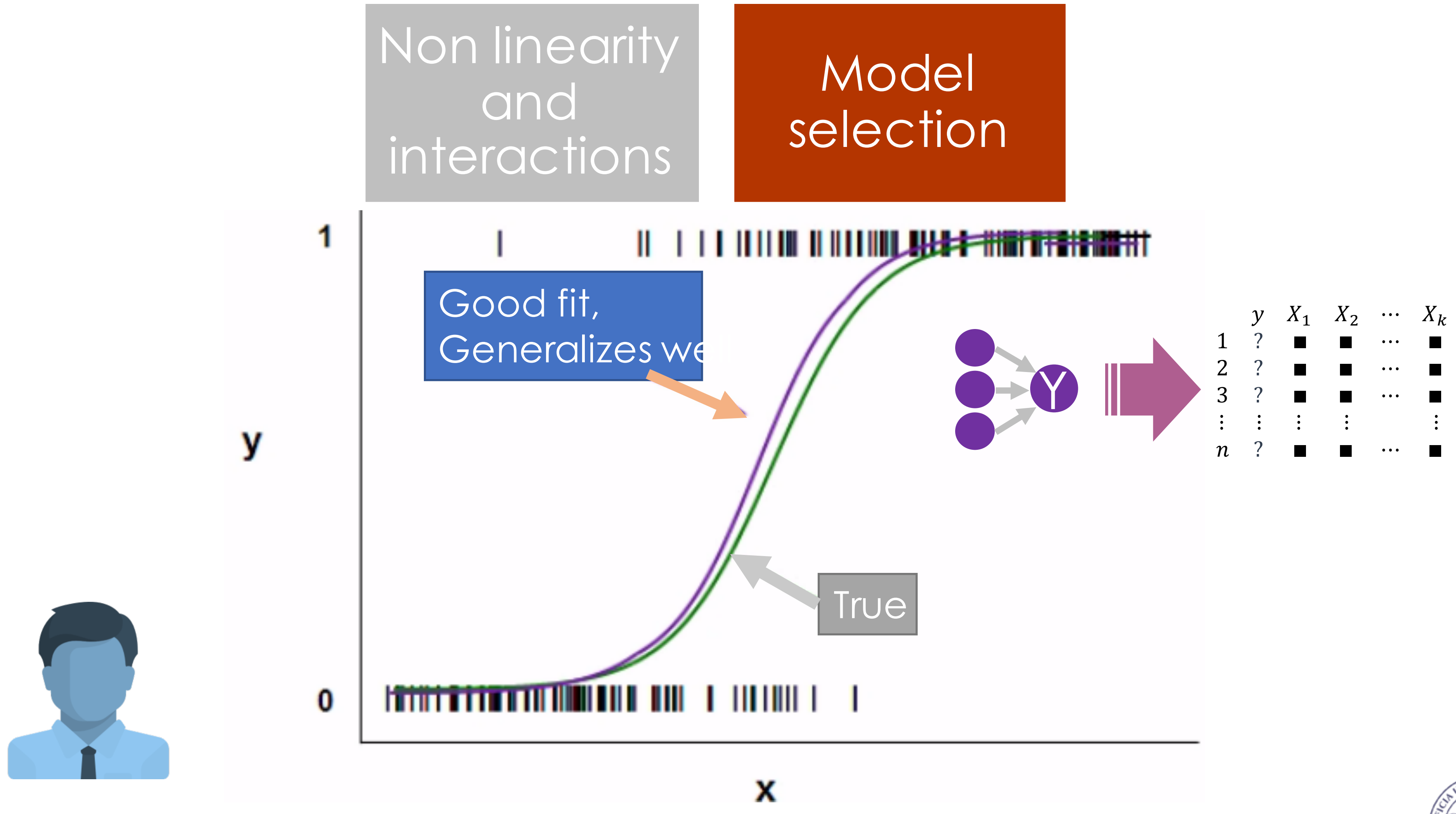
Analytical challenges

Non linearity and interactions

Model selection

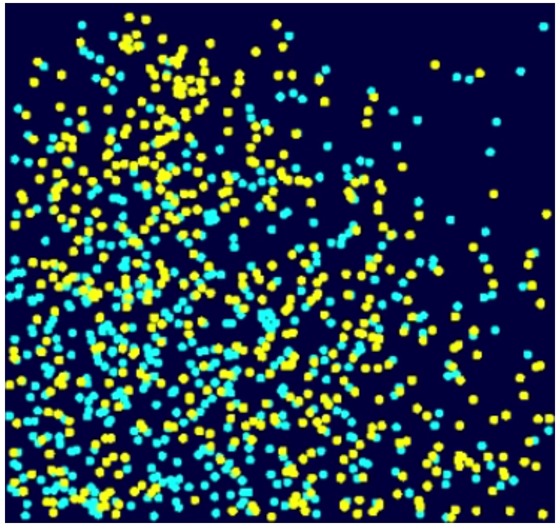


Analytical challenges

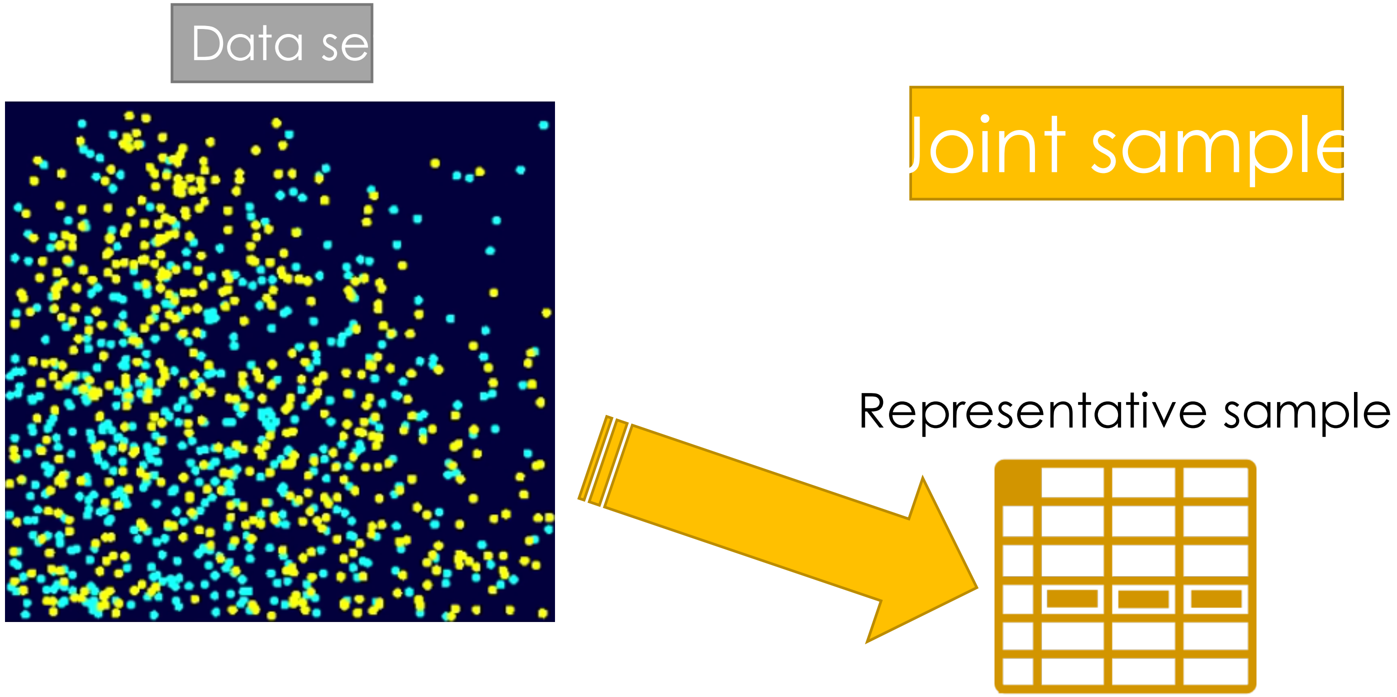


Separate sample

Data se

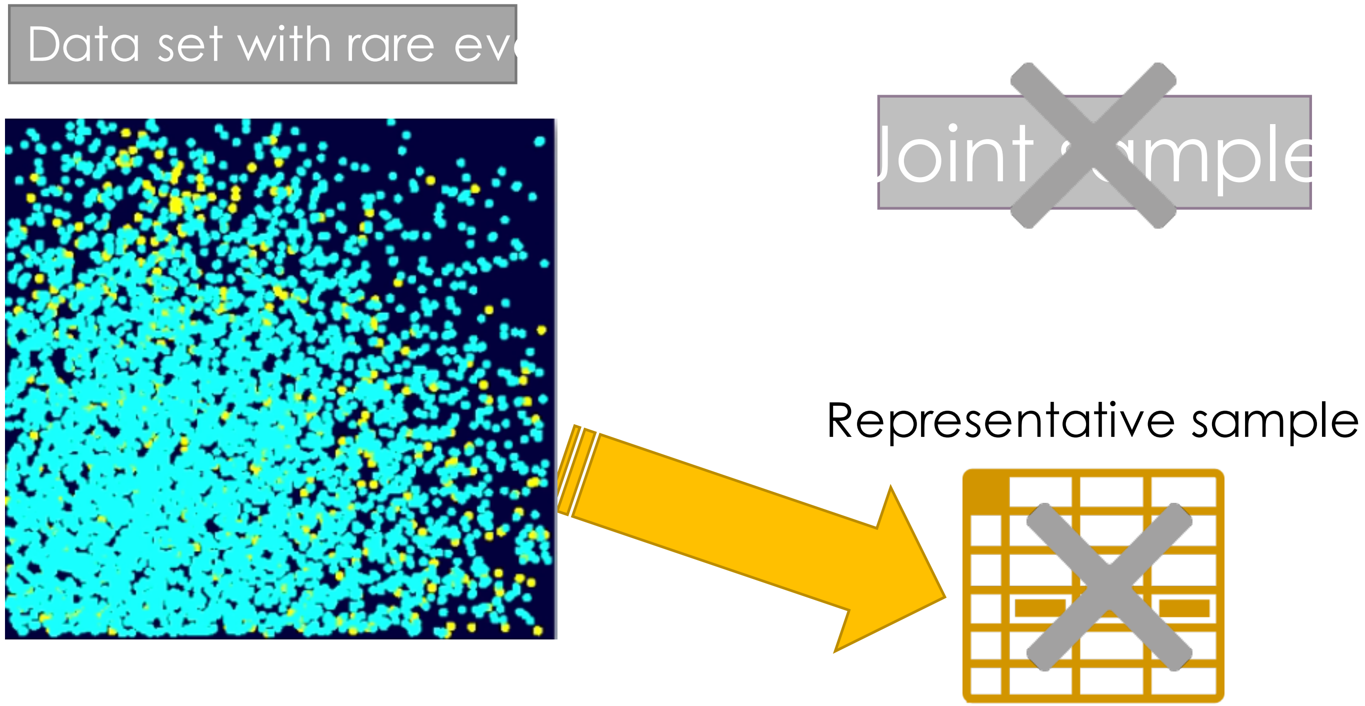


Separate sample

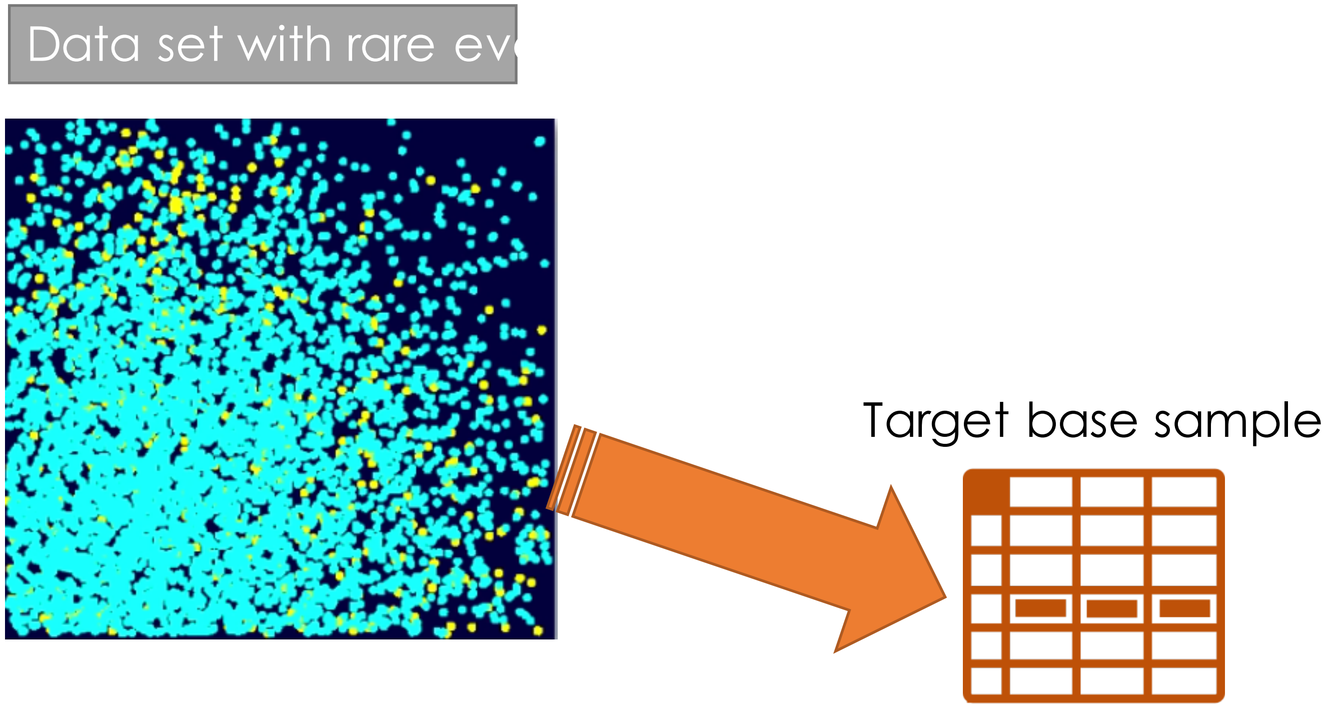


Data set with equal proportion of events to non-events

Separate sample

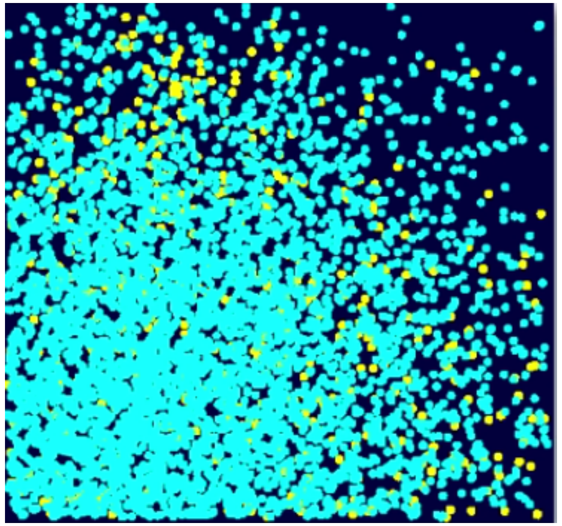


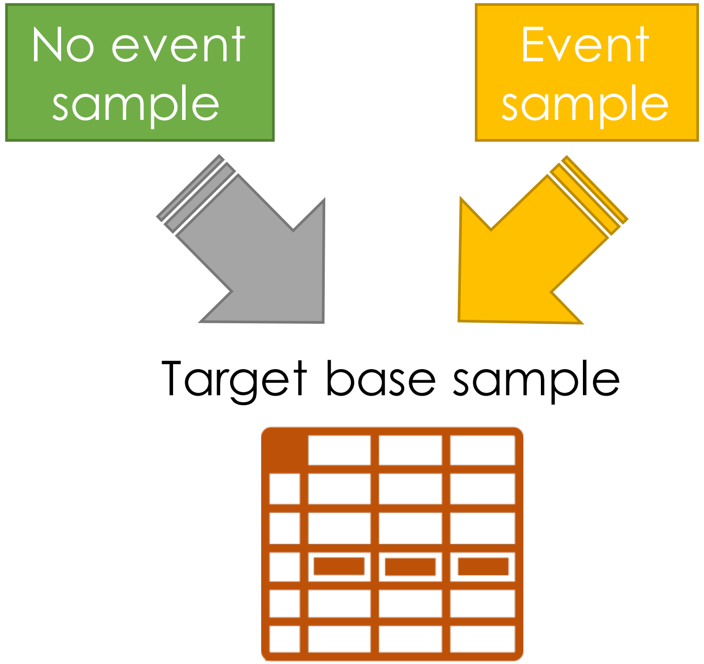
Separate sample



Separate sample

Data set with rare ev

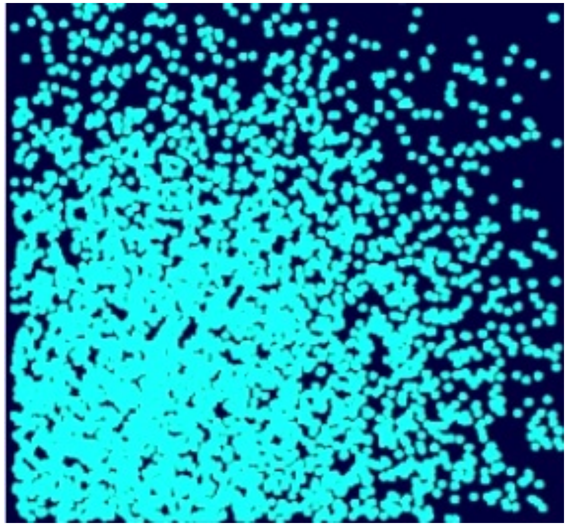




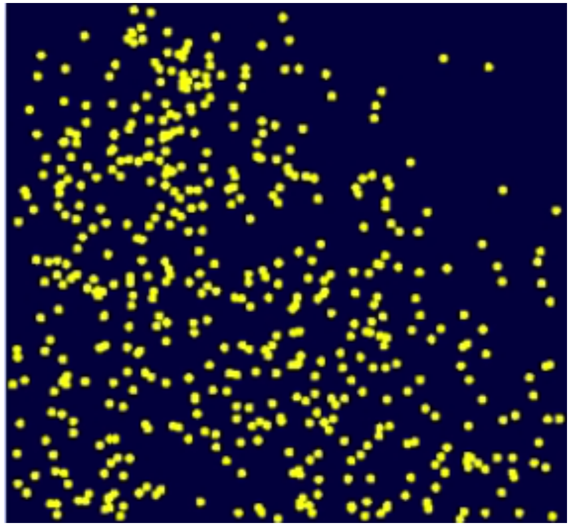
PONTIFICIA UNIVERSIDAD CATÓLICA DE VALPARAÍSO

Separate sample

Secondary outcome (Non-e

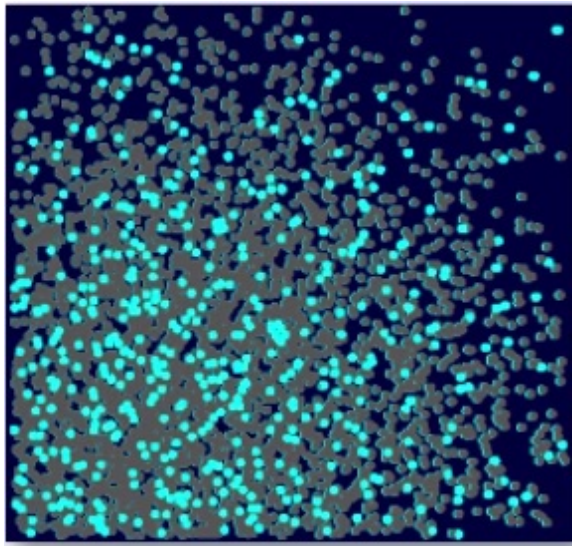


Primary outcome (eve



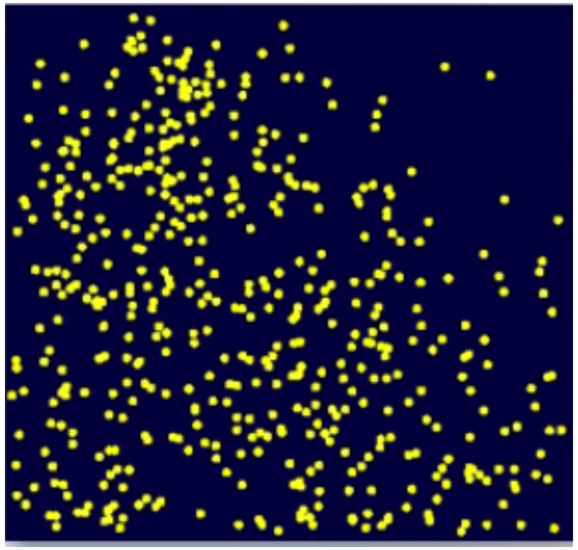
Separate sample

Secondary outcome (Non-e



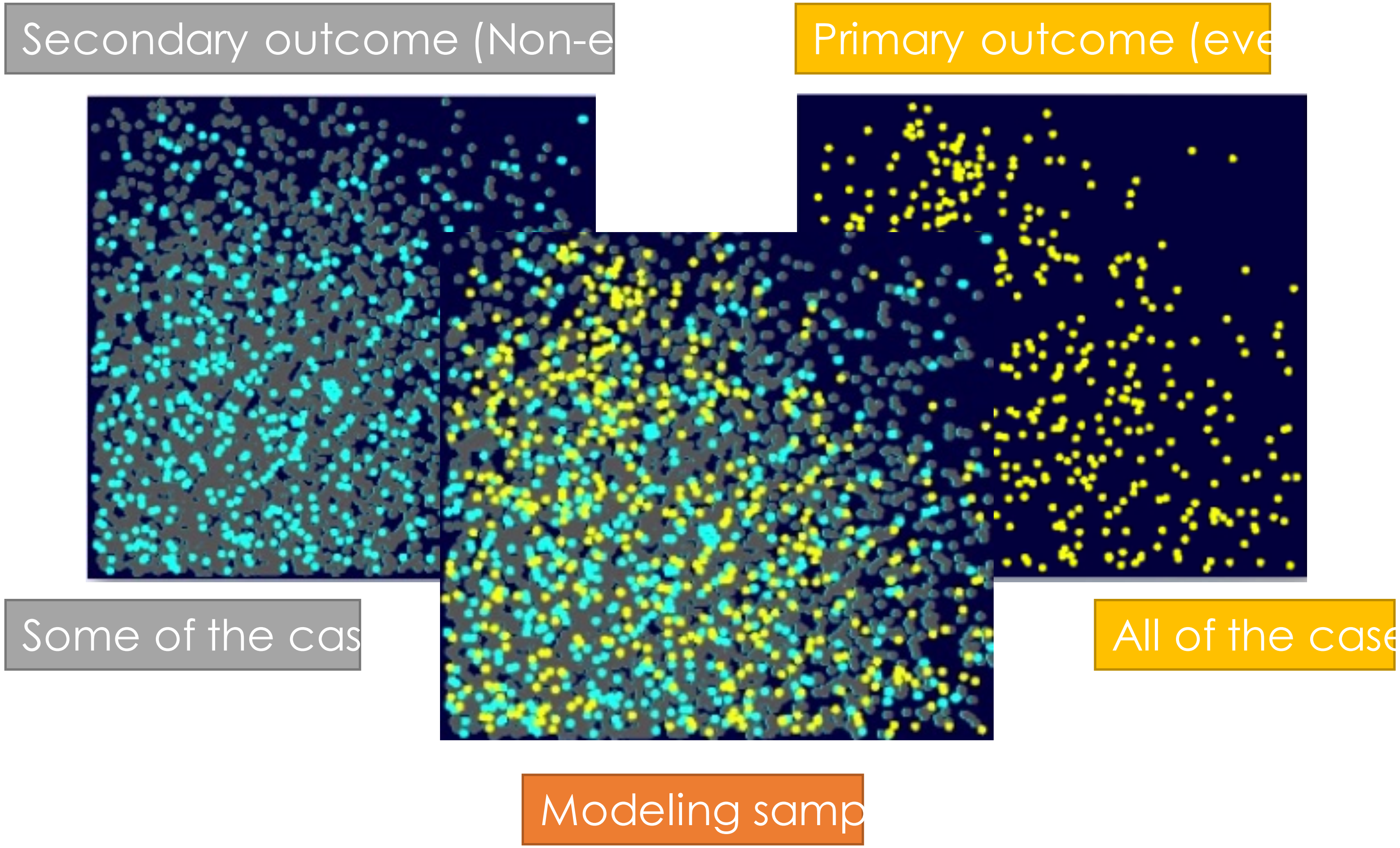
Some of the cas

Primary outcome (eve



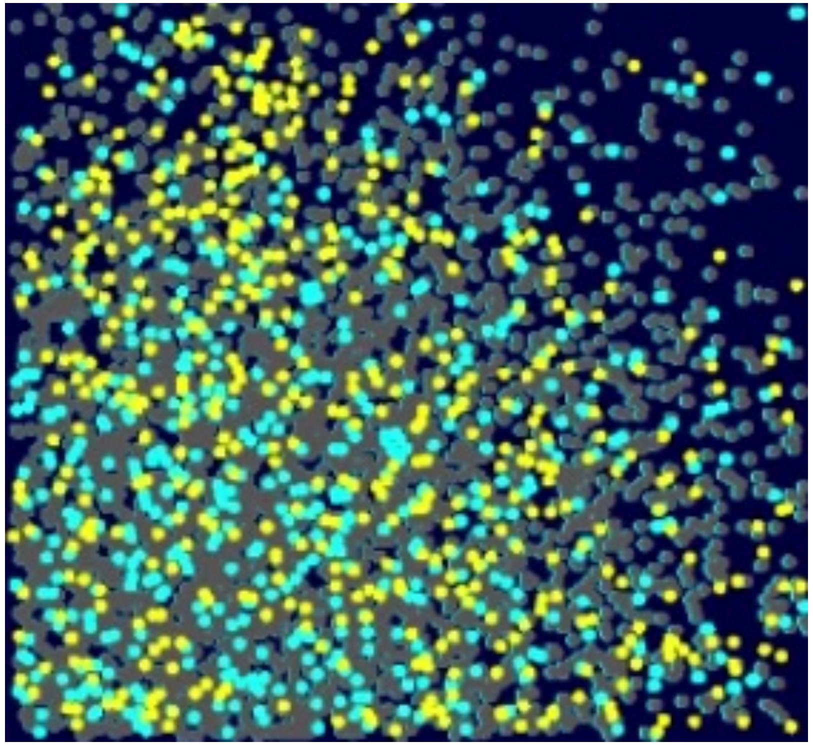
All of the case

Separate sample



Separate sample

Modeling samp



Efficient

But generate bi

Separate sample

Which of the following data scenarios lends itself most to oversampling the target?

a. a data set that consists of 100 events and 5,000 non-events

b. a data set that consists of 50 events and 10,000 non-events

c. a data set that consists of 5,000 events and 25,000 non-events

d. a data set that consists of 1,000 events and 5,000,000 non-events

Separate sample

Which of the following data scenarios lends itself most to oversampling the target?

a. a data set that consists of 100 events and 5,000 non-events

b. a data set that consists of 50 events and 10,000 non-events

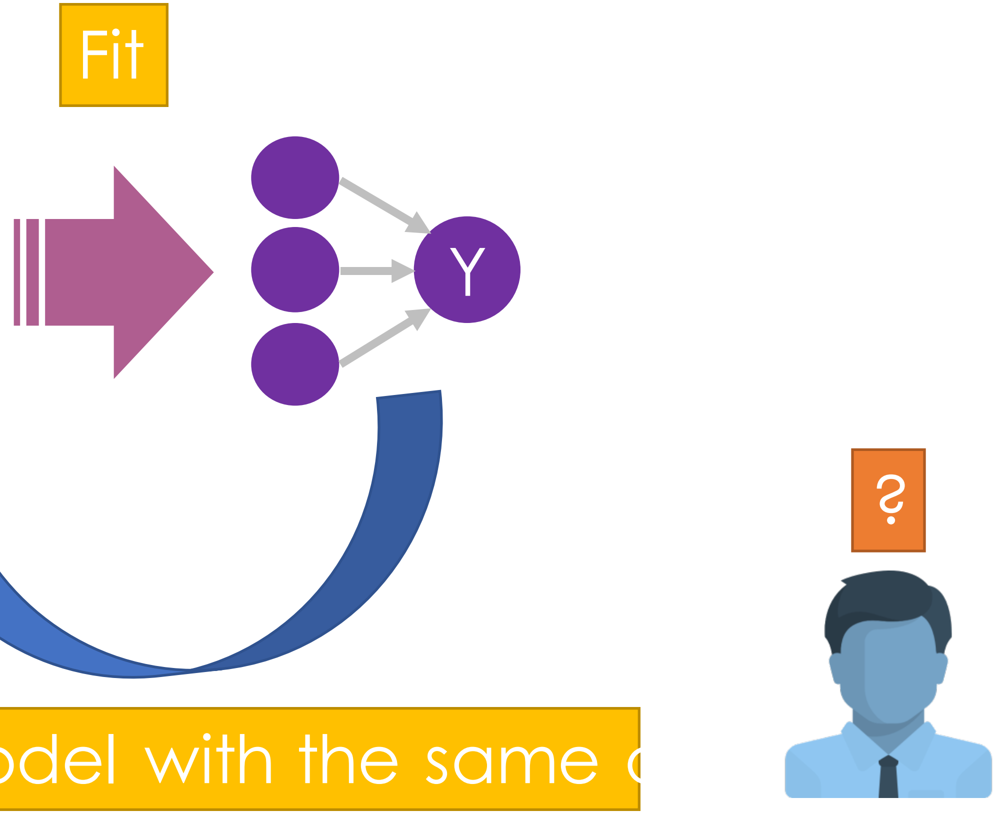
c. a data set that consists of 5,000 events and 25,000 non-events

d. **a data set that consists of 1,000 events and 5,000,000 non-events**

If you have millions of cases but only a thousand events, analyzing all of the non-events is inefficient. In scenario *d*, the ratio of non-events to events is 5000 to 1, which is larger than the ratios for the other scenarios.

Avoiding the Optimism Bias: Honest assessment

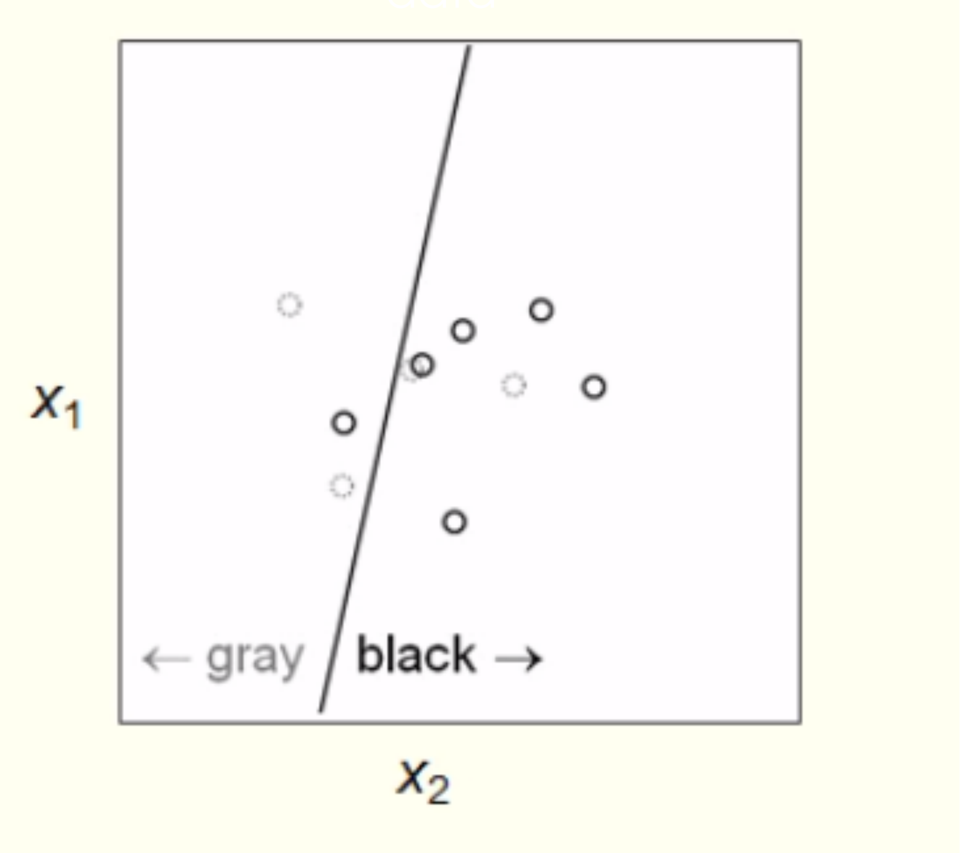
	y	X_1	X_2	$\cdots$	X_k
1	0	■	■	$\cdots$	■
2	1	■	■	$\cdots$	■
3	0	■	■	$\cdots$	■
$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$
n	1	■	■	$\cdots$	■



Avoiding the Optimism Bias: Honest assessment

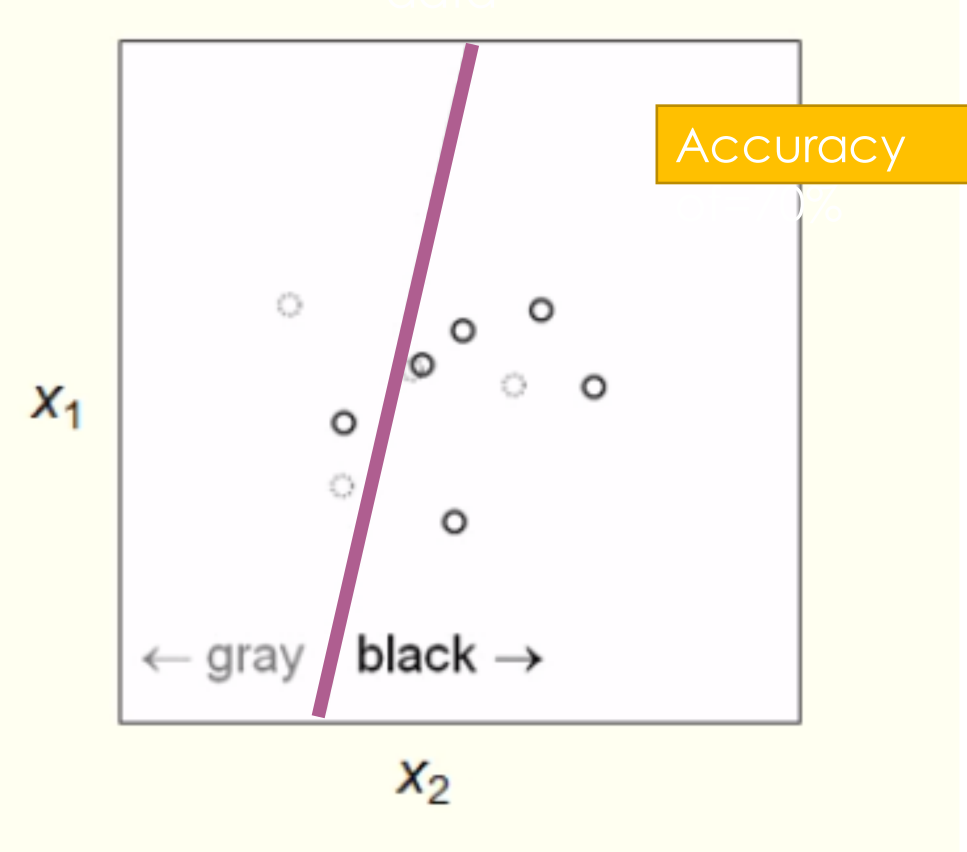
Model assessed on development

10 case data set with two



Avoiding the Optimism Bias: Honest assessment

Model assessed on development



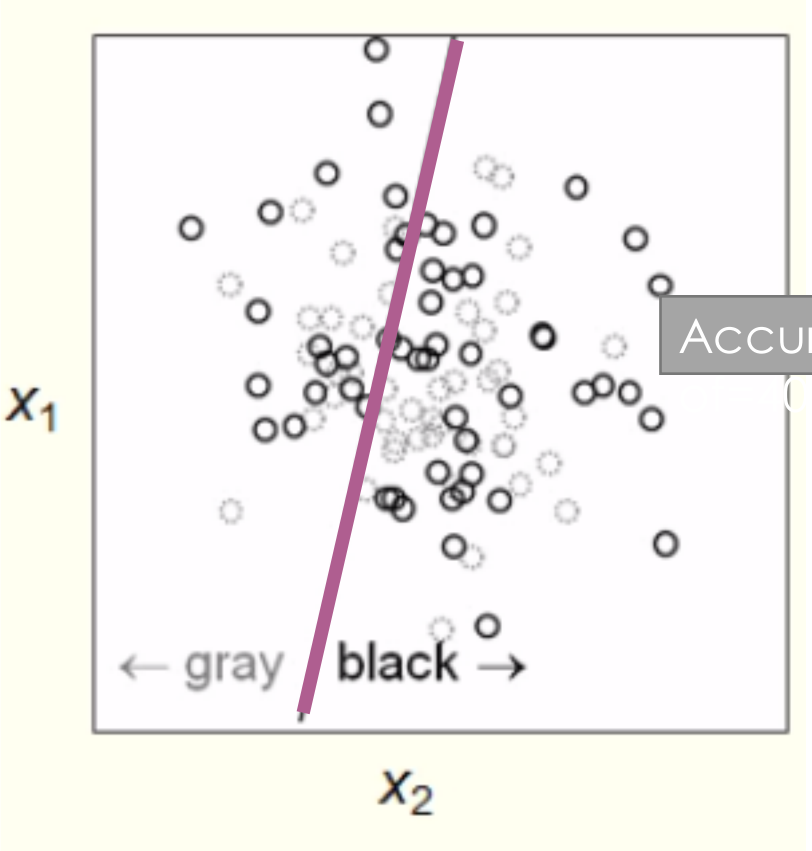
Accuracy

10 case data set with two

Generalize well to new data?

Avoiding the Optimism Bias: Honest assessment

Model assessed on 100 new cases

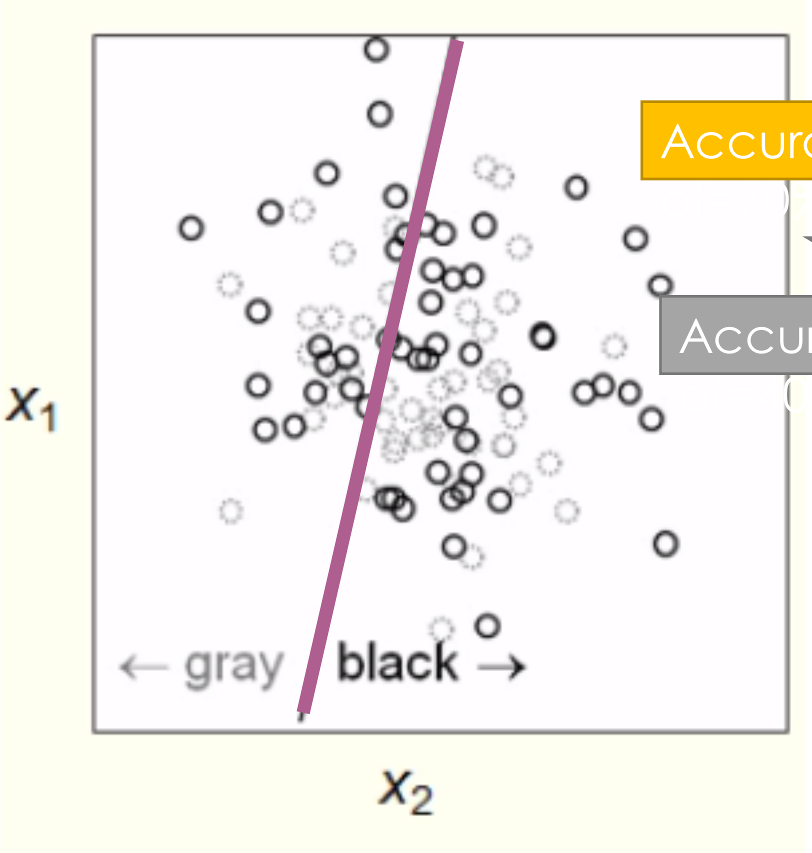


Accuracy

Generalize well to new data?

Avoiding the Optimism Bias: Honest assessment

Model assessed on 100 new cases



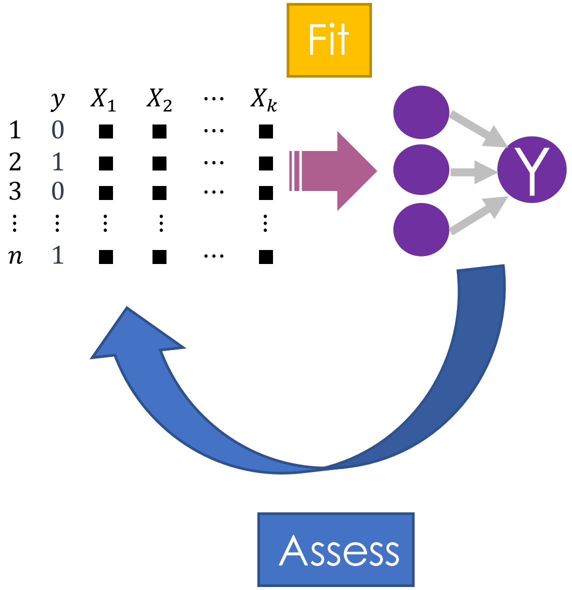
Accuracy

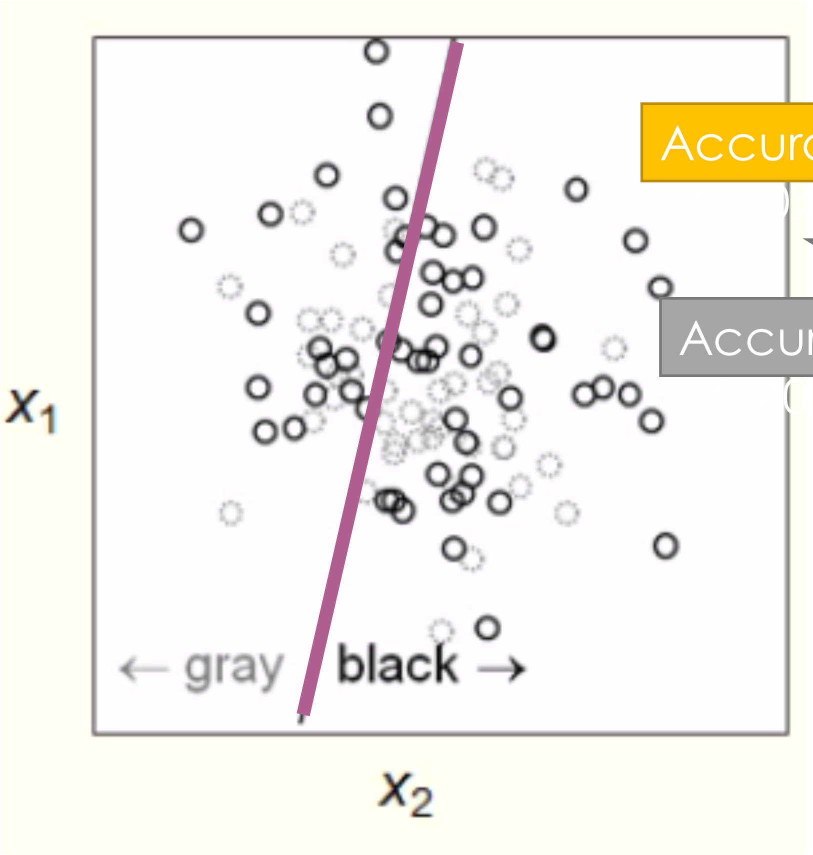
Accuracy

This indicate that you overfit or underfit the

Avoiding the Optimism Bias: Honest assessment

Model assessed on 100 new cases





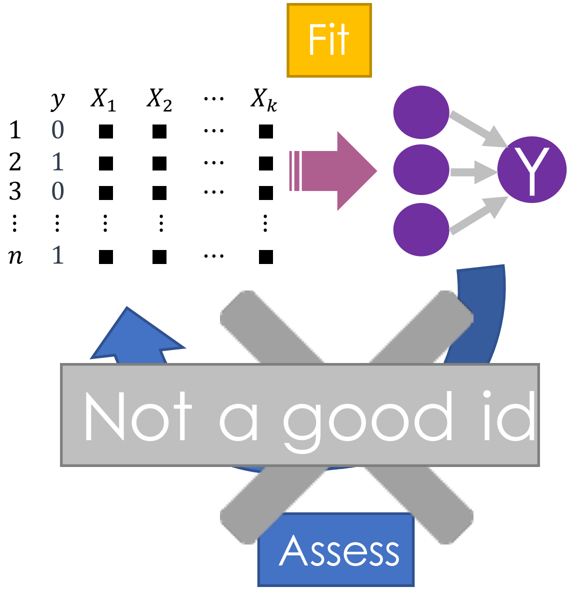
Accuracy

Accuracy

This indicate that you overfit or underfit the

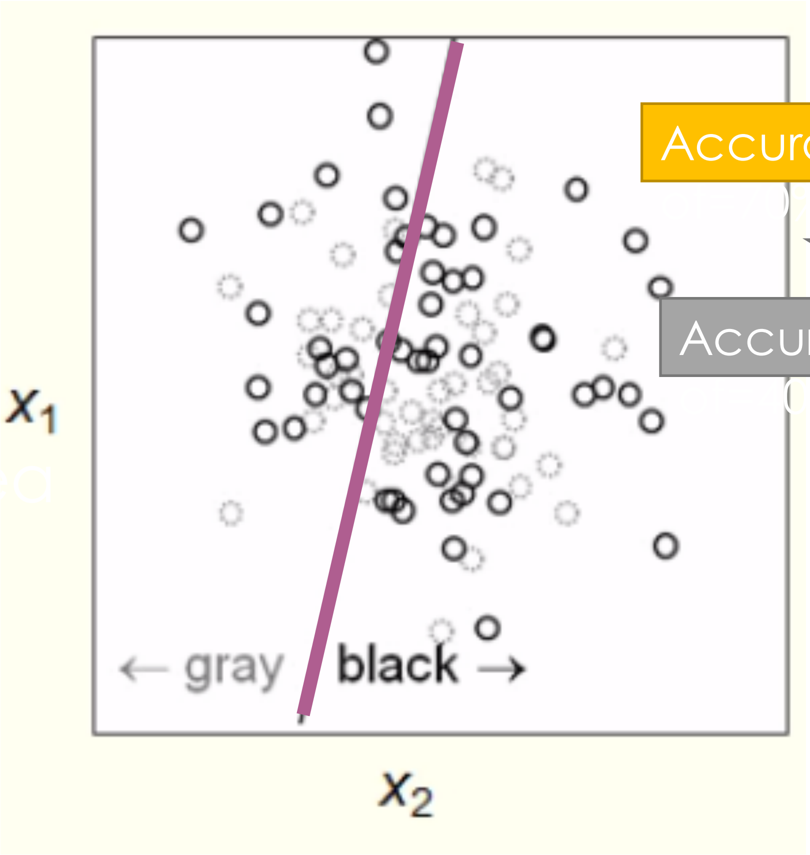
Avoiding the Optimism Bias: Honest assessment

Model assessed on 100 new cases



Not a good id

Optimism bi



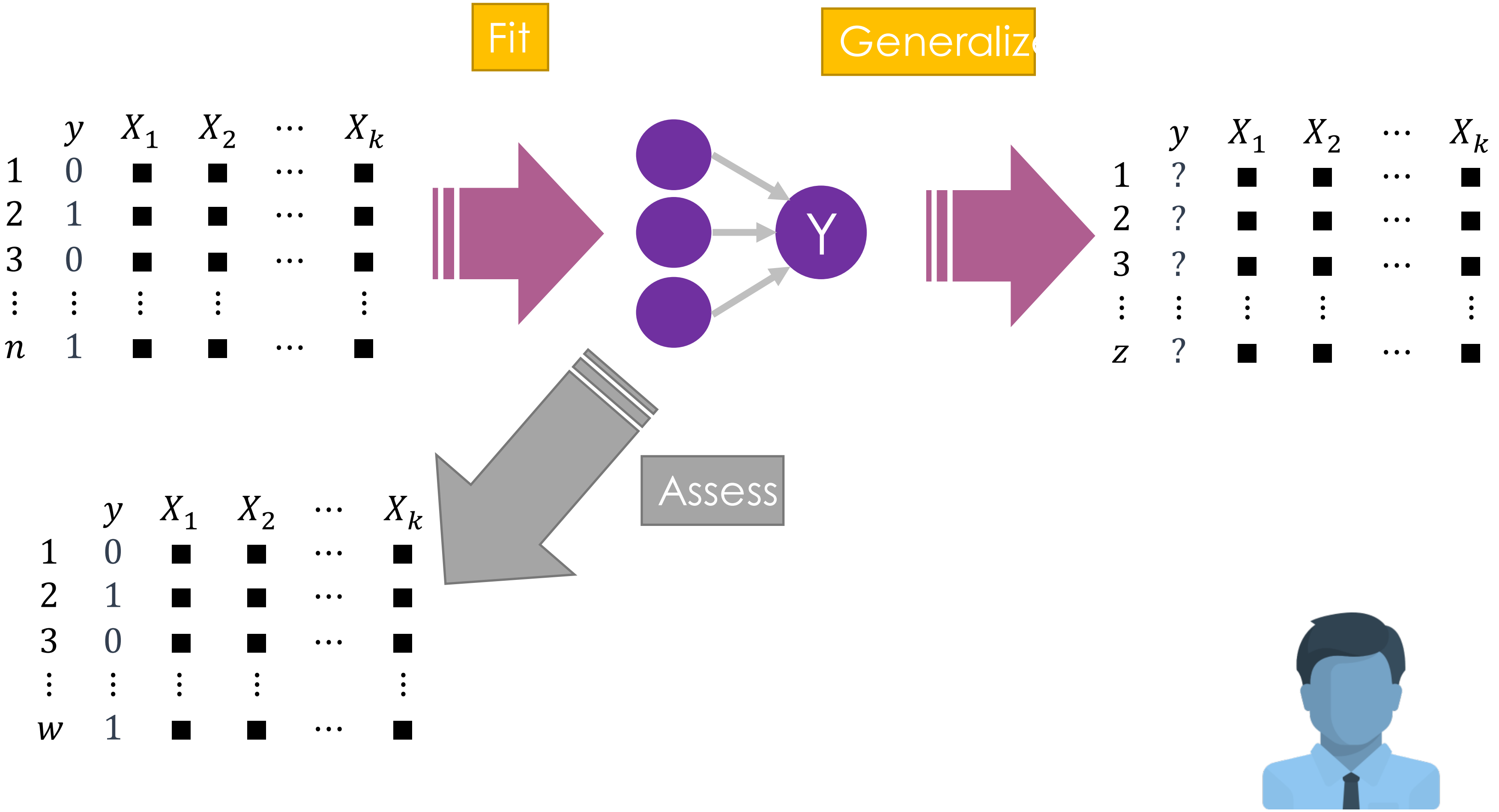
Accuracy

Accuracy

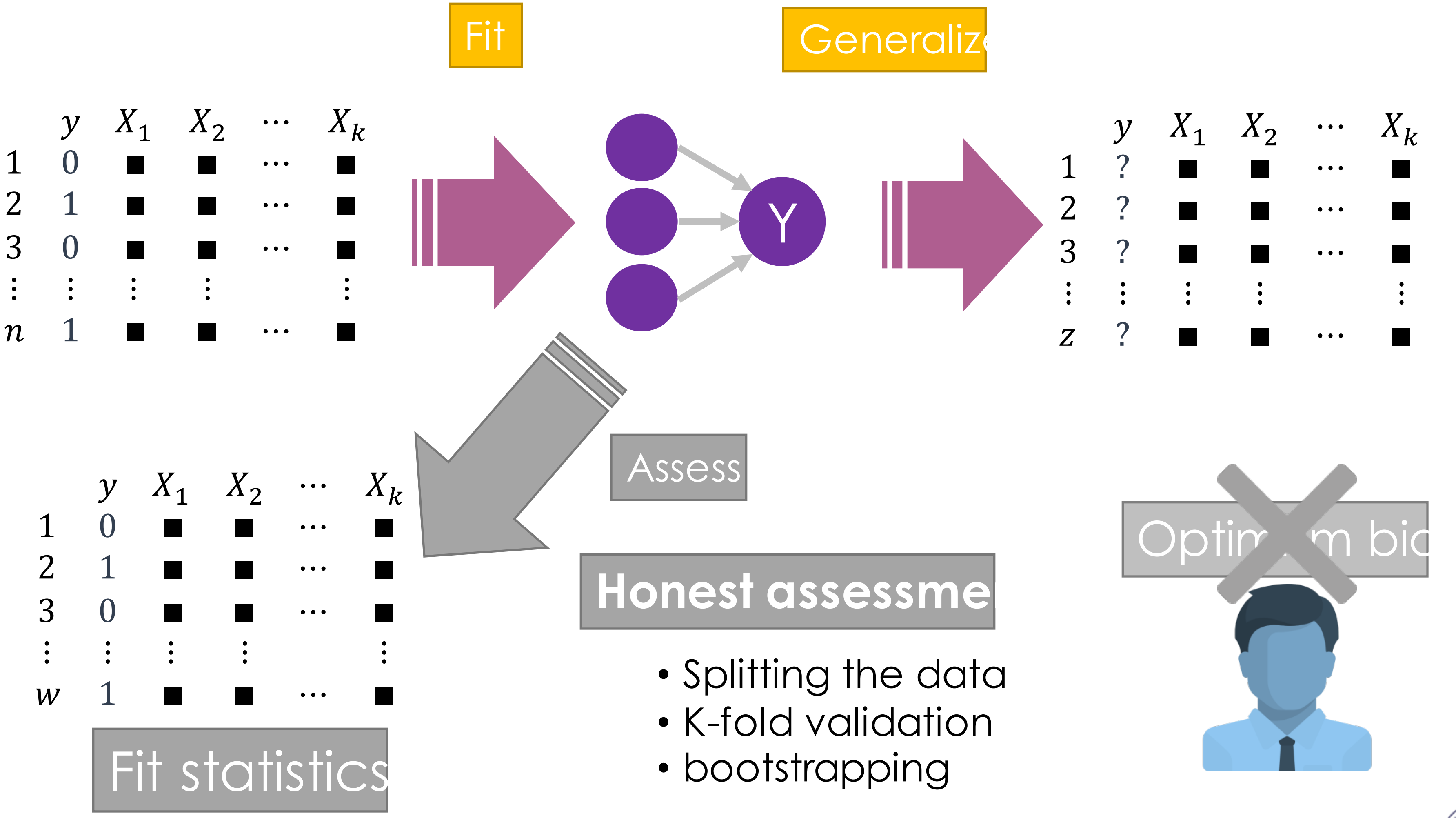
This indicate that you overfit or underfit the

PONTIFICIA UNIVERSIDAD CATÓLICA DE VALPARAÍSO

Avoiding the Optimism Bias: Honest assessment



Avoiding the Optimism Bias: Honest assessment



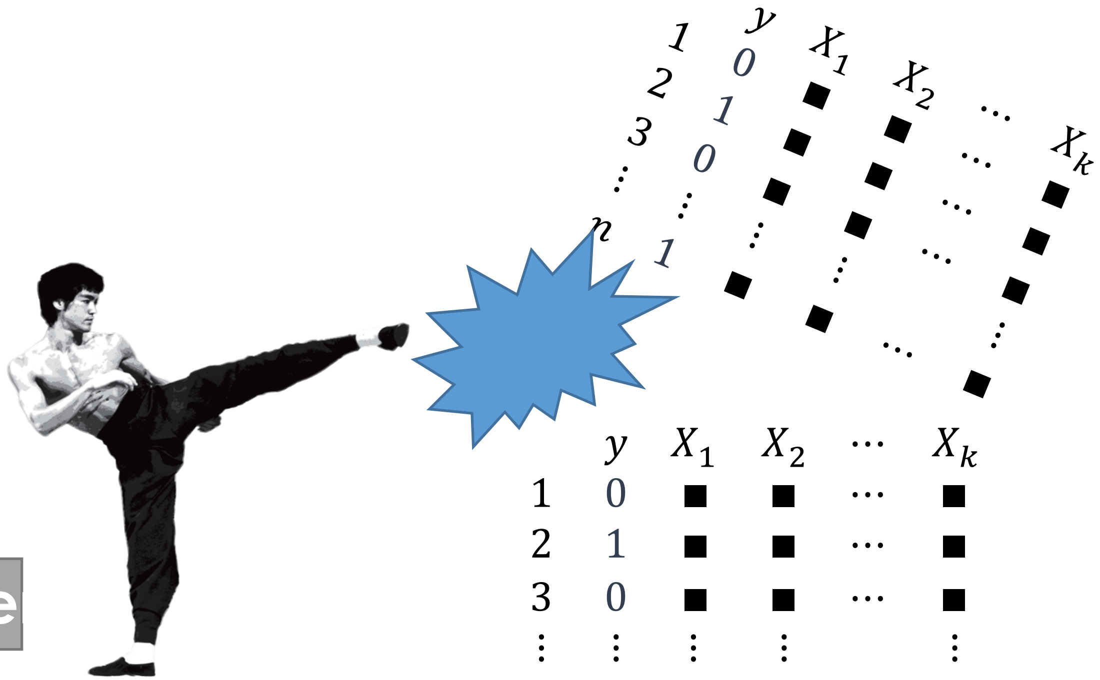
PONTIFICIA UNIVERSIDAD CATÓLICA DE VALPARAÍSO

Splitting the Data for model training and assessment

	y	X_1	X_2	$\cdots$	X_k
1	0	■	■	$\cdots$	■
2	1	■	■	$\cdots$	■
3	0	■	■	$\cdots$	■
4	1	■	■	$\cdots$	■
5	0	■	■	$\cdots$	■
6	1	■	■	$\cdots$	■
7	0	■	■	$\cdots$	■
8	0	■	■	$\cdots$	■
$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$
m	1	■	■	$\cdots$	■

Honest assessme

Splitting the Data for model training and assessment



	y	X_1	X_2	$\cdots$	X_k
1	0	■	■	$\cdots$	■
2	1	■	■	$\cdots$	■
3	0	■	■	$\cdots$	■
$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$
w	1	■	■	$\cdots$	■

Honest assessme

PONTIFICIA UNIVERSIDAD CATÓLICA DE VALPARAÍSO

Splitting the Data for model training and assessment

Training data s

Fit the mode

	y	X_1	X_2	$\cdots$	X_k
1	0	■	■	$\cdots$	■
2	1	■	■	$\cdots$	■
3	0	■	■	$\cdots$	■
$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$
n	1	■	■	$\cdots$	■

Validation data

Holdout portio

Honest assessme

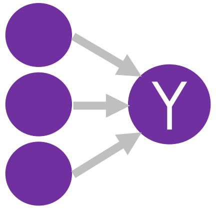
	y	X_1	X_2	$\cdots$	X_k
1	0	■	■	$\cdots$	■
2	1	■	■	$\cdots$	■
3	0	■	■	$\cdots$	■
$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$
w	1	■	■	$\cdots$	■

Splitting the Data for model training and assessment

Training data s

Fit the mode

	y	X_1	X_2	$\cdots$	X_k
1	0	■	■	$\cdots$	■
2	1	■	■	$\cdots$	■
3	0	■	■	$\cdots$	■
$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$
n	1	■	■	$\cdots$	■

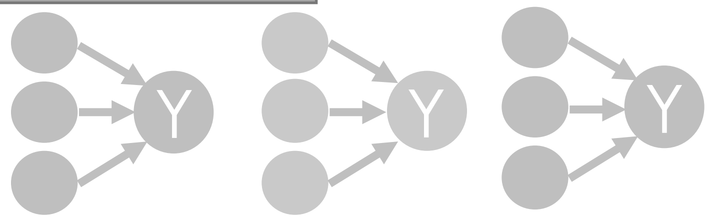


Train the model

Validation data

Holdout portio

	y	X_1	X_2	$\cdots$	X_k
1	0	■	■	$\cdots$	■
2	1	■	■	$\cdots$	■
3	0	■	■	$\cdots$	■
$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$
w	1	■	■	$\cdots$	■



Assess and compare models

Honest assessme

Splitting the Data for model training and assessment

Training data s

	y	X_1	X_2	$\cdots$	X_k
1	0	■	■	$\cdots$	■
2	1	■	■	$\cdots$	■
3	0	■	■	$\cdots$	■
$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$
n	1	■	■	$\cdots$	■

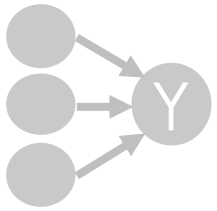
Test data se

	y	X_1	X_2	$\cdots$	X_k
1	0	■	■	$\cdots$	■
2	1	■	■	$\cdots$	■
3	0	■	■	$\cdots$	■
$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$
x	1	■	■	$\cdots$	■

Validation data

	y	X_1	X_2	$\cdots$	X_k
1	0	■	■	$\cdots$	■
2	1	■	■	$\cdots$	■
3	0	■	■	$\cdots$	■
$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$
w	1	■	■	$\cdots$	■

Honest assessme



Final assessment on the selected model

Splitting the Data for model training and assessment

Training data s

$$
\begin{array}{cccccc}
 & y & X_1 & X_2 & \cdots & X_k \\
1 & 0 & \blacksquare & \blacksquare & \cdots & \blacksquare \\
2 & 1 & \blacksquare & \blacksquare & \cdots & \blacksquare \\
3 & 0 & \blacksquare & \blacksquare & \cdots & \blacksquare \\
\vdots & \vdots & \vdots & \vdots & & \vdots \\
n & 1 & \blacksquare & \blacksquare & \cdots & \blacksquare
\end{array}
$$

Percentage of ca

Rule of thumb

2/3 or 66,66%

Validation data

Honest assessme

$$
\begin{array}{cccccc}
 & y & X_1 & X_2 & \cdots & X_k \\
1 & 0 & \blacksquare & \blacksquare & \cdots & \blacksquare \\
2 & 1 & \blacksquare & \blacksquare & \cdots & \blacksquare \\
3 & 0 & \blacksquare & \blacksquare & \cdots & \blacksquare \\
\vdots & \vdots & \vdots & \vdots & & \vdots \\
w & 1 & \blacksquare & \blacksquare & \cdots & \blacksquare
\end{array}
$$

1/3 or 33,33%

Splitting the Data for model training and assessment

Training data s

	y	X_1	X_2	$\cdots$	X_k
1	0	■	■	$\cdots$	■
2	1	■	■	$\cdots$	■
3	0	■	■	$\cdots$	■
$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$
n	1	■	■	$\cdots$	■

Percentage of co

Rule of thumb

2/3 or 66,66%

Random samplin

Validation data

	y	X_1	X_2	$\cdots$	X_k
1	0	■	■	$\cdots$	■
2	1	■	■	$\cdots$	■
3	0	■	■	$\cdots$	■
$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$
w	1	■	■	$\cdots$	■

1/3 or 33,33%

Honest assessme

Splitting the Data for model training and assessment

Training data s

	y	X_1	X_2	$\cdots$	X_k
1	0	■	■	$\cdots$	■
2	1	■	■	$\cdots$	■
3	0	■	■	$\cdots$	■
$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$
n	1	■	■	$\cdots$	■

Validation data

	y	X_1	X_2	$\cdots$	X_k
1	0	■	■	$\cdots$	■
2	1	■	■	$\cdots$	■
3	0	■	■	$\cdots$	■
$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$
w	1	■	■	$\cdots$	■

Honest assessme

Percentage of ca

Rule of thumb

2/3 or 66,66%

~~Random ampling~~

1/3 or 33,33%

PONTIFICIA UNIVERSIDAD CATÓLICA DE VALPARAÍSO

Splitting the Data for model training and assessment

Stratified random samp

Strata

	Training (66.67%)	Validation (33.33%)	
Event	7,451 (35%)	3,724 (35%)	11,175 (35%)
Non-event	14,061 (65%)	7,028 (65%)	21,089 (65%)
	21,512 (100%)	10,752 (100%)	32,264 (100%)

Honest assessme

$$
\begin{array}{cccccc}
 & y & X_1 & X_2 & \cdots & X_k \\
1 & 0 & \blacksquare & \blacksquare & \cdots & \blacksquare \\
2 & 1 & \blacksquare & \blacksquare & \cdots & \blacksquare \\
3 & 0 & \blacksquare & \blacksquare & \cdots & \blacksquare \\
\vdots & \vdots & \vdots & \vdots & & \vdots \\
n & 1 & \blacksquare & \blacksquare & \cdots & \blacksquare
\end{array}
$$

Develop data s

32,264 (100%)

Splitting the Data for model training and assessment

Which of the following statements is true regarding model assessment?

a. Data splitting can be used only on data with continuous targets.
b. The validation data set is used to calculate the parameter estimates and validate the model.
c. Assessing the performance of the model on the data that you used to fit the model usually leads to an optimistically biased assessment.
d. Small differences in performance on the training data set versus the validation data set usually indicate overfitting.

Splitting the Data for model training and assessment

Which of the following statements is true regarding model assessment?

a. Data splitting can be used only on data with continuous targets.
b. The validation data set is used to calculate the parameter estimates and validate the model.
c. Assessing the performance of the model on the data that you used to fit the model usually leads to an optimistically biased assessment.
d. Small differences in performance on the training data set versus the validation data set usually indicate overfitting.

Answer *a* is **incorrect** because data splitting can be used on data with any type of target.
Answer *b* is **incorrect** because the validation data set is used to validate the model. The training data set is used to calculate the parameter estimates.
Answer *d* is **incorrect** because large differences in performance on the training data set versus the validation data set usually indicate overfitting.

Splitting the Data for model training and assessment

For the target marketing project at the bank, we want to split the develop data set into a training data set and a validation data set. In this demonstration, we do the following: Use a stratified sample to select the records for the training and validation data sets, and create two data sets: train and valid. Let's look at the code.

Develop data s

	y	X_1	X_2	$\cdots$	X_k
1	0	■	■	$\cdots$	■
2	1	■	■	$\cdots$	■
3	0	■	■	$\cdots$	■
4	1	■	■	$\cdots$	■
5	0	■	■	$\cdots$	■
6	1	■	■	$\cdots$	■
7	0	■	■	$\cdots$	■
8	0	■	■	$\cdots$	■
$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$
m	1	■	■	$\cdots$	■

Stratifie
d
random

Training data s

	y	X_1	X_2	$\cdots$	X_k
1	0	■	■	$\cdots$	■
2	1	■	■	$\cdots$	■
3	0	■	■	$\cdots$	■
$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$
n	1	■	■	$\cdots$	■

Validation data

	y	X_1	X_2	$\cdots$	X_k
1	0	■	■	$\cdots$	■
2	1	■	■	$\cdots$	■
3	0	■	■	$\cdots$	■
$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$
w	1	■	■	$\cdots$	■

Splitting the Data for model training and assessment

```
proc sort data=work.develop out=work.develop_sort;
   by ins;
run;

proc surveyselect noprint data=work.develop_sort
                  samprate=.6667 stratumseed=restore
                  out=work.develop_sample
                  seed=44444 outall;
   strata ins;
run;

proc freq data=work.develop_sample;
   tables ins*selected;
run;


data work.train(drop=selected SelectionProb SamplingWeight)
     work.valid(drop=selected SelectionProb SamplingWeight);
   set work.develop_sample;
   if selected then output work.train;
   else output work.valid;
run;
```

Splitting the Data for model training and assessment

What would happen if you split the data by taking a simple random sample in PROC SURVEYSELECT? Assume that, as in the previous demonstration, you split the data into two data sets (a training data set and a validation data set) and

a. The results would be the same as in the demonstration.
b. The proportion of the SELECTED=1 cases (cases in the training data set) would be different from the corresponding results in the demonstration.
c. The proportion of the events in the training data set would probably be different from the proportion of events in the validation data set.

Splitting the Data for model training and assessment

What would happen if you split the data by taking a simple random sample in PROC SURVEYSELECT? Assume that, as in the previous demonstration, you split the data into two data sets (a training data set and a validation data set) and

a. The results would be the same as in the demonstration.
b. The proportion of the SELECTED=1 cases (cases in the training data set) would be different from the corresponding results in the demonstration.
c. The proportion of the events in the training data set would probably be different from the proportion of events in the validation data set.

Unlike a stratified random sample, a simple random sample does not guarantee an equal percentage of events in the training and validation data sets. However, because the sampling rate is the same as in the demonstration (0.6667), the training data set (SELECTED=1) will contain 66.67 percent of the observations regardless of the sampling method.