

# Data Mining en Economía y Finanzas



UBA

Maestría en Explotación de datos y Descubrimiento del Conocimiento  
2019

# Tipos de Posgrados

CONEAU

# Doctorado

El doctorado tiene por objeto la formación de posgraduados que puedan lograr *aportes originales* en un area de conocimiento -cuya universalidad deben procurar-, dentro de un marco de excelencia académica, a traves de una formación que se centre fundamentalmente en torno a la investigación desde la que se procurará realizar dichos *aportes originales*.

# Maestría Académica

La maestría académica se vincula específicamente con la investigación en un campo del saber disciplinar o interdisciplinar.

A lo largo del desarrollo, profundiza tanto en temáticas afines al campo como en la metodología de la investigación y la producción de conocimiento en general y en dicho campo.

# Maestría Profesional



La maestría profesional se vincula específicamente con el fortalecimiento y consolidación de competencias propias de una profesión o un campo de actuación profesional.

A lo largo del proceso de formación profundiza en competencias en vinculación con marcos teóricos disciplinares o multidisciplinarios que amplían y cualifican las capacidades de desempeño en un campo de acción profesional o de varias profesiones.

# Tipos Maestrías Profesionales

- Especialistas 

- Generalistas

# Oferta Académica Argentina

Historia

Oferta  
Académica



# Oferta Académica, Historia

- 2004/5 - [UBA](#), Maestría en Explotación de Datos y Descubrimiento del Conocimiento
- 2006 - [Universidad Austral](#), Maestría en Explotación de Datos y Gestión del Conocimiento (Data Mining)
- 2008-2010 [Universidad di Bologna](#), Maestría en Investigación de Mercado y Data Mining
- 2014 - [ITBA](#), Diplomatura en Big Data

# Avances en imagenes, speech

# The Breakthrough

2012 Large Scale Visual Recognition Challenge

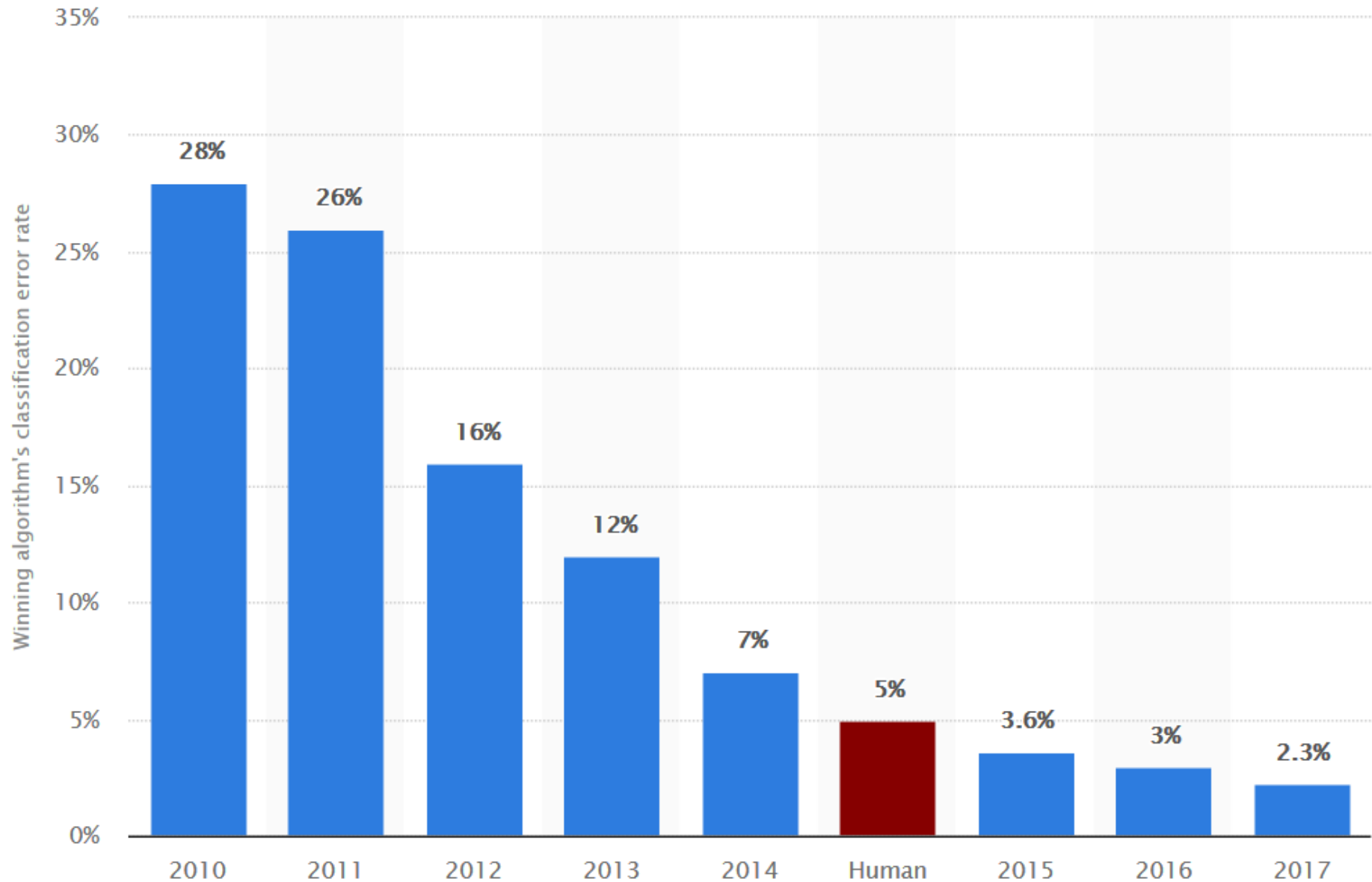
based on Imagenet repository containing millions of labeled images built in 2010

Alex Krizhevsky, Ilya Sutskever, and Geoff Hinton  
presented a solution based on  
GPUs x50 speedup

dropped the error rate from 26% to 16%

Supervised Learning

# ILSVRC Error Rate Evolution



# year 2013, 1 week to train a DNN

## GOOGLE DATACENTER



1,000 CPU Servers  
2,000 CPUs • 16,000 cores

**600 kWatts**  
**\$5,000,000**

## STANFORD AI LAB



3 GPU-Accelerated Servers  
12 GPUs • 18,432 cores

**4 kWatts**  
**\$33,000**

# The Digital Mammography DREAM Challenge



Out of every 1000 women screened, only 5 will have breast cancer. But 100 will be recalled for further testing.

We can do better.

Build a model to help reduce the recall rate for breast cancer screening.

**Calling all coders to join the Challenge.**

Up to a **\$1,000,000** in cash prizes for winning models.

**May the best model win.**



Altmetric: 21

[More detail >>](#)

Article | [OPEN](#)

# Detecting and classifying lesions in mammograms with Deep Learning

Dezső Ribli , Anna Horváth, Zsuzsa Unger, Péter Pollner & István Csabai

*Scientific Reports* **8**, Article number: 4165  
(2018)

doi:10.1038/s41598-018-22437-z

[Download Citation](#)

Breast cancer

Cancer imaging

Computer science

Image processing

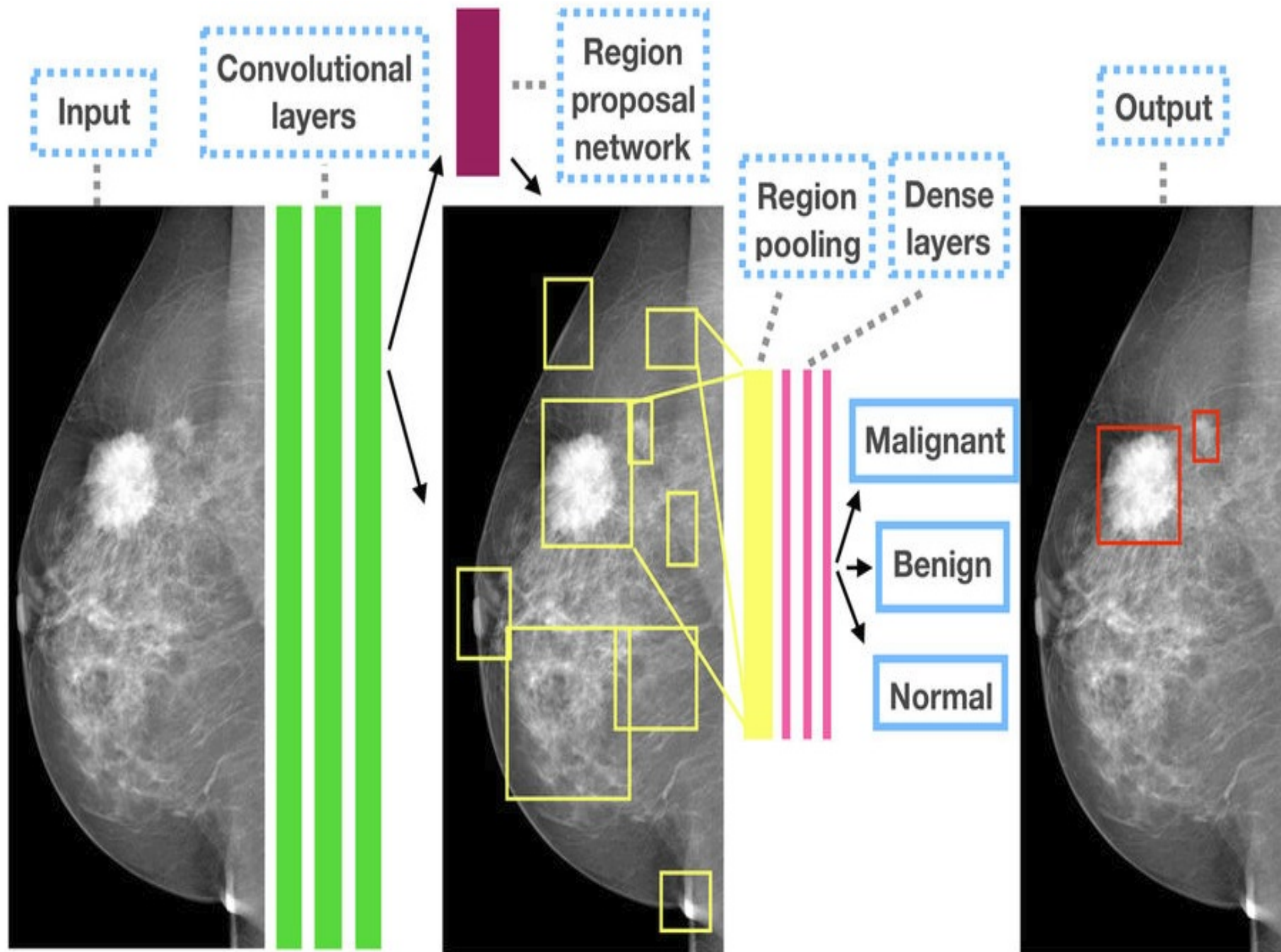
Radiography

Received: 17 November 2017

Accepted: 21 February 2018

Published online: 15 March 2018







Bronze sponsors



aidence

**behold.ai**

**DesAcc**  
Enterprise Data Discovery and Delivery

EverlightRadiology

**MEDICA**  
GROUP

Silver sponsors

**AGFA**   
HealthCare

**TMC**  
RADIOLOGYREPORTING

ARTIFICIAL INTELLIGENCE IN IMAGING -  
THREAT OR OPPORTUNITY?

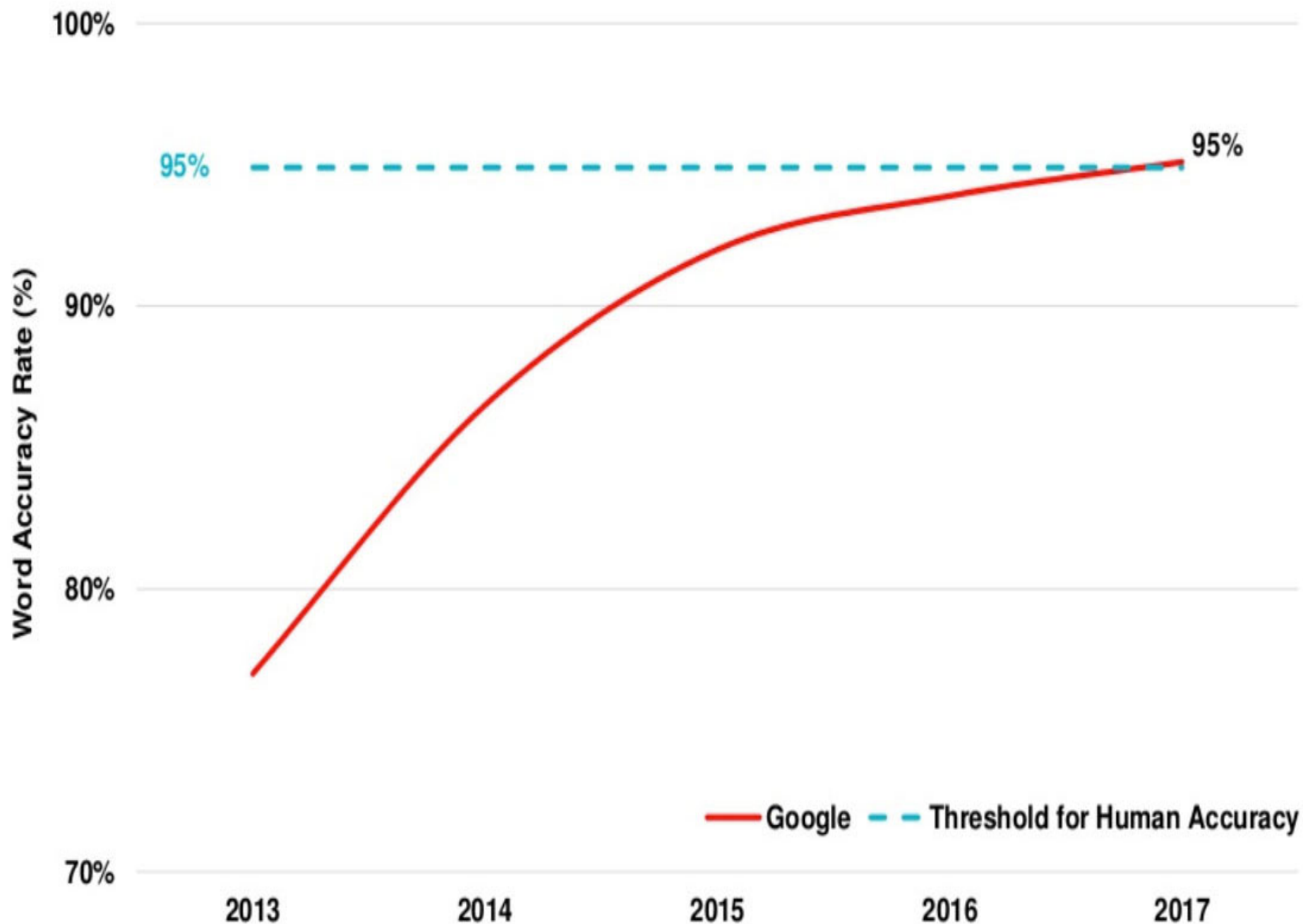
Venue: etc. venues St Paul's, London

CPD: 6 CREDITS



# Google Machine Learning

Achieving Higher Word Accuracy, 2013-2017





# Google Pixel Buds

\$159

Pre-order today

Outside of the US and Canada, this device is a prototype unit and cannot be marketed, sold, leased or distributed until it complies with applicable essential requirements and obtains required legal authorizations.

**G**  
#madebygoogle

Ok Google, will I need  
an umbrella today?



No, rain is not expected today.





Ok Google, find me a plumber.

OK, I can recommend some plumbers. To find a good fit, I need a few details. What do you need help with?

My drain is clogged.

OK. Is this for your home at 340 South Pearl Street?

Yes.

Sure, would you like to connect to a plumber who's free to call you now, or get a list of recommended plumbers?



# Amazon Staff Are Listening To Alexa Conversations -- Here's What To Do



**Kate O'Flaherty** Senior Contributor ①

[Cybersecurity](#)

*I'm a cybersecurity journalist.*



Amazon employs thousands of people around the world to listen to voice recordings captured in Echo users' homes and offices. An Amazon Echo multimedia smart speaker, taken on November 28, 2016. (Photo by Joby Sessions/T3 Magazine via Getty Images) GETTY

Over 100 million Amazon Echos have [been sold](#) as of the start of

## TECNO

# Científico de datos: por qué es la profesión más sexy del siglo XXI

En el mundo de la tecnología es el perfil que más se comenzó a demandar en el último tiempo. Cómo impacta el crecimiento de la automatización en este empleo y en qué tipo de industrias se los necesita

---

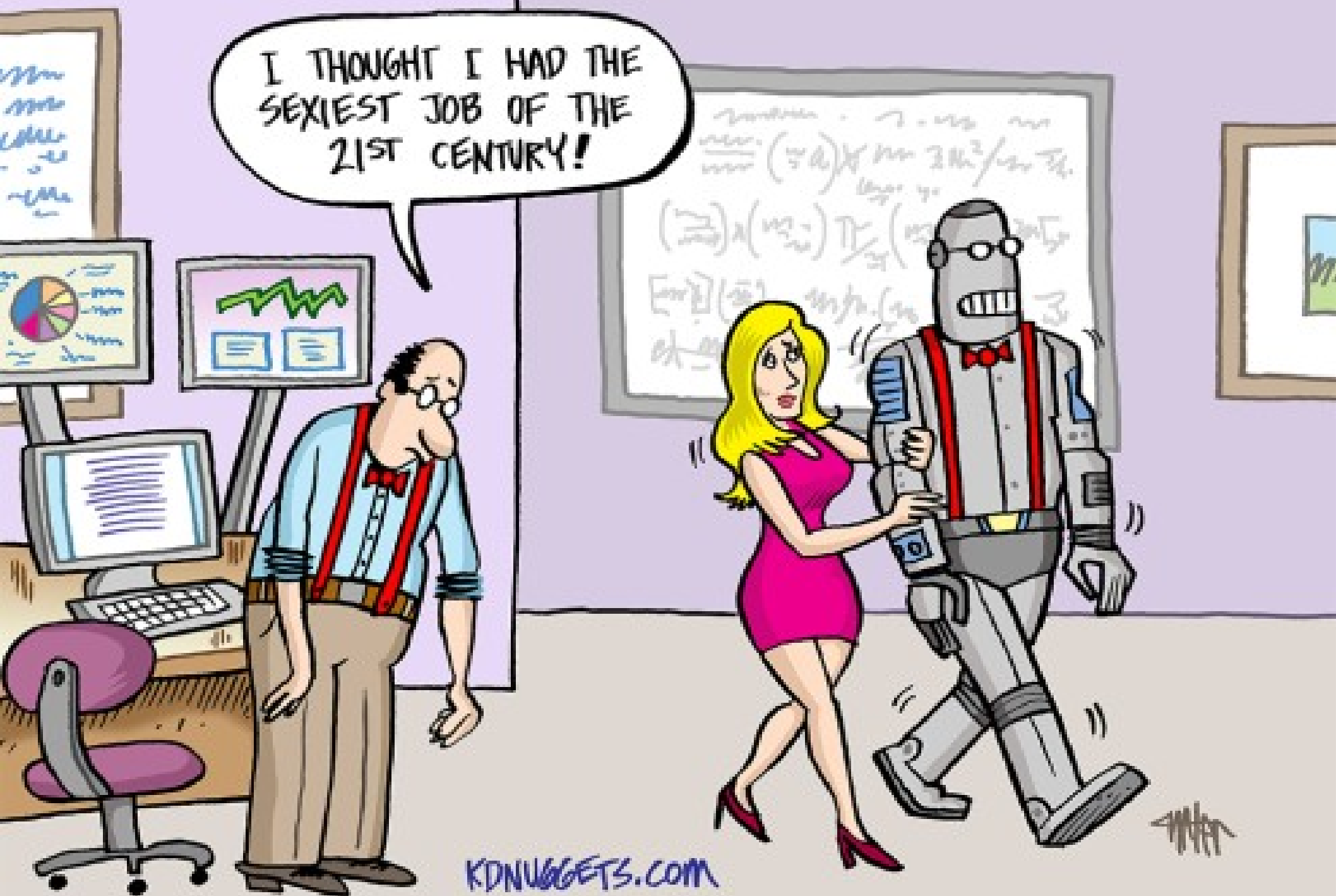


Por **Desirée Jaimovich** | 9 de mayo de 2017

[djaimovich@infobae.com](mailto:djaimovich@infobae.com)



I THOUGHT I HAD THE  
SEXIEST JOB OF THE  
21ST CENTURY!



# Oferta Académica

año 2019

# Maestrías y Especializaciones

- Maestría en Análisis y Gestión de Negocios, [Universidad Torcuato di Tella](#)
- Maestría en Explotación de Datos y Gestión del Conocimiento (Data Mining), [Universidad Austral](#)
- Maestría en Minería de Datos, [UTN Paraná](#)
- Especialización en Ciencia de Datos, [ITBA](#)
- Maestría en Inteligencia de Datos orientada a Big Data, [UNLP](#)

# Maestrías y Especializaciones

- Especialización en Métodos Cuantitativos para la Gestión y Análisis de Datos en Organizaciones, [UBA Económicas](#)
- Maestría en Explotación de Datos y Descubrimiento del Conocimiento, [UBA Exactas - Ingeniería](#)



YOU  
ARE  
HERE

# Diplomaturas

- Diplomatura en Big Data & Business Analytics, [Universidad Siglo 21](#)
- Diplomatura en Business Intelligence, [UTN Buenos Aires](#)
- Data Mining y Big Data, [UCA](#)
- Big Data y Analytics, [Universidad de Palermo](#)
- Diplomatura en Big Data, [ITBA](#)

# Diplomaturas

- Fundamentos de Métodos Analíticos Predictivos, [UBA Económicas](#)
- Análisis de Datos para Negocios, Finanzas e Investigación de Mercado, [UAI](#)

# Oferta Académica, UTN

- Experto Universitario en Estadística Aplicada a la toma de decisiones
- Introducción a Probabilidades y sus Distribuciones para la toma de Decisiones
- 1° Nivel de Profundización: “Introducción al Muestreo y a la Estadística Inferencial para la Toma de Decisiones”
- 2° Nivel de Profundización: “Introducción a Regresión Lineal y a Estadística no Paramétrica”

# Oferta Académica, UTN

- Diplomatura en Análisis de Negocios (Business Analysis)
- Especialista en Big Data con Apache Hadoop
- Diplomatura en Análisis de Negocios (Business Analysis)
- Introducción a Probabilidades y sus Distribuciones para la Toma de Decisiones



# Oferta Académica, Digital House

- Data Analytics Immersion programa ejecutivo
- Data Analytics curso
- Data Science curso
- Inteligencia Artificial curso

# Oferta Académica, EANT

- Data Scientist 135hs
- Data Analytics 96hs
- Big Data Analytics 60hs
- Machine Learning 72hs
- Big, Small & Open Data Analytics 30hs
- Python para Ciencias de Datos 30hs

# Oferta Académica, Math

- Maestría en Generación y Análisis de Información Estadística [UNTREF](#)
- Maestría en Estadística Aplicada [Universidad Nacional de Tucumán](#)
- Maestría en Estadística Aplicada [Universidad Nacional de Córdoba](#)
- Maestría en Estadística Matemática [UBA](#)
- Maestría en Estadística Aplicada [Universidad Nacional de Rosario](#)

# Oferta Académica, GRADO

- año 2019 Licenciatura en Analítica Empresarial y Social ITBA

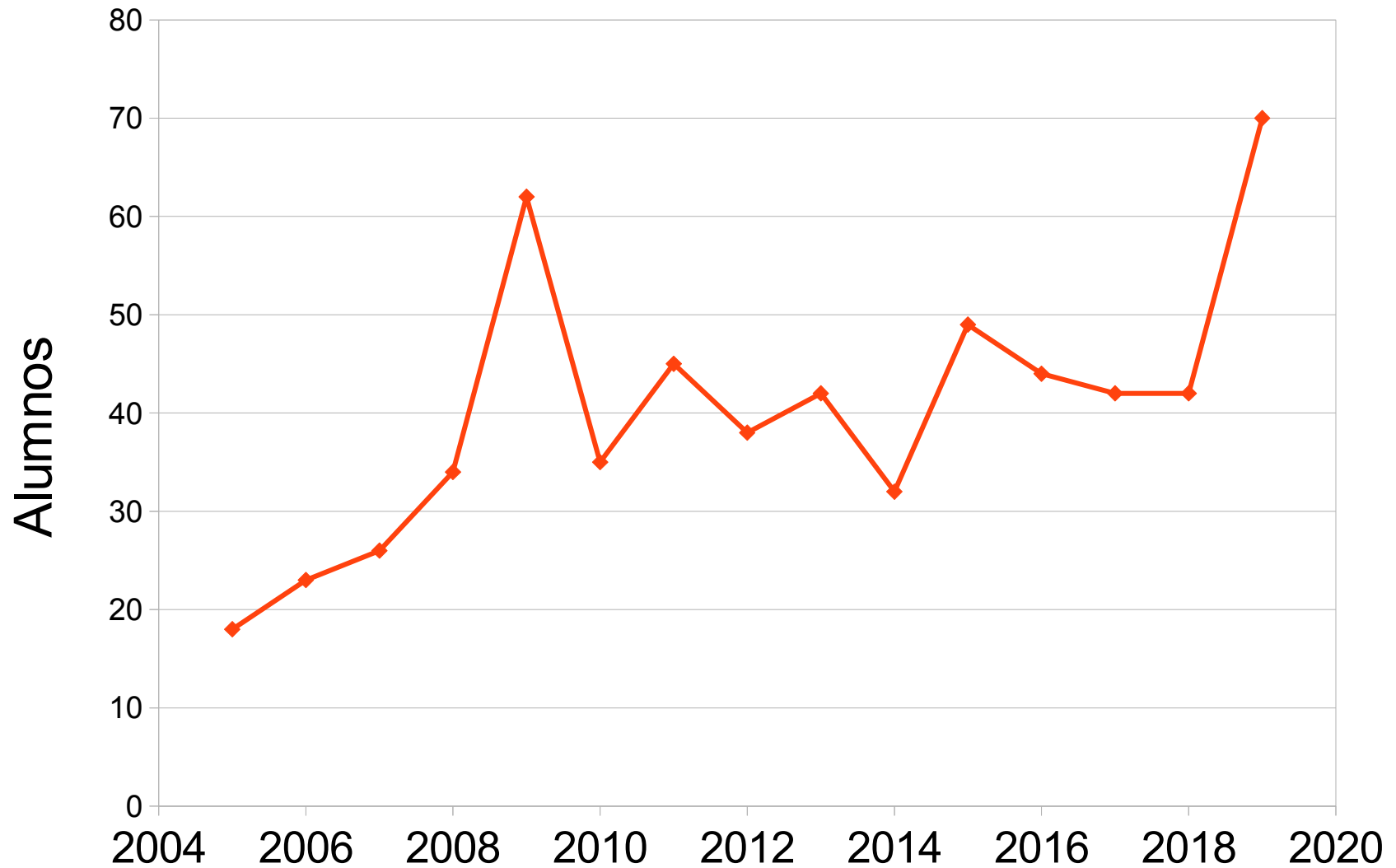
Formación en gestión y análisis de datos para la dirección estratégica de empresas y organizaciones. Una carrera con foco en el análisis de los datos, con el fin de extraer conclusiones para la toma de decisiones.

El Licenciado en Analítica Empresarial y Social, tendrá la capacidad de comprender tendencias económicas, sociales y culturales para implementar soluciones que mejoren la calidad de vida de la sociedad.

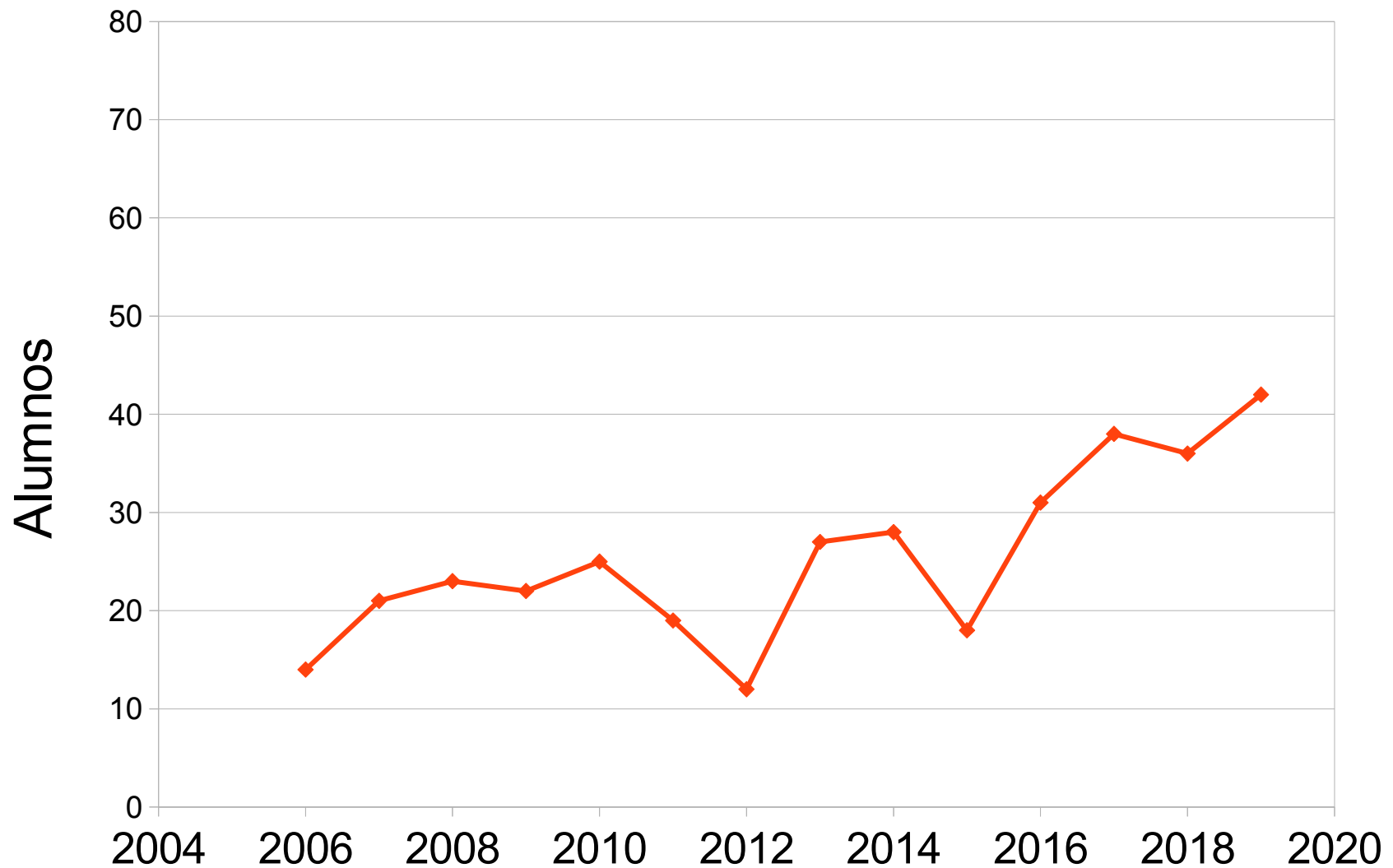
Historia  
Maestría  
Explotación de Datos  
y  
Descubrimiento del Conocimiento

UBA

# UBA Evolucion Cantidad Alumnos



# U. Austral Evolucion Cantidad Alumnos



# Una carrera de la UBA con trabajo asegurado

Genera gran demanda y buenos sueldos

Nora Bär

SEGUIR

29 de enero de 2008



Comentar  
(62)



Me gusta



Compartir

**S**andra Zabala tiene 37 años, es licenciada en física y trabaja en una agencia de medios. Juan José Lloret tiene 60, es matemático y tiene una consultoría de marketing. Gustavo Markel tiene 30, es ingeniero en sistemas y trabaja en un banco.



“Estaba trabajando en una empresa, pero el sueldo era paupérrimo (cuenta Sandra Zabala). Ahora soy directora de investigación y desarrollo de una agencia de medios.

En lo económico resultó mejor de lo que había pensado. Multipliqué mi sueldo por 20 o más.”

# Hitos

- 2005 primera camada regular
- 2010 efecto “multipliqué mi sueldo x20”
- 2014 heterogeneidad alumnos ↑
- 2015 último año sin curso nivelación
- 2016 curso nivelación + 2 comisiones
- 2018 uniformidad lenguaje R 1er año,  
Python 2do año

## SOCIEDAD

# Publicó su currículum en Twitter y ya coordinó nueve entrevistas de trabajo

Nicolás Abuchar publicó sus datos en su cuenta y recibió casi 50 propuestas. "Voy a aprovechar y disfrutar lo que viene. Para eso hice todo esto", dijo a Infobae

---



**Nico**

@nico\_\_ab

Seguir



Soy economista y busco trabajo:

-Cursando Maestría en Data Mining (UBA)

-Licenciado en Economía (UBA)

Experiencia:

-Investigación y consultoría económica.

Lenguajes:

-R

-SQL

Intereses: finanzas, riesgo, scoring, BI. Busco mucha data y análisis. Gracias!

nicolas.abuchar@gmail.com

13:06 - 16 oct. 2018

639 Retweets 419 Me gusta



47

639

419

# Heterogeneidad Alumnos

5% Matemática/Estadística

5% Actuario

**29% Sistemas**

**33% Economía (Contador, Administracion )**

**16% Ingenieros**

**8% Ing. Electrónicos**

6% Psicólogo/Lic Ciencias Políticas/Farmacéutico

# Heterogeneidad Alumnos

El 10% está cursando o planea cursar un doctorado.

El 25% quisiera que la maestría de la UBA fuera académica, más rigurosa, más científica, demostraciones de teoremas.

El 35% quiere una maestría más práctica, en donde se enseñen herramientas con salida laboral para progresar jerárquicamente.

# Heterogeneidad Alumnos

5% hace 2+ años que ya trabaja con R

10% trabaja en Data Science con R

20% ya trabaja en algun grado de Data Science

70% jamas vio R antes de la maestría

50% saben programar, pero no R, ni data mining

20% no sabe ni programar, ni R, ni data mining, ni trabaja en nada parecido ni en su empresa se hace nada similar. ( y por eso vino a aprender ! )

# Heterogeneidad Alumnos

Menos del 5% de los alumnos que terminan el primer año completan la tesis y obtienen el título de Magister.  
... sin embargo hay otras formas de ver la tesis ...

Menos del 30% de los alumnos estará trabajando en Data Science en los proximos 5 años. Esto se cumple aún para los sub-35



# Homogeneidad Alumnos

Tan solo 1 de 60 alumnos le interesa ser docente de la maestría .

# Heterogeneidad Alumnos

Algunos le dedican 40 horas extra cursada, otros más de 200 porque “se obsesionan con la competencia”, a lo largo de los 4 meses de la materia.

El 70% de los alumnos (equipos) jamas envia un email con consultas.

El 5% de los alumnos envia mas de 25 emails con consultas.

Lo que los alumnos creen saber  
Lo que realmente saben  
Lo que desean aprender

La que busca  
el  
mercado laboral  
argentino

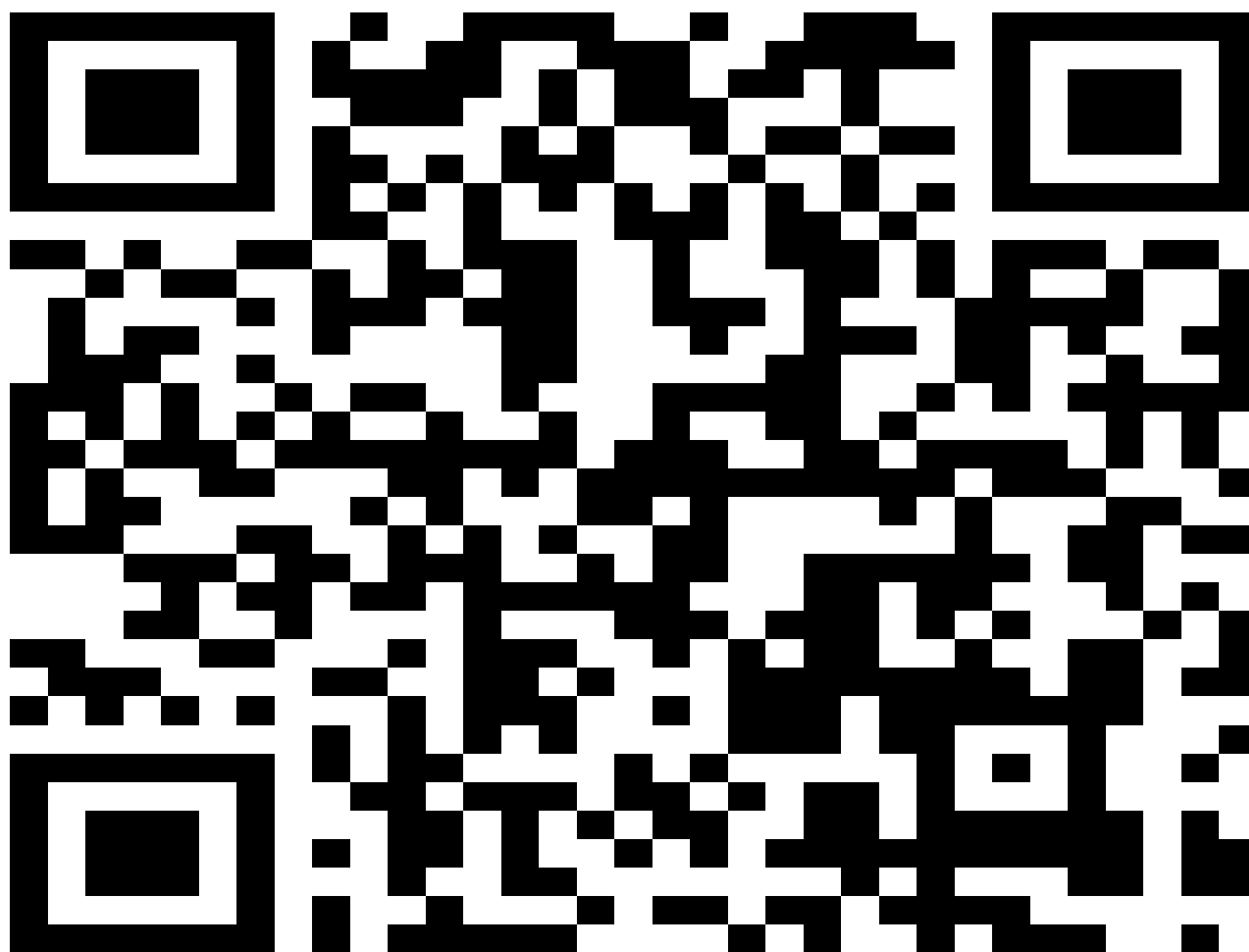
- Sentido Común
  - Negocio
  - Urgencia del negocio
  - Trabajo en equipo
  - “Storytelling”
  - Venta interna/externa de proyectos
- 
- Programar en R/Python



A close-up photograph of a computer keyboard. The central focus is a large, rectangular, light blue key. On this key, there is a white icon of an airplane in flight, angled upwards and to the right. Below the icon, the words "check-in" are printed in a white, lowercase, sans-serif font. The key is surrounded by other standard white keyboard keys, including one with a "K" and another with a small upward-pointing triangle. The background is a blurred grey keyboard frame.

# La Materia

# Generalidades






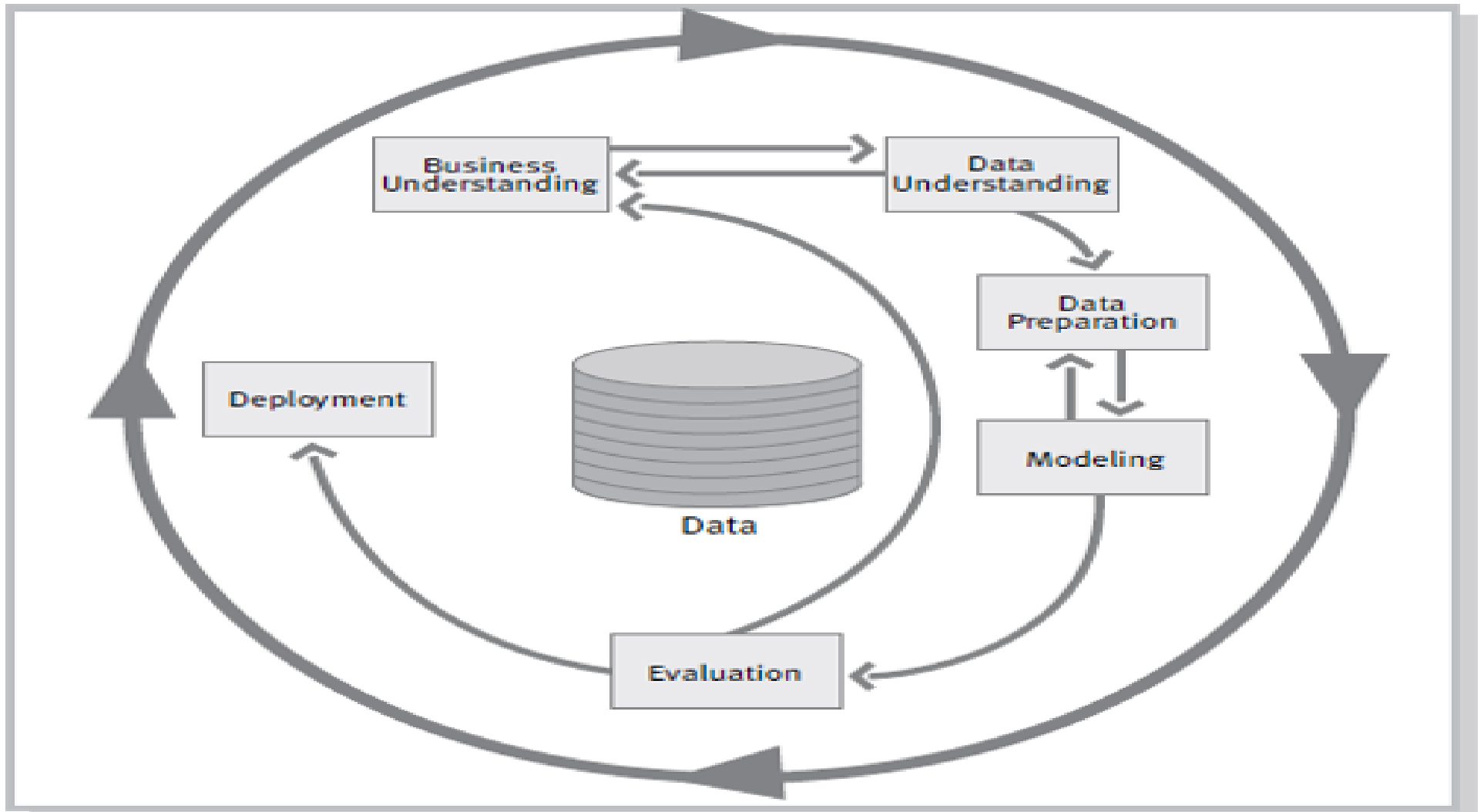
Se presentará un caso de estudio con dataset de clientes de empresa financiera .

Se guiará a los alumnos a desarrollar un modelo predictivo de attrition que sea competitivo con los de nivel profesional existentes en el mercado argentino.

# Data Science Tasks

- Description
- Prediction
  - Structured Data 
  - Unstructured Data
- Causal Inference

# Nuestra GRAN limitación



# Fechas Importantes

<b>03-dic-2019</b>	<b>Entrega IDs predicción materia</b>
<b>05-dic-2019</b>	<b>Examen individual escrito</b>
<b>12-dic-2019</b>	Ultima Clase conclusiones
<b>16-dic-2019</b>	Recuperatorio Datasets
<b>31-mar-2020</b>	Entrega Ids recuperatorio

# Modalidad de Dictado

Se hará el planteo del problema entregando el dataset y comentando en gran detalle su historia y las motivaciones, precisando claramente el objetivo a resolver.

Luego se acompañará a los alumnos en la resolución, presentando los elementos teóricos y conceptuales en el justo tiempo que van siendo requeridos para solucionar las distintas partes del problema.

# Puntos destacables

- Es una competencia, se debe maximizar la ganancia económica de una campaña preventiva de retención
- Se trabaja con datos reales, se dispone de 36 meses de historia de mas de 180k clientes, con 170 atributos.
- Se procesará en la nube
- Se trabaja en lenguaje R
- Se utilizarán los algoritmos XGBoost y LightGBM **el estado del arte**
- Los alumnos *dicen* dedicarle 80+ horas extra clase

# Puntos destacables

La entrega es una lista de IDs de clientes

La corrección de la competencia es 100% objetiva, consiste en calcular cuanto dinero obtiene esa lista de IDs

# Esta materia **NO** se trata de :

- Stock Price Prediction
- Social Network Analytics, Computational Propaganda, Cambridge Analytica
- Time Series Forecasting
- SVM, Logistic Regression



# Esta materia **NO** se trata de :

- Deep Learning
- Image Recognition
- Speech Recognition Alexa, Siri
- Chatbots
- SAS, Be Smart, IBM Watson, Anaconda, Hadoop

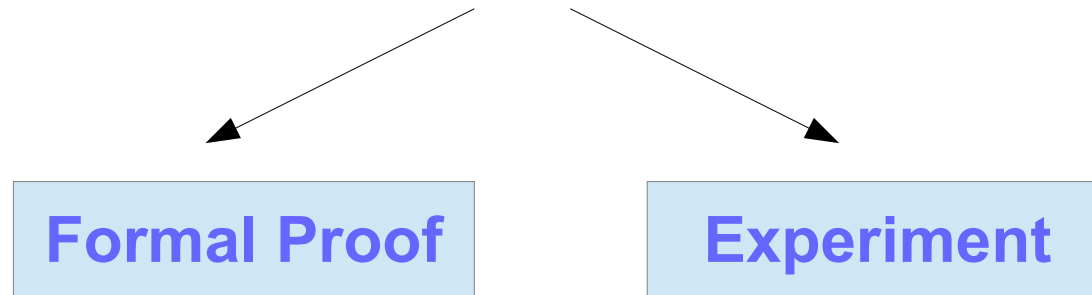
# Lamentavelmente **NO** vemos:

- IoT
- Reinforcement Learning, Alpha GO , Alpha Zero
- Causal Data Science

# Filosofía y Ética

Gustavo Denicolay

# In God we trust, all others must bring data.



W. Edwards Deming

# Science Enemy

Let us return to what was, and ever should be...  
the office of this abbey:  
The preservation of knowledge.  
Preservation, I say.

Not search for...  
because there is no progress in the history of  
knowledge...merely a continuous and *sublime*  
*recapitulation*.

The name of the Rose, Umberto Eco

Don't believe in something  
just because it is trendy.

You get more prestige by doing good  
science rather than by doing popular  
science.

Try to become expert in one topic  
rather than learning many things  
partially.

Donald Erving Knuth

Errare Humanum  
est  
Perseverare  
Diabolicum

The art of teaching is the  
art of assisting discovery .

Mark Van Doren



# **COGNITIVE DISSONANCE**

**THIS IS WHY PEOPLE GET UPSET WHEN  
THEIR BELIEFS ARE CHALLENGED**



**A MENTAL CONFLICT OCCURS WHEN BELIEFS ARE CONTRADICTED BY NEW INFORMATION. THIS CONFLICT ACTIVATES AREAS OF THE BRAIN INVOLVED IN PERSONAL IDENTITY AND EMOTIONAL RESPONSE TO THREATS. THE BRAIN'S ALARMS GO OFF WHEN A PERSON FEELS THREATENED ON A DEEPLY PERSONAL AND EMOTIONAL LEVEL CAUSING THEM TO SHUT DOWN AND DISREGARD ANY RATIONAL EVIDENCE THAT CONTRADICTS WHAT THEY PREVIOUSLY REGARDED AS 'TRUTH'**

Un docente es alguien que *inspira*  
a que el otro se transforme.

Dario Sztajnszrajber

How, when and why students learn meaningfully, or just regurgitate facts and deploy procedures and algorithms (or possibly don't manage even those).

We only think deeply about things we care about.

**Emotion is the rudder of thought.**

Meaningful learning is actually about helping students to connect their isolated algorithmic skills to abstract, *intrinsically emotional*, subjective and meaningful experiences.

Mary Helen Immordino-Yang

The important thing is not triumph,  
but the struggle . ( Pierre de Coubertin )

Struggle.verb /'strʌg.əl/  
to experience difficulty and make a very great  
effort in order to do something

I *am here* to push people beyond  
what's expected of them.  
I believe that is an absolute  
necessity.

Terence Fletcher  
Whiplash

# Free Speech Statement, April 2018

The free and open exchange of ideas and information is fundamental to the educational mission of *Association of American Universities*. The robust discussions and debates that occur at research universities have been central to the advancement of democracy, the creation of new knowledge, the fostering of educational excellence, and the promotion of social progress. As heads of these institutions we are unequivocally committed to preserving and honoring this proud heritage.

While we may deem some speech to be odious, disgraceful, and antithetical to our values, our campuses are and should remain places where ideas can be expressed free of disruption, intimidation, and violence.

Few academics challenge censorship that emerges from students. It is important that more do, because a culture that restricts the free exchange of ideas encourages self-censorship and leaves people afraid to express their views in case they may be misinterpreted. This risks destroying the very fabric of democracy.

An open and democratic society requires people to have the courage to argue against ideas they disagree with or even find offensive. At the moment there is a real risk that students are not given opportunities to engage in such debate.

A generation of students is being denied the opportunity to test their opinions against the views of those they don't agree with.

# Tacticas de la batalla



# Los ejes de avance

- Classifier Learning
- Model Comparison
- Hyperparameter tuning
- Fast and Scalable Processing
- Dataset Engineering
- Meta classifiers

# Classifier Learning

- ~~Decision~~ Probability Trees

- Rpart

- Ensembles

- Bagging , Random Forests

- Ranger

- Boosting

- XGBoost, LigthGBM

# Classifier Learning

Deep Learning and  
Time Series  
are NOT covered in this  
course

# Hyperparameter Tunning

- Grid Search
- Bayesian Optimization
  - `mlrMBO`

# Model Comparison

- Accuracy
- ROC Curve, AUC
- Custom Metrics

# Model Comparison

- Holdout Sampling (training – testing )
- Montecarlo Validation ( repeated random sub-sampling validation )
- k-fold Cross Validation
- Training – Testing in different datasets

# Fast and Scalable Processing

- Local CPU Single Core
- Local CPU Multi Core
- Cloud Computing, CPU Multiple Core
  - Google Cloud Computing Engine
- Cloud Computing, GPU
- Cloud Multinode – CPU Multicore
- TPU

# Dataset Engineering

- Class engineering
- New variables (intra month)
  - `dplyr`
- New historical variables (inter month)
  - `dplyr`, `Rcpp`



# Meta Classifiers

- Ensembles
- Stacking
- Bayesian Model Combination

# Aclaraciones Finales

# Aclaraciones

Esto no es un resolver un ejercicio en donde siguiendo un numero finito de pasos provistos por el profesor se obtiene el óptimo; esto es resolver un **problema**.

Se espera que los equipos **experimenten** formas alternativas a las que se muestran en clase de encarar el problema, logrando los mas sagaces obtener mayores ganancias.

# Aclaraciones

Esta materia demanda al alumno dedicarle una gran cantidad de horas extra a las horas de clase presenciales.

# Aclaraciones

Se sugiere fuertemente que los alumnos lean la bibliografía obligatoria hasta comprenderla íntegramente, comenzando el primer día de clase.

Esta bibliografía es lo elemental para poder comprender conceptos más complejos.

Realmente los alumnos deben lograr entenderla.

# Aclaraciones

El código R que se entrega es una simple guía, es indispensable para lograr un verdadero aprendizaje que los alumnos desarrollen su propio código.

Con el fin lograr que los alumnos hagan suyo el código, se presenta el código R en complejidad creciente.

Los alumnos deben aprender a leer la documentación de las funciones de las librerías que se utilizan; la disciplina es muy dinámica y estamos en una explosión cámbica.

# Solicitud a Alumnos

Si el código R provisto no le es de utilidad, hágaselo saber cuanto antes al profesor, que lo guiará para que usted aprenda a desarrollar su propio código desde cero.

# Recomendación

Antes de aprender como funciona Gradient Boosting, debe ser **completamente experto** en como funciona un arbol de probabilidad (decisión).

Aproveche las primeras semanas del curso, aprender como se elige el mejor corte en un nodo de un arbol es **el** concepto fundamental de un patrón en data mining.



# Ciencia

Si usted posee conocimientos previos y con su técnica obtiene mejores resultados, aplíquela y estará en el podio de ganadores.

Si su técnica obtiene peores resultados, no suponga que al resto le va mejor solo por azar o que está frente a una paradoja, diseñe experimentos para entender el porqué la nueva funciona mejor, bajo que situaciones. Evite ser prisionero de sesgos cognitivos.

**FIN**