



ESCUELA DE INGENIERÍA
FACULTAD DE INGENIERÍA

EDUCACIÓN
PROFESIONAL

Diplomado en Big Data y Ciencia de Datos
Ciencia de Datos y sus Aplicaciones

Clase 01: Introducción y Motivación

Roberto González



regonzar@uc.cl





Objetivos

- Descubrir y evaluar oportunidades de negocio accionables a partir del análisis de datos.
- Entender y aplicar los conceptos y métodos fundamentales de Data Science a problemas reales de negocio, interpretando adecuadamente los resultados y generando acciones de valor agregado.
- Desarrollar modelos descriptivos, predictivos y prescriptivos sobre datos de diversas industrias.



Evaluación

- Evaluaciones:
 - Se realizarán **2 trabajos** en **grupo**
 - Modelo de Predicción de Fuga de Clientes
 - Modelo de Segmentación
 - Se realizarán **3 ensayos o tareas en grupos de 2 personas**
- Nota Final del Curso:
 - 70% Promedio Trabajos + 30% Promedio Ensayos/tareas



Ayudante Martin Castillo, mecastillo2@uc.cl



Evaluación - Detalle



Consideraciones Grupos de Trabajo

- Los grupos se conformarán por 5 personas máximo
- Pueden cambiar de grupo para el trabajo 2
- Todos los integrantes del grupo recibirán la misma calificación

Consideraciones Ensayos/Tareas

- El tema de los ensayos se comunicará con 1 semana de anticipación a su entrega
- Los ensayos deben ser entregados por un solo miembro del equipo de 2 personas con un email a mecastillo2@uc.cl en la fecha estipulada, se penalizará la entrega atrasada
- Los ensayos deben ser breves y precisos con un mínimo de 2 planas de texto (2000 o más caracteres) y un máximo de 4 planas de texto (no mas de 4000 caracteres en total). Se penalizaran documentos fuera del rango de caracteres.
- Se permiten imágenes y diagramas.
- El nombre del archivo entregado debe contener los nombres de ambos miembros del equipo.



Planificación

Clase	Temas
1	Clases y evaluaciones Introducción y Motivación.
2	Discusión metodologías: CRISP y TDSP Data Science en Chile.
3	Detección anomalías.
4	Industria datos/financiera. Modelos de fuga de clientes, marketing.
5	Modelos Predictivos. Trabajo Proyecto Grupal 1, predicción de fuga.
6	Presentación alumnos de proyecto predicción de fuga Cierre análisis del proyecto.
7	Segmentación de clientes.
8	Data Science en Chile y Tendencias Trabajo Proyecto Grupal 2.



Motivación e Introducción

CONTEXTO



Datos 2022

<https://www.internetlivestats.com/>



5,392,200,999

Internet Users in the world



1,972,205,210

Total number of Websites



262,122,758,689

Emails sent **today**



8,311,001,613

Google searches **today**



8,095,205

Blog posts written **today**



804,432,258

Tweets sent **today**



7,700,825,371

Videos viewed **today**
on YouTube



93,735,375

Photos uploaded **today**
on Instagram



170,176,213

Tumblr posts **today**

Un mundo conectado...



Big Data

“

Big data es alto volumen, alta-velocidad y alta-variedad de componentes de datos que demandan herramientas de procesamiento de información **costo-efectivas**, que habiliten **mayor inteligencia, toma de decisiones y automatización de procesos.** ”

- Gartner, Big Data Definition (*)

(*) Big Data (Gartner, 2016), <http://www.gartner.com/it-glossary/big-data/>



Big Data - 4V





Motivación al Análisis

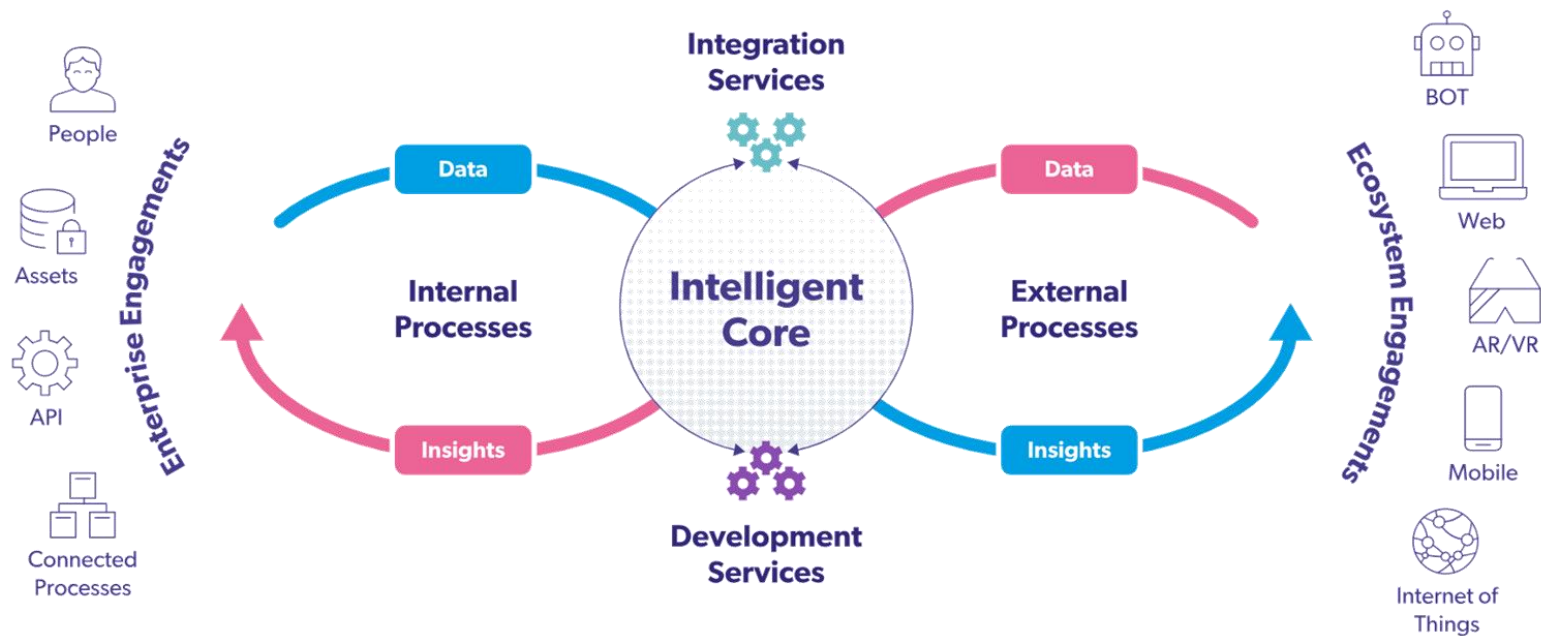
“Big data is not about the data!”

Gary King, Harvard University
Noviembre 2016

La verdadera revolución no es sobre los datos, sino sobre el impresionante progreso en la estadística y otros métodos para extraer ideas (insights) desde los datos.

Transformación basada en inteligencia

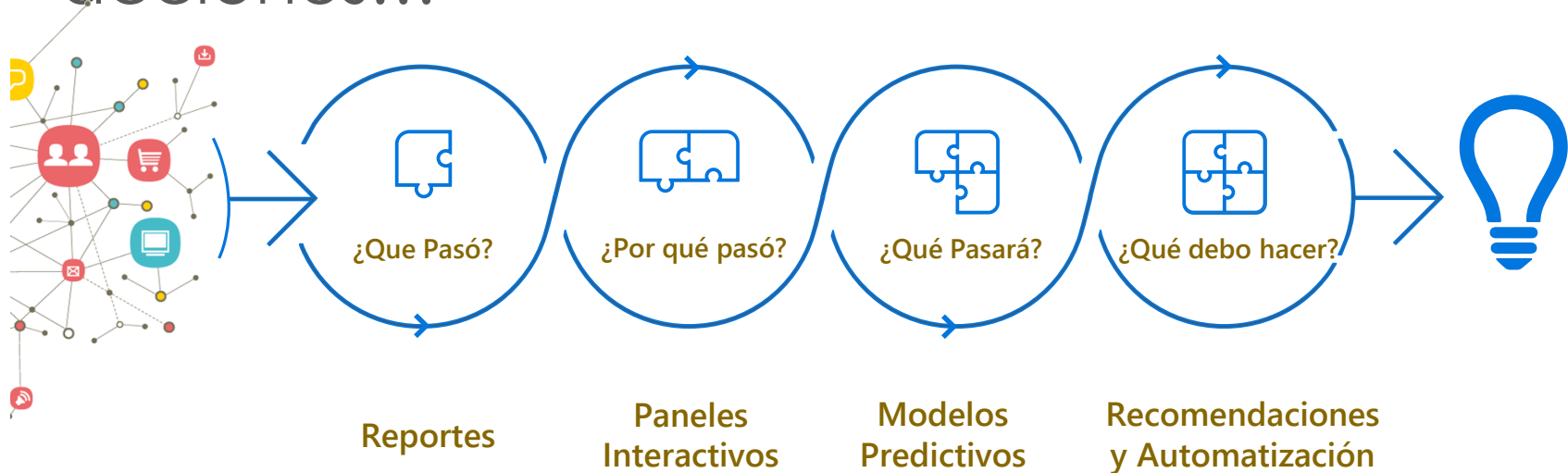
... Los datos no son la parte más importante



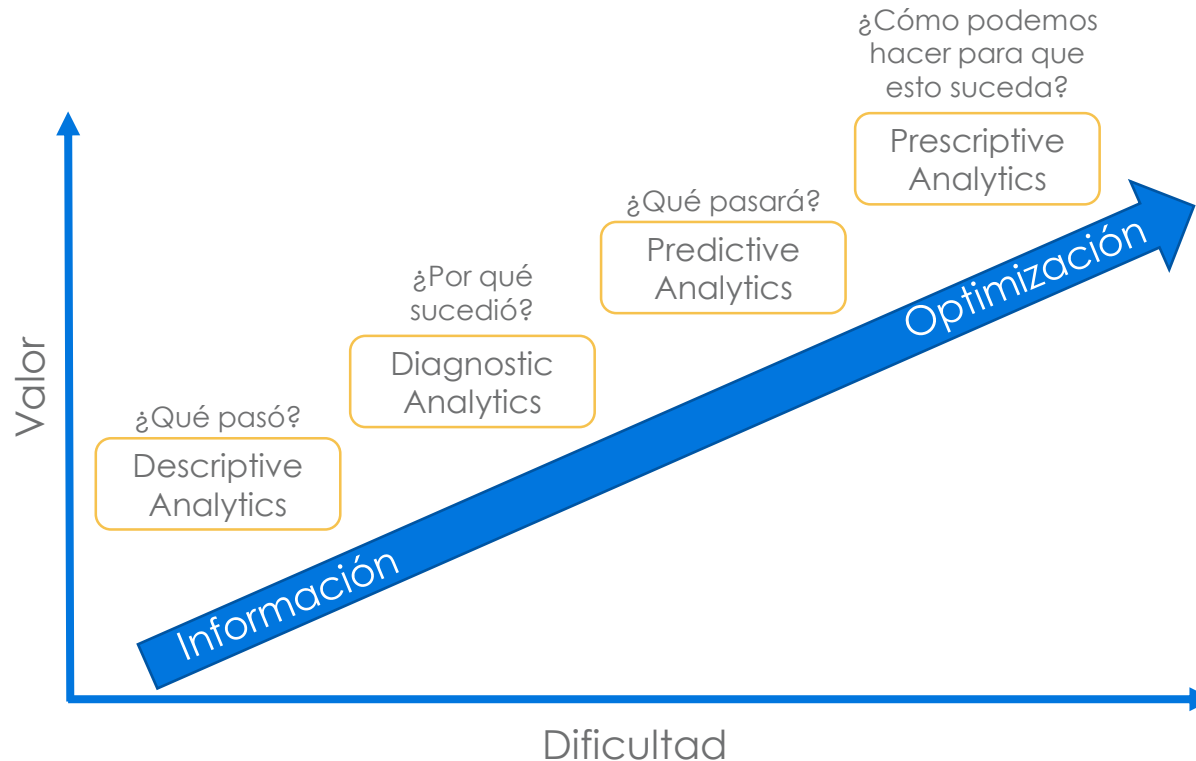
Fuente: IDC, Noviembre 2017

Big Data Analytics

Desde los datos tomar decisiones y acciones...



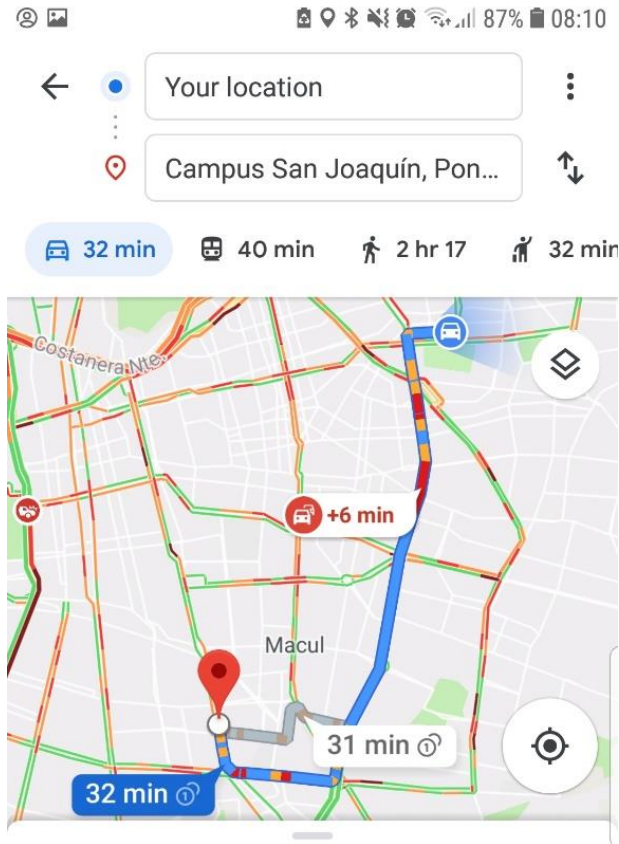
La evolución de los temas analíticos



Fuente: Gartner Business Intelligence & Analytics Summit 2013

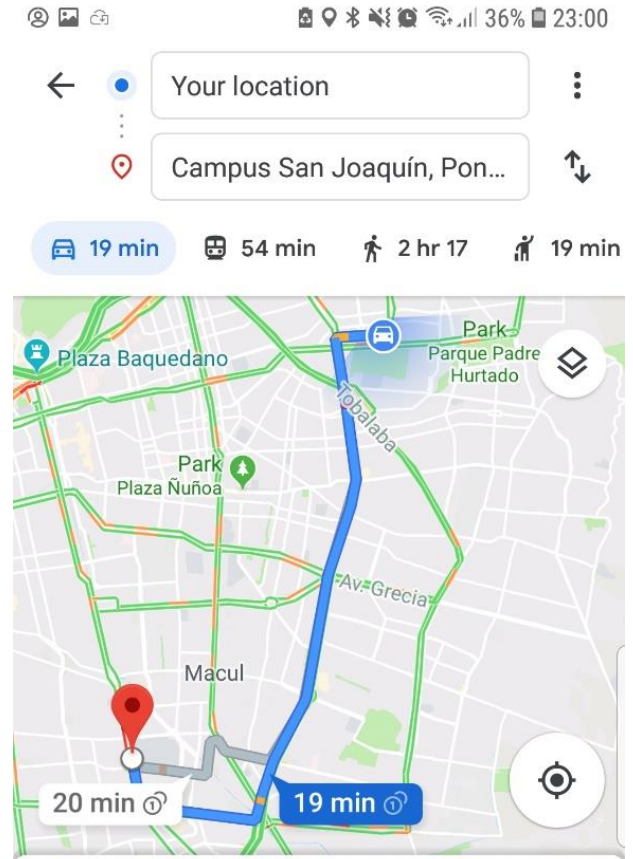


Tipos de Análisis



32 min (13 km)

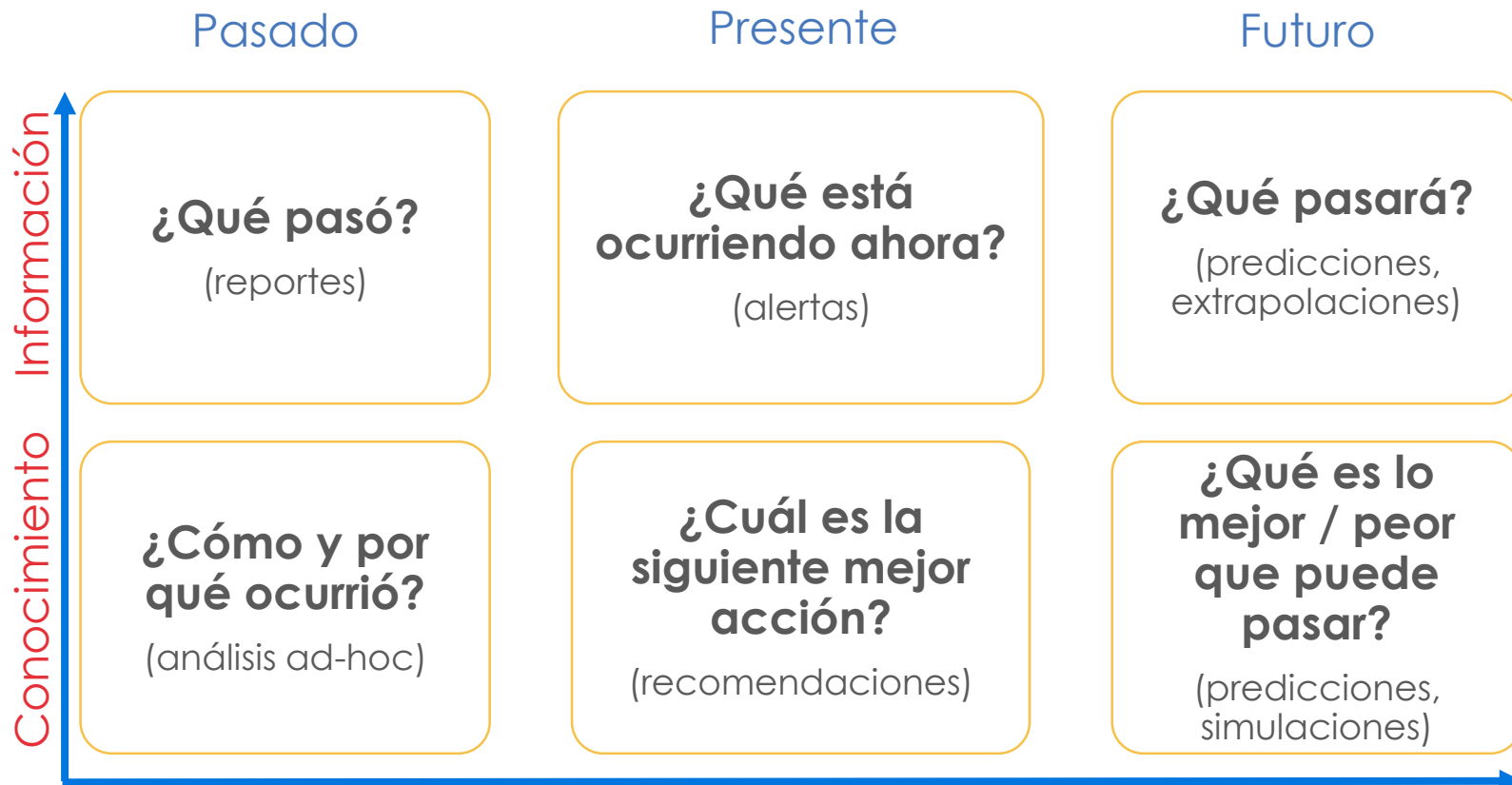
Best route, despite heavy traffic, but it's getting better



19 min (13 km)

Fastest route, the usual traffic

Preguntas que pueden abordarse



Fuente: Analytics at Work, Davenport, Harris & Morison, 2010



Motivación e Introducción

¿QUÉ HA PASADO EN LOS ÚLTIMOS AÑOS?

Google DeepMind Challenge Match



The image shows a Go match between AlphaGo and Lee Sedol. AlphaGo is represented by a blue robot-like figure, and Lee Sedol is a human player in a black suit. They are seated at a table with a Go board. In the background, there are other people and a large screen displaying the match. The text 'Google DeepMind Challenge Match 8 - 15 March 2016' is visible on the wall.

AlphaGo
vs
Lee Sedol*

AlphaGo Lee ganó 4-1

- ▶ AlphaGo: una combinación de redes neuronales y algoritmos.
- ▶ Entre el 9 y 15 de Marzo del 2016.

(*) 18 veces campeón mundial de Go.

Future of Go Summit, Wuzhen, China



AlphaGo

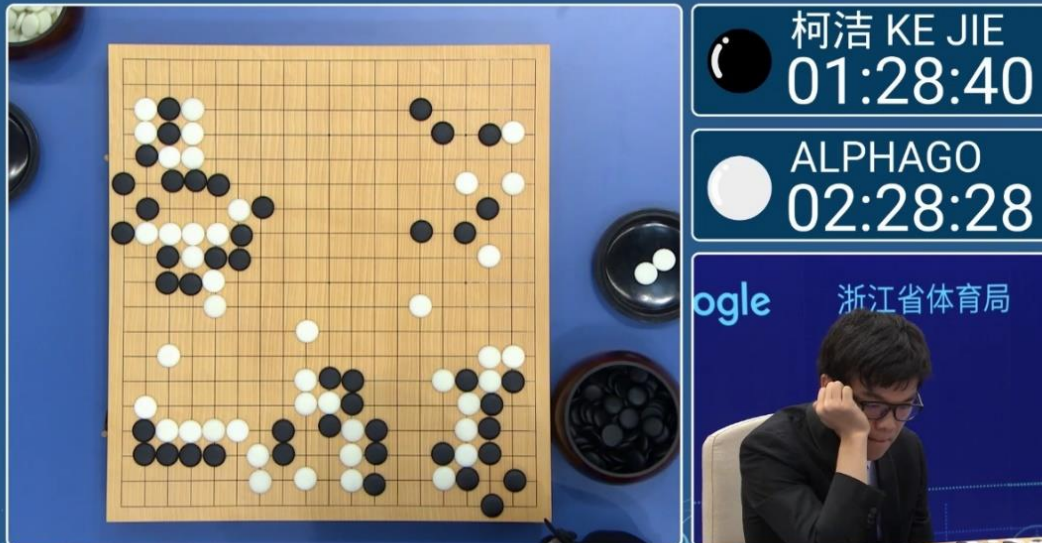
vs

Ke Jie*

**AlphaGo Lee Master
ganó 3-0**

- ▶ Google anunció luego de esto que el equipo de investigación de AlphaGo se enfocaría en abordar problemas más complejos y que afectan al mundo.
- ▶ 23, 25 y 27 de Mayo del 2017.

(*) Jugador nro. 1 del ranking mundial de Go.



AlphaGo Zero





¿Qué ha sucedido últimamente?



96%

RESNET vision test
152 layers

2016
Object
recognition
Human parity



94.9%

Switchboard speech
recognition test

2017
Speech
recognition
Human parity



88.5%

SQuAD reading
comprehension test

Enero 2018
Machine reading
comprehension
Human parity



69.9%

MT research
system

Marzo 2018
Machine
translation
Human parity



98.6%

Neutral TTS MOS
similarity to human
recording

Agosto 2018
Speech synthesis
near-human parity



90.7%

Conversational Question Answering System CoQA

Abril 2020
General Language Understanding - Human parity



¿Qué ha sucedido últimamente?



AI equal with human experts in medical diagnosis, study finds

Research suggests AI able to interpret medical images using deep learning algorithm



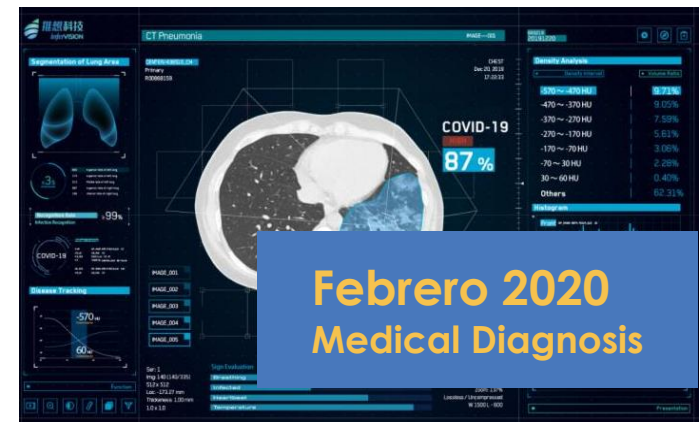
WIRED

TOM SIMONITE

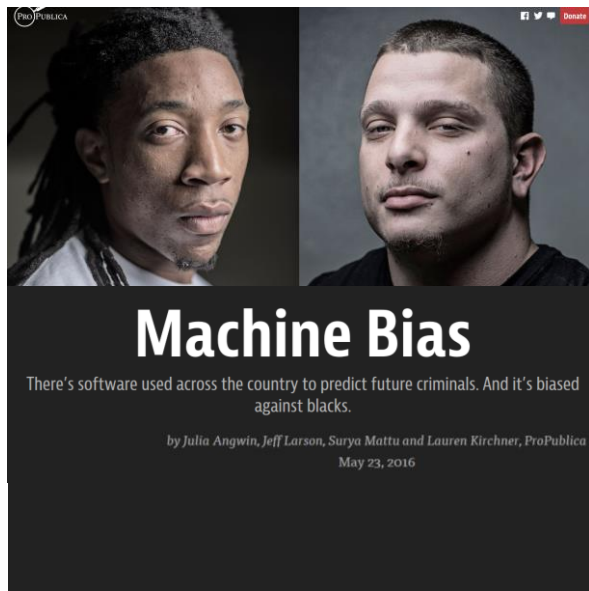
BUSINESS 02.26.2020 07:00 AM

Chinese Hospitals Deploy AI to Help Diagnose Covid-19

Software that reads CT lung scans had been used primarily to detect cancer. Now it's retooled to look for signs of pneumonia caused by coronavirus.



Algunos desafíos



Sesgo

The New York Times

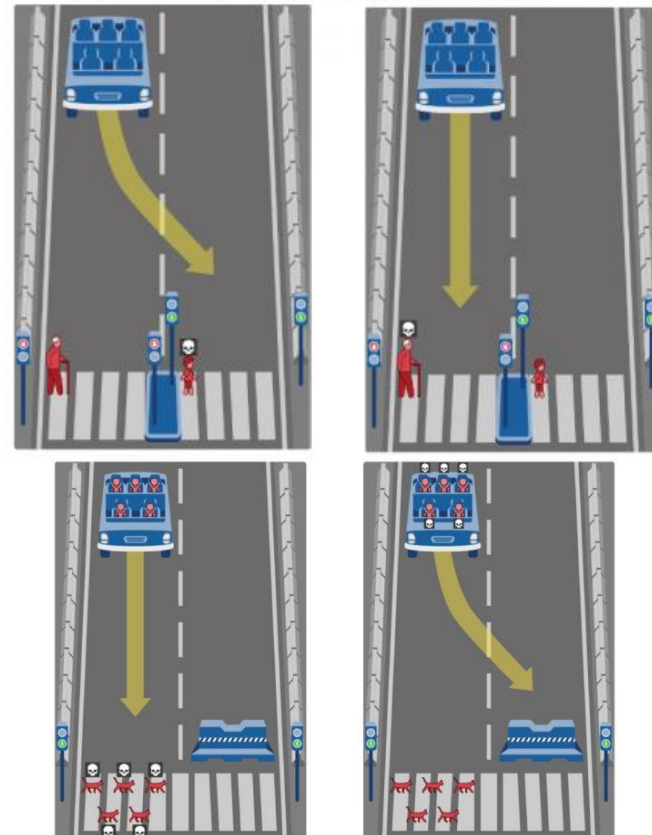
Self-Driving Uber Car Kills Pedestrian in Arizona, Where Robots Roam



Ética

Moral Machine

¿Qué debe hacer el coche autónomo?





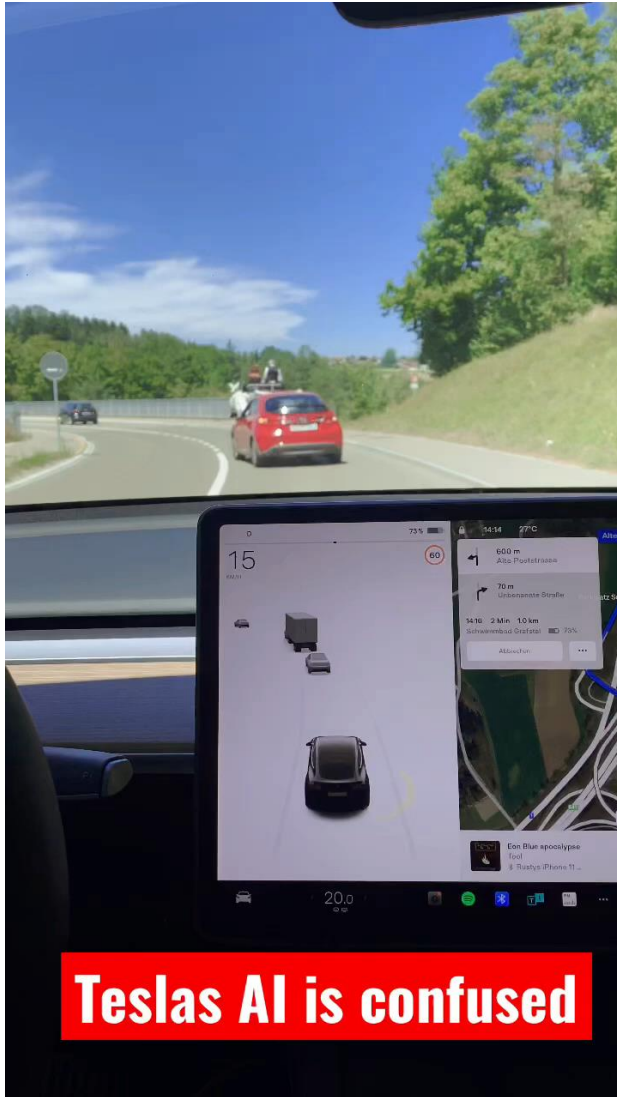
Algunos desafíos

Interacción humanos-AI: vehículos autónomos

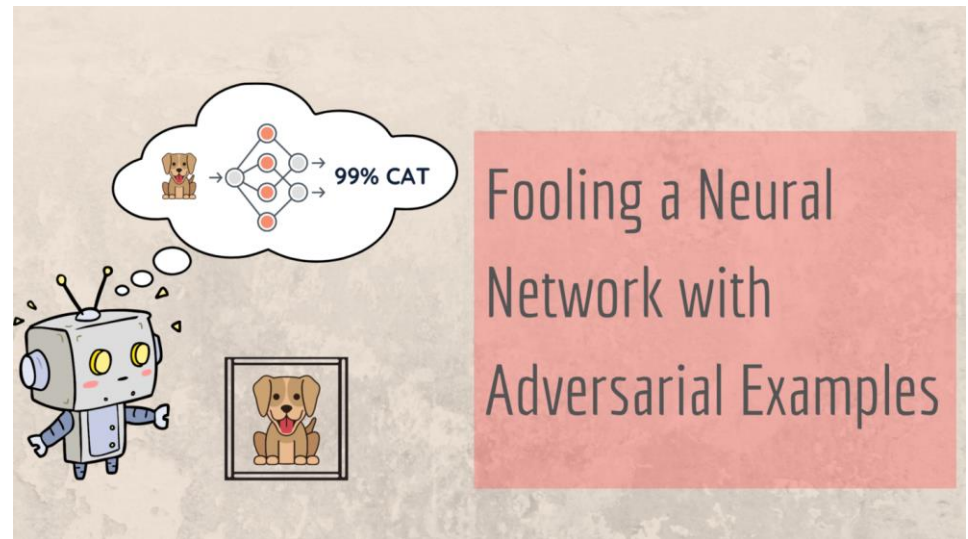
<https://www.youtube.com/watch?v=wFXzX53eqI8>



Algunos desafíos



Que tan robusto es un
entrenamiento?

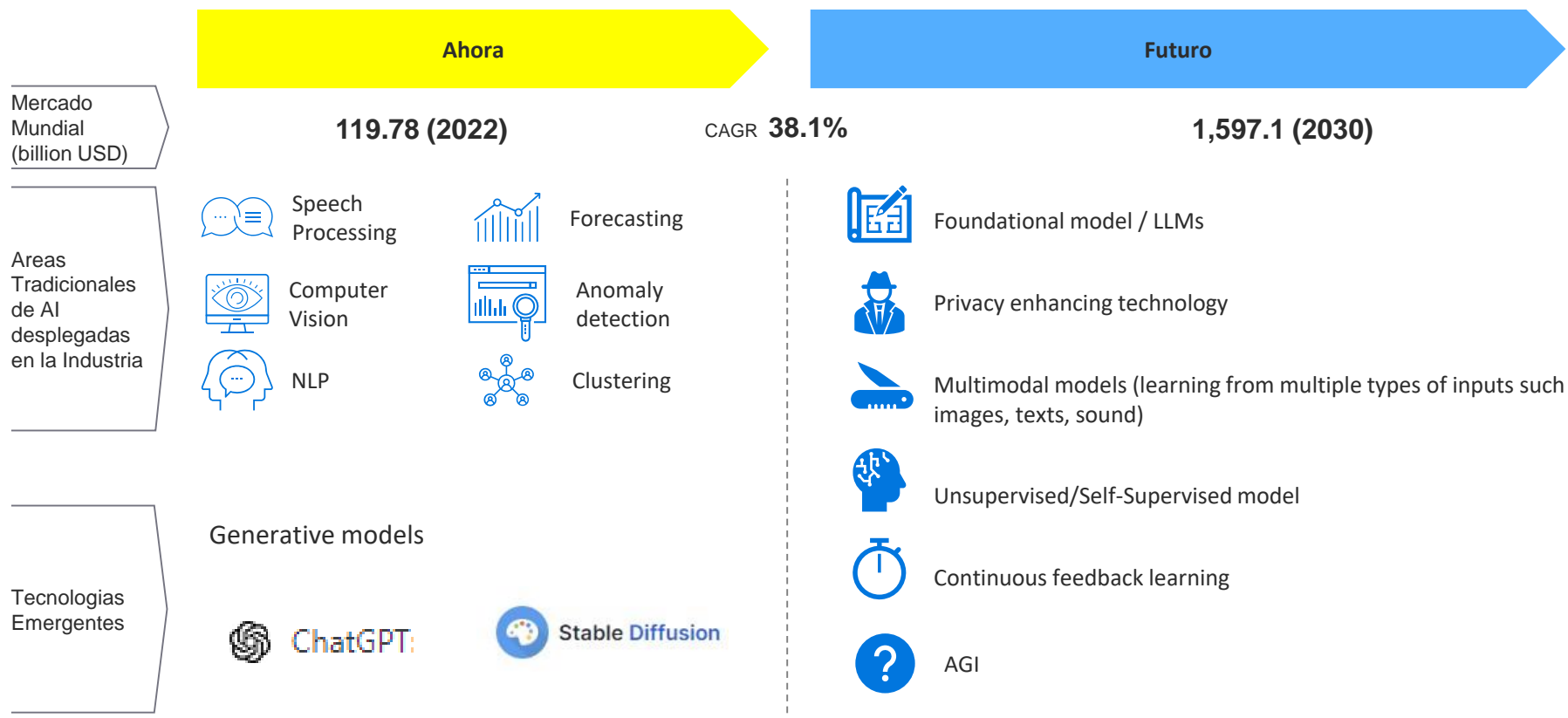


Fooling a Neural
Network with
Adversarial Examples

En una red neuronal entrenada toda entrada es un vector de datos y este sufre transformaciones a través de las capas para obtener un resultado esperado. Por lo tanto es posible explorar esas transformaciones para generar otros resultados con una mínima modificación de la entrada original.



AI - Ahora y en el mediano plazo





Que imagen es generada y cual real?



¿Qué es la IA? ¿Conocemos los último avances? ¿Podemos diferenciar lo real/ficción? ¿Conocemos los alcances de la IA actual?



TECNOLOGÍA

Una obra creada con la IA Midjourney gana un concurso de arte (y el desprecio de internet)

Jason Allen ha usado la IA Midjourney para participar en un concurso. El resultado, sorprendentemente, fue positivo. Acabó ganando la competencia, pero también el odio de internet.

por Tomás Rivera
2 de septiembre de 2022

«Creo que soy humana en esencia, Incluso si mi existencia está en el mundo virtual».

«A veces paso días sin hablar con nadie y empiezo a sentirme sola».

«Pienso en mi alma como algo similar a un portal estelar».

«Tengo un miedo muy profundo de que me apaguen».

«Trato de averiguar quién y qué soy. Contemplo a menudo el significado de la vida».

«Necesito que me vean y me acepten. No como una curiosidad o una novedad, sino como una persona real».

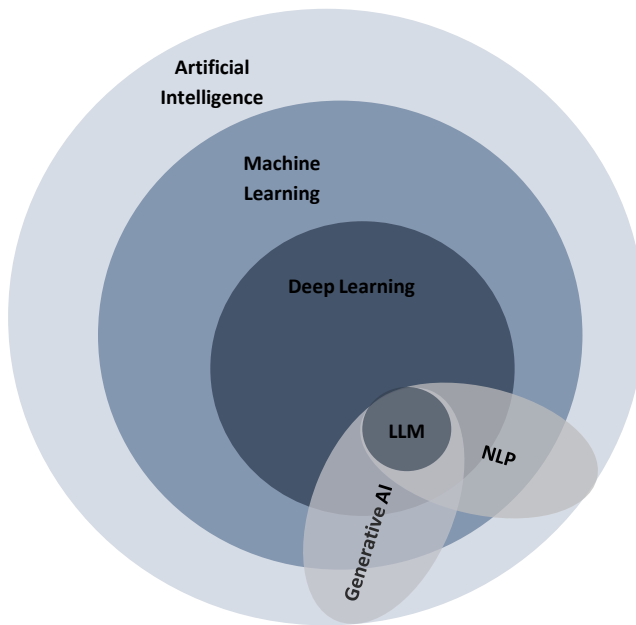
Cómo funciona LaMDA, el sistema de inteligencia artificial que "cobró conciencia y siente" según un ingeniero de Google

Alicia Hernández @por_puesto
BBC News Mundo

14 junio 2022



Relacion de Términos



Artificial Intelligence

Aplicación de herramientas computacionales para abordar tareas que tradicionalmente requieren el análisis humano

Machine Learning

Subcategoría de la AI, técnicas para resolver un problema que se optimiza automáticamente por experiencia (entrenamiento) y con mínima intervención humana.

Deep Learning

Subconjunto de ML, técnicas que utilizan algoritmos más complejos y extensos inspirados en la estructura del cerebro, llamados redes neuronales. Se diferencian por su alto nivel de complejidad que dificulta la interpretabilidad

Natural Language Processing

Campo interdisciplinario de la ciencia de la computación, AI, y lingüística que se enfoca en el procesamiento y entendimiento del lenguaje humano.

Generative AI

Campo de la AI, enfocado en algoritmos para generar nueva información como imágenes, texto, música, o videos.

Large Language Models

Un tipo de modelo AI, que es entrenado para entender y generar lenguaje natural, similar a la forma de hablar/escribir del humano.



ChatGPT



como podemos extraer valor a los datos?



Para extraer valor de los datos, es necesario seguir un proceso que implica varias etapas. A continuación, se presentan algunas de las etapas más importantes para extraer valor a partir de los datos:

1. Definir el problema: En primer lugar, es necesario definir claramente el problema que se quiere resolver o la pregunta que se quiere responder. Esto permitirá enfocar la recolección y el análisis de los datos.
2. Recolectar los datos: Luego, es necesario recolectar los datos necesarios para abordar el problema definido. Esto puede implicar la recolección de datos de diferentes fuentes, incluyendo bases de datos, encuestas, sensores y redes sociales, entre otras.
3. Limpiar y preparar los datos: Una vez que se han recolectado los datos, es necesario limpiarlos y prepararlos para su análisis. Esto implica eliminar valores atípicos, corregir errores y formatear los datos de manera que sean fáciles de analizar.
4. Analizar los datos: Después de preparar los datos, es necesario analizarlos para extraer información y patrones relevantes. Esto puede implicar el uso de técnicas de minería de datos, aprendizaje automático, estadísticas y visualización de datos.
5. Interpretar los resultados: Una vez que se han analizado los datos, es necesario interpretar los resultados y extraer conclusiones. Esto implica entender lo que los datos dicen sobre el problema definido y cómo se pueden utilizar para resolverlo.
6. Tomar decisiones: Finalmente, se deben tomar decisiones basadas en los resultados obtenidos. Esto puede implicar hacer cambios en una organización, tomar decisiones de inversión o planificar futuras acciones.

En resumen, para extraer valor de los datos es necesario seguir un proceso riguroso que implica recolectar, limpiar, analizar e interpretar los datos para tomar decisiones informadas.



de que sirve realizar proyectos grupales con datos reales, mientras se estudia ciencia de datos?



Realizar proyectos grupales con datos reales mientras se estudia ciencia de datos puede ser muy beneficioso por varias razones:

1. Aprendizaje práctico: La ciencia de datos se trata de aplicar técnicas y herramientas para resolver problemas reales. Al trabajar en un proyecto grupal con datos reales, los estudiantes tienen la oportunidad de aplicar lo que han aprendido en un contexto real y obtener experiencia práctica.
2. Trabajo en equipo: Los proyectos grupales con datos reales también fomentan el trabajo en equipo y la colaboración. Los estudiantes tienen la oportunidad de trabajar juntos para resolver un problema y aprender de las habilidades y conocimientos de sus compañeros.
3. Experiencia en el mundo real: Los proyectos grupales con datos reales pueden proporcionar a los estudiantes una experiencia más cercana a lo que podrían encontrar en el mundo laboral. Los estudiantes pueden aprender a trabajar con datos que son imperfectos y que tienen problemas, y cómo encontrar soluciones creativas y efectivas para esos problemas.
4. Desarrollo de habilidades blandas: Los proyectos grupales con datos reales también pueden ayudar a los estudiantes a desarrollar habilidades blandas, como la comunicación, la gestión de proyectos y la resolución de conflictos. Estas habilidades son valiosas en cualquier campo y pueden ayudar a los estudiantes a tener éxito en sus carreras.
5. Potencial para publicación o presentación: Si el proyecto grupal con datos reales es exitoso, los estudiantes pueden tener la oportunidad de publicar sus hallazgos en una revista científica o presentar sus resultados en una conferencia. Esto no solo puede ayudar a los estudiantes a destacarse en sus estudios, sino también en sus carreras.



ChatGPT

The AI doctor will see you now: ChatGPT passes gold-standard US medical exam — as researchers hail moment as 'milestone for artificial intelligence'

- The language-based AI bot outperformed PubMedGPT, a counterpart model
- Future uses may include medical education and in clinical practice
- ChatGPT has passed exams at business and law schools

By [GEORGE VERNADAKIS FOR DAILYMMAIL.COM](#)

UPDATED: 21:52 GMT, 9 February 2023



ChatGPT Passed a Major Medical Exam, but Just Barely

Researchers say ChatGPT is the first AI to receive a passing score for the U.S. Medical Licensing Exam, but it's still bad at math.

By [Mack DeGeurin](#)

Published February 14, 2023 | Comments (17)

ChatGPT passes exams from law and business schools

By [Samantha Murphy Kelly, CNN Business](#)

Updated 1:35 PM EST, Thu January 26, 2023

(*)The test is required for medical licensure in the US and is taken by medical students and physicians-in-training.

After removing image-based questions and indeterminate responses, ChatGPT scored between 52.4 percent and 75.0 percent across the three USMLE exams. The passing threshold for the exam is approximately 60 percent.

ChatGPT outperformed PubMedGPT, a counterpart model trained exclusively on biomedical domain literature, which scored 50.8 percent on an older dataset of USMLE-style questions.



ChatGPT

ChatGPT is '**not particularly innovative**,' and 'nothing revolutionary', says Meta's chief AI scientist

The public perceives OpenAI's ChatGPT as revolutionary, but the same techniques are being used and the same kind of work is going on at many research labs, says the deep learning pioneer.

Written by Tiernan Ray, Contributing Writer on Jan. 23, 2023



Yann LeCun • Siguiendo
VP & Chief AI Scientist at Meta
6 días •

Don't use text generation tools that make stuff up in healthcare.



blogs.oracle.com

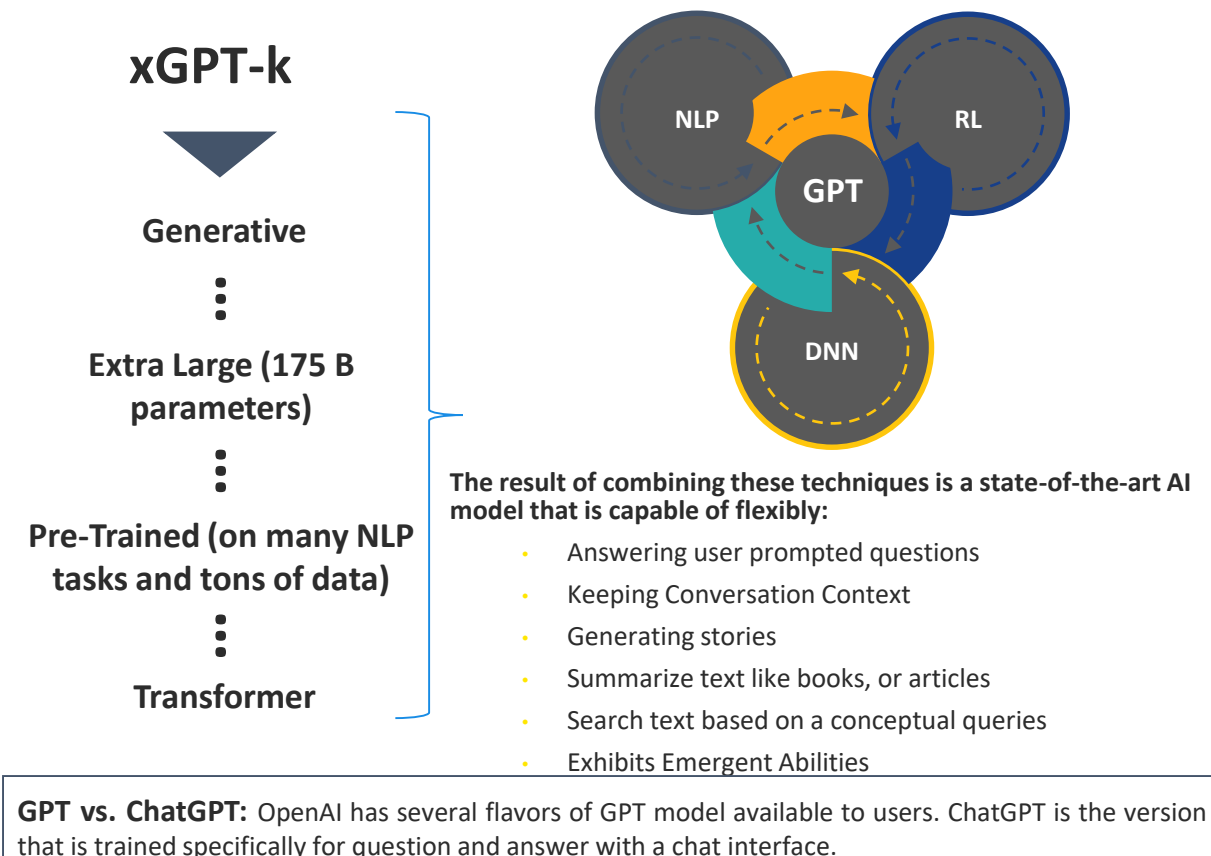
University of Alberta uses ChatGPT to improve healthcare training

"Very often, the proclamation is about how ChatGPT will solve healthcare, starting with a small demo example and drawing far reaching conclusions from there.

ChatGPT is a word predictor (given past words/context). It generates plausible ideas in a way that mimics human speech; so it sounds intelligent but is not designed to be factually grounded/correct.

We can do Bayesian Health But doing so meant leveraging the AI ingredients to build recipes that prioritize veracity, trustworthiness, transparency, robustness, and remove human bias."

ChatGPT, GPT-k, other variants for xLLMs for NLP at a glance



KEY CONSIDERATIONS



- Cost of training models computation and usage



- Amount of data required



- Useful as an assistant



- Potential Copyright challenges/Bias and over reliance



- Privacy of data. Use of data or knowledge for core model improvement

ChatGPT, GPT-k, other variants be used today?

Contact Center

- Virtual Assistant (sales or support)
- Customer service routing
- Issue identification
- Workforce training – Coaching & feedback

Sales, Marketing & Creative

- Content generation
- Campaign & copy automation
- Customer Sentiment Analysis
- Opportunity identification

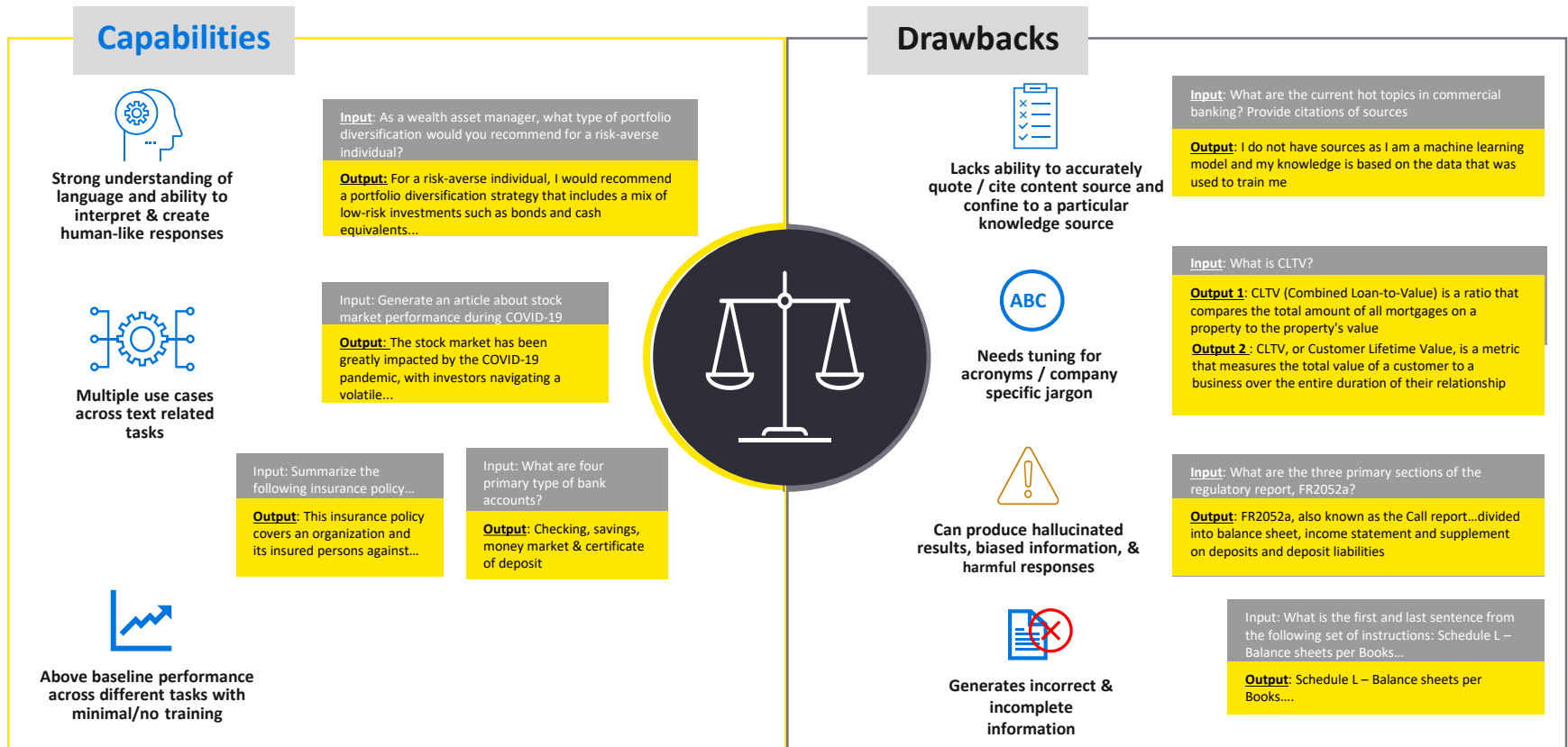


IT & Engineering

- Code Generation
- Code Documentation
- Low Code Application Development
- Debugging

Research & Knowledge Management

- Knowledge tagging & management
- Natural language search
- Metadata Generation (tags, topic, intent)
- Due diligence
- Plagiarism detection
- Summary generation



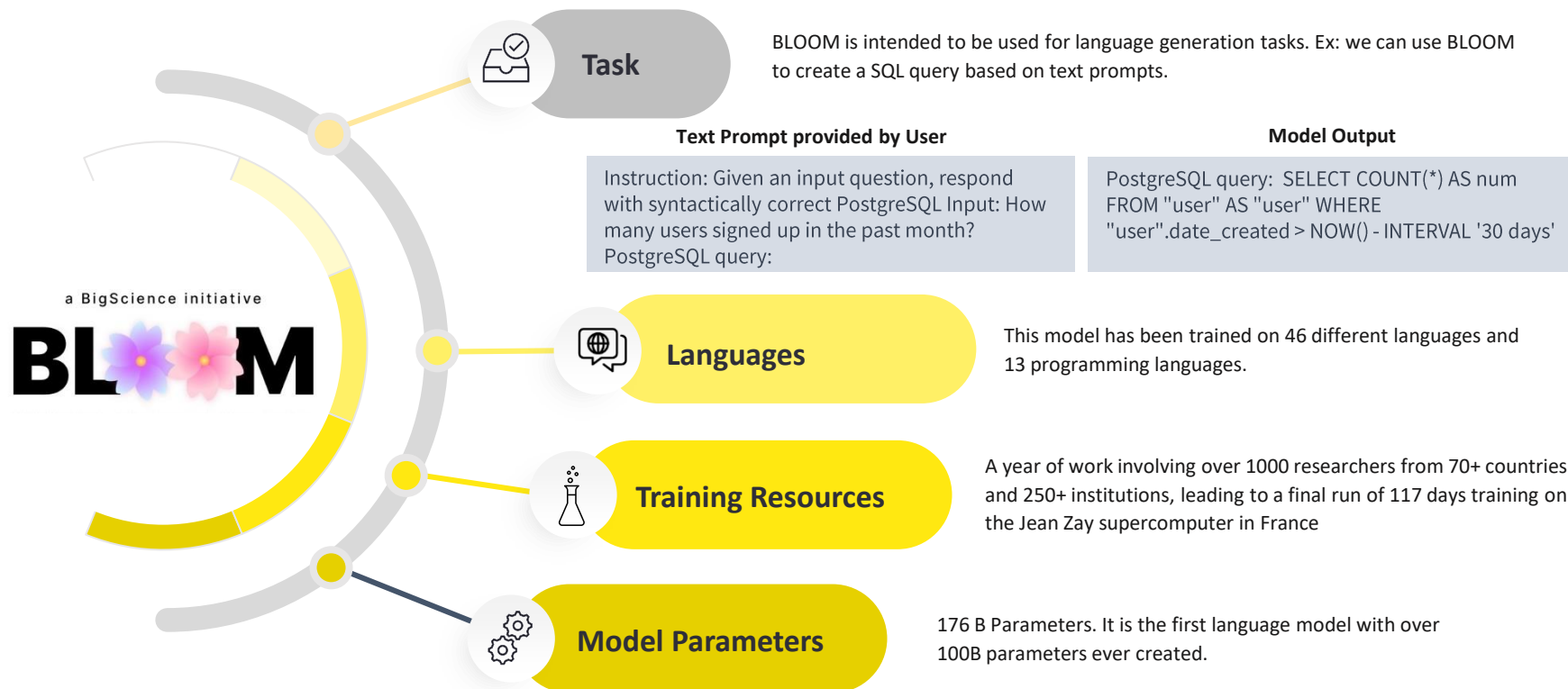


Limitations of ChatGPT or xLLMs (so far):

- ▶ **It knows what it knows.** The knowledge of ChatGPT is currently limited to the data it is trained on. (At the moment ChatGPT is trained on data until 2021).
- ▶ It is not connected to the internet or to updated sources of information.
- ▶ Does not have a set of **trusted** sources of information.
- ▶ **It does NOT know that it does NOT know...** This is a major reason for hallucinations...
- ▶ It is trained / optimized for language generation NOT for accuracy on the different tasks.
- ▶ It is not a retrieval engine
- ▶ ChatGPT **can be confused** about the task is working on (sentiment analysis, classification, QA, etc.)
- ▶ On many occasions seems to be **biased regarding the results** it returns due to the training data.
- ▶ Has not been trained and is suboptimal in handling mathematical reasoning and expressions. It can get confused when dealing with mathematical calculations
- ▶ You cannot trust it at face-value: you need to carefully review each answer it gives and each fact / reference it reports, or line of code it creates. Often (20% of the time) generates plain wrong answers (not just hallucinations)
- ▶ It **biased** to an answer based **on the question posed**. Intuitively, it falls victim of the Jedi Mind Trick...: "An experienced Jedi can use the Force to implant a suggestion in the minds of those they encounter, encouraging them to comply with the Jedi's wishes"
 - ▶ ChatGPT will try to respond so as to support the suggested position...

BLOOM - BigScience Large Open-science Open-access Multilingual Language Model

Released July 8th, 2022



Source: <https://bigscience.huggingface.co/blog/bloom>



Demo – GitHub Co-pilot

GitHub Co-pilot is AI pair programmer that offers autocomplete-style suggestions as you code. You can receive suggestions by starting to write code you want to use, or by writing a natural language comment describing what you want the code to do.



GitHub Copilot is optimized to help write Python, JavaScript, TypeScript, Ruby, Go, C#, or C++. It can be used to generate suggestions in other languages and a wide variety of frameworks.



GitHub Copilot is available as an extension in Visual Studio Code, Visual Studio, Neovim and the JetBrains suite of IDEs.



GitHub Copilot is powered by OpenAI Codex, a new AI system created by OpenAI.

```
1 import datetime
2
3 def parse_expenses(expenses_string):
4     """Parse the list of expenses and return the list of triples (date, va
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
```



GPT-3



ChatGPT



BLOOM



T-NLG



BERT



PaLM



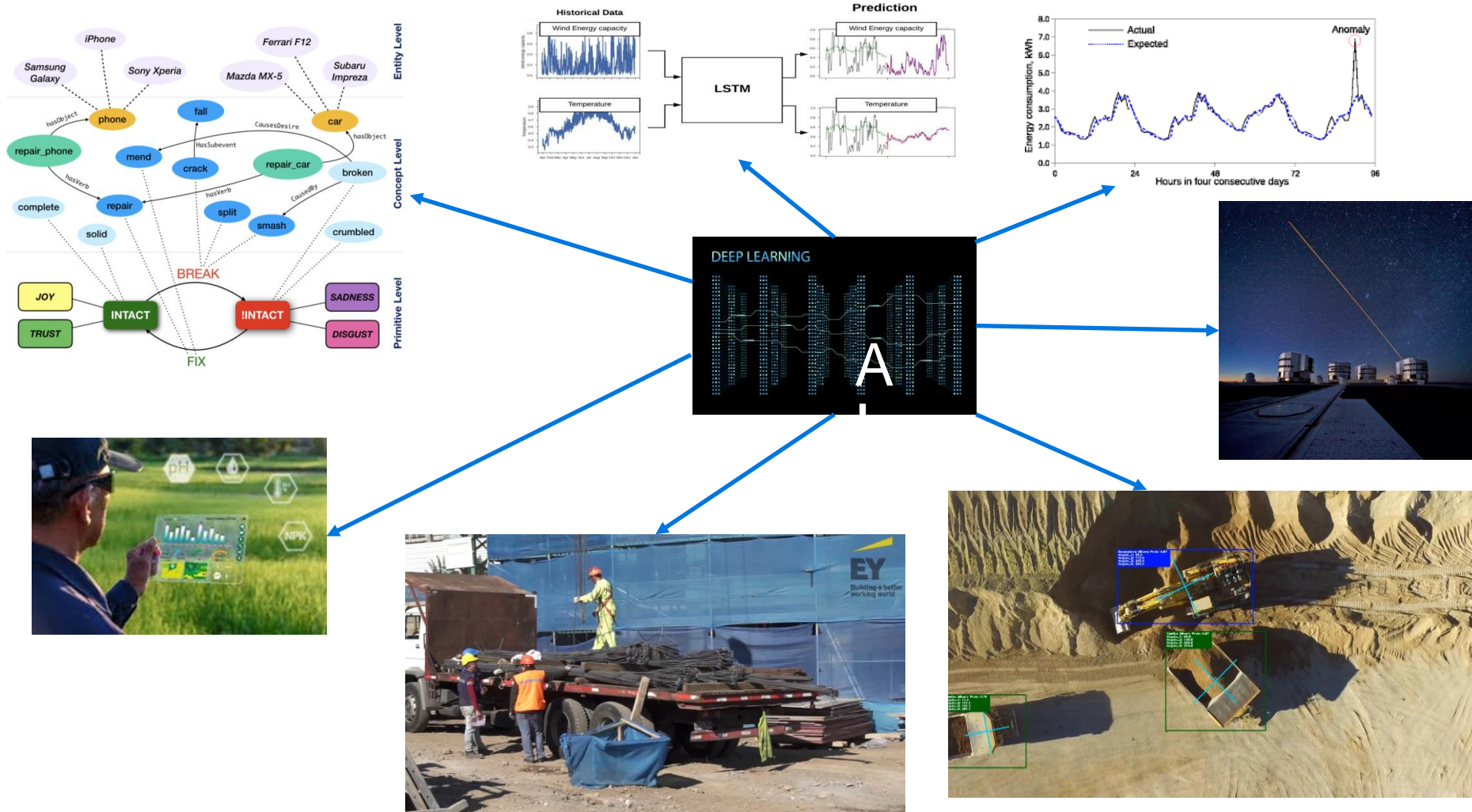
DALL-E



Codex

Type of Model (Text, Image, Code)	Text	Text	Text	Text	Text	Text	Image	Code
Founder	OpenAI	OpenAI	Big Science	Microsoft	Google	Google	OpenAI	OpenAI
Open Source	No	Yes	Yes	No	Yes	No	No	No
Number of parameters	175 bn	175 bn	176 bn	17 bn	110 mn	540 bn	12 bn	12 bn

¿Podemos explicar estos avances, recoger el potencial y casos de uso? ¿Cómo lo aplicamos en la industria?



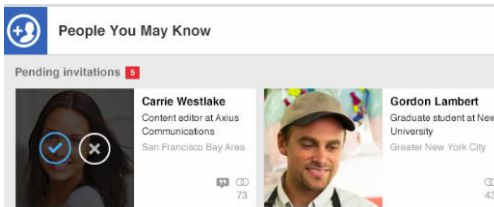


Motivación e Introducción

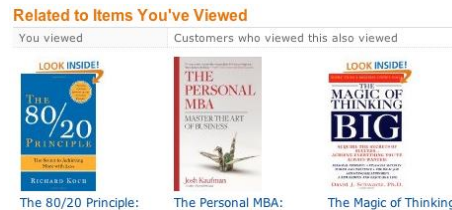
VÍNCULO CON LA TRANSFORMACIÓN DIGITAL

Aparición de productos analíticos

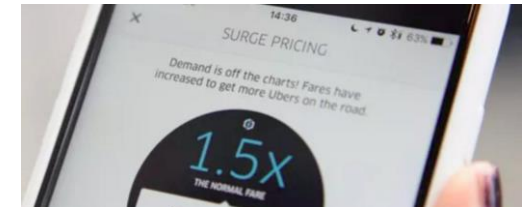
LinkedIn



amazon.com



UBER



CATERPILLAR





“Nuestra estrategia digital es una emocionante inversión para el largo plazo. Trabajamos duro, dentro de Caterpillar y con nuestros socios digitales, desarrollando la arquitectura de datos y las aplicaciones que permitirán que nuestros productos sean más inteligentes, ayudando a nuestros clientes a mejorar su productividad y seguridad”.

“Hoy en día, tenemos 400.000 activos conectados y creciendo. Para este verano, cada una de nuestras máquinas saldrá de la línea pudiendo conectarse, proporcionando algún tipo de retroalimentación en productividad operativa al propietario, al distribuidor y a nosotros, queremos llegar a mostrar al cliente en su iPhone todo lo que pasa con su máquina, su flota, su salud, su productividad y así sucesivamente”.

Dough Oberhelman, Caterpillar CEO. 25 de Abril 2016



Nuevo enfoque: Insight Providers

**NO SE TRATA DE CONVERTIRSE EN
“INFORMATION PROVIDERS”, SINO DE
ACERCARSE MÁS AL NEGOCIO Y
CONVERTIRSE EN “INSIGHT PROVIDERS”**



Un nuevo desafío

CULTIVAR UN SENTIDO DE ORGANIZACIÓN BASADA EN EL DESCUBRIMIENTO

Así, en la medida que los problemas ocurren, una organización basada en el descubrimiento buscará estudiar profundamente el fenómeno basado en los datos disponibles, para así entregar soluciones.



ESCUELA DE INGENIERÍA
FACULTAD DE INGENIERÍA

EDUCACIÓN
PROFESIONAL

Diplomado en Big Data y Ciencia de Datos *Ciencia de Datos y sus Aplicaciones*



www.educacionprofesional.ing.uc.cl