FACULTAD de
CIENCIAS ECONÓMICAS

Introducción al curso

Alan Reyes-Figueroa

Elements of Machine Learning (Aula 01) 12.enero.2023

Motivación

El curso Elements of Machine Learning es una introducción a los métodos estadísticos, matemáticos y computacionales para extraer información basada en datos. Incluye técnicas provenientes áreas como: estadística, reconocimiento estadístico de patrones (*pattern recognition*), aprendizaje estadístico o aprendizaje de máquina (*machine learning*), ciencia de datos.

Motivación

El curso Elements of Machine Learning es una introducción a los métodos estadísticos, matemáticos y computacionales para extraer información basada en datos. Incluye técnicas provenientes áreas como: estadística, reconocimiento estadístico de patrones (*pattern recognition*), aprendizaje estadístico o aprendizaje de máquina (*machine learning*), ciencia de datos.

Este es un curso integrador. Haremos uso de

- estadística e inferencia estadística,
- álgebra lineal (espacios, autovalores, descomposición matricial),
- optimización contínua,
- reconocimiento de patrones y aprendizaje estadístico,
- programación y algoritmos.

¿Qué no es ciencia de datos?

Ciencia de datos $\neq$ *machine learning* (ML)

¿Qué no es ciencia de datos?

Ciencia de datos $\neq$ *machine learning* (ML)

- El aprendizaje automático involucra, matemática, computación y estadística, pero tradicionalmente no trata sobre cómo resolver preguntas científicas.
  El aprendizaje automático tiene un enfoque más de algoritmos.

¿Qué no es ciencia de datos?

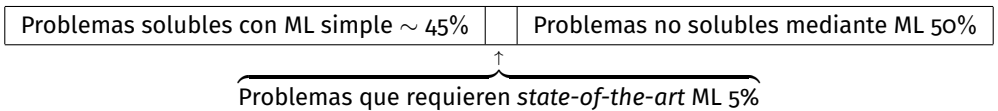
Ciencia de datos $\neq$ *machine learning* (ML)

- El aprendizaje automático involucra, matemática, computación y estadística, pero tradicionalmente no trata sobre cómo resolver preguntas científicas.
  El aprendizaje automático tiene un enfoque más de algoritmos.
- Algunas veces, la mejor forma de resolver un problema es visualizando los datos.

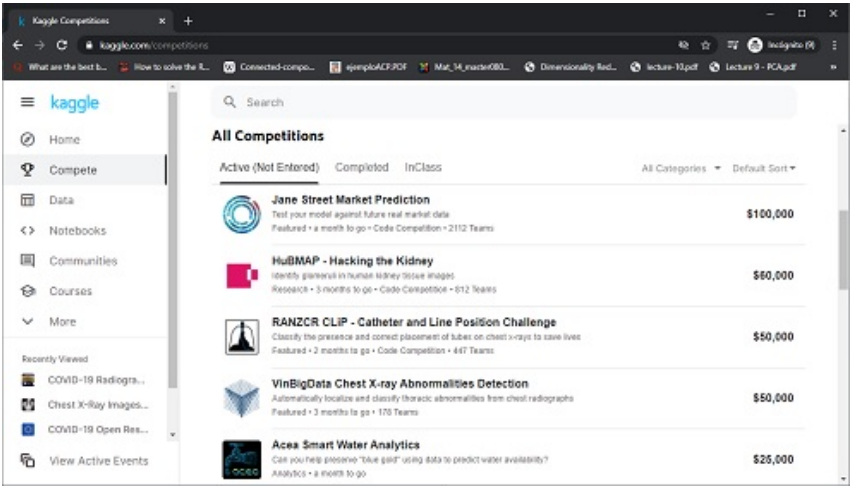
¿Qué no es ciencia de datos?

Ciencia de datos $\neq$ *machine learning* (ML)

- El aprendizaje automático involucra, matemática, computación y estadística, pero tradicionalmente no trata sobre cómo resolver preguntas científicas.
  El aprendizaje automático tiene un enfoque más de algoritmos.
- Algunas veces, la mejor forma de resolver un problema es visualizando los datos.



¿Qué no es ciencia de datos?



UFM UNIVERSIDAD FRANCISCO MARROQUÍN

¿Qué no es ciencia de datos?

Data science $\neq$ competencias o concursos.

¿Qué no es ciencia de datos?

Data science $\neq$ competencias o concursos.

- Concursos de ciencia de datos, *e.g.* Kaggle, usualmente requieren optimizar una métrica sobre un conjunto de datos fijo.

¿Qué no es ciencia de datos?

Data science $\neq$ competencias o concursos.

- Concursos de ciencia de datos, *e.g.* Kaggle, usualmente requieren optimizar una métrica sobre un conjunto de datos fijo.
- Esto, en última instancia, no resuelve un problema científico o aplicado.

¿Qué no es ciencia de datos?

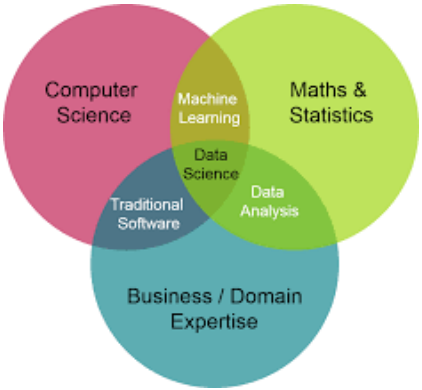
Data science $\neq$ competencias o concursos.

- Concursos de ciencia de datos, *e.g.* Kaggle, usualmente requieren optimizar una métrica sobre un conjunto de datos fijo.
- Esto, en última instancia, no resuelve un problema científico o aplicado.
- La ciencia de datos es un ciclo iterativo en el que se plantea un problema, y se busca diseñar mecanismos o algoritmos para resolverlo (o determinar que no es posible), y evaluar qué aportes pueden generar estos algoritmos sobre la pregunta en onsideración.

¿Qué no es ciencia de datos?

Ciencia de datos $\neq$ estadística

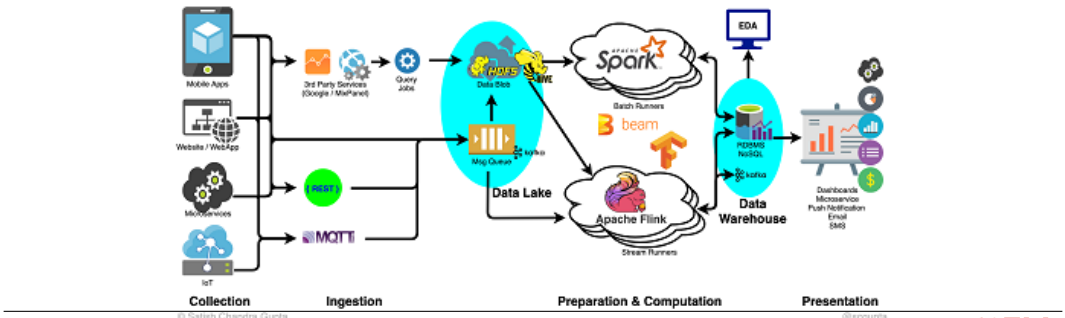
- Estadística (al menos en un sentido académico), ha evolucionado al punto de probar teoremas. Hacer teoría estadística.
- En este curso veremos algunos pocos teoremas, pero no vamos a hacer teoría. La idea principal es que este sea un curso aplicado.



¿Qué no es ciencia de datos?

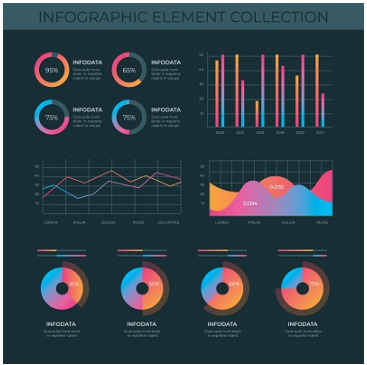
Ciencia de datos $\neq$ *big data*

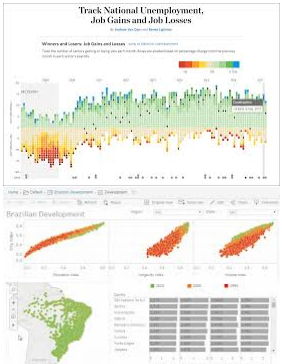
- El término *big data* está más relacionado con la ingeniería de software. Se refiere más al tratamiento de grandes cantidades de datos, o a las técnicas, metodologías o desarrollo de *pipelines* o *workflows* para el procesamiento de datos.

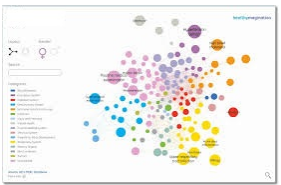


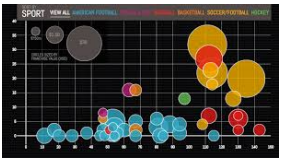
¿Qué no es ciencia de datos

Ciencia de datos $\neq$ visualización









«The greatest value of a picture is when it forces us to notice what we never expected to see.» –John Tukey

UFM UNIVERSIDAD FRANCISCO MARROQUÍN

¿Qué es ciencia de datos?

Algunas posibles definiciones.

- Es la aplicación de técnicas estadísticas y computacionales para obtener o ganar entendimiento de un problema en el mundo real, mediante datos.
- Ciencia de datos = estadística + procesamiento (minería) de datos + aprendizaje automático + investigación científica + visualización de datos + inteligencia de negocio + *big data* + . . .

¿Qué es ciencia de datos?

Algunas posibles definiciones.

- Es la aplicación de técnicas estadísticas y computacionales para obtener o ganar entendimiento de un problema en el mundo real, mediante datos.
- Ciencia de datos = estadística + procesamiento (minería) de datos + aprendizaje automático + investigación científica + visualización de datos + inteligencia de negocio + *big data* + . . .

- A criterio personal, aún no hay una definición concreta, cada persona hace su propia definición según su experiencia y punto de vista.

¿Qué es ciencia de datos?

Algunas posibles definiciones.

- Es la aplicación de técnicas estadísticas y computacionales para obtener o ganar entendimiento de un problema en el mundo real, mediante datos.
- Ciencia de datos = estadística + procesamiento (minería) de datos + aprendizaje automático + investigación científica + visualización de datos + inteligencia de negocio + *big data* + . . .

- A criterio personal, aún no hay una definición concreta, cada persona hace su propia definición según su experiencia y punto de vista.
- Lo que está claro, es que es un tema que mezcla y usa herramientas de muchas áreas del conocimiento.

¿Qué es ciencia de datos?

- Recientemente hay mucha demanda por científicos de datos.

¿Qué es ciencia de datos?

- Recientemente hay mucha demanda por científicos de datos.
- En 2018, US esperimentará una demanda de 190,000 científicos de datos, y 1.5 millones de gerentes y analistas capaces de generar información útil mediante datos.
  Ref. Susan Lund *et al.*, "Game Changers: Five Opportunities for US Growth and Renewal," McKinsey Global Institute Report, July 2013.

UFM UNIVERSIDAD FRANCISCO MARROQUÍN

¿Qué es ciencia de datos?

- La ciencia de datos y el aprendizaje automático no son nada nuevo, pero la tendencia actual continúa impulsando las tecnologías hacia el centro de atención.

¿Qué es ciencia de datos?

- La ciencia de datos y el aprendizaje automático no son nada nuevo, pero la tendencia actual continúa impulsando las tecnologías hacia el centro de atención.
- Creciente interés (y exageración) en torno a la inteligencia artificial (IA), impulsado por el *marketing* y combinada con la comprensible confusión de términos: IA, ML, DC.

¿Qué es ciencia de datos?

- La ciencia de datos y el aprendizaje automático no son nada nuevo, pero la tendencia actual continúa impulsando las tecnologías hacia el centro de atención.
- Creciente interés (y exageración) en torno a la inteligencia artificial (IA), impulsado por el *marketing* y combinada con la comprensible confusión de términos: IA, ML, DC.
- Escasez de talento en ciencia de datos y aprendizaje automático.

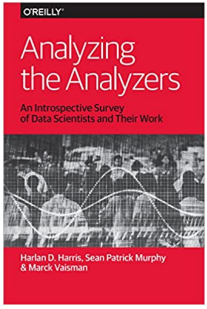
¿Qué es ciencia de datos?

- La ciencia de datos y el aprendizaje automático no son nada nuevo, pero la tendencia actual continúa impulsando las tecnologías hacia el centro de atención.
- Creciente interés (y exageración) en torno a la inteligencia artificial (IA), impulsado por el *marketing* y combinada con la comprensible confusión de términos: IA, ML, DC.
- Escasez de talento en ciencia de datos y aprendizaje automático.
- Aumento de la capacidad y potencia informática y la disponibilidad de arquitecturas avanzadas. (Estos avances han alimentado la publicidad y el interés en torno al aprendizaje profundo (*deep learning*)).

¿Qué es ciencia de datos?

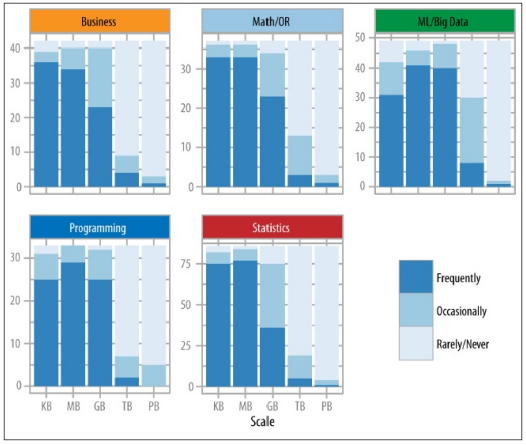
- La ciencia de datos y el aprendizaje automático no son nada nuevo, pero la tendencia actual continúa impulsando las tecnologías hacia el centro de atención.
- Creciente interés (y exageración) en torno a la inteligencia artificial (IA), impulsado por el *marketing* y combinada con la comprensible confusión de términos: IA, ML, DC.
- Escasez de talento en ciencia de datos y aprendizaje automático.
- Aumento de la capacidad y potencia informática y la disponibilidad de arquitecturas avanzadas. (Estos avances han alimentado la publicidad y el interés en torno al aprendizaje profundo (*deep learning*)).
- Aumento y popularidad de herramientas y bibliotecas de código abierto para ciencia de datos y aprendizaje automático.

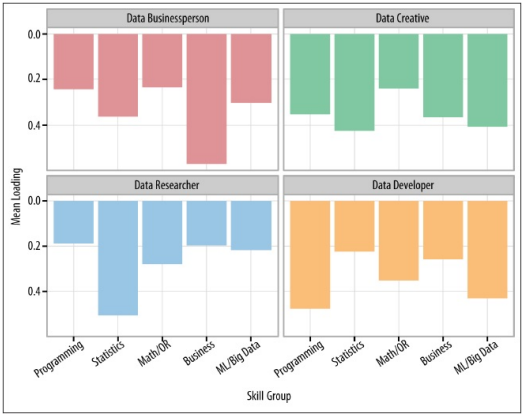
¿Qué hace un científico de datos?



Data Developer	Developer	Engineer	
Data Researcher	Researcher	Scientist	Statistician
Data Creative	Jack of All Trades	Artist	Hacker
Data Businessperson	Leader	Businessperson	Entrepeneur

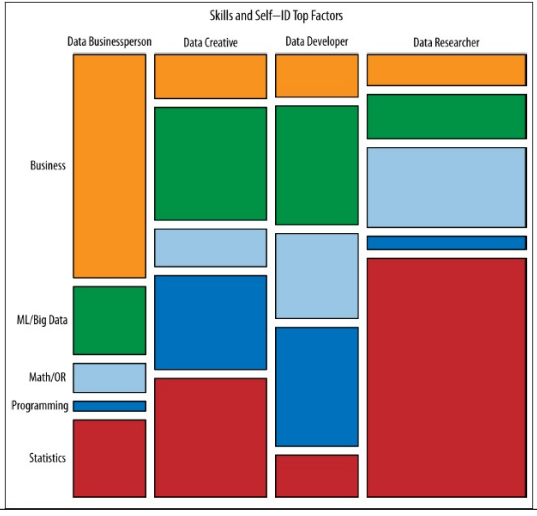
¿Qué hace un científico de datos?



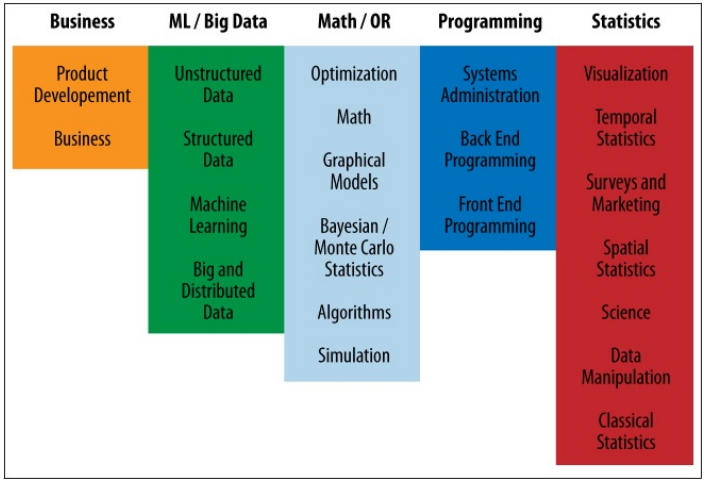


UFM UNIVERSIDAD FRANCISCO MARROQUÍN

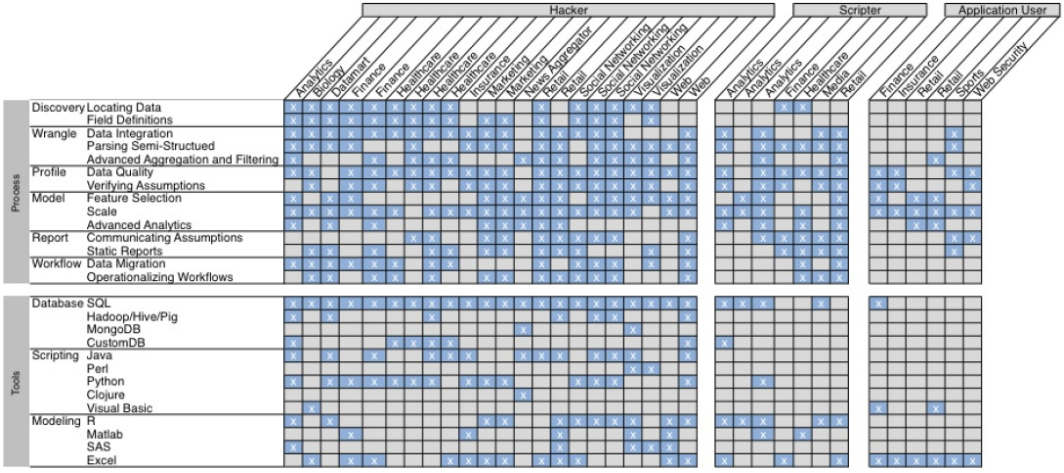
Habilidades



Habilidades

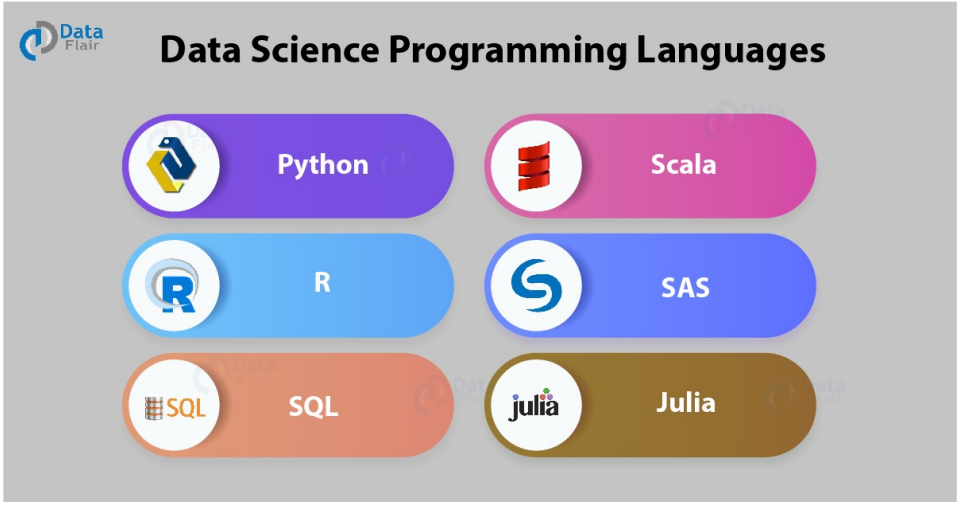


Tareas

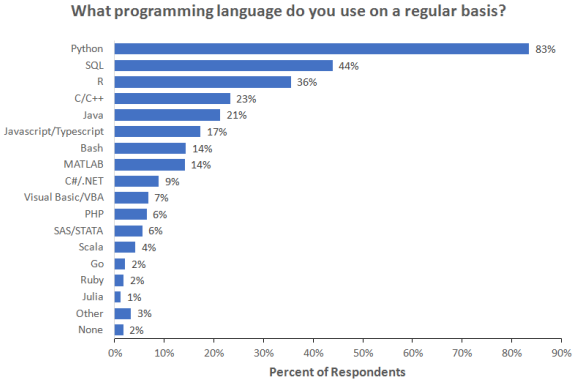


UFM UNIVERSIDAD FRANCISCO MARROQUÍN

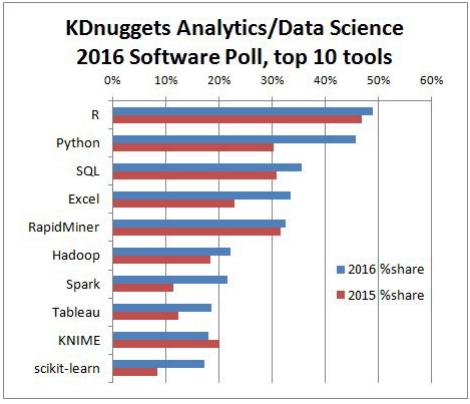
Lenguajes de programación



Lenguajes y herramientas



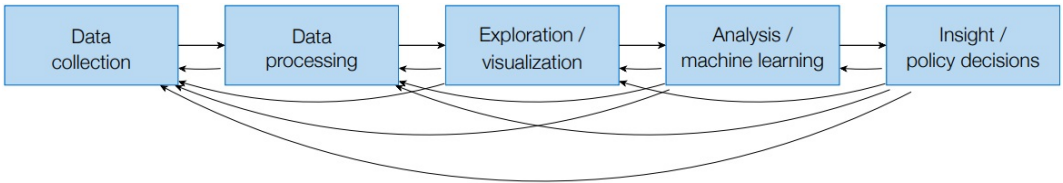
Note: Data are from the 2018 Kaggle Machine Learning and Data Science Survey.



UFM UNIVERSIDAD FRANCISCO MARROQUÍN

Haciendo ciencia de datos

Hacer ciencia de datos es un proceso que conlleva varias etapas y que integra habilidades diversas, y colaboración entre disciplinas, profesionales y enfoques diversos.



Haciendo ciencia de datos

Por ejemplo, Ben Fry, propone el siguiente modelo de ciencia de datos:

1. Acquire
2. Parse
3. Filter
4. Mine
5. Represent
6. Refine
7. Interact

Haciendo ciencia de datos

En contraste, Jeff Hammerbacher porpone este esquema para hacer ciencia de datos:

1. Identify problem
2. Instrument data sources
3. Collect data
4. Prepare data (integrate, transform, clean, filter, aggregate)
5. Build model
6. Evaluate model
7. Communicate results

Haciendo ciencia de datos

Digging Around in Data

Clean, prep

Hypothesize Model

Large Scale Exploitation

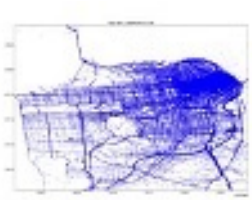
Evaluate Interpret

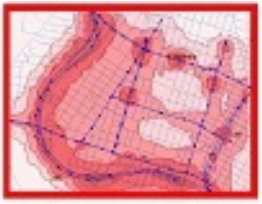
La parte difícil

¿Qué parte es difícil a la hora de hacer ciencia de datos?

- Overcoming assumptions
- Making ad-hoc explanations of data patterns
- Overgeneralizing
- Communication
- Not checking enough (validate models, data pipeline integrity, etc.)
- Using statistical tests correctly
- Mathematical models fail (who do you ask?)
- Prototype - Production transitions
- Data pipeline complexity (who do you ask?)

En resumen





Crowdsourcing + physical modeling + sensing + data assimilation

to produce:

Algunos ejemplos

¿De qué va este curso?

Como este es un curso de matemática, la idea es hacer una introducción a la ciencia de datos, desde un punto de vista más matemático.

- Más orientado a *machine learning*, patrones y análisis de datos.

¿De qué va este curso?

Como este es un curso de matemática, la idea es hacer una introducción a la ciencia de datos, desde un punto de vista más matemático.

- Más orientado a *machine learning*, patrones y análisis de datos. Veremos algoritmos, y su fundamento matemático (no vamos a hacer teoría, pero sí vamos a mencionar teoremas importantes, y mostrar algunos de ellos).

¿De qué va este curso?

Como este es un curso de matemática, la idea es hacer una introducción a la ciencia de datos, desde un punto de vista más matemático.

- Más orientado a *machine learning*, patrones y análisis de datos.
  Veremos algoritmos, y su fundamento matemático (no vamos a hacer teoría, pero sí vamos a mencionar teoremas importantes, y mostrar algunos de ellos).
  Fundamentos en optimización, estadística, cálculo y álgebra lineal (herramientas).

¿De qué va este curso?

Como este es un curso de matemática, la idea es hacer una introducción a la ciencia de datos, desde un punto de vista más matemático.

- Más orientado a *machine learning*, patrones y análisis de datos.
  Veremos algoritmos, y su fundamento matemático (no vamos a hacer teoría, pero sí vamos a mencionar teoremas importantes, y mostrar algunos de ellos).
  Fundamentos en optimización, estadística, cálculo y álgebra lineal (herramientas).
- Veremos una parte computacional: implementar algoritmos.

¿De qué va este curso?

Como este es un curso de matemática, la idea es hacer una introducción a la ciencia de datos, desde un punto de vista más matemático.

- Más orientado a *machine learning*, patrones y análisis de datos.
  Veremos algoritmos, y su fundamento matemático (no vamos a hacer teoría, pero sí vamos a mencionar teoremas importantes, y mostrar algunos de ellos).
  Fundamentos en optimización, estadística, cálculo y álgebra lineal (herramientas).
- Veremos una parte computacional: implementar algoritmos.
  Laboratorios

¿De qué va este curso?

Como este es un curso de matemática, la idea es hacer una introducción a la ciencia de datos, desde un punto de vista más matemático.

- Más orientado a *machine learning*, patrones y análisis de datos.
  Veremos algoritmos, y su fundamento matemático (no vamos a hacer teoría, pero sí vamos a mencionar teoremas importantes, y mostrar algunos de ellos).
  Fundamentos en optimización, estadística, cálculo y álgebra lineal (herramientas).
- Veremos una parte computacional: implementar algoritmos.
  Laboratorios
  Ejercicios sobre algoritmos (teórico), analizar datos (aplicado).

¿De qué va este curso?

Como este es un curso de matemática, la idea es hacer una introducción a la ciencia de datos, desde un punto de vista más matemático.

- Más orientado a *machine learning*, patrones y análisis de datos.
  Veremos algoritmos, y su fundamento matemático (no vamos a hacer teoría, pero sí vamos a mencionar teoremas importantes, y mostrar algunos de ellos).
  Fundamentos en optimización, estadística, cálculo y álgebra lineal (herramientas).
- Veremos una parte computacional: implementar algoritmos.
  Laboratorios
  Ejercicios sobre algoritmos (teórico), analizar datos (aplicado).
- Análisis de datos reales.
  Proyectos aplicados

Detalles importantes

- Requisitos:

Detalles importantes

- Requisitos:
  - Cálculo, álgebra lineal

Detalles importantes

- Requisitos:
  - Cálculo, álgebra lineal
  - Al menos un curso de estadística

Detalles importantes

- Requisitos:
  - Cálculo, álgebra lineal
  - Al menos un curso de estadística
  - Al menos un curso de programación (Python)

Detalles importantes

- Requisitos:
  - Cálculo, álgebra lineal
  - Al menos un curso de estadística
  - Al menos un curso de programación (Python)

- Horario de atención.

Detalles importantes

- Requisitos:
  - Cálculo, álgebra lineal
  - Al menos un curso de estadística
  - Al menos un curso de programación (Python)
- Horario de atención.
- ¿Qué han visto en otros cursos? (*e.g. big data*)

Detalles importantes

- Requisitos:
  - Cálculo, álgebra lineal
  - Al menos un curso de estadística
  - Al menos un curso de programación (Python)
- Horario de atención.
- ¿Qué han visto en otros cursos? (*e.g. big data*)