

Modelación estadística con datos de redes sociales en Python

5° Coloquio de Estadística
Universidad Nacional Autónoma de México
Facultad de Estudios Superiores Acatlán

Mtro. José Gustavo Fuentes Cabrera



Facultad de Estudios Superiores

Acatlán

Agenda

1. Instructor
2. Planteamiento del problema
3. Tecnología actual
4. Data Science
5. Software de trabajo
6. Alta como desarrollador
7. Extracción de datos
8. Modelación Matemática
9. Interpretación de resultados

A decorative background featuring a network diagram. It consists of numerous nodes, represented by circles of varying sizes and shades of gray, connected by thin, light gray lines. Some nodes are highlighted with a solid blue dot, and others are enclosed in a blue outline. The network is distributed across the slide, with a denser cluster on the left side and a more sparse arrangement on the right.

1. Instructor

Instructor

Trayectoria Académica

- ◎ **Lic. en Matemáticas Aplicadas y Computación.**
- ◎ **Especialidad en Minería de Datos**
- ◎ **Maestría en Inteligencia Analítica**
- ◎ **Catedrático UNAM**
 - *Actuaría: Algoritmos y Programación, Bases de Datos, Análisis Numérico, ED, Temas selectos de computación y Análisis Multivariado.*
 - *MAC: Cálculo III, ED II, Estadística II*
- ◎ **Catedrático Universidad Anáhuac**

Trayectoria Profesional

- ◎ **Fundador y Director General Técnico - Nabla Estrategia Analítica.**
- ◎ **Director Data Science-Conlana Capital**
- ◎ **SD Risk & Product Analytics - GFNorte**
- ◎ **SD Risk Analytics Minorista - GFNorte**
- ◎ **Gerente Riesgo TDC - GFNorte**
- ◎ **Consultor Datamining - BBVA**
- ◎ **Analista modelos de decisión- GFS**



g.fuentes@nablaea.com
<https://github.com/JGFuentesC>
<https://forrealanalytics.wordpress.com/>




2. Planteamiento del problema

Planteamiento del problema

A decorative network diagram in the top right corner, featuring a series of interconnected nodes and lines, resembling a social network or data structure.

Vivimos en una nueva era donde la digitalización de nuestras vidas es una realidad, las redes sociales han cambiado al mundo y la interacción que tenemos en estas plataformas en todos los ámbitos produce una cantidad masiva de datos listos para ser analizados y explotados. El gran reto para analizar estos datos está en su naturaleza no estructurada. Sentimientos, emociones, gustos se encuentran inmersos en textos, videos, imágenes, geo-referencias, etc.

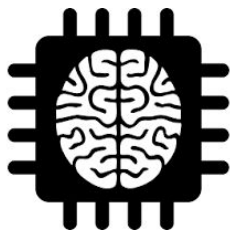
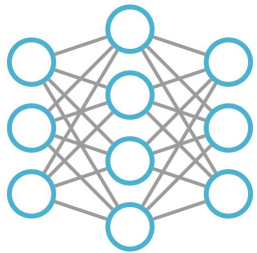
En este taller abordaremos de manera introductoria (Hello world) la explotación de dicha información mediante modelación matemática y computacional en el estado del arte.

A decorative network diagram in the bottom left corner, featuring a series of interconnected nodes and lines, resembling a social network or data structure.

A decorative network diagram in the top-left corner, featuring a complex web of interconnected nodes and lines. Some nodes are highlighted with blue circles, and others with blue dots. The lines are thin and gray, creating a mesh-like structure.

3. Tecnología Actual

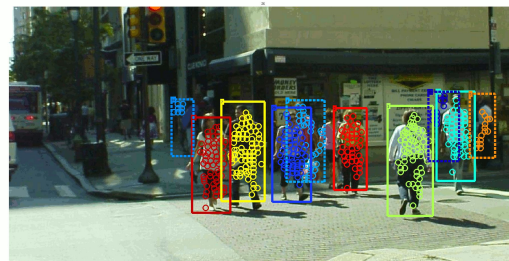
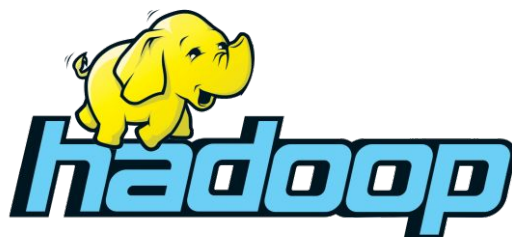
A decorative network diagram in the bottom-right corner, similar to the one in the top-left. It shows a network of nodes and lines, with some nodes highlighted by blue circles and others by blue dots.



TensorFlow



Google Cloud Platform



A decorative background featuring a network diagram. It consists of numerous nodes, represented by circles of varying sizes and shades of gray, connected by thin, light gray lines. Some nodes are highlighted with a blue outline, and a few are solid blue dots. The network is more densely packed on the left and right sides of the slide, with the central area being mostly white space containing the title.

4. Data Science

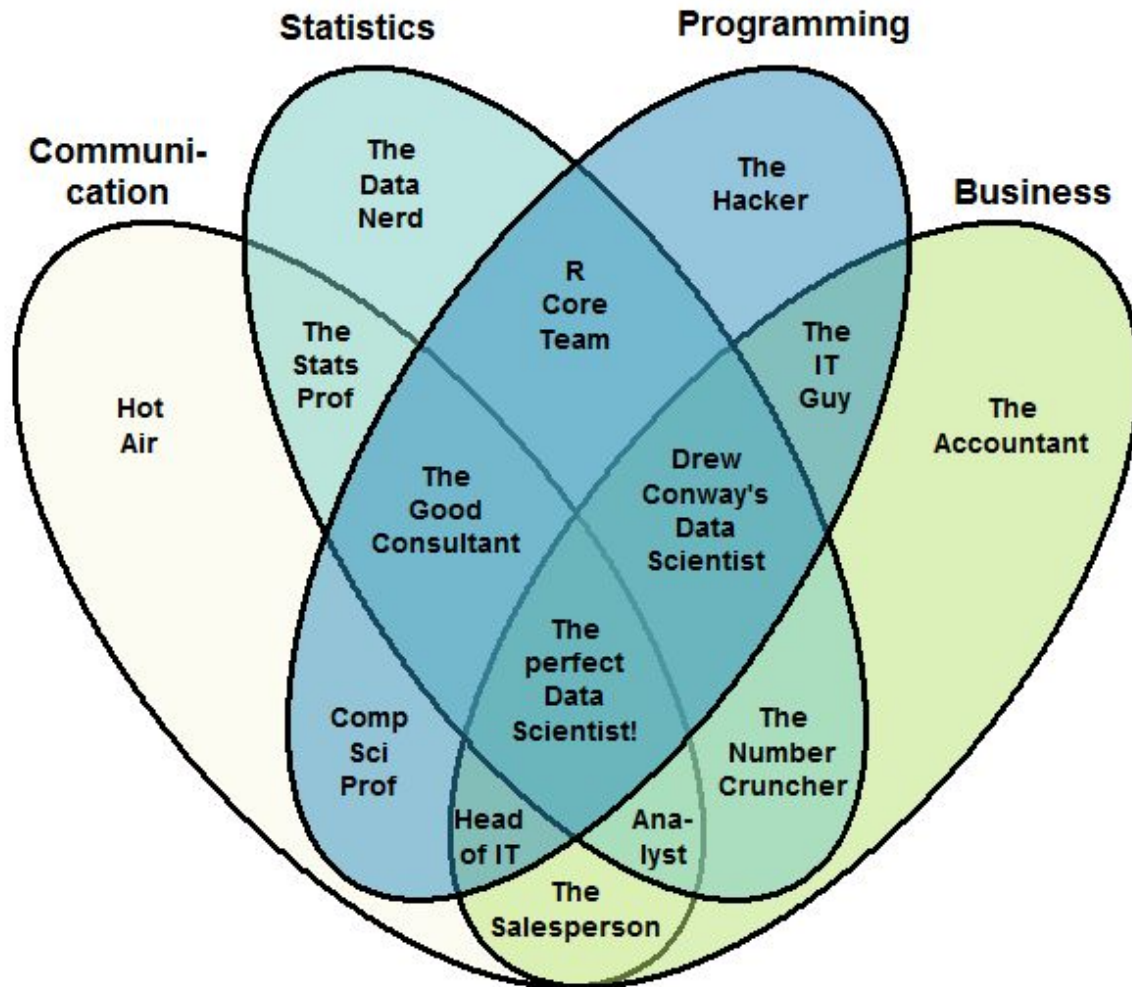
Data science

Data Science (Ciencia de datos) se considera una evolución multidisciplinaria en los campos de análisis de negocio, ciencias de la computación, modelación matemática, estadística, analítica y minería de datos.



Data science

The Data Scientist Venn Diagram



Data science

MODERN DATA SCIENTIST

Data Scientist, the sexiest job of 21st century requires a mixture of multidisciplinary skills ranging from an intersection of mathematics, statistics, computer science, communication and business. Finding a data scientist is hard. Finding people who understand who a data scientist is, is equally hard. So here is a little cheat sheet on who the modern data scientist really is.

MATH & STATISTICS

- ☆ Machine learning
- ☆ Statistical modeling
- ☆ Experiment design
- ☆ Bayesian inference
- ☆ Supervised learning: decision trees, random forests, logistic regression
- ☆ Unsupervised learning: clustering, dimensionality reduction
- ☆ Optimization: gradient descent and variants

DOMAIN KNOWLEDGE & SOFT SKILLS

- ☆ Passionate about the business
- ☆ Curious about data
- ☆ Influence without authority
- ☆ Hacker mindset
- ☆ Problem solver
- ☆ Strategic, proactive, creative, innovative and collaborative



PROGRAMMING & DATABASE

- ☆ Computer science fundamentals
- ☆ Scripting language e.g. Python
- ☆ Statistical computing package e.g. R
- ☆ Databases SQL and NoSQL
- ☆ Relational algebra
- ☆ Parallel databases and parallel query processing
- ☆ MapReduce concepts
- ☆ Hadoop and Hive/Pig
- ☆ Custom reducers
- ☆ Experience with xaaS like AWS

COMMUNICATION & VISUALIZATION

- ☆ Able to engage with senior management
- ☆ Story telling skills
- ☆ Translate data-driven insights into decisions and actions
- ☆ Visual art design
- ☆ R packages like ggplot or lattice
- ☆ Knowledge of any of visualization tools e.g. Flare, D3.js, Tableau

MarketingDistillery.com is a group of practitioners in the area of e-commerce marketing. Our fields of expertise include: marketing strategy and optimization; customer tracking and on-site analytics; predictive analytics and econometrics; data warehousing and big data systems; marketing channel insights in Paid Search, SEO, Social, CRM and brand.

Marketing
DISTILLERY



5. Software de trabajo

Software de Trabajo

En este taller, usaremos por comodidad la distribución Anaconda de python desarrollada por Continuum Analytics Inc. Se recomienda para principiantes ya que tiene (casi todas) las herramientas necesarias para Data Science preinstaladas y “Right out of the box”. Está disponible tanto para plataformas Basadas en Unix como para Windows.

Es necesario también tener cuenta en Facebook y Twitter para darse de alta como desarrollador y poder acceder a las API's de acceso a datos.



A decorative background featuring a network diagram with nodes and connecting lines. The nodes are represented by circles of varying sizes and colors (blue, grey, and white), and the lines are thin and grey. The diagram is positioned in the top-left and bottom-right corners of the slide.

6. Alta como desarrollador

Alta como desarrollador

Twitter: <https://dev.twitter.com/>, <https://apps.twitter.com/>

Facebook : <https://developers.facebook.com/>

Se requiere dar de alta una aplicación en cada una de las plataformas para obtener las credenciales de autenticación.

A decorative network diagram in the top-left corner, featuring a complex web of interconnected nodes and lines. Some nodes are highlighted with blue circles or dots, while others are grey.

7. Extracción de datos

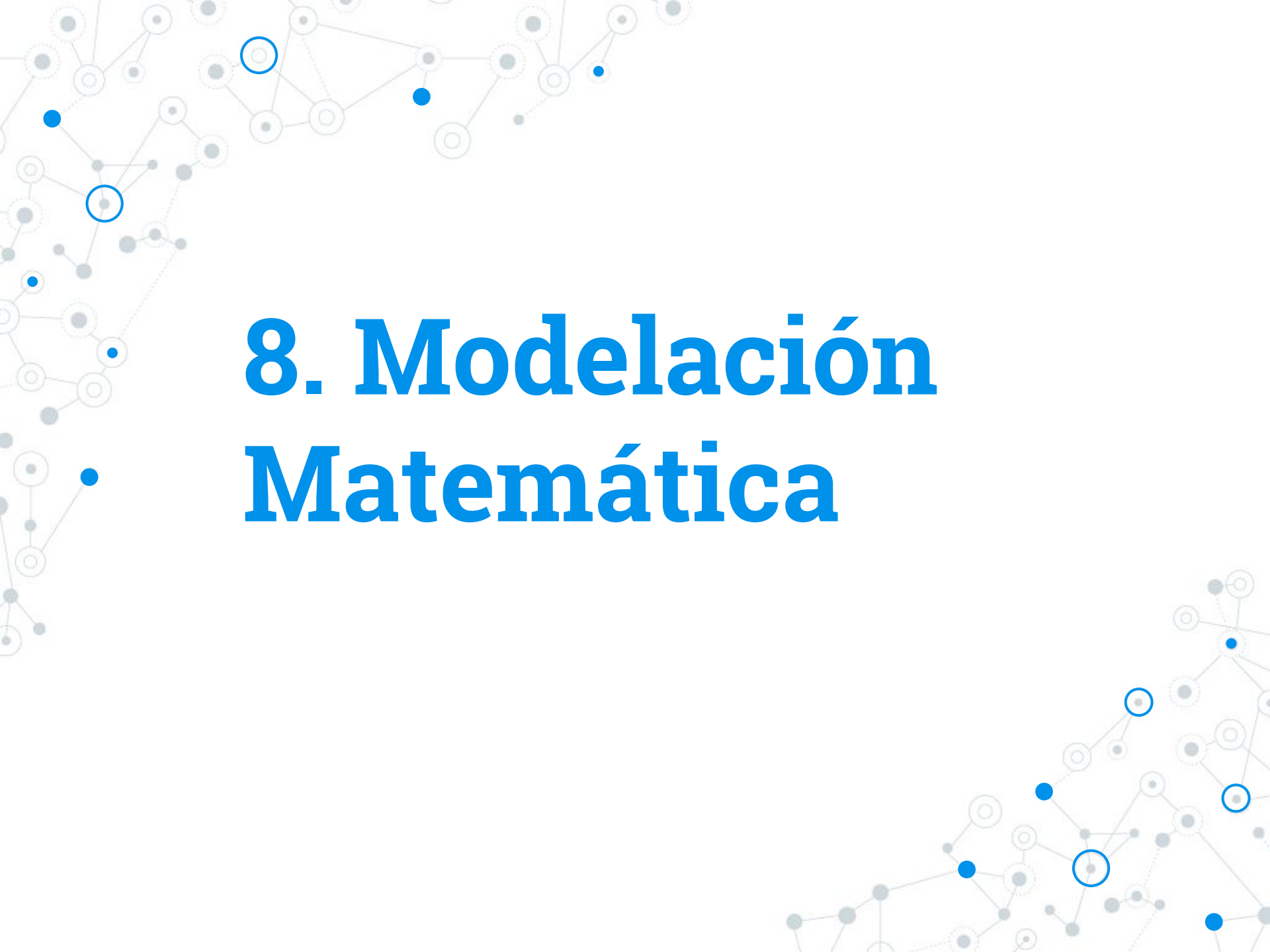
A decorative network diagram in the bottom-right corner, similar to the one in the top-left, with a web of nodes and lines, some highlighted in blue.

Extracción de datos Twitter

1. Autenticarse con tweepy
2. Requisitar user timeline o búsqueda al API
3. Organizar contenido de tweets
4. Análisis

Extracción de datos Facebook

1. Autenticarse
2. Requisitar al API información de una URL particular
3. Retorno en formato JSON
4. Atomizar JSON
5. Organizar
6. Análisis

A decorative background featuring a network diagram with nodes and connecting lines. Some nodes are highlighted with blue circles or dots. The diagram is composed of various sized circles connected by thin lines, creating a complex web-like structure.

8. Modelación Matemática

Modelación matemática

- © Modelación no supervisada para etiquetar contenido (Modelos Gaussianos Mixtos)
- © Modelación supervisada para predecir grupo de pertenencia de cada contenido (Naive Bayes Multinomial, Redes Neuronales, KNN, etc.)