

Capítulo 1: Fundamentos en Python

Capítulo 2: Machine Learning en Python

Capítulo 3: NLP en Python

# Machine Learning en Python

Data Science for Developer



# Objetivos

- Identificar las habilidades de un científico de datos.
- Reconocer los componentes de Machine Learning.
- Explicar una metodología de Data Science.
- Utilizar distintas herramientas de Python para Data Science.



# Agenda

- Definiciones de Data Science.
- Habilidades del Científico de Datos.
- Machine Learning (ML).
- Uso de Python en ML.
- Metodología de Ciencia de Datos.
- Herramientas de Python:
  - Scraping
  - Data Analysis
  - ML Models
  - Visualization
  - Data Product



# Algunas Definiciones de la Ciencia de Datos (Data Science)



La **ciencia de datos** es un campo interdisciplinario que involucra **métodos científicos, procesos y sistemas** para extraer conocimiento o un mejor entendimiento de datos en sus diferentes formas, ya sea estructurados o no estructurados.



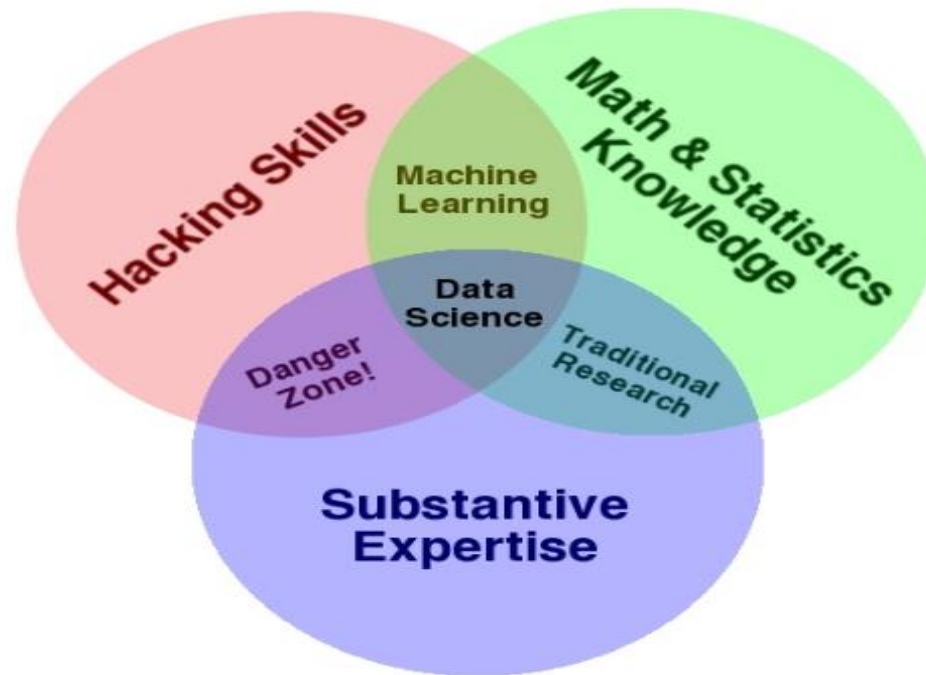
La **ciencia de datos** es un campo interdisciplinar que **combina Machine Learning, estadísticas, análisis avanzado y programación**. Es una nueva forma de arte que revela información oculta y saca el máximo partido de los datos en la era cognitiva.



La **ciencia de datos** involucra **métodos automatizados para analizar datos** y extraer conocimiento de ellos.



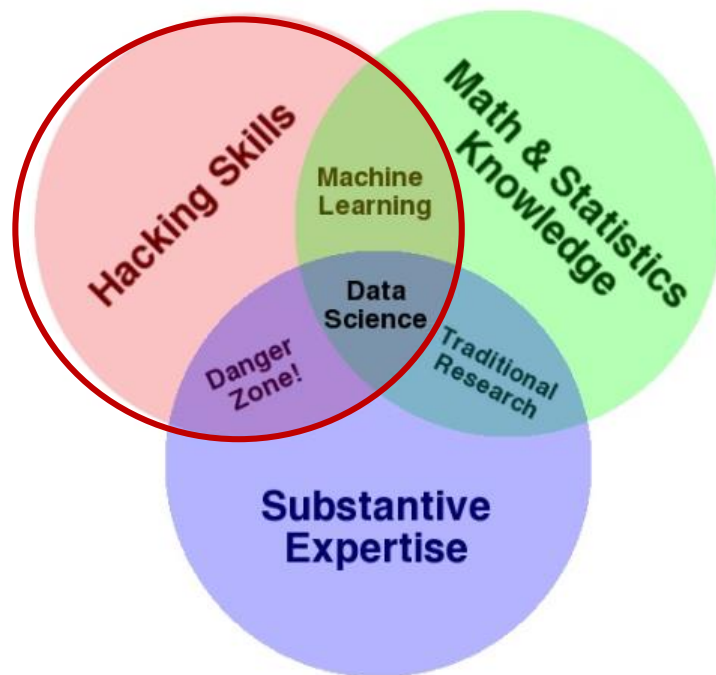
# Data Science se encuentra en la intersección de distintas materias



Fuente: drewconway.com



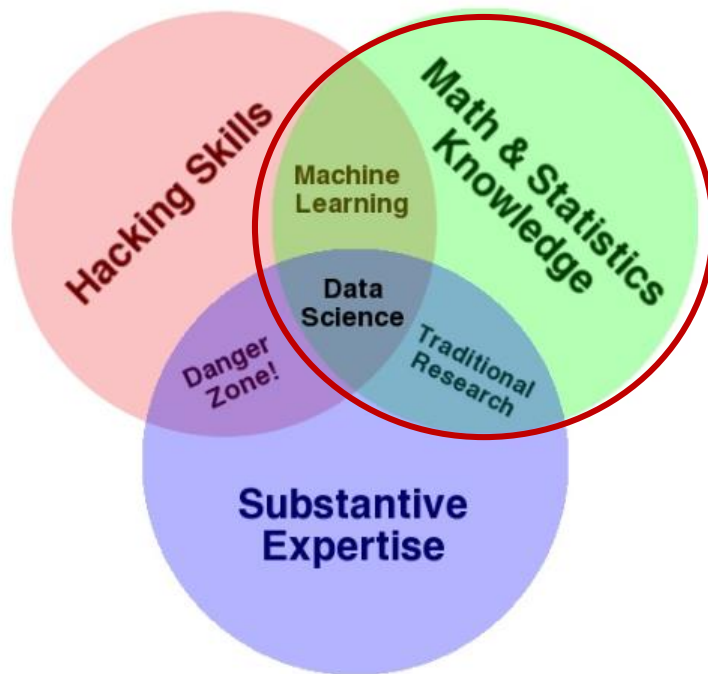
# Hacking Skills



- Programación que incluye entre otros lenguajes.
- C
- C++
- Java
- .Net
- Spark
- Scala
- Javascript
- R
- Python



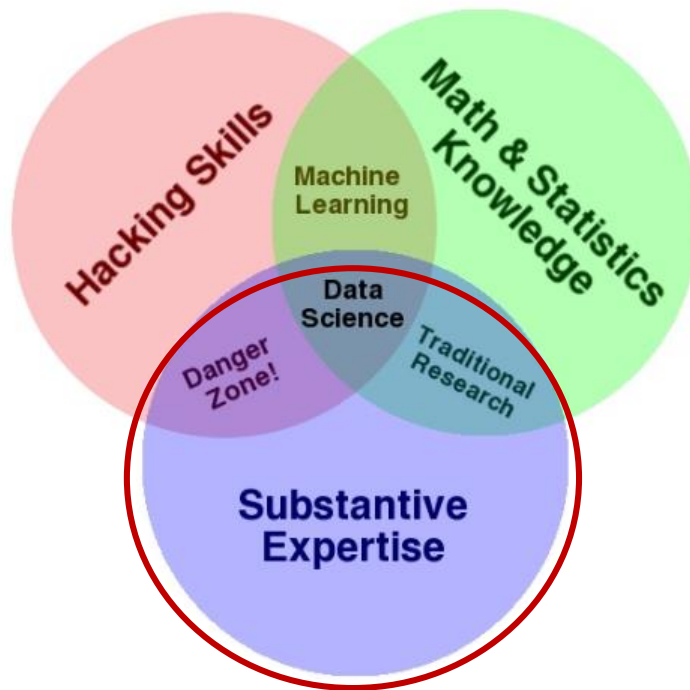
# Matemáticas y Estadística



- Matemáticas
  - Álgebra lineal
  - Cálculo
  - Probabilidades
- Estadística
  - Descriptiva
  - Inferencia
  - Distribuciones
  - Correlación
  - Test de hipótesis
  - Etc.



# Conocimiento del Negocio



- Consultoría
- Negociación
- Storytelling
- Presentación
- Visualización de datos
- Gestión de proyectos
- Finanzas





# Es sexy ser científico de datos

## Data Scientist: *The Sexiest Job of the 21st Century*

**Meet the people who  
can coax treasure out of  
messy, unstructured data.**

*by Thomas H. Davenport  
and D.J. Patil*

**W**hen Jonathan Goldman arrived for work in June 2006 at LinkedIn, the business networking site, the place still felt like a start-up. The company had just under 8 million accounts, and the number was growing quickly as existing members invited their friends and colleagues to join. But users weren't seeking out connections with the people who were already on the site at the rate executives had expected. Something was apparently missing in the social experience. As one LinkedIn manager put it, "It was like arriving at a conference reception and realizing you don't know anyone. So you just stand in the corner sipping your drink—and you probably leave early."

70 Harvard Business Review October 2012



# El científico de datos posee muchas habilidades

## 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

### 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

### DOMAIN KNOWLEDGE & SOFT SKILLS

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

### COMMUNICATION & VISUALIZATION

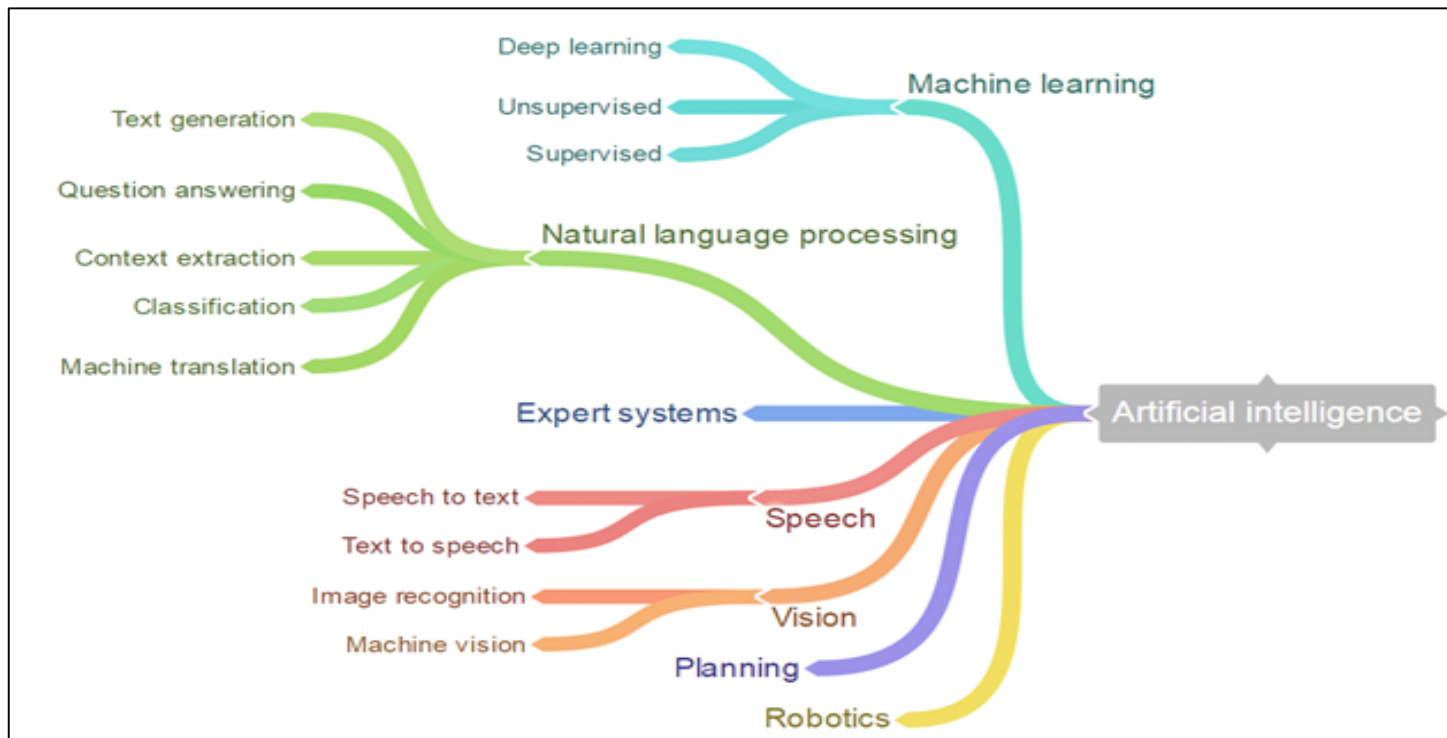
- ☆ Able to engage with senior management
- ☆ Story telling skills
- ☆ Translate data-driven insights into decisions and actions
- ☆ Visual art design
- ☆ R packages like ggplot or lattice
- ☆ Knowledge of any of visualization tools e.g. Flare, D3.js, Tableau

**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**



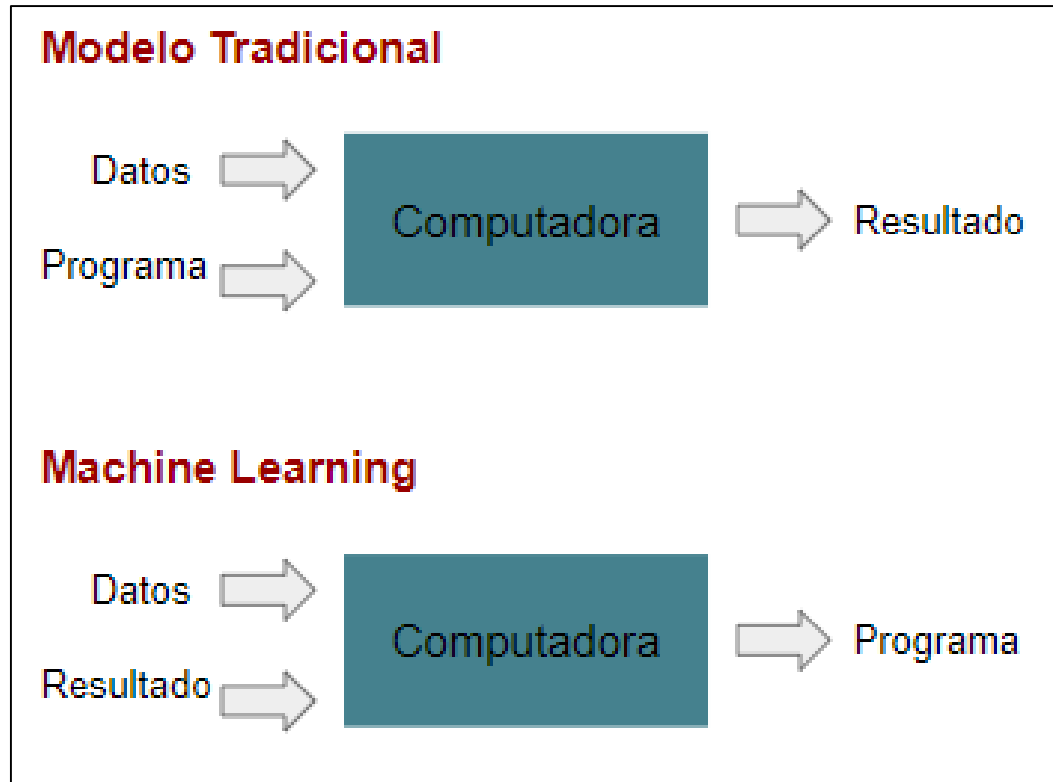
# Machine Learning es Inteligencia Artificial



Fuente: <https://hackernoon.com/jump-start-to-artificial-intelligence-f6eb30d624ec>



# Machine Learning permite aprender de los Datos



# ML se usa en varias Industrias



Machine Learning applications across industries

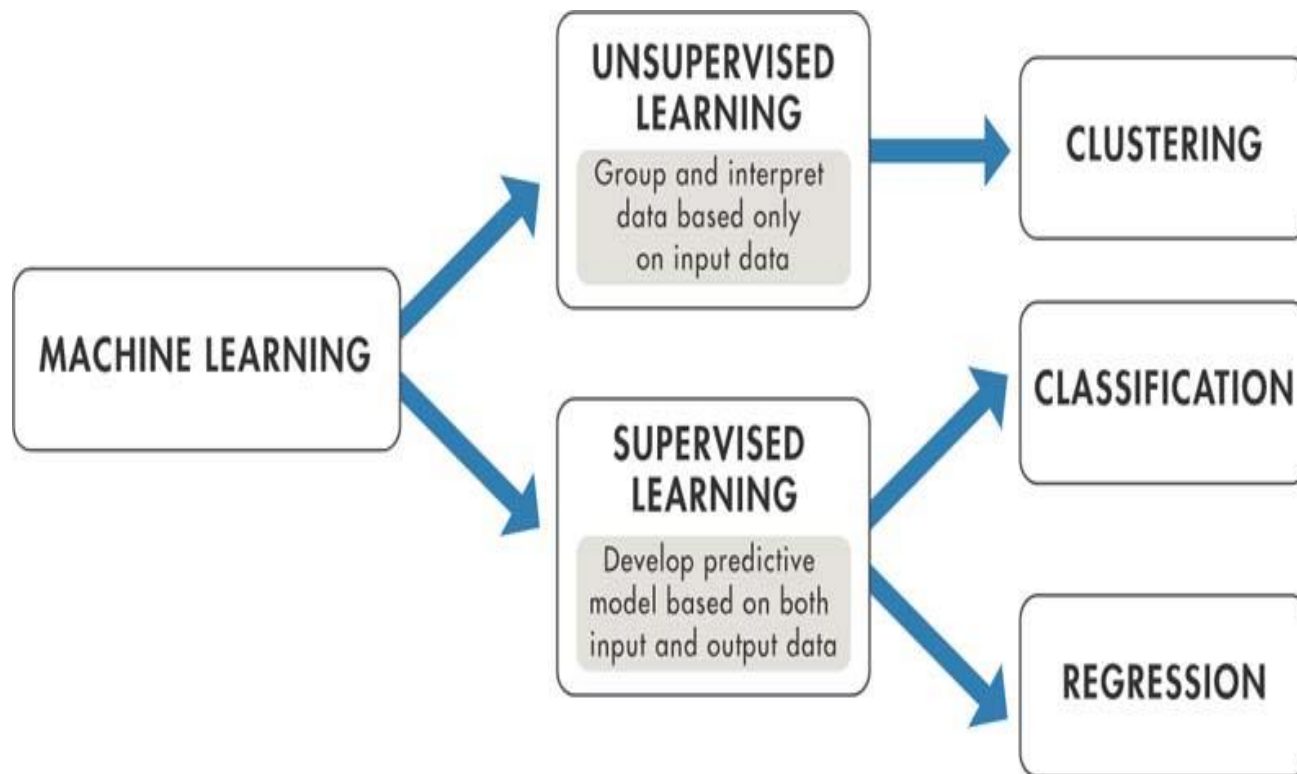
- Calificación de Demanda
- Análisis de Speech to Text
- Generación automática documentos
- Búsqueda inteligente de documentos
- Predicción de sentencias

**Judicial**

Fuente: <https://becominghuman.ai/lets-talk-about-advanced-analytics-a-brief-look-at-artificial-intelligence-bf1c7a7d3f96>



# Machine Learning: Supervisado vs No Supervisado



Fuente: <https://ecmapping.com/2018/02/21/the-10-machine-learning-algorithms-to-master-for-beginners>





# Machine Learning: Clasificación, Regresión y Clustering

Categoría ¿Pertenece a A o B?	Número ¿25,000 soles?	Agrupación ¿Hay relación de datos?
<b>CLASSIFICATION</b>	<b>REGRESSION</b>	<b>CLUSTERING</b>
Support Vector Machines	Linear Regression, GLM	K-Means, K-Medoids Fuzzy C-Means
Discriminant Analysis	SVR, GPR	Hierarchical
Naive Bayes	Ensemble Methods	Gaussian Mixture
Nearest Neighbor	Decision Trees	Neural Networks
	Neural Networks	Hidden Markov Model

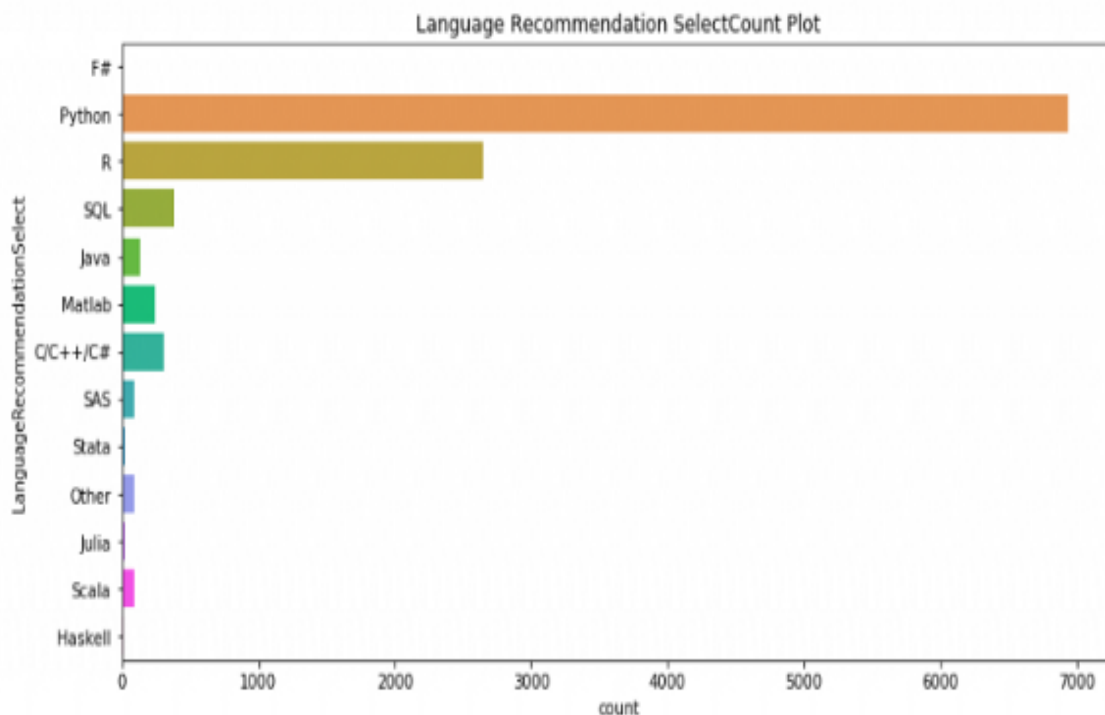
Fuente: <https://ecmapping.com/2018/02/21/the-10-machine-learning-algorithms-to-master-for-beginners/>



# Uso de Python en ML

```
In [54]: plt.figure(figsize=(12,6))  
plt.title('Language Recommendation SelectCount Plot')  
sns.countplot(data=mc, y='LanguageRecommendationSelect')
```

```
Out[54]: <matplotlib.axes._subplots.AxesSubplot at 0x7f289330aba8>
```



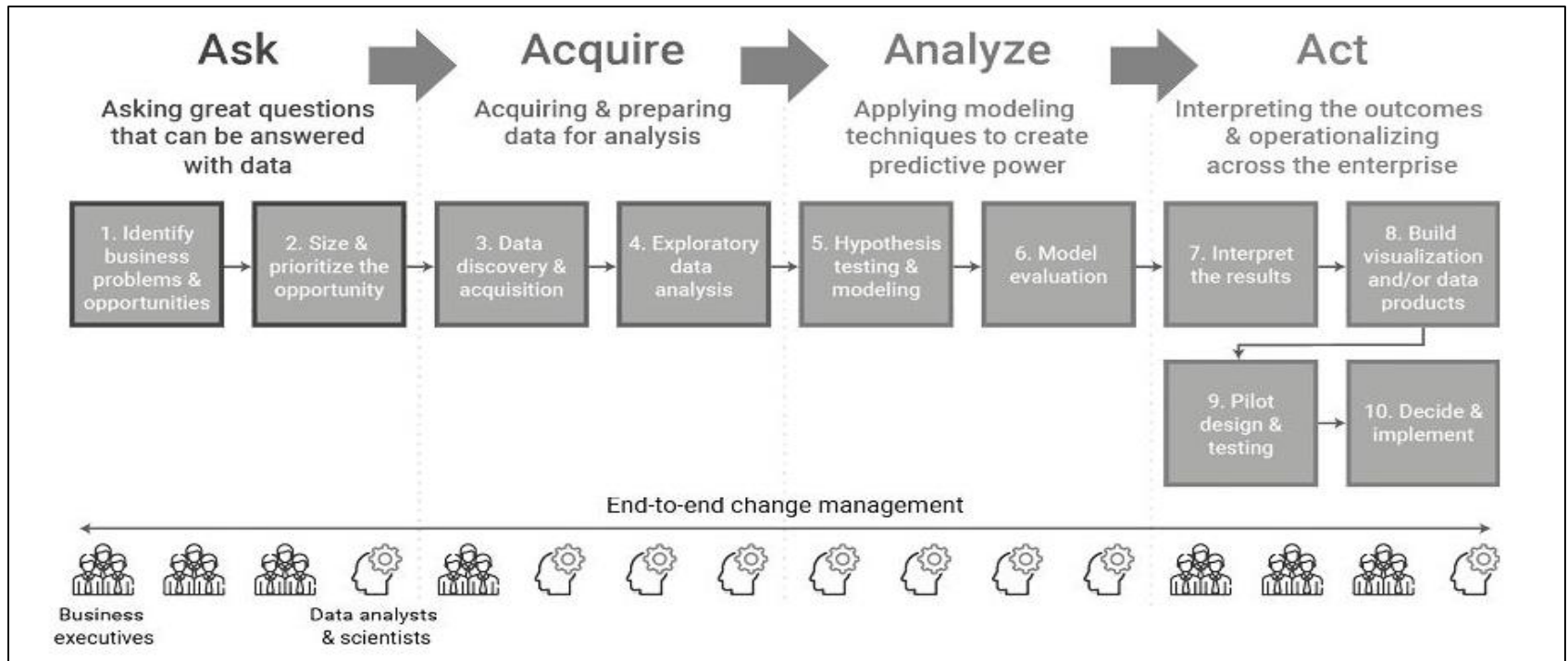
## Algunas Razones

- Rapid prototyping
- Lenguaje sencillo
- Uso de librerías (packages)
- Scikit-learn
- Pandas
- Open source
- Comunidad





# Metodología de Data Science



Fuente: Kaldero "Data Science for Executives", 2018 Lioncrest



# Herramientas de Python: Scraping

- Adquisición de datos.
- Scraping - obtención de datos por sitios web.
- Bots de Scrapy permiten atravesar páginas de HTML e interpretar los tags con xpath.
- Descarga en formatos CSV, JSON.
- Servicios de scraping: scrapinghub.



# Herramientas de Python: Data Analysis

- Exploratory Data Analysis.
- Data Frames (hoja de cálculo en memoria).
- Operaciones con los datos.



# Herramientas de Python: ML Models

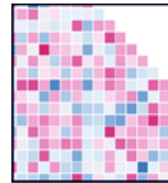
- Hypothesis Testing & Model Evaluation.
- Machine Learning.
- NLP.
- Algoritmos supervisados y no supervisados.



# Herramientas de Python: ML Models

- Build visualization.
- Distintos tipos de gráficos.
- Se puede interactuar con las estructuras de datos.

matplotlib



Seaborn



# Herramientas de Python: Data Product

- Data products & pilot design.
- Workbooks – código interactivo.
- Desarrollo de web apps.



**django**



## Ejercicio N° 2: Titanic

Al finalizar el laboratorio, el alumno logrará:

- Demostrar competencias básicas en machine learning usando el caso Titanic.



# Resumen

En este capítulo, usted aprendió que:

- La ciencia de datos requiere competencias en diversas materias.
- Machine learning, como rama de la inteligencia artificial, tiene un número de usos en la industria.
- Python cubre con sus herramientas varios elementos del proceso de ciencia de datos.





# Tarea N° 2: Algoritmos de Machine Learning

Investigar algoritmos de machine learning y enviar un informe de dos hojas máximo.

- Elegir un algoritmo
  - Regresión lineal
  - Árboles de decisión
- Contexto histórico
- Conceptos básicos
- Funcionamiento
- Fortalezas y debilidades

Enviar por correo al instructor.

