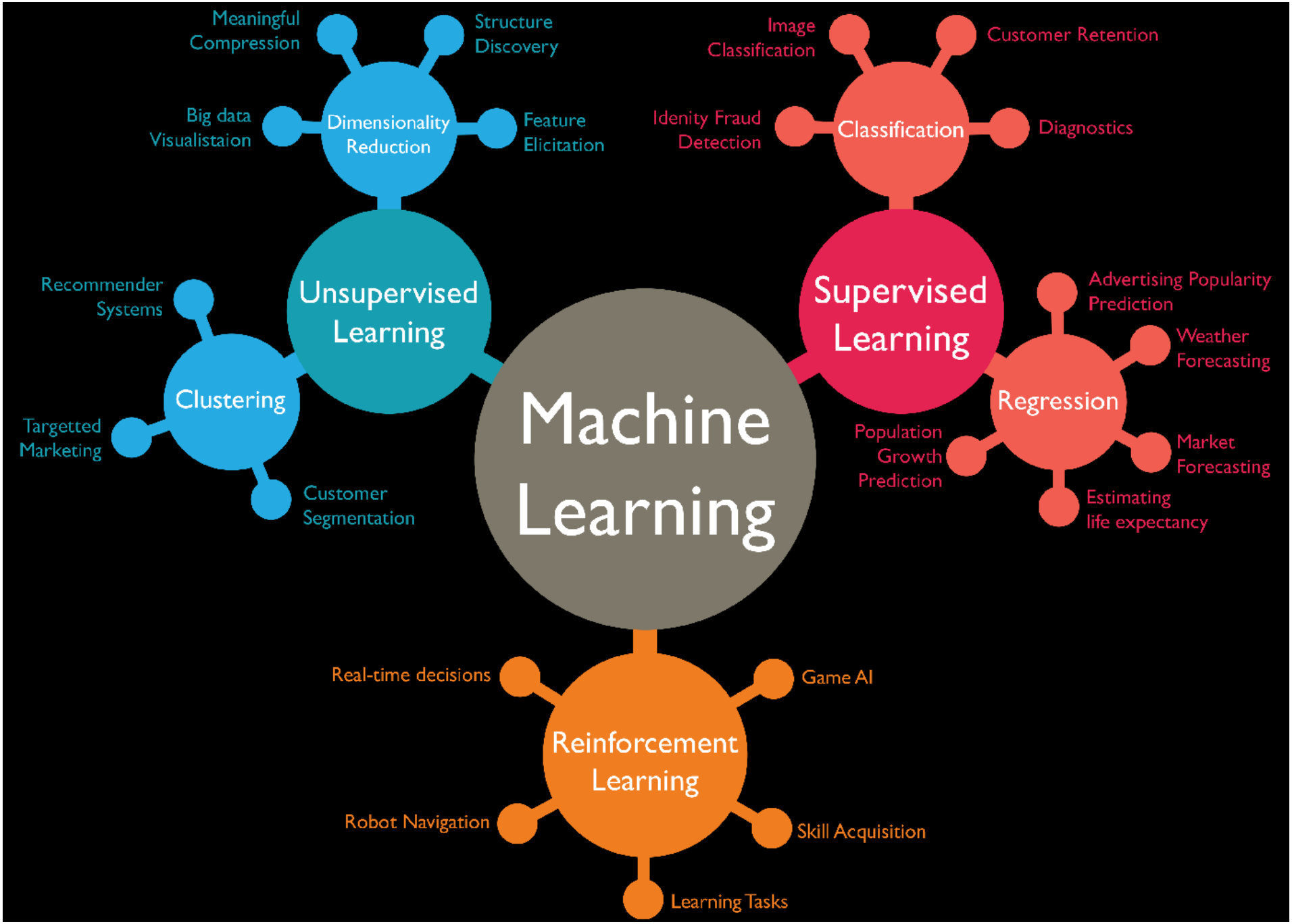
Machine Learning.
Introducción

Machine learning

- Expresión de años 60 derivada de la inteligencia artificial.
- Se pretendía crear modelos y herramientas para que las máquinas "aprendieran" por sí solas.
- La traducción habitual al español es "aprendizaje automático".

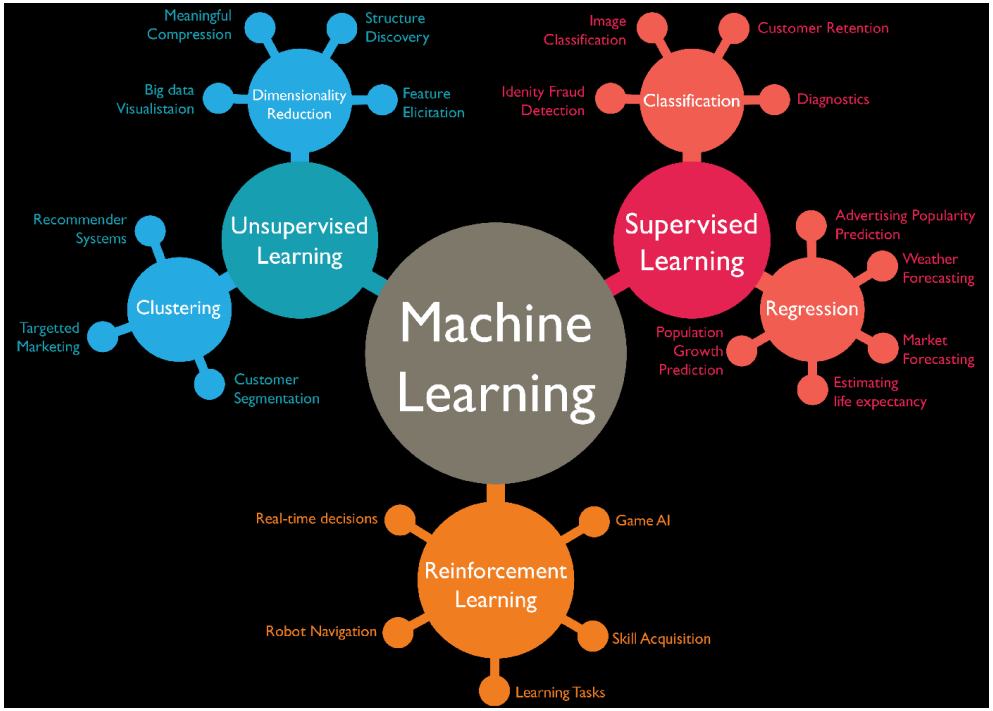

Machine Learning
Unsupervised Learning
Dimensionality Reduction
Meaningful Compression
Structure Discovery
Big data Visualistaion
Feature Elicitation
Clustering
Recommender Systems
Targetted Marketing
Customer Segmentation
Supervised Learning
Classification
Image Classification
Customer Retention
Idenity Fraud Detection
Diagnostics
Regression
Advertising Popularity Prediction
Weather Forecasting
Population Growth Prediction
Market Forecasting
Estimating life expectancy
Reinforcement Learning
Real-time decisions
Game AI
Robot Navigation
Skill Acquisition
Learning Tasks

Prosaicamente, Machine Learning es una batería de herramientas-algoritmos para solucionar:

1) Problemas predictivos supervisados: predicción de variable continua (regresión), predicción de variable binaria, o multiclase (clasificación).

Algoritmos más conocidos y eficientes:

- Regresión
- Regresión logística
- Redes neuronales
- Árboles de decisión, Random forest y gradient boosting
- Support Vector Machines (SVM)



Ejemplos básicos en Business Analytics

- Predicción de ventas (regresión)
- Predicción de fuga de clientes-churn (clasificación)
- Predicción de fraude (clasificación)

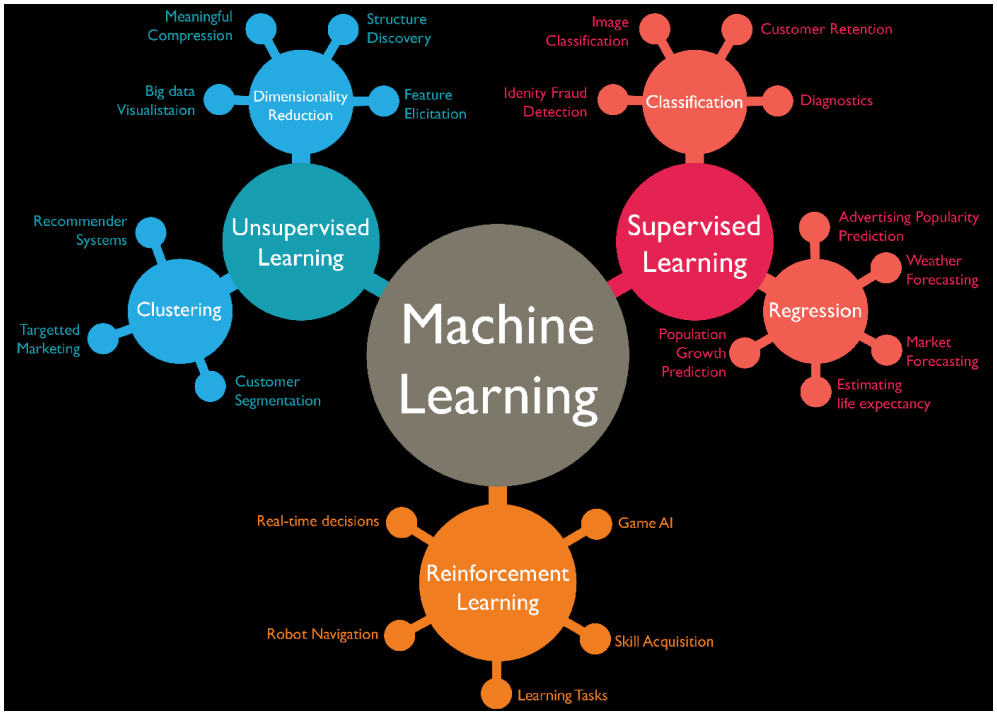
2) Problemas predictivos no supervisados: clustering, segmentación y reducción de dimensiones, búsqueda de estructuras

Algoritmos más conocidos y eficientes:

- Algoritmos de clustering (k-means, jerárquicos, etc.)
- Análisis factorial, análisis de correspondencias

Ejemplos básicos en Business Analytics

- Clustering y segmentación de clientes
- Análisis de elección discreta
- Estructura de variables de venta

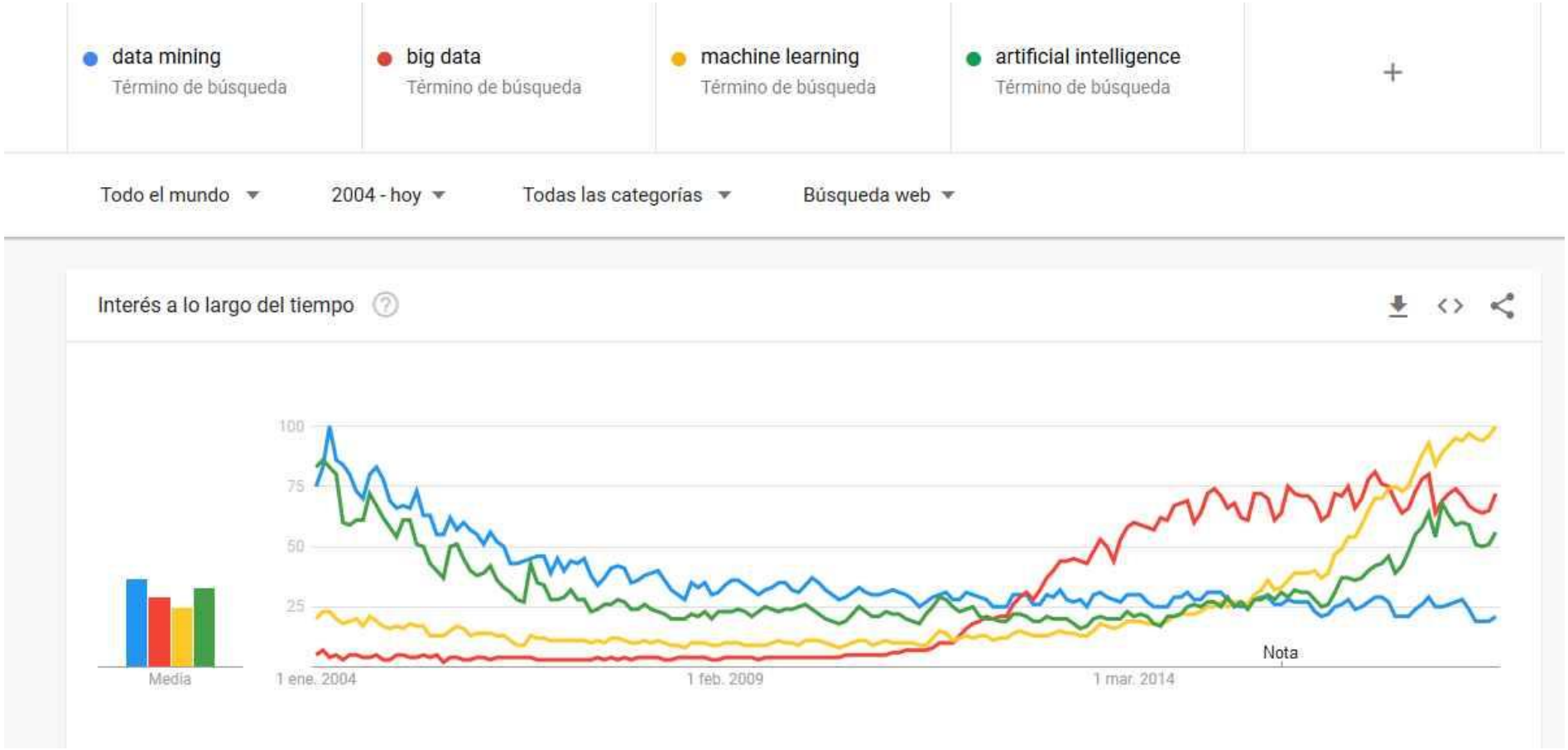


Nota

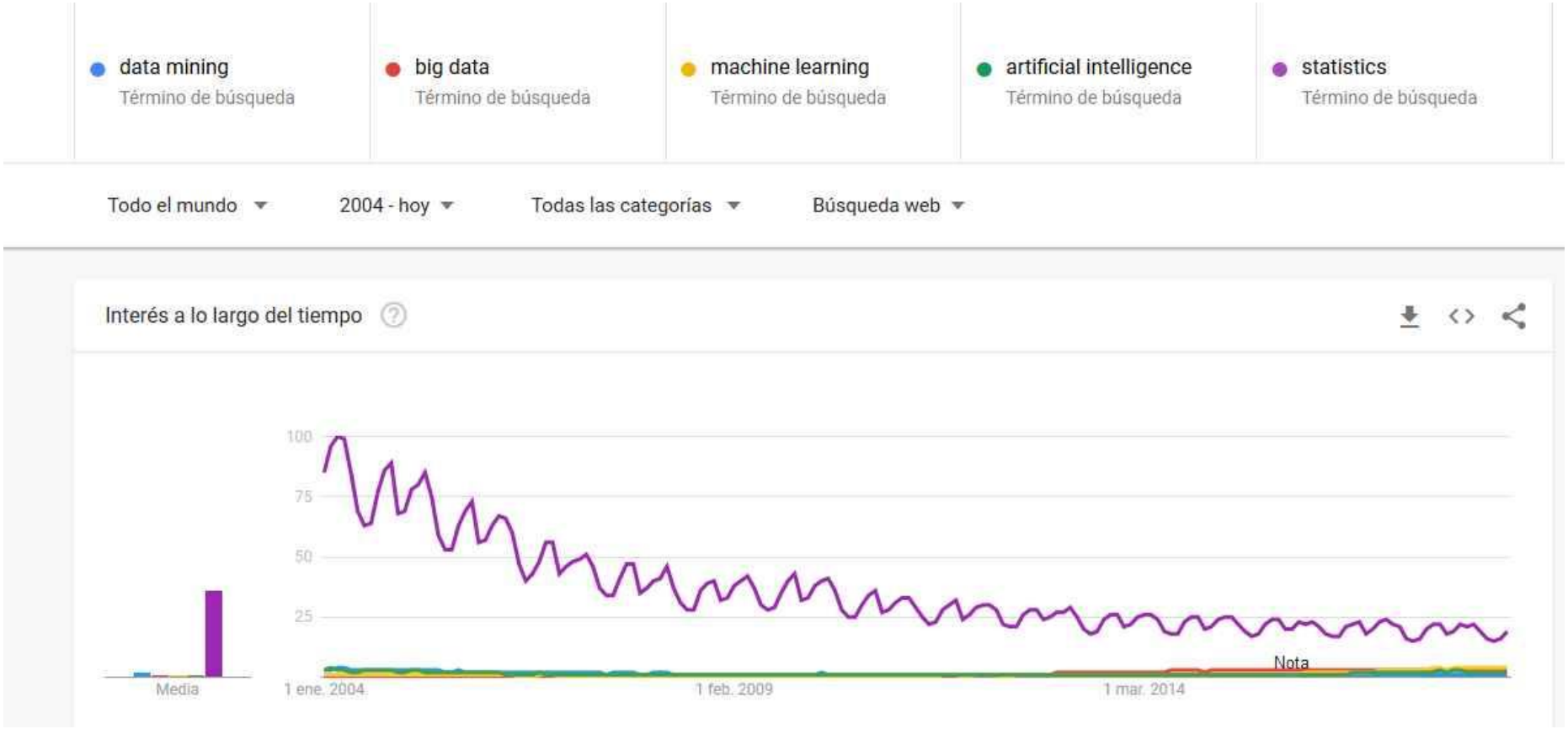
Es una gran contradicción que las técnicas de predicción de series temporales o el tratamiento de datos temporales-espaciales no aparecen o no estén tradicionalmente englobados en el conjunto de técnicas pertenecientes al ámbito del machine learning.

Es probable que sea debido a que los modelos estadísticos clásicos todavía son muy competitivos en este ámbito.

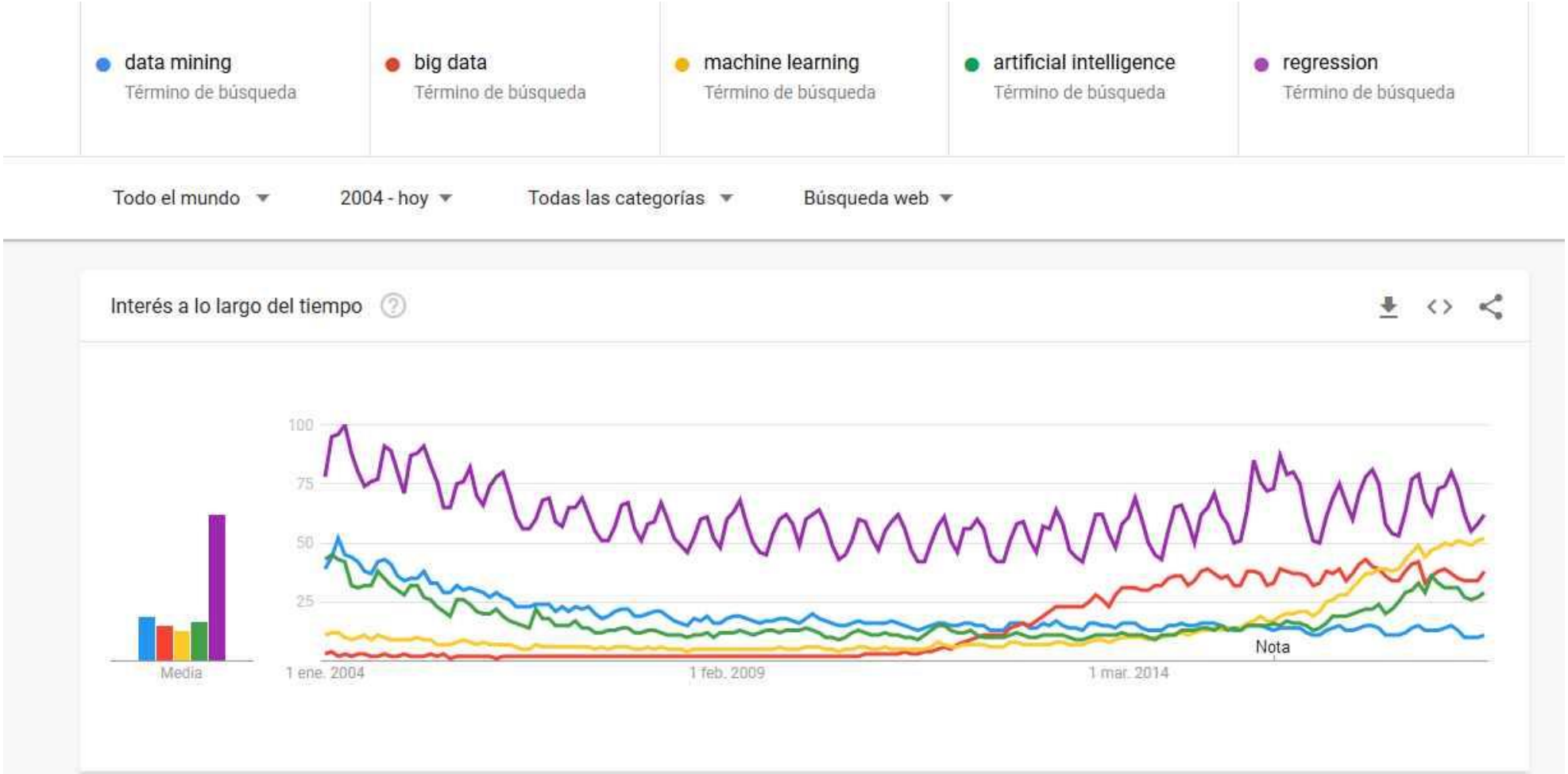
Google Trends: búsquedas de 2004 hasta hoy: **machine learning** está de moda



El término **Statistics** es más buscado, pero porque se busca en otros contextos (tablas de estadística pública tipo INE, por ejemplo)



Si ponemos el modelo estadístico más popular, **Regression**, se aprecia mejor la comparación.



Software más utilizado para Machine learning

Lenguajes puros de programación más utilizados

- R
- Python
- Java
- C, C++
- SQL

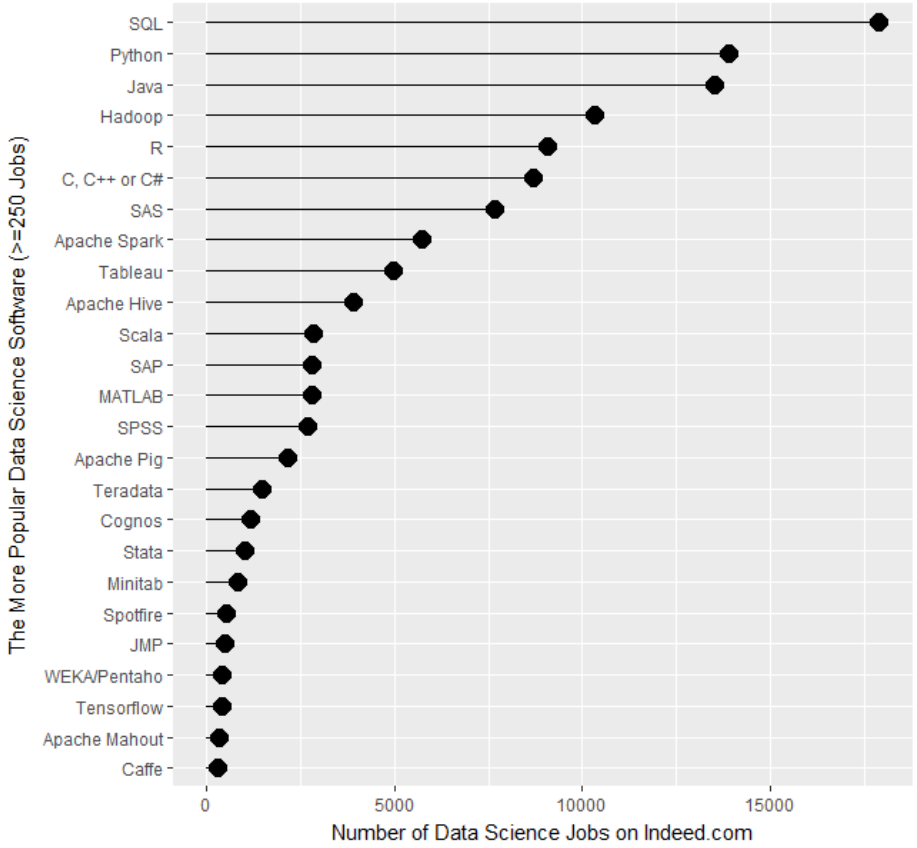
Entornos o paquetes-plataformas, en general con lenguaje de programación integrado o con compilador de otros lenguajes (en cursiva los comerciales)

- *SAS*
- *SPSS*
- *MATLAB*
- *EXCEL*
- Spark
- H2o
- Rapidminer
- Knime
- Tableau

http://r4stats.com/articles/popularity/

Software requerido en ofertas de empleo en Data Science

2017



2019

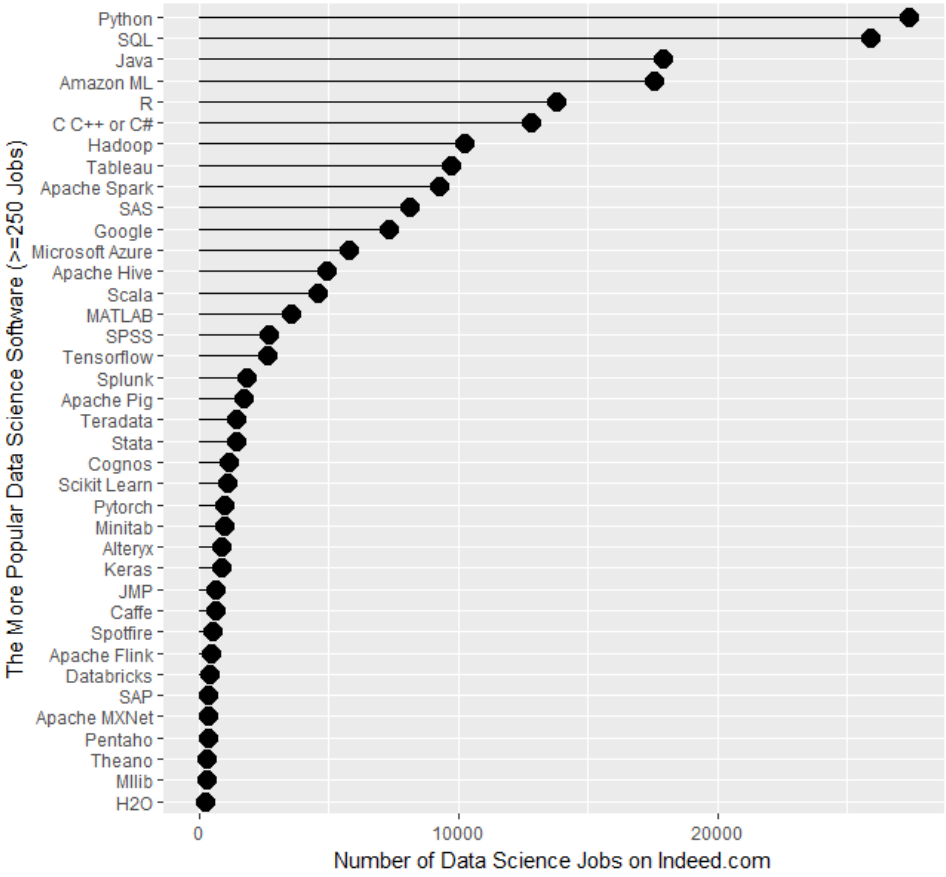
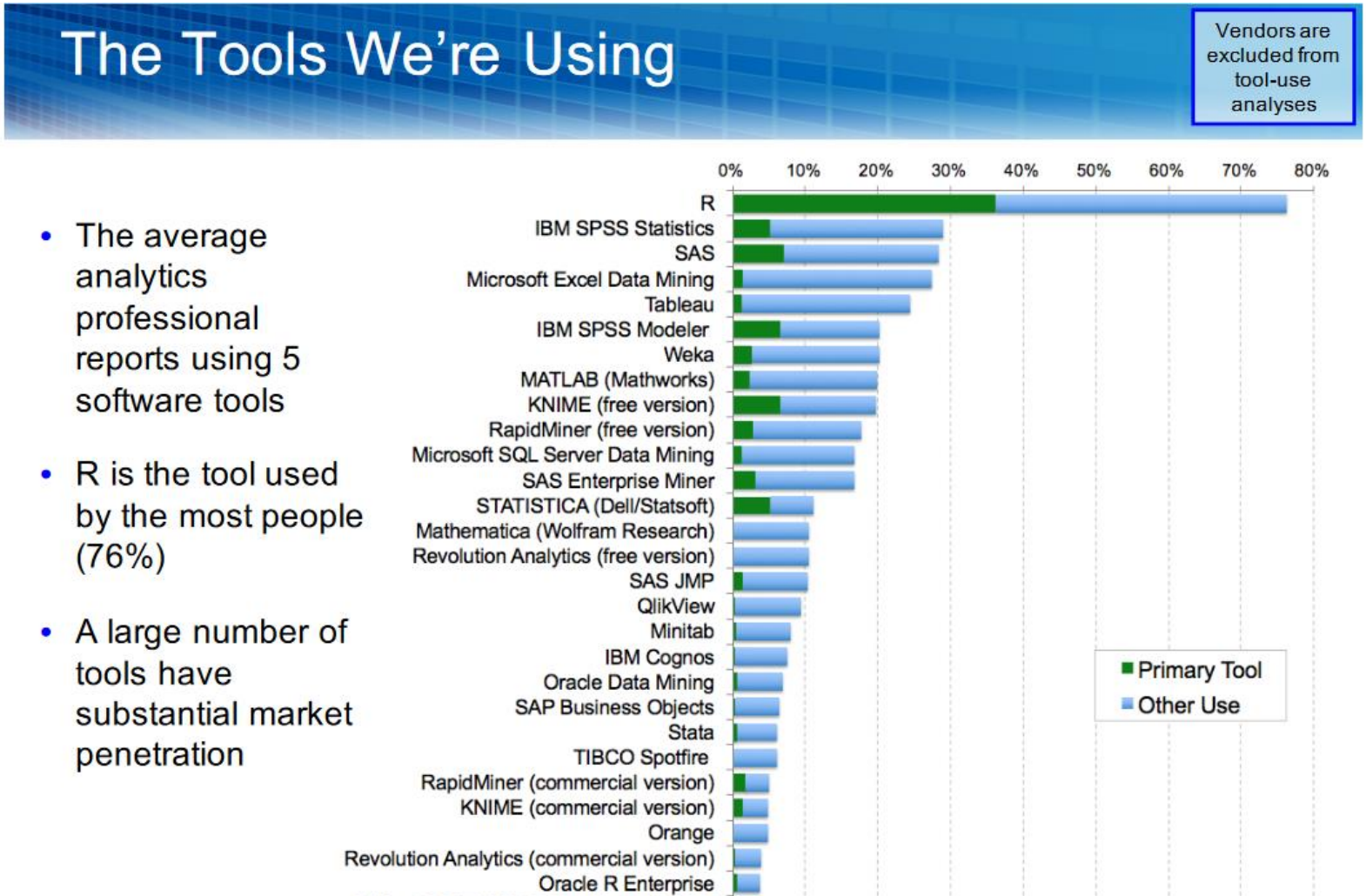


Figure 1a. The number of data science jobs for the more popular software (those with 250 jobs or more, 2/2017).

Datos de 2015-16, tomados de Rexer Analytics survey
Encuesta a usuarios



Datos tomados de KDnuggets

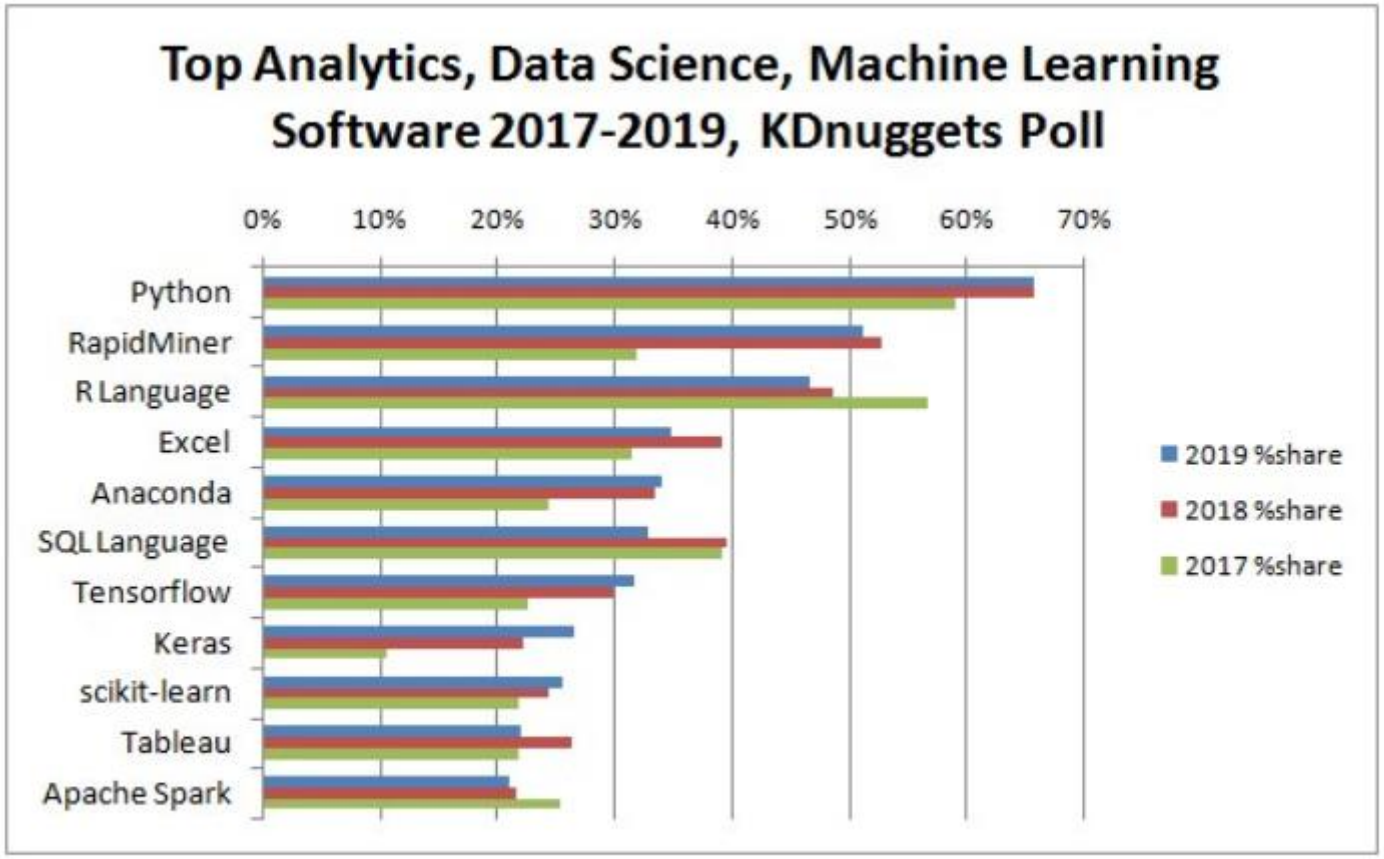
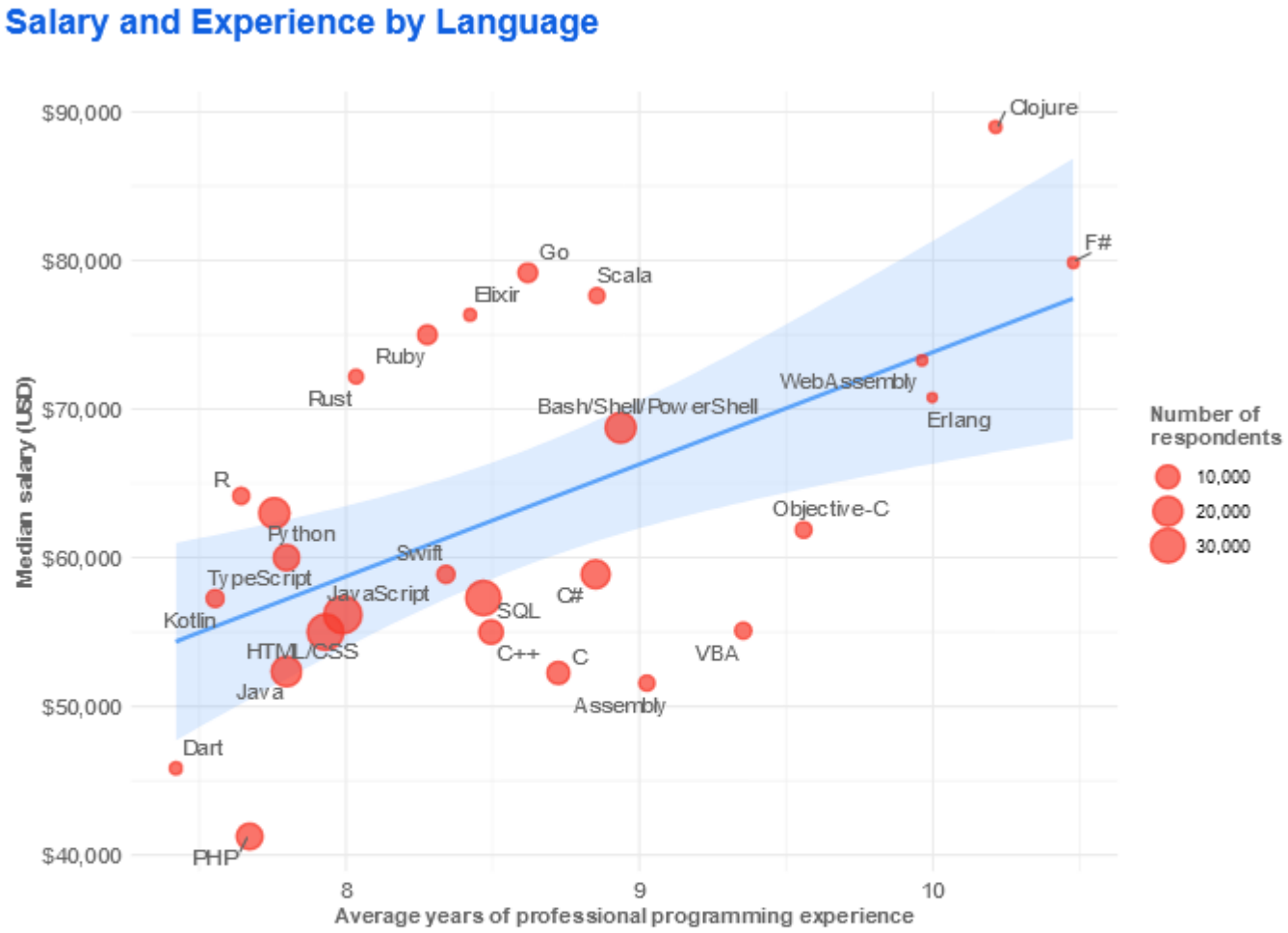


Fig 1: KDnuggets Analytics/Data Science 2019 Software Poll: top tools in 2019, and their share in the 2017, 2018 polls

Table 1: Top Analytics/Data Science/ML Software in 2019 KDnuggets Poll

Software	2019 % share	2018 % share	2017 % share
Python	**65.8%**	65.6%	59.0%
RapidMiner	**51.2%**	52.7%	31.9%
R Language	**46.6%**	48.5%	56.6%
Excel	**34.8%**	39.1%	31.5%
Anaconda	**33.9%**	33.4%	24.3%
SQL Language	**32.8%**	39.6%	39.2%
Tensorflow	**31.7%**	29.9%	22.7%
Keras	**26.6%**	22.2%	10.7%
scikit-learn	**25.5%**	24.4%	21.9%
Tableau	**22.1%**	26.4%	21.8%
Apache Spark	**21.0%**	21.5%	25.5%

Encuesta Stack Overflow developers 2019



Developers using languages that appear above the line in this chart, such as Clojure, Scala, Go, Rust, and R, are being paid more even given how much experience they have. Developers using languages below the line, like PHP, Assembly, and VBA, however, are paid less even given years of experience. The size of the circles in this chart represents how many developers are using that language compared to the others.

Temario

1. Introducción al Machine Learning y modelización predictiva avanzada.Conceptos generales.

2. Redes neuronales.

2.1 Introducción general y aplicaciones. Planteamiento del modelo para variable continua. Parametrización y estimación. Ejemplos básicos con R.

2.2 Ajuste y sobreajuste. Técnicas de remuestreo.

2.3 Linealidad y no linealidad. Monitorización del número de nodos.

2.4 Funciones de combinación y activación. Técnicas de optimización más frecuentes.

2.5 Regularización y otras técnicas de control del sobreajuste. Preliminary training y early stopping.

2.6 El problema de la selección de variables en redes

2.7 Redes para variable dependiente binaria

2.7 Redes con R. Paquetes, ejemplos y código. El intercambio sesgo-varianza. Aplicaciones y correcta evaluación del modelo.

3. Deep Learning (este tema se dejaría para el final)

3.1 Problemas generales con las redes
3.2 CNN y conceptos fundamentales asociados: pooling, dropout, dropconnect, ReLu, Maxout
3.3 Autoencoders y RBM
3.4 LSTM y RNN
3.5 Aplicaciones y actualidad

5. Bagging y Random Forest

5.1 Bagging. Conceptos básicos. Monitorización. Variable dependiente continua y binaria.
5.2 Random Forest. Monitorización. Ventajas y desventajas.
5.3 Aplicaciones con R

6.Gradient Boosting

6.1 Conceptos básicos. Monitorización. Variable dependiente continua y binaria.
6.2 Ventajas y desventajas.
6.3 Aplicaciones con R. glm y xgboost.
6.4 Historia reciente de RF y GB

7 SVM, No free lunch theorem

7.1 SVM. Construcción y aplicaciones con R
7.2 Comparación de algoritmos: No free lunch Theorem

8 Métodos de ensamblado

8.1 Introducción y ejemplos. Variable dependiente continua y binaria.
8.2 Justificación teórica básica
8.3 Técnicas de ensamblado sencillas
8.4 Desarrollo con R y recomendaciones prácticas

Evaluación

1) Un trabajo largo de comparación de algoritmos predictivos sobre una variable dependiente binaria, utilizando redes, algoritmos basados en árboles (RF; GBM), SVM, y métodos de ensamblado.

Bibliografía básica

- **FAQ que cubre todos los aspectos importantes de las redes neuronales (son 7 FAQ html):**

  ftp://ftp.sas.com/pub/neural/FAQ.html

- **Referencia clásica**

  - Bishop, C.M. (1995), Neural Networks for Pattern Recognition, Oxford: Oxford University Press.

- **Con SAS, bastante bueno**

  - Neural Network Modeling using SAS Enterprise Miner. Randall Matignon (2005)

- **Libros disponibles en PDF para Gradient Boosting, SVM y otros algoritmos**
-

  - Hastie, Tibshirani: The Elements of Statistical Learning (PDF)

  (En la web hay más información)
  http://statweb.stanford.edu/~tibs/ElemStatLearn/
  - Hastie, Tibshirani: An Introduction to Statistical Learning with Applications in R (PDF)

  (básicamente el mismo que el anterior, pero para R)
  http://www-bcf.usc.edu/~gareth/ISL/data.html/
  - Machine Learning, Neural and Statistical Classification D. Michie, D.J. Spiegelhalter, C.C. Taylor (PDF)

- **Recopilación de amplia bibliografía sobre redes comentada y bien estructurada**

ftp://ftp.sas.com/pub/neural/FAQ4.html#questions