Jornada de Innovación

Big Data y Análisis del Aprendizaje

Dr. Esteban García-Cuesta

Profesor Asociado en el Departamento de Ciencias y Tecnologías de la Información

Universidad Europea de Madrid

«la llegada de grandes volúmenes de datos y la inteligencia de la máquina en todas partes»

Eric Schmidt, 2013 (Predictions for 2014 Bloomberg interview)

Cantidad

1247,6 millones de euros

Internet & WWW

Variedad

Volumen

Tiempo real

Big Data

Análisis del aprendizaje

Predicción de opinión

Marketing online

Inteligencia Artificial

Datos ubicuos

"Se estima que para 2017 el mercado de este sector alcance los 50.000 millones de dólares en todo el mundo" ***(Periódico ABC Mayo, 2013)***

Impacto social

Recomendación

Primera aparición del término Big Data

Application-Controlled Demand Paging for Out-of-Core Visualization

1997

Michael Cox
MRJ/NASA Ames Research Center
Microcomputer Research Labs, Intel Corporation
<mbc@nas.nasa.gov>

David Ellsworth
MRJ/NASA Ames Research Center
<ellswort@nas.nasa.gov>

Abstract

In the area of scientific visualization, input data sets are often very large. In visualization of Computational Fluid Dynamics (CFD) in particular, input data sets today can surpass 100 Gbytes, and are expected to scale with the ability of supercomputers to generate them. Some visualization tools already partition large data sets into segments, and load appropriate segments as they are needed. However, this does not remove the problem for two reasons: 1) there are data sets for which even the individual segments are too large for the largest graphics workstations, 2) many practitioners do not have access to workstations with the memory capacity required to load even a segment, especially since the state-of-the-art visualization tools tend to be developed by researchers with

1 Introduction

Visualization provides an interesting challenge for computer systems: data sets are generally quite large, taxing the capacities of main memory, local disk, and even remote disk. We call this the problem of *big data*. When data sets do not fit in main memory (*in core*), or when they do not fit even on local disk, the most common solution is to acquire more resources. This *write-a-check* algorithm has two drawbacks. First, if visualization algorithms and tools are worth developing, then they are worth deploying to more production-oriented scientists and engineers who may have on their desks machines with significantly less memory and disk. Some researchers have noted that their software tools were not used in practice for several years after development because the tools required more power and memory than were available on the

MRJ Technology Solutions at NASA Ames Research Center

Primera aparición del término Big Data

Big Data - Ahora

1997

2014

Esteban García-Cuesta – Departamento de Ciencias Y Tecnologías de la Información

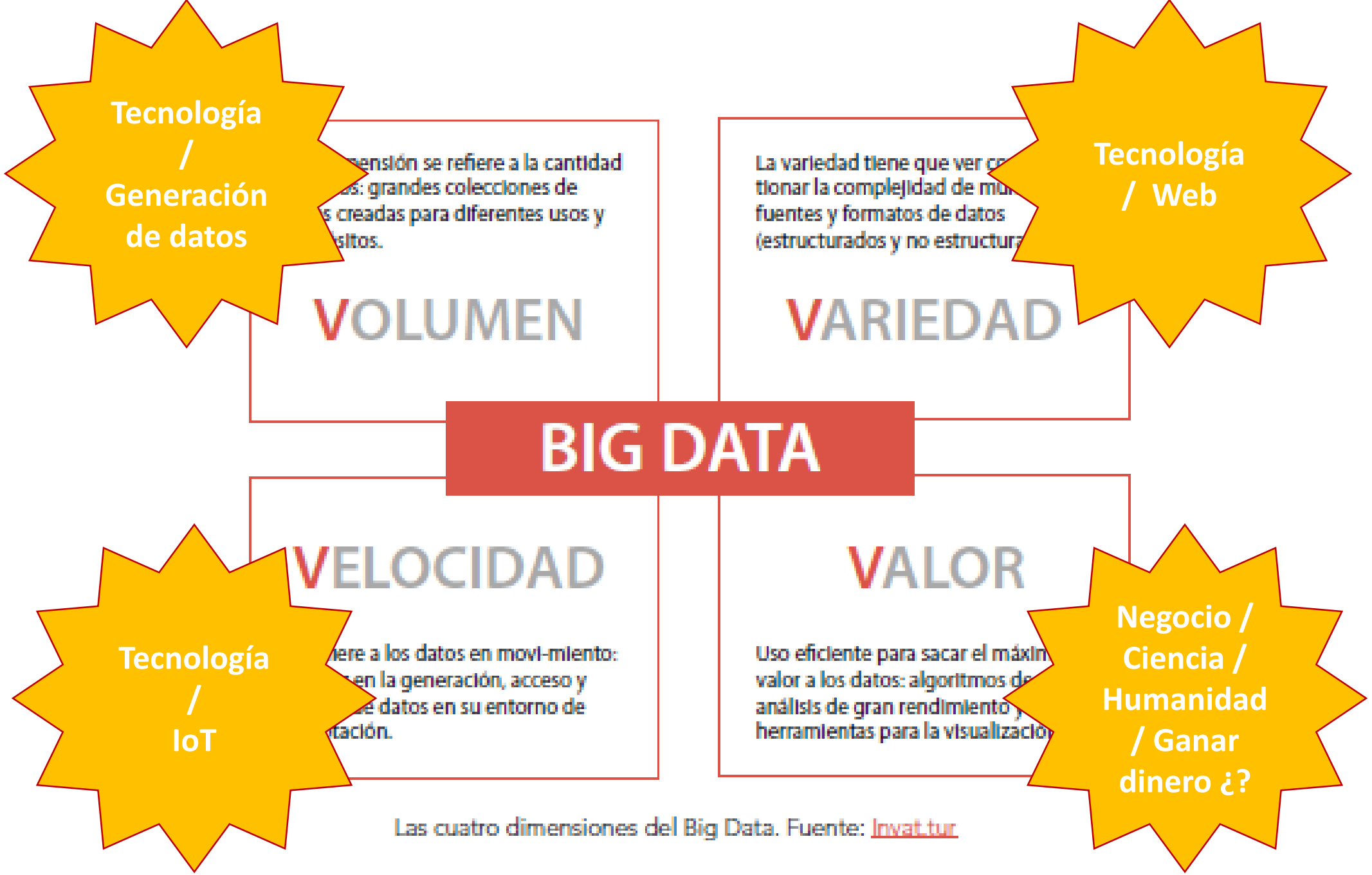
Mayor variedad

http://blog.qmee.com/wp-content/uploads/2013/07/Qmee-Online-In-60-Seconds21.png

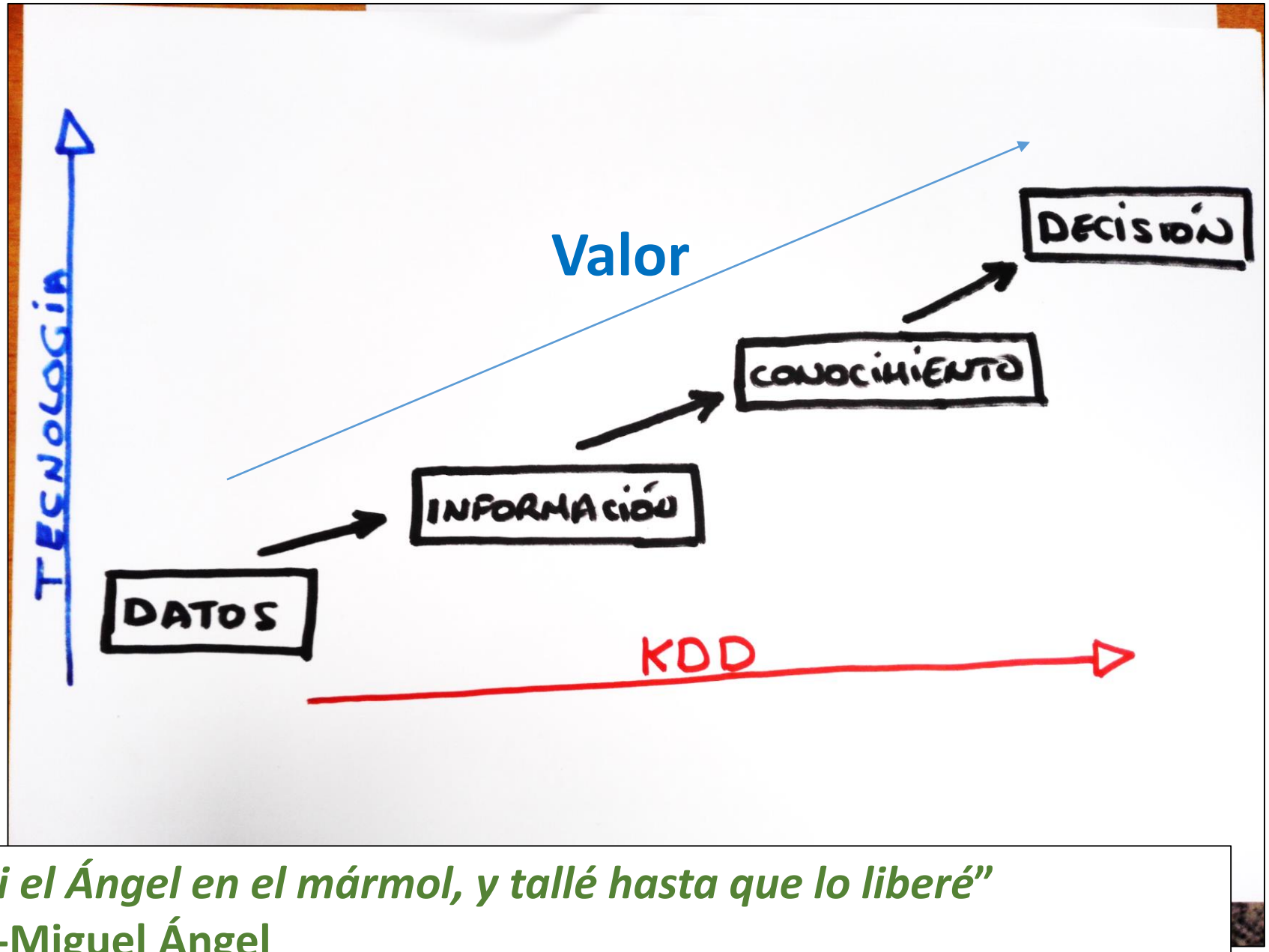
Mayor volumen

Mayor velocidad

Mayor valor

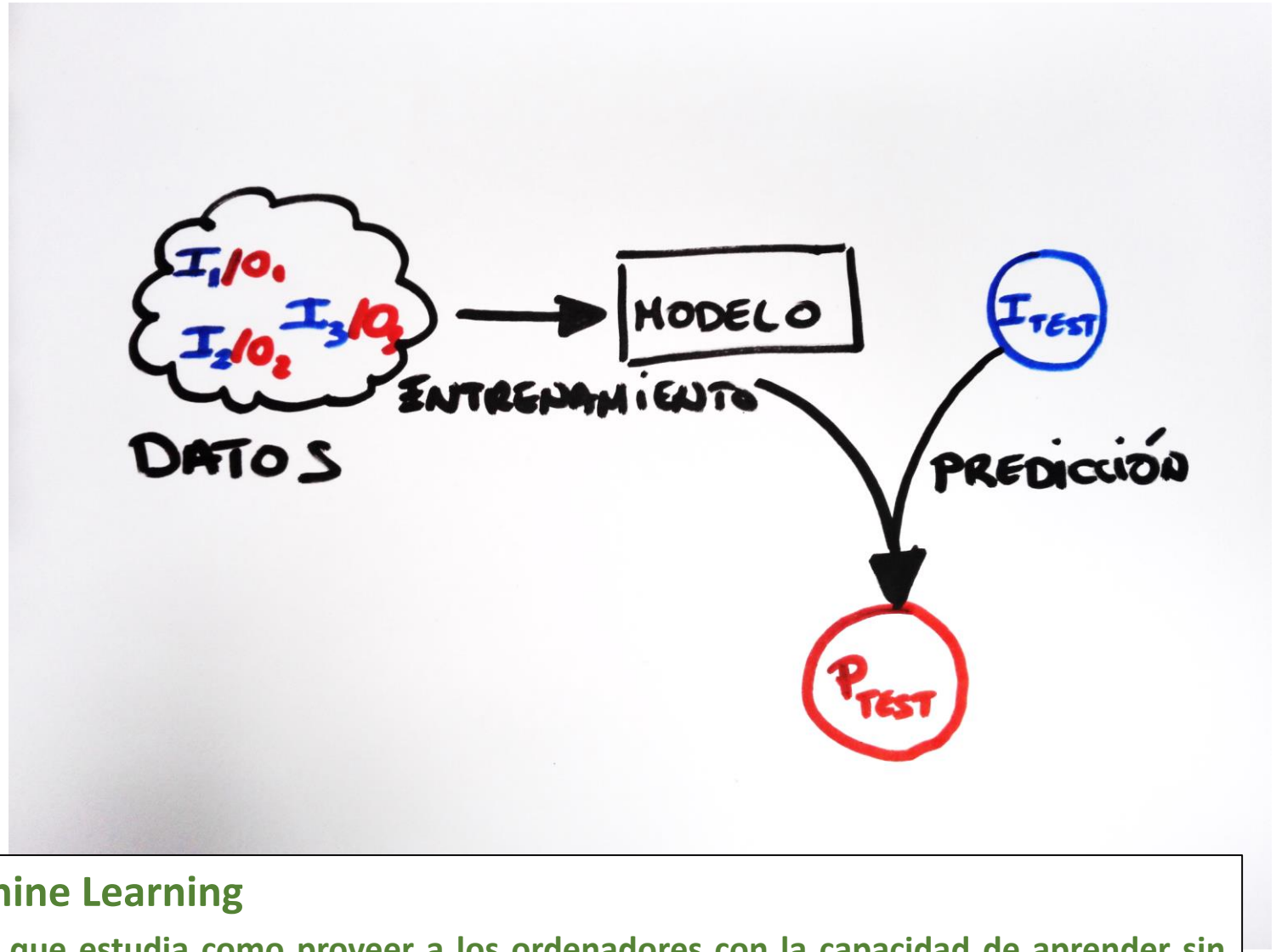


Las cuatro dimensiones del Big Data. Fuente: Invat.tur



1996

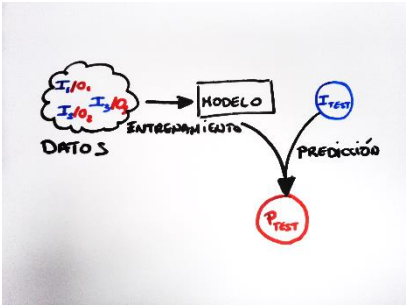
"Vi el Ángel en el mármol, y tallé hasta que lo liberé"
--Miguel Ángel



Machine Learning

"Área que estudia como proveer a los ordenadores con la capacidad de aprender sin ser programados de manera explícita. " Arthur Smauel

Machine Learning



- Dada una gran cantidad de datos → Descubrir patrones y modelos que son:
  - **Válidos:** manteniendo nuevos datos con cierta credibilidad
  - **Usables**: siendo posible actuar sobre el elemento
  - **Inesperados**: no siendo obvios para el sistema
  - **Inteligibles**: los humanos deberían ser capaces de interpretar el patrón

Big Data en el entorno educativo

Esteban García-Cuesta – Departamento de Ciencias Y Tecnologías de la Información

Cosas a tener en cuenta…

- **Enseñar es complicado**
- **Alcanzar el conocimiento y curriculum es más complicado**
- **Mejorar es todavía más complicado**

Cosas a tener en cuenta…

- Enseñar es un **balance entre ciencia y arte**
  - No es necesario ser disruptiva
  - No tiene que cambiar constantemente

- **Tecnología y desarrollar software también tiene parte de arte**
  - Es muy disruptiva
  - Cambio constante

640K estudiantes

6 millones de cuestionarios

coursera

14 millones de videos

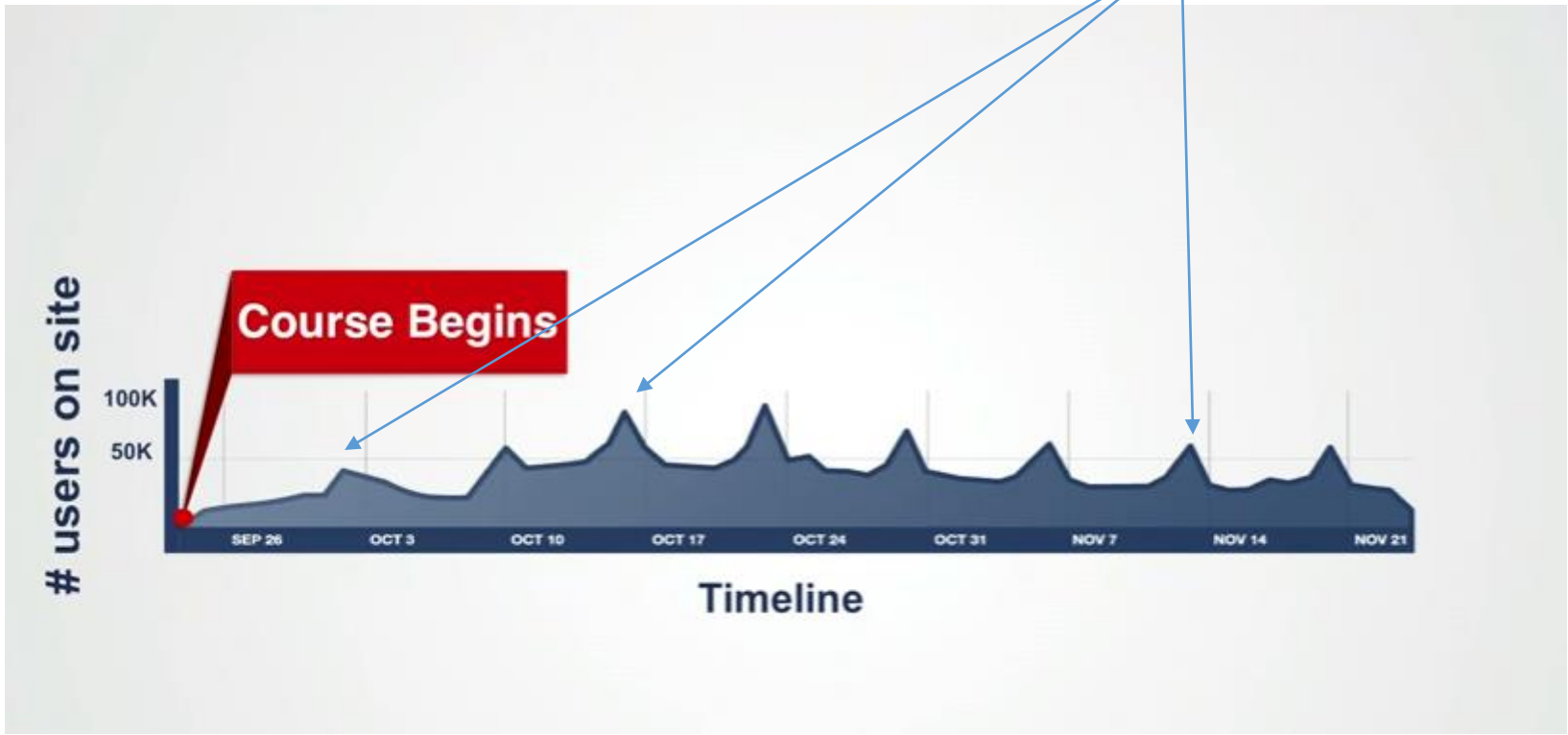
190 países

Educación a escala

Esteban García-Cuesta – Departamento de Ciencias Y Tecnologías de la Información

coursera

Entregas planificadas



Daphne Koller; Ted Talk “What we’re learning from online education”

coursera

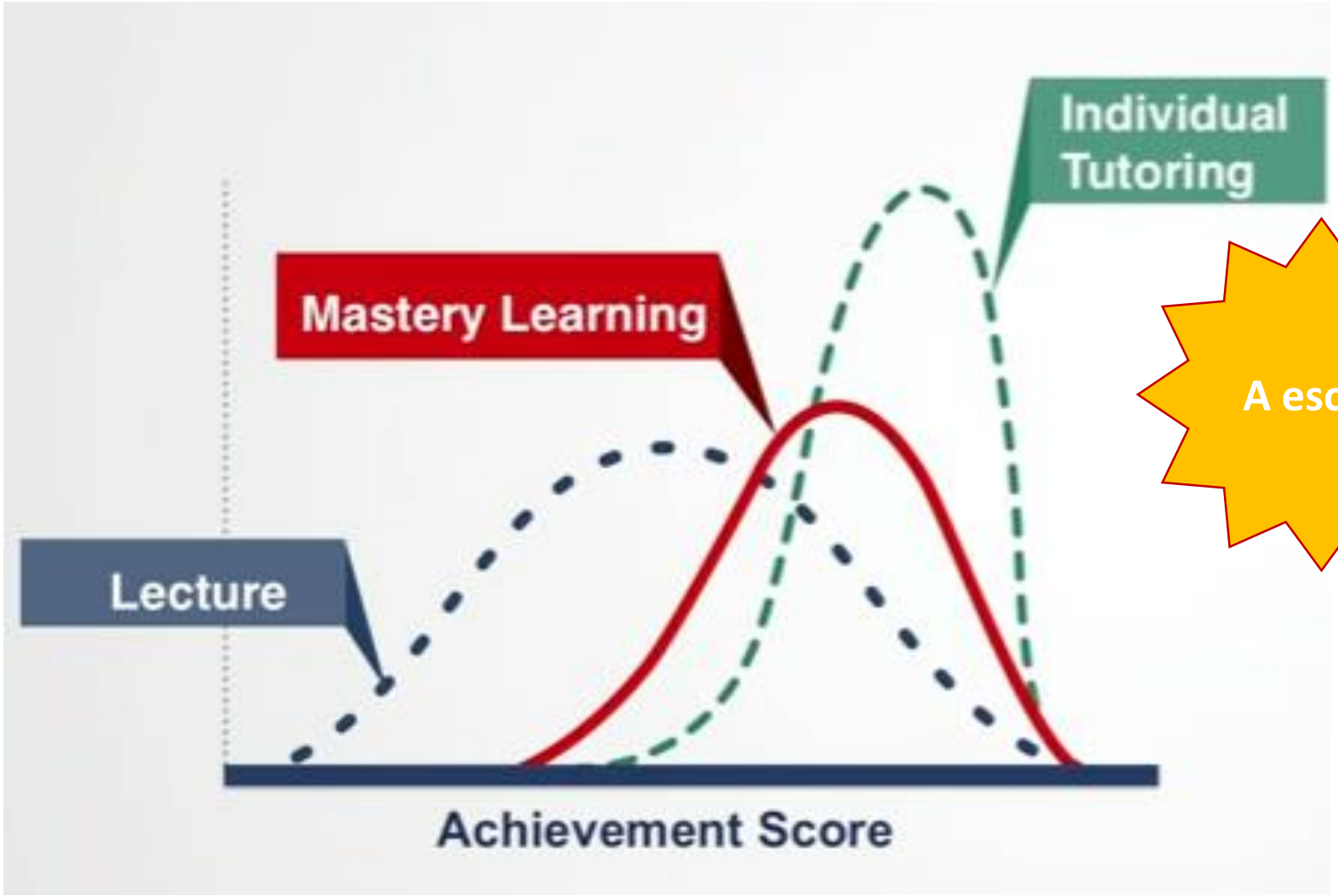
- Se puede recolectar cada:
  - click
  - Entrega
  - Post
  - Respuesta incorrecta
  - …

Para mejorar el proceso de aprendizaje

Posibles aplicaciones

- Análisis y visualización de datos
- Proveer retroalimentación a los profesores
- **Recomendaciones/personalización** para los estudiantes
- Modelado de estudiantes
- Detección de comportamientos no deseados
- Análisis de las redes de estudiantes
- Planificación y agenda

Personalización



1984

A escala

"The 2 Sigma Problem: The Search for Methods of Group Instruction as Effective as One-to-One Tutoring" B. Bloom, Educational Research, 1984

Los datos dan la información pero el profesor es quien tiene que tomar la decisión

?

El descubrimiento e interacción con el estudiante es automático dejando al profesor como diseñador

Aplicación en entornos educativos

- Predicción de abandonos
- Predicción de rendimiento

Motivación

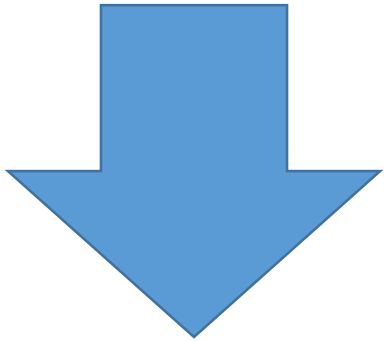
- ¿Por qué un estudiante tiene un rendimiento bajo?
- ¿Podemos predecir con antelación el bajo rendimiento o el abandono?
- ¿Podemos predecir el rendimiento para una tarea concreta?

Predicción de abandonos (Niemi, 2012)

- Estudio realizado con **60.000 estudiantes y 1.000 cursos online**.
- Los **datos** incluyen:
  - Medidas de aprendizaje
  - Satisfacción del estudiante
  - Rendimiento del profesor
  - Éxito de los graduados

- Usa regresión logística
- La causa principal del **abandono** es una **tendencia de rendimiento negativo**

David Niemi y Elena Gitin; "Using Big Data to Predict Student Dropouts: Technology Affordances for Research"



Predicción de rendimiento de un estudiante como medida para evitar su abandono

Esteban García-Cuesta – Departamento de Ciencias Y Tecnologías de la Información

Predicción de rendimiento de estudiantes (Cortez, 2008)

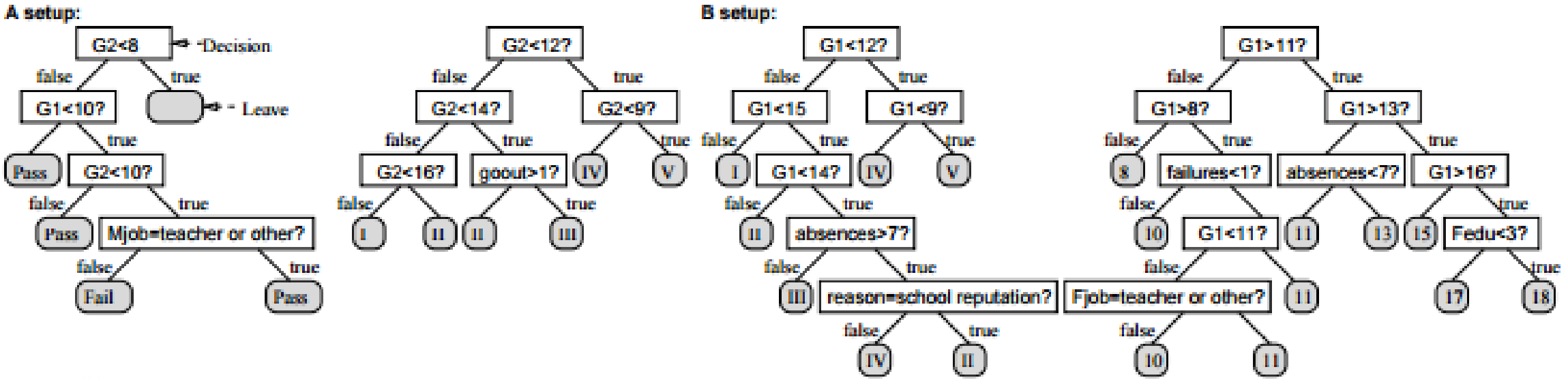
- Datos provenientes de escuelas de secundaria que incluyen
  - Variables **socio demográficas** (edad, género, tamaño de familia, etc.)
  - **Notas anteriores** (p.ej. Primer periodo, segundo periodo, final de años anteriores)
  - Asignaturas de matemáticas y portugués

- 33 Atributos y 649 instancias
- Clasificación

  - Binaria
  - [0-5]
  - Regresión

Predicción de rendimiento de estudiantes

Las notas anteriores son buenos predictores tanto para el éxito como para el fracaso

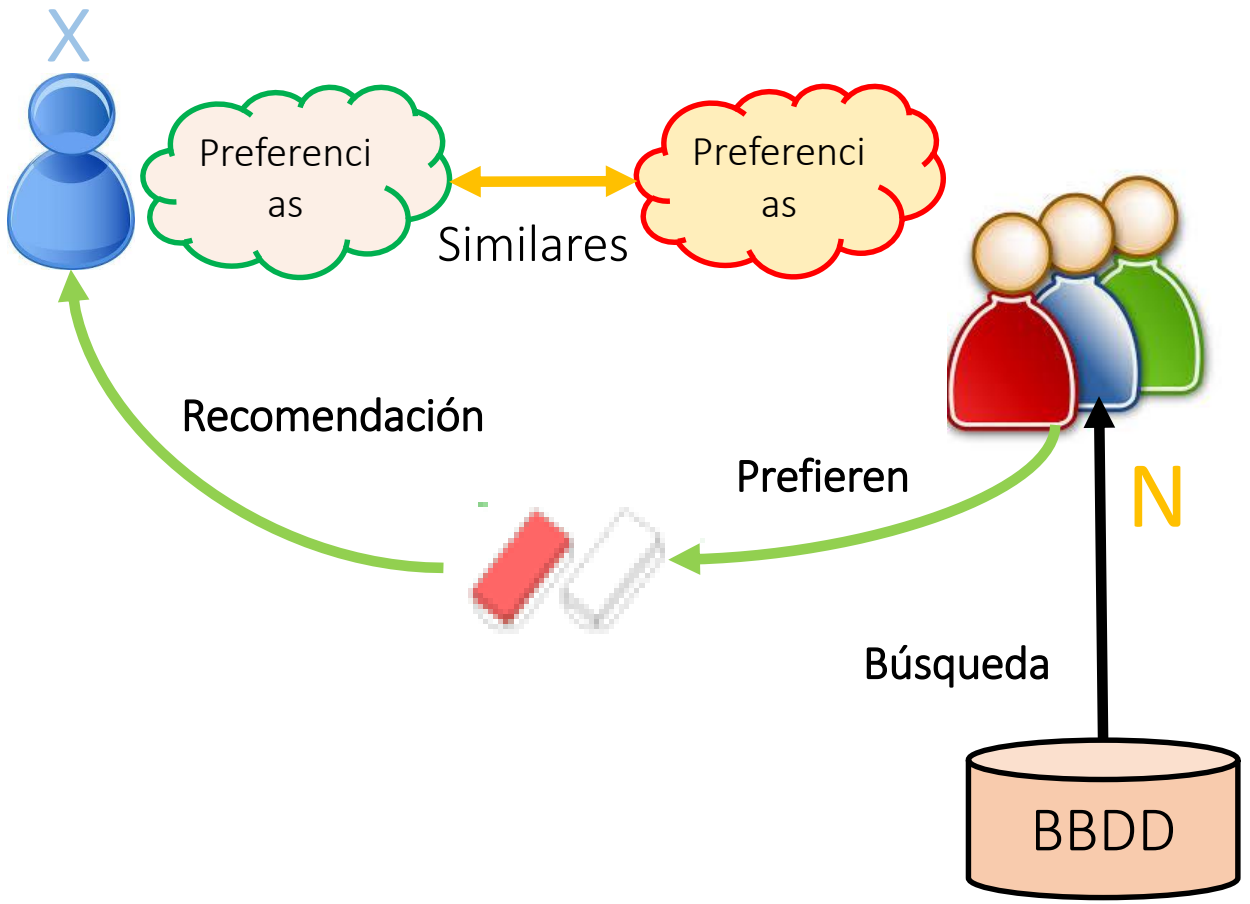


P. Cortez and A. Silva. Using Data Mining to Predict Secondary School Student Performance Proceedings of 5th FUture BUsiness TEChnology Conference (FUBUTEC 2008)

Filtrado colaborativo

Filtrado colaborativo

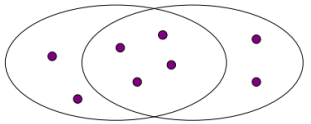
- Dado un usuario X
- Encontrar un subconjunto de usuarios N que son parecidos
- Estimar las valoraciones de X's en función de esos usuarios N



Filtrado colaborativo (encontrar usuarios similares)

- Sea r_x el vector de valoraciones del usuario x
- **El criterio de similitud de Jaccard mide la distancia entre perfiles como un conjuntos de valores**
  - **Problema :** Ignora el valor de la evaluación

$sim(D_1, D_2) = |C_1 \cap C_2| / |C_1 \cup C_2|$



- **El criterio de similitu de del coseno mide la distancia entre los vectores del perfil de usuario**

$$\text{sim}(\boldsymbol{x}, \boldsymbol{y}) = \cos(\boldsymbol{r_x}, \boldsymbol{r_y}) = \frac{r_x \cdot r_y}{||r_x|| \cdot ||r_y||}$$

  - **Problema:** trata los valores no definidos en el perfil como negativos
- **El criterio de coeficiente de correlación de Pearson**
  - S_{xy} = elementos que son valorados por los usuarios x e y

$$\boldsymbol{sim}(x, y) = \frac{\sum_{s \in S_{xy}} (r_{xs} - \overline{r_x})(r_{ys} - \overline{r_y})}{\sqrt{\sum_{s \in S_{xy}} (r_{xs} - \overline{r_x})^2} \sqrt{\sum_{s \in S_{xy}} (r_{ys} - \overline{r_y})^2}}$$

Filtrado colaborativo (Planteamiento)

$$r_{ix} = \frac{\sum_{j \in N(i;x)} s_{ij} \cdot r_{jx}}{\sum s_{ij}}$$

tareas

estudiantes	1	2	3	4	5	6	7	8	9	10	11	12	sim(1,m)
1	1		3		2.6	5			5		4		1.00
2			5	4			4			2	1	3	-0.18
3	2	4		1	2		3		4	3	5		**0.41**
4		2	4		5			4			2		-0.10
5			4	3	4	2					2	5	-0.31
6	1		3		3			2			4		**0.59**

J. Leskovec, A. Rajaraman, J. Ullman: Mining of Massive Datasets, http://www.mmds.org

La recomendación se hace ponderando ambos elementos $r_{1.5}$ = (0.41*2 + 0.59*3) / (0.41+0.59) = 2.6

Ventajas e inconvenientes del filtrado colaborativo

- **+ Funciona con cualquier tipo de elemento**
  - No es necesario seleccionar características
- **- Sufre el problema de cold-start (arranque en frio)**
  - Necesita un conjunto suficientemente grande de usuarios con valoraciones para poder realizar el “matching” entre usuarios
- **- Sufre de dispersidad**
  - La matriz de usuarios/valoraciones es dispersa
  - Dificultad para encontrar usuarios que hayan valorado los mismos elementos
- **- Primera recomendación**
  - No se pueden recomendar elementos que no hayan sido previamente recomendados.

- Sesgo en la recomendación

- Tendencia a recomendar elementos populares

Utilización de métodos híbridos

- **Implementar diferentes recomendadores y combinarlos**
  - P.e. realizar una combinación lineal

- **Combinar métodos basados en contenidos con el filtrado colaborativo**
  - Realización de perfiles de elementos para resolver el problema de un elemento nuevo
  - Utilización de información demográfica para resolver el problema de un usuario nuevo

KDD Cup 2010 Educational Data Mining Challenge

- ¿Cómo aprende un estudiante?
- ¿Qué significa que dos problemas sean parecidos?
- ¿Es posible extraer predecir el rendimiento y conocimiento de un estudiante sin intervención humana?

UNIVERSIDAD COMPLUTENSE MADRID

KDD Cup 2010 Educational Data Mining Challenge

- Data sets:
  - Contienen información de logs de interacción del estudiante con el curso
  - P.ej. Cuantas veces ha visto el problema, número de intentos fallidos, número de veces que el estudiante ha solicitado información relacionada, etc.
  - Algebra 2008-2009 y Bridge to Algebra 2008-2009

- El objetivo es predecir si el usuario resolverá la tarea o el problema en el primer intento

Recomendación aplicada a la predicción de rendimiento de un estudiante

El mapeo entre la aproximación presentada y este problema es claro

Estudiante → **usuario**

Éxito al primer intento → **valoración**

[0 -1] **[0 -5]**

Recomendación aplicada a la predicción de rendimiento de un estudiante

Los contenidos están organizados en:

- **PH Jerarquía del problema**
- **PN Nombre del problema**
- **SN Nombre del paso**
- **PV Vista del problema**

El mapeo de las tareas se hace a diferentes niveles:

- **A nivel individual (PH+PN+SN+PV)**
- **Componente de conocimiento (KC)**

Recomendación aplicada a la predicción de rendimiento de un estudiante

Table 3: Root mean squared error (RMSE) for different methods using different sets of attributes as items

Technique	Item	Algebra	Bridge	Average
Global Average	-	0.34316	0.33199	0.33757
User Average	-	0.33892	0.32843	0.33367
Logistic Regression	A	0.32226	0.30456	0.31341
Logistic Regression	B	0.32444	0.30589	0.31517
Logistic Regression	A + B	0.32354	0.30498	0.31426
Logistic Regression	A + B + C	0.31988	0.30583	0.31286
Matrix Factorization	PN	0.33752	0.31515	0.32633
Matrix Factorization	PG	0.34316	0.33199	0.33757
User-Item Collaborative Filtering	PH, PN, SN, PV	0.32240	0.29817	0.31029
Matrix Factorization + Global Average	PH, PN, SN, PV	0.31817	0.29825	0.30821
Matrix Factorization + User Average	PH, PN, SN, PV	0.31812	0.29865	0.30837
Matrix Factorization + User-Item Collaborative Filtering	PH, PN, SN, PV	0.31787	0.29804	0.30796
Matrix Factorization + User-Item Collaborative Filtering	KC-rules	*0.30228*	0.29804	**0.30016**

The best result is bold faced. A: (Student-Average, Step-Average); B: (Student-PV-Average, Step-PV-Average); C: (PG-Average, PN-Average, Student-PG-PV-Average)

Nguyen Thai-Nghe, et. Al; “Recommender System for Predicting Student Performance”

Reflexiones finales

Analítica del aprendizaje ⟺ Minería y automatización del aprendizaje

Tendencia a la automatización y globalización de la educación

Detección automática de fallos de comprensión comunes

Comportamientos de estudiantes homogéneos y creación de grupos

Aprendizaje automático de mejores rutas de aprendizaje

} **APPS**

Reflexiones finales

*"**Fomentar el rol** del profesor como **mentor/guía** integrando sus actividades cotidianas con las analíticas del aprendizaje y métodos de predicción para aplicar un método activo que mejore la **atención**, el **interés** y finalmente el **aprendizaje** del estudiante."*

GRACiAS

Esteban García-Cuesta
Email: esteban.garcia@universidadeuropea.es
LinkedIn: https://es.linkedin.com/in/egarciacuesta
Twitter: egarciacuesta

Bibliografía

- UCI Student Performance Data Set https://archive.ics.uci.edu/ml/datasets/Student+Performance
- P. Cortez and A. Silva. Using Data Mining to Predict Secondary School Student Performance. In A. Brito and J. Teixeira Eds., Proceedings of 5th FUture BUsiness TEChnology Conference (FUBUTEC 2008) pp. 5-12, Porto, Portugal, April, 2008, EUROSIS, ISBN 978-9077381-39-7.
- Daphne Koller; Ted Talk "What we're learning from online education" https://www.ted.com/talks/daphne_koller_what_we_re_learning_from_online_education#t-289381 (último acceso 13/11/2016)
- Paulo Cortez; UCI Machine Learning Repository "Student Performance Dataset" https://archive.ics.uci.edu/ml/datasets/Student+Performance
- KDD Cup 2010 Educational Data Mining Challenge https://pslcdatashop.web.cmu.edu/KDDCup/ (último acceso 13/11/2016)
- Nguyen Thai-Nghe, et. Al; "Recommender System for Predicting Student Performance" 1st Workshop on Recommender Systems for Technology Enhanced LEarning, Procedia Computer Science 1(2010) 2811-2819
- David Niemi y Elena Gitin; "USING BIG DATA TO PREDICT STUDENT DROPOUTS: TECHNOLOGY AFFORDANCES FOR RESEARCH" IADIS International Conference on Cognition and Exploratory Learning in Digital Age.
- García-Cuesta E.; Iglesias J.A.;"User Modeling: Through Statistical Analsiysis and Subspace Learning" Expert Systems with application 39(5) 5243-525, April 2012. https://www.researchgate.net/profile/Jose_Iglesias2/publication/220217117_User_modeling_Through_statistical_analysis_and_subspace_learning/links/0a85e533d5f8e1d606000000.pdf?origin=publication_detail