Grado en Ingeniería Informática

Búsqueda y minería de información

Teoría: Pablo Castells

Prácticas: Javier Sanz-Cruzado, Pablo Castells

Escuela Politécnica Superior

Universidad Autónoma de Madrid

Curso 2019 – 2020

Presentación

- Introducción al curso
- Temario
- Bibliografía
- Actividades, evaluación, puesta en marcha

¿De qué trata la asignatura?

Acceso a & análisis de **información no estructurada**

Problemas que no se pueden resolver directamente con una base de datos

Papel central del **factor humano** en la definición del problema

Definición de criterios, subjetividad, lenguaje, comportamiento, etc.

Orientado a **gran escala**, entornos abiertos

Técnicas específicas para conseguir escalabilidad

- **Motores de búsqueda** (recuperación de información, IR)
- **Sistemas de recomendación** (RS)
- **Análisis de redes sociales** (SNA)

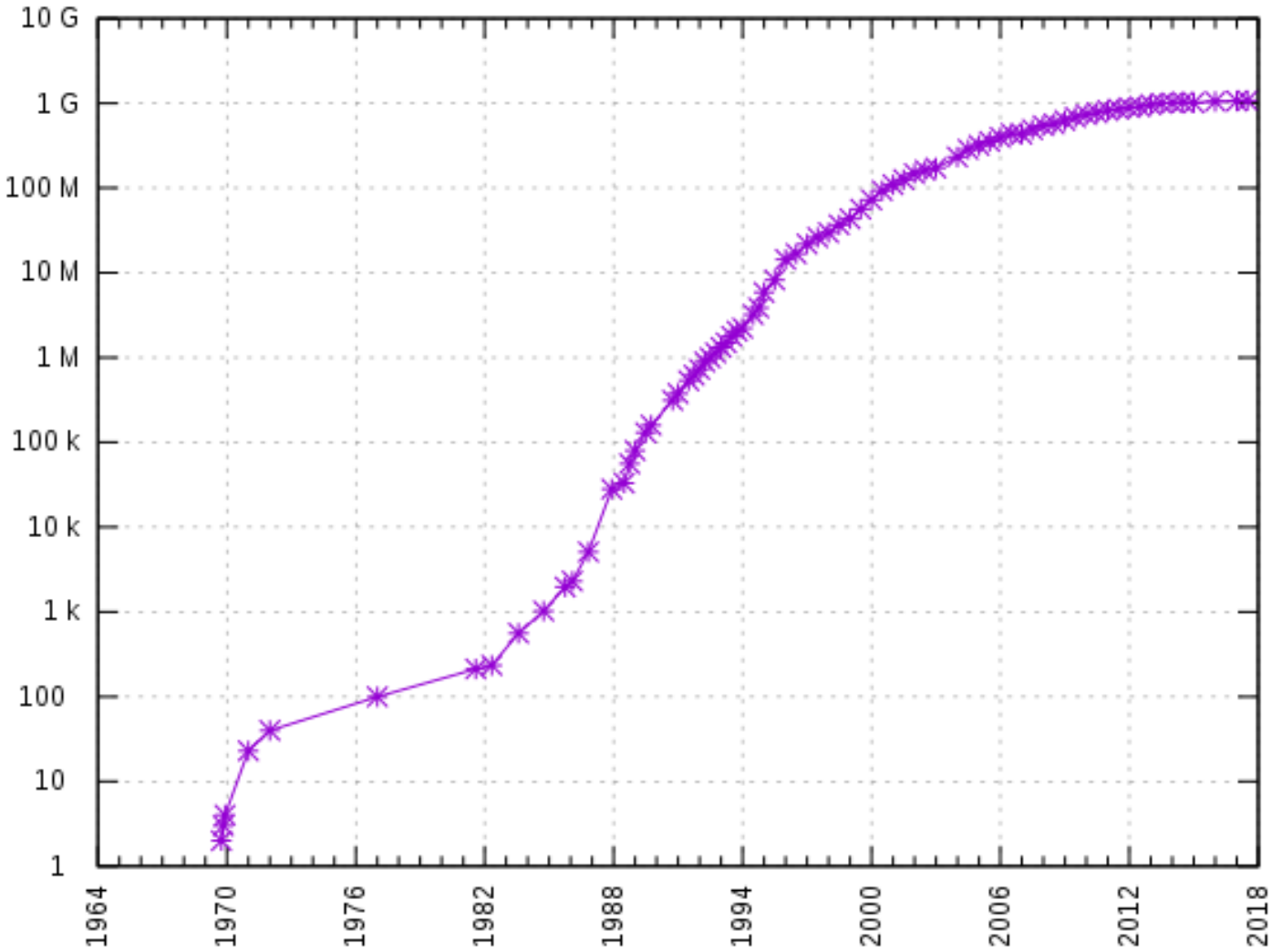
IR	RS	SNA

Otras áreas más allá del temario: análisis de opinión, detección de entidades, NLP, búsqueda & análisis MM…

Escalas de información que manejamos hoy día

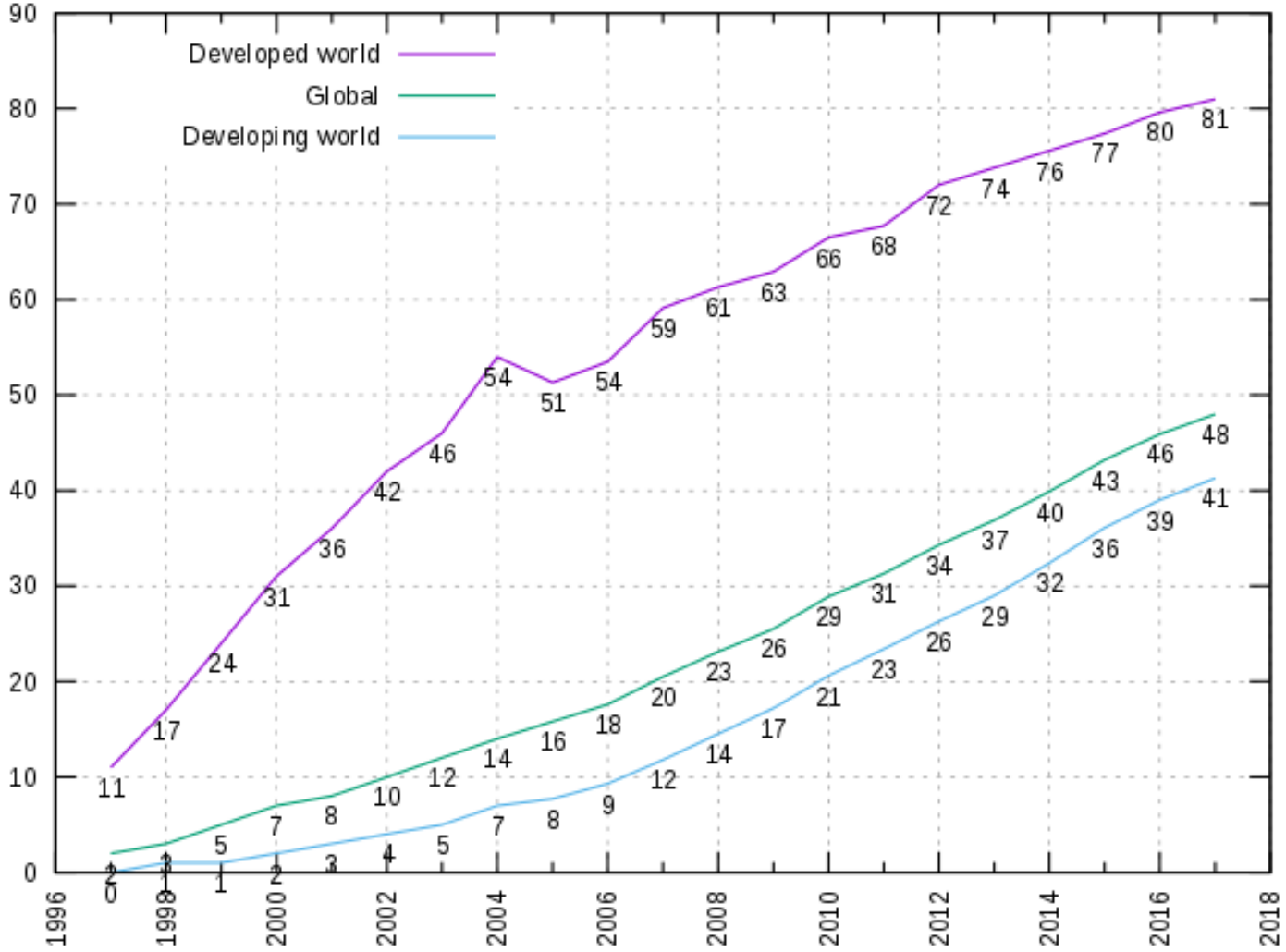
- Información accesible en la Web
  - ~1 billón de dominios, cientos de billones de URLs, miles de PB de información
  - 70% contenido creado por usuarios finales
- Más de la mitad de la población mundial utiliza Internet
  - Más del 97% de la comunicación mundial es vía Internet
  - 10-15% del comercio mundial es electrónico
  - 50.000+ búsquedas por segundo en Google
  - > 2.000M usuarios (~la mitad de usuarios de Internet) en Facebook
- ~80% de la "información mundial" es no estructurada
  - Texto, a/v, objetos
  - Datos de interacción de usuarios con objetos y usuarios
  - Estructuras (redes) e interacción sociales
- Tendencia: las cifras son sólamente ilustrativas porque seguimos en crecimientos exponenciales

Nº de dominios Internet

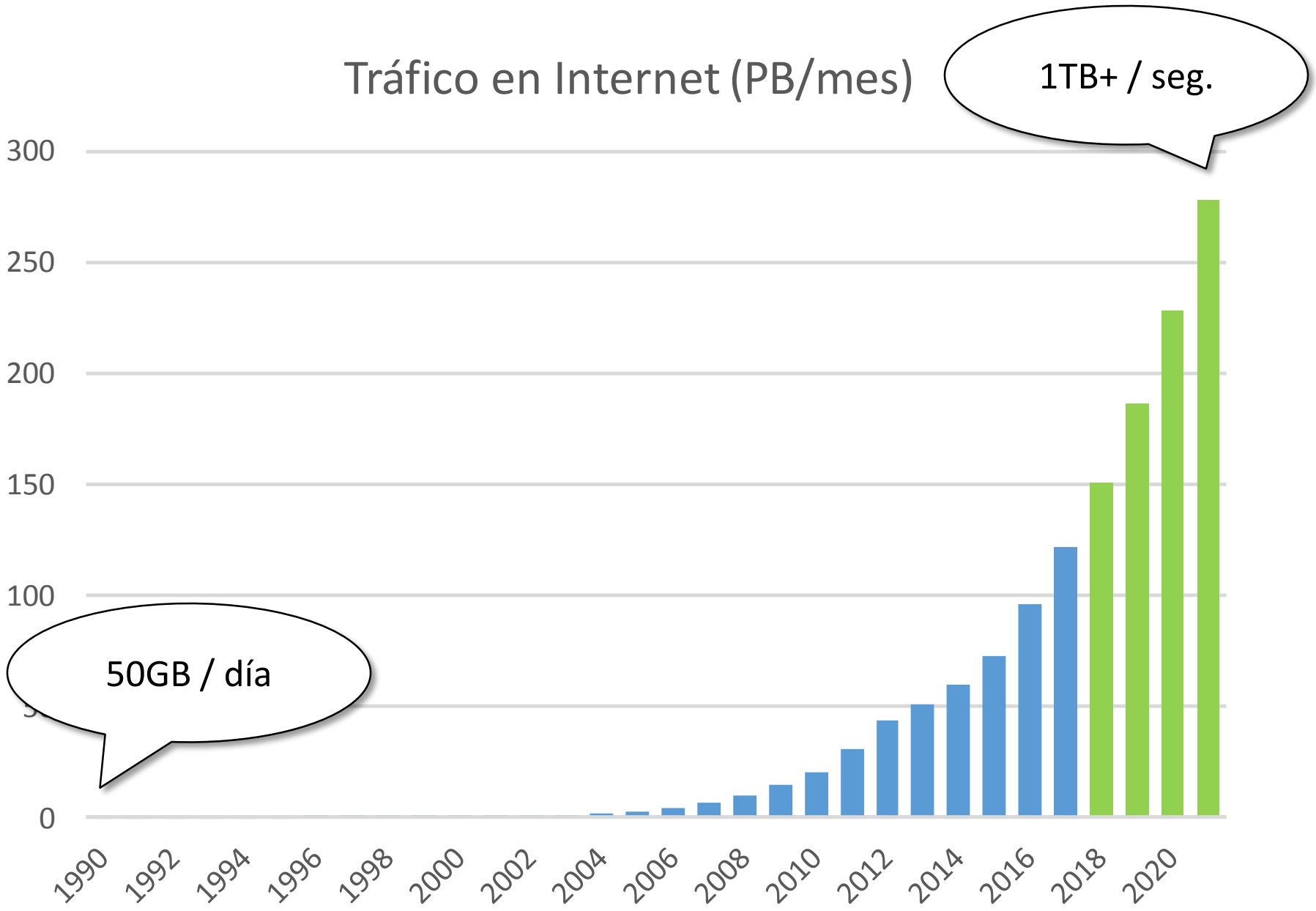


Internet Domain Survey. Internet Systems Computing, July 2017
http://ftp.isc.org/www/survey/reports/current

% usuarios de Internet en la población

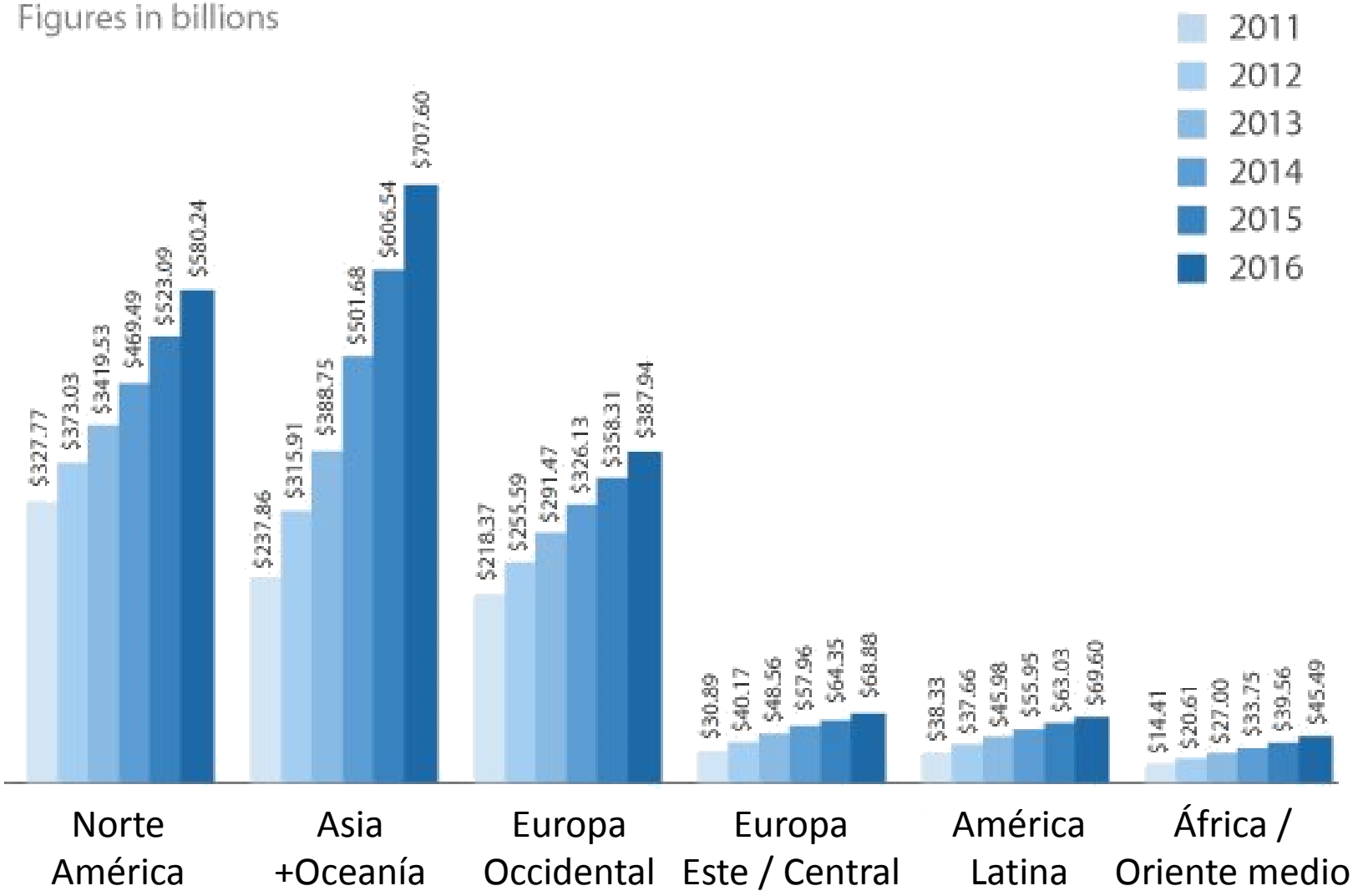


https://en.wikipedia.org/wiki/Global_Internet_usage



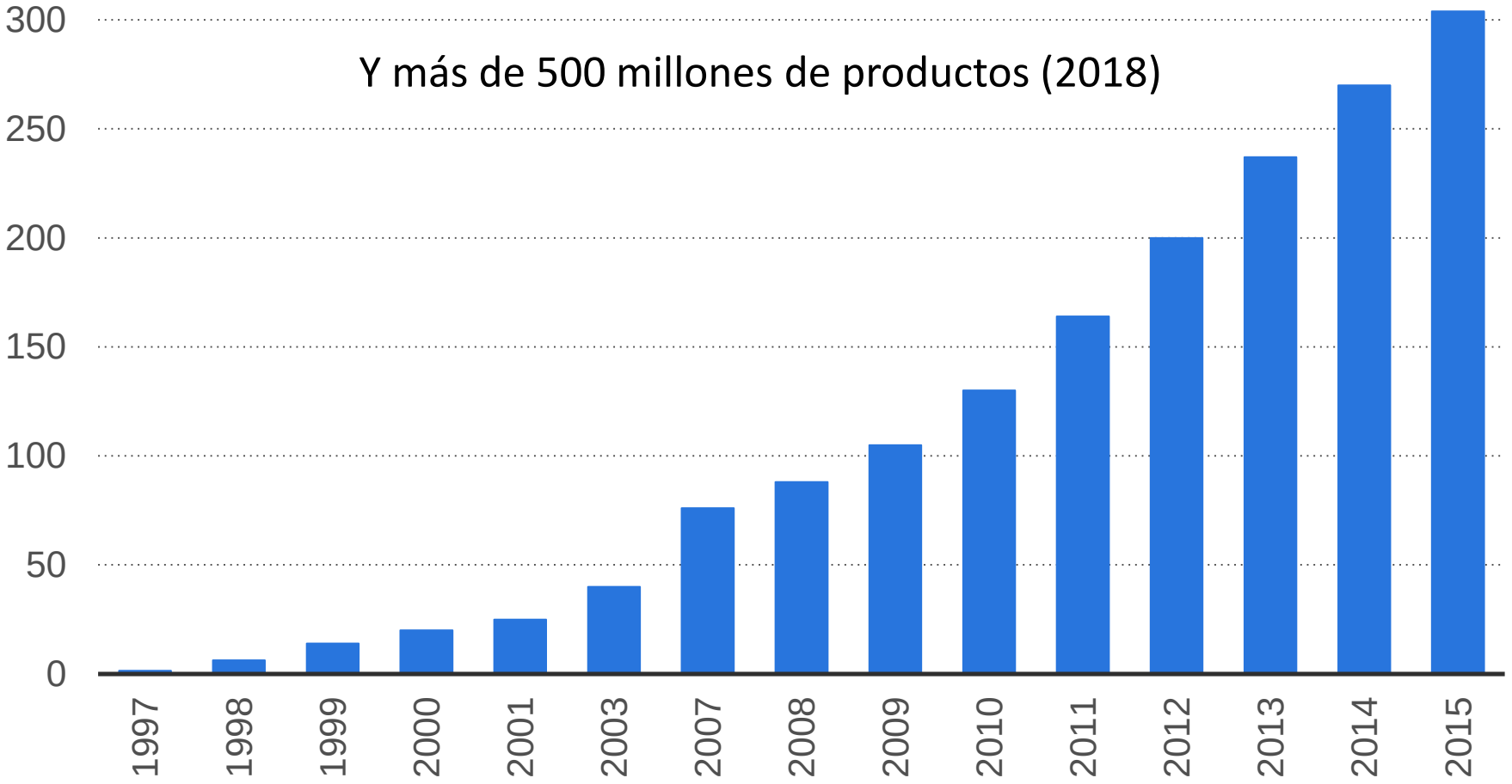
The Zettabyte Era: Trends and Analysis. Cisco Systems, June 2017

B2C Ecommerce sales growth, worldwide



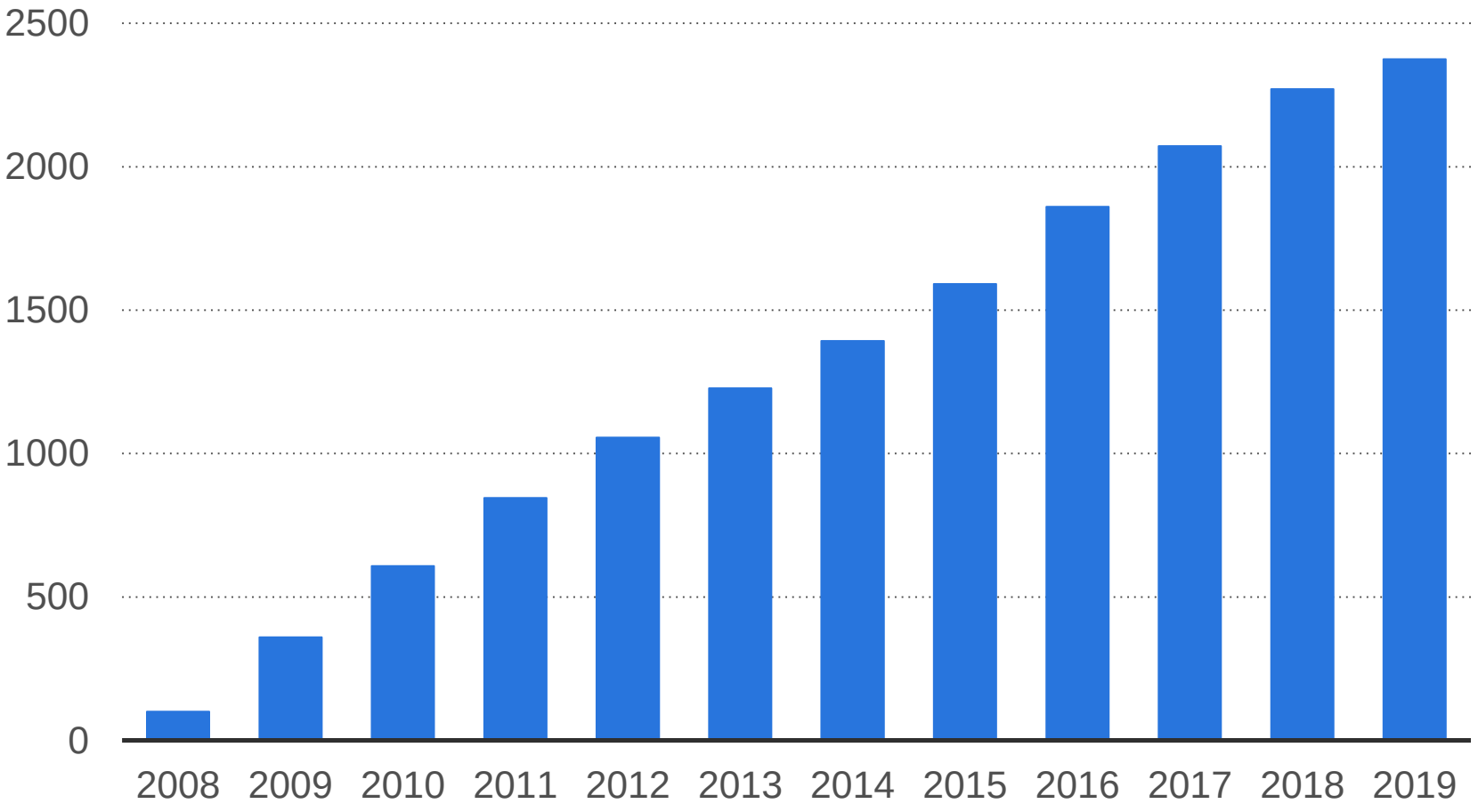
http://anewdomain.net/2017-internet-statistics-the-state-of-the-internet-web-growth

Amazon: nr. cuentas activas anuales (millones)



https://www.statista.com/statistics/237810/number-of-active-amazon-customer-accounts-worldwide

Facebook: nr. usuarios (millones)

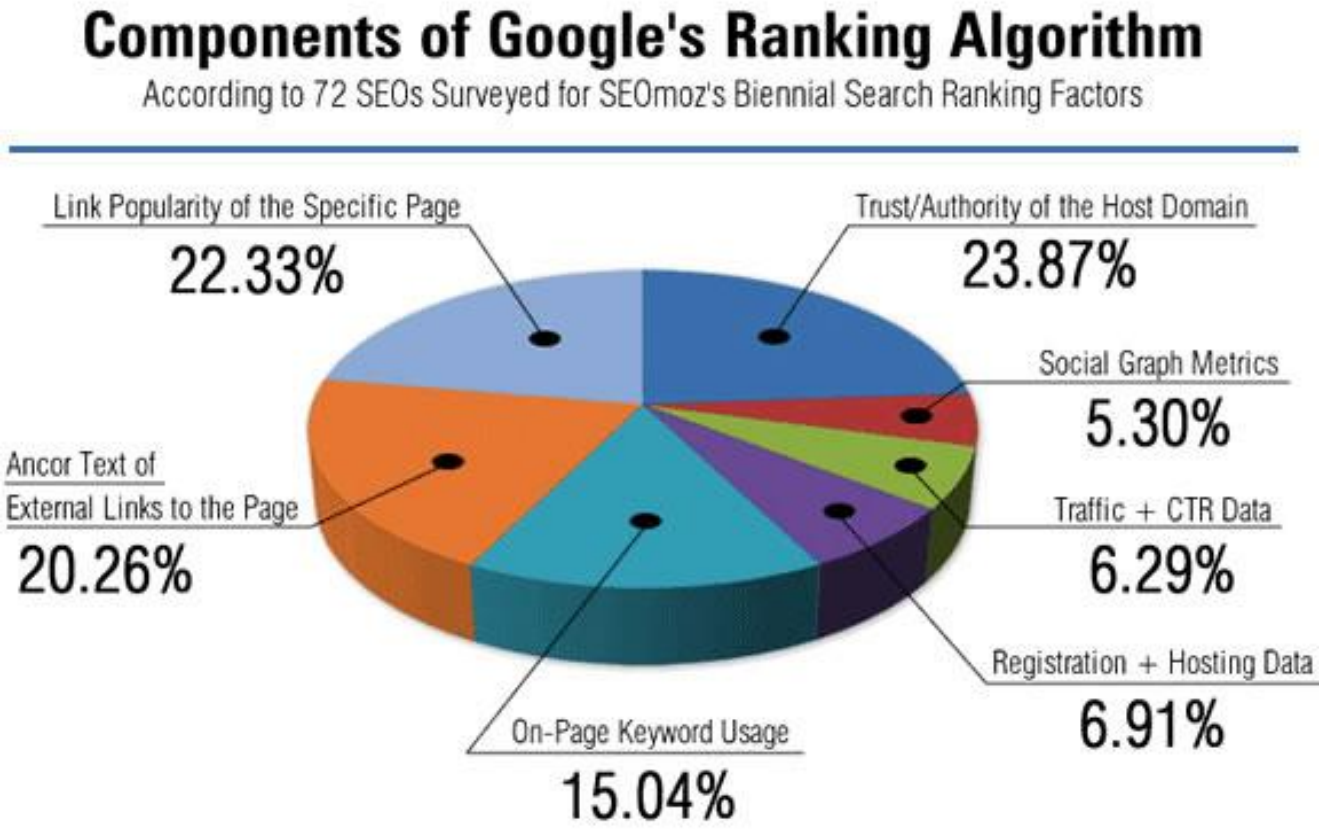


- Conocimiento
- Noticias
- Producción audiovisual
- Sensores
- Actividad de la población
- ...

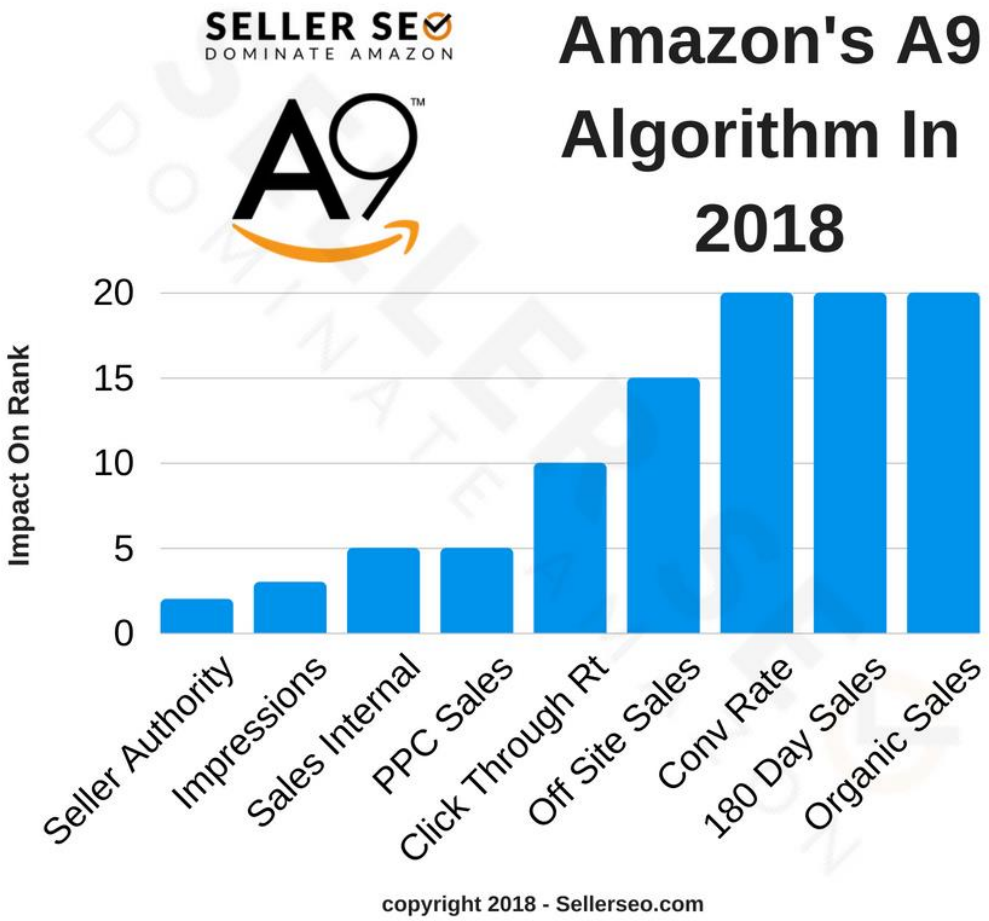
- Todo empezó en los 90…
- Disponemos de una réplica digital del “mundo”
- No se puede estructurar completamente en BDs
  - No da tiempo
  - …y no se puede
  - ⇒ La información se genera y se amontona sin estructurar
- Accesible y analizable de forma automatizada!

Escala de información


LET'S SOLVE THIS PROBLEM BY USING THE BIG DATA NONE OF US HAVE THE SLIGHTEST IDEA WHAT TO DO WITH



https://www.lyfemarketing.com/blog/google-ranking-factors

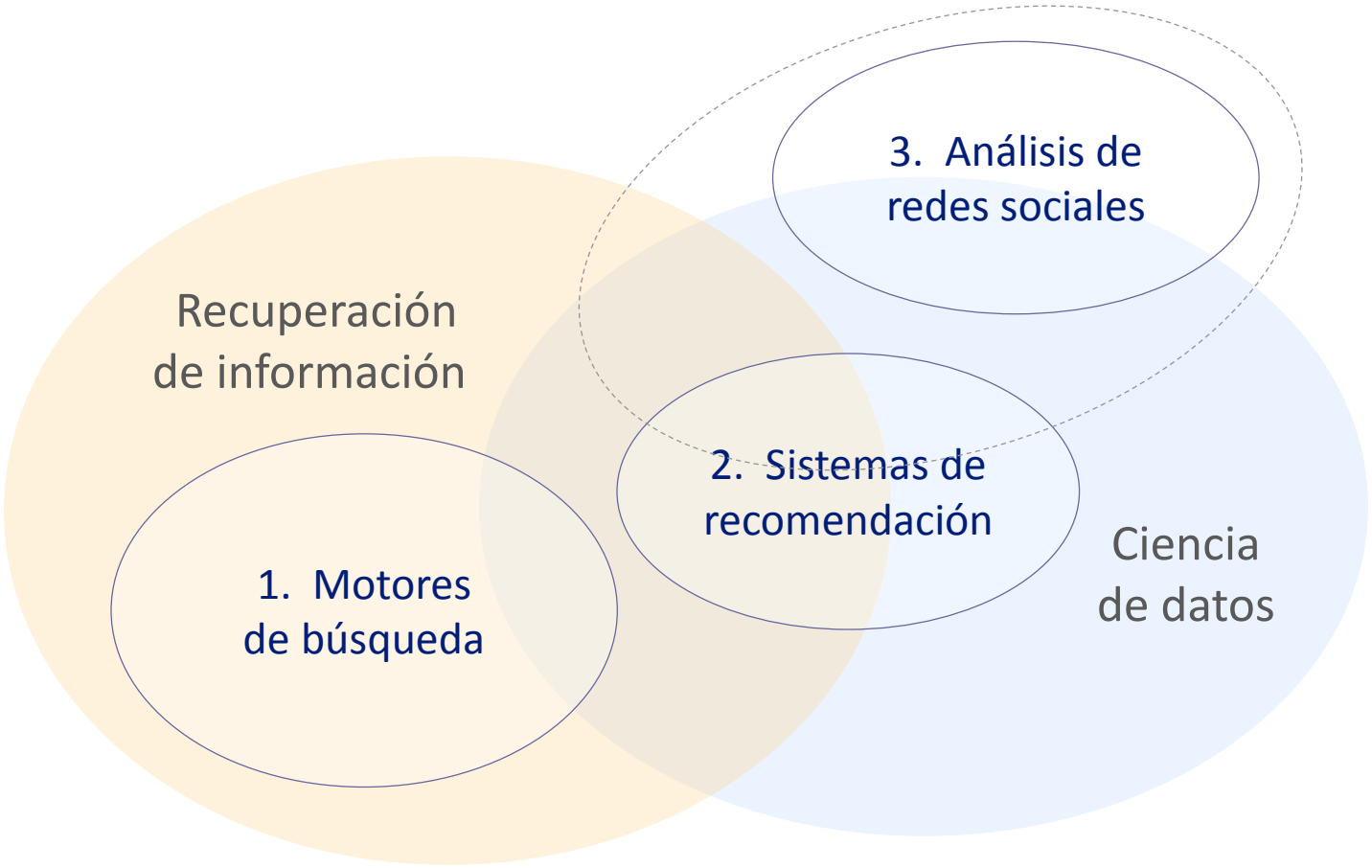


https://sellerseo.com/amazon-a10-algorithm-2018

Explosión de información

- **Acceso** a la información
  - Ayudar a personas a encontrar información de interés
- **Análisis** de actividad **social**: describir, explicar, predecir
  - Fenómenos globales en la relación entre personas

Relación de áreas cubiertas



Oportunidad de la materia

- Toca áreas de amplio impacto en tecnología básica de uso diario
  - Mercados tecnológicos entre los más amplios y activos
- Desarrollo y madurez de las tecnologías
  - Un corpus de conocimiento propicio para el estudio sistemático (propio de grado)
- Interés intrínseco de la materia
  - Desde un punto de vista teórico y científico: formulación, algoritmia, etc.
  - Desde un punto de vista de ingeniería: solución de problemas técnicos
- Continua innovación a corto, medio y previsible largo plazo

Necesidades / oportunidades

- Ayudar al usuario a **encontrar** lo que necesita
  - Buscar
  - Recomendar
- **Detectar** elementos, estructuras y situaciones de interés sin búsqueda directa
  - Encontrar significados, relaciones significativas, tendencias…
  - **Describir, explicar y predecir** fenómenos
- En esta asignatura nos centramos en:
  - Motores de búsqueda (texto) } Recuperación de información
  - Sistemas de recomendación } Minería social
  - Redes sociales } Minería social

Aplicaciones

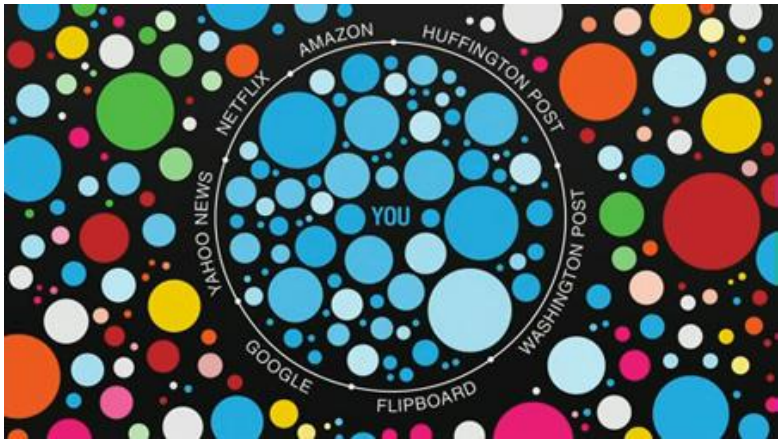
- Recuperación de información
  - Tecnología de búsqueda Web
  - Búsqueda corporativa
  - Búsqueda local en sitios Web
  - Clasificación de documentos, filtrado de spam
  - …

- Sistemas de recomendación
  - Comercio electrónico
  - Música
  - Apps
  - Cine, ocio
  - Noticias
  - Publicidad
  - Dating
  - …

- Redes sociales
  - Gestión y análisis
  - Recomendación de contactos
  - Recomendación de grupos
  - Difusión de información
  - Marketing
  - Análisis de opinión
  - …?


Riesgos éticos

- Ayuda no deseada
- Invasión de la privacidad
- Burbujas de opinión, empobrecimiento del pensamiento
- Malinterpretación del usuario
- Exprimir al usuario
  - Como fuente de información
  - Como agente monetizable
- Sesgos, discriminación
- Manipulación social



Temario

1. Motores y modelos de búsqueda
   - Principios fundamentales
   - Componentes de un motor de búsqueda
   - Modelos: booleano, vectorial, proximal
   - Indexación
   - Evaluación
2. Búsqueda en la Web
   - Aspectos específicos
   - Crawling e indexación
   - Métodos basados en enlaces: PageRank
3. Sistemas de recomendación
   - Conceptos fundamentales
   - Métodos basados en contenido
   - Filtrado colaborativo
   - Evaluación
4. Análisis de redes sociales
   - Análisis y métricas
   - Detección de comunidades
   - Modelos de red social
   - Redes de mundo pequeño
   - Procesos de difusión

Enfoque

Práctico y aplicado

- Algoritmos
- Detalles de implementación
- Tratamiento de problemas prácticos
- Soluciones viables

Teórico

- Conceptos y “aspectos científicos”
- Modelos, ecuaciones, fórmulas
- Soluciones abstractas

Relación con otras asignaturas

- Estructuras de datos
  - Índices
- Inteligencia artificial & Fund. de aprendizaje automático
  - Clasificación
- Análisis de algoritmos
  - Algoritmos de grafos

Máster I2ICSI

- Recuperación de información
  - Ampliación de métodos de búsqueda, recomendación, evaluación
- Minería Web
  - Ampliación de crawling, clasificación, redes sociales

Bibliografía

- Information Retrieval: Implementing and Evaluating Search Engines

  S. Büttcher, C. L. A. Clarke, G. V. Cormack, 2010

- Introduction to Information Retrieval

  C. D. Manning, P. Raghavan, H. Schütze, 2008

- Modern Information Retrieval, 2nd ed

  R. Baeza-Yates, B. Ribiero-Neto, 2011

- Recommender Systems Handbook

  F. Ricci, L. Rokach, B. Shapira, P. B. Kantor (eds.), 2011

- Networks: An Introduction

  M. Newman, 2010

Bibliografía auxiliar

- Search Engines: Information Retrieval in Practice

  W. B. Croft, D. Meltzer, T. Strohman, 2010

- Google's PageRank and Beyond: The Science of Search Engine Rankings

  Amy N. Langville and Carl D. Meyer, 2006

- Recuperación de Información: un enfoque práctico y multidisciplinar

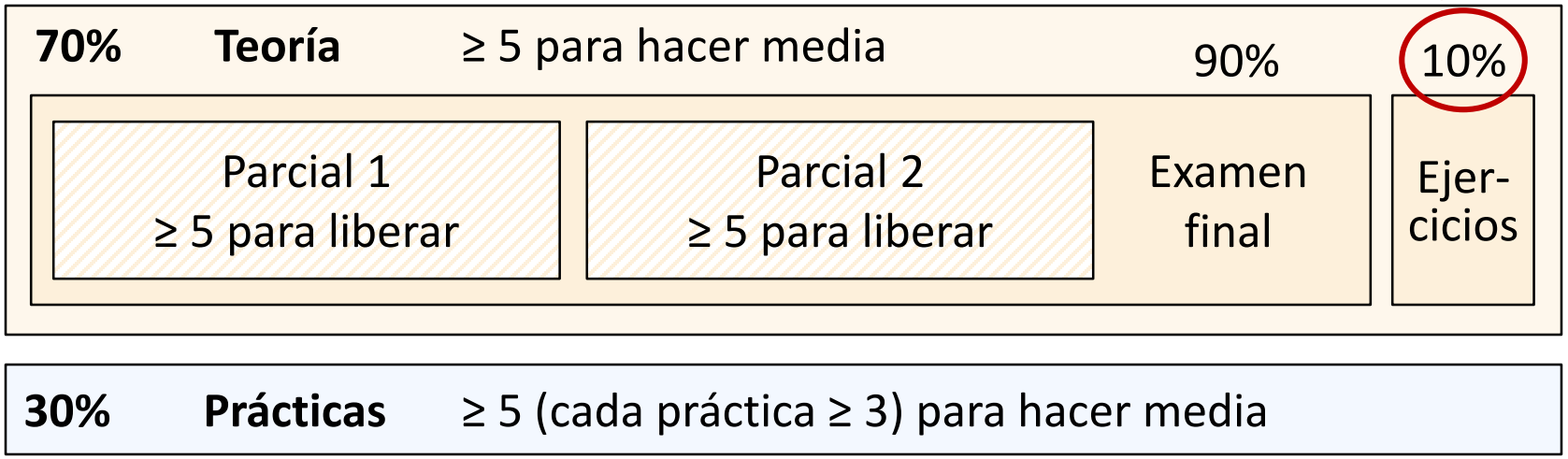
  F. Cacheda, J. M. Fernández Luna, J. Huete (eds.), 2011

- Networks, Crowds, and Markets

  D. Easley, J. Kleinberg, 2010

Evaluación

Sólo si sube la nota



- Dos pruebas intermedias liberatorias
- La nota del parcial liberado se traslada a la nota del examen final, escalada a la puntuación de la parte correspondiente
- Previsiblemente, los parciales cubrirán entorno a un 70% de la materia
- Si se repite la parte de un parcial en el examen final, y la nueva puntuación fuese inferior a la del parcial, se aplicará la media de ambas

Ejercicios

- Se publicarán a lo largo del curso (~70 aprox)
- Entrega del 60% para optar a la máxima puntuación
- La realización de ejercicios marcados con * se podrá valorar como un plus en la nota, a criterio del profesor
- Entrega vía Moodle, el último día de clase

Prácticas

- Dos temas principales

  a) Motor de búsqueda (dividido en 3 entregas previstas)

  b) Sistemas de recomendación + redes sociales
  (1-2 entregas previstas)

- Se utilizará un repositorio GitLab
  - Commits al final de cada clase de prácticas
  - Seguimiento del progreso

- Formación de grupos

Calendario previsto

- Prácticas
  - P1: 5/7 febrero → entrega 18/20 febrero
  - P2: 19/21 febrero → entrega 9/13 marzo
  - P3: 11/13 marzo → entrega 31 marzo / 2 abril
  - P4: 1/3 abril → entrega martes 12 mayo
- Parciales
  - Jueves 5 de marzo de 15 a 17h
  - Jueves 16 de abril de 15 a 17h
- Miércoles 15 de abril
  - Clase de teoría → se recupera el miércoles 13 de mayo 18-19h
  - Clase de prácticas → no la recuperamos (2 viernes festivos)
- Examen final: lunes 18 de mayo 15h