What's Big, Doc?

(con perdón de Bugs Bunny)

Fernando Das Neves

Gerente de I+D, Snoop Consulting

Universidad Austral

fernando.das.neves@snoopconsulting.com

Nuestra experiencia

- Agregación y procesamiento masivo de logs distribuidos (1000 eventos/seg por servidor)
- Extracción de información a partir de procesamiento del texto de historias clínicas.

¿Qué es big?

- Lo que sea big en este momento (volumen, variedad, velocidad de cambio)
  - ➔ probabilistic counting fue aplicado por primera vez en 1981 en System R.
- En la práctica, muchísimos de los datasets "big" reales se pueden procesar en 1 máquina grande: *.
  - ➔ *"The majority of real world analytic jobs processes take less than 100 GB of input"*
  - ➔ *"internally [...] found median job sizes to be less than 14 GB"*

* *"Nobody ever got fired for buying a cluster"*, MS Research, Enero 2013.

- Migración de "batch, todo puede fallar, vuelvo en un rato" (Hadoop, Pig) a procesamiento en memoria (Spark, Impala, Dremel).

Oportunidades de impacto general

- Sistemas (p.ej. en la dirección de AmpLab de UC Berkeley: BlinkDB, Spark) apoyados en casos reales.
- Simplificar/automatizar el ciclo de vida de aplicaciones de machine learning.
  - *"Scaling Big Data Mining Infrastructure: The Twitter Experience"*, SIGKDD Explorations, Diciembre 2012 .
- Formación de profesionales con experiencia **práctica**-teorica (cruzar el contenido de "Mining of Massive Datasets" con experiencia práctica con Cassandra, Spark).
- Entrenamiento semi-supervisado sólo se pone peor en big data: siempre vamos a tener 10, 100, 1000 veces más datos sin etiquetar.
- **Implicaciones éticas no son solo "un problema"**: ¿dónde está el límite entre "se puede" y "se debe"? ¿Poner sensores de MACs en los cestos de basura del subte?
  - *"Cyber spies in London recycle bins told to move on"* http://www.theguardian.com/media-network/partner-zone-infosecurity/cyber-spies-london-recycle-bins

Colaboración entre Universidad e Industria.

- ¿Seguir el modelo de Research and Technology Organizations (Fraunhofer, GTS System, etc)? *
  - ➔ Investigación más orientada a objetivos que las universidades
  - ➔ Explícitamente prohibido competir con empresas privadas en las partes de su negocio que reciben financiamiento básico.
  - ➔ Parte integral de la innovación abierta de empresas.
- De Fraunhofer: “Living Lab” con software de big data, para que empresas prueben y ganen familiaridad con nuevas formas de recolección y análisis de datos, primero sobre datos de prueba, y luego sobre los suyos.

* *“International Comparison of Five Institute Systems”*
http://www.teknologiportalen.dk/NR/rdonlyres/5A0D95D9-FD3B-4A44-B7C2-AB619D20C98A/3664/IntComparison2.pdf