Aprendizaje sobre Grandes Volúmenes de Datos y el Sistema Watson Jeopardy

Facultad de Ingeniería Eléctrica, Universidad Tecnológica de Panamá

Pablo Ariel Duboue

Erudite Science, Inc.
1176 Rue Bishop,
Montréal, QC H3G 2E3, Quebec
Canada

14 de septiembre, 2015

Outline

Outline

¿Qué es Bigdata?

- Es un término comercial
  - Sirve para describir productos y servicios relacionados con el manejo de datos
  - Según el interés de la persona en vender productos y servicios, son los límites de lo que es bigdata
- Es la progresión natural en manejo de datos
  - Base de datos
  - Datawarehouse
  - Soluciones de bigdata
- En el caso del aprendizaje automático, soluciones para grandes volúmenes de datos se utilizan cuando los datos no pueden entrar en la memoria y disco de una sola máquina

El valor está en los datos

- Actualmente más y más empresas y particulares se dan cuenta del valor de los datos
- El acopio de datos comienza muy antes de la búsqueda de valor en esos datos
- Las soluciones de bigdata permiten extraer valor de dichos datos

Las computadoras como humanizadoras

- Nací en mediados de los ’70
- La mitad de todos los humanos que han existido están vivos en este momento
- Ya no es posible el tipo de personalización que es natural para los humanos
- El análisis de grandes volúmenes de datos permite el tipo de personalización que nos hace falta

La democratización del cómputo

- Algunas ideas inspiradas en la presentación de Alistair Croll durante la semana de Bigdata en Montreal
  - http://www.slideshare.net/Tiltmill/cycle-time-trumps-scale-big-data-as-the-organizational-nervous-system-montreal-big-data-week-2014
- Computo, lleva a automatizar cosas, las redes llevan a interconectar pero el gran volúmen de datos lleva a predecir y cambiar cosas
- Antes había que elegir sólo dos de entre volúmen, velocidad y variedad
  - Bibliotecas: gran cantidad de datos variados pero lentas
  - Máquina de ordenar monedas: gran cantidad de monedas y rápido pero no son variadas

Los resultados inesperados de la abundancia

- Los estudios y algoritmos que estamos usando no son nuevos
  - Pero su uso indiscriminado lo es
- Antes existían soluciones específicas para grandes volúmenes de datos, a un costo muy elevado
  - Censo
  - Bancos
- Eficiencia $\Longrightarrow$ menores costos $\Longrightarrow$nuevos usos $\Longrightarrow$ $\Longrightarrow$mayor demanda $\Longrightarrow$mayor consumo.
  - Con más poder de cómputo, las necesidades de procesamiento de grandes volúmenes de datos están disparándose
  - La gente tiene necesidad de acceder a tecnología antes reservada para gobiernos y empresas multinacionales

Data Science

- Las soluciones de tipo bigdata son interdisciplinarias e involucran:
  - Hardware
  - Software
  - Análisis de datos
- Es el surgimiento del profesional especializado en data science

Conceptos de Bigdata

- Algunos conceptos que serán útiles:
  - Storage distribuido: para manejar grandes volúmenes de datos, es necesario poder almacenar datos en una red de computadoras
    - El más conocido es HDFS
  - Arquitectura de cómputo distribuido: utilizar la red de computadoras de manera eficaz
    - El más mencionado es Hadoop
    - Existe un abanico de soluciones, en esta charla vamos a hablar de ActiveMQ

Pasos del proceso de Bigdata

1. Adquicisión de datos
2. Limpieza de datos
3. Análisis de datos
4. Uso en predicción

Ejemplos Paradigmáticos

1. Construcción de un cluster Hadoop
2. Set-up de adquisión de datos en Hadoop (carga a HDFS)
3. Análisis específicos o a la espera

Acerca del presentador

- Licenciatura en Computación, Universidad Nacional de Córdoba, Argentina
  - Trabajo Final: “Desarrollo de un Parser Funcional para el Lenguaje Castellano”, presentado Ago. 1998.
- Columbia University
  - Generación de Texto
  - PhD Thesis: “Indirect Supervised Learning of Strategic Generation Logic”, defendida Ene. 2005.
- IBM Research Watson
  - Question Answering
  - Deep QA - Watson
- Viviendo en Montreal (Canadá)
  - Erudite Science, Inc.
  - Colaboración con Université de Montreal
  - Proyectos de Software Libre y consultoría para PyMES

Outline

Curso en UNC-FAMAF

- Aprendizaje Automático en Grandes Volúmenes de Datos
- El audio de las clases está grabado y junto con las presentaciones están disponibles gratis en el sitio Web de la materia:
  - http://aprendizajengrande.net
- El material didáctico está disponible bajo licencia CC-BY-SA.

Qué es el aprendizaje automático sobre grandes volúmenes de datos

- Aprendizaje Automático: un nuevo paradigma de programación
- Esta materia: cuando los datos y modelos no entran en RAM / disco de una sola máquina
- Importante para América latina porque no hay muchas máquinas / recursos

A quiénes está dirigida esta materia

- Estudiantes avanzados de carreras de grado
- Estudiantes de posgrado
- Profesionales del campo
- Prerequisitos:
  - Conocimientos de programación
  - Álgebra (particularmente álgebra matricial).
  - Probabilidad y Estadística
  - Redes y Sistemas Distribuídos (o similar, al menos Sistemas Operativos).

Estructura del curso

Tres partes:

1. Aprendizaje Automático (teórico)
2. Computo Distribuido (teórico)
3. Práctica (mahout/hadoop)

Parte I

- Modelos, Ingeniería de Features.
- Clasificación
  - Árboles de decisión
  - Regresión logística
  - SVMs
- Clustering
  - kMeans
  - Clustering estadístico
- Recomendación

Parte II

- Conceptos de Cómputo Distribuido
  - Map/Reduce
  - Teorema CAP
  - Operaciones Matriciales Distribuidas
  - Gradiente
  - Búsqueda distribuida
  - Algoritmos actualizables
  - Colas, shared memory
- Paralelizando algoritmos de Aprendizaje Automático

Parte III

- Implantación
  - Hadoop
    - Map
    - Reduce
  - Mahout
    - Recomendación
    - Clustering
    - Clasificación
  - ActiveMQ e Híbridos
- Casos de estudio

Casos de estudio

- Delicado equilibrio entre lo factible y lo útil
  - Datos disponibles
  - Problemas interesantes
- Clasificación: nombres para métodos compilados (http://keywords4bytecodes.org)
- Recomendación: git commits
- Clustering: entidades similares en DBpedia

Outline

El problema

from wikipedia

teaser day 1

Preguntas de Ejemplo

Categoría: "J.P."

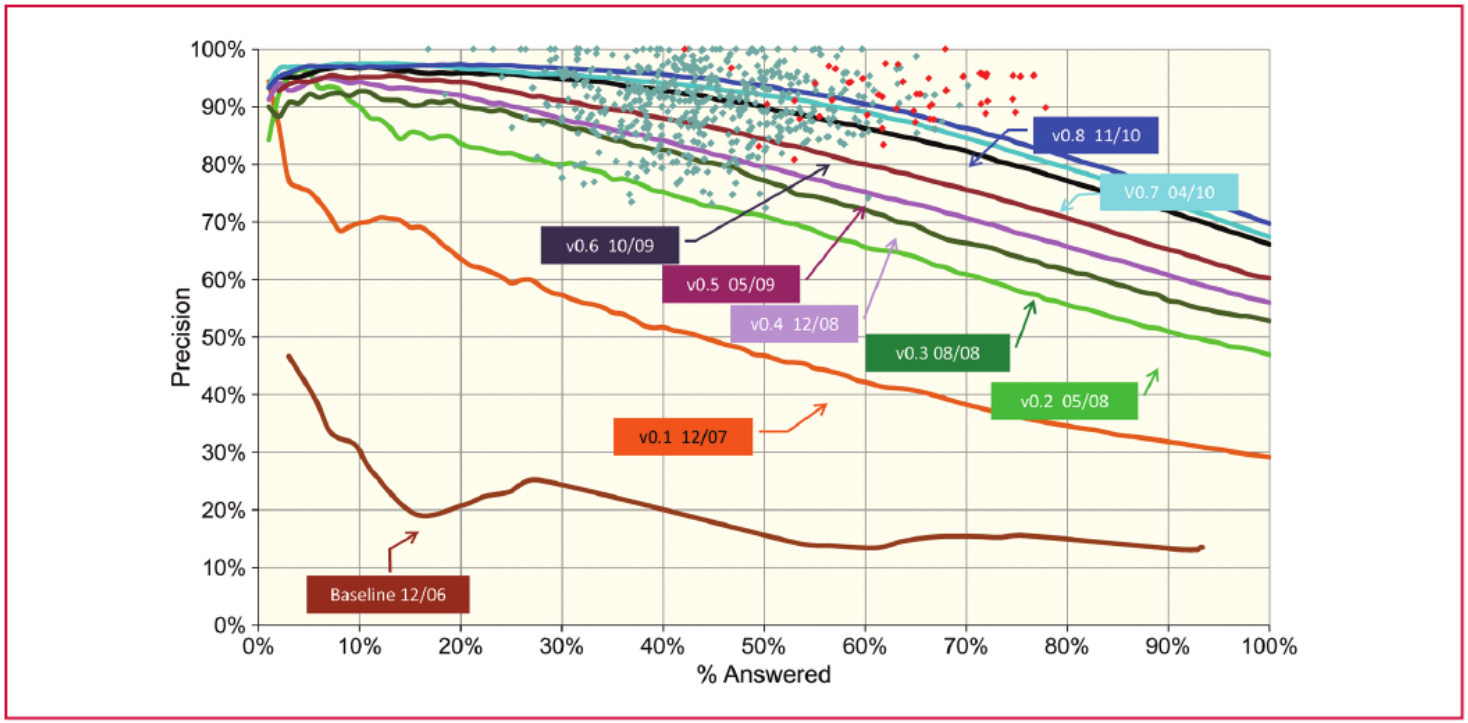
He played Duke Washburn, Curly's twin brother, in "City Slickers II".

- Respuesta: Jack Palance

Los Desafíos de un Equipo de Investigación

- Velocidad de desarrollo inusitadamente alta
  - Un *turn-around* experimental no es una propiedad *"nice to have"*, es clave
- Dead code
- Sin documentación
- Reproducibilidad de los resultados

Resultados



Progreso incremental desde junio del 2007 a noviembre del 2010, adaptado de Ferrucci (2012)

Los desafíos de un Grand Challenge

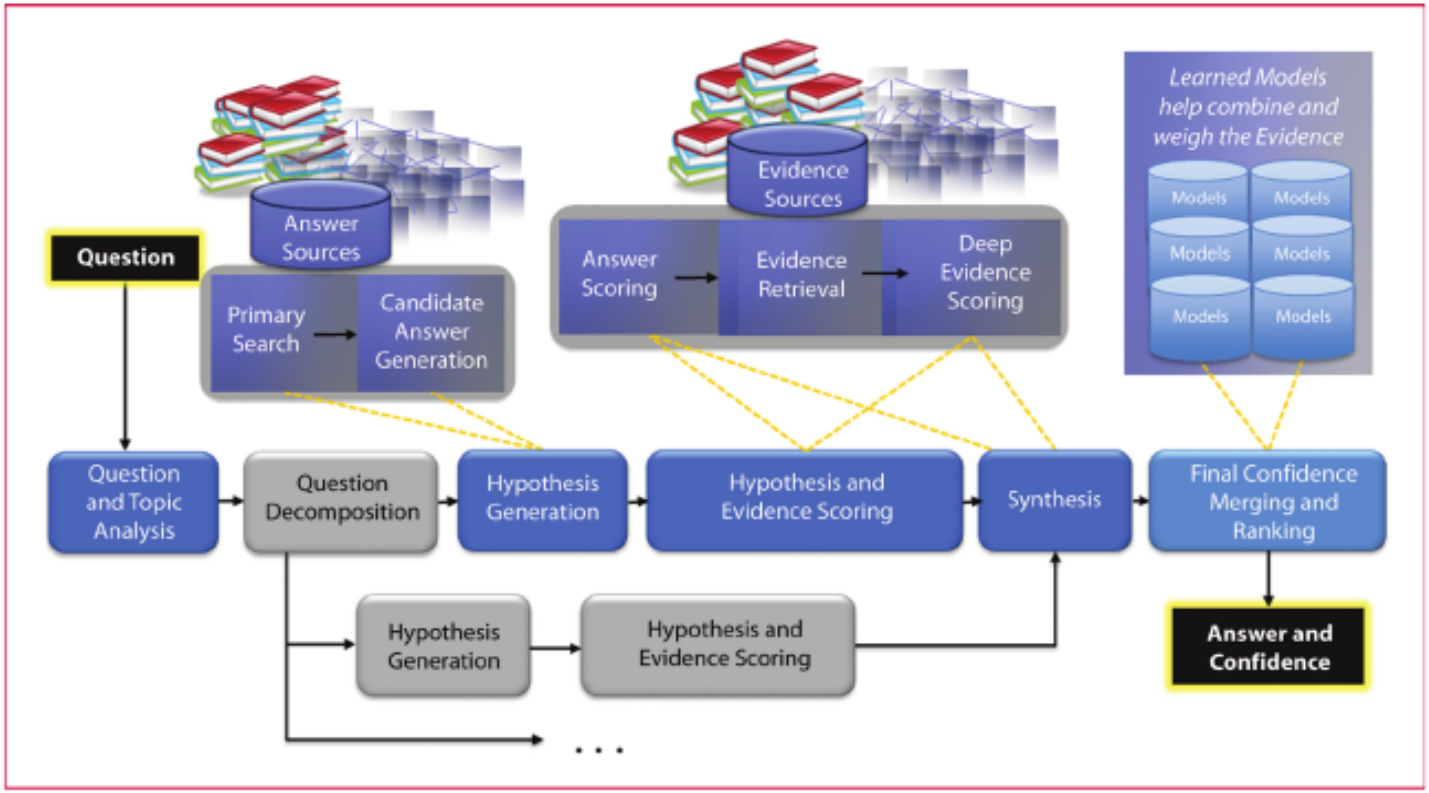
- Muy costoso.
- Todo el tiempo al borde de ser cancelado.
- Muchas cuestiones fuera del control del equipo de investigación.

Outline

Nuestra solución

- Mantener todas las interpretaciones abiertas hasta el final
  - No decidirse a algo antes de tiempo (*overcommit*)
- Proponer respuestas candidatas haciendo búsquedas
- Conseguir evidencia de soporte haciendo una búsqueda para cada respuesta candidata (!)
- Analizar todo esta cornucopia de información en paralelo
- *Scoring* y *ranking* centralizado usando Aprendizaje Automático

Arquitectura



DeepQA Architecture, adaptada de Ferrucci (2012)

Descripción de los componentes

Question Analysis. Extrae palabras clave y las asigna a clases dadas, expande entidades.

Primary Search. Busca documentos relevantes a la pregunta.

Candidate Answer Generation. Extrae de los documentos respuestas candidatas.

Evidence Retrieval and Scoring. Busca pasajes (oraciones) que contenga las respuestas y keywords relevantes, después valora el candidato en contexto.

Final Confidence Merging. Aplica un modello entrenado basado en la evidencia.

Outline

Qué es UIMA

- UIMA es un framework, una forma de integrar componentes analíticos para texto u otro tipo de información no estructurada.
- Es una implementación de referencia para Java, C++ y otros.
- Es un proyecto Open Source parte de la Apache Foundation.

Frameworks

- Los frameworks permiten:
  - Compartir y Colaborar
  - Crecer
  - Deployment e implementaciones de gran escala
  - Adopción
- Frameworks necesitan:
  - Mantenimiento (ningún software está 100% “terminado”)
  - Documentación (para ayudar con la colaboración / adopción)
  - Neutralidad (con respecto a las aplicaciones desarrolladas)
  - Ownership (por parte de los desarrolladores del mismo)
  - Publicidad (para un uso amplio)

Permitiendo Compartir y Colaborarar

- Compartir dentro de una organización
  - El código es la documentación
  - Compartir de manera ágil
  - Convention-over-configuration
- Compartir con el mundo
  - Un mundo mejor, sin pagar un alto precio (soporte, pérdida de ventures potenciales)
- Compartir con socios nuevos o potenciales
  - Educando gente nueva rápidamente
  - Atrayendo talento

Permitiendo Crecimiento

- Nuevos fenómenos
  - De análisis sintáctico a análisis semántico
  - De análisis de oraciones a anásis de tráfico USB
- Nuevos artefactos
  - De texto a voz
- Nuevas arquitecturas
  - De comprensión a generación

Permitiendo el Deployment e implementaciones de gran escala

- Multiples arcitecturas
  - Windows, Linux
- On-line vs. off-line
  - Procesado por lotes vs. servicios Web orientados al usuario
- Nuevos lenguajes de programación (y también viejos y eficaces)
- Nuevos lenguajes humanos

Analytics Frameworks

- Encuentre todos los números telefónicos en un texto
  - `(((\([0-9]{3}\))|[0-9]{3})-?[0-9]{3}-?[0-9]{4}`
- Funciona pero...
  - ¿Cómo seguimos procesando a partir de ahí?
  - ¿Cómo encontramos nombre propios en formato no-standard?
  - ¿Cómo integramos tecnología de vendors externos, proyectos open source, etc?

Anotaciones In-line

- Modificar el texto para incluir anotaciones
  - This/DET happy/ADJ puppy/N
- Se complica mucho muy rápido.
  - (S (NP (This/DET happy/ADJ puppy/N) (VP eats/V (NP (the/DET bone/N)))
- Y las anotaciones pueden cruzarse con otras fácilmente
  - He said **<confidential>**the project can't go on. The funding is lacking.**</confidential>**

Anotaciones Standoff

- Anotaciones Standoff
  - No modifican el texto
  - Mantienen el offset en el texto original
- La mayor parte de los frameworks de analytics usan anotaciones standoff.
- UIMA es construido a partir de anotaciones standoff.
- Ejemplo:

```
He said the project can't go on.  The funding is lacking.
---------------------------------------------------------
0123456789012345678901235678901234567890123456789012345678901234567
```

  - Sentence Annotation: 0-32, 35-57.
  - Confidential Annotation: 8-57.

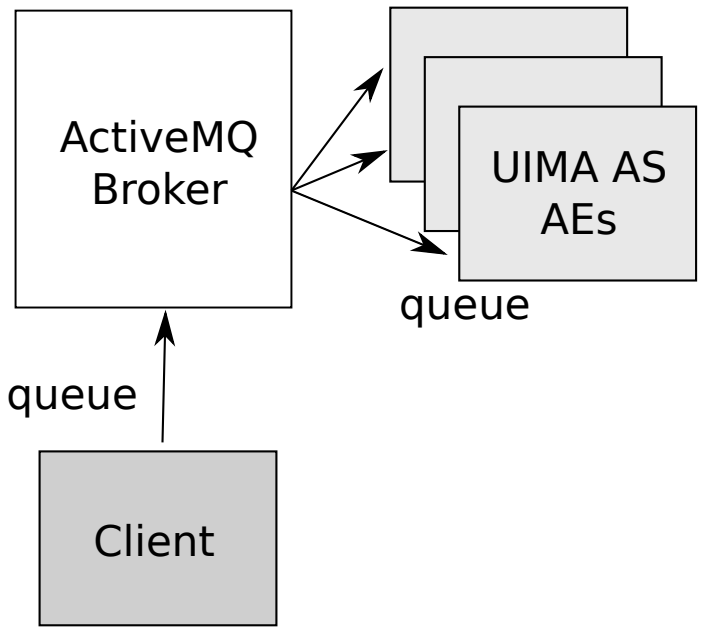
Sistemas de Tipos

- La clave para integrar paquetes de analytics desarrollados por terceros.
- Metadata clara acerca de
  - Entradas esperadas
    - Tokens, sentences, nombres propios, etc
  - Salidas producidas
    - Parse trees, opinions, etc
- El framework genera tu sistema de tipos unificado para los anotadores que se están ejecutando.

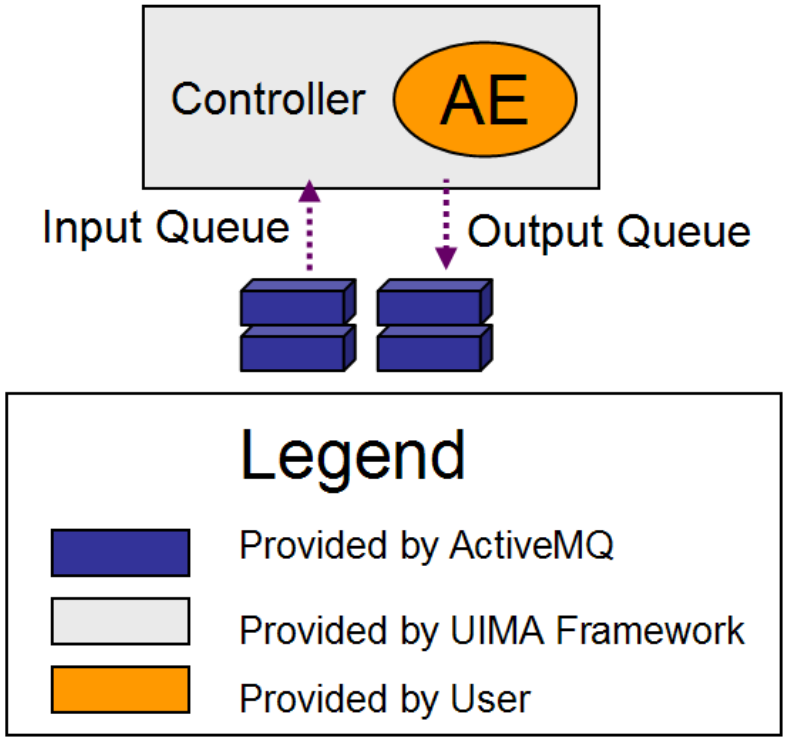
Ventajas de UIMA

- CAS
  - Eficiente uso de Memoria
  - Índices
- Tipos
- Interoperabilidad
- Protocolo de serializacion lean
  - UIMA AS envía y recupera de los nodos en la red sólo la información requerida
  - (la serialización XMI por defecto no es lean)

UIMA AS: ActiveMQ



UIMA AS: Wrapping Primitive AEs



UIMA AS: Ventajas

- Muy flexible en térmios de dividir la carga de trabajo entres los nodos
  - Tienes control total sobre como dividir las colas en sub-colas, etc.
- Muy eficiente en términos de *overhead* en la red
  - Una CAS que va a ser dividida y procesada varias veces (en partes distintas) es enviada sólo una vez.
  - Sólo las anotaciones **requeridas** son enviadas y las anotaciones **nuevas** son devueltas.
    - Archivos de metadata (descriptores) son clave para que ésto funcione

UIMA AS: más información

- http://uima.apache.org/doc-uimaas-what.html
- http://svn.apache.org/viewvc/uima/uima-as/trunk/README?view=markup
- http://uima.apache.org/d/uima-as-2.4.2/uima_async_scaleout.html

Muchos frameworks

- Aparte de UIMA
  - `http://uima.apache.org`
- LingPipe
  - `http://alias-i.com/lingpipe/`
- Gate
  - `http://gate.ac.uk/`
- NLTK
  - `http://www.nltk.org/`

Ventajas de UIMA

- Licencia Apache
- Código de calidad empresarial
- Escalabilidad demonstrada
- Desarrollado por expertos en construir frameworks
  - No expertos en el área (e.g., NLP)
- Interoperable (C++, Java, otros)

¿Qué tan difícil es aprender UIMA?

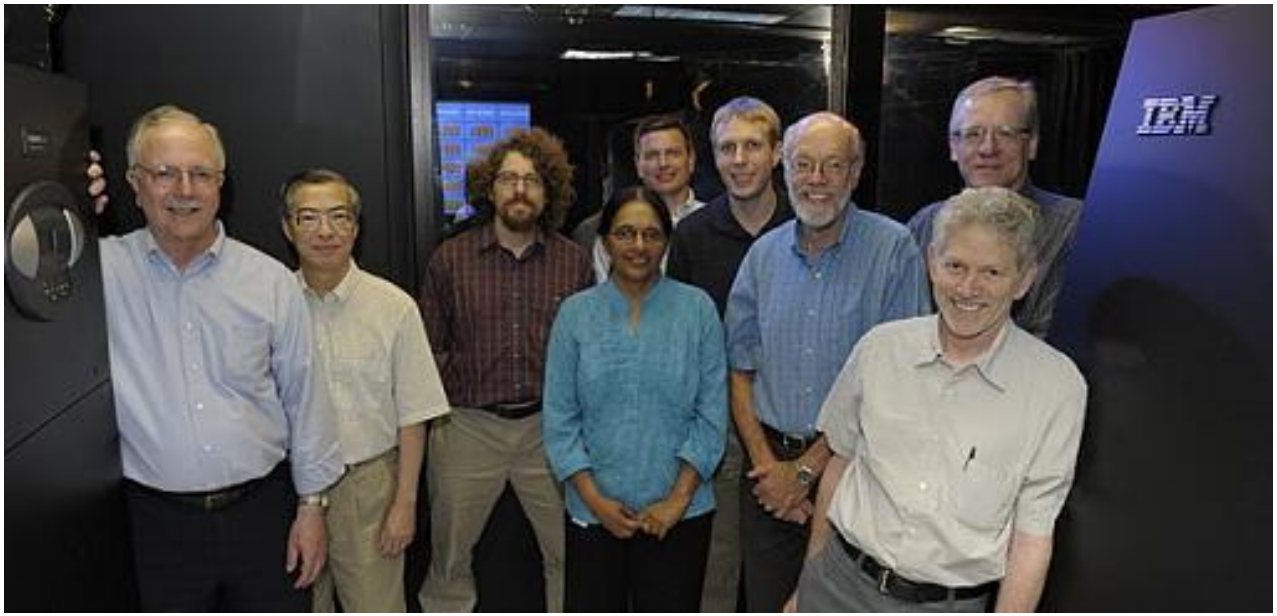
- Es bien difícil.
- La documentación es muy buena pero muy extensa.
  - Si pueden tomarse el tiempo de leerla de punta a punta, es de fácil lectura.
- Usen las herramientas de Eclipse cuando sea posible.
- Aprendan primero uimaFIT, después JCas, y CAS sólo si hace falta.
- Enfoquense en los *"goodies"*:
  - Apache UIMA Ruta – anotación basada en reglas
  - OpenNLP – modelos ya entrenados para POS, NER, etc., y bien fácil de entrenar tus propios modelos
  - ClearTk – un *wrapper* para librerias de aprendizaje automático

Outline

Mis Contribuciones al sistema Watson

- Sources Team
- Internal Tooling
- Machine learning

Systems Team

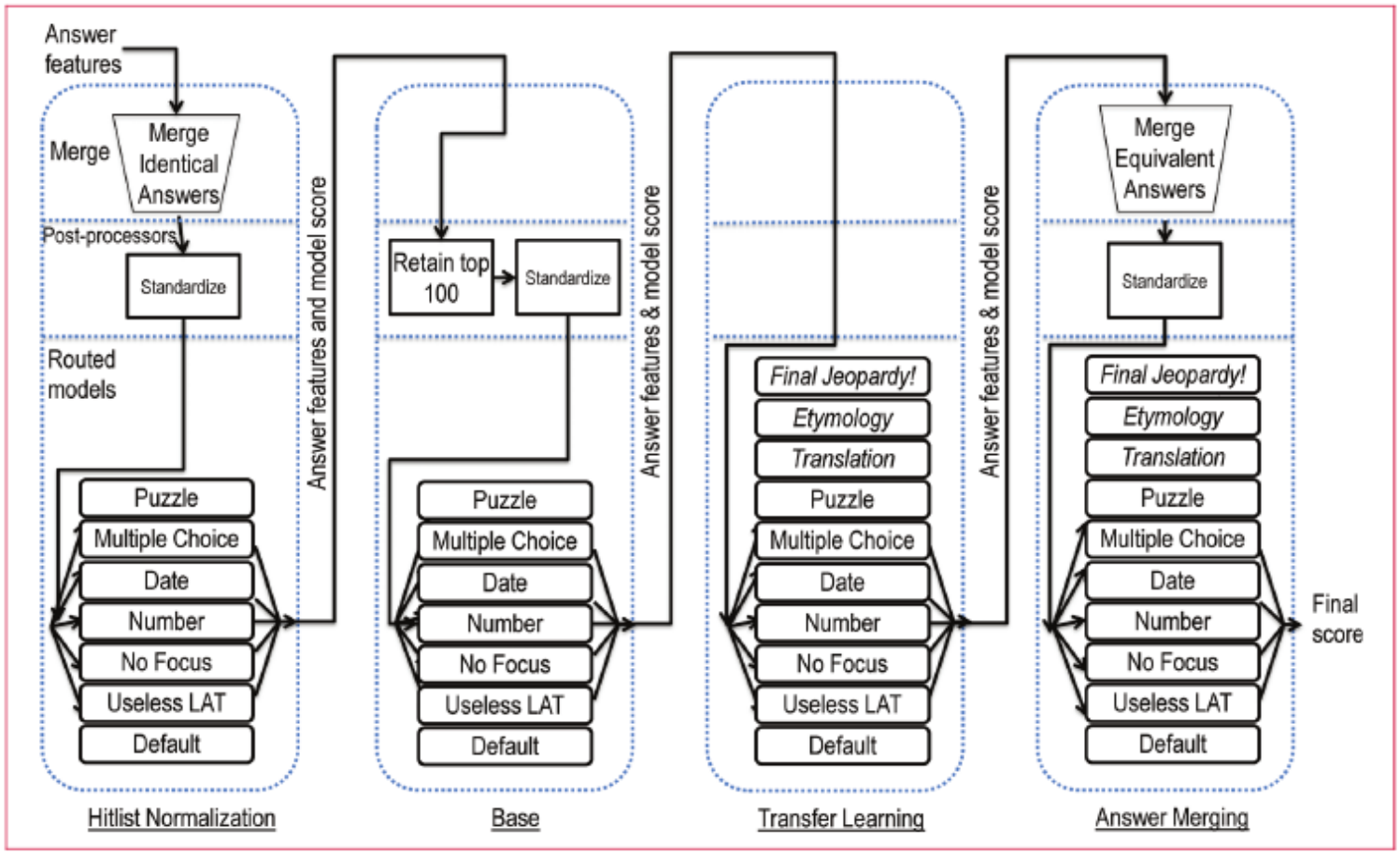


Systems Team, from `https://www.research.ibm.com/deepqa/`.

Aprendizaje Automático en Watson

- Múltiples fases de Regresión Logística
- Ingeniería de Features
- DSL para Ingeniería de Features

First Four Phases of Merging and Ranking



de Gondek, Lally, Kalyanpur, Murdock, Duboue, Zhang, Pan, Qiu, Welty (2012)

Outline

Después Watson

- Erudite Science, Inc. y Consultoria
- Trabajo Académico
  - Dictado de cursos
  - Hunter Gatherer
  - Thoughtland
- Free Software

Consultoría

- MatchFWD: datos LinkedIn
- UrbanOrca: datos Facebook
- KeaText: datos legales
- Radialpoint: datos de tech support
- Contact me at `http://duboue.net`

Trabajo Académico

- Dictado de la materia “Aprendizaje Automático sobre Grandes Volúmenes de datos” (http://aprendizajengrande.net).
- Dictado la materia “Generación de Lenguaje Natural” para el programa de doctorado en FAMAF-UNC.
- Algunas publicaciones recientes:
  - **Pablo Duboue**, Martin Dominguez and Paula Estrella. *“Evaluating Robustness of Referring Expression Generation Algorithms”*. MICAI (2015), to appear.
  - **Pablo Duboue**, Jing He and Jian-Yun Nie. *“Hunter Gatherer: UdeM at 1Click-2”*. NTCIR (2013).
  - Pablo Duboue. *“On the Feasibility of Automatically Describing n-dimensional Objects”*. EWNLG (2013).
  - Jing He, **Pablo Duboue**, and Jian-Yun Nie. *“Bridging the Gap between Intrinsic and Perceived Relevance in Snippet Generation””*. COLING (2012).
  - Fabian Pacheco, **Pablo Duboue**, and Martin Dominguez. *“On The Feasibility of Open Domain Referring Expression Generation Using Large Scale Folksonomies (short paper)”*. NAACL (2012).

Hunter Gatherer

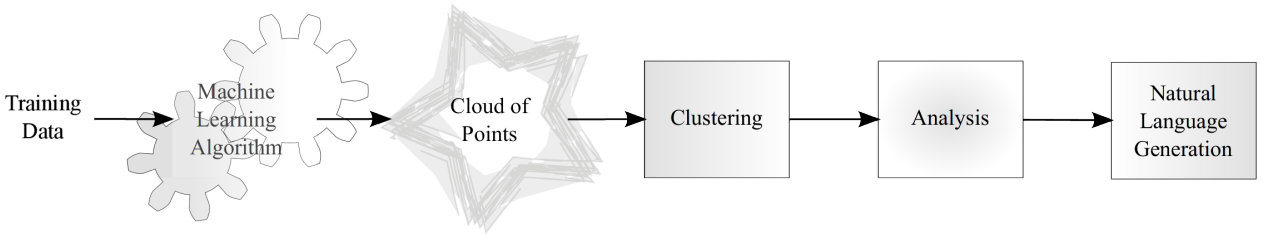
- ¿Qué es? 1-Click Search
  - Entrada: Query y 200 páginas Web en órden
  - Salida: un resúmen de 1,000 caracteres
    - Resúmen debe contener toda la información relevante a la query en las páginas
- Una research challenge parte de NTCIR
- Las queries pertenecen a 8 tipos (celebrities, how to, location, etc)
  - El tipo no es explícito

Hunter Gatherer Approach

- Utilizar la arquitectura DeepQA a la tarea 1-Click
  - No utilizar el tipo de query de manera explícita
- “Hunt nuggets, gather evidence”
  1. Buscar text nuggets en pasajes relevantes
  2. Acopiar pasajes de evidencia que contienen los nuggets y los términos de la query
  3. Generar un score para cada nuggets basado en la evidencia
  4. La salida final contiene oraciones con nuggets de alto score

`https://github.com/DrDub/hunter-gatherer`

Thoughtland



Thoughtland: Input

- A small data set from the UCI ML repo, the Auto-Mpg Data:

```
@relation auto_mpg
@attribute mpg numeric
@attribute cylinders numeric
@attribute displacement numeric
@attribute horsepower numeric
@attribute weight numeric
@attribute acceleration numeric
@attribute modelyear numeric
@attribute origin numeric

@data
18.0,8,307.0,130.0,3504.,12.0,70,1
14.0,8,455.0,225.0,3086.,10.0,70,1
24.0,4,113.0,95.00,2372.,15.0,70,3
22.0,6,198.0,95.00,2833.,15.5,70,1
27.0,4,97.00,88.00,2130.,14.5,70,3
26.0,4,97.00,46.00,1835.,20.5,70,2
```

... +400 more rows

Thoughtland: Output

- MLP, 2 hidden layers (3, 2 units), acc. 65%:

  There are four components and eight dimensions. Components One, Two and Three are small. Components One, Two and Three are very dense. ***Components Four, Three and One are all far from each other.*** *The rest are all at a good distance from each other.*

- MLP, 1 hidden layer (8 units), acc. 65.7%:

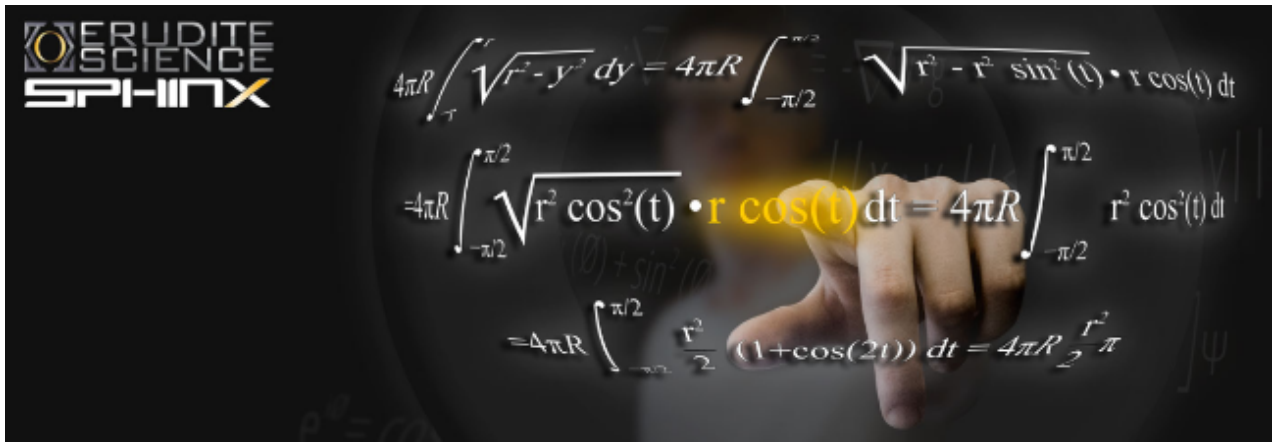
  There are four components and eight dimensions. Components One, Two and Three are small. Components One, Two and Three are very dense. ***Components Four and Three are far from each other.*** *The rest are all at a good distance from each other.*

(la diferencia está ***marcada***)

- MLP, 1 hidden layer (1 unit), acc. 58%:

  There are five components and eight dimensions. Components One, Two and Three are small and Component Four is giant. Components One, Two and Three are very dense. Components One and Four are at a good distance from each other. Components Two and

Erudite Science, Inc.

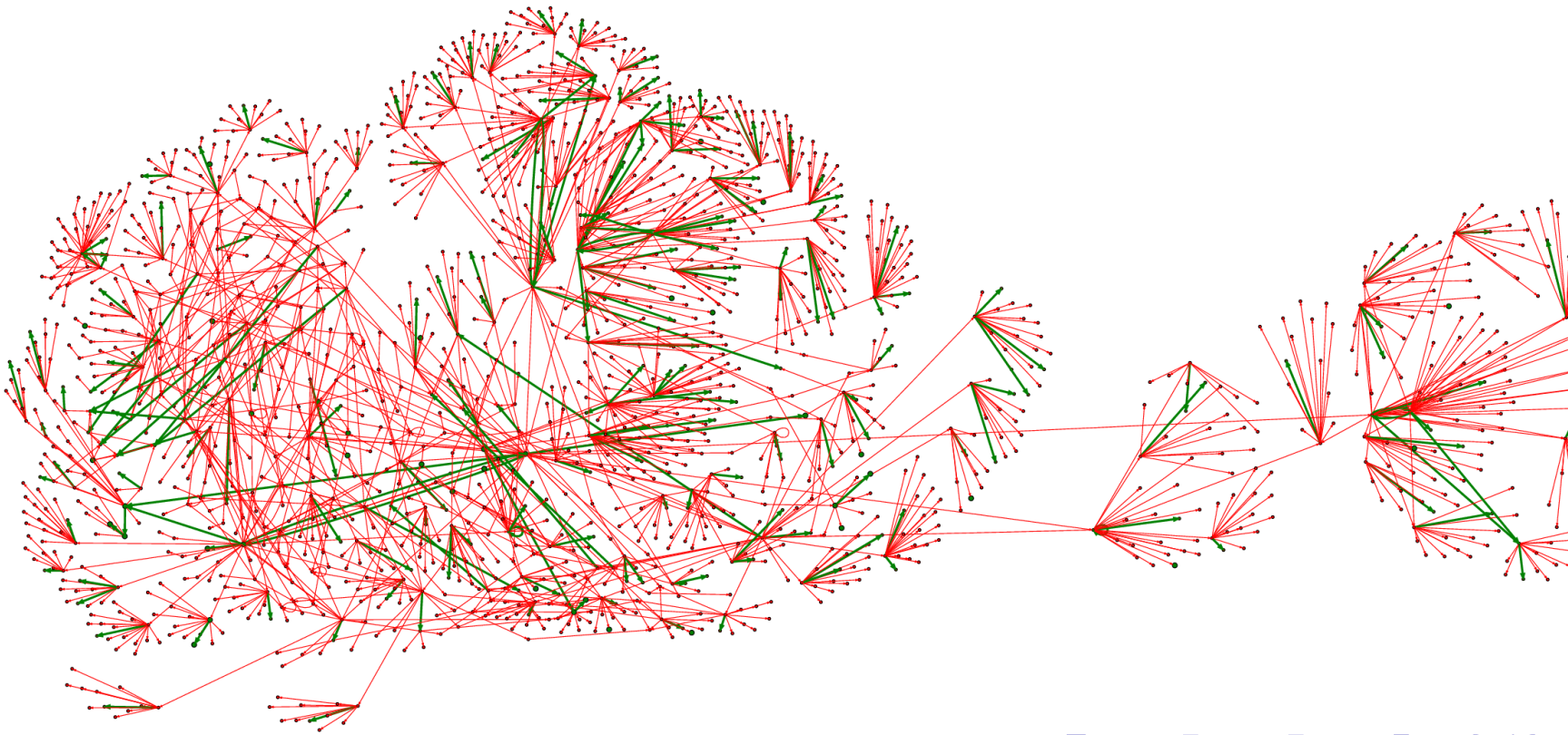


- Fundada en 2013
- Mejorar la educación matemática usando tecnología

Hacer de la tutoría personalizada un hecho para todos los alumnos, cuando y donde la necesiten, salvando las distancias entre alumnos, educadores y las aulas.

Nuestro Producto

- Sphinx: paso a paso tutor para resolución de expresiones formulaicas
- [Demo]



Resúmen

- Con Bigdata se puede acceder a tecnología antes reservada para gobiernos y empresas multinacionales
- Para el proceso de información no estructurada se puede usar el framework UIMA
  - UIMA es un *framework* para procesamiento de información no-estructurada listo para usar en producción.
    - Permite procesamiento por lotes o con muy baja latencia.
- Para aprender más de aprendizaje automático y bigdata: http://aprendizajengrande.net