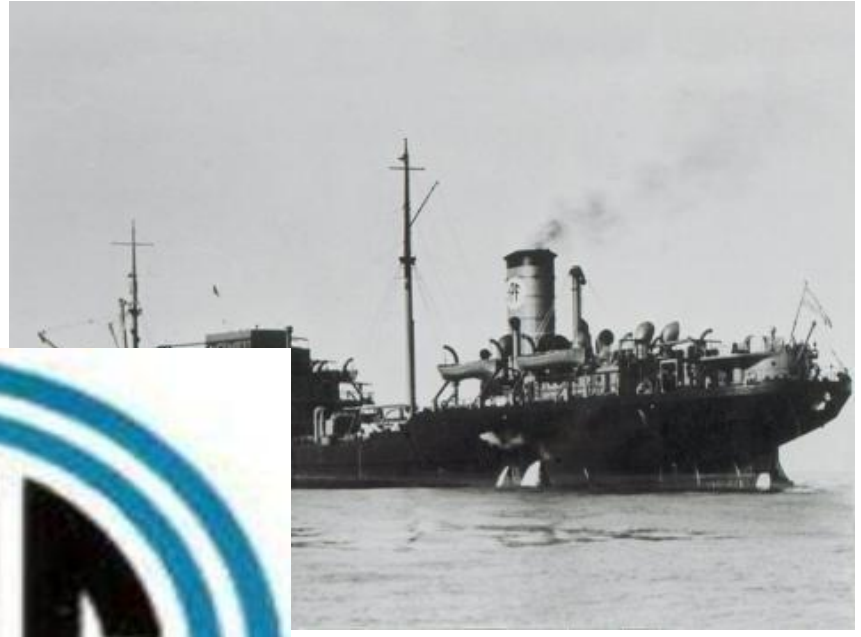
SOBRE LAS NUEVAS TENDENCIAS Y DISRUPCIONES EN LAS TICS

GRANDES DATOS EN ARGENTINA

Gabriel Baum – Setiembre de 2013



YPF

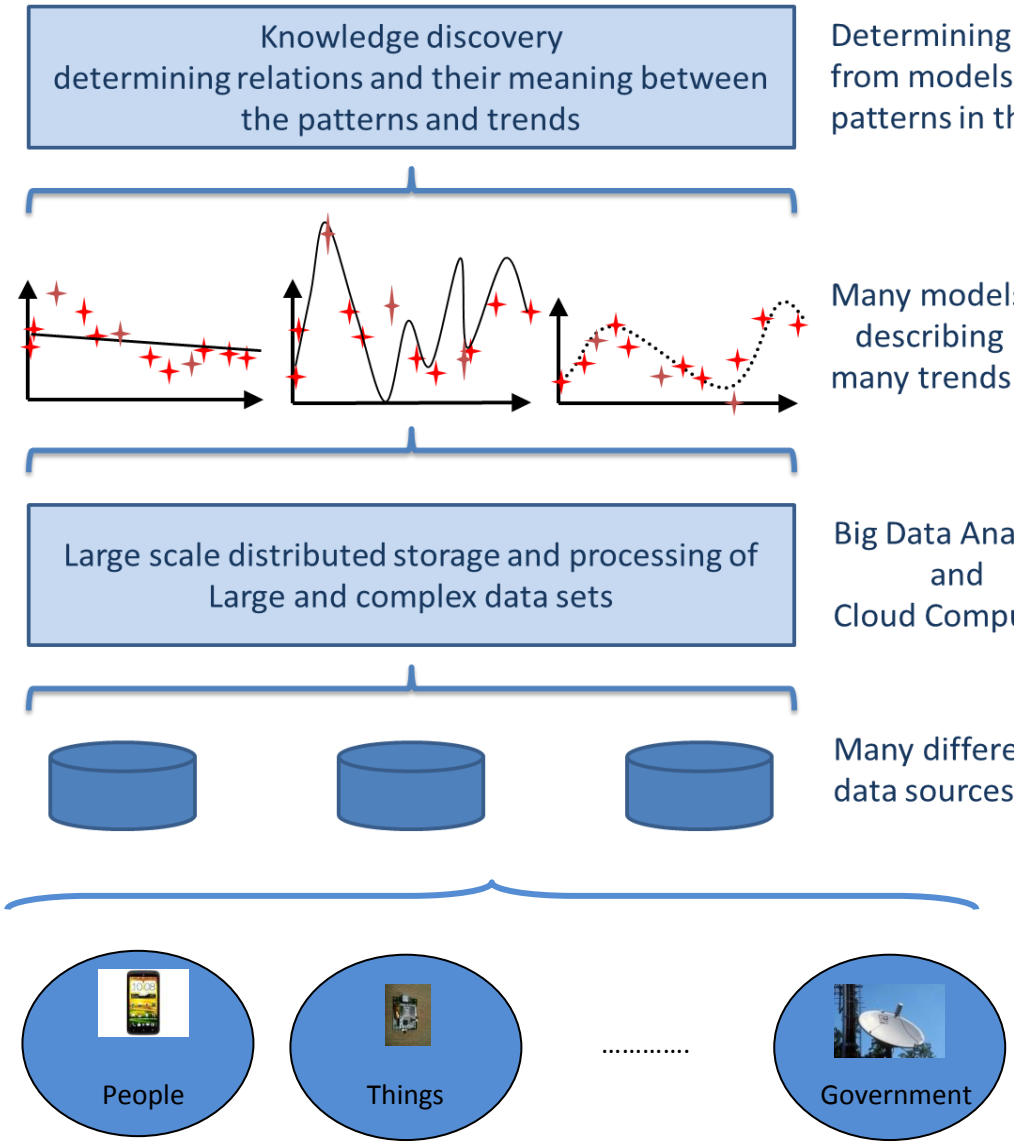
Prospectiva para las TICs en el marco de la planificación estratégica de MINCYT

Areas Técnológicas y de Aplicación Seleccionadas
(novedades respecto del Libro Blanco de la Prospectiva en TICs)

- **Big Data, Pervasive and Trusted Computing and Networks**
- **Componentes y Sistemas**
- **TICs para la salud**
- **TICs para Energía y medio ambiente**
- **e-Government**
- **Contenidos Digitales y Culturales**
- **Infraestructura para la TICs**

Escenario ingenuo de convergencia de las tecnologías[i]
[Cloud computing+Big Data] + [Cheap Phones + IoT]

"**All models are wrong**, but some are useful."
Box, George E. P.; Norman R. Draper (1987). *Empirical Model-Building and Response Surfaces*, p. 424, Wiley



"**All models are wrong**, and increasingly you can succeed without them."
Peter Norvig, Google's research director

(Deep Learning)

[i] Versión modificada de *2013: Convergence of Big Data and Cloud Computing*, posted by Anton Heijs, dec 31, 2012, http://treparel.com/news/convergence-big-data-cloud-computing/

The Model Has Changed…

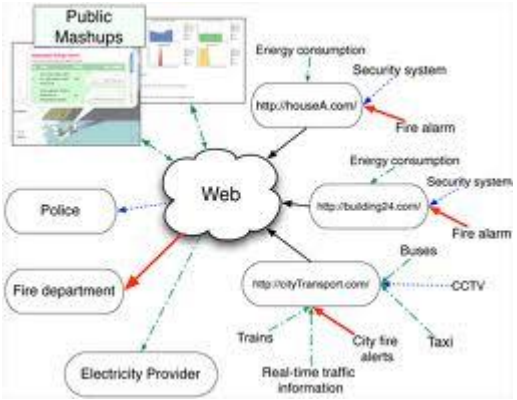
- **The Model of Generating/Consuming Data has Changed**

Old Model: Few companies are generating data, all others are consuming data

New Model: all of us are generating data, and all of us are consuming data

Internet of theThings (IoT) a big producer of Data

- HP envisions 1 trillion sensors in use around the world
- There are many types of sensors
  - temperature, pressure, level, humidity
  - speed, motion, distance
  - light or the presence/absence



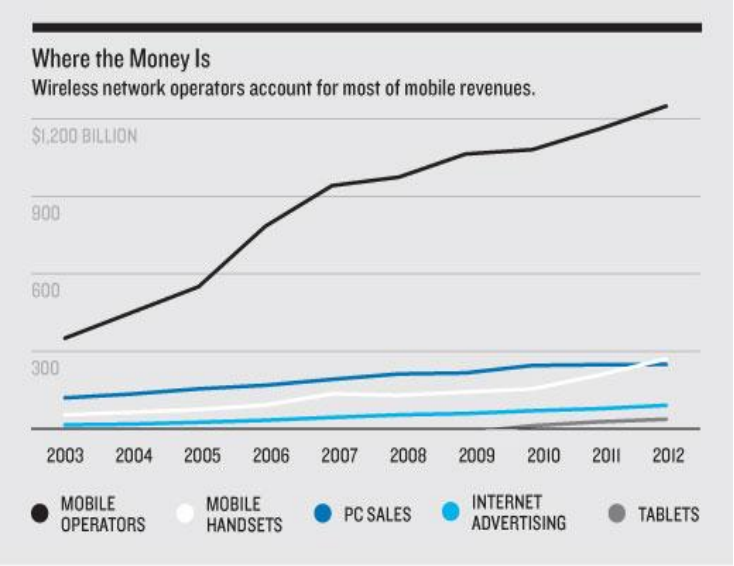
Datos, datos y más datos…datos en todas partes…los datos son importantes…

En 2015, 3 billones de personas estarán on-line, llevando los datos creados y compartidos hasta cerca de 8 zettabytes.

- 30 billones de piezas de contenidos se sumaron a Facebook en un mes generados por 600 millones de usuarios.
- Zynga procesa 1 petabyte de contenidos de jugadores cada día
- Más de 2 billones de videos fueron mirados en YouTube … ayer.
- El “adolescente promedio” envia 4762 SMSs por mes.
- 32 billones de búsquedas por mes … en Twitter.
- El tráfico mundial sobre IP se cuadruplicará en 2015

http://smartdatacollective.com/tusharrn/42826/data-data-and-more-data

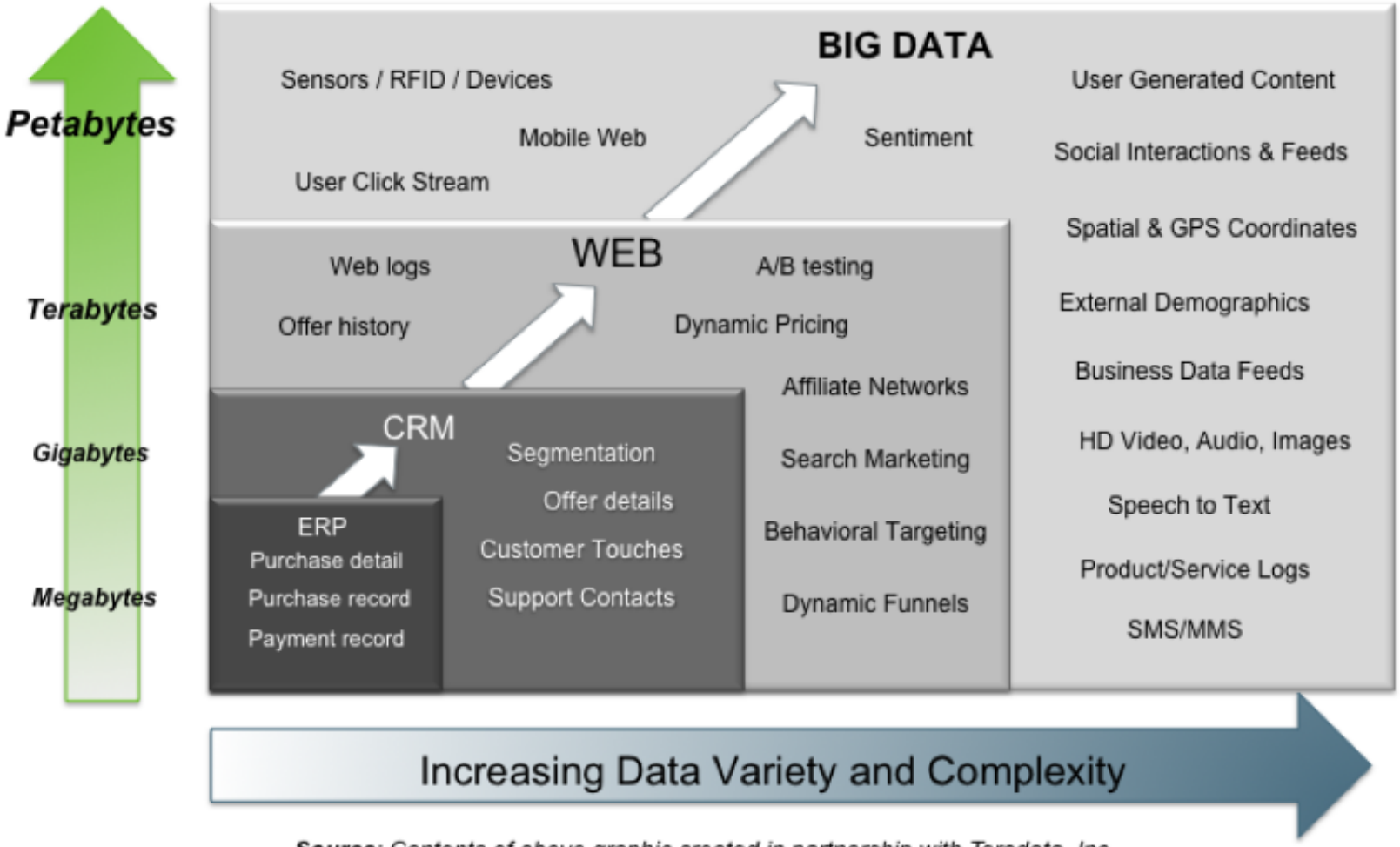
La expansión mundial de la conectividad (a Internet) móvil, que alcanza ya más de 1.100 millones de personas en el planeta, un mercado inmenso y casi virgen
Además, 5.800 millones de personas aún no están conectados, de los cuales 4.500 millones nunca han accedido a Internet. Es decir, el negocio de la computación móvil recién comienza (o tal vez aún no comenzó), la gran tajada de la torta se la llevan las compañías de comunicaciones móviles (wireless carriers): globalmente 900 de ellos ganan U$S1.3 trillones anuales, alrededor de 4 veces la ganancia sumada de Google, Apple, Microsoft e Intel, los famosos grandes ganadores de las TICs, aún cuando los constructores de dispositivos –especialmente Apple- consiguen más ganancias individualmente.



Source: Benedict Evans, Enders Analysis

Las aplicaciones y servicios aún representan el rubro de menor facturación en computación móvil. Los avisos comerciales para móviles se estiman en solamente U$S9 billones, sin embargo es en este punto donde aparecen las mayores oportunidades. Facebook tiene una audiencia mensual nunca antes alcanzada (alrededor de 1000 millones de personas), y en enero de 2013 se estima que, por primera vez, la audiencia de dispositivos móviles fue mayor que la de PCs.

Big Data = Transactions + Interactions + Observations



***Source**: Contents of above graphic created in partnership with Teradata, Inc.*

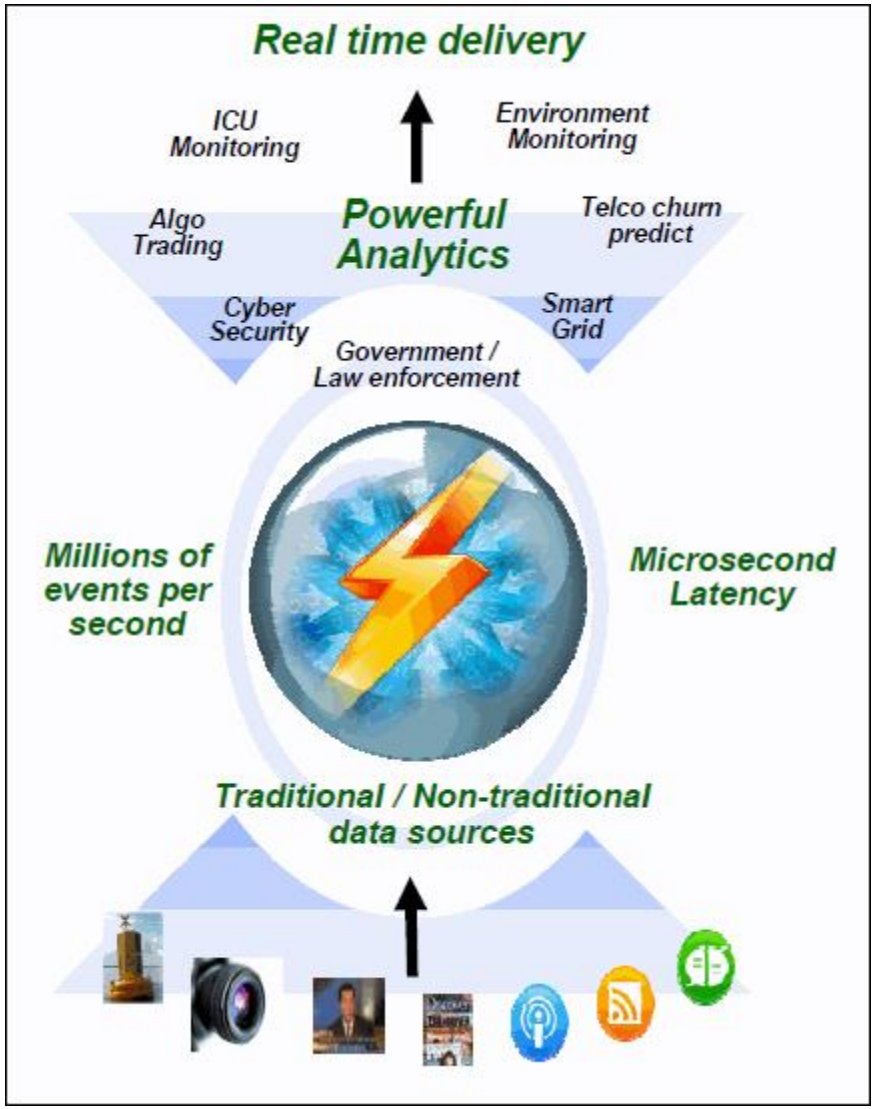
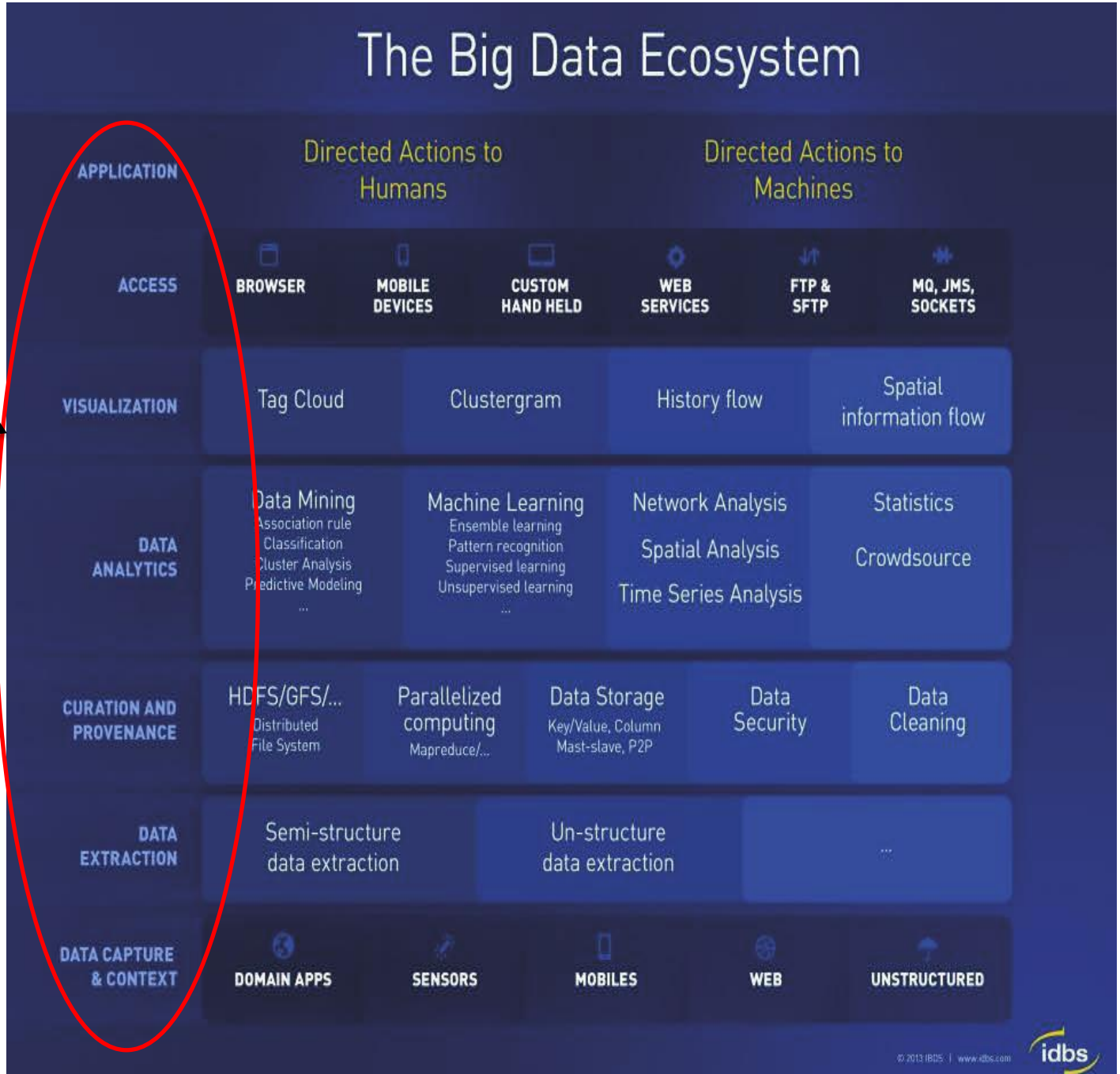
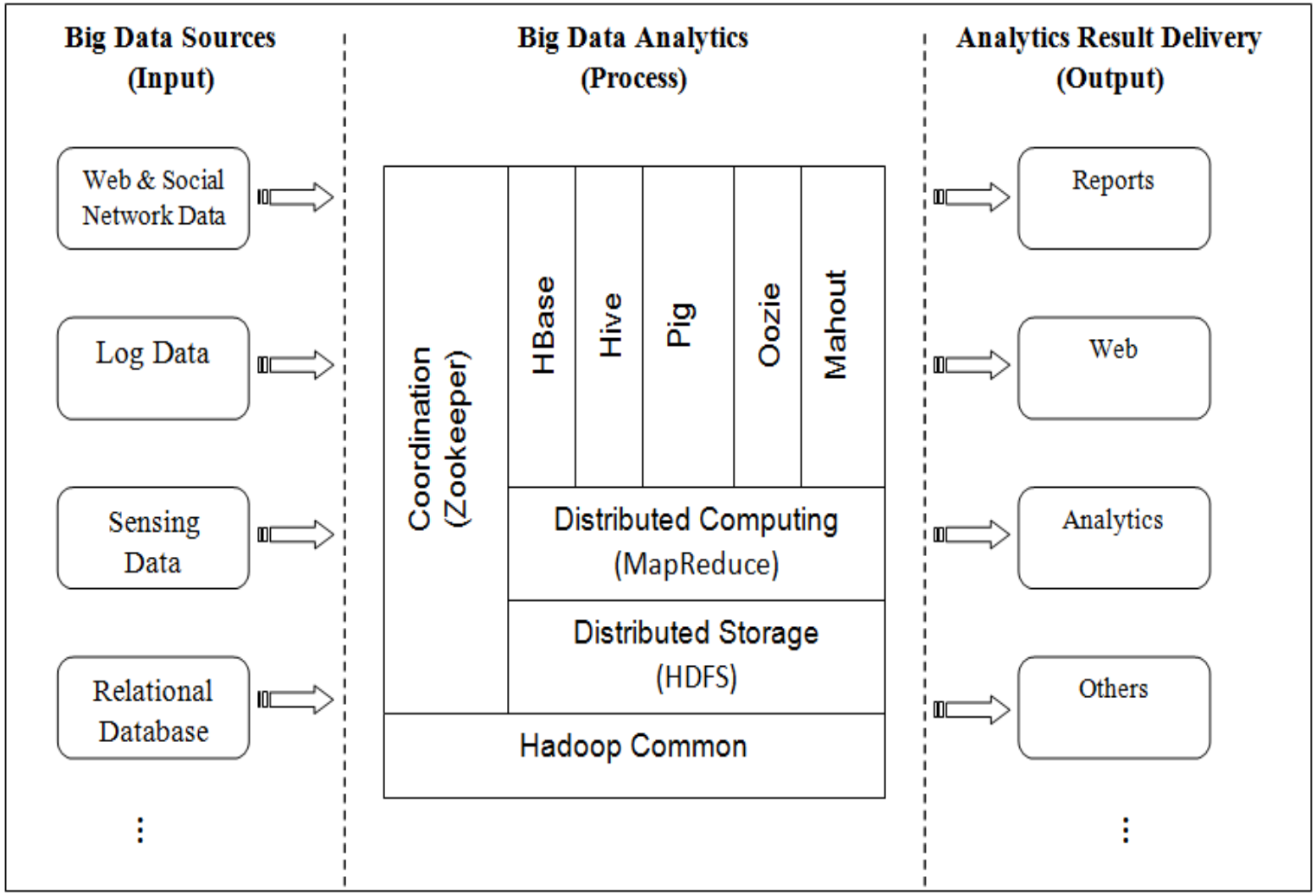


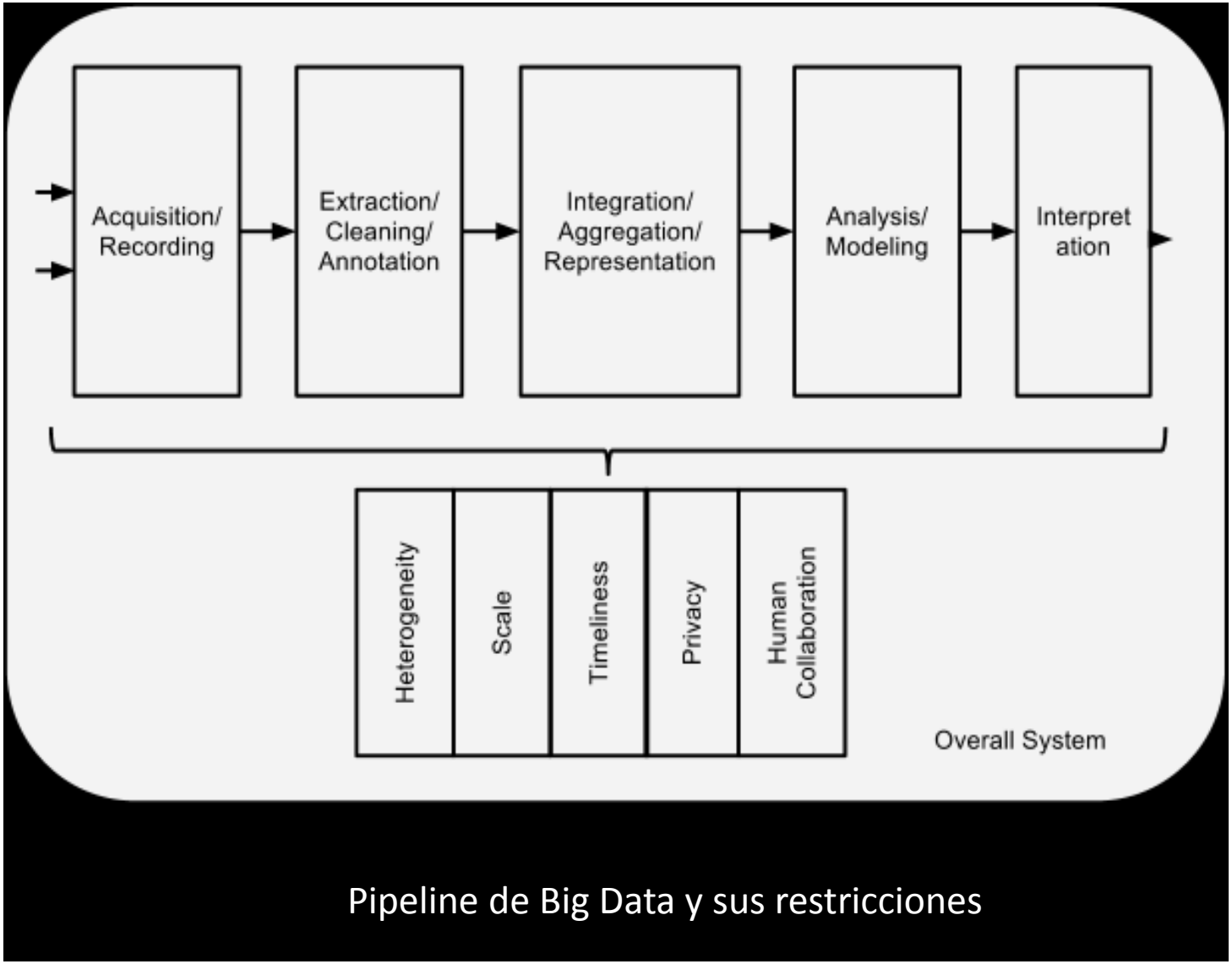
Figure 2. InfoSphere Streams solution

Pipeline of Big Data



Tecnología (software) para Big Data





Pipeline de Big Data y sus restricciones

A number of challenges in both data management and data analysis require new approaches to support the big data era.

- Dealing with highly distributed data sources,
- Tracking data provenance, from data generation through data preparation,
- Validating data,
- Coping with sampling biases and heterogeneity,
- Working with different data formats and structures,
- Developing algorithms that exploit parallel and distributed architectures,
- Ensuring data integrity,
- Ensuring data security,
- Enabling data discovery and integration,
- Enabling data sharing,
- Developing methods for visualizing massive data,
- Developing scalable and incremental algorithms, and
- Coping with the need for real-time analysis and decision-making.

….But the challenges for massive data go beyond the storage, indexing, and querying that have been the province of classical database systems (and classical search engines) and, instead, hinge on the ambitious goal of *inference.*

Inference is the problem of turning data into knowledge, where knowledge often is expressed in terms of entities that are not present in the data per se but are present in models that one uses to interpret the data.

In summary, the major research areas are :

• Data representation, including characterizations of the raw data and transformations that are often applied to data, particularly transformations that attempt to reduce the representational complexity of the data;

• Computational complexity issues and how the understanding of such issues supports characterization of the computational resources needed and of trade-offs among resources;

• Statistical model-building in the massive data setting, including data cleansing and validation;

• Sampling, both as part of the data-gathering process but also as a key methodology for data reduction; and

• Methods for including humans in the data-analysis loop through means such as crowdsourcing, where humans are used as a source of training data for learning algorithms, and visualization, which not only helps humans understand the output of an analysis but also provides human input into model revision.

Big Data: un nuevo microscopio

Hace más de un siglo Frederick Taylor puso en la agenda la "cientifización" de las organizaciones (Principles of Scientific Management). Pero este concepto fue esencialmente entendido como la mejora continua en el mangement.

Big Data promete un cambio fundamental, similar al advenimiento del microscopio, estaríamos en el comienzo de una nueva era comparable con el inicio de la ciencia moderna hace 400 años.

El "microscopio" digital , que está siendo inventado en este momento para tener Big Data, nos permitiría rápidamente analizar y predecir eventos de todo tipo, mucho más exactamente y científicamente . Descubrir estructuras y conexiones que permitirán que posibilitarán obtener impensadas tendencias, patrones, insights y soluciones, y tomar mejores decisiones. Nuevas técnicas: Data Driven Decisions y Predictive Analysis.

Big Data como una nueva era: Data Science

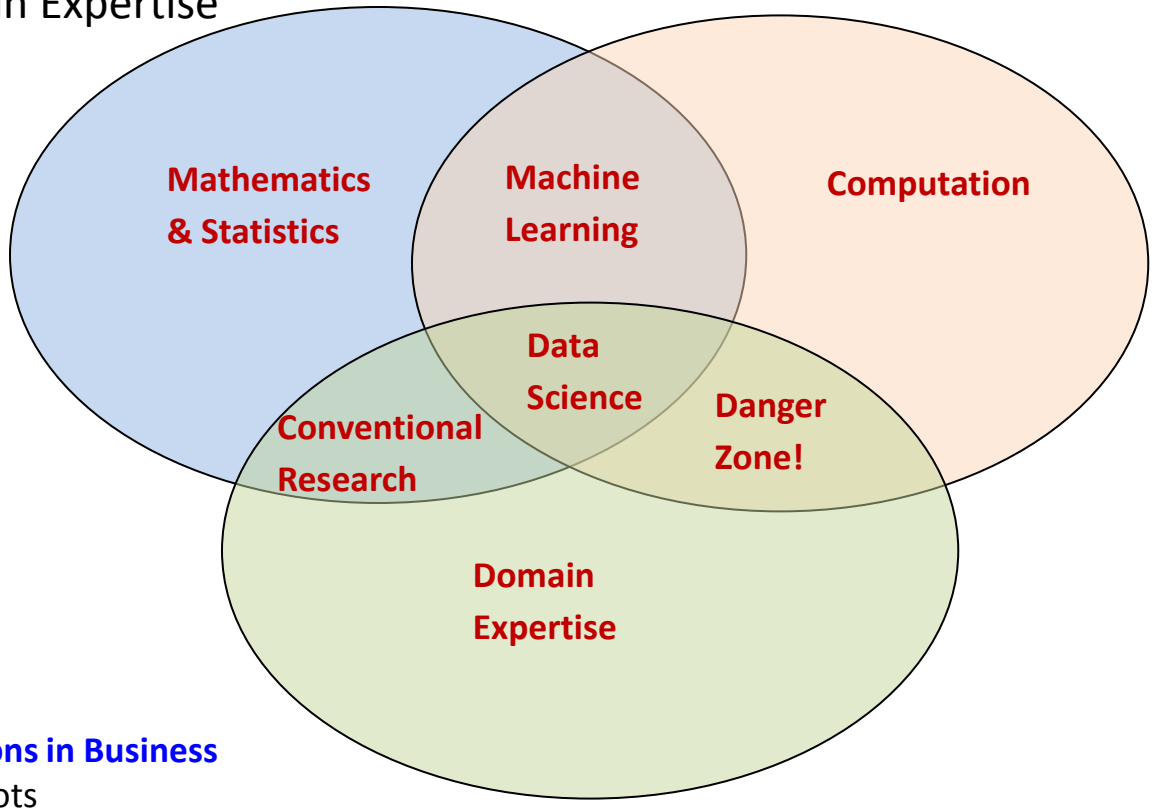
What is Data Science?

Data Science: automatically extracting knowledge from data

Mathematics & Statistics

Machine Learning

Domain Expertise



Applications in Business

Lots and lots

Applications in the Sciences

Astronomy, Cosmology

High-energy Physics

Biology, Genomics

Neuroscience

The Social Sciences

Medicine

Government

(Yann LeCun - NYU - 2013)

Perfil de un Data Scientist (Chirag Metha y otros):

- Comprensión profunda de los datos y las fuentes de datos a explorar, y descubrir los patrones que ocurren en los mismos (los datos y sus fuentes)
- Alto nivel tanto teórico como práctico (herramientas) en estadística y algoritmos estadísticos avanzados, sistemas complejos, visualización, machine learning e inteligencia artificial.
- Conexión estratégica con la organización en todos los niveles para poder comprender amplia y profundamente los desafíos de la organización y poder traducirlos en términos de diseño de experimentos con los datos.
- Diseñar e instrumentar el ambiente y las aplicaciones necesarias para generar y reunir nuevos datos, y establecer una estrategia organizacional de los datos. Materializar la noción de que Big Data es “Total Data”, es decir, no dejar ningún dato “atrás” y no tener “silos”

EN TERMINOS ARGENTINOS….

Superficie sembrada de la Argentina: 30.000.000 de hectáreas =

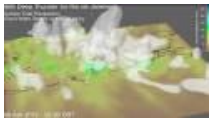
$3x10^7 Ha = 3x10^{11}\ m^2$

Si queremos tener informacíon por cada m^2 (por ejemplo, 100 bytes por m^2, información biológica sensada, satelital, climática, etc) cada dia del año, tenemos:

$$3x10^{11}x10^2x4x10^2 = 12x10^{15} \text{ bytes} = 12 \text{ petabytes anuales}$$

Es factible almacenar esta información en Argentina. Si fuera posible generarla y procesarla se podría tener entender la dinámica de cada metro de la superficie, pero lo más interesante es que se podría comprender la dinámica de todo el sistema. Esto es Agro TICs en la era de BigData.

Caso de Uso : Agricultura de Precisión



Clima

Web Logs

Sensores

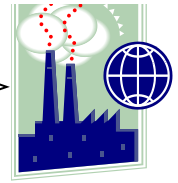
Redes sociales

facebook

Big Data Analytics

Decision Making

Solución integrada de la cadena de suministros



Múltiples fuentes de información y grandes masas de datos.

Varias fuentes de información (temperatura, humedad, sensores de suelo, sensores de cultivos, etc.).

Varios tipos de sensores (digital, análogo, cableado, celular)
Varios tipos de datos (cualitativos, numéricos)
Diferentes escalas temporales (casi continua, horaria, diaria, semanal, etc.)
Pronóstico del climatico
Demanda del mercado
Análisis de "consumer-sentiment"

Aplicar análisis sobre los datos para extraer modelos e "insights"

Crear nuevos modelos complejos y dinámicos

Utilizar modelos preexistentes para algunos subsistemas (demanda de agua, predicción del clima)

Múltiples objetivos, a veces contradictorios:
- Maximizar rendimiento de los cultivos
- Minimizar utilización de agua
- Minimizar deshechos

Los expertos y tomadores de decisiones usan la solución para recomendar los cultivos adecuados para plantar, cantidades y tiempos

La solución comprende modelos para cada cultivo, para varios cultivos que comparten propiedades (p.ej., cultivos en áreas secas, cultivos que requieren grandes cantidades de agua, etc.)
Resultados: altos rendimientos y mínimos recursos

Y ahora les cuento una anécdota de Delia Beltrán:

estaba dando un taller durante su exilio en México a unos muchachos indios (así los describe ella) y propuso un ejercicio de clasificación de semillas de porotos de varios colores y granos de maíz también diferentes entre ellos. Cuando se acerca a ver el resultado de la clasificación se encuentra con que habían armado 7 montoncitos de semillas mezcladas, que no era precisamente lo que se puede esperar de un trabajo de clasificación; fue un muchacho mazahua el que dijo "tiempo de siembra, Maestra" y esa simple explicación expuso que los conocimientos técnicos no tienen valor frente a la sabiduría que otorga el respeto por la naturaleza, ya que la mejor manera de sembrar es maíz y porotos juntos, porque se dan mutuamente lo que les falta, la fuerte caña del maíz sirve de sostén al poroto que trepa buscando el sol y que provee a la tierra de los nitratos que necesita el maíz, y el zapallo sembrado entre ellas crece arrastrándose aprovechando la materia orgánica acumulada, protegido de las heladas.

De esta forma, las plantas no sólo se apoyan entre ellas en una perfecta simbiosis sino que aseguran una dieta con suficientes hidratos de carbono y proteínas para grupos humanos que consumían y consumen poca carne.

Propuesta

AGranDAr

Alianza por los Grandes Datos de Argentina

Una estrategia de Ciencia de los Datos para la Argentina
2013-2018

MINCYT – Fundación Sadosky

AGranDAr - (Alianza por los Grandes Datos de Argentina)

Una estrategia de Ciencia de los Datos para la Argentina 2013-2018
Conclusiones generales de las Jornadas de Planificación Estratégica

✓*Grandes Datos* (Big Data) representa un cambio de paradigma en la forma en que se descubre, extrae y genera conocimiento, constituye una disrupción en el modo de resolver ciertos problemas en la ciencia, la producción y los negocios.

✓El desarrollo en esta temática es una cuestión de soberanía nacional y regional

✓Se trata de un área que se encuentra todavía en los momentos iniciales de su desarrollo. Por lo tanto es posible convertir a la Argentina en líder regional en este campo.

✓*Grandes Datos* presenta una gran oportunidad para el país ya que en el contexto actual es posible generar mucho valor a partir de insumos disponibles: datos, infraestructura, recursos humanos, conocimientos. La Ciencia de Datos es, por su naturaleza, multidisciplinar e interdisciplinaria.

✓En resumen, *"Grandes Datos es una Gran Oportunidad"*.

Diagnóstico de la Situación en Argentina

✓ Existe ya un pequeño número de compañías surgidas a partir de la calidad y buena formación de nuestros recursos humanos, y de la disponibilidad de algunos insumos y recursos, tanto datos como hardware o infraestructura de comunicaciones.

✓ En lo que refiere al sector académico, hay varios grupos que están trabajando en temáticas relacionadas (bioinformática, optimización, web y data mining, procesamiento de lenguaje natural, etc) pero no existe ningún grupo o institución que esté dando un enfoque abarcativo al tema.

✓ A futuro, se vislumbra una limitación en relación a los Recursos Humanos. Si se forma una masa crítica de profesionales se estará generando la base para aprovechar la oportunidad y generar mayor valor.

✓ Es necesario encarar un esfuerzo de formación direccionado hacia un conjunto de capacidades al mismo tiempo que generar iniciativas y proyectos desde el Estado Nacional que traccionen la actividad, potenciando las empresas emergentes y generando nuevas que permitan sinergias y la generación de nuevo valor.

✓ Si en cambio no se generan suficientes conocimientos, puestos de trabajo, proyectos y negocios, el riesgo es que la base ya existente y lo que se vaya generando sirva simplemente para alimentar las necesidades de los países que han tomado el liderazgo en este tema.

AGranDAr - (Alianza por los Grandes Datos de Argentina)

Una estrategia de Ciencia de los Datos para la Argentina 2013-2018

Borrador de trabajo

Visión: Transformar a la Argentina, en los próximos 5 años, en un actor relevante –como país "periférico"- líder en la región sudamericana, de la "revolución de Big Data".

Estrategia: Basarse en los datos y en las capacidades locales para construir una plataforma de datos y de conocimientos propios que permitan el desarrollo de una escuela de pensamiento y acción originales en torno a Big Data. La construcción de esta plataforma potenciará el avance en las ciencias, las tecnologías, el desarrollo de servicios para la economía y la sociedad, el crecimiento y la creación de nuevas compañías del conocimiento, y la competitividad en los mercados internacionales. La plataforma debenecesariamente reunir al sector público, con las empresas y la academia, de otro modo sería imposible su desarrollo; Big Data es, por su naturaleza, multidisciplinar e interdisciplinaria.

La originalidad –unicidad- de esta plataforma está dada por los propios datos y por los objetivos y aplicaciones específicas, provenientes de la realidad argentina y adecuados a las necesidades nacionales.

Una parte relevante de esta plataforma estará conformada por ***datos geolocalizados provenientes de áreas estratégicas para el país*** (recursos naturales, actividad económica, salud, educación, seguridad, defensa, innovación, etc.), también deberán integrarse a la misma ***datos e investigaciones biológicas y bioinformáticas***, así como datos y aplicaciones de ***gobierno electrónico***, preservando en todos los casos la privacidad y seguridad de los mismos.

Como parte de la misma visión y estrategia, con el fin de complementar los objetivos principales, reforzar y acelerar el desarrollo, la obtención de valor y la visibilidad, en forma concomitante a la construcción de la plataforma se impulsará el ***desarrollo institucional, de infraestructura, de instancias de formación*** y la ejecución de otros proyectos que podrán o no integrarse a la plataforma.

Esquema Organizacional de AGranDAr

Estratégico:
Funciones: Orquestación. Dirección estratégica. Prospectivas. Soporte Institucional
Integración: Comité de expertos y asesores, integrado por profesionales y académicos del país o del exterior, representantes del MINCYT y de los principales organismos gubernamentales involucrados (ArSAT, MinPlan, IGN, etc.). Coordinación: Fundación Sadosky

Gestión:
Funciones: Dirección de proyectos. Construcción de la comunidad.
Integración: Directores de proyectos, investigadores destacados, empresarios y funcionarios pioneros. Fundación Sadosky como "oficina de proyectos"

Ejecución:
Funciones: Ejecución de proyectos. Formación y capacitación. Difusión
Integrantes: Investigadores, empresarios, profesionales, estudiantes.

Algunas iniciativas (muy en borrador):

Plataforma de Datos Geolocalizados de la Argentina

Plataforma Nacional de Bioinformática

Creación de un Centro Nacional de Ciencia de los Datos (CNCD)

Apalancar la iniciativa gubernamental del Portal de Datos Públicos

Fomentar la creación de un Cloud Nacional

Creación de un cluster de empresas y laboratorios en Ciencia de Datos

Formación de posgrado en Ciencias de los datos

Escuelas anuales en Ciencias de los datos

Jornadas Argentinas de Big Data