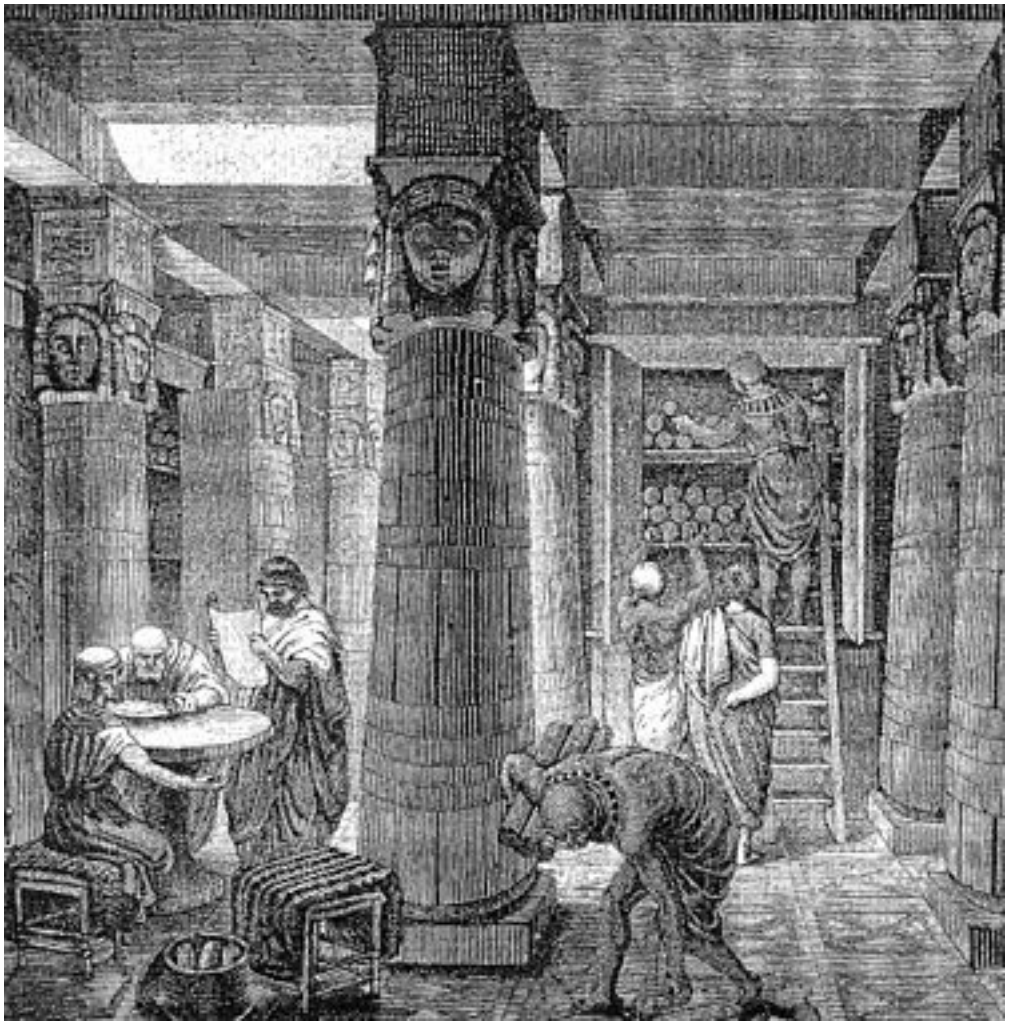
“Conceptos de Big Data”

Rogelio Ferreira Escutia

Evolución del Almacenamiento de la Información

Primera Biblioteca - Alejandría



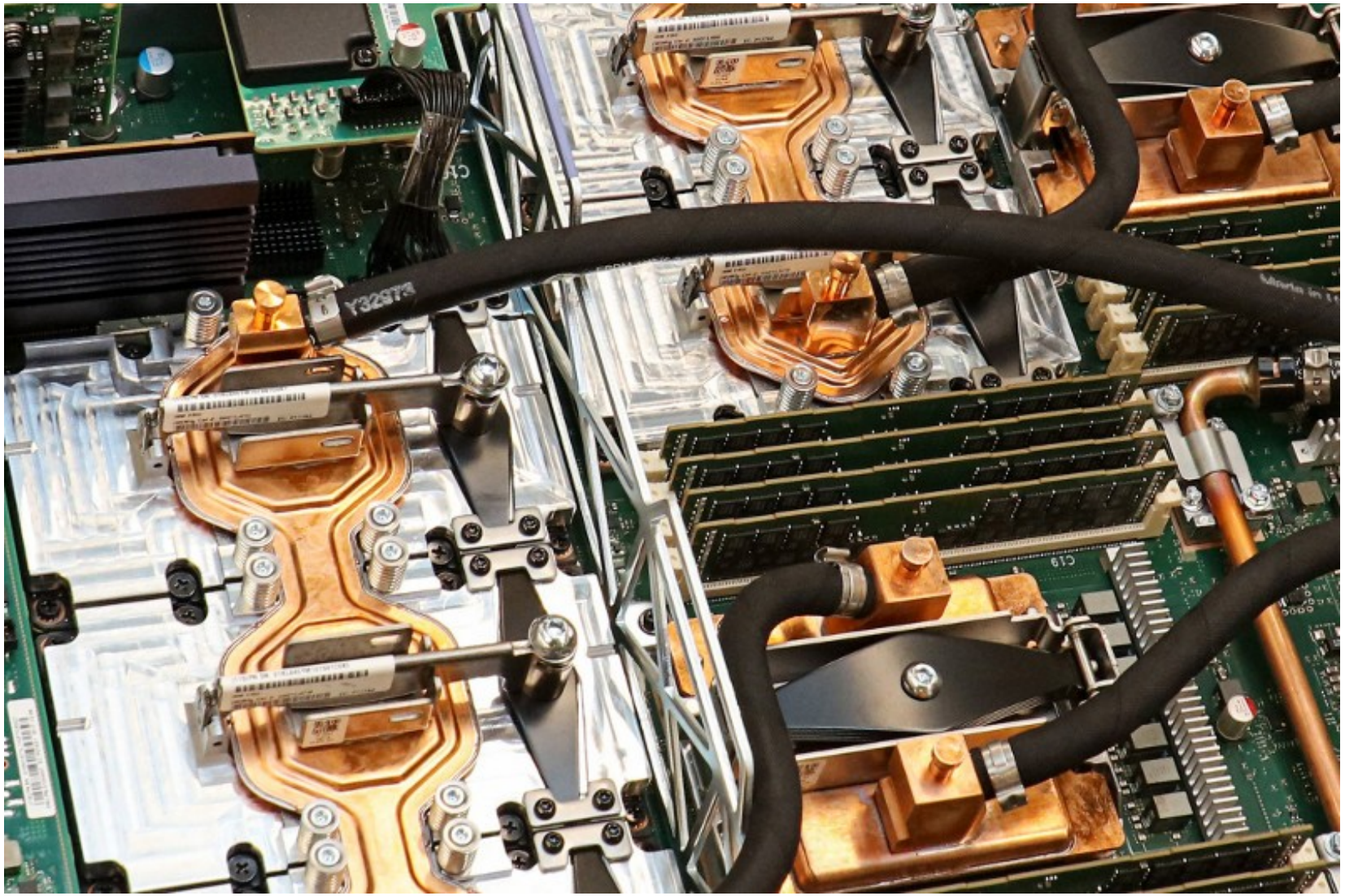
"Biblioteca de Alejandría", https://es.wikipedia.org/wiki/Biblioteca_de_Alejandr%C3%ADa 2015

Castillos – Edad Media

http://musicayocio.com/noticias/fin-de-semana-medieval-en-el-castillo-de-manzanares-el-real/, 2015

Mi Biblioteca – Tec de Morelia

Computadora IBM Summit (USA)
2,414,592 Cores – 148,600 Teraflops



"Top500", https://www.top500.org/lists/2019/06/, junio 2019

Crecimiento de los 2 últimos años

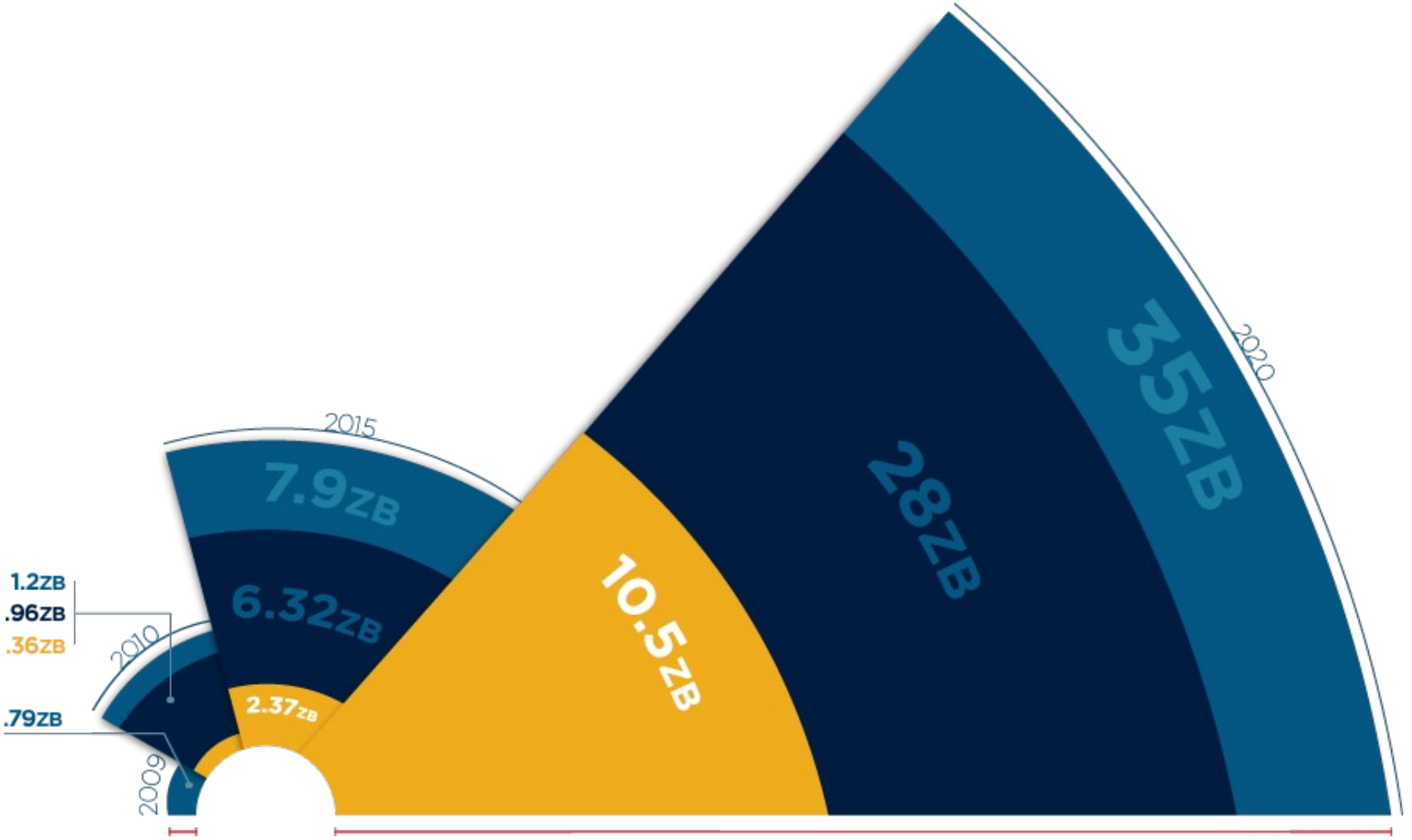
- **“90% de la información existente a lo largo del planeta en toda su historia se ha generado en los últimos dos años”**

Aitor Moren
Responsable de Inteligencia Artificial de Ibermática

“Big Data”, http://www.ibermatica.com/sala-de-prensa/opinion/big-data-el-termino-de-moda-en-el-mundo-de-la-informatica

Crecimiento estimado

- **Se estima un crecimiento del 4300% en la generación de datos anuales para 2020.**

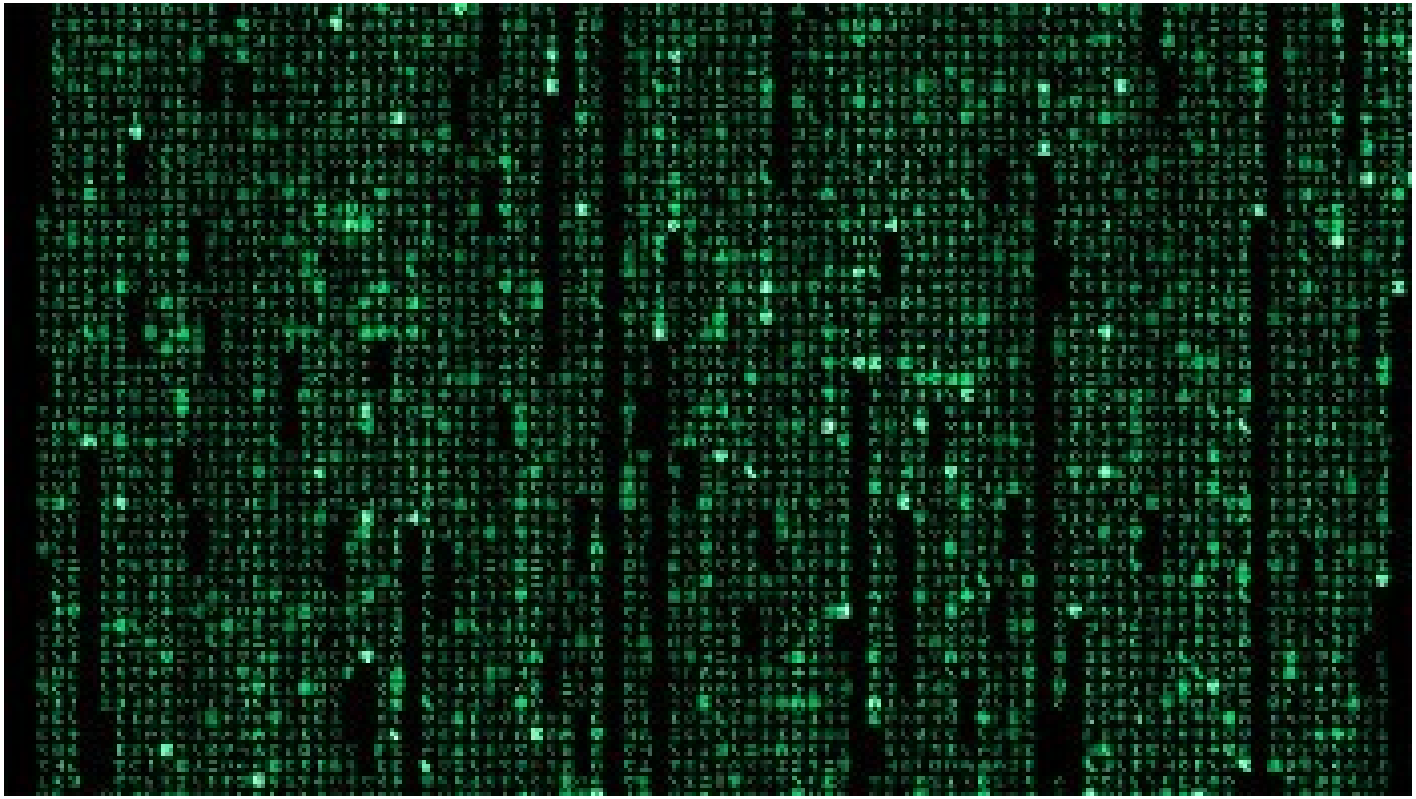


http://www.winshuttle.es/big-data-historia-cronologica/

¿Qué es el Big Data?

¿Qué es Big Data?

- **“Big Data” es una cantidad de datos muy grande, que excede las capacidades convencionales de los sistemas de Base de Datos.**



“Big Data Now”, O'Reilly Media 2012

Características del Big Data

- **Difícil de recolectar (distribuidos en toda la red)**
- **Difícil de almacenar (zetabytes = 1 x 10 E21).**
- **Difícil de analizar (gran cantidad de información).**
- **Difícil de procesar (se buscan patrones)**

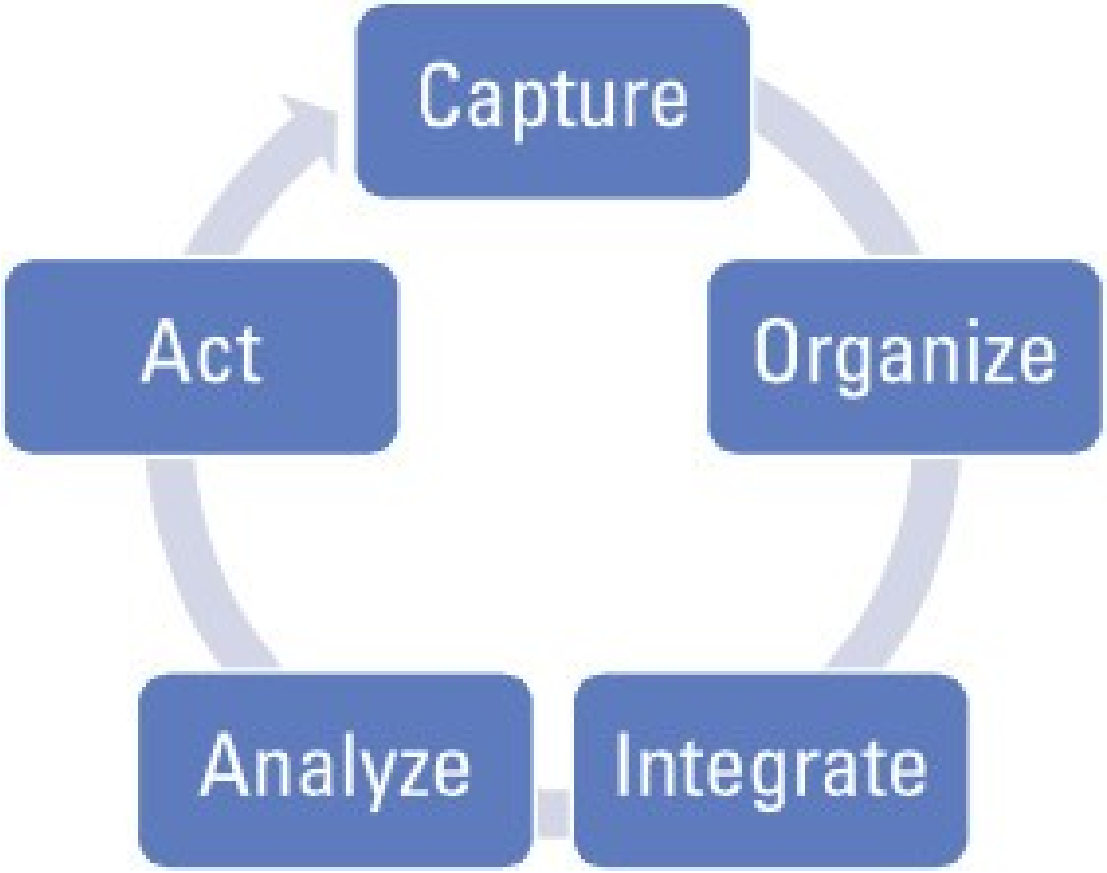
"Big Data Now", O'Reilly Media 2012

Procesamiento del Big Data

DATOS MODELADO SIMULACION OPTIMIZACION INFORMACION

"Big Data Now", O'Reilly Media 2012

Procesamiento del Big Data



"Big Data", Judith Hurwitz2013

Internet y Big Data

- **Cada cosa que se conecta a Internet**
- **Internet está lleno de datos.**
- **Por lo tanto, cada cosa que conectamos es parte del Big Data.**

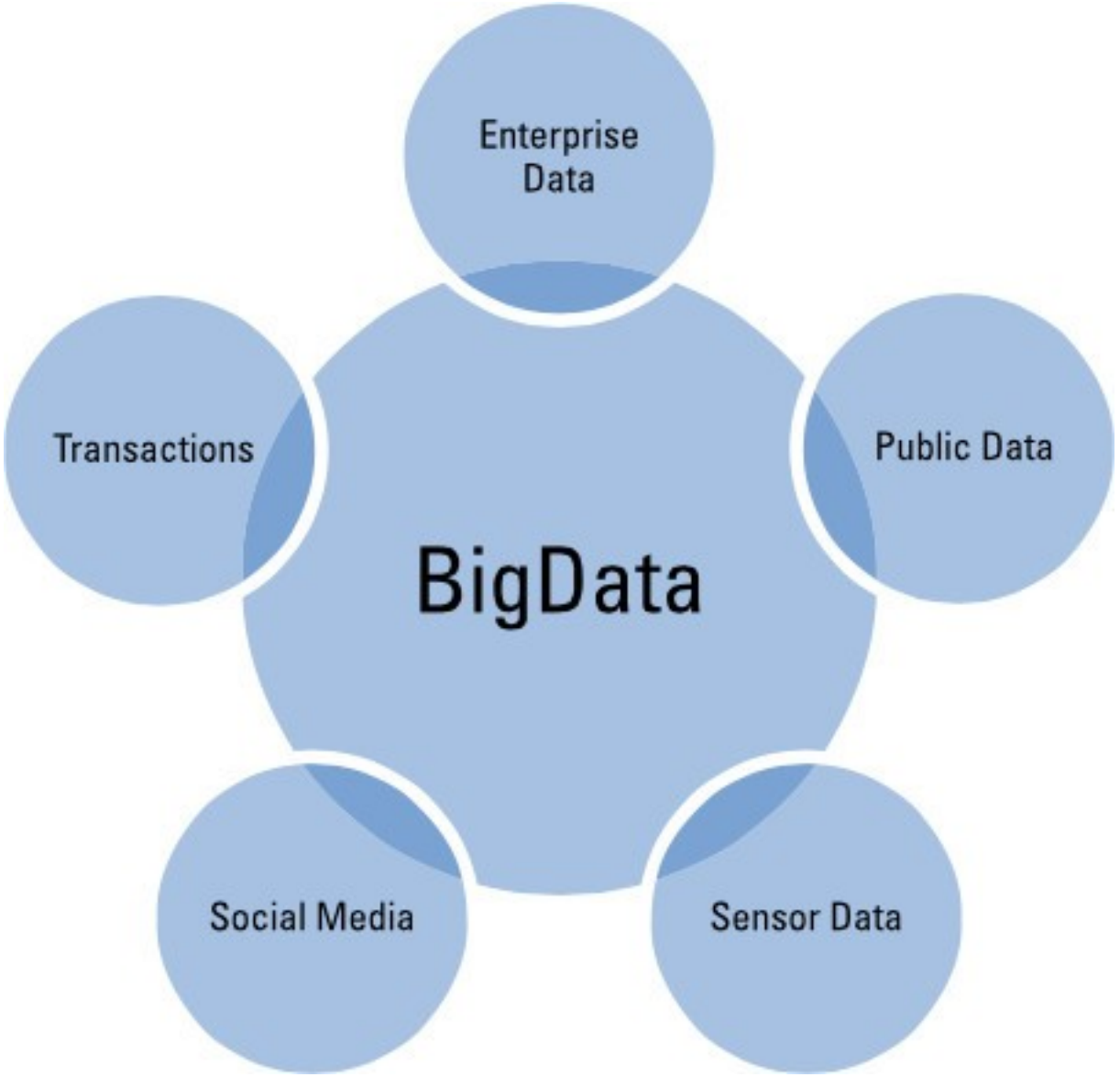
"Big Data Now", O'Reilly Media 2012

Tipos de Datos

- **Estructurados: como las bases de datos (fáciles de recolectar).**
- **Datos sin estructura: Como la mayoría de las páginas Web (complejos de recolectar).**
- **Datos SemiEstructurados: Como los documentos, que llevan un cierto formato (difíciles de recolectar).**

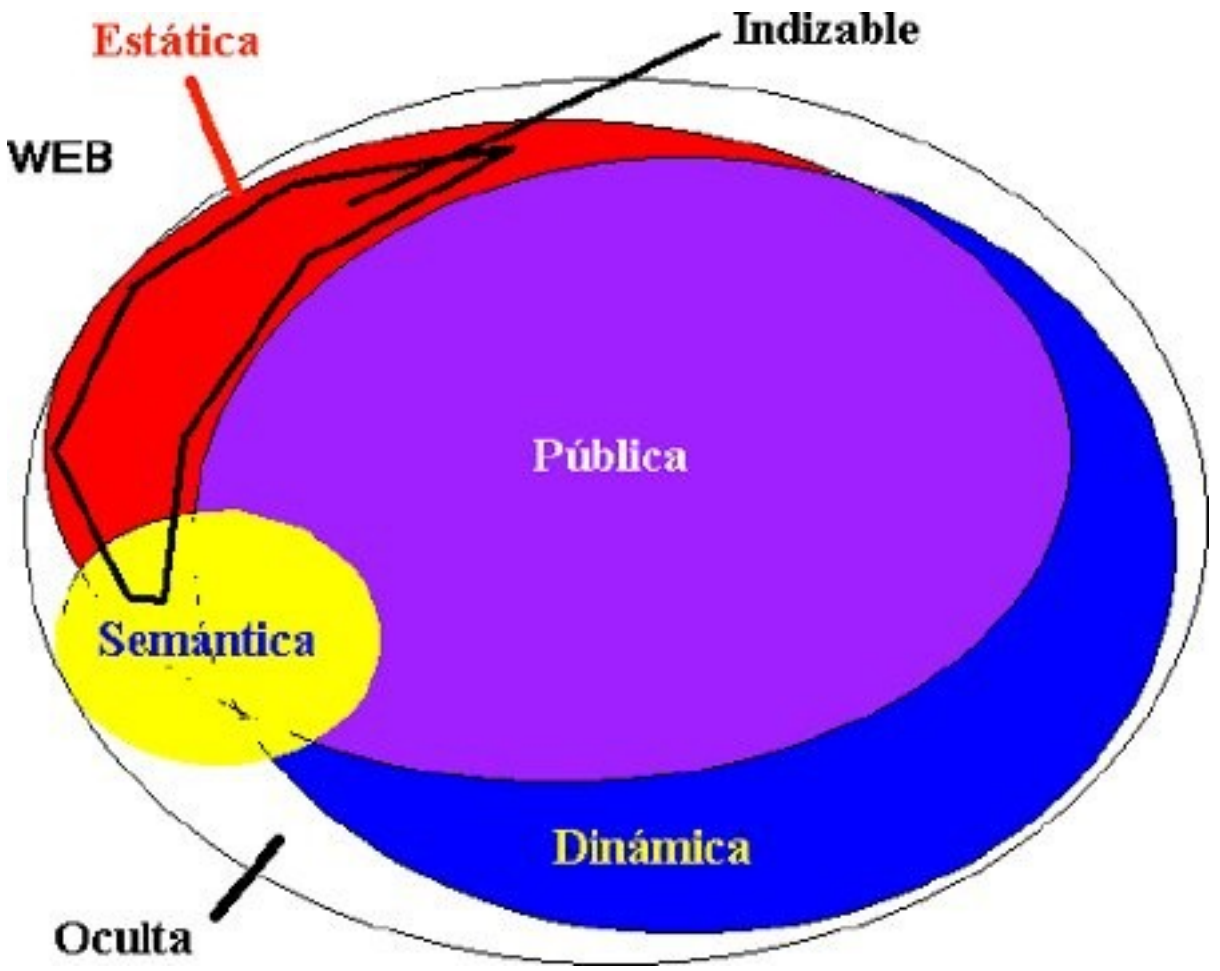
"Data Science", Lillian Pierson, Jhon Wiley & Sons 2015

Origen de los Datos



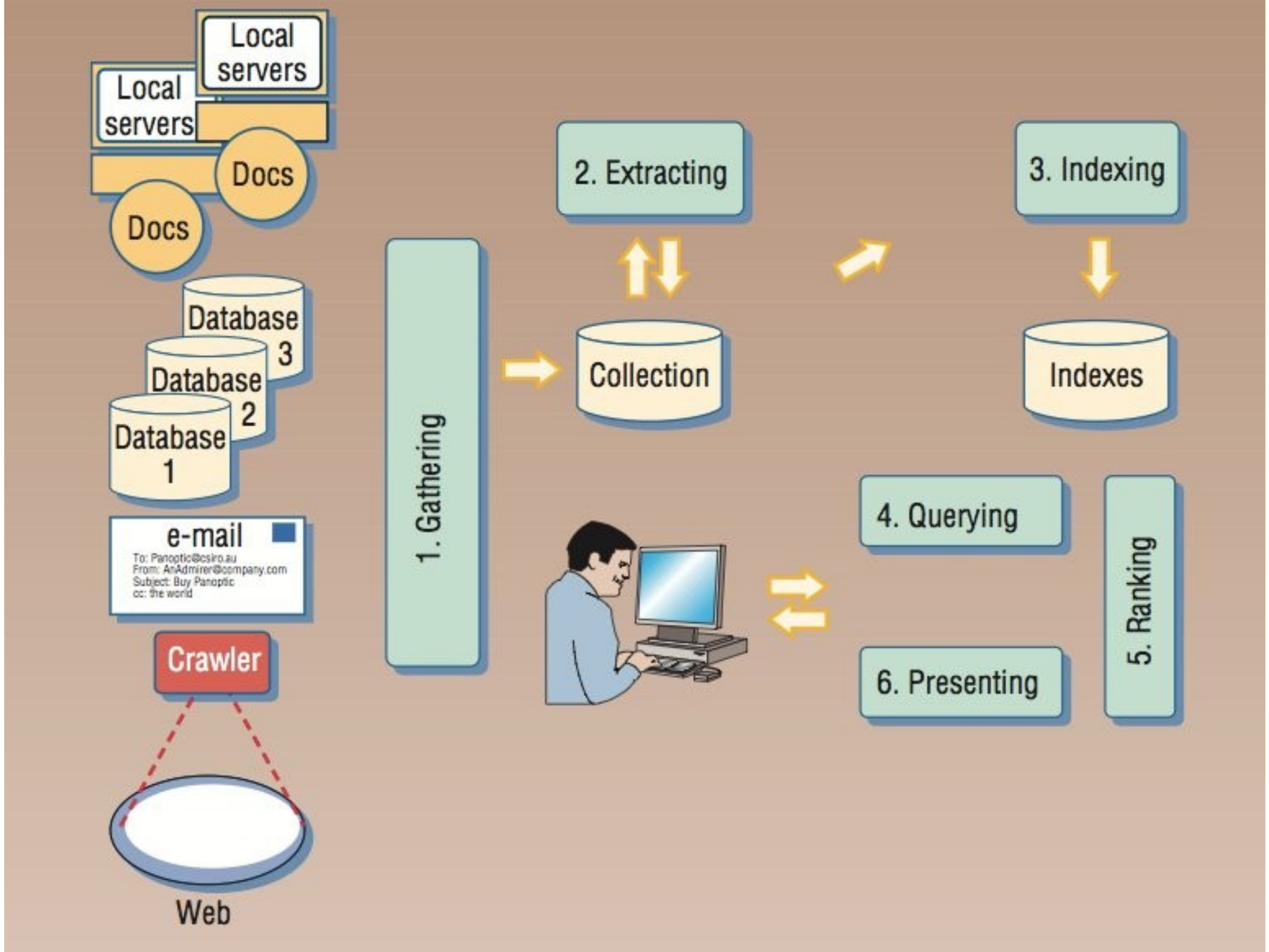
"Data Science", Lillian Pierson, Jhon Wiley & Sons 2015

Anatomía de la Web

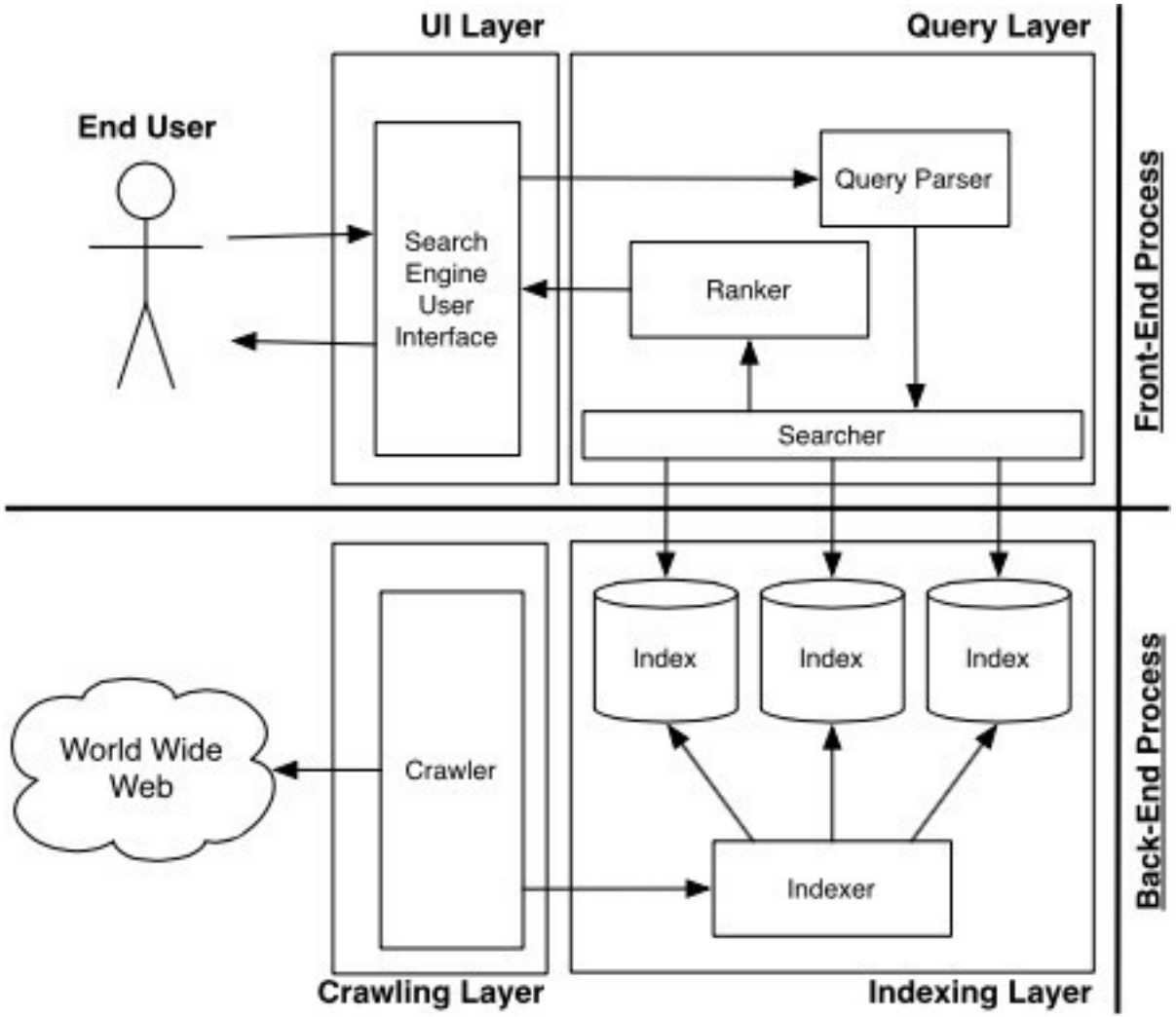


”Cómo funciona La Web”, http://www.ciw.cl/libroweb, 2008

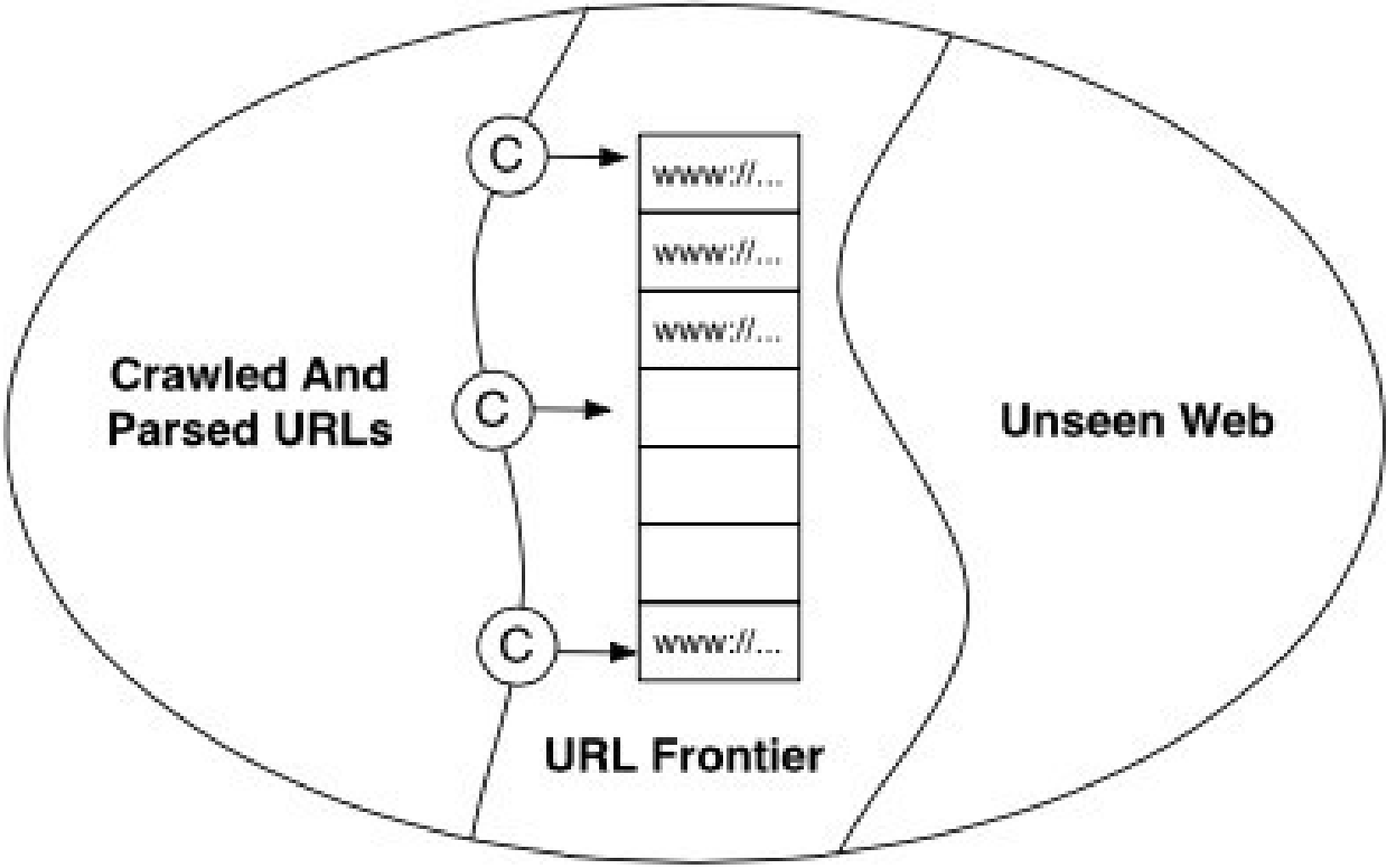
“Data Science”, Lillian Pierson, Jhon Wiley & Sons 2015


Local servers
Local servers
Docs
Docs
Database 3
Database 2
Database 1
e-mail
To: Panoptic@csiro.au
From: AnAdmirer@company.com
Subject: Buy Panoptic
cc: the world
Crawler
Web
1. Gathering
2. Extracting
Collection
3. Indexing
Indexes
4. Querying
6. Presenting
5. Ranking

Motor de búsqueda completo



La Deep Web


Crawled And
Parsed URLs
C
C
C
www://...
www://...
www://...
www://...
Unseen Web
URL Frontier

Búsqueda de Palabras

Word	Document ID
once	1
upon	1
a	1
time	1
there	1
lived	1
a	1
beautiful	1
princess	1
the	1
witch	1
cast	1
a	1
terrible	1
spell	1
on	1
the	1
princess	1
the	2
witch	2
hunted	2
the	2
dragon	2
down	2
The	2
dragon	2
fought	2
back	2
but	2
the	2
witch	2
was	2
stronger	2

(a) map

Word	Document ID
a	1
a	1
a	1
back	2
beautiful	1
but	2
cast	1
down	2
dragon	2
dragon	2
fought	2
hunted	2
lived	1
on	1
once	1
princess	1
princess	1
spell	1
stronger	2
terrible	1
the	1
the	1
the	2
the	2
the	2
the	2
time	1
there	1
upon	1
was	2
witch	1
witch	2
witch	2

(b) sort

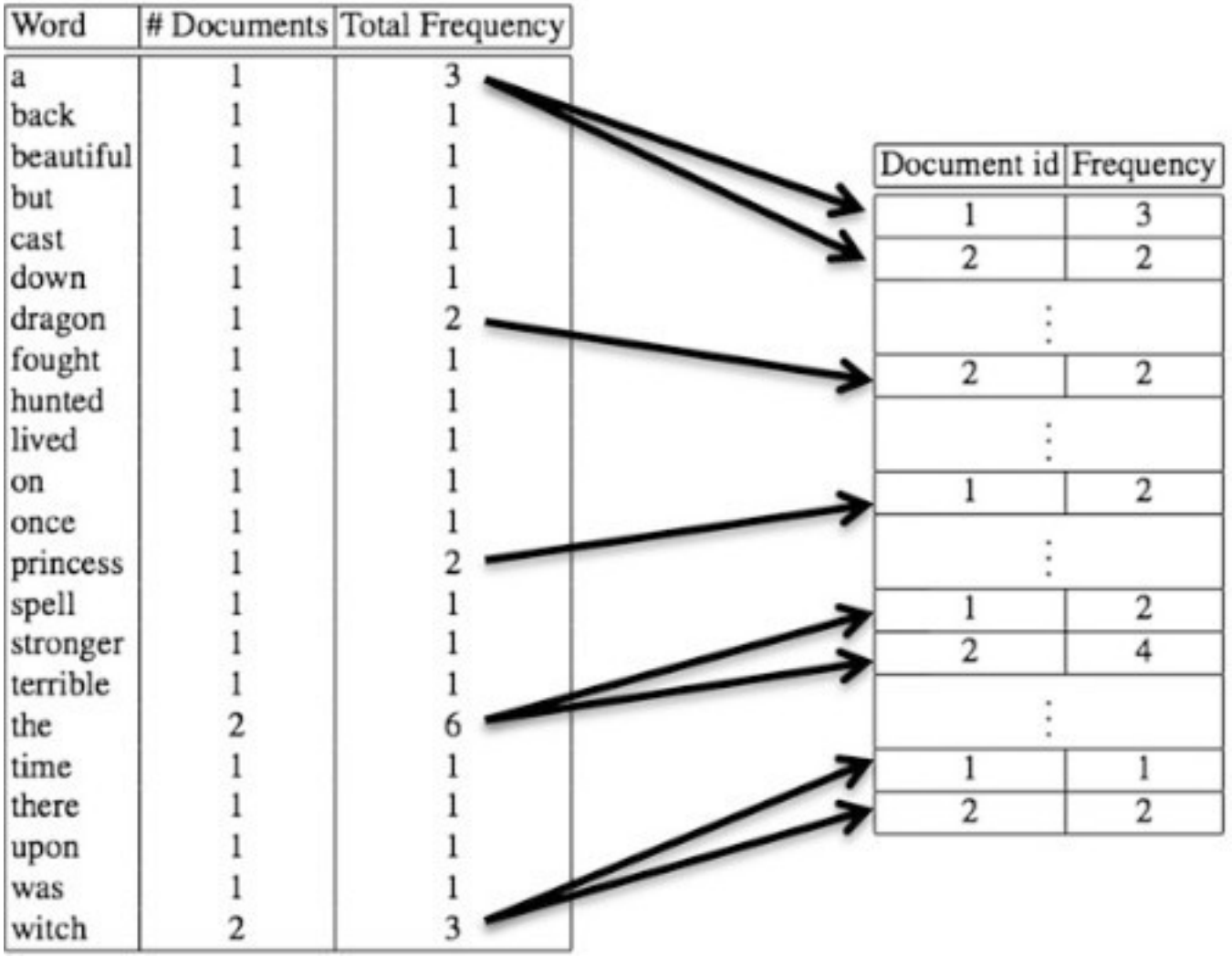
Word	Document ID	Frequency
a	1	3
back	2	1
beautiful	1	1
but	2	1
cast	1	1
down	2	1
dragon	2	2
fought	2	1
hunted	2	1
lived	1	1
on	1	1
once	1	1
princess	1	2
spell	1	1
stronger	2	1
terrible	1	1
the	1	2
the	2	4
time	1	1
there	1	1
upon	1	1
was	2	1
witch	1	1
witch	2	2

(c) merge

Indexado del contexto

Document ID	sentence ID	text
1	1	Once upon a time there lived a beautiful princess
		⋮
1	19	The witch cast a terrible spell on the princess
2	34	The witch hunted the dragon down
		⋮
2	39	The dragon fought back but the witch was stronger

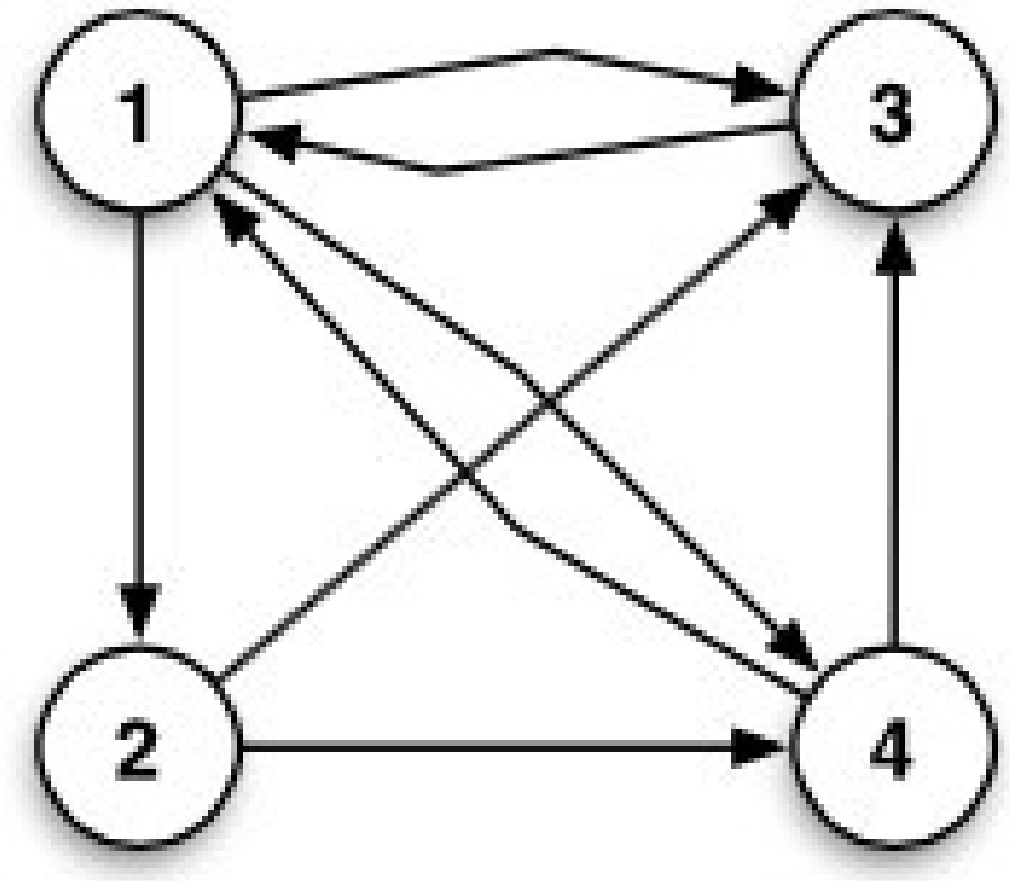
Ordenamiento y Ranking



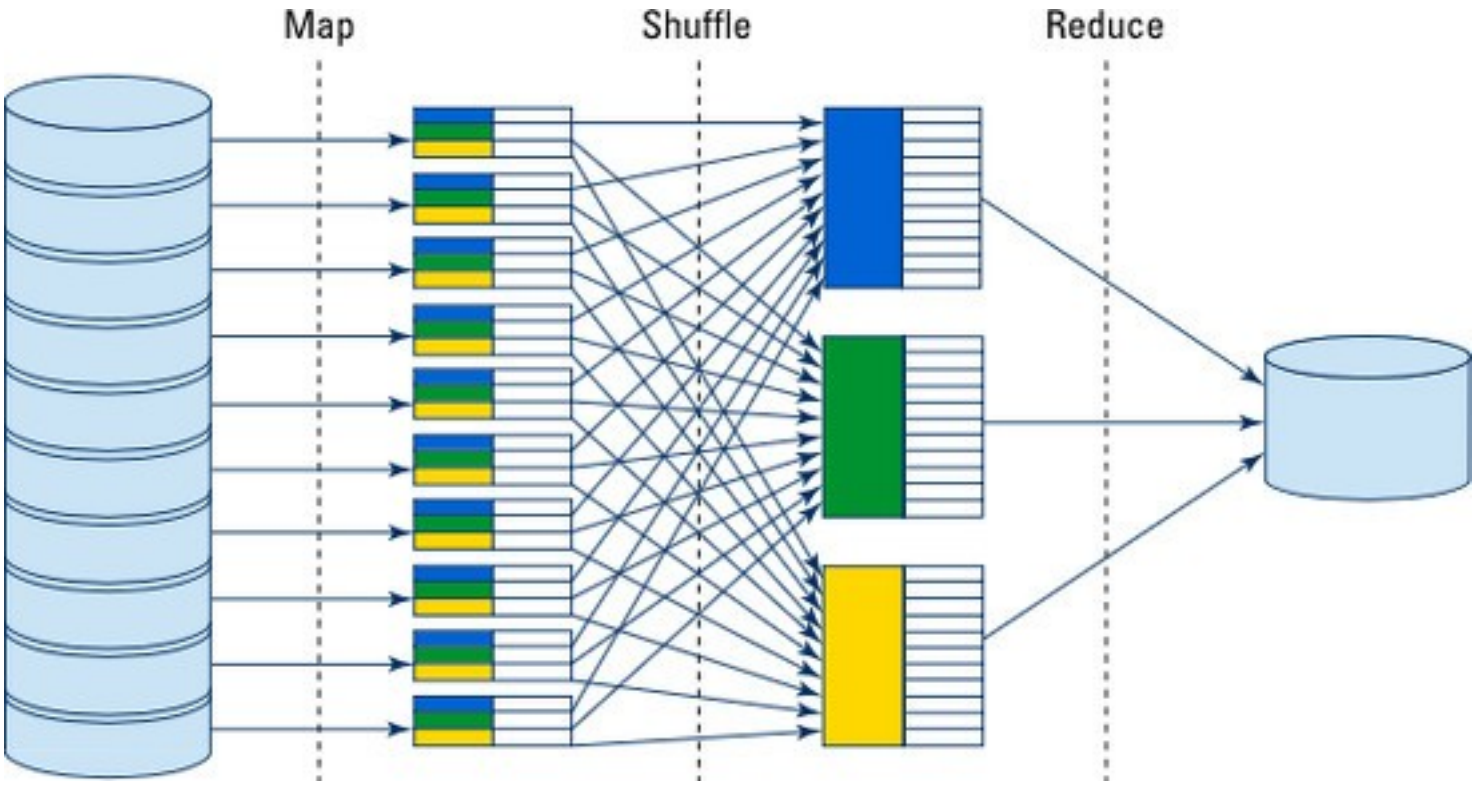
Word	# Documents	Total Frequency
a	1	3
back	1	1
beautiful	1	1
but	1	1
cast	1	1
down	1	1
dragon	1	2
fought	1	1
hunted	1	1
lived	1	1
on	1	1
once	1	1
princess	1	2
spell	1	1
stronger	1	1
terrible	1	1
the	2	6
time	1	1
there	1	1
upon	1	1
was	1	1
witch	2	3

Document id	Frequency
1	3
2	2
⋮	
2	2
⋮	
1	2
⋮	
1	2
2	4
⋮	
1	1
2	2

Page Rank de Google

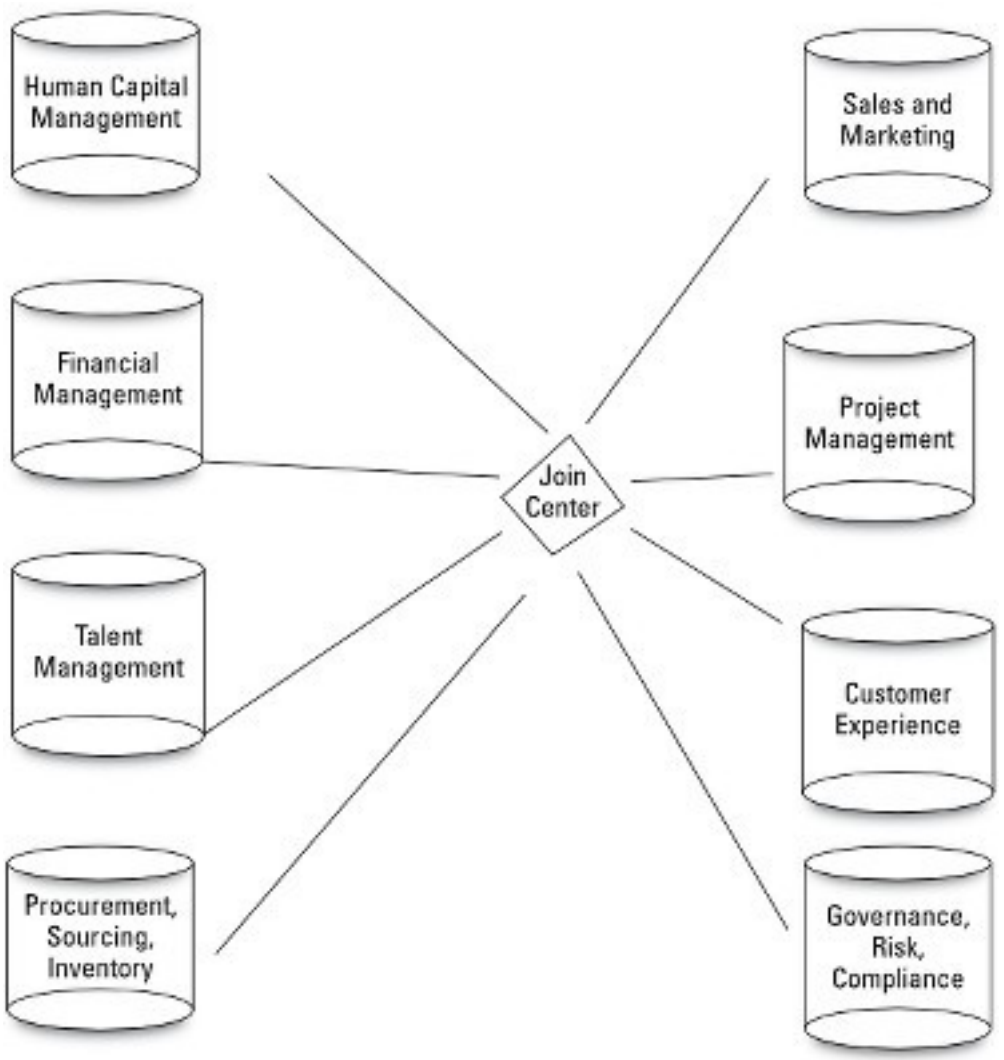


Reducción de los Datos



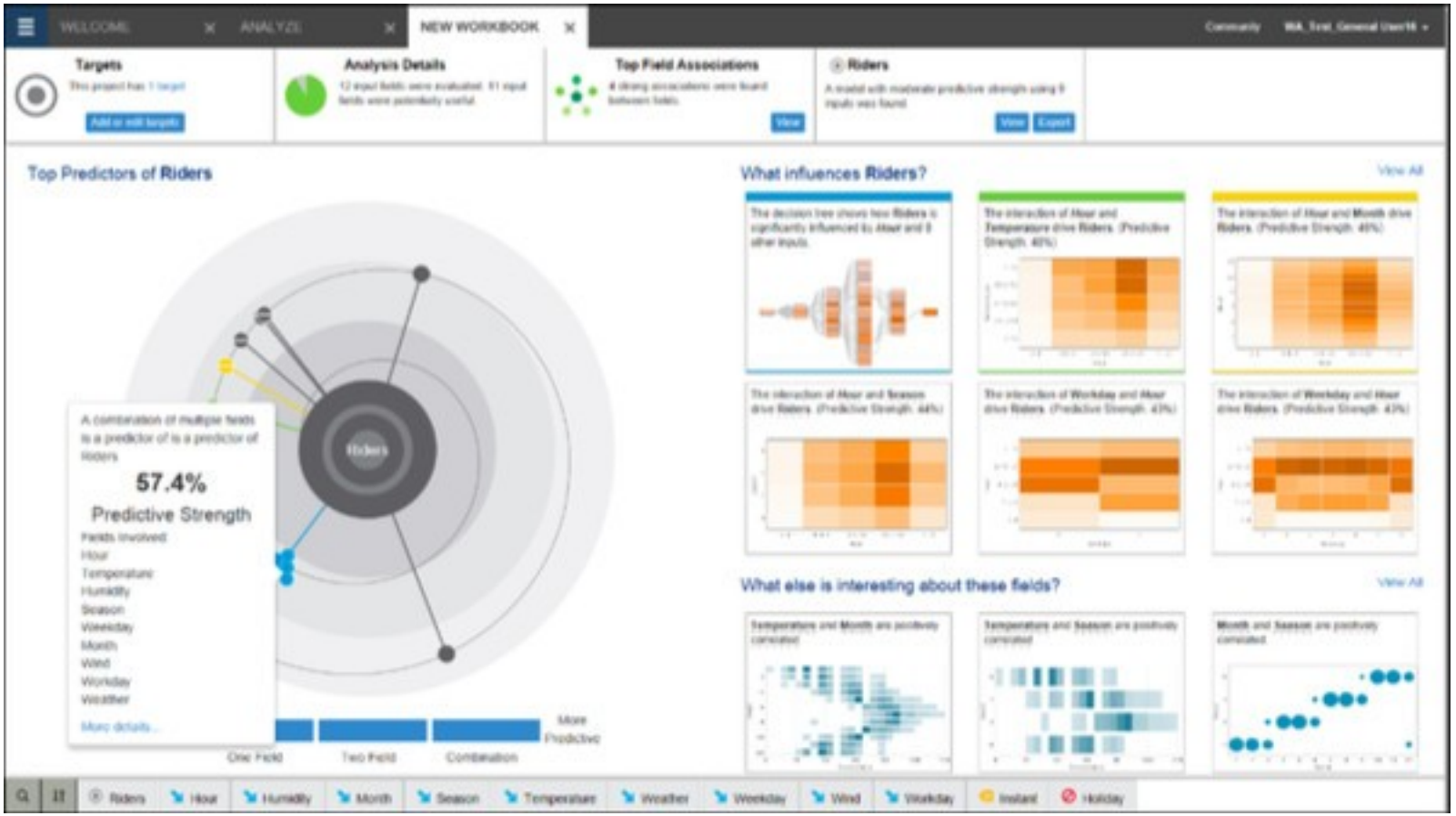
"Data Science", Lillian Pierson, Jhon Wiley & Sons 2015

Fusión e Interpretación de los Datos

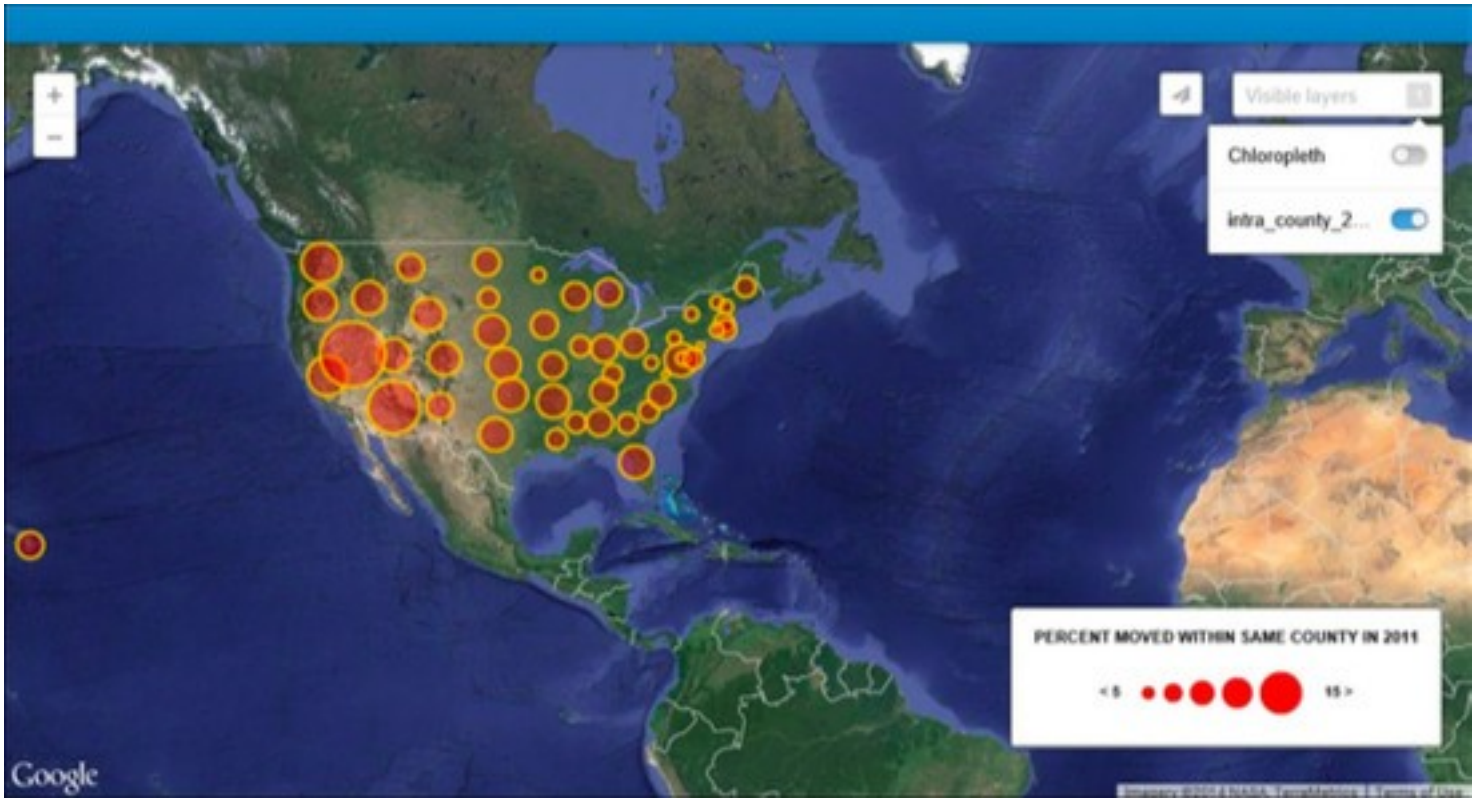


"Data Science", Lillian Pierson, Jhon Wiley & Sons 2015

Visualización de los Datos



Origen de los Datos



La Era del Big Data

Caso K Mart: Detección de Tendencias

- **Se detectó un patrón de compra en la década de los 90s en las tiendas K Mart:**

Hombres casados y con hijos pequeños compraban "cervezas y pañales"

Fuente: Cisco IBSG, abril de 2011

Avión Airbus A380

Avión Airbus A380

- **1000 millones de líneas de código.**
- **Cada motor genera 10 TB cada 30 minutos.**
- **640 TB en un vuelo Londres Nueva York.**

¿Hay oportunidades en al área de IoT y Big Data?

Se requieren mas de 100,000 Ingenieros en los próximos 5 años!

Ya me dió el dolor de caballo…

No vuelvo a cenar enchiladas…

Si hubiera entrenado mas…

¿cuánto quedó el Morelia?

Ya no voy a ir al antro…

Ahorita le meto un codazo…

Competitividad Mundial

Nunca dejar de Soñar...

Tu puedes ser el mejor!!!

Rogelio Ferreira Escutia

Instituto Tecnológico de Morelia
Departamento de Sistemas y Computación

Correo: ***rogelio@itmorelia.edu.mx***
rogeplus@gmail.com

Página Web: ***http://sagitario.itmorelia.edu.mx/~rogelio/***
http://www.xumarhu.net/

Twitter: ***http://twitter.com/rogeplus***
Facebook: ***http://www.facebook.com/groups/xumarhu.net/***