Actuarial Science

vs

Data Science

Report from the Working Party on Data & Technology

CAS Antitrust Statement

The Casualty Actuarial Society is committed to adhering strictly to the letter and spirit of the antitrust laws. Seminars conducted under the auspices of the CAS are designed solely to provide a forum for the expression of various points of view on topics described in the programs or agendas for such meetings.

Under no circumstances shall CAS seminars be used as a means for competing companies or firms to reach any understanding – expressed or implied – that restricts competition or in any way impairs the ability of members to exercise independent business judgment regarding matters affecting competition.

It is the responsibility of all seminar participants to be aware of antitrust regulations, to prevent any written or verbal discussions that appear to violate these laws, and to adhere in every respect to the CAS antitrust compliance policy.

Data & Technology Working Party

…research and define the knowledge and skills required for actuaries to *successfully partner* with IT to participate in the Data and Analytics revolution - including:

- Data Quality
- Databases
- Business Intelligence
- Data Science

Two Levels of Knowledge to Consider

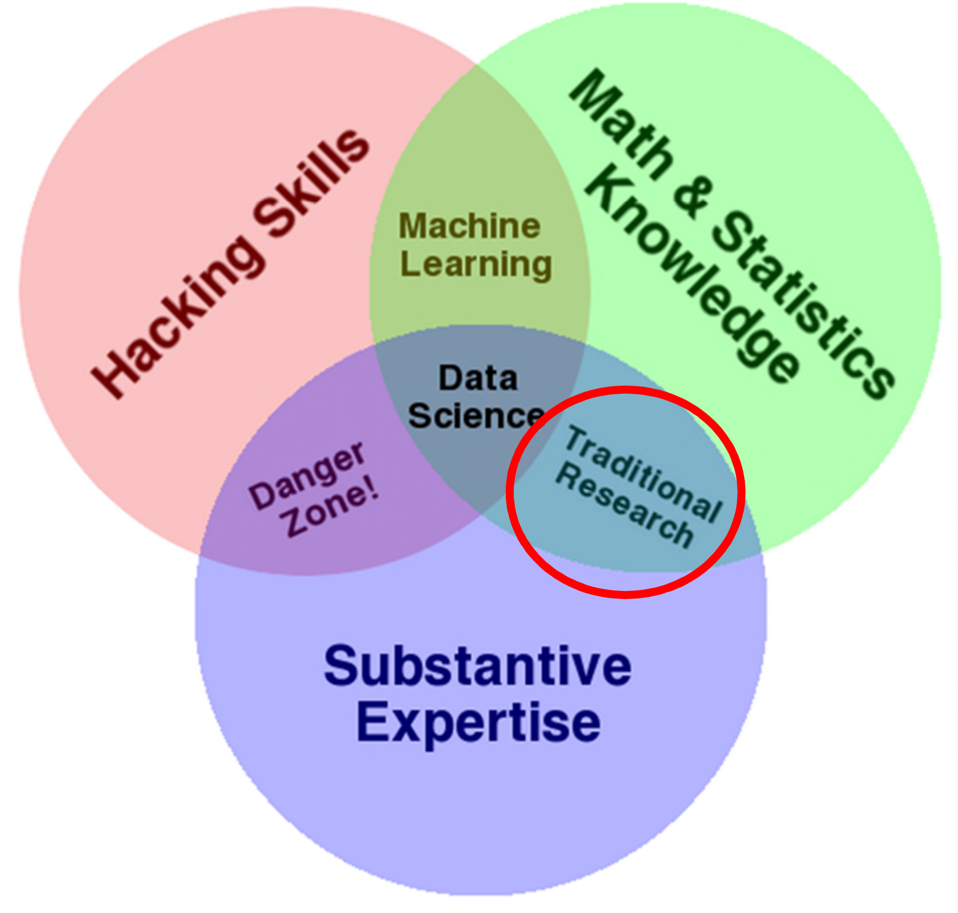
Level 1 - Partner

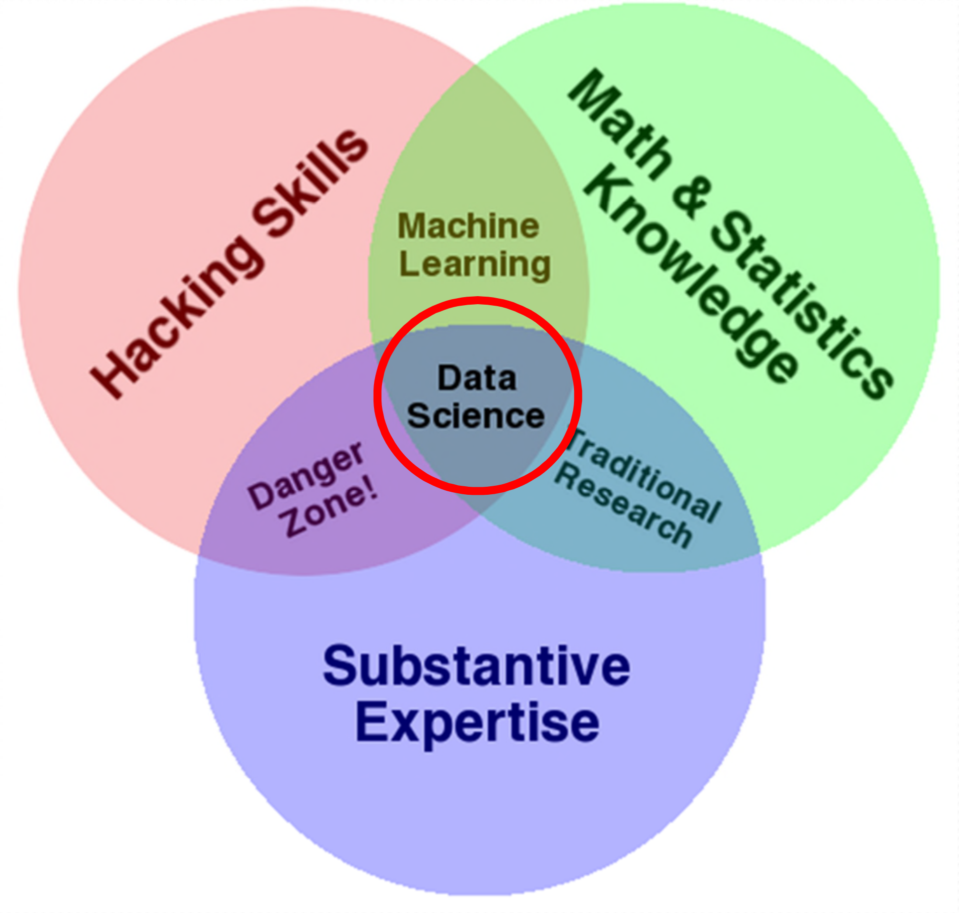
- Knowledgeable
- Conversant
- Reading books & papers
- Data & Tech Working Party papers

Level 2 - Practitioner

- Skilled & Experienced
- Capable
- Courses, degrees & certifications
- iCAS

Actuarial vs Data Science?





http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram

The D&T Working Party Papers

Data Quality

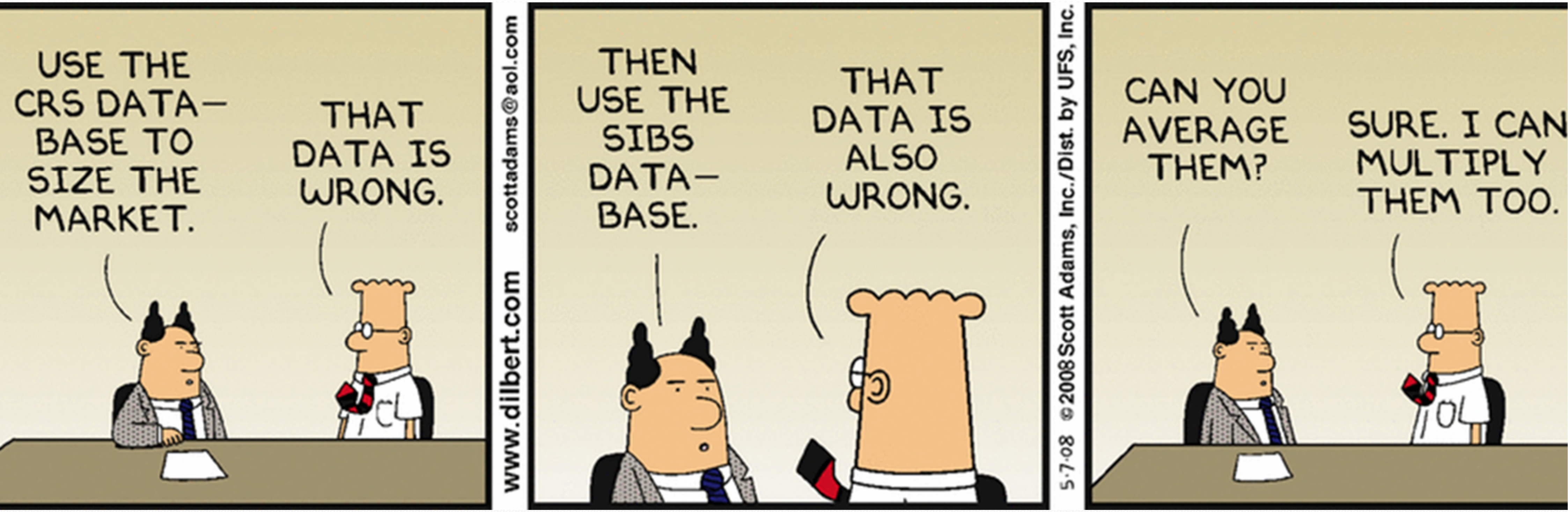
Databases

Business Intelligence

Data Science

Data Quality Principles

Data quality – Is it good enough for analysis


USE THE CRS DATA-BASE TO SIZE THE MARKET.
THAT DATA IS WRONG.
scottadams@aol.com
www.dilbert.com
THEN USE THE SIBS DATA-BASE.
THAT DATA IS ALSO WRONG.
5·7·08 © 2008 Scott Adams, Inc./Dist. by UFS, Inc.
CAN YOU AVERAGE THEM?
SURE. I CAN MULTIPLY THEM TOO.

Data Quality Terms You Should Know

Data Governance – defining and enforcing data quality policies

Data Stewardship – Ownership and accountability for data quality

Metadata – documentation that helps both IT and the data analyst

Lineage – Where the data originates and what happens along the way

Valid Values – What the data should contain

Profiles – What the data does contain

Master Data Management (MDM) – the process of reconciling critical data that is shared across the organization (e.g. customer)

How it is transformed from system to system across the enterprise?

What data exists in my enterprise?

How does one data element relate to another?

The answer? Data Governance

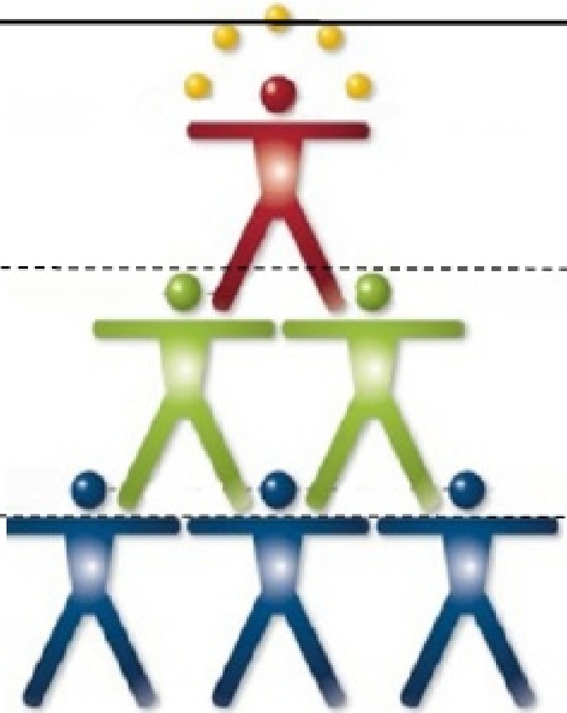
Where does it come from?

How should business terms and data elements be defined?

Can I trust my data?

Who understands what our data means?

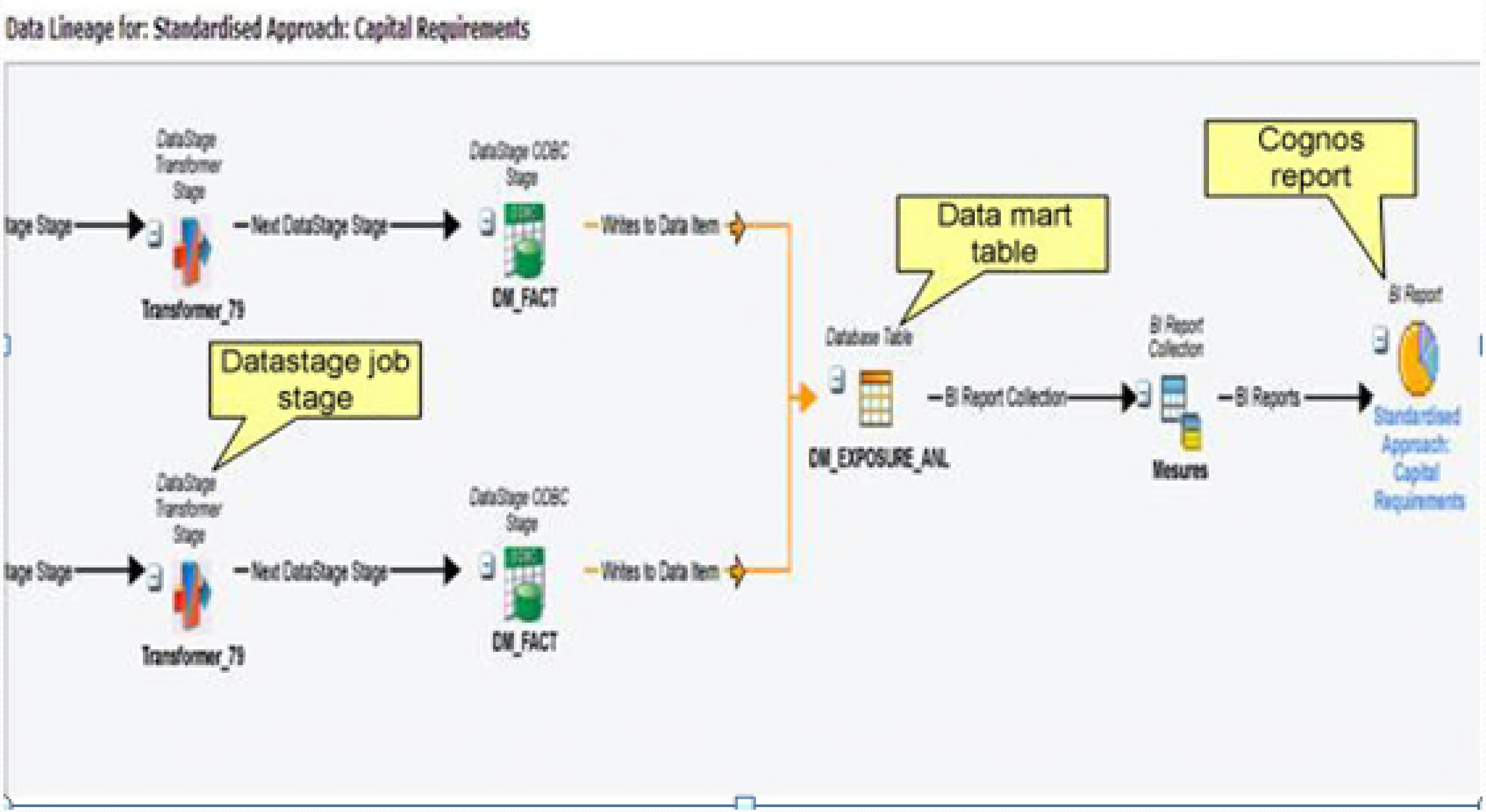
What does Data Stewardship look like?



Level of Authority	Roles & Responsibilities
CIO	- Sponsorship - Executive accountability - Data strategy - Governance
Subject Area Steward	- Enterprise data models - Enterprise Glossary - Subject area accountability - Data reuse
Data Owners	- Subject area data model - Data quality - Data profiling - Data requirements - Metadata management

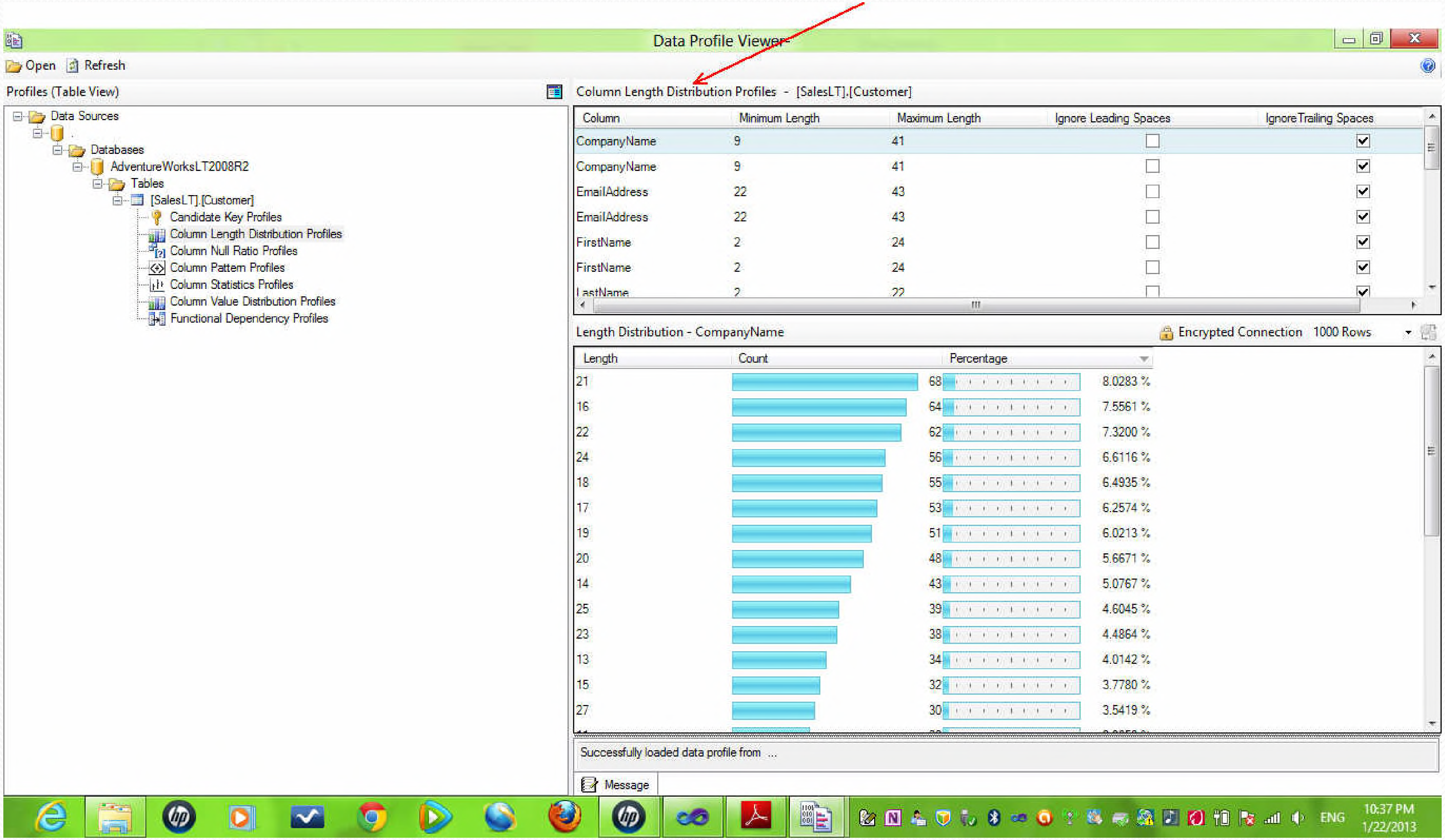
And then I told them –
we're only collecting your metadata

Data Lineage



Data Lineage for: Standardised Approach: Capital Requirements

Data Profiling



Data Profile Viewer

Profiles (Table View)

- Data Sources
  - .
    - Databases
      - AdventureWorksLT2008R2
        - Tables
          - [SalesLT].[Customer]
            - Candidate Key Profiles
            - Column Length Distribution Profiles
            - Column Null Ratio Profiles
            - Column Pattern Profiles
            - Column Statistics Profiles
            - Column Value Distribution Profiles
            - Functional Dependency Profiles

Column Length Distribution Profiles - [SalesLT].[Customer]

Column	Minimum Length	Maximum Length	Ignore Leading Spaces	IgnoreTrailing Spaces
CompanyName	9	41	□	☑
CompanyName	9	41	□	☑
EmailAddress	22	43	□	☑
EmailAddress	22	43	□	☑
FirstName	2	24	□	☑
FirstName	2	24	□	☑
LastName	2	22	□	☑

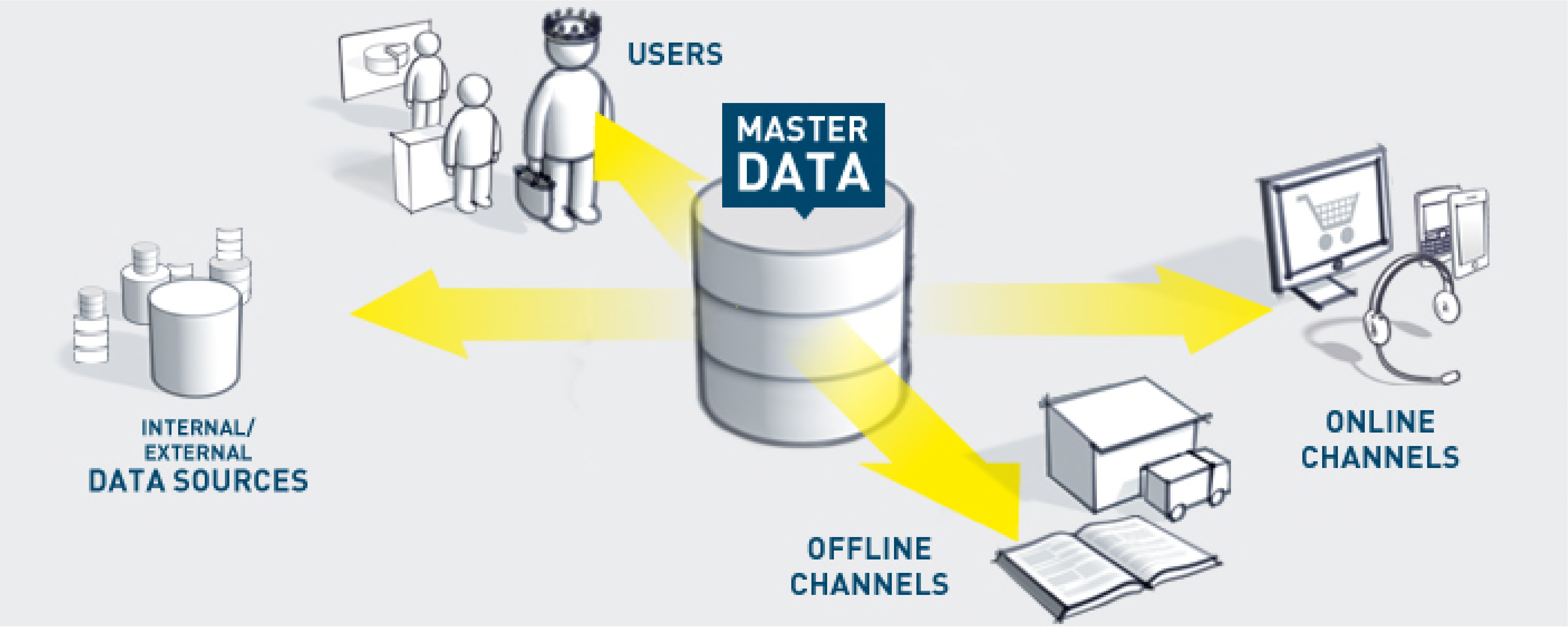
Length Distribution - CompanyName

Encrypted Connection 1000 Rows

Length	Count	Percentage
21	68	8.0283 %
16	64	7.5561 %
22	62	7.3200 %
24	56	6.6116 %
18	55	6.4935 %
17	53	6.2574 %
19	51	6.0213 %
20	48	5.6671 %
14	43	5.0767 %
25	39	4.6045 %
23	38	4.4864 %
13	34	4.0142 %
15	32	3.7780 %
27	30	3.5419 %

Successfully loaded data profile from ...

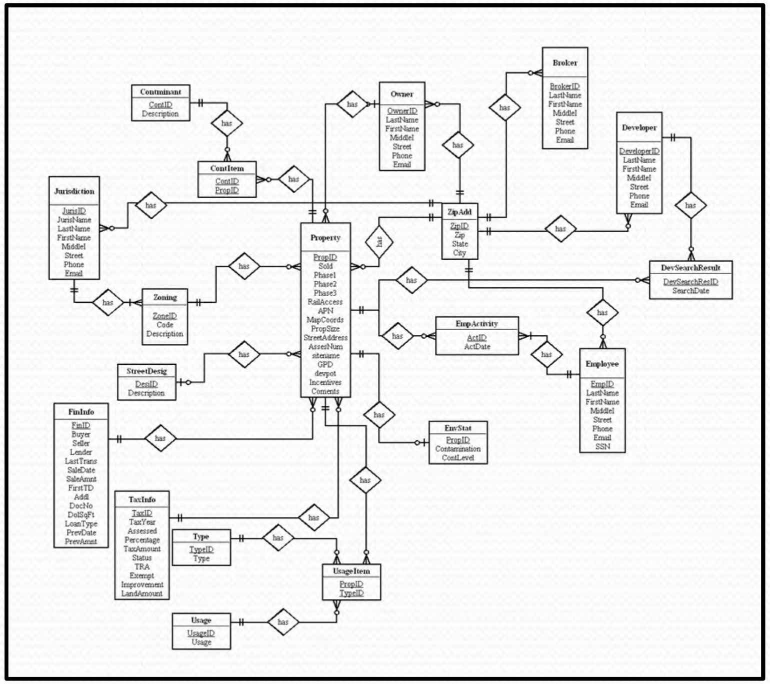
Master Data Management



Databases

Want data?

Go get it…



Does an actuary know ***how*** ?

Data scientists understand how the data is stored, updated and transmitted – with this knowledge they know how and where to hack the data they need

Databases (humor?)

I THINK WE SHOULD BUILD AN SQL DATABASE.
UH-OH
E-mail: SCOTTADAMS@AOL.COM
S.Adams
DOES HE UNDERSTAND WHAT HE SAID OR IS IT SOMETHING HE SAW IN A TRADE MAGAZINE AD?
© 1995 United Feature Syndicate, Inc.(NYC)
11/17
WHAT COLOR DO YOU WANT THAT DATABASE
I THINK MAUVE HAS THE MOST RAM.

Database Terms You Should Know

Application Databases vs Analytic Databases – transactional vs batch

Database Structures – Flat and Wide vs Relational vs Columnar vs Graph vs NoSQL vs Unstructured

Keys – a field used to join data across tables

Fact Tables – the tables that store the metrics of interest to an organization (e.g premium)

Dimension Tables – the tables that describe the context of the facts (e.g. line o business)

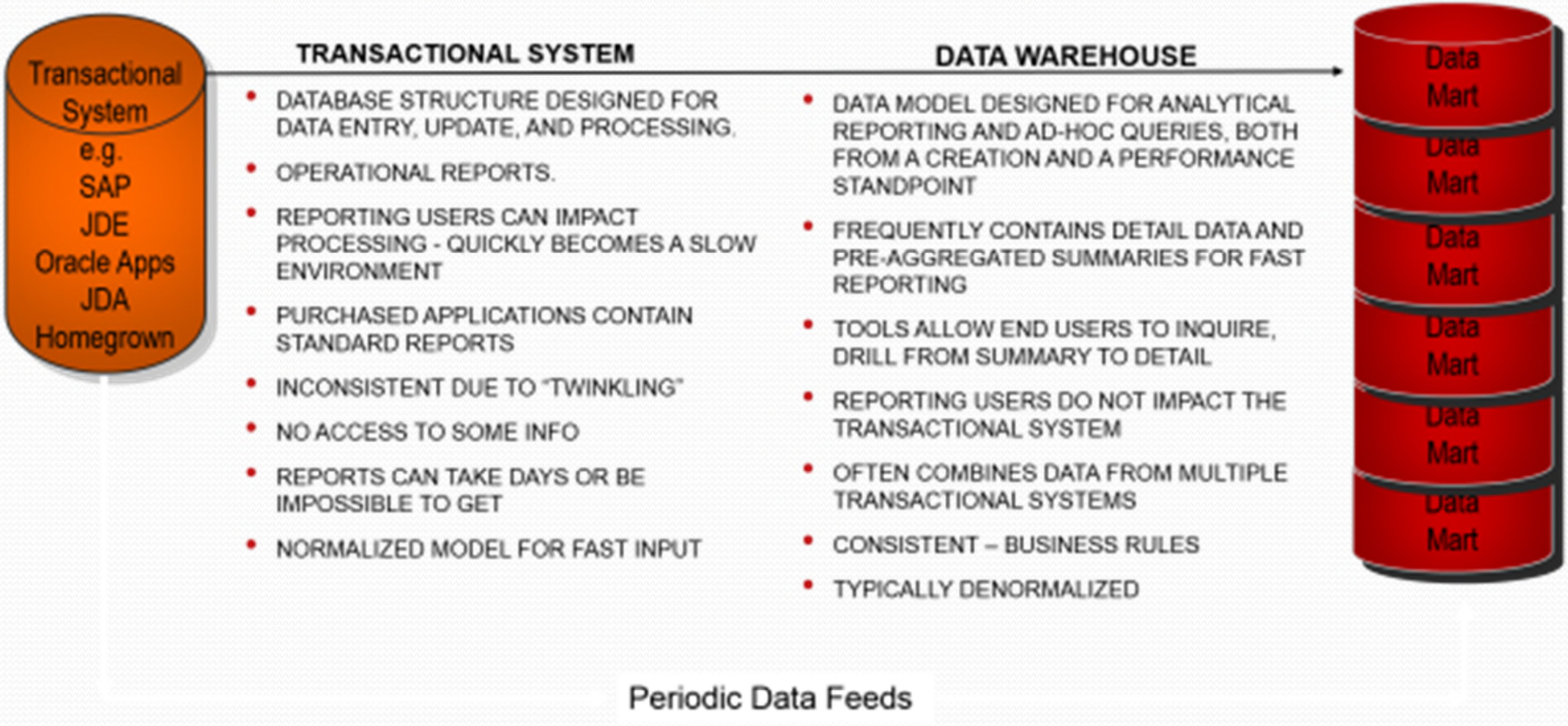
Grain – the level of detail in a database

Structured Query Language (SQL) – The universally accepted standard

Extract, Transform & Load (ETL) – The process of preparing data for analysis

Conformed – consistent fact and dimension definitions across tables

TRANSACTIONAL VS. ANALYTICAL REPORTING



Transactional System
e.g.
SAP
JDE
Oracle Apps
JDA
Homegrown

TRANSACTIONAL SYSTEM

- DATABASE STRUCTURE DESIGNED FOR DATA ENTRY, UPDATE, AND PROCESSING.
- OPERATIONAL REPORTS.
- REPORTING USERS CAN IMPACT PROCESSING - QUICKLY BECOMES A SLOW ENVIRONMENT
- PURCHASED APPLICATIONS CONTAIN STANDARD REPORTS
- INCONSISTENT DUE TO "TWINKLING"
- NO ACCESS TO SOME INFO
- REPORTS CAN TAKE DAYS OR BE IMPOSSIBLE TO GET
- NORMALIZED MODEL FOR FAST INPUT

DATA WAREHOUSE

- DATA MODEL DESIGNED FOR ANALYTICAL REPORTING AND AD-HOC QUERIES, BOTH FROM A CREATION AND A PERFORMANCE STANDPOINT
- FREQUENTLY CONTAINS DETAIL DATA AND PRE-AGGREGATED SUMMARIES FOR FAST REPORTING
- TOOLS ALLOW END USERS TO INQUIRE, DRILL FROM SUMMARY TO DETAIL
- REPORTING USERS DO NOT IMPACT THE TRANSACTIONAL SYSTEM
- OFTEN COMBINES DATA FROM MULTIPLE TRANSACTIONAL SYSTEMS
- CONSISTENT – BUSINESS RULES
- TYPICALLY DENORMALIZED

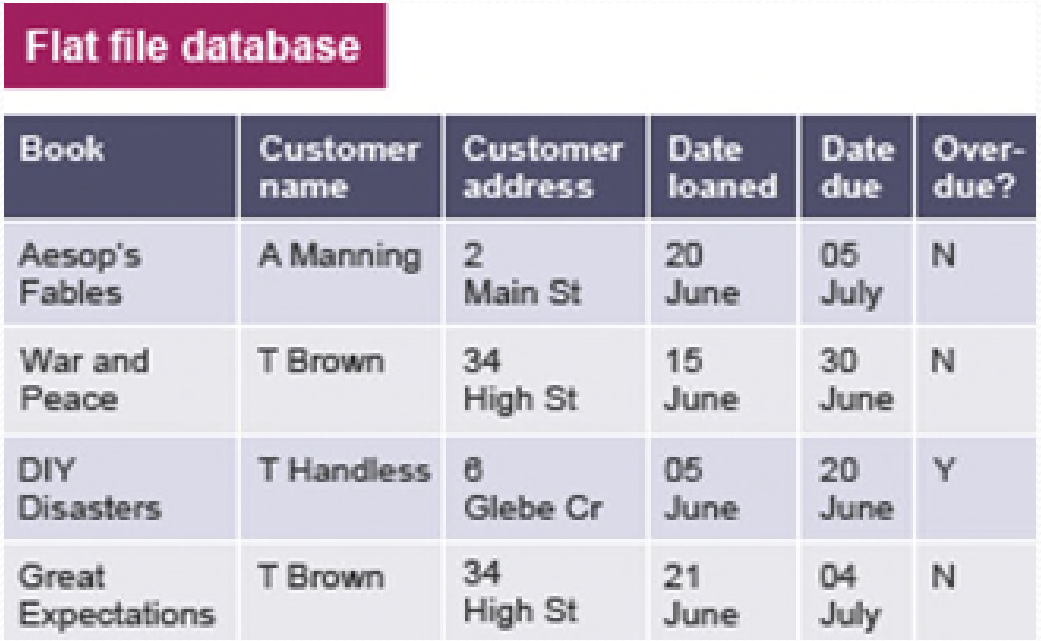
Data Mart
Data Mart
Data Mart
Data Mart
Data Mart
Data Mart

Periodic Data Feeds

Property of Relational Solutions, Inc. By Janet Dorenkott June, 2013,

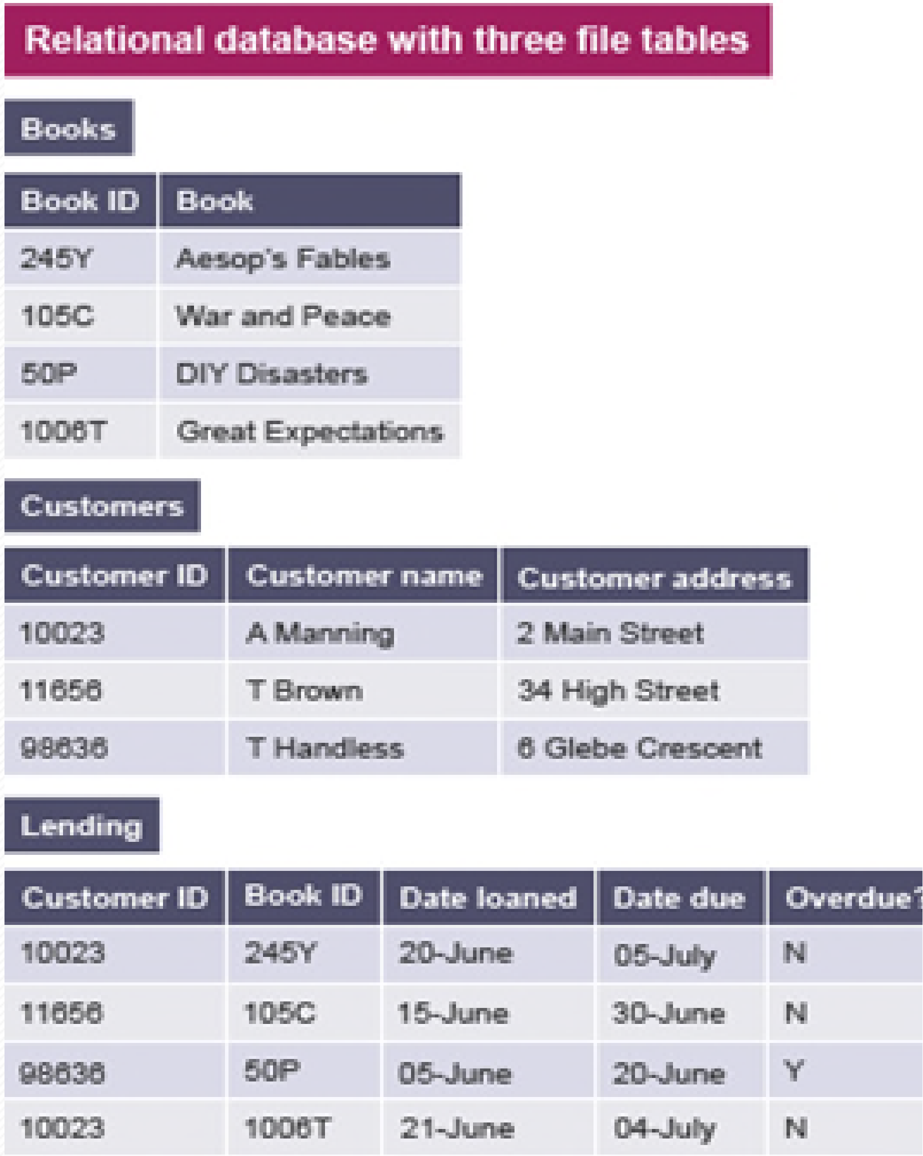
Relational Solutions

Database Types



Flat file database

Book	Customer name	Customer address	Date loaned	Date due	Over-due?
Aesop's Fables	A Manning	2 Main St	20 June	05 July	N
War and Peace	T Brown	34 High St	15 June	30 June	N
DIY Disasters	T Handless	6 Glebe Cr	05 June	20 June	Y
Great Expectations	T Brown	34 High St	21 June	04 July	N



Relational database with three file tables

Books

Book ID	Book
245Y	Aesop's Fables
105C	War and Peace
50P	DIY Disasters
1006T	Great Expectations

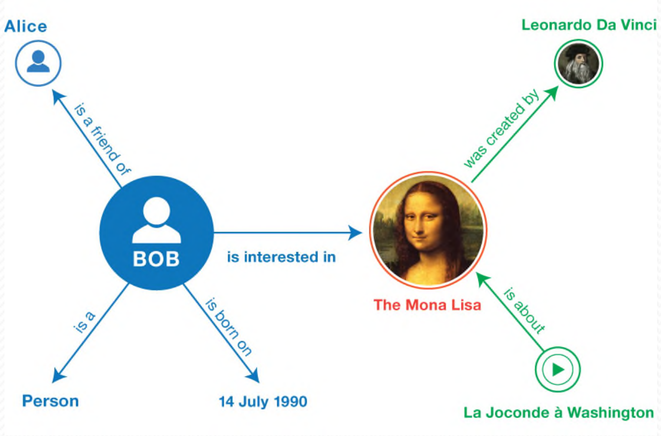
Customers

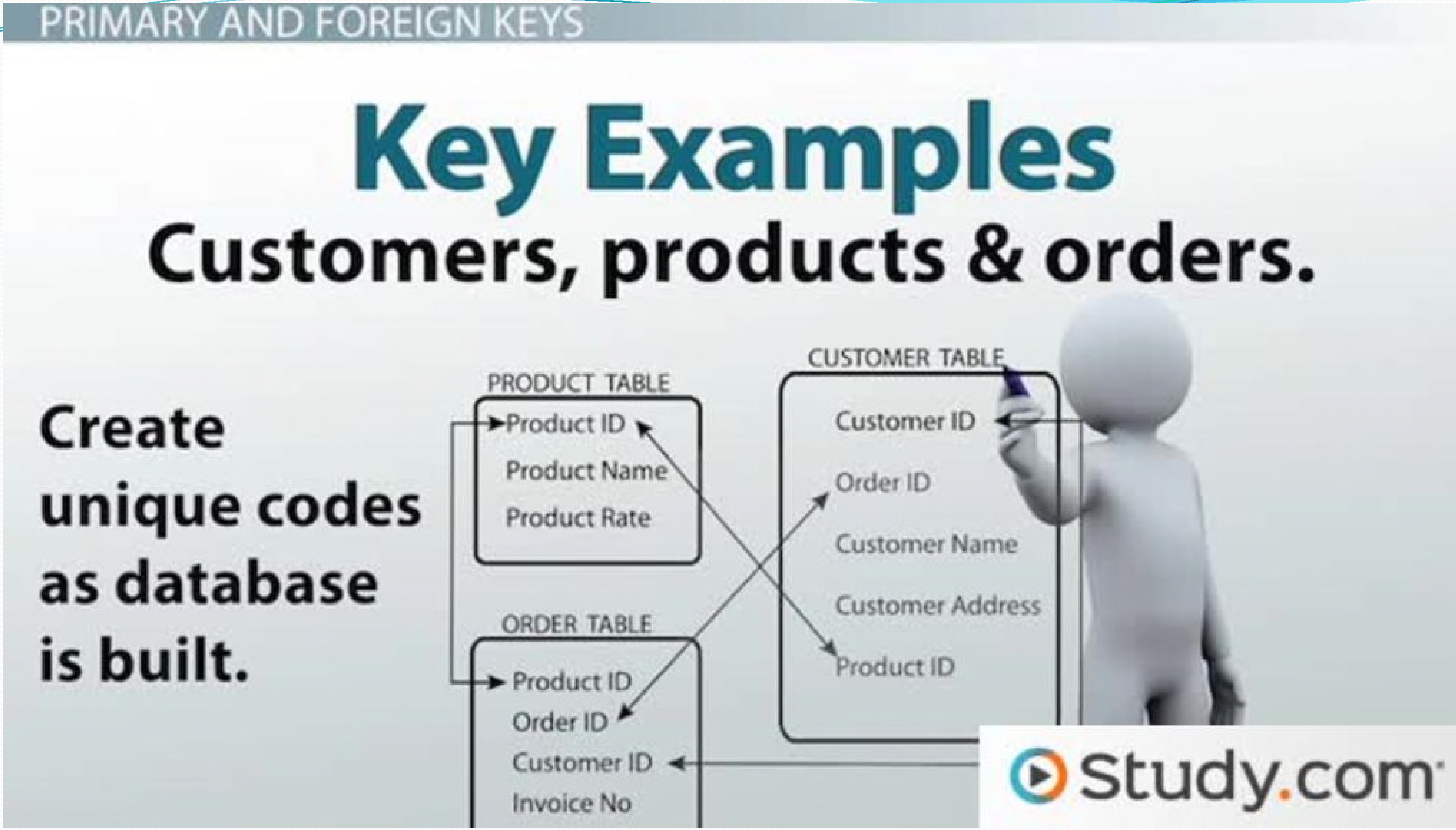
Customer ID	Customer name	Customer address
10023	A Manning	2 Main Street
11656	T Brown	34 High Street
98636	T Handless	6 Glebe Crescent

Lending

Customer ID	Book ID	Date loaned	Date due	Overdue?
10023	245Y	20-June	05-July	N
11656	105C	15-June	30-June	N
98636	50P	05-June	20-June	Y
10023	1006T	21-June	04-July	N

Graph Database





PRIMARY AND FOREIGN KEYS

Key Examples

Customers, products & orders.

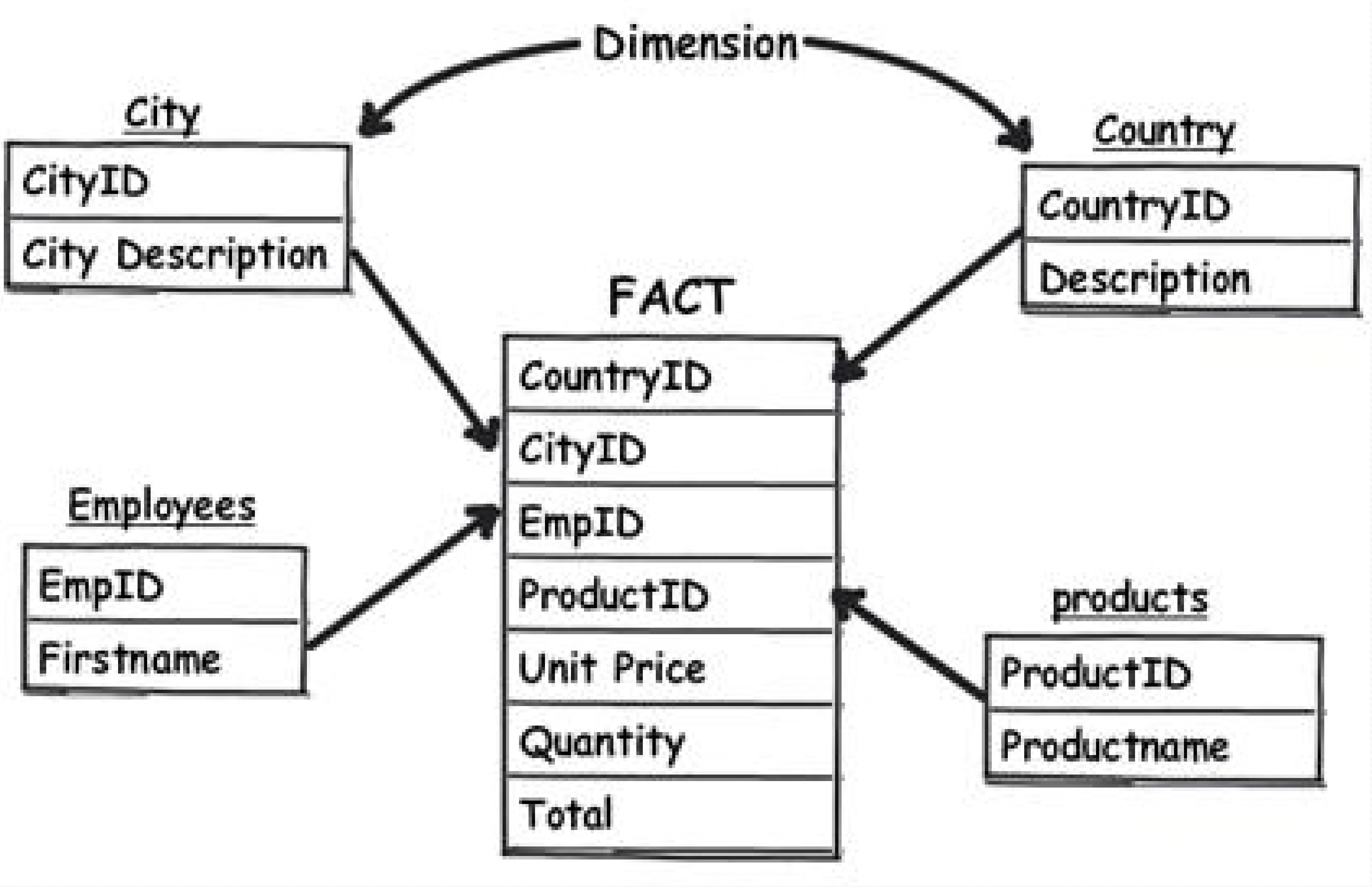
Create unique codes as database is built.

PRODUCT TABLE
- Product ID
- Product Name
- Product Rate

ORDER TABLE
- Product ID
- Order ID
- Customer ID
- Invoice No

CUSTOMER TABLE
- Customer ID
- Order ID
- Customer Name
- Customer Address
- Product ID

Fact and Dimension Tables



Dimension

City
CityID
City Description

Country
CountryID
Description

FACT
CountryID
CityID
EmpID
ProductID
Unit Price
Quantity
Total

Employees
EmpID
Firstname

products
ProductID
Productname

Insurance Grain Examples

Transaction grain

transaction_time_key
policy_key
customer_key
agent_key
coverage_key
covered_item_key
transaction_key
amount

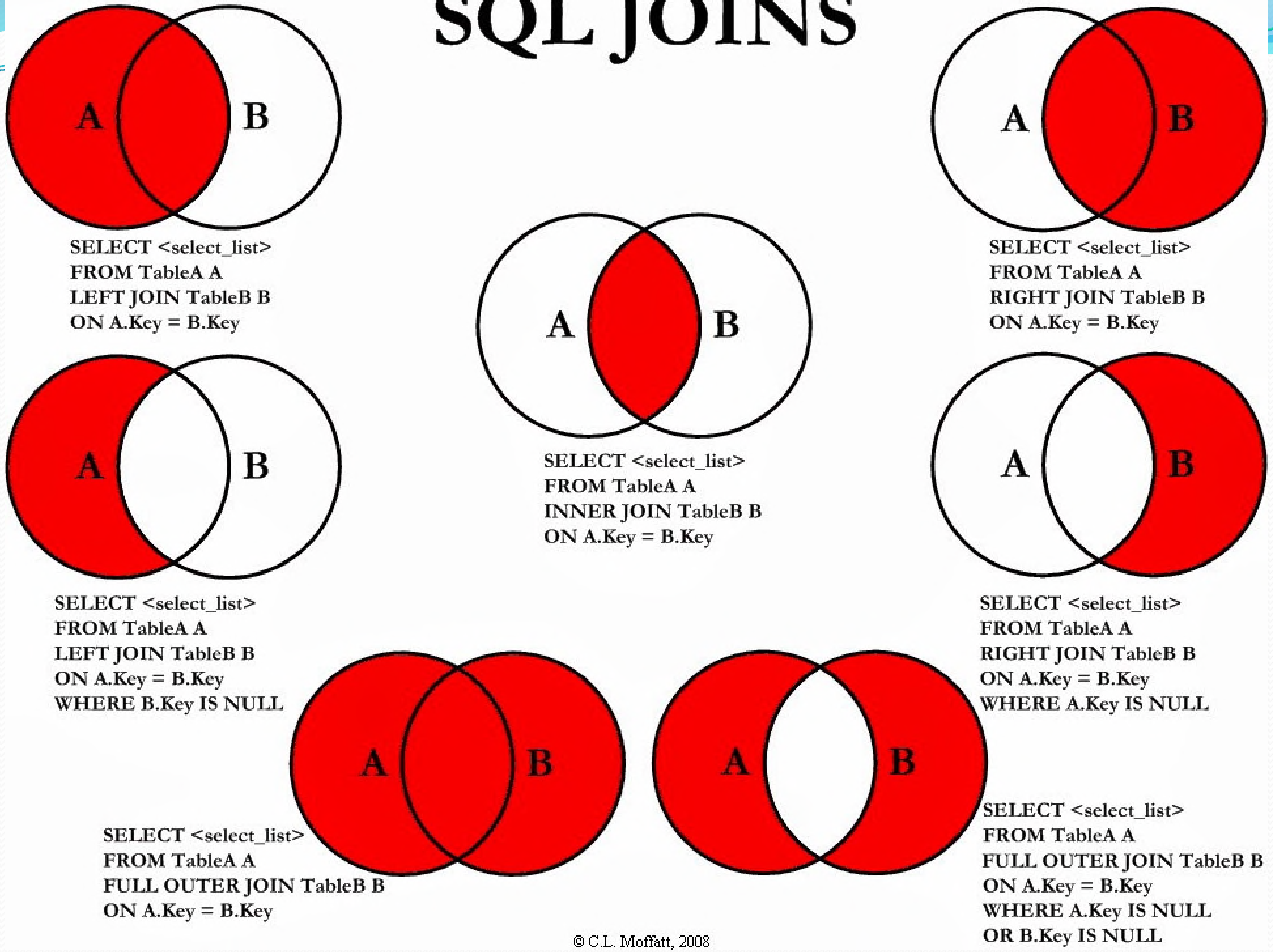
Periodic Snapshot grain

reporting_month_key
policy_key
customer_key
agent_key
coverage_key
covered_item_key
status_key
earned_premium
incurred_claims
change_in_reserve
reserve_balance
number_transactions

Accumulating Line Item grain

effective_date_key
expiration_date_key
first_claim_date_key
last_payment_date_key
policy_key
customer_key
agent_key
coverage_key
covered_item_key
status_key
earned_premium_to_date
number_claims_to_date
claims_payments_to_date

SQL JOINS



```
SELECT <select_list>
FROM TableA A
LEFT JOIN TableB B
ON A.Key = B.Key
```

```
SELECT <select_list>
FROM TableA A
INNER JOIN TableB B
ON A.Key = B.Key
```

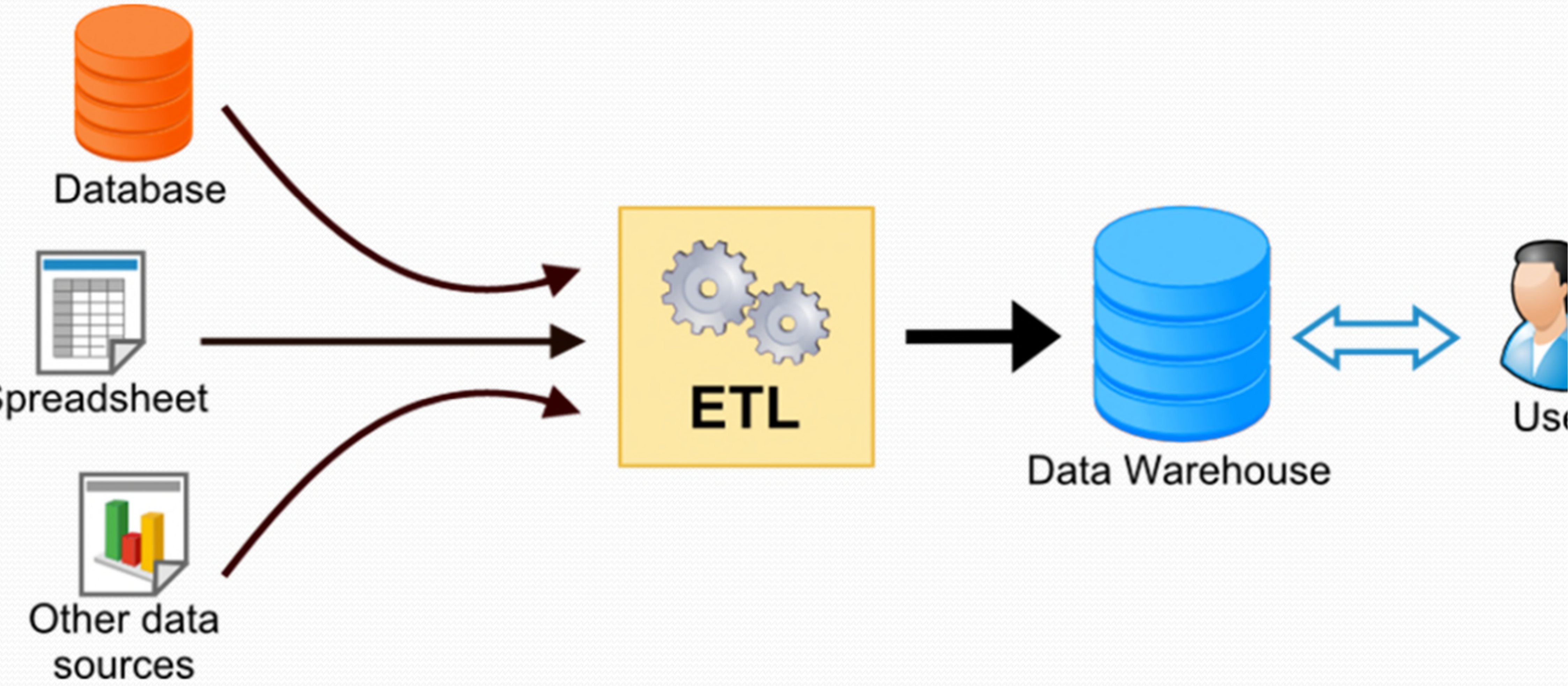
```
SELECT <select_list>
FROM TableA A
RIGHT JOIN TableB B
ON A.Key = B.Key
```

```
SELECT <select_list>
FROM TableA A
LEFT JOIN TableB B
ON A.Key = B.Key
WHERE B.Key IS NULL
```

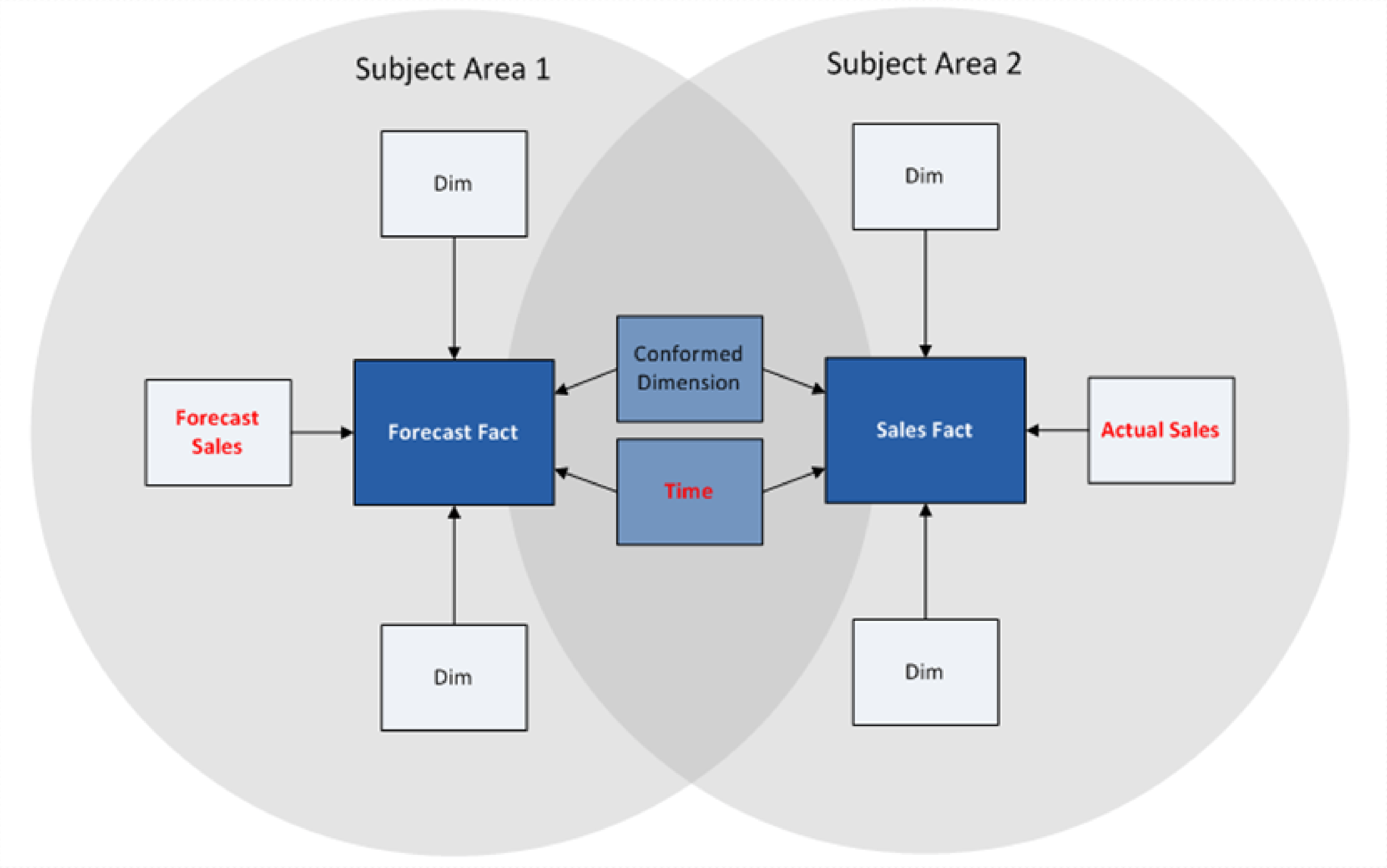
```
SELECT <select_list>
FROM TableA A
RIGHT JOIN TableB B
ON A.Key = B.Key
WHERE A.Key IS NULL
```

```
SELECT <select_list>
FROM TableA A
FULL OUTER JOIN TableB B
ON A.Key = B.Key
```

```
SELECT <select_list>
FROM TableA A
FULL OUTER JOIN TableB B
ON A.Key = B.Key
WHERE A.Key IS NULL
OR B.Key IS NULL
```


Database
Spreadsheet
Other data sources
ETL
Data Warehouse
Use

Conformed Dimensions



Business Intelligence

BI aims to enable fast and easy access to information in support of decision analysis

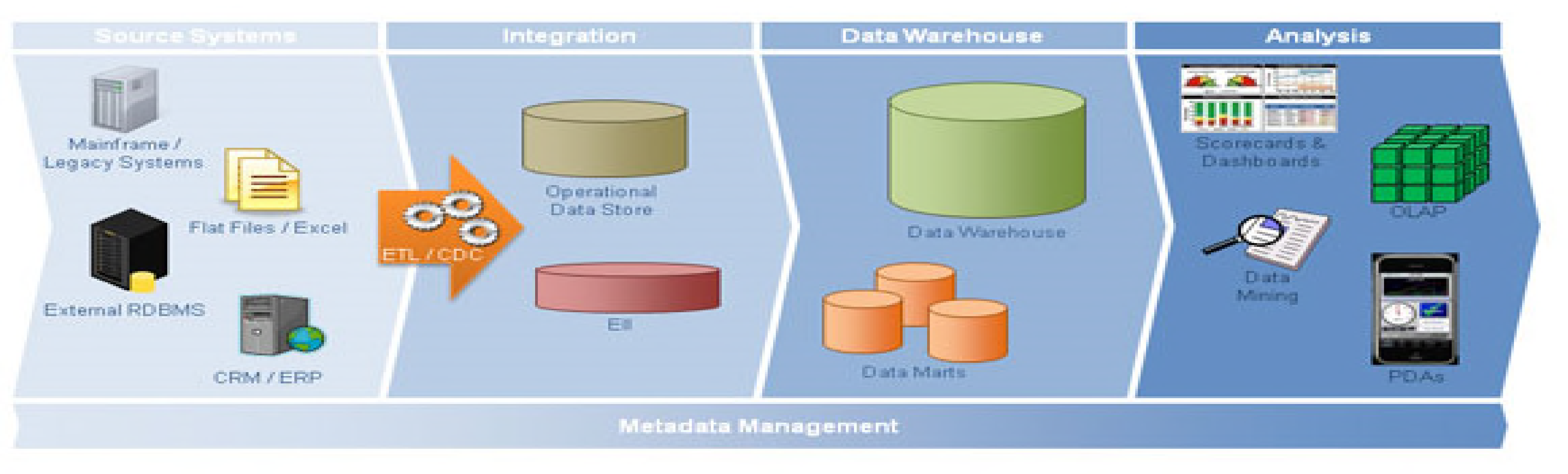
The data scientist uses BI to tell a story with data

Conclusions should be easy to understand and compelling

Knowledge of business intelligence concepts improves both insight and communication

BI Defined

A "…set of strategies, processes, applications, data, products, technologies and technical architectures which are used to support the collection, analysis, presentation and dissemination of business information" - Wikipedia



BI Terms You Should know

Business Intelligence – integrated applications and databases

Self-Serve BI – BI designed to be intuitive and safe

BI Applications – Tools used to organize and present the data

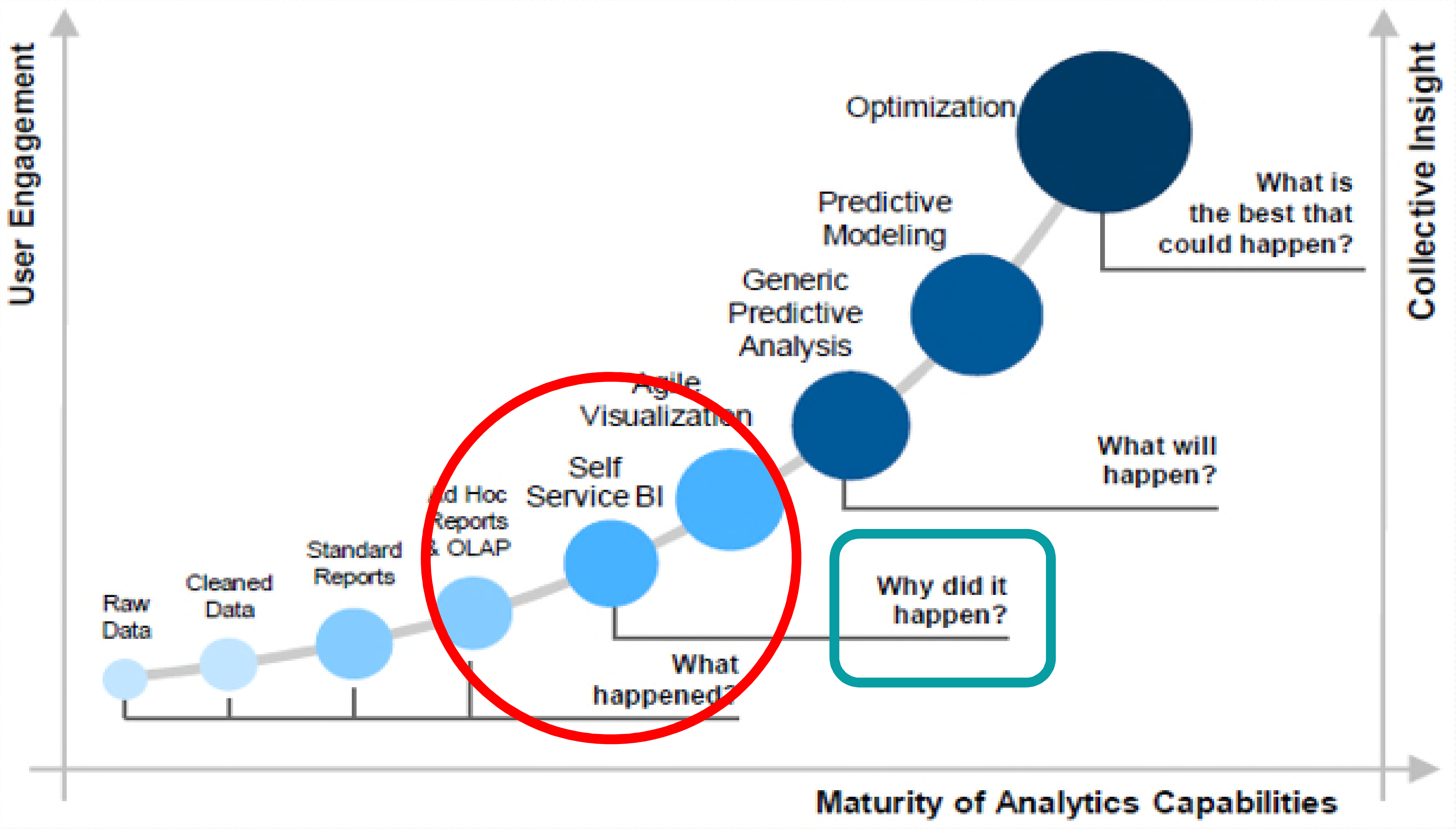
BI Database – Data Warehouses, Data Marts, OLAP Cubes, etc.

Data Visualization – a picture is worth a thousand tables

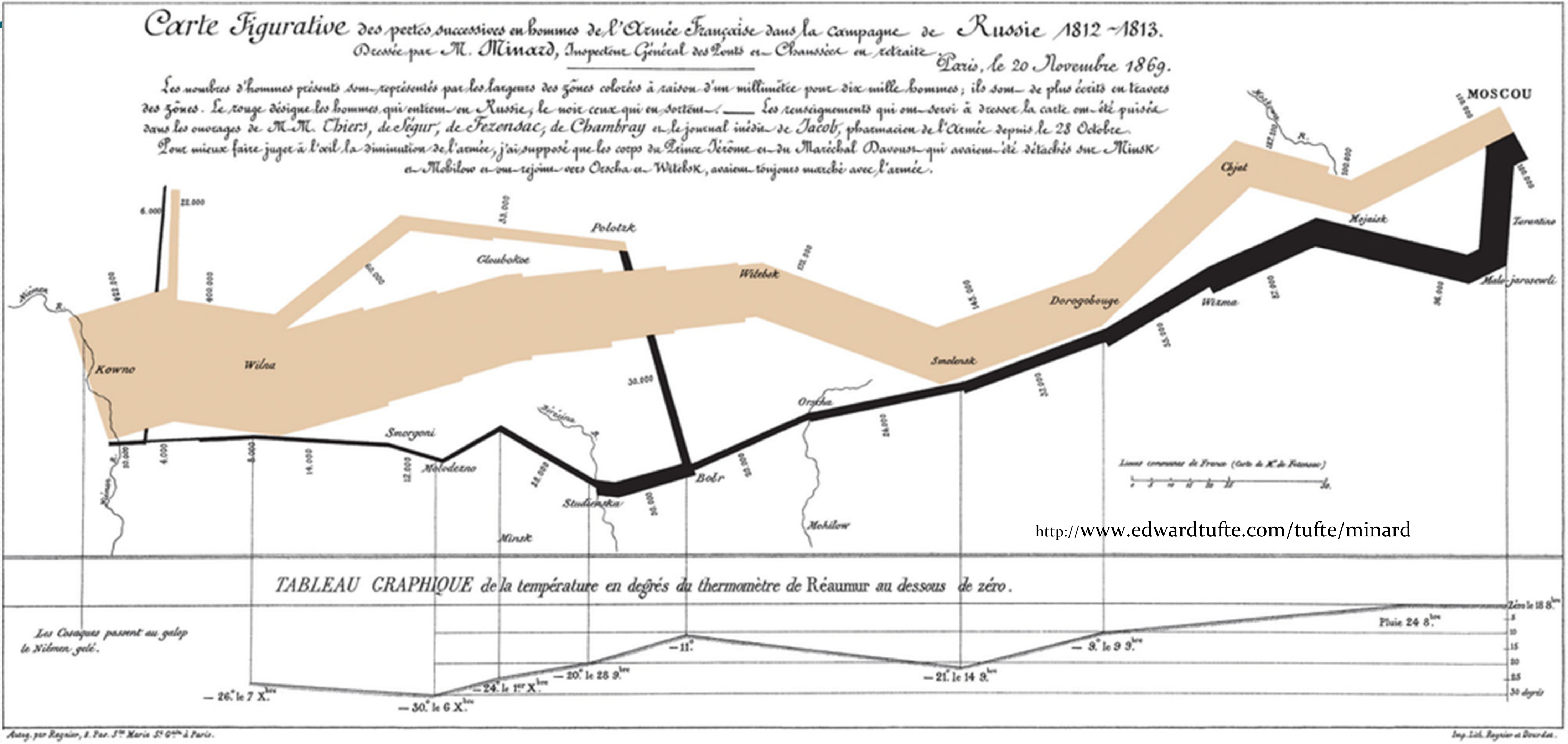
Agile – An iterative approach to BI development

Pixel Perfect Reports – reports that have to be consistent

Self-Service BI



Data Visualization



Carte Figurative des pertes successives en hommes de l'Armée Française dans la campagne de Russie 1812-1813.
Dressée par M. Minard, Inspecteur Général des Ponts et Chaussées en retraite.
Paris, le 20 Novembre 1869.

Les nombres d'hommes présents sont représentés par les largeurs des zônes colorées à raison d'un millimètre pour dix mille hommes; ils sont de plus écrits en travers des zônes. Le rouge désigne les hommes qui entrent en Russie, le noir ceux qui en sortent. —— Les renseignements qui ont servi à dresser la carte ont été puisés dans les ouvrages de MM. Thiers, de Ségur, de Fezensac, de Chambray et le journal inédit de Jacob, pharmacien de l'Armée depuis le 28 Octobre.
Pour mieux faire juger à l'œil la diminution de l'armée, j'ai supposé que les corps du Prince Jérôme et du Maréchal Davoust qui avaient été détachés sur Minsk et Mohilow et ont rejoint vers Orscha et Witebsk, avaient toujours marché avec l'armée.

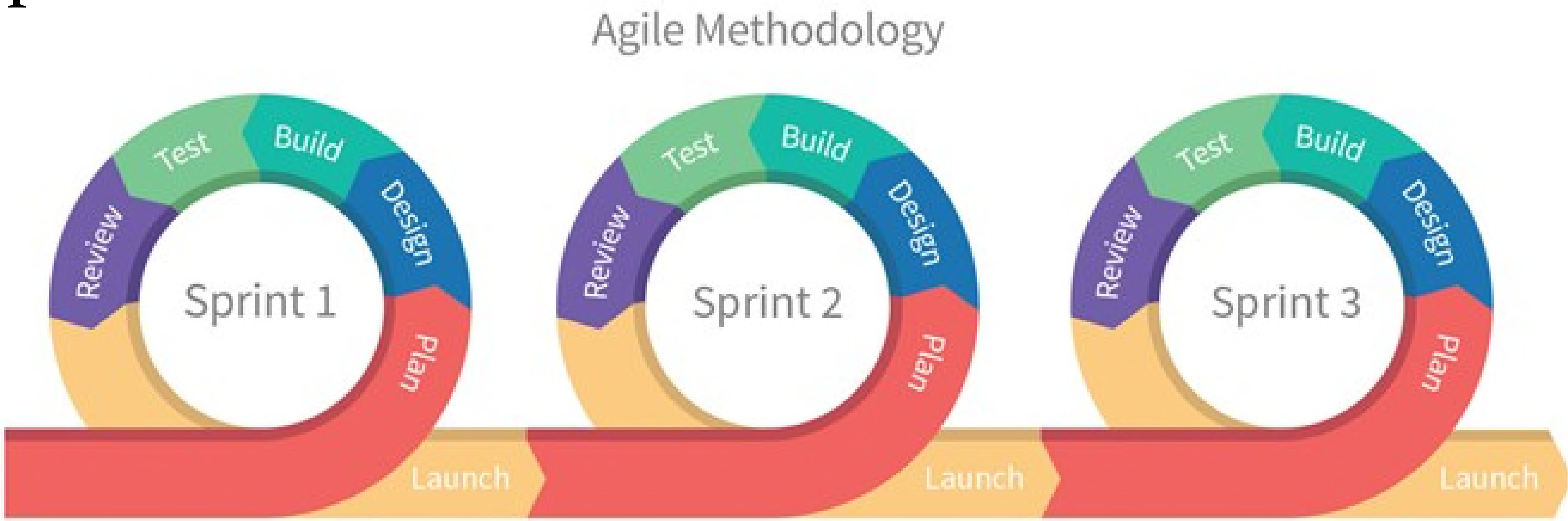
TABLEAU GRAPHIQUE de la température en degrés du thermomètre de Réaumur au dessous de zéro.

Les Cosaques passent au galop le Niémen gelé.

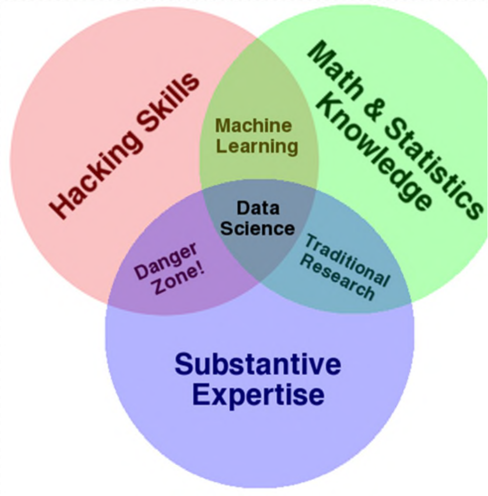
http://www.edwardtufte.com/tufte/minard

Agile

a set of principles for software development under hich requirements and solutions evolve through the llaborative effort of self-organizing cross-functional ams” - Wikipedia



Data Science

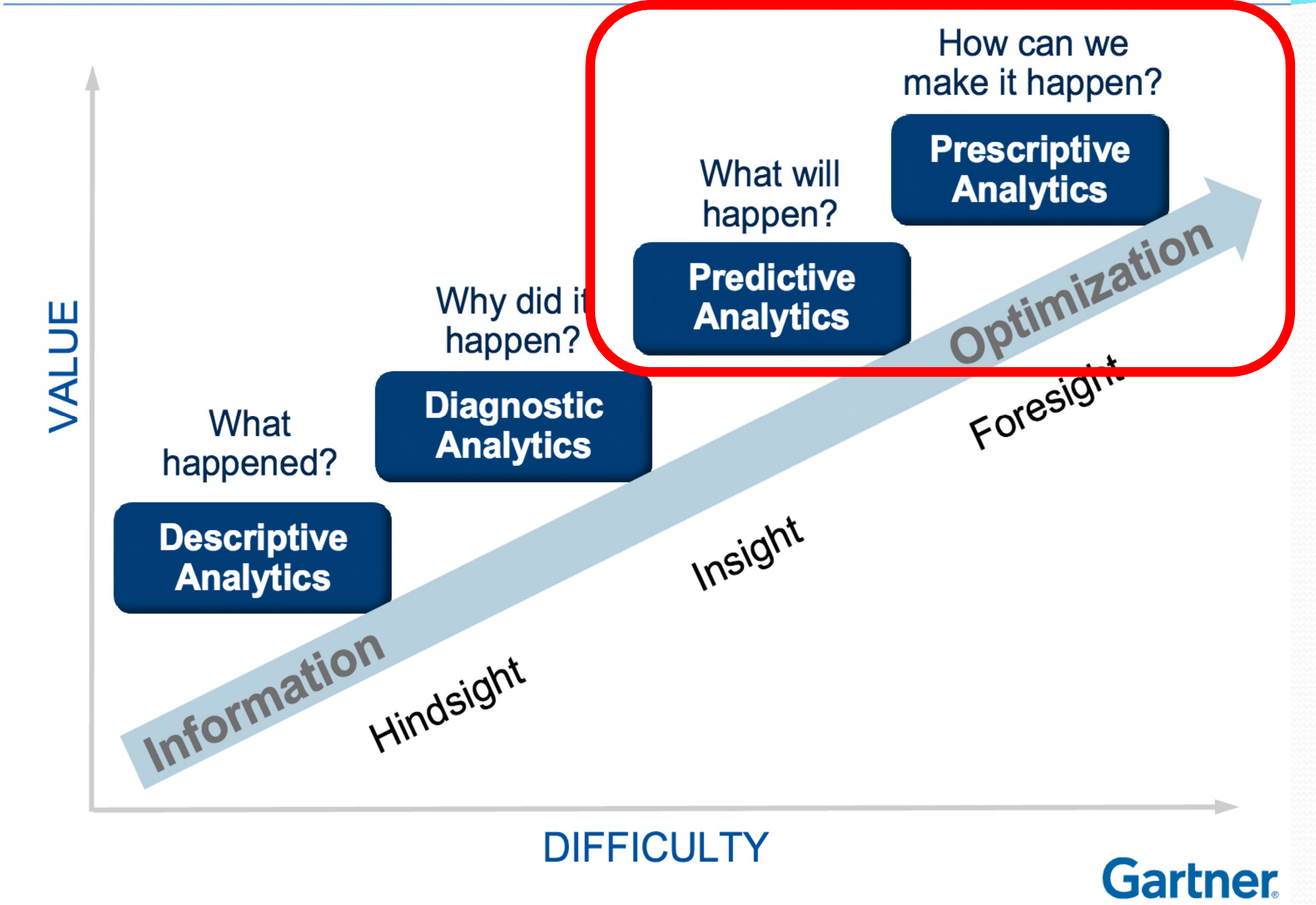


Data Science is...

- methodologies to extract knowledge and insights..
- from data in various forms, either structured or unstructured
- and a continuation of data analysis fields such as data mining, operations research, and predictive analytics..

The depth of domain knowledge and analytic rigor defines the difference between dangerous, misleading hacking and true data science

Gartner Analytic Ascendancy Model



Data Science Terms You Should Know

Big Data - Velocity, Variety, Volume, Veracity precludes traditional analysis

Hadoop – generally refers to a computing environment used for Big Data

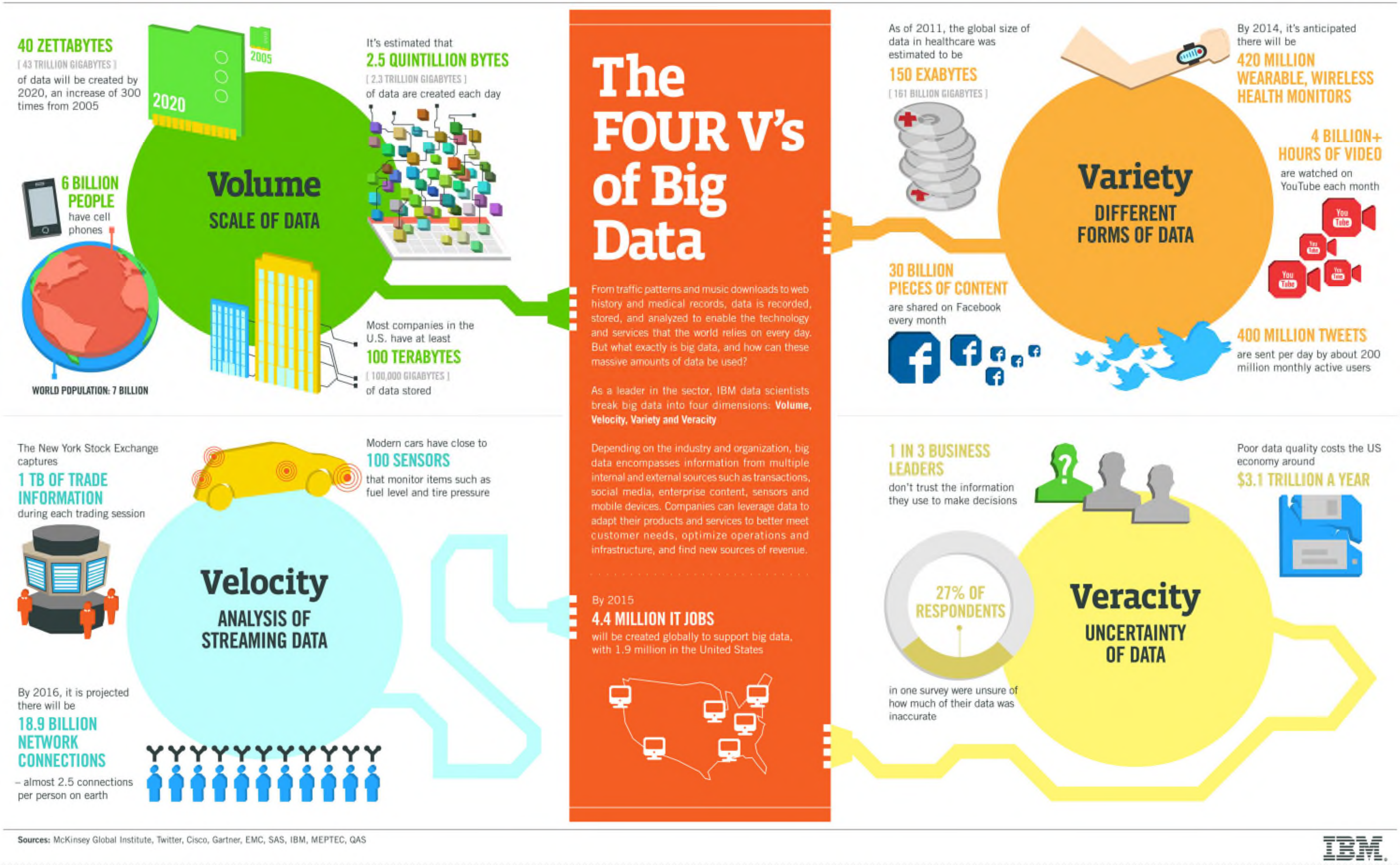
Test and Learn – A/B testing, designed experimentation accelerate insight

Unstructured Data – Text mining, Telematics, IoT, audio recordings, video

Best Practices

- Define the **Target Variable** & identify the potential **Independent Variables**
- Prepare the data
- **Train, Test & Validate**
- **Implementation** & **Adoption**
- **Monitoring**

Big



The FOUR V's of Big Data

From traffic patterns and music downloads to web history and medical records, data is recorded, stored, and analyzed to enable the technology and services that the world relies on every day. But what exactly is big data, and how can these massive amounts of data be used?

As a leader in the sector, IBM data scientists break big data into four dimensions: **Volume, Velocity, Variety and Veracity**

Depending on the industry and organization, big data encompasses information from multiple internal and external sources such as transactions, social media, enterprise content, sensors and mobile devices. Companies can leverage data to adapt their products and services to better meet customer needs, optimize operations and infrastructure, and find new sources of revenue.

By 2015 **4.4 MILLION IT JOBS** will be created globally to support big data, with 1.9 million in the United States

Volume

SCALE OF DATA

40 ZETTABYTES [43 TRILLION GIGABYTES] of data will be created by 2020, an increase of 300 times from 2005

It's estimated that **2.5 QUINTILLION BYTES** [2.3 TRILLION GIGABYTES] of data are created each day

6 BILLION PEOPLE have cell phones

WORLD POPULATION: 7 BILLION

Most companies in the U.S. have at least **100 TERABYTES** [100,000 GIGABYTES] of data stored

Velocity

ANALYSIS OF STREAMING DATA

The New York Stock Exchange captures **1 TB OF TRADE INFORMATION** during each trading session

Modern cars have close to **100 SENSORS** that monitor items such as fuel level and tire pressure

By 2016, it is projected there will be **18.9 BILLION NETWORK CONNECTIONS** – almost 2.5 connections per person on earth

Variety

DIFFERENT FORMS OF DATA

As of 2011, the global size of data in healthcare was estimated to be **150 EXABYTES** [161 BILLION GIGABYTES]

By 2014, it's anticipated there will be **420 MILLION WEARABLE, WIRELESS HEALTH MONITORS**

4 BILLION+ HOURS OF VIDEO are watched on YouTube each month

30 BILLION PIECES OF CONTENT are shared on Facebook every month

400 MILLION TWEETS are sent per day by about 200 million monthly active users

Veracity

UNCERTAINTY OF DATA

1 IN 3 BUSINESS LEADERS don't trust the information they use to make decisions

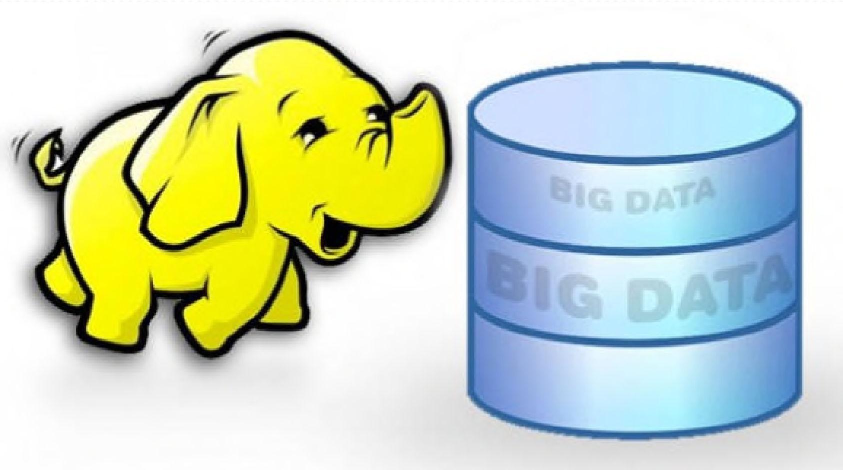
Poor data quality costs the US economy around **$3.1 TRILLION A YEAR**

27% OF RESPONDENTS in one survey were unsure of how much of their data was inaccurate

Sources: McKinsey Global Institute, Twitter, Cisco, Gartner, EMC, SAS, IBM, MEPTEC, QAS

IBM.

Hadoop



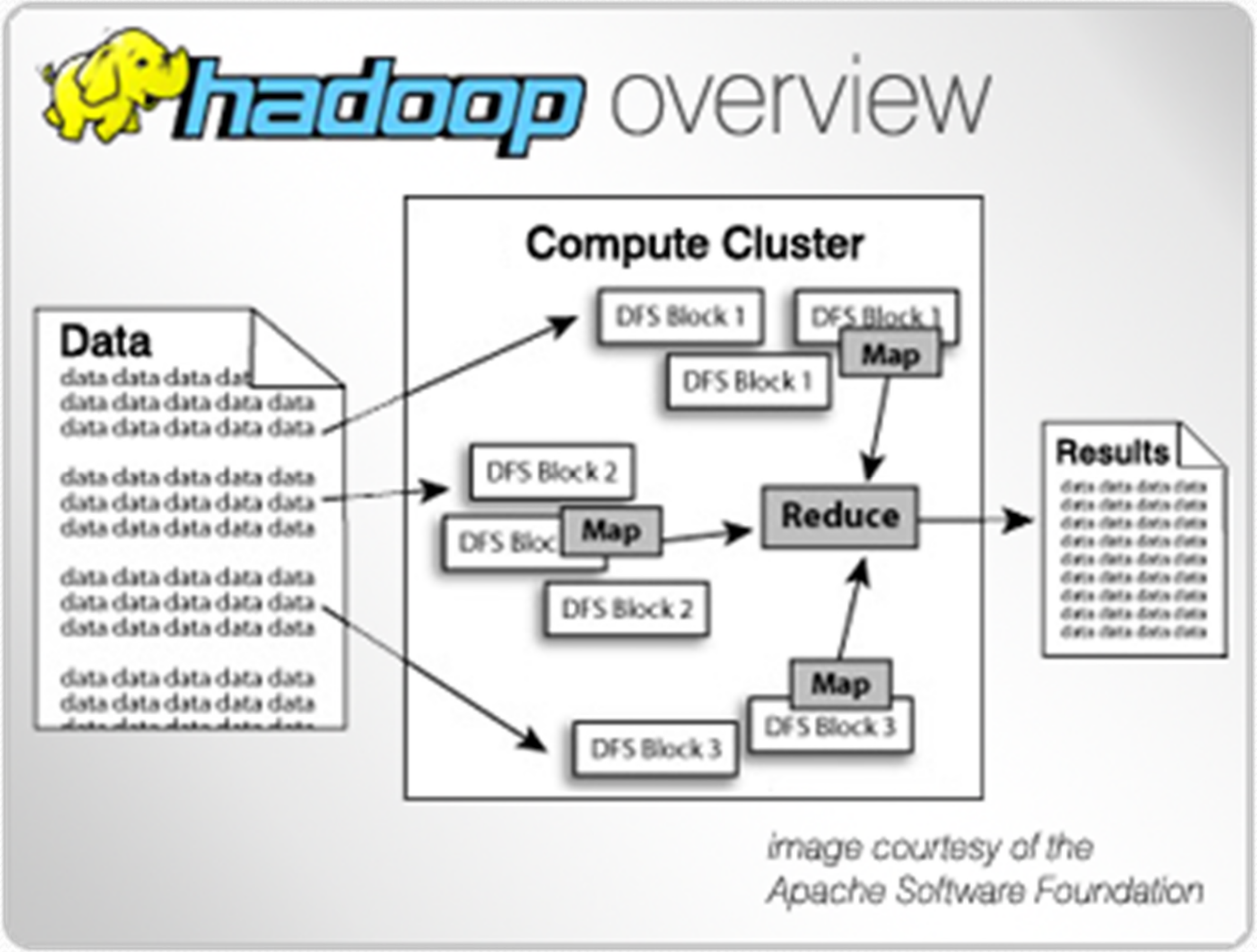
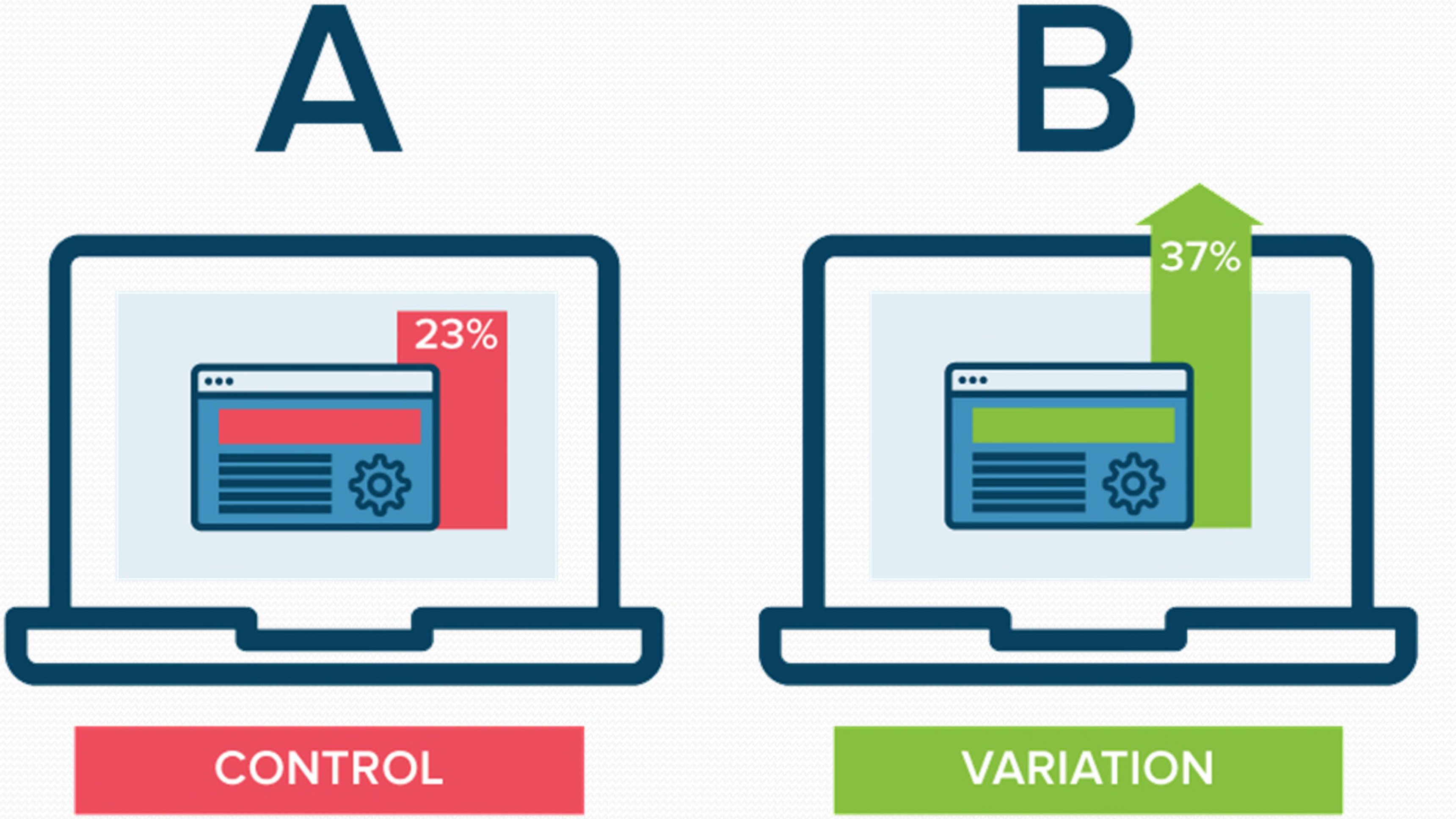
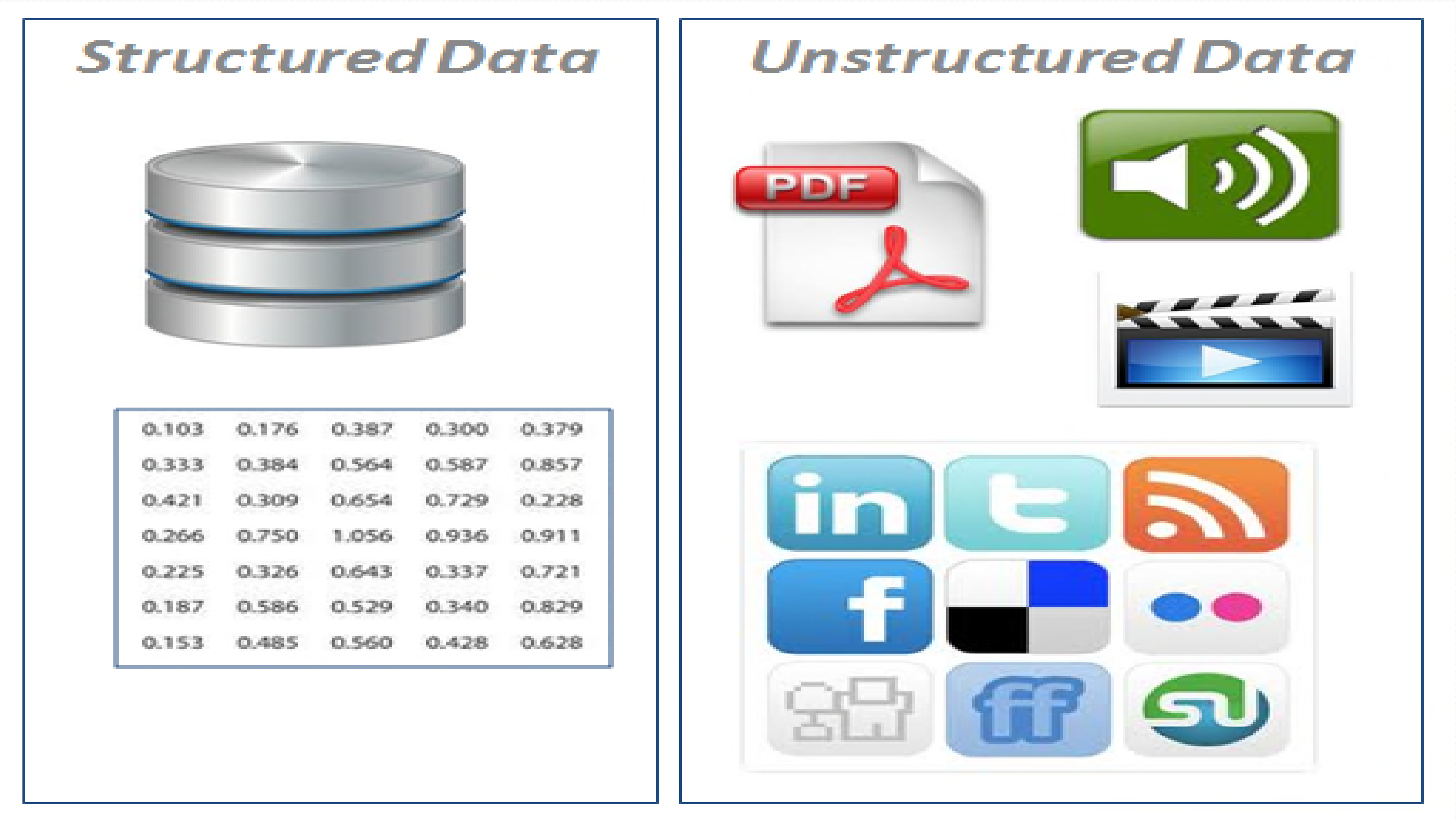


image courtesy of the Apache Software Foundation

Test and Learn



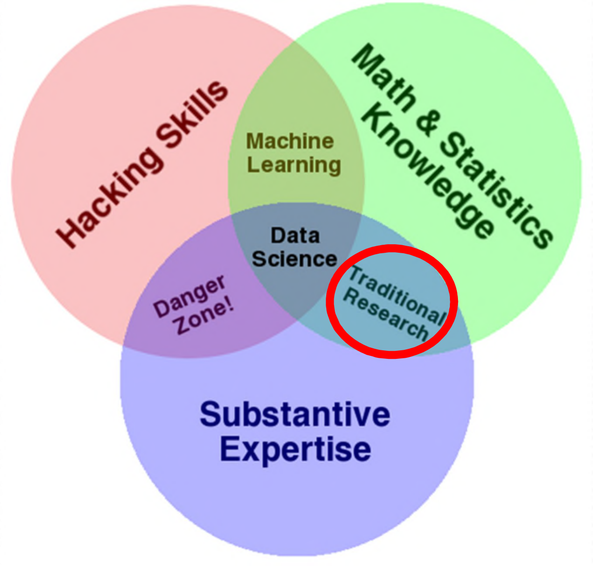
Unst



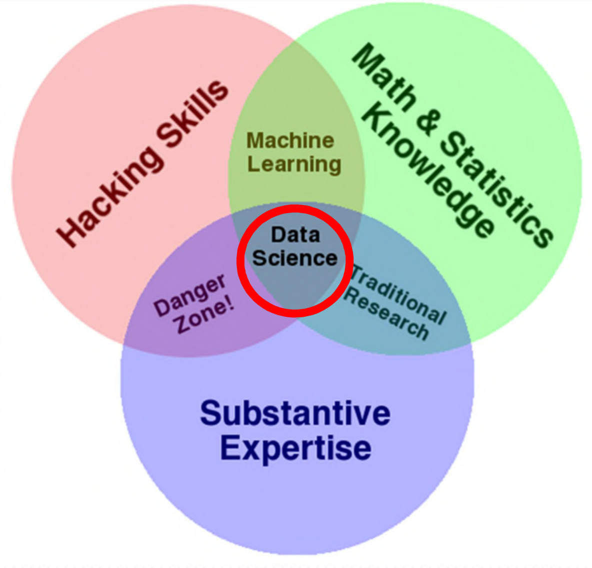
0.103	0.176	0.387	0.300	0.379
0.333	0.384	0.564	0.587	0.857
0.421	0.309	0.654	0.729	0.228
0.266	0.750	1.056	0.936	0.911
0.225	0.326	0.643	0.337	0.721
0.187	0.586	0.529	0.340	0.829
0.153	0.485	0.560	0.428	0.628

Actuary as Data Scientist

In insurance, the actuary has a tremendous head start in domain knowledge and analytical rigor



Actuary + Data & Tech = Data Scientist



The Working Party's papers are aimed at beginning to fill any remaining gaps in data and technology knowledge

Working Party Members

- Pete Bothwell, Co-Chair
- Mary Jo Kannon, Co-Chair
- Benjamin Avanzi
- Joe Izzo
- Stephen Knobloch
- Ray Nichols
- James Norris
- Andrea Pan
- Dimitri Semenovich
- Linda Waite
- Dom Yarnell
- Cheri Widowski
- Tracy Spadola
- Michele Wetzel