



Converting High Volume Data challenges to Relevant Clinical Data Insight

**Navneet Kumar
Manager , CDM
Icon Clinical Research plc**

Introduction

Focus Area



✓ Introduction

Why Data is important?

✓ Data Challenges

Changing Paradigm in Industry ; Data Challenges Types

✓ Overcoming Data Challenge

Architecture Framework; Data Scaling; Data Wrangling, Data Lakes, Clinical text Mining

✓ New Approach to Clinical Data Management

Data Slicing ; Aggregate Data Review; Risk Based Data Quality Management

Data Types

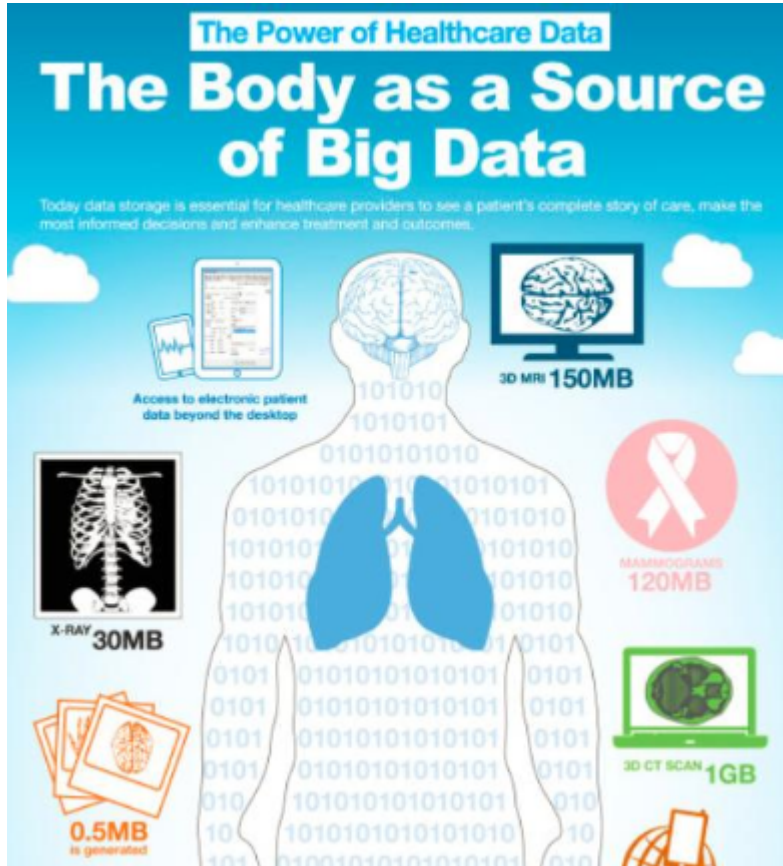
Epidemiology data

- ✓ The Surveillance Epidemiology and End Results Program (SEER)at NIH.
- ✓ Publishes cancer incidence and survival data from population-based cancer registries covering approximately 28% of the population of the US.
- ✓ Collected over the past 40 years (starting from January 1973 until now)
- ✓ Contains a total of 7.7M cases and >350,000 cases are added each year.
- ✓ Collect data on patient demographics, tumor site, tumor morphology and
- ✓ stage at diagnosis, first course of treatment, and follow-up for vital status.

Genomic Data

- ✓ Human Genome consists of 3 billion pairs of bases and particular order of As, Ts, Cs, and Gs is extremely important
- ✓ Size of single human is about 3 GB
- ✓ Whole genome sequence data is being currently annotated but not many analytics applied on this relatively new data

Image Data is really big

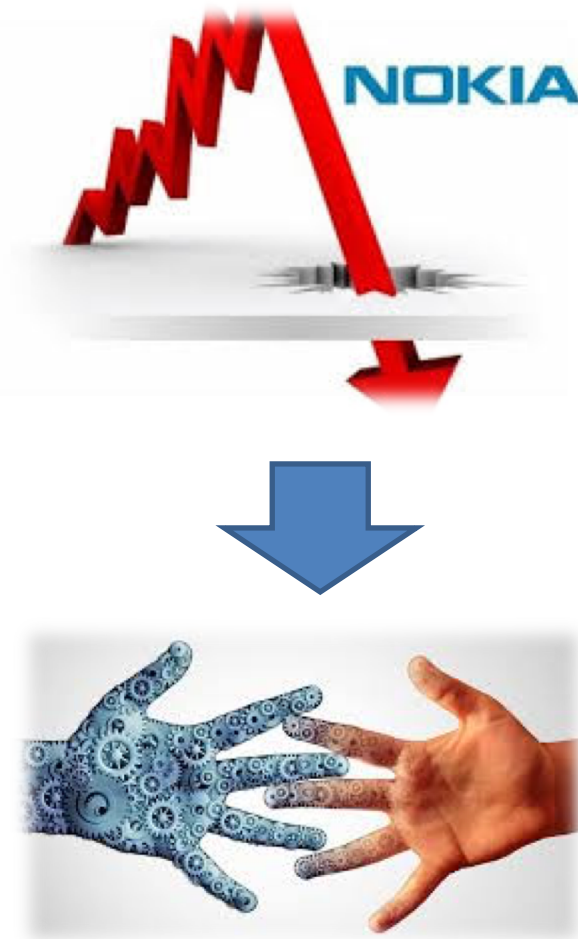


- ✓ Average hospitals will have two thirds of petabytes (665 terabytes) of patient data, of which 80% of data will be unstructured image data
- ✓ Medical imaging archives are increasing by 20%-40%

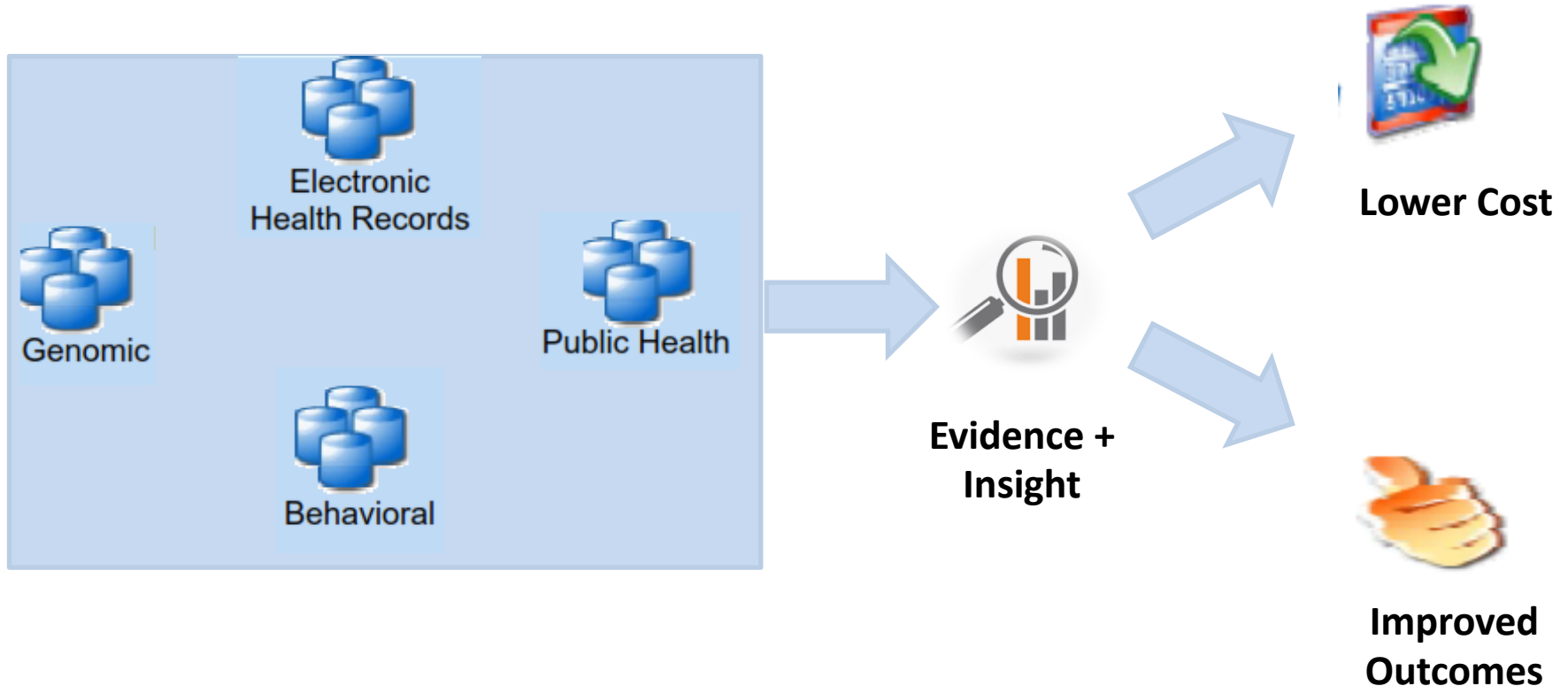


Better integrating big data, healthcare could save as much as [\\$300 billion a year](#) — that's equal to reducing costs by \$1000 a year for every man, woman, and child.

For a typical Fortune 1000 company, just a 10% increase in data accessibility will result in more than [\\$65 million additional net income](#).



What we want to Achieve



Data Challenges

Changing Paradigm

Shift Towards the Patient

Technology

- Mobile Health
- Precision Medicines



Regulation



- Affordable Health care
- Access to medicine

Expectations



- Faster Treatment
- 24*7 Personalized Care

Precision Care



1-1 Relationship



Real time insightful decision making

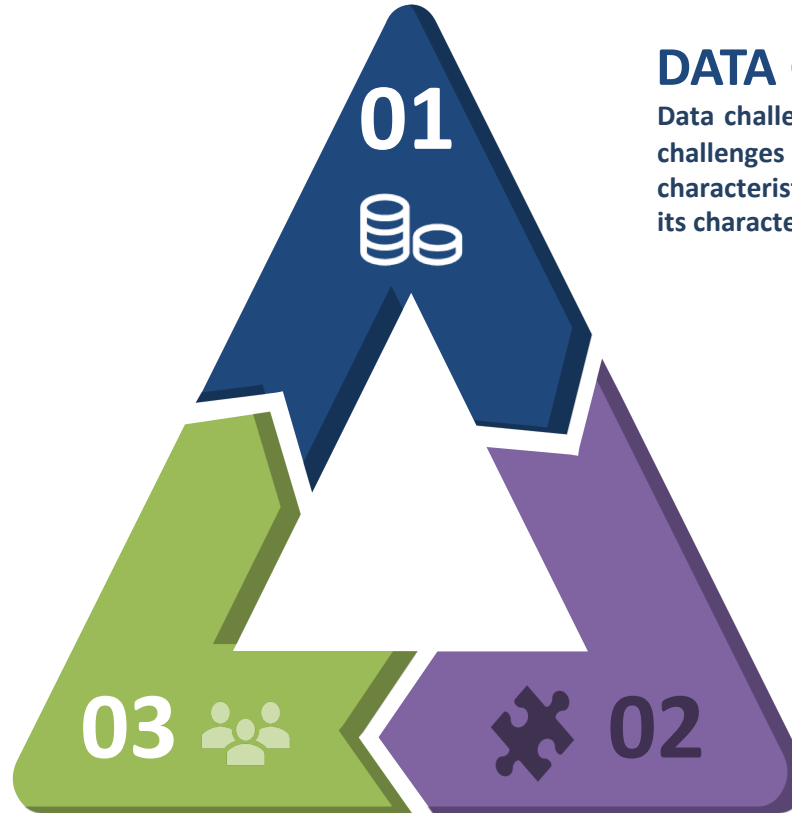


Care every ware

Roadblock to convert Data into Insight

MANAGEMENT CHALLENGES

This concerns the legal and ethical issues related to accessing data.



DATA CHALLENGES

Data challenges are the group of the challenges pertains to the characteristics of the data itself and its characteristics

PROCESS CHALLENGES

All the challenges encountered while processing the Big Data; starts with capture step and ends with presenting the output to clients, to understand the overall picture (PDF) Big Data Challenges

Data Challenges

VOLUME

Complex trials, EHRs, Insurance penetration Surveillance data etc.

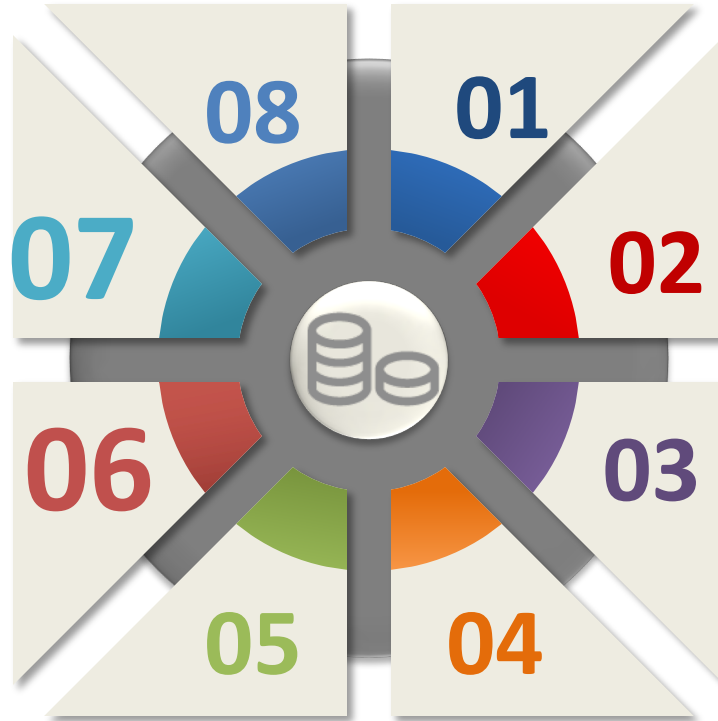
VARIETY

Data type , format, sensors and smart devices etc.

Only about 20% of data can be processed by current traditional systems and the remaining 80% are not analyzed and thereby not utilized for decision making and insight processes.

VELOCITY

Capacity of the current software application to handle and process data stream generated continuously and constantly at a pace which becomes critical due to the short shelf-life of the data which need to be analyzed in near real time if we plan to find insight in that data.



VERACITY

Biases, uncertainties, impression, untruths and missing values in the data.

QUALITY

All data is updated, free of any data issues, data is available per request and data is up to date

DISCOVERY

To identify right data for our analysis

VOLATILITY

Data Validity, duration to keep the data

DOGMATISM

Enhance domain understanding , look for things happening around us

Process Challenges

Medical data

Social Data

Payment data

Legacy data

Video/images

Data Acquisition

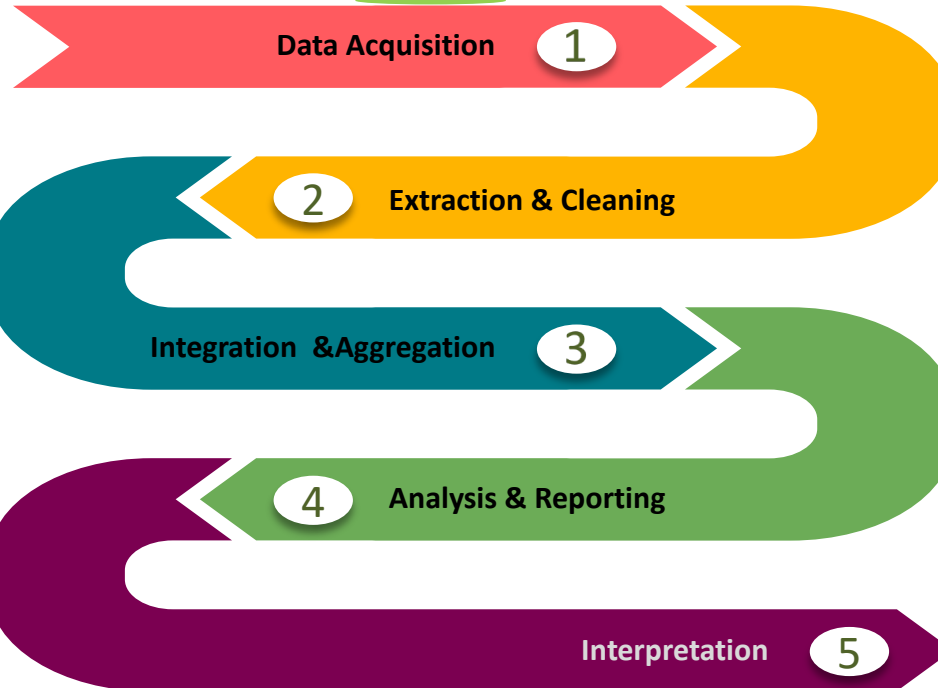
- Smart filters
- Date reduction
- Automatic Meta data generation
- Data Fidelity

Data Cleaning

- Converting structure less data to analytics friendly format
- Extracting right information
- Adequate error models

Data Aggregation

- Heterogeneity of data
- Automation of data integration and aggregation



Data Analysis

- Noisy, untrustworthy , heterogeneous data
- Integrating DB systems & analytical system
- Analytics on the fly

Data Insight

- Wrong modeling
- Erroneous data used

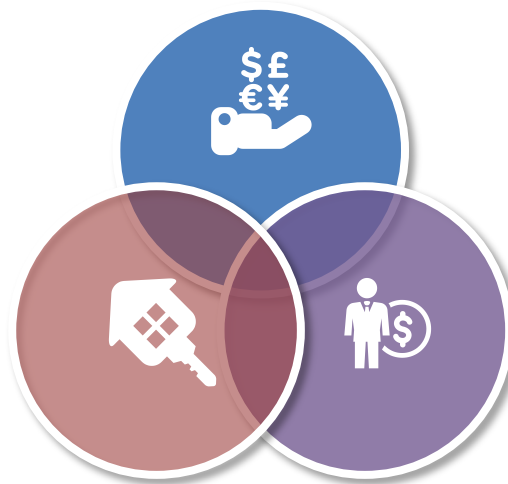
Management Challenges

Security

Variety, velocity and volume attributes of big data amplifies the security management challenges, Distributed nature of data

Privacy

There is an increasing fear of inappropriate use of personal data especially when combining this data from multiple sources.



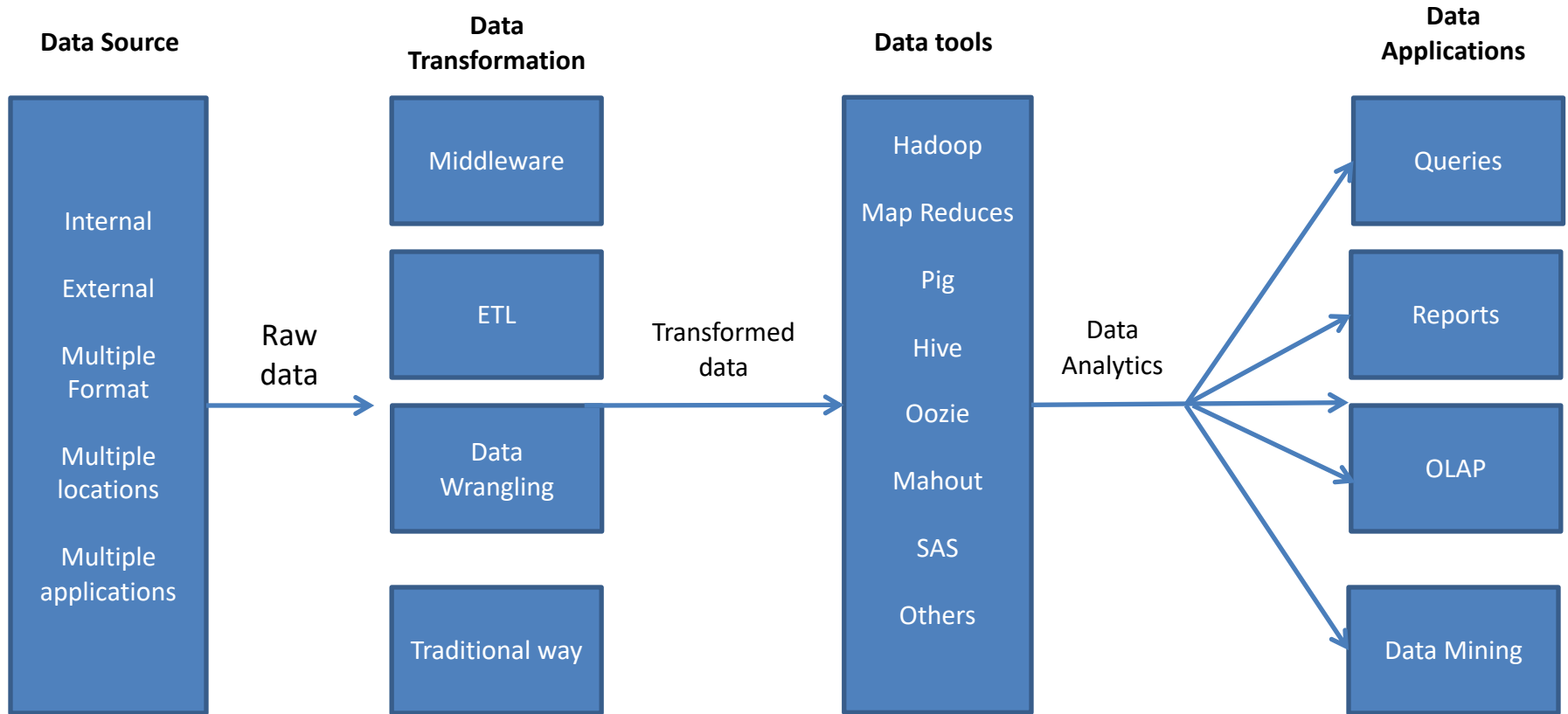
Governance

To make decision with confidence, to plan accurately for future , to avoid costs resulted from low quality data and need to re-do the work again, and provide big data reporting compatible with government standards

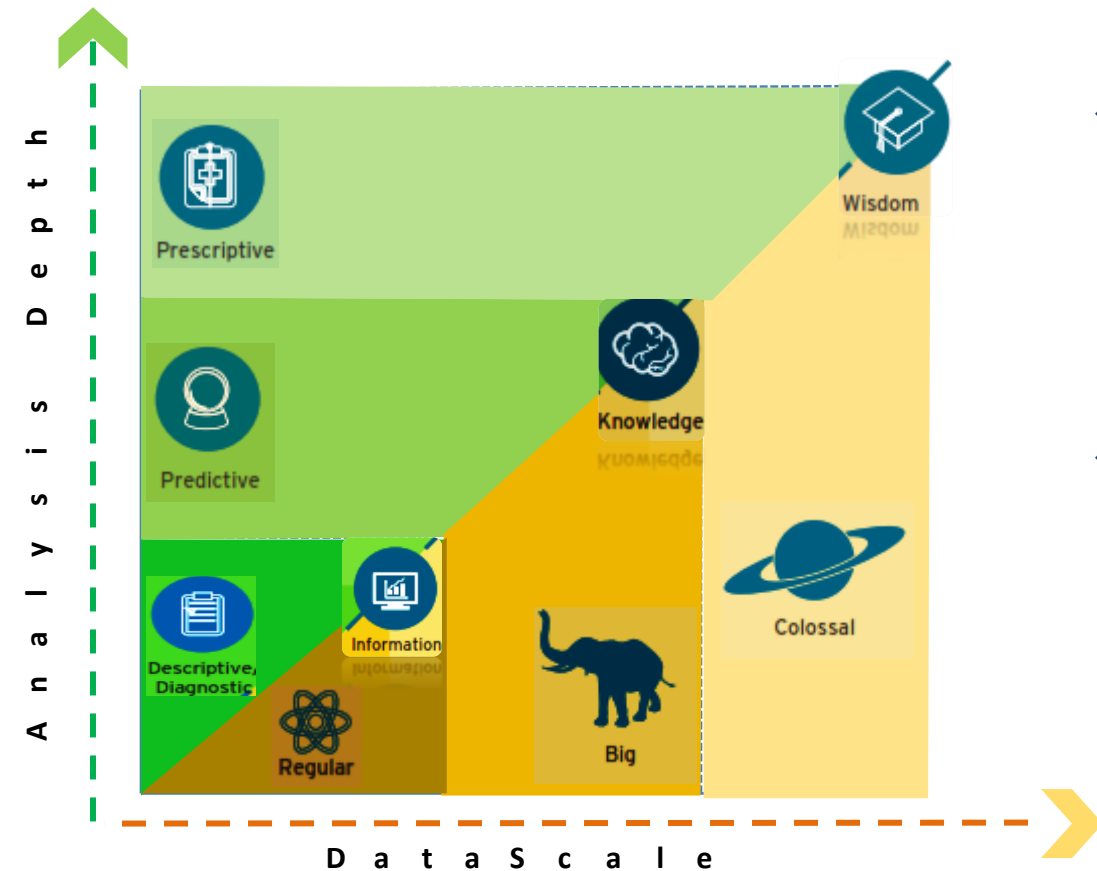
Legal and Ethical aspect of data

Overcoming Data Challenges

Framework to Manage Data Volumes



Level of Detailing



- ❖ Every Piece of data has value
 - ✓ Information
 - ✓ Knowledge
 - ✓ Wisdom
- ❖ Depth of analysis
 - ✓ Descriptive
 - ✓ Diagnostic
 - ✓ Predictive
 - ✓ Prescriptive

Data Wrangling



Use case-I:

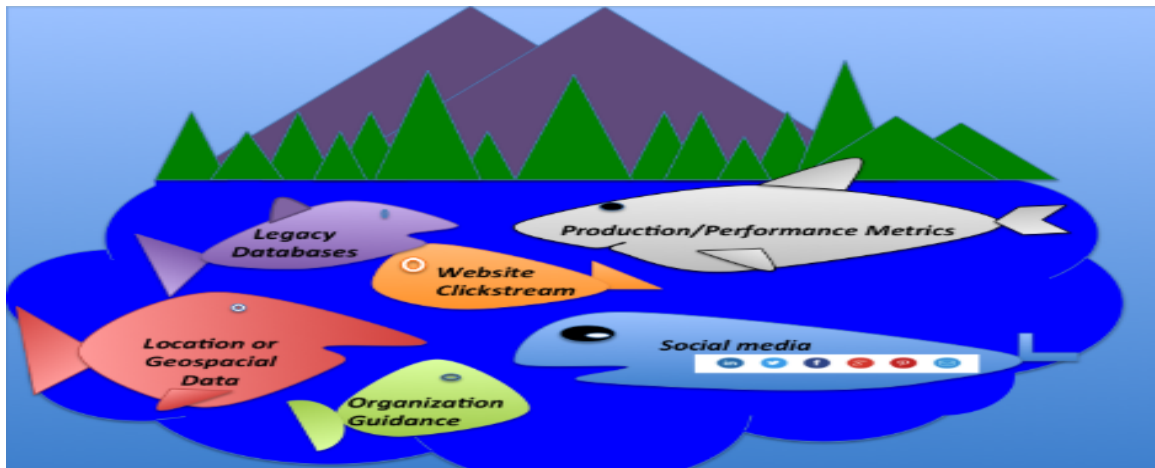
Sanofi Accelerated the Standardization of Clinical Trial, Marketing and Commercial Data to Deliver New Insights on Consumer Health and Drug Development using Data Wrangling software Trifacta

Use Case –II

Accelerating Detection of Adverse Drug Reaction in pharmacovigilance

- ✓ Better collaboration
- ✓ Provide right information to agencies, healthcare providers and patients
- ✓ Improve response times
- ✓ Resolve drug safety concerns quickly

Data Lakes



ISASA

Ingest **S**tore **A**nalyze **S**urface **A**ct



❖ Build Application



❖ Flexibility & Accessibility



❖ Data Authenticity



❖ Speed



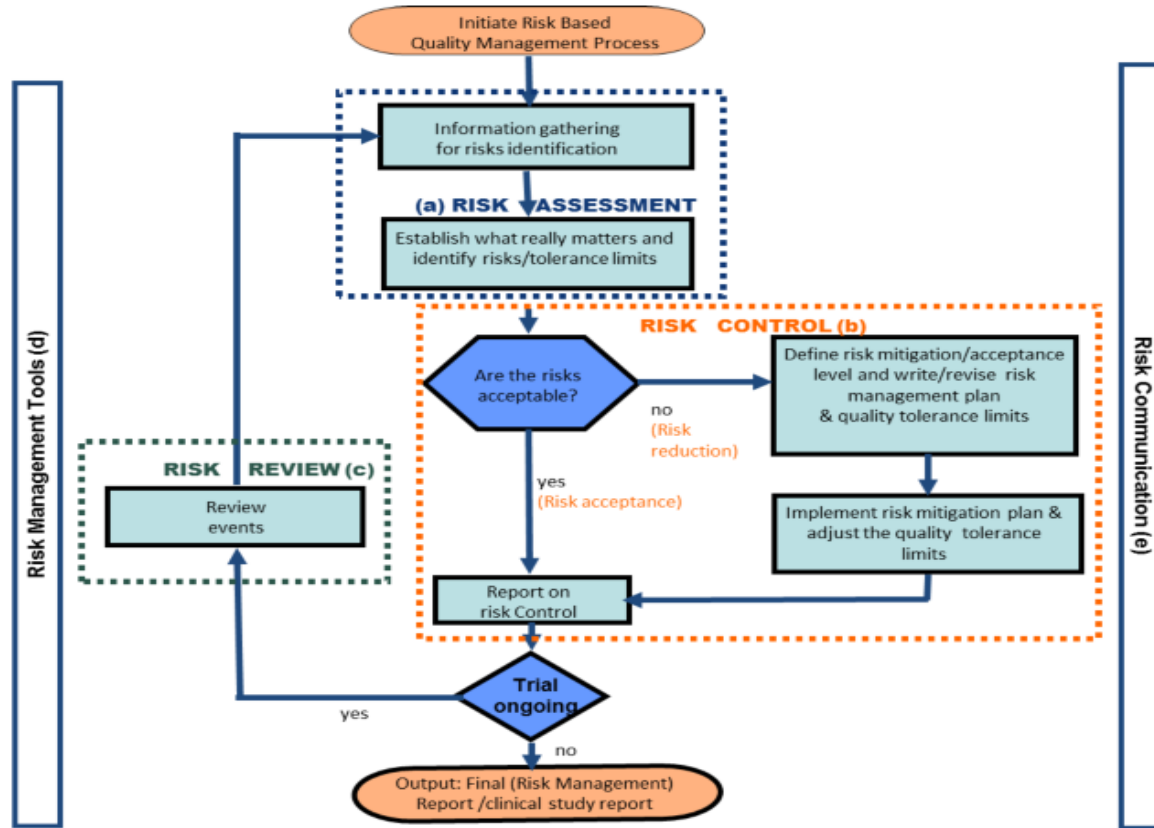
❖ Explore and Analysis

Clinical Text Mining

- ✓ Text Mining
 - Information Extraction
 - ✓ Name entity recognition
 - Informational retrieval
 - Index of words
 - Ranking of matching documents
- ✓ Clinical text vs Biomedical text
 - Biomedical Text- medical literatures
 - Clinical text: Clinical notes
- ✓ Auto encoding
 - Extracting codes from clinical text
- ✓ Context Analysis-Negation
 - NegEx
 - NegExpander
 - NegFinder
- ✓ Context Analysis-Temporality

New Approach to Clinical Data Management

Risk based Data Quality Management



- ✓ Monitor data taking into account risk factors and categories in order to track study progression and solve critical situations.
- ✓ Focus on data directly impacting primary and secondary objectives.
- ✓ Develop Data checks based on data peculiarity

Focusing on Trend and Fraud

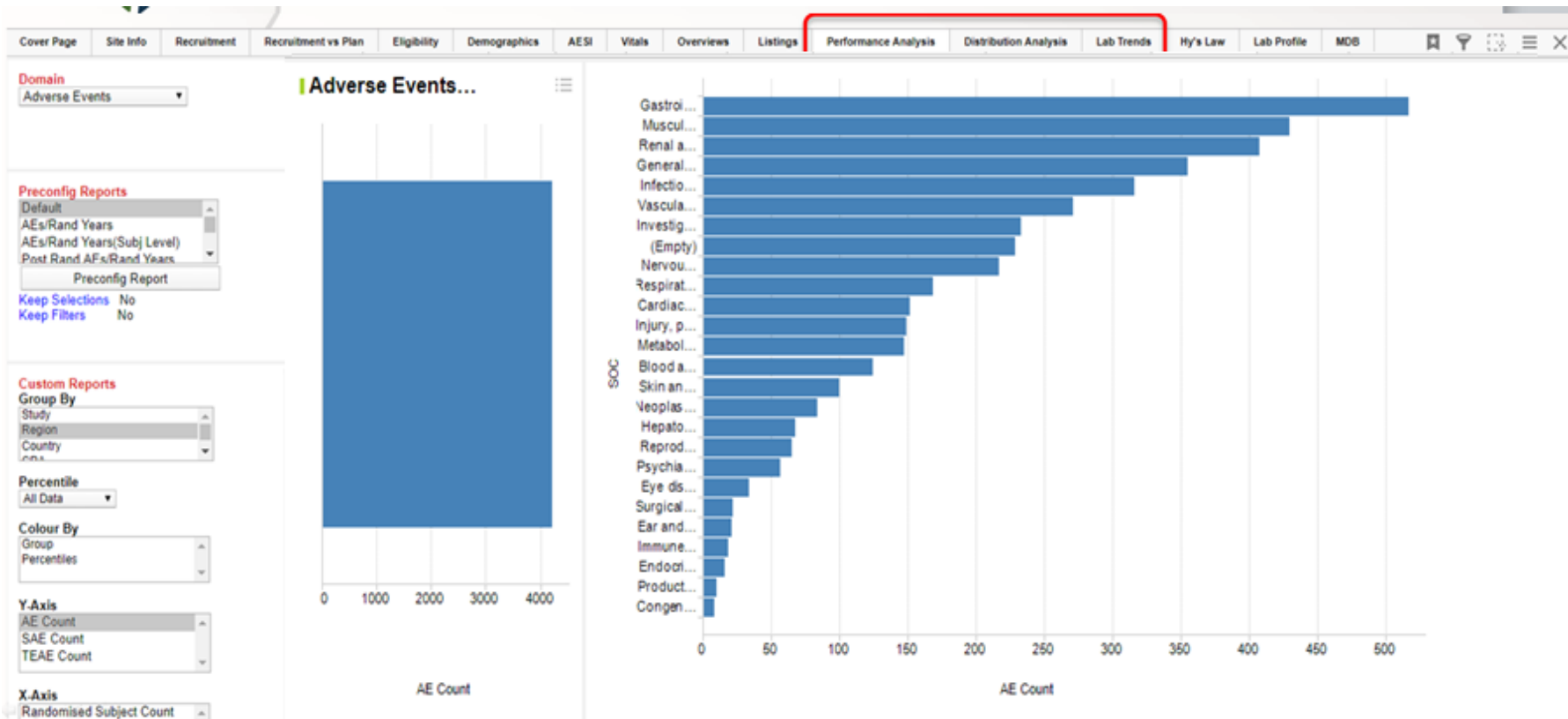
Country	Site	#	#	# of AEs at Site	%	Avg. # AEs per Subject per Site	Avg. #	Expected #	Ratio Actual vs
UNITED STATES	001	15	15	134	6.667	8.933	20.541	308.119	43.49
UNITED STATES	003	17	17	318	0	18.706	20.541	349.202	91.065
UNITED STATES	005	2	2	58	0	29	20.541	41.083	141.179
UNITED STATES	006	23	23	373	0	16.217	20.541	472.45	78.95
UNITED STATES	008	12	12	150	0	12.5	20.541	246.495	60.853

Site Level Analysis

Country Level Analysis

Country	# Site	# Screened Subjects in Country	# Randomised Subjects in Country	# of AEs in Country	% Randomised Subject without AE	Avg. # AEs per Subject per Country	Avg. # AEs per Subject in Study	Expected # AEs in Country	Ratio Actual vs Expected
UNITED STATES	38	327	313	6717	0.319	20.541	20.541	6717	100

New Approach to Data Management



Summary

- ✓ Health care and life sciences are a **data rich domain**.
- ✓ Unraveling huge data complexities can provide many insights about making **the right decisions at the right time for the patents**
- ✓ Efficiently utilizing the colossal data can help in **improving patient outcome and also reducing cost**

