What's Big Data?

No single definition; here is from Wikipedia:

Big data is the term for a collection of data sets so large and complex that it becomes difficult to process using on-hand database management tools or traditional data processing applications.

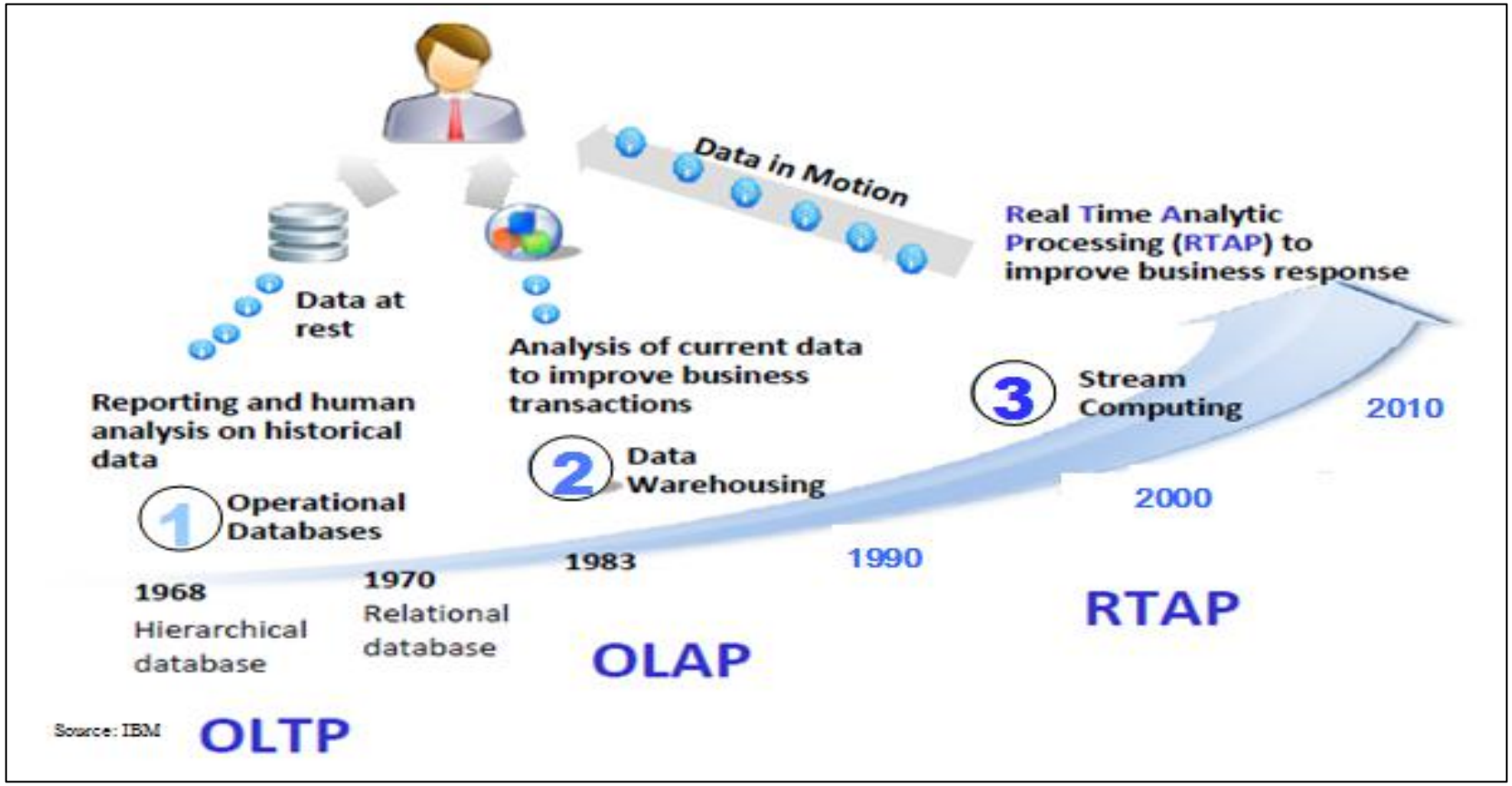
The challenges include ***capture, curation, storage, search, sharing, transfer, analysis, and visualization***.

The trend to larger data sets is due to the additional information derivable from analysis of a single large set of related data, as compared to separate smaller sets with the same total amount of data, allowing correlations to be found to

"**spot business trends, determine quality of research, prevent diseases, link legal citations, combat crime, and determine real-time roadway traffic conditions.**"

Embracing Big Data

Harnessing Big Data



- **OLTP:** Online Transaction Processing (DBMSs)
- **OLAP:** Online Analytical Processing (Data Warehousing)
- **RTAP:** Real-Time Analytics Processing (Big Data Architecture & technology)

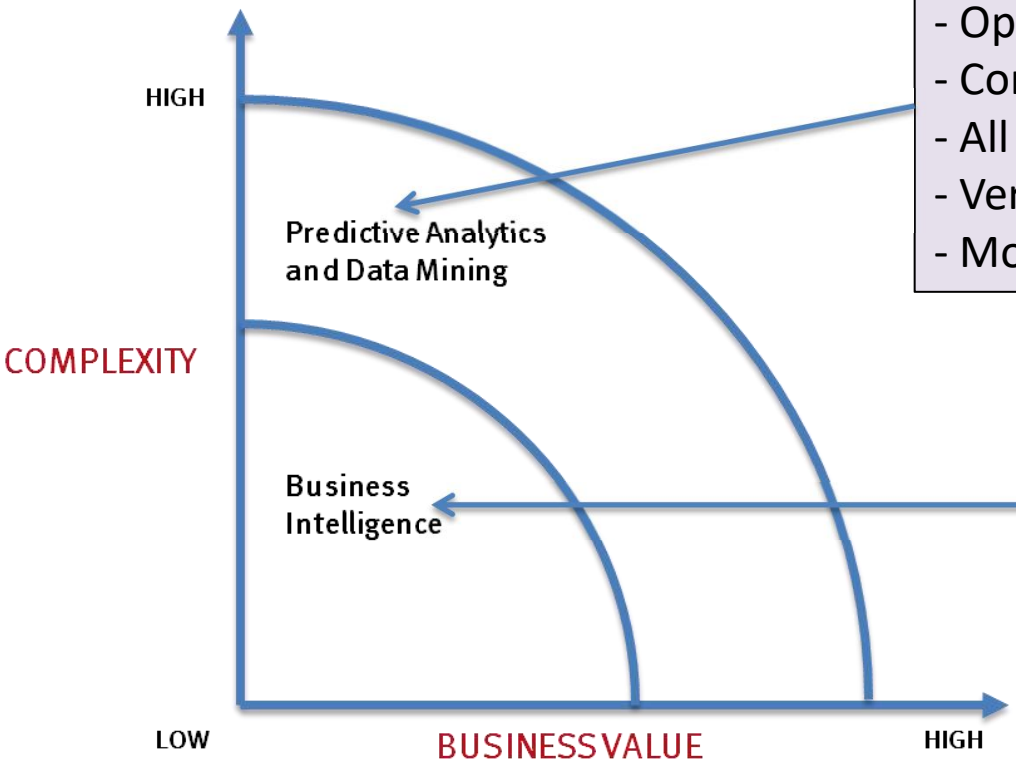
The Model Has Changed…

- **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

What's driving Big Data to Analytics



- Optimizations and predictive analytics
- Complex statistical analysis
- All types of data, and many sources
- Very large datasets
- More of a real-time

- Ad-hoc querying and reporting
- Data mining techniques
- Structured data, typical sources
- Small to mid-size datasets

Structuring Big Data

- In simple terms, is arranging the available data in a manner such that it becomes easy to study, analyze, and derive conclusion format.
- ***Why is structuring required?***

In our daily life, you may have come across questions like,

- How do I use to my advantage the vast amount of data and information I come accross?
- Which news articles should I read of the thousands I come accross?
- How do I choose a book of the millions available on my favourate sites or stores?
- How do I keep myself updated about new events, sports, inventions, and discoveries taking place across the globe?

Today, solution to such questions can be found by information processing systems.

Types of Data

- Data that comes from multiple sources, such as databases, ERP systems, weblogs, chat history, and GPS maps so varies in format. But primarily data is obtained from following types of data sources.
- ***Internal Sources : Organisational or enterprise data***
  - ***CRM, ERP, OLTP, products and sales data..... (Structured data)***
- ***External sources: Social Data***
  - ***Business partners, Internet, Government, Data supliers..... (Unstructured or unorganised data)***

Types of Data (cont..)

- On the basis of the data received from the source mentioned, big data is comprises;
  - **Structure Data**
  - **Unstructured Data**
  - **Semi-structured Data**

BIG DATA = **Structure Data** + **Unstructure Data** + **Semi-structure Data**

Structure Data

- It can be defined as the data that has a defined repeating pattern.
- This pattern makes it easier for any program to sort, read, and process the data.
- Processing structured data is much faster and easier than processing data without any specific repeating pattern.

Structure Data (cont..)

- Is organised data in a prescribed format.
- Is stored in tabular form.
- Is the data that resides in fixed fields within a record or file.
- Is formatted data that has eities and their attributes are properly mapped.
- Is used in query and report against predetermined data types.

- Sources: DBMS/RDBMS, Flat files, Multidimensional databases, Legacy databases

Structure Data (cont..)

Structured Data at a Glance

Characteristics of Structured Data

- High organized
- Clearly defined
- Easy to access
- Easy to analyze

Examples of Structured Data

- Name
- Age
- Gender
- Address
- Phone number
- Currency
- Dates
- Billing info

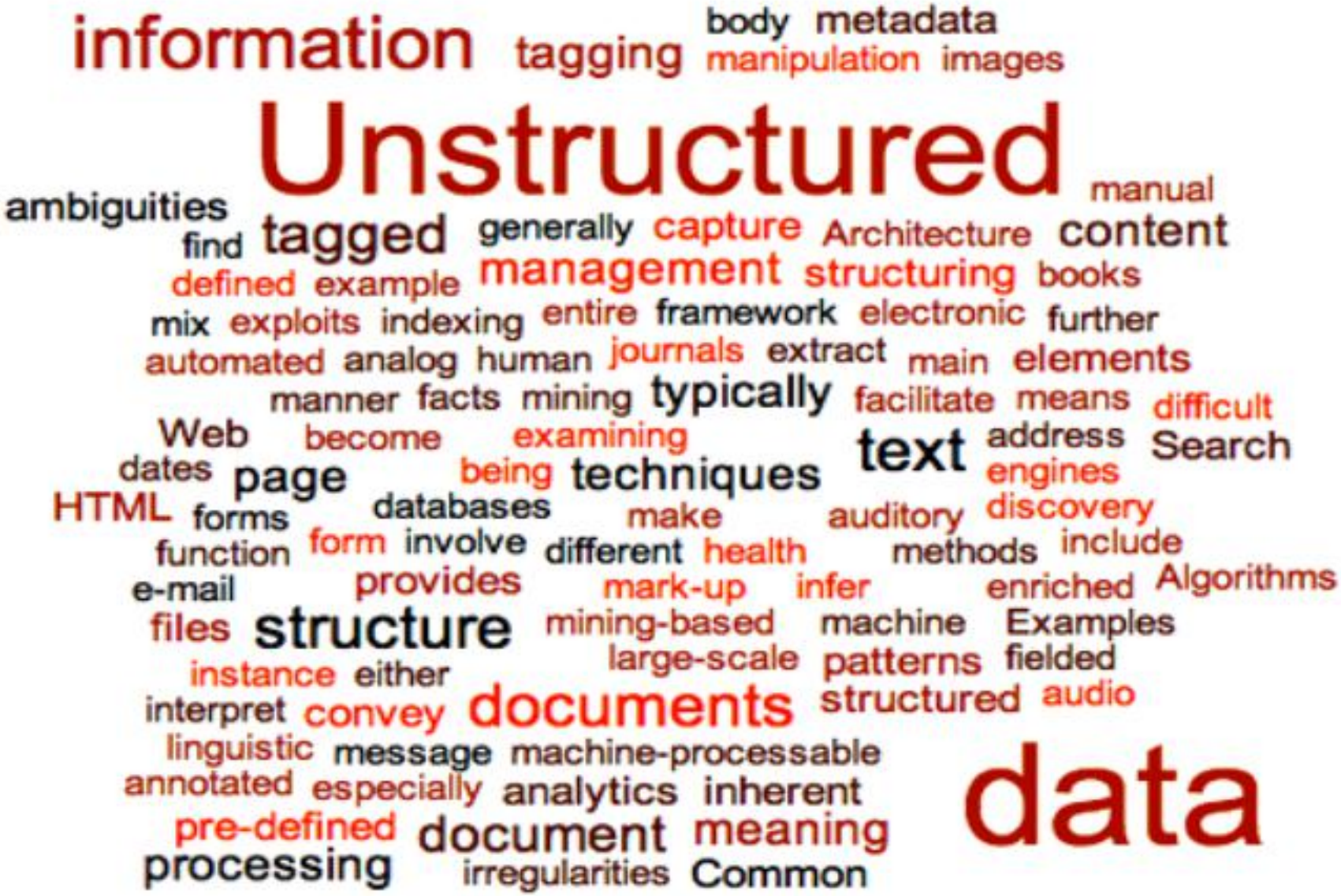
Sources of Structured Data

- SQL databases
- Spreadsheets
- Sensors
- Medical Devices
- Online Forms
- Point of Sales Systems
- Web and Server Logs

Unstructure Data

- It is a set of data that might or might not have any logical or repeating patterns.
- Typically of metadata, i.e, the additional information related to data.
- Inconsistent data (files, social media websites, satalities, etc.)
- Data in different format (e-mails, text, audio, video or images.
- **Sources: Social media, Mobile Data, Text both internal & external to an organzation**

Where Does Unstructured Data Come From?



Differences in Approaches to Measuring Attitudes

	Structured	Unstructured
Intent	• Data generated to measure specific construct	• Data not generated to measure specific construct
Source	How satisfied are you with company? Extremely Dissatisfied — Neither Satisfied Nor Dissatisfied — Extremely Satisfied 0 1 2 3 4 5 6 7 8 9 10	emails, social media, support calls, movie reviews, tweet content, transcripts of comments
Score	• Customer-generated	• Algorithm-generated

Structured vs. Unstructured Data

Structured Data

- Data that resides in a fixed field within a record or file
- Ex: data in a database table
- Easy to enter, store, and analyze

Unstructured Data

- Does not reside in a traditional database
- Ex: e-mail, videos, audio files, web pages, presentations
- Difficult and costly to analyze

Data Definition Framework

Data Format

Data Source	Structured	Unstructured
Internal	**Human-Generated** • Survey ratings • Aptitude testing **Machine-Generated** • Web metrics from Web logs • Product purchase from sales Records • Process control measures	**Human-Generated** • Emails, letters, text messages • Audio transcripts • Customer comments • Voicemails • Corporate video/communications • Pictures, illustrations • Employee reviews
External	**Human-Generated** • Number of Retweets, Facebook likes, Google Plus +1s • Ratings on Yelp • Patient ratings ratings **Machine-Generated** • GPS for tweets • Time of tweet/updates/postings	**Human-Generated** • Content of social media updates • Comments in online forums • Comments on Yelp • Video reviews • Pinterest images • Surveillance video