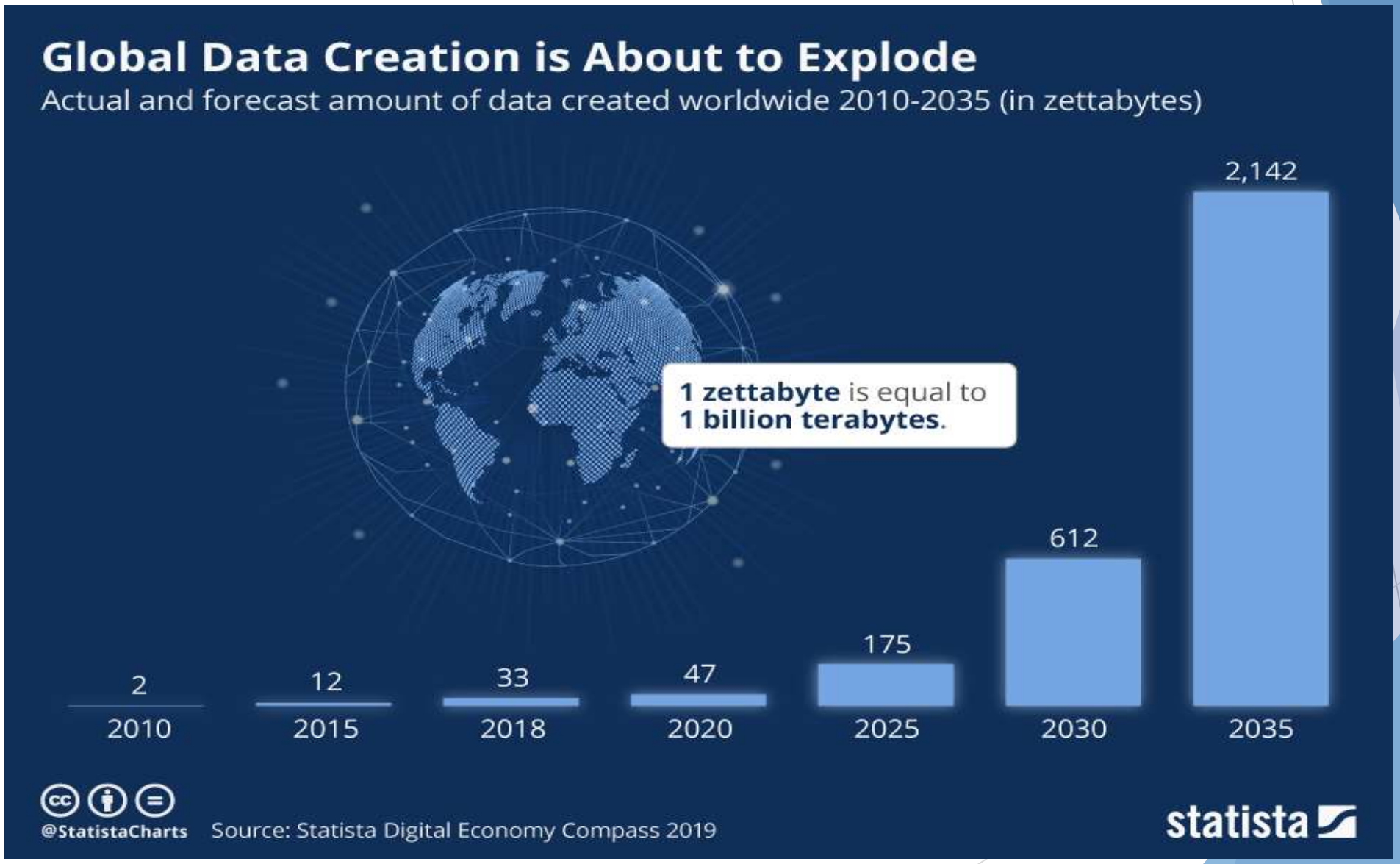
Introduction to Information Retrieval

UNIT-1

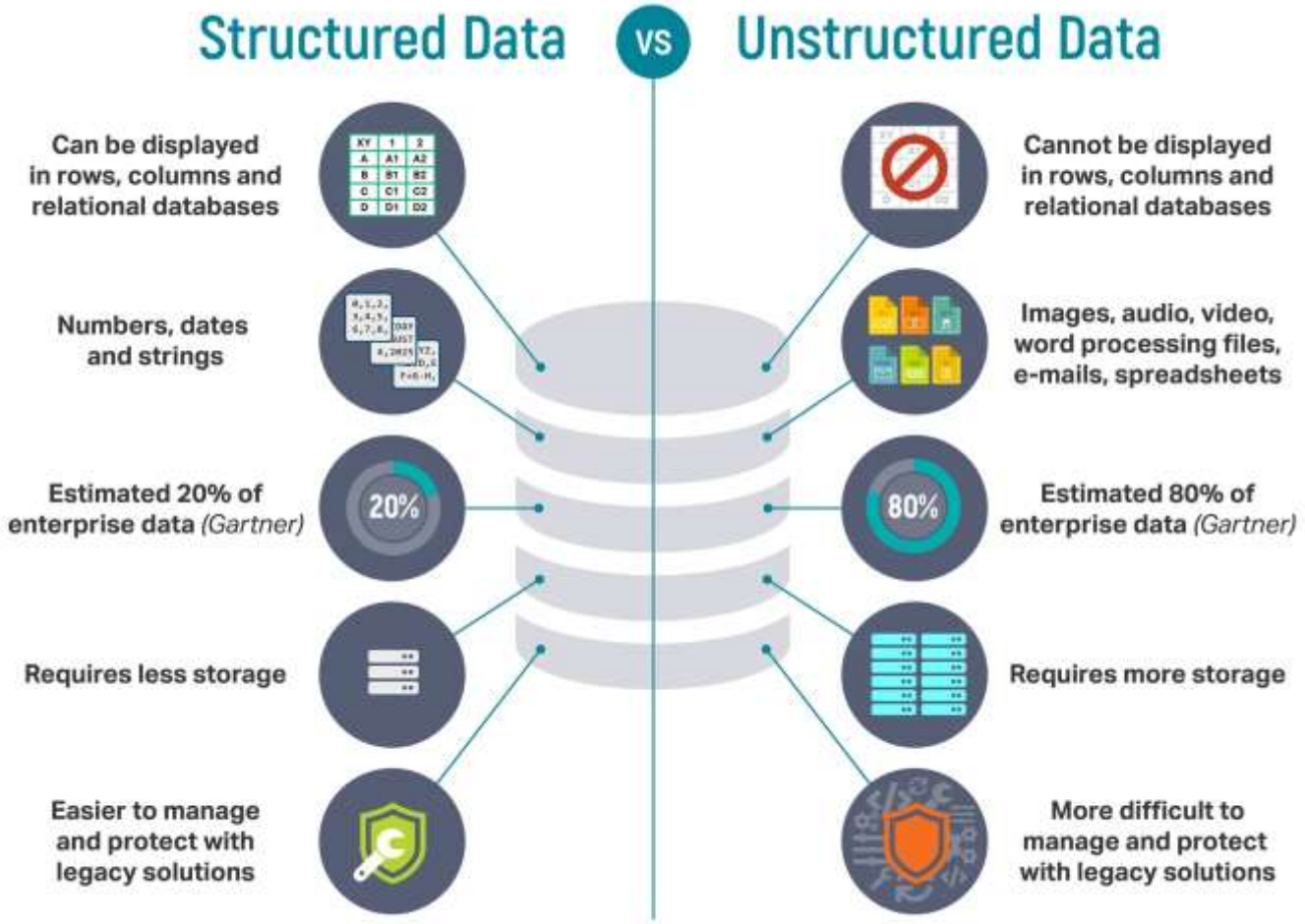
What is Information Retrieval?

- Information retrieval (IR) is the process of obtaining information from a large repository or database, typically in the form of documents or data, that is relevant to a user's query or information need.
- Information Retrieval (IR) is finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers).
- These days we frequently think first of web search, but there are many other cases:
  - E-mail search
  - Searching your laptop
  - Corporate knowledge bases
  - Legal information retrieval

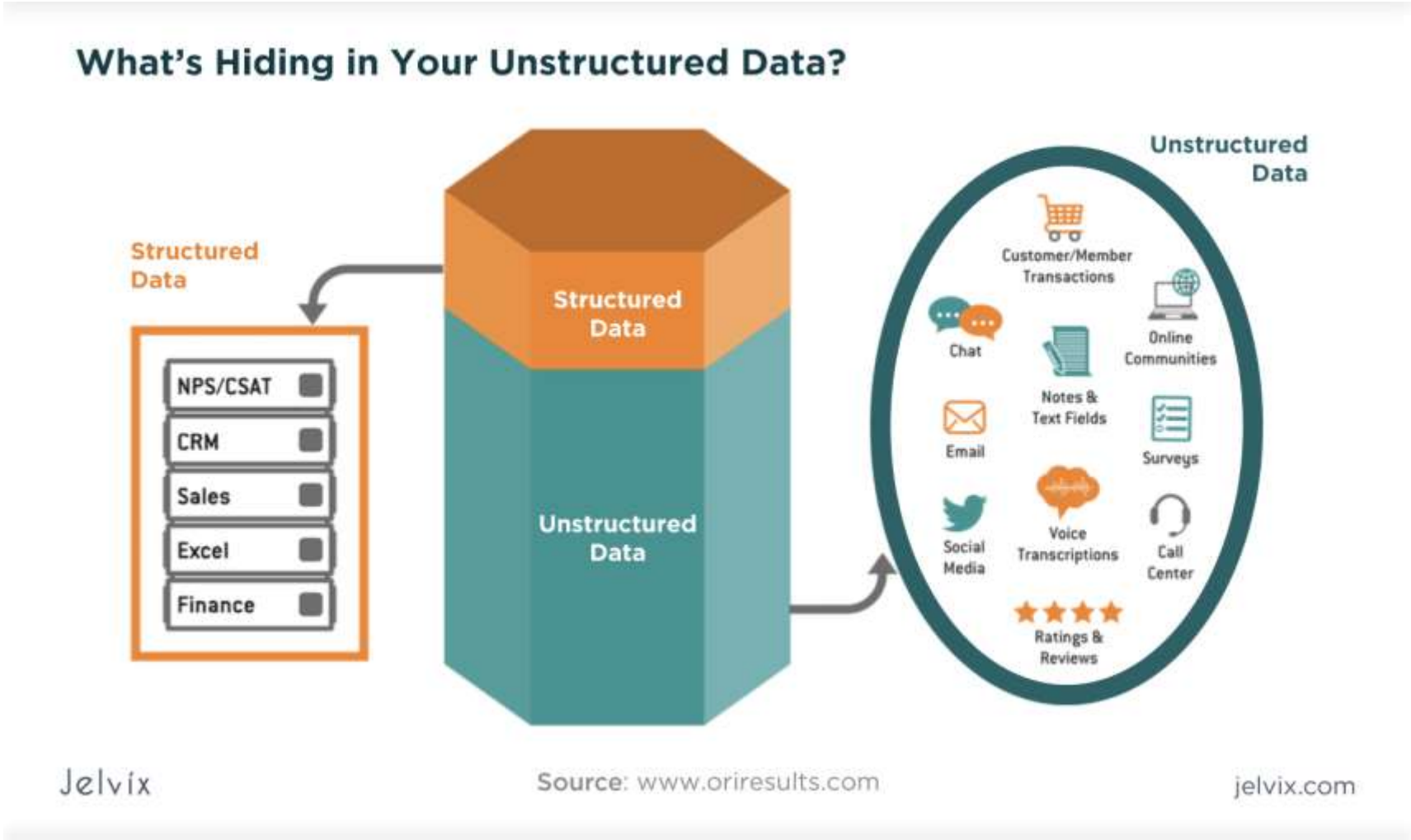
Growth in Data



Unstructured (text) vs. structured (database) data



Example of Structured Data and Unstructured Data



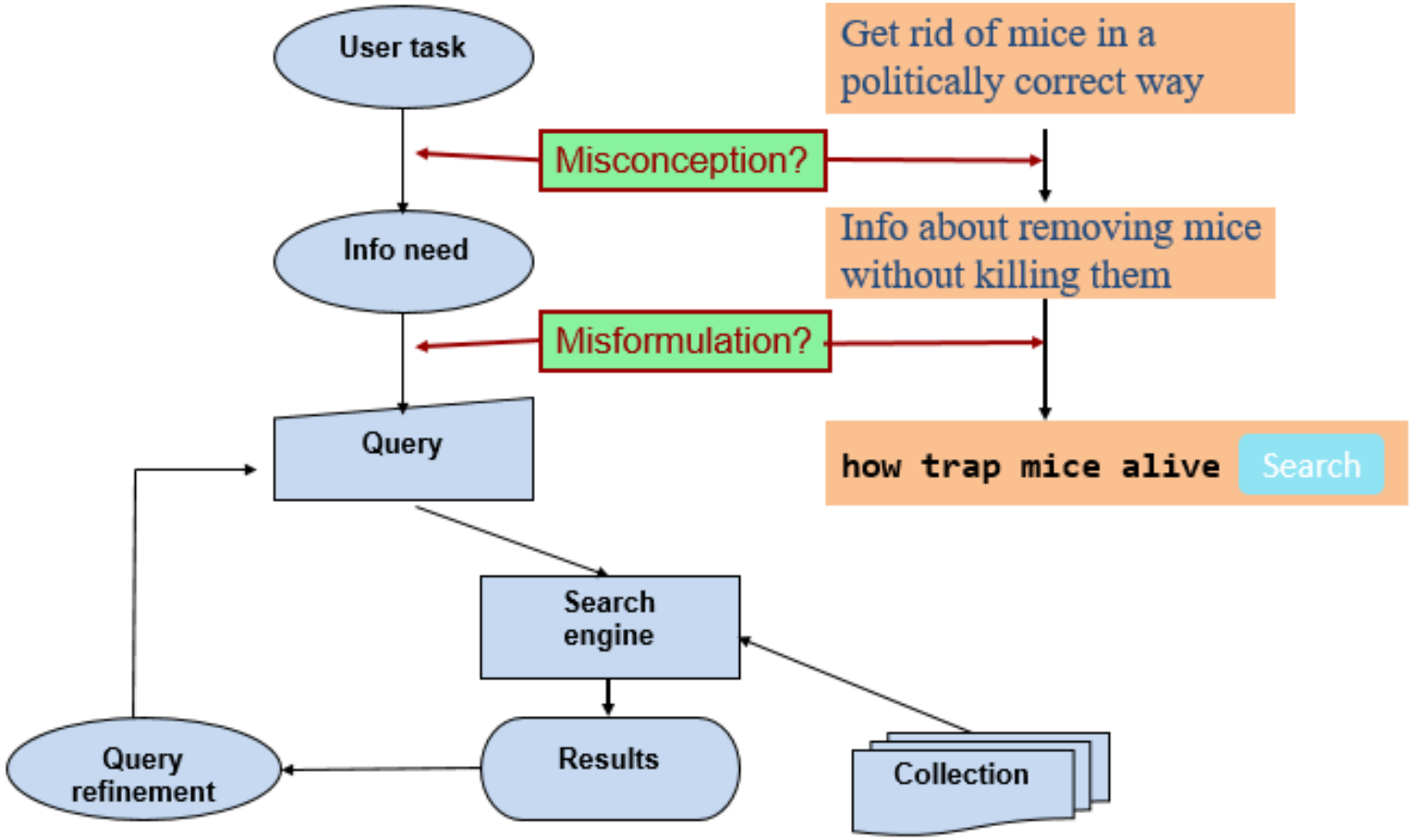
Semi-Structure Data

- Semi-structured data is a hybrid of both structured and unstructured data.
- It has some organizational framework but does not have the complete structure that is required to fit in a relation database.
- Semi-structure data has a self-describing structure that contains tags or attributes to separate various entities within data.
- Example: XML data.

Basic assumptions of Information Retrieval

- Collection: A set of documents
  - Assume it is a static collection for the moment
- Goal: Retrieve documents with information that is relevant to the user's information need and helps the user complete a task

The classic search model



How good are the retrieved docs?

- Precision : Fraction of retrieved docs that are relevant to the user's information need
- Recall : Fraction of relevant docs in collection that are retrieved

- More precise definitions and measurements to follow later

Example of Information Retrieval Problem

- Which plays of Shakespeare contain the words Brutus AND Caesar but NOT Calpurnia?
- One could grep all of Shakespeare's plays for Brutus and Caesar, then strip out lines containing Calpurnia?
- Why is that not the answer?
  - Slow (for large corpora)
  - NOT Calpurnia is non-trivial
  - Other operations (e.g., find the word Romans near countrymen) not feasible
  - Ranked retrieval (best documents to return)
  - Later lectures

Term-document incidence matrices

	Antony and Cleopatra	Julius Caesar	The Tempest	Hamlet	Othello	Macbeth
Antony	1	1	0	0	0	1
Brutus	1	1	0	1	0	0
Caesar	1	1	0	1	1	1
Calpurnia	0	1	0	0	0	0
Cleopatra	1	0	0	0	0	0
mercy	1	0	1	1	1	1
worser	1	0	1	1	1	0

1 if play contains word, 0 otherwise

Brutus** AND **Caesar** BUT NOT **Calpurnia

Incidence vectors

- So we have a 0/1 vector for each term.
- To answer query: take the vectors for Brutus, Caesar and Calpurnia (complemented) ▫ bitwise AND.
  - 110100 AND
  - 110111 AND
  - 101111 =
  - **100100**

	Antony and Cleopatra	Julius Caesar	The Tempest	Hamlet	Othello	Macbeth
Antony	1	1	0	0	0	1
Brutus	1	1	0	1	0	0
Caesar	1	1	0	1	1	1
Calpurnia	0	1	0	0	0	0
Cleopatra	1	0	0	0	0	0
mercy	1	0	1	1	1	1
worser	1	0	1	1	1	0

Answers to query

- Antony and Cleopatra, Act III, Scene ii
- *Agrippa [Aside to DOMITIUS ENOBARBUS]: Why, Enobarbus,*
- *When Antony found Julius Caesar dead,*
- *He cried almost to roaring; and he wept*
- *When at Philippi he found Brutus slain.*

- Hamlet, Act III, Scene ii
- *Lord Polonius: I did enact Julius Caesar I was killed i' the*
- *Capitol; Brutus killed me.*

Bigger collections

- Consider N = 1 million documents, each with about 1000 words.
- Avg 6 bytes/word including spaces/punctuation
- 6GB of data in the documents.
- Say there are M = 500K distinct terms among these.

Can’t build the matrix

- 500K x 1M matrix has half-a-trillion 0’s and 1’s.
- But it has no more than one billion 1’s. (Why?)
  - matrix is extremely sparse.
- What’s a better representation?
  - We only record the 1 positions.