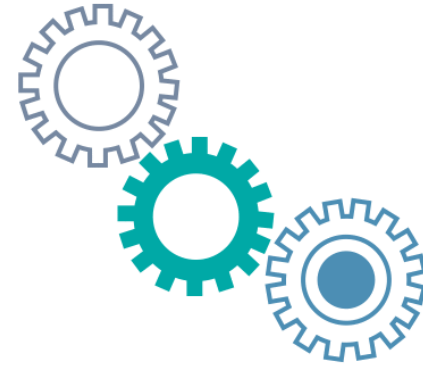


Introduction to Data Analytics

Data Analytics



Module – 1 Syllabus

- **Data-Information**
- **characteristics of data**
- **data munging**
- **Scraping**
- **Sampling**
- **Cleaning**
- **importance of data analytics**

Data Analysis

- Data analysis is a process of obtaining raw data and converting it into information useful for decision-making by users.



Data Analytics



- Data Analytics the science of examining raw data with the purpose of drawing conclusions about that information

The background of the slide features a series of thin, curved lines in light gray and white, creating a sense of motion and depth. These lines are more prominent on the left side and fade towards the right.

Applications of Analytics

- In commercial industries, to enable organizations to make more-informed business decisions and by scientists
- By researchers, to verify or disprove scientific models, theories and hypotheses.



Analysis vs Analytics

- Data analytics is a broader term and includes data analysis as necessary subcomponent.
- Analytics defines the science behind the analysis.

A red speech bubble graphic with a white outline, containing the text 'Analysis vs Analytics'. The bubble has a tail pointing towards the bottom left.

Analysis vs Analytics

- The science means understanding the cognitive processes an analyst uses to understand problems and explore data in meaningful ways.
- Analytics also include data extract, transform, and load; specific tools, techniques, and methods; and how to successfully communicate results.

Data Science???



(((Josh Wills)))

@josh_wills



Follow

Data Scientist (n.): Person who is better at statistics than any software engineer and better at software engineering than any statistician.



scott vokes

@silentbicycle

scott vokes



Follow

"What is a 'Data Scientist'? An analyst who lives in California." -

@edmundjackson #clojure_conj

7:30 PM - 16 Nov 2012



19

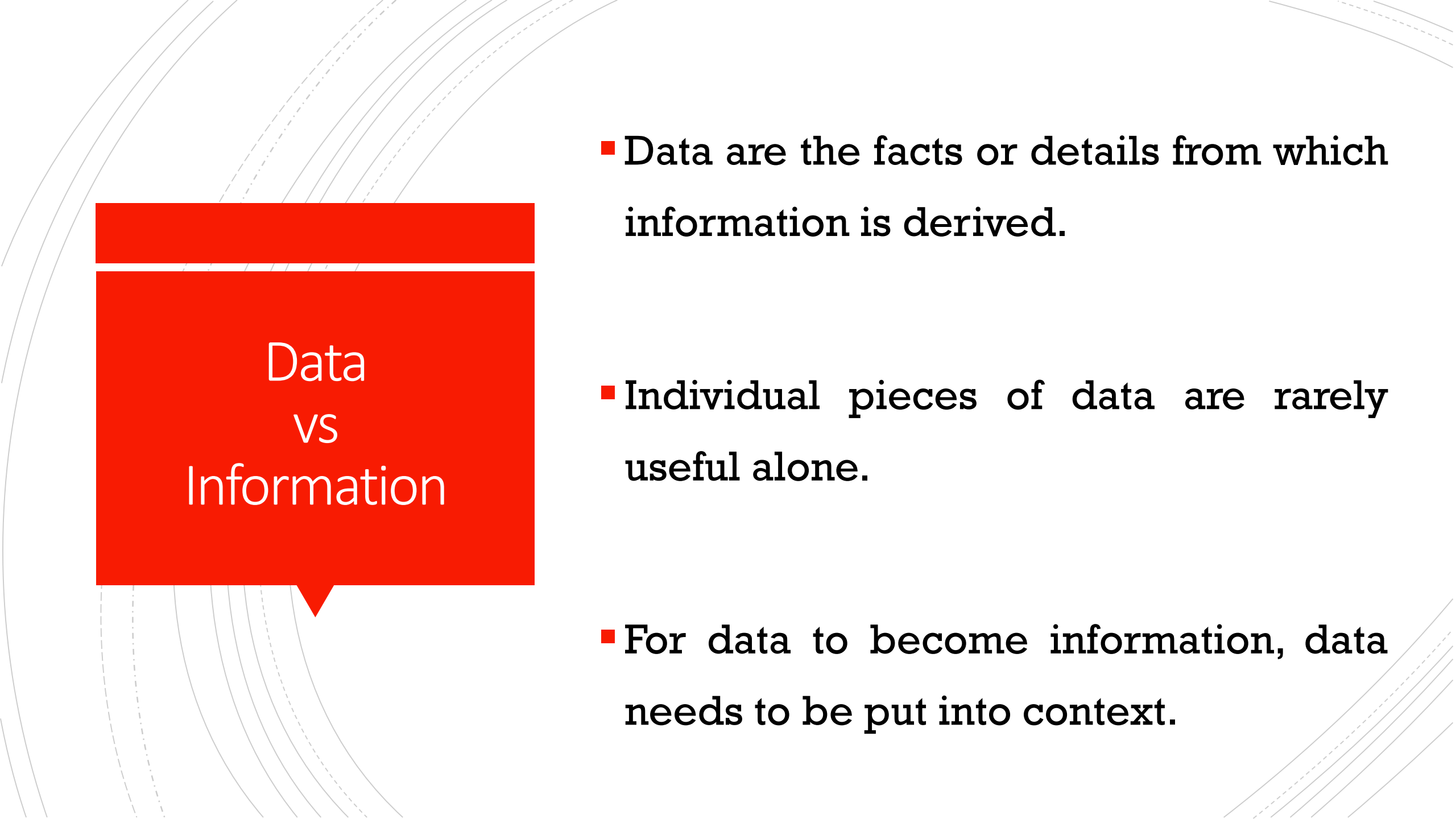


8



Data Science

- Dealing with unstructured and structured data, Data Science is a field that comprises of everything that related to data cleansing, preparation, and analysis.
- Involves in creation of new algorithms



Data vs Information

- Data are the facts or details from which information is derived.
- Individual pieces of data are rarely useful alone.
- For data to become information, data needs to be put into context.

Data vs Information

Comparison Chart

| BASIS FOR COMPARISON | DATA | INFORMATION |
|----------------------|---|---|
| Meaning | Data means raw facts gathered about someone or something, which is bare and random. | Facts, concerning a particular event or subject, which are refined by processing is called information. |
| What is it? | It is just text and numbers. | It is refined data. |
| Based on | Records and Observations | Analysis |
| Form | Unorganized | Organized |
| Useful | May or may not be useful. | Always |
| Specific | No | Yes |
| Dependency | Does not depend on information. | Without data, information cannot be processed. |

The background of the slide features a series of light gray, concentric curved lines that sweep across the frame, creating a sense of motion and depth. On the left side, there is a prominent red speech bubble with a tail pointing towards the bottom left. Inside this bubble, the text 'Characteristics of data' is written in white. To the right of the bubble, the main content of the slide is presented in black text. At the top right, a heading states 'The seven characteristics that define data quality are:'. Below this heading is a numbered list of seven items, each starting with a red number followed by a period. The list items are: 1. Accuracy and Precision, 2. Legitimacy and Validity, 3. Reliability and Consistency, 4. Timeliness and Relevance, 5. Completeness and Comprehensiveness, 6. Availability and Accessibility, and 7. Granularity and Uniqueness.

Characteristics of data

The seven characteristics that define data quality are:

1. Accuracy and Precision
2. Legitimacy and Validity
3. Reliability and Consistency
4. Timeliness and Relevance
5. Completeness and Comprehensiveness
6. Availability and Accessibility
7. Granularity and Uniqueness



Characteristics of data

Accuracy and Precision: This characteristic refers to the exactness of the data.

Legitimacy and Validity: Requirements governing data set the boundaries of this characteristic.



Characteristics of data

Reliability and Consistency: Regardless of what source collected the data or where it resides, it cannot contradict a value residing in a different source or collected by a different system.

Timeliness and Relevance: Data collected too soon or too late could misrepresent a situation and drive inaccurate decisions.



Characteristics of data

Completeness and Comprehensiveness:
Incomplete data is as dangerous as inaccurate data.

Availability and Accessibility: This presumes that the data exists and is available for access to be granted.



Characteristics of data

Completeness and Comprehensiveness: Incomplete data is as dangerous as inaccurate data.

Availability and Accessibility: This presumes that the data exists and is available for access to be granted.

Granularity and Uniqueness: The level of detail at which data is collected is important, because confusion and inaccurate decisions can otherwise occur.