

DATA:

In computing, data is information that has been translated into a form that is efficient for movement or processing

DATA SCIENCE:

Data science is an evolutionary extension of statistics capable of dealing with the massive amounts of data produced today. It adds methods from computer science to the repertoire of statistics.

BENEFITS AND USES OF DATA SCIENCE:

Data science and big data are used almost everywhere in both commercial and non-commercial Settings

Commercial companies in almost every industry use data science and big data to gain insights into their customers, processes, staff, completion, and products.

- Many companies use data science to offer customers a better user experience, as well as to cross-sell, up-sell, and personalize their offerings.
- Governmental organizations are also aware of data's value. Many governmental organizations not only rely on internal data scientists to discover valuable information, but also share their data with the public.
- Nongovernmental organizations (NGOs) use it to raise money and defend their causes.
- Universities use data science in their research but also to enhance the study experience of their students. The rise of massive open online courses (MOOC) produces a lot of data, which allows universities to study how this type of learning can complement traditional classes.

FACETS OF DATA

In data science and big data you'll come across many different types of data, and each of them tends to require different tools and techniques. The main categories of data are these:

- Structured
- Unstructured
- Natural language
- Machine-generated
- Graph-based
- Audio, video, and images
- Streaming

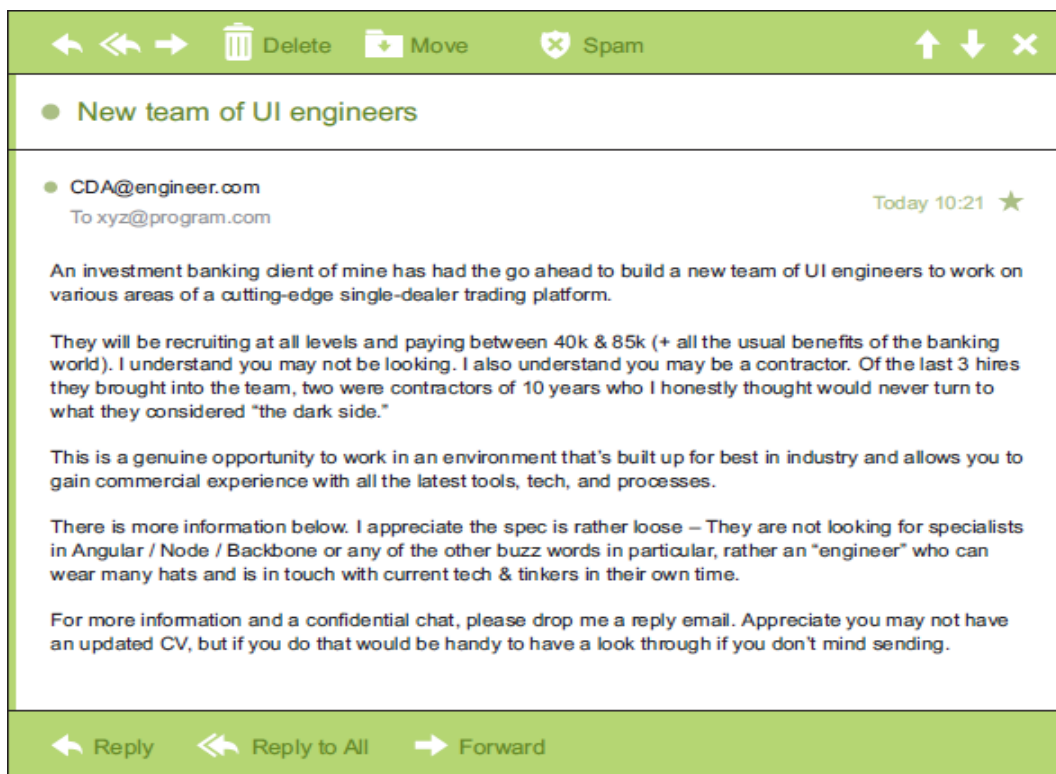
Structured data

- Structured data is data that depends on a data model and resides in a fixed field within a record. As such, it's often easy to store structured data in tables within databases or Excel files

- SQL, or Structured Query Language, is the preferred way to manage and query data that resides in databases.

Indicator ID	Dimension List	Timeframe	Numeric Value	Missing Value Flag	Confidence Int
214390830	Total (Age-adjusted)	2008	74.6%		73.8%
214390833	Aged 18-44 years	2008	59.4%		58.0%
214390831	Aged 18-24 years	2008	37.4%		34.6%
214390832	Aged 25-44 years	2008	66.9%		65.5%
214390836	Aged 45-64 years	2008	88.6%		87.7%
214390834	Aged 45-54 years	2008	86.3%		85.1%
214390835	Aged 55-64 years	2008	91.5%		90.4%
214390840	Aged 65 years and over	2008	94.6%		93.8%
214390837	Aged 65-74 years	2008	93.6%		92.4%
214390838	Aged 75-84 years	2008	95.6%		94.4%
214390839	Aged 85 years and over	2008	96.0%		94.0%
214390841	Male (Age-adjusted)	2008	72.2%		71.1%
214390842	Female (Age-adjusted)	2008	76.8%		75.9%
214390843	White only (Age-adjusted)	2008	73.8%		72.9%
214390844	Black or African American only (Age-adjusted)	2008	77.0%		75.0%
214390845	American Indian or Alaska Native only (Age-adjusted)	2008	66.5%		57.1%
214390846	Asian only (Age-adjusted)	2008	80.5%		77.7%
214390847	Native Hawaiian or Other Pacific Islander only (Age-adjusted)	2008	DSU		
214390848	2 or more races (Age-adjusted)	2008	75.6%		69.6%

Unstructured data



Unstructured data is data that isn't easy to fit into a data model because the content is context-specific or varying. One example of unstructured data is your regular email.

Natural language

- Natural language is a special type of unstructured data; it's challenging to process because it requires knowledge of specific data science techniques and linguistics.

- The natural language processing community has had success in entity recognition, topic recognition, summarization, text completion, and sentiment analysis, but models trained in one domain don't generalize well to other domains.
- Even state-of-the-art techniques aren't able to decipher the meaning of every piece of text.

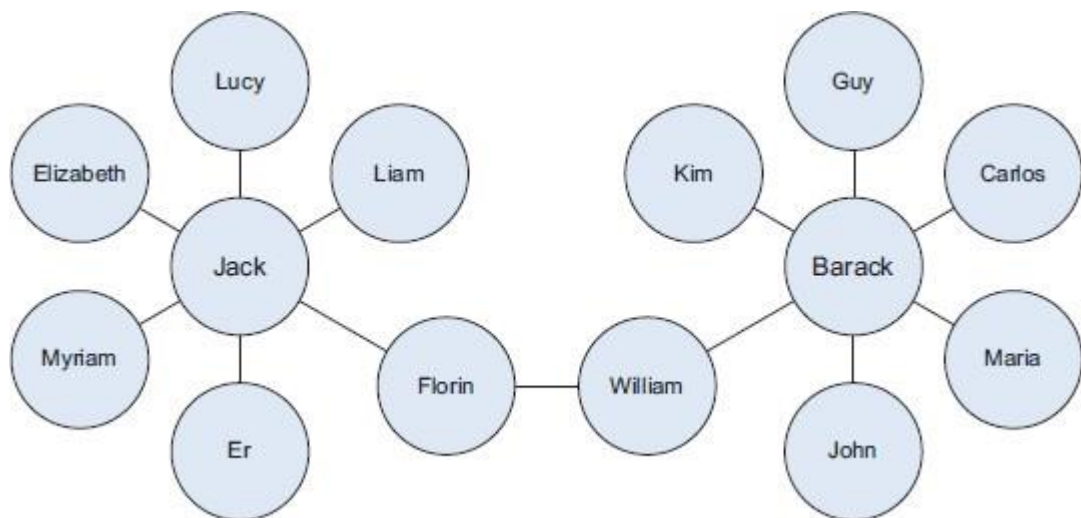
Machine-generated data

- Machine-generated data is information that's automatically created by a computer, process, application, or other machine without human intervention.
- Machine-generated data is becoming a major data resource and will continue to do so.
- The analysis of machine data relies on highly scalable tools, due to its high volume and speed. Examples of machine data are web server logs, call detail records, network event logs, and telemetry.

```

CSIPERF:TXCOMMIT:313236
2014-11-28 11:36:13, Info          CSI    00000153 Creating NT transaction (seq
69), objectname [6]"(null)"
2014-11-28 11:36:13, Info          CSI    00000154 Created NT transaction (seq 69)
result 0x00000000, handle @0x4e54
2014-11-28 11:36:13, Info          CSI    00000155@2014/11/28:10:36:13.471
Beginning NT transaction commit...
2014-11-28 11:36:13, Info          CSI    00000156@2014/11/28:10:36:13.705 CSI perf
    
```

Graph-based or network data



- "Graph data" can be a confusing term because any data can be shown in a graph.
- Graph or network data is, in short, data that focuses on the relationship or adjacency of objects.
- The graph structures use nodes, edges, and properties to represent and store graphical data.
- Graph-based data is a natural way to represent social networks, and its structure allows you to calculate specific metrics such as the influence of a person and the shortest path between two people.

Audio, image, and video

- Audio, image, and video are data types that pose specific challenges to a data scientist.
- Tasks that are trivial for humans, such as recognizing objects in pictures, turn out to be challenging for computers.
- MLBAM (Major League Baseball Advanced Media) announced in 2014 that they'll increase video capture to approximately 7 TB per game for the purpose of live, in-game analytics.
- Recently a company called DeepMind succeeded at creating an algorithm that's capable of learning how to play video games.
- This algorithm takes the video screen as input and learns to interpret everything via a complex process of deep learning.

Streaming data

- The data flows into the system when an event happens instead of being loaded into a data store in a batch.
- Examples are the "What's trending" on Twitter, live sporting or music events, and the stock market