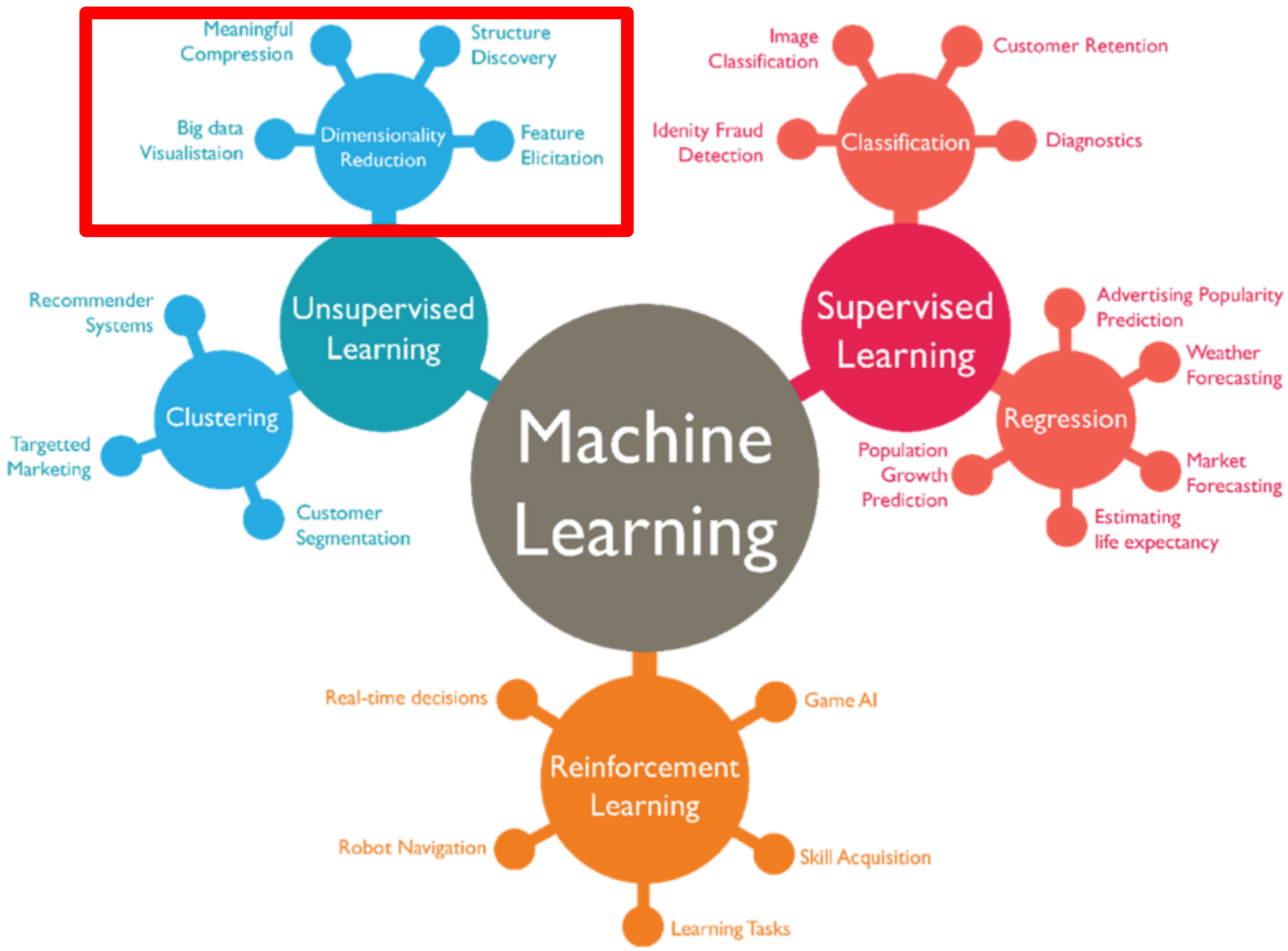


Unsupervised Learning: Topic Modeling



Meaningful Compression

Structure Discovery

Dimensionality Reduction

Feature Elicitation

Big data Visualistaion

Image Classification

Customer Retention

Classification

Diagnostics

Identity Fraud Detection

Unsupervised Learning

Clustering

Recommender Systems

Targetted Marketing

Customer Segmentation

Supervised Learning

Regression

Advertising Popularity Prediction

Weather Forecasting

Market Forecasting

Estimating life expectancy

Population Growth Prediction

Reinforcement Learning

Real-time decisions

Game AI

Robot Navigation

Skill Acquisition

Learning Tasks

Documents cover multiple topics

Topics

Topic 1
Basketball
LeBron
NBA
...

Topic 2
NFL
Football
American
...

Topic 3
Trump
President
Clinton
...

Documents

LeBron James says President Trump 'trying to divide through sport'

Basketball star LeBron James has praised the American football players who have protested against Donald Trump, and accused the US president of "using sports to try and divide us".

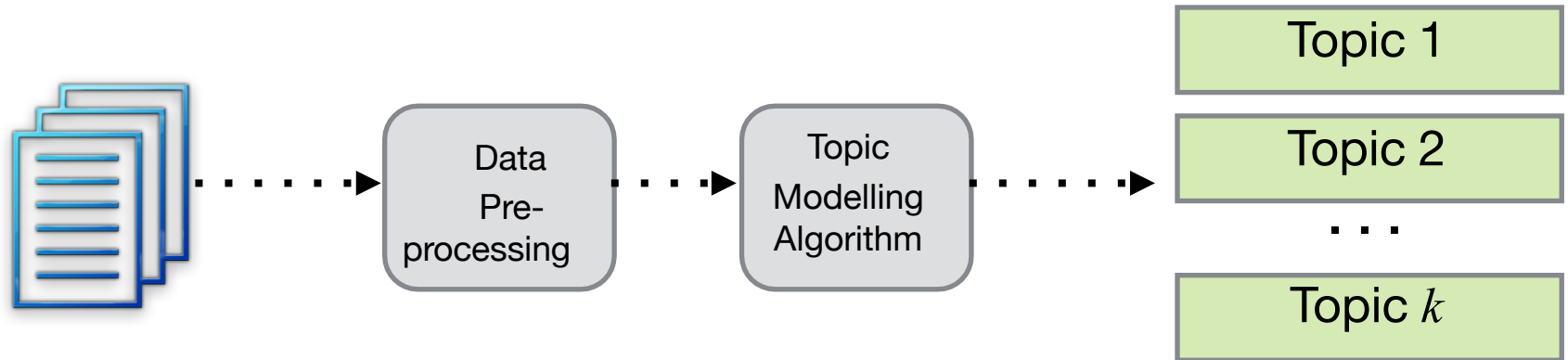
Trump said that NFL players who fail to stand during the national anthem should be sacked or suspended.

James praised the players' unity, and said: "The people run this country."

James, who plays for the Cleveland Cavaliers and has won three NBA championships, campaigned for Hillary Clinton, Trump's rival, during the 2016 presidential election campaign.

A document is composed of terms related to one or more topics.

Topic Modeling



- Topic Modeling **induces** a set of topics from a document collection based on their words
- **Output:** A set of k topics, each of which is represented by
 - A descriptor, based on the top-ranked terms for the topic
 - Associations for documents relative to the topic.

Topic Modeling

- If we want five topics for a set of newswire articles, the topics might correspond to politics, sports, technology, business & entertainment
- Documents are represented as a vector of numbers (between 0.0 & 1.0) indicating how much of each topic it has
- Document similarity is measured by the cosign similarity of their vectors

Document-term matrix

- Given collection of documents, find all the unique words in them
 - Eliminate common stopwords (e.g., the, and, a) that carry little meaning and very infrequent words
- Represent each word as an integer and construct document-term matrix
- Cell values are term frequency (tf), number of times word occurs
- Alternatively: use tf-idf to give less weight to very common words

10,000 words

	W1	W2	W3	<u>W_n</u>
D1	0	2	1	3
D2	1	4	0	0
D3	0	2	3	1
<u>D_n</u>	1	1	3	0

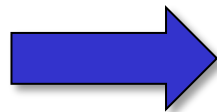
1000 documents

Dimensionality reduction

- A dimensionality-reduction algorithm converts this matrix into the product of two smaller matrices
 - *Documents to topics and topics to words*
- Document represented as a vector of topics
- Understand what K_3 is about by looking at its words with the highest values
- Documents about topic K_3 are those with high values for K_3
- Documents similar to D_{43} will have similar topic vectors (use cosine similarity)

	W1	W2	W3	<u>W_n</u>
D1	0	2	1	3
D2	1	4	0	0
D3	0	2	3	1
<u>D_n</u>	1	1	3	0

n documents x m words



	K1	K2	K3	K
D1	1	0	0	1
D2	1	1	0	0
D3	1	0	0	1
<u>D_n</u>	1	0	1	0

n documents x k topics









x

	W1	W2	W3	<u>W_m</u>
K1	0	1	1	1
K2	1	1	1	0
K3	1	0	0	1
K	1	1	0	0

k topics x m words

Topic modeling with sklearn

See and try the notebooks and data in this [github repo](#)

 data	8 minutes ago
 NMFtm.ipynb	8 minutes ago
 README.md	8 minutes ago
 articles-model-nmf-k10.pkl	8 minutes ago
 articles-raw.pkl	8 minutes ago
 articles-tfidf.pkl	8 minutes ago
 preprocessing.ipynb	8 minutes ago
 stopwords.txt	8 minutes ago

Dimensionality reduction

- There are many dimensionality-reduction algorithms with different properties
- They are also used for word embeddings
- General idea: represent a thing (i.e., document, word, node in a graph) as a relatively short (e.g., 100-300) vector of numbers between 0.0 and 1.0
- Some information lost, but the size is manageable

Topic Modeling Summary

- Topic Modeling is an efficient way for identifying latent topics in a collection of documents
- The topics found are ones that are specific to the collection, which might be social media posts, medical journal articles or cybersecurity alerts
- It can be used to find documents on a topic, for document similarity metrics and other applications