Foundations of Social and Cultural Data Analysis

Dr. Nanne van Noord & Dr. Melvin Wevers

Assignment

- Make sure that the code runs!
  - Restart the kernel and re-run all cells
  - Include git cloning, pip imports, and untar the data!
- Read the assignments carefully!

Each record, then, should be a `list` with four elements: (i) the year of publication, (ii) the title, (iii) the name of the author, and (iv) the name of the publisher.

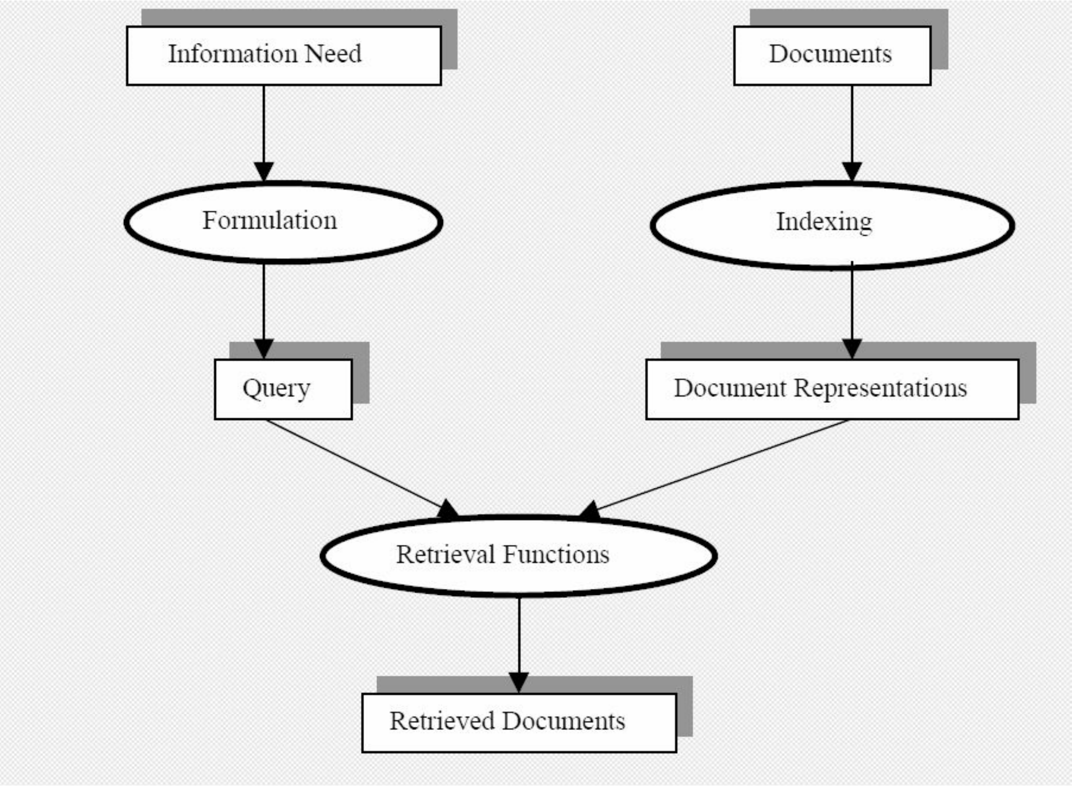
```
('year', 'title', 'author', 'publisher')
('1642', 'Hiervsalem verwoest. Trevrspel.', 'Joost van den Vondel 1587-1679', 'Matthijsz, Paulus Amsterdam')
('1641', "Gysbrecht van Aemstel, d'ondergangh van zijn stad en zijn ballingschap. Treurspel.", 'Joost van den Vondel 1587-1679', 'Houthaeck, Dirck Cornelisz Amsterdam')
('1720', 'Joseph in Egypten. Trevrspel.', 'Joost van den Vondel 1587-1679', 'Oosterwyk, Johannes van Amsterdam')
('17XX', 'Lucifer. Treurspel.', 'Joost van den Vondel 1587-1679', 'Wees, Abraham de I Amsterdam')
```

  - Are there additional questions to be answered?
- Try to avoid redundant or needless code.
  - Functions can be very useful to reduce redundancy and increase readability!

Recap

- Data search/retrieval
- Variable types / measurement scales

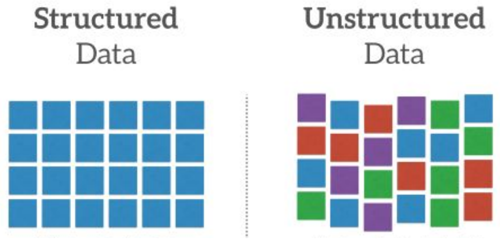
Sharing/Search Relation

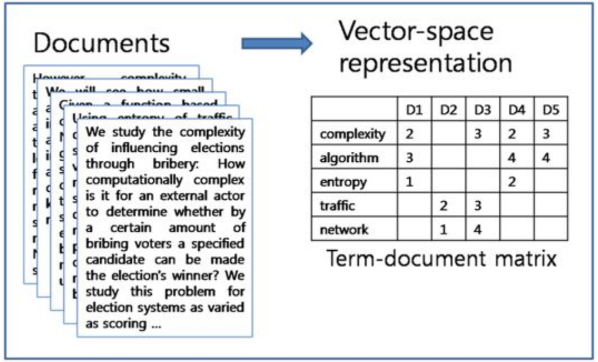


Representation/Encoding

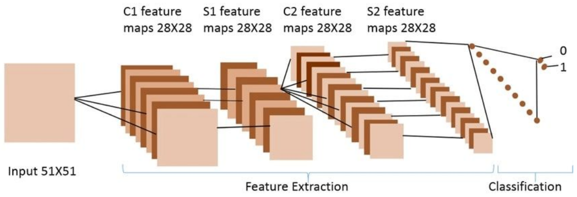
Machine/Human readable is a representation choice

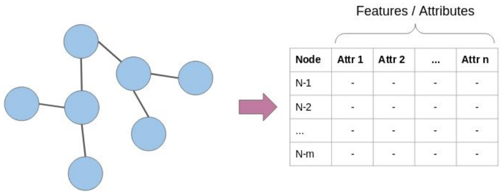
- Possible to have parallel representations





	D1	D2	D3	D4	D5
complexity	2		3	2	3
algorithm	3			4	4
entropy	1			2	
traffic		2	3		
network		1	4		

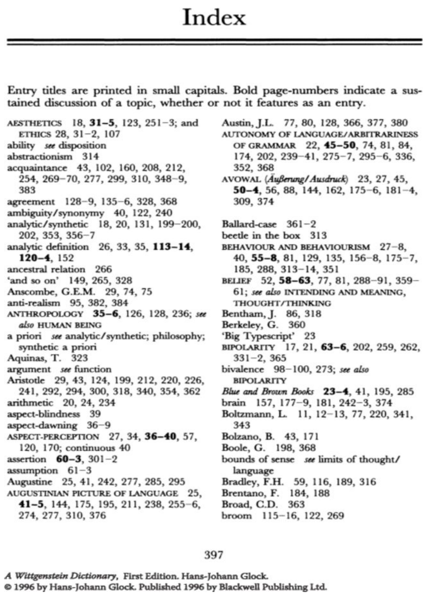




Node	Attr 1	Attr 2	...	Attr n
N-1	-	-	-	-
N-2	-	-	-	-
...	-	-	-	-
N-m	-	-	-	-

Indexing

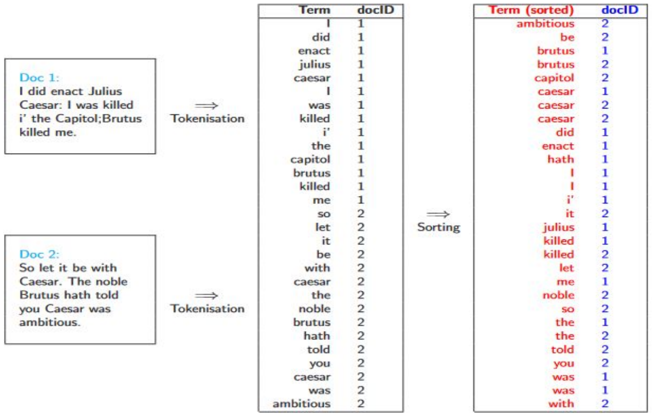
To aid search we may want to construct a representation of the data that is specifically tailored to enable search.



Index

Entry titles are printed in small capitals. Bold page-numbers indicate a sustained discussion of a topic, whether or not it features as an entry.

A Wittgenstein Dictionary, First Edition. Hans-Johann Glock.



Term	docID
I	1
did	1
enact	1
julius	1
caesar	1
I	1
was	1
killed	1
i'	1
the	1
capitol	1
brutus	1
killed	1
me	1
so	2
let	2
it	2
be	2
with	2
caesar	2
the	2
noble	2
brutus	2
hath	2
told	2
you	2
caesar	2
was	2
ambitious	2

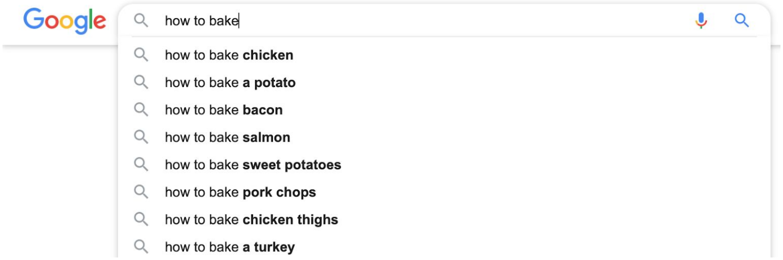
Term (sorted)	docID
ambitious	2
be	2
brutus	1
brutus	2
capitol	2
caesar	1
caesar	2
caesar	2
did	1
enact	1
hath	1
I	1
I	1
i'	1
it	2
julius	1
killed	1
killed	2
let	2
me	1
noble	2
so	2
the	1
the	2
told	2
you	2
was	1
was	1
with	2

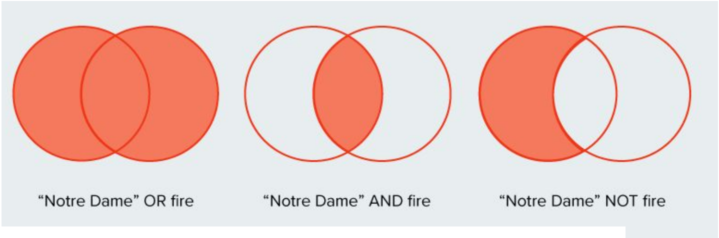
Querying

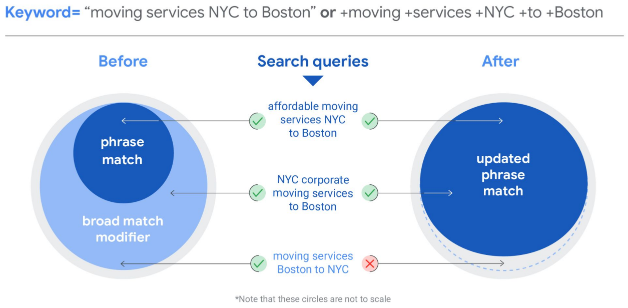
How to enable querying depends on needs and expertise of users.

Various types of queries:

- Keyword queries
  - Most common
  - Keywords implicitly connected by AND
- Boolean queries
  - Allow range of logical operators (AND, OR, NOT)
- Phrase queries
  - Search for exact multi-word match

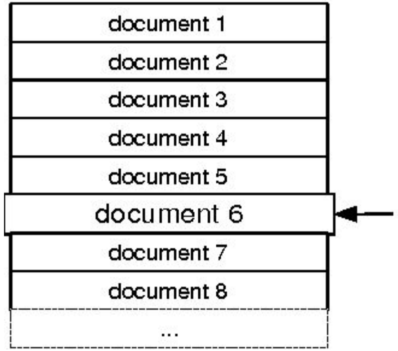


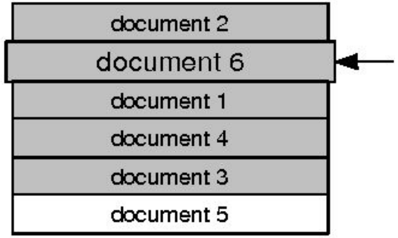


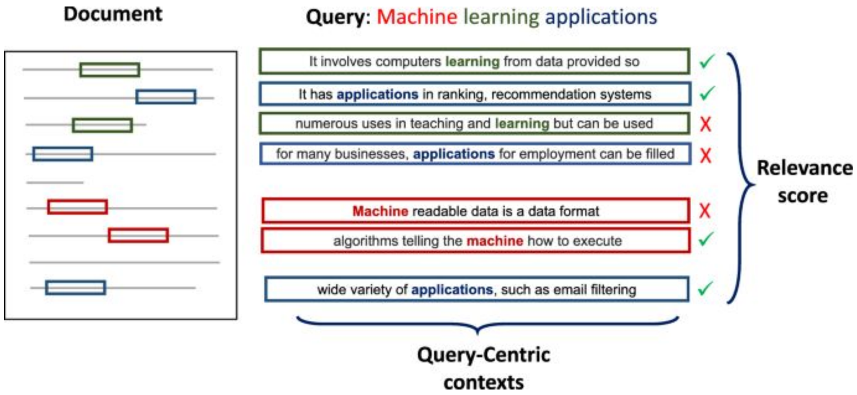


Ranking

Give a score to each indexed document based on query and return in order

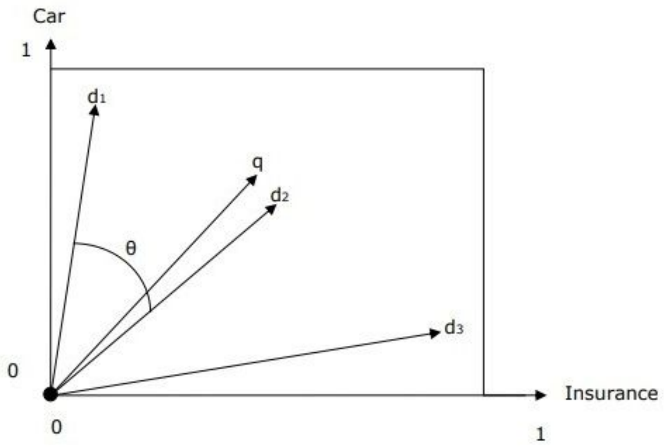




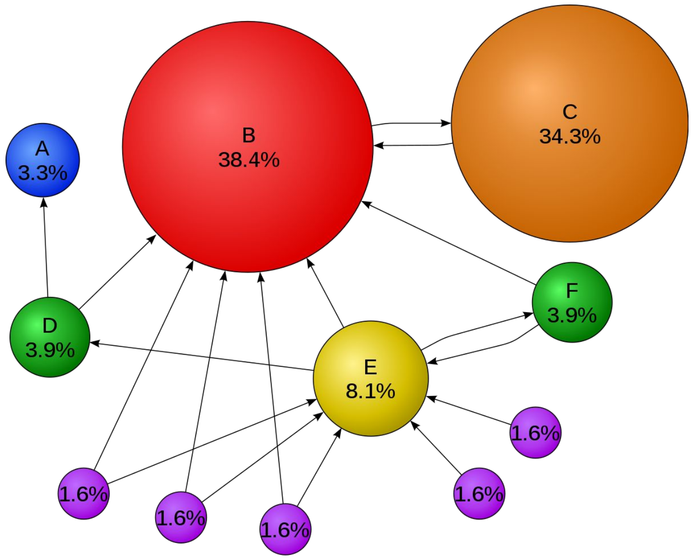


Ranking - Relevance Score

Vector Space Model



PageRank

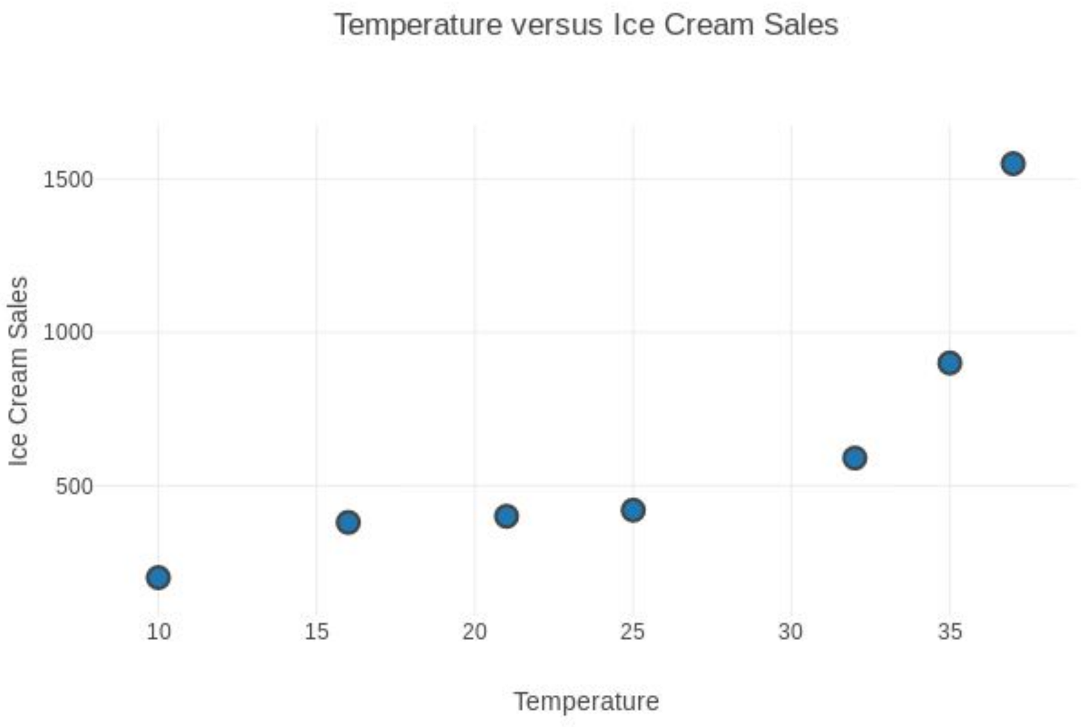


Measurement Scales

- Another aspect of a variable we can use to describe it is the measurement scale
- The measurement scale tells us:
  - How to interpret the values on the scale in relation to the other values
  - How to compare the values

Relations between variables

Association

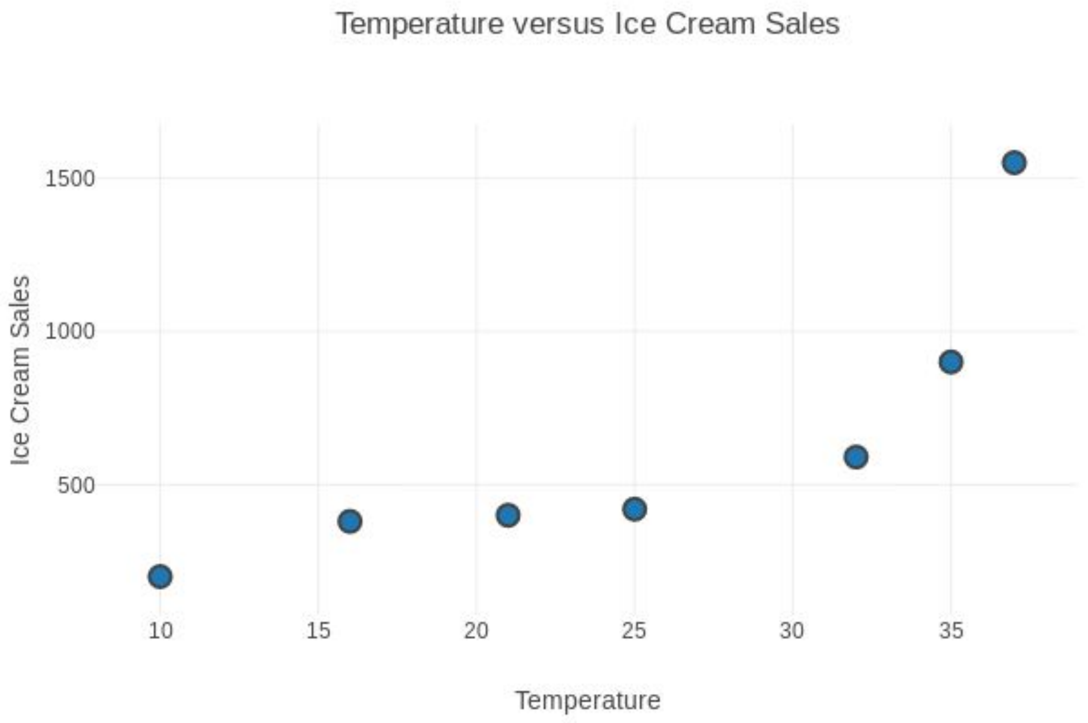


Covariance

- Measure of joint variability of two variables
  - If X_i and Y_i are both high, or both low, then covariance is positive
  - If X_i is high and Y_i is low, or other way around, then covariance is negative
  - Magnitude depends on values of variables
- Covariance:

$$cov(X,Y) = \frac{\sum_{i=1}^{N} (X_i - \mu_X)(Y_i - \mu_Y)}{N}$$

Correlation

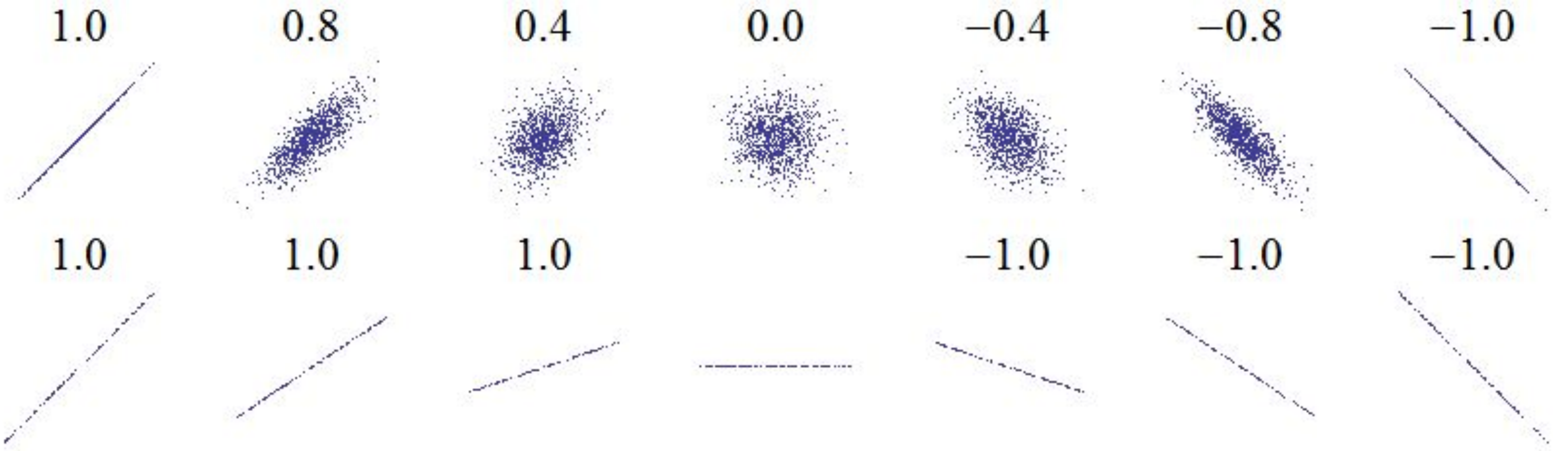


Correlation

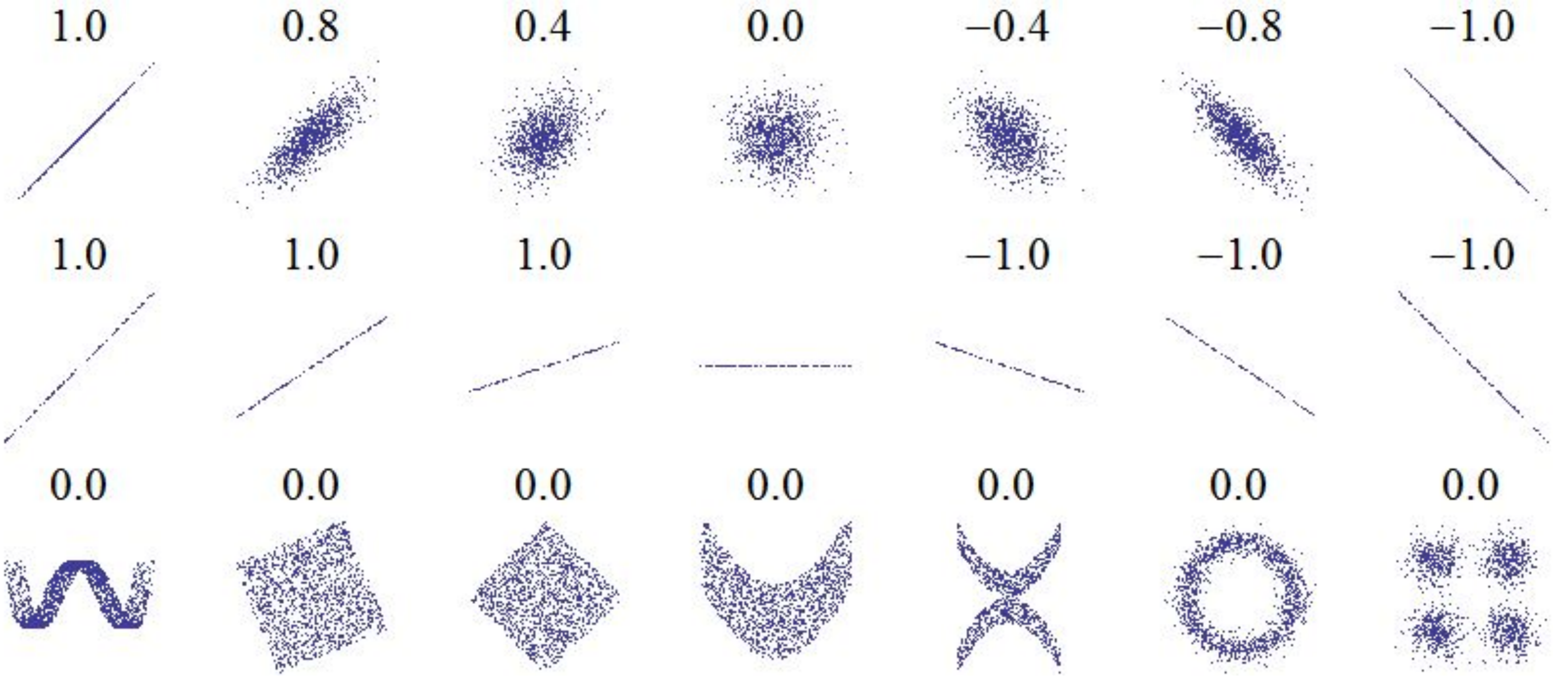
- Statistical relationship between two variables
- Normalized covariance: magnitude shows strength of relationship
- Popular measure is (sample) Pearson's r correlation coefficient:

$$r_{xy} = \frac{cov(X,Y)}{\sigma_X \sigma_Y} = \frac{\sum_{i=1}^{N} (X_i - \mu_X)(Y_i - \mu_Y)}{\sqrt{\sum_{i=1}^{N} (X_i - \mu_X)^2} \sqrt{\sum_{i=1}^{N} (Y_i - \mu_Y)^2}}$$

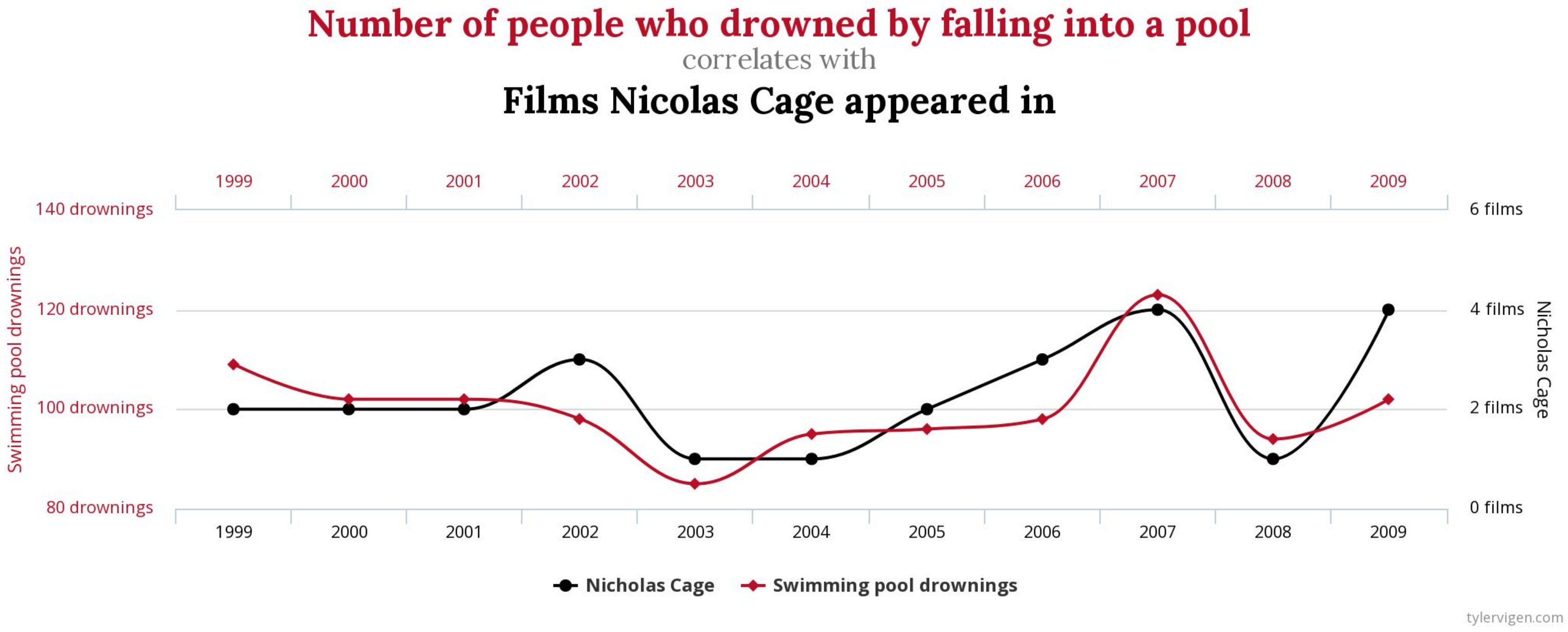
Visualising correlations



Visualising correlations

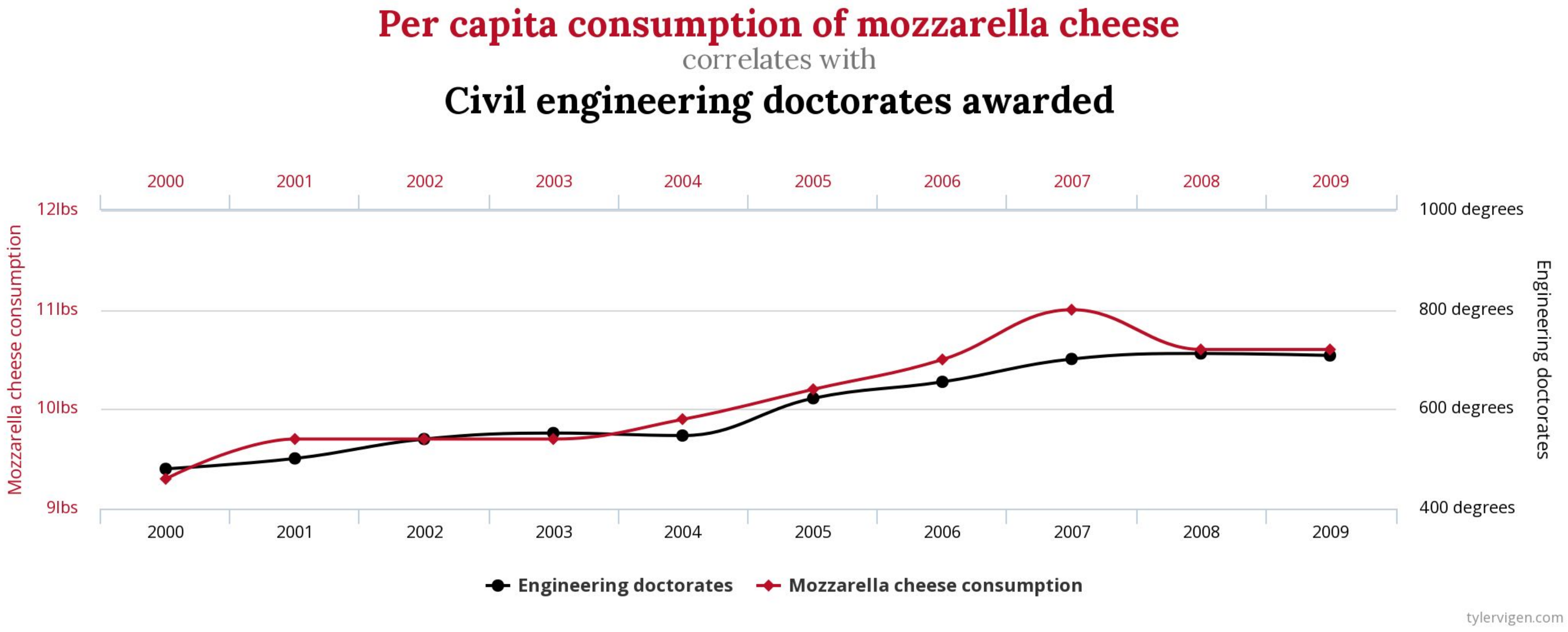


Spurious correlations



http://www.tylervigen.com/spurious-correlations

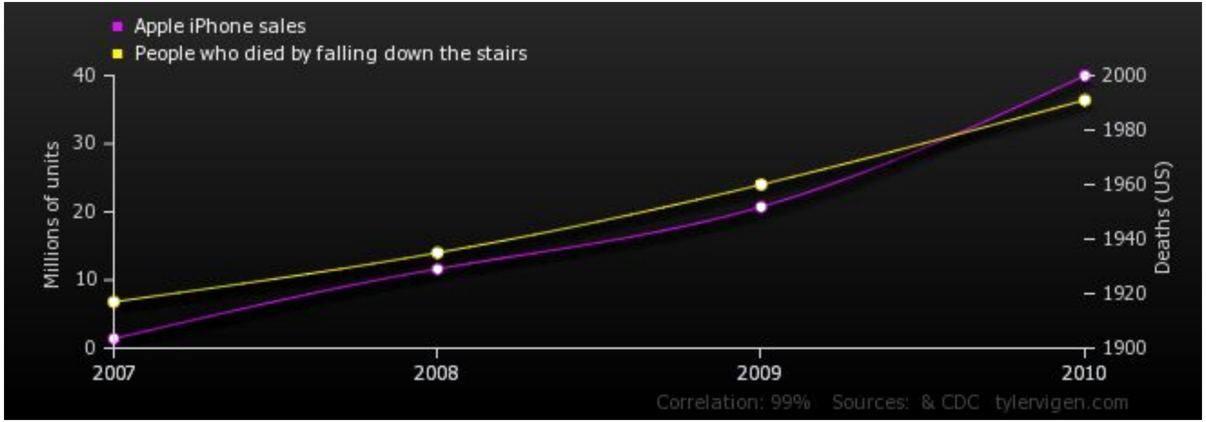
Spurious correlations



http://www.tylervigen.com/spurious-correlations

Spurious correlations

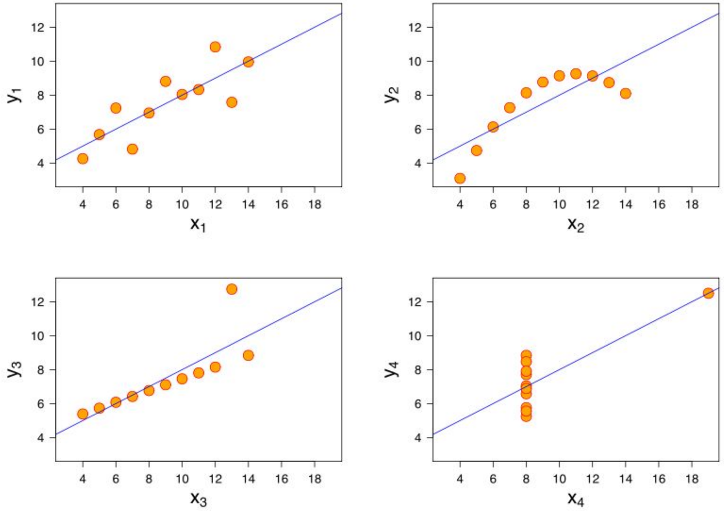
Correlation: 0.994751



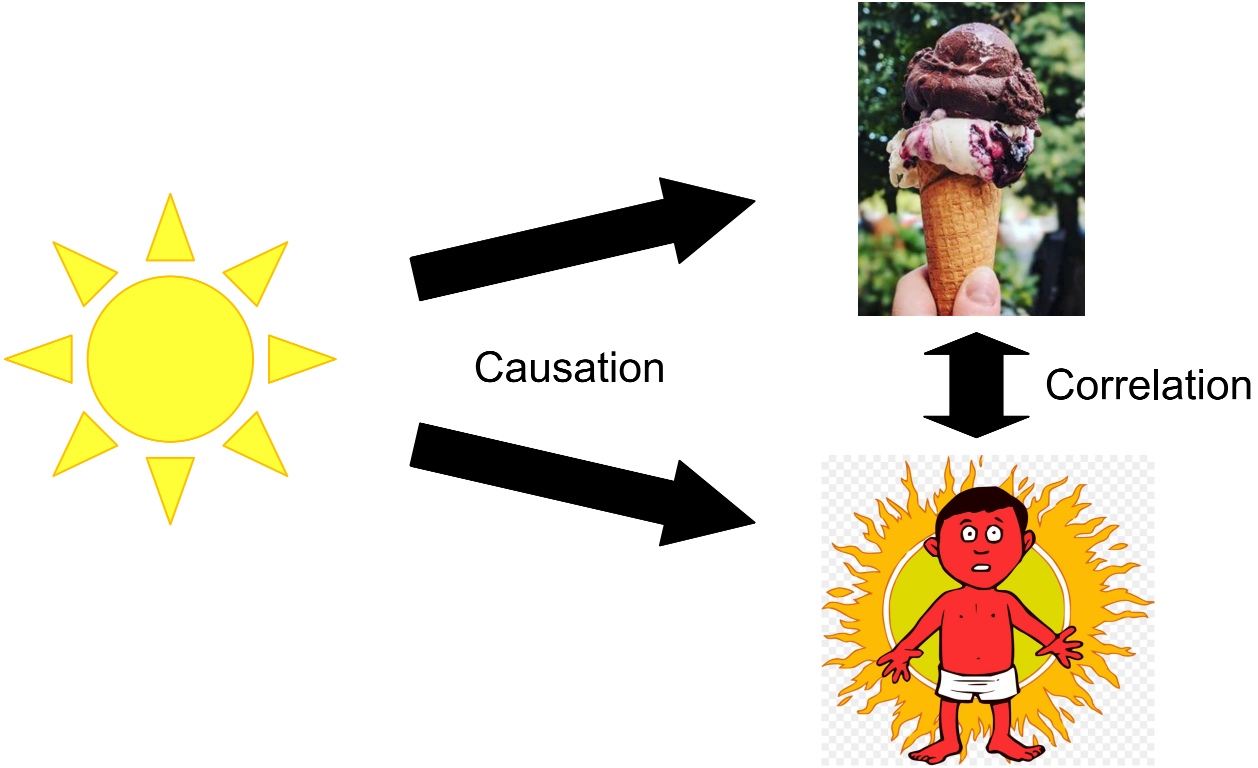
https://tylervigen.com/view_correlation?id=28669

Linear relation

- Correlations describes a linear relationship between variables
- But can be high even if relation isn’t linear
- Can be low even if there is an obvious relationship



Causation



Relations between instances

How to compare rows?

Univariate

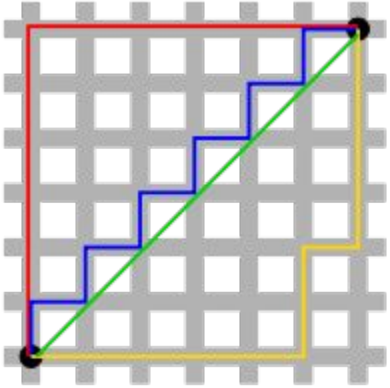
Index	Variable 1	Target Variable
..	..	...
..	..	..

Multivariate

Index	Variable 1	Variable 2	Variable n	Target Variable
..	..	..	..	...
..	..	..	..	..

Distance Metrics

- Basic properties
  - Positive separation
    - $D(x, y) > 0, \forall x \neq y$
    - $D(x, y) = 0$, i.f.f., $x = y$
  - Symmetry
    - $D(x, y) = D(y, x)$
  - Triangle inequality
    - $D(x, y) \leq D(x, z) + D(z, y)$



Dot product

$$\mathbf{a} = [a_1, a_2, \cdots, a_n]$$

$$\mathbf{b} = [b_1, b_2, \cdots, b_n]$$

$$\mathbf{a} \cdot \mathbf{b} = \sum_{i=1}^{n} a_i b_i = a_1 b_1 + a_2 b_2 + \cdots + a_n b_n$$

```python
a = [1, 2, 3, 4]
b = [4, 5, 6, 7]
```
[1] ✓ 0.0s

```python
result = 0
for i in range(len(a)):
    result += a[i] * b[i]
display(result)
```
[2] ✓ 0.0s

```
60
```

```python
result = 0
for ai, bi in zip(a, b):
    result += ai * bi
display(result)
```
[4] ✓ 0.0s

```
60
```

```python
result = sum([ai * bi for ai, bi in zip(a,b)])
display(result)
```
[5] ✓ 0.0s

```
60
```

```python
import numpy as np
np.dot(a,b)
```
[7] ✓ 0.0s

```
60
```

Cosine

$$\text{cosine similarity} = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \cdot \sqrt{\sum_{i=1}^{n} B_i^2}},$$

```
[15]
np.dot(a, b) / (np.linalg.norm(a) * np.linalg.norm(b))
✓ 0.0s
0.9759000729485332
```

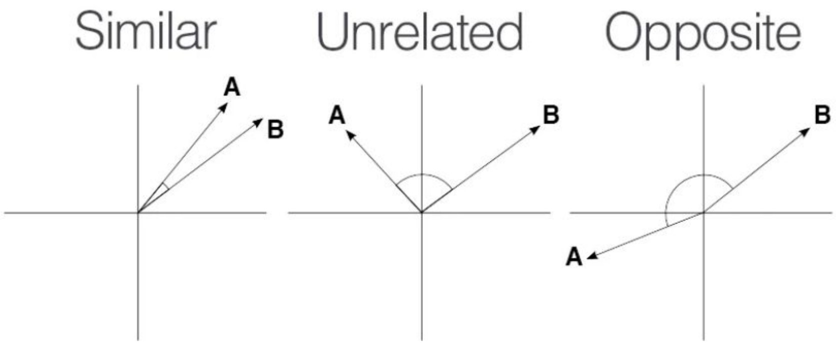
```
[16]
a_norm = a / np.linalg.norm(a)
b_norm = b / np.linalg.norm(b)
np.dot(a_norm,b_norm)
✓ 0.0s
0.9759000729485331
```

```
[17]
import scipy.spatial.distance as dist

dist.cosine(a,b)
✓ 0.0s
0.024099927051466796
```

```
[18]
1 - np.dot(a, b) / (np.linalg.norm(a) * np.linalg.norm(b))
✓ 0.0s
0.024099927051466796
```

Angle and magnitude




```
a = [10, 20, 30, 40]
```
[21] ✓ 0.0s

```
import numpy as np
np.dot(a,b)
```
[25] ✓ 0.0s

```
600
```

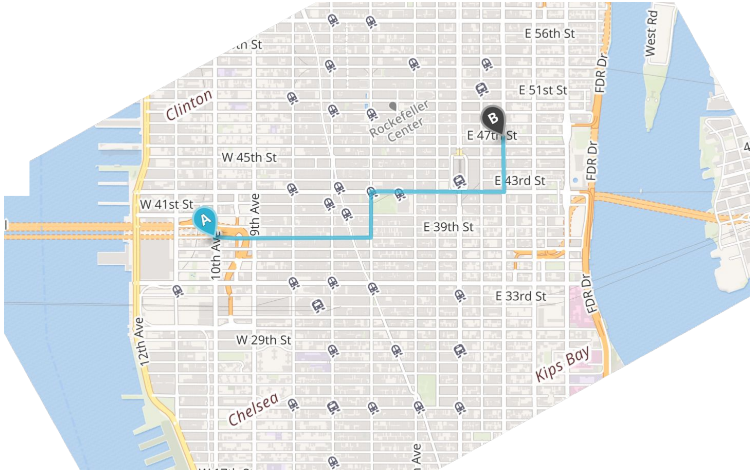
```
import scipy.spatial.distance as dist

dist.cosine(a,b)
```
[28] ✓ 0.0s

```
0.024099927051466907
```

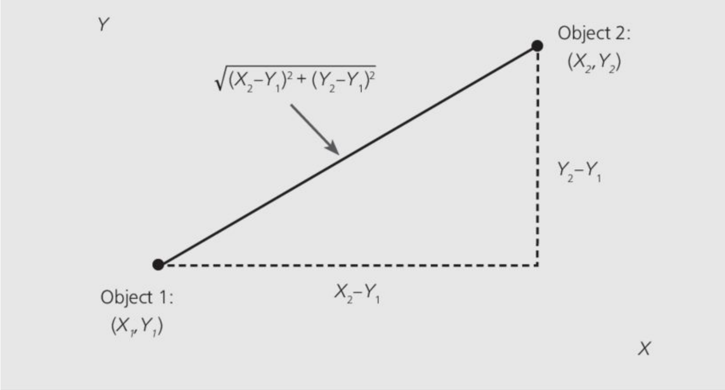
Manhattan

$$d_{\mathrm{T}}(\mathbf{p}, \mathbf{q}) = \|\mathbf{p} - \mathbf{q}\|_{\mathrm{T}} = \sum_{i=1}^{n} |p_i - q_i|$$



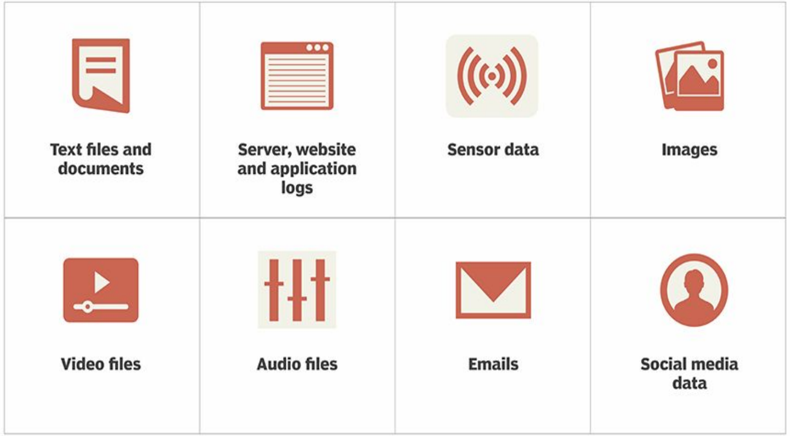
Euclidean

$$d_2(\vec{a}, \vec{b}) = \sqrt{\sum_{i=1}^{n} (a_i - b_i)^2}$$



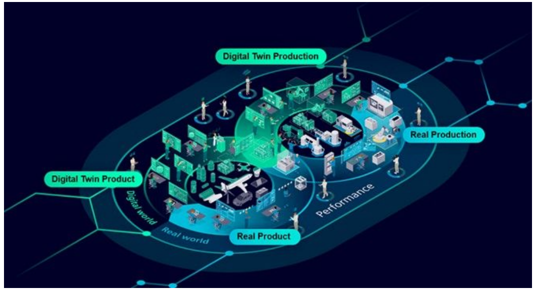
Representation Learning

Unstructured data



- Cannot simply measure the distance
- Need to build a model for how to represent

Modelling

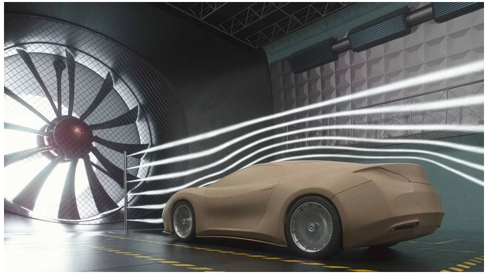


S = System

M = Model

E = Experiment

$E(S) \approx E(M)$



shutterstock.com · 2134374559

What gets counted?

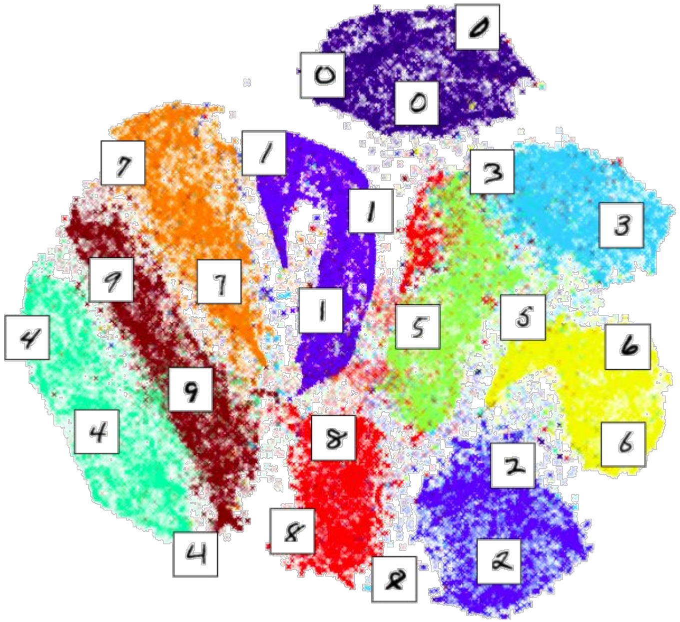
- Which data we count/store already depends on a model
  - What we count reflects how we see the world
- A data model is less nuanced than reality
  - Requires assumptions and abstractions

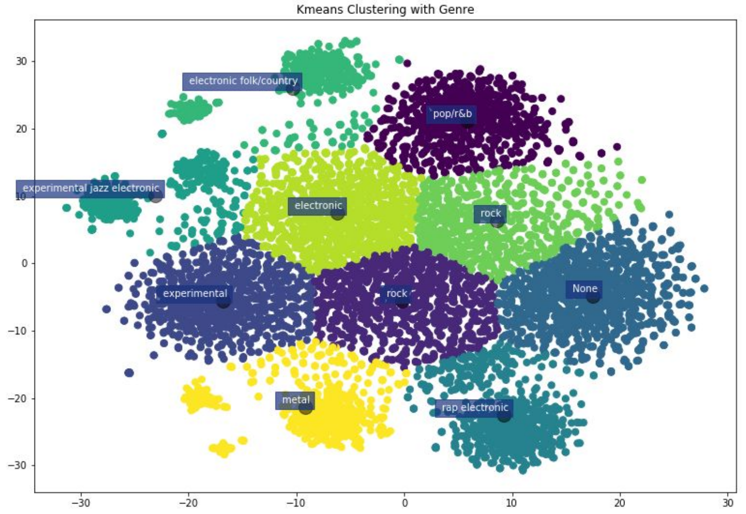
if

S = Gender
M = Binary

Then $E(S) \approx E(M)$?

Data Model


0
0
0
1
1
1
3
3
7
7
9
9
4
4
4
5
5
6
6
8
8
8
2
2


Kmeans Clustering with Genre
electronic folk/country
pop/r&b
experimental jazz electronic
electronic
rock
experimental
rock
None
metal
rap electronic
30
20
10
0
−10
−20
−30
−30
−20
−10
0
10
20
30

Bag of Words

Document D1	*The child makes the dog happy* the: 2, dog: 1, makes: 1, child: 1, happy: 1
Document D2	*The dog makes the child happy* the: 2, child: 1, makes: 1, dog: 1, happy: 1

↓

	child	dog	happy	makes	the	BoW Vector representations
D1	1	1	1	1	2	[1,1,1,1,2]
D2	1	1	1	1	2	[1,1,1,1,2]

What is the model?

```python
from collections import Counter

corpus = [
    "the child makes the dog happy",
    "the dog makes the child happy",
    "my child is happy when playing with another child"
]

tokenized_corpus = [d.split() for d in corpus]

vocabulary = Counter()
for document in tokenized_corpus:
    vocabulary.update(document)
vocabulary = sorted(vocabulary)

def BoW(doc, vocabulary):
    bow = [0] * len(vocabulary)
    for word, cnt in Counter(doc).items():
        bow[vocabulary.index(word)] = cnt
    return bow

bows = [BoW(d, vocabulary) for d in tokenized_corpus]
```

[34] ✓ 0.0s

```python
dot_distances = dist.squareform(dist.pdist(bows, metric=np.dot))
np.fill_diagonal(dot_distances, np.inf)
display(dot_distances)
```

[35] ✓ 0.0s

```
array([[inf,  8.,  3.],
       [ 8., inf,  3.],
       [ 3.,  3., inf]])
```

```python
cos_distances = dist.squareform(dist.pdist(bows, metric='cosine'))
np.fill_diagonal(cos_distances, np.inf)
display(cos_distances)
```

[36] ✓ 0.0s

```
array([[           inf, 2.22044605e-16, 6.80198925e-01],
       [2.22044605e-16,            inf, 6.80198925e-01],
       [6.80198925e-01, 6.80198925e-01,            inf]])
```

```python
dot_distances.argmin(1), cos_distances.argmin(1)
```

[37] ✓ 0.0s

```
(array([2, 2, 0]), array([1, 0, 0]))
```

Working with data is modelling

- How we see the worlds influences how we model it
- How we model the world influences how we see it