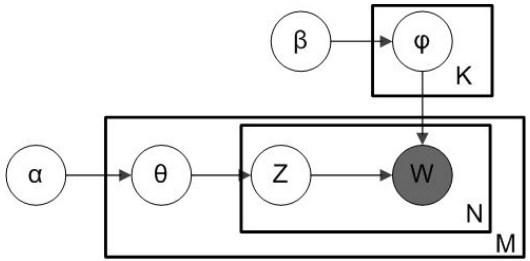
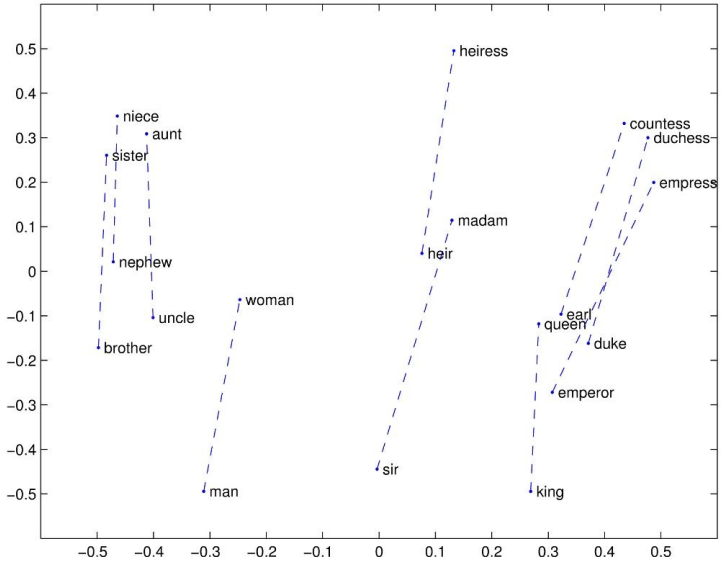
Practical Unsupervised Learning

INFO/CS 4300, Spring 2016

Jack Hessel

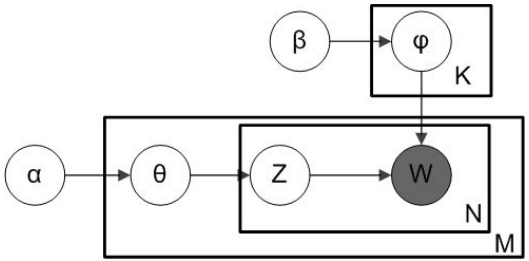
Unsupervised Learning is Cool!

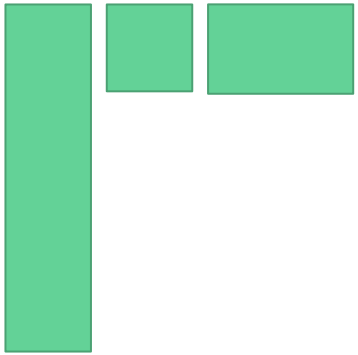




But how can we use this in our projects?

1.





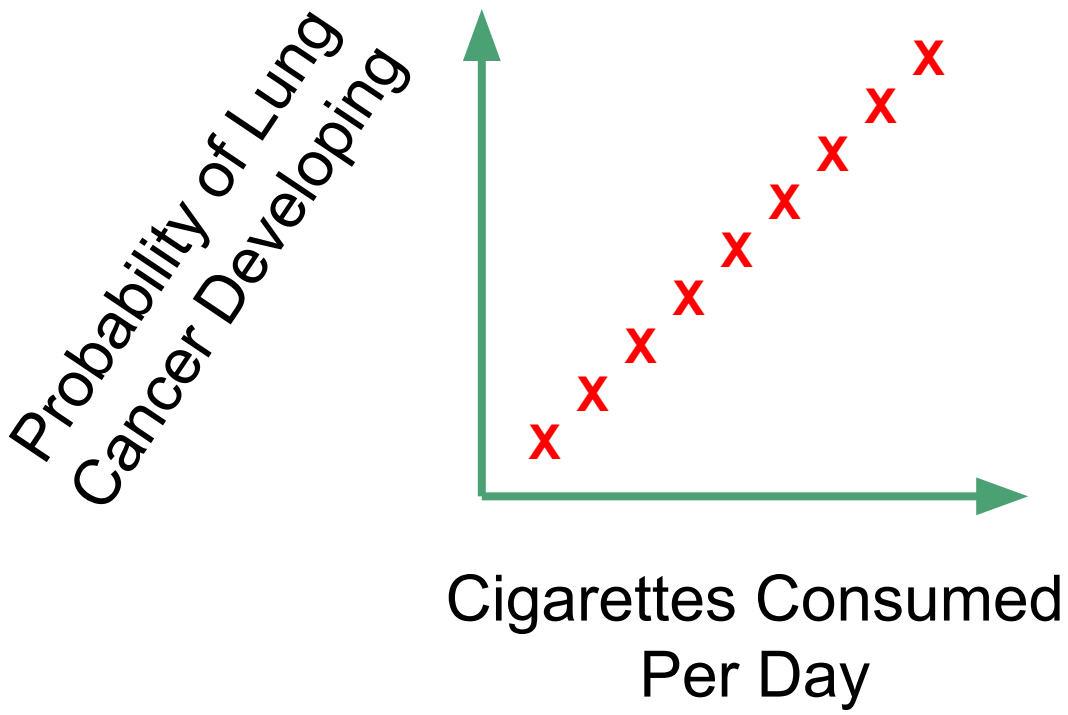
2.

KICKSTARTER

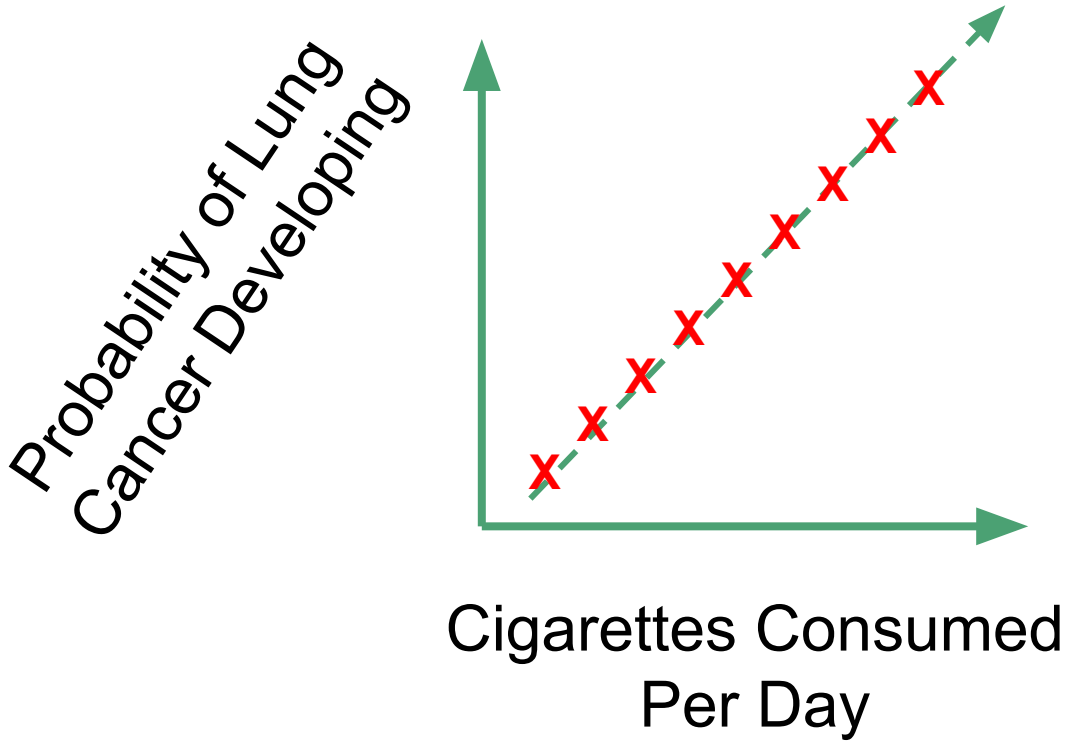
But first, let’s look at our dataset…

KICKSTARTER

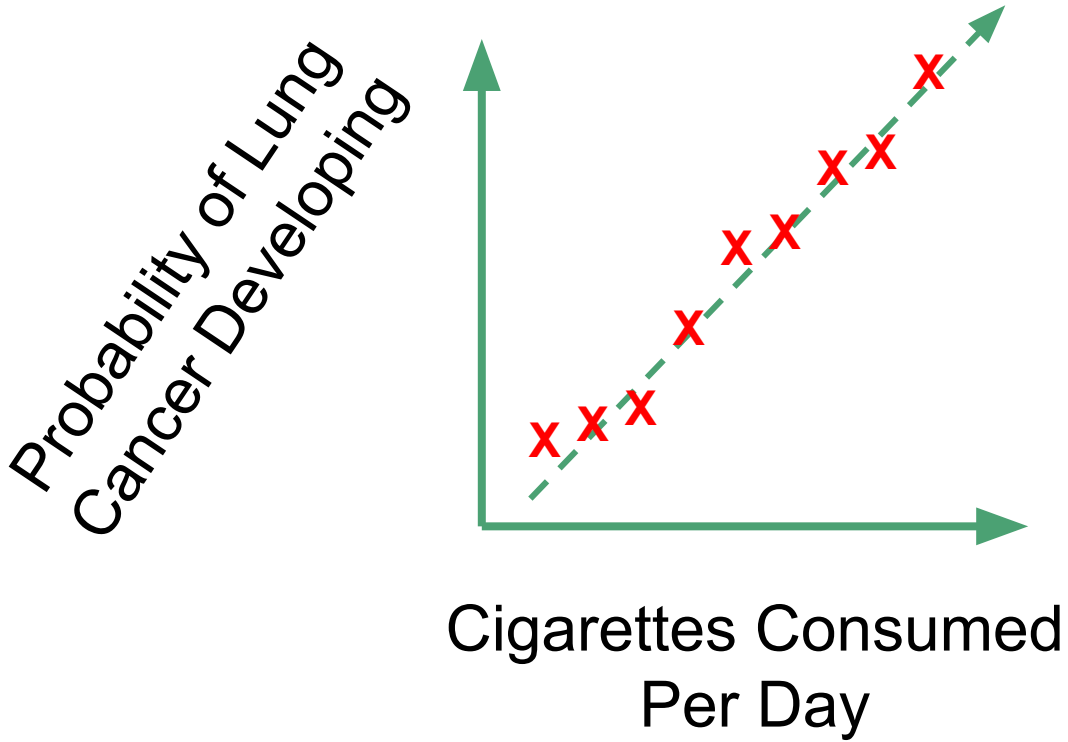
Data Dimensionality!



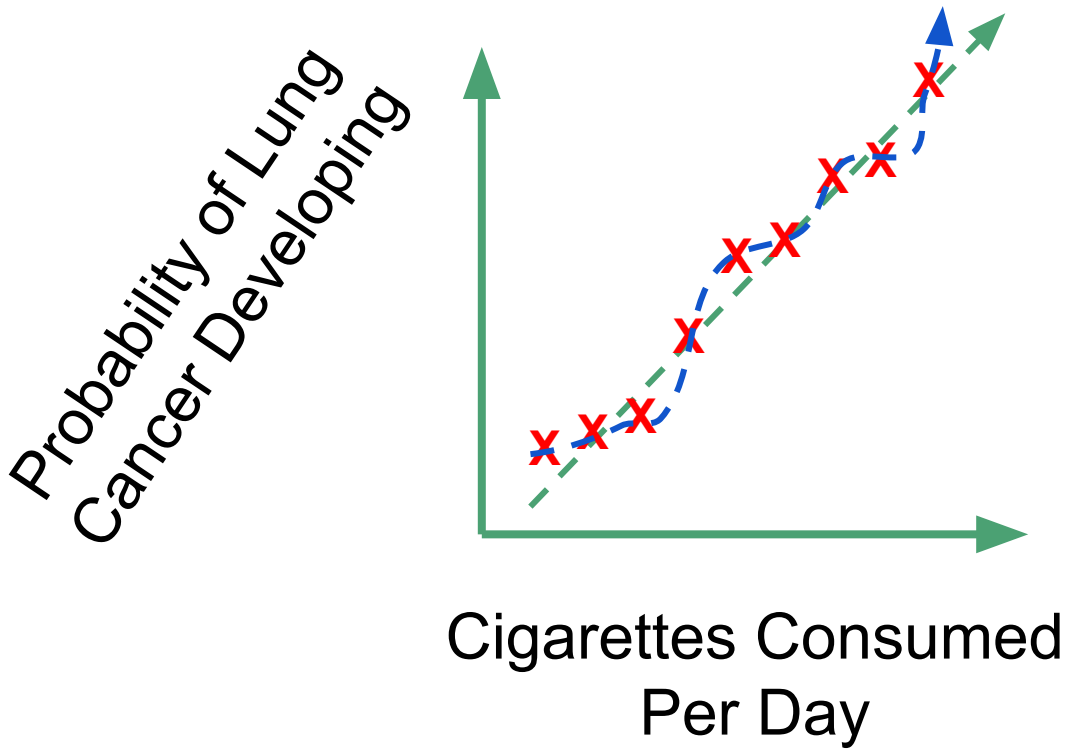
Data Dimensionality!



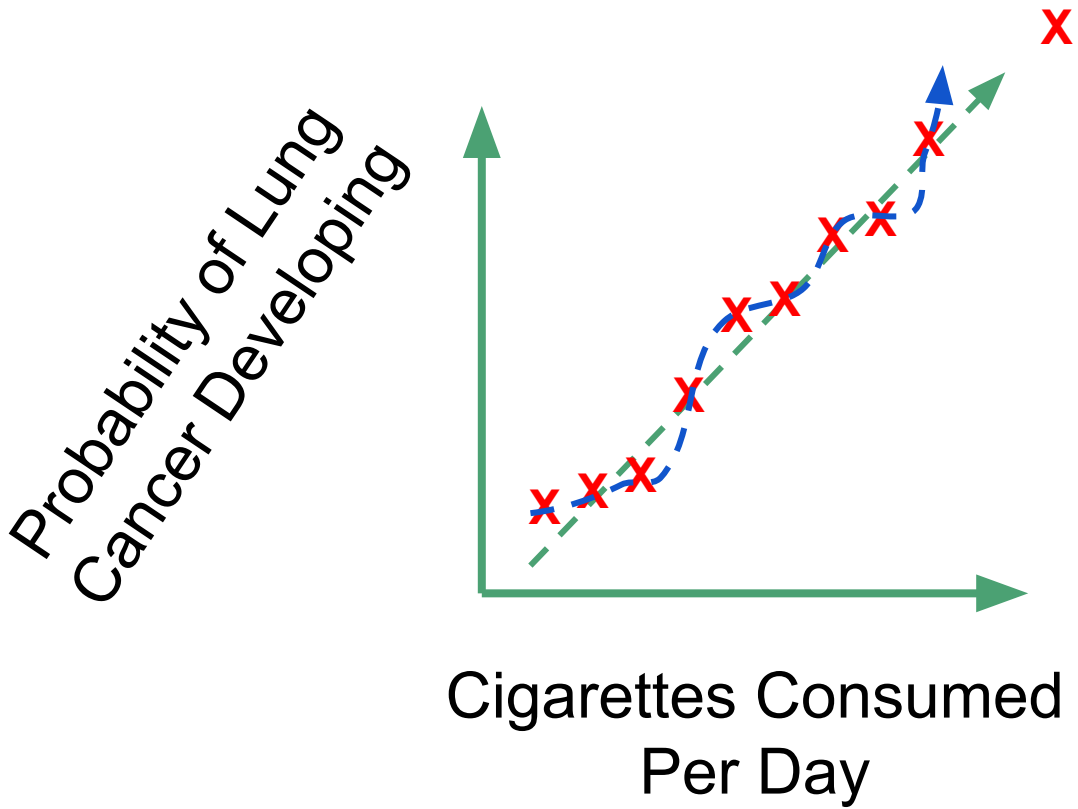
Data Dimensionality!



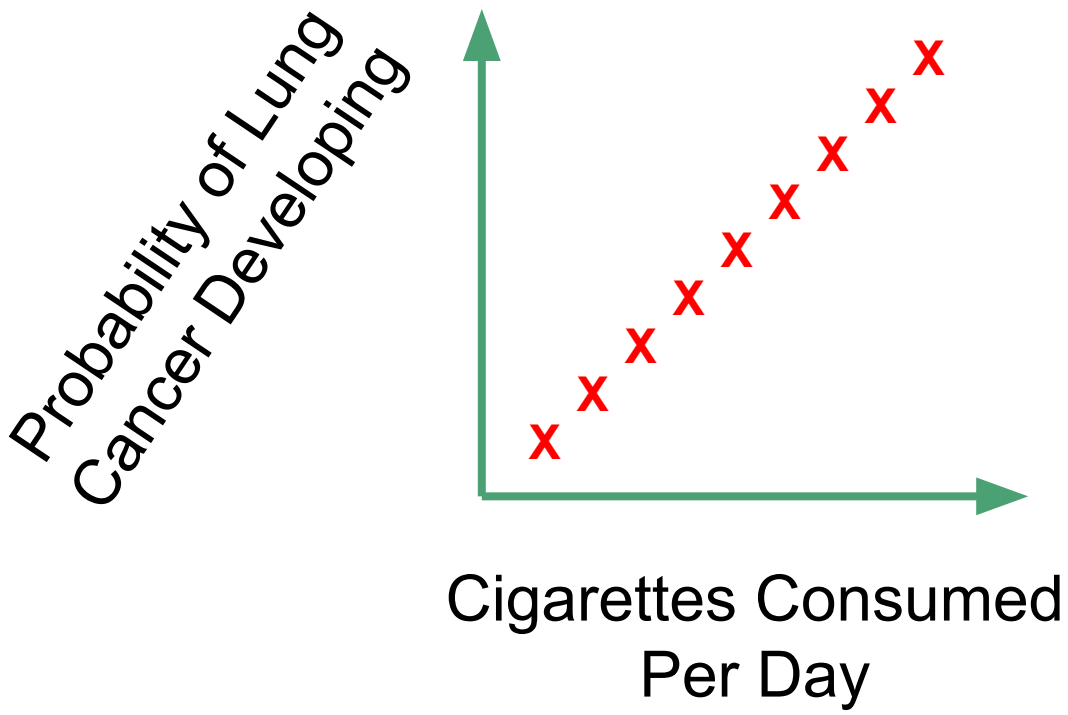
Data Dimensionality!



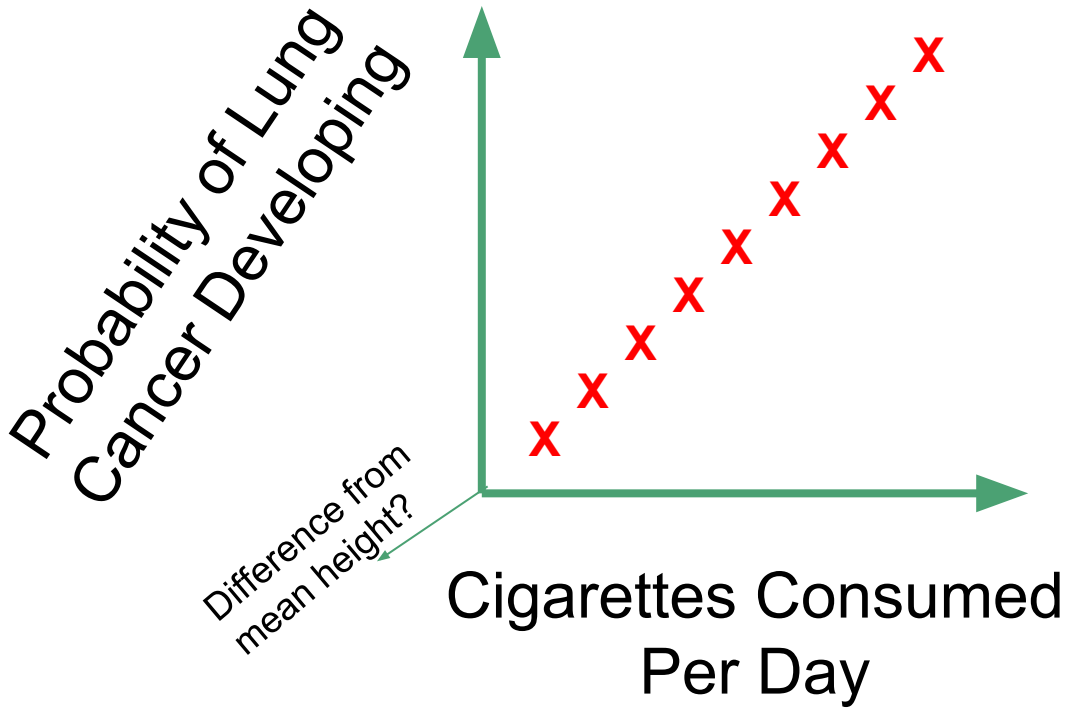
Data Dimensionality!



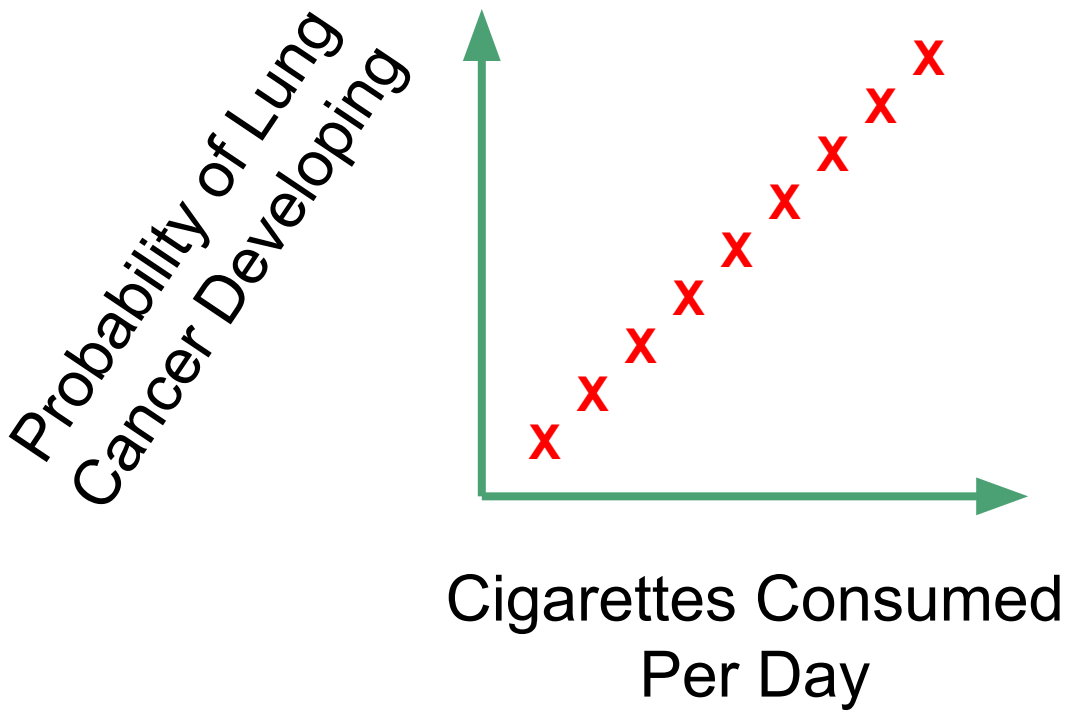
Data Dimensionality!



Data Dimensionality!

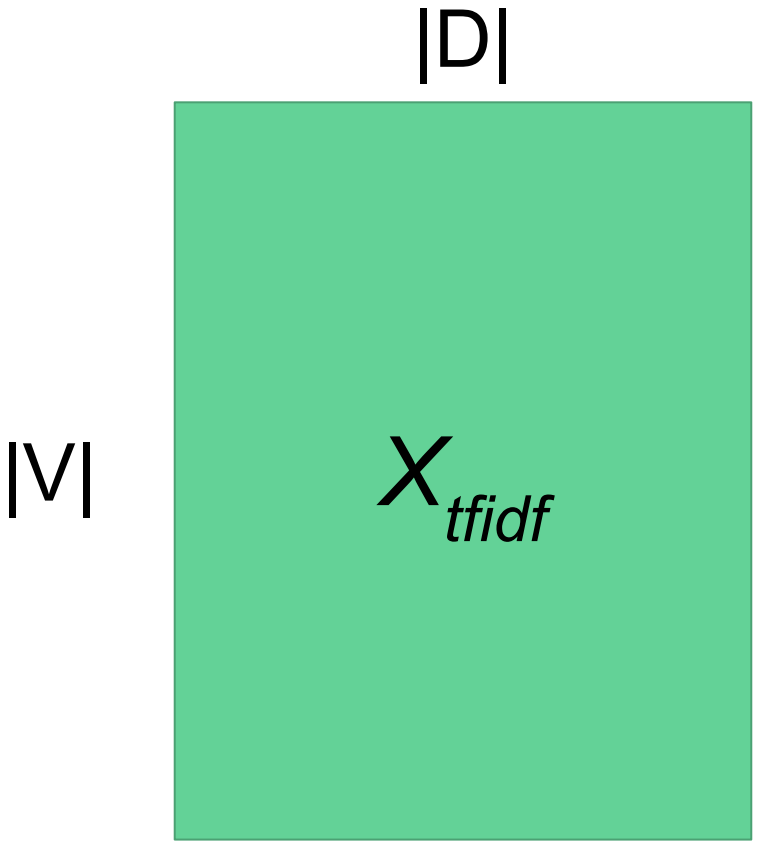


Data Dimensionality!

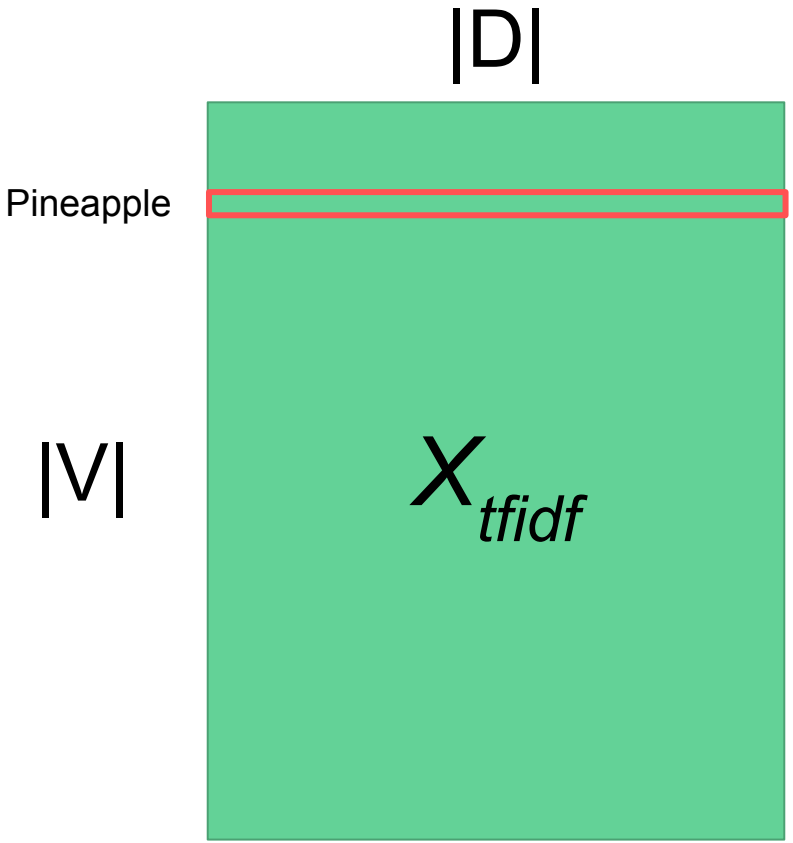


1 Dimension!

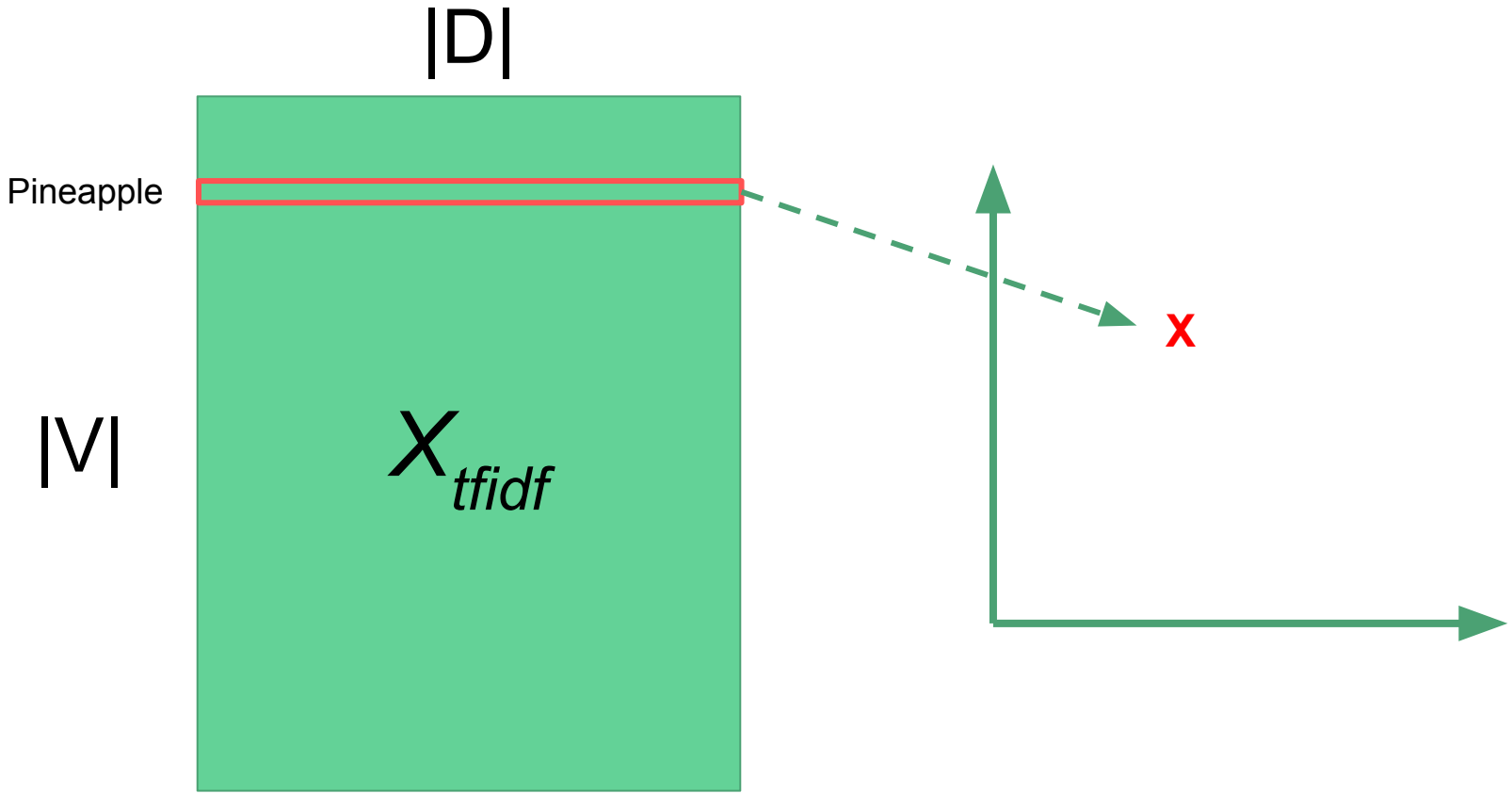
Words and documents are the same way…



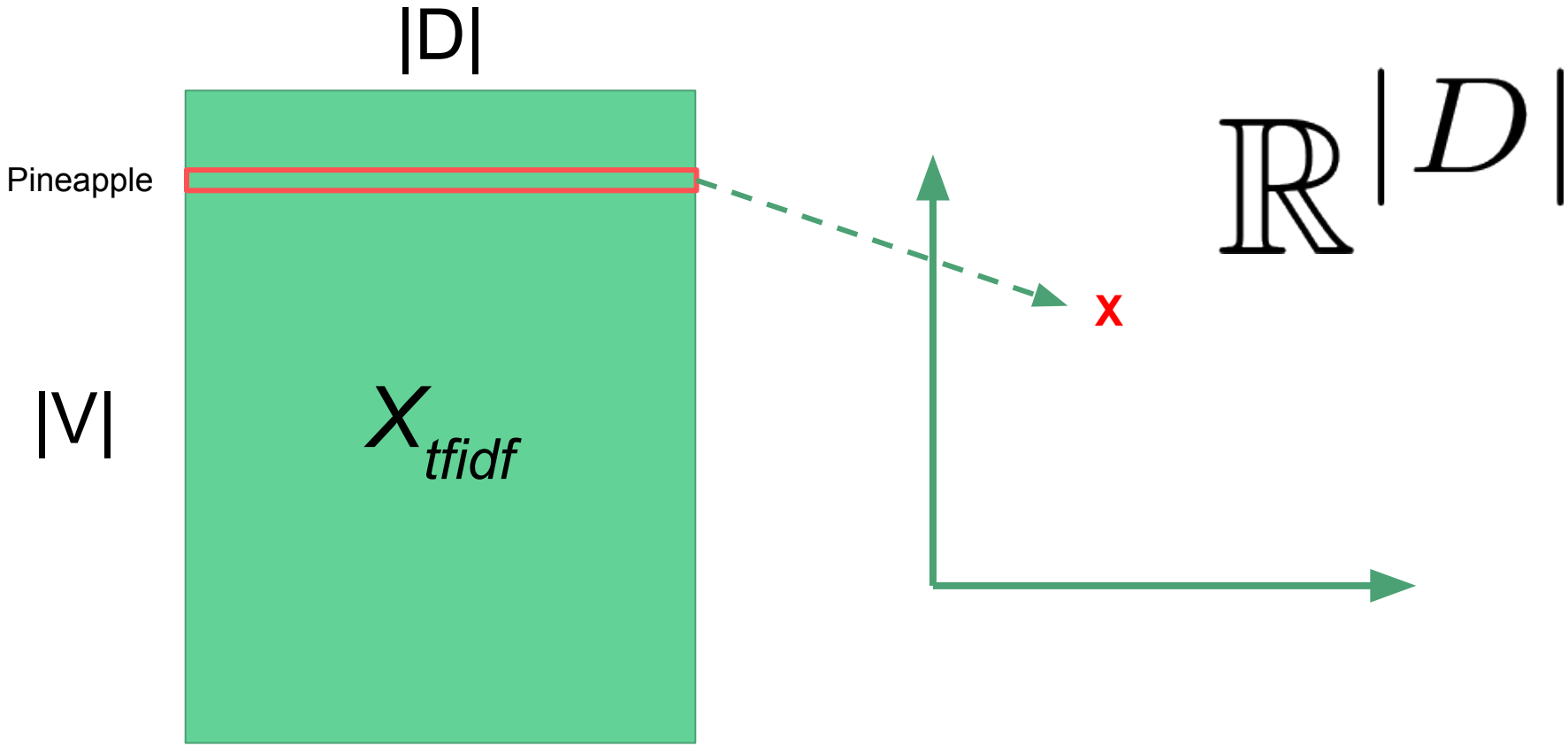
Words and documents are the same way...



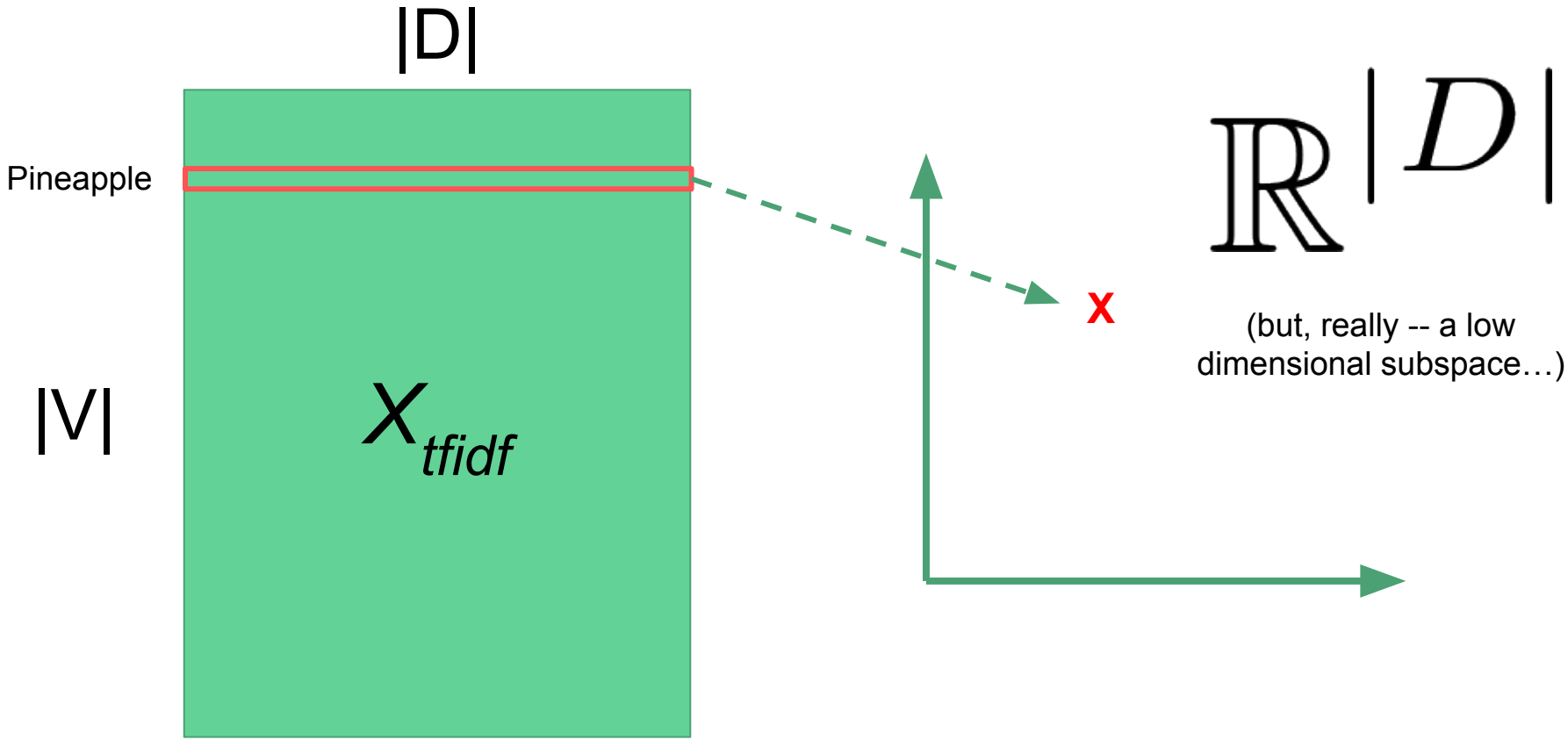
Words and documents are the same way...



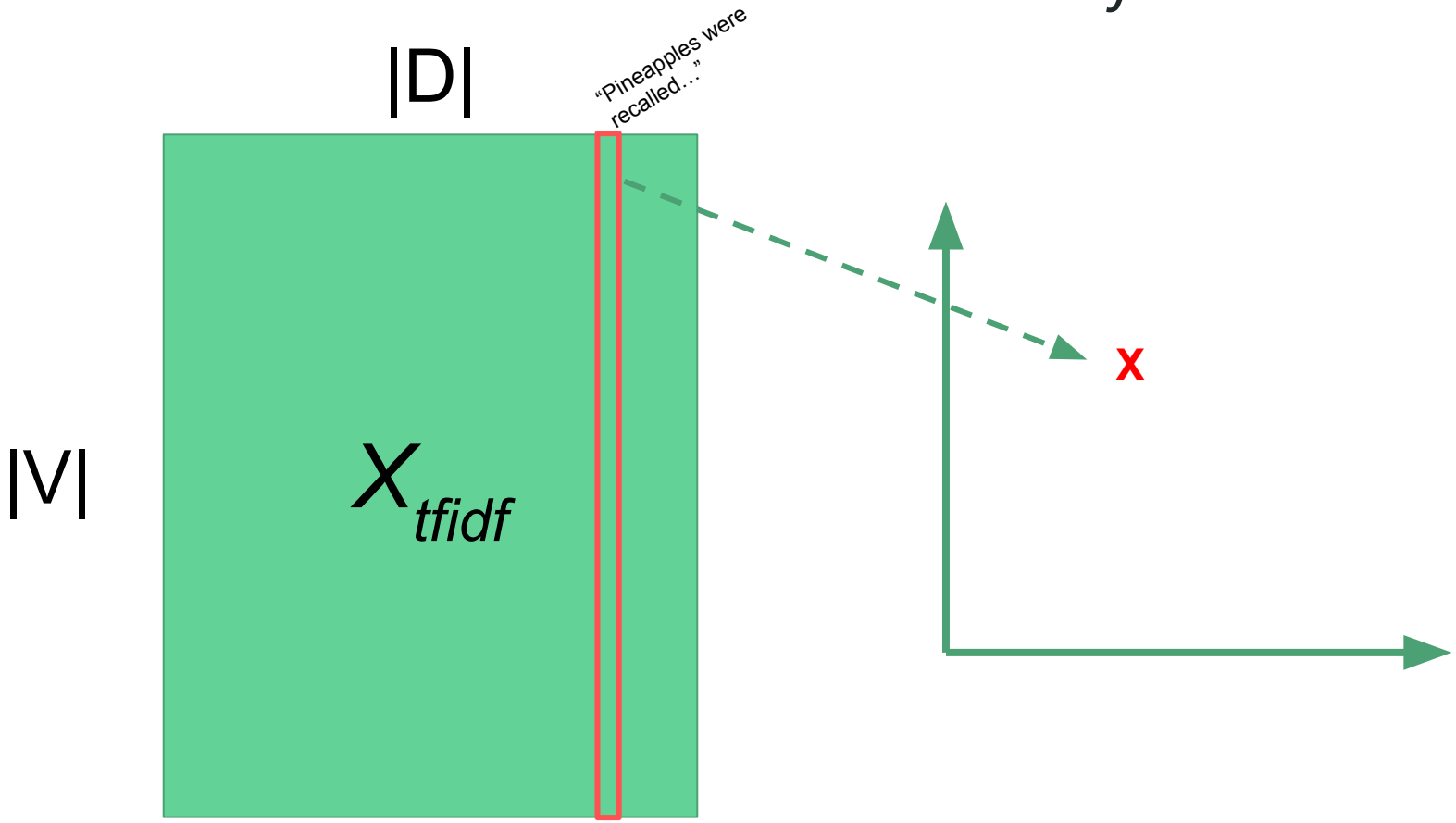
Words and documents are the same way...



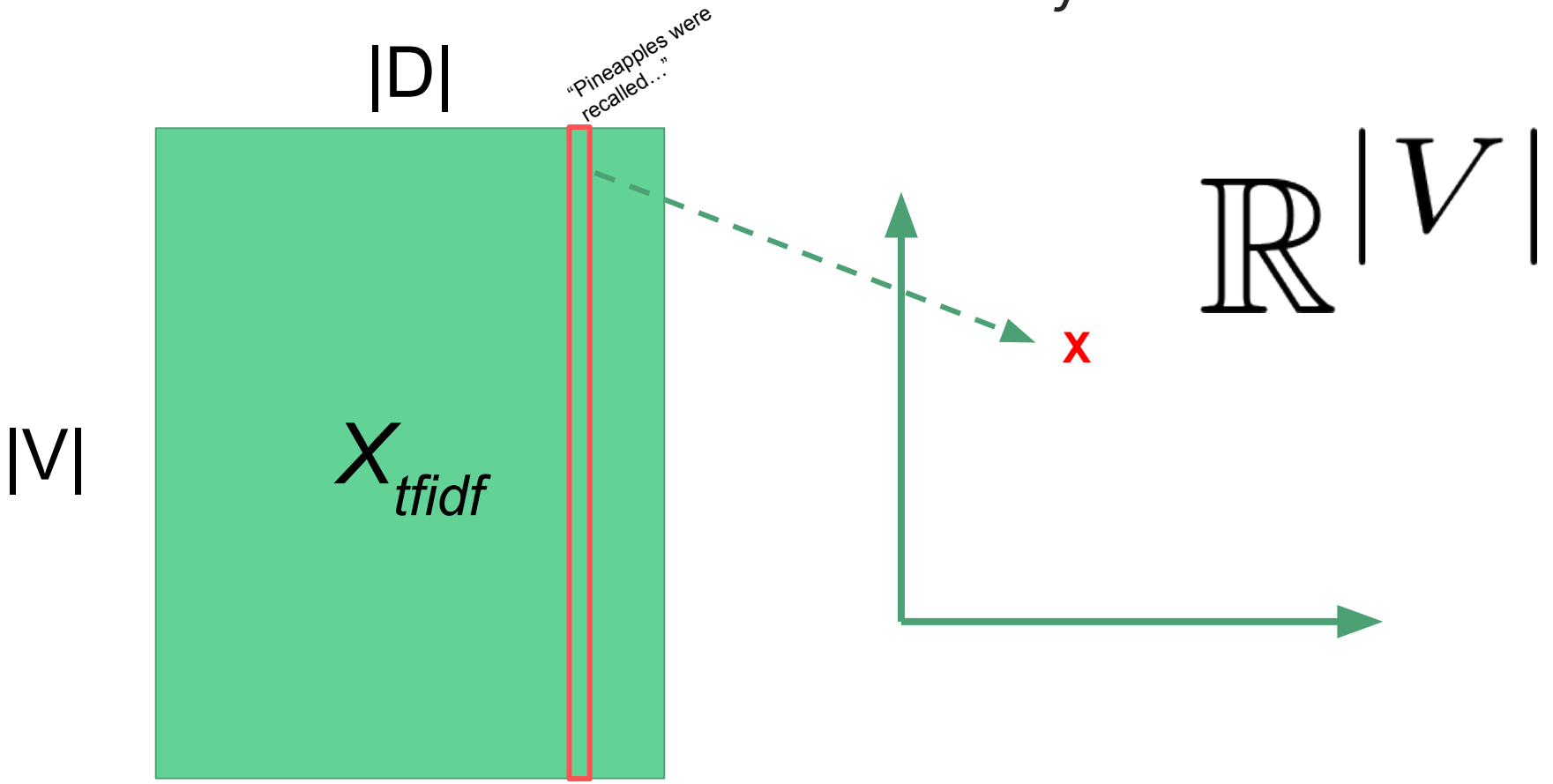
Words and documents are the same way…



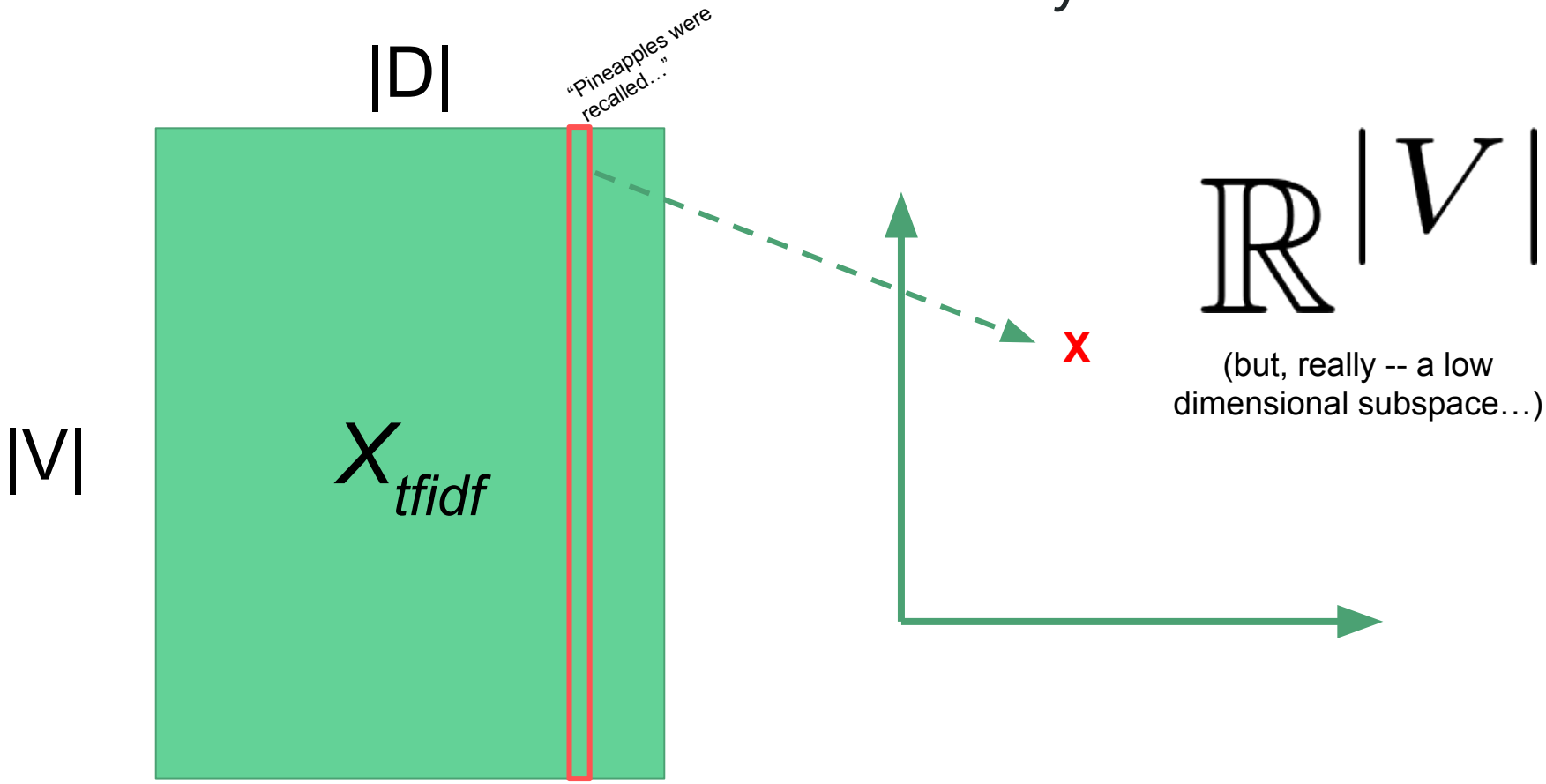
Words and documents are the same way…



Words and documents are the same way…



Words and documents are the same way…



Key questions in unsupervised NLP:

Key questions in unsupervised NLP:

1. How many dimensions does our dataset actually live in?

Key questions in unsupervised NLP:

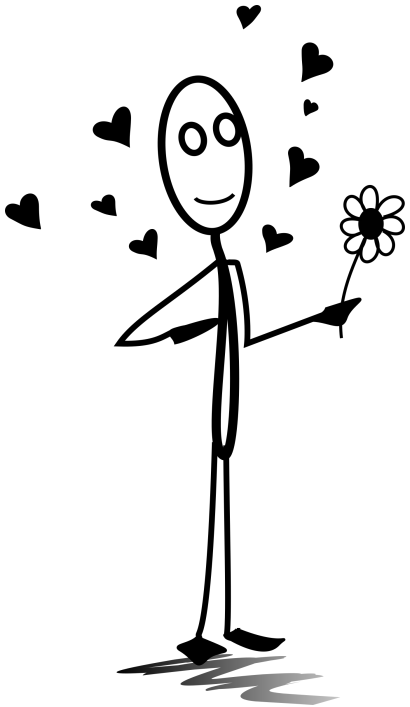
1. How many dimensions does our dataset actually live in?
2. How do we project our data down to those dimensions?

Key questions in unsupervised NLP:

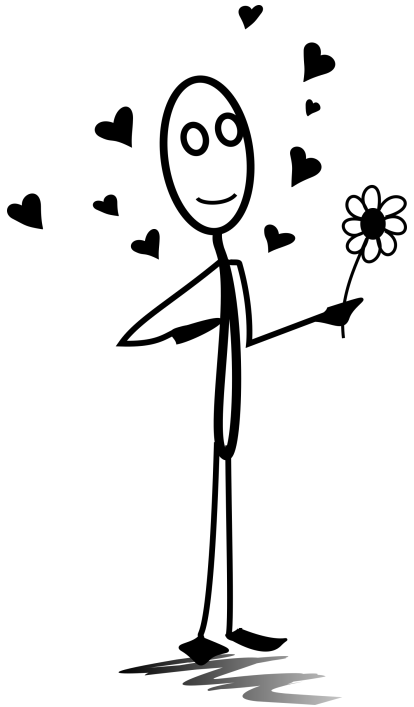
1. How many dimensions does our dataset actually live in?
2. How do we project our data down to those dimensions?
3. Does any of this stuff actually do anything for our projects?

Key tool in Linear Algebra, NLP, Machine Learning, Data Science, Computer Vision, Algorithms, Matrix Computations, Optimization, Statistics...

Key tool in Linear Algebra, NLP, Machine Learning, Data Science, Computer Vision, Algorithms, Matrix Computations, Optimization, Statistics...

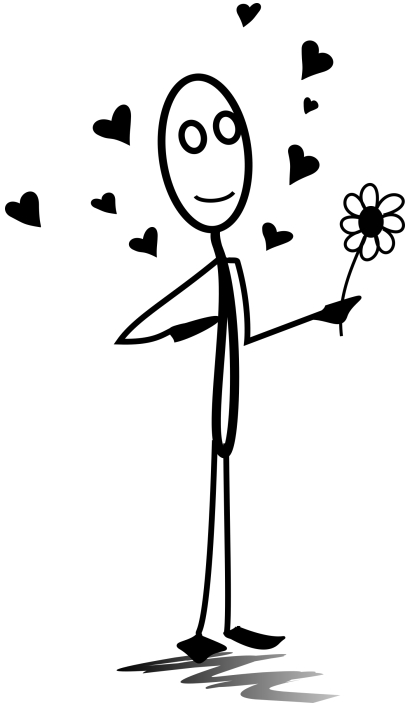


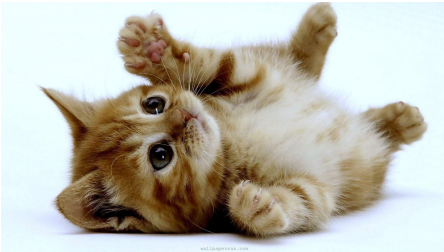
Key tool in Linear Algebra, NLP, Machine Learning, Data Science, Computer Vision, Algorithms, Matrix Computations, Optimization, Statistics...



$$X = U\Sigma V^T$$

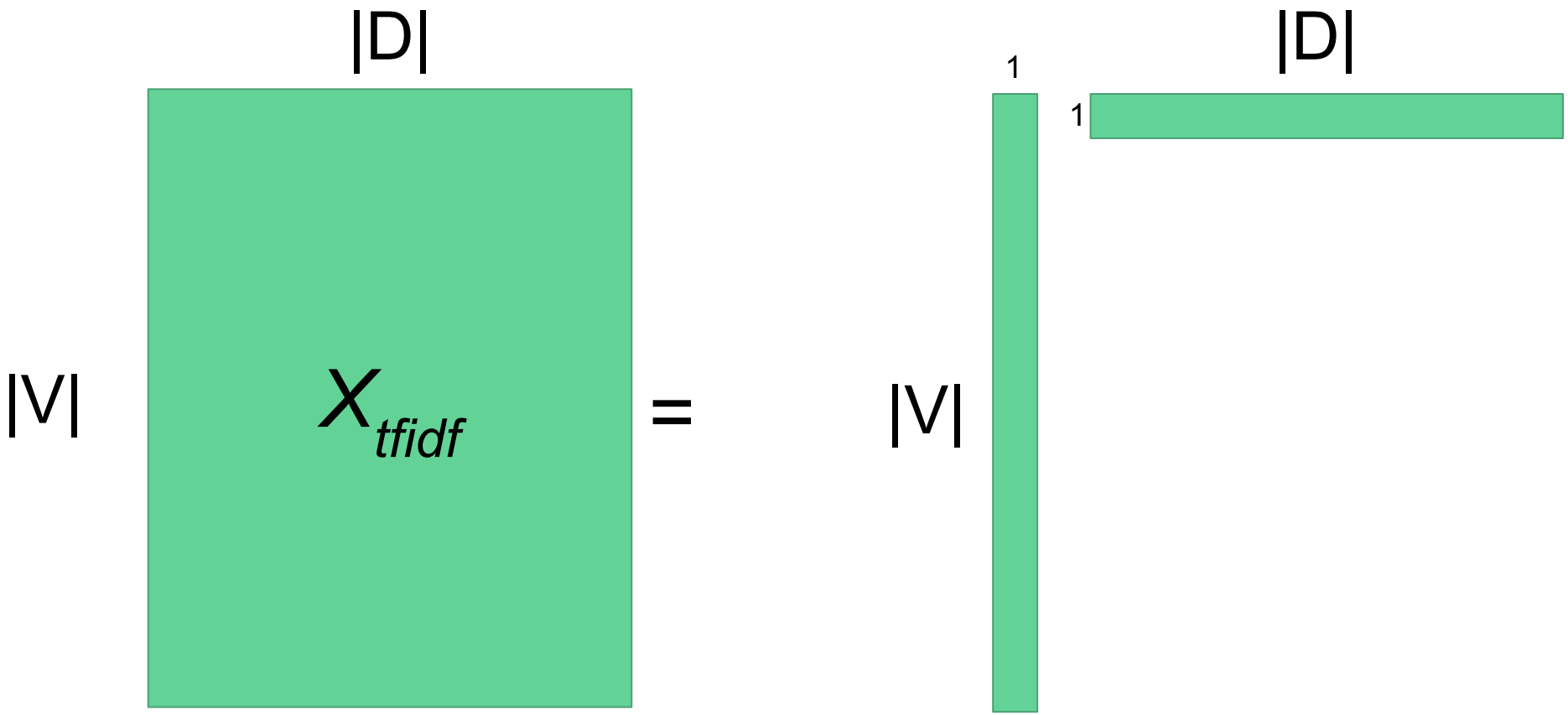
Key tool in Linear Algebra, NLP, Machine Learning, Data Science, Computer Vision, Algorithms, Matrix Computations, Optimization, Statistics...

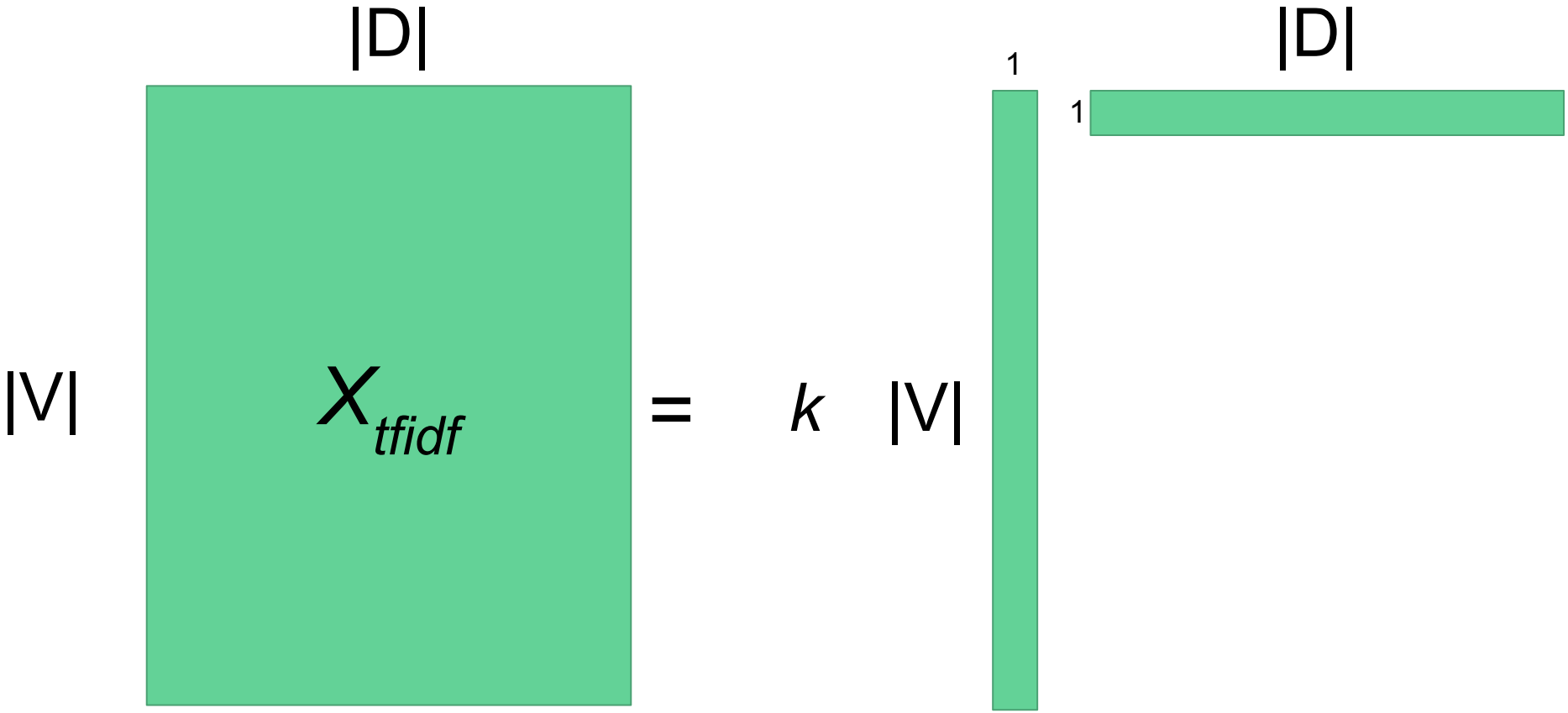


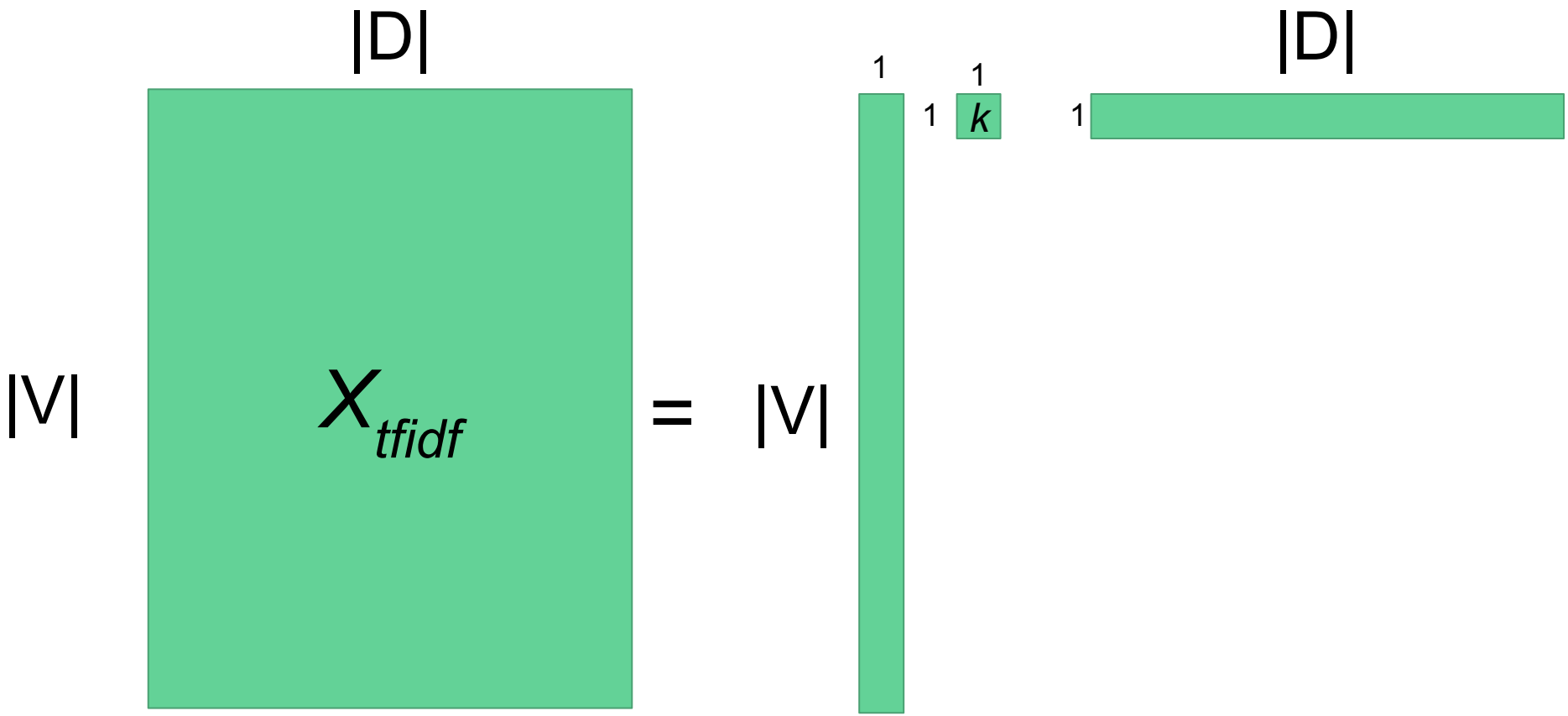


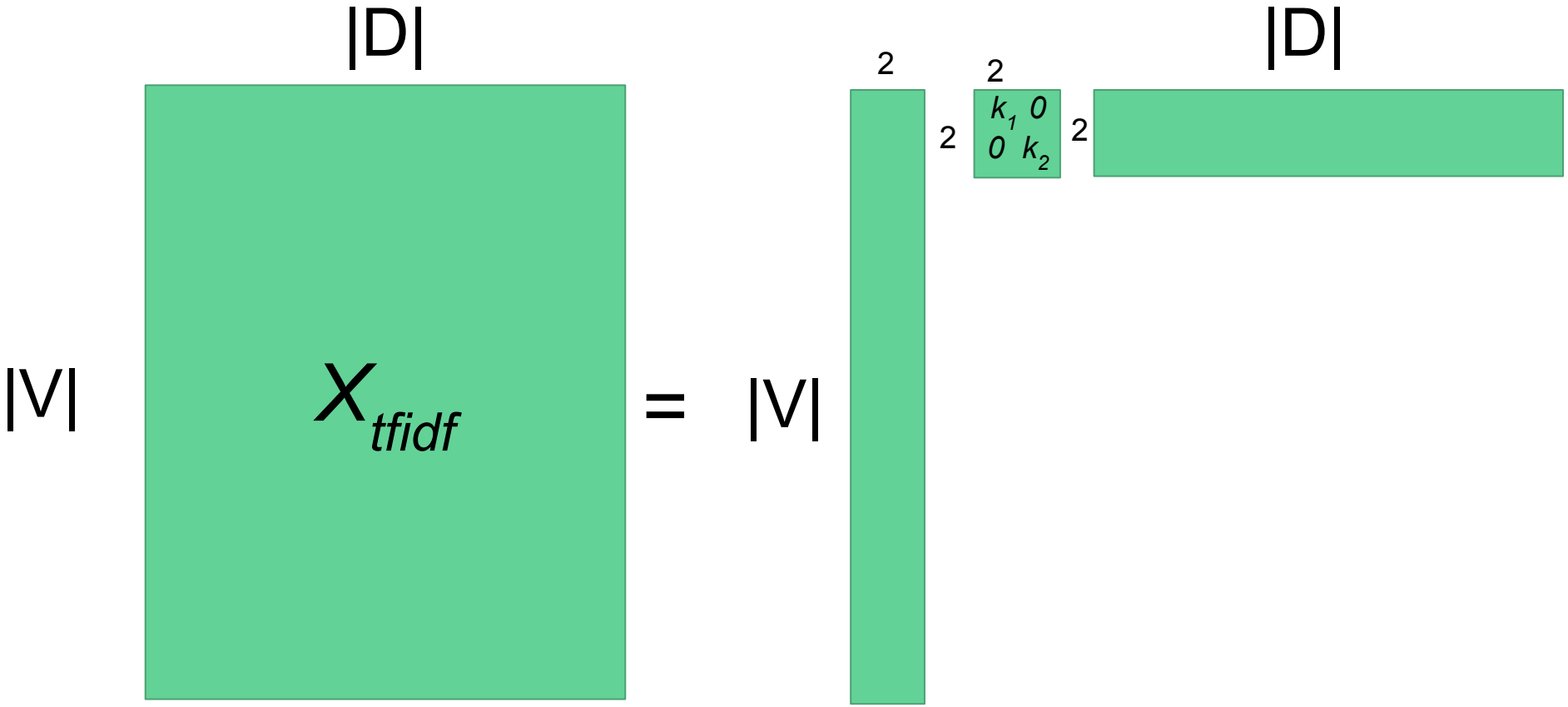
$$X = U\Sigma V^T$$

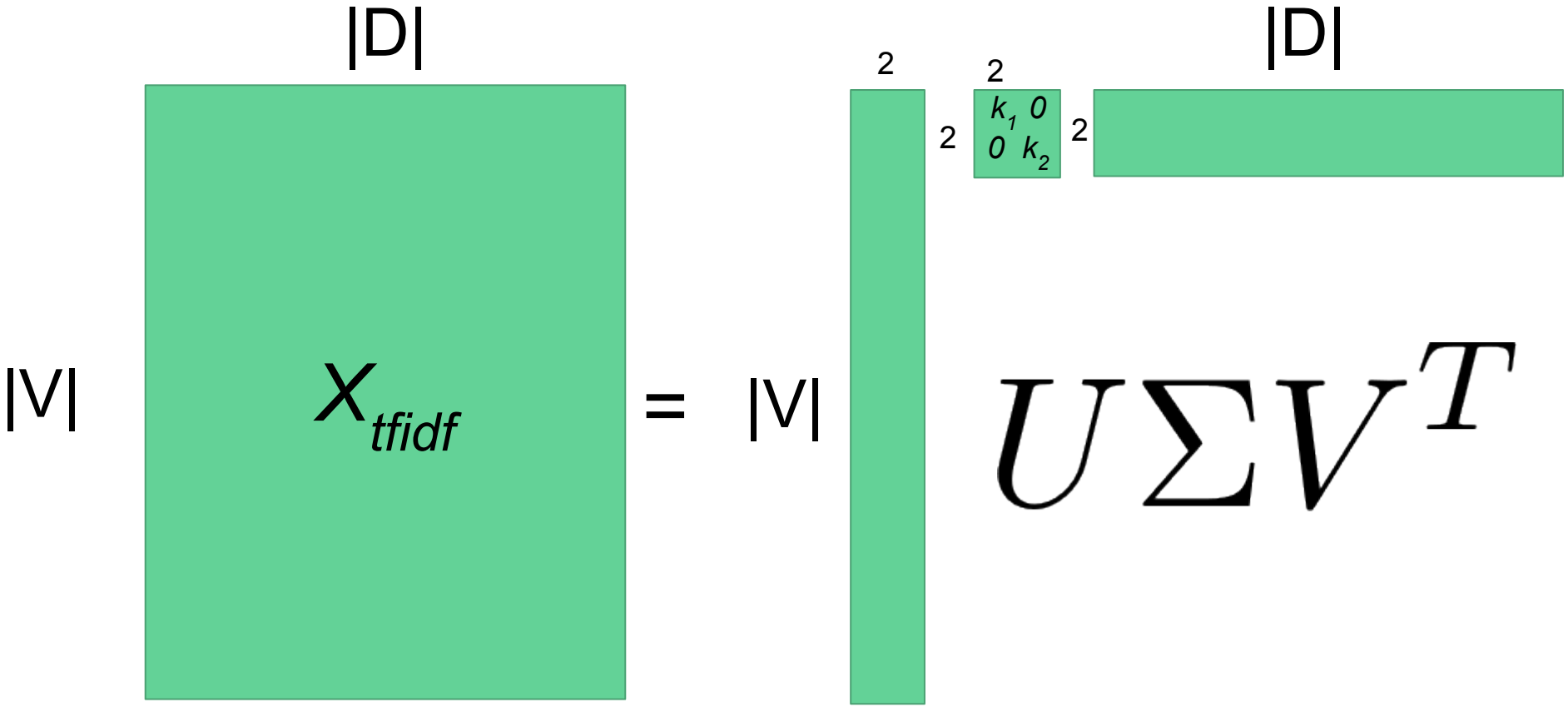
$$|V| \; \overset{|D|}{\boxed{\quad X_{tfidf} \quad}} \; =$$


|D|
|V|
X_{tfidf}
=
1
|D|
1
|V|

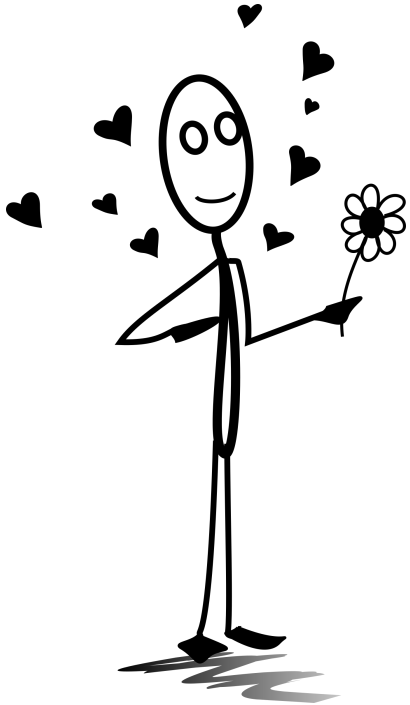

|D|
|V|
X_{tfidf}
=
k
|V|
1
1
|D|

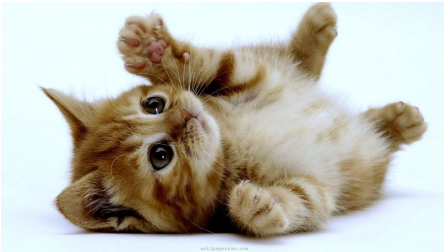

|D|
|V|
X_{tfidf}
=
|V|
1
1
1
k
1
|D|


|D|
|V|
X_{tfidf}
=
|V|
2
2
2
2
k_1 0
0 k_2
|D|



$$U \Sigma V^T$$





$$X = U \Sigma V^T$$

Key questions in unsupervised NLP:

1. How many dimensions does our dataset actually live in?

$$X = U\boxed{\Sigma}V^T$$

2. How do we project our data down to those dimensions?

$$X = \boxed{U}\Sigma\boxed{V^T}$$

Enough talk, time for some magic.

|D|

|V| X_{tfidf} + $U\Sigma V^T$ =

|D|

|V|

$$+ \; U \Sigma V^T \; =$$

Latent Semantic Indexing (LSI)

|D|

|V|

$$+ \; U \Sigma V^T \; =$$

Latent Semantic Indexing (LSI)

=

Latent Semantic Analysis (LSA)

As a side note...

Great first NLP paper to read! Highly accessible :-)

Scott Deerwester

“Indexing by latent semantic analysis” - Deerwester et al. 1990

What are topic models?

What are topic models?

Latent semantic indexing
(Deerwater et al. 1990)

What are topic models?

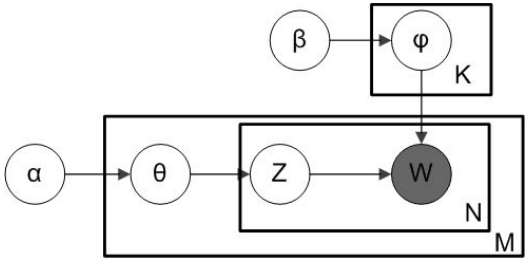
$$U\Sigma V^T$$

Latent semantic indexing
(Deerwater et al. 1990)

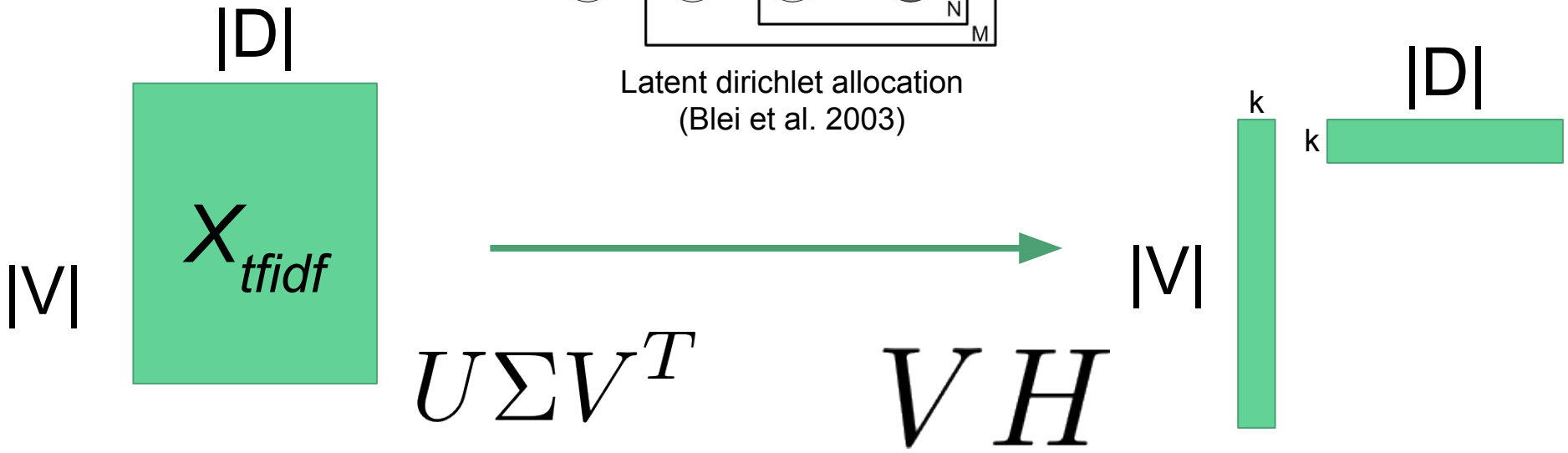
$$VH$$

Non-negative matrix factorization
(Lee and Seung 1999)

What are topic models?



Latent dirichlet allocation
(Blei et al. 2003)



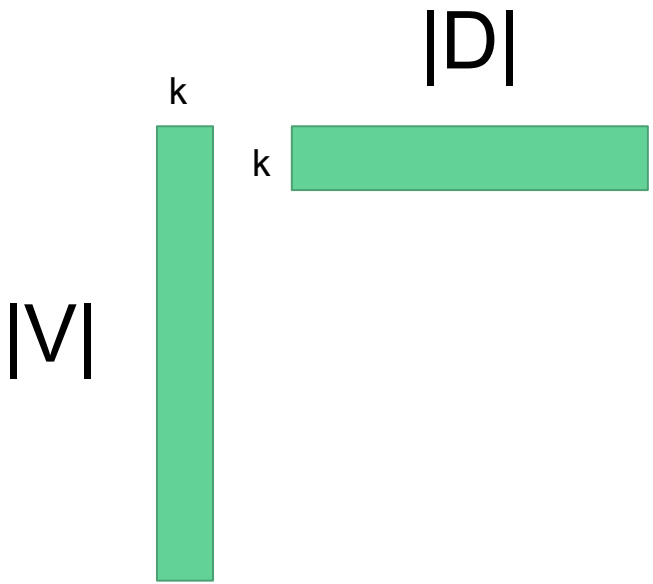
$$U\Sigma V^T$$

Latent semantic indexing
(Deerwater et al. 1990)

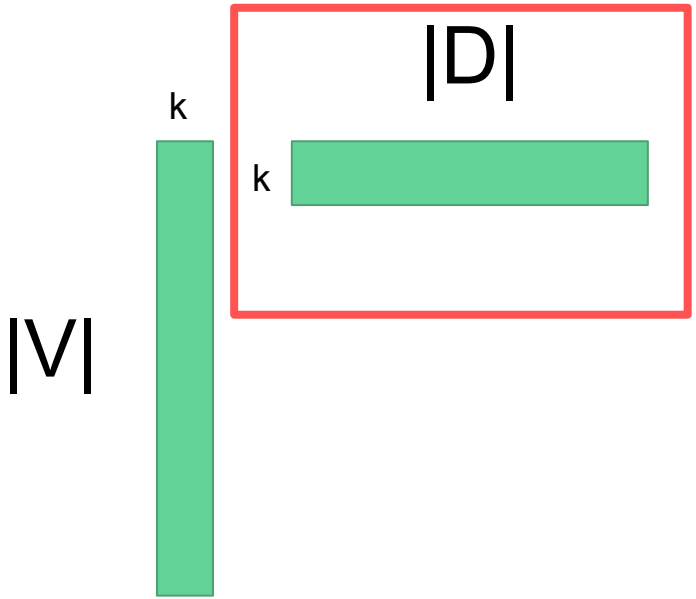
$$VH$$

Non-negative matrix factorization
(Lee and Seung 1999)

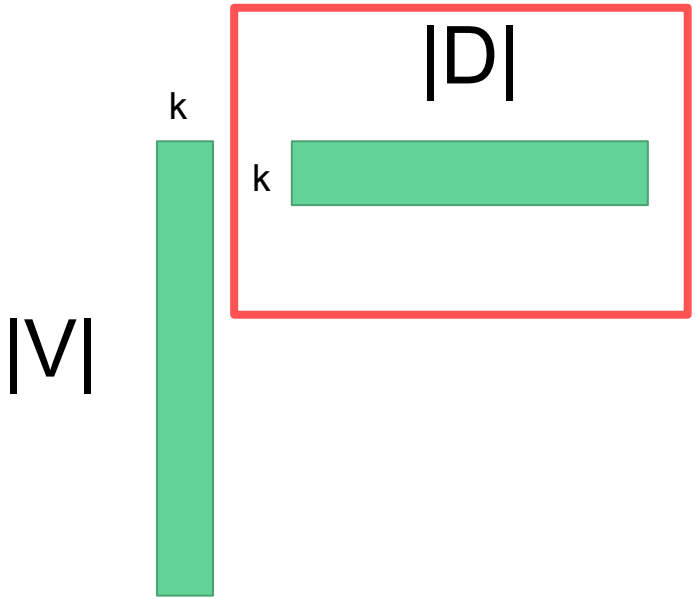
Why do we care??



Why do we care??

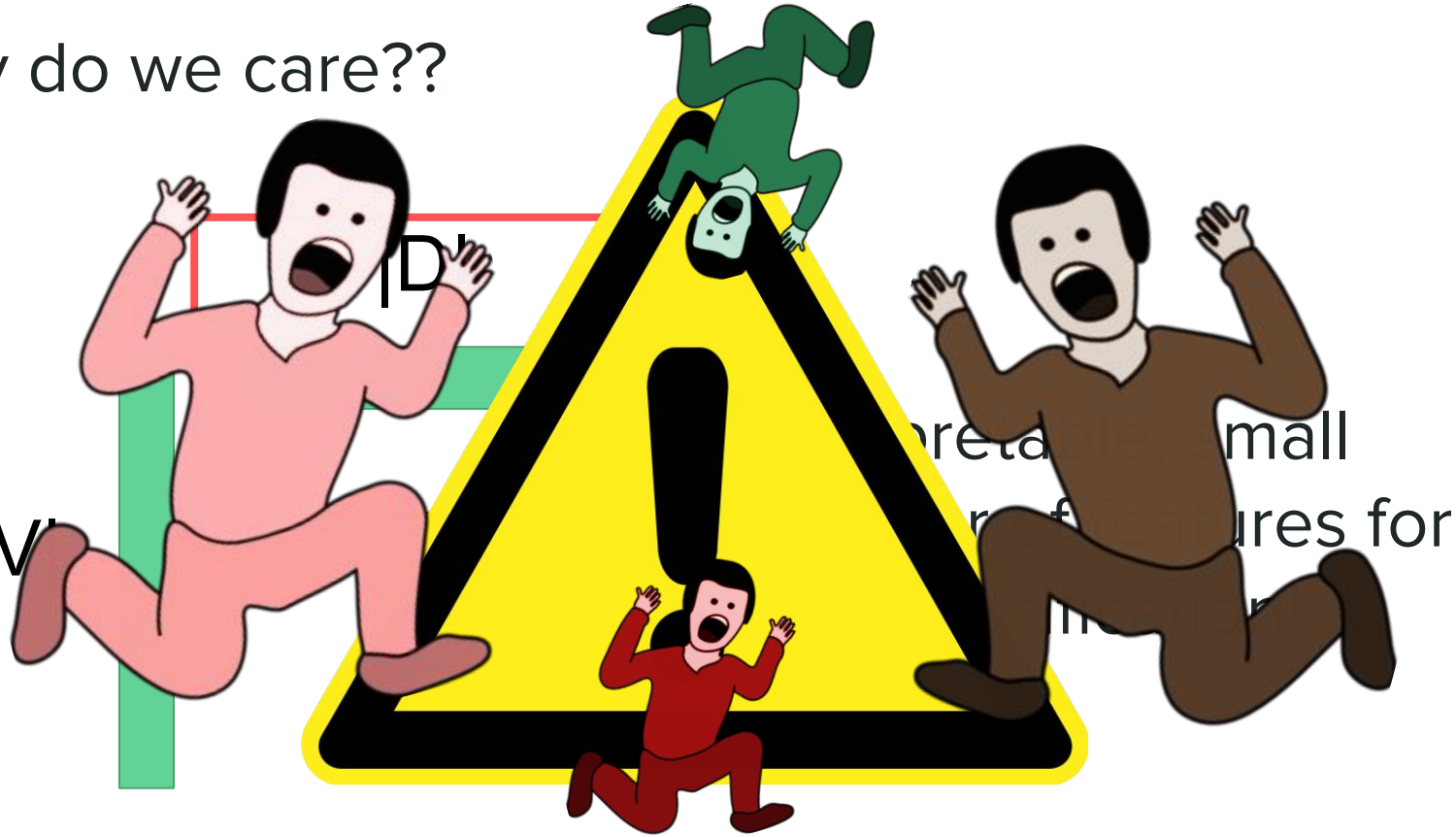


Why do we care??



Interpretable, small number of features for text classification!

Why do we care??



ID

IV

mall

ures for

Document length *matters a lot*

Document length *matters a lot*

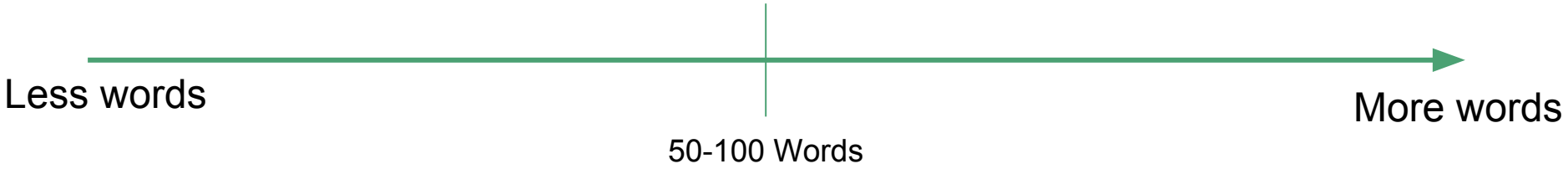
Different regimes of supervised NLP

(Jack’s opinions only! Lots of caveats!)

Document length *matters a lot*

Different regimes of supervised NLP

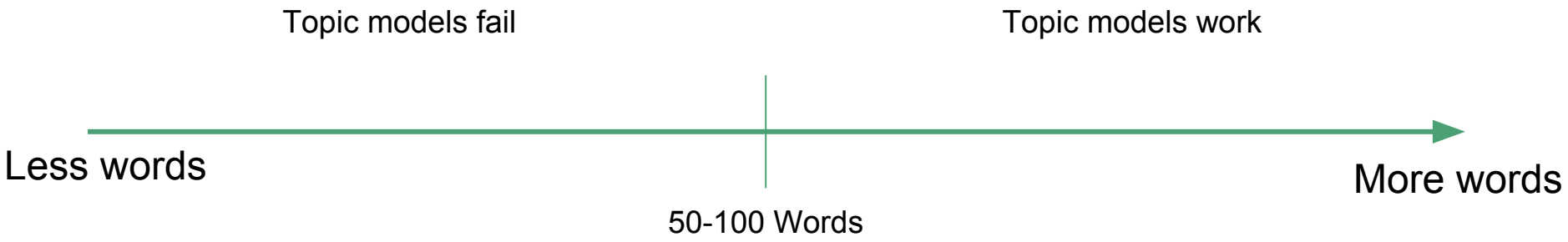
(Jack's opinions only! Lots of caveats!)



Document length *matters a lot*

Different regimes of supervised NLP

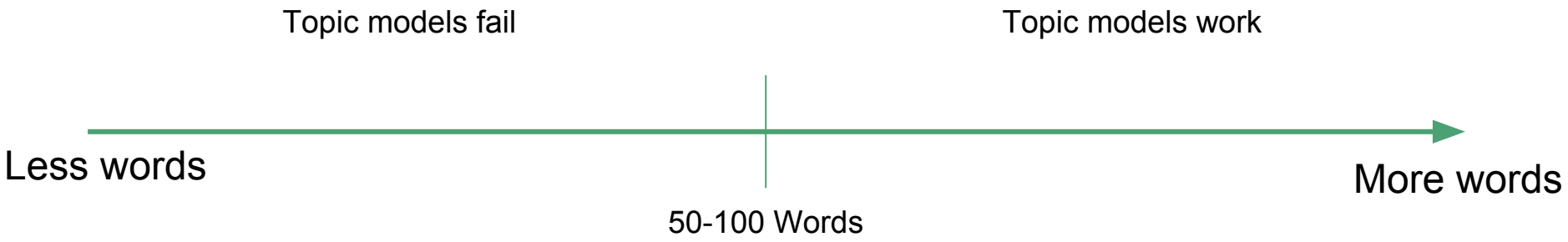
(Jack's opinions only! Lots of caveats!)



Document length *matters a lot*

Different regimes of supervised NLP

(Jack's opinions only! Lots of caveats!)



Naive bayes, n-gram features + linear classifier are almost always pretty good in practice :-)

So can we see if our Kickstarter will be successful?

KICKSTARTER