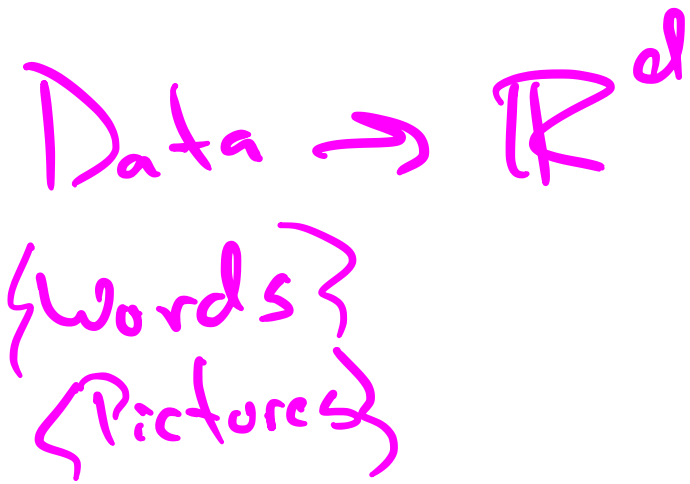
L7: Approximate Nearest Neighbors



Jeff M. Phillips

January 31, 2018

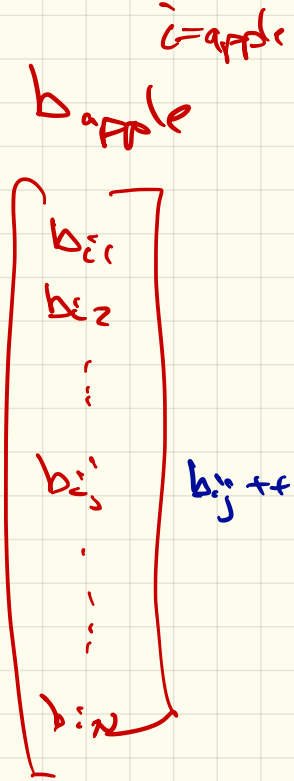
Word Vector Embeddings

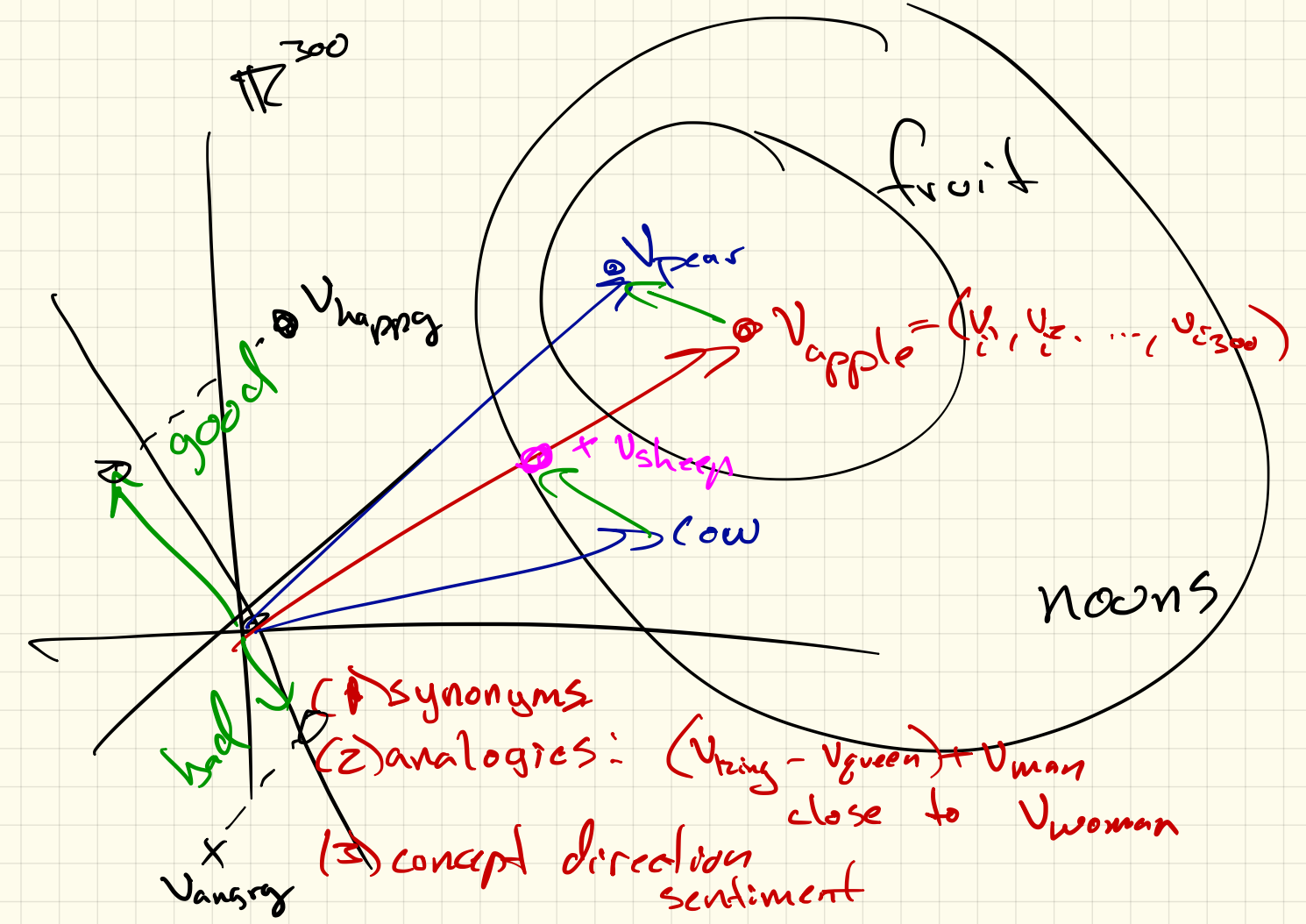
2013

Text
(all of Wikipedia)

"... I ate an apple. It was good ..."

continuous bag of words
CBOW




$\mathbb{R}^{300}$
fruit
v_{pear}
v_{happy}
$v_{apple} = (v_1, v_2, \ldots, v_{300})$
good
+ v_{sheep}
cow
nouns
bad
(1) synonyms
(2) analogies: $(v_{king} - v_{queen}) + v_{man}$
close to v_{woman}
(3) concept direction
sentiment
v_{angry}

PPMI vectors $\mathbb{R}^N$

$N = 1$ million

positive, pointwise mutal inform

b_i vector

$b_i = (b_{i1}, b_{i2}, \ldots b_{ij} \ldots b_{iN})$

$v_i = (v_{i1}, v_{i2}, \ldots v_{ij}, \ldots v_{iN})$

b_{ij} # occurences of w_j in KB(w) of w_i

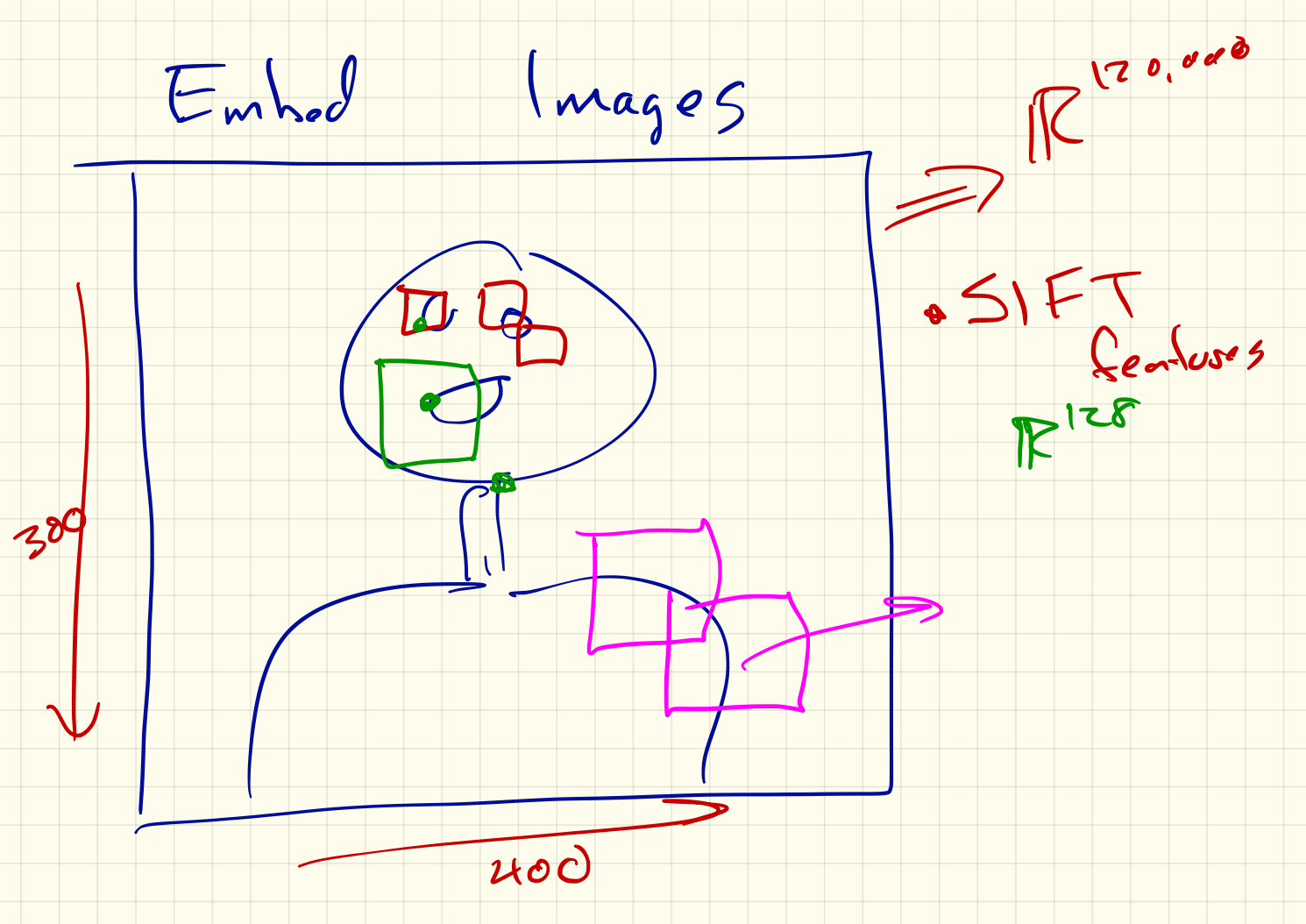
M = Total # words in corpus.

$M = \sum_i n_i$

n_i = # times word i occurs

$p(i) = \frac{n_i}{M}$, $p(i,j) = \frac{b_{ij}}{M}$

$$v_{ij}' = \max\left\{0, \log\left(\frac{p(i,j)}{p(i) \cdot p(j)}\right)\right\}$$

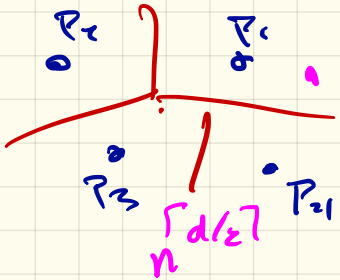

Embed Images
$\mathbb{R}^{120,000}$
• SIFT features
$\mathbb{R}^{128}$
300
400

Images and SIFT Features

N1	N2	N3
N8	X	N4
N7	N6	N5

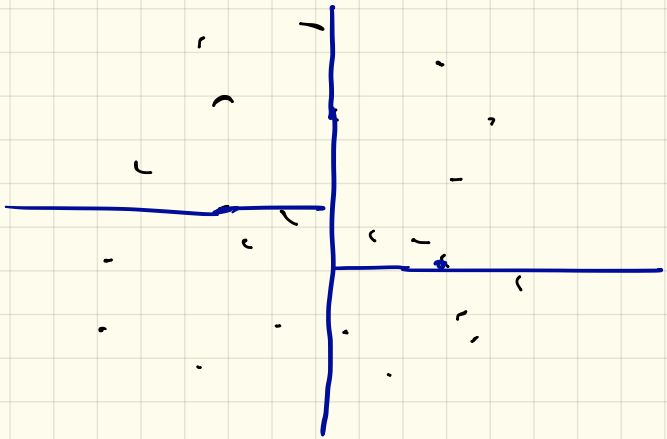
Approx Nearest Neighbors

$\mathbb{R}^1$ Sort Data $\rightarrow$ BST



$\mathbb{R}^{2-3}$ Voronoi Diagram

$\mathbb{R}^{3-10}$ kd Tree

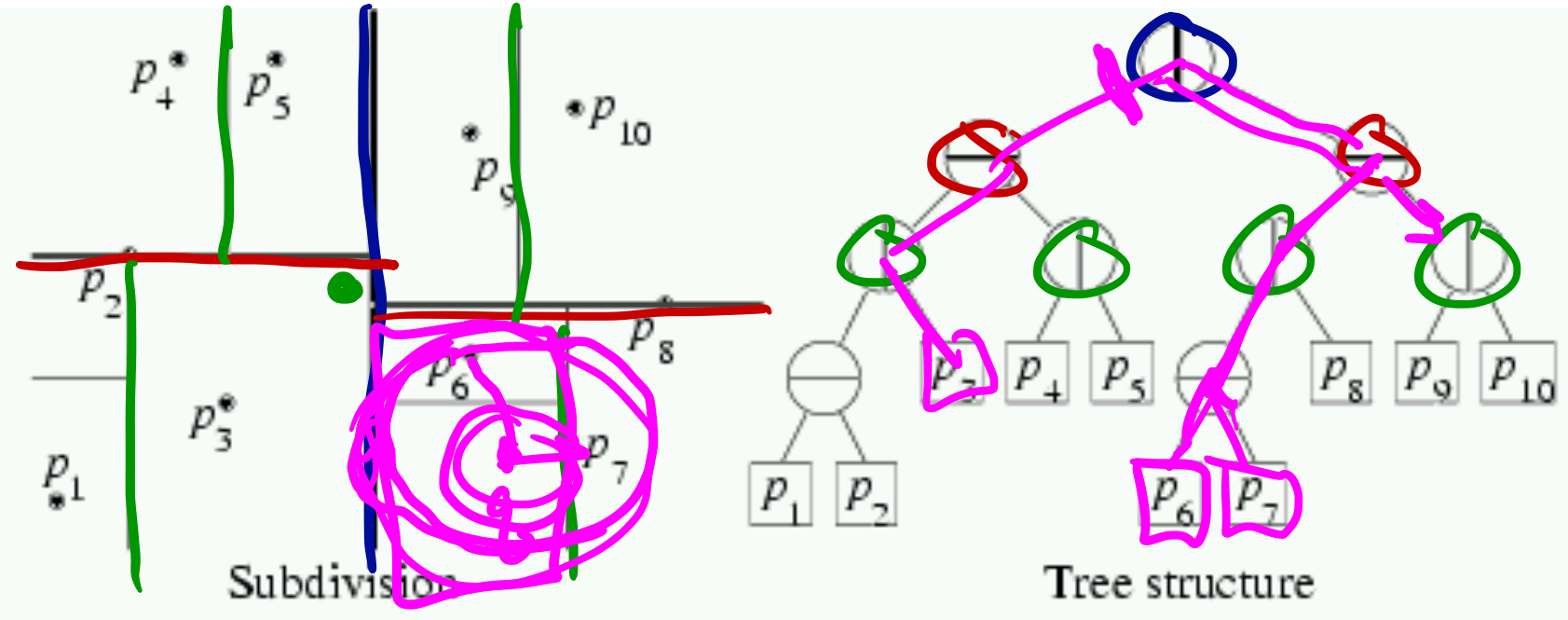


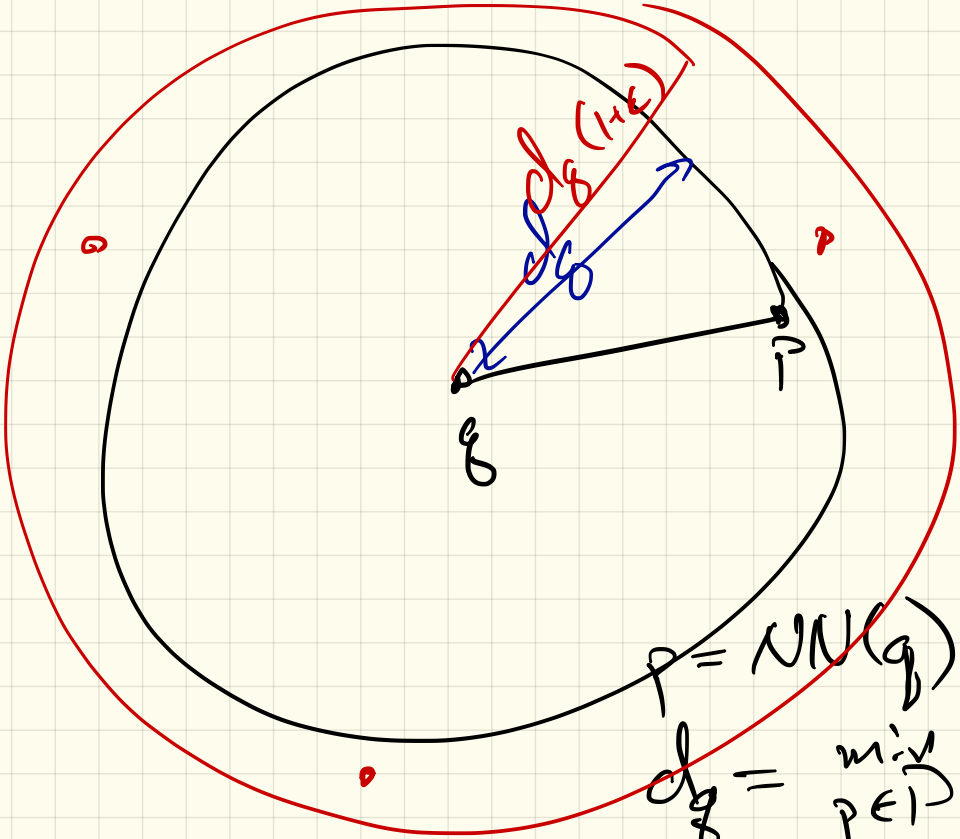
$\mathbb{R}^{10 \rightarrow 100}$

$\downarrow$

LSH

kD-Tree

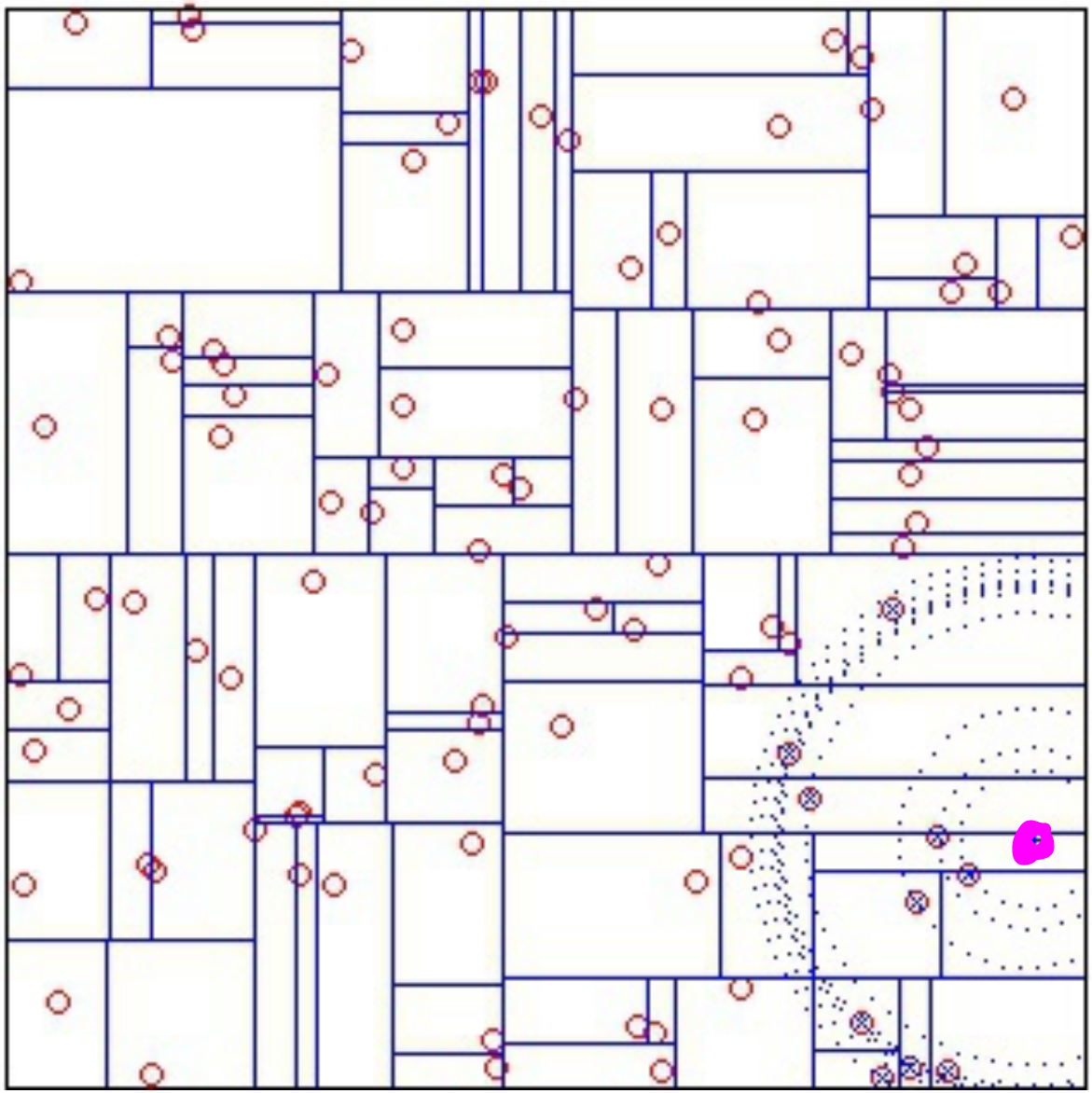

p_4
p_5
p_{10}
p_9
p_2
p_8
p_6
p_3
p_7
p_1
Subdivision
p_1
p_2
p_3
p_4
p_5
p_6
p_7
p_8
p_9
p_{10}
Tree structure



$$p = NN(q)$$

$$d_q = \min_{p \in D} d(p, q)$$

Approximate Queries on kD-Tree



kd-Tree

Split dependent on data.

- PCA → 1 dim
- 2-means cluster

Ethics of Using Word Embeddings

$\mathbb{R}^{300} \rightarrow \mathbb{R}^{299}$