The author gratefully acknowledges support from the taxpayers of Québec and of Canada who have paid his salary and research grants while this work was developed at McGill University, and from his students (who built the implementations and investigated the data structures and algorithms) and their funding agencies.

The "Curse of Dimensionality"

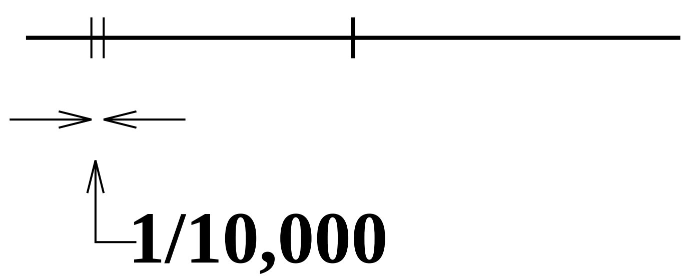
(OLAP, Feature Vectors, ..)

What happens to small activities in many dimensions?

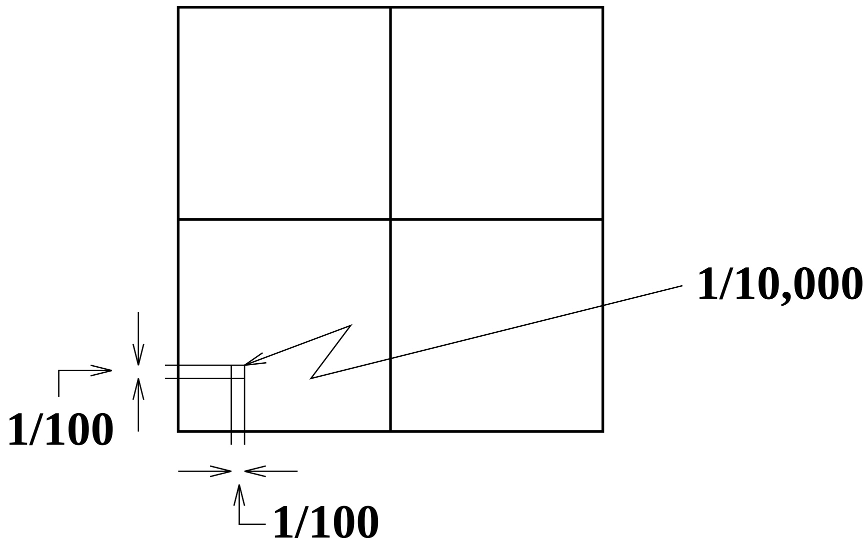
Say $a = 0.0001 = \frac{1}{10{,}000}$
Say $f = 2$ for each dimension.

In 1-D effective activity is 0.5:

$\frac{1}{10{,}000}$



In 2-D effective activity is 0.25: $\frac{1}{100} \times \frac{1}{100}$



In 4-D effective activity is 0.0625: $\frac{1}{10} \times \frac{1}{10} \times \frac{1}{10} \times \frac{1}{10}$

In 16-D effective activity is 1**!**:

$0.56 \times 0.56 \times 0.56 \times 0.56 \times 0.56 \times 0.56 \times 0.56 \times 0.56 \times$
$0.56 \times 0.56 \times 0.56 \times 0.56 \times 0.56 \times 0.56 \times 0.56 \times 0.56$

Note that $a = 0.0001$ is a breakeven activity, e.g., for $R = 100, \rho = 1,000,000$. *Any* a_{eff} over this means use sequential!

Above assumes

1. The range query has same *shape* as the data space.

2. $f_i = f$ and space is hypercube of side 1.

3. The data distribution is the product of the axial distributions.

Can be calculated generally using "fractional ceiling",

ceil$(f, x) = g/f$, where $0 \leq (g-1)/f < x \leq g/f \leq 1$:

$a_{\text{eff}} = (\text{ceil}(f, a^{1/d}))^d$

Activity blowup:
Applies to any d-dim. paging that partitions the axes. Assumes (1) data distribution is Cartesian product, (2) range query, space are hypercubes.

		$f = 2$		$f = 5$		$f = 10$	
d	$a^{1/d}$	n	a_{eff}	n	a_{eff}	n	a_{eff}
1	.0001	2	.5	5	.2	10	.1
2	.01	4	.25	25	.04	100	.01
4	.1	16	.06	625	.002	$1_{10}4$	.0001
8	.31	256	.004	$3.9_{10}5$	.0007	$1_{10}8$	.0007
16	.56	64K	1	$1.5_{10}11$	.0003		.0003
32	.75	$4.3_{10}9$	1		.0008		.0008
64	.87		1		1		.001
128	.93		1		1		1
256	.96		1		1		1
512	.98		1		1		1
1024	.99		1		1		1

N.B. $f = \infty$ (or every field is key): $a_{\text{eff}} = \text{a}$

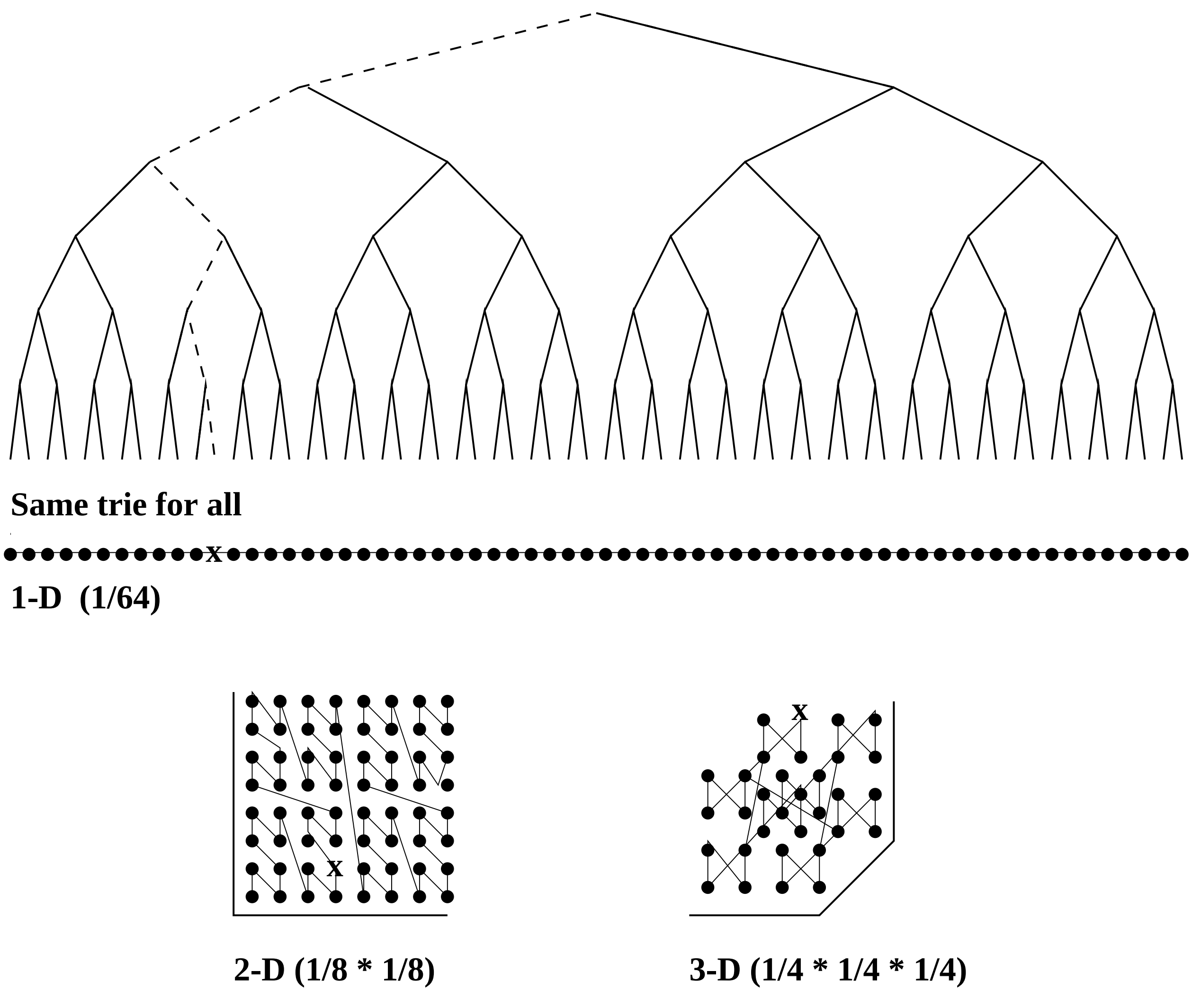
a: activity; a_{eff}: effective activity due to paging;

f: number of page partitions per axis;

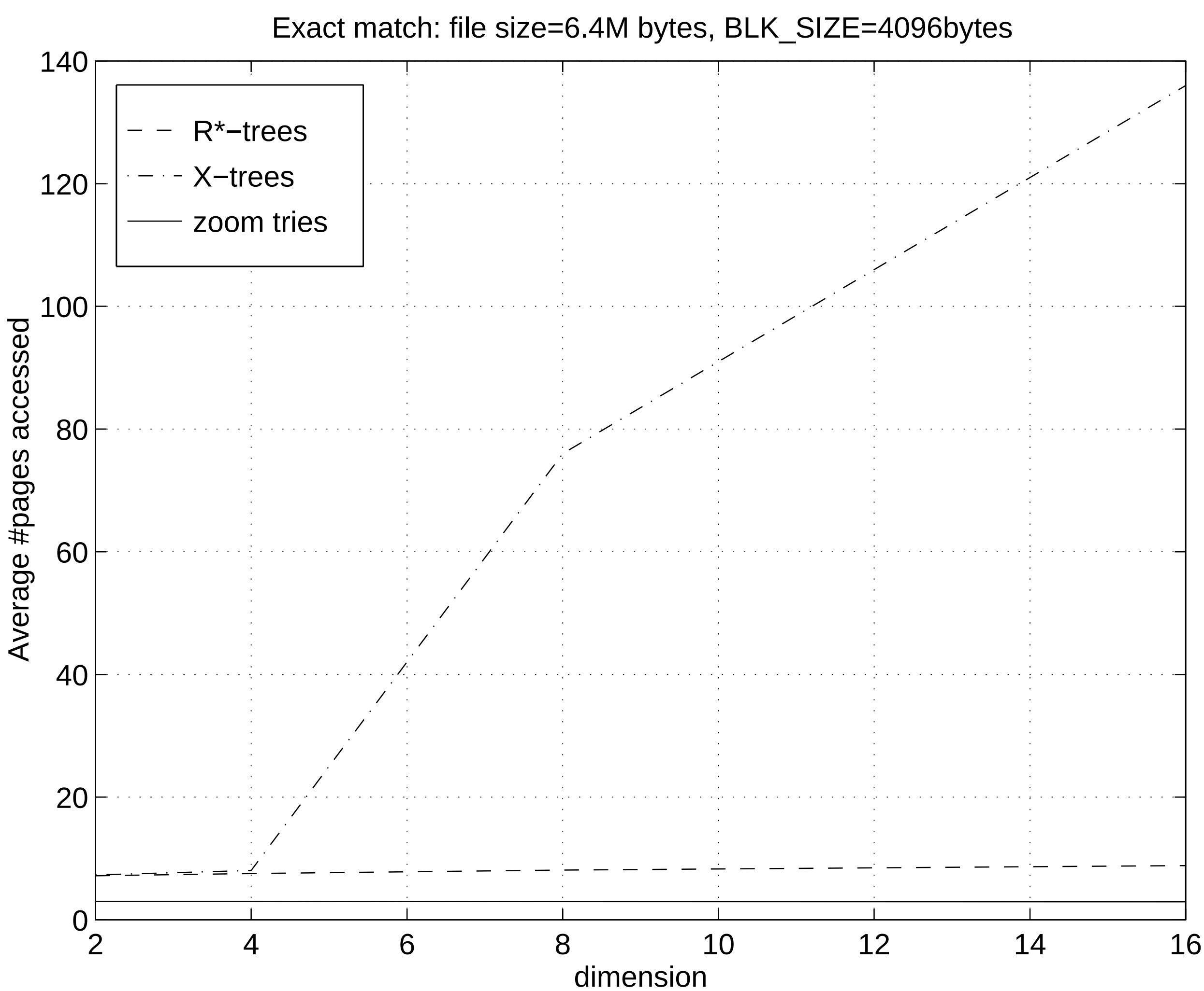
n: number of pages.

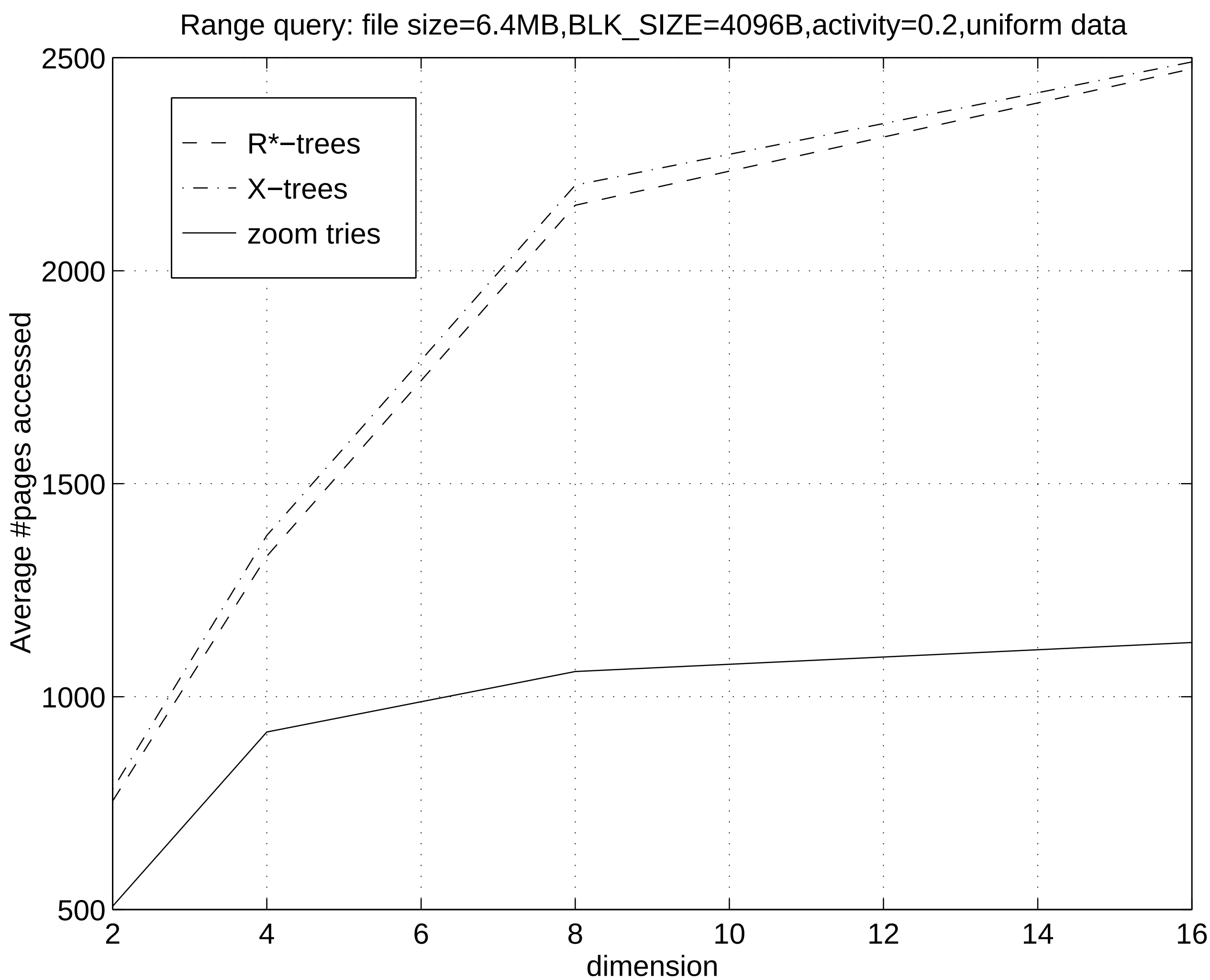
This is a danger (given the three assumptions) for *any* method involving multidimensional grids.

But not for trees. E.g., kd-tries are tries are *one*-dimensional.



Experimental Results on Data Dimensions




Range query: file size=6.4MB,BLK_SIZE=4096B,activity=0.2,uniform data
2500
2000
1500
1000
500
Average #pages accessed
R*−trees
X−trees
zoom tries
2
4
6
8
10
12
14
16
dimension