Numerical Integration Monte Carlo style

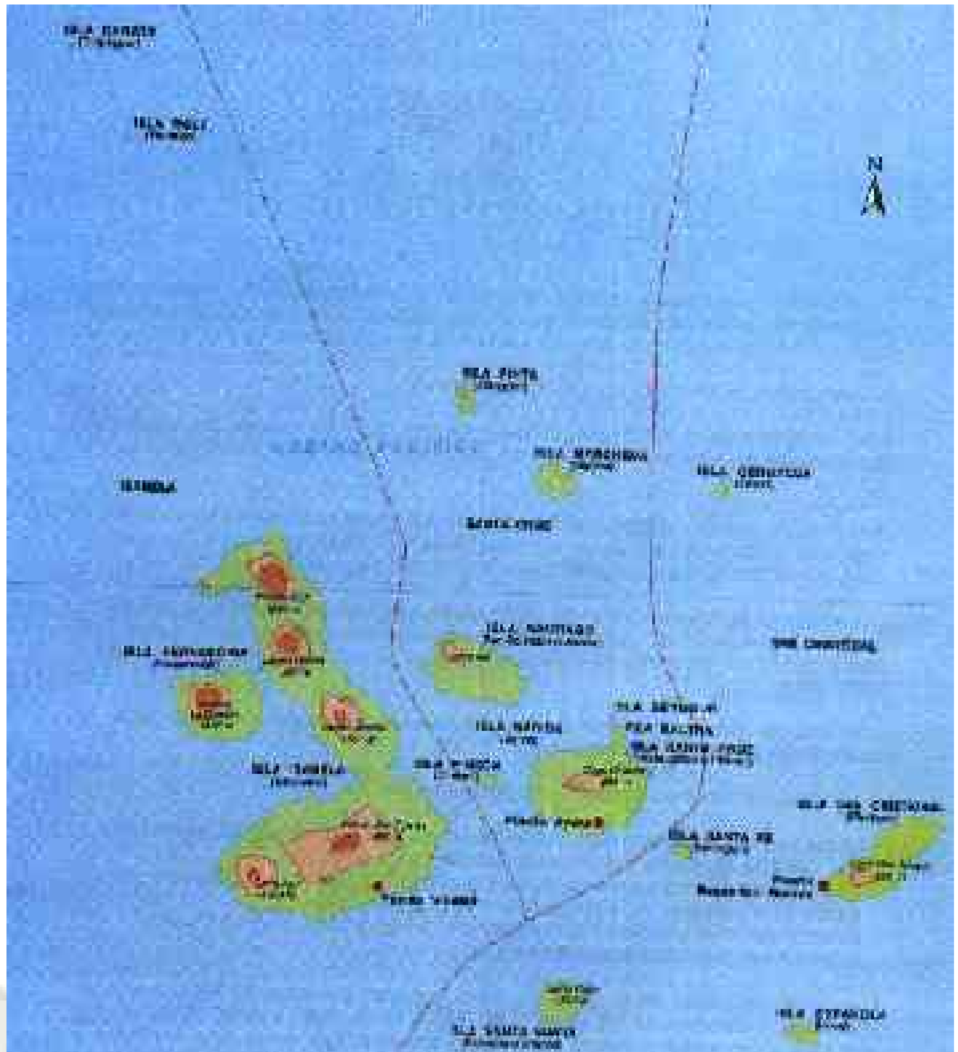
Mark Huber
Dept. of Mathematics and Inst. of Statistics and Decision Sciences
Duke University
mhuber@math.duke.edu
www.math.duke.edu/~mhuber

Integration is hard

Nature laughs at the difficulties of integration.

Pierre-Simon de Laplace

Darwin visited the Galapagos in 1835



Finches

Darwin noted 14 species of finches

(these 11 photographed by Dr. Robert Rothman)

Not all finches on all islands!

	A	*B*	*C*	*D*	*E*	...	*Sums*
large ground	0	0	1	1	1		14
medium ground	1	1	1	1	1		13
small ground	1	1	1	1	1		14
sharp-beaked	0	0	1	1	1		10
...							
sums	4	4	11	10	8		

14 types of finches, 17 islands

The Question

Is this data random?

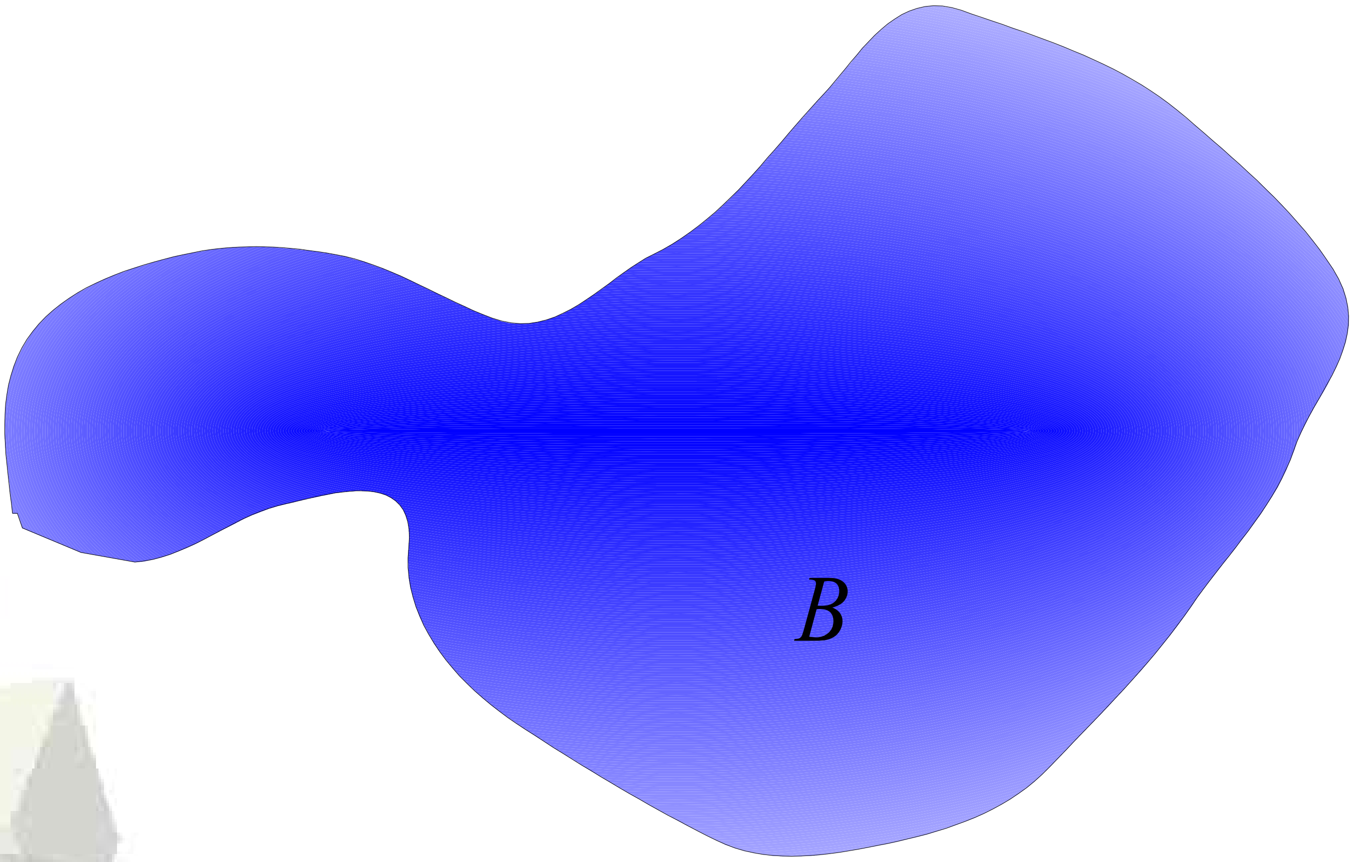
Or is it evidence of evolution?

- To answer deterministically, sum over all tables with same row and column sums

$$2.2 \times 10^{16} \text{ tables!}$$

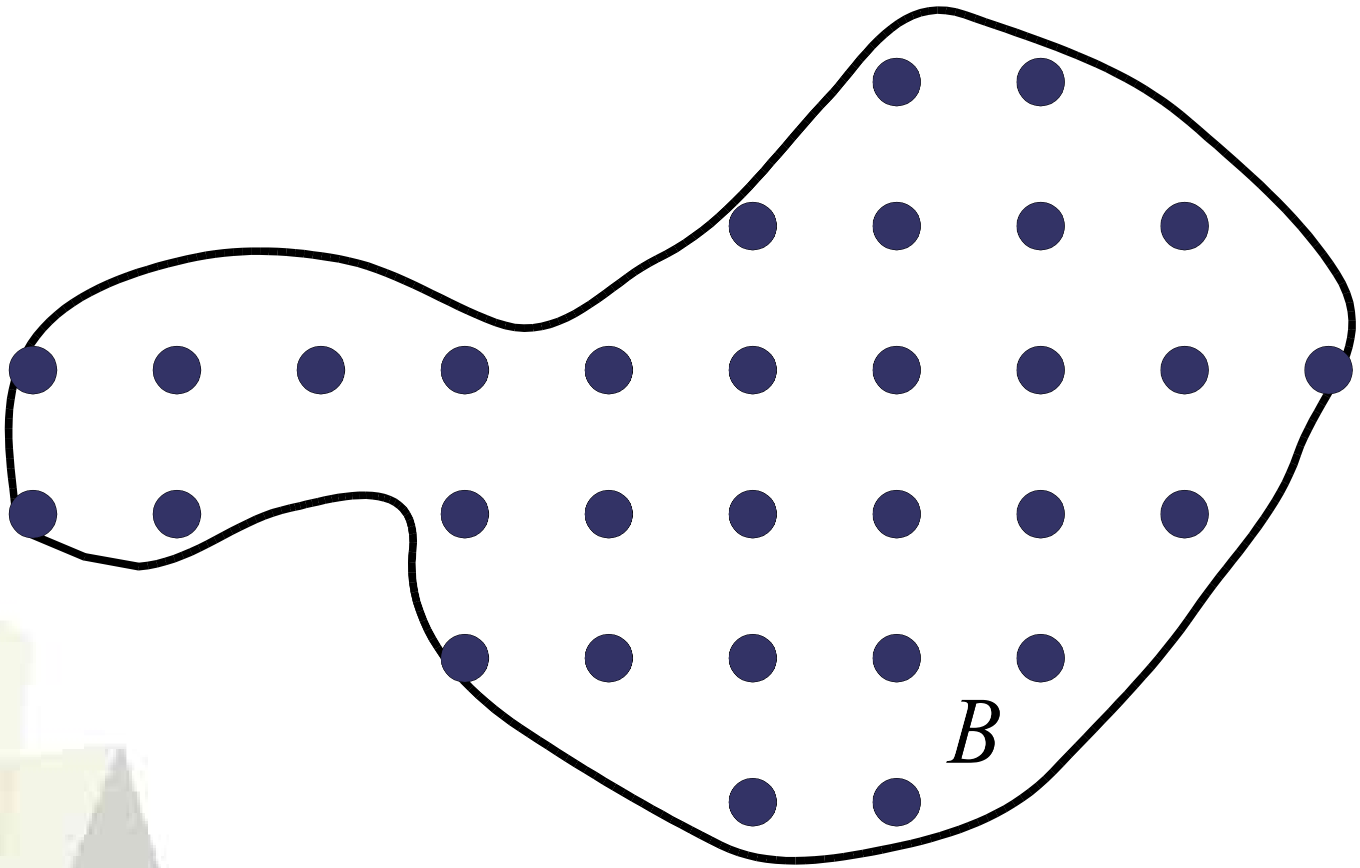
The Oldest Problem

What is the area of B *?*



Counting versus Integration

How many integer points in B *?*



Why is this hard?

These problems have very high dimension

Examples

- Statistical problems
  - dimension is number of data points
- Network (graph) problems
  - dimension is number of nodes
- Physics problems
  - dim. is number of interacting entities

Curse of Dimensionality

Deterministic methods exist

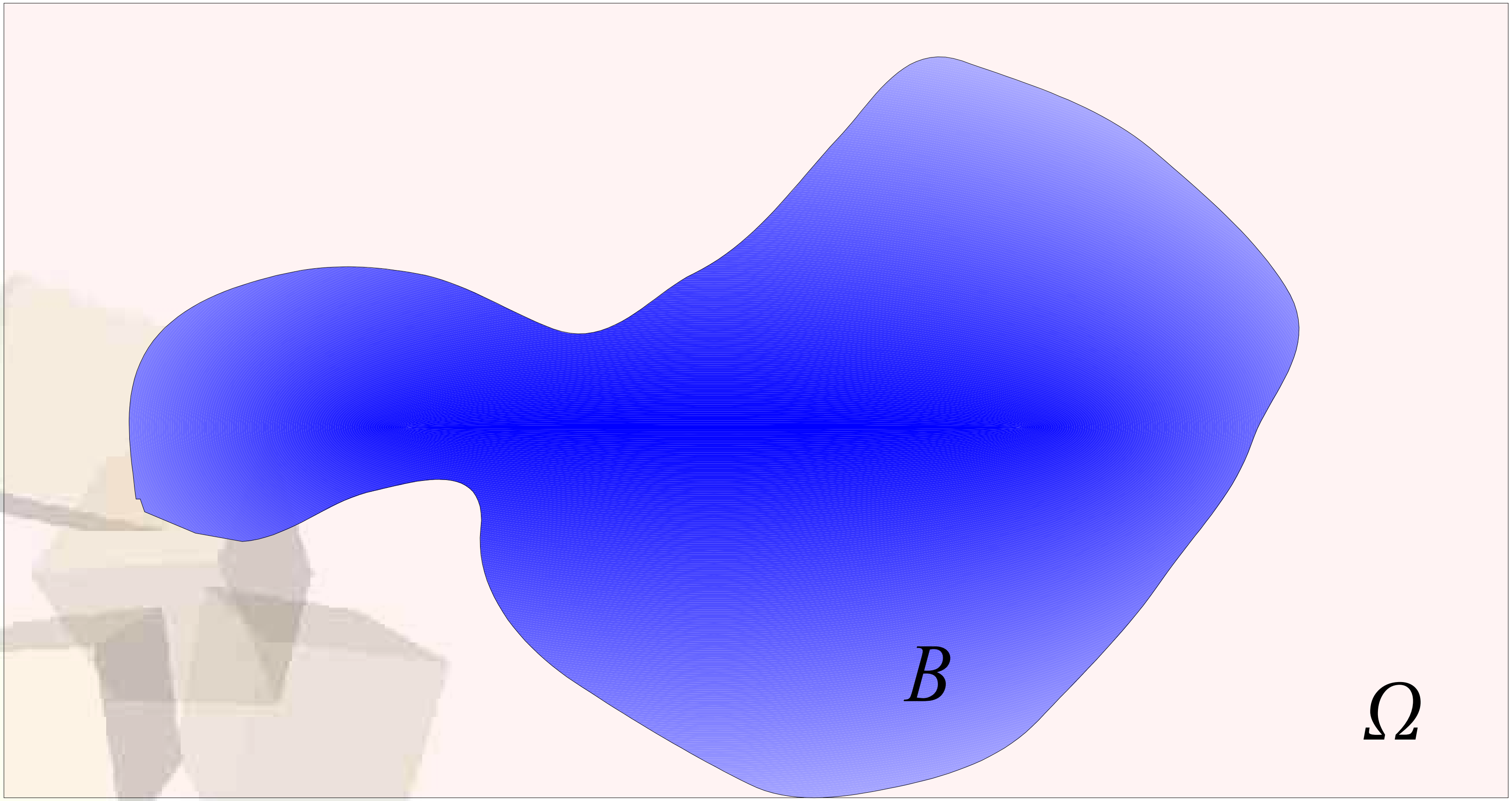
- Directly count the integer points
- Running time grows exponential with dim.
- Trapezoidal Rule, Simpson's Rule, etcetera
  - Effectively reduce dimension by 1

#*P* hard

- Counting the proper colorings of a graph
- Counting Hamiltonian cycles in a graph

Acceptance/Rejection

1) Generate samples from bounding region
2) Find percentage lie in B
3) Multiply by area of bounding region



Why rarely used

The Problem

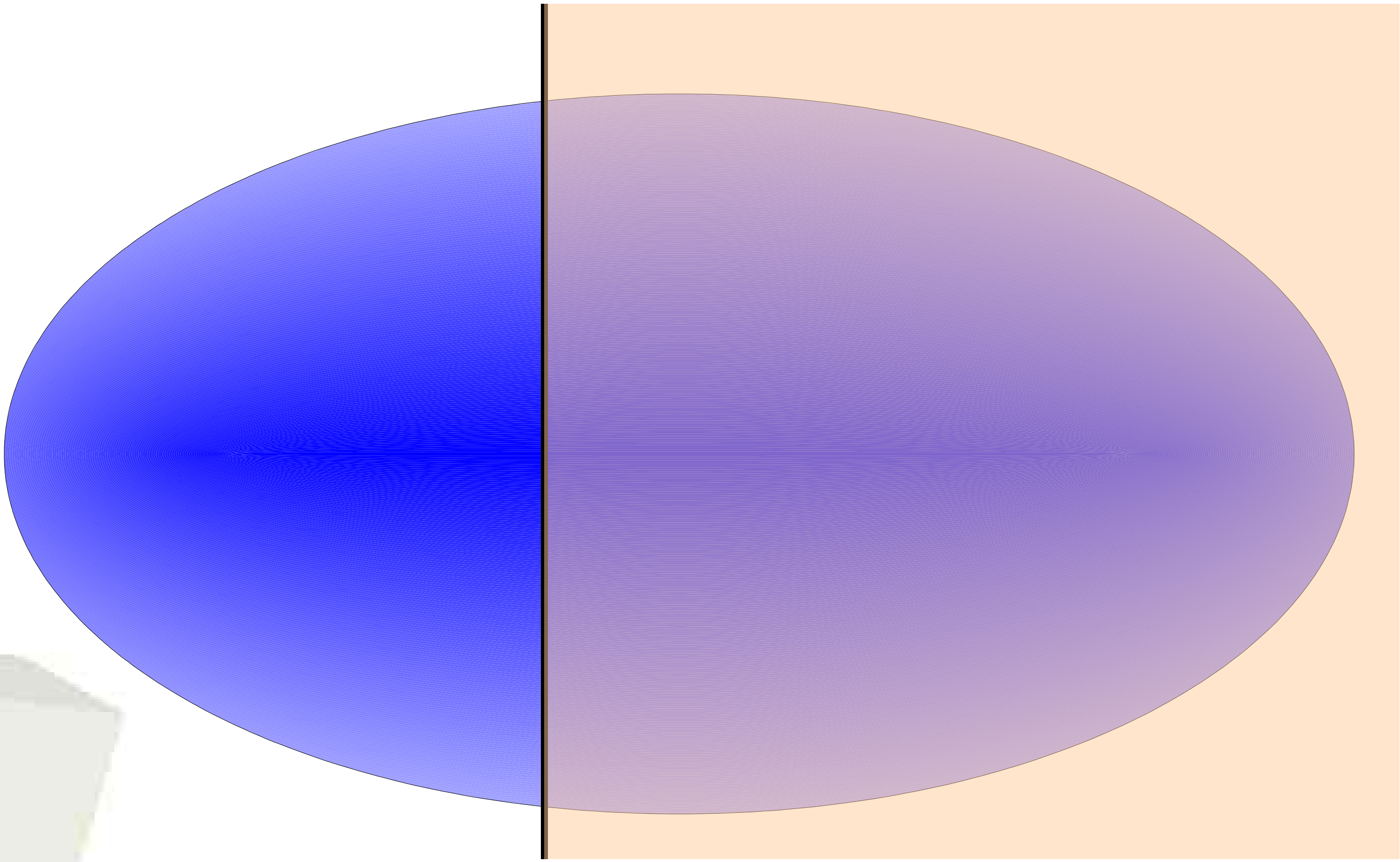
- Need “tight” bounding box
- Otherwise need lots of samples for good estimate
- Difficult to get in high dimensions

Research Area #1

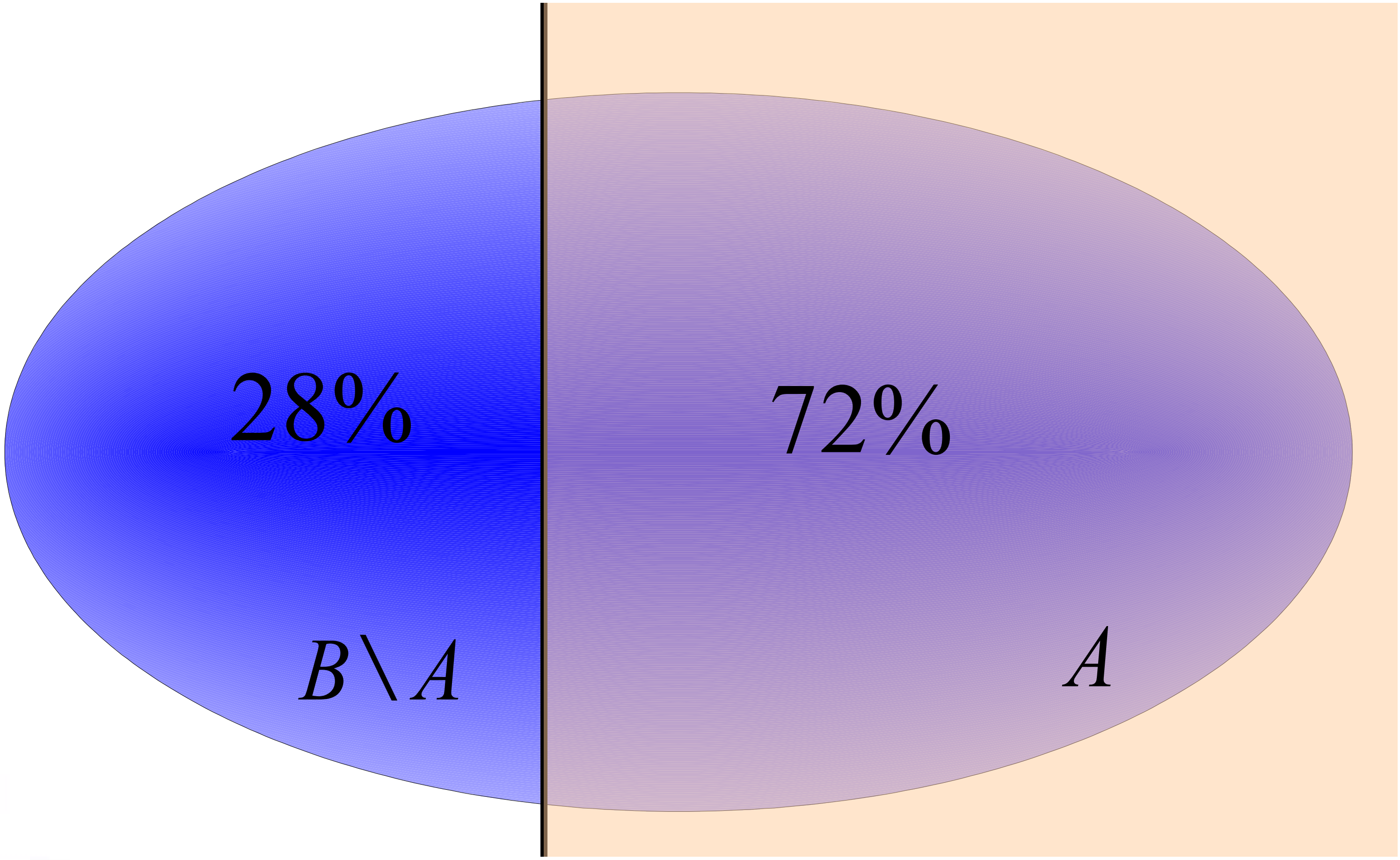
Find good bounding boxes for actual high dimensional problems of interest.

Many times, problem reducible

- Jerrum, Valiant, Vazirani, 1986
- Example: convex regions



Estimating volume

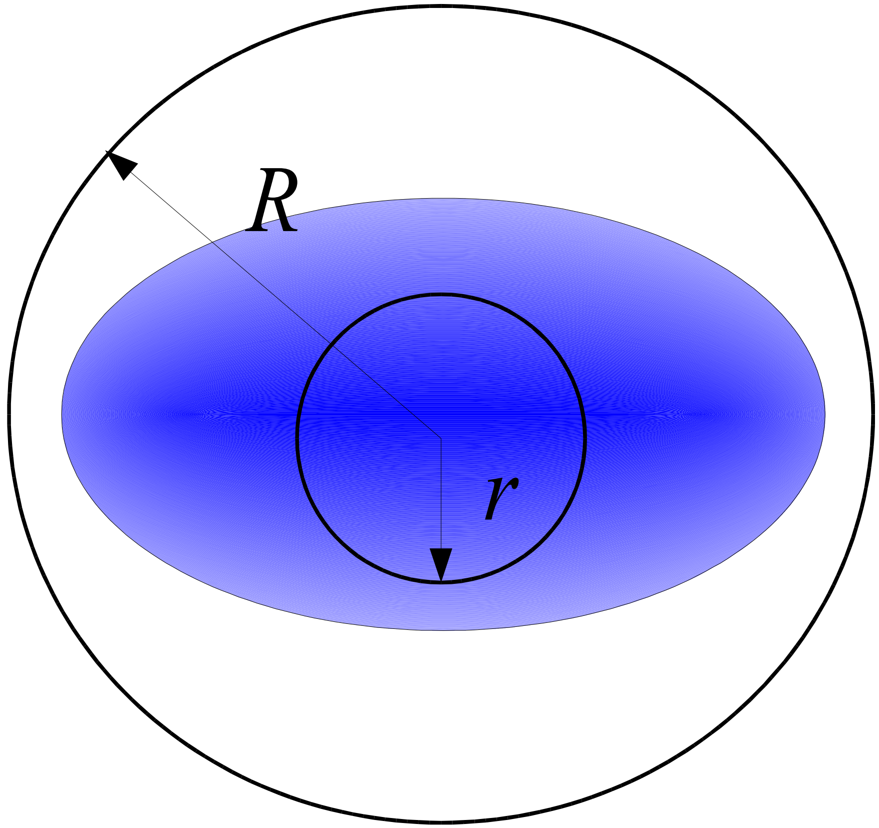


$$\mathrm{vol}(B) = \mathrm{vol}(A) \times \frac{\mathrm{vol}(B)}{\mathrm{vol}(A)}$$

Estimate $\mathrm{vol}(B)/\mathrm{vol}(A)$

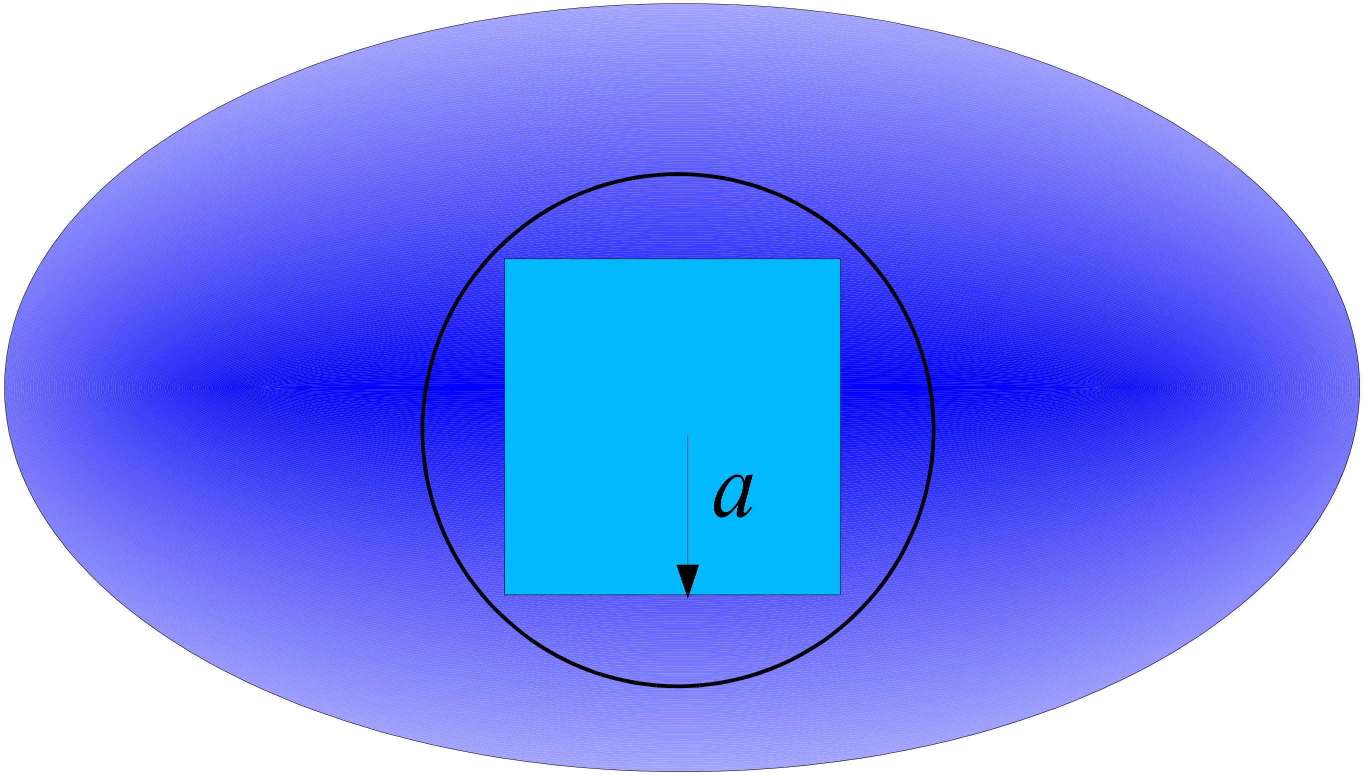
Suppose convex and fairly nice

(even with this help, can‘t come within factor of 2 efficiently with deterministic methods [Elekes 86])



$$\rho = \frac{r}{R} \text{ large}$$

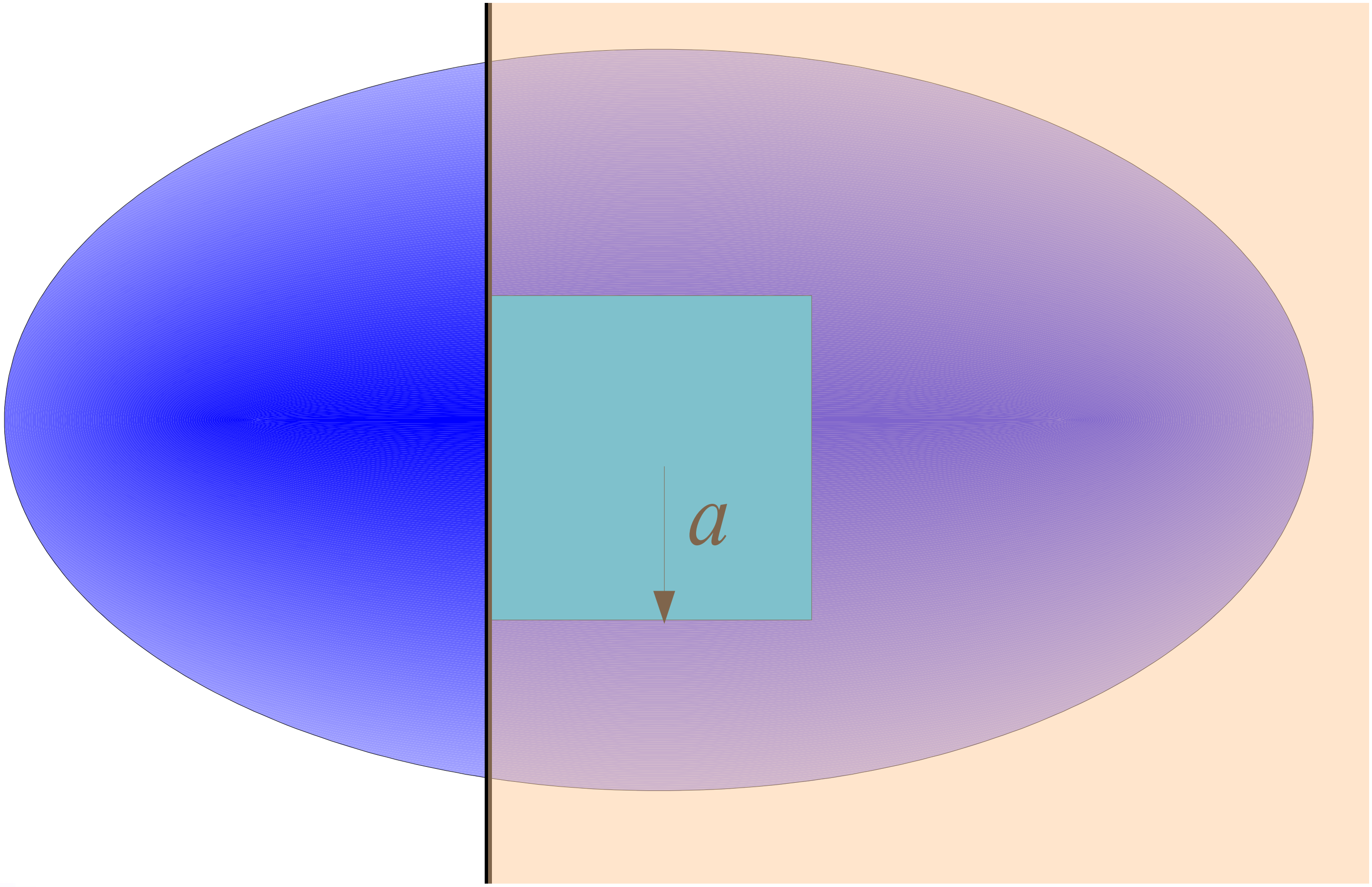
Center box



Inside inner ball, box half edge length

$$a = r / \sqrt{(\dim)}$$

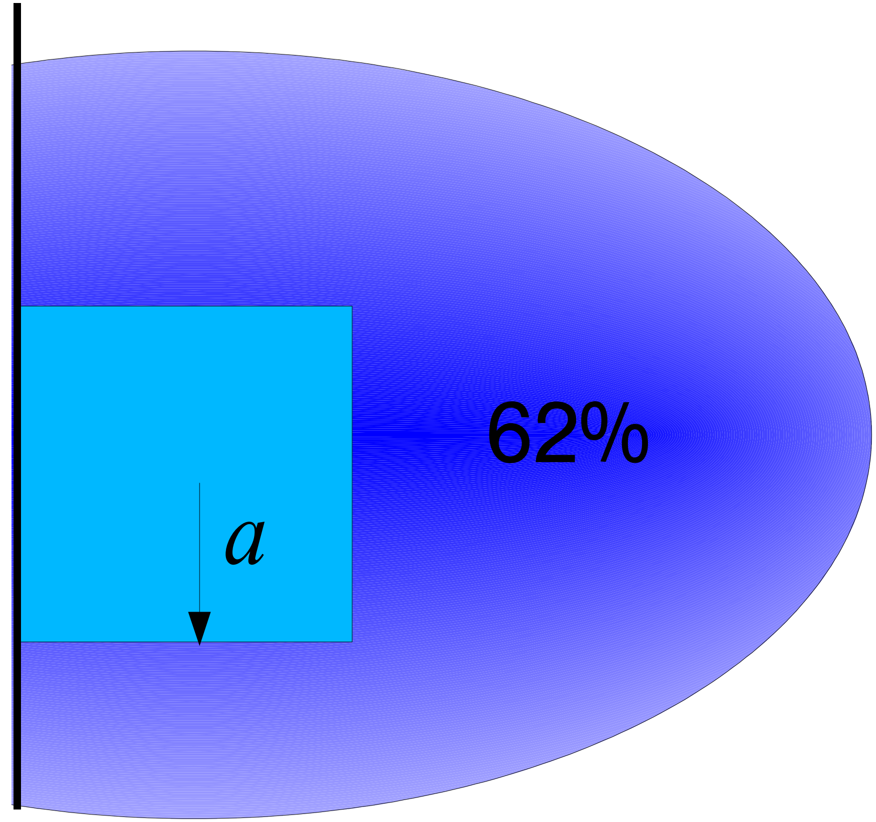
Slicin' and Dicin'



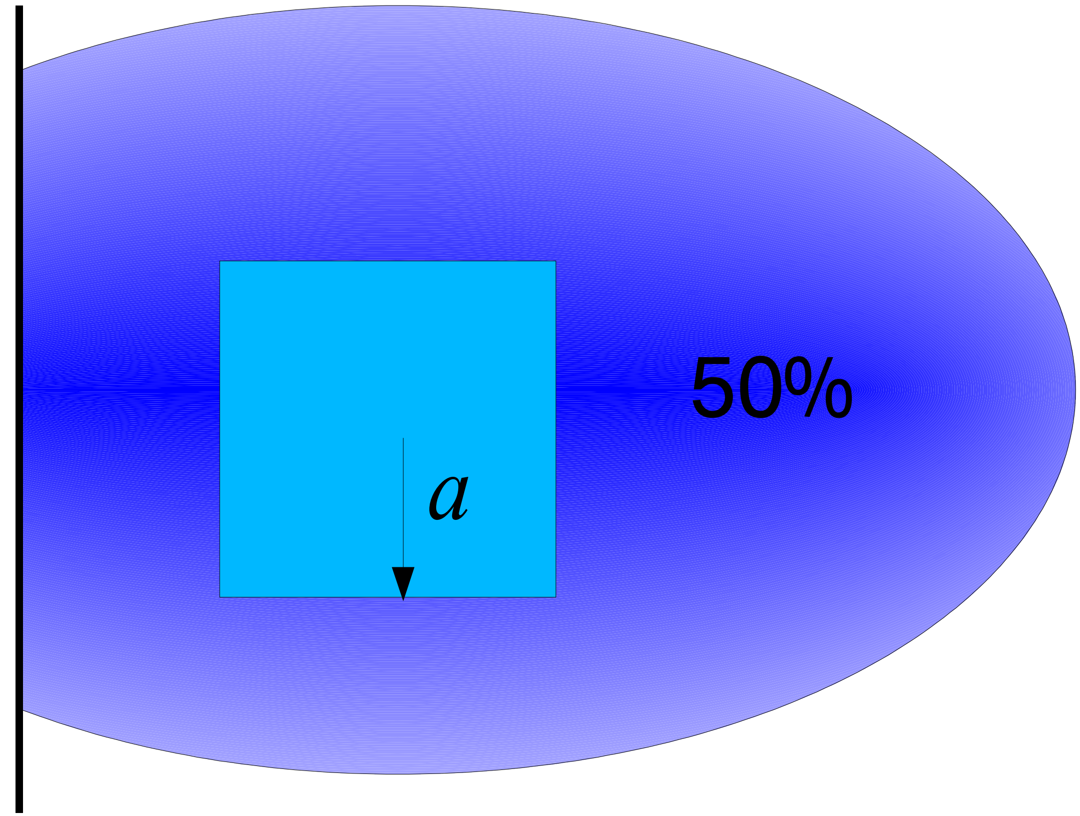
Slice off region to right of box

- Generate lots of random samples
- Estimate percent of area in sliced region

Case I

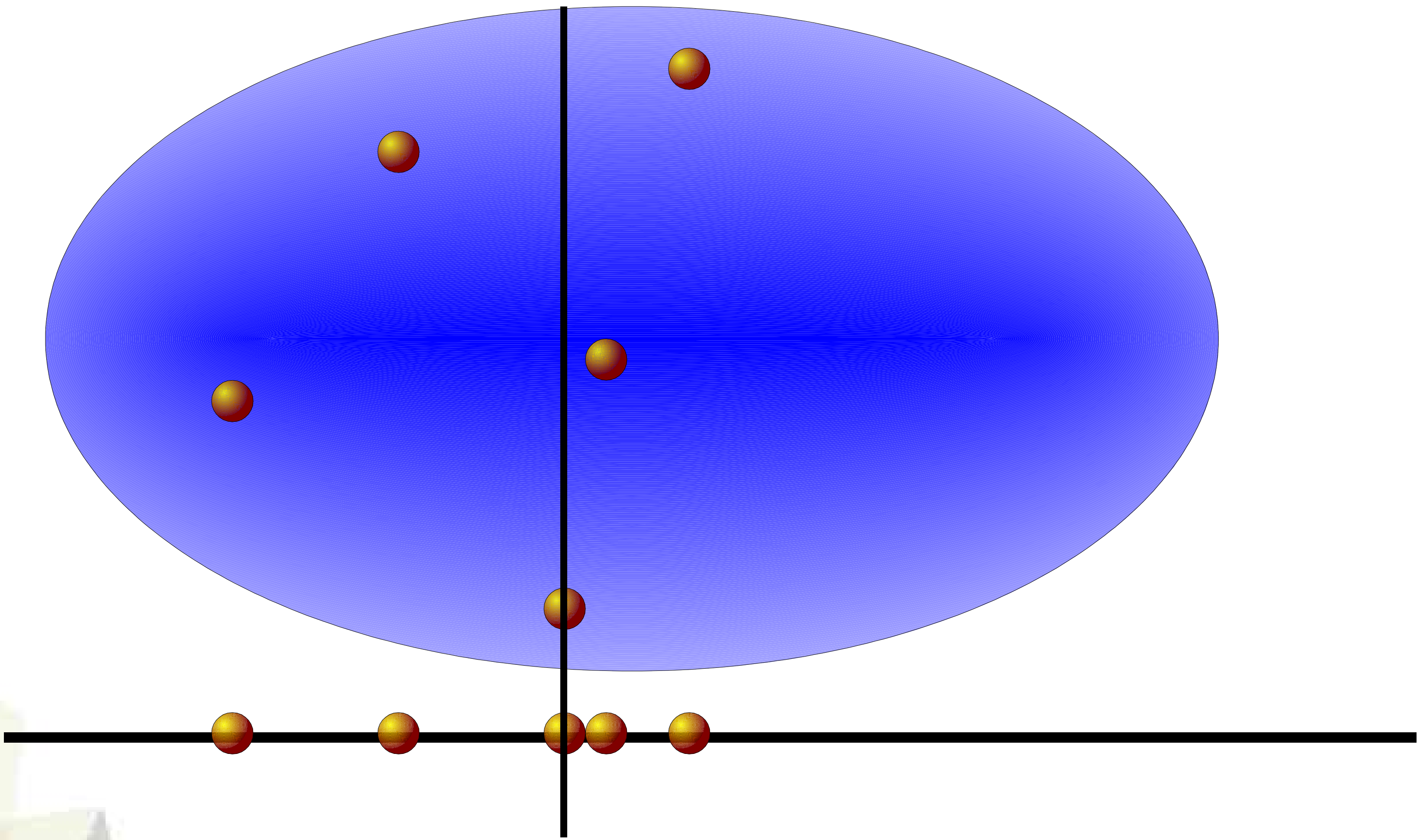


If region with box at least 50%
use as reduced problem



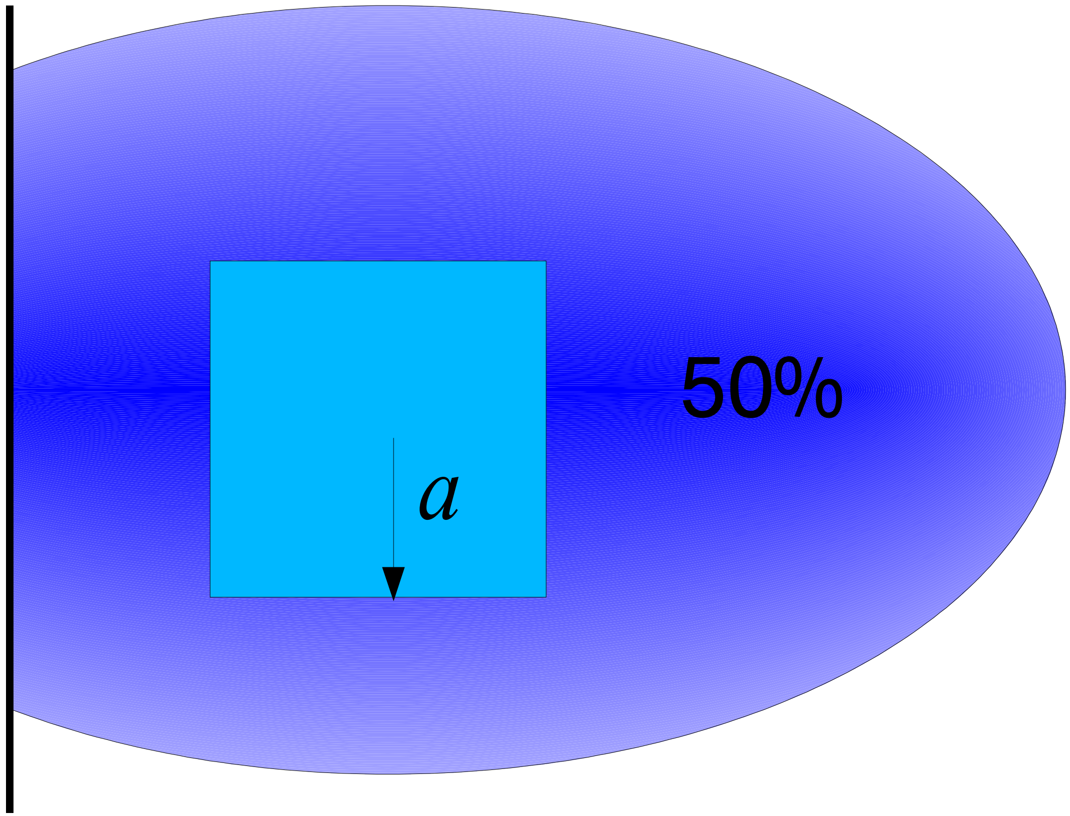
Else
find median, use that instead

Approximating the median



1) grab samples from body
2) project onto one dimension
3) take median of projections

Either way...



Either

1) Match one facet of box or
2) Volume of body reduced by 1/2

How many steps?

Note

$$(2R)^{\dim} \geq \mathrm{vol}(\text{original } B)$$

Volume of body after many steps

$$(2R)^{\dim}(1/2)^{n} \geq \mathrm{vol}(B \text{ after } n \text{ steps})$$

For center box

$$\mathrm{vol}(\text{center box}) = (2a)^{\dim} \geq \left[2R\rho/\sqrt{(\dim)}\right]^{\dim}$$

So most steps that can be taken

$$M := 2d + d(\log(d/\rho))/\log(2)$$

How many samples?

To get median need [Cohen 97][Huber 98]

$$O(\log(1/\delta)/\epsilon^2) \text{ samples}$$

To get within ϵ of answer with probability $1-\delta$

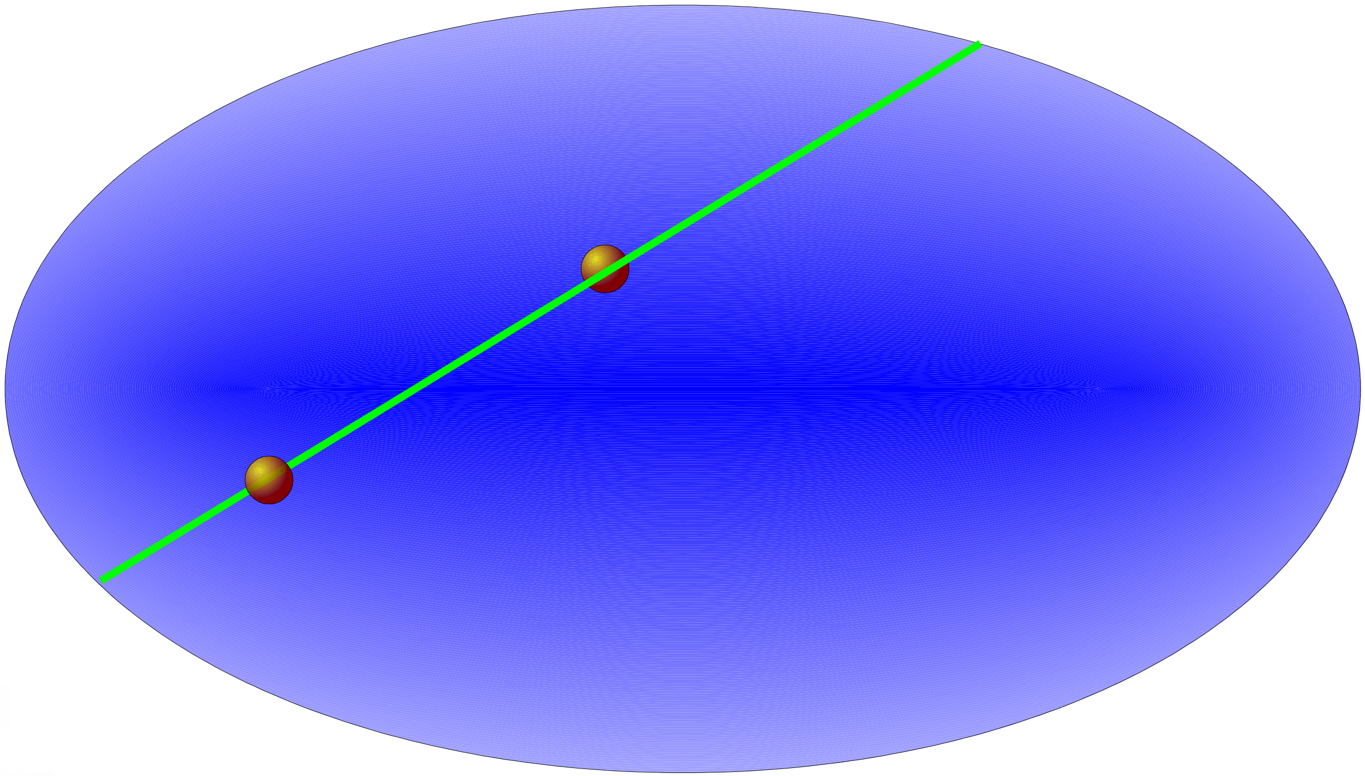
Overall, if M steps taken need $\epsilon' = \epsilon/M$

$$O(M^3 \log(M/\delta)) \text{ total samples}$$

$$O(\dim^3 \log^2(\dim/\delta)) \text{ total samples}$$

Polynomial in the dimension!

Most used method: Markov chains



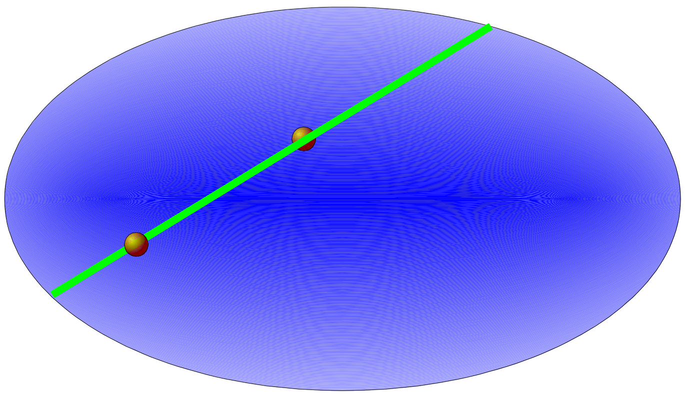
Pick a direction uniformly at random
Move to a uniform point staying inside body

$O(\dim^7)$ time [Kannen, et. al. 94]

Some questions

Research Area #2

Can bound for Markov chain be improved?



Originally $O\left(\dim^{27}\right)$ steps

Research Area #3

Can perfect sampling methods be used for this problem?

Currently working on

Some of my current research questions:

Data from unknown mixtures of distributions

(ex: responders versus nonresponders to drugs)

Perfect matchings in a graph

(ex: astronomical data is doubly truncated)

Multinormal distribution on positive orthant

Contingency tables with extra constraints

(ex: perhaps columns represent age)

The many worlds version of the Ising model

Self organizing lists

(because who has time to organize their own lists?)

The power of Monte Carlo

Monte Carlo methods are the only known way to handle high dimensional numerical integration

Many interesting questions remain:

- Better envelopes for acceptance/rejection
- Better Markov chains
- Perfect sampling algorithms instead of MC