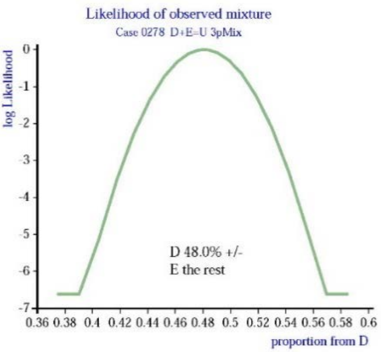
Onetime pure mathematician corrupted by exposure to APL loses moral compass and discovers, after several mis-steps, a useful numerical integration method

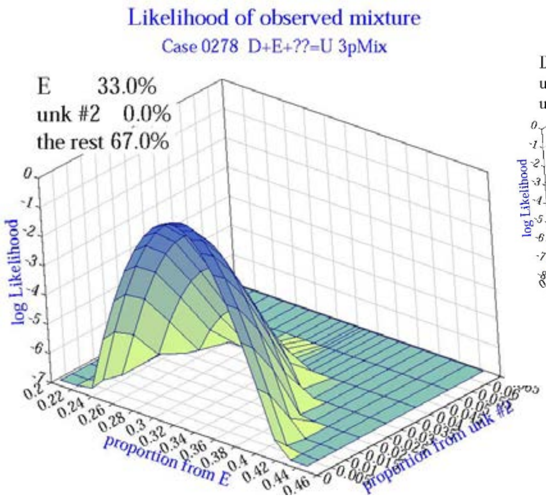
Lesson from DNA Mixture Solution™
program development

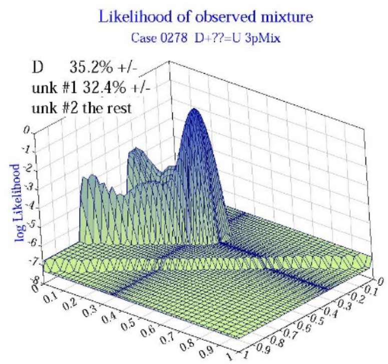
Charles Brenner, Ph.D.
Purveyor of forensic mathematics, **DNA·VIEW®**,
Senior Research Fellow at UC Berkeley Human Rights Center
http://dna-view.com c@dna-view.com **+1 510 798 7139**

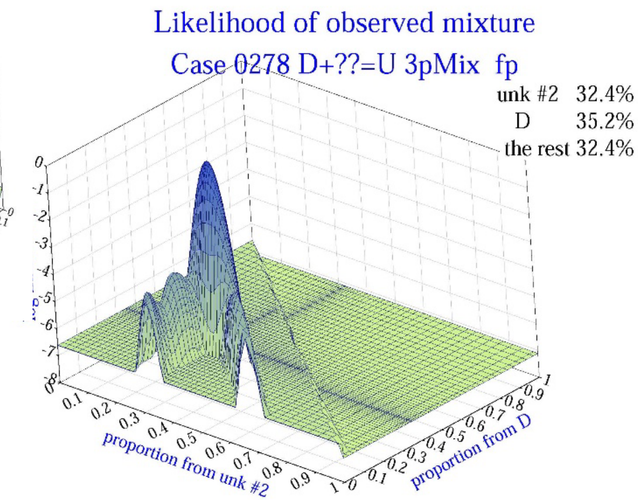
Dyalog 2022 Oct 10

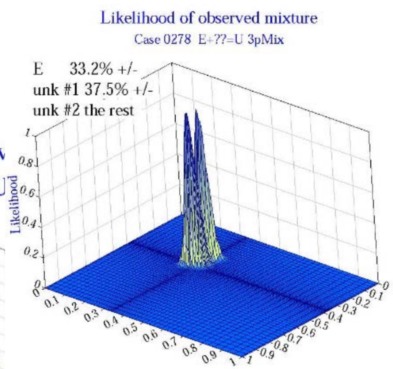
Example functions to integrate

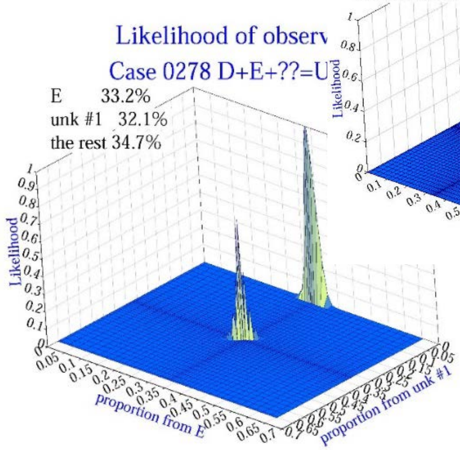












- Typical properties
  - Up to 4 dimensional domain; $\bar{x} = (x_1, x_2, x_3, x_4)$
  - Calculation of $C(\bar{x})$ is expensive
  - $\int_D C(\bar{x})$ is concentrated in a small part of D.

A little context about the DNA evidence application

Touch DNA evidence from a gun

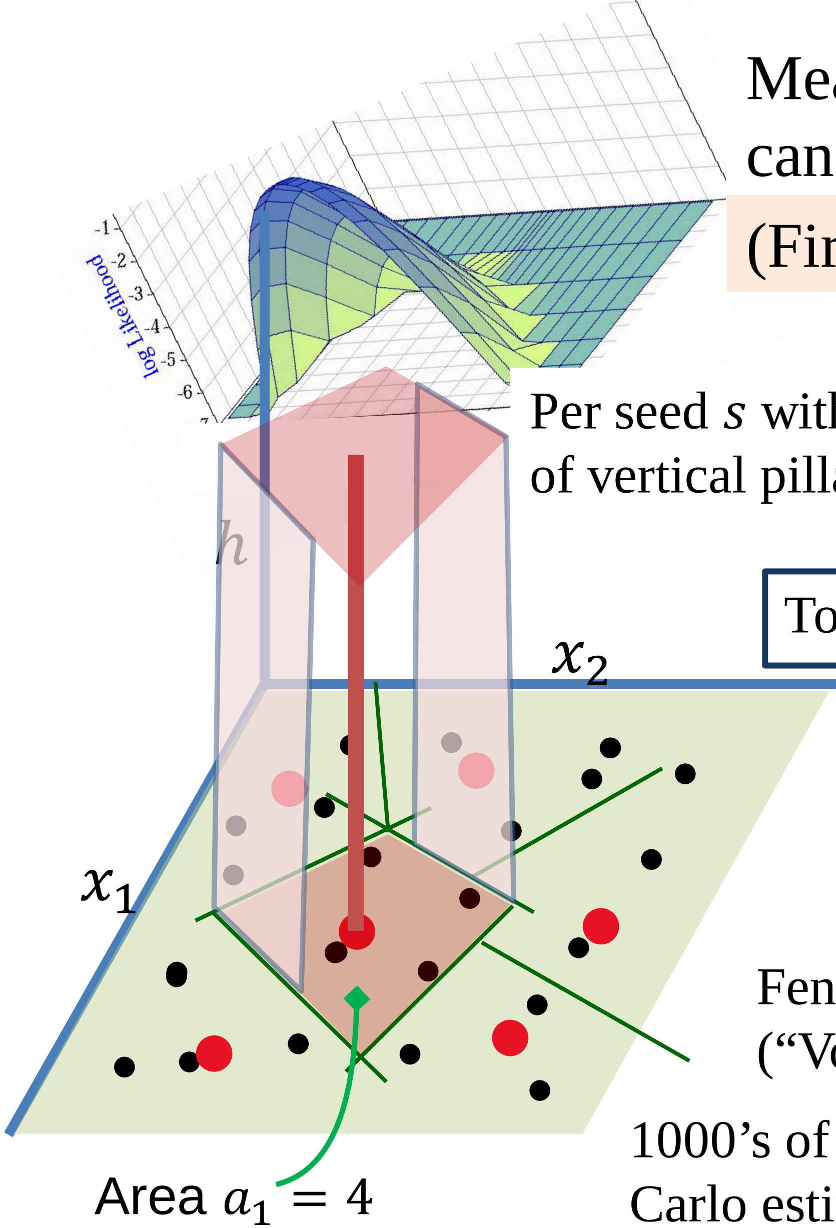
x axis: DNA location or size in genome
y axis: quantity (after lab processing)

DNA evidence overlaid with
an example partial explanation

Bar height = assumed contribution proportions of 2 color-coded people's DNA types.

Measuring volume under an irregular canopy $C(x_1, x_2)$

(First idea. Quick and dirty)



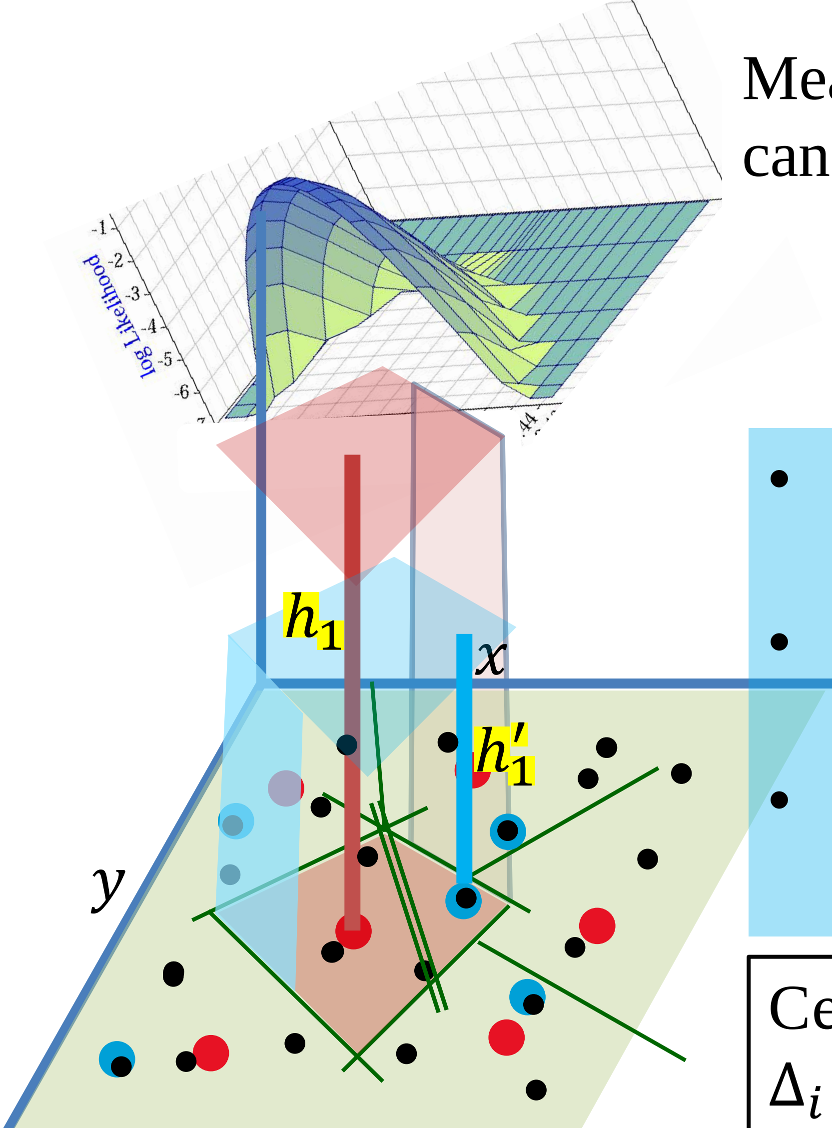
Per seed s with area a_s, compute height $h_s = C$ of vertical pillar/prism. Volume $v_s = h_s \times a_s$.

Total volume (Riemann sum) $\int C \approx \Sigma v_s$.

Choose an initial handful of seeds (big red dots) at which to compute (time consuming!) heights $h_s = C(x_{1,s}, x_{2,s})$.

Fences around each seed define its area. (“Voronoi cell”)

1000’s of random black dots give a Monte Carlo estimate of cell areas a_s.



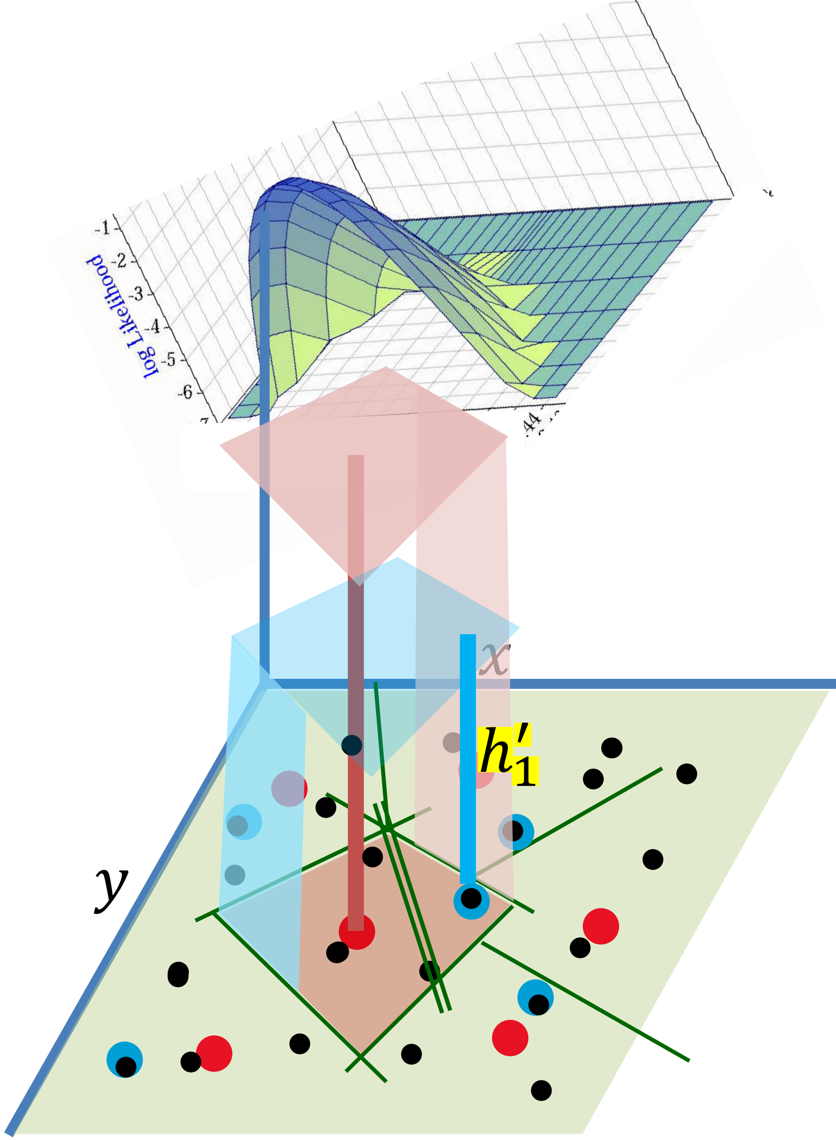
Measuring volume under an irregular canopy $C(x_1, x_2, \ldots))$

Adaptive step: Choose a pillar to split in two.

- For each cell s I'll estimate volume a 2nd time, using new points $\left(x'_{1,s}, x'_{2,s}, \ldots\right)$.
- I pick existing black (area measuring) points for the purpose.*
- Alternative heights $h'_s = C\left(x'_{1,s}, x'_{2,s}, \ldots\right)$ alternative volumes $v'_s = h'_s \times a_s$.

Cell with larger volume difference $\Delta_i = |v_s - v'_s|$ is better candidate for splitting into two cells. So split it.

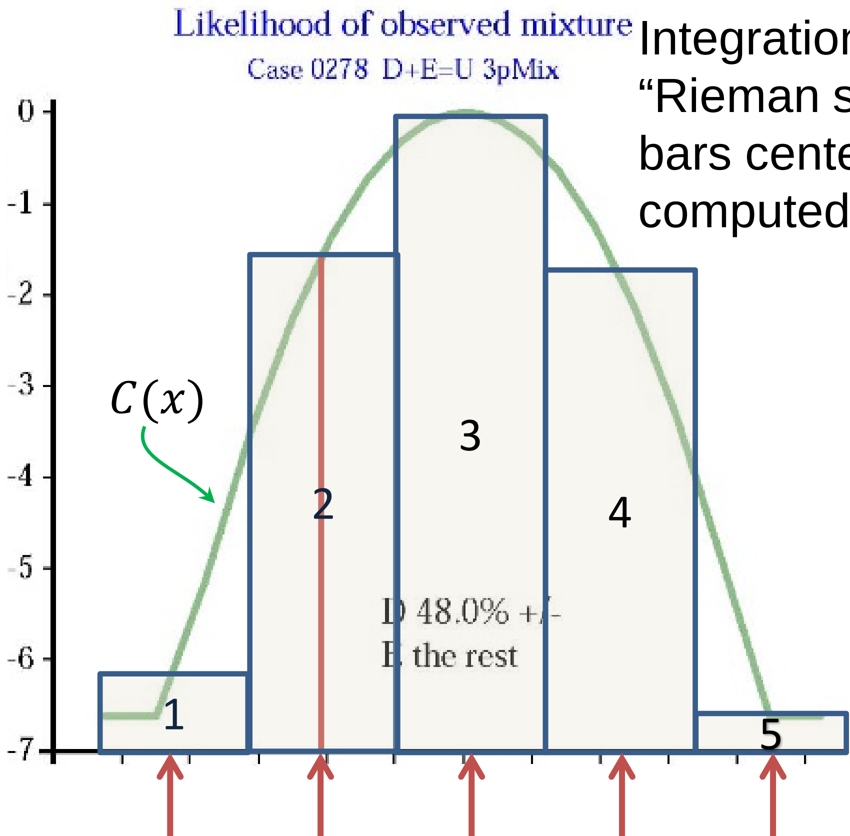
Pitfall with cell splitting:
"I pick existing black (area measuring) dots for the purpose."*



- Eventually splitting dead ends when some small cells run out of black markers to split with.
- Adding a new black dot set costs much compute time to allocate to nearest Voronoi seeds.

- But there is no simple alternative.
- Voronoi boundaries (or areas) are difficult to compute.
- Visit expert in Switzerland?

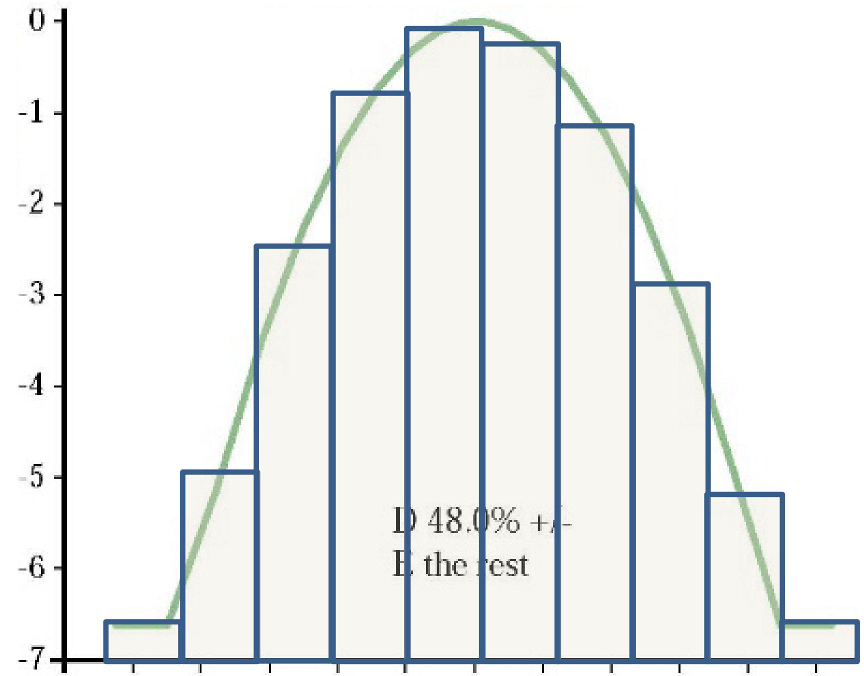
Integration by "Rieman sums" (with bars centered on computed hei



Integral = area under green curve

$\int C(x) \approx \Sigma_{\mathrm{i}}\, w \times h_i$

Refine by splitting bars

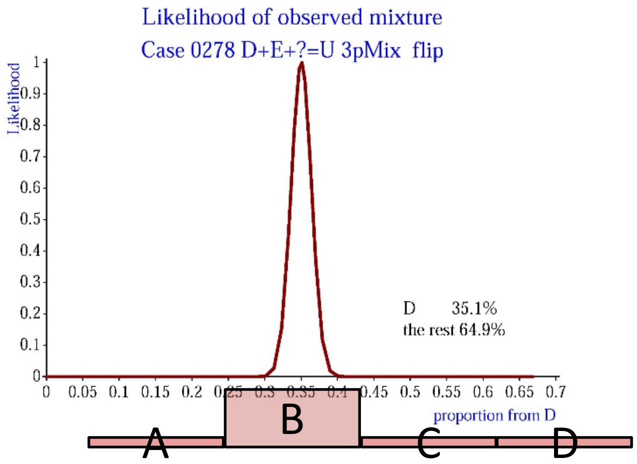


Maybe fine.

Bar splitting requires evaluating $C(x)$, maybe expensive.

Sometimes it's important to economize on splitting.

Numeric integration – area (or volume or hypervolume …) under a curve (canopy …)

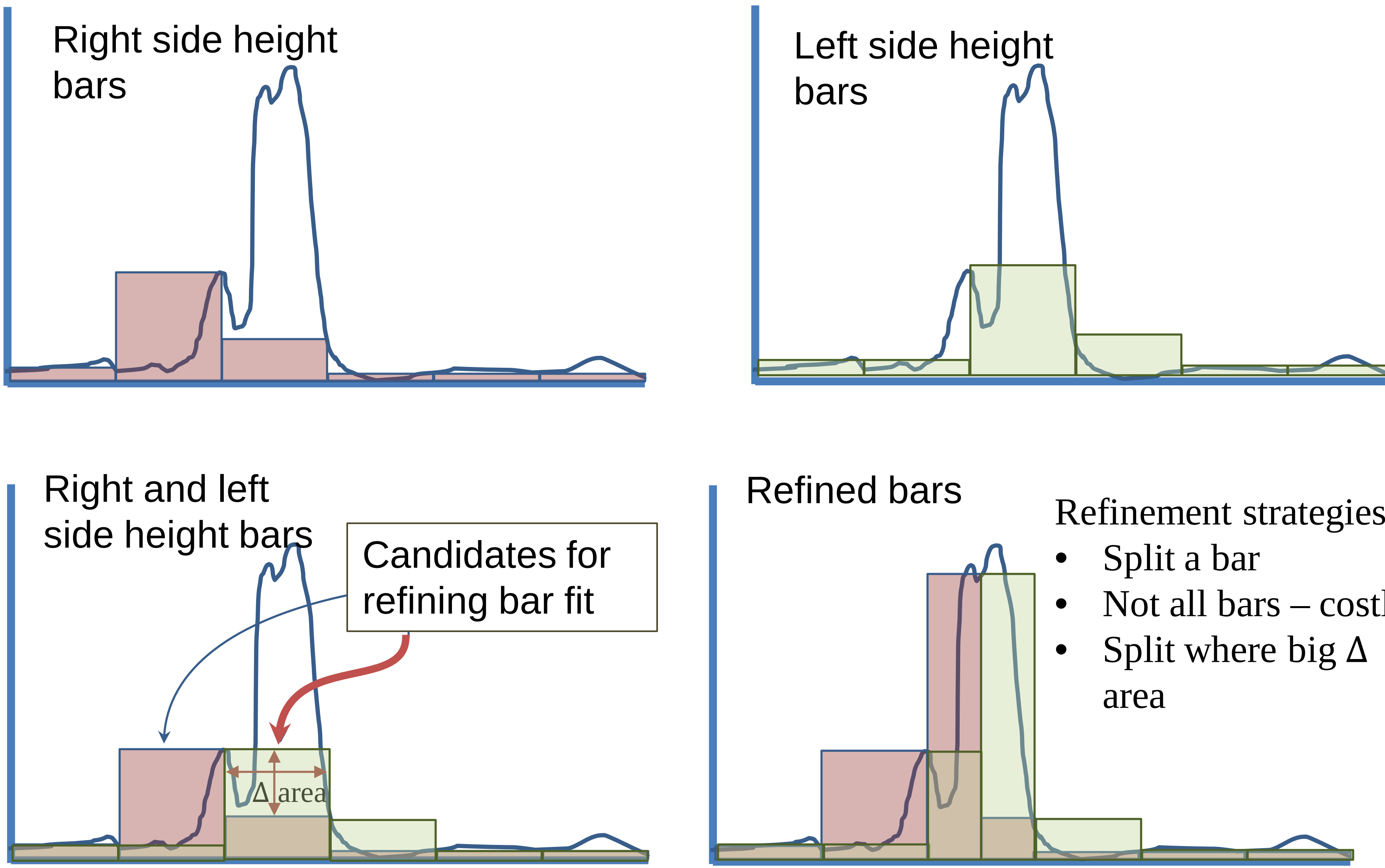


A common situation – a small fraction of the domain accounts for most of the integral.

1D domain: 10% of x-axis is 1/10 of domain
2D domain: 10% of x_1&x_2 axes is 1/100 of domain.
3D domain: 10% of each domain axis is 1/1000 of domain.

Related: Volume of hypersphere inscribed in a unit hypercube goes rapidly to 0.

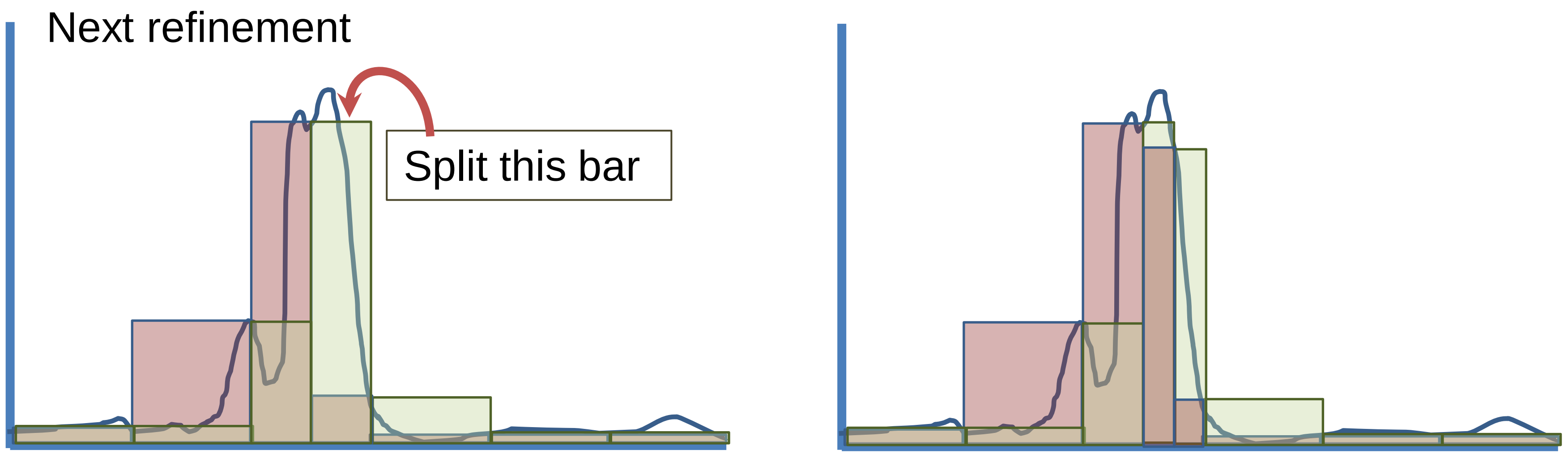
Adaptive integration



Refinement strategies:

- Split a bar
- Not all bars – costly!
- Split where big Δ area

Adaptive integration – 2^{nd} adaption



Seems like a workable method in 2 dimensions
(i.e. 1 domain dimension).

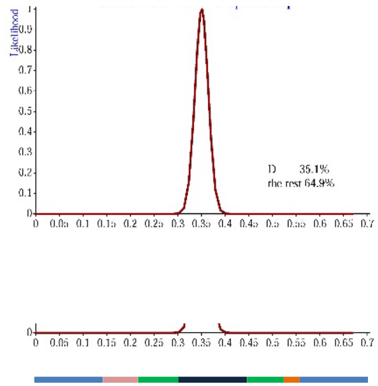
How to translate it to multiple dimension domains?

Adaptive integration summay

- (Write $\bar{x}$ for the point $(x_1, x_2, \ldots x_n)$ in an *n*-dimensional domain.)
- In each cell, compute
  - $h_i \leftarrow C(\overline{x_i})$ at at least 2 values of $\bar{x}$;
  - (hyper-)volumes $v_i \leftarrow h_i \times a$;
  - estimate of volume variation $\Delta v \leftarrow -/(\lceil/, \lfloor/)v_i$.
- Split a cell with large (largest?) Δv.

Generalize to more dimensions

Ordinate $h = C(\bar{x})$

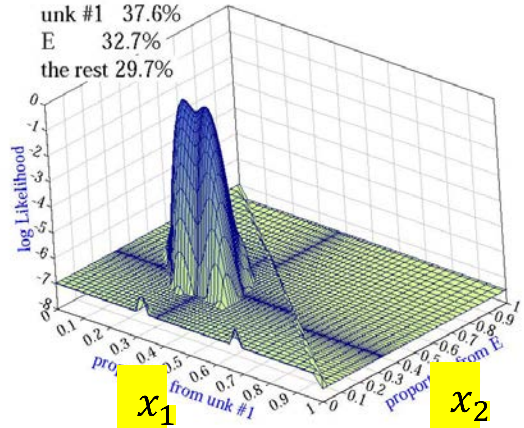


Tiled with 1 dimension line segments

Domain $\bar{x}$ of 1 dimension:

$\bar{x} = (x)$

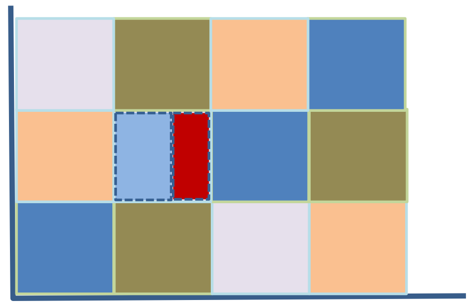
Ordinate $h = C(\bar{x})$



Domain $\bar{x}$ of 2 dimensions:

$\bar{x} = (x_1, x_2)$

x_2



x_1

Tile with squares?

Rectangles?

R-cubature

Problems: angle bias, housekeeping

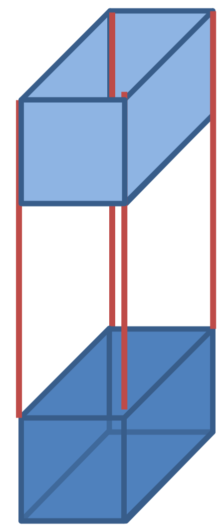
Four dimensions --

Domain $\bar{x}$ of 3 dimensions:

$\bar{x} = (x_1, x_2, x_3)$.

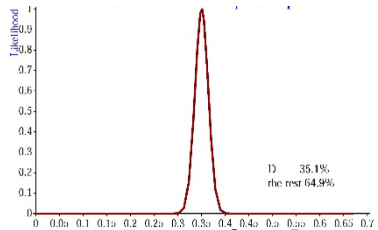
Tile with cubes/rectangular parallelepipeds?

Ordinate $h = C(\bar{x})$

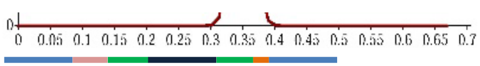


Alternative generalization – triangles etc.

Ordinate $h = C(\bar{x})$

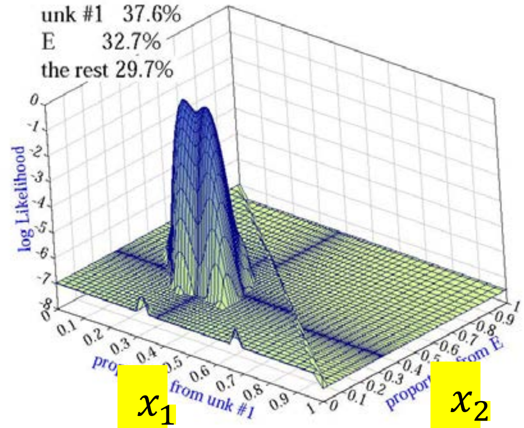


Domain $\bar{x}$ of 1 dimension:
$\bar{x} = (x)$

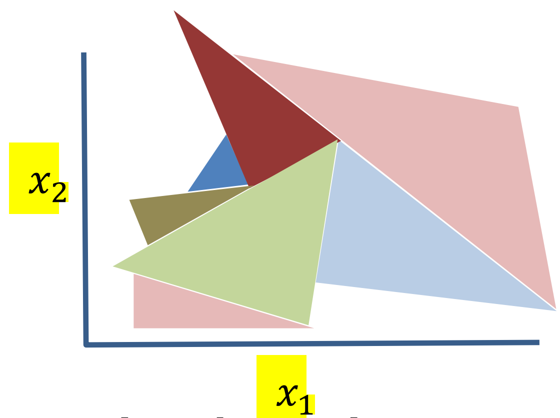


Tiled with 1 dimension line segments

Ordinate $h = C(\bar{x})$

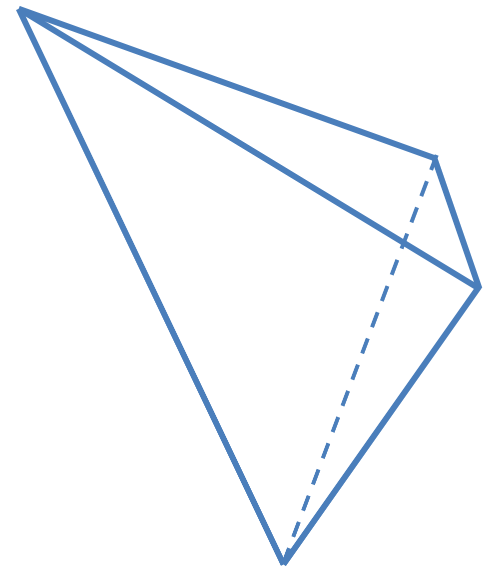


Domain $\bar{x}$ of 2 dimensions: $\bar{x} = (x_1, x_2)$

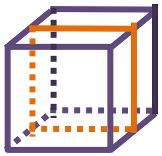


Tile with triangles

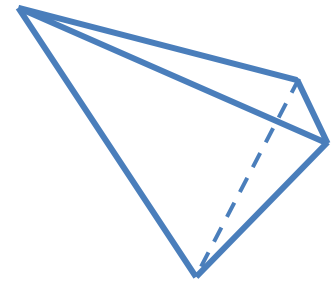
Four dimensions – (domain $\bar{x}$ of 3 dimensions):
$\bar{x} = (x_1, x_2, x_3)$.
Tile with **simplexes**



3D simplex



Pros and Comments



Hypercube cells

+ I know volume computation: $V \leftarrow \times / \bar{x}$
+ Obvious how to split
- \# of cells = # of vertices
    - Compute one $C(\bar{x})$ per new cell
+ Published papers
- Directional bias
- Keeping track of split points

Simplex cells

+ Aha! Just linear algebra: $V \leftarrow \mathrm{Det}(\bar{x})$ (Dfn by R Hui)
+ (see next slide)
+ *Huge* computing leverage, e.g. 11+ cells per vertex
▪ Simplices to *maximize* is published. But *integrating* via simplices may be new.
+ No directional bias
+ Housekeeping splits is simple

Splitting a simplex

- ***Simplex*** definition:
  - – A simplex in *n* dimensions
    - $n + 1$ points connected by
    - $2!\,n$ straight lines

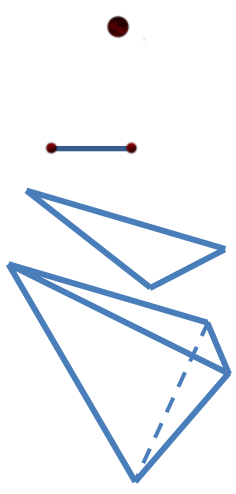
0-simplex

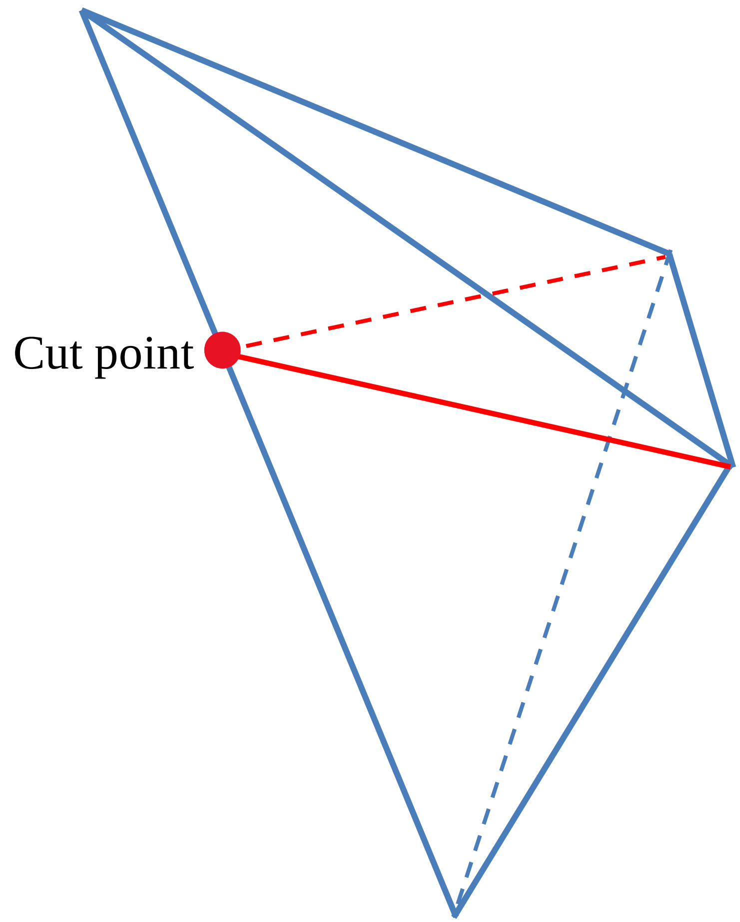
1-simplex

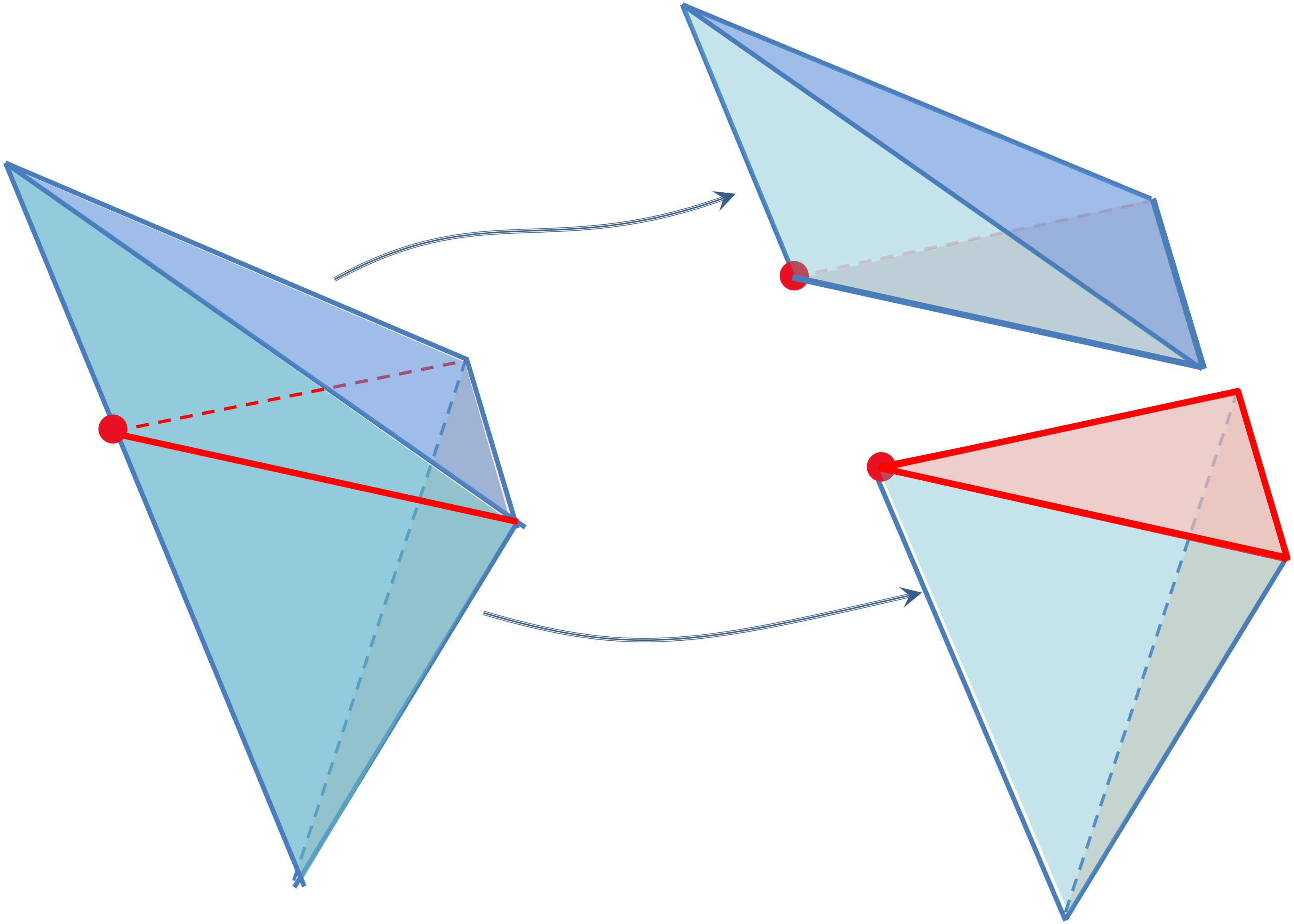
2-simplex

3-simplex

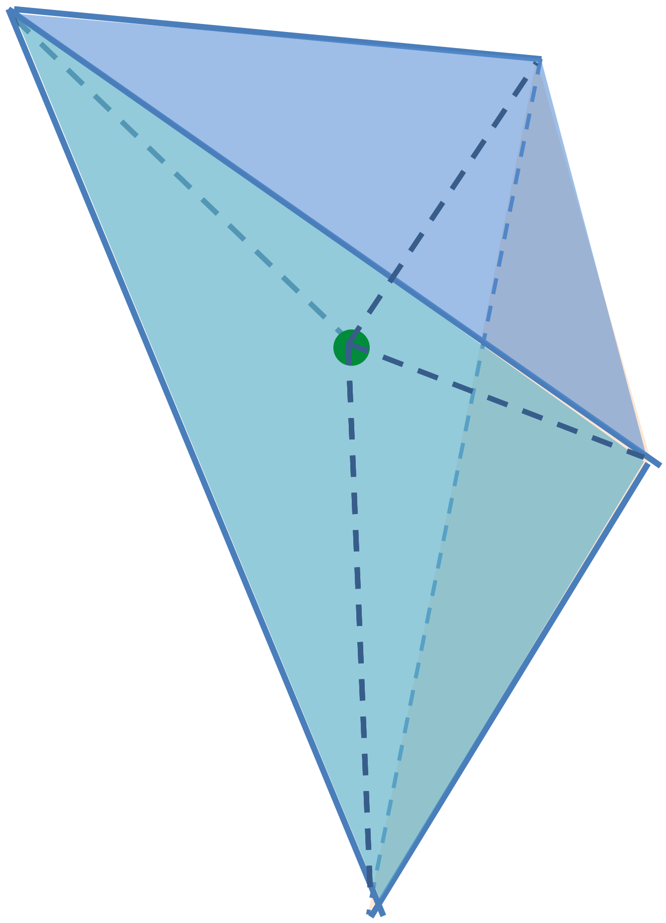
… *n*-simplex







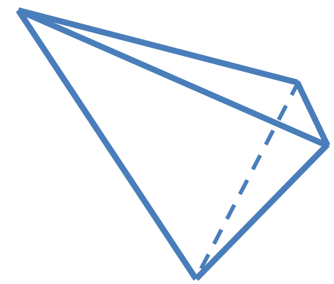
Wrong way to split a cell



- Choose an interior point.
- Connect it to all 4 vertices.
- Cell is cut into 4 cells

  with a common (new) vertex
- !!? Original edges are *never* shortened!

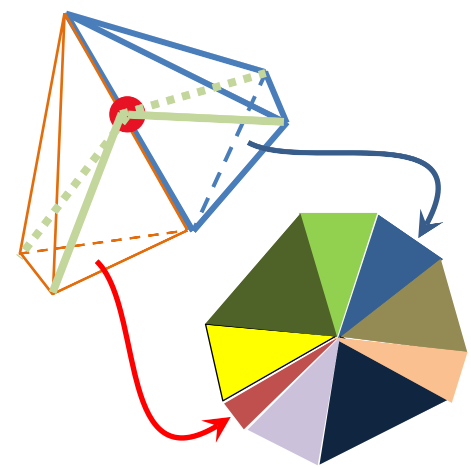
Pros and Comments



Splitting simplex cells

+ *Huge* computing leverage, e.g. 11 cells per new vertex (4D)
+ Extra cell splits are virtually free

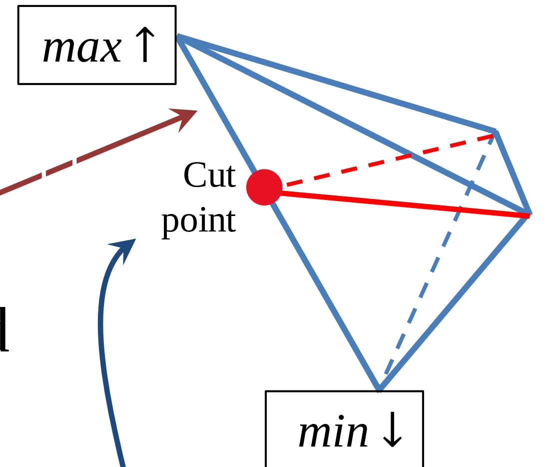
- What would the 4D geometry look like?
    - 3D already permits unlimited # of cells to share an edge



Bottom view

Integration algorithm in brief

- Initialize
  - Tesselate domain with a handful of simplex cells.
  - For each cell s and vertex $\bar{x}_{s,i}$, calculate and save all the pillar volumes $v_{s,i} = a_s \times C(\bar{x}_{s,i})$.
- Iterate
  - For each cell s, find its *extreme edge* E_s – the edge that connects the vertices $v_{s,max}$ and $v_{s,min}$ of largest & smallest of the (already-computed) pillar volumes of s.
  - Find the overall most extreme edge E among all cells.
  - Cut within E at a cleverly chosen (how?) point.
  - Split *all* cells that include edge E.

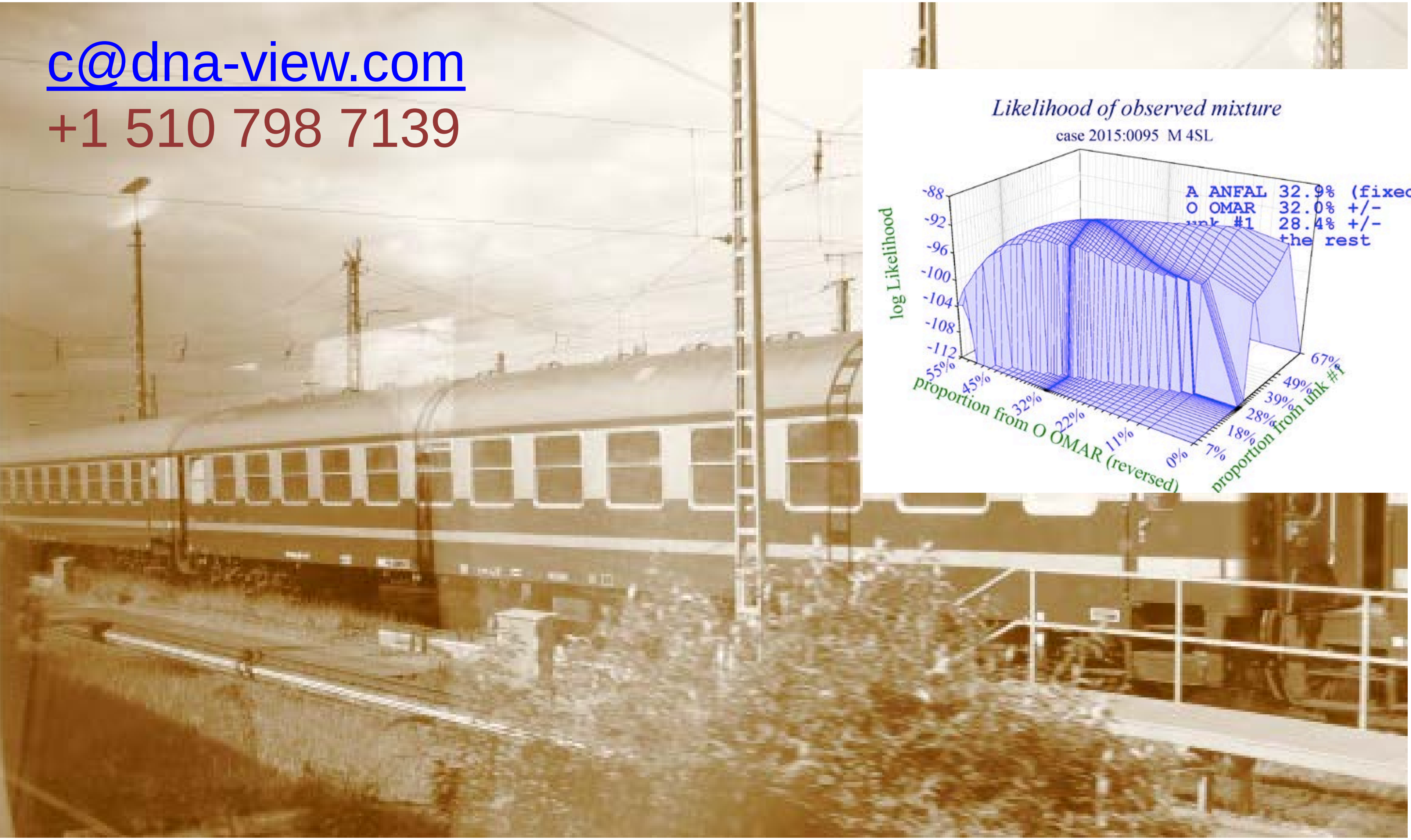


Summary & Epilogue

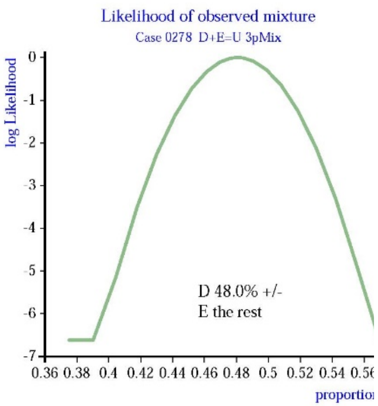
- Simplex tessellation requires 11×, 4×, or 2× **fewer** $C(\bar{x})$ calculations per cell than does cubature for 4D, 3D, or 2-dimension domain.
- Mathematically satisfying stopping rule availed by computing every vertex, comparing high-side *vs* low-side integral estimation:
  - $tolerance > \Sigma_s(v_{s,max} - v_{s,min})$
- Having decided on which edge to cut, cut where? Midway? (No!)
  - Presently: Calculate the height at some arbitrary intermediate point, then predict by quadratic interpolation with the 3 height including those of the edge ends.
  - Better idea brewing that needs a bit of housekeeping.

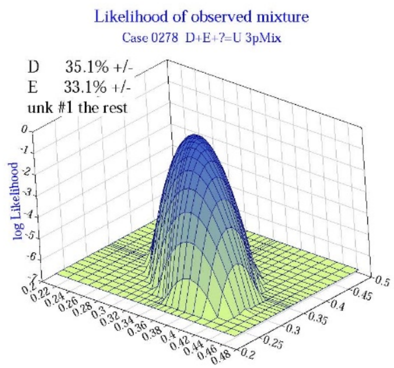
c@dna-view.com

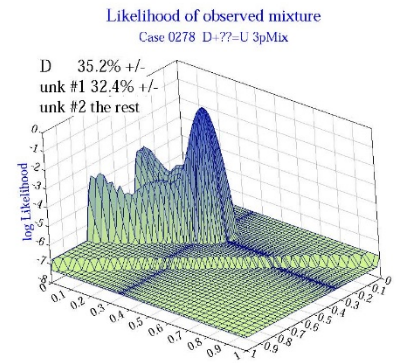
+1 510 798 7139

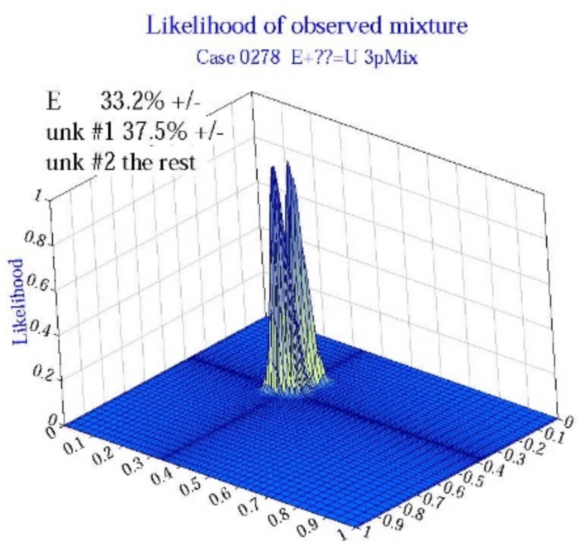


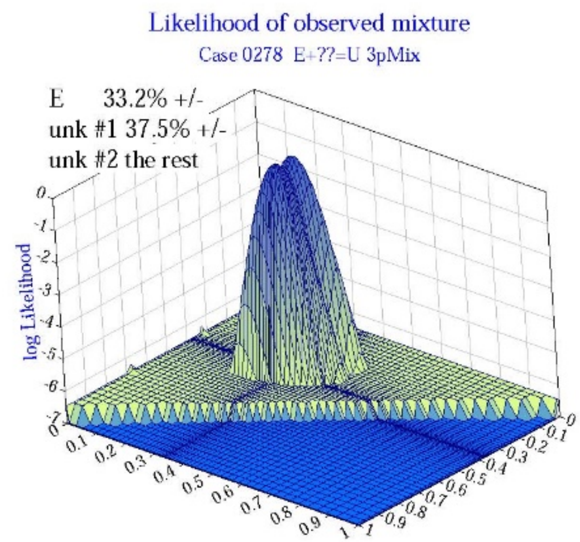
Example functions to integrate

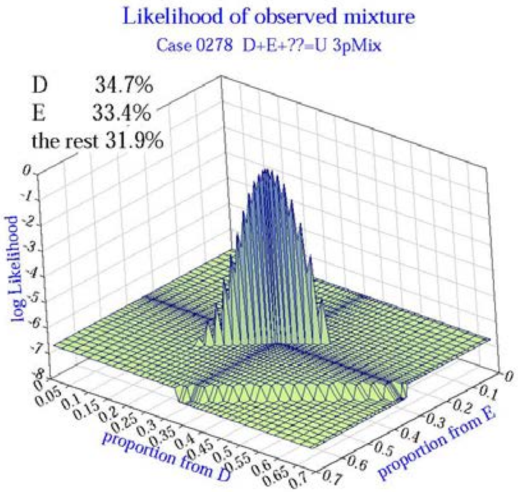


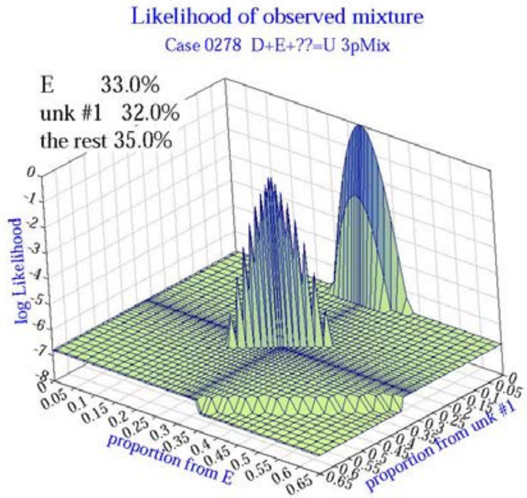


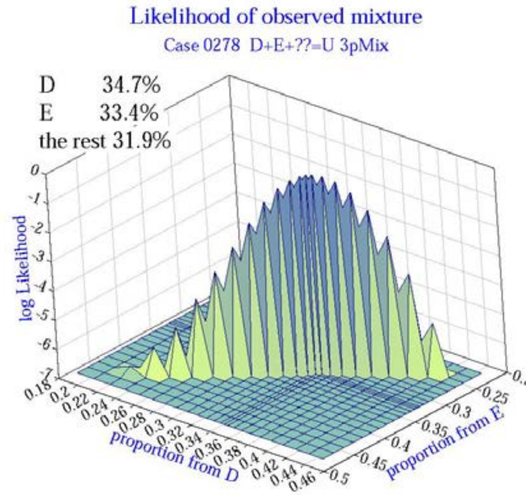


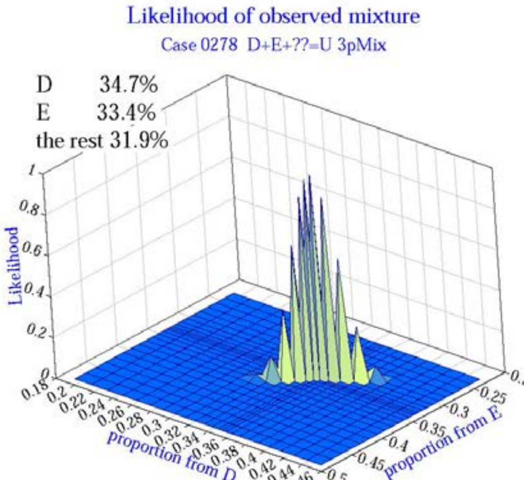


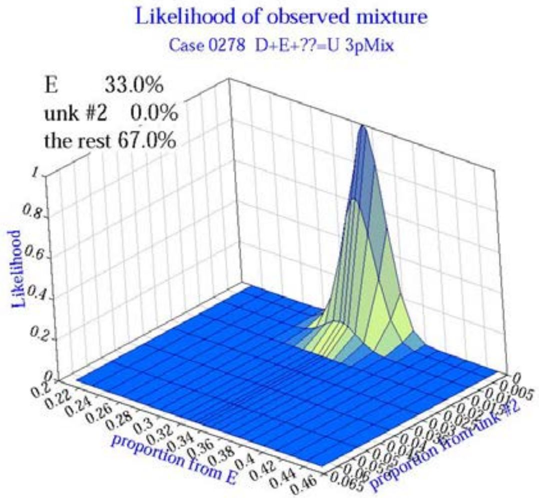


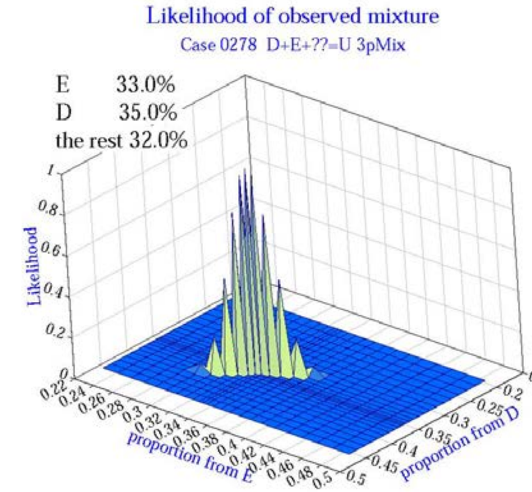

Likelihood of observed mixture
Case 0278 D+E+??=U 3pMix
D 34.7%
E 33.4%
the rest 31.9%
log Likelihood
proportion from D
proportion from E

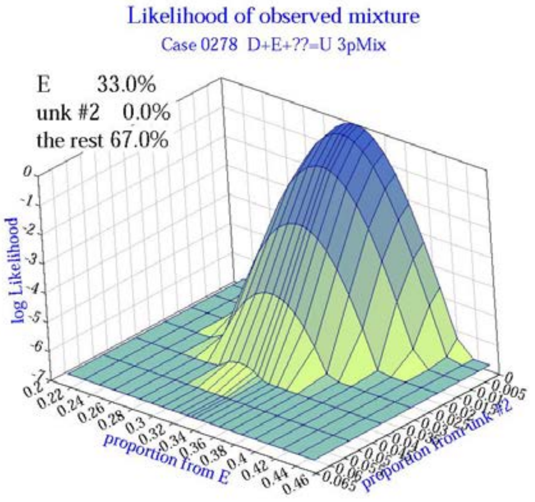

Likelihood of observed mixture
Case 0278 D+E+??=U 3pMix
E 33.0%
unk #1 32.0%
the rest 35.0%
log Likelihood
proportion from E
proportion from unk #1

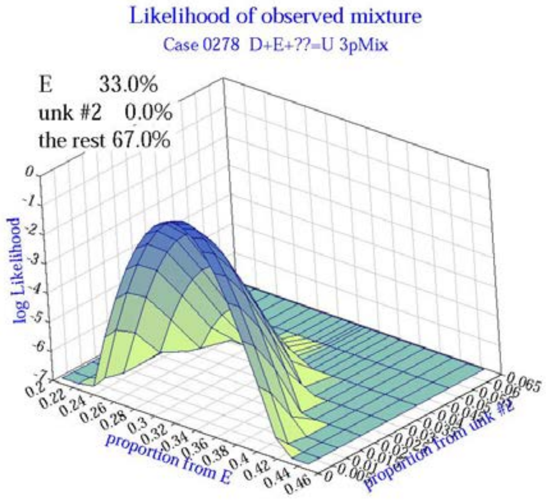

Likelihood of observed mixture
Case 0278 D+E+??=U 3pMix
D 34.7%
E 33.4%
the rest 31.9%
log Likelihood
proportion from D
proportion from E

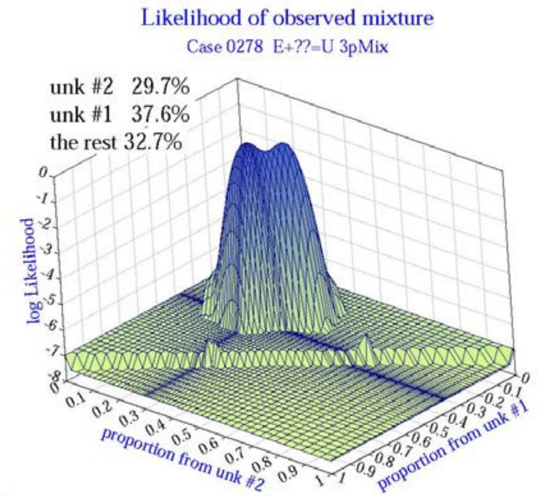

Likelihood of observed mixture
Case 0278 D+E+??=U 3pMix
D 34.7%
E 33.4%
the rest 31.9%
Likelihood
proportion from D
proportion from E

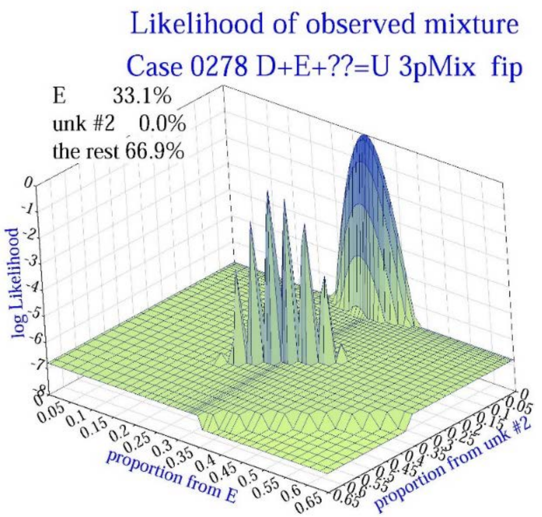

Likelihood of observed mixture
Case 0278 D+E+??=U 3pMix
E 33.0%
unk #2 0.0%
the rest 67.0%
Likelihood
proportion from E
proportion from unk #2

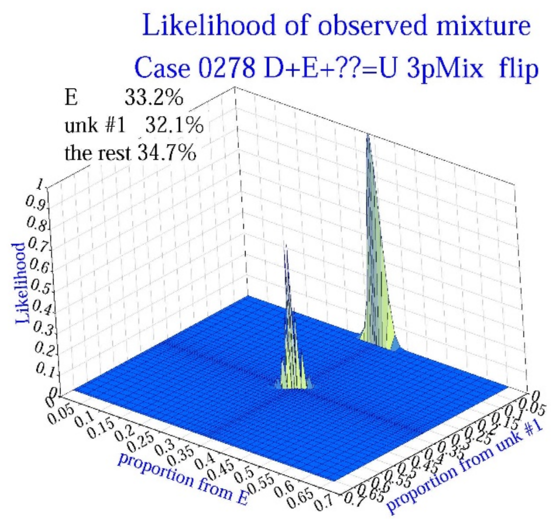

Likelihood of observed mixture
Case 0278 D+E+??=U 3pMix
E 33.0%
D 35.0%
the rest 32.0%
Likelihood
proportion from E
proportion from D

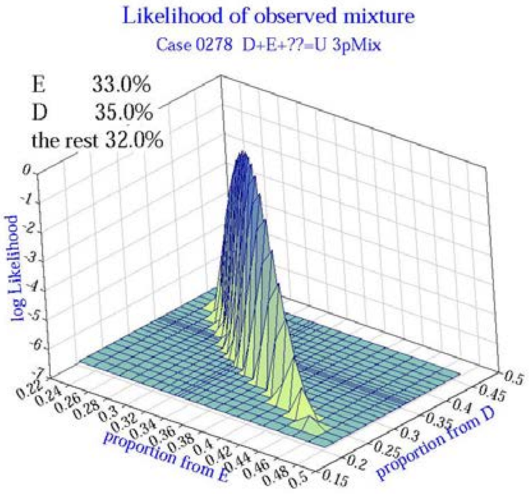

Likelihood of observed mixture
Case 0278 D+E+??=U 3pMix
E 33.0%
unk #2 0.0%
the rest 67.0%
log Likelihood
proportion from E
proportion from unk #2

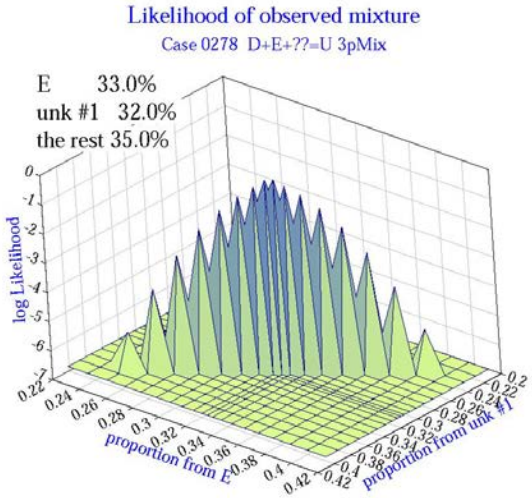

Likelihood of observed mixture
Case 0278 D+E+??=U 3pMix
E 33.0%
unk #2 0.0%
the rest 67.0%
log Likelihood
proportion from E
proportion from unk #2

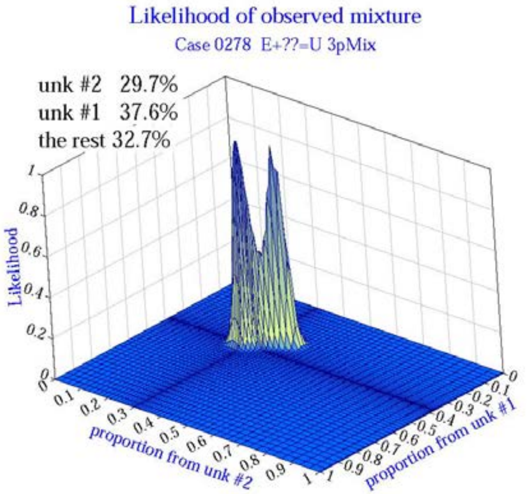

Likelihood of observed mixture
Case 0278 E+??=U 3pMix
unk #2 29.7%
unk #1 37.6%
the rest 32.7%
log Likelihood
proportion from unk #2
proportion from unk #1

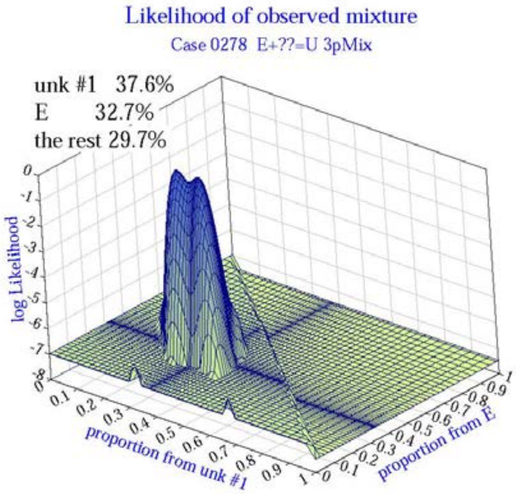

Likelihood of observed mixture
Case 0278 D+E+??=U 3pMix fip
E 33.1%
unk #2 0.0%
the rest 66.9%
log Likelihood
proportion from E
proportion from unk #2

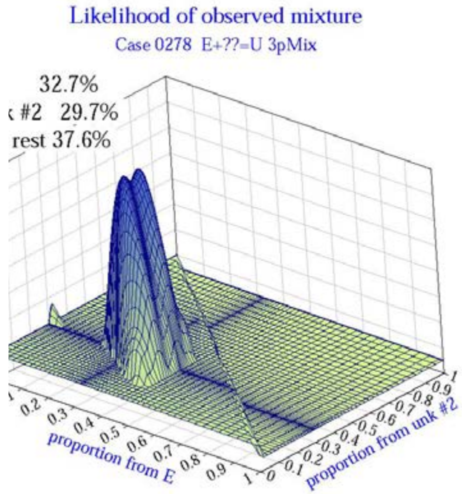

Likelihood of observed mixture
Case 0278 D+E+??=U 3pMix flip
E 33.2%
unk #1 32.1%
the rest 34.7%
Likelihood
proportion from E
proportion from unk #1

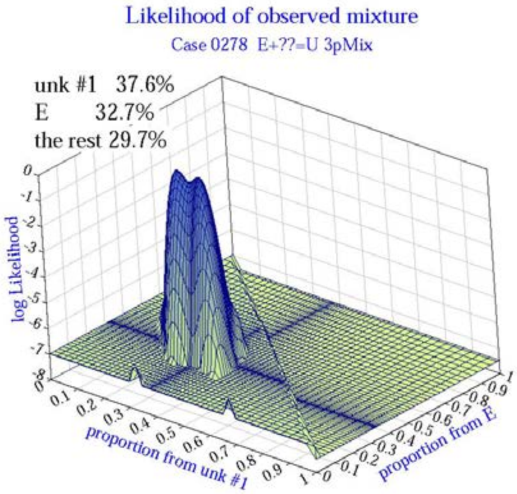

Likelihood of observed mixture
Case 0278 D+E+??=U 3pMix
E 33.0%
D 35.0%
the rest 32.0%
log Likelihood
proportion from E
proportion from D

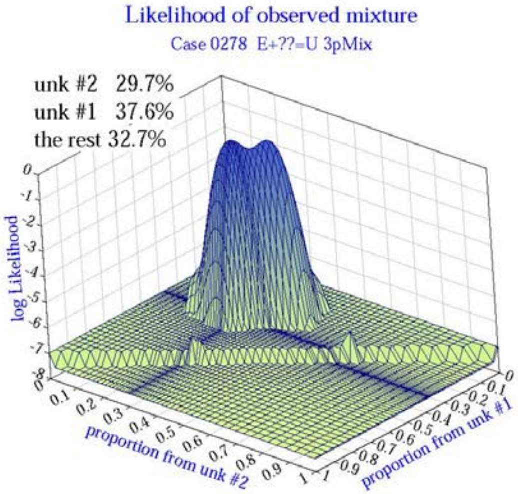

Likelihood of observed mixture
Case 0278 D+E+??=U 3pMix
E 33.0%
unk #1 32.0%
the rest 35.0%
log Likelihood
proportion from E
proportion from unk #1

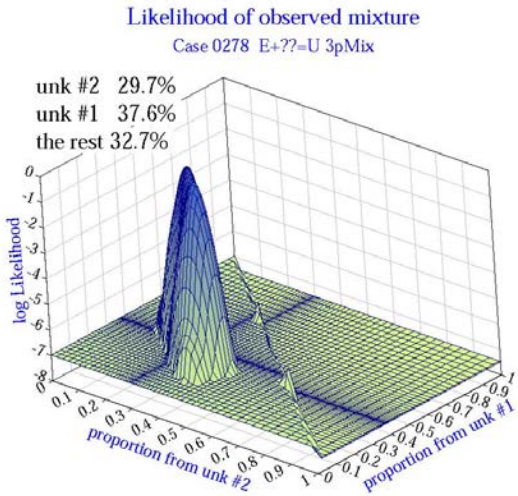

Likelihood of observed mixture
Case 0278 E+??=U 3pMix
unk #2 29.7%
unk #1 37.6%
the rest 32.7%
Likelihood
proportion from unk #2
proportion from unk #1


Likelihood of observed mixture
Case 0278 E+??=U 3pMix
unk #1 37.6%
E 32.7%
the rest 29.7%
log Likelihood
proportion from unk #1
proportion from E


Likelihood of observed mixture
Case 0278 E+??=U 3pMix
32.7%
#2 29.7%
rest 37.6%
proportion from E
proportion from unk #2


Likelihood of observed mixture
Case 0278 E+??=U 3pMix
unk #1 37.6%
E 32.7%
the rest 29.7%
log Likelihood
proportion from unk #1
proportion from E


Likelihood of observed mixture
Case 0278 E+??=U 3pMix
unk #2 29.7%
unk #1 37.6%
the rest 32.7%
log Likelihood
proportion from unk #2
proportion from unk #1


Likelihood of observed mixture
Case 0278 E+??=U 3pMix
unk #2 29.7%
unk #1 37.6%
the rest 32.7%
log Likelihood
proportion from unk #2
proportion from unk #1

How to generalize cell shape with larger # of dimensions

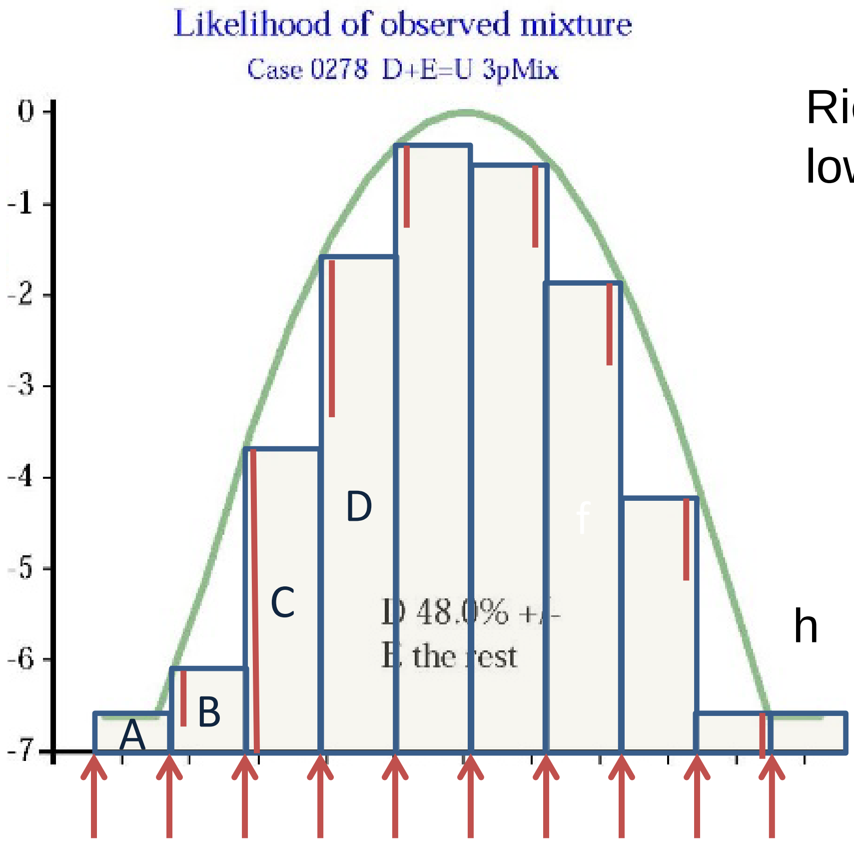
Type size 16

Type size 18

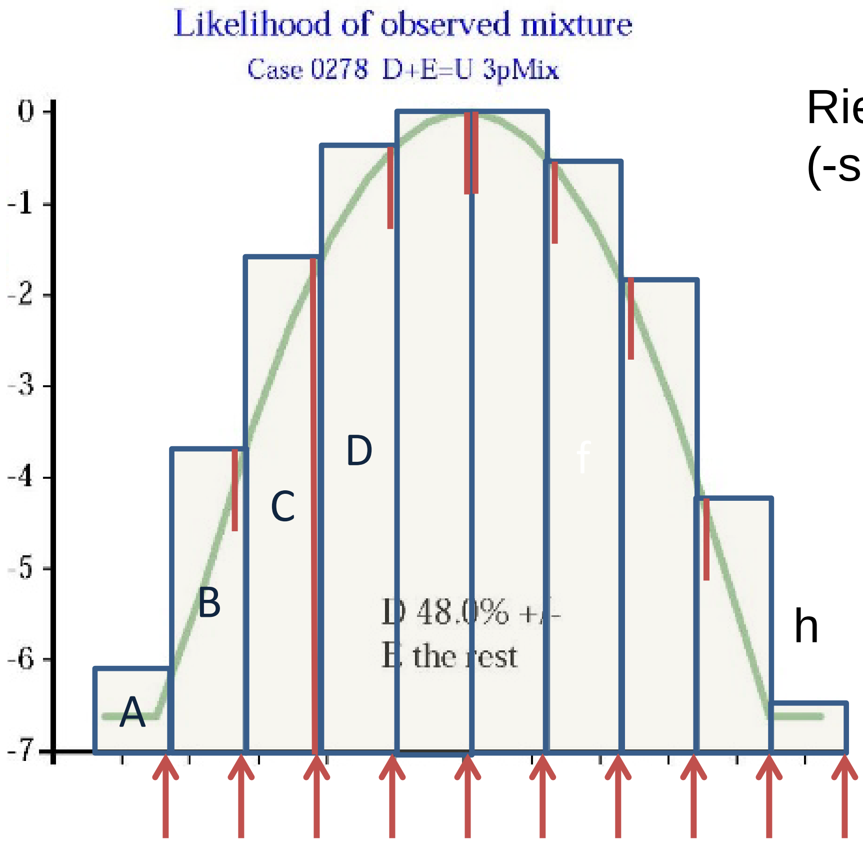
Type size 24

Type size 16

Type size 24



Riemann integration
low-side $\int$ estimate.



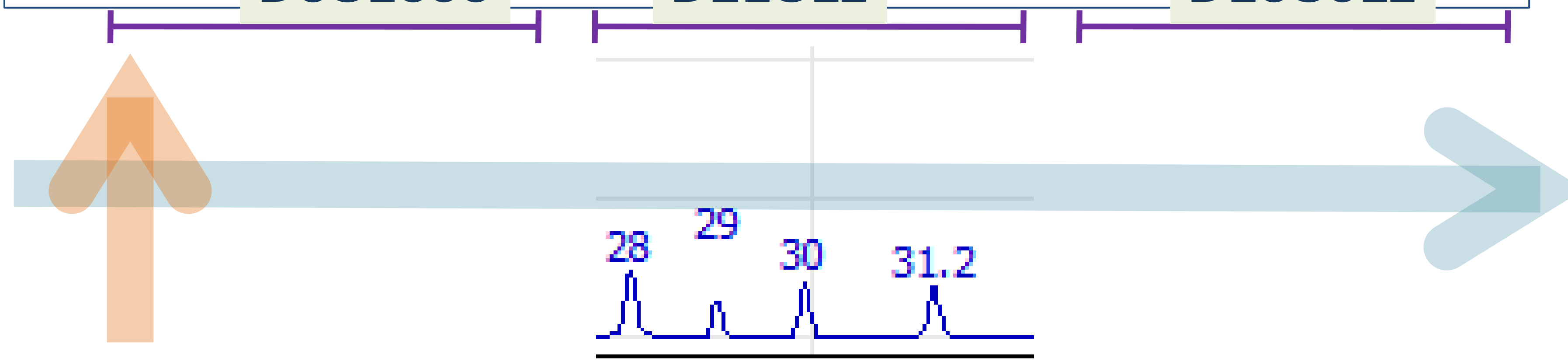
Riemann integration
(-side height)

Old & New mixture models

Then: One binary dimension – Allele size list

15 17 18 28 29 30 31.2 14 15 16 17

D3S1358 D21S11 D18S511

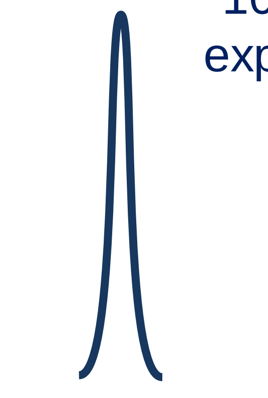


Now: Two dimensions

→ Allele sizes

↑ Peak heights – **continuous**

Stochastic variation model



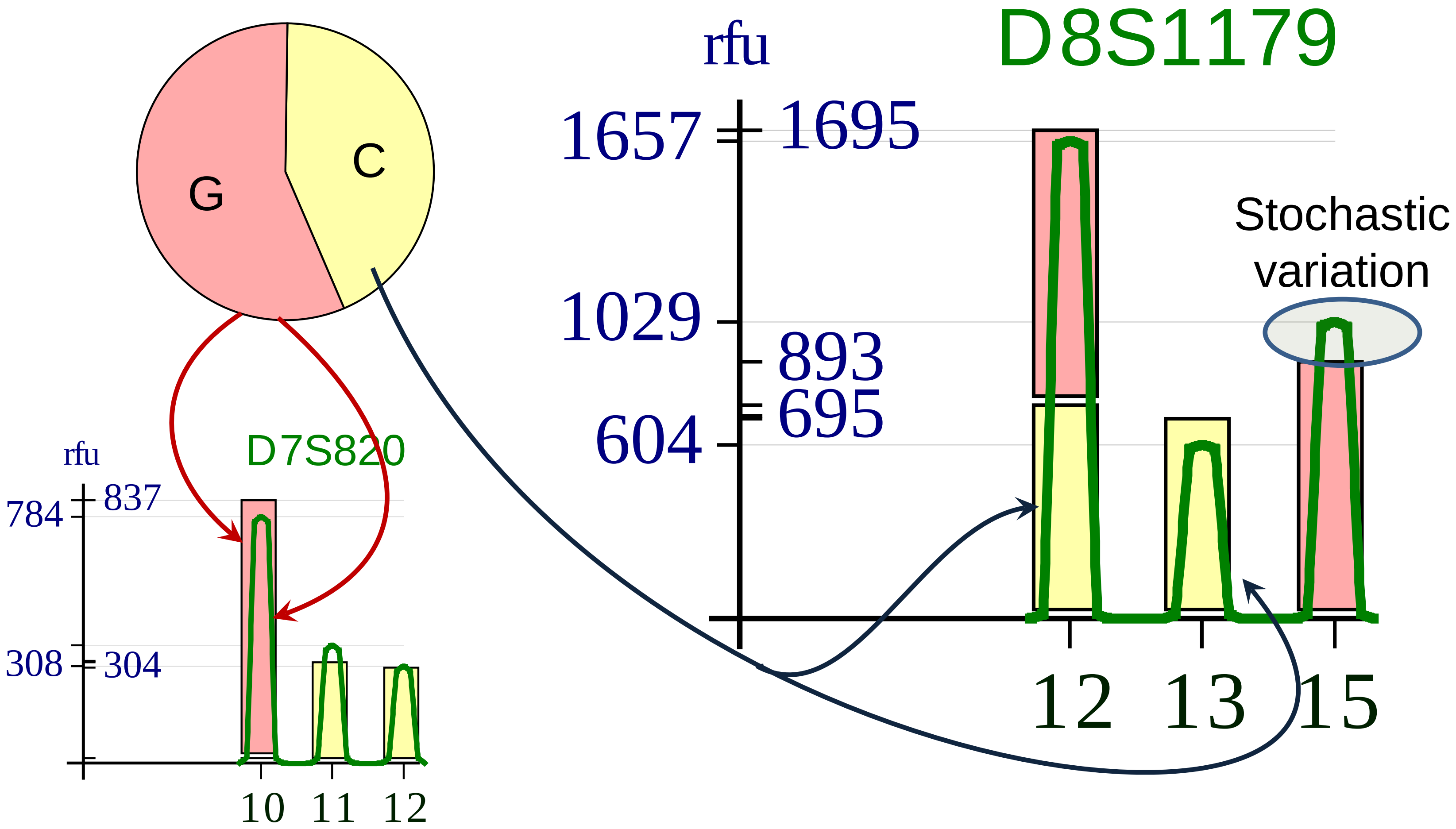
1000 rfu
expected

1122 rfu
observed

Mixture likelihood without unknowns

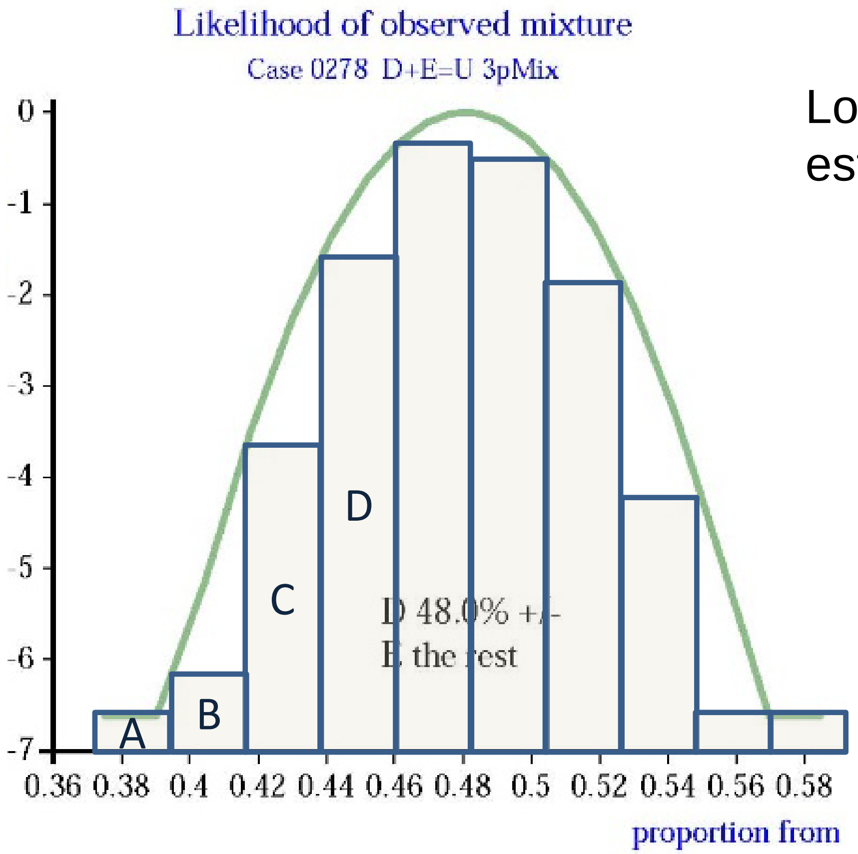
Example hypothesis:

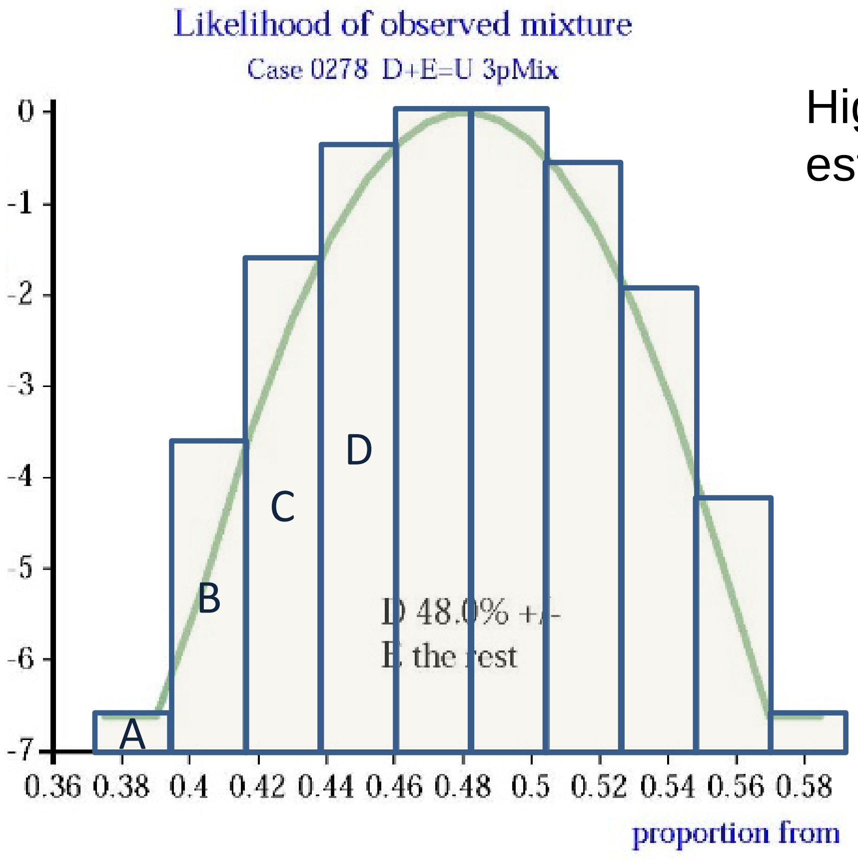
Mixture is G+C, proportion 5:4



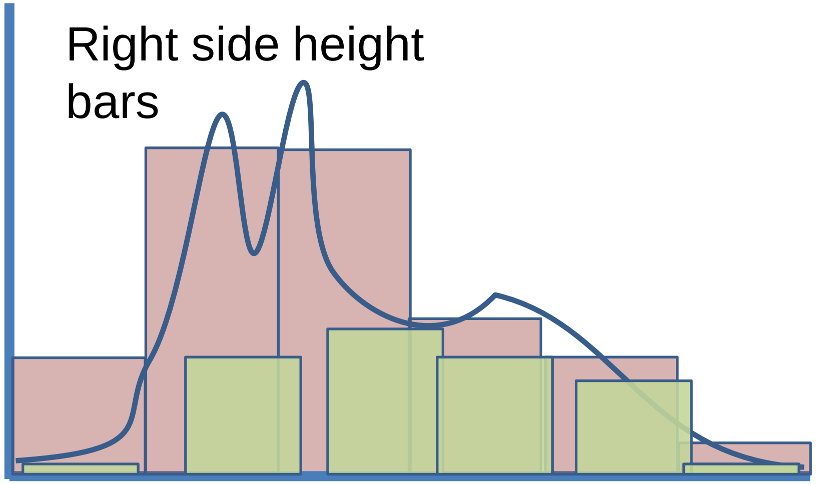
Mixture likelihood *with* unknowns

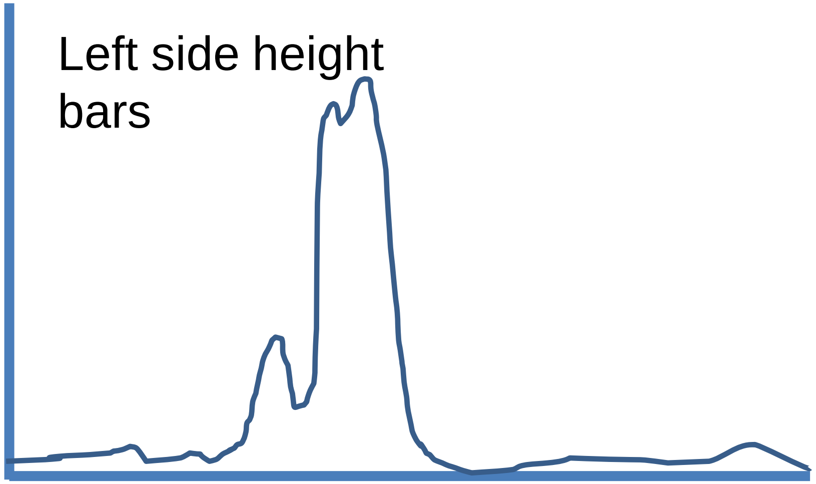
Low-side Riemann estimate





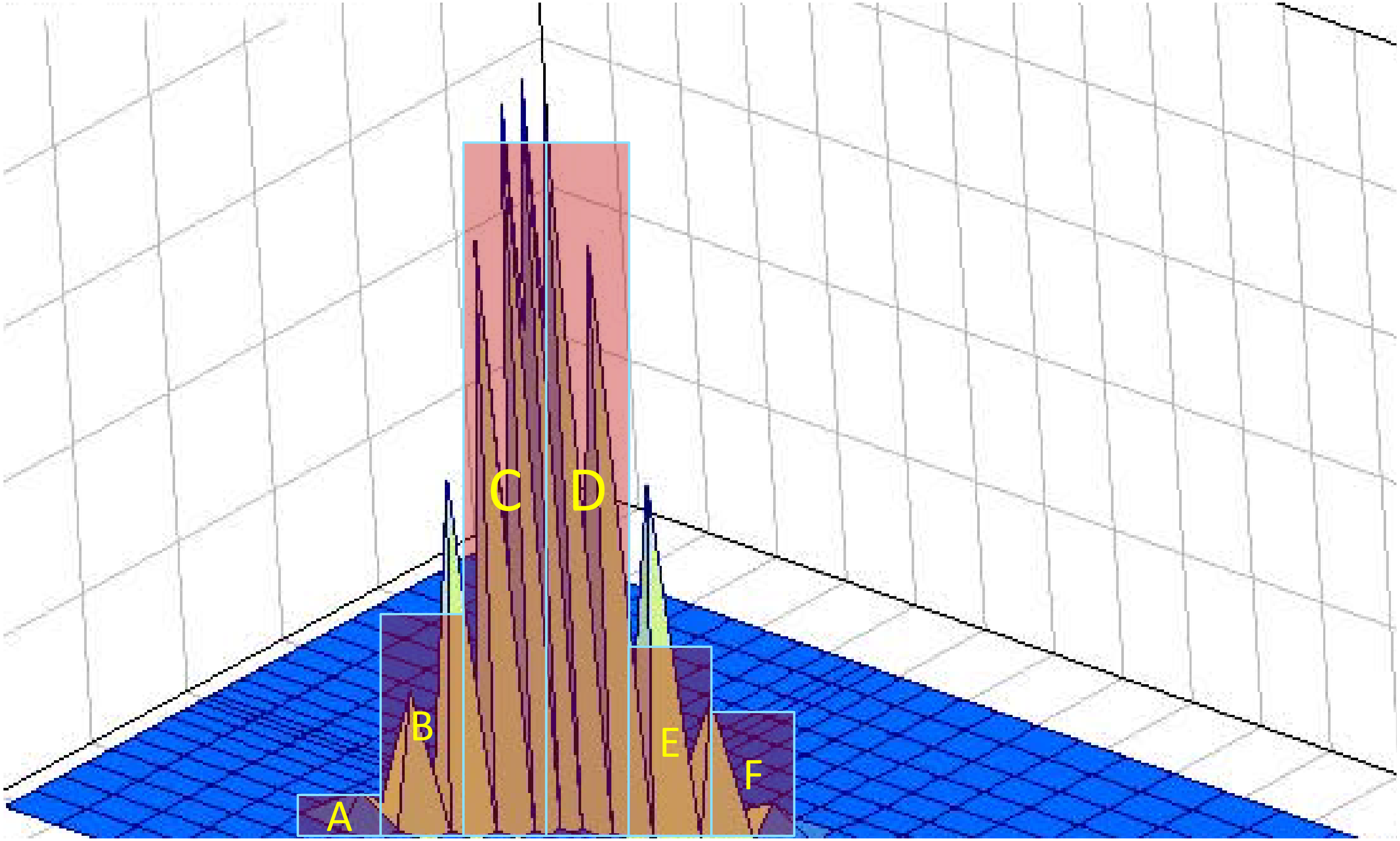
High-side Riemann estimate

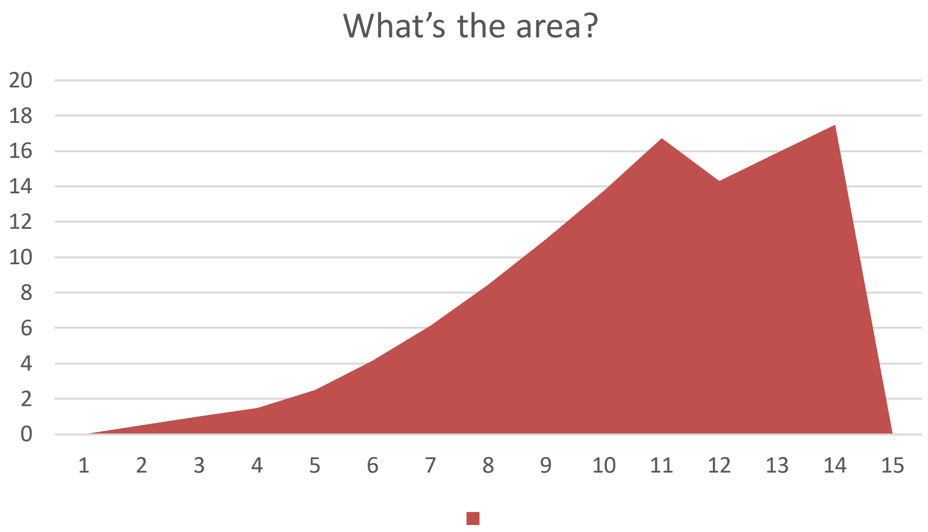


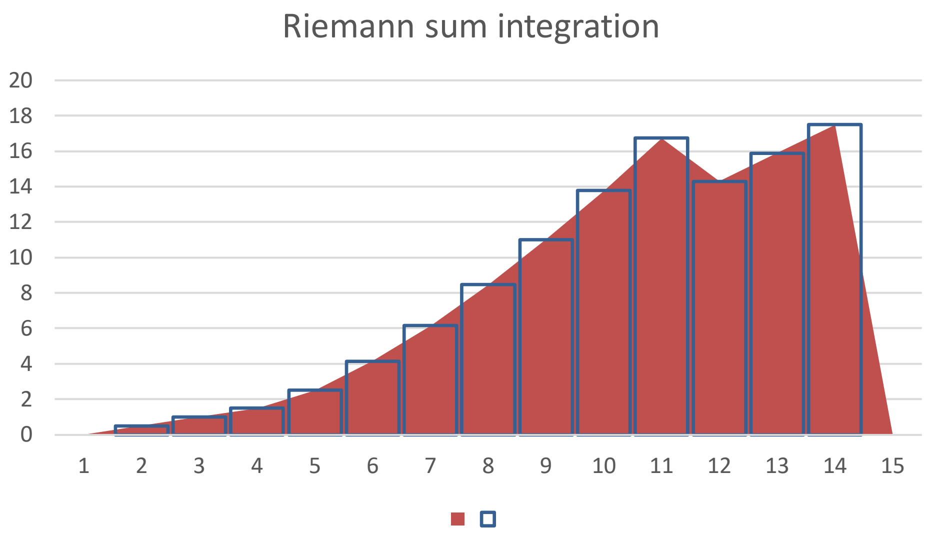


Refinement strategies:

- Split a bar
- Not all bars – costly!
- Split where big Δ area


A
B
C
D
E
F


What’s the area?
20
18
16
14
12
10
8
6
4
2
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15


Riemann sum integration
20
18
16
14
12
10
8
6
4
2
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15