Ditribution based in situ Data Modeling, Analysis,and Visualization（科学可视化）

Part1

经历

- NASA 做流场可视化的盛况
- 高性能计算，想要到美国读书的，读博士的可以考虑一下；在寻找学生；要做科学可视化的学生

在计算科学上的挑战

- very high resolution grid
- ExaFLops supercomputers 超级计算机
  - 电脑越做越大，计算越来越多，就需要把东西存到磁盘里面进去，I/O处理很麻烦，所以现在在找特征
  - 解决方案：找原味数据（很有趣的地方）；不存data，每一次迭代，看看有没有重要的东西，有重要的东西就拿下来
  - 原味可视化，你要看什么？到底哪些资料是值得我们存取的?

> 原味可视化精髓:
> 一边run，一边用
> 要把**data**适当的model起来，还要吧error model做的很好!
> 存**data**的地方，要有feature的地方
>
> 传统计算机、数学、几何有兴趣的话，科学可视化就有发挥的空间
>
> **openGL**
> **space**---->*抓几何（兴趣）*
> 商业的话，并使那么合用

> 很多学生都会平行处理，openmp,GPU一定要

```
你处理过科学可视化，你处理过big data

美国的custom 还是政府，国家实验室
```

objectives and approaches

```
### 如何把一个很大的data转换成一个分布
### 如何做sampling?
### additonal 的信息分析，这块信息有意义，那个区域和哪个区域有关系? <china
vis讲述>
```

Research step

- 1.Data reduction（ML在做）
- 2.vsiualization feature extraction based on distributioins
- 3.information theory (regions of interest)
- 4.不确定性的分析、统计分析

Local Statistical Summarization

1.怎么把数据变成机器分布，把数据做成model，存入数据

```
exaple
一个图片的云的栗子：变成一个几率分布，换成格子，然后做转换

分布表示：
Histograms、KDE、高斯分布、GMM

高斯分布：
如果一个图像里面有3个高斯分布的山峰
每一天的数据进来，我就要update
```

- model：

```
1.Histograms
2.kernel Density Estimates
3.高斯分布
4.高斯混合模型
```

non-parametric distrubutions（没有一个办法去直接 数学model化）

- 1.Histograms（比较离散，用的很多）

```
1.
不需要assume
如何去model?
用 Histograms:（用的最多的）
--
高度：区域内的样本点的个数
几率预测：i/n，曲线不平滑；切分成一千万份，就会趋近于零；很多bin都是空的，切分后的数据比原来的数据还要大
density=
Σ density=1

这个bin如何分析？1024、2048?
空间问题，10个维度，10exp10的空间，不划算---sparse 空间的论文 问题
```

- 2.kernel Density Estimates(KDE)（很重要）

```
- 连续的数据，给出一个smooth的曲线model
- 用这个方法产生出来的照片，比较光滑好看
- 根据有限的data 给training出来，可以用来做预测
- 事后evalue/分析
- 主要目的：我要产生一个soooth的几率分布，可以精确的去预测
- 不好的地方：这个方程是要run times，data都要留下来，不能丢掉。
- 必须要有一个smooth的需求

- example:
# 给几个高斯分布，然后拟合成一个curve
```

parametric distrubutions

- 高斯分布

```
- 哪里都可以查得到的
- 给value-->给几率
- 可以查一下2维/3维（似乎是球心点）的高斯分布
- 1维：u ∂
- 2维：u1 u2 ∂1 ∂2

-我们可以做什么样的分析?
```

- 高斯混合模型(Gaussian Mixture Models)(GMM)

```
- example:  一个图有2个高峰

# weight值如何确定? u如何找?
-

# Expectation Maximization (EM)
- 2个高斯线来解决
- 贝叶斯
- 高斯EM，一般可视化用k-mean的EM
```

Part2

- 如果有兴趣，可以和沈老师联系，美国是真的有这个需要

基于分布的数据分析和可视化

- 比如现在有pdf（高峰）之后，有体绘制什么的，我需要做sample---蒙特卡洛算法

如何做sampling?

- 如果给你一个值，得对应出一个值（传统科学可视化）
- 现在：底层抽掉

sampling

```
- Cumulativie Density Function (CDF)

- Sampling strategy from inverse CDFs

方法:
- inverse CDF
(概率密度函数--概率分布函数 的一个转换操作)

问题: CDF无法not-well-defined
很多函数无法inverse, 因为没有close的函数
数学家: 做了高斯的结合xxx, GMM, (name: Box-Miller)
-----


// sampling有很多方法, 查一查, 要做混合取样方法的化 就是上面那个操作可以考虑考虑
```

Uncertain Scalar Fields

- isosurface

```
等高线、等值面
完全是一个几何问题
```

Level Crossing probability

- definition

一定要openGL\ 一定要

youtube 25个video -----google science

```
入门去看一看
```

基本功

- 统计的tool