Finding “Lookmarks” for Extreme-Scale Simulation and Scientific Data

Lawrence O. Hall and Kevin W. Bowyer[1]
Computer Science & Engineering
University of South Florida
[1]University of Notre Dame

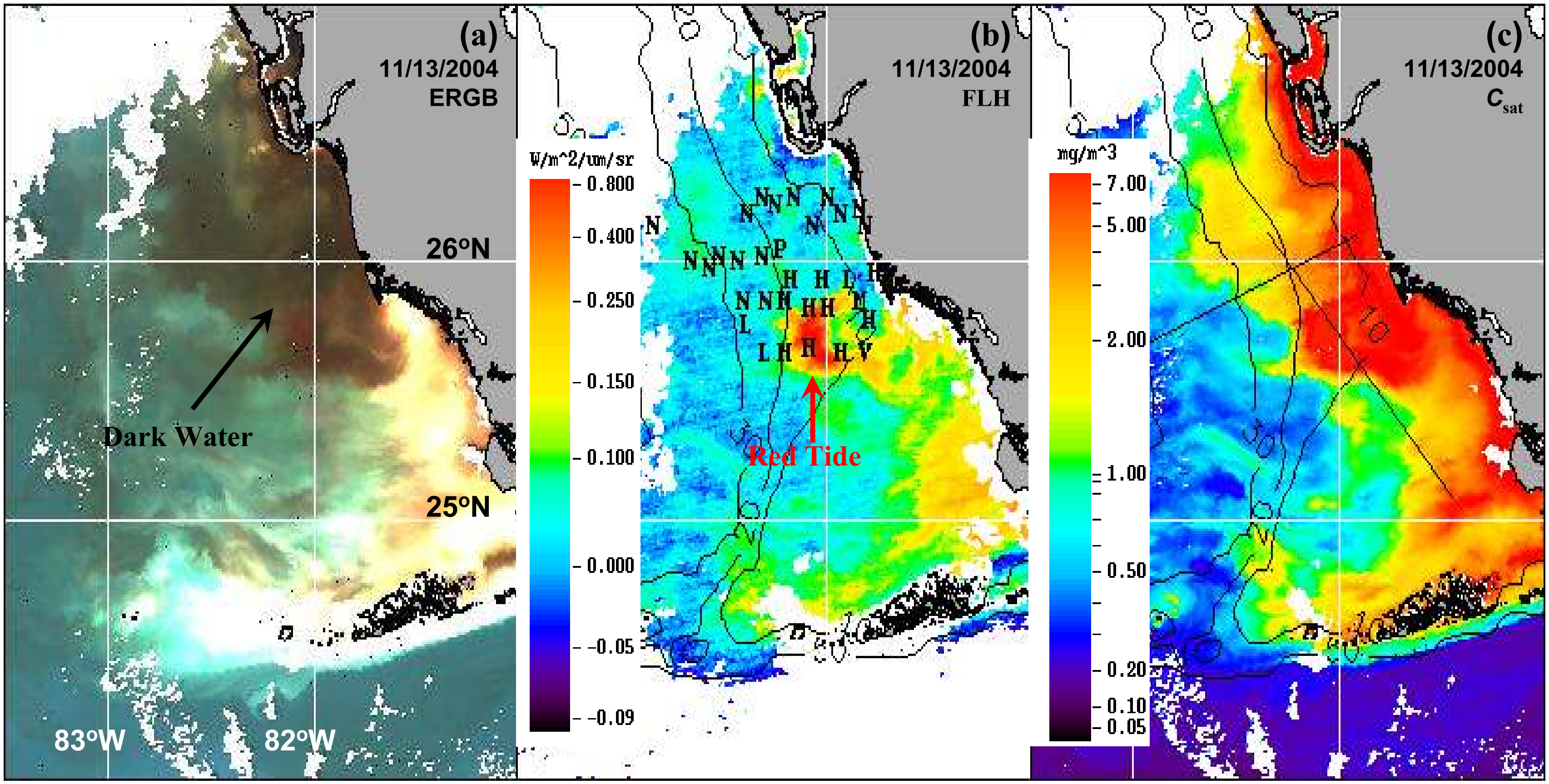
Introduction

- Petascale simulations may require significant time to debug / understand.
- Interesting regions in the simulation are generally a small part of the whole.
- “Lookmarks” that point designers and users to interesting or anomalous regions greatly increase productivity.

Introduction

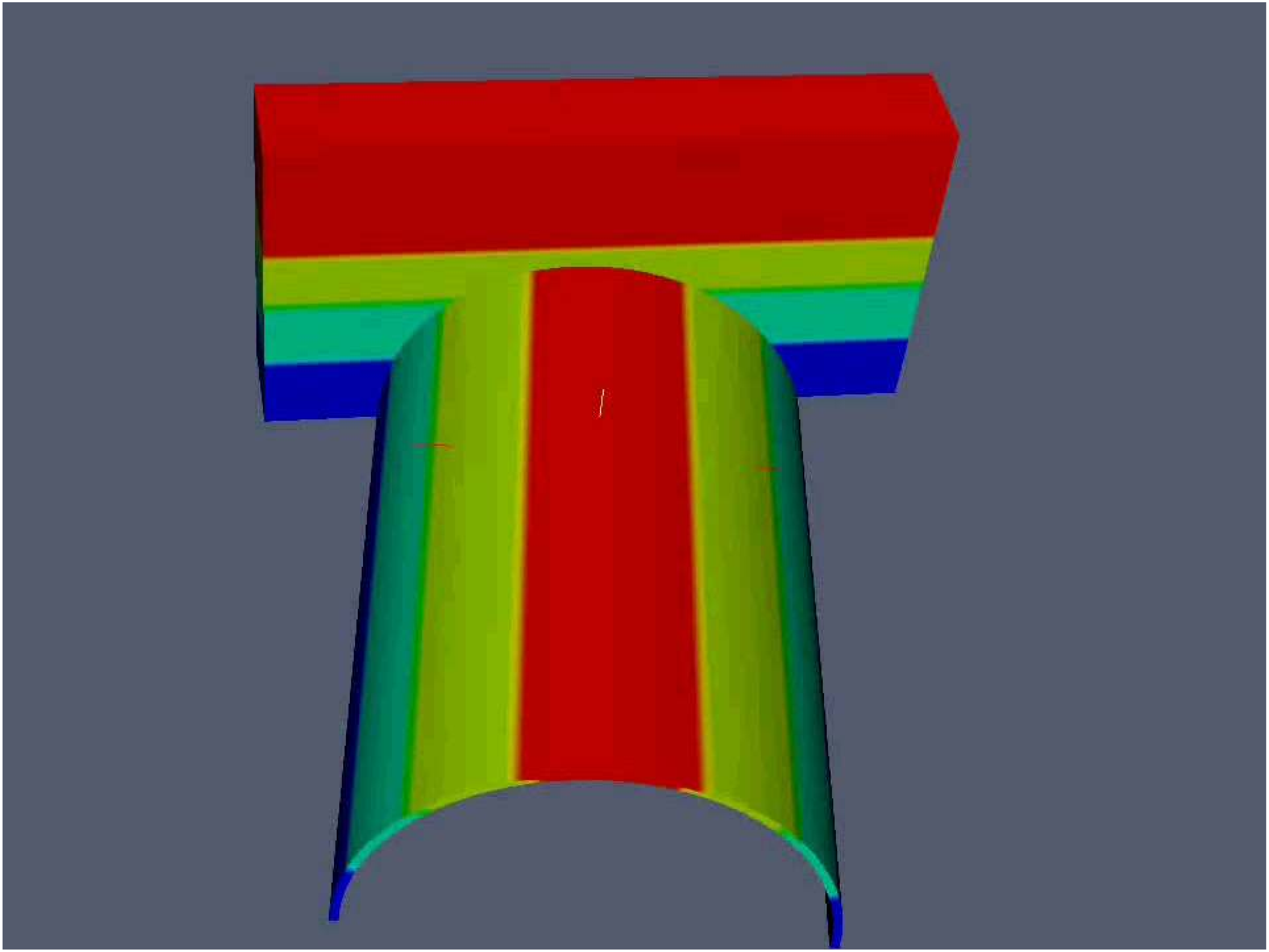
- Large-scale scientific databases are being gathered, as from astronomical observation or satellites.
- It is possible to gather more data than there is time for experts to evaluate.
- Lookmarks can point out interesting and anomalous regions; for example, red tide outbreaks in satellite images.

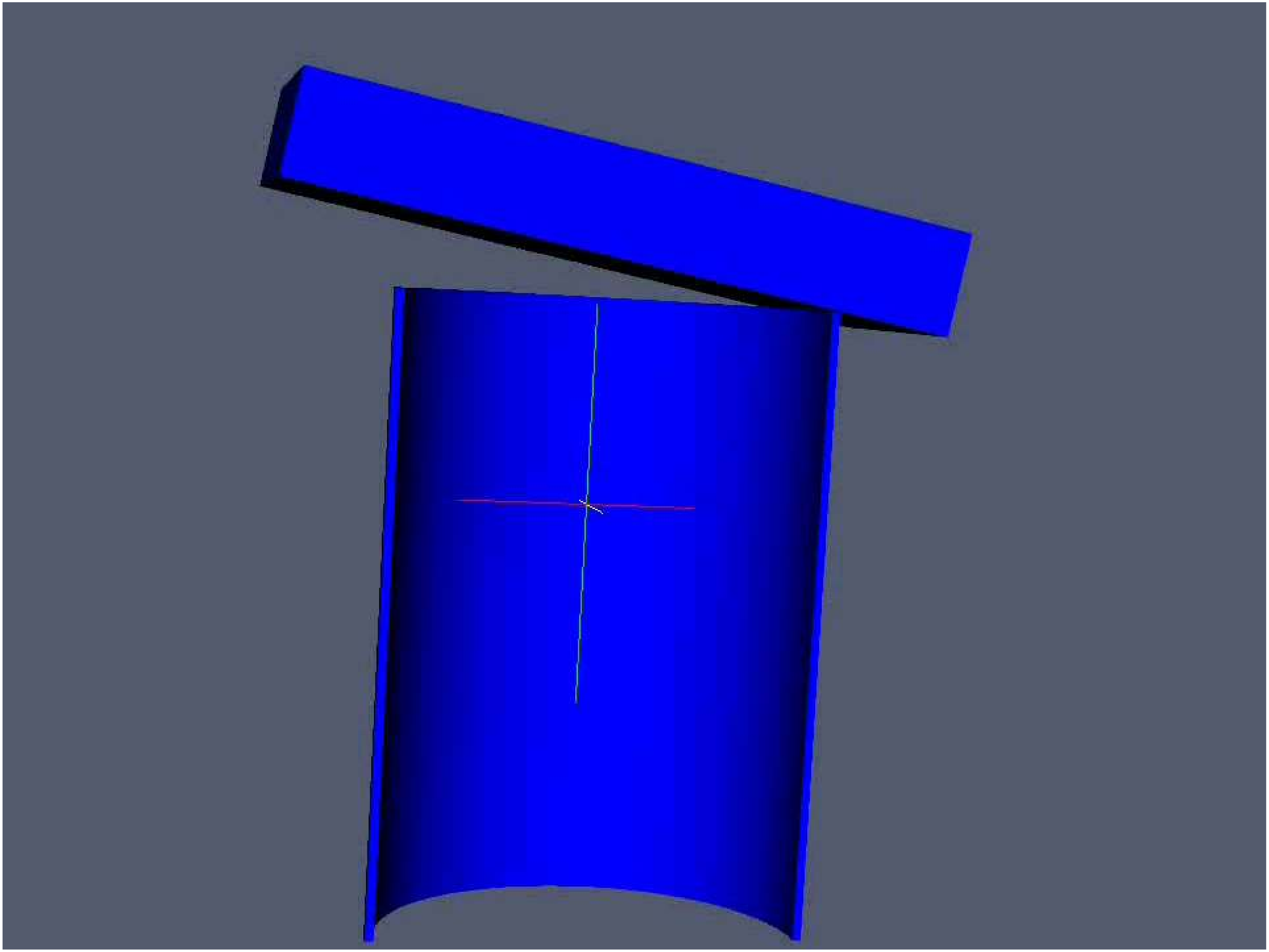
MODIS Images from Southwest Florida Coast



Some Issues

- **The data may be distributed, as in Petascale simulations, in ways that prevent it being all on one CPU.**
- **Classes of interest are typically much smaller than the inhomogeneous "*rest of the data.*"**
- **Most important for users may be to point them to interesting regions, rather than to correctly identify all examples.**





How Important is This Idea?

A “before/after lookmarks” example:

- DOE lab staff ran a particular simulation 162 times, to detect tears/breaches.
- Each run generated 876GB of data.
- 180 person-hours were spent on finding only tears in only the first 12 runs.

How Important is This Idea?

A "before/after lookmarks" example:

- Ground truth from first 12 runs, with an added feature, used to train an "avatar."
- Avatar reviews data, inserts lookmarks.
- 75 person-hours then spent checking 168 runs for tears and breaches.

<u>40% of the time to do 15 times the data!</u>

How Important is This Idea?

- Whenever you can spend about 40% as much time to do about 15x work, it is pretty important.

Challenges

- Getting labeled data to start with.
  - For simulations, for example, likely only some regions in some time steps will be labeled.
  - The “uninteresting” data is likely heterogeneous, with multiple underlying classes.

Challenges

- In region-based prompting, should the learning algorithm have a measure of regional error?
  - Does ground truth overlap define regional error?
  - Should measures of lift be used to indicate how well regions are found?

Challenges

- User-marked regions are likely imprecise.
  - Even an interesting region where a lookmark should exist may be inhomogeneous.

- Distributed data may require a distributed model be learned.

Approaches

- Active learning.
  - May be helpful to indicate the most useful examples/regions for labeling.
  - May be used to gather more salient regions for lookmark generation.

Approaches

- Semi-supervised learning.
  - On a canister simulation looking at stresses on bolts, close to 100% regional accuracy was achieved with relatively few training examples.

Approaches

- Ensembles of classifiers.
  - May be used to learn on distributed data.
  - Can be combined by weighted fusion method.

- Synthetic examples and under-sampling may be applied to address skew in data.

Approaches

- Inhomogeneous classes may be clustered into more homogenous classes to ease the task of a supervised learning algorithm.

Summary

- Advance science faster using “lookmarks” that focus evaluation of large datasets.
- Data mining to develop lookmarks likely requires distributed learning, learning under data skew, and practical ways to enhance the amount of labeled data.
- The combination of approaches needed will cause the underlying approaches to evolve.

Questions ?