INST737

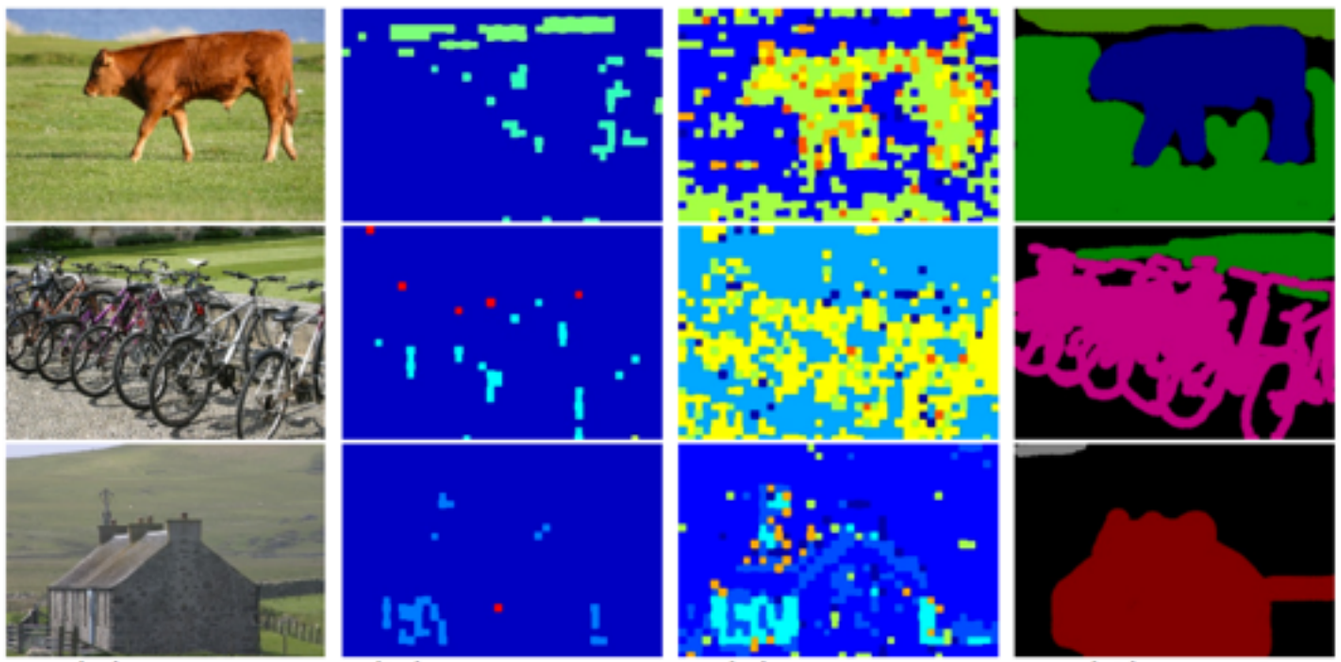
Spatial LDA in Spark

Khoa Doan
Ang Li

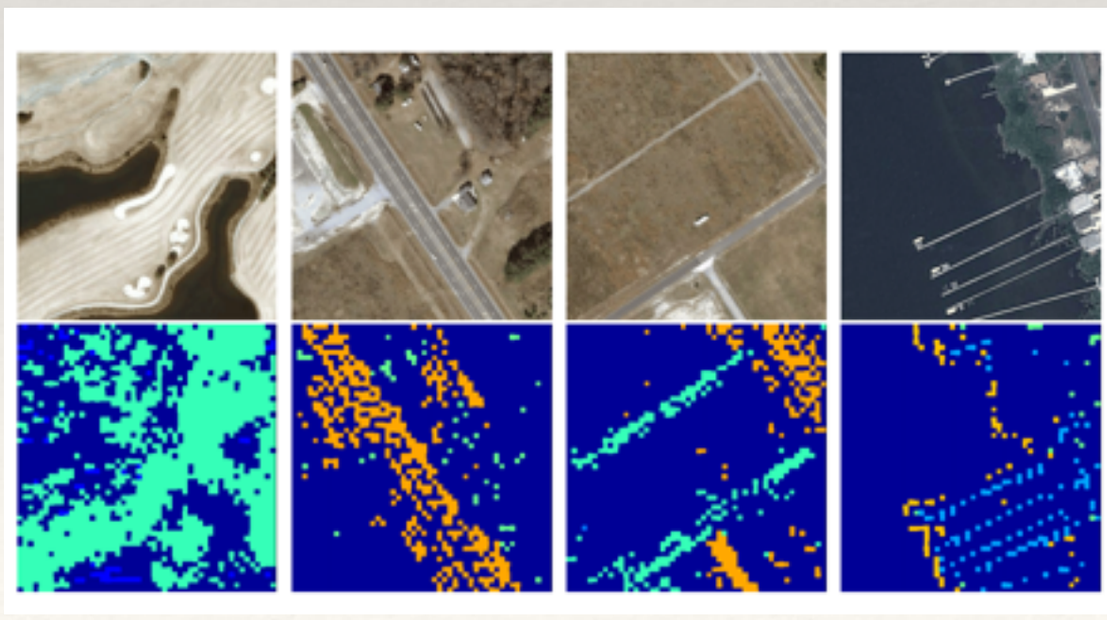
Introduction

- Latent Dirchlet Allocation
  - Topic modeling, for text data
  - Bag of words: easily groups co-occuring topics together
- Spatial Latent Dirichlet Allocation
  - Model spatial distance between pixels (or words)
  - Gibbs Sampling (non-biased, but slow), Variational Inference (faster!)
- Hadoop/MapReduce
  - Gibbs Sampling is ok. Variational Inference is **better**
- Apache Spark: the most anticipated parallel framework
  - All the bests of Hadoop.
  - And solve: multi-staged computation, more flexible high-level operators ….

Examples



(a) origin (b) 14 topics (c) 9 topics (d) truth



Preprocessing

- Many Small Images
  - Combine them into bigger key-value files.
- Transformation from images to documents
  - Pixels -> Group of pixels (Patches)
    - Grid the image, each grid cell is a patch
  - Patch -> Vector, where each component describe something about the patch.
    - Texton features
  - Vectors -> Groups -> Words
    - K-means clustering