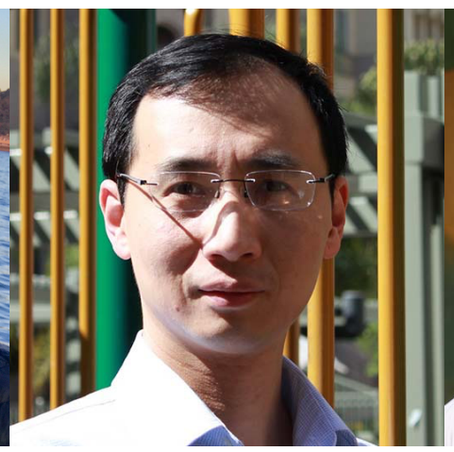
Bayesian Trees for Automated Cytometry Data Analysis

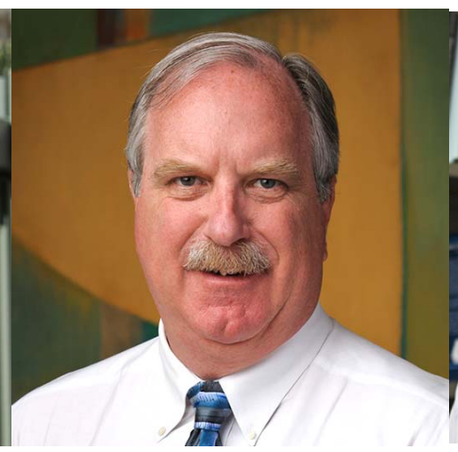
Disi Ji
Department of Computer Science, UC Irvine

Disi Ji
UC, Irvine

Eric Nalisnick
UC, Irvine



Yu Qian
J. Craig Venter Institute



Richard H. Scheuermann
J. Craig Venter Institute

Padhraic Smyth
UC, Irvine

Background

Background

- **Mass cytometry data:**

Background

- **Mass cytometry data:**
  - **High-dimensional single-cell measurements on potentially millions of cells**

Background

- **Mass cytometry data:**
  - High-dimensional single-cell measurements on potentially millions of cells
  - Increasingly used for clinical diagnosis of immunological and hematological conditions

Background

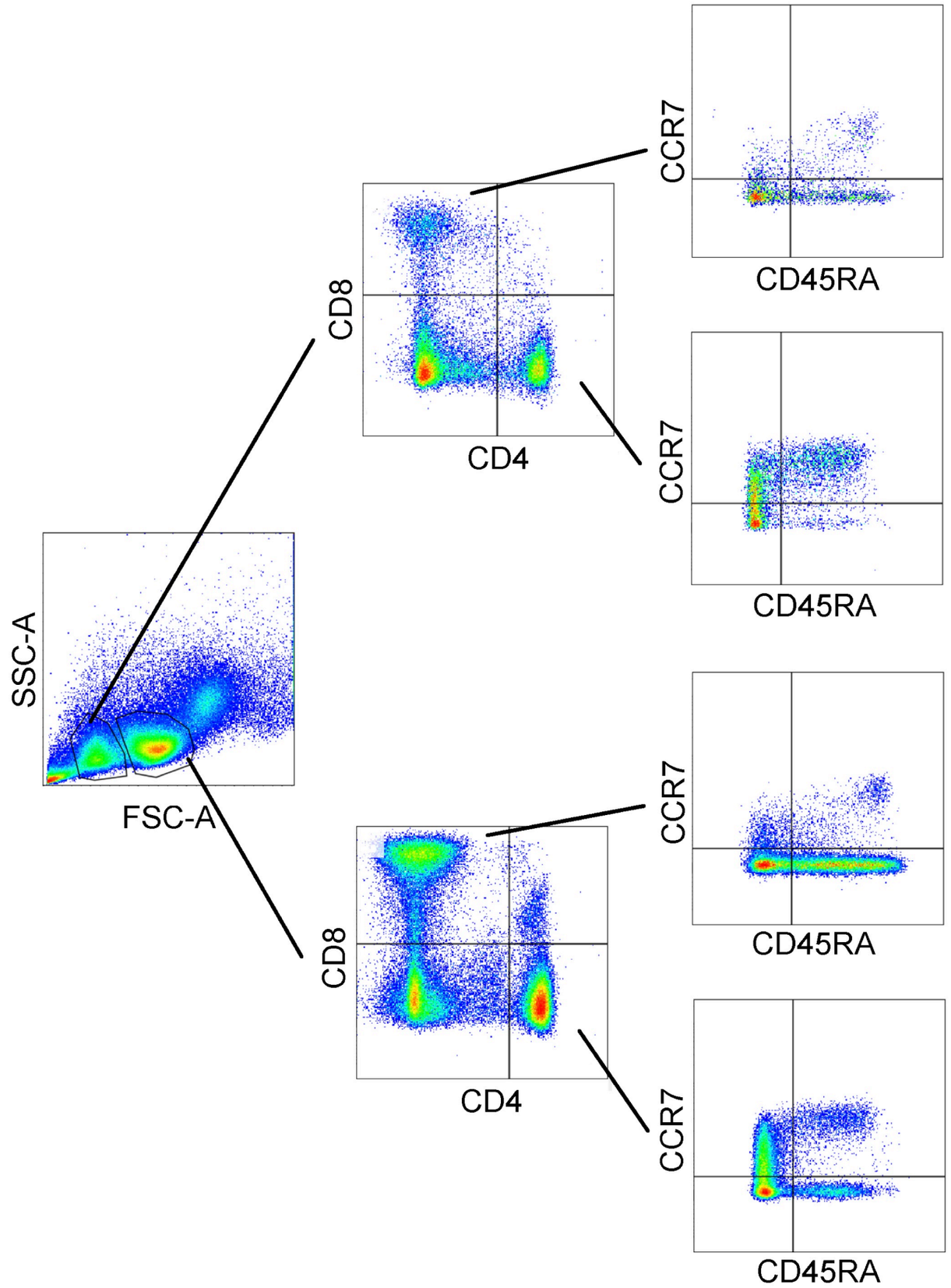
- **Mass cytometry data:**
  - High-dimensional single-cell measurements on potentially millions of cells
  - Increasingly used for clinical diagnosis of immunological and hematological conditions
- **Bottleneck:**

Background

- **Mass cytometry data:**
  - High-dimensional single-cell measurements on potentially millions of cells
  - Increasingly used for clinical diagnosis of immunological and hematological conditions
- **Bottleneck:**
  - Reliance on human classification of cells into cell types

Background

- **Mass cytometry data:**
  - **High-dimensional single-millions of cells**
  - **Increasingly used for clir and hematological condi**
- **Bottleneck:**
  - **Reliance on human class**



Contribution

Contribution

- Built a statistical machine learning model that encodes expert knowledge into a prior, and mimics the tree-structured recursive process of manual classification

Contribution

- Built a statistical machine learning model that encodes expert knowledge into a prior, and mimics the tree-structured recursive process of manual classification
- Completely unsupervised at the cell level: no cell-level labels needed

Contribution

- Built a statistical machine learning model that encodes expert knowledge into a prior, and mimics the tree-structured recursive process of manual classification
- Completely unsupervised at the cell level: no cell-level labels needed
- Comparable cell classification and disease diagnosis accuracy relative to manual classification

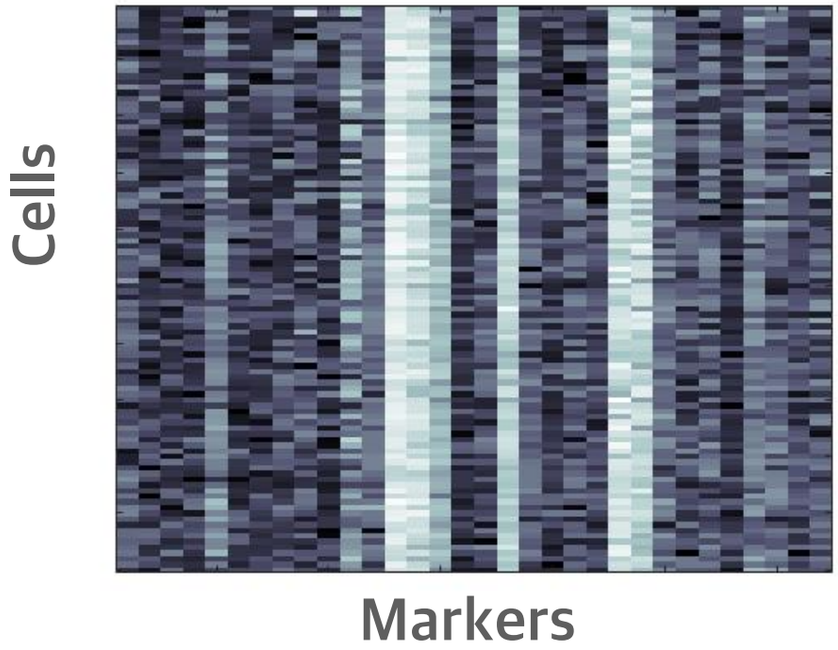
Input: Cytometry Data + Prior Knowledge

Input: Cytometry Data + Prior Knowledge

- **Cytometry Data**
  - Response of each cell to each marker
  - No cell-level label
  - Real Valued

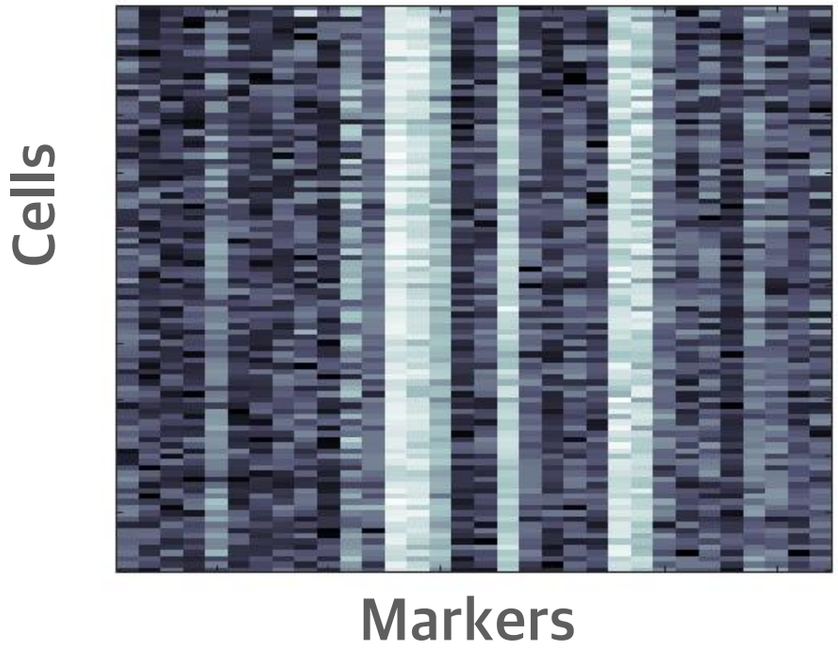
Input: Cytometry Data + Prior Knowledge

- **Cytometry Data**
  - Response of each cell to each marker
  - No cell-level label
  - Real Valued



Input: Cytometry Data + Prior Knowledge

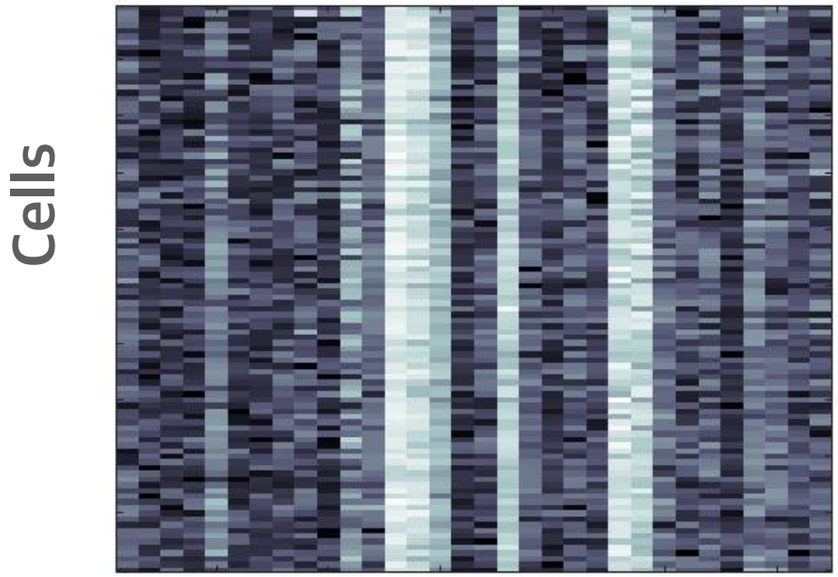
- **Cytometry Data**
  - Response of each cell to each marker
  - No cell-level label
  - Real Valued



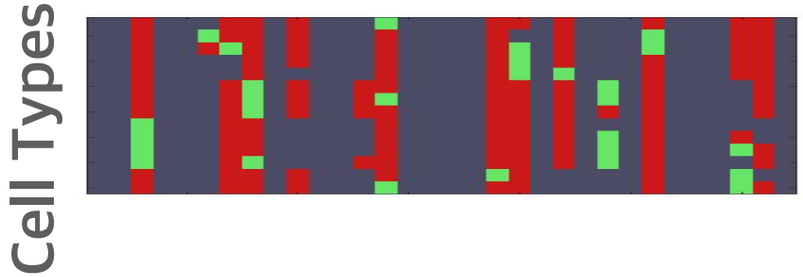
- **Prior Knowledge**
  - Expected response of a cell type to a marker
  - Each entry takes value from {+1, -1, 0}
    - High response: +1
    - Low response: -1
    - Neutral: 0

Input: Cytometry Data + Prior Knowledge

- **Cytometry Data**
  - Response of each cell to each marker
  - No cell-level label
  - Real Valued



- **Prior Knowledge**
  - Expected response of a cell type to a marker
  - Each entry takes value from {+1, -1, 0}
    - High response: +1
    - Low response: -1
    - Neutral: 0



Model: Mondrian Processes with Prior Knowledge

- **Mondrian Process(MP):**
  - A non-parametric process in which a finite region is segmented into rectangular partitions. A distribution over kd-trees T.

Model: Mondrian Processes with Prior Knowledge

- **Mondrian Process(MP):** $P(T)$
  - A non-parametric process in which a finite region is segmented into rectangular partitions. A distribution over kd-trees ***T***.

Model: Mondrian Processes with Prior Knowledge

- **Mondrian Process(MP):** $P(T)$
  - A non-parametric process in which a finite region is segmented into rectangular partitions. A distribution over kd-trees ***T***.

- **MP with Prior Knowledge:**
  - Encode biological knowledge ***K*** into Mondrian processes.
  - Intuition: placing the more discriminative features closer to the root of the tree

Model: Mondrian Processes with Prior Knowledge

- **Mondrian Process(MP):** $P(T)$
  - A non-parametric process in which a finite region is segmented into rectangular partitions. A distribution over kd-trees ***T***.

- **MP with Prior Knowledge:** $P(T \mid K)$
  - Encode biological knowledge ***K*** into Mondrian processes.
  - Intuition: placing the more discriminative features closer to the root of the tree

Model: Mondrian Processes with Prior Knowledge

- **Mondrian Process(MP):** $P(T)$
  - A non-parametric process in which a finite region is segmented into rectangular partitions. A distribution over kd-trees ***T***.

- **MP with Prior Knowledge:** $P(T \mid K)$
  - Encode biological knowledge ***K*** into Mondrian processes.
  - Intuition: placing the more discriminative features closer to the root of the tree

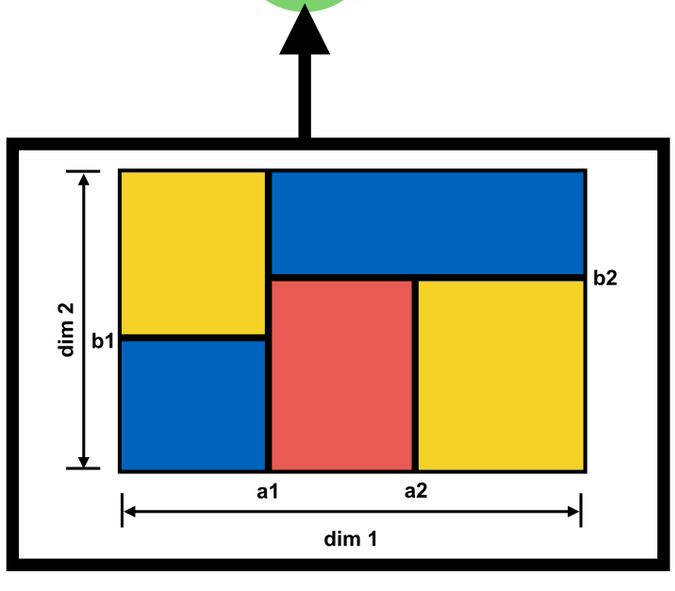
$$P(T \mid D, K) \propto P(T \mid K) \cdot P(D \mid T, K)$$

Model: Mondrian Processes with Prior Knowledge

- **Mondrian Process(MP):** $P(T)$
  - A non-parametric process in which a finite region is segmented into rectangular partitions. A distribution over kd-trees ***T***.

- **MP with Prior Knowledge:** $P(T \mid K)$
  - Encode biological knowledge ***K*** into Mondrian processes.
  - Intuition: placing the more discriminative features closer to the root of the tree

$$P(T \mid D, K) \propto P(T \mid K) \cdot P(D \mid T, K)$$

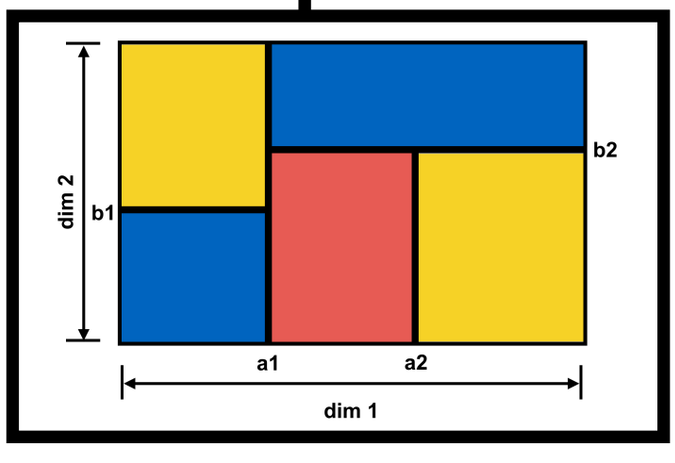


Model: Mondrian Processes with Prior Knowledge

- **Mondrian Process(MP):** $P(T)$
  - A non-parametric process in which a finite region is segmented into rectangular partitions. A distribution over kd-trees ***T***.

- **MP with Prior Knowledge:** $P(T \mid K)$
  - Encode biological knowledge ***K*** into Mondrian processes.
  - Intuition: placing the more discriminative features closer to the root of the tree

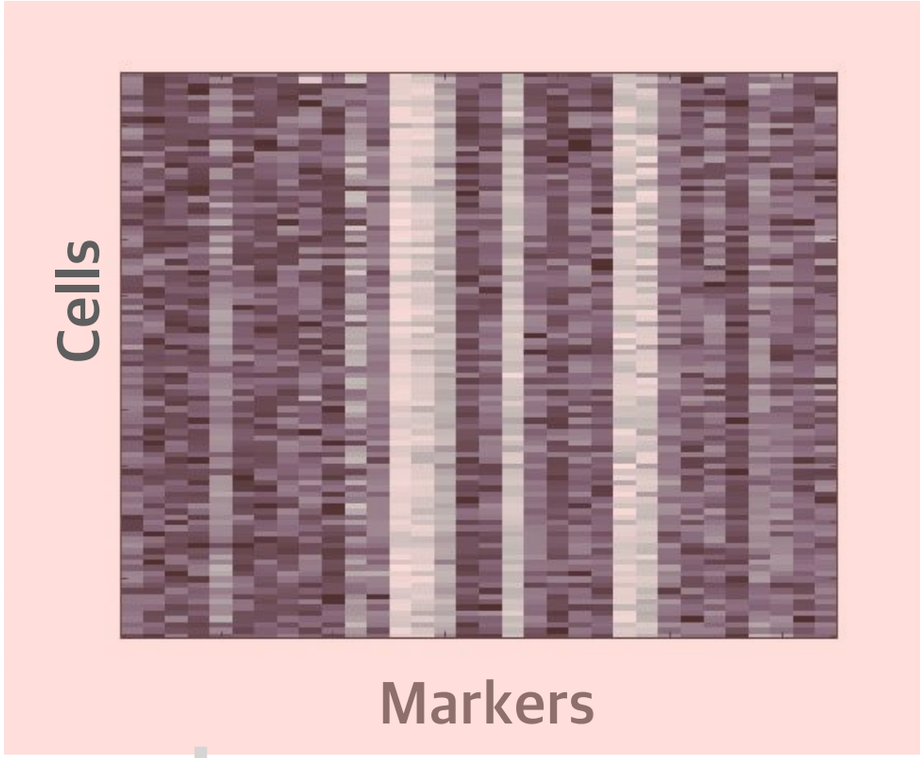
$$P(T \mid D, K) \propto P(T \mid K) \cdot P(D \mid T, K)$$



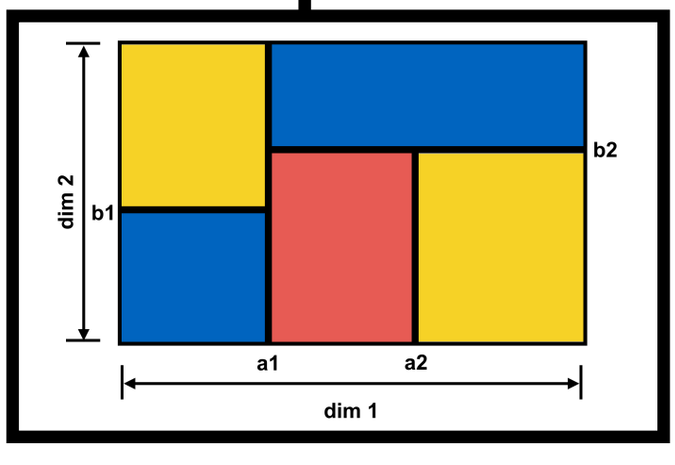
Model: Mondrian Processes with Prior Knowledge

- **Mondrian Process(MP):**
  - A non-parametric process in which a fini
    rectangular partitions. A distribution ov

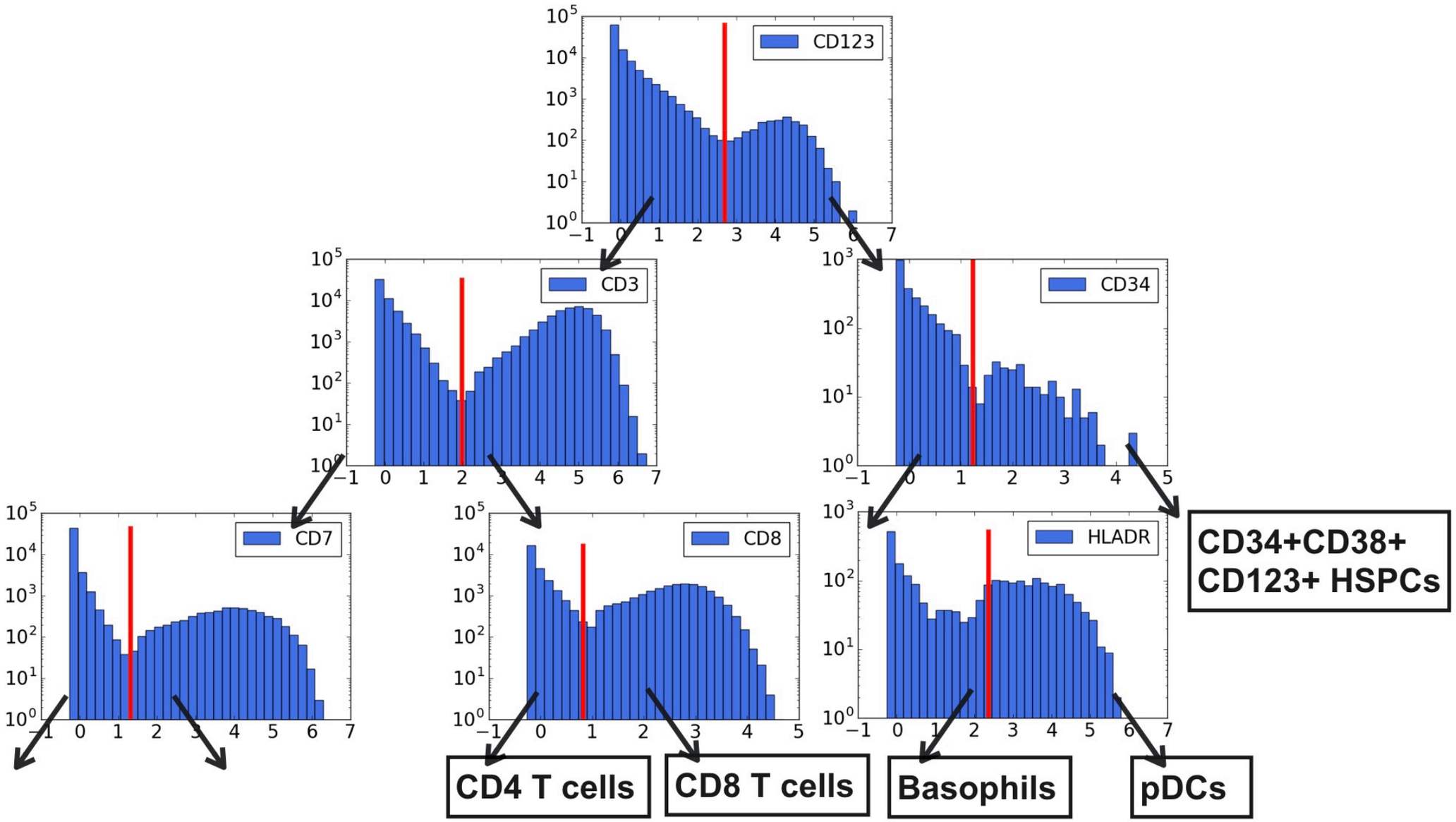
- **MP with Prior Knowledge:**
  - Encode biological knowledge ***K*** into Mc
  - Intuition: placing the more discriminati
    the tree



$$P(T \mid D, K) \propto P(T \mid K) \cdot P(D \mid T, K)$$



Output: Readily Interpretable Classifier

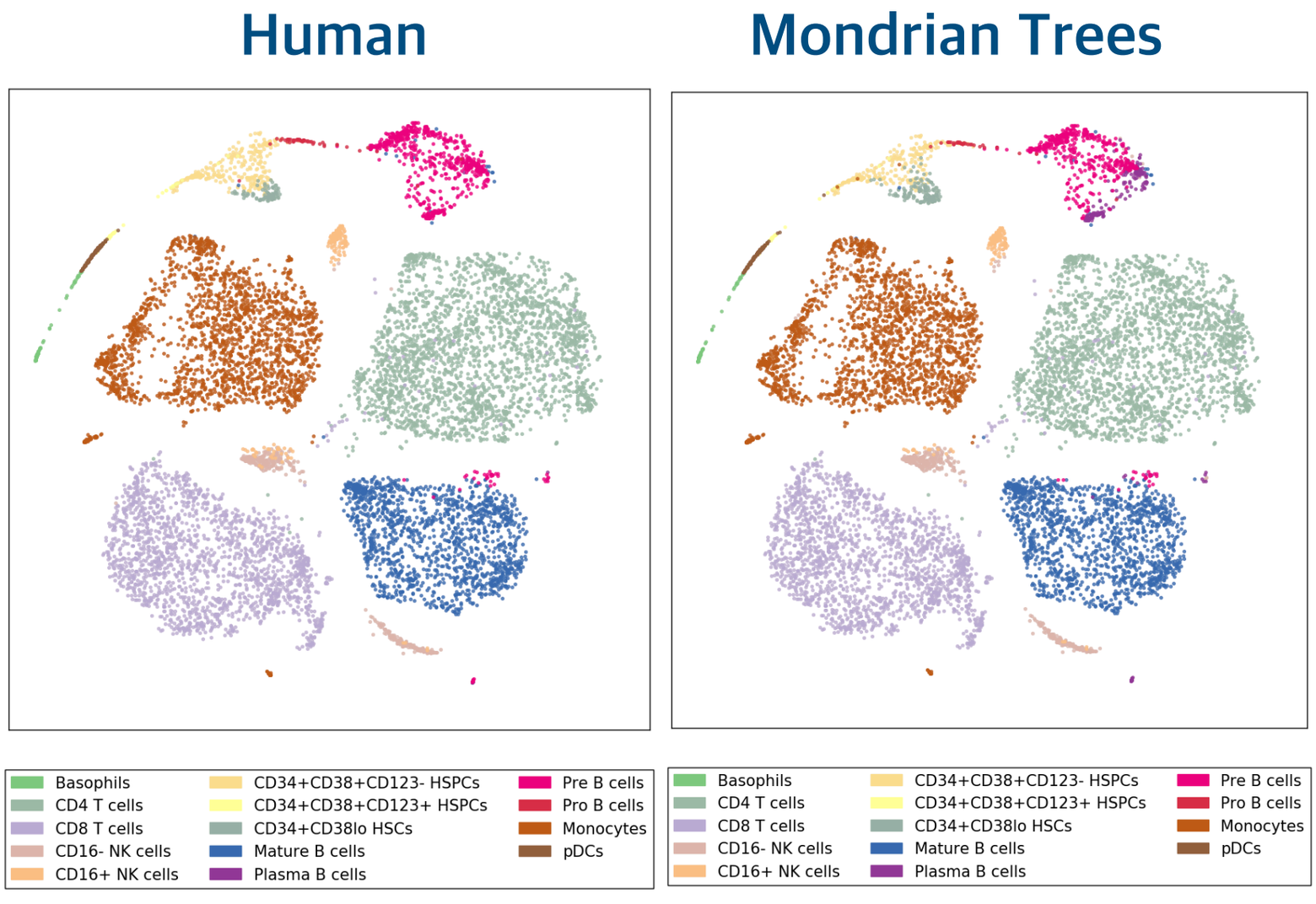


Cell Classification

- Accuracy is comparable to manual classification

Cell Classification

- Accuracy is comparable to manual classification



Learn more at our poster!

Bayesian Trees for Automated Cytometry Data Analysis

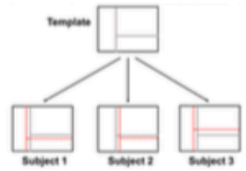
Disi Ji[1], Eric Nalisnick[1], Yu Qian[2], Richard H. Scheuermann[2], Padhraic Smyth[1]
[1]Department of Computer Science, University of California, Irvine, [2]J. Craig Venter Institute

Summary

- **Background:**
  - ✓ mass cytometry data: high-dimensional single-cell measurements on potentially millions of cells
  - ✓ increasingly used for clinical diagnosis of immunological and hematological conditions
  - ✓ bottleneck: reliance on human classification of cells into cell types
- **Goal:** Perform automated cell classification by incorporating prior information table used by mass cytometry experts.
- **Our Contributions:**
  - ✓ Built a statistical machine learning model that encodes expert knowledge into prior
  - ✓ Completely unsupervised at the cell level: no cell-level labels needed
  - ✓ Comparable cell classification and disease diagnosis accuracy relative to manual classification

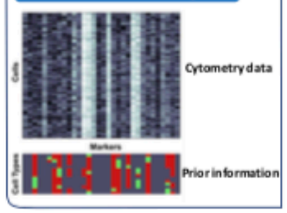
Algorithm 2: MP-RE for Disease Diagnosis

- **Idea:** Individual subjects (lower level) are modeled as instances of a Mondrian process template (upper level) plus **random effects(RE)**
- **Illustration:** The location of cuts for individual subjects are a function of both the observed data for the subject (not shown) and the template.

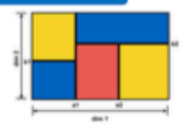


- **Step 1:** Learn template MP for healthy group jointly with REs of each healthy individual.
- **Step 2:** Repeat step 1 on disease group.
- **Step 3**: **Extract features to classify a new sample of cells**,
  - fit two Mondrian trees with RE to the labeled samples, where we estimate an MP-RE tree using the healthy Mondrian template and the other with the disease template.
  - Compute the proportion of cells assigned to each of the cell-types for each tree, resulting in two vectors, which are concatenated to create a final feature vector for prediction per sample

Data + Prior Knowledge



Mondrian Process

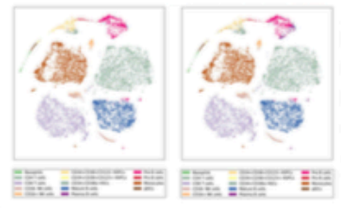


- **Mondrian Process(MP):** A non-parametric process in which a finite region is segmented into rectangular partitions. A prior over kd-trees with the tree depth determined by a lifetime parameter.

Experiments 1: Cell Classification

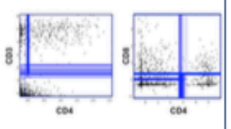
- AML data: 104k cells, 32 biomarkers, 14 cell types.
- BMMC data: 82k cells, 13 biomarkers, 19 cell types
- **Accuracy relative to manual classification:**

	AML	BMMC
Methods without Cell-Level Labels		
MP (Proposed Method)	96.9%	92.3%
MP-Prior	61.5%	85.6%
ACDC	98.2%	93.7%
Methods requiring Cell-Level Labels		
GMM	86.1%	84.1%
Phenograph	95.1%	95.0%

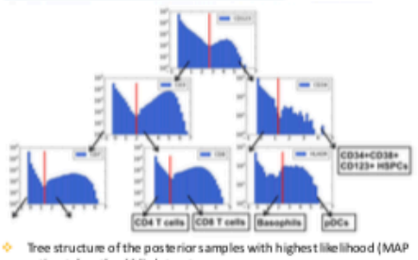


Algorithm1: MP for Automated Classification

- **Step 1: Translate the prior information into prior distributions.**
  - Dimension and position of a cut is distributed based on the set of labels observed in the corresponding column of the prior information table.
- **Step 2: Initialize an MP tree** by sampling from an MP with prior distribution obtained in step 1.
- **Step 3: Optimize joint likelihood** w.r.t. tree structure and cut locations with stochastic search.



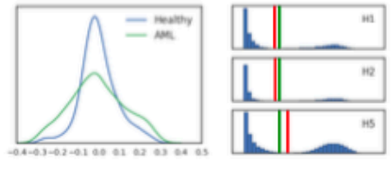
Example of Output of Algorithm 1



- Tree structure of the posterior samples with highest likelihood (MAP estimate) on the AML dataset.
- Red lines denote sampled cuts, and arrows denote the path taken by cells that fall on the left or right side of the cut. The blue rectangles denote cell type classifications.
- Variance of cut positions quantifies uncertainty.

Experiment 2: Disease Diagnosis

- **Data:** AML mass cytometry data set from Levine et al. (2015) consisting of cell-level data with 16 markers for 5 healthy subjects and 16 subjects diagnosed with AML.
- **Prior knowledge** was obtained from the expert tables provided for these markers by Lee et al. (2017)
- **Evaluation:** classification accuracy via leave-one-out cross-validation
- **Results**: MP-RE predicted the correct class label for all 21 samples.



- **Left:** The variability of random effects in the AML group is systematically greater than that of the healthy group.
- **Right**: The cut location of H5 moved towards right, because its upper component contains more data points compared to H1 and H2.