BNP models were developed for \textit{big data}.

...but inference is still slow.

We need to scale up inference for BNP to be applicable.
Our approach

- Conditioned on it, local variables are independent.

Our approach: Instantiate the latent measure.
- Atom weights.
- Atom locations.
Our approach

- Generic Scala code.
- Map/Reduce scheme.
- Two implementations (for both DP and BP):
  - Parallel implementation.
  - Distributed implementation.
    - Apache Spark.
    - Hadoop file system.
Map/Reduce Uncollapsed Gibbs Sampling for Bayesian Nonparametric Models
Experiments

Synthetic dataset (dimensionality $D = 10$)
1M observations (up to 50M)

![Graph showing log-likelihood over time for different algorithms and dataset sizes](image)

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>$N$</th>
<th>100K</th>
<th>1M</th>
<th>5M</th>
<th>50M</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sequential</td>
<td>0.1349</td>
<td>1.3963</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>Parallel</td>
<td>0.0123</td>
<td>0.1397</td>
<td>0.8736</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>Distributed (100K)</td>
<td>0.1795</td>
<td>0.1512</td>
<td>0.2143</td>
<td>1.3429</td>
<td></td>
</tr>
</tbody>
</table>

Time (minutes) per iteration
Thank you for your attention!