Hello! These slides were visual aids for a talk, and weren’t designed to be read. I’ve inserted some notes here to summarize what the points were supposed to be, and to give further references.

1. Hierarchical modelling is essential. Many models contain large numbers of ‘nuisance variables’. We need to learn how these are distributed, because if we make assumptions (including vague or so-called ‘uninformative’ ones), we’ll simply get wrong answers. The model I discussed was referring to: “Inferring the force law in the solar-system from a snapshot”, Bovy et al., 2010. http://arxiv.org/abs/0903.5308

More thoughts and references on hierarchical modelling are in my discussion http://homepages.inf.ed.ac.uk/imurray2/pub/11catchup/catchup.pdf of http://dx.doi.org/10.1111/j.1467-9868.2011.01025.x

Hierarchical models can be hard to infer. Example: http://homepages.inf.ed.ac.uk/imurray2/pub/10hypers/
2. Real-world models have a lot of messy detail:
1) complicated instrument-error distributions we may not care about; 
2) theory encoded in expensive-to-run simulations.

Machine Learning can help.
If we don’t want wrong models (point 1.) we need to learn from large amounts of data about our instruments, and from simulation data describing our theories. I’ve been involved in a series of papers on flexible black-box probabilistic models that could be used here:
http://homepages.inf.ed.ac.uk/imurray2/pub/11nade/
http://homepages.inf.ed.ac.uk/imurray2/pub/13rnade/
http://homepages.inf.ed.ac.uk/imurray2/pub/14dnade/
4. Probabilistic inference methods need extending.
Approximate inference is a heavily-mined and active area. Getting up-to-speed and finding a niche is challenging. However, work in this area is important. In deep and wide graph structures, with billions of observations in some of the plates, it’s hard to do fully Bayesian inference.

Some of my work has been on identifying common small inference problems, which are usually only part of an analysis, and developing easier-to-use inference methods for them. E.g.
http://homepages.inf.ed.ac.uk/imurray2/pub/10ess/

I’m now also interested in developing easier-to-use methods to summarize and communicate the results of local inferences across large models. I believe the way forward is fitting flexible representations of beliefs, by combining machine learning methods and approximate inference algorithms.
Acceleration law around the sun

\[ a(r) = -A \left( \frac{r}{r_0} \right)^{-\alpha} \]

From a snapshot:
8 planet positions and velocities
Graphical model

\[ n=1..N \]

\[ \omega \]

\[ x^{(n)} \]

\[ e^{(n)} \]

\[ n=1..N \]
Hierarchical graphical model

\( \omega \)

\[ x^{(n)} \]

\( e^{(n)} \)

\( \theta \)

\( n=1..N \)
Inferences about the Sun
Priors on nuisance distributions
Priors on nuisance distributions

\[ P(e) \]
\[ e \]

\[ P(e) \]
\[ e \]

\[ P(e) \]
\[ e \]

\[ P(e) \]
\[ e \]
Gravitational exponent
Figure 1: Our unified graphical model (also known as a Bayesian network [27]), for astronomical image data. It integrates in a principled framework: large-scale cosmological models of galaxy and Milky Way formation; galaxy appearance models; spectral emission models and detailed camera, sky and telescope models. The shaded oval nodes are observed variables (i.e., their values are known) while the unshaded ones are unobserved and hence will be inferred from the raw astronomical data. The square nodes represent priors, typically informed by well-understood physics models. The arrows represent dependencies between variables in the model (and the lack thereof correspond to assumptions of independence). The conditional probability distributions within the model (which detail how a particular node depends on those variables which point to it) are not shown, but will be described in the text. The rectangles refer to replications of variables, e.g. an image will contain many stars/galaxies. The realization of this model is the ultimate goal of the project, but initial work will focus on sub-pieces of the model. This figure is best viewed in color.
Machine Learning?

— **Density estimation**
  Neural networks and Gaussian processes

— **Inference methods**
  Statistical methods: MCMC, etc.
  Learning: recognition networks
  Representations: communicating results
Snapshot of the solar-system

Model for the sun: \( \omega = \{\log A, \alpha\} \)

Acceleration law, \( a(r) = -A \left[ \frac{r}{r_0} \right]^{-\alpha} \)

Model for each planet:

\[
\log \epsilon_n \sim p_\epsilon(\cdot | \theta_\epsilon) \quad \text{binding energy}
\]
\[
e_n \sim p(\cdot | \theta_\epsilon) \quad \text{radial asymmetry}
\]
\[
f_n \sim \text{Uniform}[0, 1] \quad \text{fraction of time through orbit}
\]
\[
\text{from aphelion (say), } t/T_{\text{orbit}}
\]

Observations in sky relate to:

\( r, v_r, v_t \): radial distance, radial velocity, transverse velocity
Density estimation

\[ p(x | \theta) \]

Quantiles of \( p(x | \theta) \)

Task: \( \{ x^{(n)} \} \rightarrow \theta \)
GP Density estimation

\[ p(x|f) = \frac{1}{\mathcal{Z}(f)} \Phi(f(x))\pi(x) \]

\( f \sim \mathcal{GP} \)

\( \Phi = \text{sigmoidal function} \)

\( \pi = \text{base measure} \)

Gaussian Process Density Sampler
Modelling via the Chain Rule

\[ P(x) = P(x_1) \prod_{k=2}^{K} P(x_k \mid x_{<k}) \]
Mixture Density Networks

conditional probability density

\[ p(t|x) \]

input vector

Figure stolen from Korin Richmond
Mixture of Gaussian samples
Simulation samples
Results of inference

Prior modeled with MoG

Prior modeled with AMDN
Sequential activation

\[ P(x) = P(x_1 | \theta_1) \ldots \]
Sequential activation

\[ P(x) = P(x_1 | \theta_1) P(x_2 | \theta_2(x_1)) \ldots \]
Sequential activation

\[ P(\mathbf{x}) = P(x_1 | \theta_1) P(x_2 | \theta_2(x_1)) P(x_3 | \theta_3(x_1, x_2)) \ldots \]
Sequential activation

\[ P(x) = P(x_1 | \theta_1) P(x_2 | \theta_2(x_1)) P(x_3 | \theta_3(x_1, x_2)) P(x_4 | \theta_4(x_1, x_2, x_3)) \ldots \]
NADE results

<table>
<thead>
<tr>
<th>Model</th>
<th>ADULT</th>
<th>CONNECT-4</th>
<th>DNA</th>
<th>MUSHROOMS</th>
<th>NIPS-0-12</th>
<th>OCR-LETTERS</th>
<th>RCV1</th>
<th>WEB</th>
</tr>
</thead>
<tbody>
<tr>
<td>MoB</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>± 0.10</td>
<td>± 0.04</td>
<td>± 0.53</td>
<td>± 0.10</td>
<td>± 1.12</td>
<td>± 0.32</td>
<td>± 0.11</td>
<td>± 0.23</td>
</tr>
<tr>
<td>RBM</td>
<td>4.18</td>
<td>0.75</td>
<td>1.29</td>
<td>-0.69</td>
<td>12.65</td>
<td>-2.49</td>
<td>-1.29</td>
<td>0.78</td>
</tr>
<tr>
<td></td>
<td>± 0.06</td>
<td>± 0.02</td>
<td>± 0.48</td>
<td>± 0.09</td>
<td>± 1.07</td>
<td>± 0.30</td>
<td>± 0.11</td>
<td>± 0.20</td>
</tr>
<tr>
<td>RBM mult.</td>
<td>4.15</td>
<td>-1.72</td>
<td>1.45</td>
<td>-0.69</td>
<td>11.25</td>
<td>0.99</td>
<td>-0.04</td>
<td>0.02</td>
</tr>
<tr>
<td></td>
<td>± 0.06</td>
<td>± 0.03</td>
<td>± 0.40</td>
<td>± 0.05</td>
<td>± 1.06</td>
<td>± 0.29</td>
<td>± 0.11</td>
<td>± 0.21</td>
</tr>
<tr>
<td>RBForest</td>
<td>4.12</td>
<td>0.59</td>
<td>1.39</td>
<td>0.04</td>
<td>12.61</td>
<td>3.78</td>
<td>0.56</td>
<td>-0.15</td>
</tr>
<tr>
<td></td>
<td>± 0.06</td>
<td>± 0.02</td>
<td>± 0.49</td>
<td>± 0.07</td>
<td>± 1.07</td>
<td>± 0.28</td>
<td>± 0.11</td>
<td>± 0.21</td>
</tr>
<tr>
<td>FVSBN</td>
<td>7.27</td>
<td>11.02</td>
<td>14.55</td>
<td>4.19</td>
<td>13.14</td>
<td>1.26</td>
<td>-2.24</td>
<td>0.81</td>
</tr>
<tr>
<td></td>
<td>± 0.04</td>
<td>± 0.01</td>
<td>± 0.50</td>
<td>± 0.05</td>
<td>± 0.98</td>
<td>± 0.23</td>
<td>± 0.11</td>
<td>± 0.20</td>
</tr>
<tr>
<td>NADE</td>
<td>7.25</td>
<td>11.42</td>
<td>13.38</td>
<td>4.65</td>
<td>16.94</td>
<td>13.34</td>
<td>0.93</td>
<td>1.77</td>
</tr>
<tr>
<td></td>
<td>± 0.05</td>
<td>± 0.01</td>
<td>± 0.57</td>
<td>± 0.04</td>
<td>± 1.11</td>
<td>± 0.21</td>
<td>± 0.11</td>
<td>± 0.20</td>
</tr>
</tbody>
</table>

| Normalization | -20.44 | -23.41 | -98.19 | -14.46 | -290.02 | -40.56 | -47.59 | -30.16 |

**★ Little variation when changing input ordering:**
- DNA = +/- 0.05
- MUSHROOMS = +/- 0.045
- NIPS-0-12 = +/- 0.15
### NADE results

<table>
<thead>
<tr>
<th>Model</th>
<th>Log. Like.</th>
</tr>
</thead>
<tbody>
<tr>
<td>MoB*</td>
<td>-137.64</td>
</tr>
<tr>
<td>RBM (CD1)*</td>
<td>-125.53</td>
</tr>
<tr>
<td>RBM (CD3)*</td>
<td>-105.50</td>
</tr>
<tr>
<td>RBM (CD25)*</td>
<td>-86.34</td>
</tr>
<tr>
<td>FVSBN</td>
<td>-97.45</td>
</tr>
<tr>
<td>NADE</td>
<td>-88.86</td>
</tr>
</tbody>
</table>

#### Intractable

- RBM (CD1)*
- RBM (CD3)*
- RBM (CD25)*

#### Samples

![Samples](image_url)

*Approximations taken from Salakhutdinov and Murray (2008).*
### RNADE results

<table>
<thead>
<tr>
<th>Dataset</th>
<th>dim</th>
<th>size</th>
<th>Gaussian</th>
<th>MFA</th>
<th>FVBN</th>
<th>RNADE-MoG</th>
<th>RNADE-MoL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Red wine</td>
<td>11</td>
<td>1599</td>
<td>−13.18</td>
<td>−10.19</td>
<td>−11.03</td>
<td>−9.36</td>
<td>−9.46</td>
</tr>
<tr>
<td>White wine</td>
<td>11</td>
<td>4898</td>
<td>−13.20</td>
<td>−10.73</td>
<td>−10.52</td>
<td>−10.23</td>
<td>−10.38</td>
</tr>
<tr>
<td>Parkinsons</td>
<td>15</td>
<td>5875</td>
<td>−10.85</td>
<td>−1.99</td>
<td>−0.71</td>
<td>−0.90</td>
<td>−2.63</td>
</tr>
<tr>
<td>Ionosphere</td>
<td>32</td>
<td>351</td>
<td>−41.24</td>
<td>−17.55</td>
<td>−26.55</td>
<td>−2.50</td>
<td>−5.87</td>
</tr>
<tr>
<td>Boston housing</td>
<td>10</td>
<td>506</td>
<td>−11.37</td>
<td>−4.54</td>
<td>−3.41</td>
<td>−0.64</td>
<td>−4.04</td>
</tr>
</tbody>
</table>
### RNADE results

<table>
<thead>
<tr>
<th>Model</th>
<th>Training LogL</th>
<th>Test LogL</th>
</tr>
</thead>
<tbody>
<tr>
<td>MoG $N = 50$</td>
<td>111.6</td>
<td>110.4</td>
</tr>
<tr>
<td>MoG $N = 100$</td>
<td>113.4</td>
<td>112.0</td>
</tr>
<tr>
<td>MoG $N = 200$</td>
<td>113.9</td>
<td>112.5</td>
</tr>
<tr>
<td>MoG $N = 300$</td>
<td>114.1</td>
<td>112.5</td>
</tr>
<tr>
<td>RNADE-MoG $K = 10$</td>
<td>125.9</td>
<td>123.9</td>
</tr>
<tr>
<td>RNADE-MoG $K = 20$</td>
<td>126.7</td>
<td><strong>124.5</strong></td>
</tr>
<tr>
<td>RNADE-MoL $K = 10$</td>
<td>120.3</td>
<td>118.0</td>
</tr>
<tr>
<td>RNADE-MoL $K = 20$</td>
<td>122.2</td>
<td>119.8</td>
</tr>
</tbody>
</table>

Figure 2: **Top:** 15 datapoints from the TIMIT core-test set. **Center:** 15 samples from a MoG model with 200 components. **Bottom:** 15 samples from an RNADE with 1024 hidden units and output components per dimension. On each plot, time is shown on the horizontal axis, the bottom row displays the energy feature, while the others display the filter bank features (in ascending frequency order from the bottom). All data and samples were drawn randomly.
Deep learning?
One motivation

![Graph showing the relationship between number of weights and phone error rate for different layers (128, 256, 512, 1024 units per layer). The graph demonstrates a decrease in phone error rate as the number of weights increases.](image)

Figure by Benigno Uria
Sequential deep activation

$θ_1 \theta_2 \theta_3 \theta_4 \ x_1 \ x_2 \ x_3 \ x_4 \ H \ hidden \ units \ O(H^2) \ links \ H \ hidden \ units \ x_1 \ x_2 \ x_3 \ x_4$
Sequential deep activation

\[ \theta_1 \theta_2 \theta_3 \theta_4 \]

\[ x_1 \ x_2 \ x_3 \ x_4 \]

\[ H \text{ hidden units} \]

\[ O(H^2) \text{ links} \]
Sequential deep activation

\[ \theta_1 \theta_2 \theta_3 \theta_4 \]
\[ x_1 x_2 x_3 x_4 \]
\[ H \text{ hidden units} \]
\[ \mathcal{O}(H^2) \text{ links} \]
\[ H \text{ hidden units} \]
\[ \mathcal{O}(DH^2) \]
A completing machine

$\theta_1 \theta_2 \theta_3 \theta_4 \theta_5$

$x_1 x_2 x_3 x_4 x_5$

$H$ hidden units

$\mathcal{O}(H^2)$ links
Deep NADE

Train time: $O(DH + H^2)$ per update

Test time: predict features in any order (can condition on observations)

Different orderings not consistent:
   — Seems bad, but . . .
   — have trained large ensemble
   — combining different orderings works better
Arbitrary ordering: inpainting

<table>
<thead>
<tr>
<th>-61.21</th>
<th>-36.33</th>
</tr>
</thead>
<tbody>
<tr>
<td>9</td>
<td>1</td>
</tr>
<tr>
<td>-84.40</td>
<td>-46.22</td>
</tr>
<tr>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>-96.68</td>
<td>-66.26</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>-86.37</td>
<td>-73.31</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>-93.35</td>
<td>-79.40</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>-45.84</td>
<td>-41.88</td>
</tr>
<tr>
<td>\</td>
<td>\</td>
</tr>
</tbody>
</table>
Deep ensembles — results

Small improvements across most UCI datasets

Finally beat MoG on image patches:

Table 4. Average test-set log-likelihood for several models trained on 8 by 8 pixel patches of natural images taken from the BSDS300 dataset. Note that because these are log probability densities they are positive, higher is better.

<table>
<thead>
<tr>
<th>Model</th>
<th>Test LogL</th>
</tr>
</thead>
<tbody>
<tr>
<td>MoG $K = 200$ (Zoran &amp; Weiss, 2012)</td>
<td>152.8</td>
</tr>
<tr>
<td>RNADE 1hl (fixed order)</td>
<td>152.1</td>
</tr>
<tr>
<td>RNADE 1hl</td>
<td>143.2</td>
</tr>
<tr>
<td>RNADE 2hl</td>
<td>149.2</td>
</tr>
<tr>
<td>RNADE 3hl</td>
<td>152.0</td>
</tr>
<tr>
<td>RNADE 4hl</td>
<td>153.6</td>
</tr>
<tr>
<td>RNADE 5hl</td>
<td>154.7</td>
</tr>
<tr>
<td>RNADE 6hl</td>
<td>155.2</td>
</tr>
<tr>
<td>EoRNADE 6hl 2 ord.</td>
<td>156.0</td>
</tr>
<tr>
<td>EoRNADE 6hl 32 ord.</td>
<td>157.0</td>
</tr>
</tbody>
</table>

ever reported on this task. Ensembles of RNADEs also show

Figure 5. Top: 50 examples of $8 \times 8$ patches in the BSDS300 dataset ordered by decreasing likelihood under a 6-hidden-layer NADE. Bottom: 50 samples from a 6-hidden-layer NADE.