**Combating uncertainty:**

Before you can combat uncertainty, you have to honestly quantify the uncertainty present.

Representing uncertainty in terms of probabilities is the natural framework, and allows a huge body of statistical tools to be readily applied.

It also corresponds to what we do - "quantified common sense".

"The theory of probabilities is simply the abandoned attempt to measure what we cannot measure." - Laplace

Tools now exist that allow for computation in complex statistical models.

**Probabilistic formulation:**

- Start with the distribution of what we do know (AVIRIS data), conditional on everything we don't know (NDVI, LAI, AWC)
  \[ p(\text{ndvi}, \text{lai}, \text{awc}) \]

- Invert using Bayes' Theorem
  \[ p(\text{ndvi}, \text{lai}, \text{awc}) = \frac{p(\text{aviris} | \text{ndvi}, \text{lai}, \text{awc}) \times p(\text{aviris})}{p(\text{ndvi}, \text{lai}, \text{awc})} \]

- Look at the conditional independence structure of the problem

**Case study:**

- Soil water is often a limiting factor in plant growth.
- Ecologists are therefore interested in soil water over large areas.
- Directly measuring soil water is difficult, involving extensive fieldwork and laboratory work.
- Satellite data can be used to infer plant growth.
- Under certain circumstances, plant growth can be linked to soil water.

**Inputs:**

- Satellite observations
  - AVIRIS data
  - Leaf Area Index (LAI)
- STATSGO database
  - Soil properties database
  - Provides indirect information on soil water from an unknown (and most probably sparse) and irregular sampling of the input
- Task is to relate AVIRIS data to soil water and combine with the STATSGO information.

**A Case Study:**

High Resolution Soil Water from Regional Databases and Satellite Images

Robin D. Morris, Vadim N. Smelyanskiy in collaboration with Joseph Doughan, Jennifer Doughan

Soil water is often a limiting factor in plant growth.

Ecologists are therefore interested in soil water over large areas.

Directly measuring soil water is difficult, involving extensive fieldwork and laboratory work.

Satellite data can be used to infer plant growth.

Under certain circumstances, plant growth can be linked to soil water.
Probabilistic formulation (cont):

- $p(\text{aviris} | \text{ndvi}, \text{lai}, \text{awe})$
- $p(\text{ndvi}, \text{lai}, \text{awe})$
- $p(\text{aviris} | \text{ndvi})$
- $p(\text{ndvi} | \text{lai}, \text{awe})$
- $p(\text{lai} | \text{awe})$
- $p(\text{awe})$

Sources of uncertainty:

- $p(\text{aviris} | \text{ndvi})$
- $p(\text{ndvi} | \text{lai}, \text{awe})$
- $p(\text{ndvi} | \text{awc})$
- $p(\text{awc} | \text{awe})$
- $p(\text{awe})$
- $p(\text{aviris} | \text{ndvi})$
- $p(\text{ndvi} | \text{lai})$
- $p(\text{lai} | \text{awc})$
- $p(\text{awe} | \text{awc})$

Sources of uncertainty (cont):

- More fieldwork - measuring LAI for a number of plots, and also making laboratory measurements of soil water.
- Nemanzi Running (1988) have the data points for this graph:

- Again, sample from $p(\text{A, B, } \alpha | \text{data})$
  where $\text{la} = \text{A} \times \text{awc} + \text{B} + \epsilon$
  $p(\text{la} | \text{awc}) = N(\text{A} \times \text{awc} + \text{B}, \sigma)$

Sources of uncertainty (cont):

- Nemanzi Running (1988) have the data points for this graph:

- The form of the curve, $\text{ndvi} = \text{A log}(\text{B} \times \text{la})$, comes from ecological theory.

- But what about the parameters?

Sources of uncertainty (cont):

- This comes from fieldwork - measuring LAI for a number of plots, and finding the NDVI values for those plots.
- Derived from AVIRIS image taken at the same time.
- Running et al. (1999) have the following graph:

- If we assume that $\text{ndvi} = \text{A log}(\text{B} \times \text{la}) + \epsilon$
  where $\epsilon \sim \mathcal{N}(0, \sigma)$

- Then we can write
  $p(\text{A, B, } \alpha | \text{data}) = p(\text{data} | \text{A}, \text{B}, \alpha) p(\text{A}, \text{B}, \alpha)$

- and realizations from $p(\text{A, B, } \alpha | \text{data})$ can be used to represent this distribution (more on this later)

- Blue lines: curves from (A,B) values
- Green lines: $\approx 2 \sigma$

- Slicing through this graph we can form
  $p(\text{aviris} | \text{lai}) = \mathcal{N}(\text{A log}(\text{B} \times \text{la}), \sigma)$
  where (A, B, $\alpha$) are mean values

- Blue curve gives mean values the uncertainty is in the parameters

- Nemanzi Running (1988) have the data points for this graph:

- Again, sample from $p(\text{A, B, } \alpha | \text{data})$

- This is the second source of information - the STATSGO database provides prior information on the distribution of AWC values.

For each layer we are given the mean and soil type that was considered in the sampling, and the mean and soil depth.
For each component we are given the % of the polygon that has the given component.
Assuming that the sampling misses the tails of the distribution, we form a mixture model where the mixture values are taken as $\approx 2\sigma$ for that layer.
Putting it all together:
- \( p(\text{awc}, \text{lai} | \text{ndvi}) = p(\text{ndvi} | \text{lai}) p(\text{lai} | \text{awc}) p(\text{awc}) \)
- what we're really interested in is
- \( p(\text{awc} | \text{ndvi}) = \int p(\text{awc}, \text{lai} | \text{ndvi}) \)
- This can be approximated by sampling from \( p(\text{awc}, \text{lai} | \text{ndvi}) \) and then ignoring the lai values.

Results:
- Oregon NOVI
- Polygon OR149

Results (cont):
Distribution of AWC from sampling, excluding the STATSGO prior

Results (cont):
- mean lai
- variance of lai
- mean awc
- variance of awc

Discussion:
- The awc distributions shown are for sampling \( p(\text{ndvi} | \text{lai}) p(\text{lai} | \text{awc}) \), ie not including the prior \( p(\text{awc}) \).
- Compare the distribution from sampling with the STATSGO prior.
- See the uncertainty is still on the level of the components.
- However, prior says that this component is only 40% of the polygons, but the pred map seems to suggest that all of the others are approximately the distribution.

Further refinements:
- Improving \( p(\text{ndvi} | \text{lai}) \) and \( p(\text{lai} | \text{awc}) \) can be done by incorporating other types of information, eg
  - tree species,
  - age,
  - time since last harvest etc,
  - but all of these require more fieldwork.
- Including a spatial prior will reduce variation across the polygons.
Representing distributions using samples

- We wish to represent \( p(x) \), and then to compute expectations

\[
E(f(x)) = \int f(x) p(x) \, dx
\]

- If we sample the domain of \( x \) uniformly, and compute \( p(x) \) for each \( x \) value, then the integral can be approximated by

\[
E(f(x)) = \sum_i f(x(i)) p(x(i))
\]

- However, this becomes inefficient if \( x \) is in more than a few dimensions (curse of dimensionality)

- Instead, if we sample from \( p(x) \), i.e., concentrate the samples in the high-probability regions, then we can approximate the integral by

\[
E(f(x)) = \sum_i f(x(i))
\]

Markov chain Monte Carlo:

- MCMC is a method of generating samples from \( p(x) \).

A Markov chain is a sequence of states, where the probability of the next state depends only on the current state.

- The method constructs a Markov chain that converges to the distribution of interest, \( p(x) \).

- To do this, it is necessary to determine the Transition Probabilities.

The Metropolis Algorithm is the simplest scheme for doing this:

- Initialize \( x \) randomly
- Propose a new value \( x' \), where one of the elements of \( x \) is changed by drawing it from a symmetric distribution
- Accept or reject the new value with probability

\[
P = \min \left( 1, \frac{p(x')}{p(x)} \right)
\]

- Otherwise retain the current value, \( x \)
- Store the realizations

- The realizations are an approximate sample from the posterior, from which we can compute quantiles of interest (means, variances, etc.).