**Longitudinal AIDS data** Large scale clinical trials often require extended periods of followup to evaluate the clinical efficacy of a new treatment, as measured for example by survival time.

In AIDS research, the life-threatening nature of the disease heightens the need for rapid dissemination of knowledge.

So we select an easily-measured biological marker, known to be predictive of the clinical outcome, as a surrogate endpoint.

Example: Number of CD4 lymphocytes per cubic millimeter of drawn blood: a popular marker for progression to AIDS.

CD4 counts are usually recorded at study entry and again at the 2,6,12 and 18 months visit.

Main goal is to analyze the association among CD4 count, survival time, drug group and AIDS diagnosis at study entry (an indicator of disease progression status).

(i) Correlated count data with time effect.
(ii) Presence of several missing values.
(iii) Presence of other explanatory variables (covariates): age, smoker/not, food habits.

**Statistical Maps** Disease incidence data are available as summary counts or rates for a defined region, such as a county, district, or census tract.

Within a region, suppose that counts or rates are observed for subgroups of the population defined by sociodemographic variables such as gender, race, age and ethnicity.

Suppose further that for each subgroup within each region, counts or rates are collected within regular time intervals: e.g: annually.

We have the raw data and using MLE we can create crude maps.

It fails to take advantage of the similarity of relative risks in adjacent or nearby regions.

We need to smooth this crude map as well as explain the variation in risk.

Prior distribution can work as a smoother here.

(i) Count data
(ii) Time as well as spatial effects
(iii) Presence of covariates

**Flow of fluid** Permeability predictions in reservoirs which is an important problem for petroleum engineer.

Physically permeability arises both from the existence of pores and from the average structure of the connectivity of pores.
Mathematically, fluid flow can typically be described by Darcy’s law, which states that for steady-state flow in a porous medium, 

\[ v = -\rho \frac{\delta p}{L} \mu, \]

where \( \rho \) is the permeability, \( v \) is the volume flux per surface area of some region of length \( L \), \( \mu \) is the viscosity and \( p \) is the pressure. The key role of permeability is evident from Darcy’s law.

Due to petrophysical variations rooted in digenesis, grain size variation, cementation, we observe highly and sharply heterogeneous behavior of the process at different regions of the reservoir so assumption stationarity (smoothness) is no more true.

Addition to that, we have several data sources (cores, logs, seismic traces, inter-well tracer data and pressure transient tests), so we can reduce uncertainty by using data integration from these variety of sources.

Bayesian Hierarchical models and borrowing strength principals are natural here.

**Statistical learning**

Learning from the data.

(i) Outcome measurement: quantitative (stock price) or categorical (like heart attack/not)

(ii) Wish to predict based on features (like diet and clinical measurements).

(iii) Have training set of data: we observe the outcome and feature measurements for a set of objects (such as people)

(iv) Using this data we build a prediction model or learner which will enable us to predict the outcome for new unseen object.

(v) Good learner predicts accurately.

(1) Supervised learning: Presence of the outcome variable to guide the learning process.

(ii) Unsupervised learning problem: we observe only the features but no outcome. Our task is rather to describe how the data are organized or clustered.

**DNA Expression Microarrays**

DNA: Deoxyribonucleic acid: the basic material that makes up human chromosomes.

DNA microarrays measure the expression of a gene in a cell by measuring the amount of mRNA (messenger ribonucleic acid) present for that gene.

Microarray is a slide or chip systematically dotted with DNA from thousands of genes that can serve as probes for selecting which genes are activated or expressed.
Procedure:
(i) Nucleotide sequences for a few thousands of genes are printed on a glass slide.
(ii) Target sample (cancer cell) and a reference sample (normal cell) are labeled with red and green dyes and each are hybridized with the DNA on the slide.
(iii) Through fluoroscopy, the log(red/green) intensities of RNA hybridizing at each site is measured.
(i) result is a few thousands numbers, typically ranging from say -6 to 6, measuring the expression level of each gene in the target relative to the reference sample.
(ii) Positive values indicate higher expression in the target versus the reference and vice versa for negative values.

Dataset: Expression values from a series of DNA microarray experiments, with each column representing an experiment. So several thousand rows representing individual genes and tens of columns representing samples. Example: 7000 rows and 60 columns.

Questions:
(i) Which samples are most similar to each other, in terms of their expression profile across genes?
(ii) Which genes are most similar to each other, in terms of their expression profiles across samples?
(iii) Do certain genes show very high (or low) expression for certain cancer samples?

(i) Regression Problem: two categorical predictor variables” genes and samples, response variable being the level of expression.
(ii) Unsupervised learning problem: samples as points in 6000-dimensional space: which we want to cluster together in some way.

Very high dimension with small sample: need prior knowledge.