**DS-GA 3001.001**
**Special Topics in Data Science: Probabilistic Time Series Analysis**
**Project proposal instructions**

Proposal due on 10/16 (by 6pm)

The project should be done in groups of 2 or 3 registered students (exceptions only if you have explicitly cleared it with CS first).

The **project** can take any of the following forms:

- High quality **software implementation** of algorithms related to the class that are currently not publicly available, to be released for public usage.
- Application of a machine learning model to a previously unconsidered **dataset** and to a specific scientific question. For this: find some interesting (for you) data. Keep in mind the issue of stationarity. Make sure you have enough data to be able to fit a reasonable model (the more parameters, the more data you'll need).
- **Extension** to existing method**,** or **theoretical analysis** of existing algorithm. This would likely have a scope outside the course, e.g. it could be the starting point for a longer research project.
- Multiple **models comparison** on existing dataset
- In-depth **review** of recent papers related to time series topics either covered in the lecture (e.g. generalizations of LDS, HMMs) or not (e.g. probabilistic RNNs).

If you are unsure if your idea fits into these types, talk to CS.

**Proposal:**

Write a proposal that details the question you are planning to address, which dataset are you planning to use (if applicable), the family of algorithms used for the analysis, and how you plan to evaluate these methods. The goal is to check that you have a plan, so add whatever details you have already worked out that may be relevant (within the space limits: **no more than 1 page**).

Important: explain also how the tasks will be allocated across team members (who does what).

Proposals should be uploaded on nyuclasses in **pdf**. We plan to provide some constructive feedback for each proposal by end of following week.

**Some past projects:**

- **Model comparison:** HMMs vs RNNs for anomaly detection**.**
- **Model comparison:** Automatic speech recognition **-** HMMs Acoustic models, Deep nets
- **Implementation:** RNNs for multivariate time series with missing values (GLU-R)
- **Object tracking:** implementation + model comparison
- **Dataset:** Modeling protein sequences + Model comparison HMMs vs RNNs
- **Dataset:** Forecasting Wikipedia web traffic
- **Dataset:** Prediction of Beijing pollution levels
- **New algorithm:** Gaussian Process Factor Analysis with Spectral Mixture kernels
- **Review:** latent state space models for modeling neural data
- **Voice synthesis:** speak like Pippa the pig

**Some new ideas that I'd like to see explored:**

- **Finite sample** properties of linear models *(e.g. Kley et al)*
- **Probabilistic RNNs —** variational auto-encoders for time series
- **Volatility** models**,** e.g. ARCH/GARCH
- **Nonlinear** generalizations of a model considered in class
- Find a **new dataset** with dependencies and try to model it
- **Nonstationarity:** datasets, or models, or theory

I have included some example papers in the current literature folder. If something sounds interesting, but you don't have a clear idea how to make it into a project — reach out to CS and we can figure it our together.