# Generating Time Series Data Using Probability Specifications

Alan Ismaiel (ENG 2022), Jason Shu (COL 2022)
Ivan Ruchkin, Oleg Sokolsky, Insup Lee, Kaustubh Sridhar
Penn Engineering, Department of Computer and Information Science (CIS)
Penn Undergraduate Research Mentoring Program (PURM)

**Our data generation tool takes user-inputted probability and independence statements and fully determines discrete distributions through a distinct parameterization method. Time series data is then generated from the resulting probability distributions.**

## Motivation

In an increasingly data-driven world, it's more important than ever to have sources of usable data.

- Generating accurate data (if possible) is **more cost-effective** than collecting real-world data
- **Time series data**—data points indexed in a clear time-based order—is particularly useful, but many existing tools just use deterministic ways to generate it
- By inputting independence relationships between variables in a time series, the **complexity** of its distribution is **exponentially reduced**

## Existing Approaches and Tools

- **Probabilistic graphical models** (e.g. Bayesian nets, Markov nets, factor graphs) can visually represent a certain distributions but cannot represent complex dependence and independence relationships
- Tools using an **autoregressive-moving-average model (ARMA)** can represent randomness and noise but fail to intuitively model dependencies between variables
- Tools like **Faker** and **DataGenerator** can randomly generate data, but have no built-in notion of time series
- Tools like **T Simulus**, **Pandas** (Python library), and **Pyro** can represent some specifications and time series, but cannot intuitively represent independence relations

### References

Castanon, D. and Karl, W.C., 2004, *Stochastic Processes*, Lecture Notes, Class Notes, SC505, Boston University, http://www.mit.edu/people/hmsallum/GradSchool/sc505notes.pdf

DeepDive Project at Stanford University, "Factor Graphs", n.d., http://deepdive.stanford.edu/assets/factor_graph.pdf

Gillespie, Daniel T., "Exact stochastic simulation of coupled chemical reactions", 1977, https://pubs.acs.org/doi/10.1021/j100540a008

Henderson, T.C. et al., "Probabilistic Logic for Intelligent Systems", 2018, http://www.cs.utah.edu/~tch/publications/pub302.pdf

Hu, Z. and Hong, L.J., "Robust Simulation Of Stochastic Systems With Input Uncertainties Modeled By Statistical Divergences", 2015, https://www.informs-sim.org/wsc15papers/055.pdf

Kang, Yanfei et al., "GRATIS: Generating Time Series with Diverse and Controllable Characters", 2019, https://arxiv.org/ftp/arxiv/papers/1903/1903.02787.pdf

Pandey, A. et al., May 2017, "Towards a Formal Framework for Hybrid Planning in Self-Adaptation", Presentation, Carnegie Mellon University, https://www.slideshare.net/ivanruchkin/towards-a-formal-framework-for-hybrid-planning-in-selfadaptation

## Methods

First, we define a notion of **relevant time steps** in a set $B_t$. Put simply, when conditioned on $B_t$, variables at a time t are independent of all other variables outside $B_t$. The set $B_t$ and all variables at time t define a probability system $S_t$.

We have divided the data generation problem into three separate cases, **Static**, **Time-invariant**, and **Time-variant**.

| | STATIC | TIME-INVARIANT | TIME-VARIANT |
|---|---|---|---|
| **Models distribution(s) over set of random variables at some time t** | | | |
| **Uses probability specifications (equations) to define a distribution** | | | |
| **Has relevant time steps ($B_t$ nonempty), need base case for first steps** | | | |
| **Distribution of $S_t$ is the same for all t (Stationary Assumption)** | *N/A* | | |
| **Past probabilities in $B_t$ can define those at t, e.g. $P(X_t) = 0.5*P(X_{t-1})$** | *N/A* | | |
| **Example** | *Drawing cards from a deck with replacement* | *Deciding what to wear based only on what was worn the previous day* | *Mechanical parts deteriorate over time; likelihood of failure at t is function of likelihood at t-1* |
| **Visual** | *(3 samples, time not relevant)* | 1 2 3 4 *(2 relevant time steps, invariant)* | 1 2 3 4 *(2 relevant time steps, variant)* |

*Detailed examples of all three case types can be found by following the QR code here.*

*shorturl.at/bnHQ6*

To generate these case types, our data generation tool does the following:

**PARAMETERIZING**
Converting user specifications into Mathematica syntax; translating probability symbols into generic variables

→

**SOLVING**
Determining the probability system based on valid user input; if invalid, returning "contradictory" or "underdetermined"

→

**GENERATING**
Producing data if the solved system is well-defined; using conditional probabilities to account for results from relevant past time steps

## Example

Consider a **boolean variable T**. Generate a **time-invariant** time series with ten data points (time steps) for T, where for each t > 2, steps t - 1 and t - 2 are relevant to generate data at time t.

There are **8 elementary probabilities**: $P(T_t \,\&\&\, T_{t-1} \,\&\&\, T_{t-2})$, $P(\text{not } T_t \,\&\&\, T_{t-1} \,\&\&\, T_{t-2})$, … $P(\text{not } T_t \,\&\&\, \text{not } T_{t-1} \,\&\&\, \text{not } T_{t-2})$. All events are independent, and probabilities add to 1.

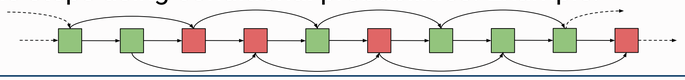
**Stationary assumption** (distributions don't change):

- $P(T_t \,\&\&\, T_{t-1}) = P(T_{t-1} \,\&\&\, T_{t-2})$
- $P(T_t \,\&\&\, \text{not } T_{t-1}) = P(T_{t-1} \,\&\&\, \text{not } T_{t-2})$
- $P(\text{not } T_t \,\&\&\, T_{t-1}) = P(\text{not } T_{t-1} \,\&\&\, T_{t-2})$

We have 4 equations, **need 4 more** to determine 8 vars:

- $P(T_t \mid T_{t-1} \,\&\&\, T_{t-2}) = .2$
- $P(T_t \mid \text{not } T_{t-1} \,\&\&\, \text{not } T_{t-2}) = .8$
- $P(T_t \mid T_{t-1} \,\&\&\, \text{not } T_{t-2}) = .5$
- $P(T_t \mid \text{not } T_{t-1} \,\&\&\, \text{not } T_{t-2}) = .5$

**Solve** for the elementary probabilities. **Generate** data for 10 time steps using conditional probabilities. Example:



## Results & Summary

Utilizing the **Mathematica** software, we have created a tool that converts a text file with specifications into a probability system that generates data.

- For now, only the Static Case is recognized, future steps for the project require implementing the time invariant and time variant cases developed here

Overall, by implementing the methods described here, our tool has the potential to **generate realistic time series data** for many unique and practical applications.