CAMBRIDGE SYSTEMATICS

Think Forward

Gaussian Process Regression for Risk Analysis of Travel Demand Forecasts

presented to

7th International Conference on Innovations in Travel Modeling

Atlanta, Georgia

June 2018

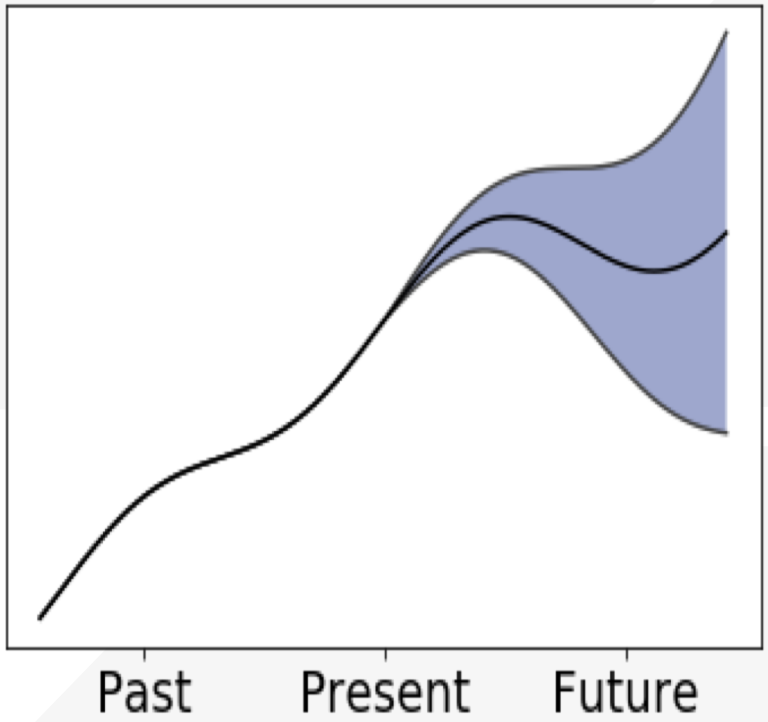
presented by

Cambridge Systematics, Inc.
Jeffrey Newman

with
Rachel Copperman, Jason Lemp, David Kurth
Boris Lipkin, *California High-Speed Rail Authority*
Matt Henley, *John Helsel, WSP*

Why Risk Analysis?

- The base forecast model generates a single **point estimate** forecast for future conditions.
- This estimate is **not precise**, because of incomplete or inaccurate representations of present or future inputs or assumptions embedded in the models.
- Robust planning and decision making processes instead will consider a **range of model forecasts**.



Risk Analysis Approach

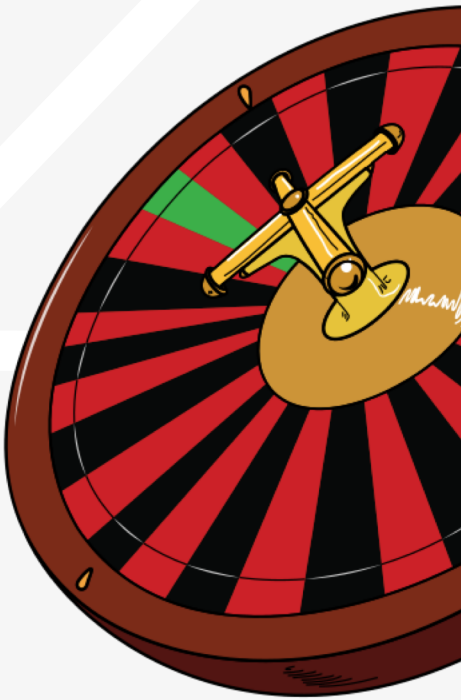
these are important but let's assume these are already done

Today's Topic

CAMBRIDGE SYSTEMATICS

Monte Carlo Simulation

- To develop a robust distribution of outcomes, we want to run our model a lot: many thousands of times
- But our simulation model is complex, it takes a long time to complete a single experiment



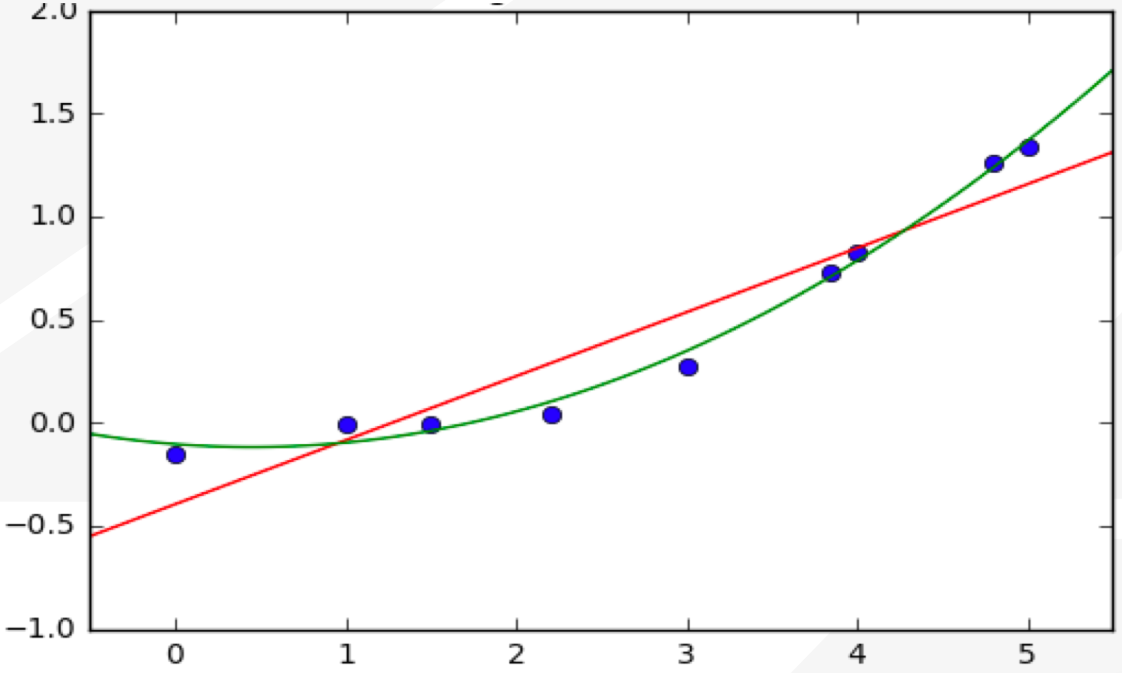
Monte Carlo Simulation

- To develop a robust distribution of outcomes, we want to run our model a lot: many thousands of times
- But our simulation model is complex, it takes a long time to complete a single experiment
- Solution: replace the model with a simpler one, which takes only fractions of a second to run

Enter the Meta-Model

- We replace the expensive simulation model with a fast regression meta-model.
- Common practice in transportation planning is to use a linear regression model:
  - Easy to implement
  - Exceptionally fast
  - Generally appears to have good fit
    - But is it really good enough?

An Illustration in One Dimension



Simple Linear Regression $R^2 = 0.90$

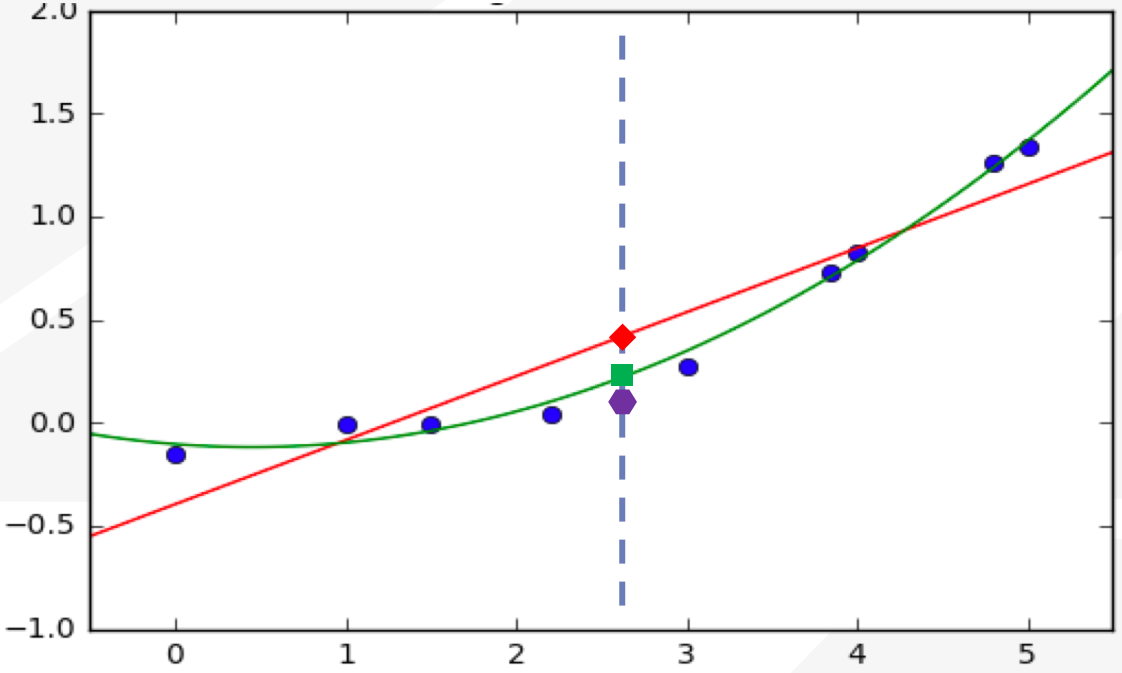
Polynomial Regression $R^2 = 0.99$

R^2 this high is typically regarded as a good fit…

CAMBRIDGE SYSTEMATICS

An Illustration in One Dimension

Suppose we want to predict at X=2.6



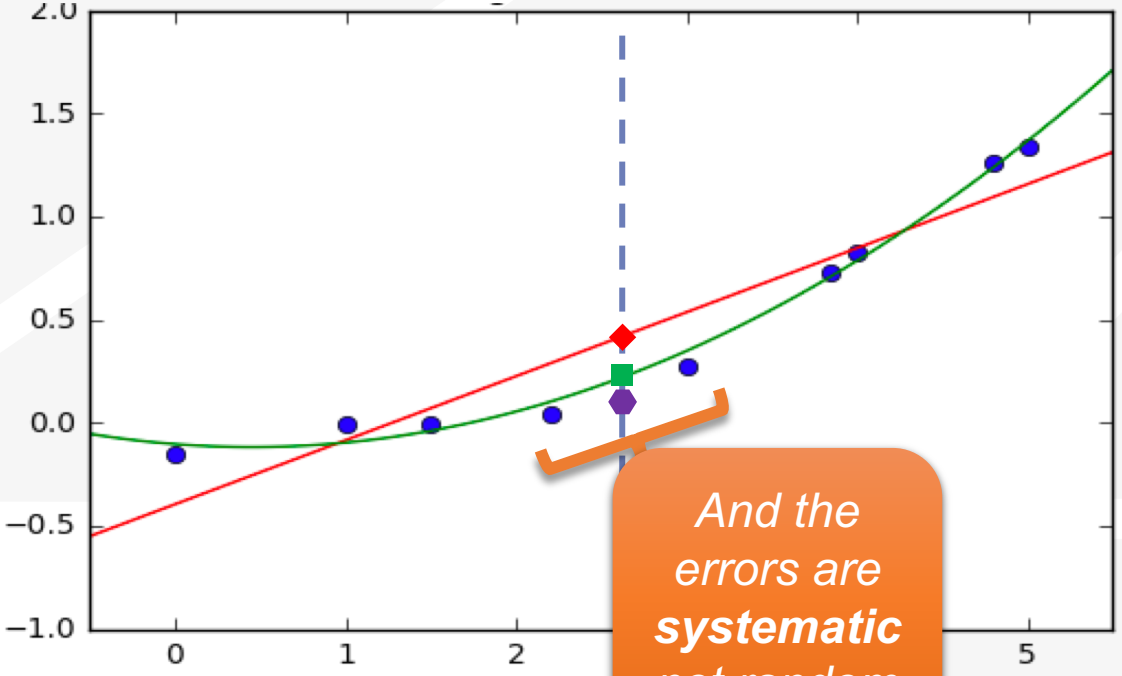
- Simple Linear Regression Y=0.43
- Polynomial Regression Y=0.26
- Probably the Real Value Y=0.21

… but there is still some remaining error

CAMBRIDGE SYSTEMATICS

An Illustration in One Dimension

Suppose we want to predict at X=2.6



*And the errors are **systematic** not random*

- Simple Linear Regression Y=0.43
- Polynomial Regression Y=0.26
- Probably the Real Value Y=0.21

CAMBRIDGE SYSTEMATICS

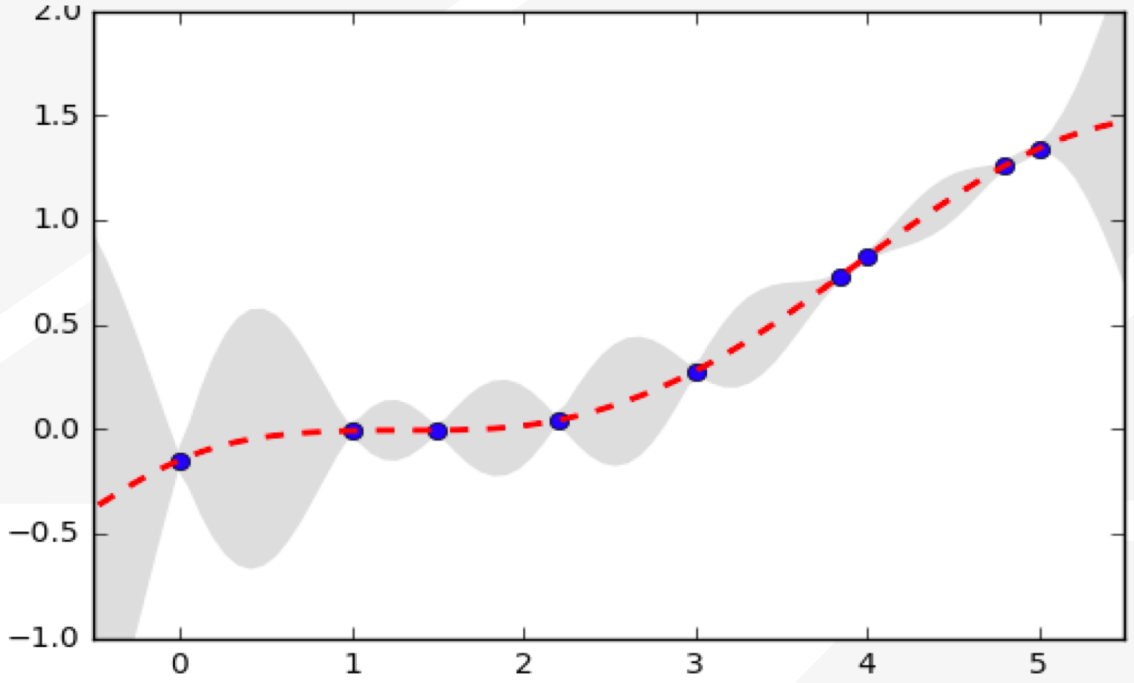
Gaussian Process Regression

- Gaussian Process Regression (GPR) is a non-parametric “machine learning” tool for regression analysis
- GPR does not impose a restriction on the functional form of the output
- Instead, just assume the output is **auto-correlated**: if the inputs are similar, then the output should also be similar
  - This auto-correlation violates the independent errors assumption in OLS linear regression

Two Important Features

- The BPM-V3 model for the California HSR has two features that make it work well with GPR:
  - **Deterministic**: Re-run the model with the same inputs, get the same output
  - **Smooth**: Re-run the model with the infinitesimally different inputs, get the only infinitesimally different output
- Conveniently, many travel demand models share these features
  - Although it makes things simpler, neither is strictly necessary for the use of GPR

GPR Illustration in One Dimension



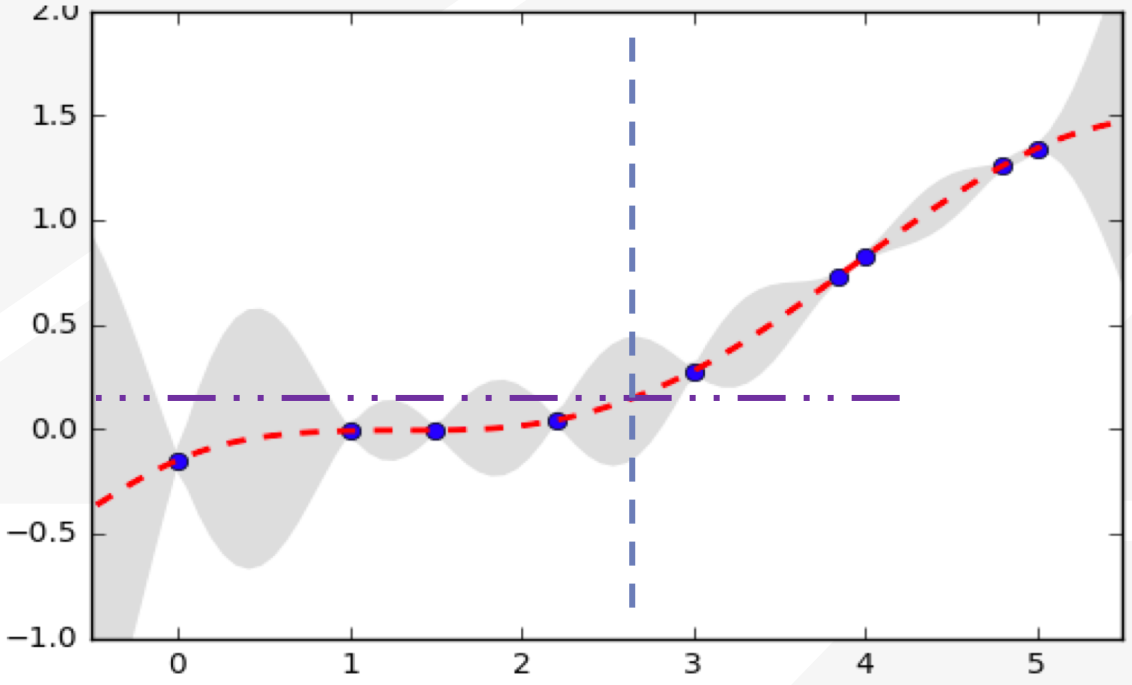
Gaussian Process Regression $R^2 = 1.00$

2 Standard Deviations

We will come back to this

CAMBRIDGE SYSTEMATICS

GPR Illustration in One Dimension

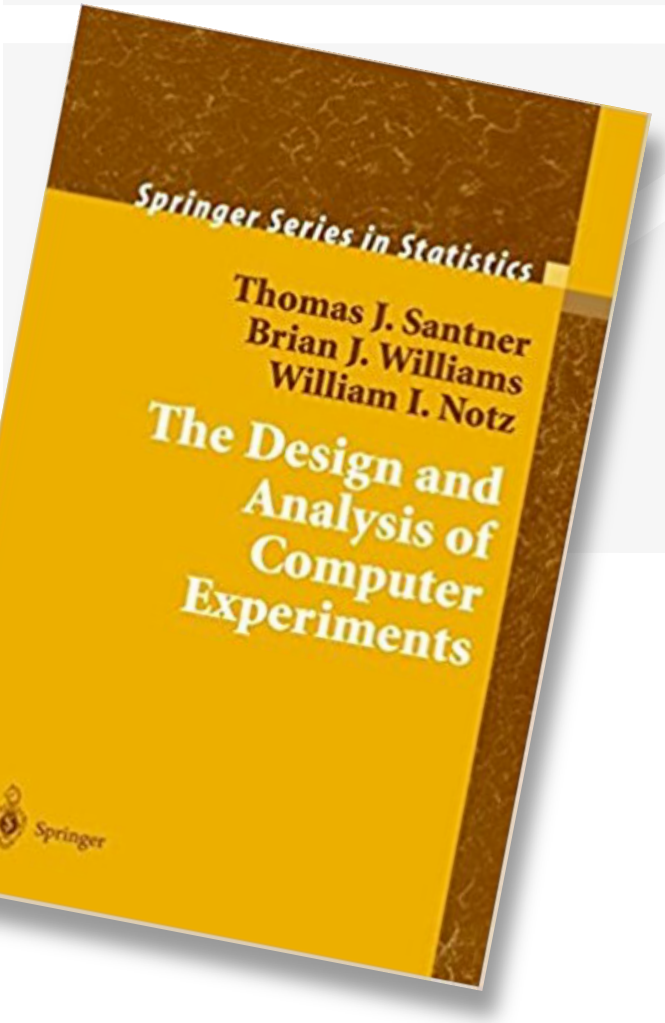


Bingo!

Gaussian Process Regression Y=0.21

CAMBRIDGE SYSTEMATICS

GPR Represents Best Practices



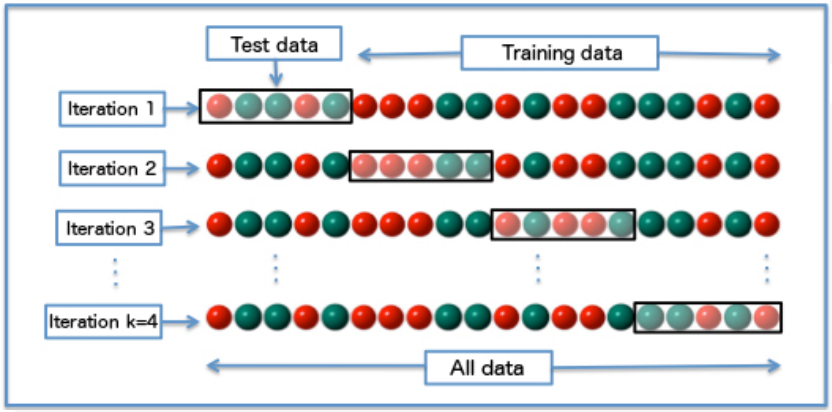
- Although not widely used for transportation planning meta-model applications, it is widely used for computer simulation meta-models in other fields
- Gaussian Process Regression is the textbook approach for modern meta-models of computer experiments
- And this is the textbook:
  Santner, T. J., Williams, B. J., & Notz, W. I. (2013). *"The design and analysis of computer experiments."* Springer Science & Business Media.

About that R^2 of 1.0

- GPR meta-models cannot be evaluated based on traditional “goodness of fit” measures derived from the estimation data, as for deterministic models they by design always fit all the estimation data perfectly
- Instead it is necessary to measure fit on a validation data set that is not used for model estimation
- Since additional data is expensive to collect, it is preferred to use k-fold cross-validation

K-fold Cross-Validation

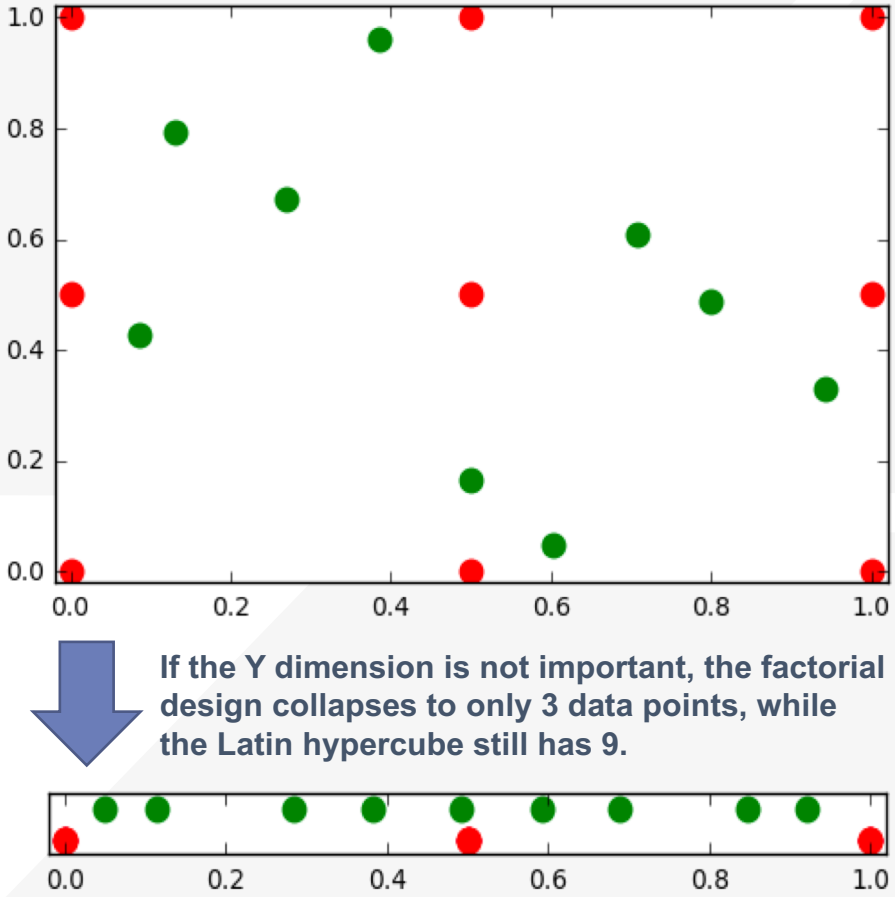
- Data is randomly split into K groups of roughly even size
- The model is fit using only K-1 groups, then evaluated based on the fit of the remaining holdout group
- Process is repeated for each of the K groups and averaged across them to create fit statistics



Source: https://commons.wikimedia.org/wiki/File:K-fold_cross_validation_EN.jpg

Design of Experiments

- Previous transportation planning applications have focused on a **factorial** or fractional factorial design of experiments (example in red)
- GPR instead is better supported by a **Latin Hypercube** design with irregular distances between experiments (example in green)
- If some dimensions are not important, the factorial design partially collapses but the hypercube design still recovers maximum information.



CAMBRIDGE SYSTEMATICS

Application: California High Speed Rail

- GPR was employed to conduct a risk analysis for the ridership and revenue forecasts for the California High Speed Rail Authority 2018 Business Plan
- The Latin Hypercube design was adopted to allow for 13 to 15 risk factors (varies by forecast year) — prior Business plans relied on a fractional factorial designs that limited the analysis to only 10 risk factors.

Risk Factors Included

- High speed rail constants
- Trip frequency constants
- Quality of connecting bus service
- Coefficient on access/egress time by distance
- Coefficient on extremely long access/egress
- Impact of Automated Vehicles
- Automobile operating cost
- Air and High speed rail fares
- High speed rail frequency of service
- High speed rail reliability
- Number and distribution of households throughout the state
- Level of visitor travel
- Level of extra induced ridership

Note: Not all risk factors are relevant for every forecast year

Still a Heavy Computation Load

- A single run of the full BPM-V3 simulation requires about 12 hours of CPU time
- One pass of this risk analysis involved conducting 150 runs for each of 3 forecast years = 5,400 CPU-hours
- We built an *ad hoc* cluster using Python and Dask with on average about 200 CPU cores available to complete the experimental runs in just a few days

Results: Revenue

	2029 – VtoV	2033 – Phase 1	2040 – Phase 1
GPR Cross Validation Score (Improvement over Linear Regression)	0.747	0.987	0.983
RMSE of Cross Validation Predictions (millions of 2017$)	$14.4	$7.1	$9.0
Long Distance HSR Revenue – 2018 Business Plan Base Runs (millions of 2017$)	$823	$2,085	$2,329
RMSE as a percent of Base Run Long Distance HSR Revenue	1.7%	0.3%	0.4%

Results: Ridership

	2029 – VtoV	2033 – Phase 1	2040 – Phase 1
GPR Cross Validation Score (Improvement over Linear Regression)	0.834	0.986	0.983
RMSE of Cross Validation Predictions (millions of annual riders)	0.25	0.16	0.19
Long Distance HSR Revenue – 2018 Business Plan Base Runs (millions of 2017$)	14.4	35.6	39.4
RMSE as a percent of Base Run Long Distance HSR Revenue	1.7%	0.4%	0.5%

http://www.hsr.ca.gov/docs/about/business_plans/2018_Business_Plan_Ridership_Revenue_Risk_Model.pdf

http://www.hsr.ca.gov/docs/about/business_plans/2018_CA_High_Speed_Rail_Business_Plan_Ridership_and_Revenue_Risk_Analysis.pdf

jnewman@camsys.com