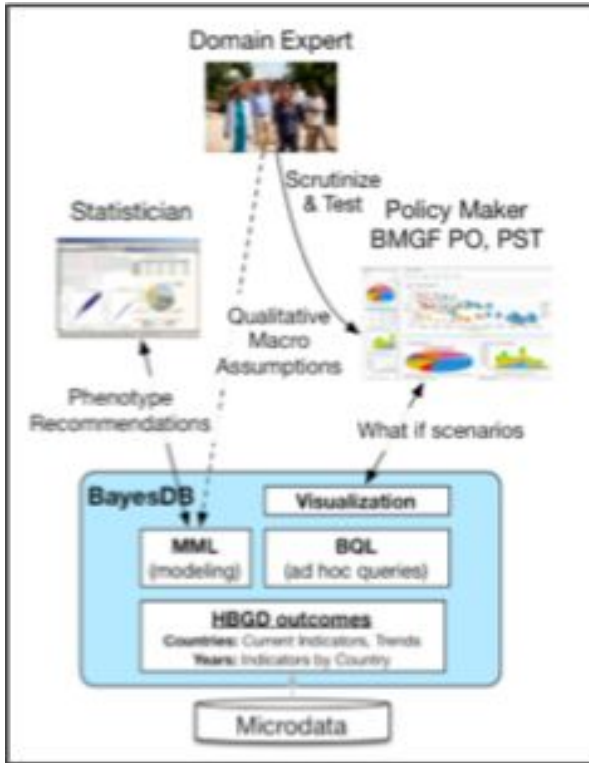


Probabilistic Computing Lab

Belhal KARIMI



The need for augmented intelligence



Policy advocate

"What are the comparable countries to Kenya in terms of everything we know about the malnutrition rate of infants?"

Domain expert

"Recent work in development economics suggests sanitation standards influence growth stunting in India but not in Africa."

Field researcher

"Here is new data on ~10,000 children in Bangladesh. Please update all relevant models and inform stakeholders."

Statistician

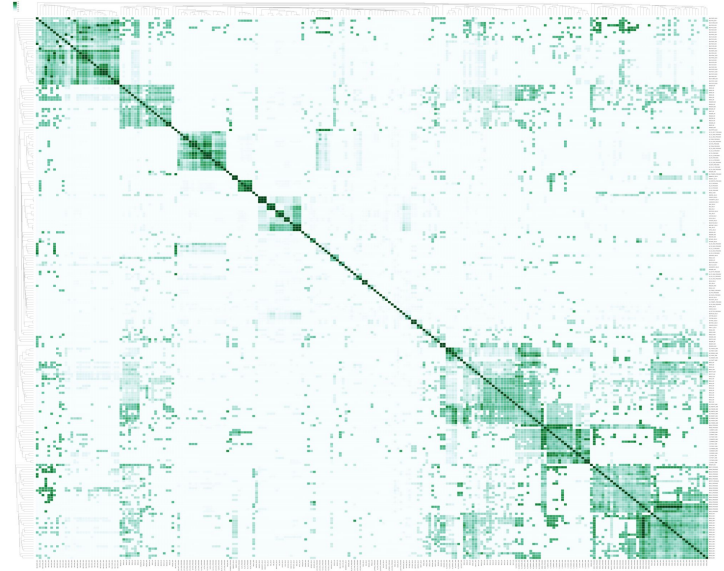
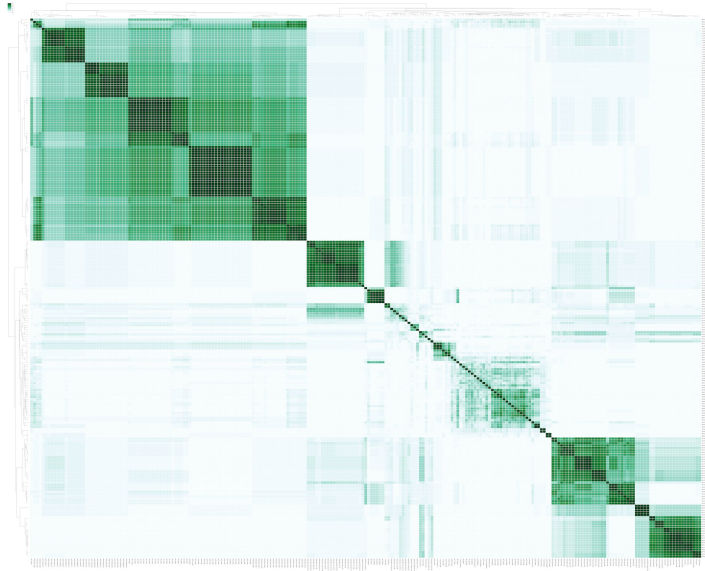
"Despite what the economists think, the p-value for this hypothesis test indicates that my mixed-effects model's finding of two different country clusters with respect to longitudinal variation in sanitation standards is not actually significant."

GUSTO (Growing up in Singapore towards healthy outcomes) analysis

- Raw Dataset: 24988 rows and 315 columns
- After processing: 1171 rows (1171 infants) and 233 columns (variables about the infants and their parents)
- Creation of a crosscat generator:
 - Analysis of 32 models for 100 iterations
- The following slides show:
 - Overall dependence probabilities between pairwise variables of the dataset
 - Zooms on subset of variables of interest

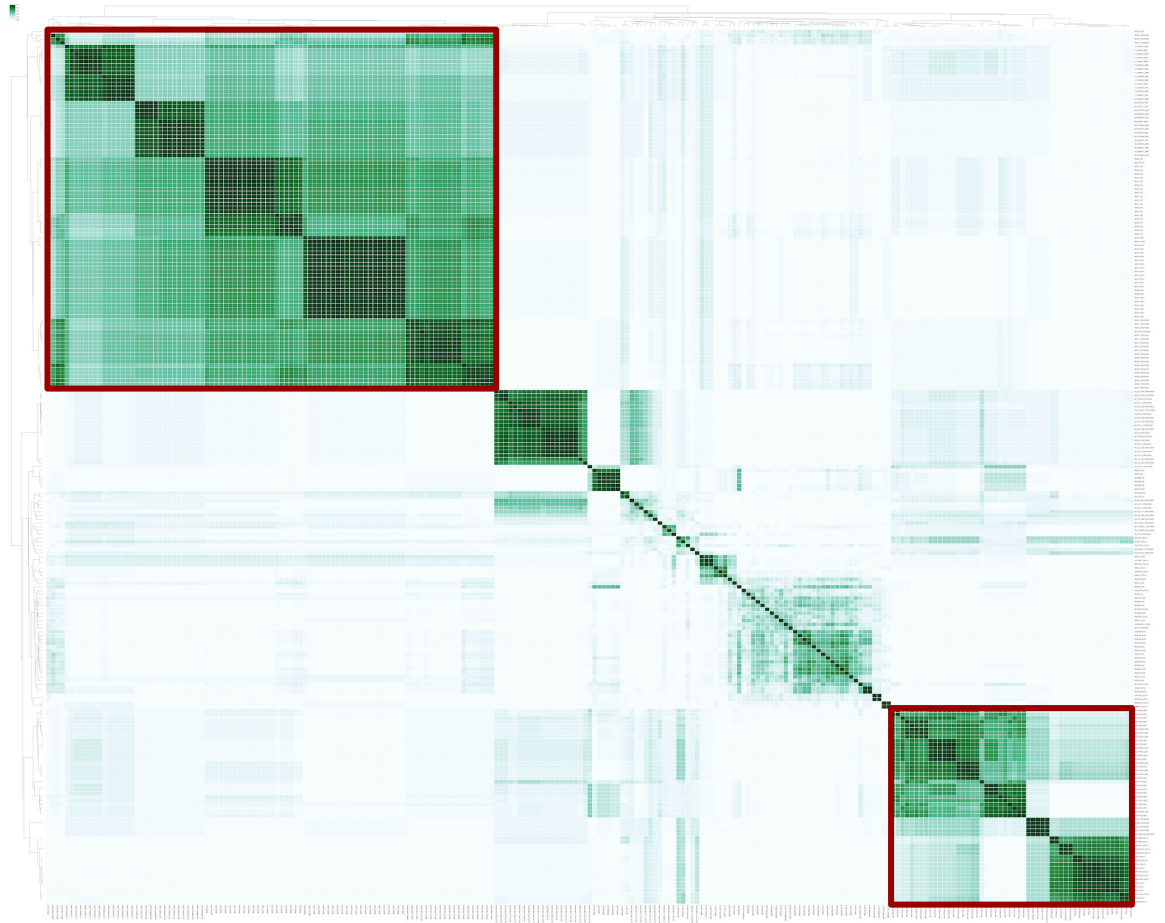
Variable label	Description
DELIV	At Delivery
PREGW26	At Pregnancy week 26
M3	At Month 3
M26	At Month 26
M48	At Month 48
M_	Mother
F_	Father

Dependence probability vs Significant Correlation



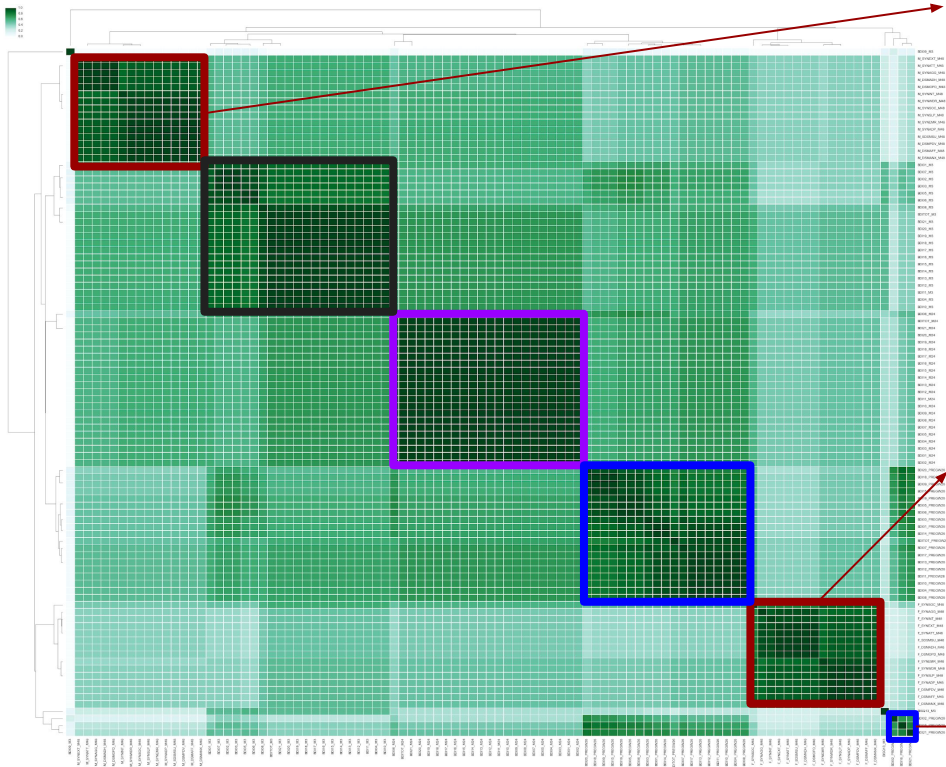
Overall dependency heatmap

- We will focus on the two largest clusters of pairwise dependent variables



Top left

Variable name	Description
BDI01	Beck Depression Inventory-Sadness
BDI02	Beck Depression Inventory-Pessimism
BDI03	Beck Depression Inventory-Past Failure
BDI04	Beck Dep Inv-Loss of Pleasure
BDI05	Beck Dep Inv-Guilty Feelings
BDI06	Beck Dep Inv-Punishment Feelings
BDI07	Beck Depression Inventory-Self-Dislike
BDI08	Beck Dep Inv-Self-Criticalness
BDI09	Beck Dep Inv-Suicidal Thoughts/Wishes
BDI10	Beck Depression Inventory-Crying
BDI11	Beck Depression Inventory-Agitation
BDI12	Beck Dep Inv-Loss of Interest
BDI13	Beck Depression Inventory-Indecisiveness
BDI14	Beck Depression Inventory-Worthlessness
BDI15	Beck Depression Inventory-Loss of Energy
BDI16	Beck Dep Inv-ChangesSleep Pattern
BDI17	Beck Depression Inventory-Irritability
BDI18	Beck Dep Inv-Changes Appetite
BDI19	Beck Dep Inv-Concentration Difficulty
BDI20	Beck Dep Inv-Tiredness or Fatigue
BDI21	Beck Dep Inv-Loss of Interest in Sex
BDITOT	Beck Depression Inventory-Total Score



- At Month 3
- At Month 24
- At Month 48
- At Pregnancy week 26

Variable name	Description
F_SYNEXT_M48	Syndr Externlz Prob Raw Tot Eval=FATH
F_SYNATT_M48	Syndr Attention Prob Raw Tot Eval=FATH
F_SYNAGG_M48	Syndr Aggr Behav Raw Tot Eval=FATH
F_DSMDH_M48	DSM AtnDefic/Hyper Prob RawTot Ev=FATH
F_DSMDOP_M48	DSM Oppos Def Prob Raw Tot Eval=FATH
F_SYNINT_M48	Syndr Intrnlz Prob Raw Tot Eval=FATH
F_SYNWDR_M48	Syndr Withdrawn Raw Tot Eval=FATH
F_SYNSOC_M48	Syndr Somatic Compl Raw Tot Eval=FATH
F_SYNSLP_M48	Syndr Sleep Prob Raw Tot Eval=FATH
F_SYNEMR_M48	Syndr Emotion Reac Raw Tot Eval=FATH
F_SYNADP_M48	Syndr Anxious/Depr Raw Tot Eval=FATH
F_SDSMSU_M48	Synd and DSM Total Score Sum Eval=FATH
F_DSMPDV_M48	DSM Pervasiv Dev Prob Raw Tot Eval=FATH
F_DSMAFF_M48	DSM Affective Prob Raw Tot Eval=FATH
F_DSMANX_M48	DSM Anxiety Prob Raw Tot Eval=FATH

Variable name	Description
M_SYNEXT_M48	Syndr Externlz Prob Raw Tot Eval=MOTH
M_SYNATT_M48	Syndr Attention Prob Raw Tot Eval=MOTH
M_SYNAGG_M48	Syndr Aggr Behav Raw Tot Eval=MOTH
M_DSMDH_M48	DSM AtnDefic/Hyper Prob RawTot Ev=MOTH
M_DSMDOP_M48	DSM Oppos Defian Prob Raw Tot Eval=MOTH
M_SYNINT_M48	Syndr Intrnlz Prob Raw Tot Eval=MOTH
M_SYNWDR_M48	Syndr Withdrawn Raw Tot Eval=MOTH
M_SYNSOC_M48	Syndr Somatic Compl Raw Tot Eval=MOTH
M_SYNSLP_M48	Syndr Sleep Prob Raw Tot Eval=MOTH
M_SYNEMR_M48	Syndr Emotion Reac Raw Tot Eval=MOTH
M_SYNADP_M48	Syndr Anxious/Depr Raw Tot Eval=MOTH
M_SDSMSU_M48	Synd and DSM Total Score Sum Eval=MOTH
M_DSMPDV_M48	DSM Pervasiv Dev Prob Raw Tot Eval=MOTH
M_DSMAFF_M48	DSM Affective Prob Raw Tot Eval=MOTH
M_DSMANX_M48	DSM Anxiety Prob Raw Tot Eval=MOTH

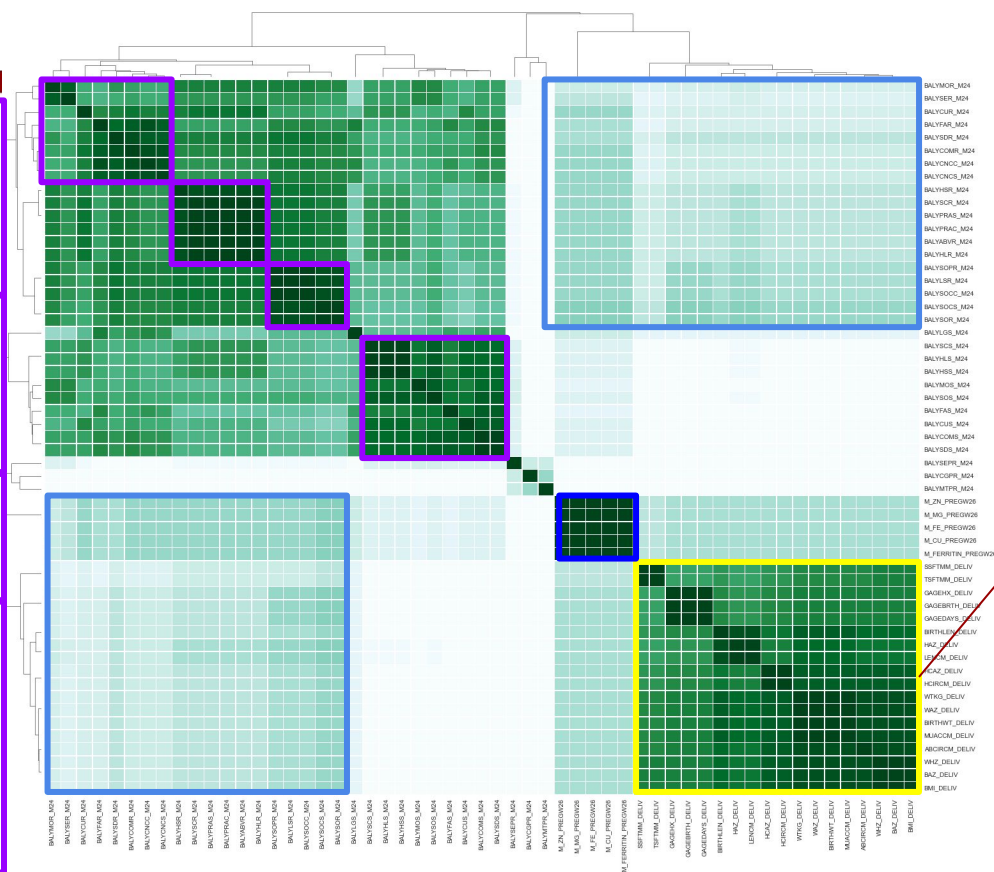
Variable name	Description
BDI02_PREGW26	Beck Depression Inventory-Pessimism
BDI16_PREGW26	Beck Dep Inv-ChangesSleep Pattern
BDI21_PREGW26	Beck Dep Inv-Loss of Interest in Sex

Bottom Right

Weak evidence of dependencies



Variable name	Description
BALYMOR	Bayley Motor - Raw Score
BALYSER	Bayley Social-Emotional - Raw
BALYCUR	Bayley Community Use - Raw Score
BALYFAR	Bayley Functional Pre-Academic - Raw
BALYSOR	Bayley Social - Raw
BALYCOMR	Bayley Communication - Raw Score
BALYCNC	Bayley Conceptual - Composite
BALYCNS	Bayley Conceptual - Scaled
BALYHSR	Bayley Health and Safety - Raw Score
BALYSCR	Bayley Self Care - Raw Score
BALYPRAS	Bayley Practical - Scaled
BALYPRAC	Bayley Practical - Compsite
BALYABPR	Bayley GAP - percentile rank
BALYHLR	Bayley Home Living - Raw Score
BALYSOPR	Bayley Social Composite percentile rank
BALYLSR	Bayley Leisure Behavior - Raw Score
BALYSOCC	Bayley Social Composite - Raw Score
BALYSOCS	Bayley Social Composite - Scaled
BALYSOR	Bayley Social - Raw
BALYLS	Bayley Language - Scaled
BALYSCS	Bayley Self Care - Scaled
BALYHLS	Bayley Home Living - Scaled
BALYHSS	Bayley Health and Safety - Scaled
BALYMOS	Bayley Motor - Scaled Score
BALYSOS	Bayley Social - Scaled
BALYFAS	Bayley Functional Pre-Academic - Scaled
BALYCUS	Bayley Community Use - Scaled
BALYCOMS	Bayley Communication - Scaled Score
BALYSEPR	Bayley Social-Emotion - Percentile rank
BALYCGPR	Bayley Cognitive - Percentile rank
BALYMPR	Bayley Motor - Percentile rank



Variable name	Description
M_ZN	Zinc
M_MG	Magnesium
M_FE	Iron
M_CU	Copper
M_FERRITIN	Ferritin

SSFTMM	Subscapular skinfold thickness (mm)
TSFTMM	Triceps skinfold thickness (mm)
GAGEHX	Gest age at end of pregnancy (days)
GAGEBRTH	Gestational age at birth (days)
GAGEDAYS	Gestational age at examination (days)
BIRTHWT	Birth weight (gm)
HAZ	Length/height for age z-score
LENCM	Recumbent length (cm)
HCAZ	Head circumference for age z-score
HCIRCUM	Head circumference (cm)
WTKG	Weight (kg)
WAZ	Weight for age z-score
BIRTHWT	Birth weight (gm)
MUACCM	Mid upper-arm circumference (cm)
ABCIRCUM	Abdominal circumference (cm)
WAZ	Weight for length/height z-score
BAZ	BMI for age z-score
BMI	BMI (kg/m**2)

- At Month 24
- At Pregnancy week 26
- At Delivery

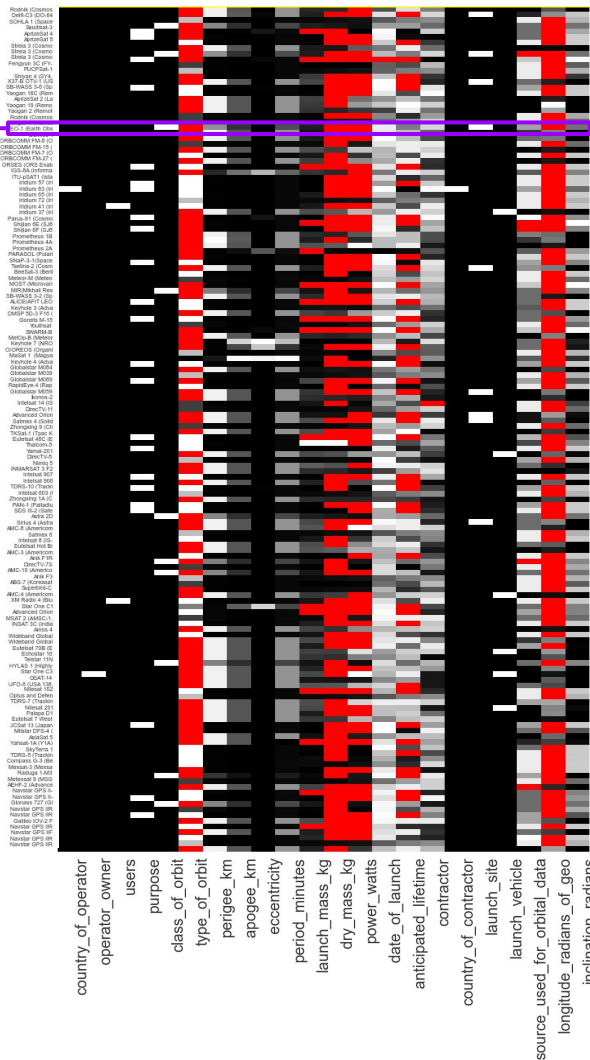
- Setting: UCS Satellites Database
 - 1167 rows (satellites) and 23 columns
 - Illustrations using 150 row subsample
 - Variables include, electrical, geopolitical, kinematic characteristics
 - Engineering note:
 - Schematics come from cleaned 'lovecat' states
 - Predictions come from 'gpmcc' states
- BayesDB capabilities illustrated:
 - Representing high-dimensional, incomplete, heterogeneously typed data
 - Estimating pairwise dependence probabilities from multiple GPMS
 - Generating simulations conditioned on hypotheticals

UCS Satellites Database: Raw Data

Data for Compass M4

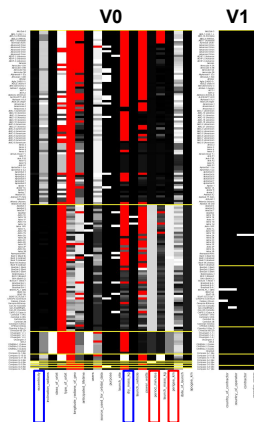
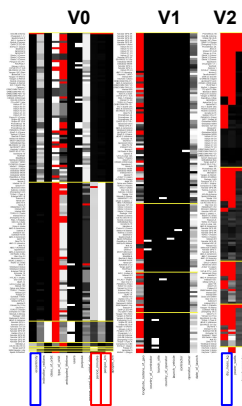
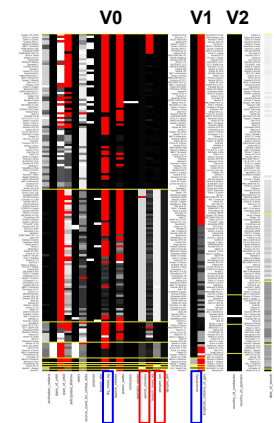
	0
Name	Compass M4 (Beidou 2-13)
Country_of_Operator	China (PR)
Operator_Owner	Chinese Defense Ministry
Users	Military
Purpose	Navigation/Global Positioning
Class_of_Orbit	MEO
Type_of_Orbit	NaN
Perigee_km	21452
Apogee_km	21603
Eccentricity	0.00271
Period_minutes	773.21
Launch_Mass_kg	2200
Dry_Mass_kg	NaN
Power_watts	NaN
Date_of_Launch	41027
Anticipated_Lifetime	8
Contractor	Space Technology Research Institute (part of C...
Country_of_Contractor	China (PR)
Launch_Site	Xichang Satellite Launch Center
Launch_Vehicle	Long March 3B
Source_Used_for_Orbital_Data	ZARYA
longitude_radians_of_geo	NaN
Inclination_radians	0.961676

Red are nans



Variable	Type
Country_of_Operator	categorical
Operator_Owner	categorical
Users	categorical
Purpose	categorical
Class_of_Orbit	categorical
Type_of_Orbit	categorical
Perigee_km	normal
Apogee_km	normal
Eccentricity	normal
Period_minutes	normal
Launch_Mass_kg	normal
Dry_Mass_kg	normal
Power_watts	normal
Date_of_Launch	normal
Anticipated_Lifetime	normal
Contractor	categorical
Country_of_Contractor	categorical
Launch_Site	categorical
Launch_Vehicle	categorical
Source_Used_for_Orbital_Data	categorical
longitude_radians_of_geo	normal
Inclination_radians	normal

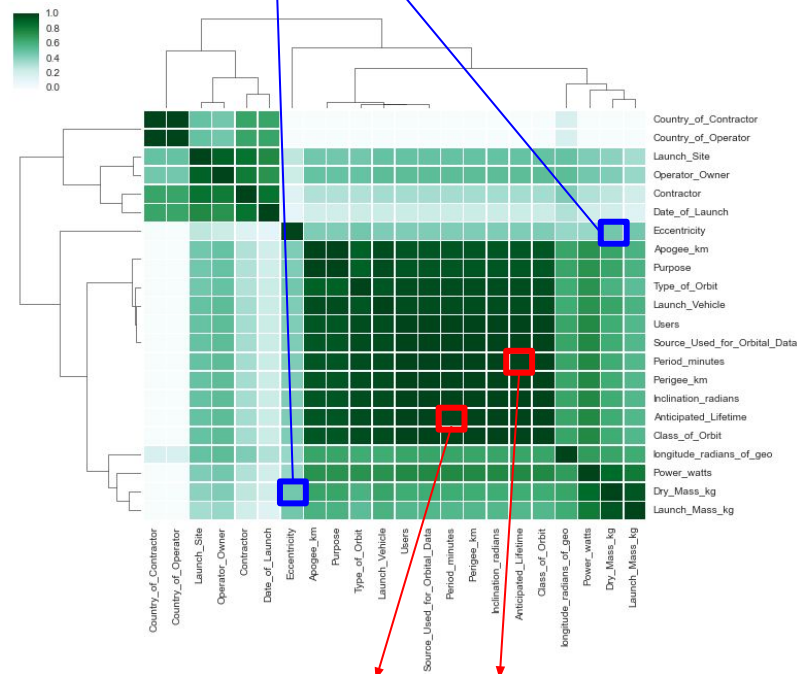
UCS Satellites Database: Relation between Dependence Probability Heatmap and clustering



- Anticipated_lifetime and Period_minutes
- Eccentricity and Dry_mass

ESTIMATE DEPENDENCE PROBABILITY FROM PAIRWISE COLUMNS OF generator

$$P(\text{eccentricity} \not\perp \text{dry_mass}) = 1/3$$

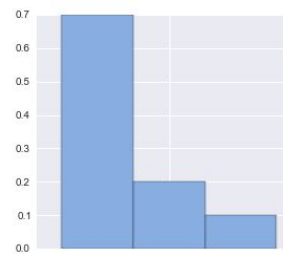


$$P(\text{Anticipated_lifetime} \not\perp \text{Period_minutes}) = 3/3$$

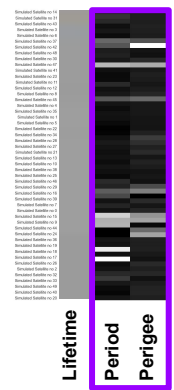
UCS Satellites Database: Generating simulations conditioned on hypotheticals

SIMULATE **Period_minutes**, **Perigee_km** FROM generator GIVEN Anticipated_Lifetime=3

$p(\text{cluster}|\text{lifetime} = 3)$



Cluster 0 Cluster 1 Others



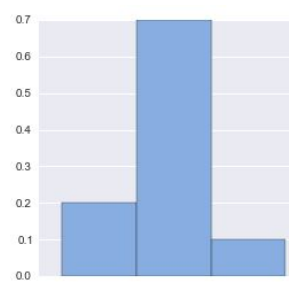
$\sim p(\text{period, perigee}|\text{lifetime}=3)$

Cluster 0

Cluster 1

SIMULATE **Period_minutes**, **Perigee_km** FROM generator GIVEN Anticipated_Lifetime=15

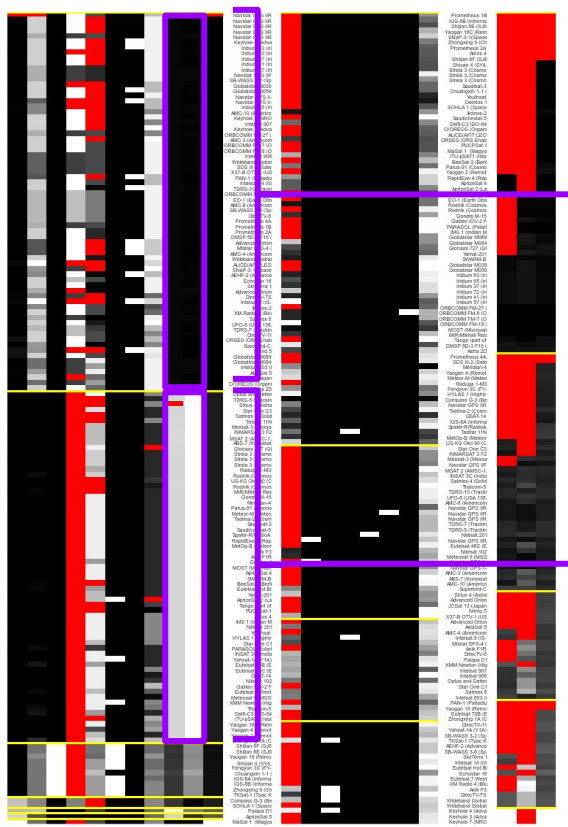
$p(\text{cluster}|\text{lifetime} = 15)$



Cluster 0 Cluster 1 Others

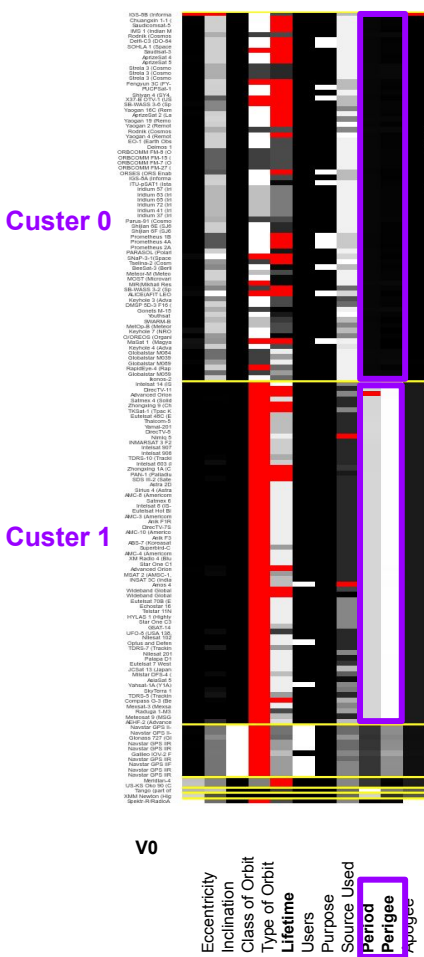


$\sim p(\text{period, perigee}|\text{lifetime}=15)$

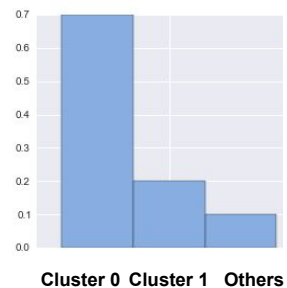


V0
 Eccentricity
 Inclination
 Class of Orbit
 Type of Orbit
 Lifetime
 Users
 Purpose
 Source Used
 Period
 Perigee
 Apogee
 V1
 Longitude
 Country of contractor
 Launch site
 Country of operator
 Launch vehicle
 Contractor
 Operator
 Date of launch
 V2
 Dry mass
 Power watts
 Launch mass

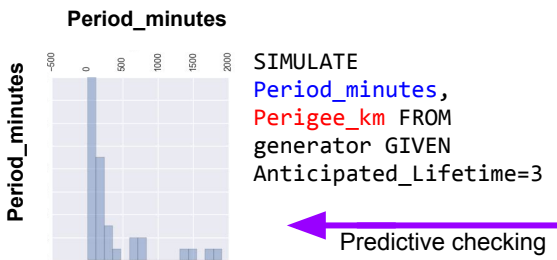
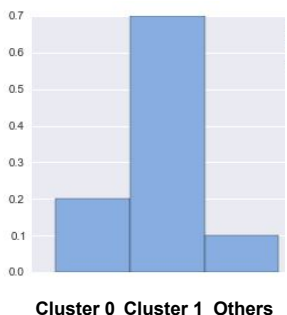
UCS Satellites Database: Posterior distribution vs. CC clustering



$p(\text{cluster}|\text{lifetime} = 3)$

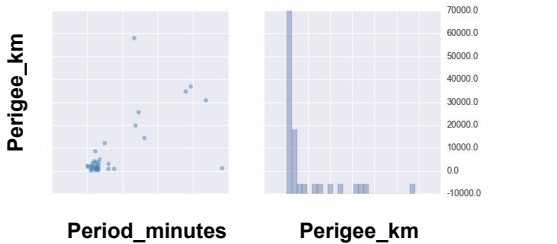


$p(\text{cluster}|\text{lifetime} = 15)$



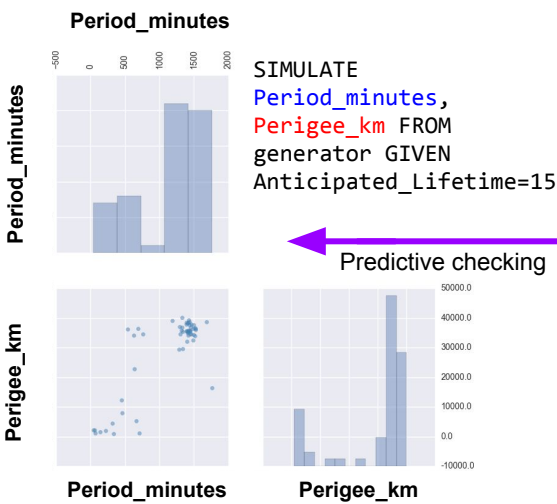
SIMULATE
Period_minutes,
Perigee_km FROM
generator GIVEN
Anticipated_Lifetime=3

Predictive checking



SIMULATE
Period_minutes,
Perigee_km FROM
generator GIVEN
Anticipated_Lifetime=15

Predictive checking



SIMULATE
Period_minutes,
Perigee_km FROM
generator GIVEN
Anticipated_Lifetime=15

Real data

Period_minutes

Period_minutes

Perigee_km

Period_minutes

Perigee_km

SELECT Period_minutes,
Perigee_km FROM table
WHERE
Anticipated_Lifetime=3

Period_minutes

Period_minutes

Perigee_km

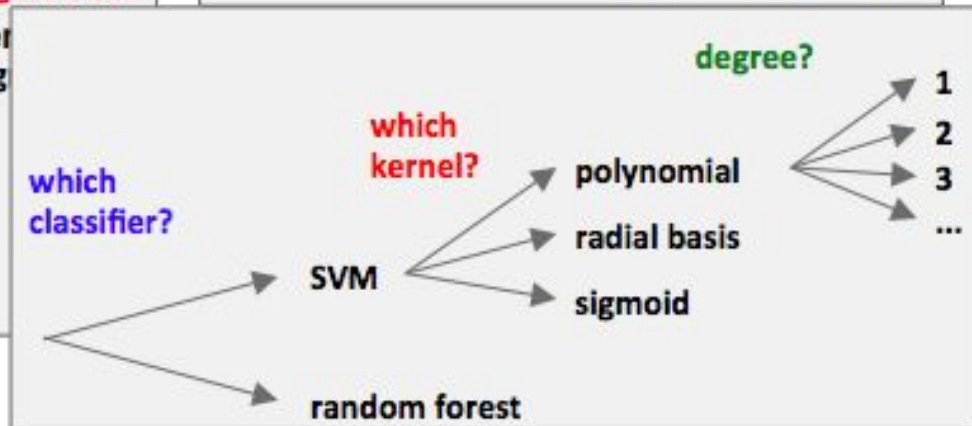
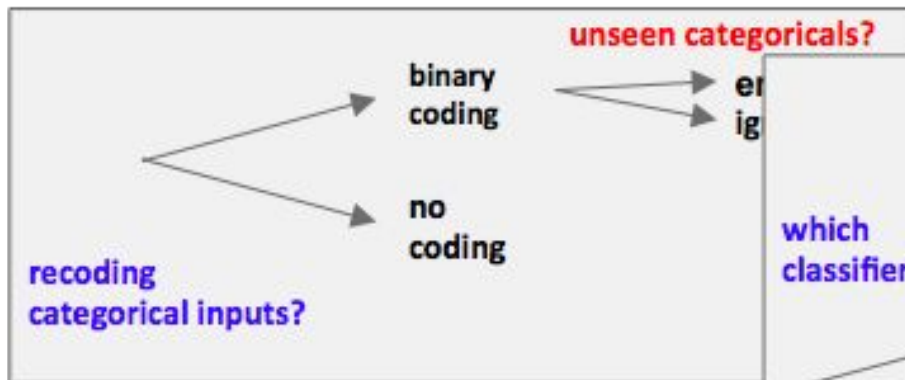
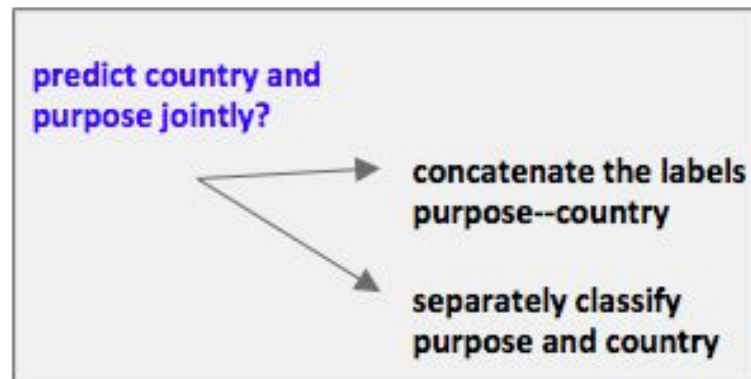
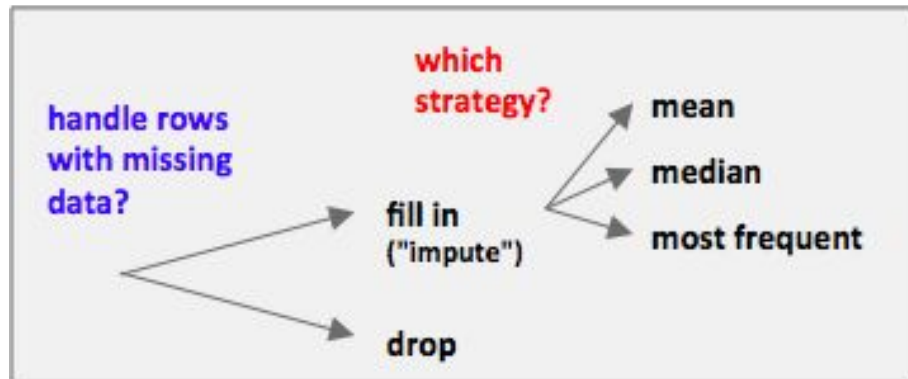
Period_minutes

Perigee_km

SELECT Period_minutes,
Perigee_km FROM table
WHERE
Anticipated_Lifetime=15

**BACKUP Slides from V. Mansinghka's Talk at Media
Lab**

Machine learning requires many decisions



Machine learning results are unstable

Approach 1

drop missing, no coding, random forest, separate classifiers

Simulations	Frequency
Egypt-Earth Science	9
Egypt-Earth/Space Science	5
Egypt-Astrophysics/Earth Science	3
Canada-Earth Science	3

Probably Egypt, definitely science

Approach 2

impute missing, binary coding, svm, joint classification

No idea

Approach 3

impute missing, no coding, random forest, separate classifiers

India-Meteorology	11
ESA-Meteorology	3
India-Communications	3
India-Earth Science	1
China (PR)-Space Physics	1
Russia-Space Physics	1

Probably India, probably science

STATISTICIAN

"Use the data from this .CSV file."

"Choose whatever data types you think are reasonable --- I don't have any knowledge about that."

"Build me a quick-and-dirty ensemble of models that gives me some ability to quantify uncertainty."

MML

```
CREATE POPULATION satellites  
FROM ucs_satellites.csv
```

```
CREATE METAMODEL ON satellites  
USING default_metamodel( GUESS(*) );
```

```
INITIALIZE 16 GENERATIVE POPULATION MODELS  
FOR satellites;  
ANALYZE satellites FOR 4 MINUTES;
```