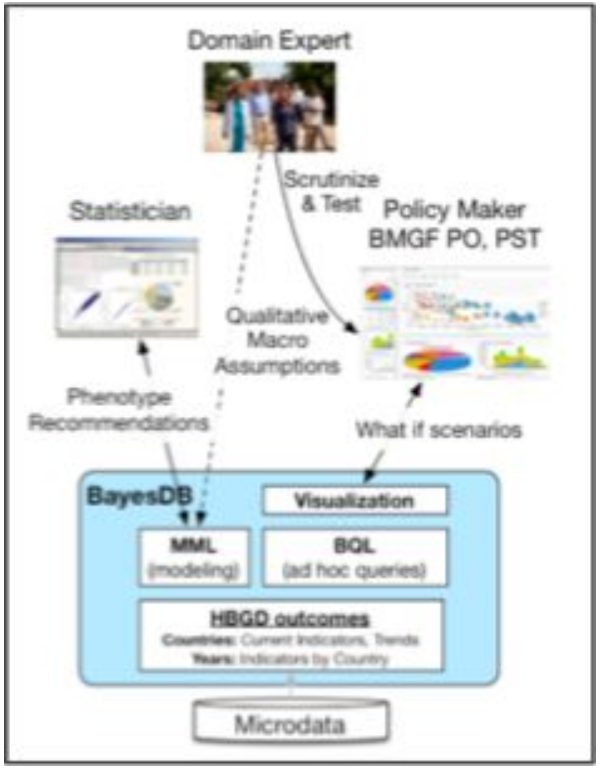
Probabilisitc Computing Lab

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."

Main Research: Convergence accuracy of an inference program

- Inference program
- Which algorithm?
- Measure of accuracy:
  - KL

    $$\mathrm{D}_{\mathrm{KL}}(Q_a(X)||P(X|y)) + \mathcal{L}(Q_a) = \log P(y)$$

  - ELBO

$$\mathcal{L}(Q_a) = \mathrm{E}_{x\sim Q_a}\left[\log \frac{\tilde{P}(x|y)}{Q_a(x)}\right]$$

Algorithm 1 Single-particle MCMC algorithm

Require: $K_1, \ldots, K_{T-1}$ are MCMC kernels and $Q_1^{aug}, \ldots, Q_T^{aug}$ are augmentation distributions compatible with problem instance defined by $\tilde{P}(\theta, z_{1:T}, y_{1:T})$

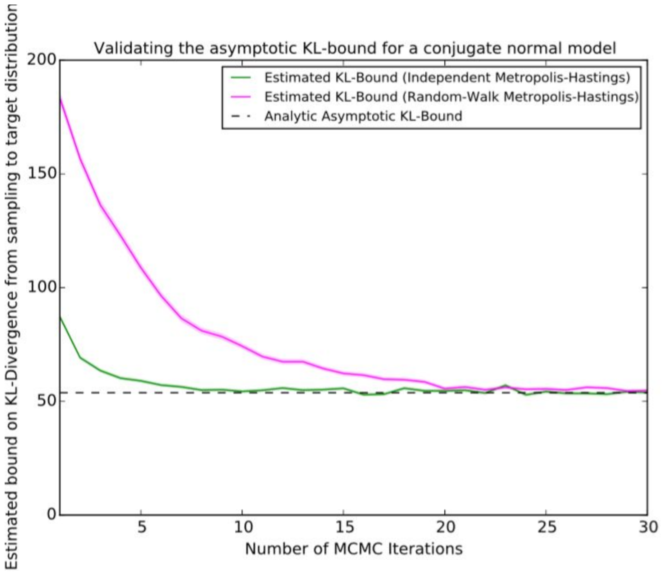
```
w ← 1
θ^(0) ~ P(Θ)                                         ▷ Sample θ from the prior
z_1^(0) ~ Q_1^aug(z_1 | θ^(0), y_1)
for t ← 1 to T − 1 do
    (θ^(t), z_{1:t}^(t)) ~ K_t(θ, z_{1:t} | θ^(t−1), z_{1:t}^(t−1))   ▷ Run the MCMC kernel K_t
    z_{t+1}^(t) ~ Q_{t+1}^aug(z_{t+1} | θ^(t), z_{1:t}^(t), y_{1:t+1})  ▷ Augment the state space
    w ← w · P(z_{t+1}^(t), y_{t+1} | θ^(t), z_{1:t}^(t), y_{1:t}) / Q_{t+1}^aug(z_{t+1}^(t) | θ^(t), z_{1:t}^(t), y_{1:t+1})
end for
return (θ^(T−1), z_{1:T}^(T−1)), w   ▷ Return the sample (θ^(T−1), z_{1:T}^(T−1)) ~ Q_a and the weight w
```

Even if log P is not known, this holds:

$$\mathcal{L}(Q_{a_1}) \geq \mathcal{L}(Q_{a_2}) \implies \mathrm{D}_{\mathrm{KL}}(Q_{a_1}(X)||P(X|y)) \leq \mathrm{D}_{\mathrm{KL}}(Q_{a_2}(X)||P(X|y))$$

Main Research: Convergence accuracy of an inference program

- Two proofs
  - KL does not increase by applying Kernel
  - Log weight is a good estimate of lower bound on ELBO
- Result
  - Independent Metropolis Hastings
  - Random walk Metropolis Hastings



Gates Foundation Case

BILL & MELINDA GATES foundation

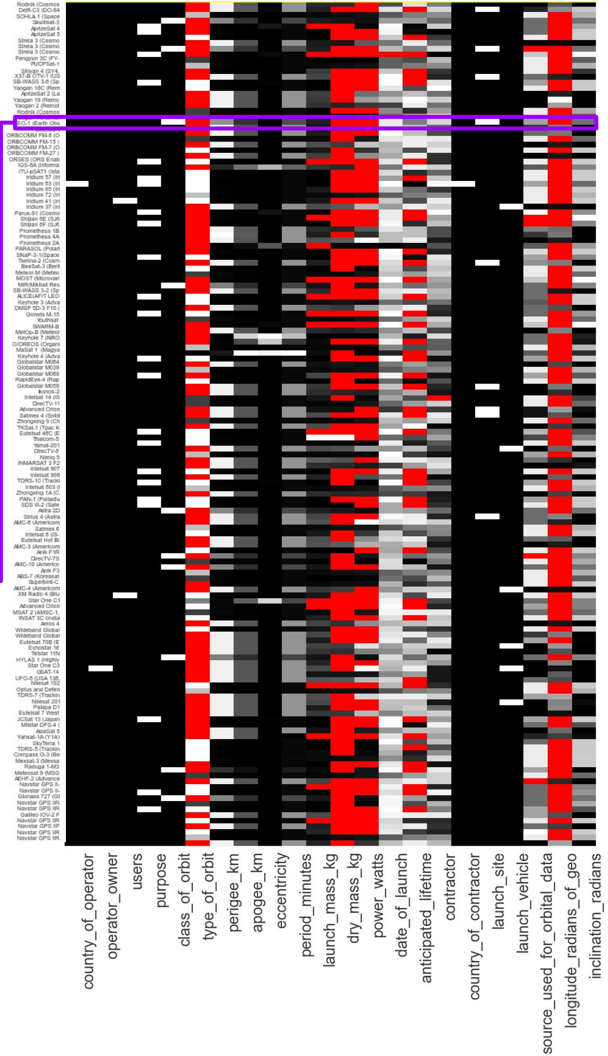
- Setting: UCS Satellites Database
    - 1167 rows (satellites) and 23 columns
    - Illustrations using 150 row subsample
    - Variables include, electrical, geopolitical, kinematic characteristics
    - Engineering note:
        - Schematics come from cleaned ‘lovecat’ states
        - Predictions 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

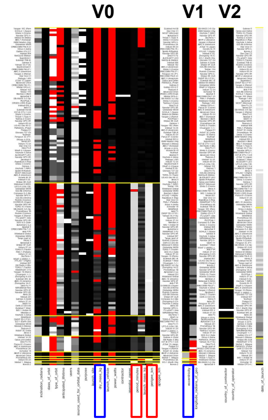


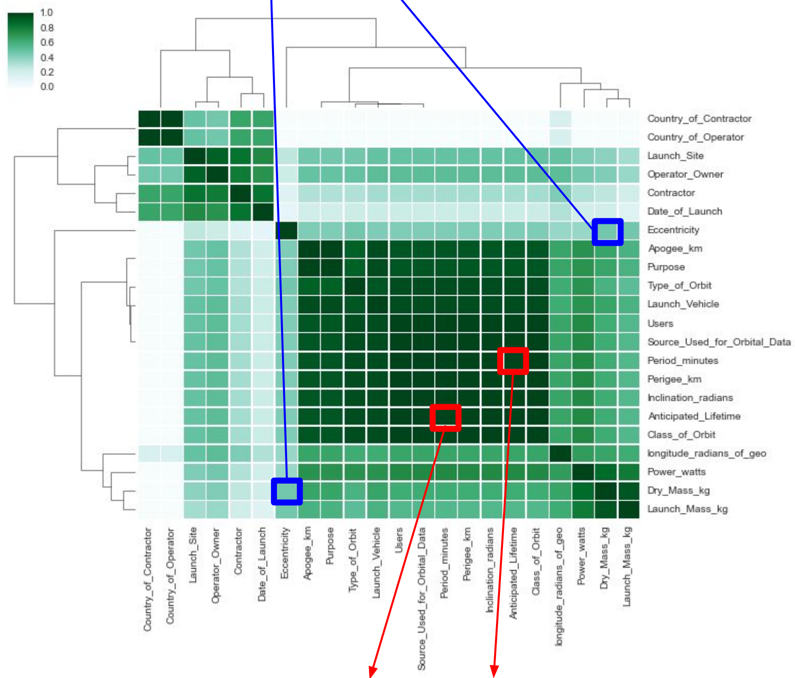
Variable	Type
Country_of_Operator	categorical
Operator_Owner	categorical
Users	categorical
Purpose	categorical
Class_of_Orbit	categorical
Type_of_Orbit	categorical
Perigee_km	normal
Apogee_km	normal
Eccentricity	normal
Period_minutes	normal
Launch_Mass_kg	normal
Dry_Mass_kg	normal
Power_watts	normal
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

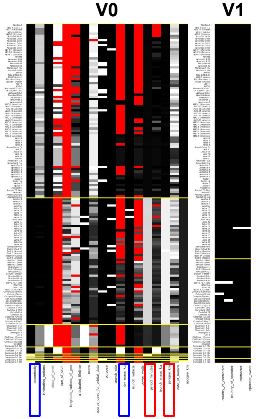
UCS Satellites Database: Relation between Dependence Probability Heatmap and clustering

```
ESTIMATE DEPENDENCE PROBABILITY
FROM PAIRWISE COLUMNS OF generator
```

$$P(\text{eccentricity} \not\perp\!\!\!\perp \text{dry\_mass}) = 1/3$$







Anticipated_lifetime and Period_minutes

Eccentricity and Dry_mass

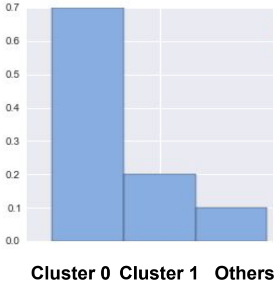
$$P(\text{Anticipated\_lifetime} \not\perp\!\!\!\perp \text{Period\_minutes}) = 3/3$$

UCS Satellites Database: Generating simulations conditioned on hypotheticals

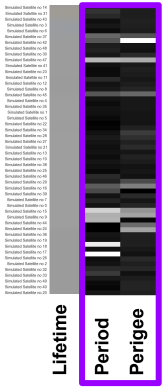
Custer 0

Cluster 1

p(cluster|lifetime = 3)

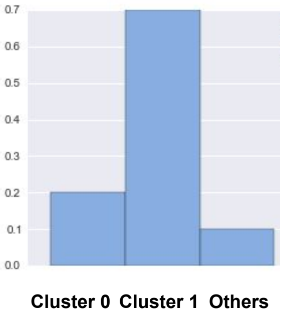



```
SIMULATE Period_minutes,
Perigee_km FROM generator GIVEN
Anticipated_Lifetime=3
```



~p(period,perigee|lifetime=3)

p(cluster|lifetime = 15)



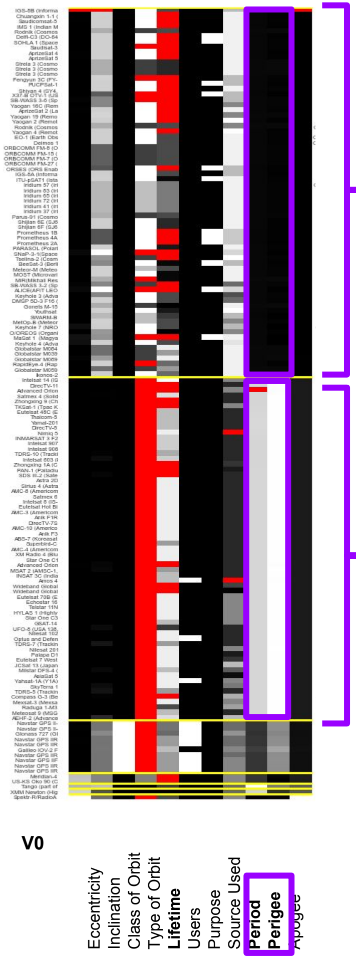

```
SIMULATE Period_minutes,
Perigee_km FROM generator GIVEN
Anticipated_Lifetime=15
```

~p(period,perigee|lifetime=15)

UCS Satellites Database:
Posterior distribution vs. CC clustering

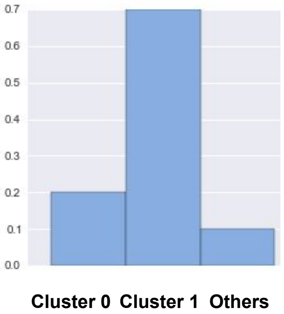
Custer 0

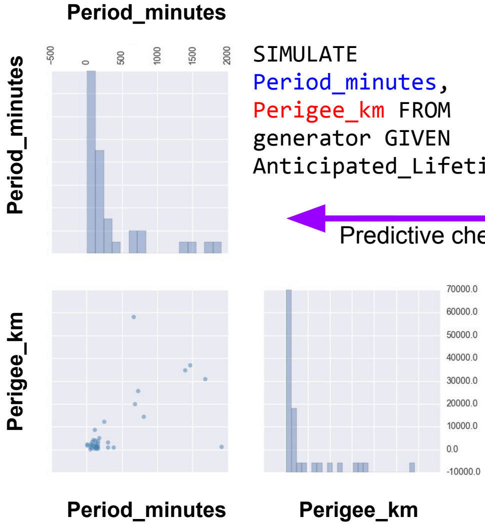
Custer 1



p(cluster|lifetime = 3)

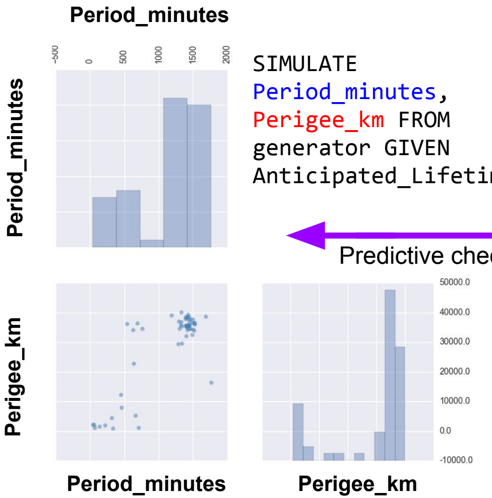
p(cluster|lifetime = 15)






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

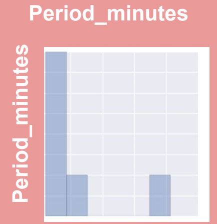
Predictive checking



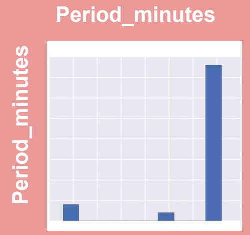

```
SIMULATE
Period_minutes,
Perigee_km FROM
generator GIVEN
Anticipated_Lifetime=15
```

Predictive checking

Real data



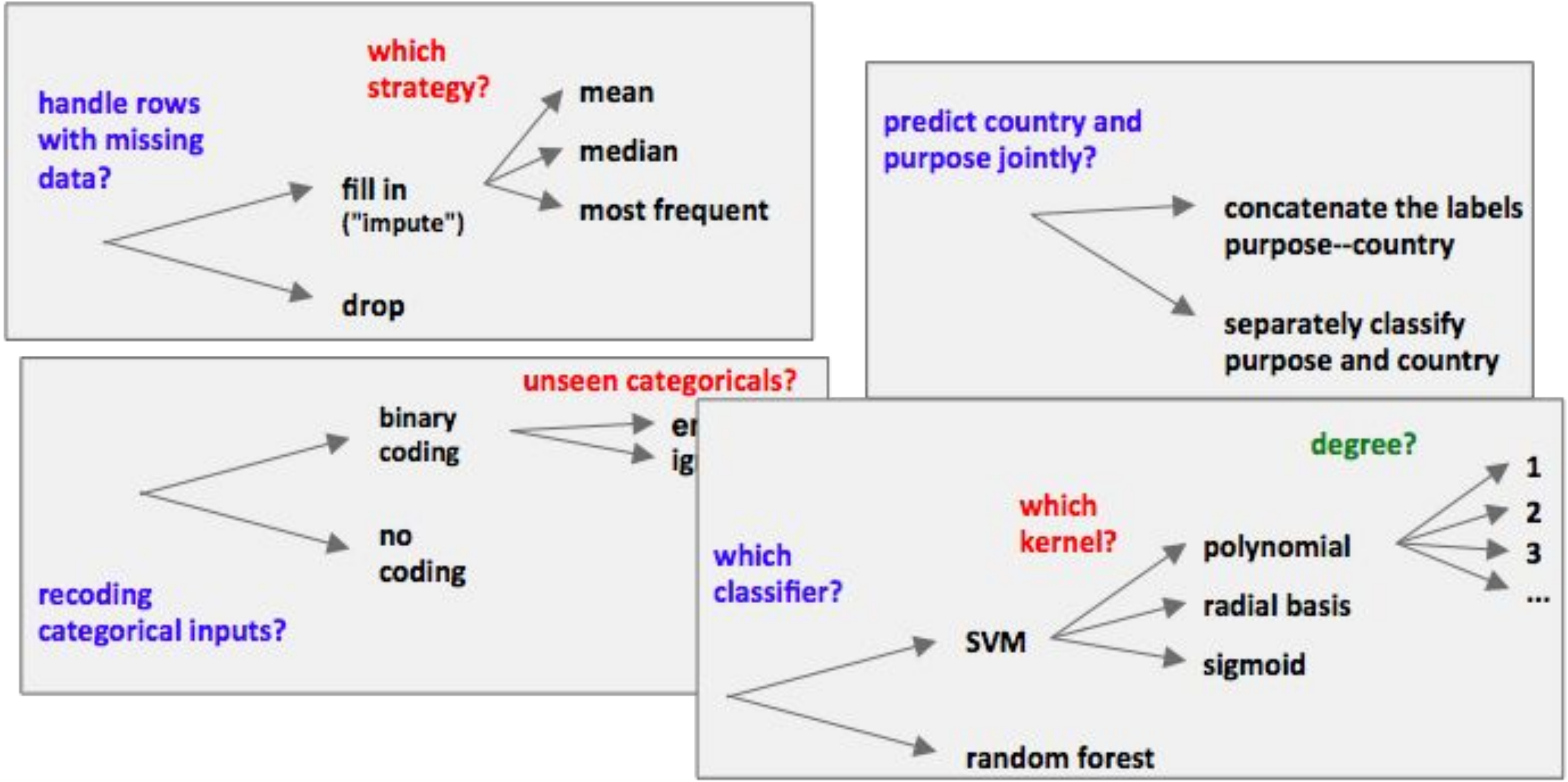

```
SELECT Period_minutes,
Perigee_km FROM table
WHERE
Anticipated_Lifetime=3
```




```
SELECT Period_minutes,
Perigee_km FROM table
WHERE
Anticipated_Lifetime=15
```

BACKUP

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;
```

Log weight is an estimate of ELBO

Since first proof

$$
\begin{aligned}
\mathrm{D}_{\mathrm{KL}}(\mathbf{Q}^{mcmc}(X^{(T)})||P_{y_{1:T}}) &\leq \mathrm{D}_{\mathrm{KL}}(\mathbf{Q}^{mcmc}(X^{(T-1)})||P_{y_{1:T}}) \\
\log p(y) - \mathcal{L}_{P_y}(\mathbf{Q}_K^{mcmc}(X^{(T)})) &\leq \log p(y) - \mathcal{L}_{P_y}(\mathbf{Q}_K^{mcmc}(X^{(T-1)})) \\
\mathcal{L}_{P_y}(\mathbf{Q}_K^{mcmc}(X^{(T-1)})) &\leq \mathcal{L}_{P_y}(\mathbf{Q}_K^{mcmc}(X^{(T)}))
\end{aligned}
$$

Note that $\mathbf{Q}_K^{mcmc}(X^{(T-1)})$ and $P_y(X^{(T-1)})$ are the marginal distributions of $X^{(T-1)}$ for $\mathbf{Q}_K^{mcmc}$ and $\mathbf{P}_{y,K}^{mcmc}$, respectively. Therefore: $\mathrm{D}_{\mathrm{KL}}(\mathbf{Q}_K^{mcmc}(X^{(T-1)})||P_y) \leq \mathrm{D}_{\mathrm{KL}}(\mathbf{Q}_K^{mcmc}||\mathbf{P}_{y,K}^{mcmc})$

So

$$
\begin{aligned}
\log p(y) - \mathcal{L}_{P_y}(\mathbf{Q}_K^{smc}(X^{(T-1)}) &\leq \log p(y) - \mathcal{L}_{\mathbf{P}_{y,K}^{smc}}(\mathbf{Q}_K^{smc}) \\
\mathcal{L}_{\mathbf{P}_{y,K}^{smc}}(\mathbf{Q}_K^{smc}) &\leq \mathcal{L}_{P_y}(\mathbf{Q}_K^{smc}(X^{(T-1)})) \\
\mathrm{E}_{x^{(0:T-1)} \sim \mathbf{Q}_K^{smc}}\left[\log \frac{\tilde{\mathbf{p}}_{y,K}^{smc}(x^{(0:T-1)})}{\mathbf{q}_K^{smc}(x^{(0:T-1)})}\right] &\leq \mathcal{L}_{P_y}(\mathbf{Q}_K^{smc}(X^{(T-1)})) \\
\mathrm{E}\left[\log w\right] &\leq \mathcal{L}_{P_y}(\mathbf{Q}_K^{smc}(X^{(T-1)}))
\end{aligned}
$$

and finally

$$
\boxed{\mathrm{E}\left[\log w\right] \leq \mathcal{L}_{P_y}(\mathbf{Q}_K^{smc}(X^{(T-1)})) \leq \mathcal{L}_{P_y}(\mathbf{Q}_K^{smc}(X^{(T)})) = \mathcal{L}_{P_y}(Q_K^{smc})}
$$