Roban Hultman Kramer

HIGH TABLE

Uncertainty versus Decisions

Some (false) dichotomies between Astrophysics and Machine Learning

robanhk@gmail.com http://roban.github.com

Astrophysics

vs.

Machine Learning

Uncertainty

is

everything

Decisions

are

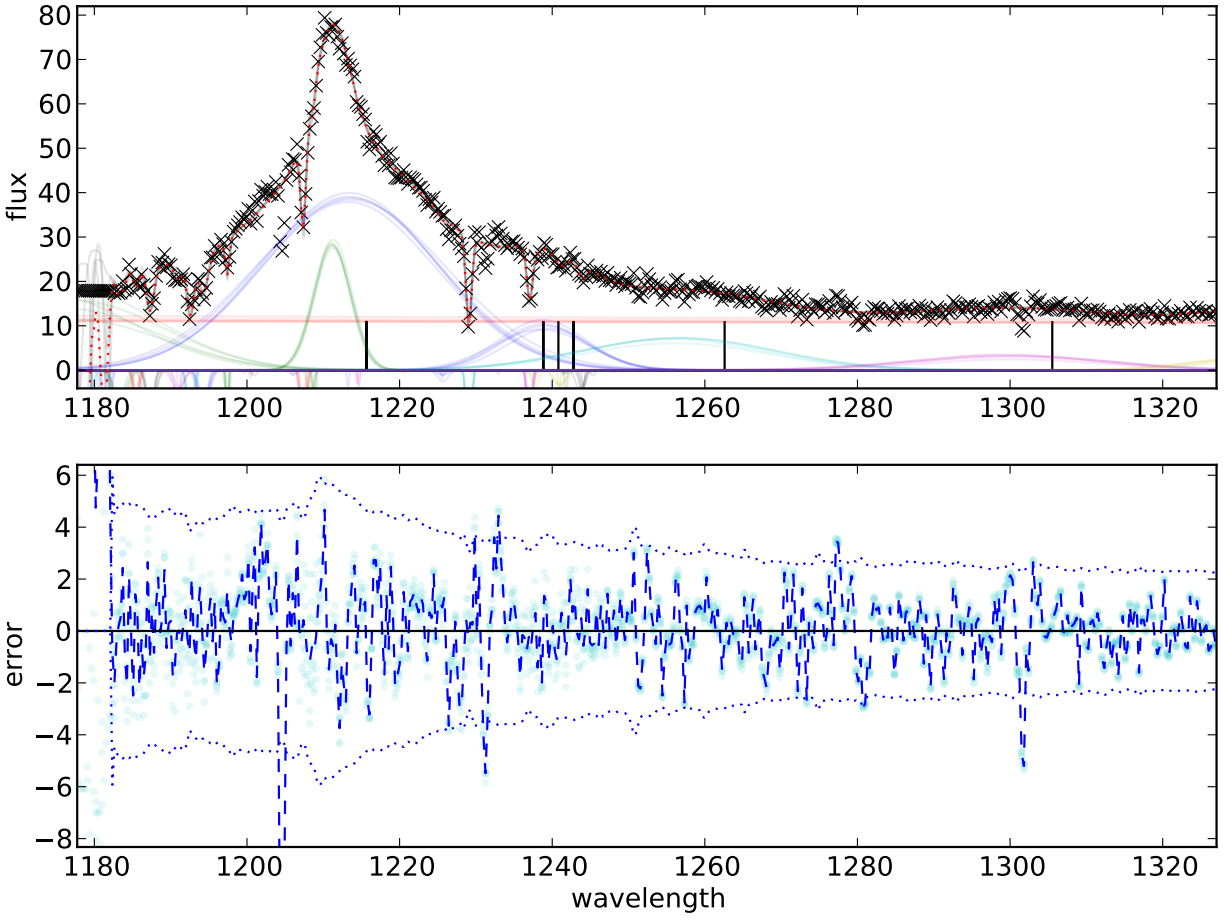
everything

Constraining Parameters

Making Predictions

Astrophysics

Uncertainty is everything

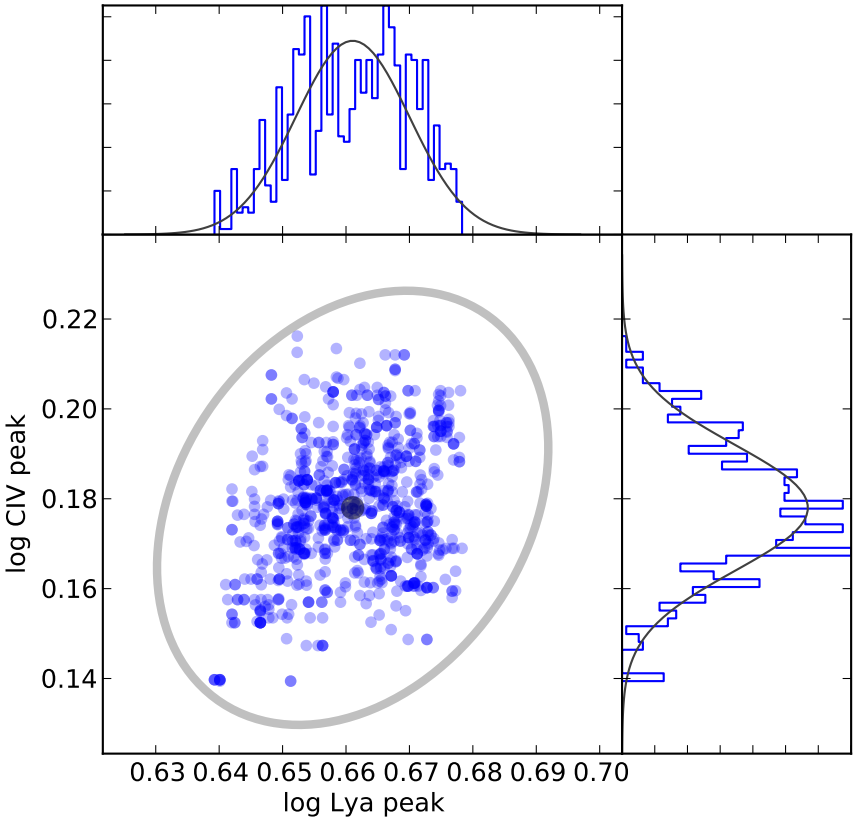


The uncertainty of the measurement is as important as the value.

Astrophysics

Uncertainty is everything

posterior distribution of parameters from MCMC

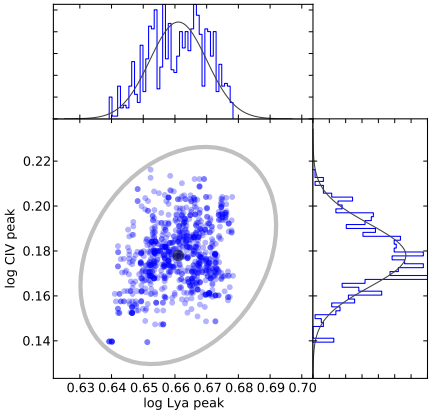


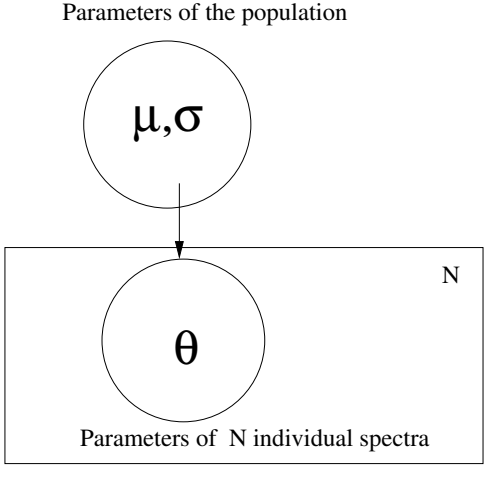
The uncertainty of the parameter is as important as the value.

Astrophysics

Uncertainty is everything

hierarchical model of the population
and individual objects





Population level:
distribution of parameter vectors

$p(\theta_i|\mu,\sigma) \sim N(\mu,\sigma)$

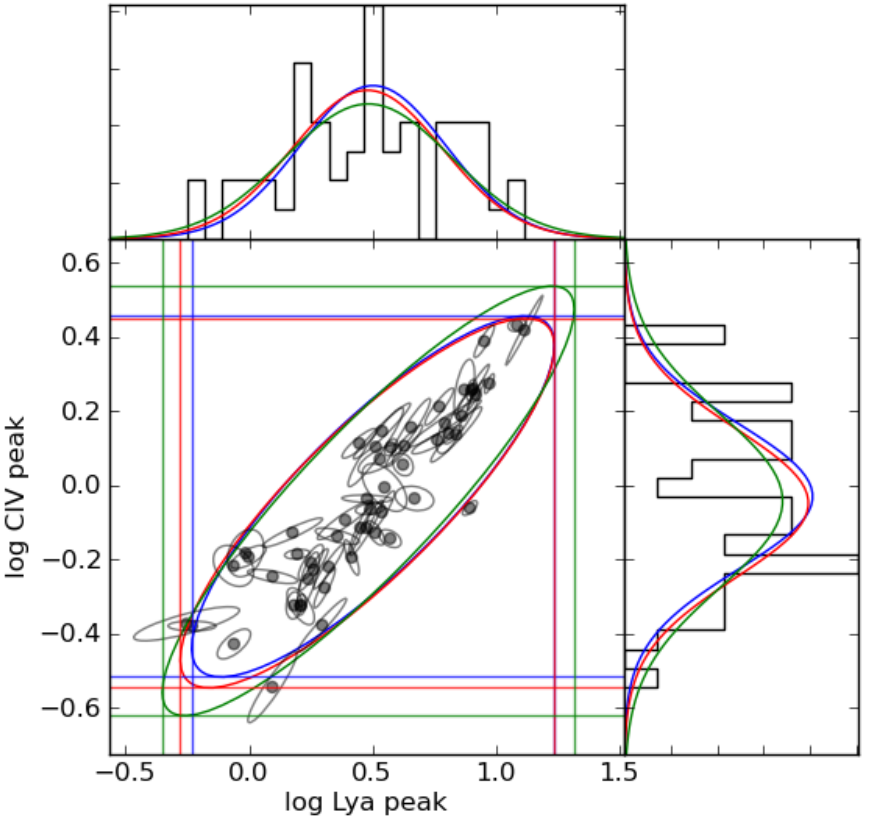
Quasar level:
likelihood of observed spectrum
given model

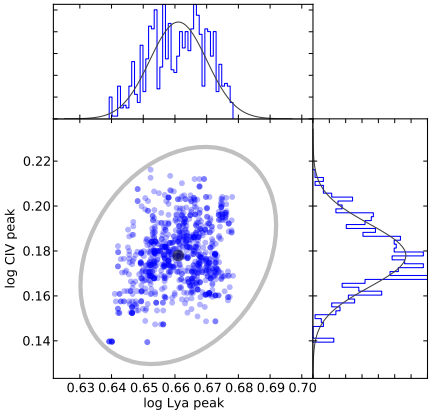
$p(F_{i,j}|\theta_i) \sim \cdots$

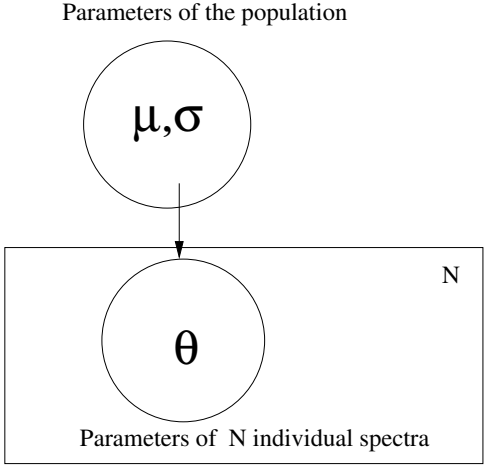
Astrophysics

Uncertainty is everything

distribution of the population







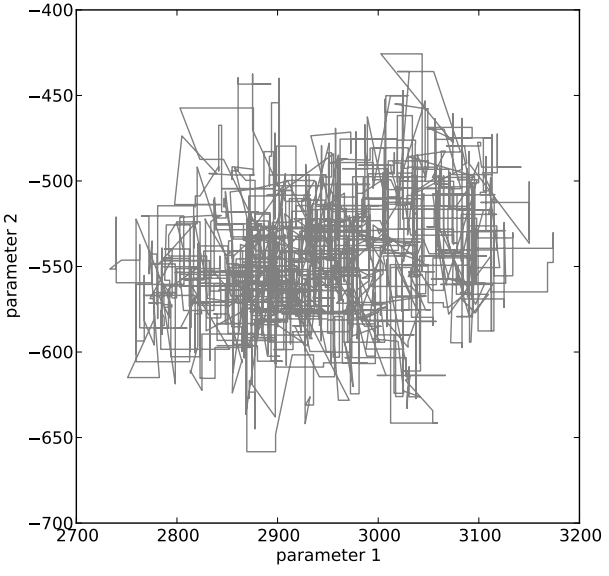
hierarchical model of the population and individual objects

Astrophysics

Uncertainty

Example: MCMC

exploring parameter space



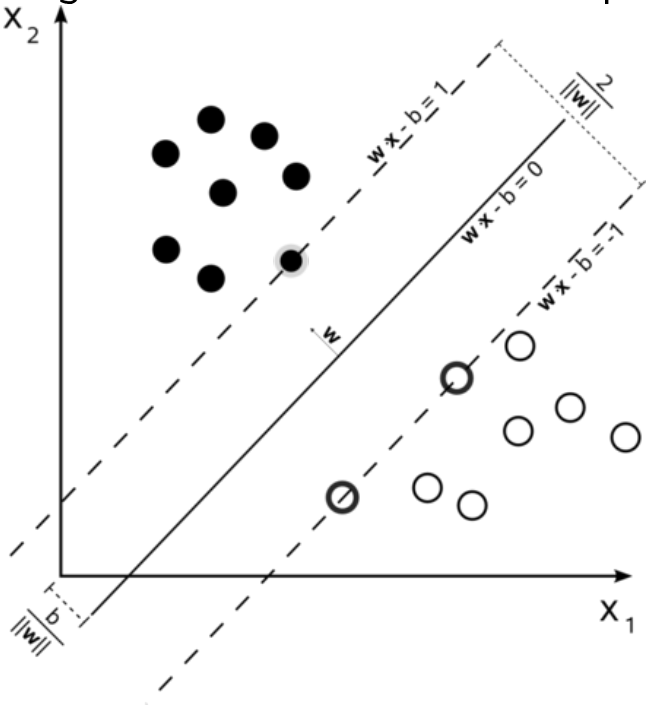
vs.

Machine Learning

Decisions

Example: SVM

finding boundaries in feature space



Credit: Wikimedia Commons
http://en.wikipedia.org/wiki/File:Svm_max_sep_hyperplane_with_margin.png

HIGH TABLE

Machine Learning

Decisions are everything

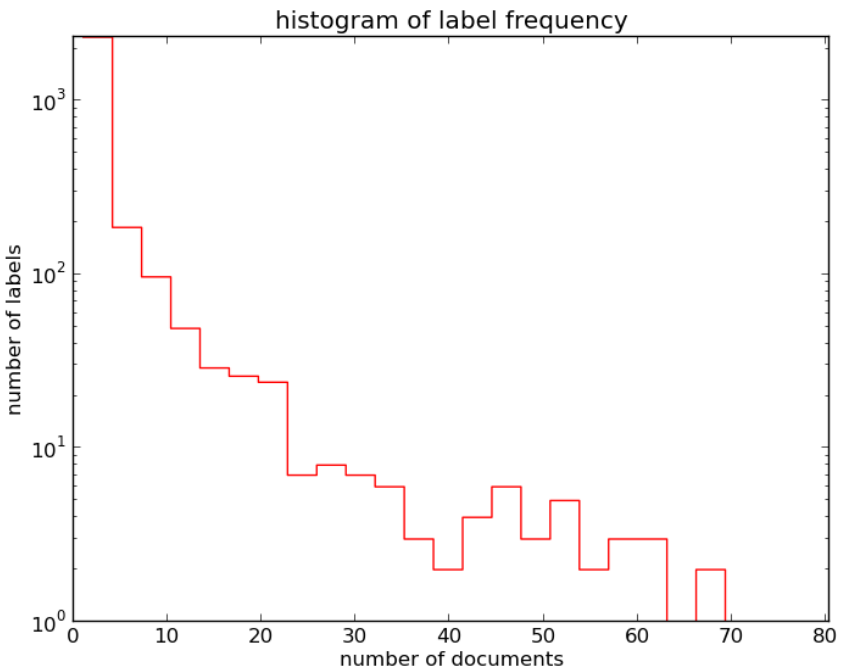
Problems:

tagging of incoming content

classification of new users

Challenges:

free-text tags are not consistent



when humans tag

HIGH TABLE

Machine Learning

Decisions are everything

Problems:

tagging of incoming content

classification of new users

Challenges:

free-text tags are not consistent

Potential Solutions:

fully-automatic tags? – not accurate enough, want human validation

pre-defined categories? – feels limiting and arduous

Machine Learning

Decisions are everything

Problems:
tagging of incoming content
classification of new users

Challenges:
free-text tags are not consistent

Potential Solutions:
fully-automatic tags? – not accurate enough, want human validation
pre-defined categories? – feels limiting and arduous

Implemented Solution:
train classifiers (e.g. SVM, logistic regression) on best tags
suggest high-confidence tags to users for validation

Forces a decision: show a tag or don't.

Astrophysics

~~Uncertainty~~

Counter Example

Decisions

vs.

Machine Learning

Decisions

planning observations
target selection

Limited by:
telescope time
instrument budgets

(Hubble oversubscribed by $\approx 600\%$)

recommendation engines
targeted marketing

Limited by:
user attention span
marketing budgets

Machine Learning

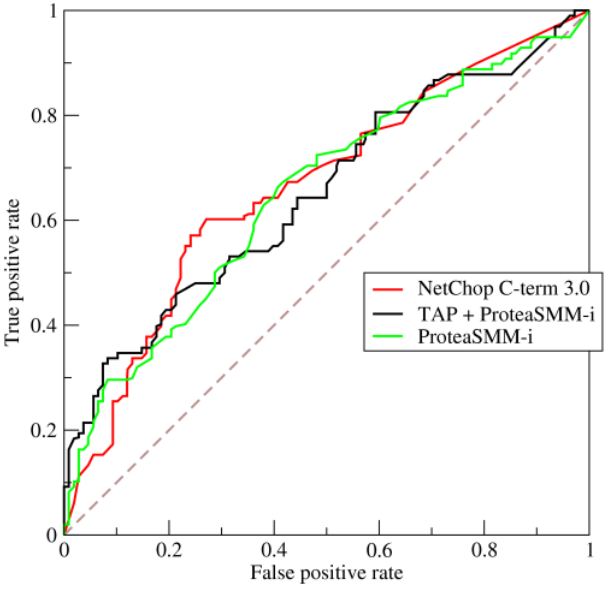
Decisions

evaluating results

F_β-scores

lift

ROC curves



Wikimedia Commons http://en.wikipedia.org/wiki/File:Roccurves.png

Astrophysics

Uncertainty

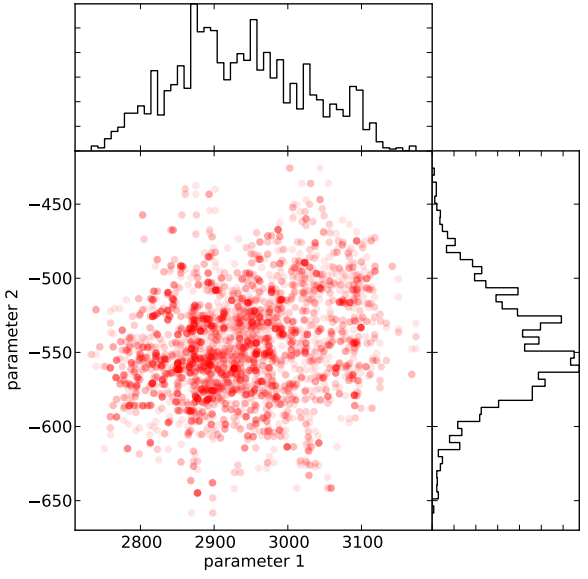
vs.

Machine Learning

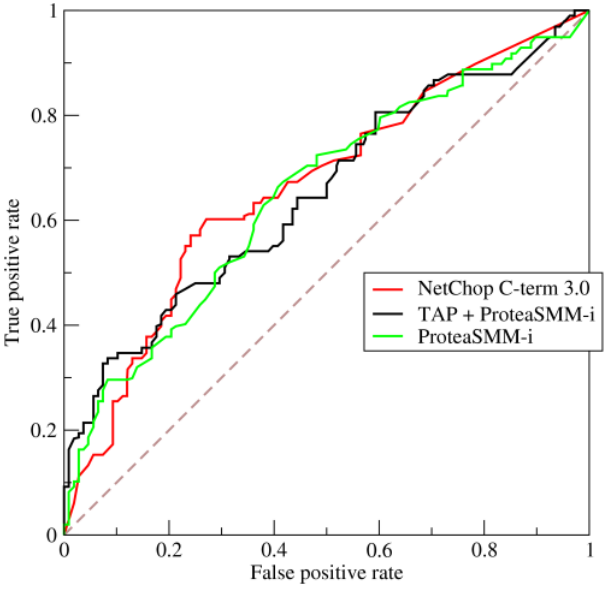
Decisions

evaluating results

error bars
p-values
posterior distributions



F_β-scores
lift
ROC curves



Wikimedia Commons http://en.wikipedia.org/wiki/File:Roccurves.png

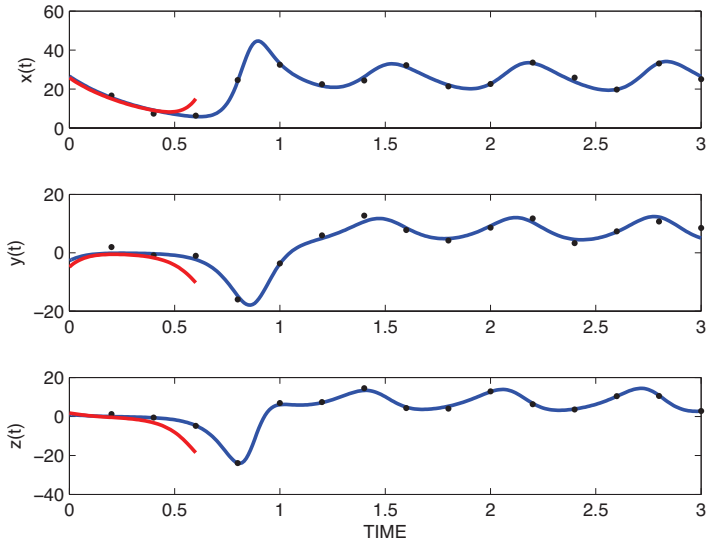
Astrophysics

Computational bottleneck: model complexity

VS.

Machine Learning

Computational bottleneck: data size



"Efficient MCMC for Climate Model Parameter Estimation: Parallel Adaptive Chains and Early Rejection" Solonen et al. *Bayesian Analysis* 7, 3 (2012), 715-736.

Astrophysics

Computational bottleneck:
model complexity

VS.

Machine Learning

Computational bottleneck:
data size

Counter Example

The Square Kilometer Array

Data Rate:
1 TB per second
after pre-processing

Computational bottleneck:
data size

"Artist's impression of the SKA dishes." Credit: SKA Organisation/TDP/DRAO/Swinburne Astronomy Productions