idark project
Utilizing Multi-parametric (model) DataSets

darksurvey visualization and database: escience center
(Faruk Diblen, Jisk Attema, Rena Bakshi, Luc Hendriks)

BSM-AI: with Jong Soo Kim, Krzysztof Rolbiecki, Roberto Ruiz de Austri, Bob Stienen,
first result is SUSY-AI: [1605.02797]
This talk is about publishing, releasing and utilizing (HEP/DM) model data (e.g. Likelihoods, constraints) in high-dimensional parameter spaces.

... proposing a tool/framework for very simple minded experiment/phenomenology people to analysis data beyond the end of the analysis (i.e. after submitting to arxiv)...

In 10 minutes I hope to have convinced you that this is useful...and hope to get some people interested...
What is Dark Matter? ...
Let’s determine the best parameter sets of a Dark Matter model?

\[ -103 \text{ GeV} < M_1 < -119 \text{ GeV}, \]
\[ 240 < M_2 < 660 \text{ GeV}, \]
\[ 108 \text{ GeV} < \mu < 142 \text{ GeV}, \]
\[ 8 < \tan \beta < 50. \]

Everybody (plots are actually from our own paper) gives 2 dimensional Information.
We can not compare solution 1 of paper X with solution 2 of paper Y
We do not get the full parameter-set of the solutions ...
Simple reminder...

Simplified models * number of models

not equal

Full model

Do not be afraid of models > 2 parameters! We need them!
Publish model evaluations in N-dim
⇒ There are 3 dimensions ;)

There are 3 dimensions ;)

X Label

Y Label

Z Label
Can we collect/store the multidimensional model datasets?

Prototypes at www.idarksurvey.com and http://54.93.46.86/

Idea:
- Provide a database (postgresql for the moment)
  for multi-dimensional theory model solutions (or HEP data)
- Connect the database with web-based visualization tools as frontend
  (developing SPOT with Dutch eScience center, also quick demo using highcharts)
- Connecting to arxiv (maybe with hepdata?)

Let us have a look....
SPOT visualisation

• “SPOT” started as tool in Summer in the City project ([https://github.com/jiskattema/spot](https://github.com/jiskattema/spot))
• plans to be useful as scientific tool: provenance, downloading of data
  • multi-dimensional data using linked, animated & interactive low dimensional plots
  • two contributors: Jisk Attema, Faruk Diblen
• Chart.js using html5 canvas (no DOM, gpu accelerated drawing)
  • Uses modern web technologies: quick development (crossfilter, node.js, ampersand.js, chart.js, vis.js)
• two possible backends: crossfilter (fully client side) and PostgreSQL
• Responsive interface: material design (google)
  • documented (jsdoc) and tested (jasmine)
• three types of data: numerical, categorial (labels or text), dates and duration
SPOT visualisation

• exploration sessions can be saved or restored as json file
• mostly a data viewer, modification not supported. Some changes for displaying purposes are possible (change strings labels, format dates)
• works with grouping, all plots performant (canvas is often OpenGL accelerated)
Why is such high-dimensional (model) data useful?

1. Data preservation
2. Exclude/Update theory predictions
3. Compare solutions of theory papers
4. “Machine” Learning limits in high dimensions
5. Training tools...
6. ...?
Example of n-dimensional model evaluations:

ATLAS released > 300000 model evaluations in the 19 dimensional MSSM

Can we “learn” the full 19-dim pMSSM exclusion range with
300000 models?
Is this enough to populate a 19 dimensional parameter space?
ATLAS analysis chain for each of the 300000 model points

\[ T = O(\text{hours}) \]
Machine Learning Applied

- Training data: model points in supersymmetric model with only phenomenologically relevant parameters (pMSSM) source: ATLAS [1508.06608]
- Testing data: independent (unseen) data

Is currently a classification algorithm using scikit-learn (a Random Forest)
Exclusion analysis

Model point

Simulate events

Simulate detector response

Event reconstruction

Calculate cross section

Analyze results

Exclusion

$T = O(\text{hours})$

Model point

WARNING

$T = < O(1 \text{ ms})$
Why use SUSY-AI / Machine Learning

- **Fast** statistical results based on earlier analyses
- **High** accuracy by learning hard-to-see relations in data
- Works also in **submodels** of the learned model (e.g. mSUGRA)
- Providing **confidence levels** on prediction

- New way to publish and recast **multivariate data**
- Creating **plots** not present in paper
- **Re-usability** and **persistence** of analysis and results
Summary

Idark project

1. idarksurvey: Database and Online data visualisation
2. Generalizing Likelihoods and Limits for the HEP community with Machine Learning (first tool is www.susy-ai.com)

3. Next steps: Discussing tools with community ...
Extra Slides
SUSY-AI (Online)

- Tool has been published https://susyai.hepforge.org
  - Python interface to classifier
  - Scikit-learn package for ML implementation
- Online interface http://susy-ai.org/
  - All functionalities except batch predictions
  - Predictions in < 2 seconds
Summary and Conclusions

- High-speed + high accuracy prediction of ATLAS exclusion
- Applicable to phenomenological supersymmetry and its submodels
- Programmatic and online interface (http://www.susy-ai.org)

- First time use of Machine Learning for publishing and extrapolating multivariate results
- More models will be done
  - Dark Matter models
  - Higgs couplings
  - Effective Field Theories
  - ...

  Development can be accelerated with more public data

  ➔ New “initiative” to release multivariate model likelihood evaluations!

More people highly welcome!

Non-polite summary: STOP WORKING ONLY ON SIMPLIFIED MODELS!
Used training data to learn classification.

It determines a confidence level of its classification using the training data.

Low confidence level

➔ Higher chance of misclassification

Confidence level: “Chance for misclassification”

➔ Can also be selected by the user

(CL by estimating signal / total as function of classifier output)
Accuracy (no confidence cut) as a function of training data points: Note that we can do better by “active learning”, i.e. adding points at difficult places.

Conclusion: Machine Learning (SUSY-AI) works already fantastic in 19 dimensions with 300000 training points (especially it knows its own confidence)

Select CL cut to >0.99
\[ \Rightarrow \text{Accuracy is } >0.99 \]
Model exclusion in Particle Physics

- Consider a model of new physics with parameters A, B, C, D and E
- Assume Supersymmetry or a model to fit the Higgs couplings
- Usually such models have many parameters (e.g. pMSSM can have 19, Higgs couplings has >5-8).

**Our Goal:** Set limits on models and its parameters
Model exclusion in Particle Physics

We are used to publish on a piece of paper... i.e. in 2 dimensions.

What we usually do in HEP?

a) Forget about the 6 dimensional model, take a “simplified” model with only A and B
b) Set parameter C = ..., D = ..., E = ... and plot A vs B
c) More sophisticated: Show projections (profiling) of the model likelihood on parameters A, B as well as C, D etc.

This is the experimentalists...
What about the theory community?