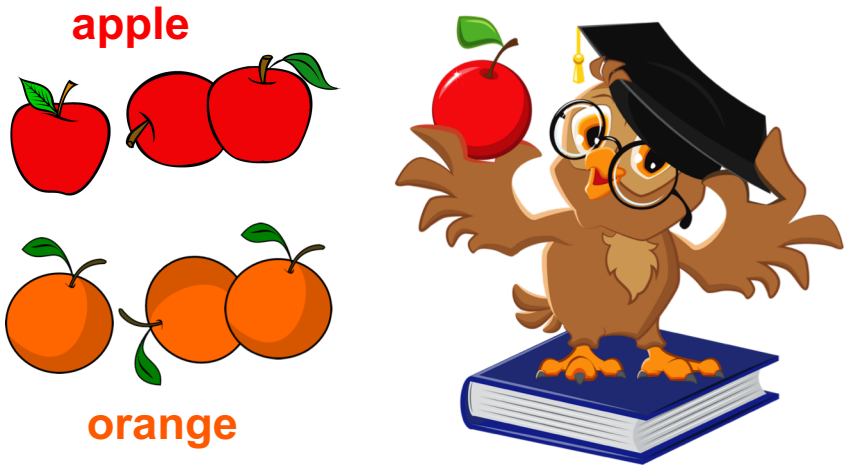
Machine Learning Background
(Tutorial at SPLC'20: Part 1)

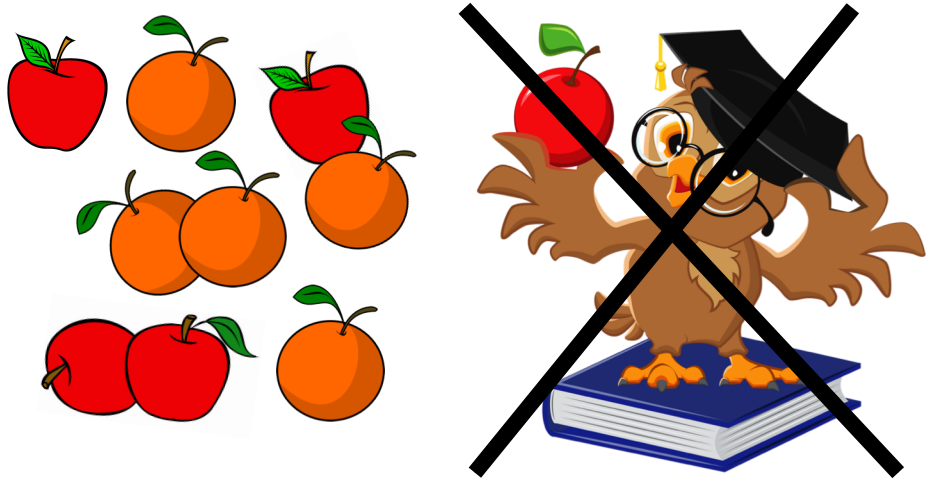
Juliana Alves Pereira, Hugo Martin,
Paul Temple, Mathieu Acher

https://github.com/VaryVary/

ML Classification

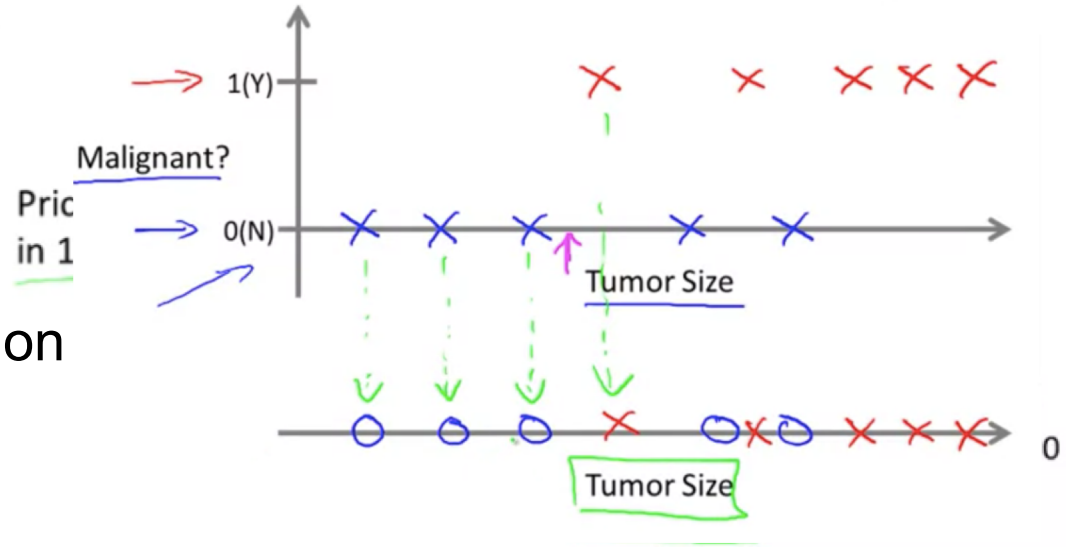
- **Supervised learning**
  - Knowledge of output: labelled data
  - Goal: teach the computer how to label new data
  - Algorithms: linear and polynomial regression, decision trees, random forest, and others
- **Unsupervised learning**
  - No knowledge of output: unlabelled data
  - Goal: let the computer learn by itself data patterns/groupings
  - Algorithms: k-means, genetic algorithms, clustering approaches, and others





ML Classification

- **Supervised learning**
  - **Regression**
    - Predict continuous valued output (price)
    - Applications: #1 x264 Performance Prediction
      #2 Kernel Size Prediction
  - **Classification**
    - Predict discrete valued output (0 or 1)
    - Applications: #3 “Smart” build infrastructure



“Smart” build infrastructure

sample

```
CONFIG_PM_WAKELOCKS=y
CONFIG_PM_WAKELOCKS_LIMIT=100
CONFIG_PM_WAKELOCKS_GC=y
CONFIG_PM=y
# CONFIG_PM_DEBUG is not set
CONFIG_PM_CLK=y
CONFIG_PM_GENERIC_DOMAINS=y
CONFIG_WQ_POWER_EFFICIENT_DEFAULT=y
CONFIG_PM_GENERIC_DOMAINS_SLEEP=y
CONFIG_PM_GENERIC_DOMAINS_OF=y
CONFIG_ENERGY_MODEL=y
CONFIG_ARCH_SUPPORTS_ACPI=y
```

build passing

```
CONFIG_VM_EVENT_COUNTERS=y
CONFIG_SLUB_DEBUG=y
# CONFIG_SLUB_MEMCG_SYSFS_ON is not set
# CONFIG_COMPAT_BRK is not set
# CONFIG_SLAB is not set
CONFIG_SLUB=y
# CONFIG_SLOB is not set
# CONFIG_SLAB_MERGE_DEFAULT is not set
# CONFIG_SLAB_FREELIST_RANDOM is not set
# CONFIG_SLAB_FREELIST_HARDENED is not set
CONFIG_SHUFFLE_PAGE_ALLOCATOR=y
CONFIG_SLUB_CPU_PARTIAL=y
CONFIG_SYSTEM_DATA_VERIFICATION=
```

```
CONFIG_PM_WAKELOCKS=y
CONFIG_PM_WAKELOCKS_LIMIT=100
CONFIG_PM_WAKELOCKS_GC=y
CONFIG_PM=y
# CONFIG_PM_DEBUG is not set
CONFIG_PM_CLK=y
CONFIG_PM_GENERIC_DOMAINS=y
CONFIG_WQ_POWER_EFFICIENT_DEFAULT=y
CONFIG_PM_GENERIC_DOMAINS_SLEEP=y
CONFIG_PM_GENERIC_DOMAINS_OF=y
CONFIG_ENERGY_MODEL=y
CONFIG_ARCH_SUPPORTS_ACPI=y
```

build passing

Classification problem:
predict the class (BUILD/FAILURE)
out of options values

Kernel Size/Time Prediction

sample

```
CONFIG_PM_WAKELOCKS=y
CONFIG_PM_WAKELOCKS_LIMIT=100
CONFIG_PM_WAKELOCKS_GC=y
CONFIG_PM=y
# CONFIG_PM_DEBUG is not set
CONFIG_PM_CLK=y
CONFIG_PM_GENERIC_DOMAINS=y
CONFIG_WQ_POWER_EFFICIENT_DEFAULT=y
CONFIG_PM_GENERIC_DOMAINS_SLEEP=y
CONFIG_PM_GENERIC_DOMAINS_OF=y
CONFIG_ENERGY_MODEL=y
CONFIG_ARCH_SUPPORTS_ACPI=y
```

build passing

Size? **176.8Mb**
Time? **54.3 min**

```
CONFIG_VM_EVENT_COUNTERS=y
CONFIG_SLUB_DEBUG=y
# CONFIG_SLUB_MEMCG_SYSFS_ON is not set
# CONFIG_COMPAT_BRK is not set
# CONFIG_SLAB is not set
CONFIG_SLUB=y
# CONFIG_SLOB is not set
# CONFIG_SLAB_MERGE_DEFAULT is not set
# CONFIG_SLAB_FREELIST_RANDOM is not set
# CONFIG_SLAB_FREELIST_HARDENED is not set
CONFIG_SHUFFLE_PAGE_ALLOCATOR=y
CONFIG_SLUB_CPU_PARTIAL=y
CONFIG_SYSTEM_DATA_VERIFICATION=
```

```
CONFIG_PM_WAKELOCKS=y
CONFIG_PM_WAKELOCKS_LIMIT=100
CONFIG_PM_WAKELOCKS_GC=y
CONFIG_PM=y
# CONFIG_PM_DEBUG is not set
CONFIG_PM_CLK=y
CONFIG_PM_GENERIC_DOMAINS=y
CONFIG_WQ_POWER_EFFICIENT_DEFAULT=y
CONFIG_PM_GENERIC_DOMAINS_SLEEP=y
CONFIG_PM_GENERIC_DOMAINS_OF=y
CONFIG_ENERGY_MODEL=y
CONFIG_ARCH_SUPPORTS_ACPI=y
```

build passing

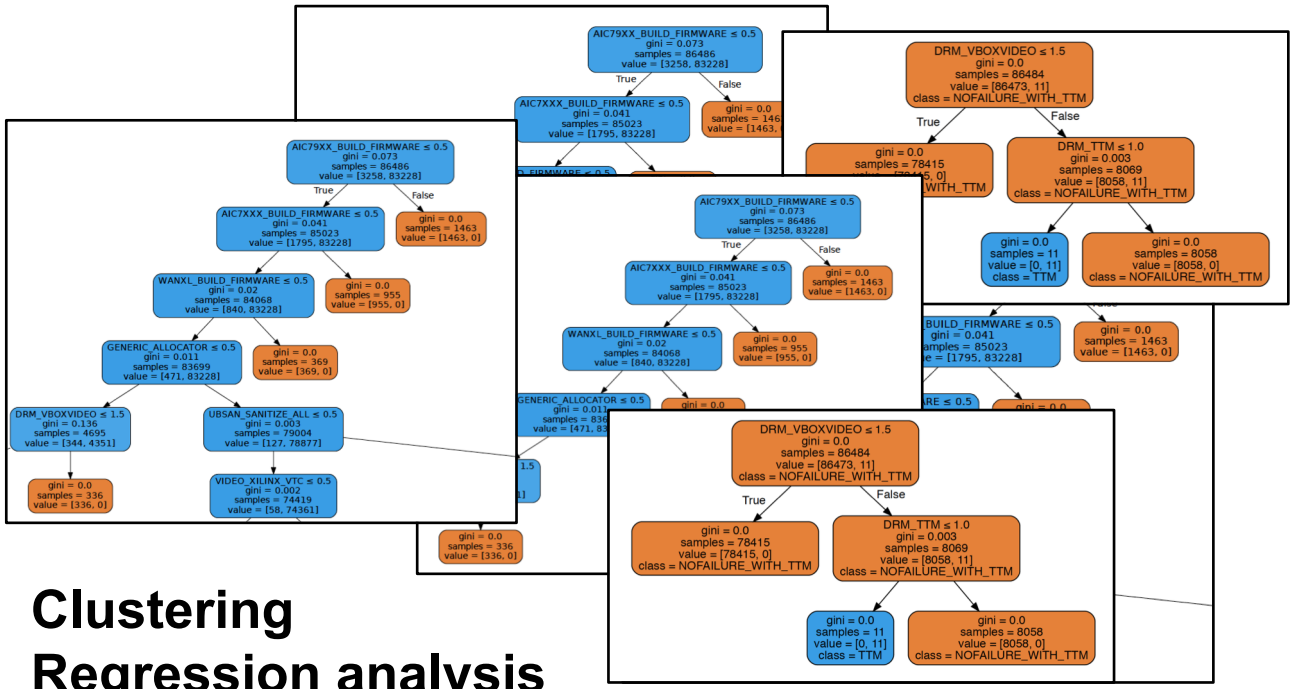
102.3Mb
10.8 min

IGRIDA Computing Grid

95,854 configurations
Costly and hard to engineer

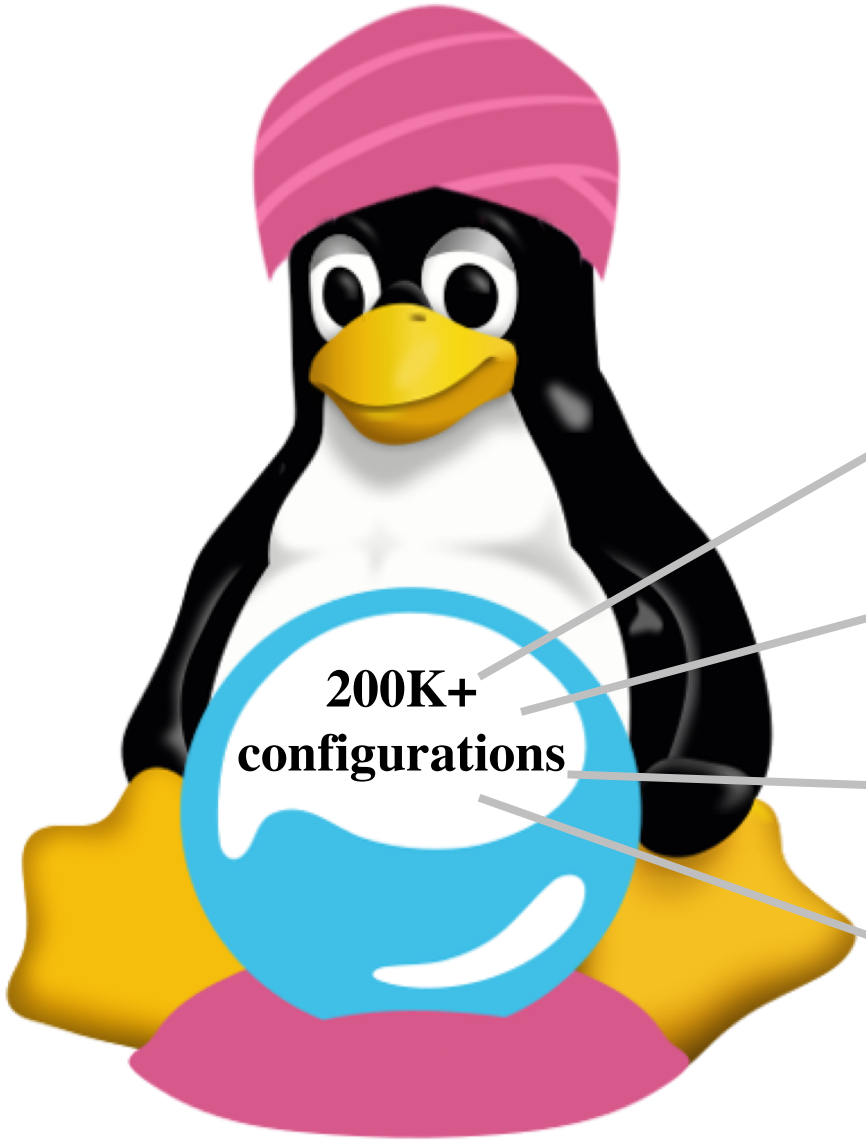
Regression problem:
predict a quantitative value
out of options values

Statistical learning can automatically pinpoint what combinations of options lead to a failure and/or influence size/time



- Clustering
- Regression analysis
- Decision tree
- Random forest
- Gradient boosting tree
- Neural networks
- and others

"Optimizer, recommender, configurator" can be built on top of TuxML



Bug reporting

Hey Linux developers, the combination of options DRM_VBOXVIDEO and GENERIC_ALLOVATOR is buggy and responsible of lots of build failures

Bug prevention

Hey KernelCI, 0-day, and testing community do not build configurations with specific values of DRM_VBOXVIDEO and GENERIC_ALLOVATOR

Size prediction

Hey Linux user, your configuration will give a 70Mb kernel. By deactivating X86_NEEDS_RELOC can save 23Mb (ask me if you want to further decrease)

Documentation improvement

Hey Linux community, we have updated the documentation to describe options that do influence size. Please, let us know if we miss some. TuxML needs you!

"Optimizer, recommender, configurator" can be built on top of TuxML

Machine Learning: Python

- **One of the easiest programming languages to learn**
- **Syntax is closely related to English**
- **There are incredible libraries**
  - Numpy (numerical analysis library)
  - Pandas (For processing CSV files)
  - Matplotlib & Seaborn (visualizations)
  - Scikit-Learn (algorithms)
  - Tensorflow and Pytorch (Deep Learning)
  - Kaggle.com (Machine Learning with Python Projects)

Open-source web application to create live code documents, equations and visualizations

Docker for a reproducible environment with tools/packages needed and Python procedures inside

Jupyter Notebooks:
https://github.com/TuxML
https://github.com/jualvespereira/ICPE2020
https://github.com/FAMILIAR-project/x264-inputsensitivity

Machine Learning Background: Decision Tree (Tutorial at SPLC'20: Part 1)

Juliana Alves Pereira, Hugo Martin,
Paul Temple, Mathieu Acher

https://github.com/VaryVary/

Demonstration of VaryLaTeX (Tutorial at SPLC'20: Part 2)

Juliana Alves Pereira, Hugo Martin,
Paul Temple, **Mathieu Acher**

https://github.com/VaryVary/