# Tool Resource Prediction for Genomic Datasets

Masterproject: Öner Aydogan

# Motivation

- **Galaxy**: web-based, open-source scientific analysis platform
- Over **5500 tools** for bioinformatics applications

**fixed amount** of memory for each tool

**Overallocation** → waste of resources

**Underallocation** → failure of tools

# Motivation

**Example** with RNA-Star:

- Uses in one job **30 GB**, in a different job **10 GB**
- Suppose **20 GB** are allocated
- **Insufficient memory** or **too much** would be assigned

→ Usage of **machine learning** methods to predict memory usage

- Automatically **learn patterns** and **decision rules** from data

# Galaxy job run dataset & Preprocessing

- Collected from 31.12.2019 to 23.05.2022 on **Galaxy Europe**
- **38 millions** entries (10 GB) → after preproccessing **26 million** (2.8 GB)
- Filter **invalid combinations** for *Filesize & Number of files*

| job_id | tool_id | state | filesize | num_files | runtime_seconds | slots | memory_bytes | create_time |
|---|---|---|---|---|---|---|---|---|
| 35222449 | toolshed.g2.bx.psu.edu/repos/devteam/concat/gops_concat_1/1.0.1 | ok | 225 | 2 | 6.0000000 | 1.0000000 | 156790784.0000000 | 2021-11-28 02:28:42.941101 |
| 35223981 | toolshed.g2.bx.psu.edu/repos/devteam/subtract/gops_subtract_1/1.0.0 | ok | 1029 | 2 | 6.0000000 | 1.0000000 | 151023616.0000000 | 2021-11-28 02:31:37.448641 |
| 35224560 | toolshed.g2.bx.psu.edu/repos/iuc/lofreq_filter/lofreq_filter/2.1.5+galaxy0 | ok | 33056 | 1 | 5.0000000 | 1.0000000 | 150167552.0000000 | 2021-11-28 02:32:51.30508 |
| 35222790 | toolshed.g2.bx.psu.edu/repos/iuc/ivar_trim/ivar_trim/1.3.1+galaxy0 | ok | 94167318 | 3 | 139.0000000 | 1.0000000 | 1131245568.0000000 | 2021-11-28 02:29:23.773031 |
| 35228785 | CONVERTER_gz_to_uncompressed | ok | 2138539188 | 1 | 189.0000000 | 1.0000000 | 4296945664.0000000 | 2021-11-28 03:48:33.163304 |
| 35227471 | toolshed.g2.bx.psu.edu/repos/devteam/vcfvcfintersect/vcfvcfintersect/1.0.0_rc3+galaxy0 | ok | 70870 | 3 | 7.0000000 | 1.0000000 | 149172224.0000000 | 2021-11-28 02:41:30.823983 |
| 35225116 | toolshed.g2.bx.psu.edu/repos/devteam/column_maker/Add_a_column1/1.6 | ok | 135 | 1 | 7.0000000 | 1.0000000 | 149745664.0000000 | 2021-11-28 02:34:00.328274 |

# Machine Learning methods

- Random Forest
- Linear Regression
- XGB (Extreme Gradient Boosting)
- SVR (Support Vector Regression)

- Baseline:
  - LR & SVR: using **default parameters**
  - RF & XGB: n_estimators = 200

# Training & evaluation

- Input features: *Filesize, Number of files* & *Slots*
- Target: *Memory bytes*

Example:

- fastqc/0.72 → train & test set (80%|20%) → 5-fold CV → evaluation on test set
- scoring method: $R^2 = 1 - \frac{\sum (y_i - \hat{y})^2}{\sum (y_i - \bar{y})^2}$

# Removing faulty data

- Initial assigned memory:
  - Given by **tool destinations file**
  - **Default value** of 1 GB
- $\max\_memory = initial\_assigned\_memory * 2^3 = initial\_assgined\_memory * 8$
- Filter entries **exceeding** *max_memory or 1 TB*
- **Validity** of the dataset **is put in question**

# Results: Removing faulty data

| Dataset | Random Forest | XGB | Linear Regression | SVR |
|---|---|---|---|---|
| fastqc/0.72 | -0.13 → 0.85 | -0.17 → 0.86 | 0.00 → 0.82 | -0.01 → 0.89 |
| ivar_trim/1.2.2 | -0.93 → 0.86 | -1.50 → 0.87 | 0.00 → 0.77 | -0.02 → 0.84 |
| ivar_removereads/1.2.2 | -0.51 → 0.64 | -0.64 → 0.68 | 0.00 → 0.74 | -0.03 → 0.70 |
| cutadapt/1.16.5 | -0.82 → 0.89 | -1.64 → 0.89 | 0.05 → 0.13 | 0.05 → 0.90 |
| mimodd_reheader/0.1.8_1 | -0.37 → -0.12 | -0.47 → -0.51 | 0.00 → 0.37 | -0.02 → 0.16 |

# Correlation analysis

- **Pearson correlation coefficient:** *Filesize* ↔ *Memory bytes*
- about 300 of total 4800 tools moderate to strong **negative correlation**

| Version | Pearson correlation Filesize ⟷ Memory bytes | nr_samples |
|---|---|---|
| 0.4 | 0.99 | 5 |
| 2.3.4.1 | 0.74 | 130 |
| 2.3.2.2 | 0.53 | 37 |
| 2.4.2 | 0.31 | 48286 |
| 2.4.5 | 0.30 | 6031 |
| 2.3.4.3 | 0.25 | 97465 |
| 2.2.6.2 | 0.24 | 161 |
| 2.3.4.2 | 0.21 | 355 |
| 0.2 | 0.01 | 16 |
| 0.3 | -0.16 | 3 |

Pearson correlation for different versions of *bowtie2*

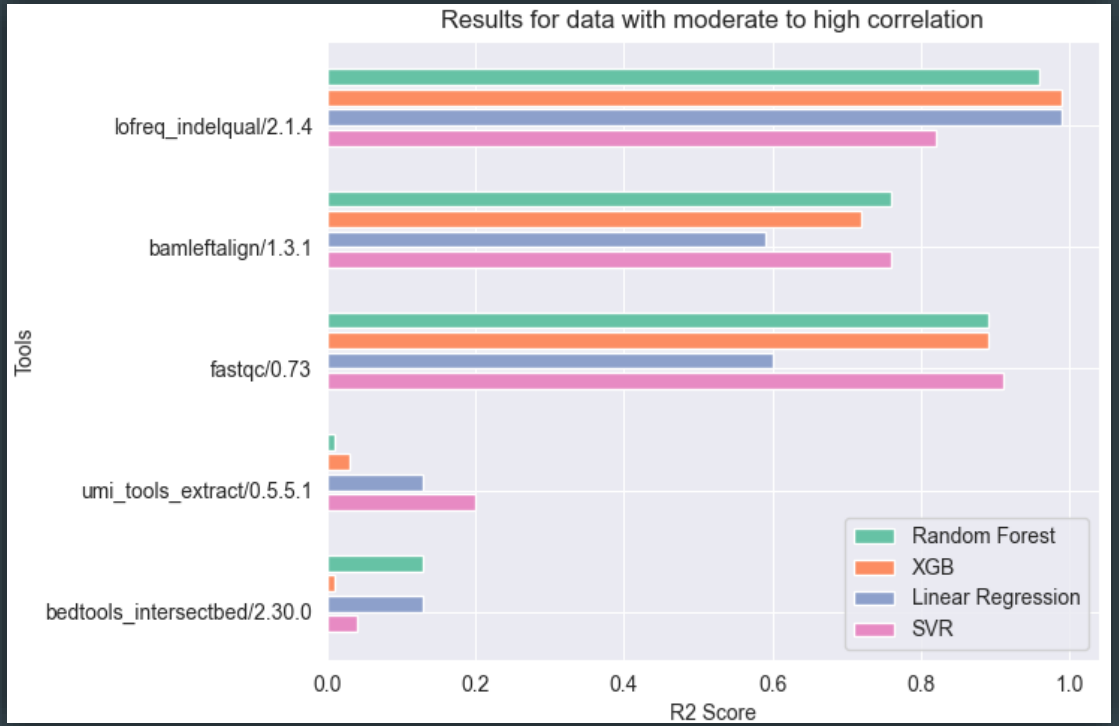
# Moderate-high correlation

| Dataset | Pearson correlation coefficient |
|---|---|
| lofreq_indelqual/2.1.4 | 0.95 |
| bamleftalign/1.3.1 | 0.74 |
| fastqc/0.73 | 0.66 |
| umi_tools_extract/0.5.5.1 | 0.55 |
| bedtools_intersectbed/2.30.0 | 0.35 |

# Results: Moderate-high correlation

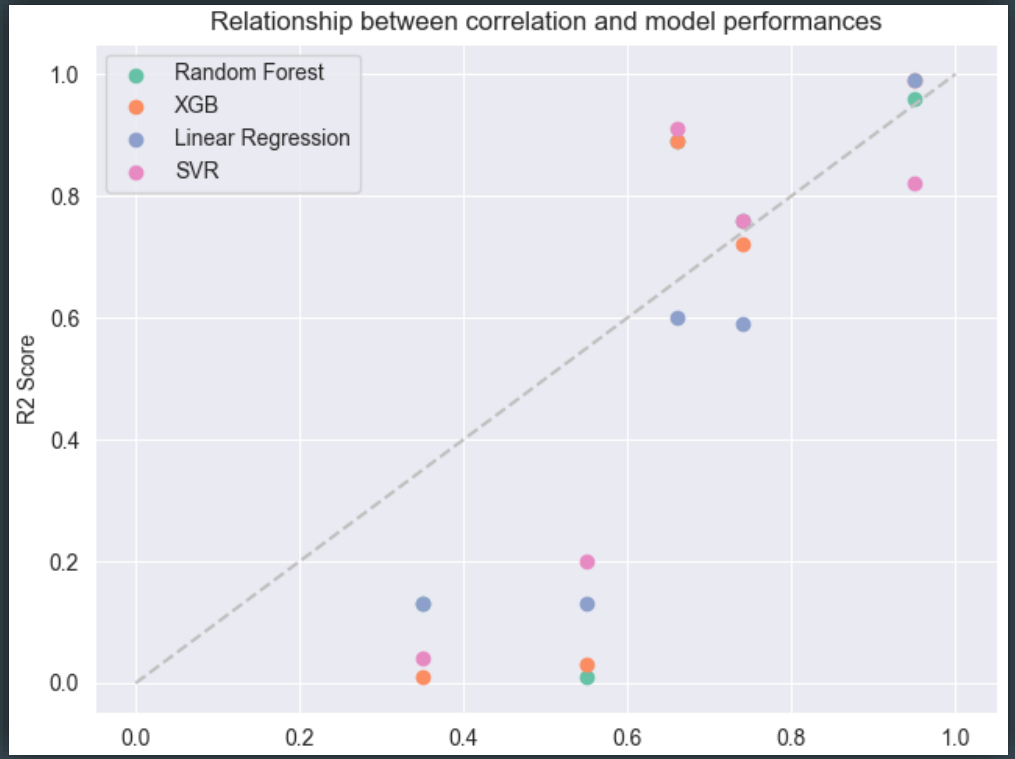
- For datasets with very high correlation
  → all models perform well

- Lower correlation
  → performance decreases



# Results: Moderate-high correlation

- Clear tendency:
  - the greater the correlation, the better all models perform

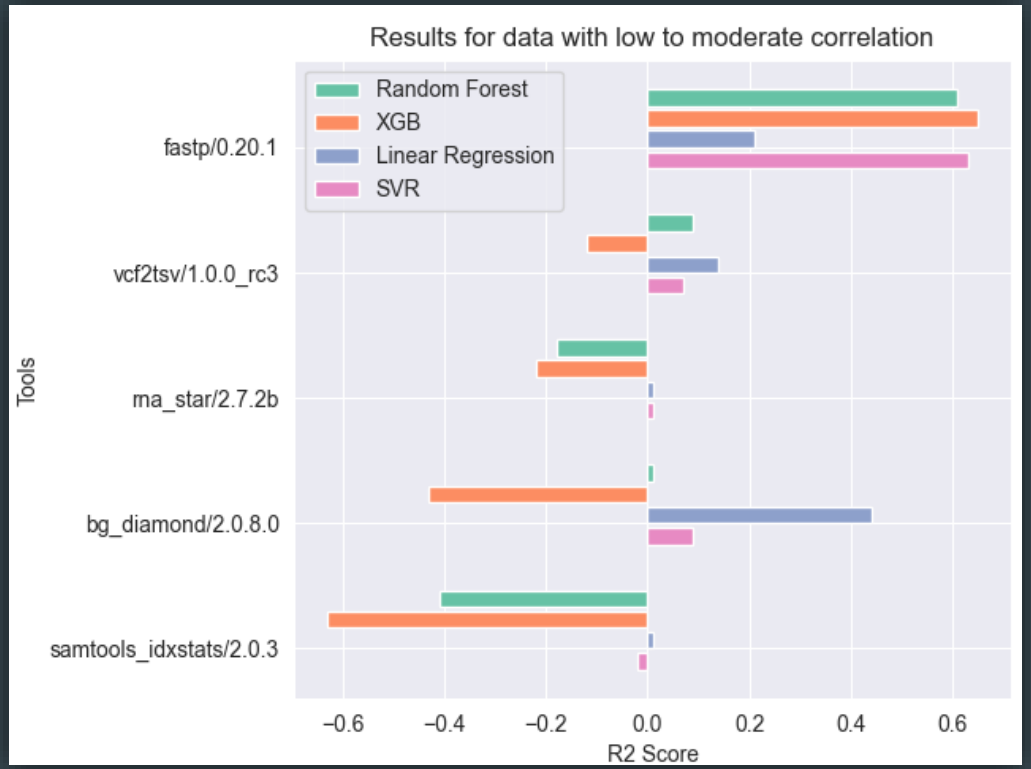


# Low-moderate correlation:

| Dataset | Pearson correlation coefficient |
|---|---|
| fastp/0.20.1 | 0.45 |
| vcf2tsv/1.0.0_rc3 | 0.31 |
| rna_star/2.7.2b | 0.16 |
| bg_diamond/2.0.8.0 | 0.07 |
| samtools_idxstats/2.0.3 | 0.01 |

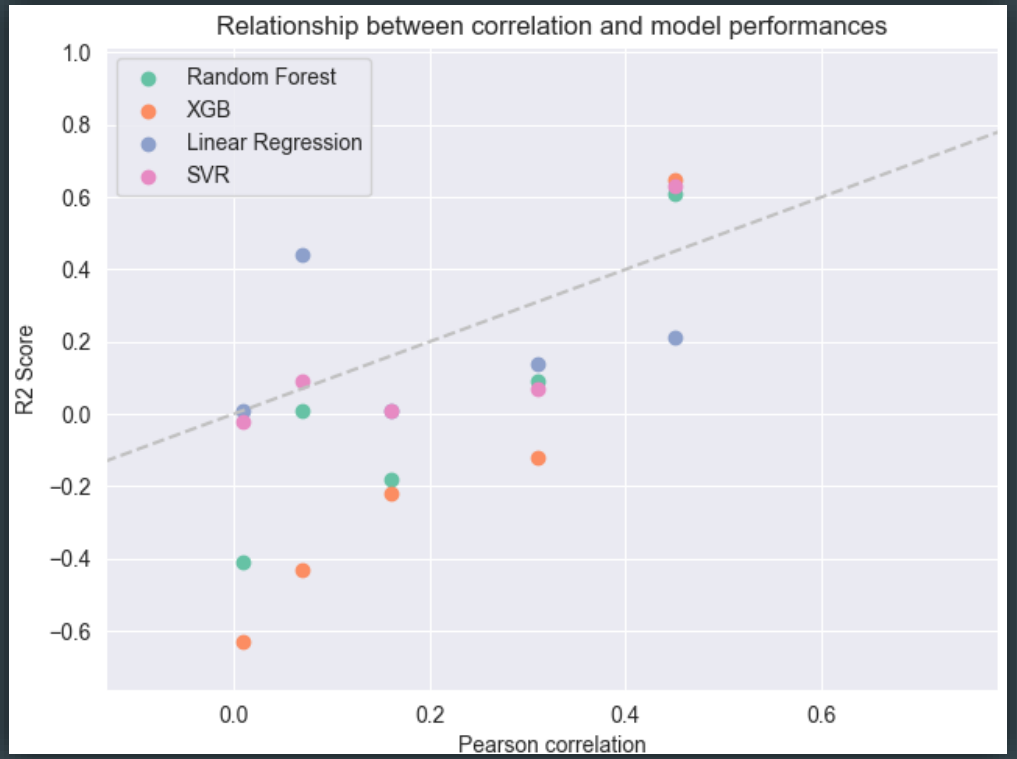
# Results: Low-moderate correlation:

- For datasets with low correlation
  →R2 score varies greatly for all models
- General performance not good
- XGB performs especially bad
  → hyperparameters need to be optimized



# Results: Low-moderate correlation:

- For increasing correlation → clear upward trend in performance
- Performance for *bg_diamond/2.0.8.0* stands out



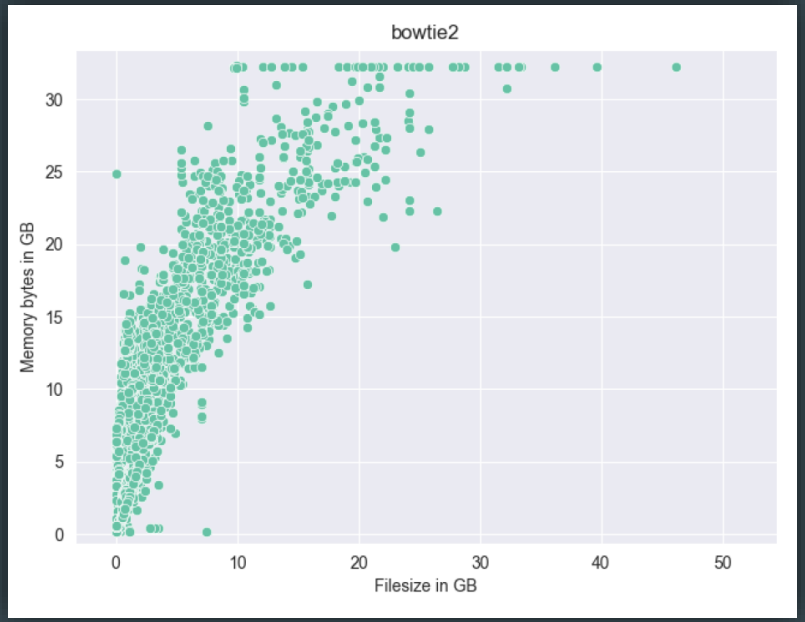
# Results: Low-moderate correlation:

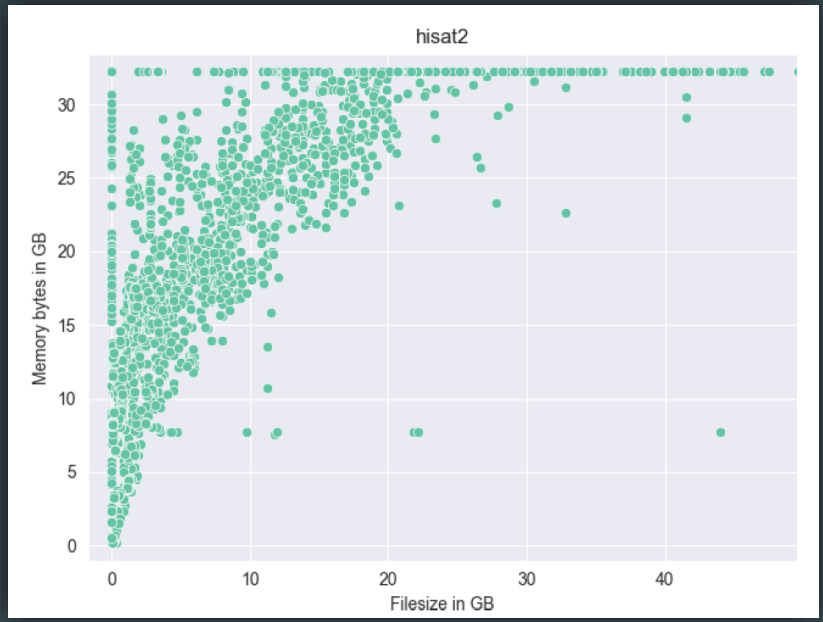
- Models **rely** mostly on *Filesize* feature
- Low correlation → models **perform bad**
- *Slot* is **redundant** most of the time
- **Problematic** because 50% of the tools show correlation < 0.5

| | Feature importance | | |
|---|---|---|---|
| **Dataset** | Filesize | Number of files | Slots |
| fastp/0.20.1 | 0.97 | 0.03 | 0.00 |
| vcf2tsv/1.0.0_rc3 | 1.00 | 0.00 | 0.00 |
| rna_star/2.7.2b | 0.97 | 0.03 | 0.00 |
| bg_diamond/2.0.8.0 | ***0.55*** | ***0.45*** | 0.00 |
| samtools_idxstats/2.0.3 | 1.00 | 0.00 | 0.00 |

# Performance on other datasets

| Dataset | Pearson correlation | Number samples |
|---|---|---|
| hisat2 (Galaxy main) | 0.66 | 2797 |
| bowtie2 (Galaxy main) | 0.62 | 3963 |



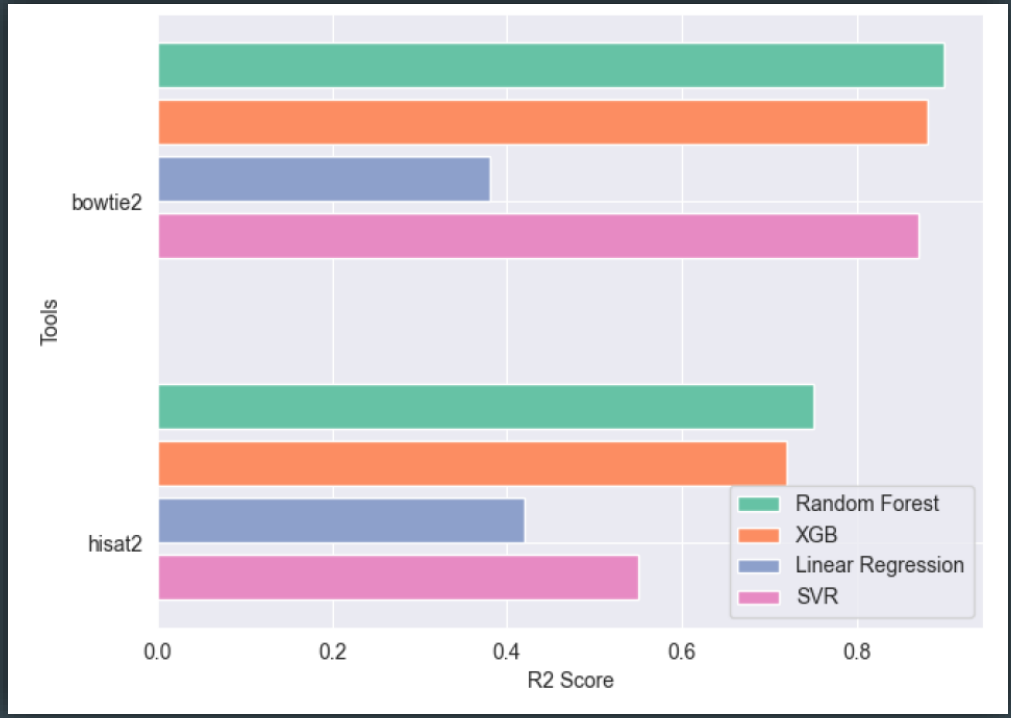



# Results: Performance on other datasets

- Linear Regressor performed **worst**

- Models performed worse for **hisat2** than for **bowtie2**

- Possible reasons:
  - **Fewer** samples (-30%)
  - Data points with **great variance**



# Hyperparameter Optimization

- **Comparison** to Galaxy's current method of memory assignment
- Models need to be **optimized** → HalvingGridSearchCV

| Random Forest | XGB | SVR |
|---|---|---|
| n_estimators = [50, 100, 200, 500] | learning_rate = 8 samples from log_space [0.003 to 0.3] | kernel = ["rbf", "sigmoid", "poly"] |
| max_depth = [None, 4, 16, 32] | n_estimators = [50, 100, 200, 500] | C = [0.01, 0.1, 0.5, 1, 2, 4] |
| min_samples_split = [2, 4, 8] | max_depth = [2, 6, 16, 32, 64] | gamma = ["scale", 0.001, 0.01, 0.1, 1] |
| min_samples_leaf = [2, 4, 8] | | |

# Results: Hyperparameter Optimization

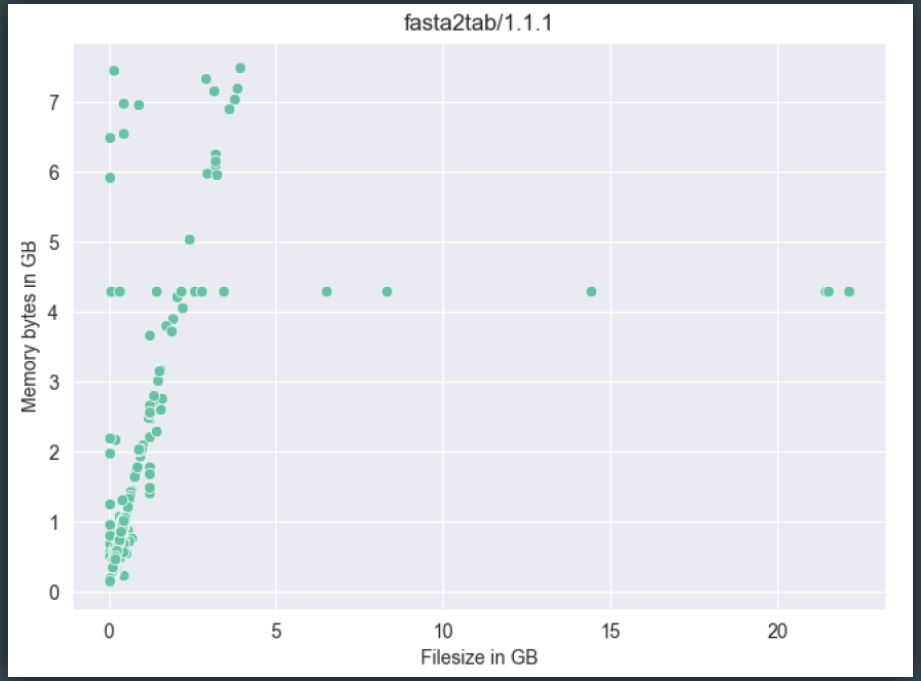
| Group | Tool | Pearson Correlation | Average Memory [GB] |
|---|---|---|---|
| High Correlation | lofreq_indelqual/2.1.4 | 0.95 | 0.21 |
| | bamleftalign/1.3.1 | 0.74 | 0.34 |
| Low Correlation | vcf2tsv/1.0.0_rc3 | 0.28 | 0.14 |
| | rna_star/2.7.2b | 0.20 | 56.72 |
| High Memory | rna_star/2.7.5b | 0.24 | 60.35 |
| | kraken2/2.1.1 | 0.10 | 36.82 |
| Low Memory | fasta2tab/1.1.1 | 0.29 | 0.20 |
| | gmx_sim/2020.4 | 0.21 | 0.50 |

| Dataset | Random Forest | XGB | Linear Regression | SVR |
|---|---|---|---|---|
| lofreq_indelqual/2.1.4 | 0.96 → 0.98 | 0.99 → 0.96 | 0.99 | 0.74 → 0.86 |
| bamleftalign/1.3.1 | 0.76 → 0.79 | 0.72 → 0.78 | 0.59 | 0.77 → 0.77 |
| vcf2tsv/1.0.0_rc3 | 0.09 → 0.15 | -0.12 → 0.07 | 0.14 | -0.01 → -0.04 |
| rna_star/2.7.2b | -0.18 → 0.04 | -0.22 → 0.04 | 0.01 | -0.01 → 0.01 |
| rna_star/2.7.5b | -0.33 → 0.09 | -0.35 → 0.09 | 0.05 | 0.07 → 0.07 |
| kraken2/2.1.1 | -0.09 → 0.18 | -0.02 → 0.17 | 0.08 | -0.01 → -0.04 |
| fasta2tab/1.1.1 | 0.69 → 0.80 | 0.72 → 0.80 | 0.25 | 0.81 → 0.78 |
| gmx_sim/2020.4 | -0.01 → 0.10 | 0.01 → 0.07 | 0.07 | 0.05 → 0.06 |

# Results: Hyperparameter Optimization

- Beside a few outliers

  → **clear pattern** visible



# Analysis on resource consumption

- Using the best model: how much memory can **be saved**?
- **Decrease of 65%/50%** memory overallocation for *rna_star/kraken2*
- **Increase** in percentage of **failed jobs**

| | | Average overallocation | | Failed jobs | |
|---|---|---|---|---|---|
| **Tool** | **Assigned by Galaxy** | ***Galaxy*** | ***RF Model*** | ***Galaxy*** | ***RF Model*** |
| rna_star/2.7.5b | 90 GB | 42.52 GB | 14.59 GB | 12.25% | 23.10% |
| kraken2/2.1.1 | 64 GB | 41.38 GB | 20.90 GB | 25.25% | 44.95% |

# Accuracy-Failure trade-off

- Idea: give algorithms **success probability** of a job
- Prediction interval: prediction is **expected to be** with particular probability
  - → Take **maximum** of interval as prediction
  - → In theory: job is **less likely to fail**

| | Average overallocation | | Failed jobs | |
|---|---|---|---|---|
| **Tool** | ***Galaxy*** | ***RF Model*** | ***Galaxy*** | ***RF Model*** |
| rna_star/2.7.5b | 42.52 GB | 14.59 GB → 28.33 GB | 12.25% | 23.10% → 13.55% |
| kraken2/2.1.1 | 41.38 GB | 20.90 GB → 28.68 GB | 25.25% | 44.95% → 35.3% |

# Discussion and Conclusion

- Existence of **invalid entries** and other characteristics
  - **Validity** of dataset?
- All models benefit from **filtering faulty data**
- For **strong correlation** → accurate and robust prediction
  - Models **rely** mostly on one *Filesize* → too **few features**
- Overallocation of memory **reduced by 50% to 65%**

# Sources