LIVE:

[Multithreading in Python for ML/AI & Multiprocessing]

AppliedAICourse.com

APPLIED COURSE

what we will cover

- Multiprocessing & Threading for Datascience, ML & DL
- code-walkthroughs
- Introductory-Session

What we will not cover

- internals of OS [IPC, PM]
- https://gate.appliedcourse.com/course/5/operating-systems
- software engineering aspects

Popular - Libraries -

sklearn.linear_model.LogisticRegression

→ joblib

class `sklearn.linear_model.` **`LogisticRegression`**(*penalty='l2', *, dual=False, tol=0.0001, C=1.0, fit_intercept=True, intercept_scaling=1, class_weight=None, random_state=None, solver='lbfgs', max_iter=100, multi_class='auto', verbose=0, warm_start=False, n_jobs=None, l1_ratio=None*) [source]

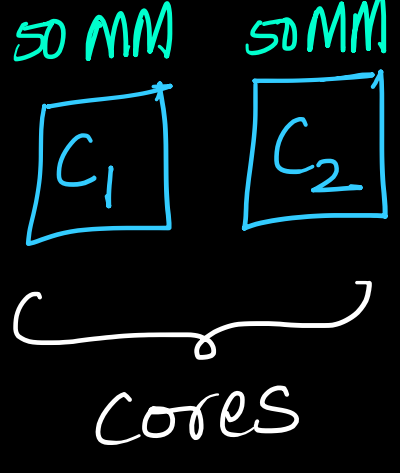
Logistic Regression (aka logit, MaxEnt) classifier.

Numpy & SCipy → BLAS

TensorFlow → multithreading, multicore, GPUs & distributed computing

APPLIED COURSE

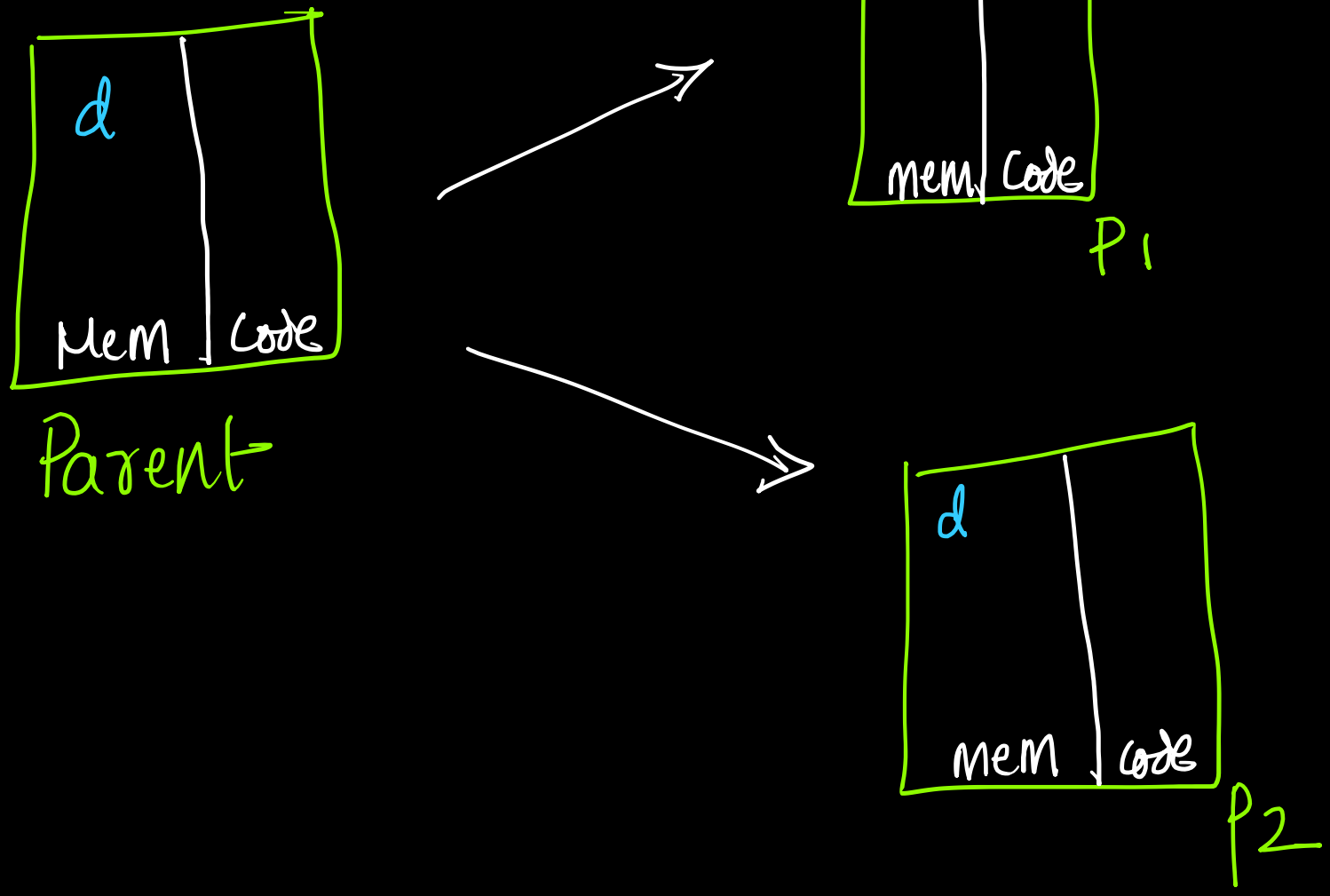
Task: mean of 100 million numbers.

1. 

↳ m_1 & m_2

2. $\frac{m_1 + m_2}{2}$ = mean

Multi-processing [code]

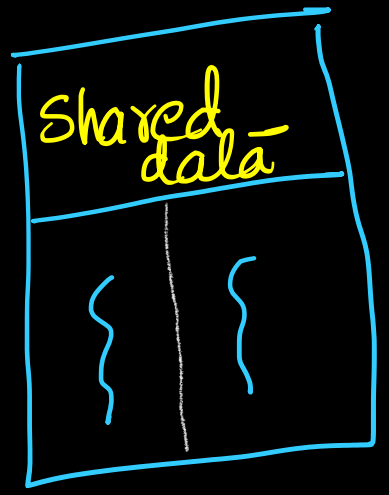


Tons of internals operating Systems

Multithreading:

2 threads per core (Intel)

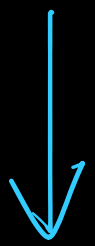
↳ 1 process per core @ a time



- lightweight
- faster context-switching

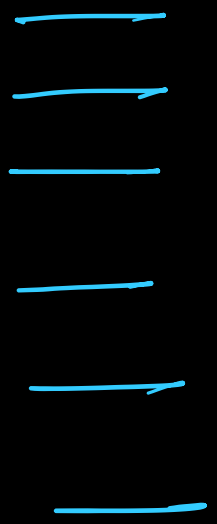
[code]

Combine multiprocessing & Threading



Careful coding

Global Interpreter Lock → worst part of Python MT



Python Interpreter

T_1

T_2

Python's memory-management is not thread-safe

Joblib

https://joblib.readthedocs.io/en/latest/

→ Simple parallel-computing in Python

→ Disk-caching of function outputs [code]

→ Widely used : scikit-Learn

Joblib : Pasallel [code]

→ Only multi-processing

→ can also use multi-Threading [GIT can slow down]

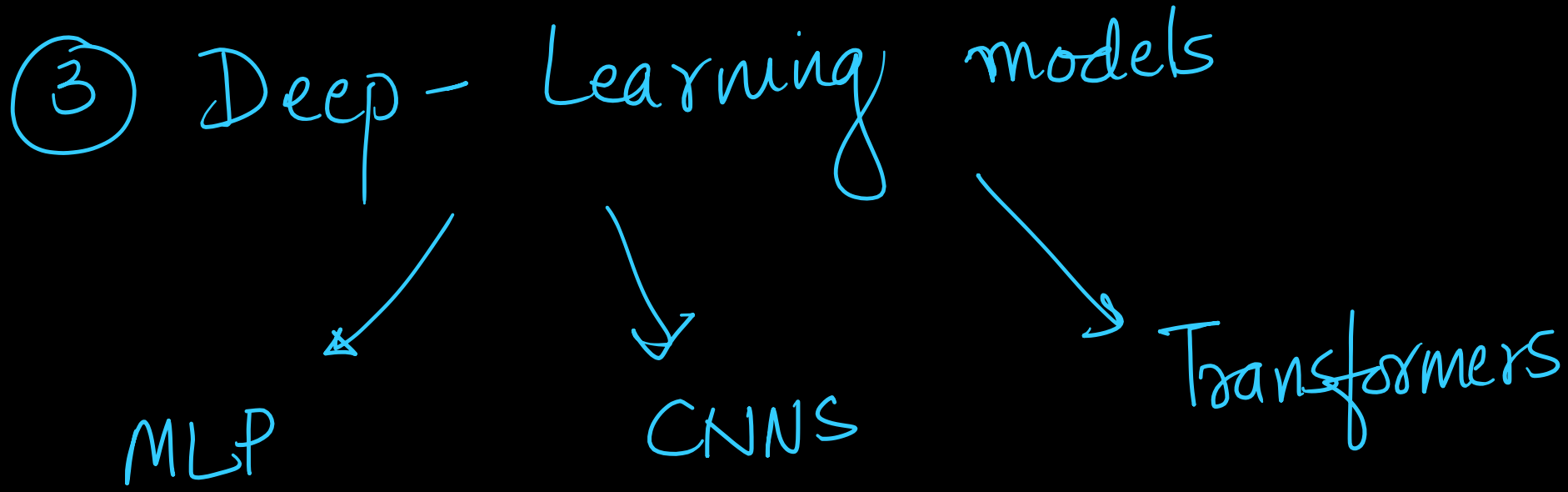
Common Datascience / ML / DL tasks for parallel processing

1. Matrix & Vector products

2. Data preprocessing

② Model-Training

- Logistic Regression
- Decision Trees
- Random Forests
- GBDT



Parallelism for productionization [C/C++/Java]

Logistic Regression

GBDT / Random Forest

Deep Learning