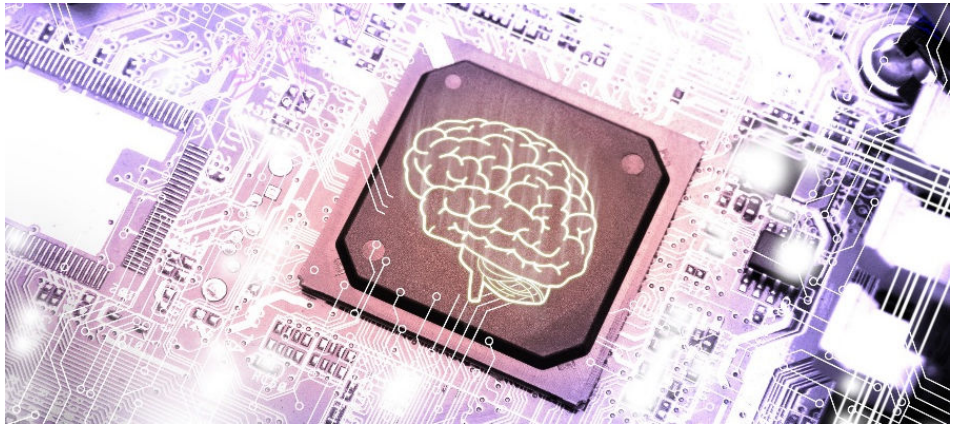
FROM RESEARCH TO INDUSTRY

cea tech



N2D2 : AN OPEN-SOURCE DESIGN ENVIRONMENT FOR DNN

CEA LIST | Nicolas VENTROUX | MPSoC | June 21st, 2017

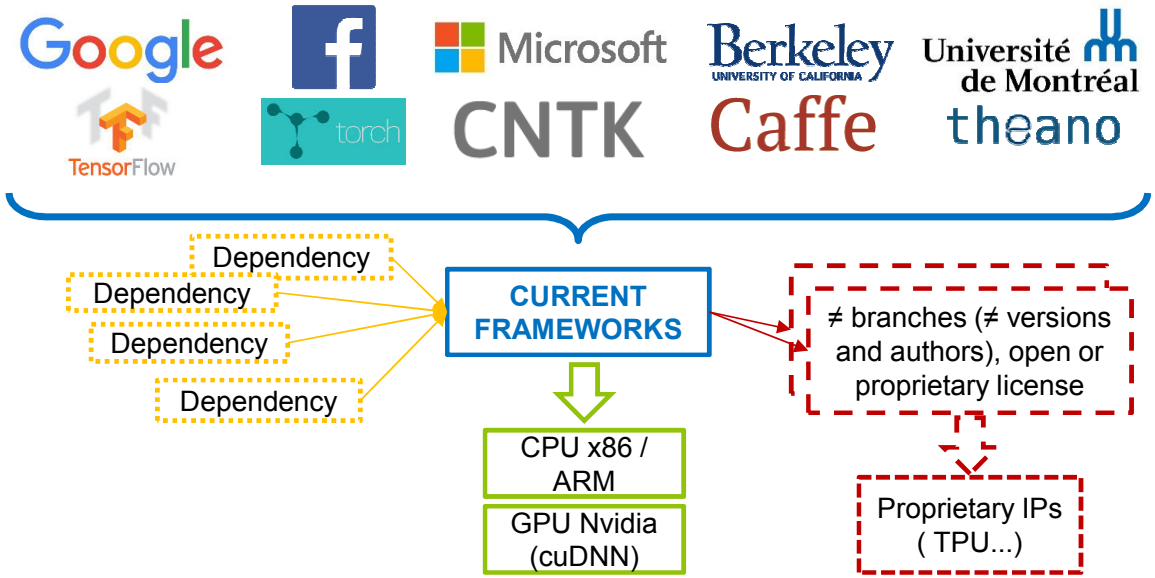
Motivations

- Deep Neural Networks (DNN) are today extremely successful in the vast majority of classification/recognition benchmarks
  - …on high-end multi-250W GPU clusters

- Embedding low-power DNN remains challenging:
  - Must adapt and simplify DNN topologies
    - Reduce layers complexity (number of operations)
    - Reduce precision (8 bit integer)
  - Must balance speed/power and applicative performances

- Need for a framework to automate DNN shrinking exploration and evaluation, performances projection, and porting on embedded platforms

Existing deep learning frameworks



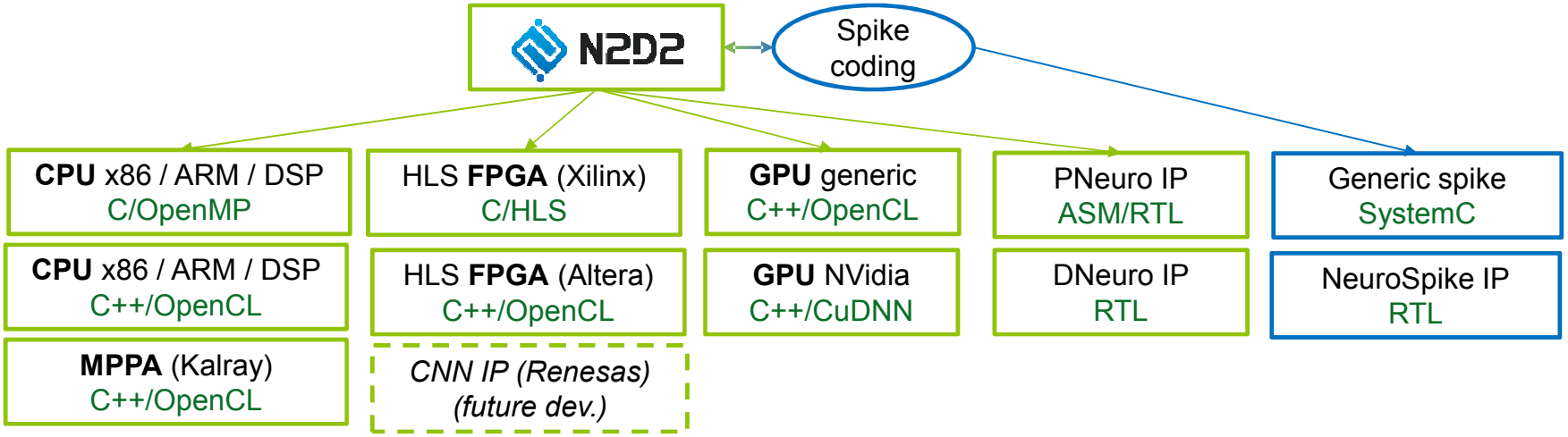
- TensorFlow (Google)
  - Popular and evolving fast, but hardware acceleration closed (TPU)
- Torch (Facebook)
  - Programmed in LUA, uncertain future (Caffe2 by Facebook!)
- Caffe (Univ. Berkeley / support from NVIDIA)
  - Many ≠ versions: NVidia, IBM, research groups (Faster-RCNN…)

N2D2

- A unique platform for the design and exploration of DNN applications
  - Full proficiency of the framework
    - Contributions, algorithms, implementations and code (no dependency except OpenCV, no third party code) in C++
    - No backdoors or Trojan horses
  - Large flexibility
    - Open to developments and specific orientations based on industrial needs
  - Unified modeling and tool flow for both formal and spike coding
  - Explore Deep Neural Network (DNN) topologies with fast simulation and efficient analysis view
    - Experiment state-of-the-art learning techniques with large databases
    - Integrated benchmarking tools (number of computing cycles, memory footprint…)
  - Easily integrate data conditioning by chaining pre-/post-processing transformations
    - Benefit from approximate computing to generate optimized DNN with reduced complexity
    - Data range adaptation tools (for 8 bits integer operations or less)

Hardware exports

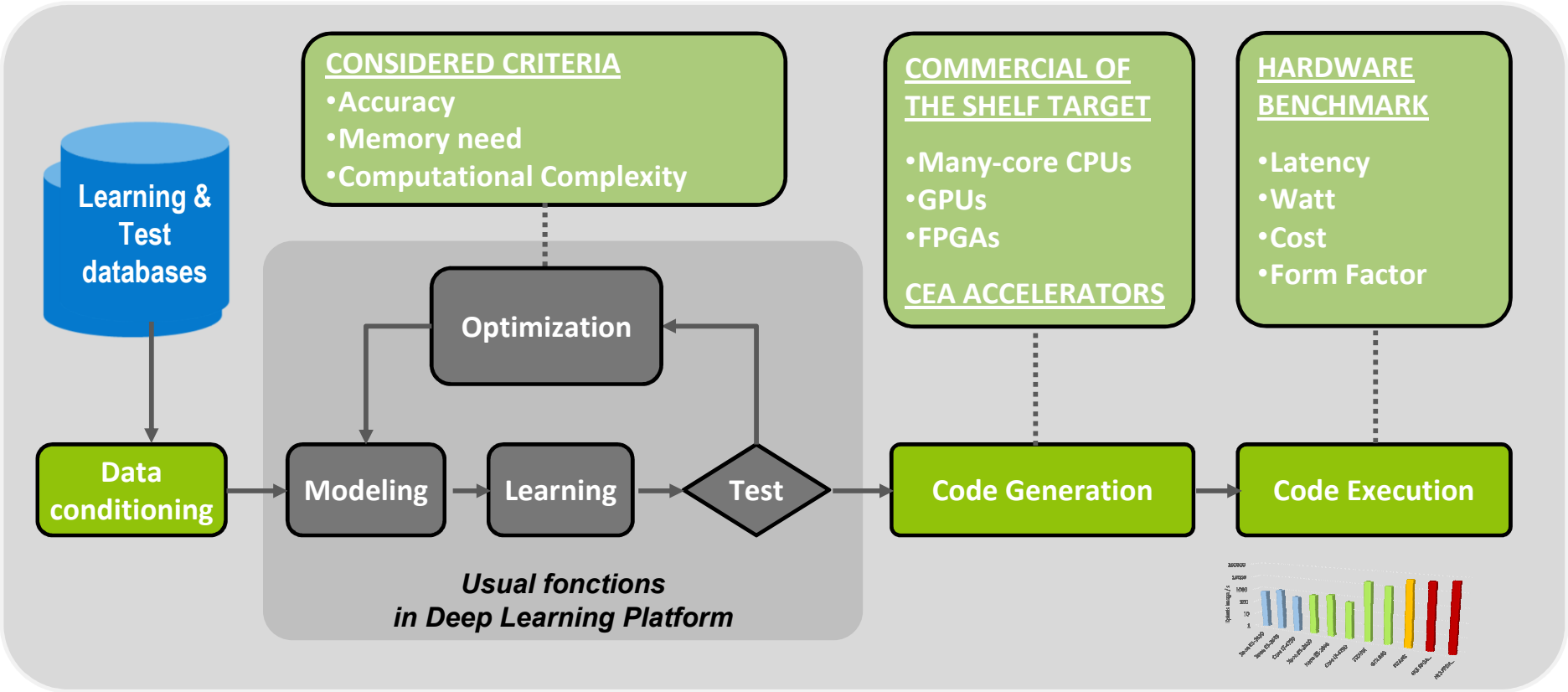
- A unified tool for multiple hardware targets
  - CPUs (C/OpenMP, C++/OpenCL)
    - Multi-/Many-core CPUs: Intel, ARM…
  - GPUs (OpenCL, CUDA, cuDNN)
    - NVidia, AMD, ARM MALI....
  - FPGAs (RTL, HLS)
    - Direct synthesis to FPGA with multiple optimization criteria (timing, throughput, area…)
    - Xilinx, Altera, etc.
  - Specific hardware (PNeuro IP, DNeuro IP and third party) – *under development*



N2D2

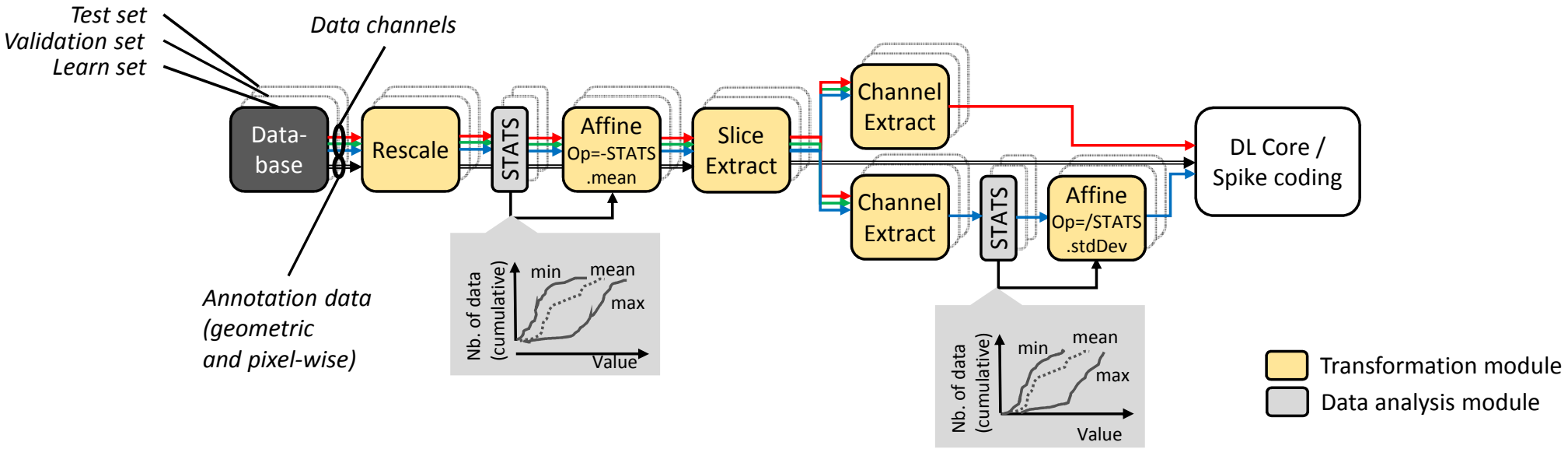
- A unique platform for the design and exploration of DNN applications

N2D2 Specific Advantages



Database conditioning and analysis

- N2D2 integrates data processing and analysis dataflow building
  - Genericity: process image and sound, 1D, 2D or 3D data
  - Associate a label for each data point, 1D or 2D labels
  - Support arbitrary label shapes (circular, rectangular, polygonal or pixel-wise defined)
  - Apply transformations to data, pixel-wise labels and geometrical labels
    - *Basic operations: rescaling, flipping, normalization, affine, filtering, DFT…*
    - *Advanced operations: elastic distortion, random slices/labels extraction, morphological reconstructions…*

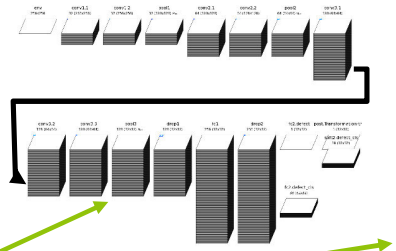


N2D2 typical outputs

N2D2 INI network description file

```
; Database
[database]
Type=MNIST_IDX_Database
Validation=0.2

; Environment
[env]
SizeX=24
SizeY=24
BatchSize=128

[env.Transformation]
Type=PadCropTransformation
Width=[env]SizeX
Height=[env]SizeY

[env.OnTheFlyTransformation]
Type=DistortionTransformation
ApplyTo=LearnOnly
ElasticGaussianSize=21
ElasticSigma=6.0
ElasticScaling=36.0
Scaling=10.0
Rotation=10.0

; First layer (convolutionnal)
[conv1]
Input=env
Type=Conv
KernelWidth=5
KernelHeight=5
NbChannels=6
Stride=2
ConfigSection=common.config

; Second layer (convolutionnal)
[conv2]
Input=conv1
Type=Conv
KernelWidth=5
KernelHeight=5
NbChannels=12
Stride=2
ConfigSection=common.config

; Third layer (fully connected)
[fc1]
Input=conv2
Type=Fc
NbOutputs=100
ConfigSection=common.config

; Output layer (fully connected)
[fc2]
Input=fc1
Type=Fc
NbOutputs=10
ConfigSection=common.config

; Softmax layer
[soft]
Input=fc2
Type=Softmax
NbOutputs=10
WithLoss=1
ConfigSection=common.config

; Common solvers config
[common.config]
WeightsSolver.LearningRate=0.05
WeightsSolver.Decay=0.0005
Solvers.LearningRatePolicy=StepDecay
Solvers.LearningRateStepSize=[sp]_EpochSize
Solvers.LearningRateDecay=0.993
```

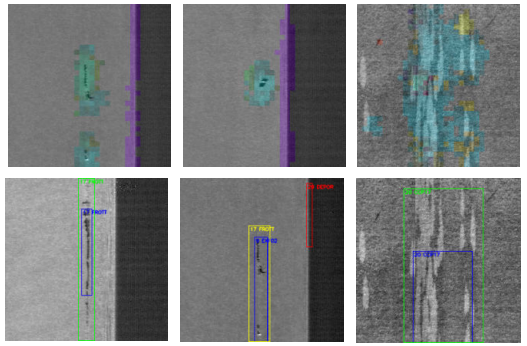
Dataflow visualization

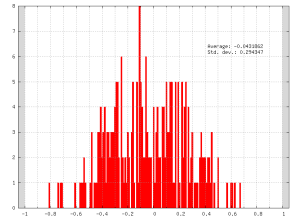


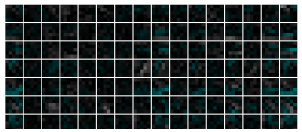
Layer-wise detailed memory and computing requirements

Results visualization:

- Pixel-wise segmentation
- ROI bounding box extraction and classification

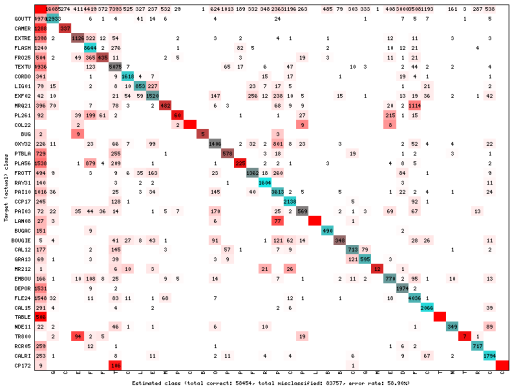






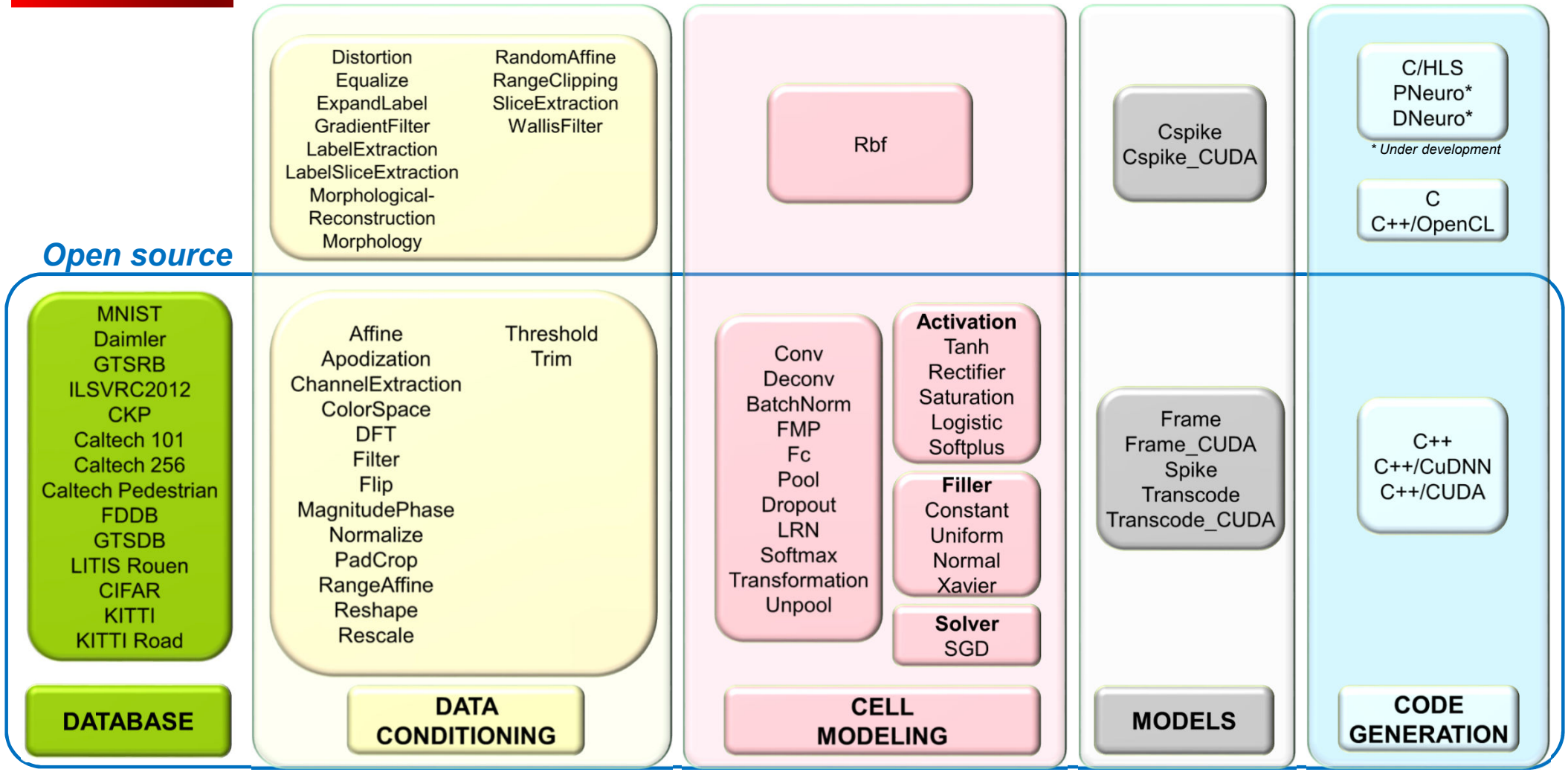
Layer-wise weights and kernels visualization, distribution and data-range analysis

Layer-wise output visualization and data-range analysis

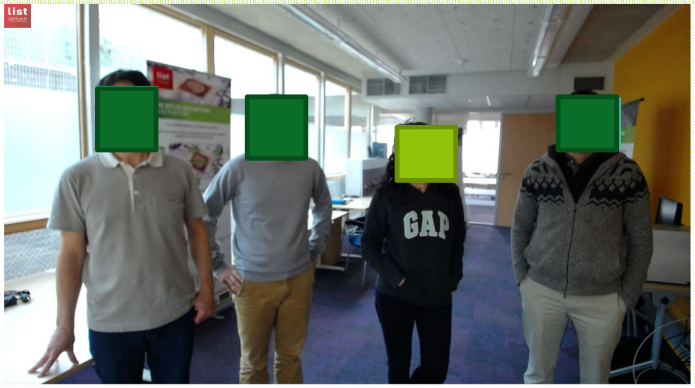


Pixel-wise and object wise confusion matrix reporting

N2D2 is an open-source framework



Try N2D2 NOW!

AppObjectRecognition/	AppFaceDetection/	AppRoadDetection/
Live object recognition application *based on ILSVRC2012 (ImageNet) dataset*	**Live face detection application, with gender recognition** *based on the IMDB-WIKI dataset*	**Simple road segmentation application** *based on the KITTI Road dataset*
		

N2D2 is available at https://github.com/CEA-LIST/N2D2/

- Smallest dependencies and requirements among major frameworks:
  Min requirements: GCC 4.4 or Visual Studio 12 / OpenCV 2.0.0
- Easily extendable with a “plug-and-play” modular system for user-made modules

leti
Centre de Grenoble
17 rue des Martyrs
38054 Grenoble Cedex

list
Centre de Saclay
Nano-Innov PC 172
91191 Gif sur Yvette Cedex