Frederick National Laboratory
for Cancer Research

CANDLE: A Scalable Infrastructure to Accelerate Machine Learning Studies

George Zaki and Andrew Weisman, Frederick National Laboratory for Cancer Research

FAES-BIOINF399, Dec 2nd, 2019

The Future is Supercomputing

"For instance, researchers at ANL, in conjunction with the National Cancer Institute, have developed the ***CANcer Distributed Learning Environment (CANDLE)*** *program to accelerate cancer research and to ultimately tailor treatment plans for individual patients."*

Rick Perry
Secretary of Energy

May, 2018

The White House

https://www.whitehouse.gov/articles/the-future-is-in-supercomputers/

Frederick National Laboratory for Cancer Research (FNLCR)

- **FNLCR is the only Federally Funded Research and Development Center (FFRDC) dedicated exclusively to biomedical research**
  - Operated in the public interest by **Leidos Biomedical Research, Inc** (formerly SAIC-Frederick) on behalf of the National Cancer Institute
- **Main campus located on 70 acres at Ft. Detrick, MD**
  - Leidos Biomed employees co-located with NCI researchers and other contractors on the NCI Campus at Frederick
  - Additional Leidos Biomed scientists at Bethesda and Rockville sites

Mission

Provide a unique national resource for the development of new technologies and the translation of basic science discoveries into novel agents for the prevention, diagnosis and treatment of cancer and AIDS.

Research & Development at FNLCR

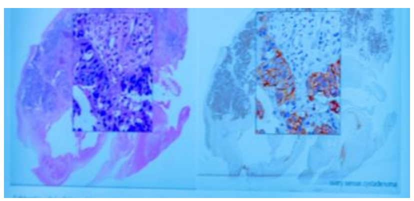
- **Research & Development**
  - **Basic Research**: New knowledge about AIDS and cancer
  - **Applied R&D**: New diagnostics and therapeutics
  - **Clinical Research:** Clinical trials and laboratory analysis
  - **cGMP manufacturing:** Biologicals and vaccine production

- **Specialties**
  - Genomics, proteomics, and metabolomics
  - Bioinformatics and imaging
  - Nanotechnology
  - Animal models
  - Tumor cell biology and virology
  - Immunology and inflammation

- ***Data science key to enabling R&D activities and specialties***

Biomedical Informatics and Data Science Directorate @ FNLCR

Leverage leading edge data science and enabling technologies skills, tools, and capabilities to accelerate translation of biomedical data to scientific discoveries, medical treatments, diagnostic and prevention tools for cancer and AIDS patients.



Analyze

Decide

Data

Insight

Action

Descriptive Analysis
What has happened?

Predictive Analysis
Why did it happen? What will happen?

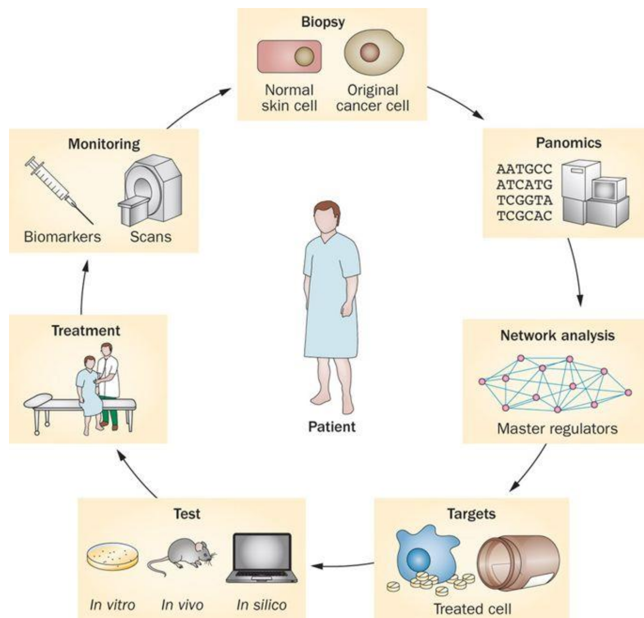
Prescriptive Analysis
What should we do?

HPC Enabling Precision Medicine

Cancer Knowledge

Patient Profile

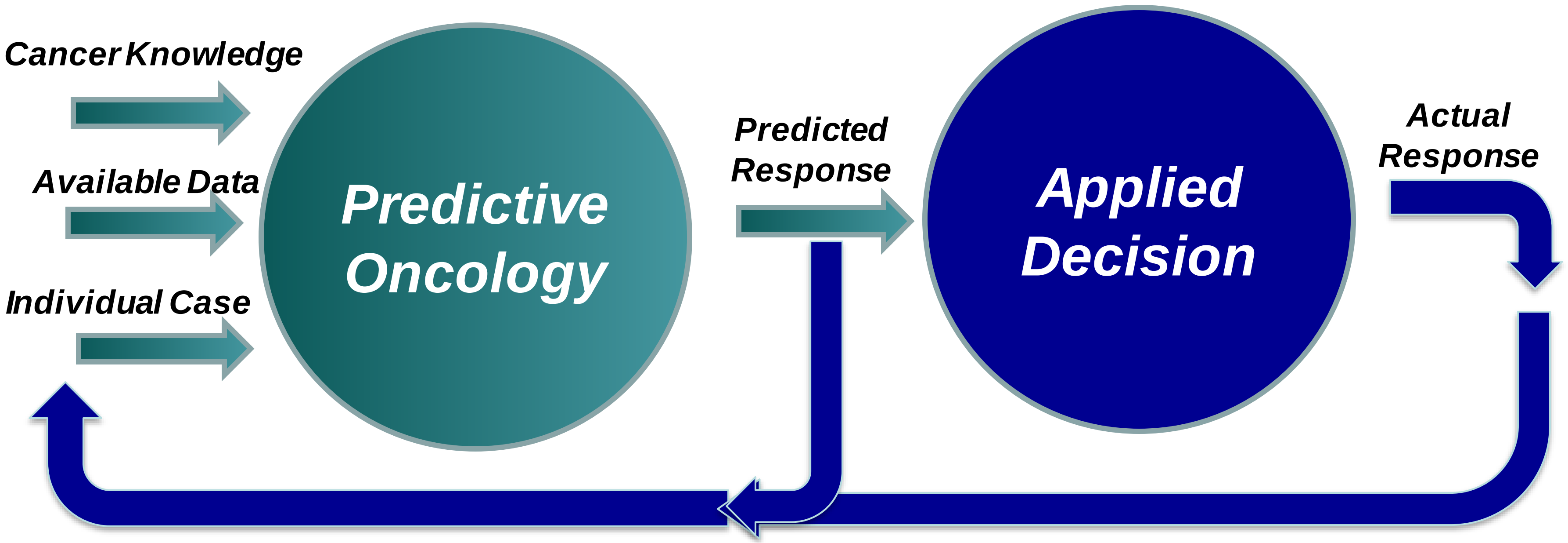
Available Data



nature.com

Predicted Outcome

Oncology Learning System



Descriptive Analysis
What has happened?

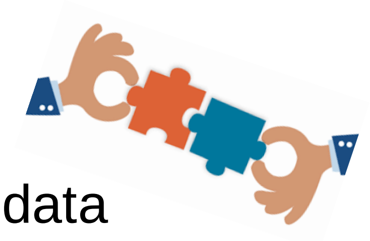
Predictive Analysis
Why did it happen? What will happen?

Prescriptive Analysis
What should we do?

Challenge Areas for Predictive Oncology

- Challenges for cancer
  - Insufficient data for describing all possibilities

    - Over 250,000 unique cancer characterizations
    - Observation gaps – absence of specific confirming data
    - Bridging molecular with preclinical and preclinical to clinical domains

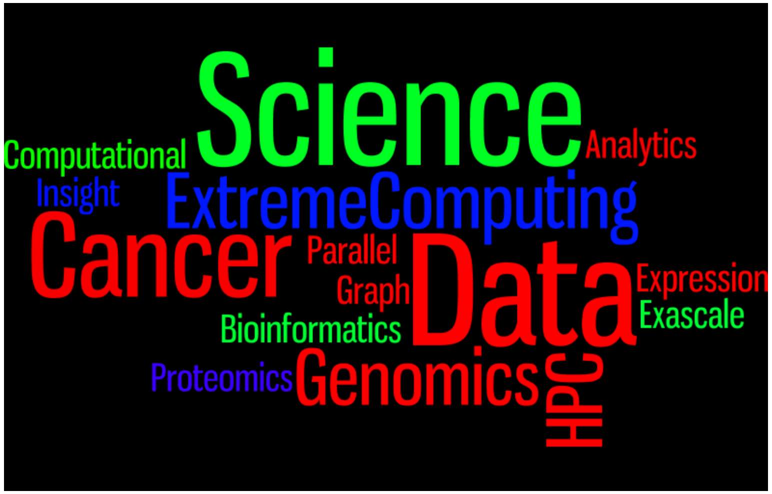


  - Data fusion and scientific credibility
    - Achieving coherence across scales and types of data
    - Achieving coherence and quality across organizations
  - Achieving reliability

    - Consistency of response for characterized conditio
    - Accounting for uncertainty of unknown factors
    - Similarity of behavior across similar models

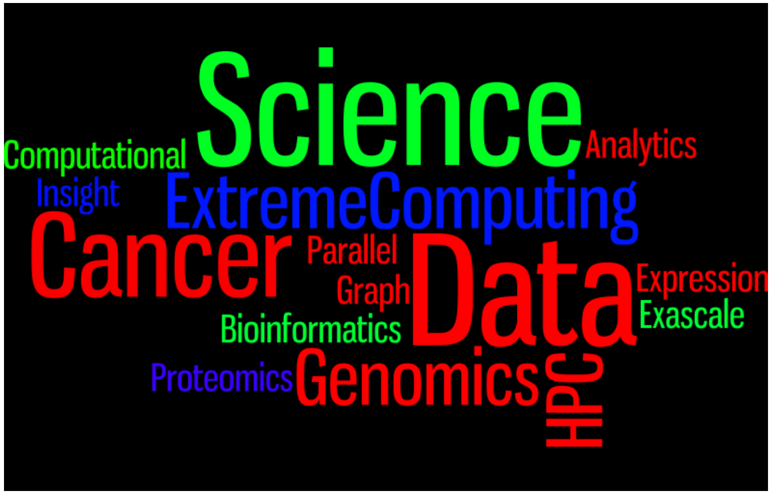
Example Biomedical Informatics and Data Science Projects and Programs

- Cancer Research Data Commons
- Clinical Trials Reporting Program
- Molecular Analysis for Therapy Choice (MATCH)
- Pediatric MATCH
- Joint Design of Advanced Computing Solutions for Cancer
- Accelerating Therapeutics for Opportunities in Medicine (ATOM)
- Systems Biology Cube
- BiodbNet
- Cancer Distributed Learning Environment (CANDLE)



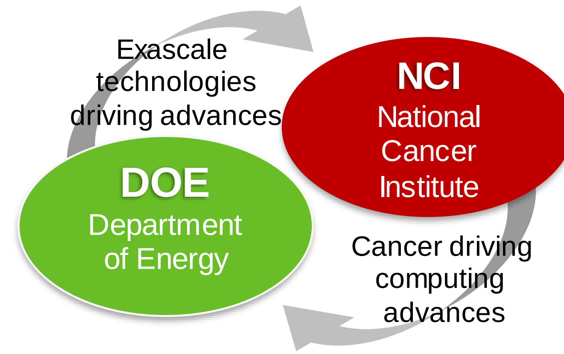
Example Biomedical Informatics and Data Science Projects and Programs

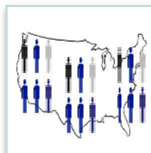
- Cancer Research Data Commons
- Clinical Trials Reporting Program
- Molecular Analysis for Therapy Choice (MATCH)
- Pediatric Match
- Joint Design of Advanced Computing Solutions for Cancer
- Accelerating Therapeutics for Opportunities in Medicine (ATOM)
- Systems Biology Cube
- BiodbNet
- Cancer Distributed Learning Environment (CANDLE)



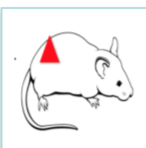
JDACS4C NCI-DOE Collaboration

- **Shared Interests**
  - Cancer scientific challenges driving advances in computing
  - Exascale technologies driving cancer advances
- **Three Pilot Efforts:**



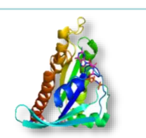


Clinical Domain – Precision oncology surveillance
Expanded SEER database information capture
Modeling patient health trajectories



Pre-clinical Domain – Improved predictive models
Computational/hybrid predictive models of drug response
Improved experimental design

250,000 cancer types



Molecular Domain – Multiscale biological models
Models for RAS-RAS complex interactions
Insight into RAS related cancers

1000s of drugs, millions of combinations

4 Billions core hours per simulation

Joint Design of Advanced Computing Solutions for Cancer

JDACS4C

Exascale technologies driving advances

NCI
National Cancer Institute

DOE
Department of Energy

Cancer driving computing advances

Initiatives Supported NSCI and PMI

NIH NATIONAL CANCER INSTITUTE

Argonne NATIONAL LABORATORY

OAK RIDGE National Laboratory

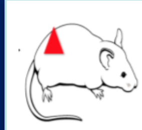
Lawrence Livermore National Laboratory

Los Alamos NATIONAL LABORATORY EST. 1943

Frederick National Laboratory for Cancer Research
sponsored by the National Cancer Institute

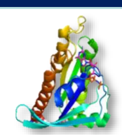
Integrated Precision Oncology

Molecular | Pre-clinical | Population



Pre-clinical Domain – Improved predictive models
Computational/hybrid predictive models of drug response
Improved experimental design

Clinical Domain – Precision oncology surveillance
Expanded SEER database information capture
Modeling patient health trajectories

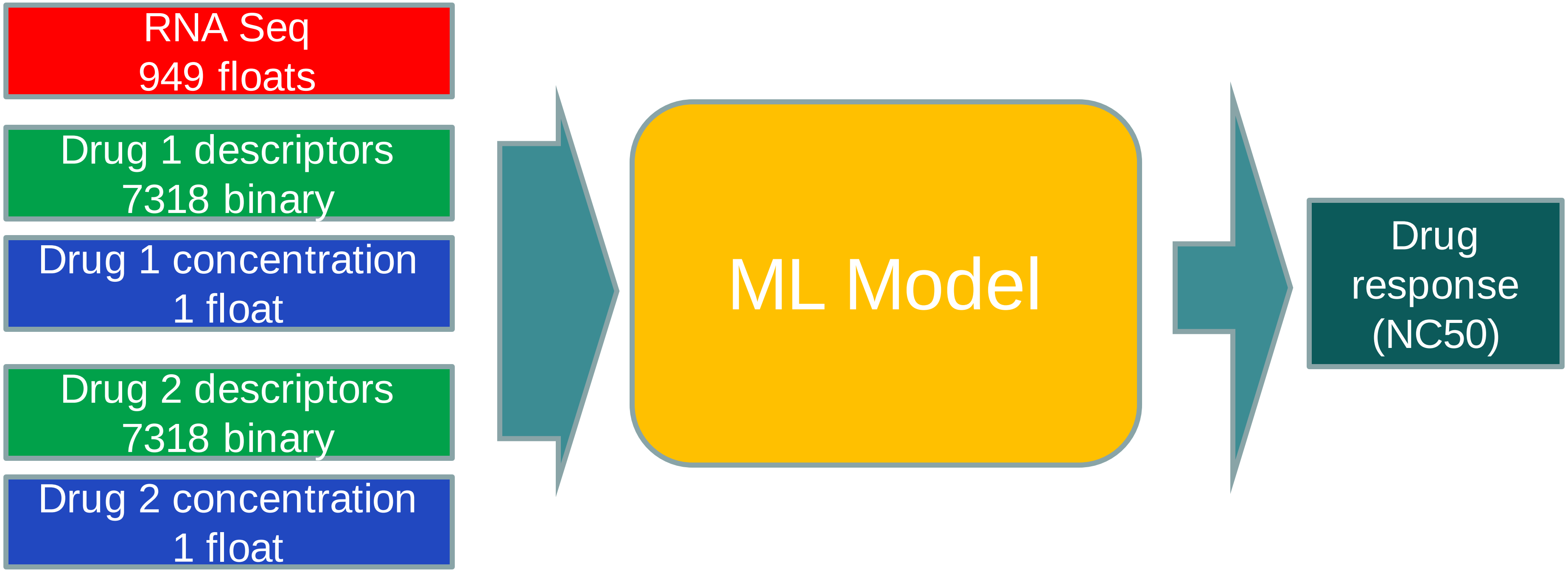


Molecular Domain – Multiscale biological models
Models for RAS-RAS complex interactions
Insight into RAS related cancers

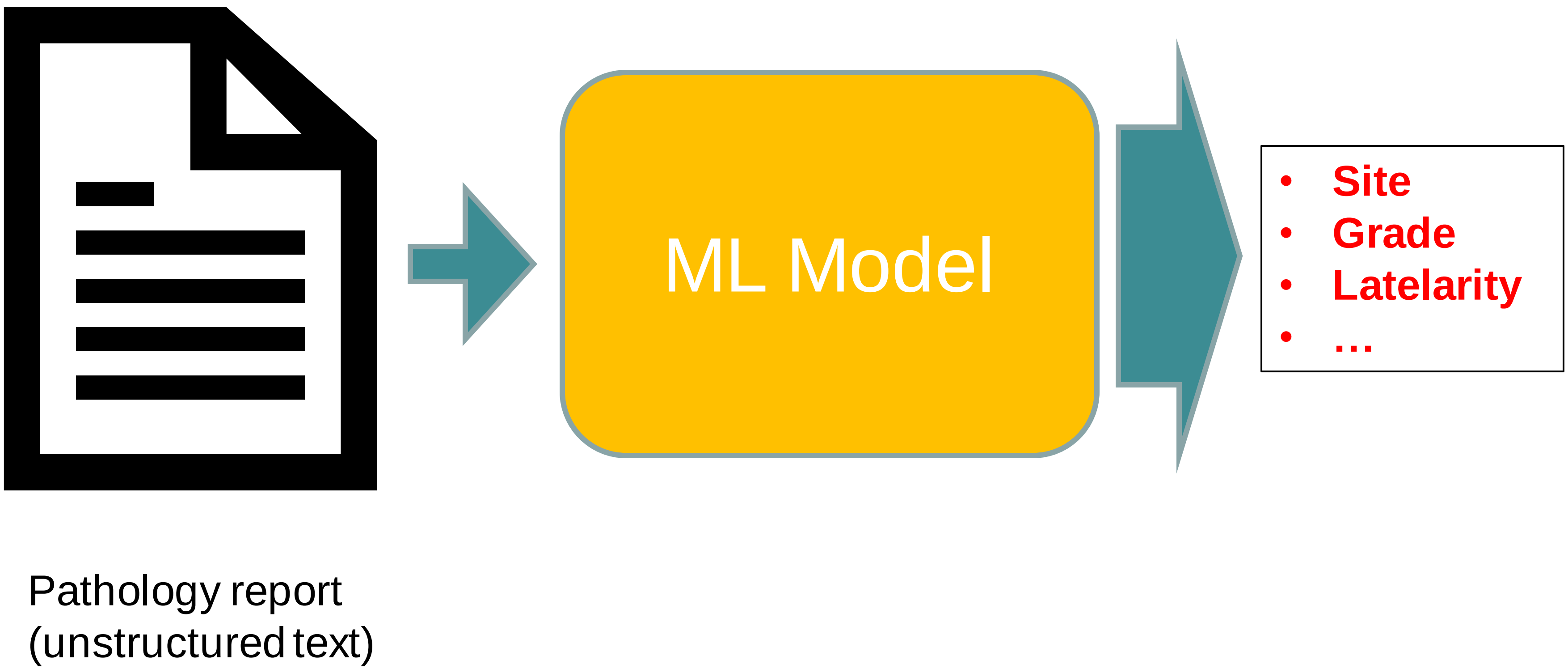
CANcer Distributed Learning Environment (CANDLE)
Scalable Deep Learning for Cancer

JDACS4C established June 27, 2016 with signed MOU between NCI and DOE

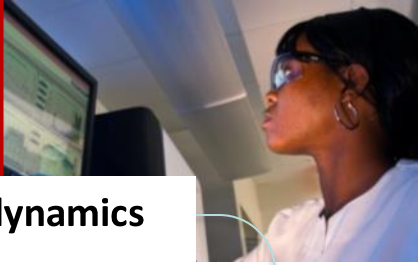
Pilot 1 Example: Drug Response Prediction



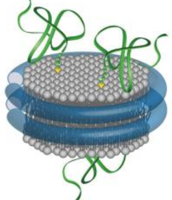
Pilot3 Example: Pathology Report Multitask Classifier



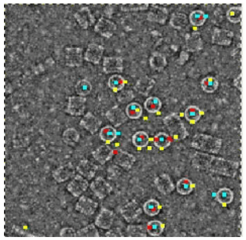
RAS proteins in membranes



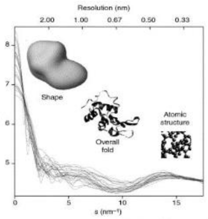
RAS activation experiments at NCI/FNL



Experiments on nanodisc



CryoEM imaging



X-ray/neutron scattering

Multi-modal experimental data, image reconstruction, analytics

Protein structure databases

New adaptive sampling molecular dynamics simulation codes

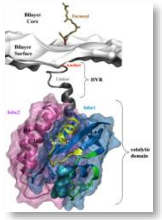
Adaptive time stepping

Phase Field | Coarse-Grain MD | Classical MD

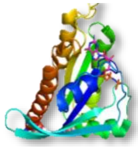
Adaptive spatial resolution

High-fidelity subgrid modeling

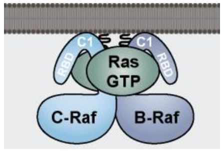
Predictive simulation and analysis of RAS activation



Granular RAS membrane interaction simulations

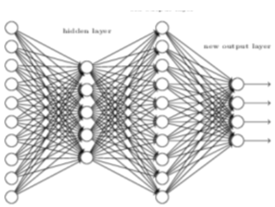


Atomic resolution sim of RAS-RAF interaction

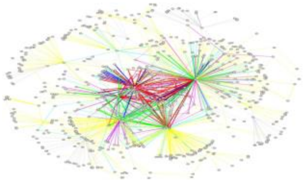


Inhibitor target discovery

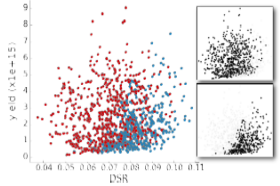
Machine learning guided dynamic validation



Unsupervised deep feature learning



Mechanistic network models

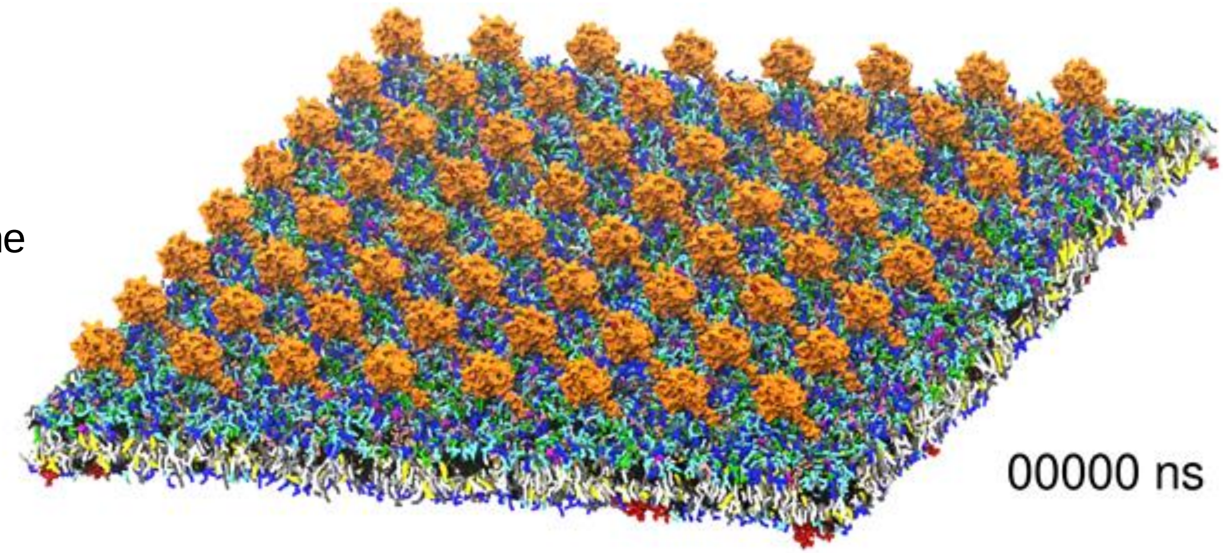


Uncertainty quantification

KRAS4b in plasma membrane – MD simulation

- 20,000 lipids (70x70 nm)
- 40 µs pre-equilibration
- 64 Ras proteins cluster readily

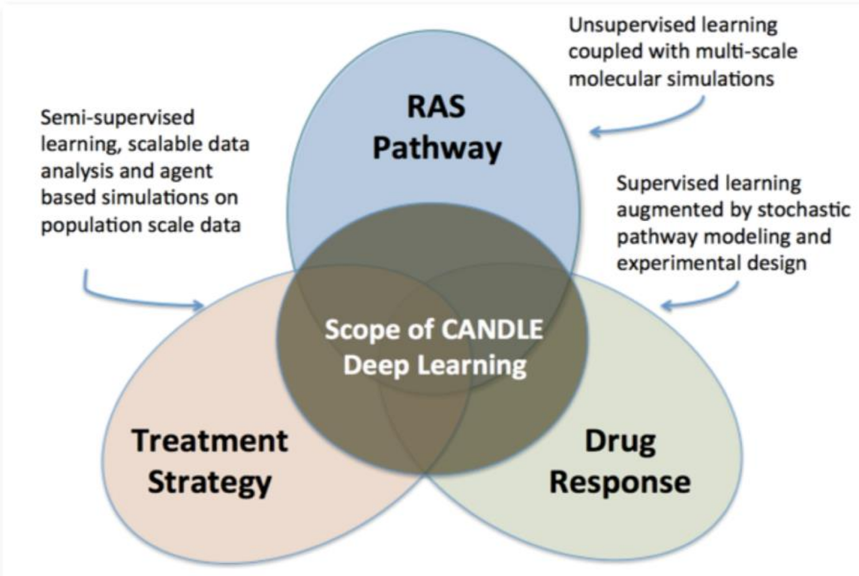
- Associates with and aggregates charged lipids in the membrane



Helgi Ingólfsson, LLNL

CANDLE – Deep Learning Across JDACS4C

ECP-CANDLE Project : CANcer Deep Learning Environment



CANDLE Goals

Develop an exscale deep learning environment for cancer

Building on open source Deep learning frameworks

Optimization for CORAL and exascale platforms

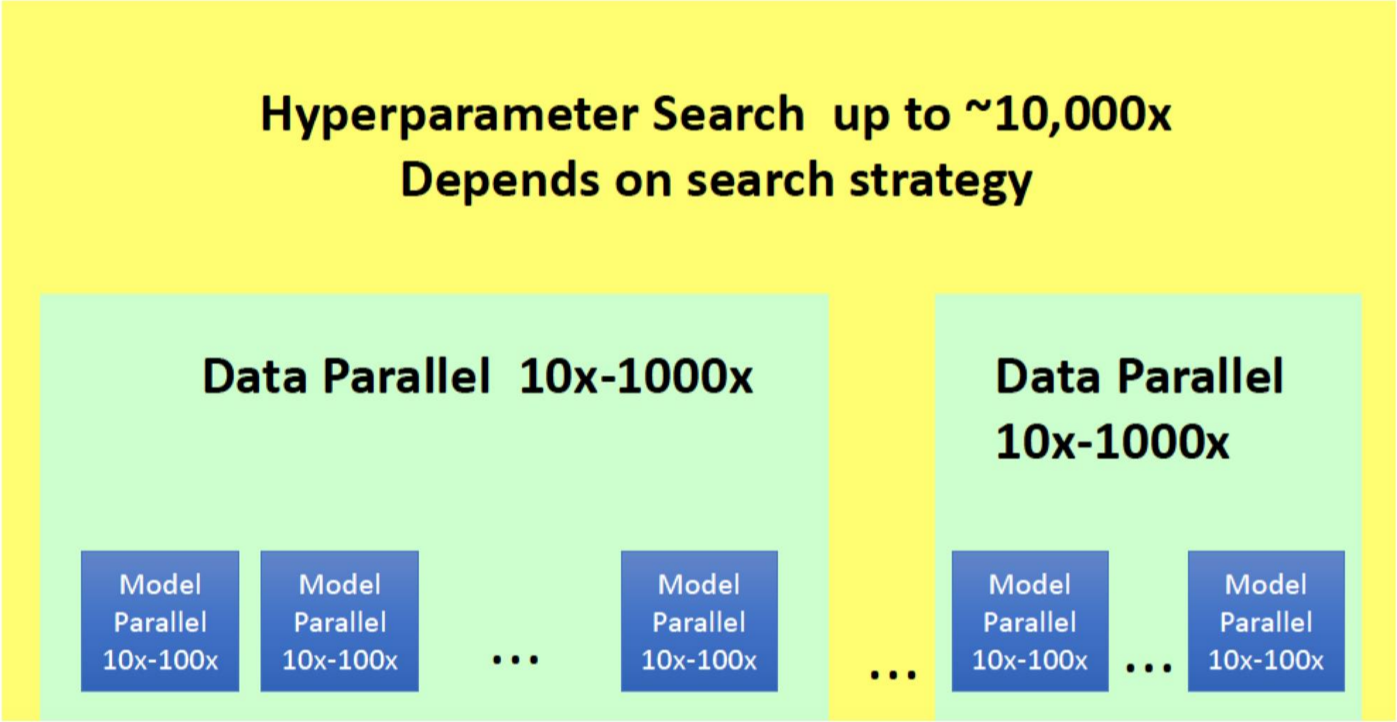
Support all three pilot project needs for deep dearning

Collaborate with DOE computing centers, HPC vendors and ECP co-design and software technology projects

CANDLE - Multi-level Parallelism on HPC Systems

Parallelism Targets in CANDLE

10,000 x 10-1000 x 10-100 = 1M – 1000M “cores”



Hyper-parameter Optimization (HPO)

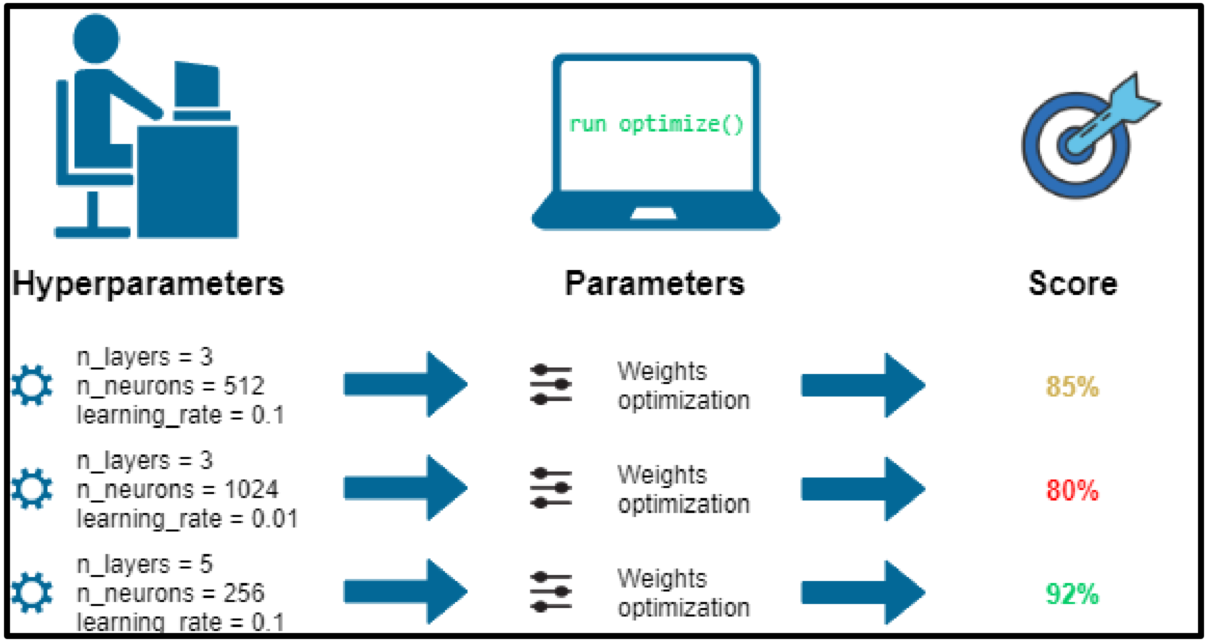
- Many empirical studies do not give a good direction for insight to build knowledge.
- **Hyper-parameter search** is very important once you get something that basically works.
- Many recent incremental advances can reproduce the same result as prior art if a *good* hyper-parameter search in deep learning research is used.

WINNER'S CURSE?
ON PACE, PROGRESS, AND EMPIRICAL RIGOR

D. Sculley, Jasper Snoek, Ali Rahimi, Alex Wiltschko
{dsculley, jsnoek, arahimi, alexbw}@google.com
Google AI

What are hyperparameters?

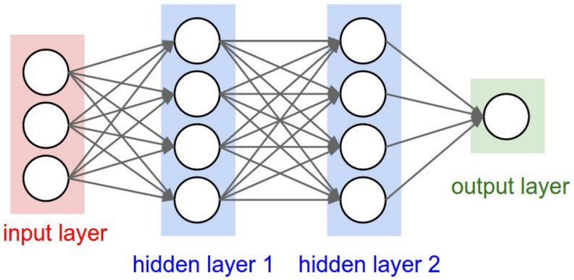
- Parameters of your system with no straightforward method on how to set their values:
  - Usually set before learning process
  - Is not directly estimated from the data

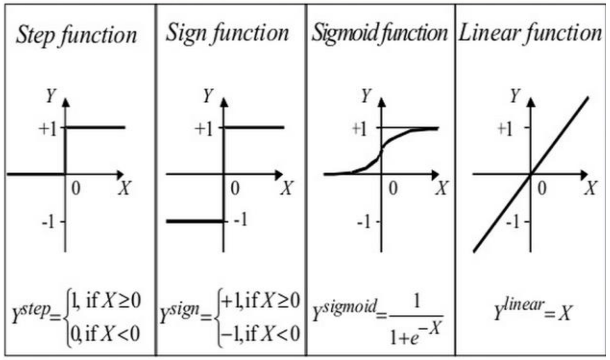


deepai.org

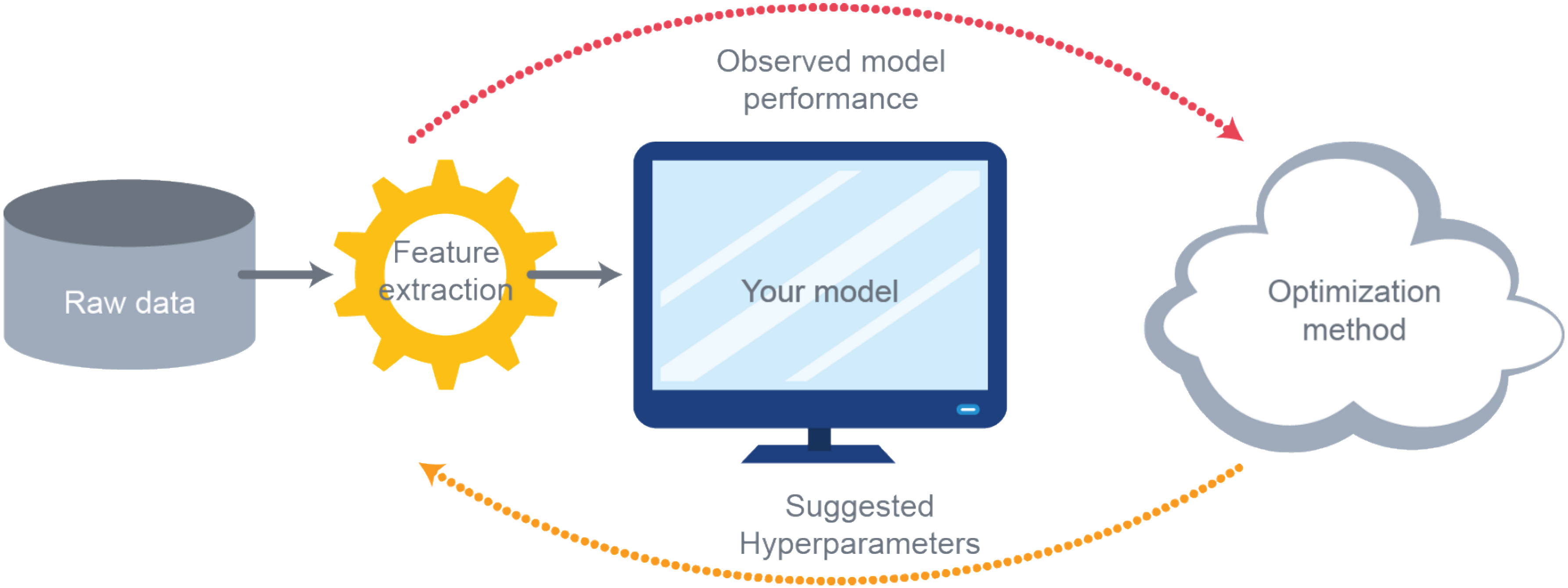
Examples of Hyperparameters

- The depth of a decision tree
- Number of trees in a forest
- Number of hidden layers and neurons in a neural network,
- Degree of regularization to prevent overfitting
- K in K-means
- Learning rate schedule in Stochastic Gradient Descent (SGD)
- ….



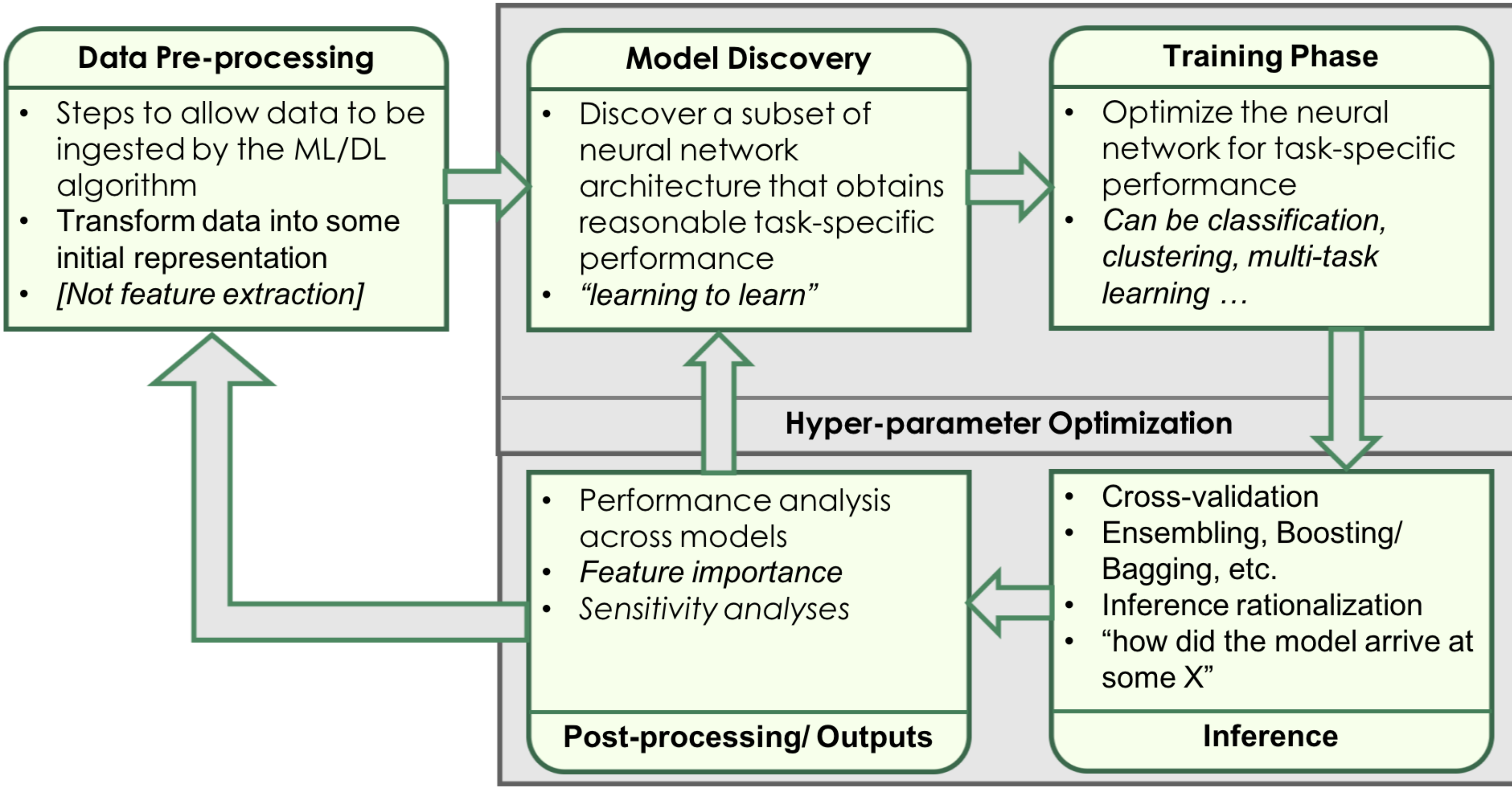


Generalized Machine Learning Workflow



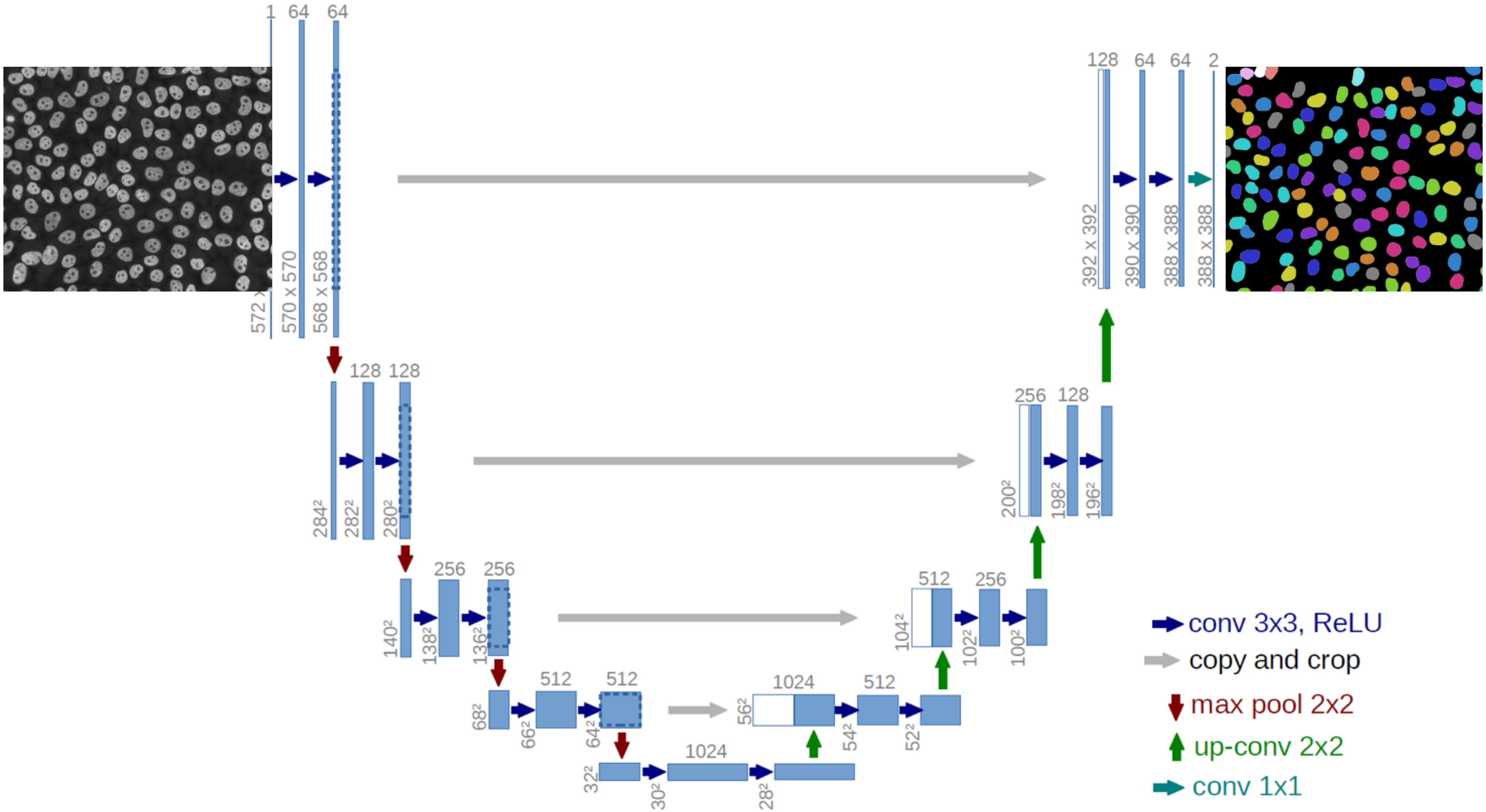
https://sigopt.com/blog/common-problems-in-hyperparameter-optimization/

Generalized Machine Learning workflow



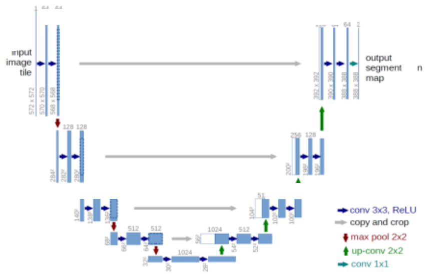
https://github.com/ECP-CANDLE/Tutorials/tree/master/2019/ECP

Evaluation: HPO for U-Net



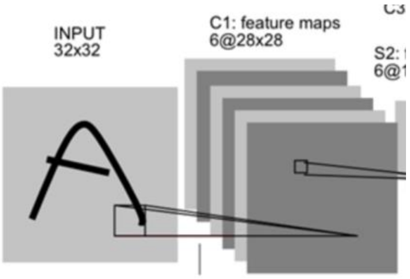
Source: Ronneberger et. al,., MICAAI, 2015

U-Net Hyper parameters



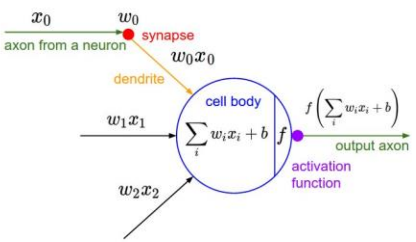
ONLY 2 Levels of Max-Pooling

N_layers = {2,3,4,5}



How many convolution filters?

Num_filters= {16,32,64}



What is the activation function?

Activation= {relu, softmax, tanh

0	0	0
1	1	1
0	0	0

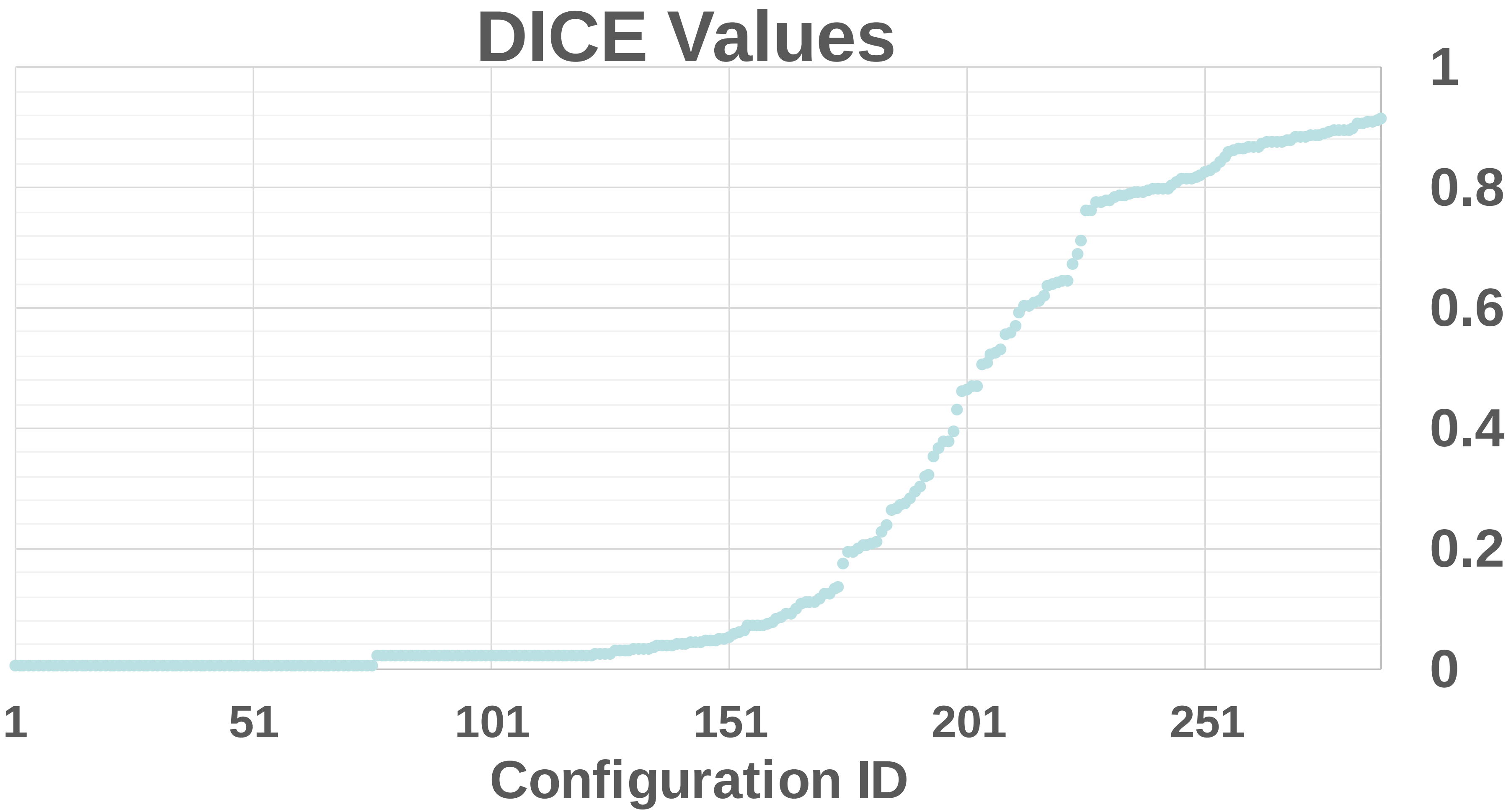
Size of conv filter?

Filter_size = {3x3, 5x5}

Drop out some results to avoid overfitting?

Drop_out = {0, 0.2, 0.4, 0.6, 0.8}

Hyper parameters sweep



WHAT IS HYPERPARAMETER OPTIMIZATION

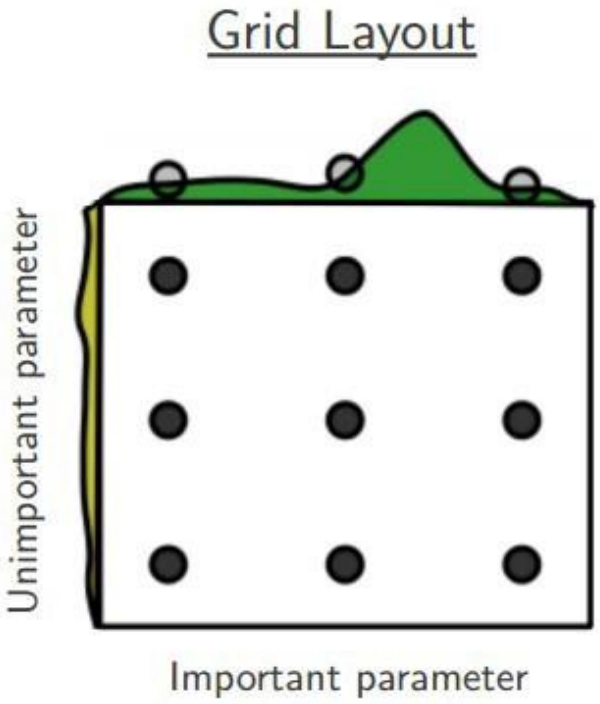
Hyperparameter optimization (tuning) = HPO

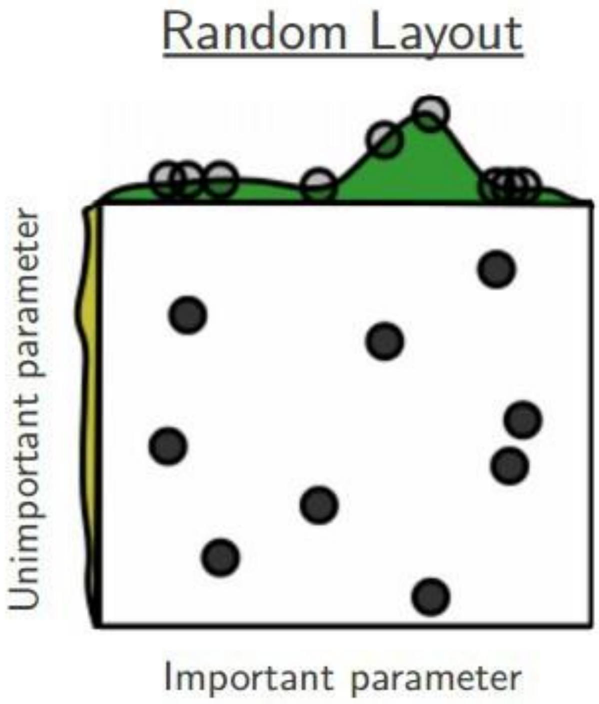
- Neural networks have a large number of possible configuration parameters, called *hyperparameters*
  - Avoids collision with NN *weights*, which are sometimes called *parameters*
- Applying optimization can automate part of the design of the neural network
- Involves two problem:
  - How to set the values of the hyperparameters?
  - How to manage multiple evaluations on compute resources?

Basic HPO Strategies

- **Grid search**
- **Random search**
- **Generic optimization**
  - Evolutionary algorithms
  - Baysian Optimzation
  - Gradient-Based Optimization
  - Model-based optimization (mlrMBO in R)

Baseline Methods: Grid Search & Random Search



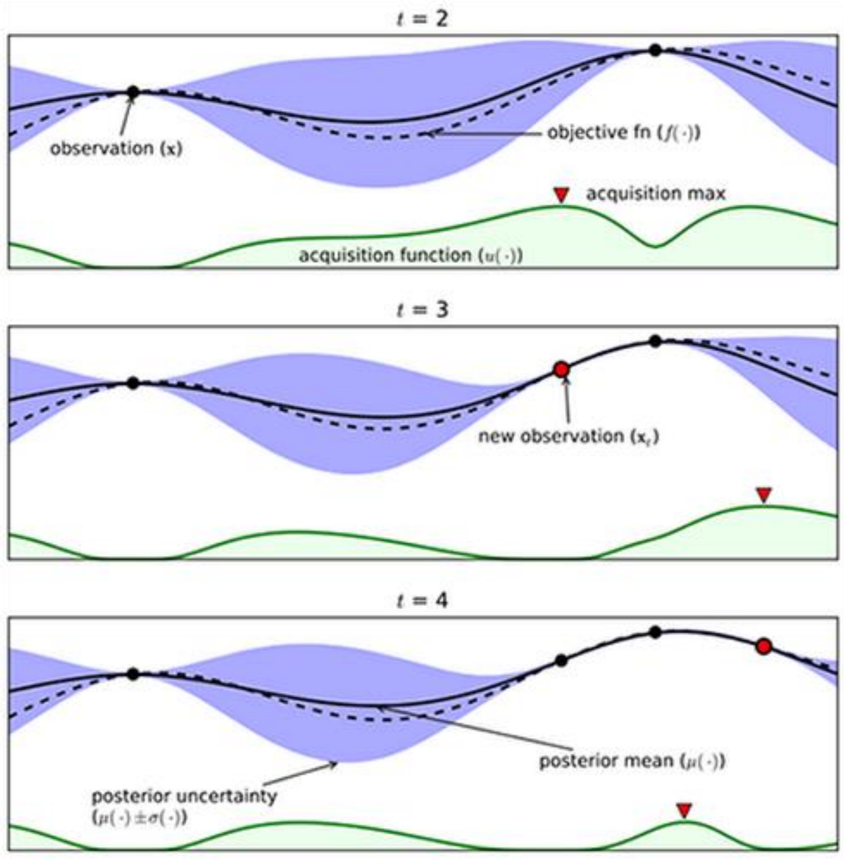


- Embarrassingly parallel
- Curse of dimensionality

- Embarrassingly parallel
- Does not learn from history

Bayesian Optimization

- Initially select random configurations to evaluate
- Build a gaussian process approximation of the objective function based on seen evaluations (posteriory distribution)
- Select good configurations to evaluate next based on a surrogate function (acquisition function) of your real objective.
- Balance exploration versus exploitation



Gaussian process approximation of objective function from Eric Brochu, Cora and Freitas 2010

HPO packages

- **Python:**
  - **Hyperopt**
  - **scikit-optimize**
  - **Spearmint**
- **R:**
  - **mlrMBO**
- **Cloud**:
  - Google's Hypertune
  - Amazon's SageMaker
- NN hyperparameter-specific optimization
  - NEAT, Optunity, …

HPO and HPC

- HPO required good amount of compute resources:
- Used to manage large-scale training runs
  - Hyperparameter searches $O(10^4)$ jobs
  - Cross validation (5-fold, 10-fold, etc.)
  - Data encodings (log2, Z-score, percent, etc.)
  - Low-level optimizations (tensor backends)
- Locate and transform input data
- Manage caching on local NV store
  - Internal joins, batching management, epochs
- Each job could be 100's to 1000's of nodes
- Driver scripts manage runs of 1K >10M core/hrs

Deep Learning for Life Science Users

Focus on what matters:

- Define the the deep learning model
- Define the Hyper-Parameters (HP)
- Choose a HP optimization algorithm
- Select resources (GPUs, time,)

Run this workflow on personal computer, commodity clusters, and supercomputers.

References

- https://cloud.google.com/blog/products/gcp/hyperparameter-tuning-cloud-machine-learning-engine-using-bayesian-optimization

- https://docs.aws.amazon.com/sagemaker/latest/dg/automatic-model-tuning-how-it-works.html

- https://roamanalytics.com/2016/09/15/optimizing-the-hyperparameter-of-which-hyperparameter-optimizer-to-use/

- https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/tune-model-hyperparameters

Thank you!

george.zaki@nih.gov