CIMVHR
Canadian Institute for Military and Veteran Health Research

ICRSMV
L'Institut canadien de recherche sur la santé des militaires et des vétérans

Machine Learning for Health Research

Sunday October 20th, 2019

9:00 am – 12:00 pm

Krieghoff room

WORKSHOP AGENDA

8:30 am	Breakfast available for attendees
9:00 am	Welcoming remarks and Introductions Mr. John Whitnall, IBM
9:15 am	Machine Learning and the Analytics Process Dr. Patrick Martin, Queen's University
10:15 am	Coffee break
10:30 am	Hands-on tutorial Dr. Mohamed Sami Rakha, Queen's University
11:45 am	Closing remarks and wrap up

CIMVHR | ICRSMV

INTRO AND BACKGROUND

- IBM is the sponsor of the first major research program at CIMVHR to take advantage of advanced computing research featuring Machine Learning
- Using Machine Learning and Big Data Analytics in research will lead to many research breakthroughs that were not previously possible
- Most researchers in Canada do not have access to the powerful specialized computers (such as large GPU clusters) needed to do this research
- At present these computers only exist in a few universities in Canada, but that is changing quickly
- Dr. Pat Martin and Dr. Sami Rahka are running a special project with CIMVHR to try to unravel the significant medical data challenges facing CIMVHR researchers today

CIMVHR | ICRSMV

MACHINE LEARNING AND THE ANALYTICS PROCESS

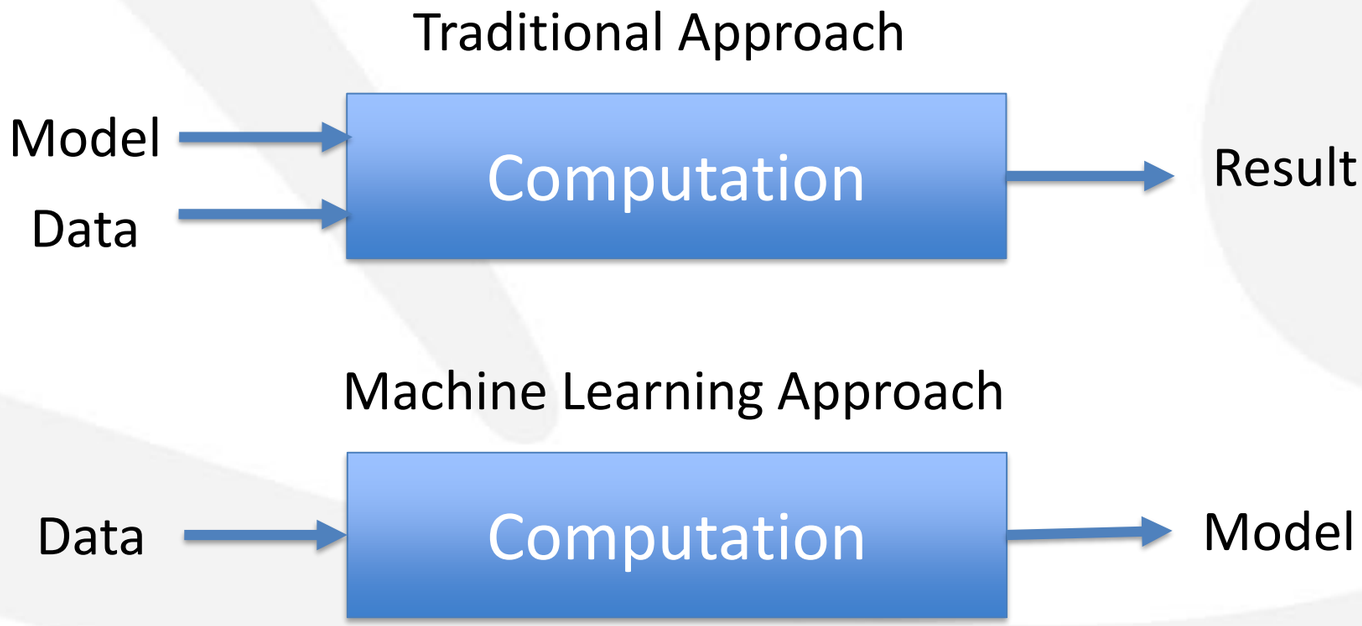
OUTLINE

- What is Machine Learning?
- Use Cases in Health Research
- The Analytics Process
- Final Thoughts

CIMVHR | ICRSMV
Canadian Institute for Military and Veteran Health Research | L'Institut canadien de recherche sur la santé des militaires et des vétérans

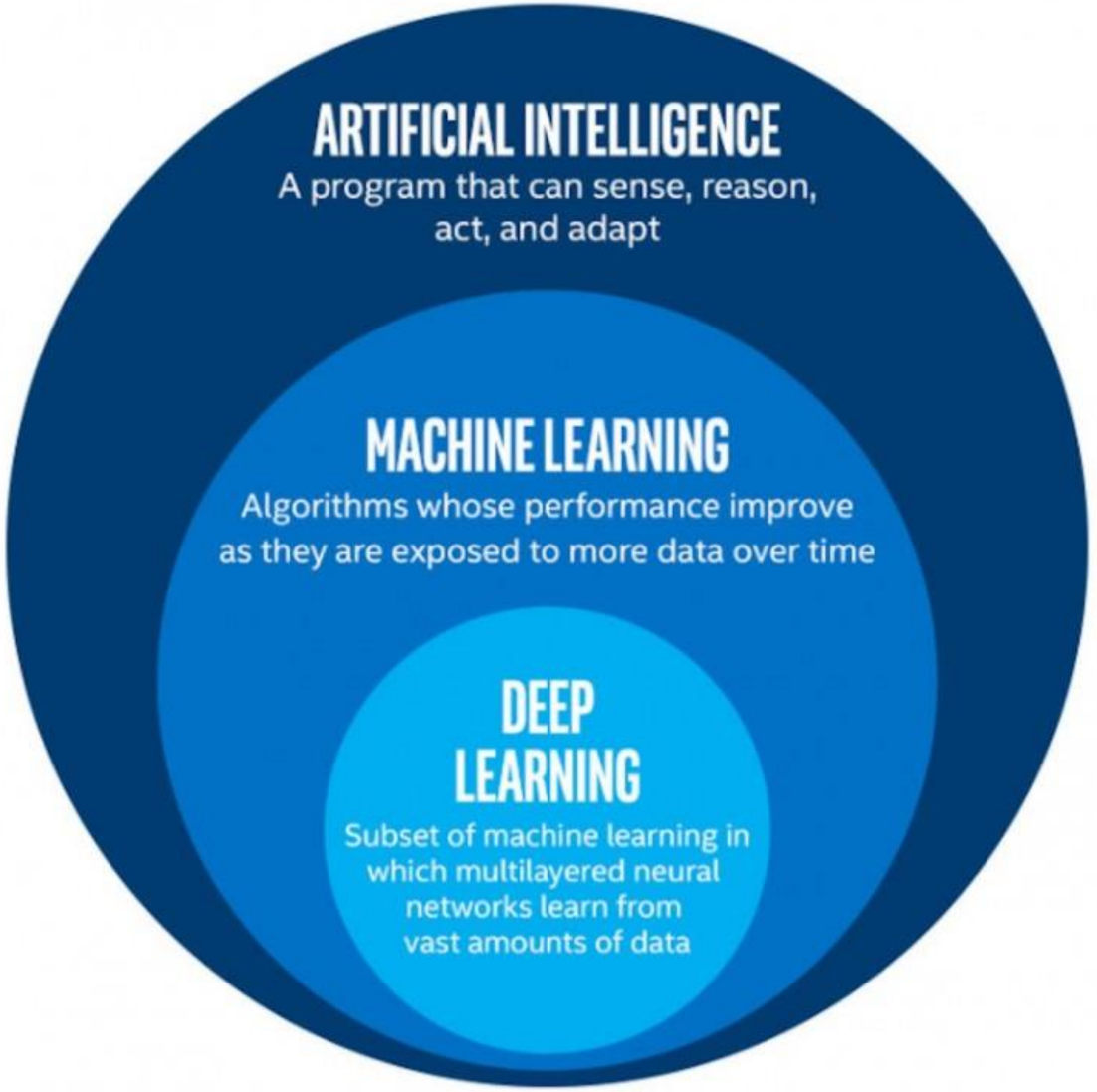
WHAT IS MACHINE LEARNING?

- *“Machine learning is the science of getting computers to act without being explicitly programmed.”*



CIMVHR | ICRSMV
Canadian Institute for Military and Veteran Health Research
L'Institut canadien de recherche sur la santé des militaires et des vétérans

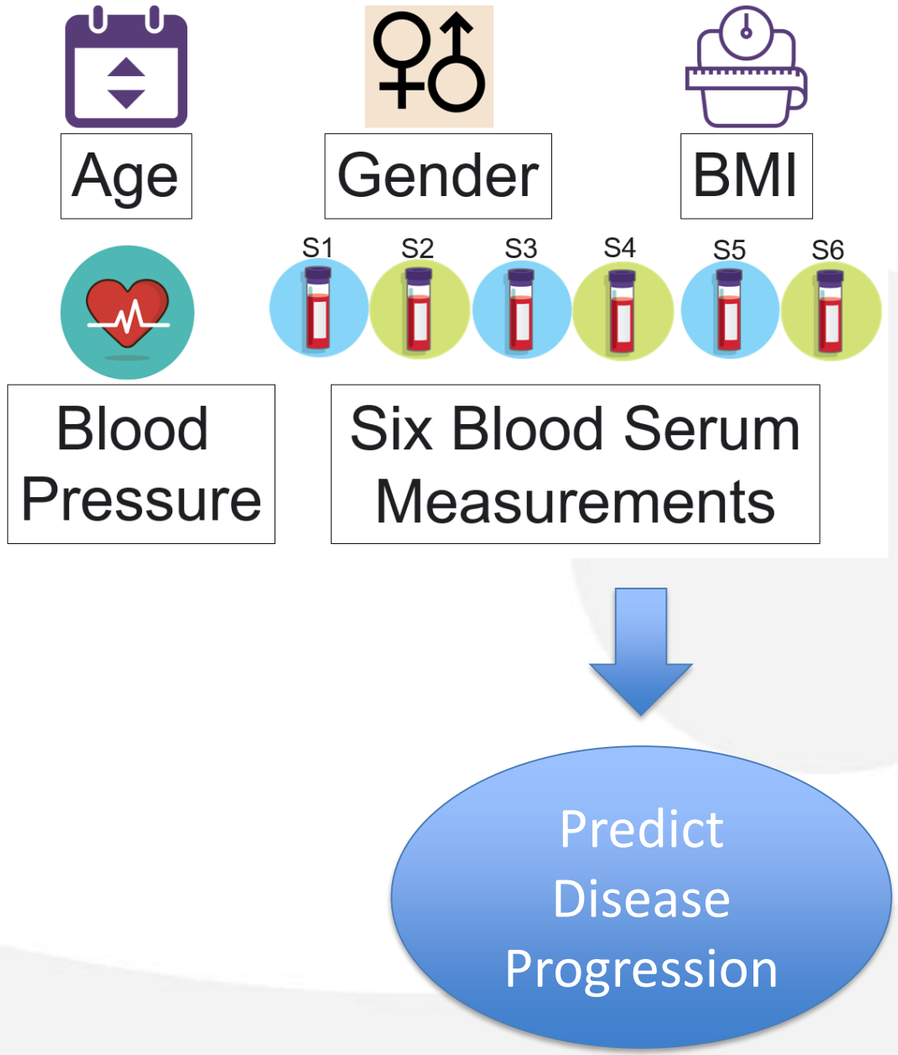
MACHINE LEARNING VS AI



CIMVHR | ICRSMV
Canadian Institute for Military and Veteran Health Research
L'Institut canadien de recherche sur la santé des militaires et des vétérans

TYPES OF MACHINE LEARNING

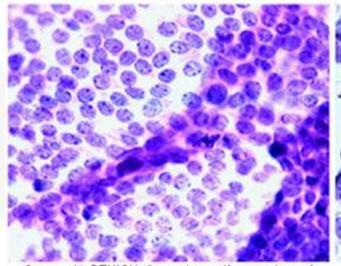
- **Supervised Learning**: model is *trained* on known input and output data so that it can predict future outputs



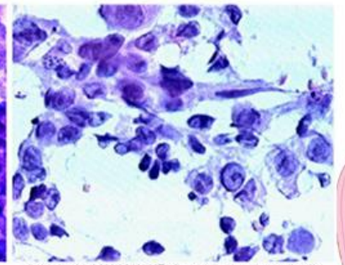
Quantitative Measure of the Disease

TYPES OF MACHINE LEARNING - 2

- **Unsupervised Learning**: model created based on hidden patterns or groupings in input data.



Smear with BENIGN diagnosis – uniform nucleus of cells, symmetrical, homogeneous, with areas within normal size



Smear with MALIGNANT diagnosis – nucleus of cells without uniformity, asymmetrical, not homogeneous (multiple sizes) and with areas above normal size

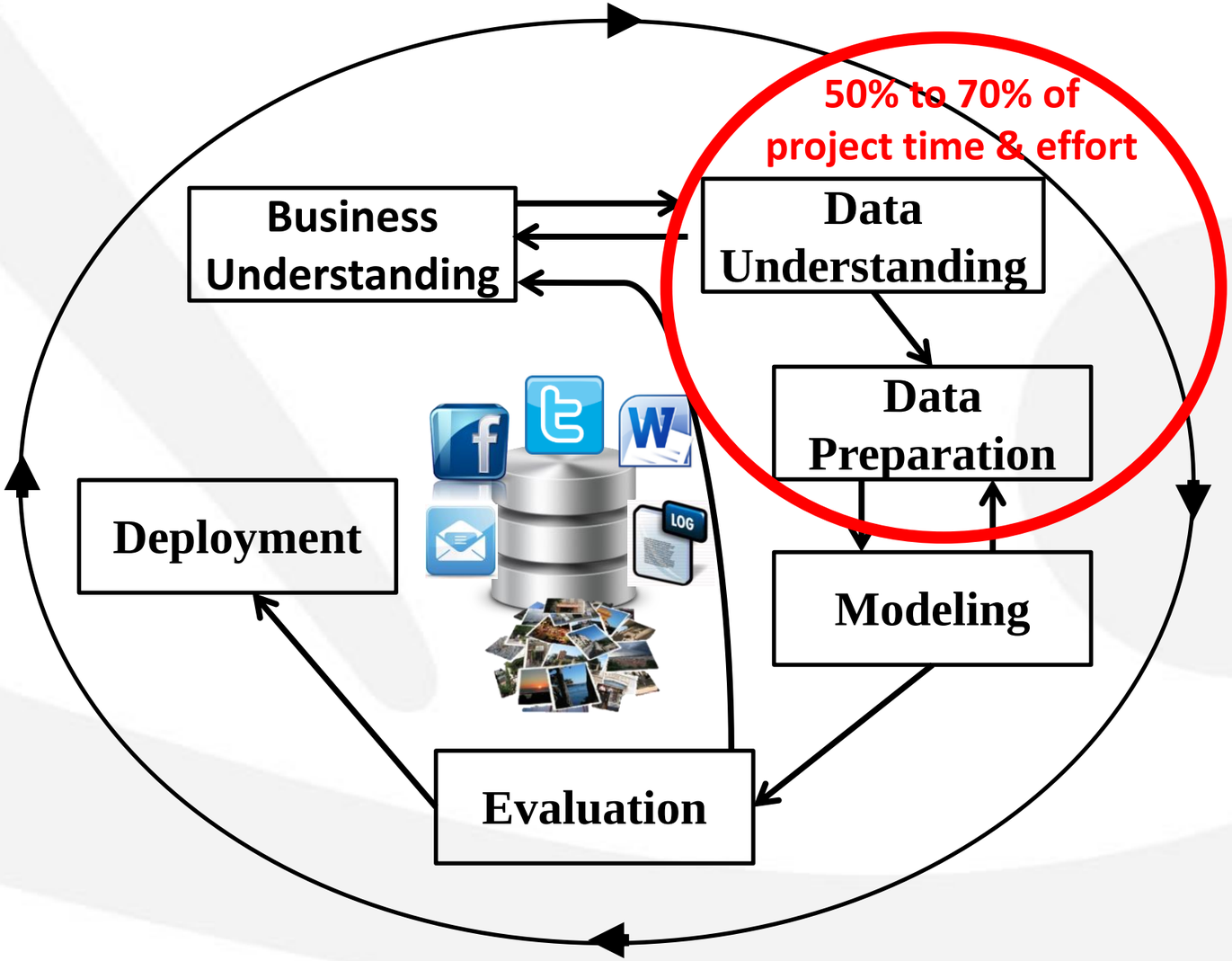
radius	compactness	radius	compactness	radius	compactness
texture	concavity	texture	concavity	texture	concavity
perimeter	concave points	perimeter	concave points	perimeter	concave points
area	symmetry	area	symmetry	area	symmetry
smoothness	fractal dimension	smoothness	fractal dimension	smoothness	fractal dimension

Identify Benign and Malignant Groups

SOME USE CASES OF MACHINE LEARNING

- **Diagnosis**
  - Automatic delineation of tumors as well as healthy anatomy in 3D radiological images.
  - Detecting diabetic retinopathy from retinal photographs
- **Treatment**
  - Predicting treatment outcomes from fMRI brain images and clinical data for PTSD patients
- **Healthcare epidemiology**
  - Predicting risk of healthcare associated infections
  - Predicting spread of infectious diseases

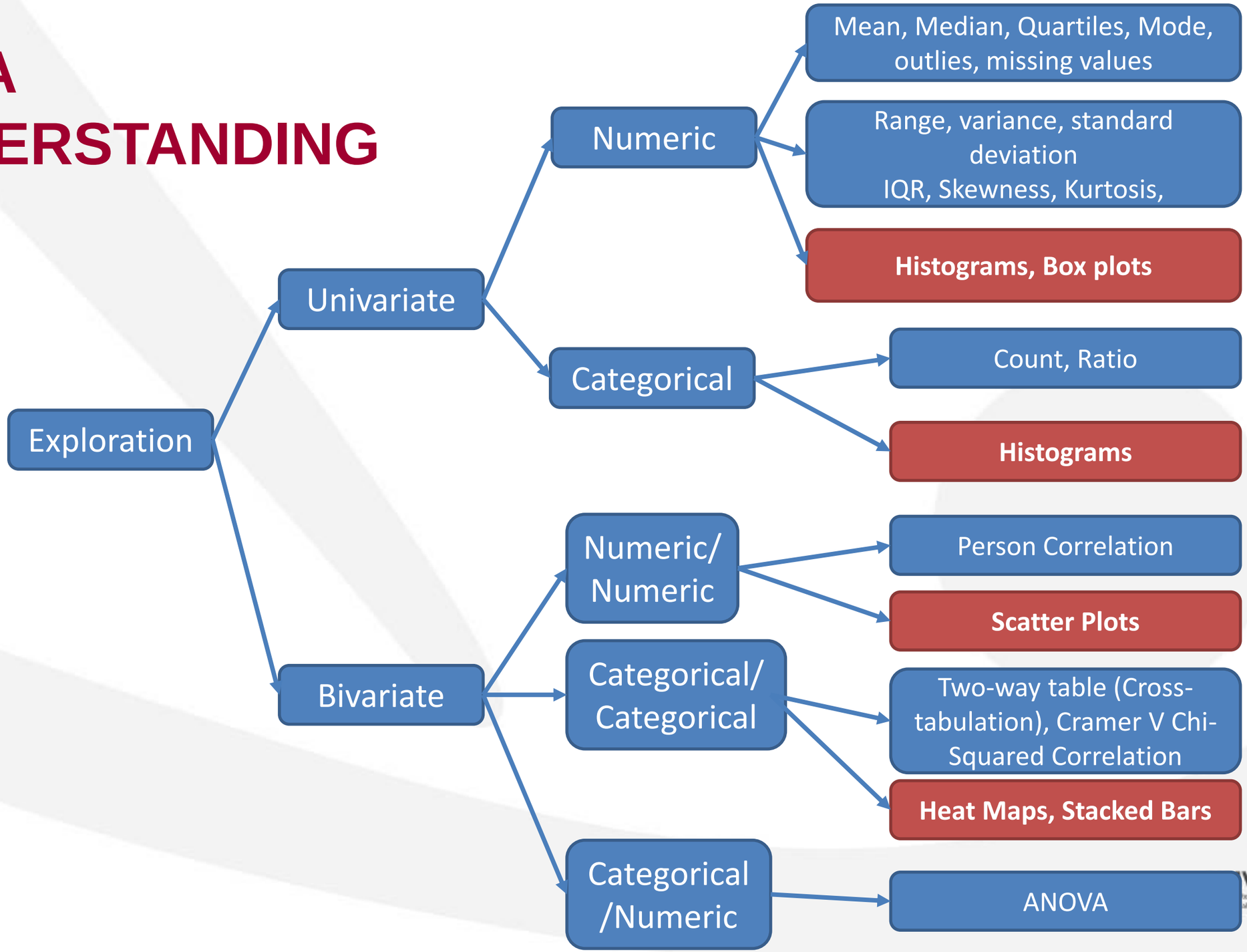
DATA ANALYTICS PROCESS(CRISP-DM)



BUSINESS UNDERSTANDING

- **Acquire domain knowledge**
- **Identify the questions**.
- **Identify the type of Analytics**
  - *Descriptive* (Insight into the past)
  - *Predictive* (Understanding the future)
  - *Prescriptive* (Advise on possible outcomes)

DATA UNDERSTANDING



DATA PREPARATION

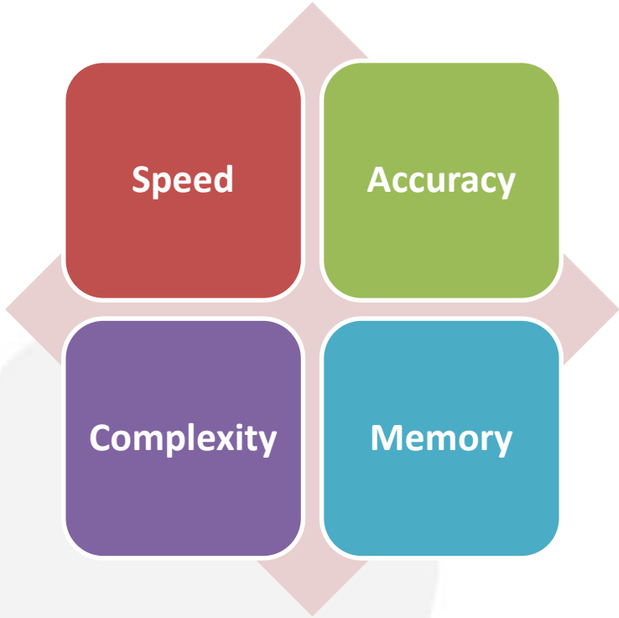
- The process of preparing data for modeling (Creating a predictive model) consists of three stages:
  - **Data Cleaning**: the process of detecting and correcting (or removing) incomplete, corrupt or inaccurate values from data.
  - **Feature Engineering**: the process of creating new features (attributes/columns) from existing ones to improve the predictive model performance.
  - **Feature Selection**: the process of selecting a subset of relevant features (variables, predictors) for use in the model construction.

MODELING

- Analytic Models **learn from experience**.
  - Analytic Models are computational programs that learn information directly from data without relying on a predetermined equation.
  - The models adaptively improve their performance as the number of samples from which they can learn increases.
- **Supervised Learning**: model is *trained* on known input and output data so that it can predict future outputs
- **Unsupervised Learning**: model created based on hidden patterns or groupings in input data.

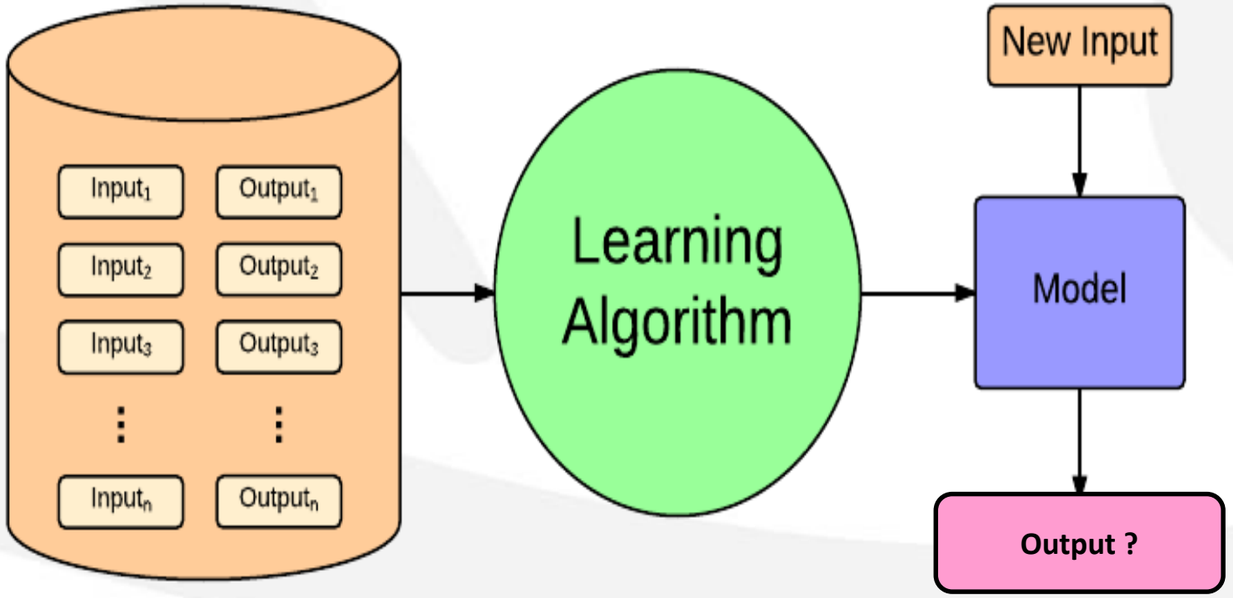
SELECTING THE MODELING ALGORITHM

- Algorithm selection depends on:
  - Data size
  - Data type (Numeric or Nominal)
  - Problem Type (Prediction or Grouping)
- Largely trial and error
  - Complex models tend to overfit data
  - Simple models miss the details and might not provide good accuracy.
- A good approach
  - Start with something simple that is faster to run and easier to interpret
  - Then move to more complex models to get better accuracy



SUPERVISED LEARNING

- Input is a set data records that consist of a set of independent attributes (features) and a known dependent attribute (target) value
- Algorithm trains a model to generate reasonable predictions for the target attribute for new data with an unknown target value.

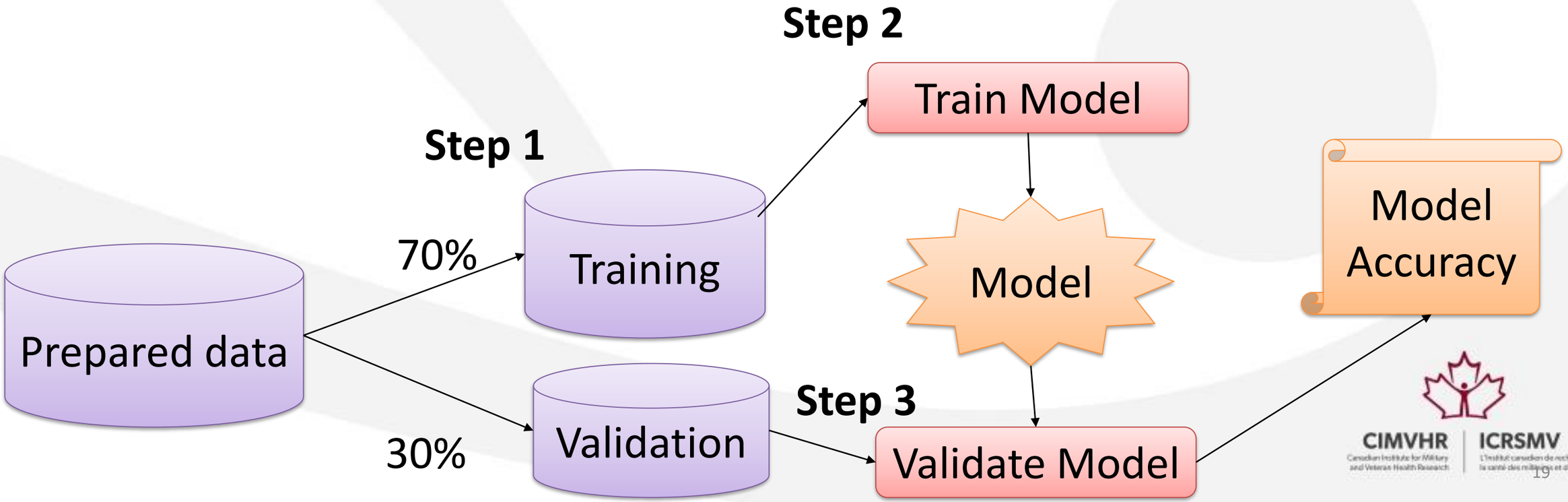


SUPERVISED LEARNING 2

- **Classification techniques**
  - Used to predict discrete target values. For example, whether a tumor is cancerous or benign.
  - Eg. K Nearest Neighbor (kNN), Decision Tree, Random Forest, Bayesian Networks and Naïve Bayes
- **Regression techniques**
  - Used to predict continuous target values. For example, predicting the length of the survival period of patients.
  - Eg. Simple and multiple linear regression, non-linear regression

VALIDATION

- Build training and test sets for your model
  - Split your dataset into two representative subsets (have records belonging to all Classes)
  - **Training**: used to train/teach your Model (usually 70% of your records)
  - **Validation**: used to calculate the accuracy of your Model (usually 30% of your records)



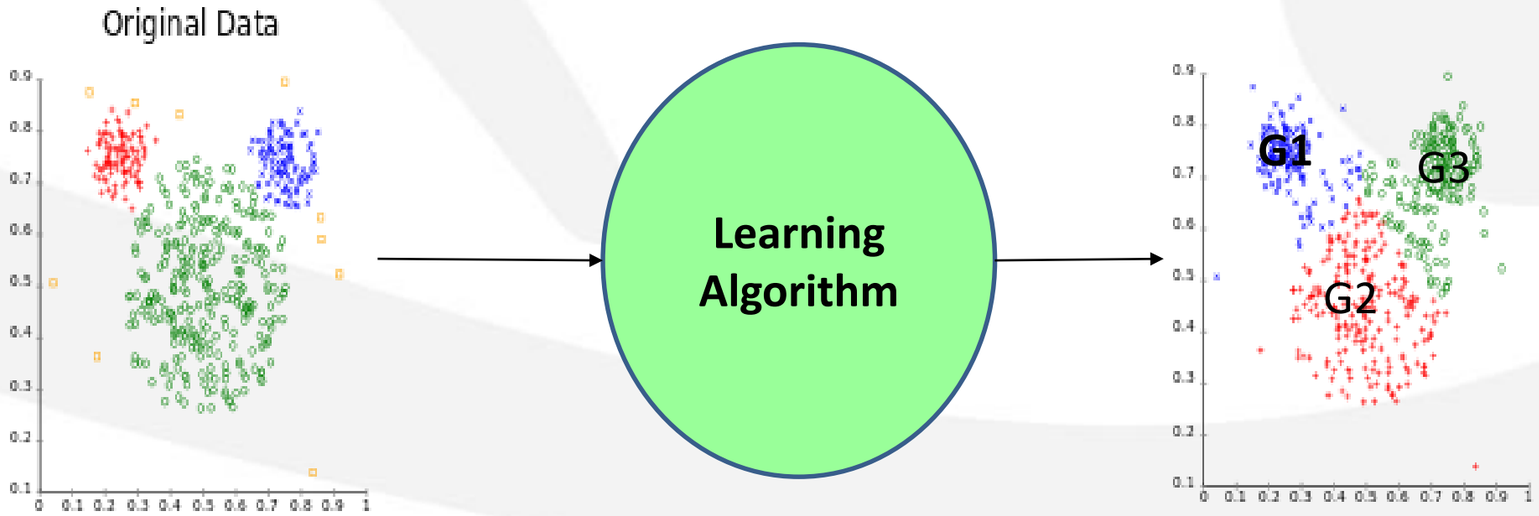
K-FOLD CROSS VALIDATON

- Simply splitting data into 70% training and 30% validation can give unreliable accuracy measurements.
- **k-fold Cross-Validation**:
  i. Split data to into *k* folds of equal size (same number of records).
  ii. Use *k-1* folds to train the classifier and the remaining one for validation. Calculate the model accuracy θ_i.
  iii. Repeat step ii *k times* leaving out a different fold each time.
  iv. Calculate the mean of the model accuracy $\widehat{\mu_\theta} = \frac{1}{k}\sum_{i=1}^{k}\theta_i$
  v. Calculate the variance of the model accuracy $\widehat{\sigma_\theta^2} = \frac{1}{k}\sum_{i=1}^{k}(\theta_i - \widehat{\mu_\theta})^2$

- Note that the *k* folds can be created using bootstrap resampling (random sampling with replacement) so that a record can exist in zero or more folds.
- *K* is typically set to 10 (The 10-fold cross validation).

UNSUPERVISED LEARNING

- Unsupervised learning **finds hidden patterns** in data.
- It is used to draw inferences from datasets consisting of input data without labeled responses (no Target attribute).
- Unsupervised learning is useful when you want to explore your data but don't yet have a specific goal or are not sure what information the data contains.



UNSUPERVISED LEARNING 2

- **Clustering** is an unsupervised technique used for finding groupings (Clusters) in data when the clusters are not known in advance.
- Clustering algorithms fall into two broad groups:
  - **Hard clustering:** each data point belongs to only one cluster. Eg. k-Means and k-Medoids, Hierarchical Clustering and Self-Organizing Map (SOM)
  - **Soft clustering:** each data point can belong to more than one cluster. Eg. Fuzzy c-Means
- Clustering objectives are:
  - **Maximizing intra-cluster similarity**: the overall similarity of data points within each cluster.
  - **Minimizing inter-cluster similarity**: the overall similarity of data points between clusters.

FINAL THOUGHTS

- Machine learning eliminates much of the human effort required to build prediction models
- BUT many of the complex approaches are just black boxes and as more control given to algorithms there is no guarantee of fairness, equitability or even veracity
- Machine learning is a valuable tool to exploit big data
- Collection and linking of data from variety of sources key to advancing health research
- Garbage In – Garbage Out is still true!

CIMVHR
Canadian Institute for Military and Veteran Health Research

ICRSMV
L'Institut canadien de recherche sur la santé des militaires et des vétérans

THANK YOU

www.cimvhr.ca