CSE-CST

Mathematics
Mathématique

(U) Directions for Data Science at CSE

Abstract. (U//FOUO) Data Science is increasingly becoming a necessary skill for many people at CSE. We feel that the most important Data Science problem to solve in the near term is continued investment in Data Science skills and tooling. This is needed at many different levels for all employees with big data problems at CSE, from analysts to Data Science specialists. Different teams and individuals require different levels of technical skill, and we all need to build communication skills so that we can act as effectively as possible. In this document, we discuss the current status of Data Science as we see it within the organization, and propose ways that to improve the overall effectiveness of Data Science at CSE in the future.

Keywords. Data science, coordination, research, requirements, pull-through, mission, outreach.

1. (U) Introduction

(C) As the size and complexity of data
continue to increase at CSE, the importance of Data Science is quickly increasing. In particular,

(C)

(U//FOUO) This document briefly describes the current state of Data Science at CSE in four key areas:

- Skills and Training
- Communication, Community and Engagement
- Pull-through, Tools and Technology

CSE Mathematics Research (LaTeX: January 5, 2018)

- Research and Requirements

(U//FOUO) Each of these areas currently the way Data Science functions at CSE. Therefore, we then recommends how these areas could evolve be in the coming years, and provides a strategy for doing so.

(U//FOUO) This document does not discuss particular Data Science techniques and their applications. Discussing CSE's Data Science is important and should also be discussed in detail; however, such dicussion is out of the scope of this document.

2. (U) What is Data Science?

(U) Let us start by briefly describing what we mean by "Data Science", in general and at CSE. Data Science aims to explore and understand patterns in data by examining the data sets in bulk. A search on the Internet will lead to many slight variants that we summarize by the following series of steps:

- Problem formulation – understanding client needs
- Data representation
- Exploratory data analysis and data cleaning
- Hypothesis formation, which could different levels of effort from simple analytics to more sophisticated statistical modelling
- Prototyping and evaluation
- Communicating results – data visualisation and presentation

(U) Data Science problems can vary considerably in complexity. We show in Figure 1 that tasks which are purely automation and data processing are considered to be at the least complex end of the Data Science spectrum. Although the coding itself may be complicated to do, we are referring to low Data Science complexity because the problem is completely defined and we know exactly what the expected outcomes will be. Data Science complexity increases as the algorithms give us new insight into the data itself.

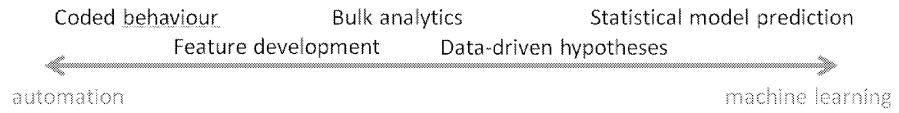


Figure 1: (U) There is a spectrum of Data Science tasks, from strict automation and data processing to the opposite extreme of machine learning and prediction from statistical models.

(C)

(U//FOUO) Complex Data Science problems require
and need joint investment from several teams. In this cases we think of Data Science as a team sport - that no one person should be expected to accomplish themselves. Success comes through the covergence of diverse skills from a mixed team. Data Science problems require individuals with programming and computer science skills, as well as individuals experienced in mathematics and statistics, and finally they require subject matter expertise for the problem at hand, as shown in Figure 2.

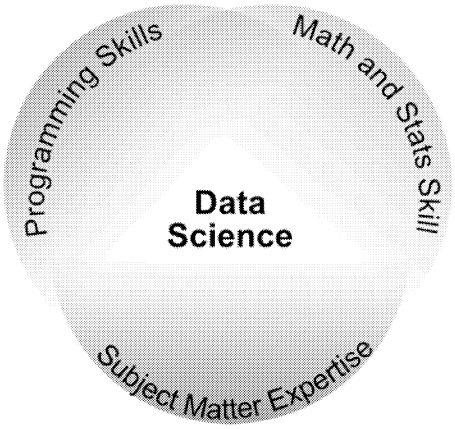


Figure 2: (U) Data Science is a team sport, requiring us to bring together individuals with different backgrounds.

(C) Data Science techniques build models to reveal patterns in data when there is too much data to examine manually. These techniques

There must be commitment on everyone's part to have appropriate training in the skills that are needed to work jointly on the projects to make this work.

3. (U) The State of Data Science at CSE

3.1. (U) Skills and Training

(S)

(C) Many teams at CSE need Data Science skills of different types. Figure 3 gives a view of how we see the nature of Data Science problems to that different CSE teams want to accomplish in terms of the operational (more immediate) to strategic (long-term).

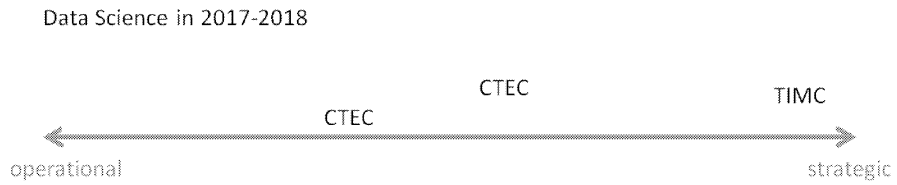


Figure 3: (C) Many teams at CSE pose Data Science problems. Here are examples of teams where the nature of the problems vary from more operational to more strategic.

(U//FOUO) Devoting time to training and refining skills is necessary to understand and adapt to new trends, to understand how the community is evolving, and to help us see the big picture and focus on longer-term general Data Science problems that can affect CSE in a broader context.

(U//FOUO)

Several blogs and spaces have been created on that employees can follow such as a Python User Group, Data Science Brain Foodies and several blogs written by Data Scientists about Data Science.

(C) Learning and Development have been working toward higher visibility of Data Science by asking to offer short lectures as part of their Onboarding program and their Operational Management Development Program (OMDP) this year;

(S//SI//REL CAN, FVEY)

3.2. (U) Communication, Community and Engagement

(C)

This type of activity gives first-hand knowledge of problems to the individuals that have background in the Data Science techniques for solving them.

(C) Another successful method we have found to gather requirements is through workshops. For example at several Big Dig workshops,

Many other workshops and surges also take place regularly with different teams varying across the organization such as to name a few.

(C) Communication to analytic clients about what is possible for data scientists is also important. We currently update pages with examples of Data Science projects that have been recently completed. We also post weekly notes for on which are followed by several analysts throughout CSE and give an idea of what data scientists are accomplishing on a regular basis, in terms of quick wins, longer term projects and events.

(C) the methods above have worked well for us and has exposed our services to a lot of the organization

We hope to find even more ways to reach out to different parts of the organization so there is more awareness of the types of problems that can be facilitated by Data Science.

3.3. (U) Pull-through, Tools and Technology

(C) Pulling research prototypes through to production

(C) We have made several efforts to stramline the process of getting a prototype from research to production. The most noteable has been

This is a tremendous win in terms of giving analysts and others the power to use our services

(C) Further, we have noticed that our analysts

(C) environments allow analysts to use written by data mining specialists, as well as writing their own notebooks. are environments that support code development with output, as well as mark-up for documentation in the same files. They are also easy to share, so a data scientist can write-up a notebook with complete code and examples along with the documentation, and an analyst can run this code directly and see right away how it is affected by any changes.

(C) Yet another way we have tried to facilitate pull-through is a project called

(C) Finally, we have also used the integration model to aid with the pull-through of solutions.

3.4. (U) Research and Requirements

(C)

(C) The prioritization of Data Science requirements is Different clients approach Data Science specialist teams with problems that they would like help completing. Since these teams have also shown that they want to work with us and commit subject-matter expert resources, we tend to work with nearly every client that comes to us. Data Science specialists may also seek out clients based on

4. (U) The Vision of Data Science at CSE

4.1. (U) Skills and Training

(C)

(C) We take particular note of

(C)

(C) We may also choose to

For example, Learning and Development is already discussing

(C)

4.2. (U) Communication, Community and Engagement

(C) Data Science literacy should facilitate communication and engagement. This would means that analysts have a better concept of what to ask about, and what is feasible. Also, we hope that analysts would not be afraid to approach the Data Science research branch even to simply have discussions or brainstorm. We would like analysts to understand and even contribute to building statistical models that represent their data, and for them to make use of these models in generating and proving/disproving hypotheses.

(C)
We take know that employees

Executives have the broadest view of all the work done at CSE and can best prioritize what is needed. Strategic data science researchers have

The best results can be achieved when we all communicate well and work together.

4.3. (U) Pull-through, Tools and Technology

(C) We hope that in the future our pull-through tasks will be

(C) We consider for a moment what technologies would be needed to facilitate Data Science activities in the future and why. Data Science is discovery of patterns in data, which can sometimes be described as finding "unknown unknowns" – we don't know what we are looking for or how to find it.

(U) The Data Science environment itself is also changing. Historically, the push for high-performance computing has been to give more power to get more accurate solutions. More recently, there has been a realization of the importance of interpretable models. There are use-cases which fit two different models and the choice will depend on the opinion of clients:

1. Models should have the best possible accuracy, but do not have to be interpretable
2. Models have reasonable accuracy, but interpretability should be considered more important than the best possible accuracy
3. Models with an interleaving of interpretability and accuracy with HPC usability

(C)

(C) As mentioned earlier, ; seem to be an effective method to help pull research prototypes through to operations in a useful and adjustable way for analysts.

(U) One way that the external community is keeping up with this ever-increasing Data Science tool development is to open-source tools. This helps alleviate the burden of maintenance, allows many other to contribute new and interesting features to projects, and assist with the time-consuming burden of training as well. Rather than hiring individuals and needing to spend time on training after they have been hired, open-sourcing the code allows them to ramp up even before they are hired and get up-to-speed as quickly as possible. A prime example of this is Google open-sourcing Tensorflow for deep learning.

(C)

(C)

(C) CSE is a small and agile organization

(C)

4.4. (U) Research and Requirements

- Craft discussion points here.

5. (U) How We Get There

5.1. (U) Research Branch for Data Science

(C)

The idea was to draw on the knowledge of these specialists to help point the analysts in the right directions.

(C)

These researchers would be expected to keep abreast of Data Science trends and to

These are the researchers that would

(C) Much like the a "place to go" with their Data Science thoughts and problems, a front desk where they can provide Data Science research requirements and ideas.

5.2. (U)

(C)

(C)

(C)

5.3. (U) **Development Resources for Pull-through**

(C)

6. (U) Dependencies

6.1. (U) Governance

6.2. (U) Support from analyst community

(C) We hope that this proposal will also been embraced by the analyst community.

(C) We also remind the readers of several "quicker wins" that Data Science may help to bring. For example,

*****

6.3. (U) Support from developer community

(C)

*****

6.4. (U) Support from Learning and Development

Proper training plans

6.5. (U) Support from Policy and Compliance

7. (U) Conclusion

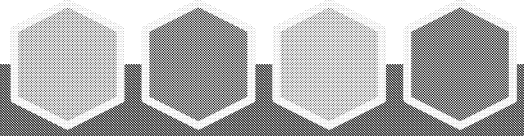
(U) References

s.15(1) - DEF

s.15(1) - IA

SECRET//SI//REL TO CAN, FVEY

Intro to Data Science



Everything you always wanted to know about data science and machine learning but were too afraid to ask

- By the end you will have some idea of:
  - What data science is
  - What machine learning is
  - How they might be used in ITS
  - What questions we should be asking

What is data science? Good question…

- **Data science** is the process of applying scientific principles to *data* in order to understand it
- **Data science** is an *interdisciplinary* field of scientific methods, processes, algorithms and systems to extract *knowledge* or *insights* from data
- **Data science** is the *intersection* of maths, statistics, machine learning, data mining, visualisation and information systems

i.e.

- **Data science** is everything to do with storing, processing and analysing data in order to understand it

It is a very loose term, don't worry about it

(everything we do is data science…?)

Why do we need it?

We have a lot of data and humans can't keep up

(smoking guns are much harder to come by)

- **Automation**: taking a task done by hand by an analyst and getting a computer to do all/most of the work

- **Analytics:** automated process to (help) answer a question/hypothesis of interest by obtaining, processing and summarizing the necessary data

What is machine learning (ML)?

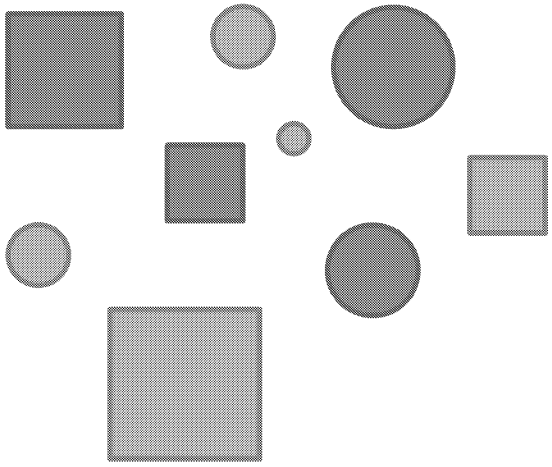
- Building algorithms that can learn the patterns and structure in data
- Can then use knowledge of those patterns to:
  - Group existing data (clustering)
  - Make predictions on new data (e.g. classification)

Side note

- **Artificial Intelligence** is another buzzword
- Usually when people talk about AI they mean machine learning (technically a subset of AI)

Example

- Clustering

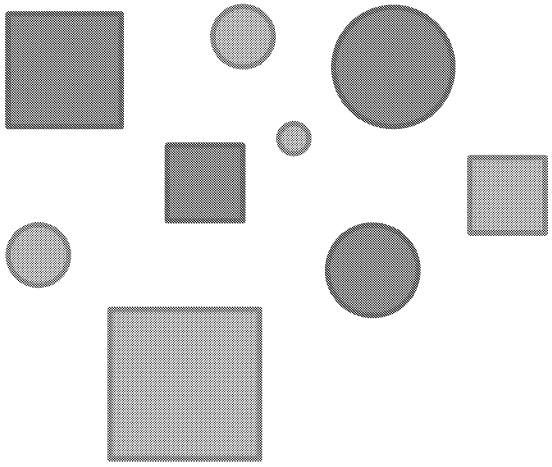


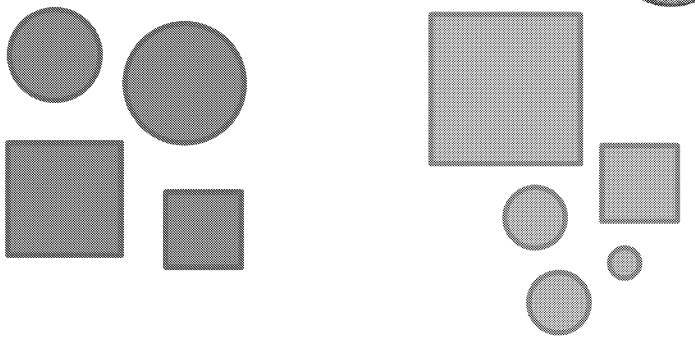
Features

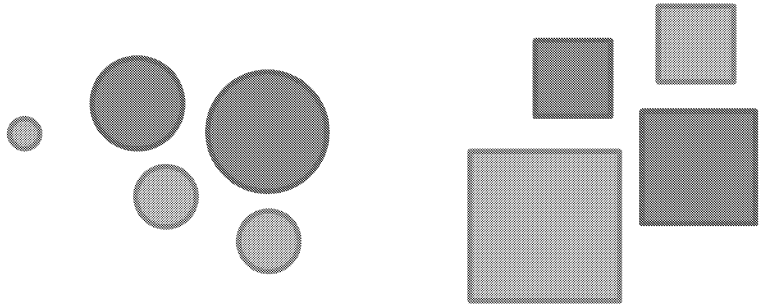
Colour	Shape	Size
Blue	Circle	5
Blue	Circle	4
Blue	Square	5
Blue	Square	3
Brown	Square	7
Brown	Square	3
Brown	Circle	2
Brown	Circle	2
Brown	Circle	1

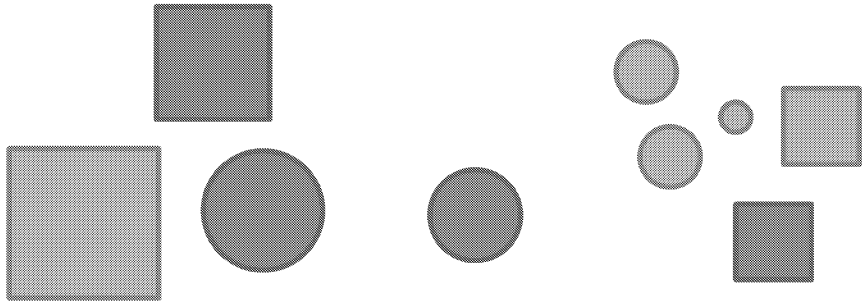
Example

- Clustering







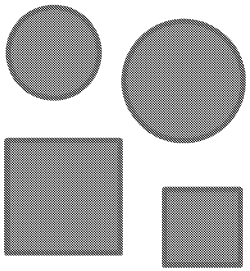


The features you provide matter!

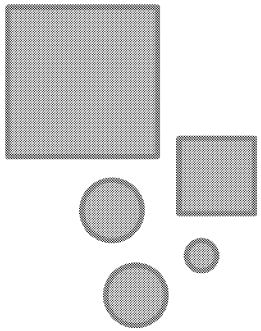
Example

- Classification

Class 1



Class 2



?

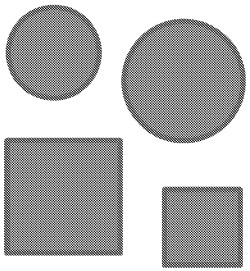
Features

Colour	Shape	Size	Label
Blue	Circle	5	1
Blue	Circle	4	1
Blue	Square	5	1
Blue	Square	3	1
Brown	Square	7	2
Brown	Square	3	2
Brown	Circle	2	2
Brown	Circle	2	2
Brown	Circle	1	2

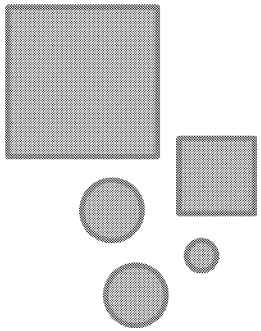
Example

- Classification

Class 1



Class 2



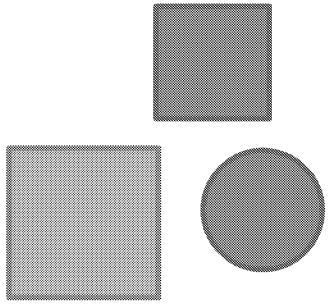
Colour	Shape	Size	Label
Blue	Circle	5	1
Blue	Circle	4	1
Blue	Square	5	1
Blue	Square	3	1
Brown	Square	7	2
Brown	Square	3	2
Brown	Circle	2	2
Brown	Circle	2	2
Brown	Circle	1	2

- Class 1

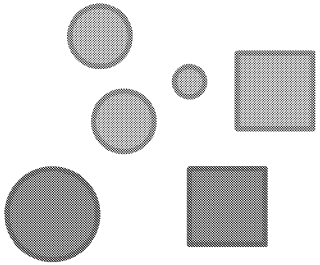
Example

- Classification

Class 1



Class 2



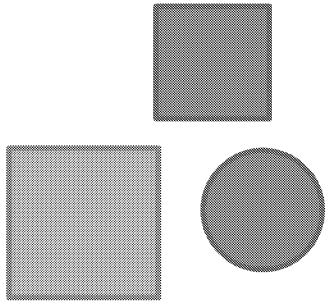
?

Colour	Shape	Size	Label
Blue	Circle	5	1
Blue	Circle	4	2
Blue	Square	5	1
Blue	Square	3	2
Brown	Square	7	1
Brown	Square	3	2
Brown	Circle	2	2
Brown	Circle	2	2
Brown	Circle	1	2

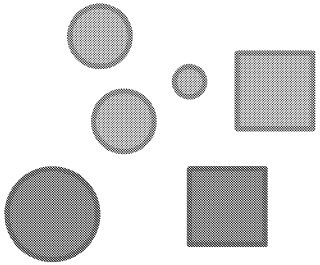
Example

- Classification

Class 1



Class 2



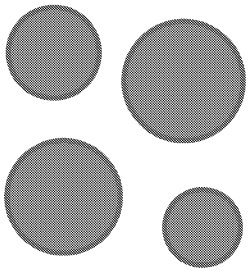
Colour	Shape	Size	Label
Blue	Circle	5	1
Blue	Circle	4	1
Blue	Square	5	1
Blue	Square	3	1
Brown	Square	7	2
Brown	Square	3	2
Brown	Circle	2	2
Brown	Circle	2	2
Brown	Circle	1	2

Class 2

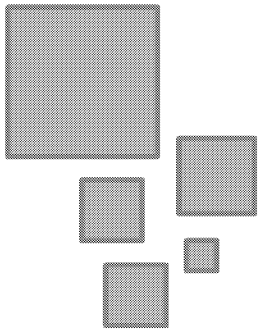
Example

- Classification

Class 1



Class 2



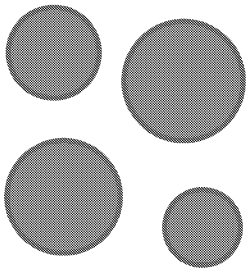
 ?

Colour	Shape	Size	Label
Blue	Circle	5	1
Blue	Circle	4	1
Blue	Circle	5	1
Blue	Circle	3	1
Brown	Square	7	2
Brown	Square	3	2
Brown	Square	2	2
Brown	Square	2	2
Brown	Square	1	2

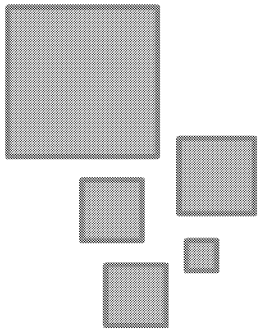
Example

- Classification

Class 1



Class 2

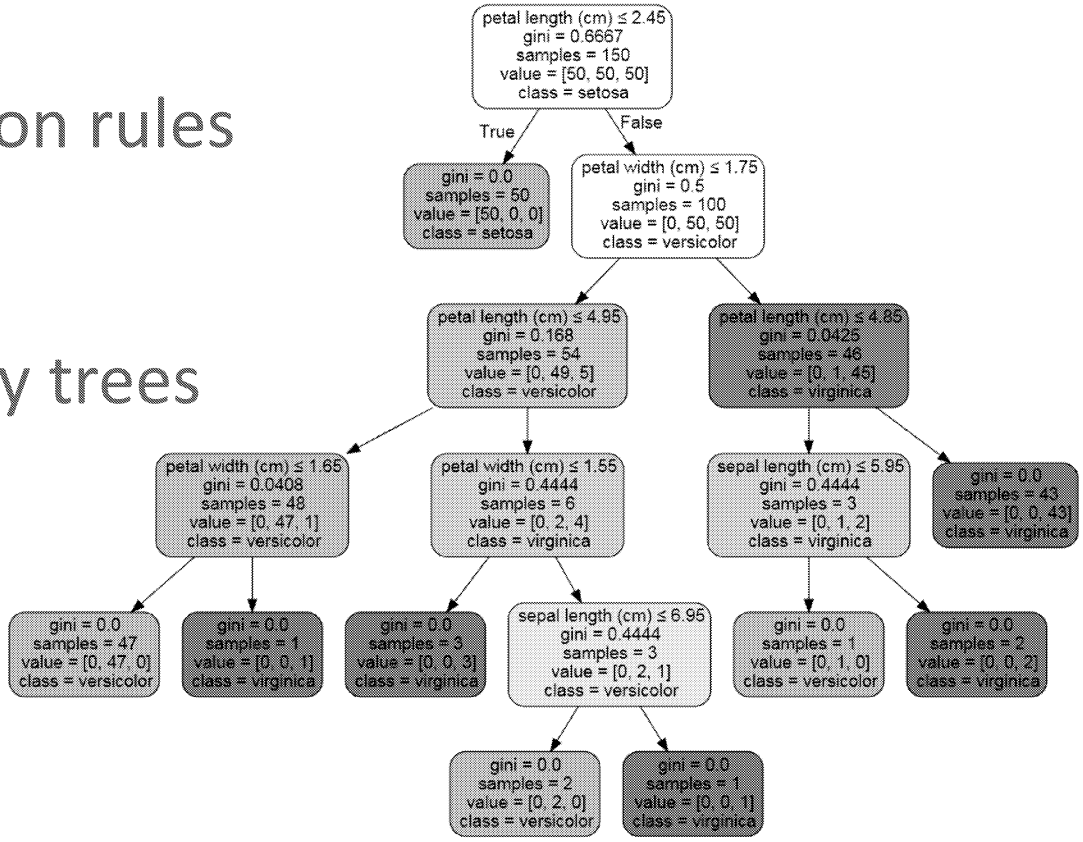


Colour	Shape	Size	Label
Blue	Circle	5	1
Blue	Circle	4	1
Blue	Circle	5	1
Blue	Circle	3	1
Brown	Square	7	2
Brown	Square	3	2
Brown	Square	2	2
Brown	Square	2	2
Brown	Square	1	2

Class 1

Some example classifiers

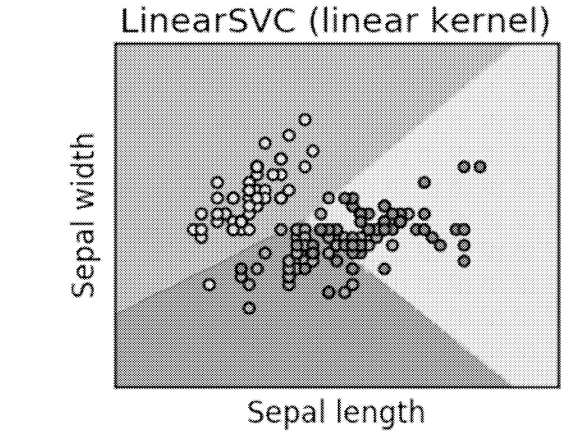
- Decision trees
  - Learn a set of decision rules
- Random forest:
  - Tallied votes of many trees
  - Each tree uses a random subset of training data and features

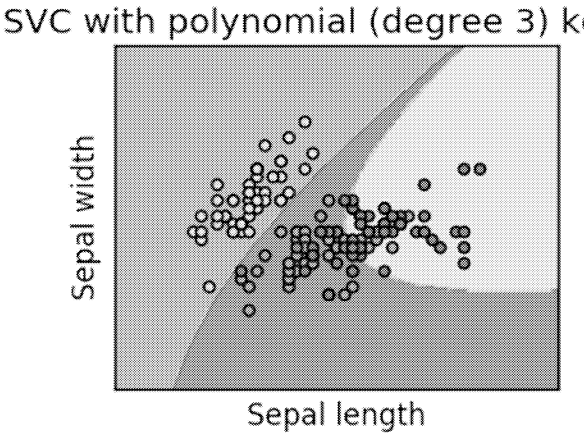


Some example classifiers

- K nearest neighbour
  - Contains all the training data (labelled)
  - Think of data as points in space
  - When a new point comes in, look at which points are closest and assign it that class
  - Can also be used for clustering (no labels)

Some example classifiers

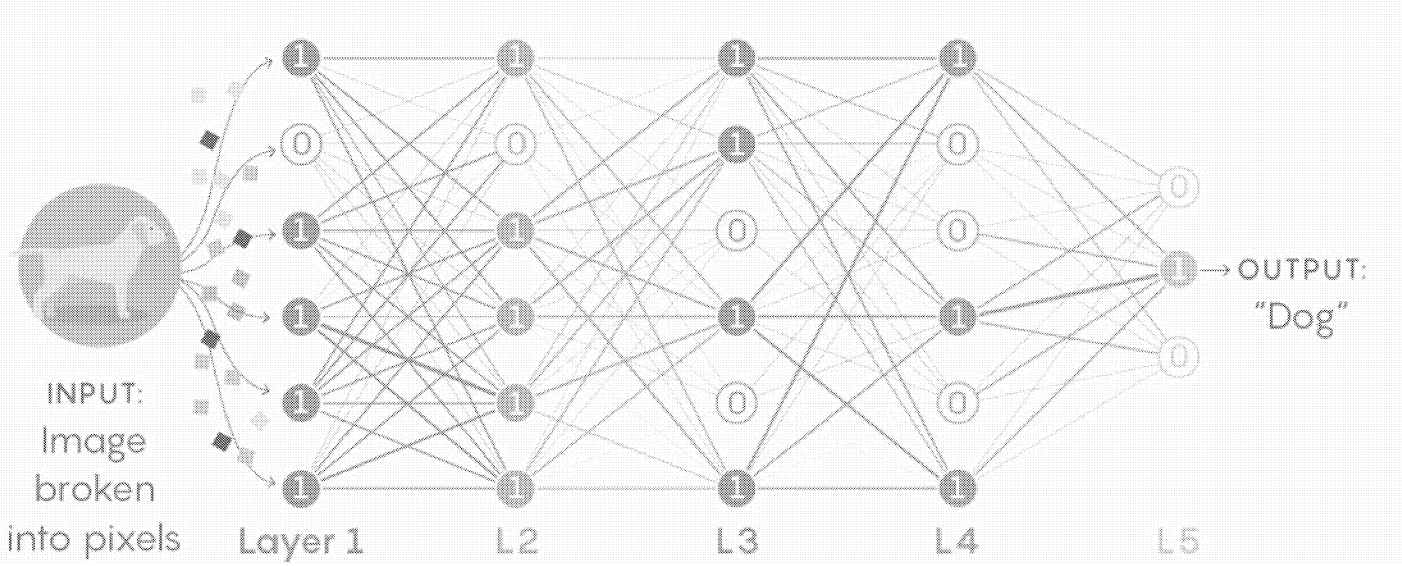




- Linear classifiers:
  - Think of data as points in space
  - Learn an equation that best separates points of different classes (labels)
  - Can also do maths tricks to end up with curved lines (decision boundaries)

Some example classifiers

- Neural networks
  - Lots of simple classifiers stacked together in layers
  - Learns input/output weights for each “neuron”



A note about training data

- The more data, the better the model
- Need positive and negative examples (i.e. examples of “thing” and “not thing”)
- Labelling takes a **lot** of time, effort and domain knowledge

Potential applications in Cyber Defence

- Clustering can help:
  - Break down pots of data into more manageable chunks;
  - Find links you didn't see before.
- Examples:

Potential applications in Cyber Defence

- Classification can help:
  - Infer context for your data;
  - Find examples of behaviour you care about;
  - Find things that don't conform to expectations.
- Examples:

I want to find weird (or normal) behaviour

Write rules based on experience

Automation

Heuristics

Basic analytics

Define behaviour using the data

"Give me groups of similar things"

e.g. clustering

unsupervised

"Give me things that look like this"

e.g. classification

supervised

"Do these groupings make sense?"

Potential questions

- Do you preserve original data in the way you derive your features?
  - E.g. if you have an input vocabulary of process names

Potential questions

- Do you preserve original data in the way you derive your features?
- Can you reconstruct original data using your model?
  - In some cases, this is an ongoing area of research!
  - Is a reconstruction/approximation original data…?

Potential questions

- Do you preserve original data in the way you derive your features?
- Can you reconstruct original data using your model?
- Does a trained model need to be retained?
  - Do you need to be able to repeat the same analytic process?
  - If so, do you need the same results (i.e. the same model)?

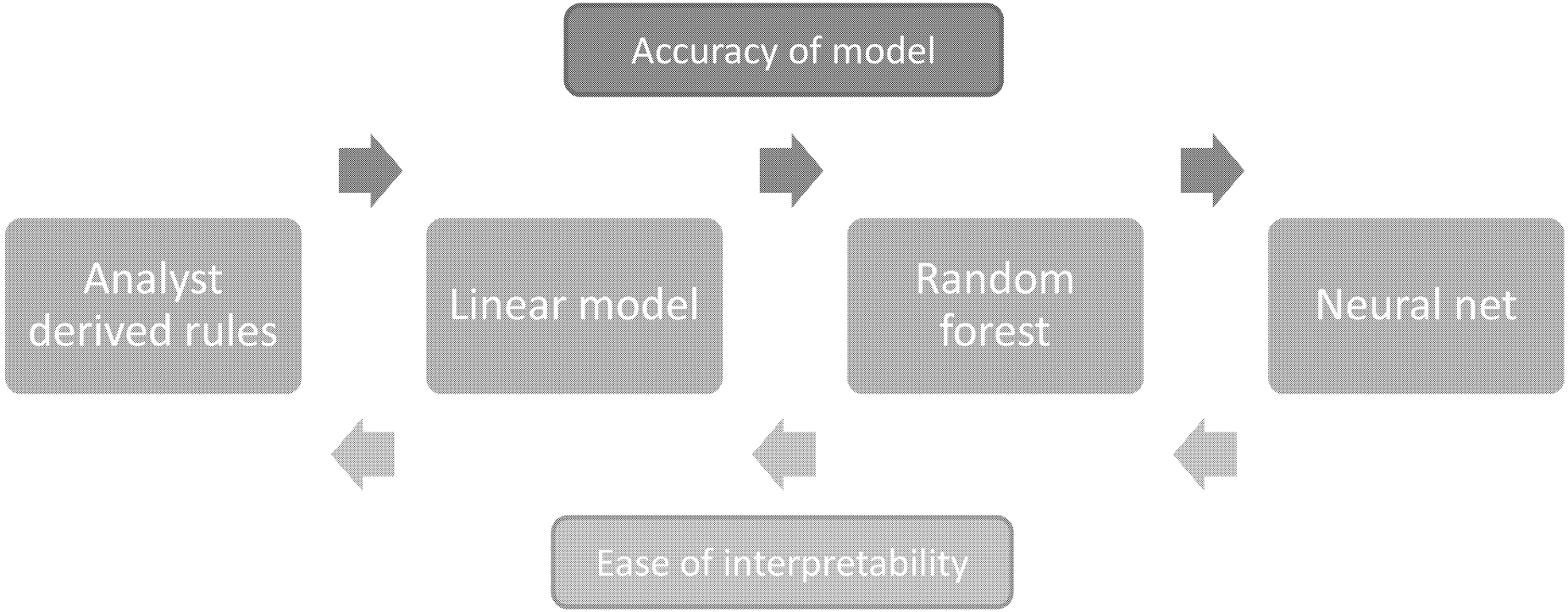
Potential questions

- Do you preserve original data in the way you derive your features?
- Can you reconstruct original data using your model?
- Does a trained model need to be retained?
- Does training data need to be retained?
  - If so, you're going to need a lot of storage…

Potential applications in Cyber Defence

- Making decisions using machine learning
  - Decide to collect more data e.g.
  - Sort alert triage ordering
  - Recommend similar alerts
- Maybe not important now, but in future?

A note about interpretability



Potential questions

- Do you preserve original data in the way you derive your features?
- Can you reconstruct original data using your model?
- Does a trained model need to be retained?
- Does training data need to be retained?
- Are your decisions compliant?
- How much needs to be explained/explainable?

Points to take away

- "Data Science" means whatever you want it to mean
- Machine learning (ML) is just automatically learning patterns/structure in data
- There are questions that we will need to answer together to use ML effectively and compliantly

Some vocabulary

Term	Definition
Features	The measureable properties of an observation which form the inputs to a machine learning problem
Feature vector	A list of feature values (in the correct order) representing an observation
Model	The mathematical description of how to transform features into an output prediction
Classification	The challenge of deciding to which group(s) a new observation belongs, based on a model of the groups derived from a set of labelled training observations
Regression	The challenge of predicting a continuous value for a quantity based on an observation, rather than a category (as in classification)
Clustering	The challenge of grouping unlabelled observations based on a measure of “similarity”
Training	The process of using labelled observations to optimise quantities in your model, in order to get the most accurate model overall (using the labels to determine accuracy)
Overfitting	Occurs when a model is trained on a training set “too closely” i.e. where a model takes into account irrelevant variations (noise) or other features peculiar to the training set

Brief on RA-070

(C) Goal

Research Activity 070 (RA-070) is the first effort prioritized by the working group associated with CSE's strategic research thrust:

The primary goal of the RA-070 is to This brief presents a possible approach to achieving this goal. The program is encapsulated by three related objectives:

Examples of include:

Examples of include:

(C) Research Guidelines

The philosophy behind this research project is to

Our objective is to publish a significant portion of our work in order to contribute publically and influence the greater research community that already care about the problem. Ideally, we'd want to

[1] R-Future – Strategy for a complex world reaching beyond human cognition:
/51595115

Our intention is

With the present research activity, we want to

(C) Context and Background

Soon after we got the confirmation that the topic of would be in our research strategy, we started to search for

We quickly realized that CSE is currently

for the upcoming expanded mandate enabled by the upcoming adoption of bill C-59.

We subsequently reached out to

After a few meetings, we agreed that CSE has a lot to offer given its unique mandate, but also from a research and technology development perspective. We also determined that

In response we organised internal brainstorming sessions between SIGINT, CCCS and ETS/Applied Research/TIMC experts and managers from various related missions including: We obviously also included the mission policy team in the discussion as we started to have a clearer idea of what we wanted to propose. The objective of the brainstorm was to agree on one project with the following characteristics:

The consensus is that this project, identified as RA-70, is the best proposed idea that satisfies the above criteria.

(C) Imagineering

In this section, we provide an overview of an example approach to tackling our objectives. It is very important to keep in mind that this proposal fits in the category of

This document is not trying to be precise and specific, but it is simply aiming at providing a conceptual and approximate picture of the proposal. The final implementation, if successful, might end up being significantly different.

The main idea is to

The following diagram illustrates the conceptual approach.

Steps

In summary, the proposed approach will

Use cases

If successful, this system will enable analysts to answer questions such as:

Operational policies

Mission policy representatives will work with researchers throughout the R&D phases. Reviews and measures will be put in place throughout the duration of the research project. For example, measures will be put in place to

Development and deployment

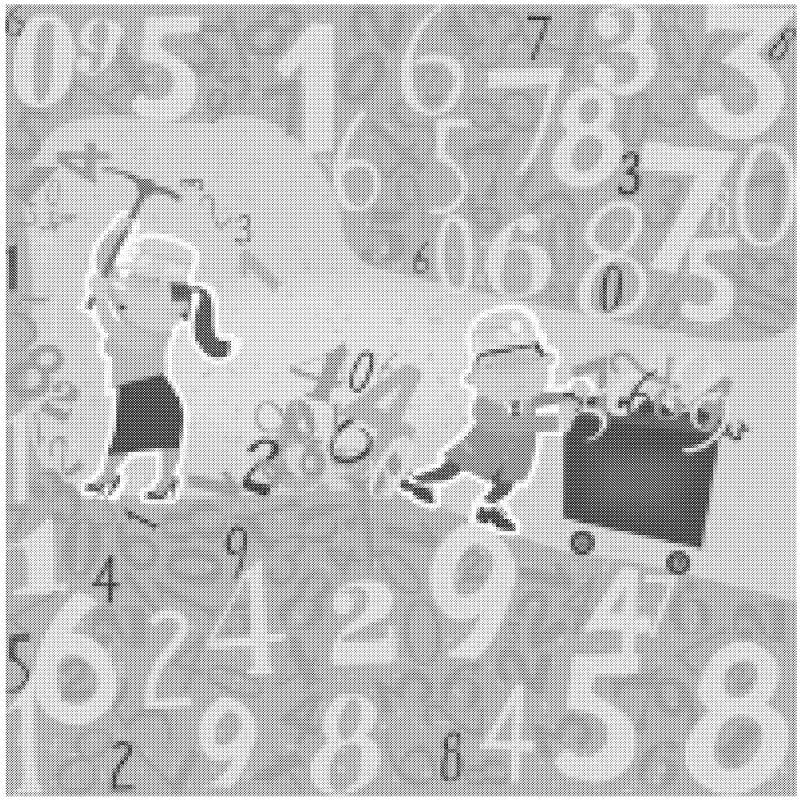
Our intention is to

For the moment, we'll keep the details of the project confidential (without publicity and without free-for-all access), only sharing details with the people involved.

(TS) Future Applications

This brief presents the main features of the research project. However, we have identified several derived benefits and future work that could be linked to this project.

(U) – Data Mining



Communications Security Establishment Canada | Centre de la sécurité des télécommunications Canada

(U) Major Functions (Skills)

- (U) Cleaning/filtering/enriching/exploring
- (U) Building/interpreting models
- (U) Finding ***known*** unknowns
- (U) Finding ***unknown*** unknowns
- (U) Presenting understandable results

SIGINT

Canada

Communications Security Establishment Canada | Centre de la sécurité des télécommunications Canada

(U) Mission Focus: DGI

- (U) Automatic image label generation (with feedback)
- (U)

SIGINT

Canada

Communications Security Establishment Canada | Centre de la sécurité des télécommunications Canada

(U) Mission Focus:

SIGINT

Canada

Communications Security Establishment Canada | Centre de la sécurité des télécommunications Canada

(U) Research Focus

- (U) Relational data: graphs and hypergraphs
- (U) Text similarity and embeddings

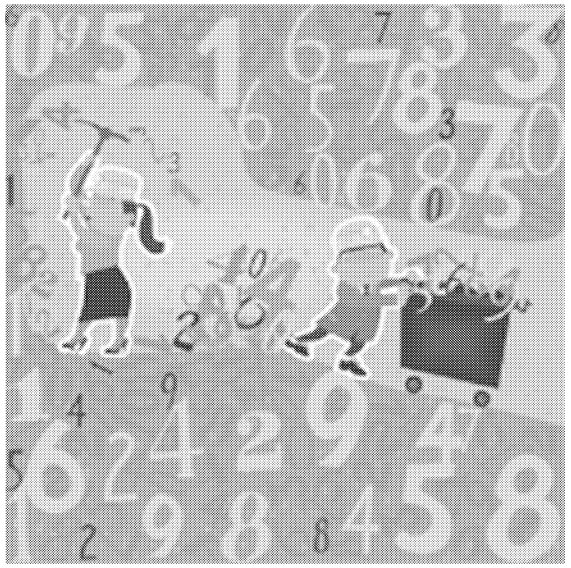
- (U) Deep learning often a tool

SIGINT

Canada

Data Science at CSE

An introduction to Big Data problems



The overall classification of this presentation is

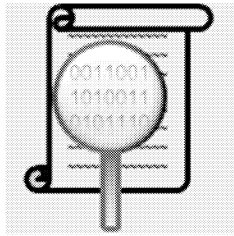
TOP SECRET//SI//REL to CAN, FVEY

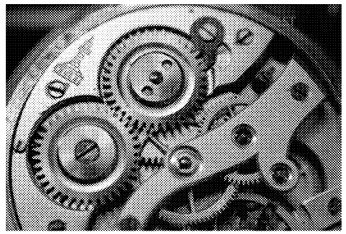
What is Big Data?

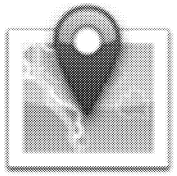
- High volume (~3.6 million Google searches/minute in 2017)
- High velocity (no time to process all data again)
- High variety (laptop, phone, fitbit, server…)
- High veracity (incompleteness, noise, duplication…)

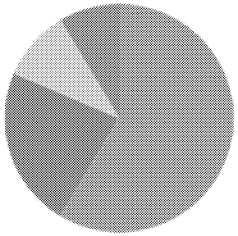
What is Data Science?

- Problem formulation
  - understanding client needs
- Data representation
- Exploratory analysis and data cleaning
- Hypothesis formulation
- Prototyping and evaluation
- Communication of results









Types of Data

- Need to understand what data to expect because machines are processing it
- **Structured data**

  [first_name, last_name, store, num_items]
- **Semi-structured data**

  { “description” : “Entry about a person”,

  “properties” : { “firstname”:“string”, “lastname”:“string” }

  }
- **Unstructured data**

Types of Algorithms

Contrasting Concerns

	Exploration	Production
Data	• Fast, unfettered access • Ease of introducing new, varied, messy datasets • Reproducibility	• Strict, governed access • Well-defined schema • Provenance & auditability
Compute Infrastructure	• High performance • Low latency, interactive • Individualized & specialized	• Scalable, high-availability • Manageable at scale • Cost amortization over many machines and users
Organization	• Individual high-achievers with lots of context & capability • Agile, able to quickly learn new skills and approaches	• Sustain operations at lowest possible cost • Robustness against unintended change

Who does Data Science at CSE

- Many teams -- analytics facilitated by tools
- AR (Research Directorate)
- TIMC (Research Directorate)
- DASI (previously CTEC)
  - Cyber defense

More on the Research Directorate later!

Pages 65 to / à 68

are withheld pursuant to section

sont retenues en vertu de l'article

15(1) - DEF

of the Access to Information

de la Loi sur l'accès à l'information

Page 69

is withheld pursuant to sections

est retenue en vertu des articles

16(2)(c), 15(1) - DEF

of the Access to Information

de la Loi sur l'accès à l'information

Page 70

is withheld pursuant to section

est retenue en vertu de l'article

15(1) - DEF

of the Access to Information

de la Loi sur l'accès à l'information

Page 71

is withheld pursuant to sections

est retenue en vertu des articles

16(2)(c), 15(1) - DEF

of the Access to Information

de la Loi sur l'accès à l'information

Any other questions about…

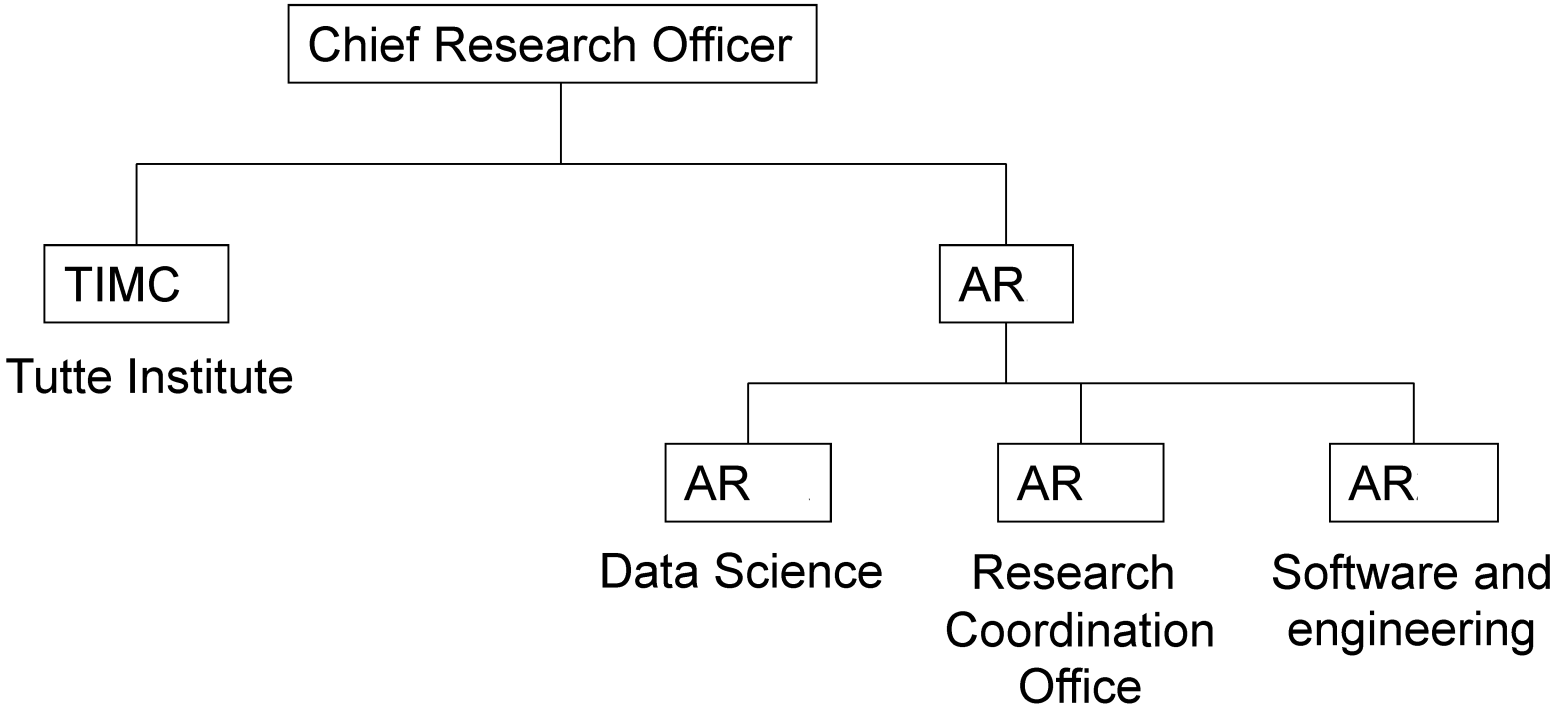
- Big Data or Data Science?
- How we get requirements from other teams?
- Tools we use?
- How we stay current?
- Our backgrounds?
- Anything else?

- Feel free to email @cse-cst.gc.ca

Research Directorate

April 2018 – created as part of Vision 2020

Research Directorate



Strategic Research

Applied Research

Strategic Research

- Removed from operational pressures
- Novel techniques and approaches
- Multi-disciplinary approach and collaboration with
-
- General results apply over a range of problems

Why create an institute?

- Tutte Institute was created (in 2009)

  "to tackle the most important scientific challenges facing the cryptologic intelligence and cyber defence communities"

Applied Research

- AR : Data Science team that we discussed earlier
- AR : Research Coordination Office (previously the Joint Research Office) to support administration of research
- AR : Software and Engineering

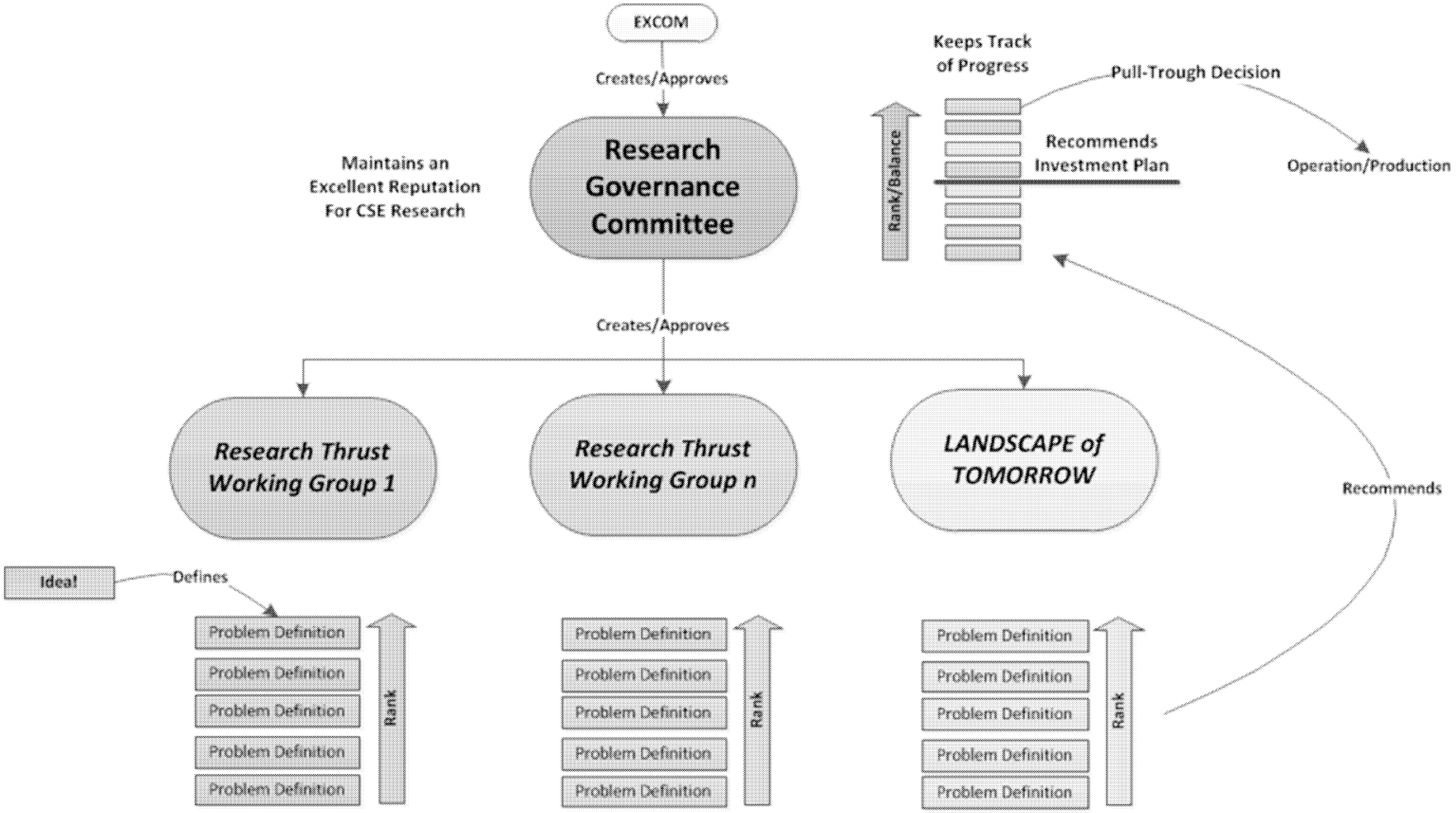
In the first year…

- Build a Governance Structure
- Revisit the Research Strategy
- Be involved with business planning for the business lines
- Provide an annual report on the progress of Research

Governance

- Research Governance Committee responsibilities:
  - Setting the research thrusts and priorities for a balanced research portfolio
  - Keeping track of progress
  - Making pull-through decisions
  - Endorsing the business/investment plan
  - Making sure CSE's research is recognized, appreciated and valued

Proposal in One Picture



Building bank of Research Activities

- May-June: Documented work currently ongoing for researchers in TIMC, AR
- July-Sept: Consultation sessions and meetings
  - themed sessions

  [crypt, computing, and cyber defense, analysis, cyber protection,]
  - "Part-B" sessions
  - individual sessions
  - Several ad hoc conversations

A few examples

Proposed Research Thrusts

- Improve **effectiveness and efficiency of analysis** to produce high impact actionable intelligence
- Extend leading edge knowledge in **secure communications and computing**

Writing the Strategy (Nov 2018)

- Thrusts play a central role
  - Diagnosis
  - Goals
  - Obstacles
  - Proposed Measures
- Landscape of Tomorrow
- Research Management

Next steps

- Validate the RAs with clients and sponsors
- Incorporate feedback from consulted staff into thrusts and strategy
- RGC approves thrusts
- Create working groups
  - Draft the TOR
  - Establish co-chairs (SIGINT and ITS)
  - Establish membership

Data Science at CSE

An introduction to Big Data problems

Communications Security Establishment | Centre de la sécurité des télécommunications

CERRID 63604442

Canada

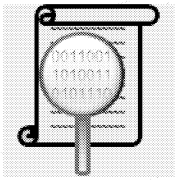
The overall classification of this presentation is
TOP SECRET//SI//REL to CAN, FVEY

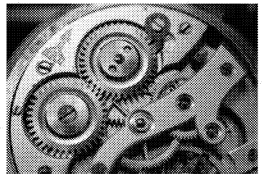
What is Big Data?

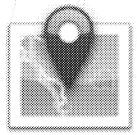
- High volume (Google Search Statistics: over 40,000 search queries every second)
- High velocity (no time to process all data again)
- High variety (laptop, phone, fitbit, server…)
- High veracity (incompleteness, noise, duplication…)

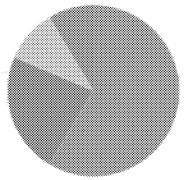
What is Data Science?

- Problem formulation
  - understanding client needs
- Data representation
- Exploratory analysis and data cleaning
- Hypothesis formulation
- Prototyping and evaluation
- Communication of results









Types of Data

- Need to understand what data to expect because machines are processing it
- **Structured data**

  [first_name, last_name, store, num_items]
- **Semi-structured data**

  { “description” : “Entry about a person”,
  “properties” : { “firstname”:“string”, “lastname”:“string” }
  }
- **Unstructured data**

Types of Algorithms

Contrasting Concerns

	Exploration	Production
Data	• Fast, unfettered access • Ease of introducing new, varied, messy datasets • Reproducibility	• Strict, governed access • Well-defined schema • Provenance & auditability
Compute Infrastructure	• High performance • Low latency, interactive • Individualized & specialized	• Scalable, high-availability • Manageable at scale • Cost amortization over many machines and users
Organization	• Individual high-achievers with lots of context & capability • Agile, able to quickly learn new skills and approaches	• Sustain operations at lowest possible cost • Robustness against unintended change

Who does Data Science at CSE

- Many teams -- analytics facilitated by tools

- Data Science researchers
  - Applied Research (AR)
  - Tutte Institute for Mathematical Computing (TIMC)
  - Data Analysis and System Integration (DASI)

More on the Research Directorate later!

Pages 93 to / à 96

are withheld pursuant to section

sont retenues en vertu de l'article

15(1) - DEF

of the Access to Information

de la Loi sur l'accès à l'information

Any other questions about…

- Big Data or Data Science?
- Tools we use?
- How we stay current?
- Our backgrounds?
- Anything else?

- Feel free to email @cse-cst.gc.ca
  Applied Research: Data Science

Research Directorate at CSE

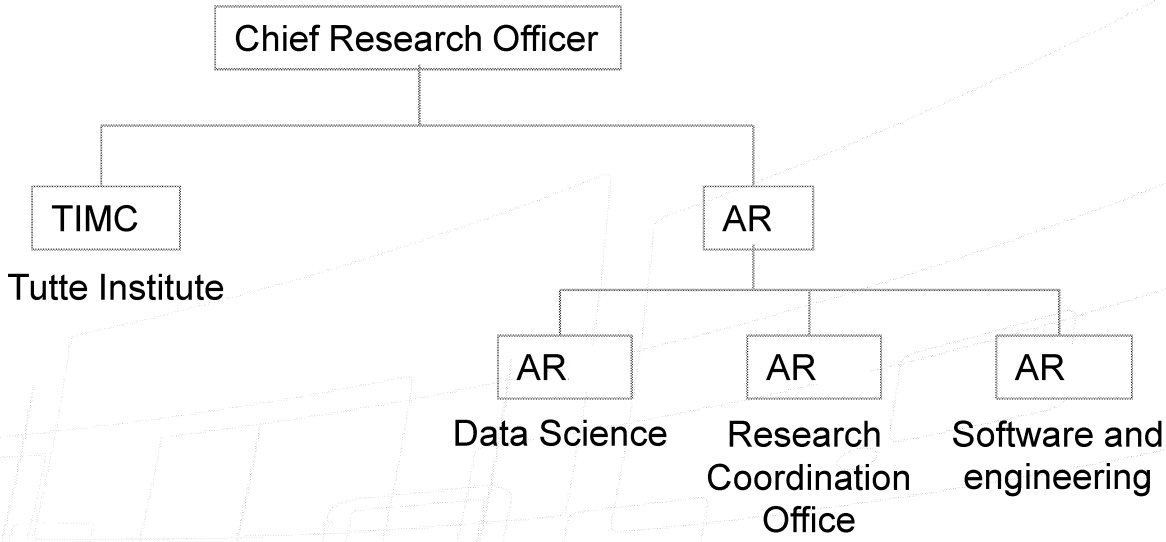
Created as part of Vision 2020 in April 2018

Communications Security Establishment | Centre de la sécurité des télécommunications

CERRID 63604442

Canada

Research Directorate



Strategic Research

Applied Research

Strategic Research

- Removed from operational pressures
- Novel techniques and approaches
- Multi-disciplinary approach and collaboration

- General results apply over a range of problems

Why create an institute?

- Tutte Institute was created (in 2009)

  "to tackle the most important scientific challenges facing the cryptologic intelligence and cyber defence communities"

s.15(1) - DEF
s.16(2)(c)

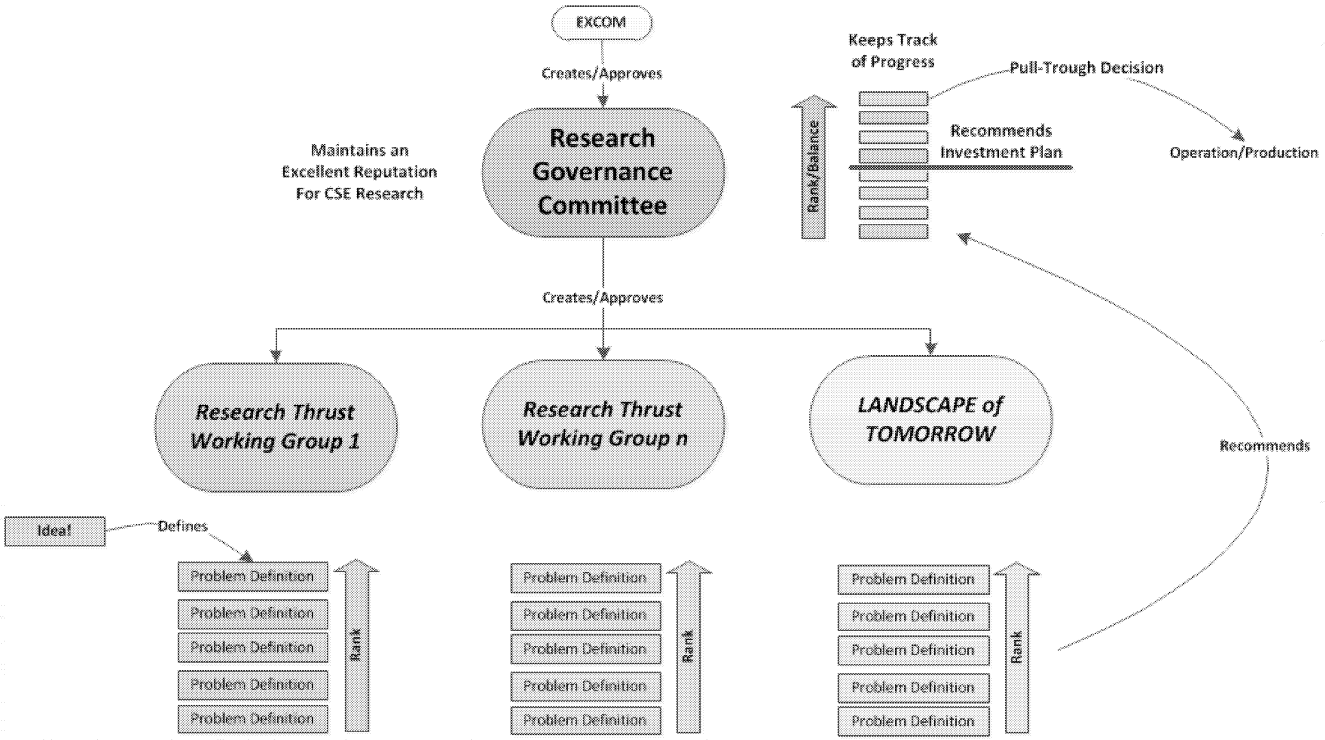
Applied Research

- AR : Data Science team that we discussed earlier
- AR : Research Coordination Office (previously the Joint Research Office) to support management of research
- AR : Software and Engineering

In the first year…

- Build a Governance Structure
- Revisit the Research Strategy
- Be involved with business planning for the business lines
- Provide an annual report on the progress of Research

Proposal in One Picture



Document Research Activities

- May-June: Documented work currently ongoing for researchers in TIMC, AR
- July-Sept: Consultation sessions and meetings – each group chose research proposals
- Grouped research activities into main themes and described in strategy

Governance of research priorities

- Research Governance Committee
  - Prioritization of Research Activities suggested Working Groups, in accordance with CSE Research Strategy

R→FUTURE | STRATEGY FOR A COMPLEX WORLD REACHING BEYOND HUMAN COGNITION

TOP SECRET//SI

RESEARCH THRUSTS

LANDSCAPE OF TOMORROW

RESEARCH ENVIRONMENT & MANAGEMENT

INTENDED OUTCOMES

Have prioritized research projects and direction.

Maintain an excellent reputation in research.

Be an integral part of a larger ecosystem.

Increase the velocity in moving research ideas to production analytics and tools.

Communications Security Establishment | Centre de la sécurité des télécommunications

Canada

Pages 108 to / à 111

are withheld pursuant to sections

sont retenues en vertu des articles

16(2)(c), 15(1) - DEF

of the Access to Information

de la Loi sur l'accès à l'information

Where are we in the process now?

- Research Activities have been written
- Strategy approved, to use as tool for governance
- Thrust WG now started and prioritized RAs
- Next step 1: WG Chairs present to governance committee for final prioritization approval
- Next step 2: Finalize annual report

Questions?

Data Science at CSE

An introduction to Big Data problems

Communications Security Establishment | Centre de la sécurité des télécommunications

CERRID 62913410

Canada

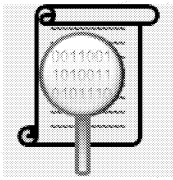
The overall classification of this presentation is
TOP SECRET//SI//REL to CAN, FVEY

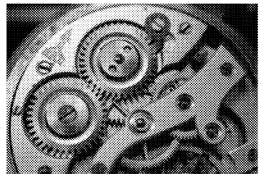
What is Big Data?

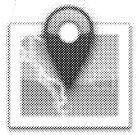
- High volume (Google currently processes 40,000 searches **per second**)
- High velocity (no time to process all data again)
- High variety (laptop, phone, fitbit, server…)
- High veracity (incompleteness, noise, duplication…)

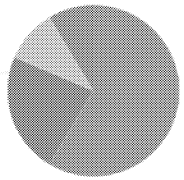
What is Data Science?

- Problem formulation
  - understanding client needs
- Data representation
- Exploratory analysis and data cleaning
- Hypothesis formulation
- Prototyping and evaluation
- Communication of results









Types of Data

- Need to understand what data to expect because machines are processing it
- **Structured data**

  [first_name, last_name, store, num_items]
- **Semi-structured data**

  { “description” : “Entry about a person”,
  “properties” : { “firstname”:“string”, “lastname”:“string” }
  }
- **Unstructured data**

Types of Algorithms

Contrasting Concerns

	Exploration	Production
Data	• Fast, unfettered access • Ease of introducing new, varied, messy datasets • Reproducibility	• Strict, governed access • Well-defined schema • Provenance & auditability
Compute Infrastructure	• High performance • Low latency, interactive • Individualized & specialized	• Scalable, high-availability • Manageable at scale • Cost amortization over many machines and users
Organization	• Individual high-achievers with lots of context & capability • Agile, able to quickly learn new skills and approaches	• Sustain operations at lowest possible cost • Robustness against unintended change

Who does Data Science at CSE

- Many teams -- analytics facilitated by tools

- Data Science researchers
  - Applied Research (AR)
  - Tutte Institute for Mathematical Computing (TIMC)
  - Data Analysis and System Integration (DASI)

Who chooses Data Science research priorities

- Research Governance Committee
  - Prioritization of Research Activities suggested Working Groups, in accordance with CSE Research Strategy

Pages 122 to / à 125

are withheld pursuant to section

sont retenues en vertu de l'article

15(1) - DEF

of the Access to Information

de la Loi sur l'accès à l'information

s.15(1) - DEF
s.16(2)(c)

Any other questions about…

- Big Data or Data Science? Applied Research?
- Governance and Prioritization?
- Tools we use?
- How we stay current?
- Our backgrounds?
- Anything else?

- Feel free to email @cse-cst.gc.ca
  Applied Research: Data Science

Extra slides

Pages 128 to / à 129

are withheld pursuant to section

sont retenues en vertu de l'article

15(1) - DEF

of the Access to Information

de la Loi sur l'accès à l'information

(U) Project Proposal on RA-070

(C)

(C) Goal

Research Activity 070 (RA-070) is the first effort prioritized by the working group associated with CSE's strategic research thrust:

The primary goal of the RA-070 is to The program is encapsulated by three related objectives:

The philosophy behind this research project is to build fundamental science and techniques in the domain of and to demonstrate the applicability and effectiveness of those techniques in a realistic application scenario. Our objective is to publish a significant portion of our work in order to contribute publically and influence the greater research community that already care about the problem. Ideally, we would want to Hopefully,

Our intention is not to

With the present research activity, we want to develop knowledge by working on innovative techniques and by making sure we put all the possible measures to respect the privacy of Canadians.

[1] R-Future – Strategy for a complex world reaching beyond human cognition:

51595115

(C) Privacy Considerations

We have several measures that we use to protect the privacy of Canadians during the course of this project.

Note for the purposes of this project,

(C) Imagineering

In this section, we provide an overview of an example approach to tackling our objectives. It is very important to keep in mind that this proposal fits in the category of This document is not trying to be precise and specific, but it is simply aiming at providing a conceptual and approximate picture of the proposal. The final implementation, if successful, might end up being significantly different. Management oversight will be maintained as the project is evolved.

The main idea is to

The following diagram illustrates the conceptual approach.

Steps

In summary, the proposed approach will

The following parties support this project proposal. They acknowledge understanding and accept the terms and conditions of this agreement.

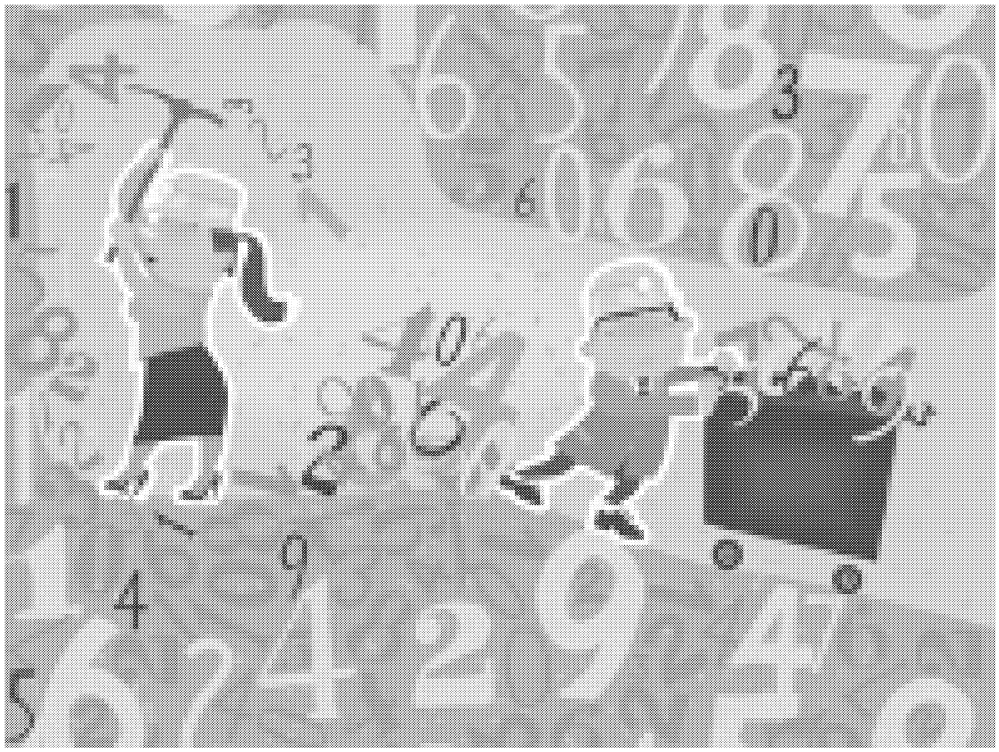
_____ _____

CSE Chief Research Officer Date

_____ _____

Dir Date

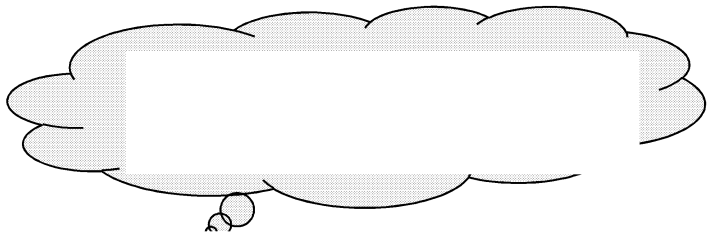
Communications Security Establishment Canada
Centre de la sécurité des télécommunications Canada



Data Science at CSE

- with focus on CTEC

Major Functions (Skills)



- Building/interpreting models

  Note mutually beneficial goals: TIMC works on deep learning for text; CTEC gets a way to

- Enriching/cleaning/filtering/exploring data
  - Triage/prioritization
  - Pattern detection
  - Grouping/clustering
  - Classification
- Telling a story with your data -- understandable and actionable results

Current Status

- Strong acknowledgment that Data Science is valued and useful in cyber defense
  - Creation several years ago of a position in CTEC that staffs with a data scientist (rotating)
  - CTEC : data science backgrounds
  - CTEC : additional Data Science tests Talked to about
  - CTEC Data Scientists approved to be part of workshops and reading group with TIMC
  - cyber analysts at training
  - BIG DIG Data Science Friends team

Current Concerns

Data Science Applied R&D team for Cyber Defence

- Team whose focus is Data Science
- Maintaining and enhancing skills in Data Science
- Understanding and adapting to trends
- Time to see big picture and work toward longer-term problems
- Applying strategic ITS work now done in TIMC
- Focus assists with recruitment and retention

- DS support team very successful at Big Dig
- Many individuals asked for DS help at GeekWeek

What about operations?

- Integrations of researchers into mission teams
- Some research officers
  – help gather research requirements
- Any team can seek advice and guidance from Applied R&D team - know where to go!
- Want collaborative vs transactional relationships
- Researchers continue training (…)
- Note:

Proposal – Option A

- Set up a Data Science Applied R&D team in CTEC

- Considerations:
  - Refining skills and rewarding research are very important
  - Can be difficult to get “permission to take time” for research and for training others under operational constraints
  - Should take into account lessons from TIMC: how to engage clients and get requirements, how to contribute to research community, establish links internally and externally, reading group,

Proposal – Option B

- Create ITS positions within a Data Science Applied R&D team with dual mandates in the proposed new Research Directorate

- (U) Considerations:
  - Could take advantage of knowledge and experience from /TIMC to contribute to Cyber Defense mission
  - Proposed Chief Research Officer would assure the research portion of job is cultivated and rewarded
  - Need to assure that there are research officers in mission as well to help define requirements, assist with pull-through, and ensure operational success
  - Already training together (TIMC, CTEC)
  - Second party interactions easier with one research area

CLASSIFICATION: SECRET

Research Activity # and Title:	Update date
RA-069	2019-03-05

Project Proposal

(U) Goal:

(U) Motivation:

Current TRL: 5 **Target TRL:** 7

(PB) CSE client:
(PB) CSE sponsor:
(U) Other clients: TBD
(U) Horizon (S/M/L): S (Short)
(U) Activity Type (SR/AR/ED): AR (Applied Research)
(U) This activity is about enabling/supporting (E/S/ES): S (Supporting)

Possible measures of success:

- (U) Addressing requirements identified needing research support
    - (U) High precision in the creation of alerts for the requirement (low "false positive" rates)
    - (U) High recall in the creation of alerts for the requirement (high rate identifying any potential issue)
    - (U) Precision/recall sufficiency evaluated by analysts in

Level of classification of research: SECRET

Research Activity # and Title:	Update date
	2019-03-05

Research Response

Avenues for investigation / Suggested Roadmap for FY 2018/19:

Skills and knowledge required:

- (U) Data mining and statistics theoretical knowledge
- (U) Practical experience applying data mining and statistical techniques
- (U) expertise (provided by
- (U) In depth knowledge of system (provided by
- (U) Cyber Defence experience – asset not strictly required

Research PoC: TBD

Possible resource allocations (FTEs, Capital/O&M):

Alternative resource allocations:

Collaboration (R/IR/J/IC/C/2P/OGD/E): J (Joint between research and client)

Challenges / dependencies:

Gaps:

s.15(1) - DEF

s.16(2)(c)

CLASSIFICATION: SECRET

Research Activity # and Title:	Update date
	2019-03-05

Detailed Project Proposal

Goal:

Context:

Potential Applicability:

Risks:

Constraints:

Data available for research:

References: