Problem 1
Dynamic Model Conditioning Problem
Imperial Oil

When we do DMC (dynamic matrix control) projects, the work process for development of the model can be summarized as follows

1) Conduct plant testing to generate a dataset. Typically use an automated tester (e.g. AspenTech Smartstep/Calibrate) which means we may have some closed loop data in the dataset.
2) Run model identification in Aspen Model software (use Subspace ID technology which is a multi-input multi-output (MIMO) ID technique versus FIR (finite impulse response)
3) Assess condition of matrix (e.g., using relative gain arrays)
4) Fix matrix to remove ill-conditioning and to impose process knowledge (e.g., parallel manipulated variables must be collinear, such as furnace pass flows or dual reboilers, .... and parallel controlled variables such as flows and valves in same line .... Plus much more)
5) Finally, we may compare the conditioned matrix with other models that are less empirical (e.g. available models used for real-time optimization).

The issue is that step 4 is an onerous and difficult process that relies on the process knowledge in the brain of the engineer (this is a rate-limiting step because it takes a lot of experience to get people to the level of being able to condition a dense matrix with many intersecting groups).

Sadly, we tend to resort to an Excel sheet where I end up calculating many ratios of rows and columns in the matrix and the many gains end up becoming dependent on other gains … .this again relies on experience of the engineer to have a strategy to do this piece of work.

When we try to explain this to new engineers, one of the sentences that we often quote is …. We are trying to rationalize the degrees of freedom for DMC to more closely match reality … e.g. we want to get rid of the numerical degrees of freedom that DMC may try to exploit when certain constraint sets become active

On the other hand, we have a bunch of applications (training simulators, Real Time Optimization models, ....) .... Is there a way to exploit the 1st principles nature of these models to extract the process knowledge that we must impose on our unconditioned matrix? Is there a smarter way to solve the problem in item 4 … there are some tools in the Aspen software and XOM tools that address this … but none take advantage of information from other models …. They all rely on the user to impart their process knowledge …. At the end of the day, we prefer to retain full control of the gain matrix and fix it manually versus allowing the tools to start adjusting the gains (we tend to end up with smaller gain moves with my method)

As you can tell, we would love to have a better way to solve this problem.
Problem 2
Sketch to 3D
Autodesk Research

The purpose of this project is to develop a system that can automatically generate a 3D object from a 2D sketch. Once an underlying neural network understands the semantics of the 2D sketch and the 3D object, user can continue editing the object in 2D or 3D: changes made in sketch will propagate to 3D, and vice versa – changes made in 3D will propagate to 2D. Because the system understands the semantics of the whole object, as well its' components, it can also help the user to autocomplete the object in 2D and 3D. The following diagrams illustrate the concept:

Step 1. User sketches a square in the “Front” view, and system generates a corresponding box in the “Perspective” view.

Step 2. User sketches a triangle on the top of the square in 2D. The neural net recognizes the object as a house, and correctly positions a triangular prism on the top of the box in 3D.

User sketches a door in the “Front” view, and the system creates the door segment in 3D. Changing the position or the size of the door in 2D is reflected in 3D, and vice versa.
There are a couple of possible directions of the research that we identified us potentially interesting.

1. To pose the problem as a multilingual translation task, where different representations (sketch, 3D mesh, csg, voxel etc) are considered as different languages. The sequence of operations are translated into internal embedding of the shape, and then reconstructed in different representation. The advantage of this approach is that changes made in one representation are automatically translated into other representations. (some of the relevant papers in language translation: https://arxiv.org/pdf/1812.10464.pdf, https://arxiv.org/abs/1611.04558)

2. To use knowledge graph for accurate classification and prediction of missing semantic segments. Once the system recognizes that user draws a house, for example, it can use the knowledge graph of the house to recognize and position the segments of the house in 3D (a relevant paper: https://arxiv.org/abs/1612.04844)

In addition to standard 3D datasets (Shapenet, ModelNet), we also have a dataset from one of pir products called TinkerCAD that provides the construction history. The dataset is very large, however, and might need additional preprocessing to prepare it for solving the problem.

The solution is targeted for architects and industrial designers, and not for biological models.
Problem 3
Gas Turbine Engine Performance Evaluation: Prediction of Long-Term Degradation Patterns
Tecsis

Background: The performance of GTE gradually deteriorates during the operation even under normal engine operating conditions. There are two major mechanisms contributing to the performance deterioration. One is the rapid (short-term) deterioration due to fouling and congestions of the particles in the air and in the combustion gases on the rotating and stationary blades. Another type of deterioration occurs due to structural degradation of the GTE parts such as flow path damage, surface erosion and corrosion, distortion, etc. The resulting performance deterioration is irrecoverable by cleaning and washing, but will get worse with usage, unless the degraded parts are repaired or replaced. The predicting of GTE degradation and future health state is highly complicated and challenging due to the structural complexity, non-stationary operating conditions and great uncertainty associated with gas turbine design, manufacturing, ambient and environment condition, etc.

The main motivation and challenge for PI project is proposing a performance index which have simple definitions from large amount of inputs and principles of thermodynamics and effectively represent engine performance, especially can provide meaningful short-term and long-term degradation patterns and trends. Another challenge is studying the feasibility of applying performance measurement indices into prediction of long-term degradation and fault modes. Several research works have been done to investigate the relation between fault modes and consequence degradation symptoms of the turbines, whereas fewer works have been performed to study the growth of the degradation symptoms over time.

Problems and challenges: In our preliminary research, performance analysis for a low power rating and partially loaded gas turbine engine, was carried out by a proposed model-free data analytics approach using available data. Several performance indexes (PI) are proposed considering the ratio of Power and EGT output to combination of parameters related to fuel consumption. The main tasks for proposed research include:

- Identify the rapid and short-term performance deterioration patterns, i.e., compressor fouling-washing effects using the proposed performance indexes.
- Examining the trends of the long-term deterioration symptoms and propose candidates for the prediction models (functions including 2nd Degree Polynomial function, Logarithmic function, Logistic-Exponential function, and ANFIS-Based Prediction Model) of deterioration patterns.
- Preparing codes which receives long-term performance deterioration symptoms as time series of possible related parameters and find the optimized coefficients of the models by using data training process (evaluating the local standard error and whole model error).
- Choosing appropriate evaluate ranking criteria and compare different prediction models for best prediction results.
Problem 4
Web Archives and Big Data for the Humanities and Social Sciences
University of Toronto Libraries

**Problem:**
Big Data, in the form of born-digital historical sources, has the potential to reshape the humanities and social sciences. The sheer volume of cultural information generated and, crucially, preserved every day presents exciting opportunities for historians, political scientists, sociologists, and other scholars. Much of this information is captured within web archives containing billions of URLs, including individual homepages, social media sites and feeds, institutional pages, and corporate sites. This material introduces important new avenues of research for historians working in diverse fields. Historians broaching topics dating back to the mid-1990s, for example, will find their projects enriched by this web data: military historians can use forum posts by soldiers; social historians can explore aspects of everyday life through blogs, homepages, comments, and guestbooks; and economic historians can explore commercial activity online.

Yet the tremendous opportunities afforded by web archives are dampened by the sheer challenge of working with all that data. We have more information now than ever before, but the scale overwhelms the scholars who seek to interrogate their contents.

So, what is there to be done? This is a big problem, but the major questions include:
- How can historians leverage new advances in statistics, natural language processing, and other approaches to working with data at scale to uncover patterns in large historical data?
- Are there clusters of similar documents that can be extracted from a web archive?
- Are there best practices for teaching historians and other social scientists about how to work with unstructured data at scale?

**Data:**
Since 2005, the University of Toronto Libraries has been actively curating, crawling, and preserving archived web pages. You can see their collections at [https://archive-it.org/organizations/75](https://archive-it.org/organizations/75). Two collections are particularly useful: the Canadian Political Parties and Interest Groups collection, which has been a quarterly crawl of every political party website registered with Elections Canada as well as an assemblage of active political interest groups; and the Canadian Labour Unions collection which is a national quarterly crawl of public and private-sector unions as well as labour federations. One can imagine many
use cases for this sort of data: scholars working on political or labour issues in any time period since 2005 could use this well-crafted information.

Yet while the collections are impressive, their use has lagged. This is for several reasons:

1. Traditional keyword search approaches do not scale for collections of this size. A search for “Stephen Harper” leads to 724,829 results, with few faceting options. Unless there is a very specific research query, scholars need to work with this data using data mining or text analysis approaches. This leads to the next problem…

2. Bulk access to content comes in the form of WebARChive (WARC) files, an ISO standard. However, few specialized tools exist for working with WARC files. Fortunately, our Archives Unleashed project can transform WARC files into more traditional research derivatives. Even then, however…

3. Humanists and social scientists are not well-equipped to work with data at scale. The plain text of this collection (just the HTML documents) is 82GB in size. This requires specialized tools and approaches, which is where humanists and social scientists are lagging.

Solutions?:
What can we do? How can we garner insights from this amount of data? Approaches to date have involved using approaches from the natural language toolkit, or leveraging the hyperlinks to find documents of interest using PageRank or other approaches. At the IPSW, attendees might be interested in using large amounts of unstructured text (or image data, or PDFs, etc.) to imagine how we can find better statistical approaches to answer questions such as: What is in a collection? What is most relevant to a given research question? Are there clusters of documents or images? The questions are nearly endless, but there may be useful general approaches to explore and thus facilitate more interdisciplinary collaboration around large datasets.
Problem 5

Mathematical methods to estimate burden of influenza-attributable complications

Sanofi Pasteur

Annual seasonal influenza epidemics remain a serious public health concern and are estimated to result in approximately 12,000 hospitalizations and 3,500 deaths per year across Canada. Adults 65 years of age and older are disproportionate affected by the serious complications of influenza, and can account for up to 70% of influenza-associated hospitalizations and 90% of influenza-associated deaths in severe influenza seasons. While it is generally accepted that acute influenza infections can trigger or exacerbate other medical conditions that can lead to hospitalization and deaths, it can be difficult to accurately measure the full scale of influenza burden due to inconsistencies in case definitions and influenza testing. In addition, recent clinical studies of influenza vaccines have demonstrated larger than expected impact that vaccination had on less specific outcomes such as cardiovascular and all-cause hospitalizations. Therefore, a model is needed to reconcile these conflicting data, and to help to understand the potential underreporting of influenza cases during routine disease surveillance to determine the proportion of hospitalizations and other serious outcomes that are attributable to influenza.

Influenza is estimated to have an annual attack rate ranging from 6-20% depending on age, and complications from influenza can lead to hospitalizations caused by direct respiratory causes (e.g. pneumonia), and indirect causes (such as cardiovascular events). A number of studies in the US have used regression models fitted to hospitalization data in various healthcare databases to estimate the proportion of these hospitalizations that are attributable to influenza. These estimates (summarized in Table 1), range from 16.8/100,000 in adolescents and young adults, to 1669.2/100,000 in the very elderly. While these rates have been used to inform policy decisions and mathematical modelling of influenza, recent clinical studies of influenza vaccines have demonstrated substantially higher rates of hospitalizations that are attributable to influenza. Randomized controlled clinical studies of a high-dose influenza vaccine indicated for adults 65 years of age and older has demonstrated reductions in all-cause hospitalizations compared to a standard-dose vaccine at a rate of 705/100,000 in community dwelling seniors and 1,456/100,000 among nursing home residents. These rates are deemed to be unexpectedly high especially in light of the fact that this is the difference in the hospitalization rate between people receiving two different influenza vaccines (i.e. not compared to unvaccinated individuals), and is higher than the total overall hospitalization rates reported in earlier studies. These findings have resulted in skepticism from some influenza experts about the validity of this and other clinical studies of influenza vaccines. It is also possible that there is underestimation of influenza-attributable hospitalizations, and that it is only when an effective intervention becomes available that the true impact of influenza becomes apparent.
Influenza is a very complex disease and it is likely that our understanding of the disease epidemiology through existing studies represents only the tip of the iceberg. This represents an opportunity for a mathematical approach to reconcile routine disease surveillance studies focused on specific outcomes, and the recent clinical studies that demonstrate additional impacts influenza can have on less specific outcomes, specifically a missing value analysis that attempts to estimate influenza-attributable burden. This work is important for providing a more accurate estimate of the influenza-attributable burden (attack rate, hospitalization rate, mortality rate), in order to guide evidence-based immunization policy-making. It is also vital to the development of realistic mathematical models of influenza, and to accurately evaluating the impact of influenza vaccines, thus enabling the design of optimal vaccination programs that can best reduce the burden of influenza.

**Table 1: Attack rates in modeling studies and meta-analyses of randomized controlled trials of influenza vaccines**

<table>
<thead>
<tr>
<th>Sources</th>
<th>Age Group (years)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>&lt;5</td>
</tr>
<tr>
<td>Moinari 2007</td>
<td>20.3%</td>
</tr>
<tr>
<td>Clements 2011 (Direct only)</td>
<td>40.0%</td>
</tr>
<tr>
<td>Clements 2011 (Direct + Indirect)</td>
<td>24.0%</td>
</tr>
<tr>
<td>Jayasundara 2014</td>
<td>15.2%</td>
</tr>
<tr>
<td>Diazgranados 2014</td>
<td></td>
</tr>
<tr>
<td>Chit 2015</td>
<td></td>
</tr>
<tr>
<td>Tokars 2017 (Statistical Estimation)</td>
<td>13.2%</td>
</tr>
<tr>
<td>Tokars 2017 (Meta-analysis)</td>
<td>12%</td>
</tr>
<tr>
<td>Somes 2018</td>
<td>13%</td>
</tr>
<tr>
<td>Somes 2018 (NA only)</td>
<td>20.8%</td>
</tr>
</tbody>
</table>

**Table 2: Influenza-associated excess hospitalization rates per 100,000, cardiorespiratory definition**

<table>
<thead>
<tr>
<th>Sources</th>
<th>Age Groups (Years)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>&lt;1</td>
</tr>
<tr>
<td>Zhou 2012</td>
<td>151</td>
</tr>
<tr>
<td>Thompson 2004</td>
<td>113.9</td>
</tr>
<tr>
<td>Mullisaly 2005 (Low risk)</td>
<td>-</td>
</tr>
<tr>
<td>Mullisaly 2005 (High risk)</td>
<td>-</td>
</tr>
</tbody>
</table>

**Table 3: Impact of high-dose influenza vaccine on hospitalization rates per 100,000 in adults 65+**

<table>
<thead>
<tr>
<th>Sources</th>
<th>Reduction in Hospitalization Rate (versus standard dose influenza vaccine)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Pneumonia</td>
</tr>
<tr>
<td>Diazgranados 2015</td>
<td>294</td>
</tr>
<tr>
<td>Gravenstein 2017</td>
<td>120.2</td>
</tr>
</tbody>
</table>
References

Problem #6: Improvements In CRISPR Based Gene Editing

Submitted by: Tesseraqt Optimization Inc.

The human genome is a sequence of nucleic acids, encoded as DNA and consists of roughly 3 million base pairs. It can be thought of as a string of 3 billion characters consisting of A,C,T or G. The human genome has approximately 20,000 regions (called “genes”). As strings, they range in length from 20 characters to 2.2 million characters long. By modifying the string of characters in one (or more) of these regions, you can alter the behaviour of the genes, and ultimately the behaviour of the organism. This is called gene editing, which is a form of genetic engineering.

Recent advances in gene editing technology have captured the imagination of many researchers. The recently developed CRISPR-Cas9 system for gene editing is quicker, cheaper, and more efficient than other existing gene editing methods, furthermore. The potential applications are enormous both in number and impact. It holds great promise in human therapy, as clinical trials using CRISPR-Cas9 are now underway in the United States, Europe and China.

CRISPR stands for “clustered regularly interspaced short palindromic repeats,” a recently discovered feature in the viral defense system of bacteria. By using a special protein (for example the CRISPR associated protein 9, “Cas9”) for cutting, and a special sequence of nucleotides (the guide RNA, “gRNA”) for targeting, scientists are able to target specific areas of the DNA to be cleaved and edited. Once the DNA is cleaved, the cell's own DNA repair machinery is used to add or remove pieces of genetic material, or to edit the DNA by replacing an existing segment with a custom DNA sequence.
The gRNA performs its guidance by binding to the target DNA, allowing the Cas9 protein to perform the cut. This means that programming CRISPR system is as simple as using a gRNA that binds to the desired area on the target DNA. One could essentially do this by using the same sequence for the gRNA as for the DNA to be targeted. The gRNA + Cas9 system scans the target DNA for areas that match the gRNA, and when a match is found, it binds and makes a cut. In practice however, it is not this simple. A “perfect match” doesn’t always imply an edit will occur, and edits can occur without a perfect match. Moreover, the CRISPR system doesn’t perform only a single edit, at one location. It will cleave wherever the gRNA binds to the target DNA. These issues are exacerbated by the fact that gRNAs can be only 20 characters long. Given the fact that DNA is 3 million characters long, with an alphabet of only 4 characters, there may be many places where a given 20 character string occurs (or is very similar). One particular type of edit that is of interest is any change that leads to a “knockout” of the gene. Essentially altering the gene in any way that disables it’s normal function.

This leads to the main focus of this problem. Given a desired area of DNA to target, can one predict the effect of using a particular gRNA in the CRISPR system? In particular, given a gRNA, one would like to know a) the probability it will perform the desired edit and b) the probability it edit elsewhere in the genome. These edits are called “on-target” and “off-target”, respectively. Off-target edits are important to predict because they may result in catastrophic changes to the organism.

If one was able to find these probabilities, one could optimize the selection of gRNA in the following steps:

1) Select the gene you wish to be edited, and the type of edit desired (insert, delete, etc)
2) Locate all 20 character strings in the gene that would be candidate binding sites for gRNAs
3) For each candidate gRNA:
   a) Predict the probability of successfully editing the target gene
   b) Search the 3 billion character string for similar 20 character strings (the potential off-target sites)
   c) Predict the probability of an edit at each of these potential off-target sites
   d) For each off-target edit, predict the expected biological outcome if that edit occurs

We solicit progress toward solutions of the following tasks:

1) Increasing the speed and accuracy of step 3b
2) Increasing the accuracy of steps 3a and 3c for “knockout” edits specifically
Although these are difficult questions, recent progress has been made.

For example, a number of solutions have been presented for the Task 1 of search speed and accuracy including Cas-OFFinder, CRISPOR, CHOP-CHOP, e-CRISPR, CRISPR-DO and CROP-IT. Each of these use different methods to perform to search and/or perform the search to various levels of completeness. As for Task 2, a few different options exist, such as the MIT web server, CRISPOR and CRISPR-DO, CHOP-CHOP, CROP-IT and the CFD web server.

Finally, recent work has been carried out to use machine learning approaches for Task 2, called Azimuth and Elevation, in a collaboration between the Broad Institute and Microsoft. Their work can be found here:


As a potential starting point, datasets and code relating to Azimuth and Elevation will be made available to those who work on the project. Many other public data sets for CRISPR experiments as well as bioinformatics will be made available. Finally, AWS access for high performance computing will be made available (free of charge) to groups working on this project.

Some other references that may be useful can be found here:
https://academic.oup.com/bioinformatics/article/34/17/i656/5093220
Strategies to Determine Off-Target Effects of Engineered Nucleases