**1. Network Analysis of Entrepreneurship Ecosystems**

Description: We are interested in implementing network analysis of global/regional/country entrepreneurship ecosystems using nearly real-time big data sources and traditional data sources (e.g., surveys, official data). We seek to identify clusters/patterns across ecosystem actors (e.g., startups, incubators, mentors) and relationships between actors vis-a-vis policy/ecosystem outcome data (e.g., growth, investment size, innovation indicators) to provide a wider evidence base for data-driven policymaking.

Prerequisites: Python or R (required); experience accessing NoSQL DBs, Spark (preferable); Azure/AWS experience, Gephi (bonus)

Contact: Prasanna Lal Das (plaldas@worldbank.org)

**2. Dynamic Topic Modelling for Entrepreneurship Ecosystems over time**

Description: We are interested in building dynamic topic models for startups/companies over time to check for industry shifts and appearance of new technologies in country entrepreneurship ecosystems. We want to experiment having a more dynamic approach using nearly real-time company data (e.g., company website, social media, search data) as "signals" for new themes/technologies in entrepreneurial ecosystems, and compare this with standard/official industry classifications in the context of the fast-growing digital space.

Prerequisites: Python or R (required); experience accessing NoSQL DBs (preferable); Azure/AWS experience (bonus)

Contact: Prasanna Lal Das (plaldas@worldbank.org)

**3. Using blockchain data of value chains, from farmers in poor areas to end consumers in developed countries, to identify and forecast bottlenecks and propose best destination market**

Description: We are piloting a DLT solution in Haiti, that offers traceability and payment, that allows farmer to the sell directly to consumers in North America. We would like to use AI to analyze the data from those transactions (we have access to anonymized data), to forecast bottlenecks (shortages in shipping capacity, customs clearance …) and, based on an algorithm that takes into account previous transactions, to propose the best market where to try to sell the products to obtain the maximum price (kind of finding the faster route, but finding the most profitable destination).

Prerequisites: Experience with Javascript (required), Ruby (preferable), Python (bonus). Bonus: Knowledge of HCI, d3 data visualization, backbone

Contact: Emiliano Duch (educh@worldbank.org)

**4. Financial crisis prediction**

Goal is to improve upon the "state of the art" crisis prediction models by keeping it very simple -- using out-of-the-box ML techniques such as gradient boosted trees and the traditional macro and financial data which is available from IFS, WDI, BIS, Bloomberg, etc. (external debt, credit gap, VIX, spreads, etc.). This should already produce superior ROC AUC results. If promising, we could offer to incorporate more innovative predictors (e.g. based on news flow, search). Ultimately such efforts will fit into the multi-dimensional crisis / early warning framework for the WBG.

The "left hand side" data can be found in this IMF paper that just came out which updates the global crisis database:
https://www.imf.org/en/Publications/WP/Issues/2018/09/14/Systemic-Banking-Crises-Revisited-46232 .
We can provide in electronic format when needed.

Contact: Erik Feyen (efeijen@worldbank.org)

**5. Autonomously identify discrepancies between drones in a blood delivery fleet**
Zipline delivers blood and medical supplies to remote hospitals via drone. The goal of this project is to use machine learning on the drone telemetry database to autonomously identify discrepancies between our drone wings, bodies, and batteries which could indicate damage, manufacturing problems, or degradation.
Prerequisites: Python (required); experience with S3 is a plus

Contact: Emma Schott (emma@flyzipline.com)

**6. Cystic Fibrosis Bubble Detection and Measurement**
Cystic fibrosis is a genetic disorder causing progressive, pulmonary failure and is universally fatal. Current outcome measures used in clinical trials for cystic fibrosis have many limitations. We have developed a bioassay that uses sweat rate in order to measure the function of the abnormal protein and thus could be used as an outcome measure to more precisely measure the efficacy of new, emerging, transformative treatments for cystic fibrosis. In order to use this bioassay, we must measure the radius of sweat bubbles in images. We are interested in creating a program that uses machine learning to accurately detect and measure these bubbles.

Contact: Carmen Strassle (strassle@stanford.edu)

**7. Hybrid Distributional and Definitional Word Vectors**
Word vectors are typically arrived at through distributional statistics (such as co-occurence), but it is surprising that the most logical source of words' meanings - dictionaries - are not leverages in the process. We want to investigate the ability to use word definitions in the process of creating word vectors.

Contact: Andrey Kurenkov (andreyk@stanford.edu)

**8. Illuminating the Druggable Genome (IDG)-DREAM Challenge**
Mapping the complete target space of drugs and drug-like compounds, including both intended ‘primary targets’ as well as secondary ‘off-targets’, is a critical part of drug discovery efforts. This DREAM (Kaggle-style) competition involves predicting the affinity of drug-kinase pairs in a collaborative filtering type setup.
Pre-requisities: a high level programming language, e.g. Python, R or Julia.

Contact: David A Knowles (dak33@stanford.edu)

**9. Crack Pattern Characterization from 3D Laser Scanner Data**
Our research goal is to leverage 3D laser scanner data (3D point cloud & RGB) of concrete samples to identify crack location and quantify its geometric parameters and severity. The ultimate ultimate goal of

the study is to create a digital twin model of an infrastructure through automated scanning. The student is expected to utilize data/image-based processing tools for this application.

Contact: Tanay Topac (ttopac@stanford.edu)

**10. Classifying lymphoma and healthy status from prospective cohort using genomics data**
We are interested in building classifiers that will label future lymphoma and future healthy status based on gene expression, DNA methylation and microRNA expression from individuals that were healthy at the moment of blood sampling but that developed cancer disease during the 17 years follow-up.

Contact: Almudena Espin-Perez (aespin@stanford.edu)

**11. Heat People, not Buildings**
We heat buildings mainly to make occupants comfortable, wasting most of the energy and related expenses. Imagine a small device that directs a gentle, focussed stream of warm air toward only the people present, sort of a "pan-tilt-zoom" (PTZ) hair dryer controlled by a person-recognition ML algorithm. Now imagine several of these in a room sharing data and coordinating their actions. Turn down your thermostat to cut heating bills, conserve energy, and reduce global warming! (Great project for an entry-level ML + robotics team.)

Contact: Jerry Kaplan(jerrykaplan@stanford.edu)

**12. Predicting mechanical properties of soft matter using machine learning methods**
We are interested in predicting the mechanical properties of complex materials from the geometrical configuration of the microstructure. Specifically, the material is a concentrated emulsion made of microscale water droplets suspended in oil (similar to mayonnaise). We have experimental data on the shape, position, velocity, and packing configuration of each droplet when they flow in a microfluidic system. We aim to use machine learning to 1) efficiently capture the broad range of complex drop shapes, 2) examine how the shapes evolve with changing flow conditions, and 3) predict when the emulsion starts to become unstable and undergo break-up.
Relevant background papers:
http://pubs.rsc.org/en/content/articlelanding/2013/sm/c3sm51843d#!divAbstract
http://aip.scitation.org/doi/full/10.1063/1.4994668

Contact: Prof. Sindy Tang for more info (sindy@stanford.edu) https://web.stanford.edu/group/tanglab/

**13. Running patterns and global health**
Use Strava API (https://labs.strava.com/) and publicly available social, economic, or health data (e.x. WHO) to mine new hypotheses on physical activity and global health (see https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5774986/ for good examples).

Contact: Rachel Jackson (rwj1@stanford.edu)

**14. Human Protein Atlas Image Classification**
This is a Kaggle competition involving predicting cellular localization of proteins from multichannel images. The most natural approach would be to use conv nets but there is plenty of opportunity for innovation given the structure of the problem.
https://www.kaggle.com/c/human-protein-atlas-image-classification

Prerequistes: Python.

Contact: David A Knowles (dak33@stanford.edu)

**15. Predicting disease progression from a few observation [theory/algorithms]**
In clinical practice and biomedical research, measurements are often collected sparsely and irregularly in time while the data acquisition is expensive and inconvenient. Modelling such data has been usually done using linear models (https://arxiv.org/abs/1809.08771). In this project the task is to investigate non-linear approaches to modeling latent representations of patients.

Contact: Łukasz Kidziński (lukasz.kidzinski@stanford.edu)

**16. Identifying different types of cells in the tumor microenvironment from pictures of stained tissue sections.**
To better understand cancer biology and , researchers look at the frequencies of certain types of cells (immune cells like macrophages, etc.) in a tissue. Computer vision methods allowing for identification of these cells at scale would enable generating new insights and therapy targets. In this project, we analyze cell properties in an unsupervised manner and correlate the types with the known features of a tissue.

Contact: Magdalena Matusiak (mmatusia@stanford.edu)

**17. Human in the loop for histopathology image annotation**
Most of modern machine learning algorithms are heavily data hungry. Acquiring high quality training data in medical domain is laborious but necessary. In this project, we are interested in developing user-friendly active learning tools to accelerate the acquisition of clinically relevant annotated features from diagnostic slides.

Contact: Magdalena Matusiak (mmatusia@stanford.edu)

**18. Classifying heavy drinkers from normal controls using neuroimaging data of the NCANDA study**
One of our goals is to use machine learning methods to differentiate 134 high-drinking adolescents from 674 minimal-drinking adolescents based on their longitudinal MRI images of the NCANDA study (http://ncanda.org/). Through this learning we would also like to identify MRI biomarkers associated with alcohol-use effects on the developing adolescent brain.
Prerequisites: Python or MATLAB, knowledge in statistical analysis (preferable).

Contact: Qingyu Zhao (qingyuz@stanford.edu)

**19. Building a Structured Knowledge Database for News Articles**
Description: We are building a processing system that extracts facts and relations from news articles, and then connects them with existing open knowledge bases, in order to relate articles of similar topics and summarize these topics.

Prerequisites: Python, Java, Javascript

Contact: Jens Gould(jens@jenserikgould.com)

### 20. Multi-State Bar Exam Challenge for AI

Multi-State Bar Exam is a multiple choice question answering task in which a case description is followed by four possible answers. An example question answer pair is shown below. The project is to baseline the performance of a ML based approach on this task. We want to find out if this task is too hard for ML, or whether it is within its reach.

A father lived with his son, who was an alcoholic. When drunk, the son often became violent and physically abused his father. As a result, the father always lived in fear. One night, the father heard his son on the front stoop making loud obscene remarks. The father was certain that his son was drunk and was terrified that he would be physically beaten again. In his fear, he bolted the front door and took out a revolver. When the son discovered that the door was bolted, he kicked it down. As the son burst through the front door, his father shot him four times in the chest, killing him. In fact, the son was not under the influence of alcohol or any drug and did not intend to harm his father.
At trial, the father presented the above facts and asked the judge to instruct the jury on self-defense.
How should the judge instruct the jury with respect to self-defense?
(A) Give the self-defense instruction, because it expresses the defense's theory of the case.
(B) Give the self-defense instruction, because the evi- dence is sufficient to raise the defense.
(C) Deny the self-defense instruction, because the father was not in imminent danger from his son.
(D) Deny the self-defense instruction, because the father used excessive force.

Contact: Vinay Chaudhri (vinay_chaudhri@yahoo.com)

### 21. Comparing Learned Analogy Models to Curated Knowledge Bases

A proportional analogy holds between two word pairs: a:a* :: b:b* (a is to a* as b is to b*) For example, Tokyo is to Japan as Paris is to France. It is possible to build learners to perform this task. See https://aclweb.org/aclwiki/Analogy_(State_of_the_art) for a survey of the state-of-the-art. We have also built an answerer for such questions based on a hand-built knowledge base which takes into account appropriate semantic relationship. We want to better understand the relative tradeoffs of these two approaches.

Contact: Vinay Chaudhri (vinay_chaudhri@yahoo.com)

### 22. Automated Glossary Construction

Given a textbook, we are interested in identifying all the terms that should go into a glossary at the back of the book. We have the content of the book, and a curated glossary. Can we train a learner to extract these terms and improve upon what is already being done manually?

Contact: Vinay Chaudhri (vinay_chaudhri@yahoo.com)

### 23. Machine Learning for Crystal Facet Determination

It has been shown that synthesizing materials with different crystal facet exposure can greatly affect the electronic, optical and catalytic properties of materials. However, very few analytical techniques are available to quantify which crystal facets are exposed in these materials and hence no quantitative relationship has been established between facet exposure and material properties. This project will focus on using machine learning on SEM images to quantify facet exposure and correlate it with catalytic activity.

Contact: Thomas Gill (tgill3@stanford.edu)

**24. Intelligent Tax Assistant Generation**
Internal Revenue Service has created intelligent assistants for a variety of income tax topics. For example, see: https://www.irs.gov/help/ita This assistants rely on a manual coding of knowledge available in a variety of income tax publications. Can we use these examples as training set to construct assistants on new topics? For example, one candidate is calculation of mortgage taxes. IRS provides a rudimentary version of the assistant. can we enrich it by taking into account all the knowledge in https://www.irs.gov/publications/p936

Contact: Vinay Chaudhri (vinay_chaudhri@yahoo.com)

**25. Quantifying Patient Response to Cancer Treatment**
We are interested in building an algorithm for quantifying patient response to cancer treatment by mining both structured and clinical text data using NLP. The product will have both clinical and research applications.

Contact: Haruka Itakura (itakura@stanford.edu)

**26. ScaleUp Entrepreneur**
Innovative entrepreneurs around the world want to learn and/or adapt to the 'entrepreneurship language' and absorb the SV mindset. The process to search and access the right resources in SV demands significant effort, time and money. Although there is a plethora of resources available online, they are usually fragmented, unstructured and disorganized. As a result, entrepreneurs may be overwhelmed, unsure if they found the best resource for their needs, and anxious due to the fear of missing out. ScaleUp Entrepreneur is a platform that will apply AI/ML to optimize entrepreneurs' search for the resources that best fit their needs.

Contact: Wolney Betiol (wolney@stanford.edu)

**27. Protein structure prediction with sequence models**
The 3D structures of the proteins are biomedically important, yet difficult to characterize with experiment. Using publicly-available datasets[1], this project aims to develop an advanced predictive model for 3D protein structures given an amino acid sequence. We would like to compare the performance of the model with that of models[2]. Reference [1]: https://www.rcsb.org/ [2]: https://www.rosettacommons.org/

Contact:Yosuke Tanigawa (ytanigaw@stanford.edu)

**28. KnowledgeBuilder**
This non-profit and open source project will build a software tool based on agents and AI-based integrated and autonomous algortihms for continuous confidence building in medical knowledge. On a daily basis a very large number of research articles and other information is published in written form; often the information is very conflicting and very hard to tease out if it can be trusted towards building the true glibal medical knowledge. Building an easy to use software tool that continiously aggregates, mines, classifies and scores the medical data sources and their content on a daily basis and autonimously builds confidence/facts in an unbiased fashion will accelerate medicine tremendiusly, hence, this will be the focus if this project.

Contact: Demir Akin (Akin1@stanford.edu)

**29. Combining Evolutionary Strategies with Policy Gradient methods in Reinforcment Learning**
In Reinforcement Learning, Evolutionary Strategies has been shown to be an effective algorithm, and has performed quite well in learning policy functions for agents, due to it's ability to stochastically explore the parameter space. However they can be less sample-efficient in comparison to Policy Gradient methods, which are typically much better at zeroing in on optimal solutions, but can suffer from local optima and lack of exploration. Hence is there a way that we can combine the Evolutionary Strategies and Policy Gradient methods together to form a hybrid algorithm, likely with ES spawning instances of PG, to create a method that can vastly explore the parameter space, and be able to zero in on numerous optimal policies, so we can pick the best one? Please talk to me for more information.

Contact: Mario Srouji (msrouji@stanford.edu)

**30.Reducing the variance of Evolutionary Strategies and exploring optimizations in Reinforcement Learning**
Evolutionary Strategies is a quite simple algorithm in terms of how it works. However efficient implementations are hard to come by, and often times the algorithm can struggle from high variance, causing it to be sample-inefficient in certain learnable environments. How can we modify the ES algorithm to reduce the variance of the updates made to the weight vector, and what sort of tricks can we do to allow ES to more efficiently zero-in on optimal policies for our agents? Talk to me for more information.

Contact: Mario Srouji (msrouji@stanford.edu)

**31. Creating a Central Pattern Generator (CPG) network for solving rhythmic-control tasks in Reinforcement Learning**
Central pattern generators (CPGs) are biological neural circuits that produce rhythmic outputs in the absence of rhythmic input. They are the source of the tightly-coupled patterns of neural activity that drive rhythmic motions like walking, breathing, or chewing in animals. There has been previous work that has been able to implement these neural circuits for use in robotics and Reinforcement Learning, however no work has explored creating a self-sustaining CPG network. They have vastly promising research implications due to the priors that they impose on Reinforcement Learning policies (they allow us to learn rhythmic motion much more efficiently and performantly than standard MLP networks). Hence how can we create a CPG network, either by using RNN's, or modeling Fourier Transforms, to create a model that can beat the performance of MLP's in Reinforcement Learning control tasks? Highly recommend that you talk to me for more information.

Contact: Mario Srouji (msrouji@stanford.edu)

**32. Diagnosis of Chronic recurrent multifocal osteomyelitis (CRMO) using MRI scan data**
Professor A. V. Ramanan, a leading medical researcher in the UK (Bristol Royal Hospital for Children & Royal National Hospital for Rheumatic Diseases, Bath) specializing in children's diseases has been doing research into a condition called Chronic recurrent multifocal osteomyelitis (CRMO), which affects children between ages 8-15. The children present with bone pain and on whole body MRI (WB-MRI) show multiple areas of inflammation (enhancing lesions on MRI).
He has collected labeled MRI scan data and would like to investigate whether a machine learning model can aid with diagnosis. He also has approval to share anonymized MRI images. Please contact Anand Rajaraman for more information.

Contact: Professor A. V. Ramanan FRCPCH, FRCP
Consultant Paediatric Rheumatologist (avramanan@hotmail.com), Suvadip Paul (suvadip@stanford.edu)