

# Matthew DongKyu Cho

MASTER'S CANDIDATE @ SEOUL NATIONAL UNIVERSITY · CAUSAL DATA SCIENCE

✉ kulupapa1127@snu.ac.kr | 🏠 umamicode.github.io | 💻 github.com/umamicode | 🌐

linkedin.com/in/dong-kyu-cho-023259176



## Personal Profile

A Master's candidate in the Graduate School of Data Science at Seoul National University, under the supervision of Professor Sanghack Lee. My field of interests include Causal Inference, Representation Learning, and Domain Generalization.

## Education

### Seoul National University

Seoul, South Korea

MSc in Data Science

March 2021 - Current

- Research Assistant at Causality Lab
- Advised by Sanghack Lee
- **Courses:** Machine Learning and Deep Learning for Data Science, Big Data and Knowledge Management Systems, Software Platforms for Data Science, Machine Learning for Visual Understanding, Text Analytics and Big Data, Causal Inference for Data Science, Special Topics in Data Science: Meta Learning, Data Science Capstone Project, Dissertation Research, Special Lecture on Data Science

### Seoul National University

Seoul, South Korea

Bachelor of Arts in Information Science and Culture / Western History (Double Major)

March 2014 - February 2021

- Graduated with Distinction (Cum Laude)
- Data-Driven Marketing Project with LOCAL STITCH

## Work Experience

### VAIV Company (former DaumSoft)

Seoul, South Korea

Data Marketing Intern

Sept 2020 - Aug 2021

- Data-Driven Marketing projects for corporate clients
- Marketing based on Sentiment Analysis of the Korean Social Media.

### Republic of Korea Armed Forces

Icheon, South Korea

Military Interpreter

March 2016 - December 2017

- Military Interpreter for the Republic of Korea Armed Forces.
- **Technical Skills:** Korean-English Interpretation

## Projects

### Stock Interrelation Research using Keyword and Supply Chain data

Seoul, South Korea

Seoul National University and NH Investment & Securities

July 2021 - August 2022

- A collaborative research with NH Investment & Securities.
- Built a Domain-adapted Language Model and its applicable pipeline using large scale Financial Corpus.
- Built an Interactive Graph Database using the Bloomberg Supply Chain Data.
- **Technical Skills:** Financial-Domain specific Language Models, Financial Data, Graph Database.
- **Soft Skills:** Teamwork, Presentation skills, Report writing.

### Towards Language Models Capable of Causal Reasoning

Seoul, South Korea

Seoul National University and LG AI Research

August 2022 - Present

- Collaborative research with LG AI Research & UNIST
- A Causal NLP Project Under Progress
- **Technical Skills:** Causal Inference, Large-scale Language Models.

### Semi-Supervised Federated Learning with Representations

Seoul, South Korea

Seoul National University - Class Project

March 2022 - July 2022

- A Class Project for *Special Topics in Data Science: Meta Learning*.
- A Semi-Supervised Approach towards federated learning using Contrastive Learning methods.
- **Technical Skills:** Meta Learning, Self-Supervised Learning, Federated Learning.
- **Soft Skills:** Report writing, Presentation Skills.

## Skills

---

**Programming** Python, R, SQL.  
**Miscellaneous** Linux, Shell (Bash/Zsh), Neo4j,  $\LaTeX$ (Overleaf/R Markdown), Tableau, Git.  
**Soft Skills** French, Korean, English-Korean Interpretation

## Achievements

---

2022 **X /340**, Graduate Record Examinations (GRE) *South Korea*  
2022 **116/120**, TOEFL *South Korea*

## Research

---

**Leveraging Style-Content Disentanglement for single source domain generalization with causal mechanisms**

*Seoul, South Korea*

Seoul National University Causality Lab

*Working Paper*

- Masters Dissertation Paper
- **Abstract:** Machine learning suffers from the frequent discrepancy in the data distribution, commonly known as domain shift. Single-source Domain Generalization (sDG) is a task designed to simulate domain shift, where the objective is to train a model that can generalize to multiple target domains from a single source domain. A common approach is to use adversarial data augmentation to simulate unseen domains and learn augmentation-invariant features. However, prior methods frequently display fluctuations in the o.o.d performance. To alleviate this issue, we connect the discoveries in the nonlinear ICA literature to revisit sDG with the concept of S-C Disentanglement. We show that augmentation-based sDG methods face an innate uncertainty derived by the unobservability of the domain. Hence we reformulate the sDG task by using a oracle model approximated with a pretrained model. We devise a novel sDG method: PROF(Progressive mutual information Regularization for Online distillation of Frozen oracle models) that functions as a stable objective for single source domain generalization. We further devise several modifications to the existing methods for effective learning.

## Languages

---

**English** Professional proficiency/ TOEFL 118 (2020.08), 116 (2022.08)  
**Korean** Native proficiency  
**French** Intermediate proficiency/ DELF B1

**References available upon request.**