Overall intro:

At MS, I lead innovation, development, and enhancement of predictive modeling capabilities across commodities and equity markets. More specifically, I drive applications of mored advanced weather and fundamental models across the Trading team, Strats, and sell-side Research.

I focus on 2 types of quant models The first is to develop advanced capabilities in weather data and models. This work is around how we can find and realize edges in weather data to pnl impacts. They are all pretty novel, (at least to us at MS), some are more operationalized, and some are exploratory.

Launching weather AI models:

- Build in-house capabilities for weather AI models (GraphCast, FourCastNet, etc.); Sent daily views on model comparisons.
- Conducted comprehensive model evaluation and backtest vs. ground truth and vendors across key ISOs/RTOs, with special focus on extreme-weather performance.
- Develop methodologies for customized model improvements, including ensembles, finetuning, and refreshed input data.

High-level: I setting up in-house running of weather AI models, evaluating their performance comparing to traditional models, and design methods to improve.

- Onboarded datasets from ECMWF, NOAA, NASA. These include historical, real-time (in production), and forecast data.
-

Exploring market edges in systematic weather shifts related to climate and teleconnections:

- Tested how variations in short-term weather, long-term climate, and ENSO phases are "priced in" using weather futures and regressions versus realized weather. Investigate links between polar vortex weakening, February cooling, and prices.

High-level: weather system is complicated – long-term climate change, S2S ENSO, short-term weather, noise (internal variability). I wanted to know what are perceived and what are not - because eventually for trading we play with other people's perceptions. I picked weather derivatives as an instrument.

- Approach: I examined the relationship between future price changes and weather "shocks" (anomalies: deviating from climatology) through regressions.
- And I could separate different types of variability by different time horizon:
  - To look at short-run capitalization, I regressed weather future's prices with 2-week short-term weather data., which shows how market prices daily anomalies, and demonstrate the peak time for weather to influence prices: medium range (8-14 days).
  - Then, to evaluate capitalization of longer-term patterns, I took the avg future prices 3-4 weeks before the before the start of each contract (so no weather info is available, only longer-term perception), and run a yearly trend with monthly FE.
    - I found that these futures track climate-model yearly trends but with a smaller magnitude. So it embeds certain level of climate change (e.g. every year +$2/month for summer, -$1/month for in winter), but smaller than realized weather.
    - And the monthly FE effects are interesting: everywhere else's winter future prices go down and expect warming, except for Feb futures price in NE going up, so market perceived and priced in the late-winter cold bursts, while it's not visible in long-term climate model results not visible in short-term weather data. So where does it come from?
    - My theory is it's related to people priced in polar vortex due to stratospheric sudden warming, which is a systematic pattern that's causing cold February for NE.
      - Mechanism: with climate change, ice sheet melts--> sea ice content reduces, more generally, SSW events --> polar vortex destabilizes and splits --> instead of cold air being locked above the polar region, it can push further south into the mid-latitudes and causing cold winters in NH.

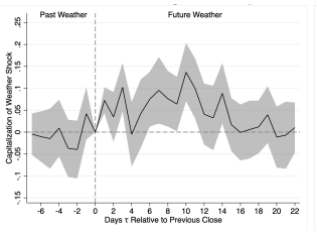
(unless further asked about teleconnection and ENSO)

- I also tested different oscillation indices, the overall takeaway is: they impact the realized volatility, but doesn't change the trend of what market priced. They are reflected in the 2-week weather forecasts. They could be more of a variability risk, but not a systematic mispricing.
- That being said, the relationship between sea ice loss, sudden stratospheric warmings (SSW), North Atlantic Oscillation (NAO), Quasi-Biennial Oscillation (QBO), and late winter cooling is complex and interlinked:
  - A valuable next step: we can probably improve late winter temperatures by incorporating more teleconnection info. Paper using a MLR model involves predictors such as autumn SST, SIC, some indicators for QBO and NAO, reaching skill of 0.5-0.6 --> can improve Feb outlook.

Details:

- Short-term weather: I used next 3 weeks' daily temp anomalies (in DDs) by day in right hand side, to regress on price daily changes. Also included last week as control- if the model is set up correctly, the coef should be 0 for past days. FE by monthly contract.

$$\Delta p_{cd} = \alpha_c + \sum_{\tau=-7}^{21} \beta_\tau \left[D(T_{c[d+\tau]}) - D(\widehat{T_{c[d+\tau]}}) \right] + \epsilon_{cd}$$



  - Plot beta with time is an inverted U shape→ Peak time for weather to influence prices: medium range (8-14 days). Too short-term everyone knows, too long term no good info. After day 14, no sig diff from 0
- Longer-term: to isolate the effects away from weather forecast – take avg. futures prices 3-4 weeks before the start of each contract (reflect market's seasonal expectations, before short-term variability available).

$$\overline{p_{amy}} = \alpha_{am} + \beta y + \epsilon_{amy}$$

  - I decomposed the time series of (monthly mean) future prices: long-term trend + monthly FE).
  - Replaced Y to be HDD/CDD from weather station (NOAA, observations), and CMIP5 climate simulations. Findings:
    - For yearly trend term, coef of future prices significantly smaller than actual observations → mkt underestimates the magnitude of long-term trend. (+$2/month/year for CDD, -$1/month/year for HDD)
    - But CMIP5 RCP8.5 results are closer to the obs → longer term climate model simulations could be useful in addition to weather forecast.
- What's more interesting: the effects by month (Feb): Weather station data is very noisy so no significant signal for winter months. CMIP5 shows reduction in HDD in all winter months, but future prices have a positive increase in Feb→ so people intentionally expect and price in cold February for NE US → I think it must come from S2S variabilties.
  - Did a simple test of the variabilities without obtaining all S2S data: add ENSO indices to the regression, tested 5 indices → (Include in month future prices model) adding yr-to-yr variation improves the model R2, but does not change long-term trend coef → proved this is what also priced in.
  - Super interesting: I chose to use CMIP5 (as the prediction years overlap starting 2005). CMIP5 hasn't incorporated the stratosphere-troposphere interactions dynamics à happened to be a good experiment isolating the effects of these impacts from the polar vortex related February cooling - a consistent pattern with climate change.
- The relationship between sea ice loss, sudden stratospheric warmings (SSW), North Atlantic Oscillation (NAO), Quasi-Biennial Oscillation (QBO), and late winter cooling is complex and interlinked:
  - Arctic sea ice loss (alter energy exchange, wave patterns) & SSW can both influence polar vortex. Sea ice loss can increase the likelihood of SSW events by altering the planetary wave activity.
  - NAO (key mode of atmospheric variability in the North Atlantic region) can be influenced by SSW (often transitioning to a negative phase after an SSW, causing colder late-winter conditions in Europe and NA) and SIC
  - QBO (a regular oscillation of equatorial stratospheric winds, can modulate the response of the polar vortex to external forcing, such as sea ice loss) Certain phases of the QBO (e.g., easterly) are more conducive to SSW events and a negative NAO. QBO adds variability, affecting the timing and intensity of these processes.

These interactions can all cause colder-than-average late winter conditions in the mid-latitudes, despite global warming. E.g,

- A few facts: 2001-2023
  - Outliers: for too large 1/10, too small: x10
  - Weather station data (NOAA, observations), CMIP6 (NASA downscaled, 0.25x0.25, multi-model mean, hist. 1995-2014)
  - Model:
    - Monthly: avg. prices = fixed effect (city, month) + beta*year +error, or replace prices with HDD, CDD.

Anomalies: a departure from long-term avg. = obs. – climatological. Allows for comparison across regions
Climatology: long-term avg of a variable.
Reanalysis data: can have 1) deficiencies in model structure, and 2) uncertainties in assimilated observations.--> have biases that's acceptable globally but not regionally/locally.

- Bias correction: rescale to local reference climatology, through a big 3D matrix for each grid and month (reference climatology / ERA5 climatology)

Better dataset that includes predictions: **Multi-Source weather (MSWX) data**, 3h, near real-time, 0.1x0.1, 7 month forecast, systematic bias corrected. Incorporates satellite, reanalysis, obs. and models.

- **Expanding use of alternative weather datasets:**

- Onboarded probabilistic storm-track forecasts for real-time hurricane monitoring; Designed backend infrastructure for flexible storage, processing, and visualization of spatial data.
- Assess value and limitations of Subseasonal-to-Seasonal forecasts across commodities and cross-asset applications.

High-level: other than the standard weather data, there are potentially edges in other types of data:

- **Storm track.** I've helped the team to build a real-time monitoring tool with probablistic storm tracks. E.g. 50 tracks of all the realizations where a hurricane likely will go. In particular, the weather AI models supposed have an edge on forecasting storms.
- We built a hexagon-based spatial system to handle spatial data.
- **Another is S2S.** Traders found it hard to use/trust, but there are some interesting explorations these days to improve those. I want to closely monitor that, and work with traders to find the right ways to use it – to get even just a bit more useful info.

- **Fundamental Supply/Demand Model Enhancements:**
  - **Developing advanced short-term load forecast model:**
    - Developed an hourly load forecast method using a flexible "like-days" sampling framework, improving transparency, reducing dependency on historical climate, and enhancing peak-load performance.
    - Optimized model performance through a hybrid statistical-ML approach, capturing non-linearity, cross-hour dependencies, and improving accuracy under extreme-temperature conditions.
    - Prototype a 2D spatial load-prediction framework incorporating high-resolution population distribution.

High-level:

In academia, I specialized in translating weather and climate dynamics into real-world impacts (crop yield, energy demand, health impacts). And now I use these experiences to build more accurate fundamental forecasts.

Because of this background, I approach load forecasting very differently from our strats' method. For example, I don't use any static thresholds like 30-year climate normals or fixed definitions of heat/cold events. Those approaches systematically miss today's extremes and doesn't take into account the climate baseline itself is shifting.

I built a fully flexible 'like-days' sampling framework for load forecasting. It dynamically adapts to atmospheric conditions, based on similar periods of year and weather profile. This approach reduces dependency on historical patterns, and improves peak-load accuracy. For example if there's an unusual cold event, it finds the similar extreme events in the past as the most relevant data to train the regressions.

And you can make it pretty sophisticated in terms of how to find "like days", can test some ML approaches.
I also demonstrated a spatial framework that overlays with high-resolution population distribution is very helpful in improving the accuracy.
Overall, my edge is integrating atmospheric science with market fundamentals to produce more accurate, climate-supply/demand forecasts.

- Derive regression models on how weather impacts energy and crop yield:

- Empirically derived regression models that estimate effects of weather/climate on commodities.
  - Topics can vary, approach transferrable: could be temp, SLP, hurricane frequency on crop yield, energy use

    1) Built a global-scale non-linear regression model of electricity consumption as a function of daily temperature and income at 30x30km grid-cell level. Results allow high precision power demand forecast for any location any day.
- Based on a paper originally published on Nature (I contributed by processing the weather data GMFD).
  - The paper estimated a nonlinear function between daily temp at each grid with monthly energy (electricity, other fuels) cmp.

    Significance: 1) offered a methodology of how you can link the socio-economic data (usually discrete in space and time), with high resolution weather data (continuous in space and time), to best exploit the granularity in weather data.
    - while we only know monthly energy, but because we observe the higher granularity temp data daily, we can recover the relationship between daily grid-cell per capita energy and daily grid-cell temp → essentially the model searched for a grid-cell level temp-energy relationship, that best matches monthly obs when applied to every grid for each daily temp.
  - I replicated this regression – so that if I have daily temperature (from weather reanalysis data, at each grid 0.25x0.25, paper was 1970s-2015, 45 years), and other variables (daily precip, gdp pc, [population dist, long-term HDDCDD]), I can predict the daily electricity use at the grid-cell level, and then aggregate to any regions (add the fixed effect matrix (by country, by month))

Model: monthly country elec per capita = country-month aggregates (across all days in a month and all grids in a country, weighted by population) of the nonlinear temperature term (βc1*linear + βc2*quadratic terms of temp) + monthly FE
→ estimates βc1, βc2, coef for each country for the linear and quadratic temp terms
→ mathematically can show it recovers the same coefficients, βc1, βc2, that describe the primitive grid cell-by-day relationship

$$E_{jtc} = \sum_{d\in t}\sum_{z\in j} w_{zj} E_{zjdtc} = \sum_{d\in t}\sum_{z\in j} w_{zj}\left[\sum_{m=1}^{M} \beta_{cm} T_{zjdtm}\right] = \sum_{m=1}^{M} \beta_{cm} \underbrace{\left[\sum_{d\in t}\sum_{z\in j} w_{zj} T_{zjdtm}\right]}_{\bar{T}_{jtm}} \cdot \underbrace{\qquad}_{f_c(\bar{T}_{jt})}$$

Our estimating equation takes the following form:

$$E_{jtc} = f_c(\bar{\boldsymbol{T}}_{jt} | \overline{LogGDPPC}_{jt}, \overline{CDD}_j, \overline{HDD}_j) + g_c(\bar{\boldsymbol{P}}_{jt}) + \alpha_{jic} + \delta_{wtc} + \varepsilon_{jtc}. \quad (3)$$

z: grid, d: day, c: country

- On top of it, allow energy-temperature responses to vary by income through interaction terms with every βc1, βc2 → address heterogeneity across income levels.
- Nonlinear U shape for elec especially for high income, L shape for other fuels.
- Thus, exploit within-country, yr-to-yr variations in daily temperatures, while the variation in income help predict how this relationship may change in future.

- With the power strats, I'm reconstructing to incorporate 1) subnational, EIA's state-level monthly data or 2) daily power consumption, to make it more useful for power desk

2) Estimated the relationship between daily temperature extremes and crop yield using two weather datasets, significantly improving the average temperature model. Predicted corn and soybean yields.

- Model part: previously we have a crop model, but outdated, performance not good especially for outside of US. I also notice that they used monthly mean temp. --> the best model needs to include daily temperature extremes because I know the effects are nonlinear. Data part: previous model does not have high solution weather/crop yield data.

- I suggested: 1) better resolution weather data with global coverage 2) upgrade crop models to account for temp extremes.
  - 1) For 1), I onboarded two global reanalysis datasets, 2) calculated the non-linear model (GDDs, sum of degrees between 2 ranges) as the predictor for crop yield in regression. I used model results in the US for evaluation, comparing it to the best model in literature (with 4x4km PRISM data) and our existing (monthly) avg temp model.

    3 findings:
    - Estimated the regression model --> almost same coefficients: robust to datasets
      - Model form: an integral of a non-linear function of heat (piecewise linear and polynomial) over lower and upper bound of heat (GDD), and calculated in a discrete way where you sum values of the func at each GDD
    - Compare out-of-sample predictions: a large increase than using the mean function (+ 10ppt), [only a slight drop in model perf. (-1~2 ppt).]
    - Main reason: Used the GDD metric, compared to a mean temperature model, much better.
  - 2) proved the robustness of the model with the global datasets→ then predicted crop yield elsewhere (interested in South America and Asia). We are still evaluating it via crop yield data and satellite vegetation proxies. The crop data could leverage the enhanced vegetation index (EVI) 0.1 x 0.1from remotely sense data.

- A few facts: 1980-2020
  - Data: both MERRA-2 and ERA5 used remotely sensed measurements for reanalysis:
    - MERRA-2: 0.5 x 0.625, hourly (1980 – present)
    - ERA5: 0.1 x 0.1, hourly (1940-present)
    - CRU: monthly, global coverage
    - PRISM: 4 km x 4 km for the paper
    - Crop: county-level yield data in the US from USDA
  - Missing value: the distance-weighted avg of the surrounding grids
  - Merge weather data to counties (grid > county): based on shares of grid intersect with the county.
  - GDD: (Tmax + Tmin)/ 2 - base temperature by species, over March to Sep growing season
  - Model: Log(yield_it) = approximate to a function summing all the 1C intervals in degree days during a growing season

$$y_{it} = \sum_{h=-4}^{41} g(h+0.5)\left[\Phi_{it}(h+1) - \Phi_{it}(h)\right] + z_{it}\delta + c_i + \epsilon_{it}$$

    - g(h) is a piecewise linear (or polynomial). To estimate piecewise linear, loop all possible threshold and slopes and find the one with best fit (least squares error)
  - Controls: z_it, includes a quadratic func of total growing season precip + state-specific quadratic time trends (e.g. tech progress) + c_i: county fixed effect (e.g. soil quality).

- Back-test/prediction: Evaluate results by randomly sample 85% of the years and predict using the rest 15%, comparing the RMS (root mean square error) drops to a baseline prediction without weather, ie. how much error can be explained by inclusion of weather data. 1000 repetitions in a cross-validation way. New models drop 10-15%, old drop 5%
- Region: Brazil and US largest importer of soybean. Main importers, China, Japan, Korea. Helpful to evaluate in South America and Asia

Built a global-scale non-linear regression model of electricity consumption as a function of daily temperature and income at 30x30km grid-cell level. Results allow high precision power demand forecast for any location any day.

- Understand how long-term climate trends and teleconnections can help predict prices:

Reduced-form models I used since in academia, GCAM + GCIMS (global change intersectional modeling system), DOE.
Climate model: a reduced-complexity climate model, but arguably by far the closest to ESMs, with ability to 1) output "high" spatial-temporal resolution based on CMIP6, 2) include uncertainty in projections based on prob. dist. of parameters, 3) run large ensembles or emulate more complex ESMs. 4) integrate with energy, hydro, land use module.

1. Hector: reduced complexity climate model with uncertainty analysis.
   - A globally resolved carbon climate model, with terrestrial and ocean CC, and active surface ocean chemistry. Model runs until all carbon pools are in equilibrium. 37 gases and exogenous radiative forcers, calculate aerosols and final RF.
     - Carbon cycle→ GHG concentrations and aerosols→ RF(39 forcings) -> use 4 boxes energy balance model-> temp
     - Parameterization: calibrated to MMM of CMIP6
     - Evaluation: in good consistence with historical data and CMIP6. Focused on dynamics.
     - Can run large ensembles or emulate to more complex ESMs.
   - In RCMIP, default annual, can be higher freq. can be run "free running" or "constrained". Base year 1750.
2. STITCHES: fast statistical emulators. Comprehensive, generate ESM like output under scenarios.
   - Spatially resolved (same with the underlying CMIP6 ESM), high frequency (monthly), include internal variability
   - Extend time sampling to build a continuous time series at any time/space in CMIP6 archive, through a look-up table for each decadal window.
   - Algorithms applied separately to each ESM in CMIP6, all models, all experiments, all ensemble members with reported monthly gridded data (mainly surface air temp and precip).
3. Matilda: quickly run probabilistic projections, by run Hector iteratively
   - Produce perturbed parameter ensembles from prior dist. using Monte Carlo, weight ensembles with obs. → produce prob. projections for climate variables--> quantify uncertainty.

Others: Xanthos: 0.5x0.5 hydrological model; land use module.

Bottom-up energy system model, coupled with the climate system

1. Bottom up, has granular inventory of energy technology & land use sector (therefore agriculture).
2. Equilibrium model: everything through SD great to look at substitutional effects across fuels, energy tech, policies.
3. Strengths to study commodities because 1) it indicates potential linkages and directionalities→ can inform further empirical models. 2) it can be flexible adjusted to run simulations and scenario analyses quickly.

   E.g. biomass fuels: 1) primary production à land/ag for biomass feedstocks, 2) and these go into the secondary production of e.g. biofuels, wood pellet, 3) the end-use market, where can estimate demand.
4. Limitation and improvement: combine with more empirical models, can flexibly link to automated modeling system as it's opensourced.

Supply and demand simulations:

- Designing scenarios regarding key uncertainties in technology and regulation, simulating using the commodity market equilibrium model, validating and calibrating to empirical data.

1) Simulated technology options (natural gas, fuel cells, diesel) to fill the power demand for supercomputers and data centers and the supply demand impacts.

- An example where I combined fundamental judgement from power analyst + simulation/realization with the models.
- Research: big topic. They estimated 1) the power demand of genAI, 2) narrowed down some tech options (For gen: aeroderivative gas turbine, fuel cells, diesel engines for backup + some PPA options such as geothermal. For storage: esp. long duration energy storage options) 3) the cost of the tech options.
- Adjust model on both demand & supply side:
  - For demand: adding projected demand to commercial-other sector in both Direct Use (by states, assumptions can be made) and increasing demand in grid.
  - For supply: 1) Forcing in additional storage 2) adding off-grid power options, gas turbine, fuel cells, diesel
- Questions to answer:
  - Grid-level: can look at the impacts on the US energy market, e.g. NG, elec prices.

- Do scenario analysis:
  - Regarding key assumptions on energy storage, can stress test.
  - (Regarding the schedule of power demand & tech phase in-> sensitivity)
- Even without all the tech building, readily doable: for supercomputers, test the impacts of co-locating with nuclear pp. We have a few guesses of locations/plants. I can simulate the impacts of by forcing a capacity drop in those and stress test the regional grid and see what fills the gap, what's the price impacts on elec.

Fact: very high time-to-power (how much DC owners want to pay to get powered earlier).

- Estimated demand for sustainable aviation fuel under IATA's rule on emissions, quantified the capacity gap and the impacts on the jet fuel market.
- Based on IATA's NZ goal, Airline's emission targets, IRA, we wanted to look at what would that mean for S/D, what need to happen, what are the key gaps
- Added the emission pathways as a constraint for aviation sector in GCAM. Simulation results showed:

  1) the growth in airline demand (based on income elasticity and further constrained with a price elasticity.

  2) supply: a 15x increase of the current capacity.

  3) pricing: because I set the emissions as a target (as opposed to through economic competition) --> price of SAF is still more than doubling the price of conventional jet fuels. à I can quantify how much subsidies are needed.

- Conclusions: The SAF economics highly depend on policies: ~ > $4/gallon environmental credits
- Interesting question: How SAF compete with Biodiesel.
  In my results I saw a rather flat trend for biodiesel, but robust across SAF scenarios à Research colleagues think they are in competition. I don't know for sure

Fact: 2030 goal (13% below 2019 baseline)

- SAF feedstocks mostly from bioenergy crop in both grass and trees, not food.
- Scope 3 emission regulations may be a big driver if more adapted.

Developing quantitative methods for emerging commodities:
Derive methods to evaluate opportunities in new areas. Evaluate vendor datasets. Explore alternative datasets (satellite, sensor data).
Vendor data:
Sensor data: Measurable. Sells solutions with sensors to show real time energy use and carbon emission in assets globally. Later I learned their energy use is based on NLP reading energy bills of the buildings they contract with.
Satellite: Fathom. High-resolution flood map based on satellite data (terrain info) + climate model+cat model.

e.g. 1) Derived region-specific marginal abatement cost curve for 17 CCS technologies, used to forecast profits for CCS projects in the voluntary carbon market and price impacts on the energy market.

- With carbon trading desk. Help clients to invest in CCS projects & sell/buy carbon offset credits. But they don't know how much to invest. Because no costs ref info relative to other places/technologies.
- So having a tech/region specific cost curve could be fundamental (x: sequestrated CO2, y: price /ton)
- I derived these curves, by taking 5-point of different carbon price (0, 25, 50, 75 ,100 $/t). Plot the CCS options by how much the sequestrations are for each tech.

Note: if asking the results for different results: need to go back and take more points in lower cost ranges.

2) Simulated EU's Carbon Border Adjustment Mechanism to examine impacts on international trade and the EU Allowances, UK Allowances markets.
EU has it's CBAM policies to tax on embedded carbon of imported goods from countries do not tax as much. After 2026, CBAM phase in, and the free allowances phase out.
2 types of outcomes we are interested:

1. Impacts on trade: simulated using an input-output model with international trade (GTAP) in the automated modeling system I built:
   - Calculated the import tariff by sectors/countries based on carbon intensity and location-specific compliance carbon price (carbon intensity: vendor data, country/sector specific. Research has a cost curve for compliance carbon price by regions.)à a large matrix by countries and sectors of tariffs
   - Fed to GTAP as input. Evaluate impacts on trade and productions in exporter countries. For example, I quantified about 1.5% decrease in trade balance for China, but e.g. 0.4% of increase for Brazil.
2. Impacts on the EUA market: they have a fundamental S/D modelà use the SD balance as x in a regression to estimate EUA forward curve.
   a. Supply: allowances (MRS, caps, auction)
   b. Demand: actual emissions. I simulate emissions from the model on the power and heat, industry, aviation, (through 3 scenarios with different levels of carbon price: bull bear base)
      Distinguishing part is that this approach incorporates the emission decline responses to the price increases of EUA in the specific sectors, as I provide 3 scenarios of carbon prices

What interests me is the response of these, e.g. traded demand would want to hedge the CBAM liabilities using EUA or futures contractsà EUA price higher. I wonder if we can transform to Short-term demand/S balance increase and apply in the futures market.

Fact: 2026 sectors: cement, iron & steel, fertilizers, aluminum, hydrogen, electricity.

Built company-level ML models to estimate sensitivity to energy transition and climate impacts for fundamentals. Conducted feature reduction and model selection from multiple algorithms (GLM, Random Forest, XGBoost, CatBoost, etc.) at different timescales, used to calculate exposure and identify mispricing of stocks.

- **Input data:** features are 5000+ variables from the integrated energy/climate/macro system, by region by quarter. Targets: global equity fundamental dataset.
- **Automated ML pipeline:**
  - **Preprocessing:**
  1) Dimensionality reduction & feature selection, multistep: 1) multicollinearity threshold 2) penalty-based shrinkage algo combined with either recursive feature elimination (RFE, pick # of features to select and algorithm to use, start with all features, give weights of importance based on the algo, discard the least important, and re-fit), or forward sequential feature selection: select the best feature set, keep this, links others after this, and retrain.
  2) Missing data imputation: mean/median imputation
  3) Categorical data encoding: one-hot (dummies, no difference, adding separate col for each category)/ordinal encoding
  4) Standardization. Remove mean, normalize by var. Thus: 1) preventing outliers from skewing the dist. 2) allows direct comparison of model coefs to help assessing feature importance
  - **Training**: 2010 to 2021, use 2022, 2023 as test set, features vary by sector but 4 lags included for each company

1) Algos: GLM, Random Forest, XGBoost, CatBoost with ensemble modeling techniques.
bagging: GLM, DRF

boosting: Extreme gradient boosting (XGBoost): sequential training approach.

2) K fold combined with forward chaining: fold 1 train [1] test [2], fold 2 train [1,2], test [3]
3) Best model: usually GLM or DRF.
4) Best model fine tuning using grid search and random search

- **Predictions:**
  feature importance analysis: for tree-based algo, take variable relative influence, for non-tree-based, take coefficient magnitude. (Feature contributions: shapley values.)
  a. Top features for real estate: cooperate borrowing rate, interest rate, electricity, gas & AC price index, private consumption.
  b. Top features for industrial sector, 3 month gov bond yields, 3 years gov bond yield, transp & storage sector value added output.

Next step: calculate a **sensitivity metric, model monitorning: drift detection, anomalies detection**

Fact:

- Best model fine tuning (hyperparameter, depth, learning rate, iterations, bootstrap types on training set):
  - grid search: define the search space as a grid of hyperparameters, and evaluate every position in the grid
  - random search: define a search space as a bounded domain of hyperparameters, and randomly sample within that
- Evaluation metrics: RMSE, MAE, MSE, R2
- Diff ensemble learning technique: create a strong model by combining several models, get a balance of learning with low var and bias
  - Bagging: GLM, DRF (ensemble of multiple trees). Trees tend to be overfitting→ form a random forest with lower variance.
    - Data sampling: create multiple subsets of the training dataset using bootstrap
    - Model training: train a separate model on each subset, independently.
    - Aggregation: combining predictions from all individual models. For regression, avg model→ reduced the variance of the model (ie learning too well), avoid overfitting, improve performance than individual models, by creating diverse models.
  - Boosting: XGBoost
    - Sequential training: each model trying to reduce errors made by the previous models, sequentially.
    - Weight adjustment: after each model is trained, weights of wrong instances become higher→ the next model focus on
    - Model combination: by weighted avg.--> reducing bias (ie learning too little, model is not related to dist.)
- GLM: apply in: the model has 1) a linear predictor (beta), ie. the relationship between beta and x, but 2) can be a non-linear link func that links the linear predictor and the parameter for prob dist. of Y.
  - overcome limitations of linear model: 1) The range of Y is restricted (e.g. binary, count) 2) relaxes the asp of linear models, e.g. var of Y depends on X. In OLS, residuals need to follow a normal dist with mean 0 and constant variance.
  - Estimation of beta: Max likelihood. Results are more interpreterable.
- Distributional Random Forest (DRF)

- Define boosting rounds ie how many rounds of small incremental. It can find the optimal # of rounds through hyperparameter tuning. Split training and test set.
- Use validation set during the training process - showing model performance after each boosting round for both the training and validation set. Stop early at the middle where highest performance on the validation set.
- Can use K fold cross validation. While the test set waits in the corner, we split the training into 3, 5, 7, or k splits or folds. Then, we train the model k times. Each time, we use k-1 parts for training and the final kth part for validation. Take the mean of the scores as the final, most realistic performance of the model. Once find the best model with the lowest score, must retrain it on the full data.
- Note: we have time series data,

Catboost (Category boosting)

Better accuracy compared to other algorithms For data with categorical features. No need to preprocess categorical features (like one-hot encoding), just specify some hyperparameter

Postdoc:

- Examined the oil and gas stock price responses to the policy lift on methane emissions. Studied emitting behaviors of 600 oil and gas facilities before and after the policy using TROPOMI satellite retrievals.
  - Goal: if there's any response in O&G companies after Trump's lift on the policy of disclosing CH4 emissions in 8/2020.
  - Approach this through 3 analyses, all using DID design to test causal effect: which looks at the relative change before and after the event for the base group and contrast group, by looking at the post x Y2020 interactions for 2020 vs 2019
    1. using TROPOMI to compare the grids with O&G facility (EIA) VS. grids without: ~ 5pb increase
    2. ultra emitters' event: detected large plumes 0.15 increase in daily count.
    3. Stock market response, compared to coal and power companies, little effects.

IIASA:

- Conducted climate model simulations. Analyzed climate model data across space and time. Attributed model uncertainty from natural variability. Results published in PNAS.
    - Simulated decarbonization scenarios to study the combined effects of climate and air quality impacts.
    - The interesting angle: 1) to see how the longer-term impacts of climate vs. shorter term impacts of AQ distribute spatially and temporally. 2) cost-benefit analysis of the co-impacts including crop, health, labor productivity. 3) Monte Carlo of different components of uncertainty
    - Attributed model uncertainty from natural variability: ScenarioMIP
        1. SSP3-7.0 has 10 initial condition ensemble members from 5 models. Looked at the ensemble spread, i.e. inter-realization standard deviation. Did it for global annual mean temp (SD between 0.1 to 0.15 C).
            - Found the ensemble spread narrows over time for temperature (a time trend). By contrast, no time trend for precip. Results are SD = XX degree C per year and mm/day/year
            - Could use the CESM2 single model large ensemble to further study (100 members, 1x1, SSP3-7.0)

Motivations:

Go deeper than broader.

Focus on commodities, than other stuff

Looking for edges, than having a deliverable.