README

Part I of this README documents all datasets generated for the paper "Hidden experts in the crowd: Using meta-predictions to leverage expertise in single-question prediction problems" by Wilkening, Martinie, and Howe (2020) and describes how to replicate the analysis. Part II describes the location and content of the files that can be used to replicate the two experiments.

I. Analysis Code

AnalysisCode.m contains the full analysis code to generate all the results reported in the main text and online appendix of Wilkening, Martinie, and Howe (2020) except for the simulations described in Appendix A. Running the code *"AnalysisCode.m"* as-is runs all the functions in the analysis and generates all the outputs; this can take anywhere between 5-15 mins to run. The code for simulations is in the "Simulations in Appendix A" folder and is discussed below the summary of the main analysis.

Datasets ("Files" subfolder)

Below is a list and description of each dataset csv/xlsx file:
1. m4data.csv – raw States dataset
2. m4data_details.csv – details on the States dataset questions
3. r2a_rawdata.csv, r2b_rawdata.csv, r2c_rawdata.csv, r2d_rawdata.csv, r2e_rawdata.csv – raw Grades dataset, collected over 5 samples. Each sample contains approximately 100 people who each answered 20 questions from each grade difficulty.
4. r2_outcomes.csv – questions and outcomes for the Grades dataset
5. GradesDataset.xlsx – cleaned version of the Grades Dataset
6. StatesDataset.xlsx – cleaned version of the States dataset
7. QuestionsList_Grades.pdf – list of all the questions in the Grades Dataset
8. QuestionsList_States.pdf – list of all the questions in the States Dataset

Once processed, each dataset is saved as a .mat file in the Files subfolder:
1. m4data.mat – States dataset
2. d1data.mat – Grades 1 (lowest difficulty) dataset
3. d2data.mat – Grades 2 dataset
4. d3data.mat – Grades 3 dataset
5. d4data.mat – Grades 4 dataset
6. d5data.mat – Grades 5 dataset
7. gsdata.mat – Grades dataset (All difficulties)

Note that the first two letters of each of the above .mat files correspond to a two-letter prefix for that dataset and this is used consistently throughout the analysis code (including the plots generated).

Auxiliary dataset files ("Files" subfolder)

There are several other data files generated throughout the analysis (where %s is the two-letter prefix indicated above):

1. %s_preds.mat – saved models' predictions for each model for each dataset
2. GenPreds_paper.mat – saved models' predictions over all datasets

Analysis Functions ("Paper" subfolder)

Below is a description of each analysis function in the analysis code, in the order that they are executed in AnalysisCode.m:

preprocess_m4.m
Generates .mat file for the States dataset.

preprocessGrades.m
Generates .mat files for the Grades datasets (difficulty 1-5).

GenPreds_paper.m
Generates models' predictions for each dataset.

WeightRegressionPlots_paper.m
Generate plots showing forecasters' posterior vs. their weights in SP/SC algorithm (Figures 3, 9).

SP_MSplit_paper.m
Plots the relevant SP responses from forecasters using a median split (Figure 4a).

PSP_MSplit_paper.m
Plots the relevant SC responses from forecasters using a median split (Figure 4b).

PSPvsSPweights_States_TF.m
Plots the weights assigned to each quartile of forecasters in the States dataset (Figure 10).

PSPvsSPweights_Grades.m
Plots the weights assigned to each quartile of forecasters in the Grades dataset (Figure 11).

ComputeContributions_gs.m
Compute differences in high-performers' vs. low-performers' contributions in the Grades dataset.

ComputeContributions_m4.m
Compute differences in high-performers' vs. low-performers' contributions in the States dataset.

PlotContributions_Datasets.m
Plots the difference in high-performers' vs. low-performers' contributions (Figure 5).

plotResults_paper_MCC.m
Generate plots for the MCC for each algorithm on each dataset (Figure 6).

computeMCCs_Results.m
Compute differences in MCC between each model for each dataset.

HybridSP.m
Computes difference in MCC for hybrid SP/SC algorithm vs. individual SP/SC algorithm.

CalibrationPlots.m
Generate calibration plots for each dataset (Figures 12, 13).

Supplementary Functions ("Paper" subfolder)
There are several supplementary functions used in the analysis. Below is a short description of each function.

acc.m
Computes the accuracy between predictions and outcomes.

BSCORE.m
Computes a transformed Brier score on a set of forecasts.

calibrationCurve_paper.m
Computes the accuracy and bins for each set of predictions and outcomes, which are then used as inputs for the calibration curve plotting function.

colours.m
Gives RGB values for a selection of colours.

CW.m
Computes a confidence-weighted prediction from forecasts.

ds.m
Computes decision similarity between forecasts and outcomes.

f1.m
Computes the f1 score between forecasts and outcomes.

fullfig.m
Maximises the figure window.

hline.m
Plots a horizontal line.

kappa.m
Computes Cohen's Kappa between forecasts and outcomes.

kappaDiff.m
Computes the difference in Cohen's Kappa between two sets of forecasts and outcomes.

ksr.m
Generates a kernel smoothing regression on a set of points.

MC.m
Computes the predictions for a max-confidence model.

MCC.m
Computes Matthew's Correlation Coefficient between forecasts and outcomes.

MCCdiff.m
Computes the difference in MCC between two sets of forecasts and outcomes.

MV.m
Generates predictions for a majority vote model from forecasts.

plotCalibrationCurve_paper.m
Plotting function called by CalibrationPlots.m for generating the calibration curve plots.

plotRegressionPrior.m
Computes the theoretical Prior for each dataset based on forecasts and probability meta-predictions.

PSP.m
Generates predictions for the Surprisingly Confident algorithm.

PSPvsSPweights_States.m
Compares the weights assigned to each quartile by the SC vs. the SP algorithms in the US states dataset used in Experiment 1.

SP.m
Generates predictions for the Surprisingly Popular algorithm.

suplabel.m
Generates a super-label for plots beyond the spacing of standard subplots.

vline.m
Plots a vertical line.

Plots ("Figures" subfolder)

Below is a description of each set of plots generated by the analysis code. As before, '%s' indicates the two-letter string for each dataset.

SP_WeightRegressionPlots_paper_%s.png
Plots showing forecasters' posterior vs. their weights in SP algorithm for each dataset (Figures 3 and 9).

PSP_WeightRegressionPlots_paper_%s.png
Plots showing forecasters' posterior vs. their weights in SC algorithm for each dataset (Figures 3 and 9).

SP_medianSplit_%s.png
Plots of the relevant SP responses from forecasters using a median split for each dataset (Figure 4a).

PSP_medianSplit_%s.png
Plots of the relevant SC responses from forecasters using a median split for each dataset (Figure 4b).

PlotContributions_Datasets.png
Plots showing the difference in high-performers' vs. low-performers' contributions in each Grades dataset (Figure 5).

plotResults_MCC.png
Plots showing for the MCC for each algorithm on each dataset (Figure 6).

PSPvsSPweights_States_TF.png
Plots showing the weights assigned to each quartile of forecasters in the States dataset (Figure 10).

PSPvsSPweights_Grades.png
Plots showing the weights assigned to each quartile of forecasters in the Grades dataset (Figure 11).

plotCalibrationCurve_paper_%s.png
Calibration plots for each dataset (Figures 12, 13).

Simulations ("Simulations in Appendix A" subfolder)

Appendix A includes a simulation exercise where the average weight and variance is calculated for the SC and SP algorithm for randomly generated information systems. The file "Appendix A simulations and macros.xlsm" contains two tabs:

- The first tab, "Asymmetric Distributions," is used to construct Figure 7 in the paper and uses the macro "Normal_Approximations" to run. Press Alt+F11 in excel to access this macro.

- The second tab, "Sample Distributions," is used for the simulation described in the main text and uses the "sample_weights" macro to run. We recommend turning off screen updating if you want to run this code. It will take about 4 hours to run.
  - The output from the simulation is large and has been saved as a separate Stata file called "RandomlyGeneratedInfoSystems.dta."
  - This data is analyzed with the Stata .do file called "Analysis of Randomly Generated Information Systems (Appendix A).do"

The simulation output code has been analyzed in Stata 15.1. To run, open Stata and change the working directory to the one where you have saved the dataset and do file. You can then run the do file. This will produce a text file called "Appendix A Analysis.txt," which matches the reported tests and numbers in the paper.

II. Experiment Replication Code ("Experiment Replication Files" subfolder)

Our main paper describes in detail how we ran each of our two experiments. In the Experiment Replication Files folder we have included five files that describe our experimental protocols and that can be used to replicate the experiments.

1. Subject Pool, Experiment Protocols, and Pilots.pdf – This document describes how we selected our subject pool, outlines the protocols used in each experiment, and reports on a pilot session.
2. QuestionsList_Grades.pdf – list of all the questions in the Grades Dataset
3. QuestionsList_States.pdf – list of all the questions in the States Dataset
4. ScreenShot_Grades.png – This is a screen shot of a question from the Grades Experiment. It shows the ordering of questions and the slider interfaces used.
5. ScreenShot_States.png – This is a screen shot of a question from the States Dataset. It shows the ordering of questions and the slider interfaces used.

For the States dataset, we do not allow participants to progress if they (i) select True and report that the probability that the event is true is below 50% or (ii) select False and report that the probability that the event is true is above 50%. This is done via a pop-up box that is generated when the continue button is pressed.