



INTEGO
GROUP

***USING THE GEOMETRIC PROPERTIES OF
DATASETS TO ANALYZE THE ACCURACY OF
COVID-19 FORECASTING MODELS***



US Connect 2022



Kostiantyn Drach, PhD

Researcher, Institute of Science and Technology Austria, Klosterneuburg, Austria/ Intego Group, Kharkiv, Ukraine

Kostiantyn graduated from Kharkiv National University with a Master's degree in Mathematics. He obtained his Ph.D. in Mathematics with a focus on geometry and topology from the Institute for Low Temperature Physics in Kharkiv, Ukraine. Over the last several years, he has participated in various research projects in pure and applied mathematics. Kostiantyn is a core external member of an R&D team of data scientists studying clinical data using methods of applied topology.

Iryna Kotenko

Site Lead, Intego Group, Kharkiv, Ukraine

Iryna graduated with a Master's in Quantitative Management in Health Analytics from Duke University. Her seventeen-year career as a statistical programmer spans numerous CROs as a clinical SAS programmer and data analyst. Over the last ten years, she has held the position of Site Lead for Clinical Programming Services, where she leads a group of over 200 statistical programmers in Ukraine, Poland, and the US. She is also heavily involved in a partnership program with Kharkiv National University and University of Central Florida.

The goals of the experiment:

- *Introduce machine-learning techniques* that rely on graphs as a fundamental approach for structuring, analyzing and forecasting COVID-19-related data
- *Build a topological model* (a graph) for the time series of COVID-19 confirmed cases in the USA
- Focusing on the geometric properties of data, *assess the prediction accuracy* of various existing forecasting models
- Using the topological model, build **dynamic ensembles** of forecasts with improved prediction accuracy
- *Compare* dynamic ensembles with the *CDC* forecasting ensemble

Experiment step-by-step:

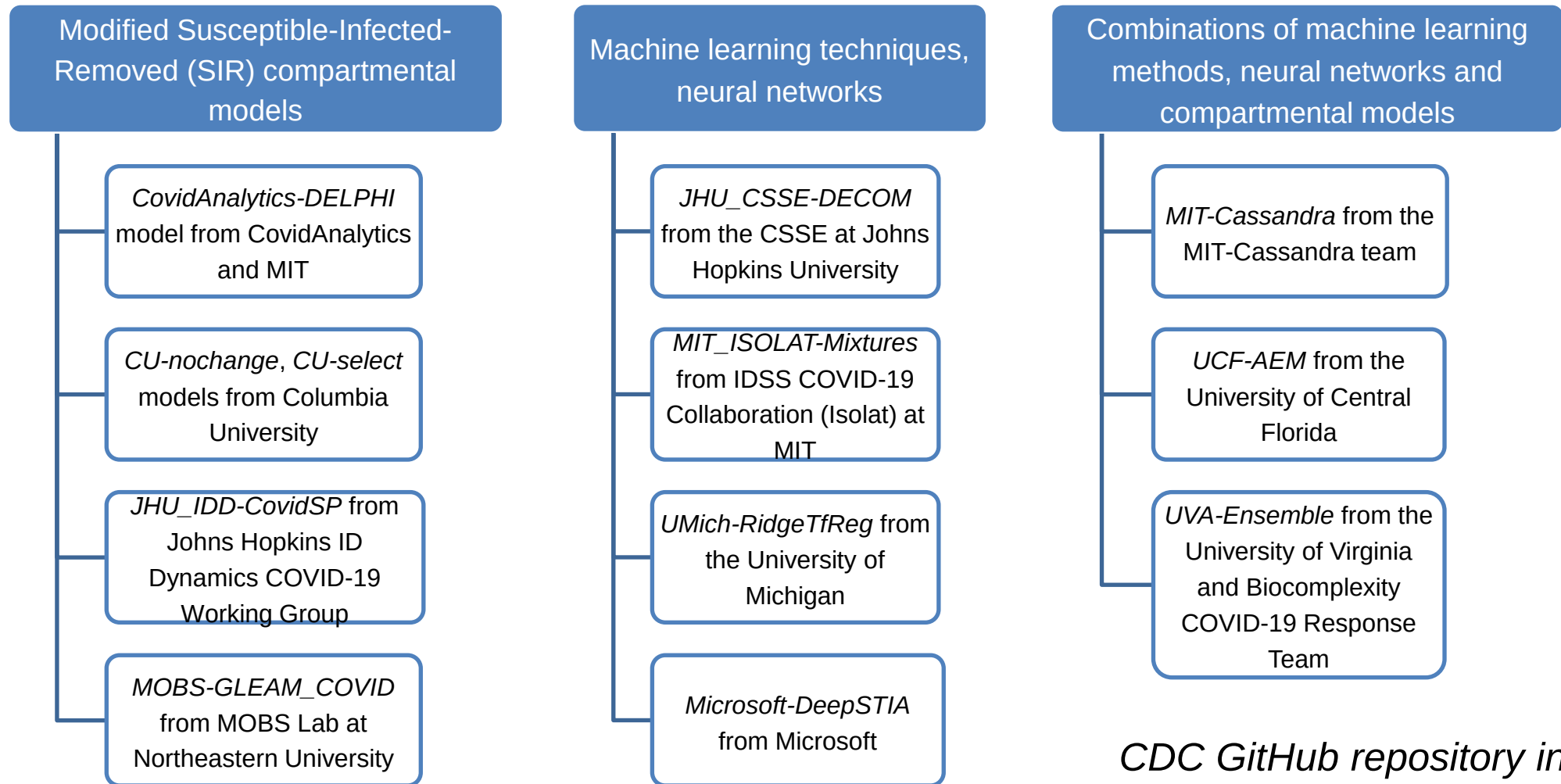
- 1) Process epidemic data (of COVID-19 daily cases within the USA)
- 2) Using *Topological Data Analysis*, build a **graph** to unveil geometric properties and hidden interdependencies of the dataset
- 3) Select the best-performing forecasting models from the *CDC-studied* list of models
- 4) Based on the TDA graph, build **dynamic ensembles** of the selected models
- 5) Evaluate and compare the dynamic ensembles with the CDC ensemble and other models



STEP 1.
PROCESS EPIDEMIC DATA

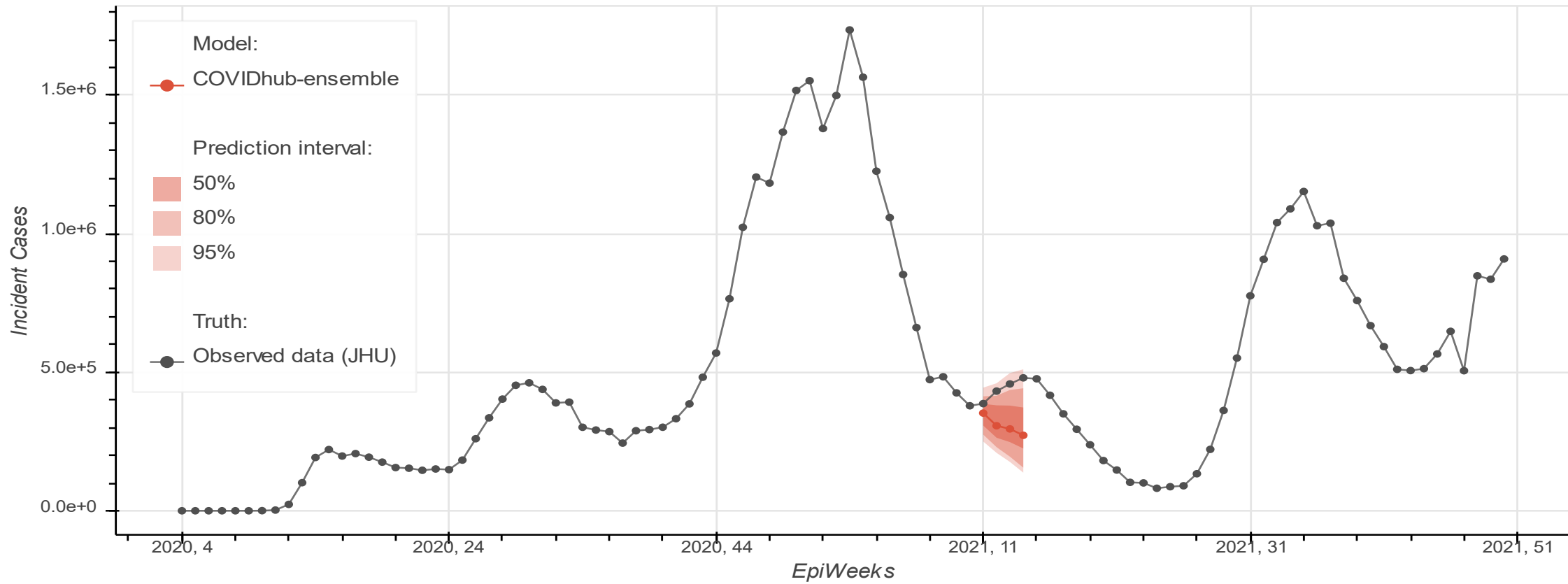


Mathematical models for forecasting COVID-19 development



CDC GitHub repository includes more than 100 models

COVID-19 time series



Evaluation of forecasts with Weighted Interval Score (WIS)

The **interval score (IS)** for $(1 - \alpha) \cdot 100\%$ prediction interval is:

$$IS_{\alpha}(F, y) = (u - \ell) + \frac{2}{\alpha} \cdot (\ell - y) \cdot \mathbf{1}(y < \ell) + \frac{2}{\alpha} \cdot (y - u) \cdot \mathbf{1}(y > u),$$

where y is a true value of the data,

ℓ , u are the $\alpha/2$ and $1 - \alpha/2$ quantiles of the prediction distribution F ,

$\mathbf{1}$ is the indicator function.

The **weighted interval score (WIS)** is a weighted sum of IS for $\alpha = 0.5, 0.2, 0.05$:

$$WIS(F, y) = \frac{1}{7} \cdot (|y - m| + 0.5 \cdot IS_{0.5}(F, y) + 0.2 \cdot IS_{0.2}(F, y) + 0.05 \cdot IS_{0.05}(F, y))$$

where m is the 50% quantile of F .

Lower WIS means that the forecast is more accurate



STEP 2.

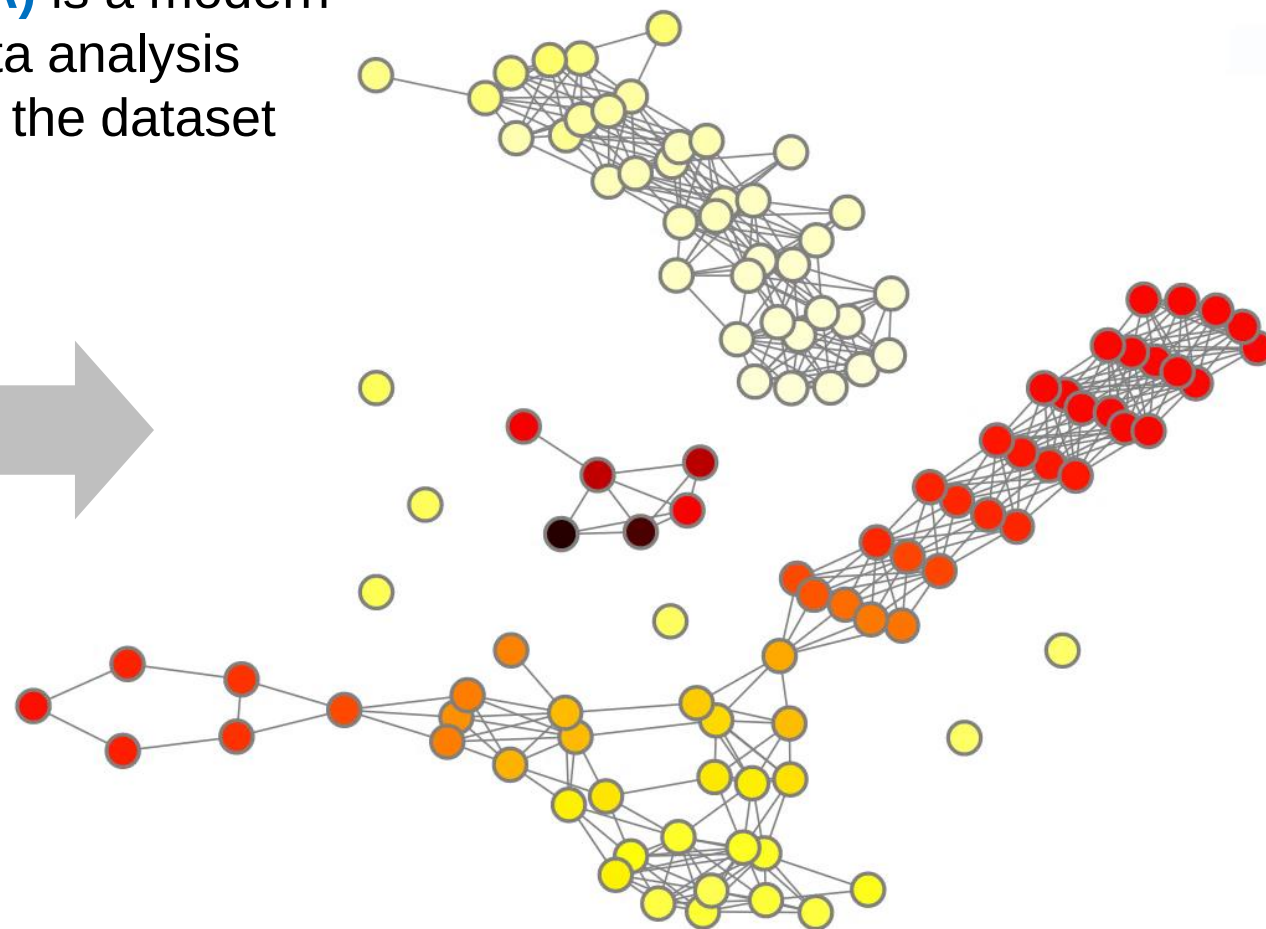
**BUILDING THE GRAPH USING
TOPOLOGICAL DATA ANALYSIS**



Topological Data Analysis (TDA) is a modern assumption-free approach for data analysis based on geometric properties of the dataset



Each **node** represents one subject



Two nodes are connected if they are **similar**, within **stratifying bins**, in terms of **outcomes**

In our experiment:

- Each node corresponds to a segment of the time series, transformed to a point in a multidimensional space.
- Transformation is done using **Takens' embedding**:

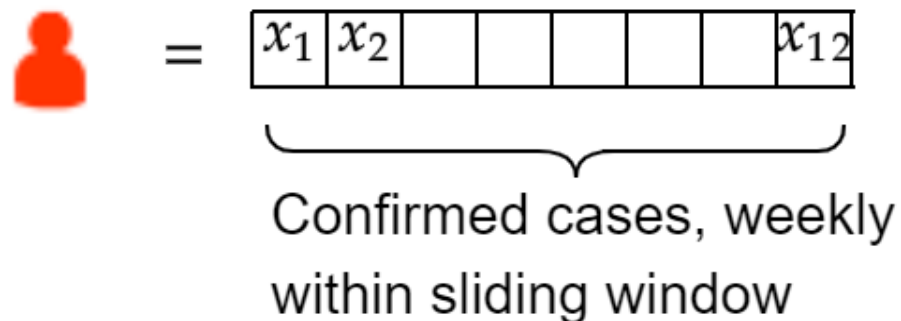
time series
 $x_1, x_2, \dots, x_N$



set of points of the form
 $(x_i, x_{i+\tau}, x_{i+2\tau}, \dots, x_{i+(d-1)\tau})$

- We select $\tau=1$ (time delay) and $d=12$ (embedding dimension) by the [method of mutual information](#) and the [method of false nearest neighbors](#), respectively.

*Thus, we encode the information of the time series as a point-cloud of outcomes in **12-dimensional space***



In our experiment:

- Stratifying bins arrange segments into several overlapping groups. Bins are chosen based on **the first eigenvector of a distance matrix** with respect to outcomes.
- Similarity is then measured within each bin separately.



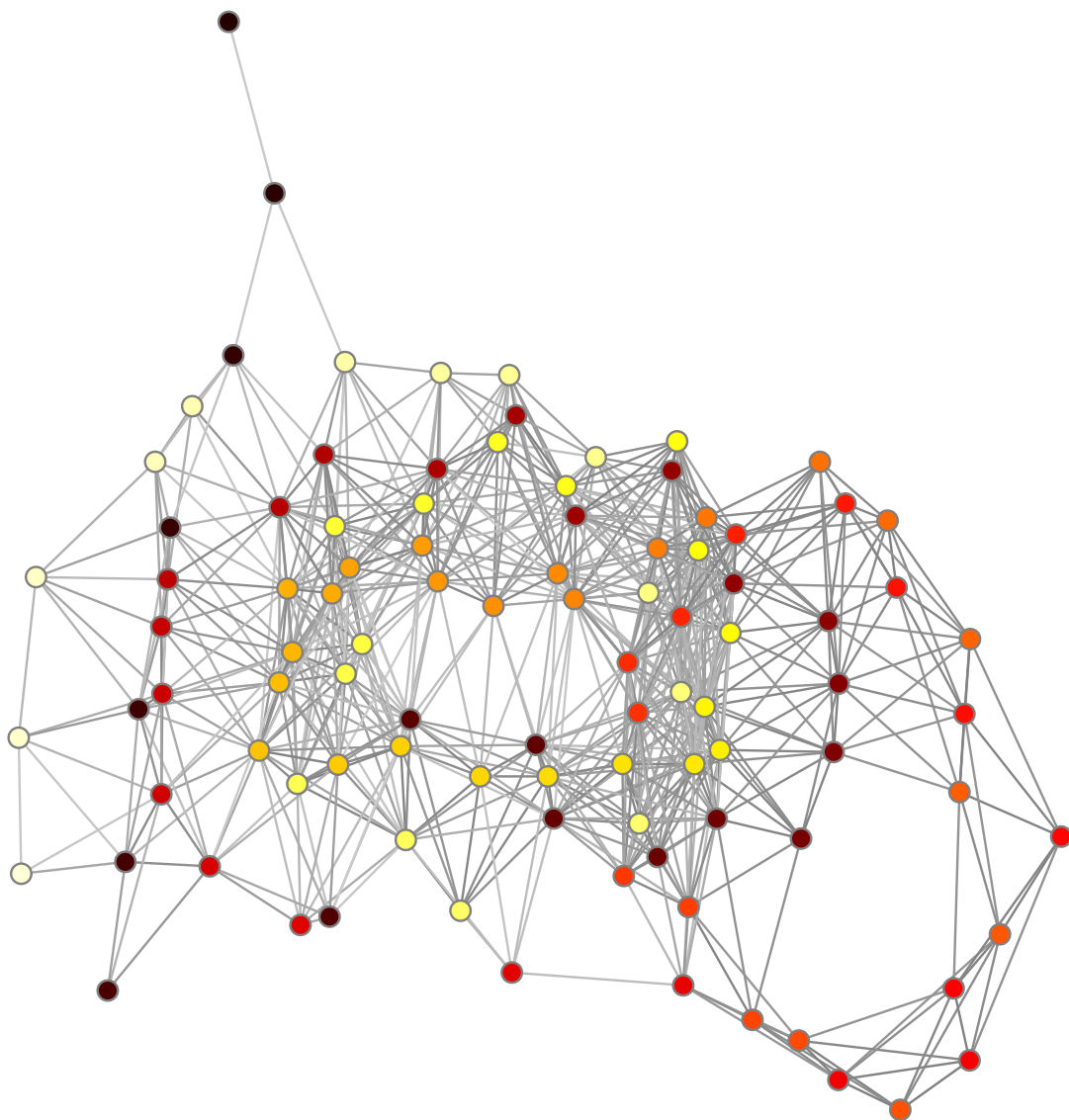
$$\text{Red Person} = \begin{bmatrix} x_1 & x_2 & & & & & x_{12} \end{bmatrix}$$

$$\text{Yellow Person} = \begin{bmatrix} y_1 & y_2 & & & & & y_{12} \end{bmatrix}$$

Cosine metric as a distance function is used to reveal the segments of time series with “similar behavior”

$$d(\text{Red Person}, \text{Yellow Person}) = 1 - \frac{\sum_{i=1}^{12} x_i y_i}{\sqrt{\sum_{i=1}^{12} x_i^2} \sqrt{\sum_{i=1}^{12} y_i^2}}$$

Topological model of COVID-19 weekly cases in the USA



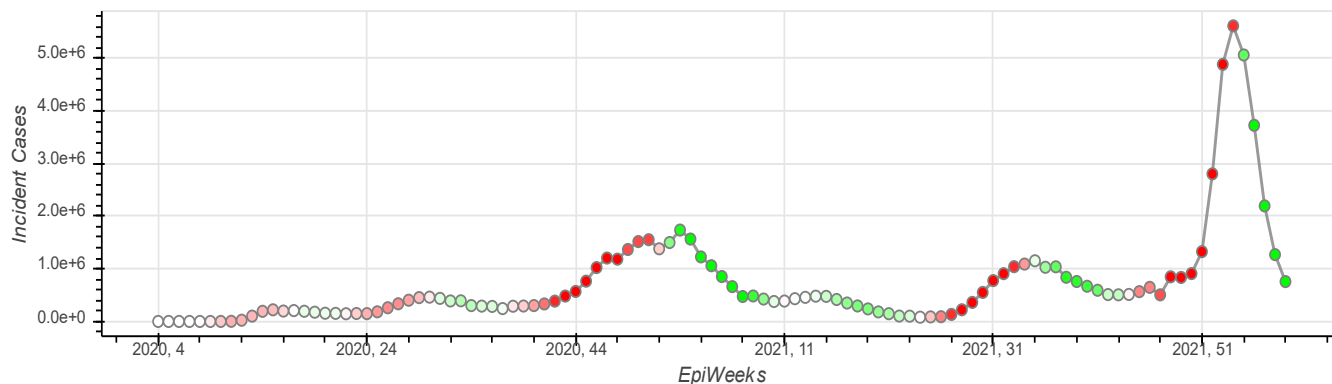
- Each node on the **graph** corresponds to a segment of length 12 weeks on the epidemic curve (total **98 nodes**)
- Two nodes are connected with an edge if the corresponding segments have similar behavior (increasing, decreasing or plateau-like)
- **Similarity** is measured by the cosine metric
- The color of nodes corresponds to the number of the week since the beginning of the pandemic: from a lighter color (yellow) to a darker color (brown)



STEP 3.
**SELECT BEST-PERFORMING
MODELS**

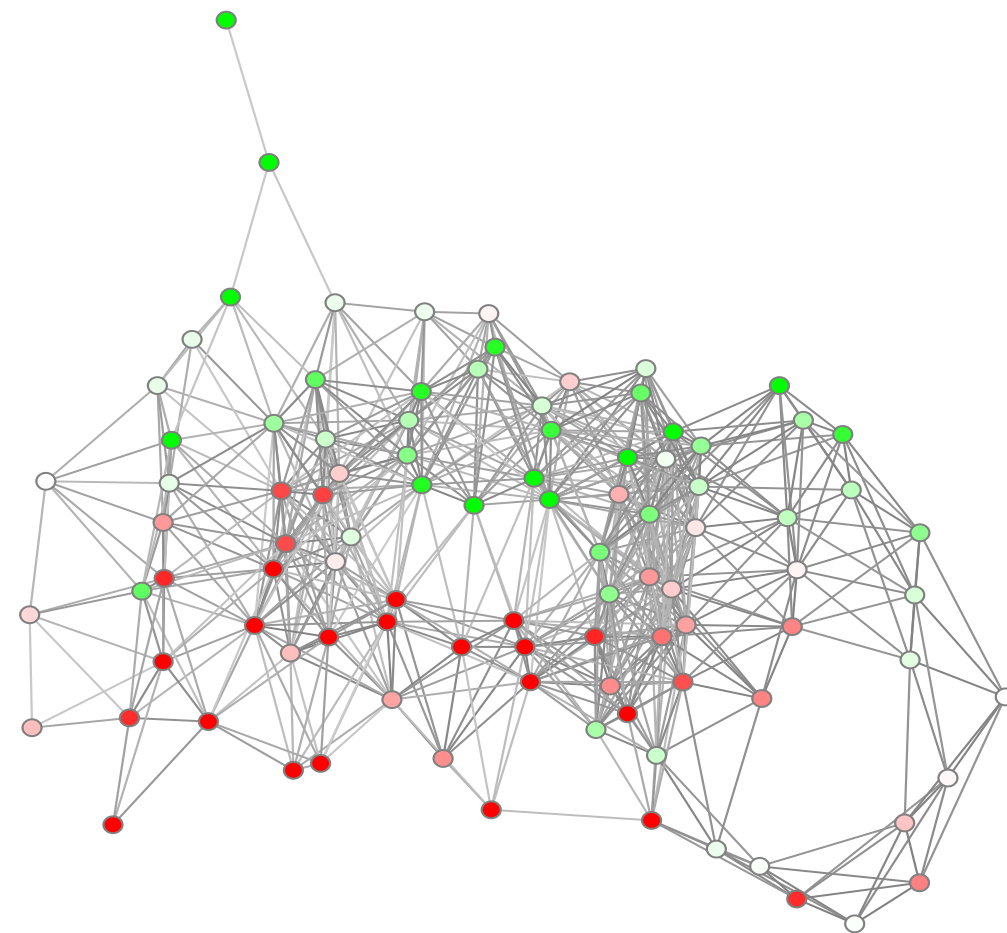


Behavior of the COVID-19 time series on the TDA graph



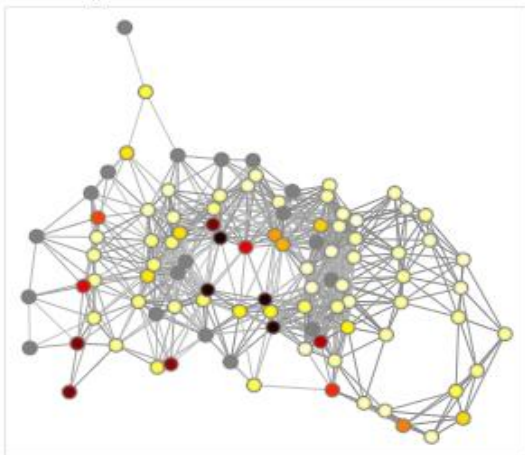
The time series for the USA and the TDA graph colored in accordance with the behavior of the curve:

- increasing (red)
- decreasing (green)
- plateau-like (white)

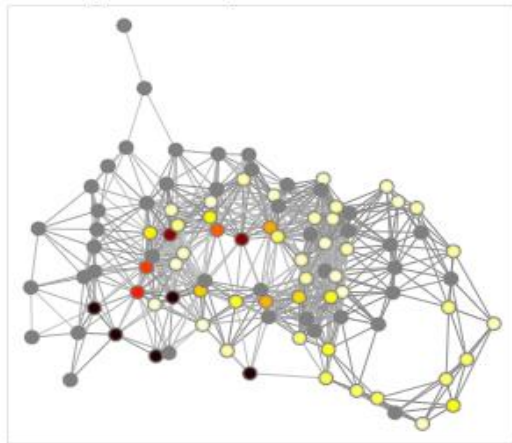


Identifying accuracy of models on the TDA graph

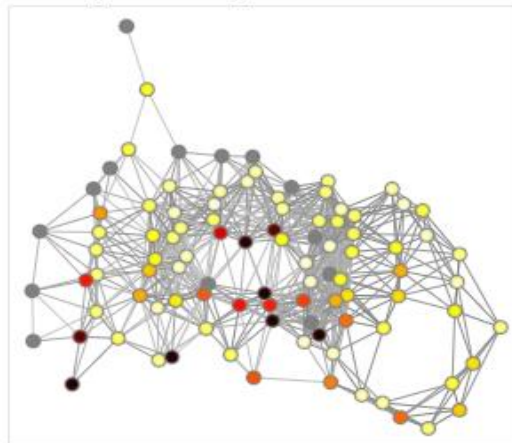
Coloring: CU-select



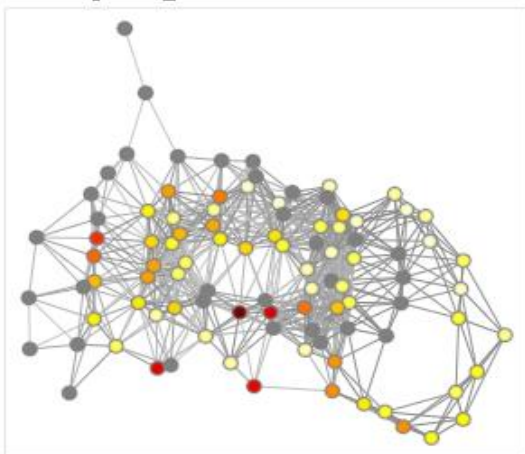
Coloring: Columbia_UNC-SurvCon



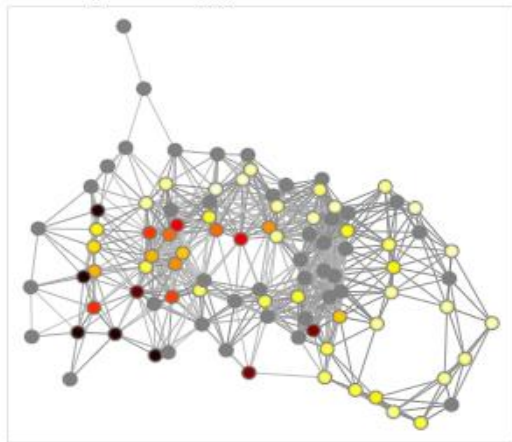
Coloring: CovidAnalytics-DELPHI



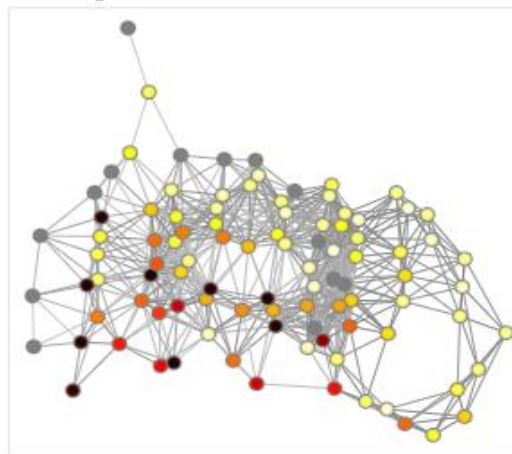
Coloring: JHU_IDD-CovidSP



Coloring: Karlen-pypm



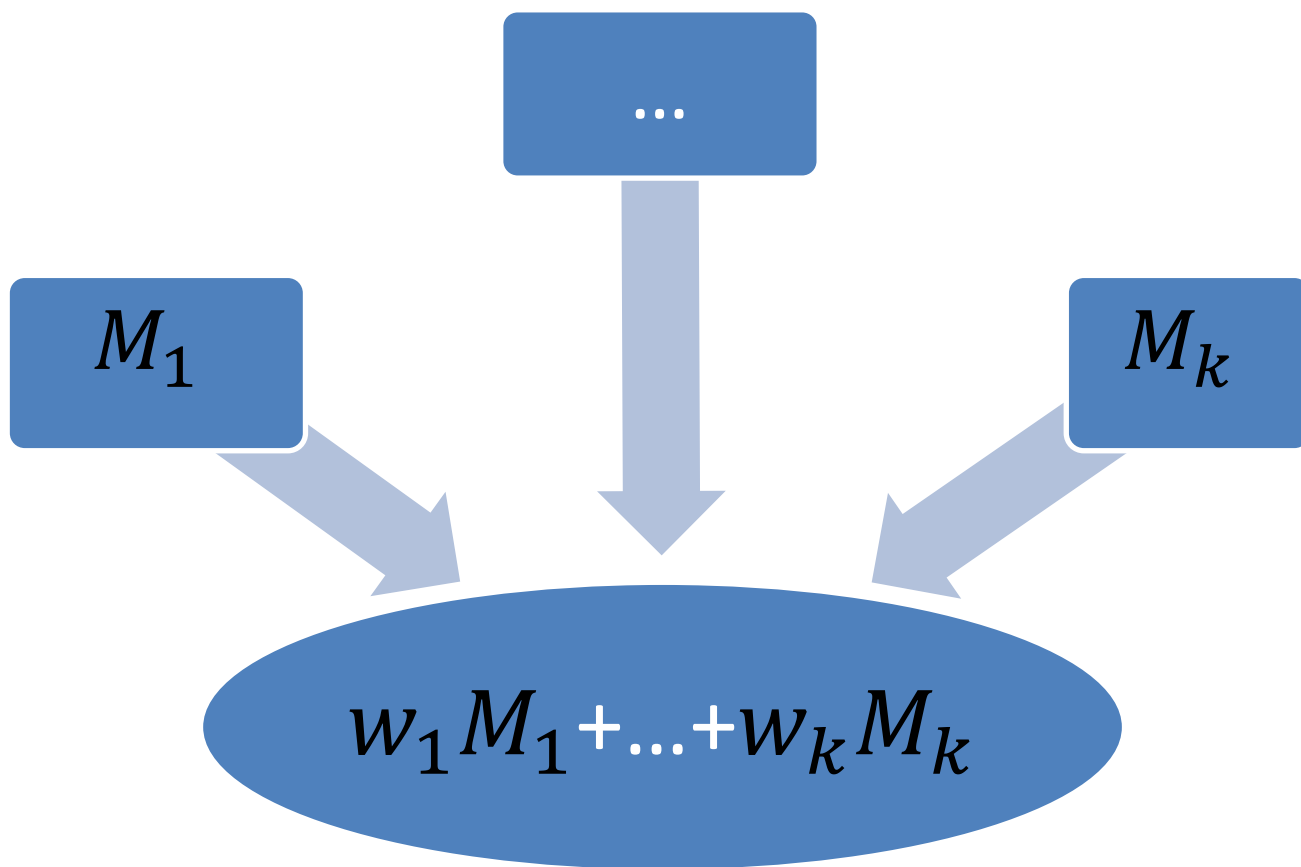
Coloring: RobertWalraven-ESG



Coloring by **WIS**:

- lighter colors → lower WIS
→ better forecasts
- darker colors → higher WIS
→ worse forecasts
- grey color means the
absence of the forecast for
this period

The idea of weighted dynamic ensemble



At a forecasting point of interest:

- Assign weights $w_1, w_2, \dots, w_k$ to models $M_1, M_2, \dots, M_k$, based on their performance in “similar” situations in the past
- “Similar” situations can be detected by the TDA graph.
- Models with better forecasts in the past are assigned greater weights at present



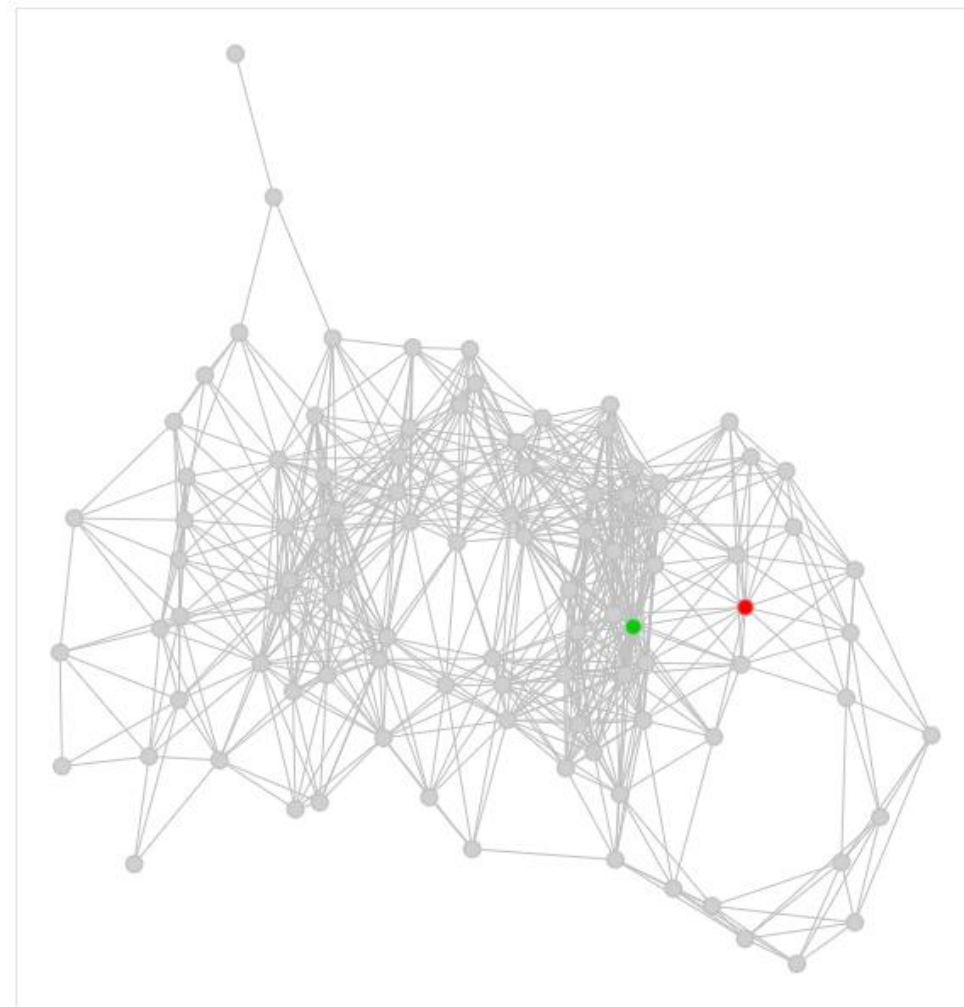
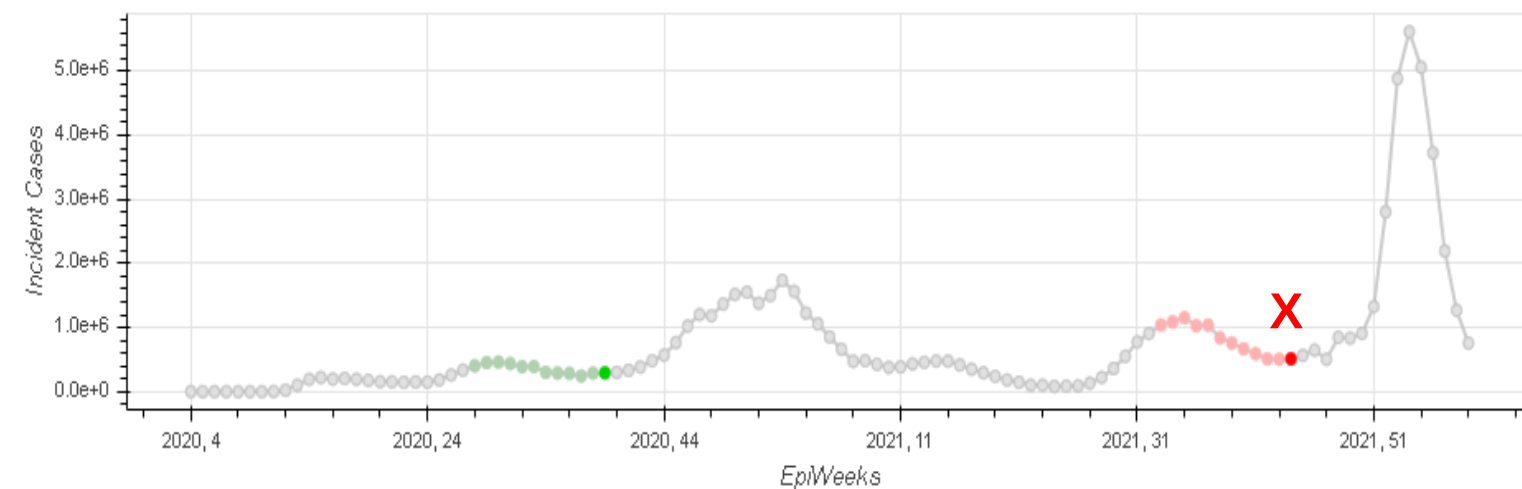
STEP 4.

BUILD DYNAMIC ENSEMBLES



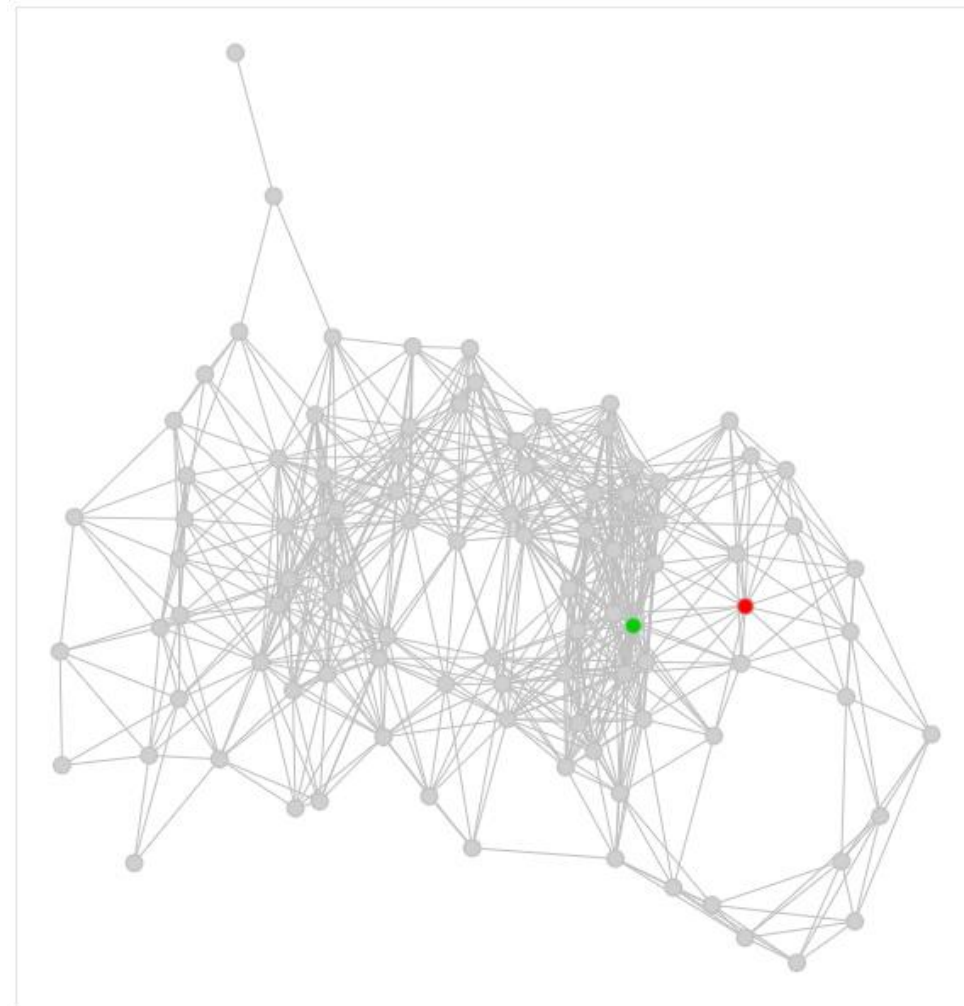
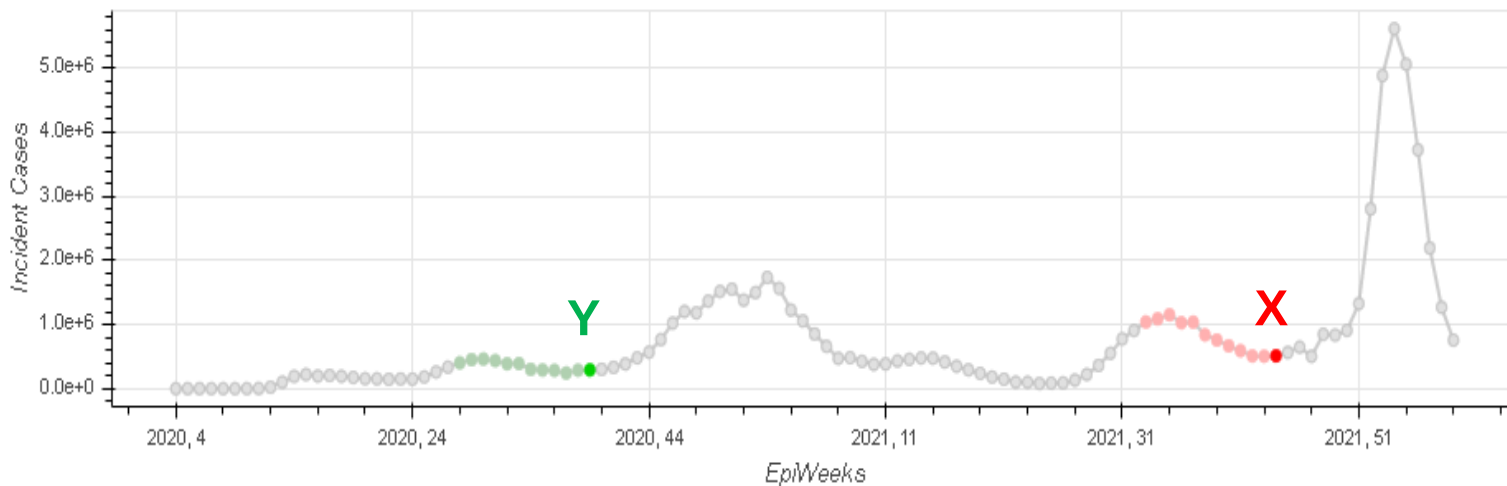
How to choose weights for the dynamic ensemble

1. Pick a forecasting point **X (red)** on the epidemic curve and the corresponding node on the graph **(red)**



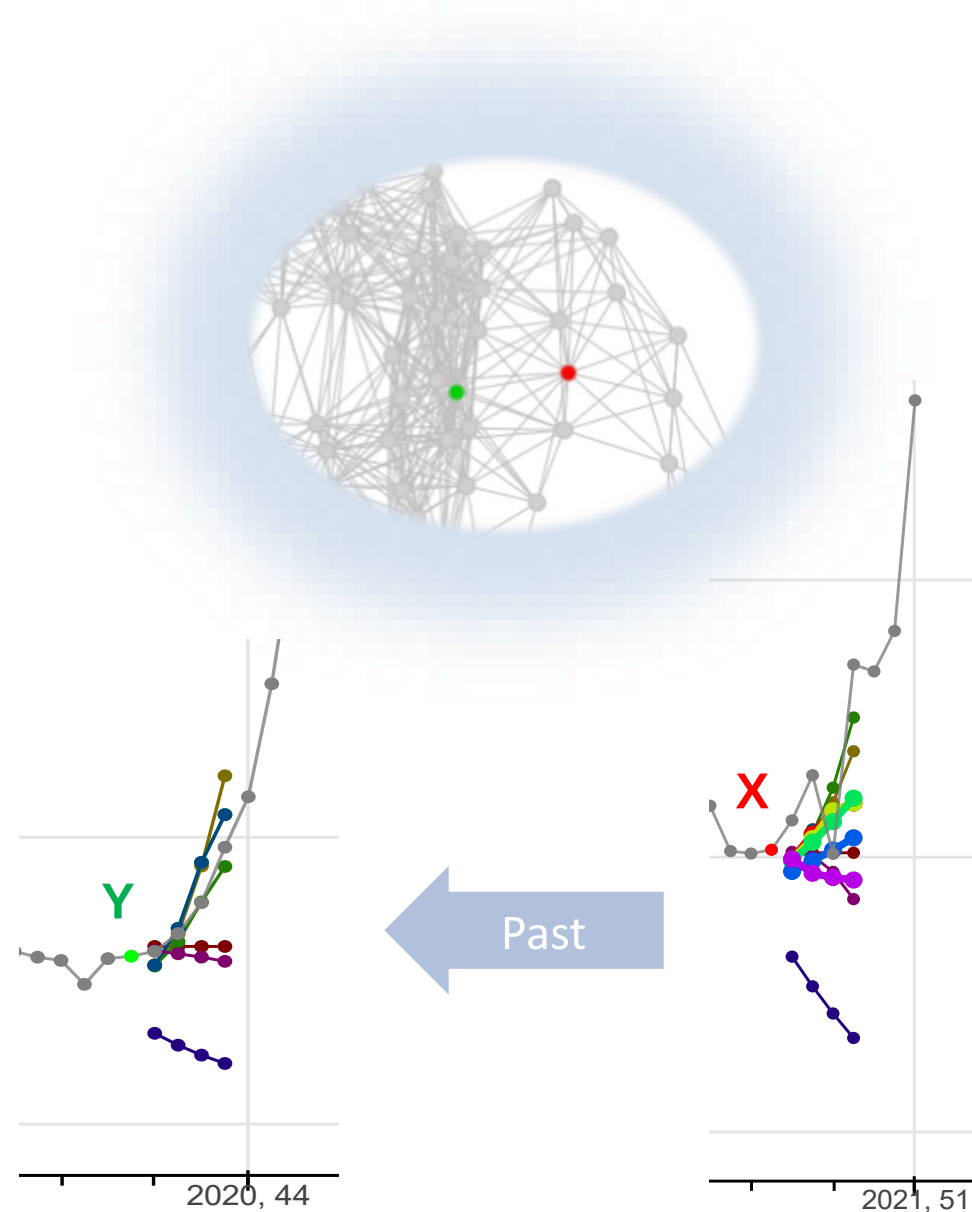
How to choose weights for the dynamic ensemble

1. Pick a forecasting point **X (red)** on the epidemic curve and the corresponding node on the graph **(red)**
2. In the graph, find the closest neighboring node **(green)** which corresponds to the point **Y (green)** located in the past w.r.t. **X** on the time series



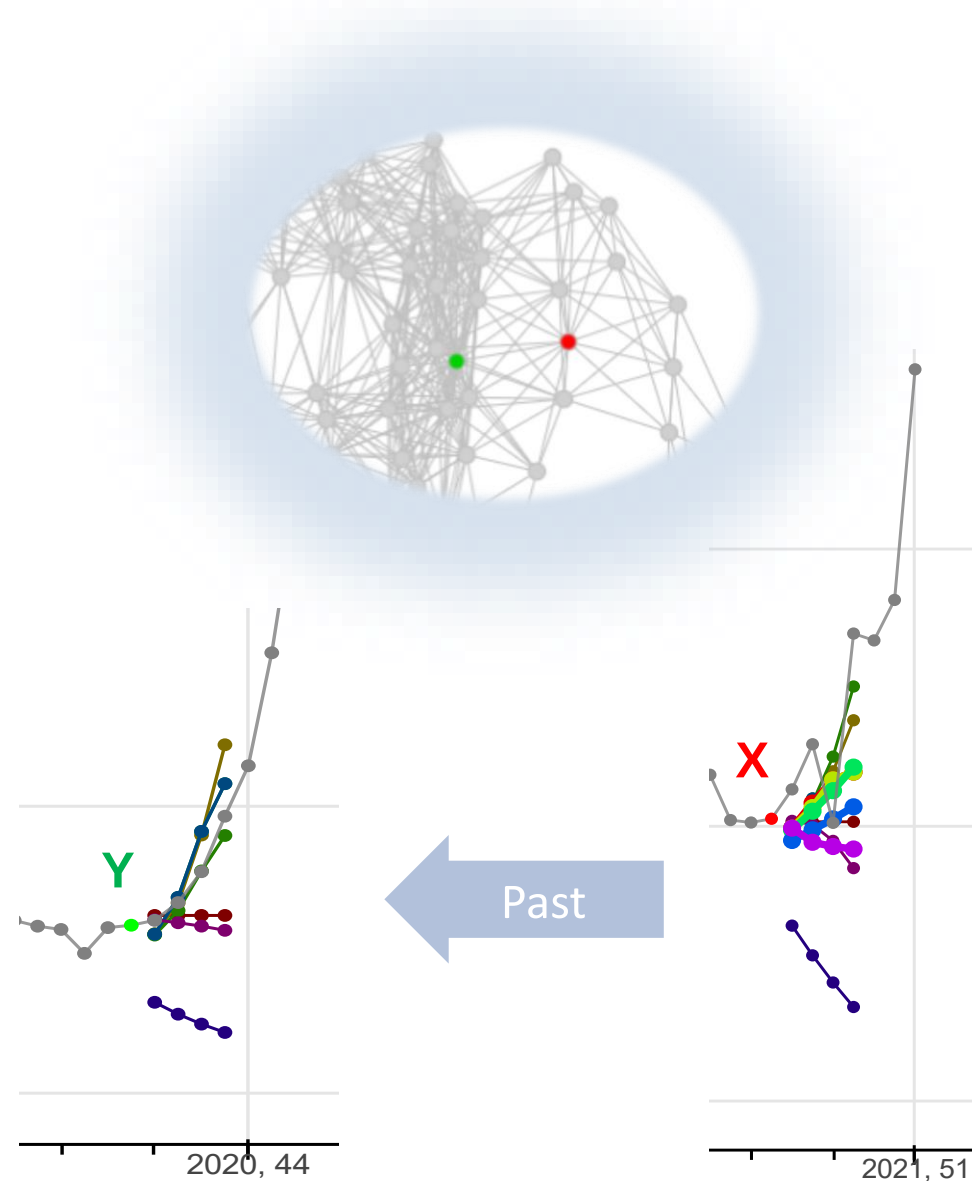
How to choose weights for the dynamic ensemble

1. Pick a forecasting point **X (red)** on the epidemic curve and the corresponding node on the graph **(red)**
2. In the graph, find the closest neighboring node **(green)** which corresponds to the point **Y (green)** located in the past w.r.t. **X** on the time series
3. Select at least three models that provide forecasts at **X** and **Y**



How to choose weights for the dynamic ensemble

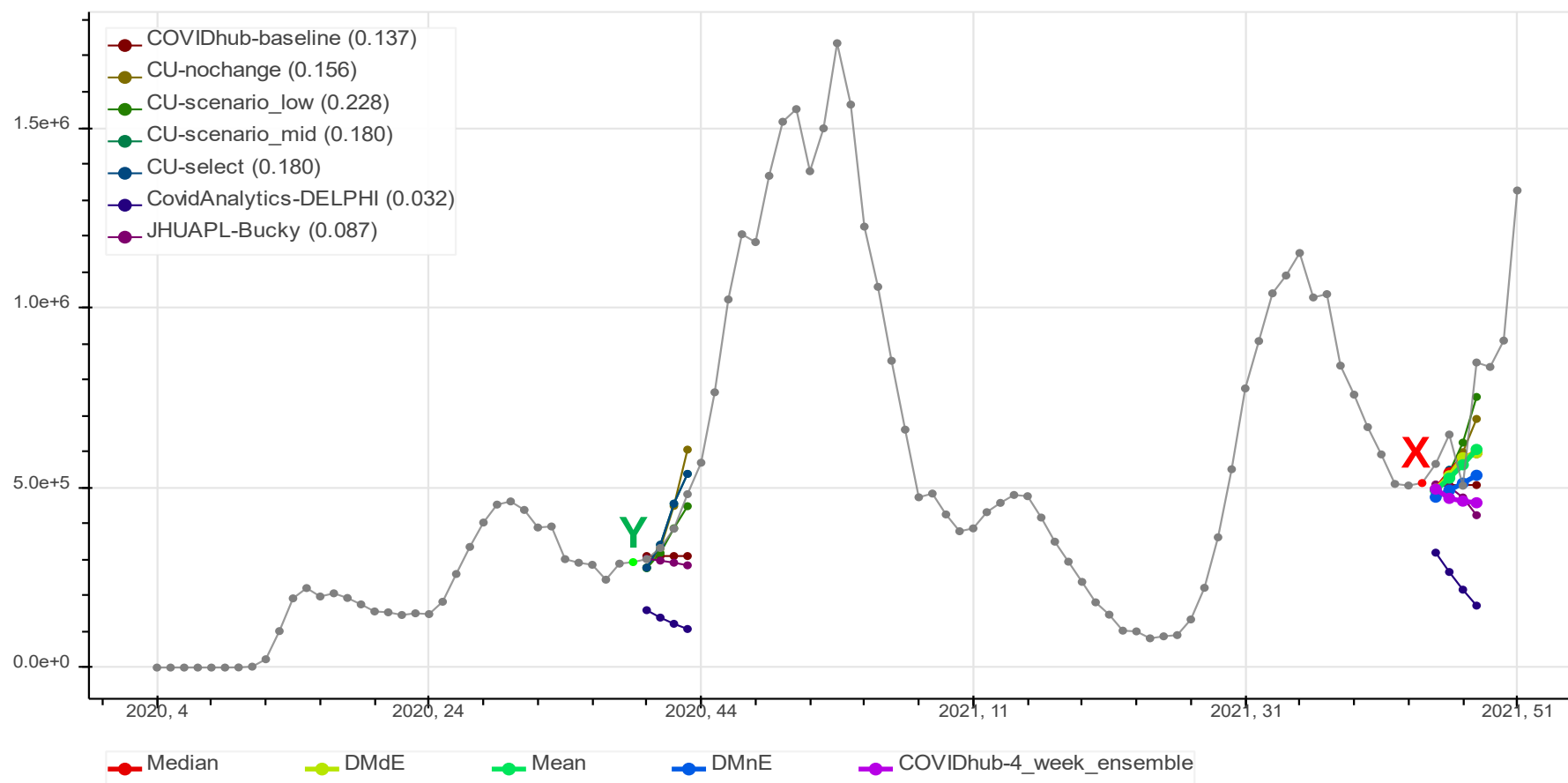
1. Pick a forecasting point **X (red)** on the epidemic curve and the corresponding node on the graph **(red)**
2. In the graph, find the closest neighboring node **(green)** which corresponds to the point **Y (green)** located in the past w.r.t. **X** on the time series
3. Select at least three models that provide forecasts at **X** and **Y**
4. As lower WIS means better forecasts, at **X** we choose the weights for the selected models equal to reverse normalized values of WIS for these models at **Y**



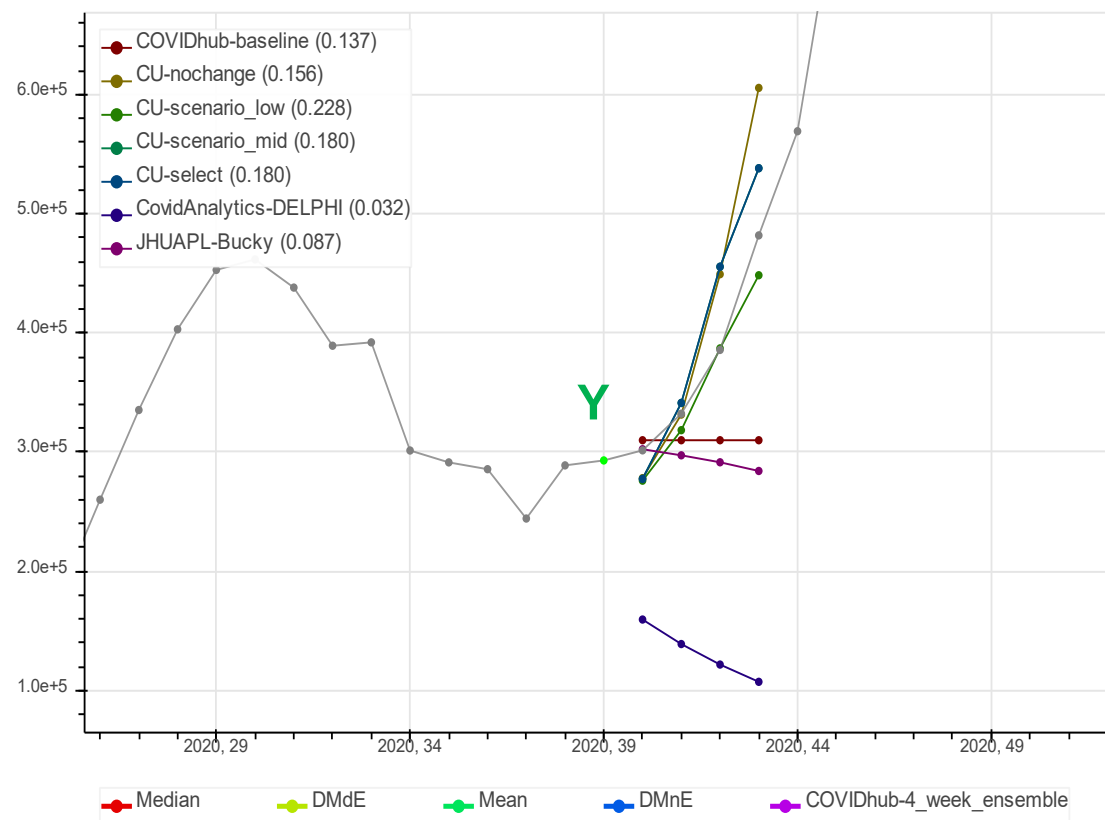
Dynamic Ensembles

We build a weighted mean ensemble and call it the *Dynamic Mean Ensemble* (DMnE).

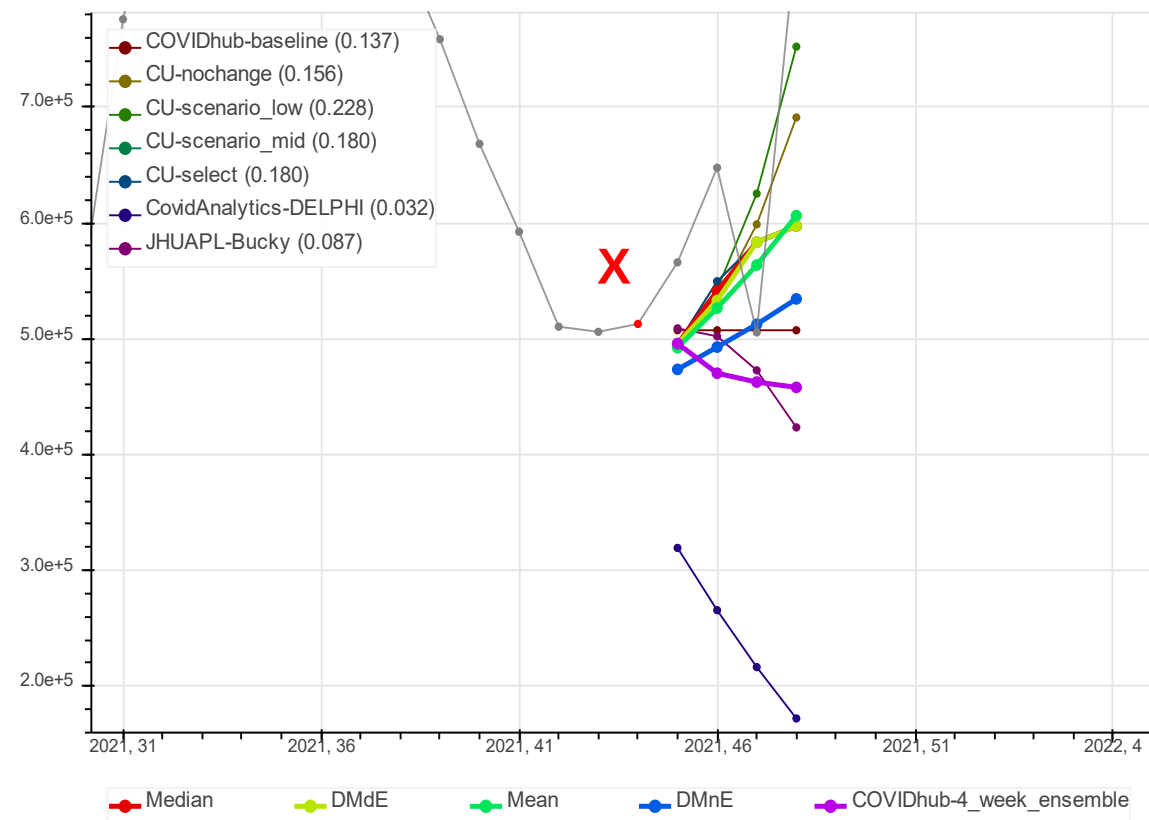
Similarly, the *Dynamic Median Ensemble* (DMdE) can be built.



Detalization of the forecast at the point X



Epiweek 39 of 2020 (similar point in the past)



Epiweek 44 of 2021 (forecasting point)



STEP 5.

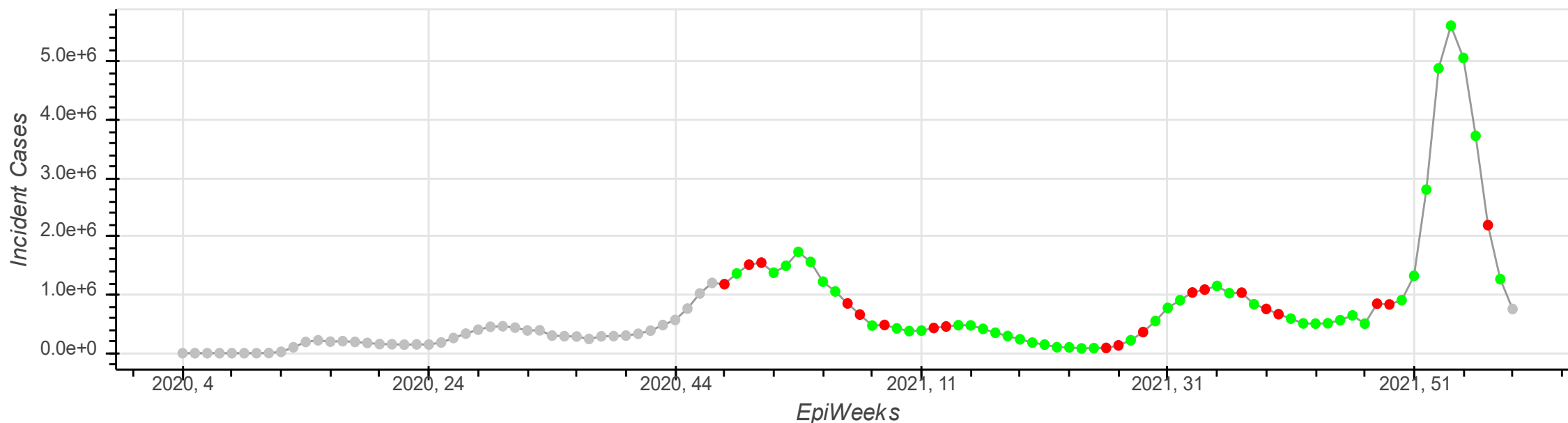
**COMPARE DYNAMIC ENSEMBLES
WITH OTHER MODELS**



Rates of fitting into the confidence intervals

Ensemble/confidence interval	50%	80%	95%
DMdE	0.360	0.616	0.732
Median	0.376	0.600	0.736
DMnE	0.412	0.684	0.800
Mean	0.384	0.640	0.792
CDC COVIDhub-4_week_ensemble	0.356	0.608	0.776

Mean Ensemble vs. Dynamic Mean Ensemble (DMnE)



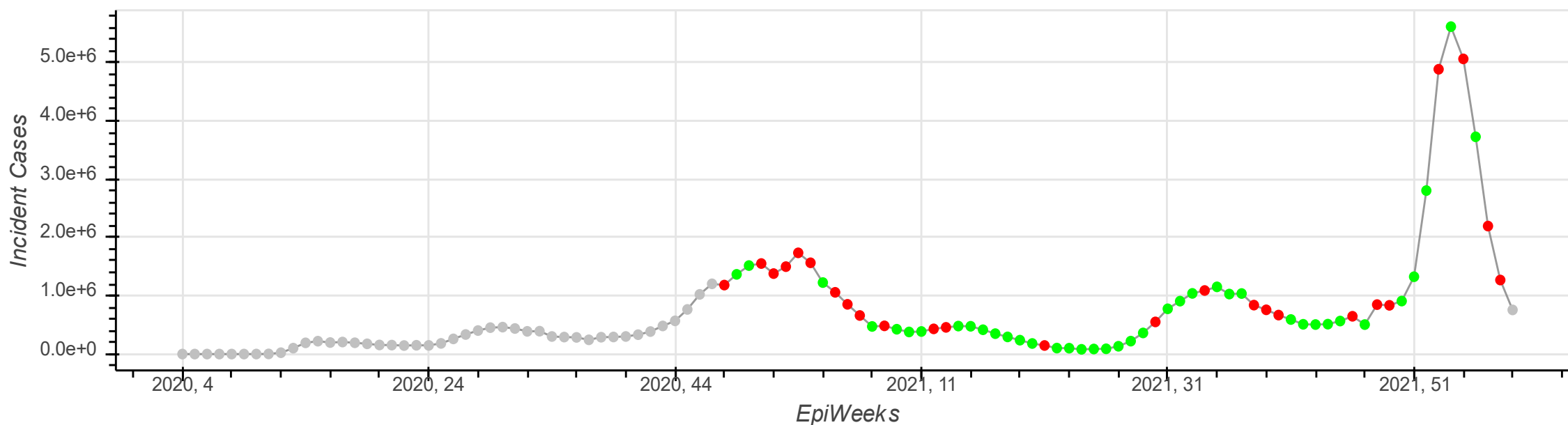
Coloring: **green** – the forecast of the DMnE is better (lower WIS)
red – the forecast of the Mean Ensemble is better

green points = 45
red points = 19



DMnE provides better forecasts in
70% cases

CDC ensemble vs. Dynamic Mean Ensemble (DMnE)



Coloring: **green** – the forecast of the DMnE is better (lower WIS)
red – the forecast of the CDC ensemble is better

green points = 39
red points = 25



**DMnE provides better forecasts in
61% cases**

Presentation takeaways:

1. **Topological Data Analysis** was used to build a model based on:
 - embedding the COVID-19 time series into 12-dimensional space with the points corresponding to the segments of the epidemic curve
 - similarity of the segment's behavior (increasing, decreasing, or plateau-like)
2. **Dynamic ensembles** of forecasts are constructed using TDA provide more accurate forecasts than the CDC ensemble and the unweighted ensemble built on the same set of forecasting models

Please [**download the paper**](#) to get more information about the design of the experiment and more findings discovered



THANK YOU

Intego Group, LLC
2300 Maitland Center Pkwy, Ste. 240,
Maitland, FL 32751
Phone: +1 (407) 512-1006
www.intego-group.com/tcdm

