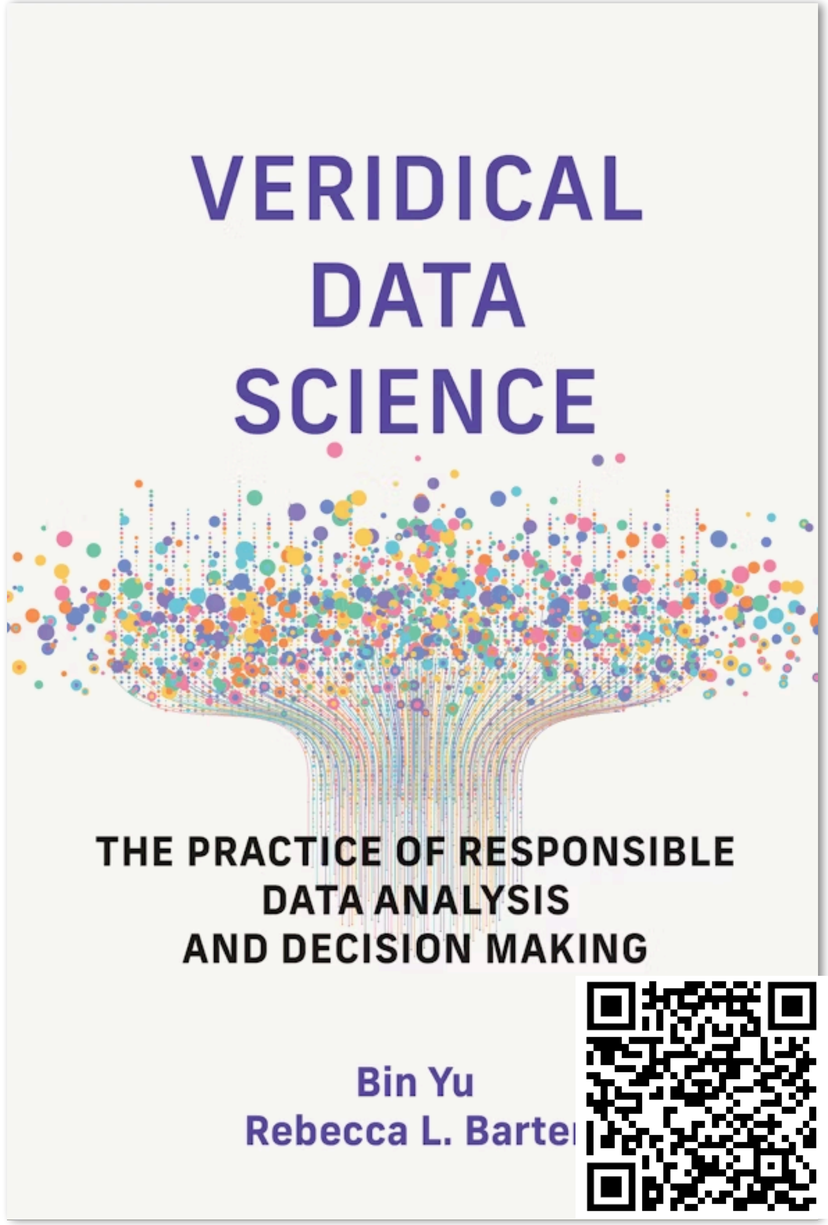
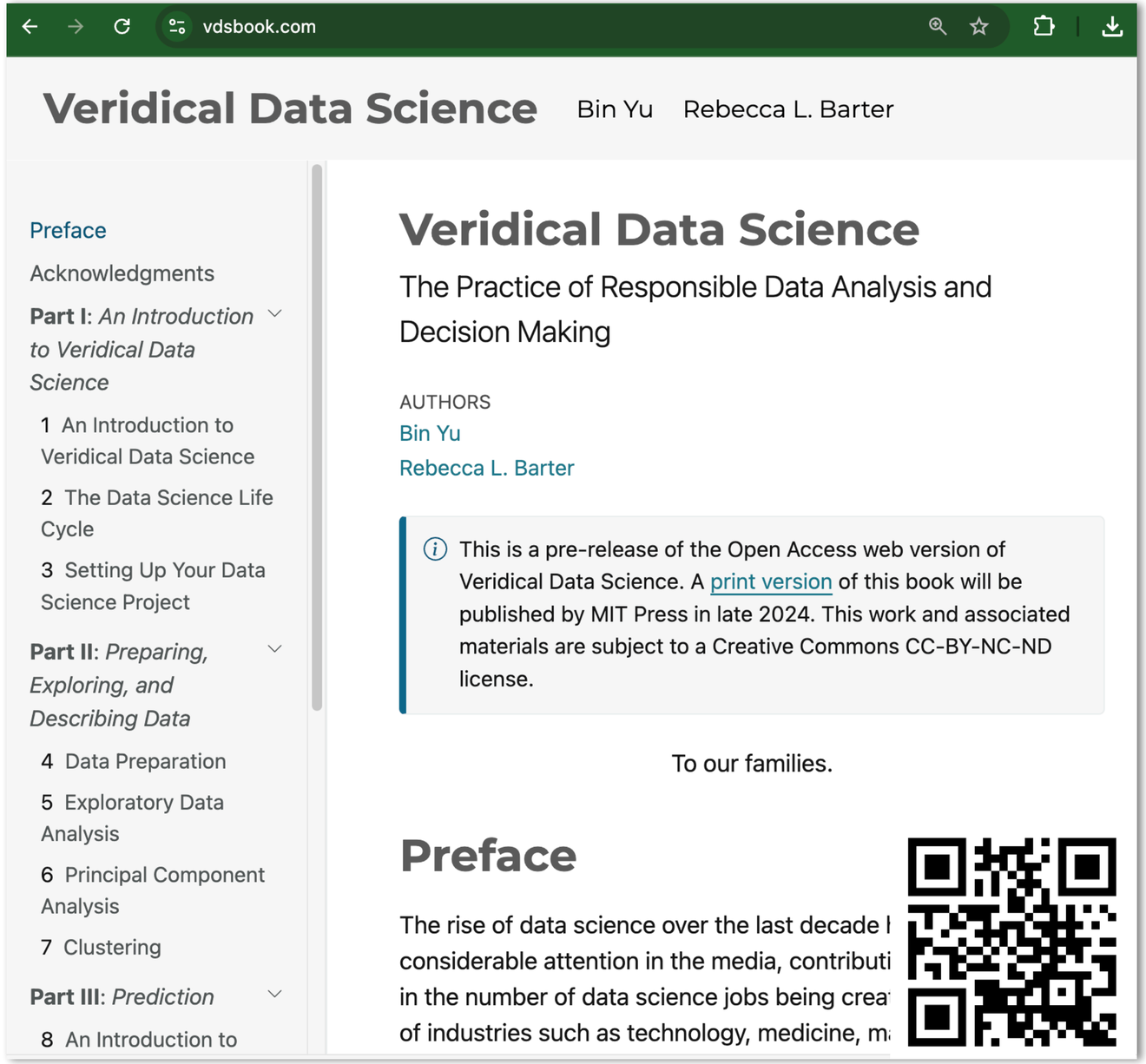
Veridical Data Science

The Practice of Responsible Data Analysis and Decision Making

Rebecca Barter





The MIT Press

Free online version: www.vdsbook.com

Veridical

Definitions

Definitions from Oxford Languages · Learn more

adjective **FORMAL**

truthful.
"Pilate's attitude to the veridical"

- coinciding with reality.
  "such memories are not necessarily veridical"

The origins of Veridical Data Science (VDS)

Veridical data science (VDS) is a philosophy and framework pioneered by Professor Bin Yu throughout her career

VDS provides a framework for **producing trustworthy data-driven** results, and **critically assessing** the **trustworthiness of data-driven results** in the context of domain science and reality

Original PNAS article: Yu and Kumbier (2020), *Veridical Data Science*

Setting the scene for Veridical Data Science

Real data (& thus data-driven results) imperfectly reflect reality

The **real world is too complex** to be perfectly captured within a dataset

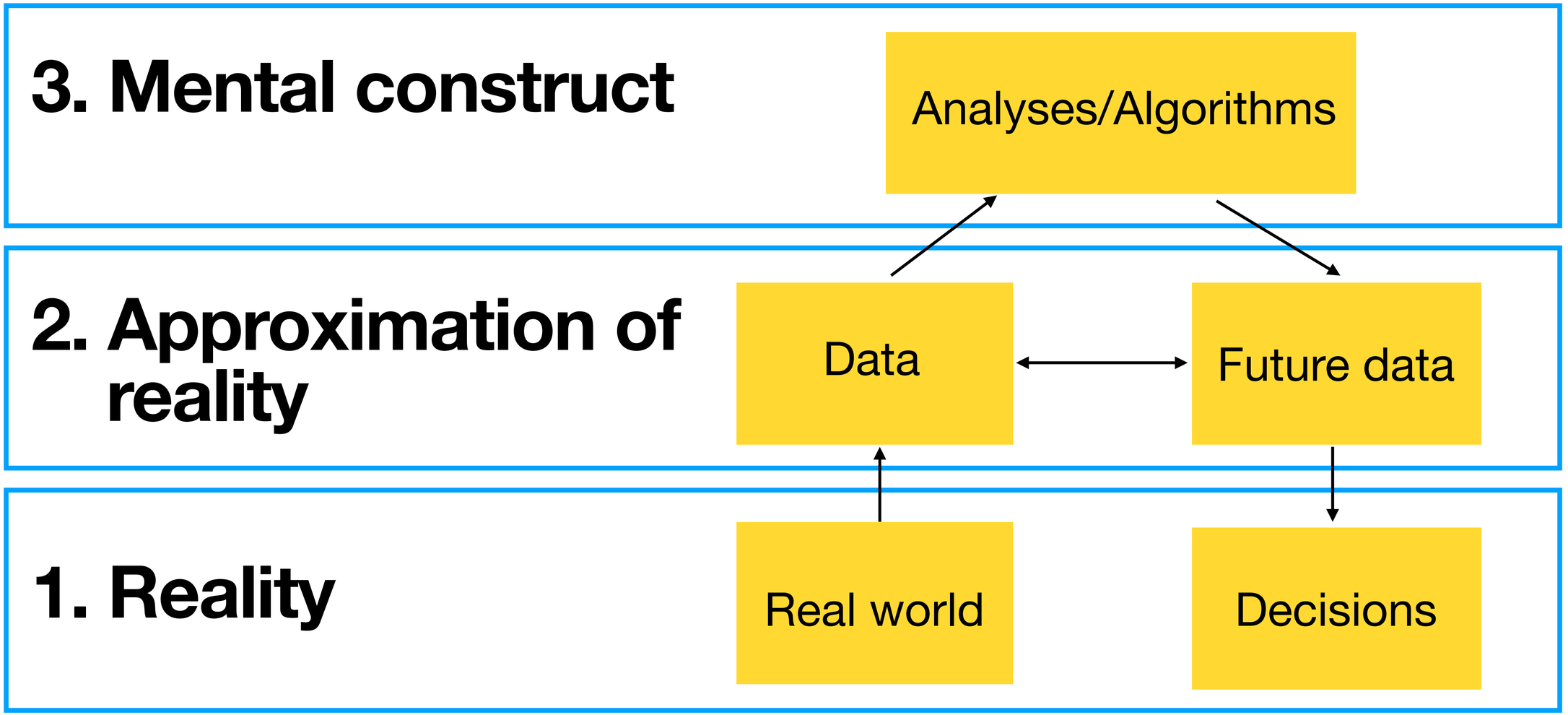
The extent that **data-driven conclusions** reflect the real world is upper-bounded by how well the underlying **data** reflects the real world

Images generated by Dall-e-3

Real data (& thus data-driven results) imperfectly reflect reality

Image generated by Dall-e-3

Algorithms are mental constructs built upon imperfect data



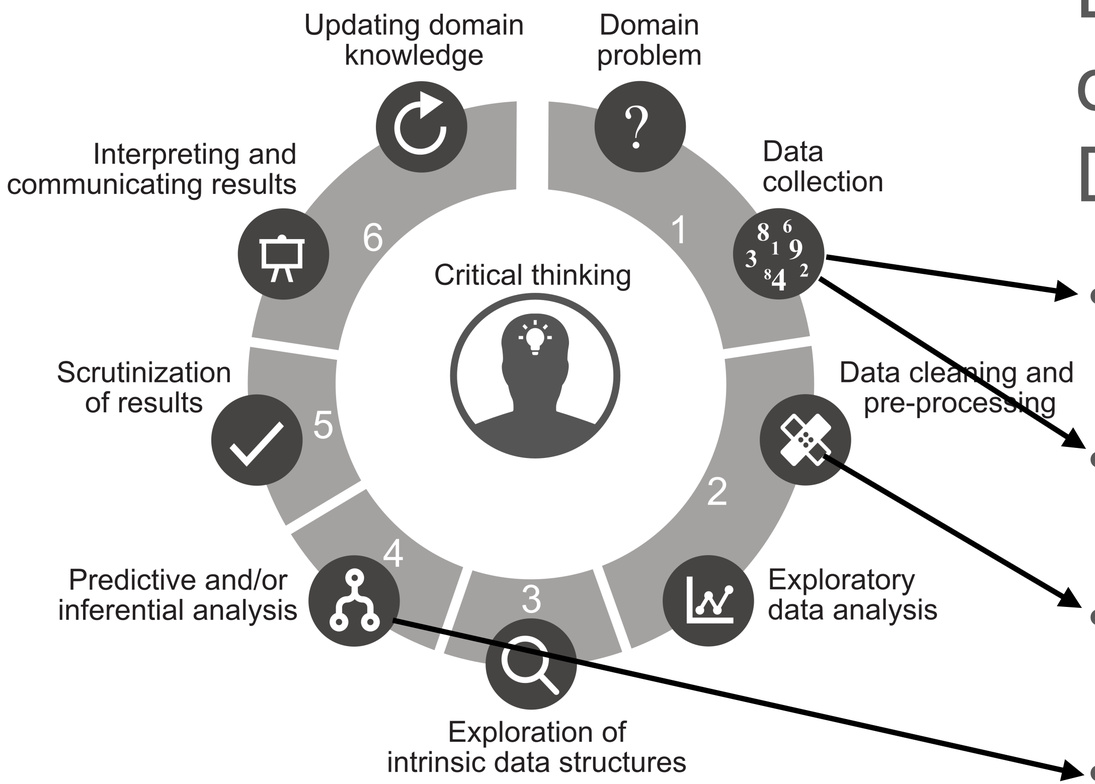
Uncertainty and judgment calls

Every data-driven result is a function of

- The **human** analyst
- The **data** collected
- The data cleaning **judgment calls**
- The **modeling/algorithm** choices

Image generated by Dall-e-3

The Data Science Life Cycle (DSLC)

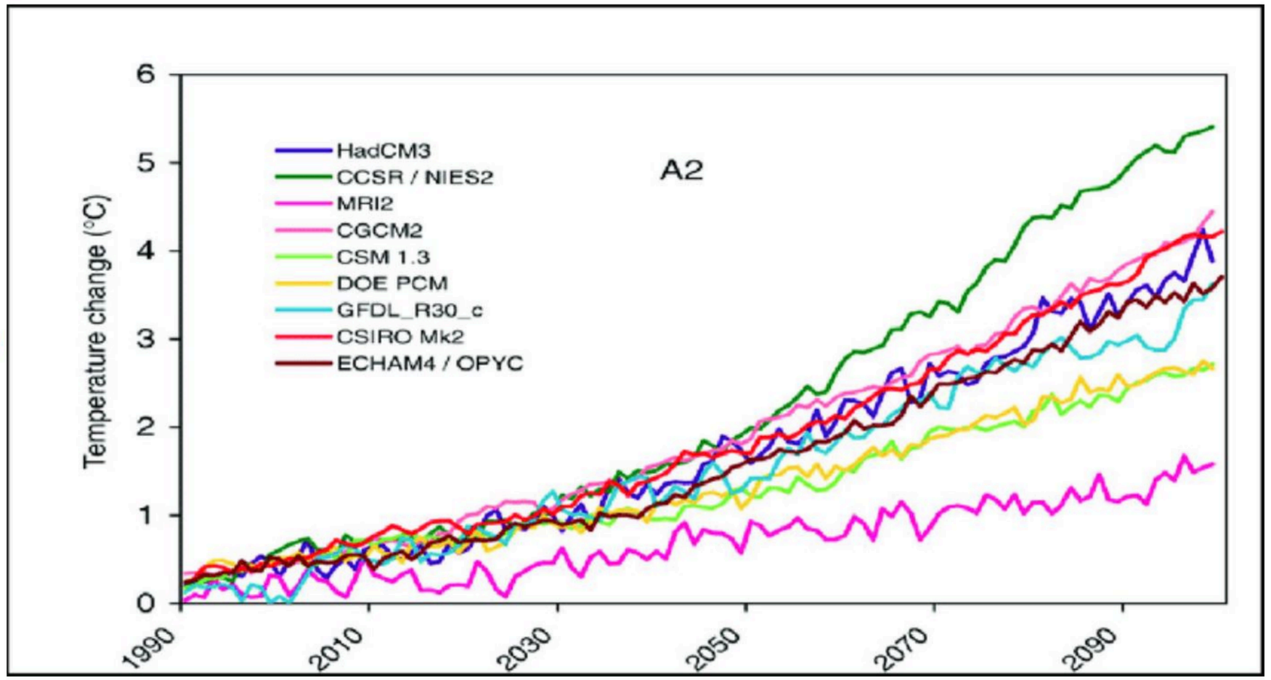


Every data-driven result is a result of the decisions made at *every* stage of the DSLC

- What if a clinician had interpreted a patient's answer differently?
- What if the dataset had included a different set of people?
- What if the data had been cleaned differently or by a different person
- What if a different model had been used to generate a prediction?

There are multiple plausible versions of almost every data-driven result

Climate scientists have generated a range of different projections of mean global temperature change



The change in global-mean temperature estimated by nine climate models forced by the SRES A2 emission scenario. (Source: IPCC TAR, Chapter 9)

The PCS Framework

Under the PCS framework, data-driven results should be:

1. **Predictable**: captures realistic phenomena and re-emerges in new/future data
2. **Computable**: computationally efficient and accessible to practitioners
3. **Stable**: remain stable to reasonable perturbations throughout the DSLC

Note: classical statistics only considers stability in the context of random sampling and random noise

Evaluating predictability of results:

A food clustering case study

Clustering foods by nutritional content

Use data from the **Food and Nutrient Database for Dietary Studies (FNDDS)** to cluster foods by nutritional content

Nutrient values for foods and beverages reported in “What We Eat in America”, the dietary intake component of the *National Health and Nutrition Examination Survey*.

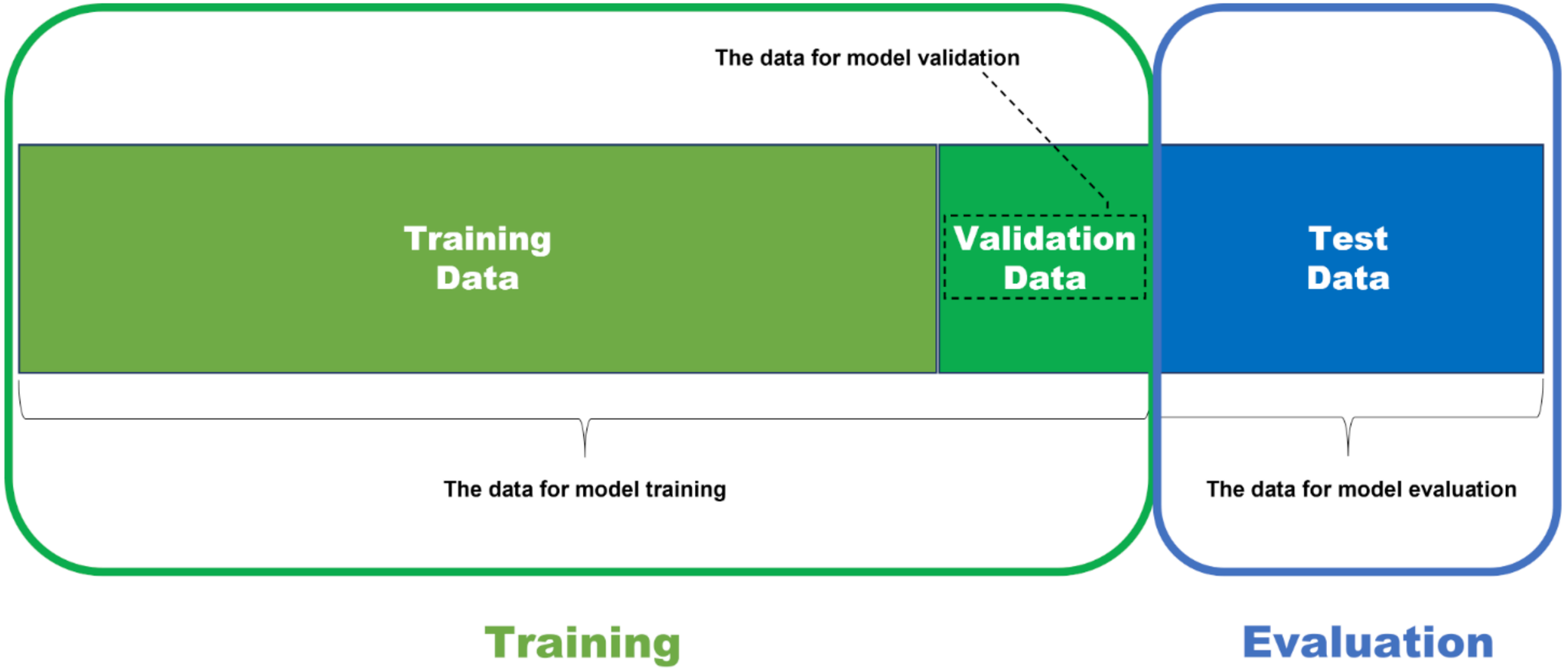
Sample foods* from the first 10 (of 30) clusters

1. Meats	2. Vegetables	3. Fish	4. Unclear	5. Alcohol
Beef	Spaghetti	Tuna	Chow	Beer
Chicken	Mixed	Mackerel	Shrimp	Gin
Bologna	Green	Herring	Fish	Beer
Chicken	Beef	Herring	Shrimp	Vodka
Meat	Macaroni	Sardines	Soupy	Whiskey
Pork	Green	Mackerel	Strudel	Long

6. Fats/oils	7. Nutri-mix	8. Infant formula	9. Vegetables	10. Cereal
Margarine	Whole	Infant	Beans	Cereal
Mayonnaise	Oatmeal	Infant	Turnip	Cereal
Coleslaw	Whey	Oatmeal	Mustard	Cereal
Safflower	Cereal	Infant	Beans	Cereal
Mayonnaise	Textured	Infant	Broccoli	Cereal
Sandwich	Gerber	Cream	Beans	Yeast

*We’re just showing the first word of each food item description

Use validation/test set and replicate results



You can do this to evaluate general results, not just for ML!

Image from https://support.cognex.com/docs/deep-learning_321/web/EN/deep-learning/Content/deep-learning-Topics/training-green-hd/prepare-train-valid-test.htm

Replicating results for *representative* external data

USDA SR Legacy

External data containing nutrient information for food items!

	Food and Nutrient Database for Dietary Studies (FNDDS)	**SR Legacy**
Definition	Data on nutrients and portion weights for foods and beverages reported in What We Eat in America, NHANES	Historic data on food components including nutrients derived from analyses, calculations, and published literature
Data Source	**USDA:** compiled based on values from FDC data types	**USDA:** based on Standard Reference originally available via the USDA National Nutrient Database (NNDB)

Your "**external data**" used for evaluation should maximally reflect the "**future data**" that you will be applying your results to in the **real world**

Replicating results for *representative* external data

Predictability evaluation

Use cluster centers from FNDDS to cluster SR Legacy food items

1. Meats	2. Vegetables	3. Fish	4. Unclear	5. Alcohol
Nuts	Soup	Fish	Potsticker	Alcoholic
Chicken	Cornmeal	Fish	Eggs	Alcoholic
Pork	Tomatoes	Fish	Turnover	Alcoholic
Pork	Soup	Fish	Fast	Alcoholic
Lamb	Babyfood	Fish	Muffin	Alcoholic
Chicken	Bananas	Fish	HOT	Alcoholic

6. Fats/oils	7. Nutri-mix	8. Infant formula	9. Vegetables	10. Cereal
Oil	Babyfood	Infant	Turnip	Cereals
Margarine	Beverages	Infant	Parsley	Formulated
Margarine	Beverages	Infant	Brussels	Cereals
Oil	Protein	Headcheese	Cress	Cereals
Oil	Cereals	Yogurt	Kale	Leavening
Oil	Beverages	Infant	Spinach	Yeast

Evaluating the stability of results to data cleaning/preprocessing judgment calls

Predicting surgical site infections

Can we identify patients who are at high risk of getting a surgical-site infection (SSI) after a surgery

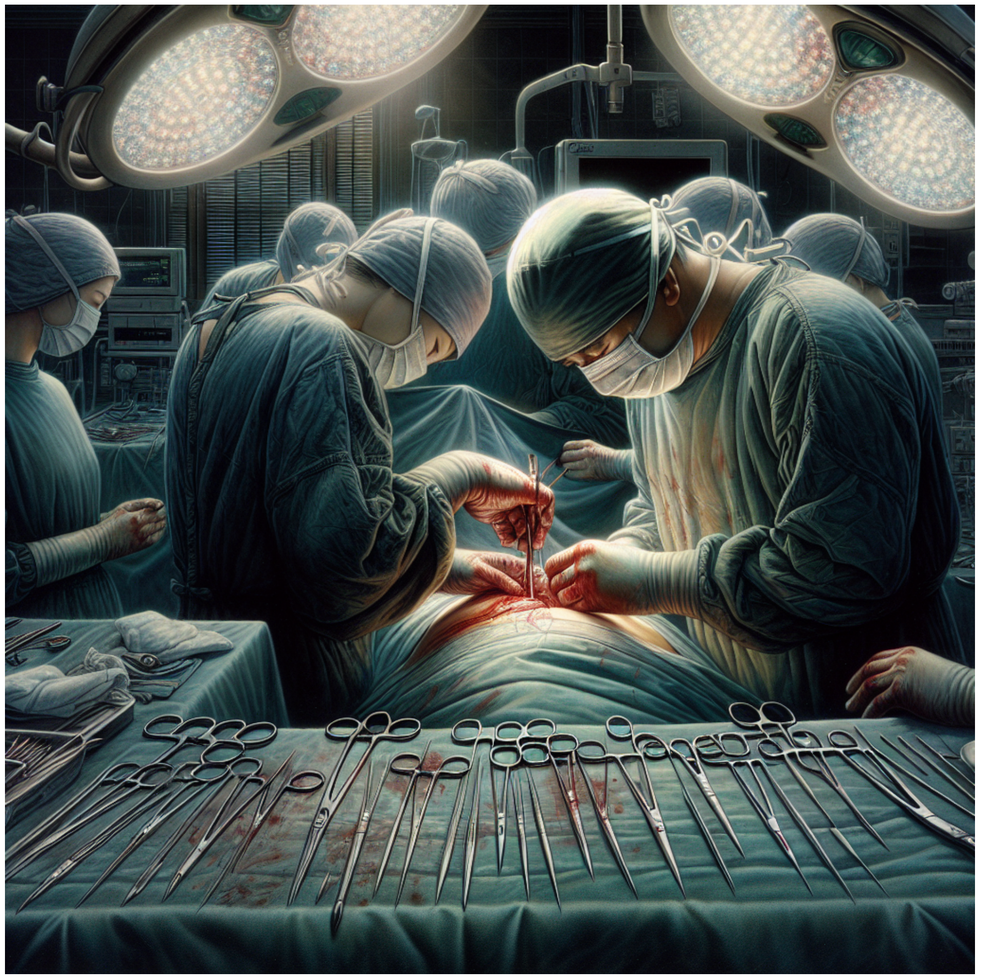
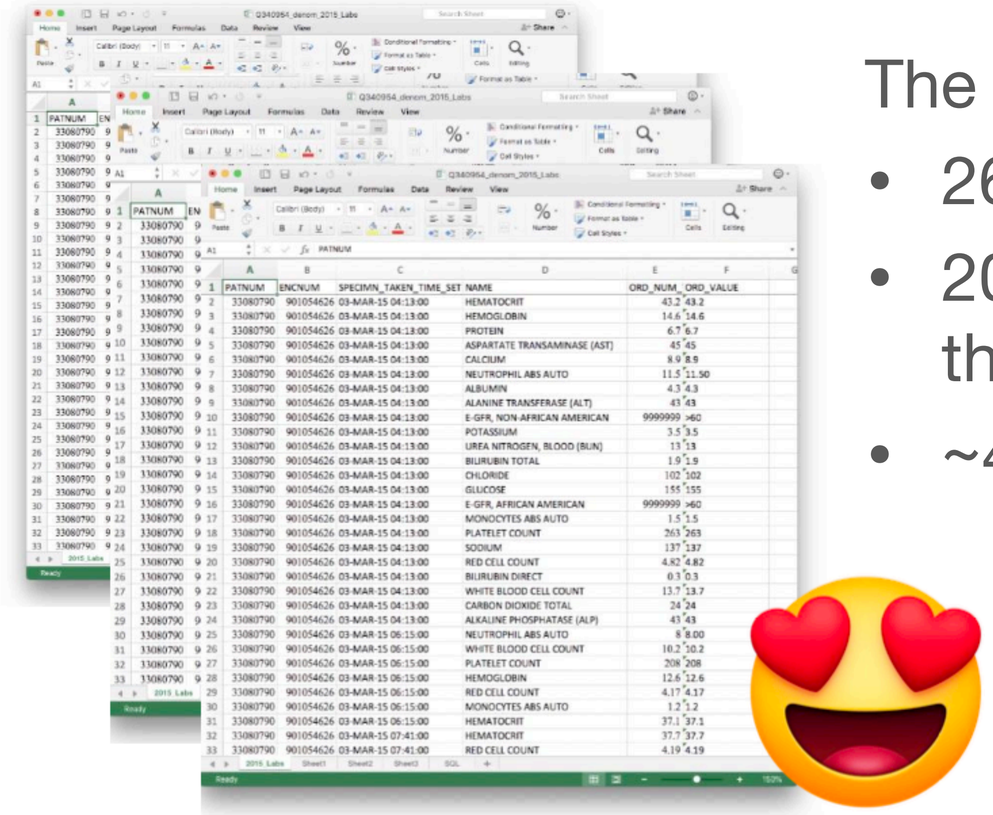


Image generated by Dall-e-3

The data is large and messy



The data consisted of:

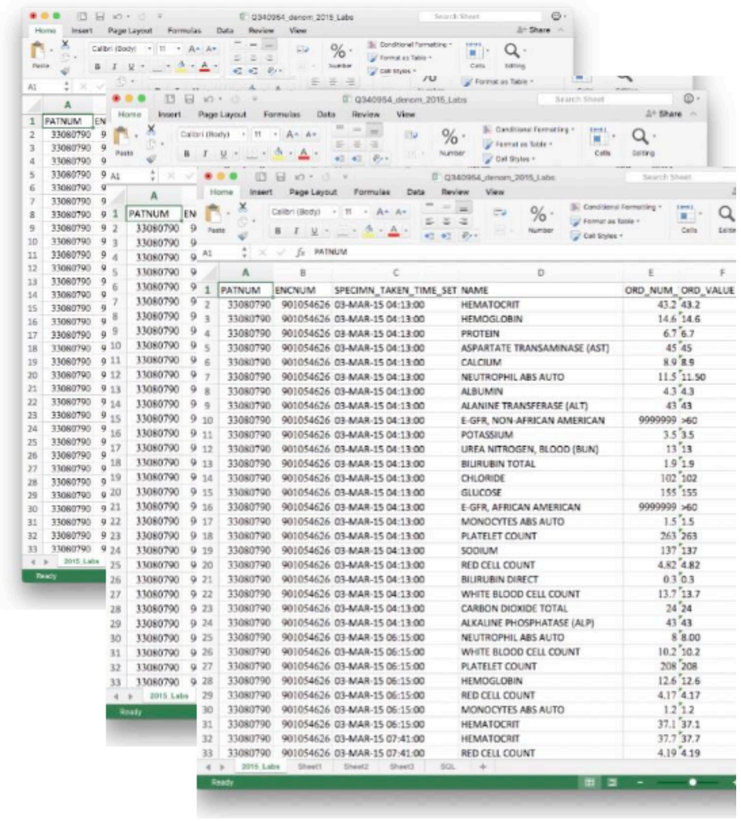
- 26 multi-sheet Excel files
- 200+ variables with inconsistent names across the files
- ~40,000 surgeries (~800 SSI cases, 2%)

There are A LOT of different ways to potentially clean/prepare this data for analysis!

Judgment calls we could have made:

- Separate variables for each different type of **antibiotic**, or a single summary antibiotic variable?
- Ever had an **autoimmune disease diagnosis** or just in the past 2 years?
- Max **temp** measurement in the first 24 hours or 48 hours? What about outpatient procedures?

There is an endless spectrum of potential “clean” datasets that we could use for our analysis



Judgment call combination 1

	A	B	C	D	E	F
1	PATNUM	ENCNUM	SPECIMN_TAKEN_TIME_SET	NAME	ORD_NUM	ORD_VALUE
2	33080790	901054626	03-MAR-15 04:13:00	HEMATOCRIT	43.2	43.2
3	33080790	901054626	03-MAR-15 04:13:00	HEMOGLOBIN	14.6	14.6
4	33080790	901054626	03-MAR-15 04:13:00	PROTEIN	6.7	6.7
5	33080790	901054626	03-MAR-15 04:13:00	ASPARTATE TRANSAMINASE (AST)	45	45
6	33080790	901054626	03-MAR-15 04:13:00	CALCIUM	8.9	8.9
7	33080790	901054626	03-MAR-15 04:13:00	NEUTROPHIL ABS AUTO	11.5	11.50
8	33080790	901054626	03-MAR-15 04:13:00	ALBUMIN	4.3	4.3
9	33080790	901054626	03-MAR-15 04:13:00	ALANINE TRANSFERASE (ALT)	43	43
10	33080790	901054626	03-MAR-15 04:13:00	E-GFR, NON-AFRICAN AMERICAN	9999999	>60
11	33080790	901054626	03-MAR-15 04:13:00	POTASSIUM	3.5	3.5
12	33080790	901054626	03-MAR-15 04:13:00	UREA NITROGEN, BLOOD (BUN)	13	13
13	33080790	901054626	03-MAR-15 04:13:00	BILIRUBIN TOTAL	1.9	1.9
14	33080790	901054626	03-MAR-15 04:13:00	CHLORIDE	102	102
15	33080790	901054626	03-MAR-15 04:13:00	GLUCOSE	155	155
16	33080790	901054626	03-MAR-15 04:13:00	E-GFR, AFRICAN AMERICAN	9999999	>60
17	33080790	901054626	03-MAR-15 04:13:00	MONOCYTES ABS AUTO	1.5	1.5
18	33080790	901054626	03-MAR-15 04:13:00	PLATELET COUNT	263	263
19	33080790	901054626	03-MAR-15 04:13:00	SODIUM	137	137
20	33080790	901054626	03-MAR-15 04:13:00	RED CELL COUNT	4.82	4.82
21	33080790	901054626	03-MAR-15 04:13:00	BILIRUBIN DIRECT	0.3	0.3
22	33080790	901054626	03-MAR-15 04:13:00	WHITE BLOOD CELL COUNT	13.7	13.7
23	33080790	901054626	03-MAR-15 04:13:00	CARBON DIOXIDE TOTAL	24	24
24	33080790	901054626	03-MAR-15 04:13:00	ALKALINE PHOSPHATASE (ALP)	43	43
25	33080790	901054626	03-MAR-15 06:15:00	NEUTROPHIL ABS AUTO	8	8.00
26	33080790	901054626	03-MAR-15 06:15:00	WHITE BLOOD CELL COUNT	10.2	10.2
27	33080790	901054626	03-MAR-15 06:15:00	PLATELET COUNT	208	208
28	33080790	901054626	03-MAR-15 06:15:00	HEMOGLOBIN	12.6	12.6
29	33080790	901054626	03-MAR-15 06:15:00	RED CELL COUNT	4.17	4.17
30	33080790	901054626	03-MAR-15 06:15:00	MONOCYTES ABS AUTO	1.2	1.2
31	33080790	901054626	03-MAR-15 06:15:00	HEMATOCRIT	37.1	37.1
32	33080790	901054626	03-MAR-15 07:41:00	HEMATOCRIT	37.7	37.7
33	33080790	901054626	03-MAR-15 07:41:00	RED CELL COUNT	4.19	4.19

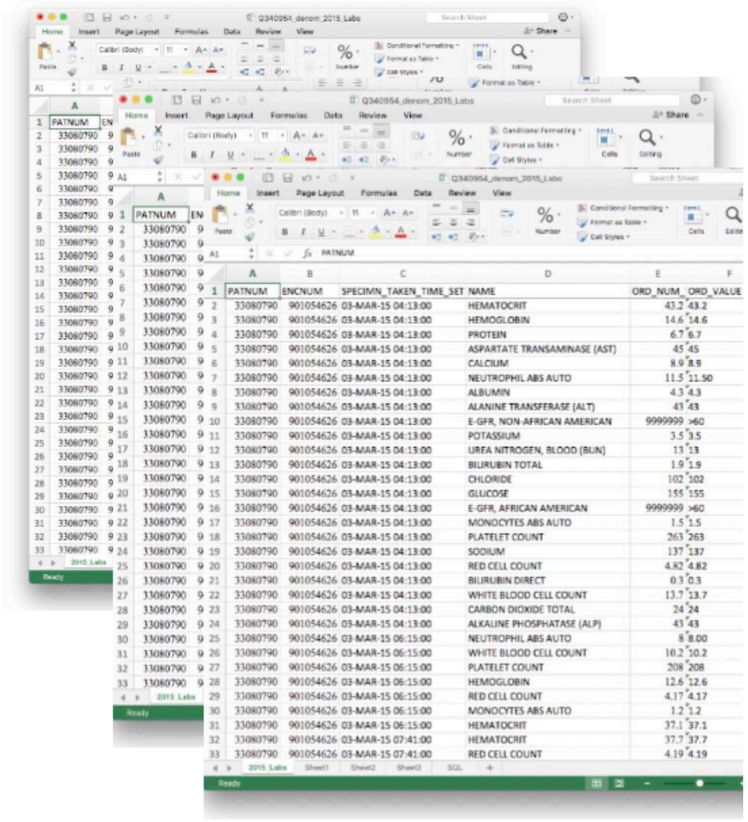
Judgment call combination 100

	A	B	C	D	E	F
1	PATNUM	ENCNUM	SPECIMN_TAKEN_TIME_SET	NAME	ORD_NUM	ORD_VALUE
2	33080790	901054626	03-MAR-15 04:13:00	HEMATOCRIT	43.2	43.2
3	33080790	901054626	03-MAR-15 04:13:00	HEMOGLOBIN	14.6	14.6
4	33080790	901054626	03-MAR-15 04:13:00	PROTEIN	6.7	6.7
5	33080790	901054626	03-MAR-15 04:13:00	ASPARTATE TRANSAMINASE (AST)	45	45
6	33080790	901054626	03-MAR-15 04:13:00	CALCIUM	8.9	8.9
7	33080790	901054626	03-MAR-15 04:13:00	NEUTROPHIL ABS AUTO	11.5	11.50
8	33080790	901054626	03-MAR-15 04:13:00	ALBUMIN	4.3	4.3
9	33080790	901054626	03-MAR-15 04:13:00	ALANINE TRANSFERASE (ALT)	43	43
10	33080790	901054626	03-MAR-15 04:13:00	E-GFR, NON-AFRICAN AMERICAN	9999999	>60
11	33080790	901054626	03-MAR-15 04:13:00	POTASSIUM	3.5	3.5
12	33080790	901054626	03-MAR-15 04:13:00	UREA NITROGEN, BLOOD (BUN)	13	13
13	33080790	901054626	03-MAR-15 04:13:00	BILIRUBIN TOTAL	1.9	1.9
14	33080790	901054626	03-MAR-15 04:13:00	CHLORIDE	102	102
15	33080790	901054626	03-MAR-15 04:13:00	GLUCOSE	155	155
16	33080790	901054626	03-MAR-15 04:13:00	E-GFR, AFRICAN AMERICAN	9999999	>60
17	33080790	901054626	03-MAR-15 04:13:00	MONOCYTES ABS AUTO	1.5	1.5
18	33080790	901054626	03-MAR-15 04:13:00	PLATELET COUNT	263	263
19	33080790	901054626	03-MAR-15 04:13:00	SODIUM	137	137
20	33080790	901054626	03-MAR-15 04:13:00	RED CELL COUNT	4.82	4.82
21	33080790	901054626	03-MAR-15 04:13:00	BILIRUBIN DIRECT	0.3	0.3
22	33080790	901054626	03-MAR-15 04:13:00	WHITE BLOOD CELL COUNT	13.7	13.7
23	33080790	901054626	03-MAR-15 04:13:00	CARBON DIOXIDE TOTAL	24	24
24	33080790	901054626	03-MAR-15 04:13:00	ALKALINE PHOSPHATASE (ALP)	43	43
25	33080790	901054626	03-MAR-15 06:15:00	NEUTROPHIL ABS AUTO	8	8.00
26	33080790	901054626	03-MAR-15 06:15:00	WHITE BLOOD CELL COUNT	10.2	10.2
27	33080790	901054626	03-MAR-15 06:15:00	PLATELET COUNT	208	208
28	33080790	901054626	03-MAR-15 06:15:00	HEMOGLOBIN	12.6	12.6
29	33080790	901054626	03-MAR-15 06:15:00	RED CELL COUNT	4.17	4.17
30	33080790	901054626	03-MAR-15 06:15:00	MONOCYTES ABS AUTO	1.2	1.2
31	33080790	901054626	03-MAR-15 06:15:00	HEMATOCRIT	37.1	37.1
32	33080790	901054626	03-MAR-15 07:41:00	HEMATOCRIT	37.7	37.7
33	33080790	901054626	03-MAR-15 07:41:00	RED CELL COUNT	4.19	4.19

Judgment call combination 500

	A	B	C	D	E	F
1	PATNUM	ENCNUM	SPECIMN_TAKEN_TIME_SET	NAME	ORD_NUM	ORD_VALUE
2	33080790	901054626	03-MAR-15 04:13:00	HEMATOCRIT	43.2	43.2
3	33080790	901054626	03-MAR-15 04:13:00	HEMOGLOBIN	14.6	14.6
4	33080790	901054626	03-MAR-15 04:13:00	PROTEIN	6.7	6.7
5	33080790	901054626	03-MAR-15 04:13:00	ASPARTATE TRANSAMINASE (AST)	45	45
6	33080790	901054626	03-MAR-15 04:13:00	CALCIUM	8.9	8.9
7	33080790	901054626	03-MAR-15 04:13:00	NEUTROPHIL ABS AUTO	11.5	11.50
8	33080790	901054626	03-MAR-15 04:13:00	ALBUMIN	4.3	4.3
9	33080790	901054626	03-MAR-15 04:13:00	ALANINE TRANSFERASE (ALT)	43	43
10	33080790	901054626	03-MAR-15 04:13:00	E-GFR, NON-AFRICAN AMERICAN	9999999	>60
11	33080790	901054626	03-MAR-15 04:13:00	POTASSIUM	3.5	3.5
12	33080790	901054626	03-MAR-15 04:13:00	UREA NITROGEN, BLOOD (BUN)	13	13
13	33080790	901054626	03-MAR-15 04:13:00	BILIRUBIN TOTAL	1.9	1.9
14	33080790	901054626	03-MAR-15 04:13:00	CHLORIDE	102	102
15	33080790	901054626	03-MAR-15 04:13:00	GLUCOSE	155	155
16	33080790	901054626	03-MAR-15 04:13:00	E-GFR, AFRICAN AMERICAN	9999999	>60
17	33080790	901054626	03-MAR-15 04:13:00	MONOCYTES ABS AUTO	1.5	1.5
18	33080790	901054626	03-MAR-15 04:13:00	PLATELET COUNT	263	263
19	33080790	901054626	03-MAR-15 04:13:00	SODIUM	137	137
20	33080790	901054626	03-MAR-15 04:13:00	RED CELL COUNT	4.82	4.82
21	33080790	901054626	03-MAR-15 04:13:00	BILIRUBIN DIRECT	0.3	0.3
22	33080790	901054626	03-MAR-15 04:13:00	WHITE BLOOD CELL COUNT	13.7	13.7
23	33080790	901054626	03-MAR-15 04:13:00	CARBON DIOXIDE TOTAL	24	24
24	33080790	901054626	03-MAR-15 04:13:00	ALKALINE PHOSPHATASE (ALP)	43	43
25	33080790	901054626	03-MAR-15 06:15:00	NEUTROPHIL ABS AUTO	8	8.00
26	33080790	901054626	03-MAR-15 06:15:00	WHITE BLOOD CELL COUNT	10.2	10.2
27	33080790	901054626	03-MAR-15 06:15:00	PLATELET COUNT	208	208
28	33080790	901054626	03-MAR-15 06:15:00	HEMOGLOBIN	12.6	12.6
29	33080790	901054626	03-MAR-15 06:15:00	RED CELL COUNT	4.17	4.17
30	33080790	901054626	03-MAR-15 06:15:00	MONOCYTES ABS AUTO	1.2	1.2
31	33080790	901054626	03-MAR-15 06:15:00	HEMATOCRIT	37.1	37.1
32	33080790	901054626	03-MAR-15 07:41:00	HEMATOCRIT	37.7	37.7
33	33080790	901054626	03-MAR-15 07:41:00	RED CELL COUNT	4.19	4.19

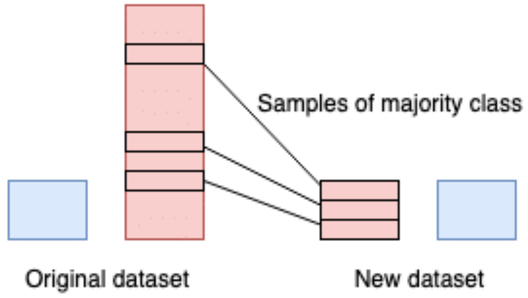
Traditionally… we just make one set of judgment calls



Judgment call combination 1

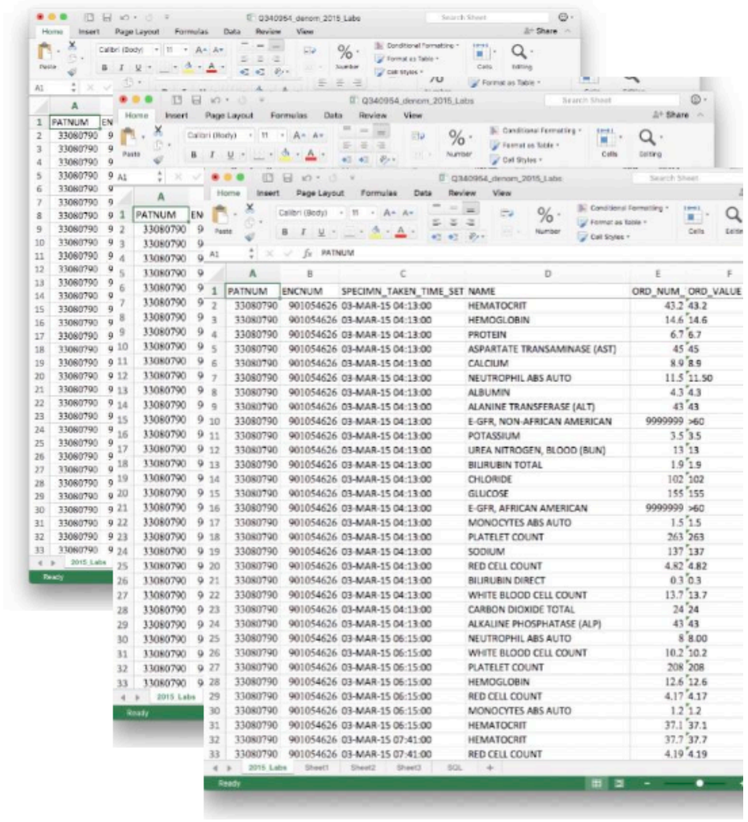
	PATNUM	ENCNUM	SPECIMN_TAKEN_TIME_SET	NAME	ORD_NUM	ORD_VALUE
2	33080790	901054626	03-MAR-15 04:13:00	HEMATOCRIT	43.2	43.2
3	33080790	901054626	03-MAR-15 04:13:00	HEMOGLOBIN	14.6	14.6
4	33080790	901054626	03-MAR-15 04:13:00	PROTEIN	6.7	6.7
5	33080790	901054626	03-MAR-15 04:13:00	ASPARTATE TRANSAMINASE (AST)	45	45
6	33080790	901054626	03-MAR-15 04:13:00	CALCIUM	8.9	8.9
7	33080790	901054626	03-MAR-15 04:13:00	NEUTROPHIL ABS AUTO	11.5	11.50
8	33080790	901054626	03-MAR-15 04:13:00	ALBUMIN	4.3	4.3
9	33080790	901054626	03-MAR-15 04:13:00	ALANINE TRANSFERASE (ALT)	43	43
10	33080790	901054626	03-MAR-15 04:13:00	E-GFR, NON-AFRICAN AMERICAN	9999999	>60
11	33080790	901054626	03-MAR-15 04:13:00	POTASSIUM	3.5	3.5
12	33080790	901054626	03-MAR-15 04:13:00	UREA NITROGEN, BLOOD (BUN)	13	13
13	33080790	901054626	03-MAR-15 04:13:00	BILIRUBIN TOTAL	1.9	1.9
14	33080790	901054626	03-MAR-15 04:13:00	CHLORIDE	102	102
15	33080790	901054626	03-MAR-15 04:13:00	GLUCOSE	155	155
16	33080790	901054626	03-MAR-15 04:13:00	E-GFR, AFRICAN AMERICAN	9999999	>60
17	33080790	901054626	03-MAR-15 04:13:00	MONOCYTES ABS AUTO	1.5	1.5
18	33080790	901054626	03-MAR-15 04:13:00	PLATELET COUNT	263	263
19	33080790	901054626	03-MAR-15 04:13:00	SODIUM	137	137
20	33080790	901054626	03-MAR-15 04:13:00	RED CELL COUNT	4.82	4.82
21	33080790	901054626	03-MAR-15 04:13:00	BILIRUBIN DIRECT	0.3	0.3
22	33080790	901054626	03-MAR-15 04:13:00	WHITE BLOOD CELL COUNT	13.7	13.7
23	33080790	901054626	03-MAR-15 04:13:00	CARBON DIOXIDE TOTAL	24	24
24	33080790	901054626	03-MAR-15 04:13:00	ALKALINE PHOSPHATASE (ALP)	43	43
25	33080790	901054626	03-MAR-15 06:15:00	NEUTROPHIL ABS AUTO	8	8.00
26	33080790	901054626	03-MAR-15 06:15:00	WHITE BLOOD CELL COUNT	10.2	10.2
27	33080790	901054626	03-MAR-15 06:15:00	PLATELET COUNT	208	208
28	33080790	901054626	03-MAR-15 06:15:00	HEMOGLOBIN	12.6	12.6
29	33080790	901054626	03-MAR-15 06:15:00	RED CELL COUNT	4.17	4.17
30	33080790	901054626	03-MAR-15 06:15:00	MONOCYTES ABS AUTO	1.2	1.2
31	33080790	901054626	03-MAR-15 06:15:00	HEMATOCRIT	37.1	37.1
32	33080790	901054626	03-MAR-15 07:41:00	HEMATOCRIT	37.7	37.7
33	33080790	901054626	03-MAR-15 07:41:00	RED CELL COUNT	4.19	4.19

Downsample to deal with class imbalance



“Women over the age of 71 whose surgery lasted more than 6 hours and whose temperature reaches 100F in the 24 hours post-surgery are at highest risk of infection”

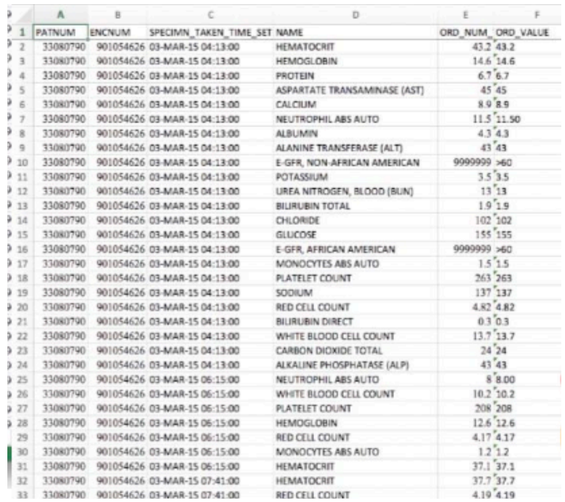
But there are so many others that we could have made



PATNUM	ENCNUM	SPECIMN_TAKEN_TIME_SET	NAME	ORD_NUM	ORD_VALUE
33080790	901054626	03-MAR-15 04:13:00	HEMATOCRIT	43.2	43.2
33080790	901054626	03-MAR-15 04:13:00	HEMOGLOBIN	14.6	14.6
33080790	901054626	03-MAR-15 04:13:00	PROTEIN	6.7	6.7
33080790	901054626	03-MAR-15 04:13:00	ASPARTATE TRANSAMINASE (AST)	45	45
33080790	901054626	03-MAR-15 04:13:00	CALCIUM	8.9	8.9
33080790	901054626	03-MAR-15 04:13:00	NEUTROPHIL ABS AUTO	11.5	11.50
33080790	901054626	03-MAR-15 04:13:00	ALBUMIN	4.3	4.3
33080790	901054626	03-MAR-15 04:13:00	ALANINE TRANSFERASE (ALT)	43	43
33080790	901054626	03-MAR-15 04:13:00	E-GFR, NON-AFRICAN AMERICAN	9999999	>60
33080790	901054626	03-MAR-15 04:13:00	POTASSIUM	3.5	3.5
33080790	901054626	03-MAR-15 04:13:00	UREA NITROGEN, BLOOD (BUN)	13	13
33080790	901054626	03-MAR-15 04:13:00	BILIRUBIN TOTAL	1.9	1.9
33080790	901054626	03-MAR-15 04:13:00	CHLORIDE	102	102
33080790	901054626	03-MAR-15 04:13:00	GLUCOSE	155	155
33080790	901054626	03-MAR-15 04:13:00	E-GFR, AFRICAN AMERICAN	9999999	>60
33080790	901054626	03-MAR-15 04:13:00	MONOCYTES ABS AUTO	1.5	1.5
33080790	901054626	03-MAR-15 04:13:00	PLATELET COUNT	263	263
33080790	901054626	03-MAR-15 04:13:00	SODIUM	137	137
33080790	901054626	03-MAR-15 04:13:00	RED CELL COUNT	4.82	4.82
33080790	901054626	03-MAR-15 04:13:00	BILIRUBIN DIRECT	0.3	0.3
33080790	901054626	03-MAR-15 04:13:00	WHITE BLOOD CELL COUNT	13.7	13.7
33080790	901054626	03-MAR-15 04:13:00	CARBON DIOXIDE TOTAL	24	24
33080790	901054626	03-MAR-15 04:13:00	ALKALINE PHOSPHATASE (ALP)	43	43
33080790	901054626	03-MAR-15 06:15:00	NEUTROPHIL ABS AUTO	8	8.00
33080790	901054626	03-MAR-15 06:15:00	WHITE BLOOD CELL COUNT	10.2	10.2
33080790	901054626	03-MAR-15 06:15:00	PLATELET COUNT	208	208
33080790	901054626	03-MAR-15 06:15:00	HEMOGLOBIN	12.6	12.6
33080790	901054626	03-MAR-15 06:15:00	RED CELL COUNT	4.17	4.17
33080790	901054626	03-MAR-15 06:15:00	MONOCYTES ABS AUTO	1.2	1.2
33080790	901054626	03-MAR-15 07:41:00	HEMATOCRIT	37.1	37.1
33080790	901054626	03-MAR-15 07:41:00	HEMATOCRIT	37.7	37.7
33080790	901054626	03-MAR-15 07:41:00	RED CELL COUNT	4.19	4.19

Judgment call combination 500

Upsample to deal with class imbalance



PATNUM	ENCNUM	SPECIMN_TAKEN_TIME_SET	NAME	ORD_NUM	ORD_VALUE
33080790	901054626	03-MAR-15 04:13:00	HEMATOCRIT	43.2	43.2
33080790	901054626	03-MAR-15 04:13:00	HEMOGLOBIN	14.6	14.6
33080790	901054626	03-MAR-15 04:13:00	PROTEIN	6.7	6.7
33080790	901054626	03-MAR-15 04:13:00	ASPARTATE TRANSAMINASE (AST)	45	45
33080790	901054626	03-MAR-15 04:13:00	CALCIUM	8.9	8.9
33080790	901054626	03-MAR-15 04:13:00	NEUTROPHIL ABS AUTO	11.5	11.50
33080790	901054626	03-MAR-15 04:13:00	ALBUMIN	4.3	4.3
33080790	901054626	03-MAR-15 04:13:00	ALANINE TRANSFERASE (ALT)	43	43
33080790	901054626	03-MAR-15 04:13:00	E-GFR, NON-AFRICAN AMERICAN	9999999	>60
33080790	901054626	03-MAR-15 04:13:00	POTASSIUM	3.5	3.5
33080790	901054626	03-MAR-15 04:13:00	UREA NITROGEN, BLOOD (BUN)	13	13
33080790	901054626	03-MAR-15 04:13:00	BILIRUBIN TOTAL	1.9	1.9
33080790	901054626	03-MAR-15 04:13:00	CHLORIDE	102	102
33080790	901054626	03-MAR-15 04:13:00	GLUCOSE	155	155
33080790	901054626	03-MAR-15 04:13:00	E-GFR, AFRICAN AMERICAN	9999999	>60
33080790	901054626	03-MAR-15 04:13:00	MONOCYTES ABS AUTO	1.5	1.5
33080790	901054626	03-MAR-15 04:13:00	PLATELET COUNT	263	263
33080790	901054626	03-MAR-15 04:13:00	SODIUM	137	137
33080790	901054626	03-MAR-15 04:13:00	RED CELL COUNT	4.82	4.82
33080790	901054626	03-MAR-15 04:13:00	BILIRUBIN DIRECT	0.3	0.3
33080790	901054626	03-MAR-15 04:13:00	WHITE BLOOD CELL COUNT	13.7	13.7
33080790	901054626	03-MAR-15 04:13:00	CARBON DIOXIDE TOTAL	24	24
33080790	901054626	03-MAR-15 04:13:00	ALKALINE PHOSPHATASE (ALP)	43	43
33080790	901054626	03-MAR-15 06:15:00	NEUTROPHIL ABS AUTO	8	8.00
33080790	901054626	03-MAR-15 06:15:00	WHITE BLOOD CELL COUNT	10.2	10.2
33080790	901054626	03-MAR-15 06:15:00	PLATELET COUNT	208	208
33080790	901054626	03-MAR-15 06:15:00	HEMOGLOBIN	12.6	12.6
33080790	901054626	03-MAR-15 06:15:00	RED CELL COUNT	4.17	4.17
33080790	901054626	03-MAR-15 06:15:00	MONOCYTES ABS AUTO	1.2	1.2
33080790	901054626	03-MAR-15 06:15:00	HEMATOCRIT	37.1	37.1
33080790	901054626	03-MAR-15 07:41:00	HEMATOCRIT	37.7	37.7
33080790	901054626	03-MAR-15 07:41:00	RED CELL COUNT	4.19	4.19

“Cancer patients over the age of 60 undergoing abdominal surgery lasting more than 2 hours have the highest risk of SSI”

Build your judgment calls into your data preparation process!

```r
prepareData <- function(raw_data,
                        max_temp_hours = 24,
                        antibiotic_var_type = c("aggregate", "separate"),
                        balance_method = c("downsample", "upsample", "smote")) {

  if (antibiotic_var_type == "aggregate") {
    # code to create a single aggregated antibiotic variable
    ...
  } else if (antibiotic_var_type == "separate") {
    # code to create a separate variable for each type of antibiotic
    ...
  }

  if (balance_method == "downsample") {
    # code to create a downsampled version of the data
    ...
  } else if (balance_method == "upsample") {
    # code to create a upsampled version of the data
    ...
  }

  return(processed_data)
}
```

```r
data_clean <- prepareData(
  data_orig,
  antibiotic_var_type = "aggregate",
  max_temp_hours = 24,
  balance_method = "smote"
)
```

Build your judgment calls into your data preparation process!

```r
prepareData <- function(raw_data,
                        max_temp_hours = 24,
                        antibiotic_var_type = c("aggregate", "separate"),
                        balance_method = c("downsample", "upsample", "smote")) {

  if (antibiotic_var_type == "aggregate") {
    # code to create a single aggregated antibiotic variable
    ...
  } else if (antibiotic_var_type == "separate") {
    # code to create a separate variable for each type of antibiotic
    ...
  }

  if (balance_method == "downsample") {
    # code to create a downsampled version of the data
    ...
  } else if (balance_method == "upsample") {
    # code to create a upsampled version of the data
    ...
  }

  return(processed_data)
}
```

```r
data_clean <- prepareData(
  data_orig,
  antibiotic_var_type = "separate",
  max_temp_hours = 24,
  balance_method = "upsample"
)
```

Build your judgment calls into your data preparation process!

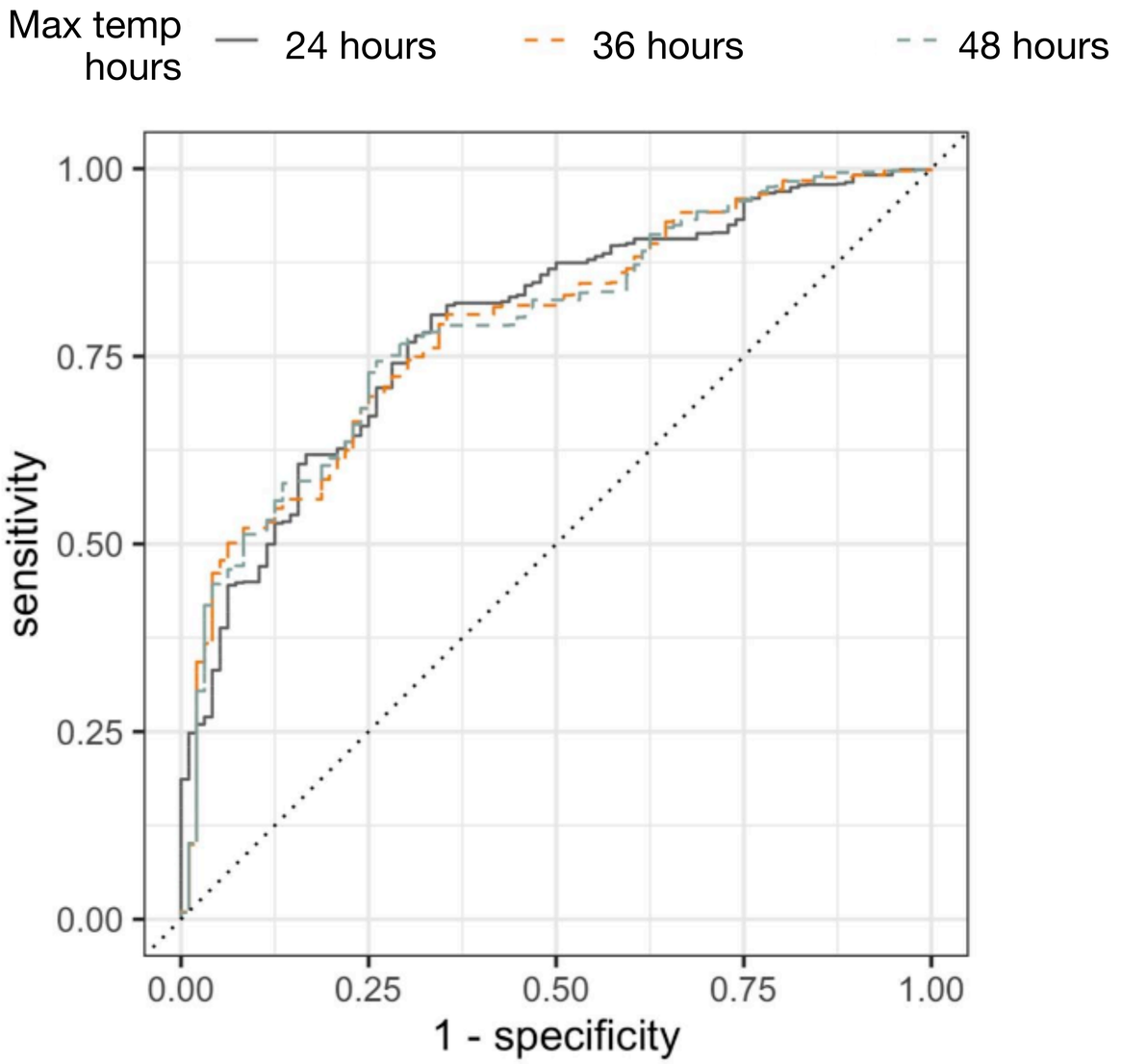
```r
prepareData <- function(raw_data,
                        max_temp_hours = 24,
                        antibiotic_var_type = c("aggregate", "separate"),
                        balance_method = c("downsample", "upsample", "smote")) {

  if (antibiotic_var_type == "aggregate") {
    # code to create a single aggregated antibiotic variable
    ...
  } else if (antibiotic_var_type == "separate") {
    # code to create a separate variable for each type of antibiotic
    ...
  }

  if (balance_method == "downsample") {
    # code to create a downsampled version of the data
    ...
  } else if (balance_method == "upsample") {
    # code to create a upsampled version of the data
    ...
  }

  return(processed_data)
}
```

```r
data_clean <- prepareData(
  data_orig,
  antibiotic_var_type = "separate",
  max_temp_hours = 48,
  balance_method = "downsample"
)
```

Then evaluate the stability of results to your judgment calls

Judgment call

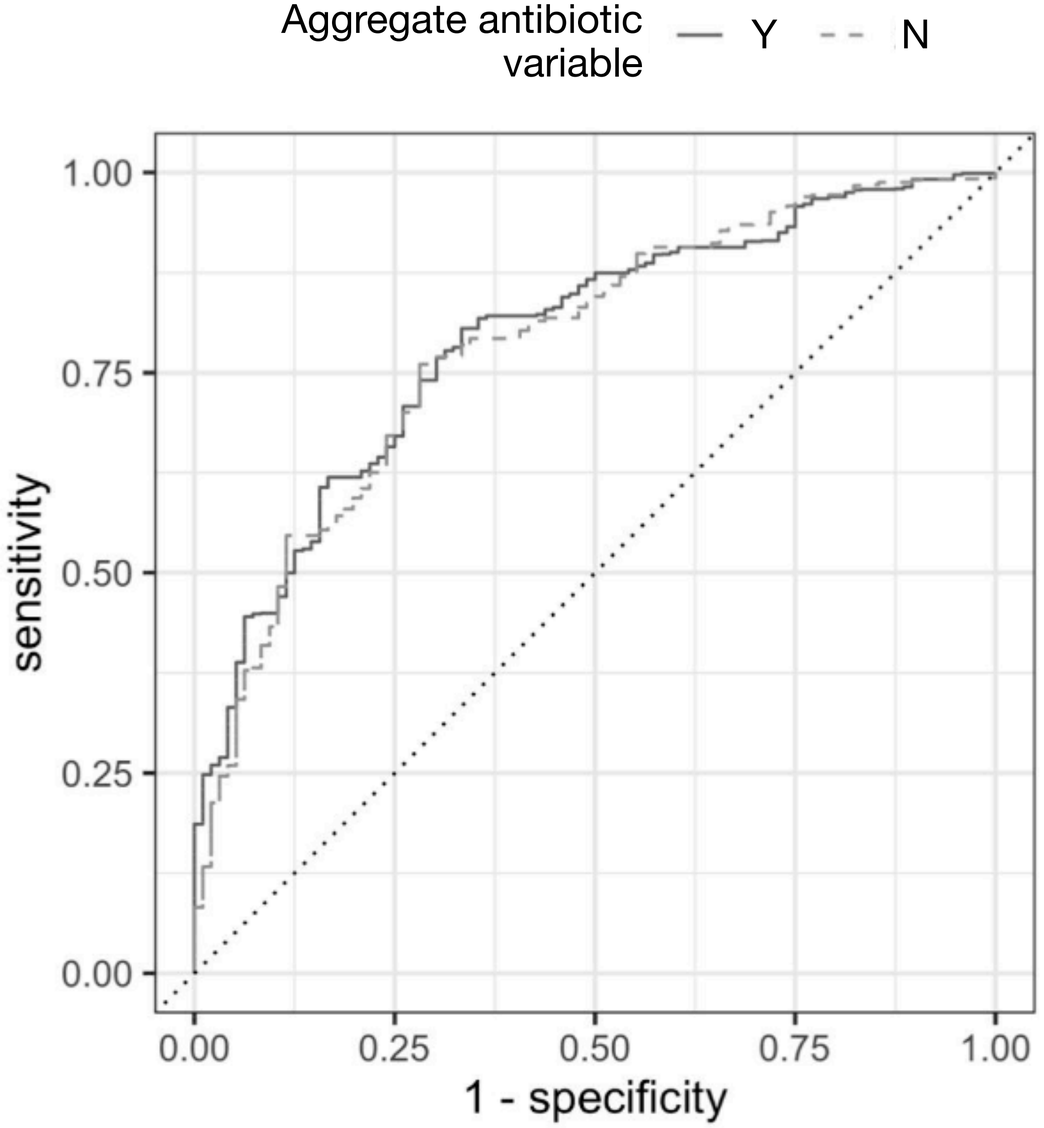
Maximum temperature in the first 24, 46, or 48 hours?



Then evaluate the stability of results to your judgment calls

Judgment call

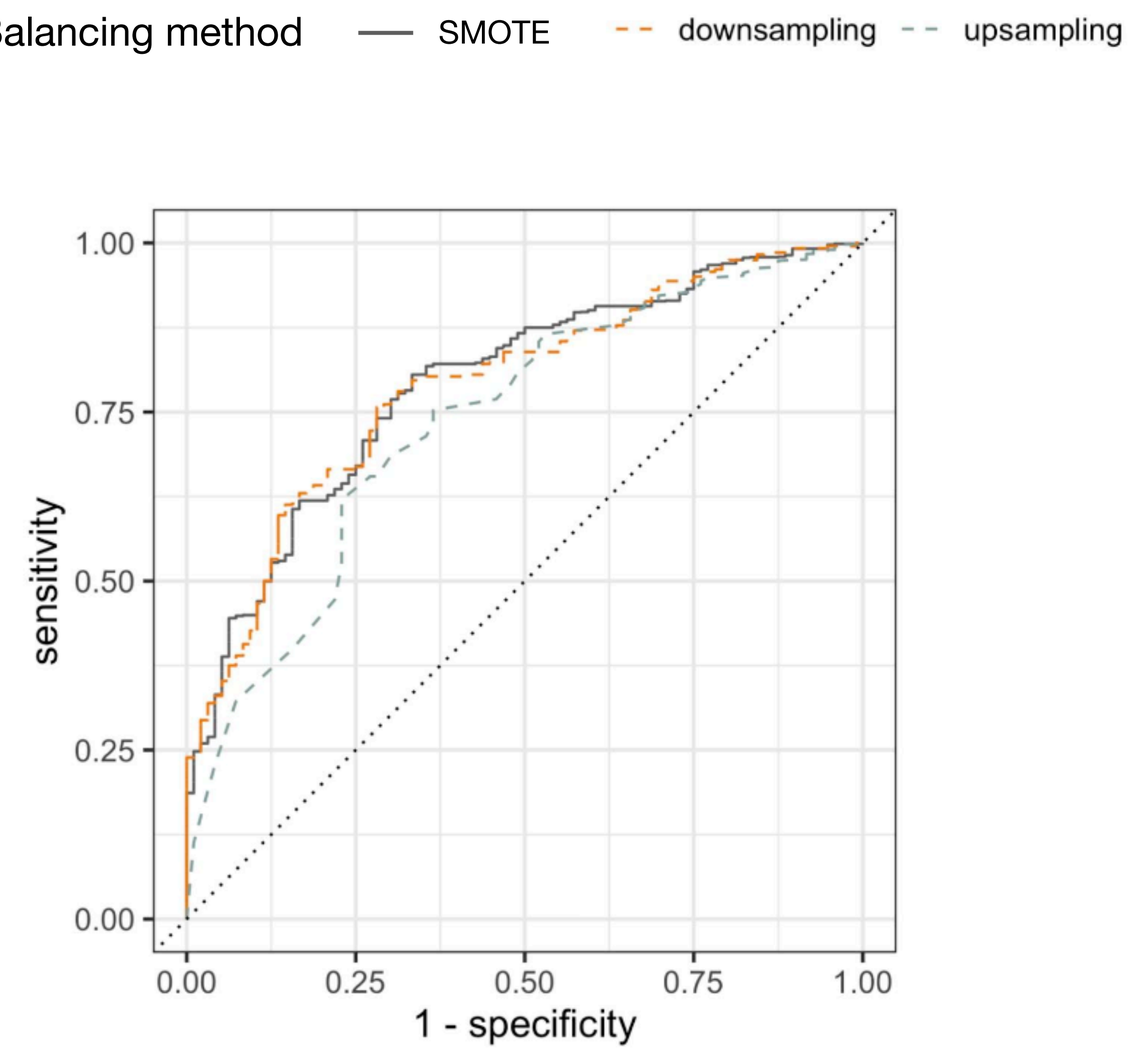
Create aggregated antibiotics variable or use separate variables for each antibiotic?



Then evaluate the stability of results to your judgment calls

Downsampling technique

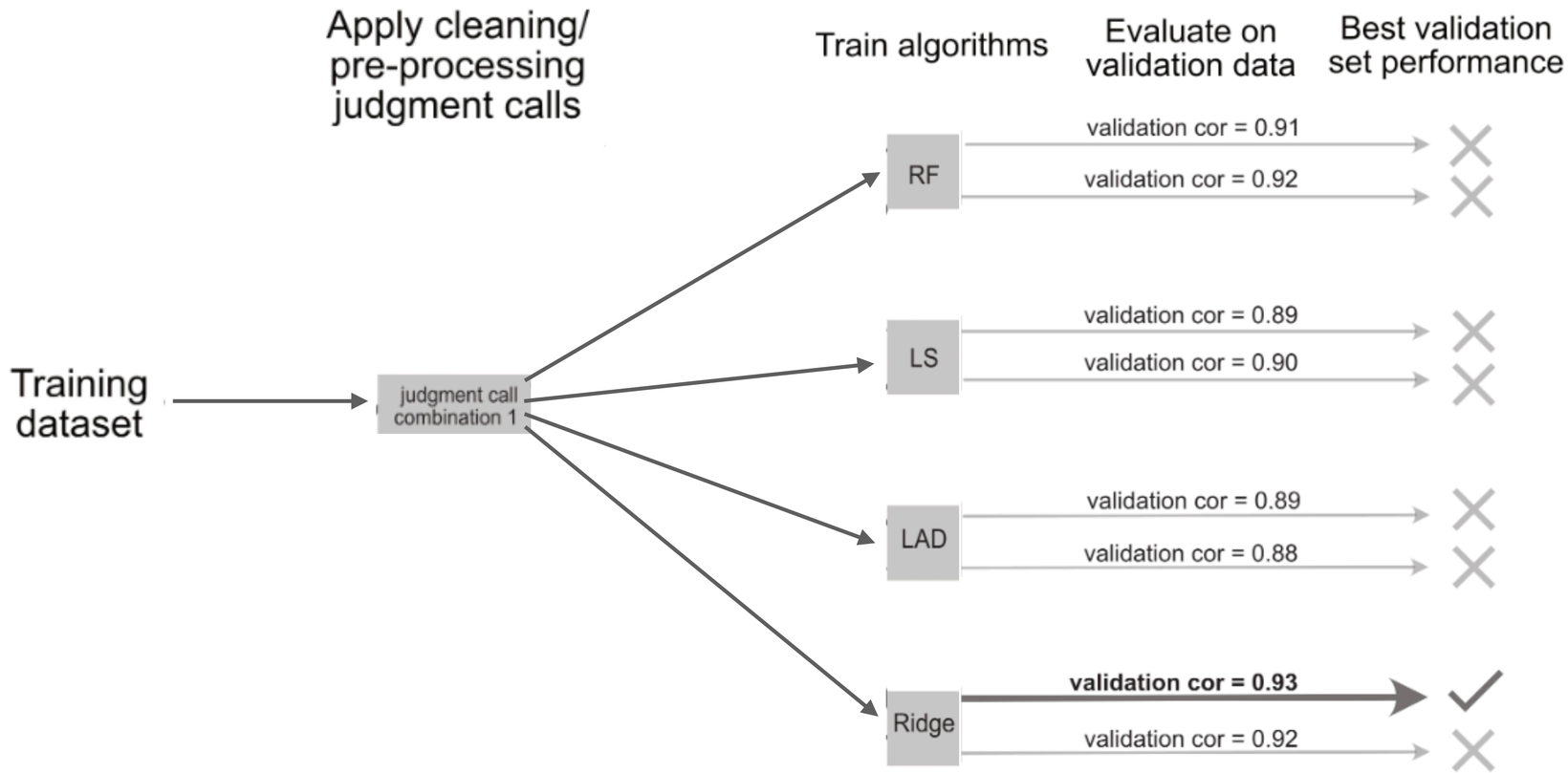
Use downsampling, upsampling, or SMOTE to balance data before fitting predictive algorithms?



Incorporating uncertainty into data-driven results (using prediction as a case-study)

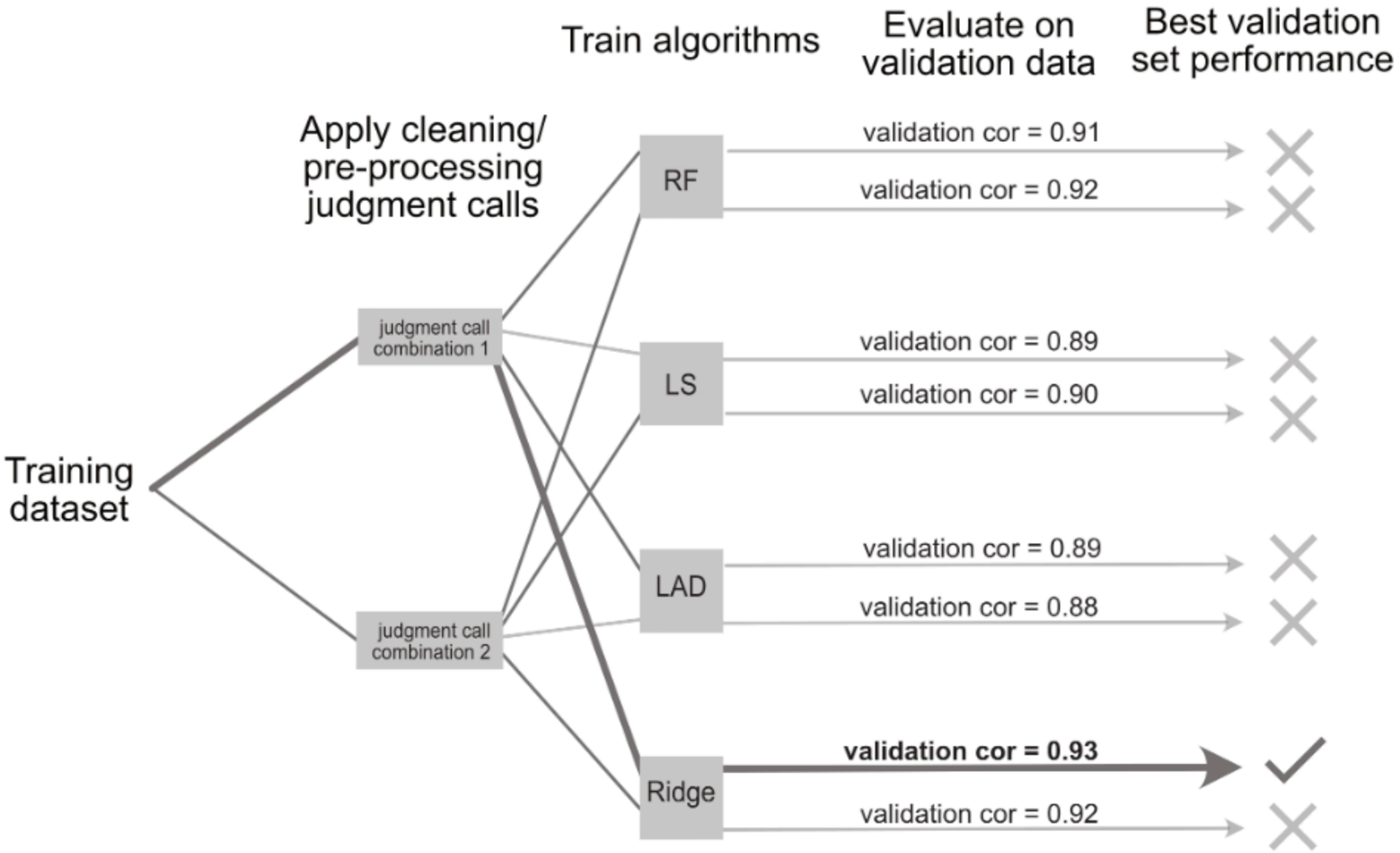
Traditional approach to choosing the predictive algorithm

Traditional approach: Train algorithm based on ***one* training dataset** and choose the **best algorithm** based on **validation set** performance



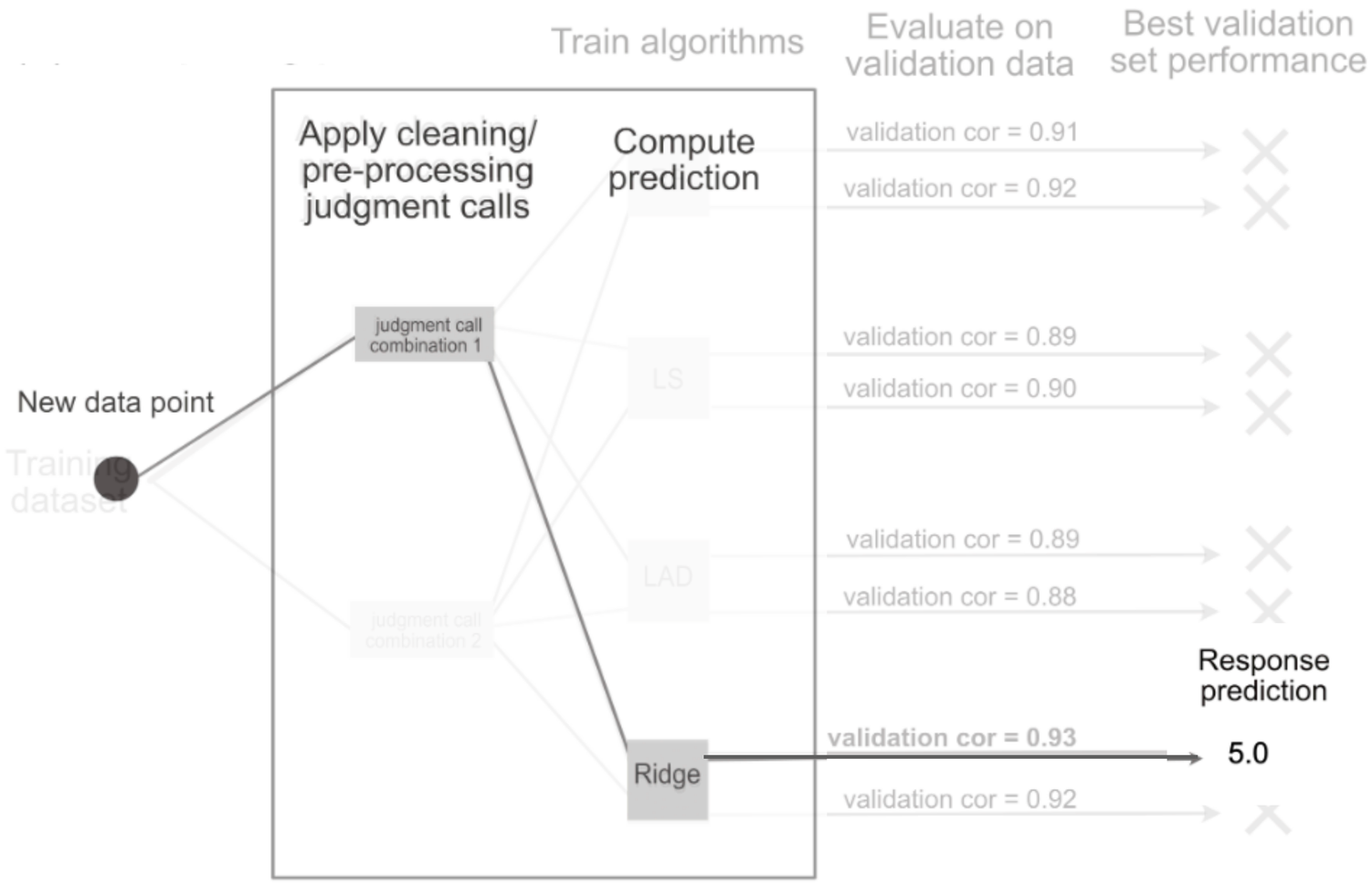
VDS approach to predictive algorithm

Approach 1: Single best PCS fit



VDS approach to predictive algorithm

Approach 1: Single best PCS fit

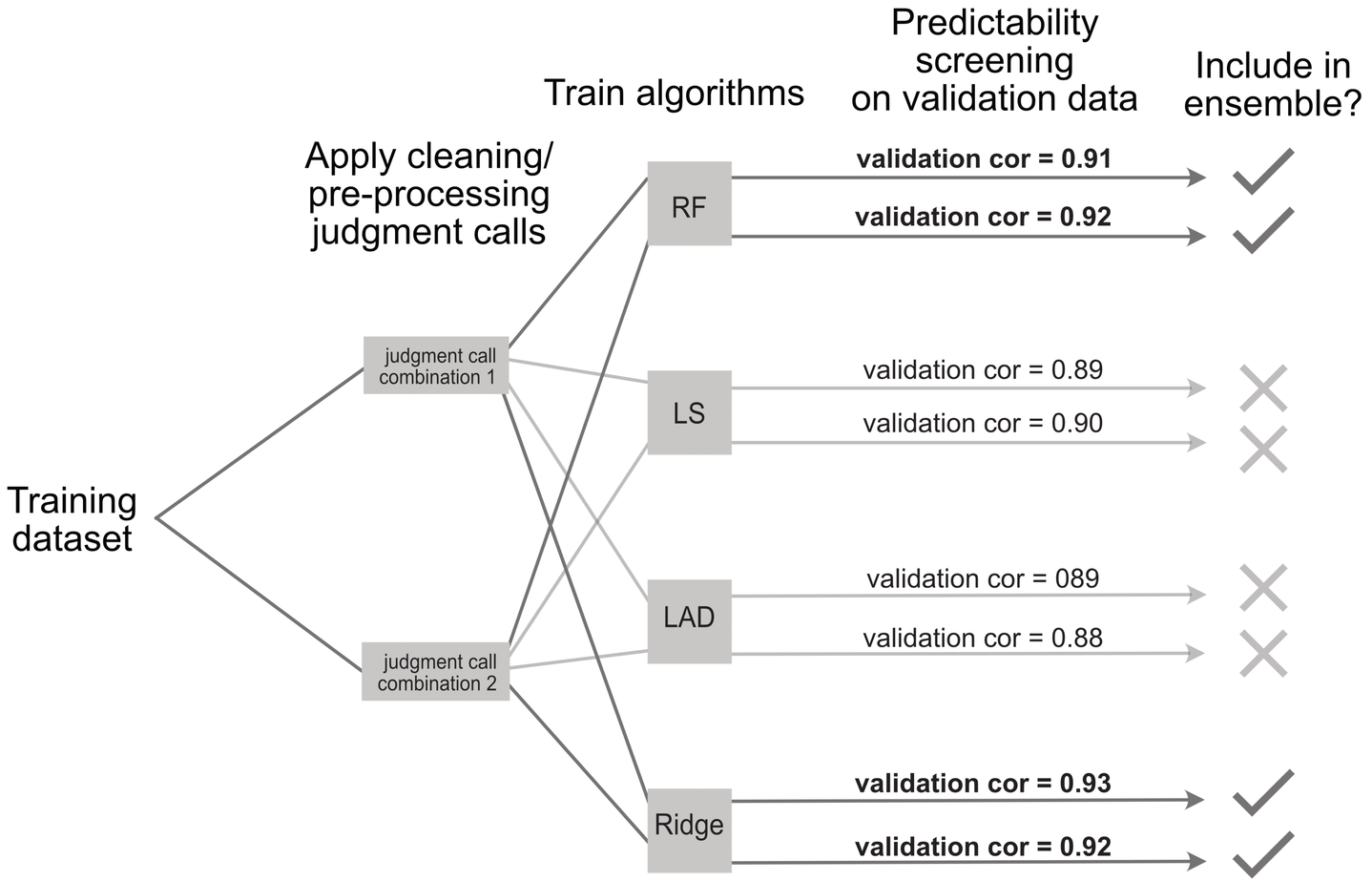


Prediction incorporates **algorithmic** and **judgment call** uncertainty

But it hides these from the end-user…

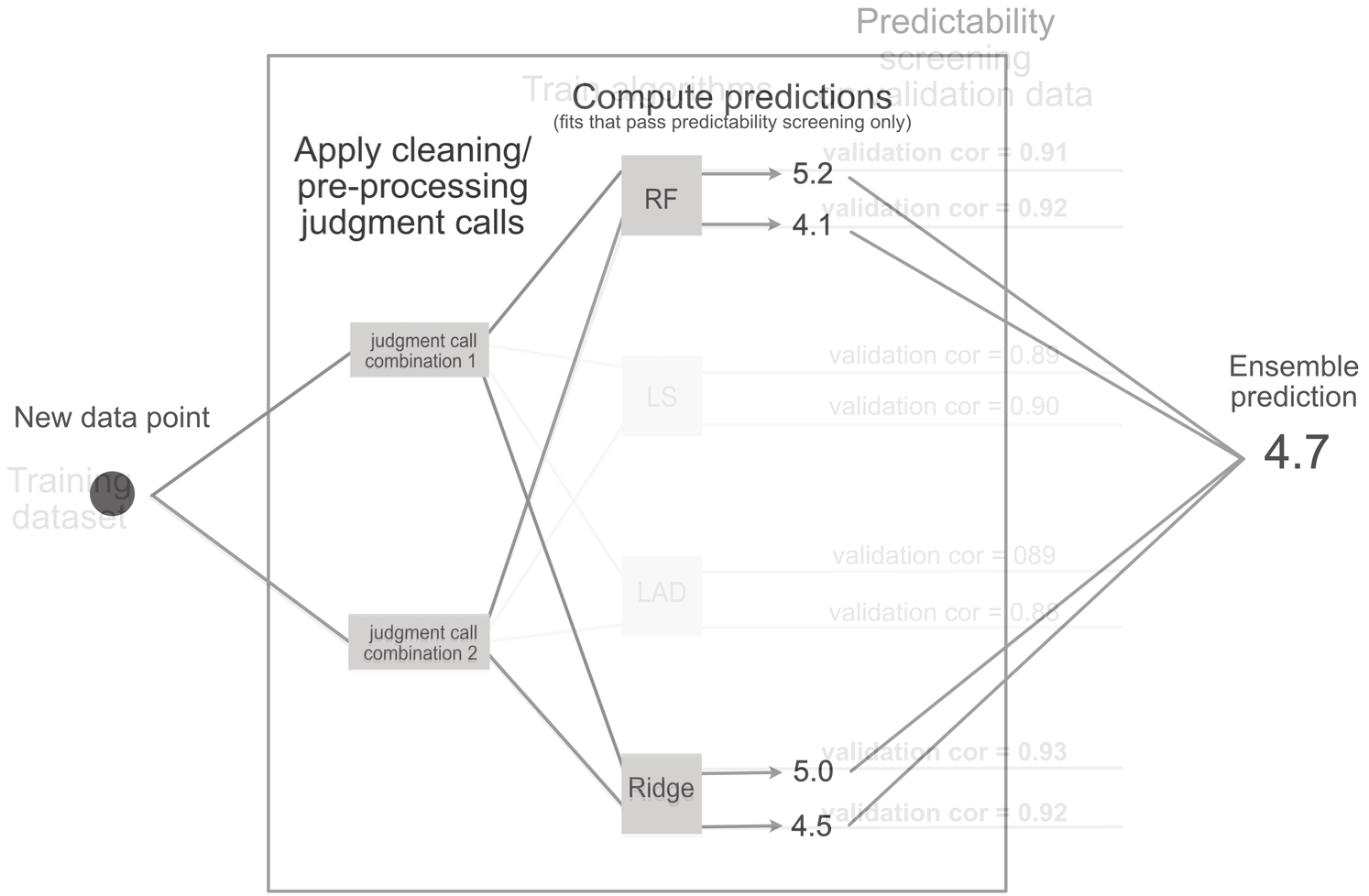
VDS approach to predictive algorithm

Approach 2: Ensemble fit



VDS approach to predictive algorithm

Approach 2: Ensemble fit

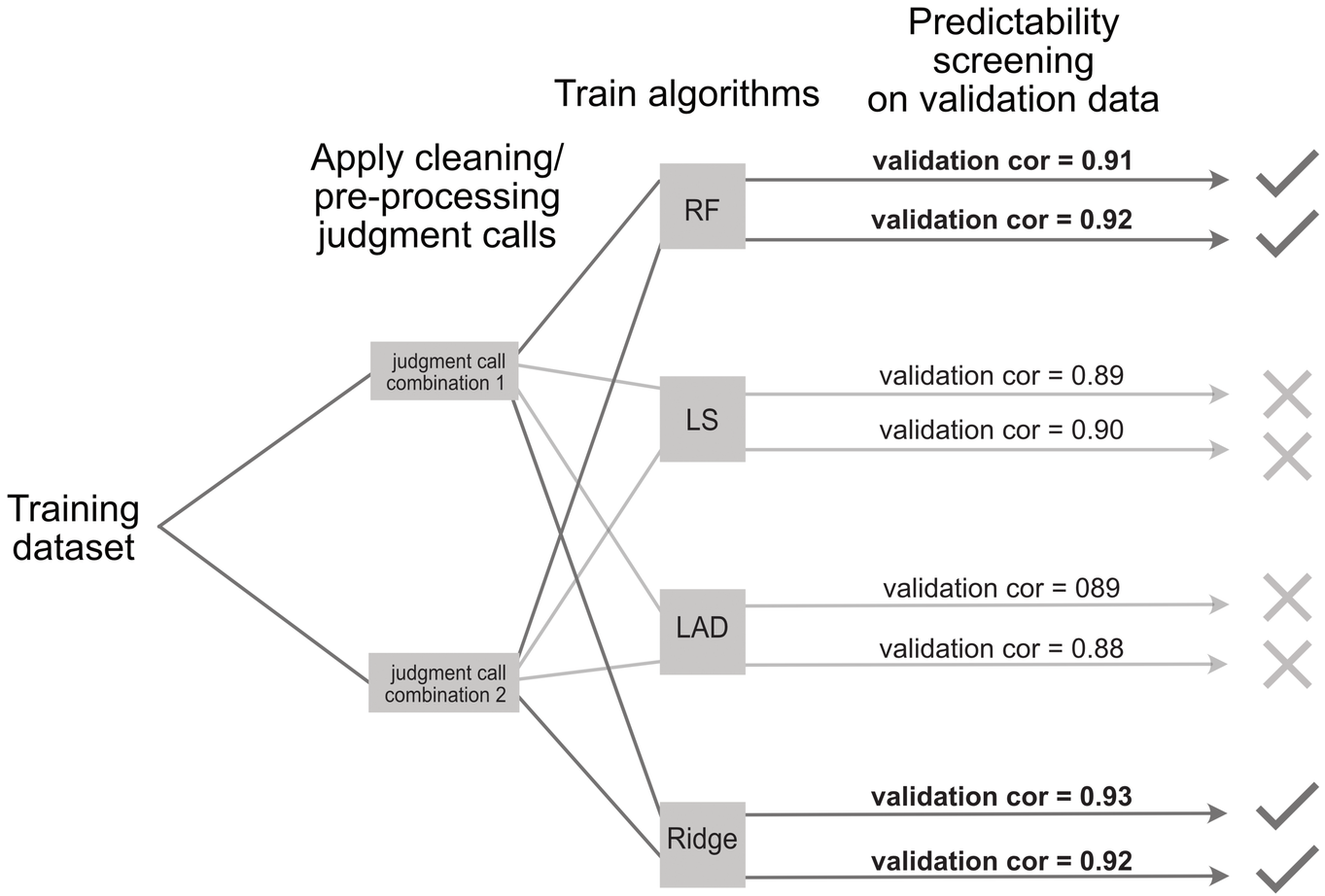


Prediction incorporates **algorithmic** and **judgment call** uncertainty

But it hides these from the end-user…

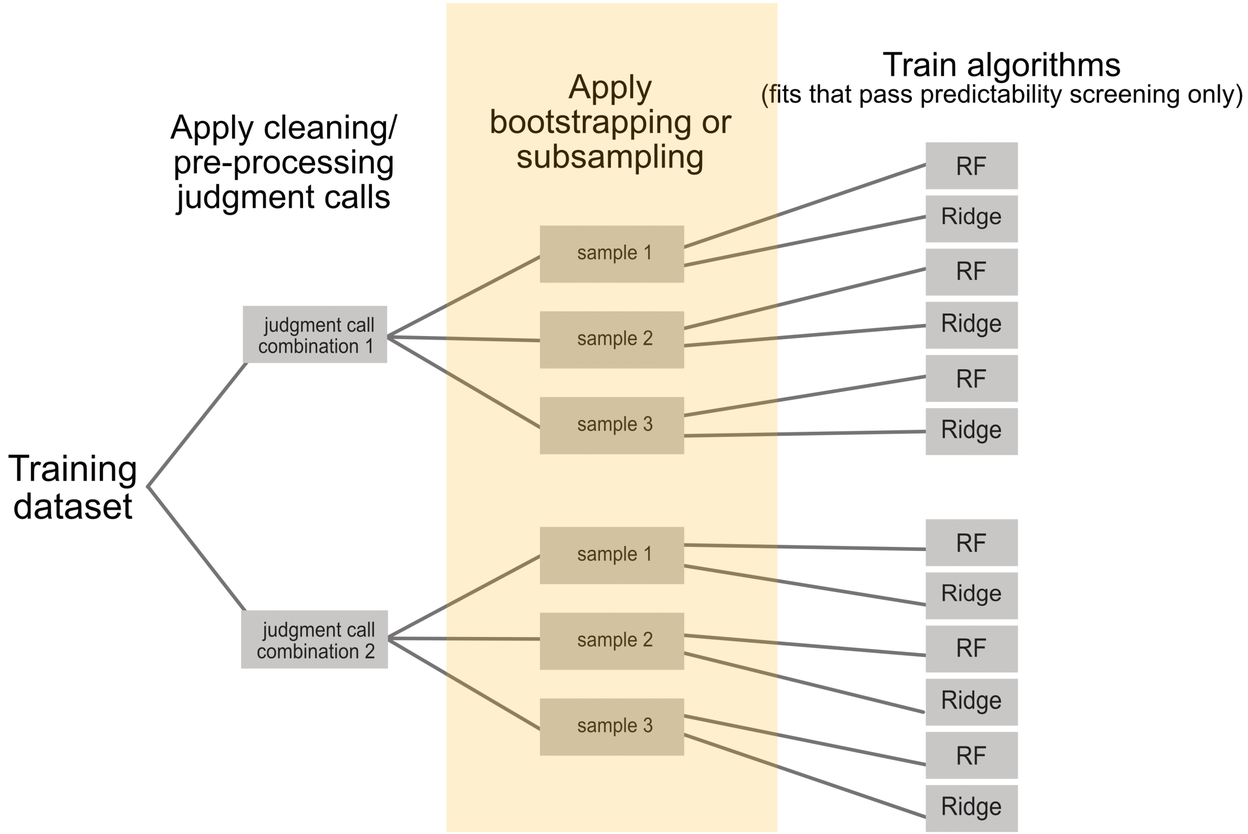
VDS approach to predictive algorithm

Approach 3: Perturbation Prediction Intervals (PPI) [continuous only]



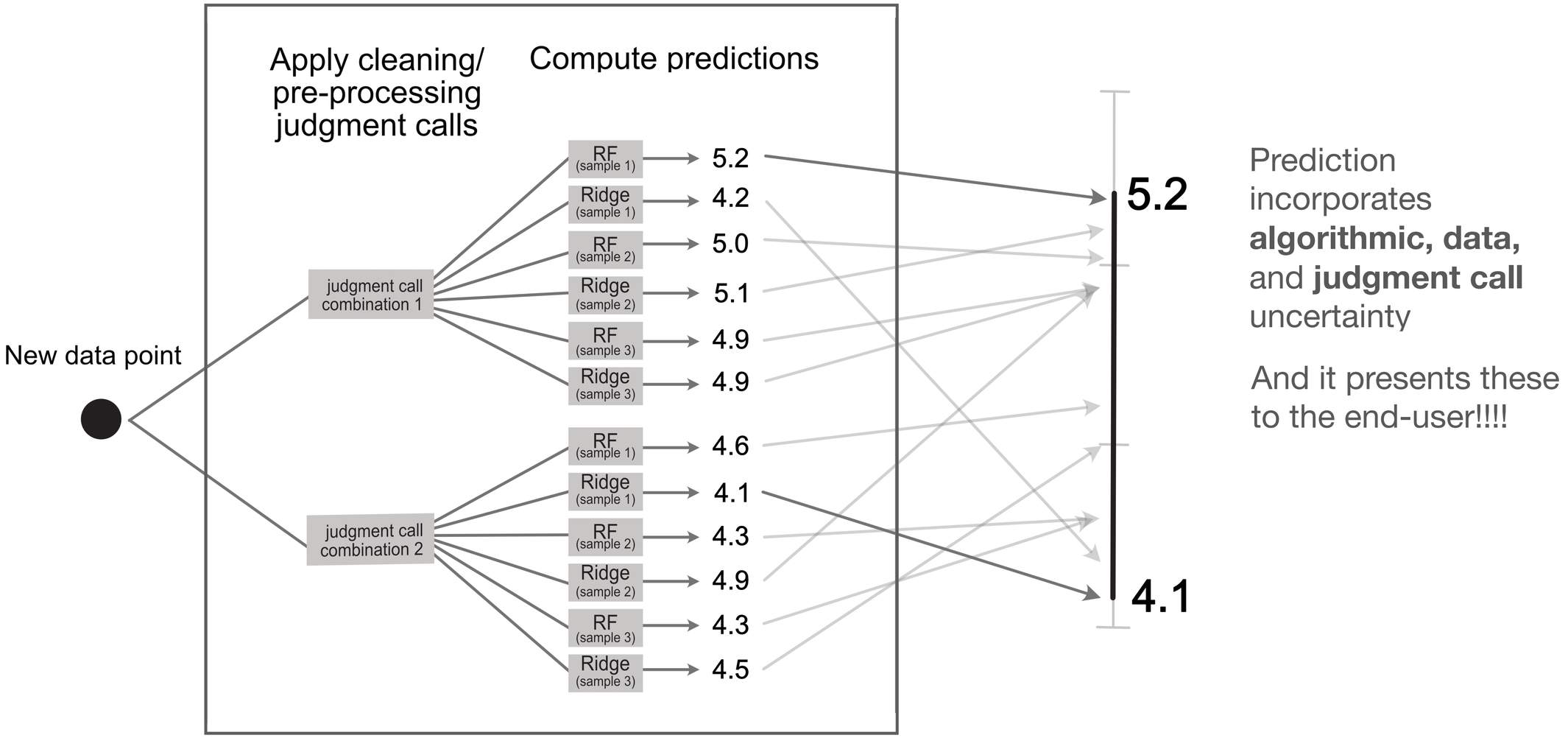
VDS approach to predictive algorithm

Approach 3: Perturbation Prediction Intervals (PPI) [continuous only]



VDS approach to predictive algorithm

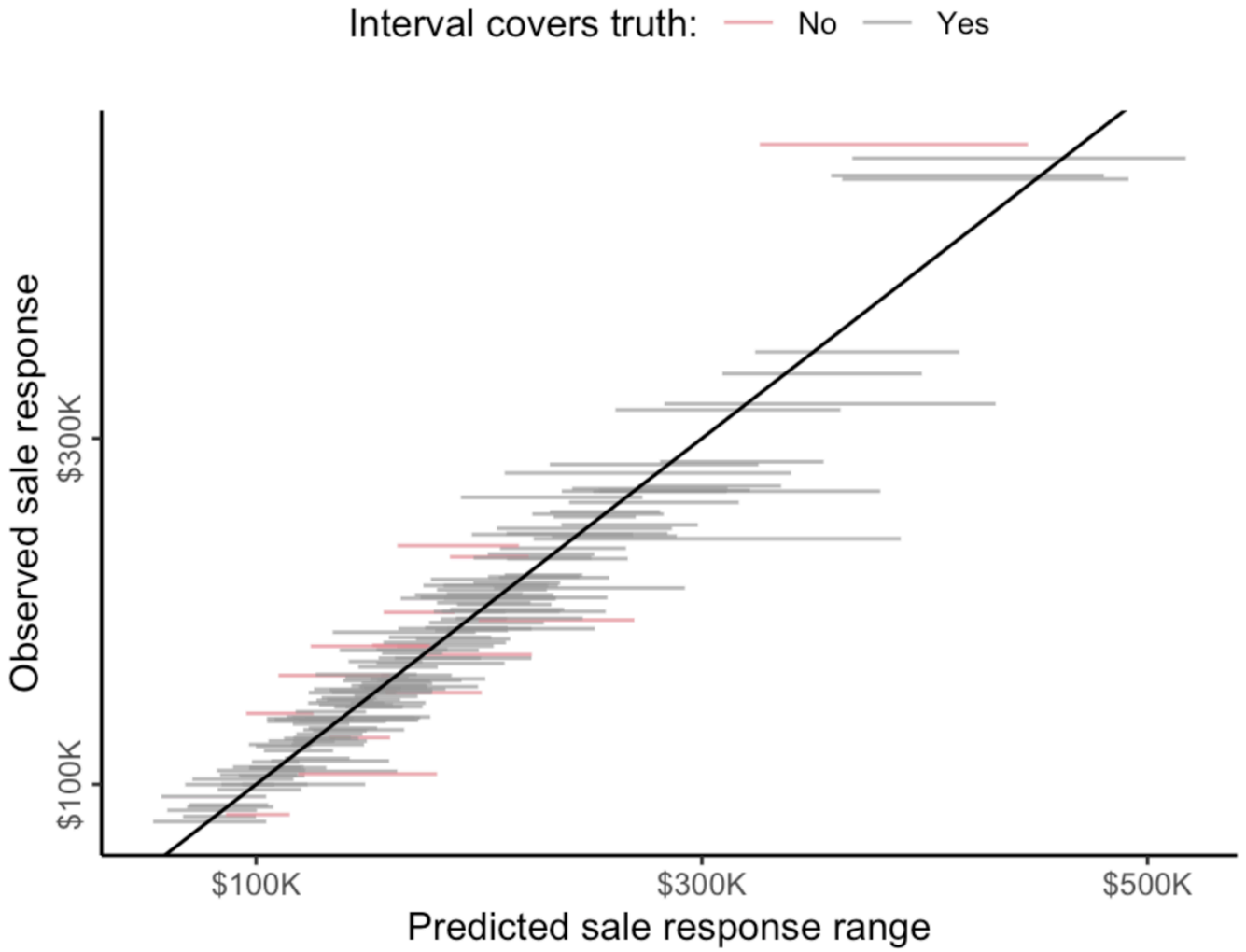
Approach 3: Perturbation Prediction Intervals (PPI) [continuous only]



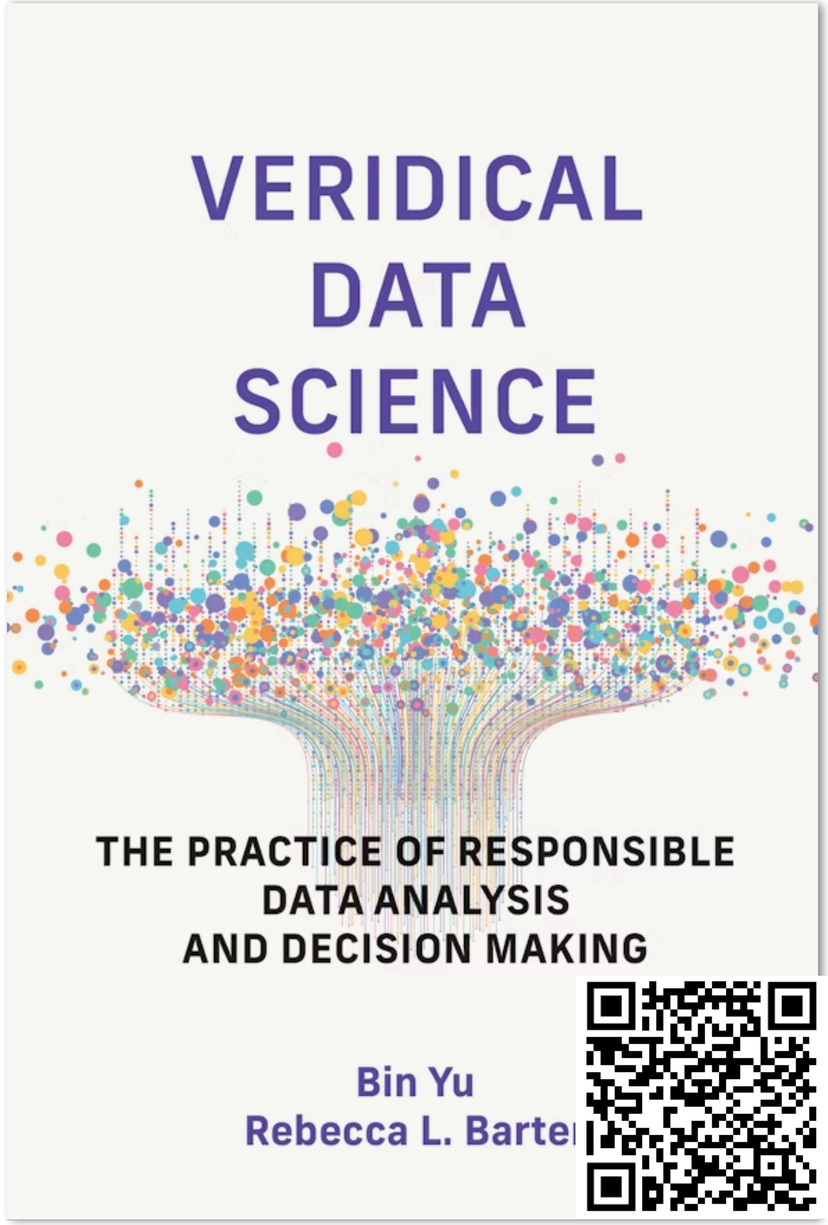
Prediction incorporates **algorithmic, data,** and **judgment call** uncertainty

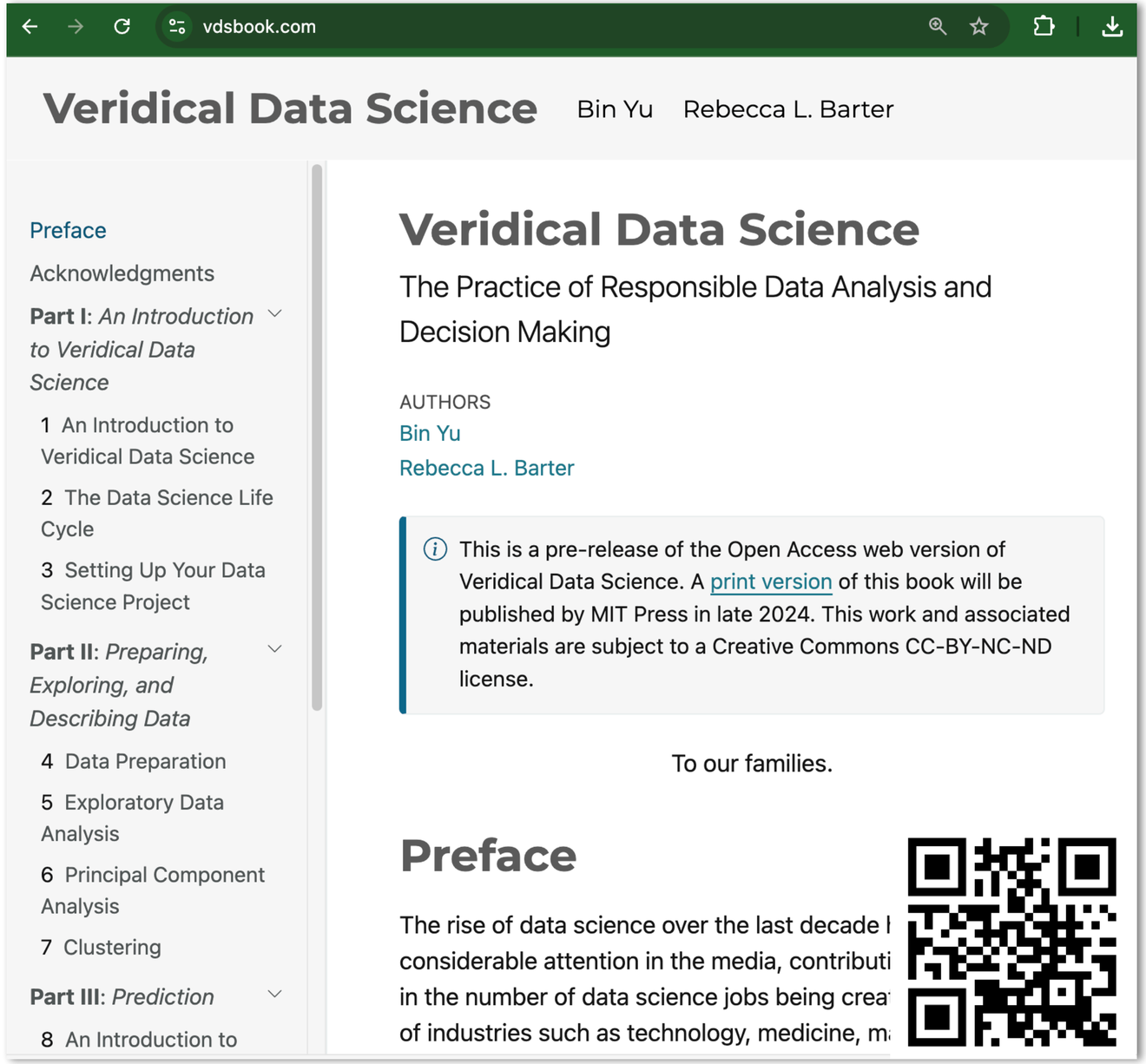
And it presents these to the end-user!!!!

Visualizing predictive performance for PPIs



Evaluate performance using (calibrated) coverage





The MIT Press

Free online version: www.vdsbook.com