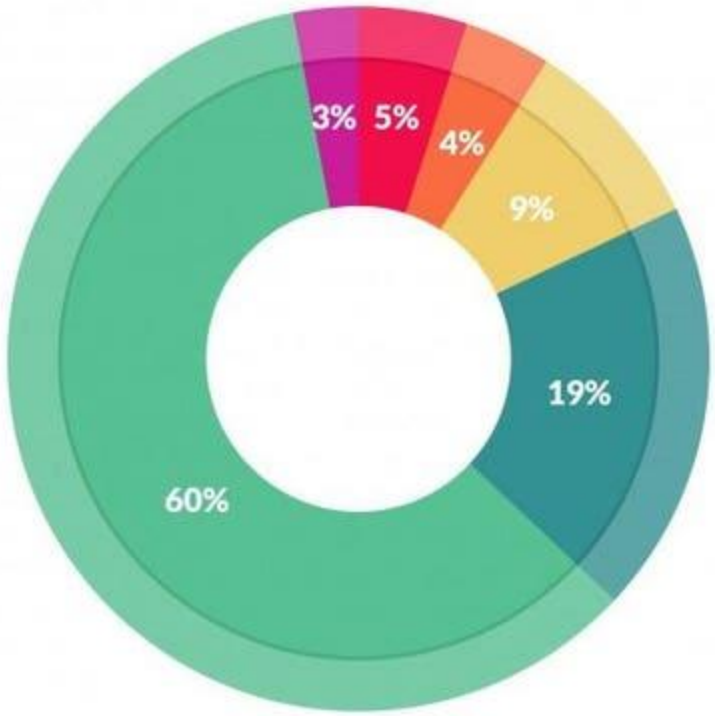
From AutoML to AutoDS

Andreas Mueller

Grey Systems Lab

Microsoft

The Cliché (since 2016)

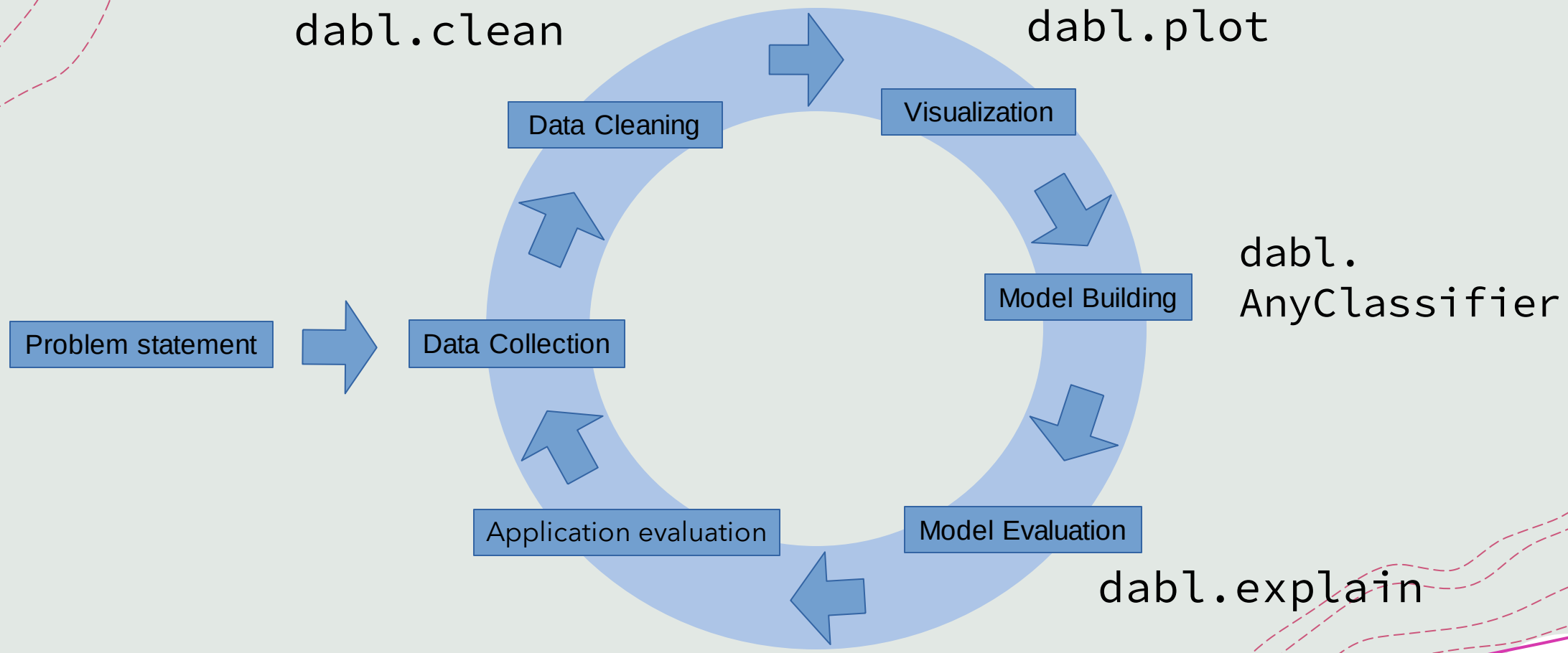


What data scientists spend the most time doing

- *Building training sets: 3%*
- *Cleaning and organizing data: 60%*
- *Collecting data sets; 19%*
- *Mining data for patterns: 9%*
- *Refining algorithms: 4%*
- *Other: 5%*

Cleaning Big Data: Most Time-Consuming, Least Enjoyable Data Science Task, Survey Says (forbes.com)

Dabl: an exploration



https://dabl.github.io

```
detect_types(data)
```

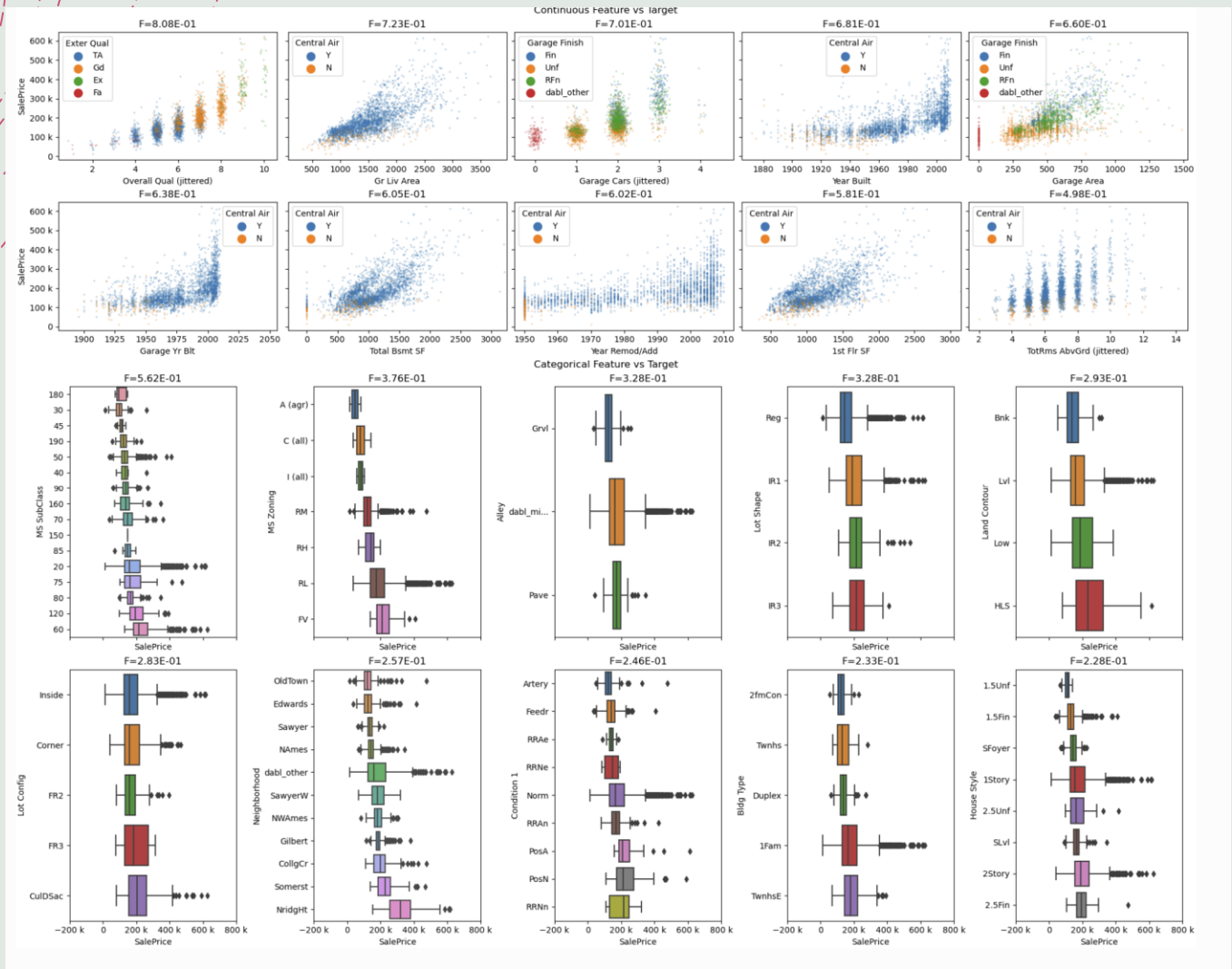
	continuous	dirty_float	low_card_int_ordinal	low_card_int_categorical	categorical	date	free_string	useless
Order	False	False	False	False	False	False	False	True
PID	True	False	False	False	False	False	False	False
MS SubClass	False	False	False	True	False	False	False	False
MS Zoning	False	False	False	False	True	False	False	False
Lot Frontage	True	False	False	False	False	False	False	False
...	...	...	...	...	...	...	...	...
Mo Sold	False	False	True	False	False	False	False	False
Yr Sold	False	False	False	False	True	False	False	False
Sale Type	False	False	False	False	True	False	False	False
Sale Condition	False	False	False	False	True	False	False	False
SalePrice	True	False	False	False	False	False	False	False

82 rows × 8 columns

Thanks to Vraj Shah and Arun Kumar (SortingHat)

```
data = load_ames()

plot(data, target_col='SalePrice')
```



Also see Lux by Doris Lee

We've tried automating the easiest (and fastest) part of model creation

What's so complicated about data collection and preparation?

+ 1) discovery and systems integration. Master data management. Not in this talk but check out Kitana (Huang et al).
+ 2) The kinds & shape of data
+ 3) Forcing things into a classification / regression problem

Typical Data Science Tasks

Here's a lake now:

+Build a (and deploy) churn model.

+Why did the sales for product X in region Y drop?

+Why is our website slow today?

+Which of the students are most at risk of not graduating?

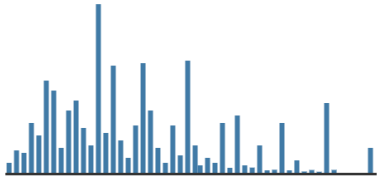
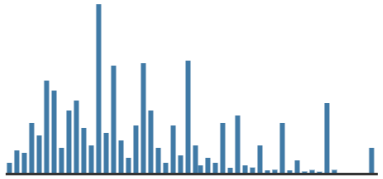
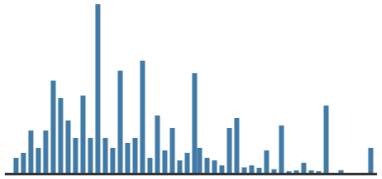
+What subpopulation is most at risk of MPox?

(most) Important Data Is Tabular

loan.csv (1.19 GB)

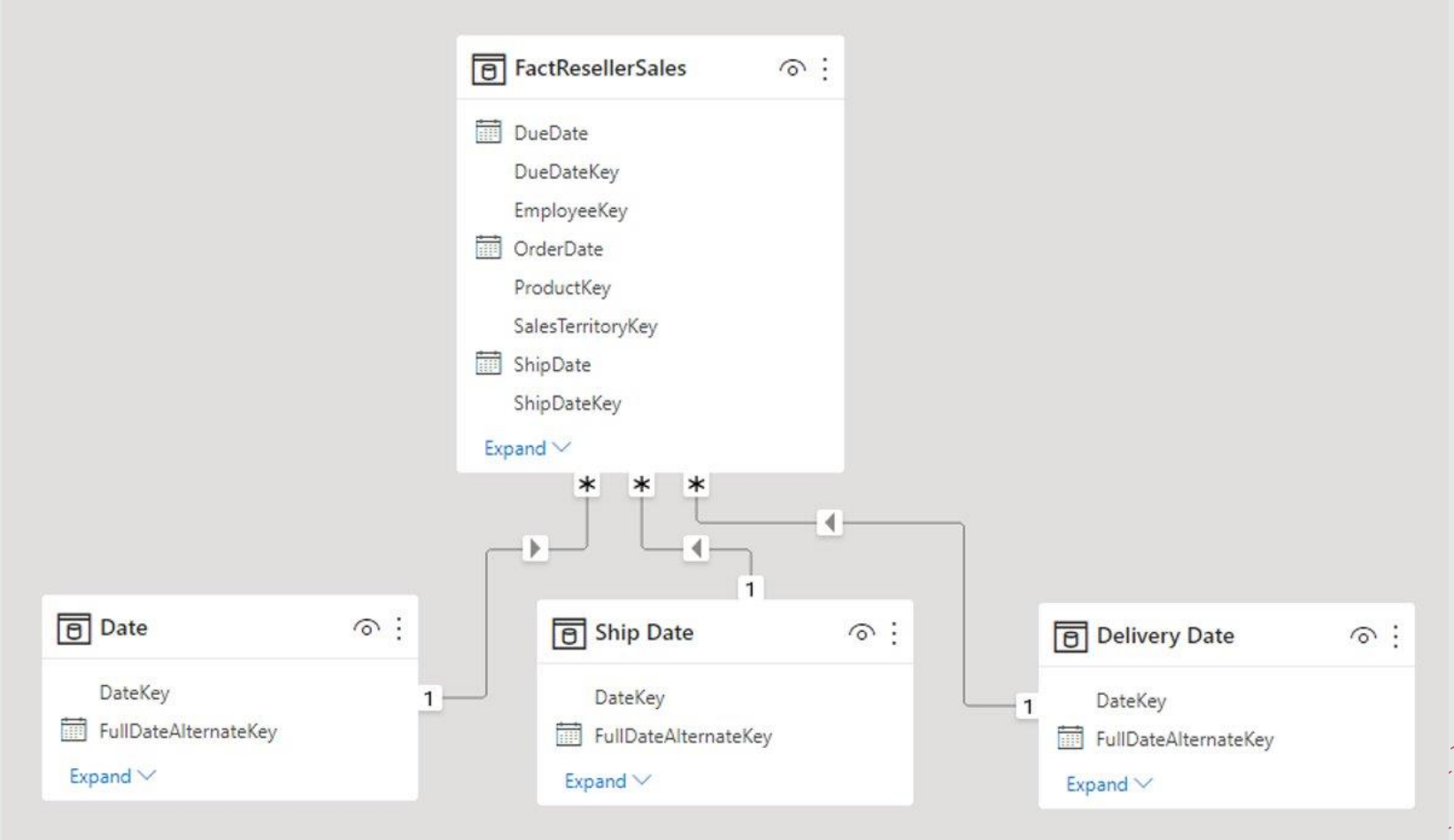
Detail Compact Column

10 of 145 columns

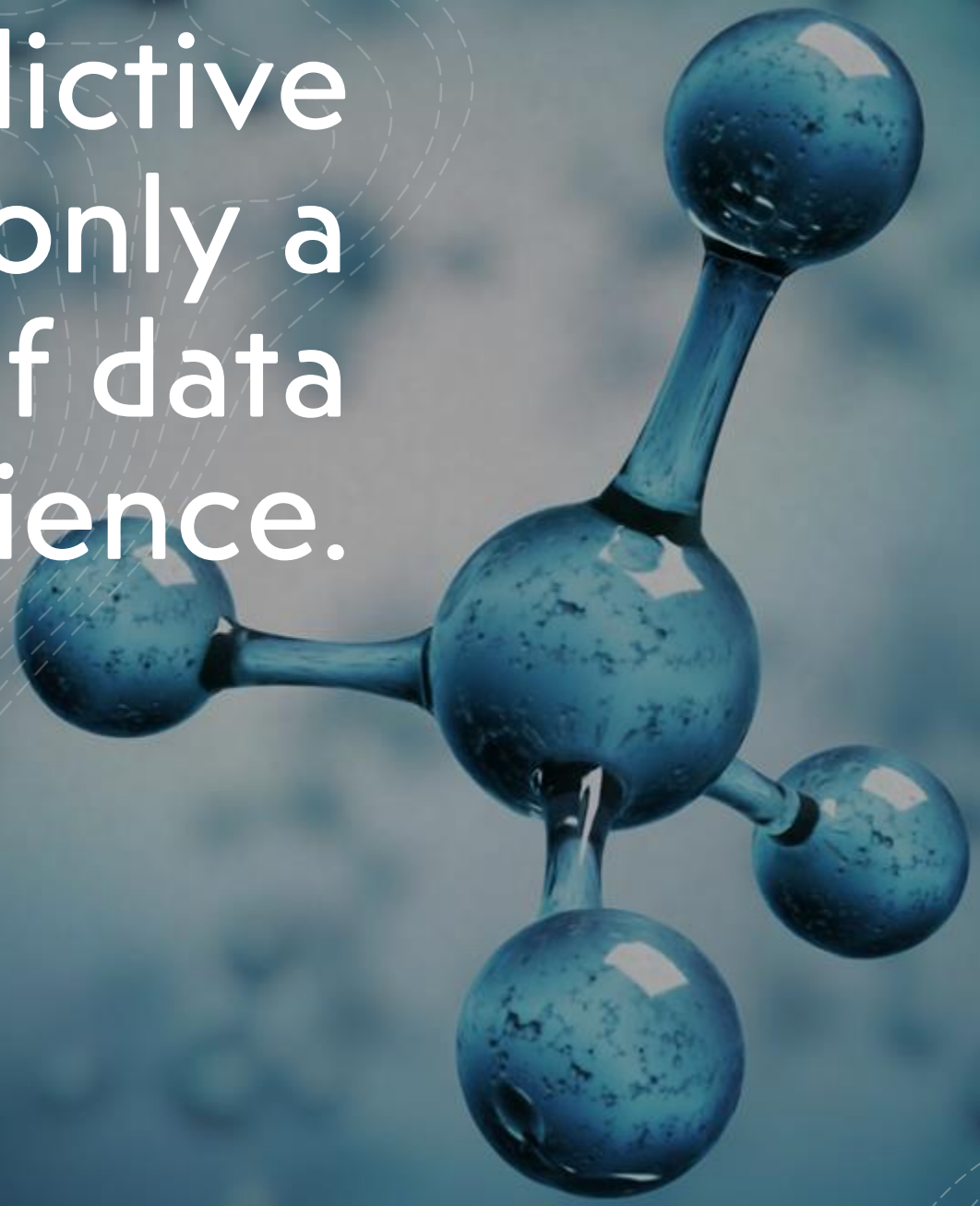
id	member_id	loan_amnt	funded_amnt	funded_amnt_inv	term
[null] 100%	[null] 100%	 500 40.0k	 500 40.0k	 0 40.0k	36 months 60 months
		2500	2500	2500	36 months
		30000	30000	30000	60 months
		5000	5000	5000	36 months
		4000	4000	4000	36 months

All data is time series data

(nearly) all data is relational



Predictive modeling is only a fraction of data science.

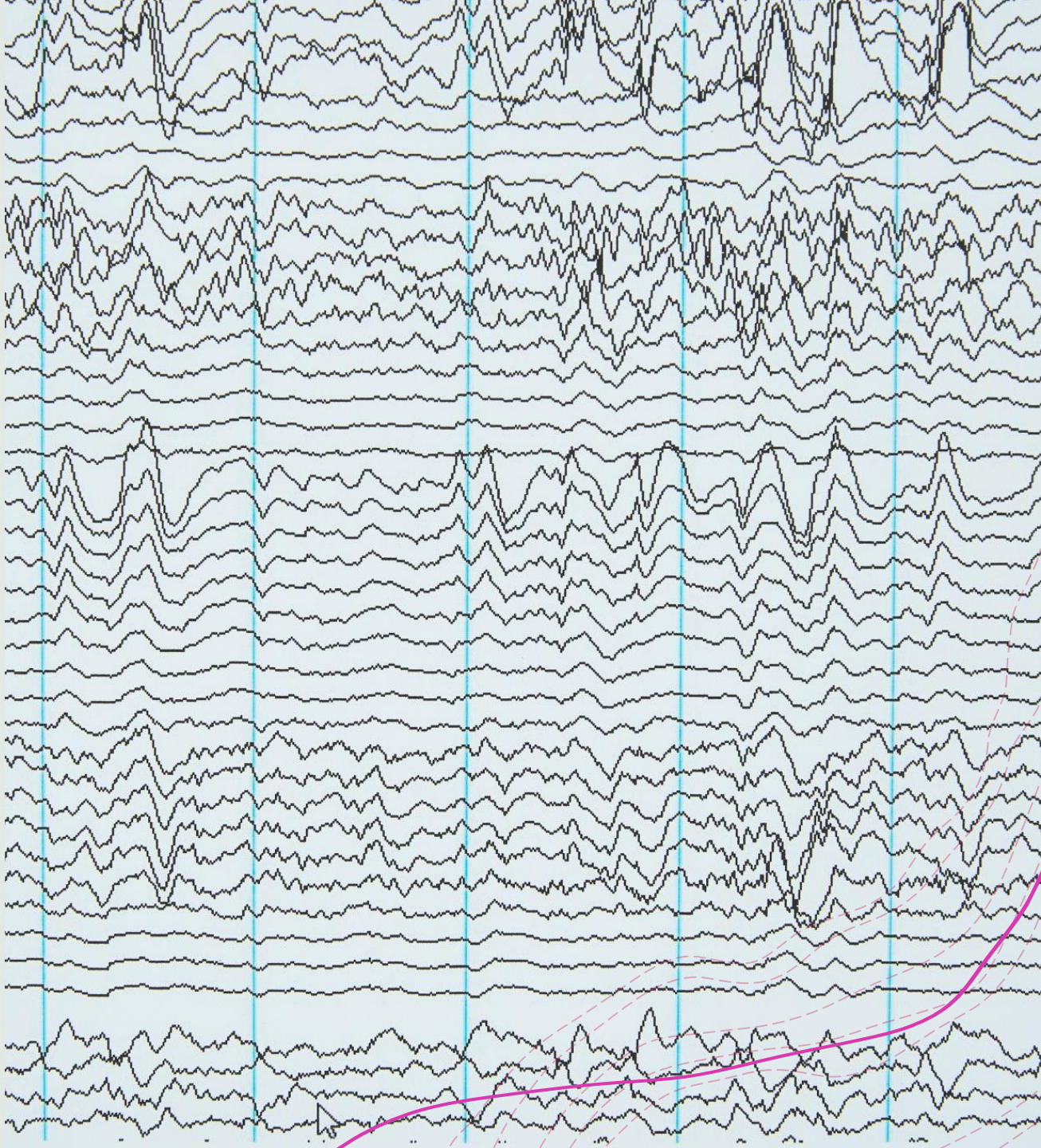


What is the output of Data Science?

- "typical" ML: a deployed model.
- IRL often: a graph / number / insight

We need to talk to BI & Stats & Vis more.

- The two standard tools for causal modeling in practice are:
  - A single scatter plot
  - Random Forest feature importances (or SHAP)



Expanding AutoML to AutoDS

We don’t even have the data

+No “as of” datasets?

+No datasets on feature engineering

+Few relational datasets

+Unclear datasets/tasks for EDA

+Few Datasets for causal analysis / outlier detection

+Hypothesis creation benchmarks?

+No (?) Distributional drift based on deployed model

Teaser: Learning on Relational and Semantic Data

- Getting to closer to the source will make understanding the data EASIER!
- The source might have additional meta-data

See https://www.semanticlayersummit.com/

- Getml.com (propositionalization)
- Kumo.ai (graph NNs)
- AtScale.com (focus on infrastructure)

Take Away

More realistic data

- Dirty tables
  (there's some progress here)
- Time dependent
  (including as-of)
- Relational
  (including encoded semantics)

More realistic tasks

1. Interpretable models
2. Root cause analysis
3. Causal models
4. Decision support & inference
5. **Hypothesis generation**