Microsoft

Data Science Approach / Philosophy

Danielle Dean

TL;DR

Data science isn't magic,
but there are tricks
and secrets.

You can’t use just any data.

Recipe

Crispy crust
Marinara sauce
Fresh mozzarella
Thick cut pepperoni
Fresh tomato

Recipe

Data that is

- Relevant
- Connected
- Accurate
- Enough

and a Sharp Question

Irrelevant data

Price of milk ($/gal)	Red Sox batting avg.	Blood alcohol content (%)
3.79	.304	.03
3.45	.320	.09
4.06	.259	.01
3.89	.298	.05
4.12	.332	.13
3.92	.270	.06
3.23	.294	.10

Relevant data

Body mass (kg)	Margaritas	Blood alcohol content (%)
103	3	.03
67	5	.09
87	1	.01
52	2	.05
73	5	.13
79	3	.06
110	7	.10

[data points] [rows] [samples] [features] [columns] [attributes] [table] [database]

Disconnected data

Grill temperature (F)	Weight of beef patty (lb)	Burger rating (out of 10)
	.33	8.2
	.24	5.6
550		7.8
725	.45	9.4
600		8.2
625		6.8
	.49	4.2

[missing values]

Connected data

Grill temperature (F)	Weight of beef patty (lb)	Burger rating (out of 10)
575	.33	8.2
550	.24	5.6
550	.69	7.8
725	.45	9.4
600	.57	8.2
625	.36	6.8
550	.49	4.2

Disconnected data

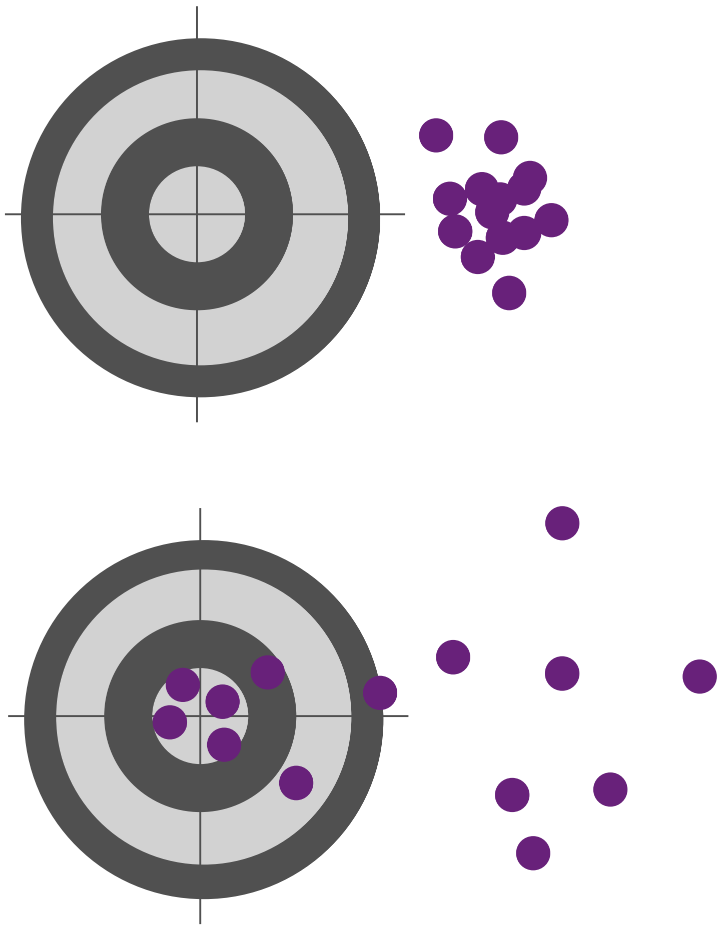
Grill temperature (F)	Weight of beef patty (lb)	Burger rating (out of 10)
	.33	8.2
	.24	5.6
550		7.8
725	.45	9.4
600		8.2
625		6.8
	.49	4.2

[missing values]

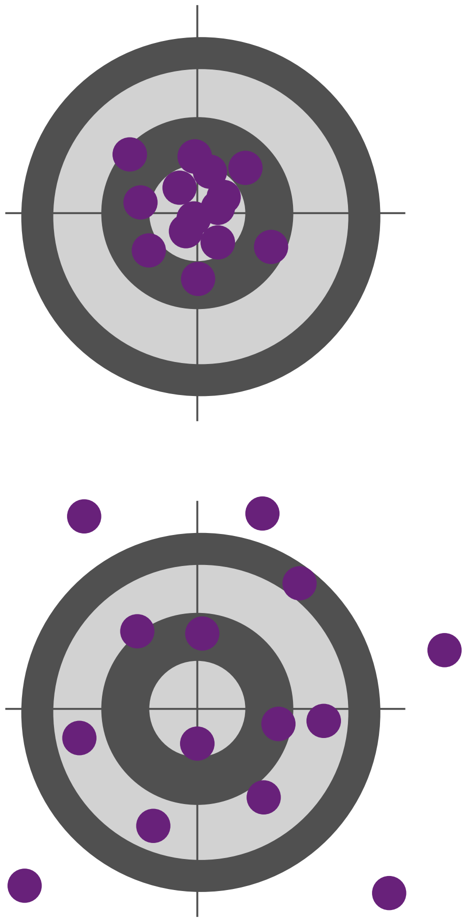
Connected data

Grill temperature (F)	Weight of beef patty (lb)	Burger rating (out of 10)
575	.33	8.2
550	.24	5.6
550	.69	7.8
725	.45	9.4
600	.57	8.2
625	.36	6.8
550	.49	4.2

Inaccurate data

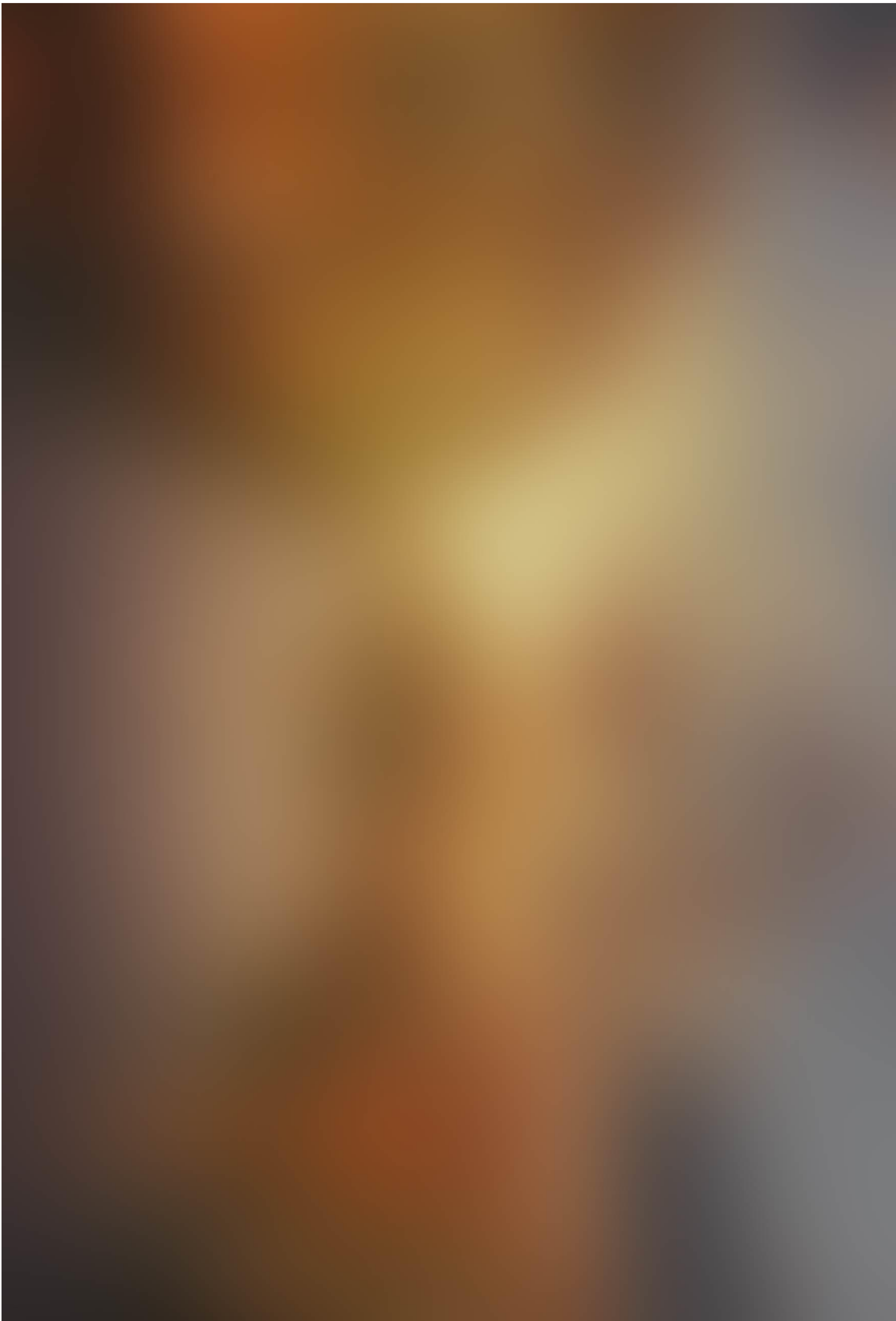


Accurate data

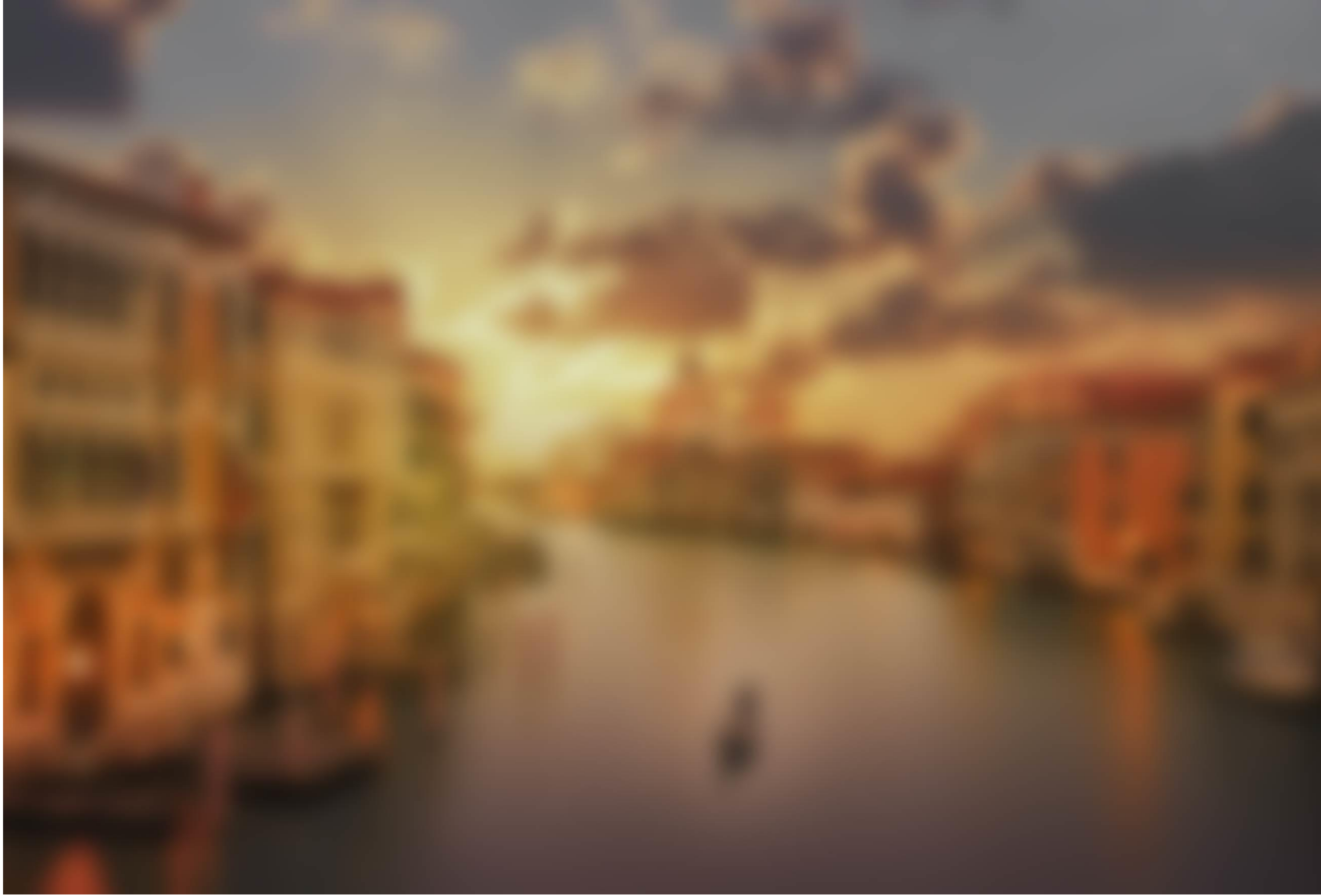


[precision] [accuracy] [bias]

Not enough data



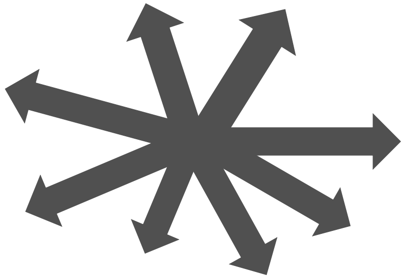
Barely enough data



Enough data

Vague questions vs. Sharp questions

Vague questions



Can't be answered with a name or a number

What can my data tell me about my business?

What should I do?

How can I increase my profits?

Sharp questions

Can be answered with a name or a number.

How many Model Q Gizmos will I sell in Montreal during the third quarter?

Which car in my fleet is going to fail first?

Are we ready for ML?

Question is sharp.	Data measures what they care about.	Data is accurate.	Data is connected.	A lot of data.
E.g. Predict whether component X will fail in the next Y days	E.g. Identifiers at the level they are predicting	E.g. Failures are really failures, human labels on root causes	E.g. Machine information linkable to usage information	E.g. Will be difficult to predict failure accurately with few examples

Predictive Maintenance – Data Science Approach

Qualification Criteria

Goal is to create a **generalizable model**

For ML-based solution:

1. Problem is predictive in nature
2. Clear path of action if potential failures detected
3. Data with sufficient quality
    - For predicting time left to failure, do you have failures or some proxy recorded?
    - Do you have enough failures to be able to model?
    - Is the “non-IoT” data in usable format?
    - Can the domain knowledge, such as timing of maintenance recordings, be translated into usable data for modeling?

Data Sources

FAILURE HISTORY

The failure history of a machine or component within the machine.

REPAIR HISTORY

The repair history of a machine, e.g. previous maintenance records, components replaced, maintenance activities performed. Maintenance types.

MACHINE CONDITIONS

The operation conditions of a machine, e.g. data collected from sensors.

MACHINE FEATURES

The features of machine or components, e.g. production date, technical specifications.

OPERATING CONDITIONS

Environmental features that may influence a machine's performance, e.g. location, temperature, other interactions.

OPERATOR ATTRIBUTES

The attributes of the operator who uses the machine, e.g. driver.

Feature Engineering

The process of creating features that provide better or additional predictive power to the learning algorithm.

id	cycle	setting1	setting2	setting3	s1	s2	s3	...	s19	s20	s21
1	1	-0.0007	-0.0004	100	518.67	641.82	1589.7		100	39.06	23.419
1	2	0.0019	-0.0003	100	518.67	642.15	1591.82		100	39	23.4236
1	3	-0.0043	0.0003	100	518.67	642.35	1587.99		100	38.95	23.3442
...	...										
1	191	0	-0.0004	100	518.67	643.34	1602.36		100	38.45	23.1295
1	192	0.0009	0	100	518.67	643.54	1601.41		100	38.48	22.9649
2	1	-0.0018	0.0006	100	518.67	641.89	1583.84		100	38.94	23.4585
2	2	0.0043	-0.0003	100	518.67	641.82	1587.05		100	39.06	23.4085
2	3	0.0018	0.0003	100	518.67	641.55	1588.32		100	39.11	23.425
...	...										
2	286	-0.001	-0.0003	100	518.67	643.44	1603.63		100	38.33	23.0169
2	287	-0.0005	0.0006	100	518.67	643.85	1608.5		100	38.43	23.0848

a1	a2	...	a21	sd1	sd2	...	sd21	RUL	label1	label2

Other potential features: change from initial value, velocity of change, frequency count over a predefined threshold

Good to utilize domain knowledge, often better than “auto-featurizing”

Example Feature Engineering Methods

1- Rolling aggregates:

For each labelled record of an asset, pick a rolling window of size w, compute rolling aggregate features for the periods before the labelling date and time of that record.

2- Lag features for short term:

For each labelled record of an asset, pick a window of size w and use tumbling windows to create aggregate features for the periods before the labelling date and time.

3- Lag features for long term:

For each labelled record, find aggregated features for a larger window than w reflecting the long term effects.

Create features that capture degradation over time.

Modelling Techniques

BINARY CLASSIFICATION

Predict failures within a future period of time

REGRESSION or SURVIVAL ANALYSIS

Predict remaining useful life, the amount of time before the next failure

MULTICLASS CLASSIFICATION

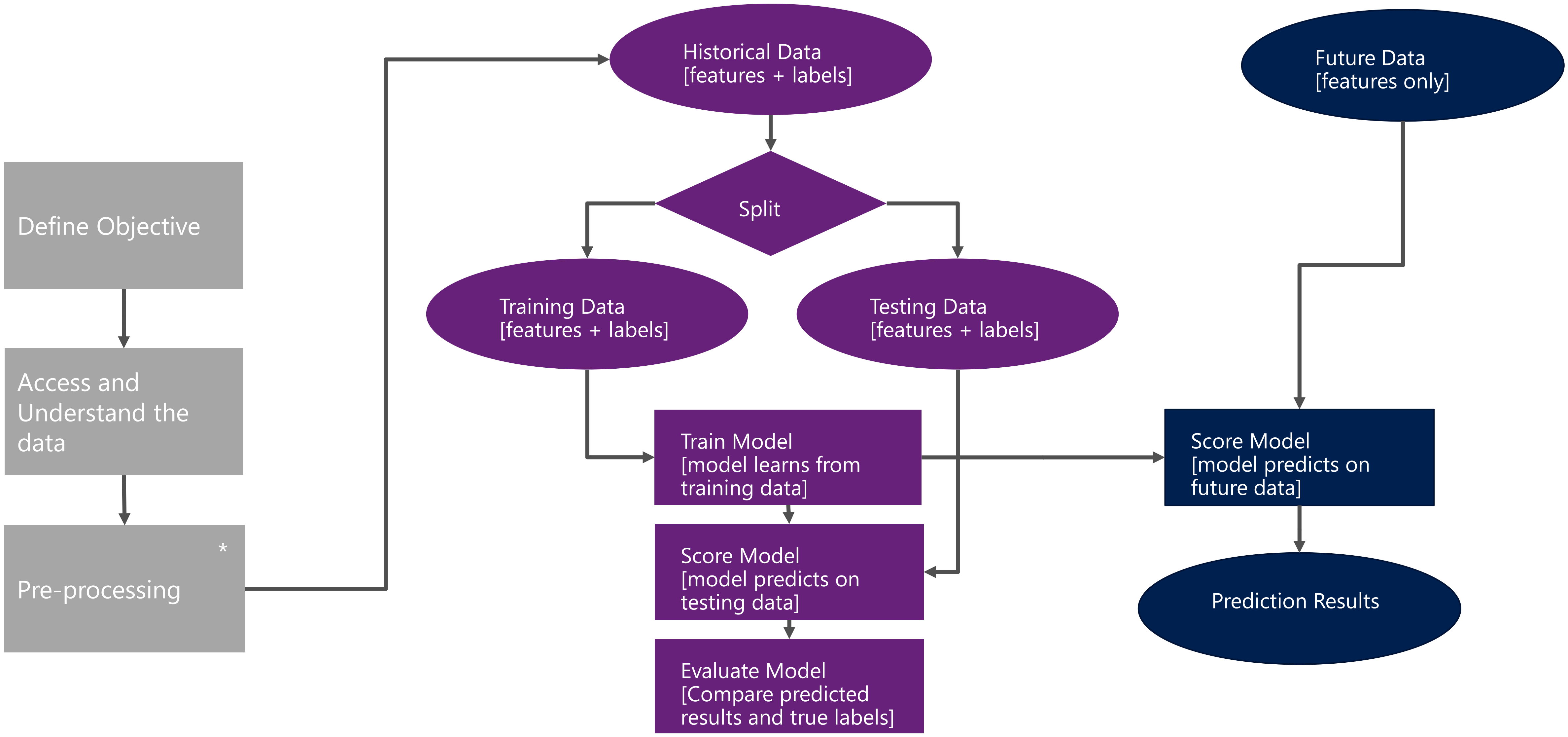
Predict failures with their causes within a future time period.

Predict remaining useful life within ranges of future periods

ANOMALY DETECTION

Identify change in normal trends to find anomalies

ML Process


Define Objective
Access and Understand the data
Pre-processing
*
Historical Data [features + labels]
Split
Training Data [features + labels]
Testing Data [features + labels]
Train Model [model learns from training data]
Score Model [model predicts on testing data]
Evaluate Model [Compare predicted results and true labels]
Future Data [features only]
Score Model [model predicts on future data]
Prediction Results