Train In Data

Probability Distributions

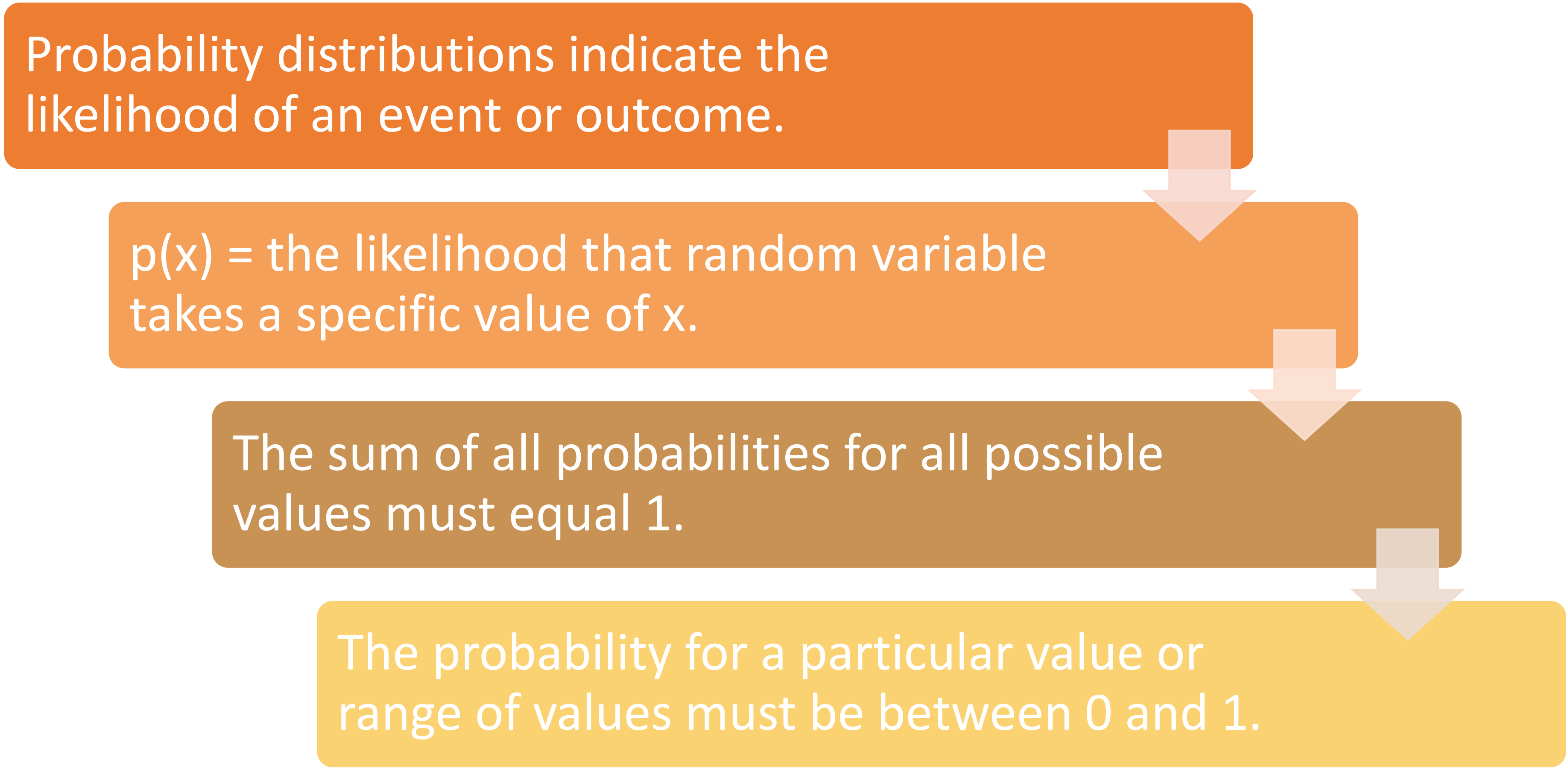
What is a probability distribution?

- A probability distribution is a function that describes the likelihood of obtaining the possible values that a variable can take.

- For the variable height, the probability distribution describes how often we can get a value of 161 cm, or 174 cm, or 200 cm, etc.

- As you can infer from the previous, it is more likely to obtain values between 161 – 170 cm, than values around or bigger than 200 cm.

Properties of probability distributions



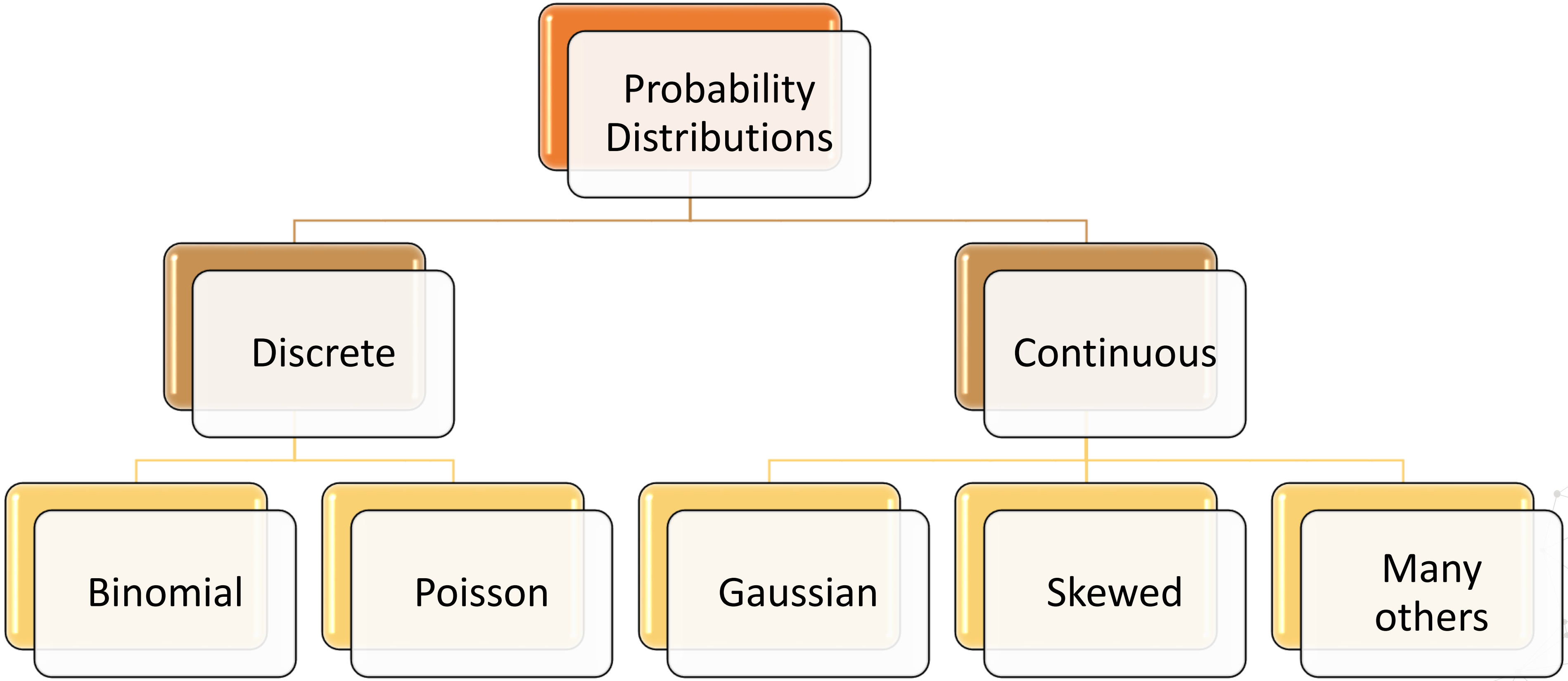
Probability distributions indicate the likelihood of an event or outcome.

p(x) = the likelihood that random variable takes a specific value of x.

The sum of all probabilities for all possible values must equal 1.

The probability for a particular value or range of values must be between 0 and 1.

Different probability distributions



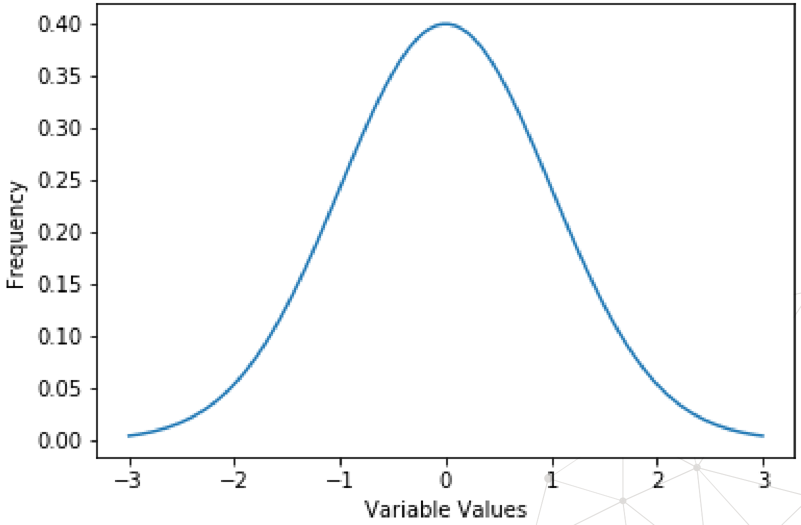
Gallery of probability distributions

Follow this link for more probability distributions.

The Normal distribution

- Many natural phenomena follow a normal distribution
  - Height, blood pressure, etc.

- Symmetric:
  - Most of the observations occur around the central peak
  - Probabilities for values further away from the centre decrease equally in both directions.
  - Extreme values in both tails of the distribution are similarly unlikely.

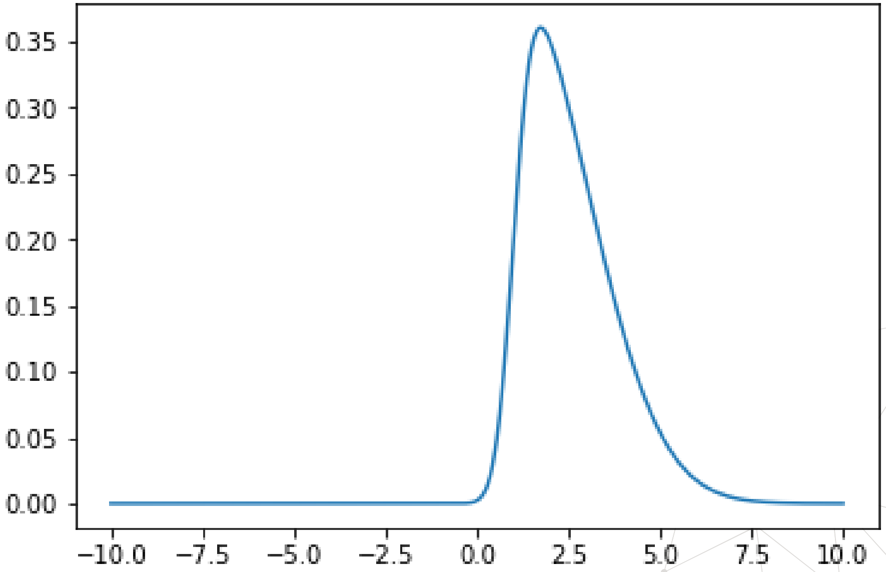


Skewed distributions

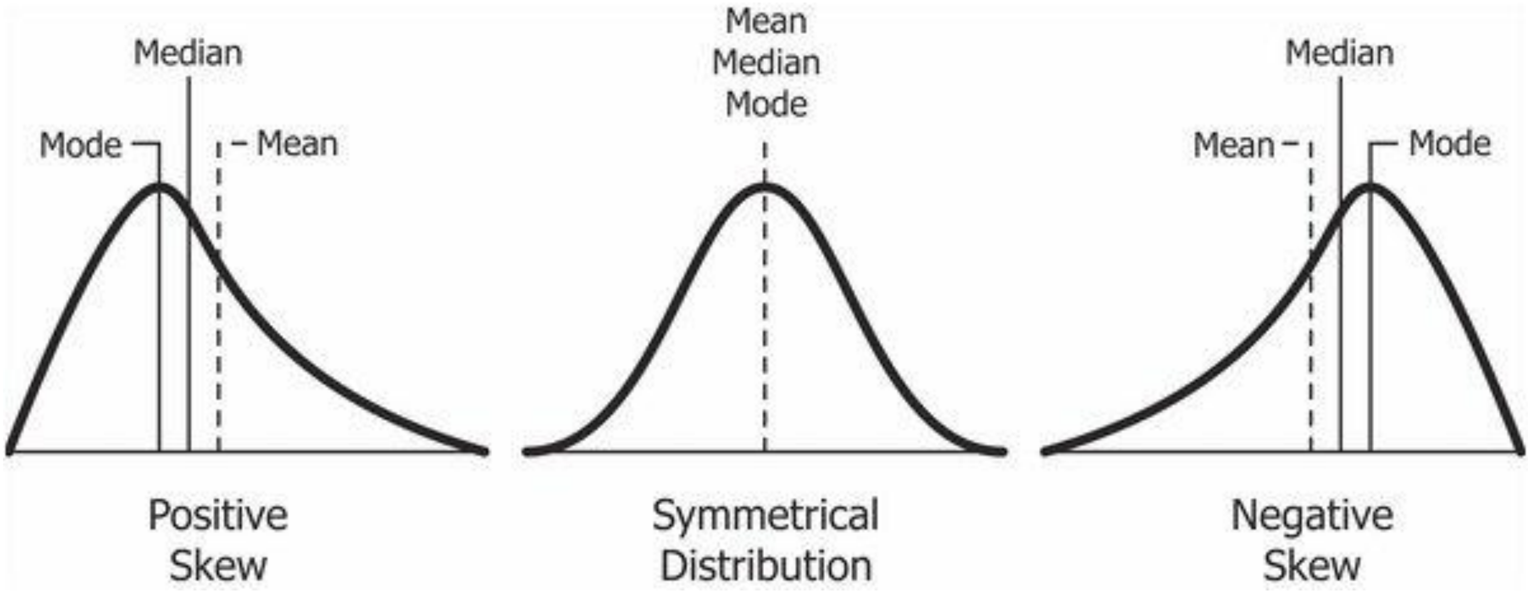
- A distribution is skewed if one of its tails is longer than the other

- A left-skewed distribution has a long left tail. Also called negatively-skewed distributions.

- A right-skewed distribution shows a long right tail. Also called positive-skew distributions.



Gaussian vs Skewed distributions

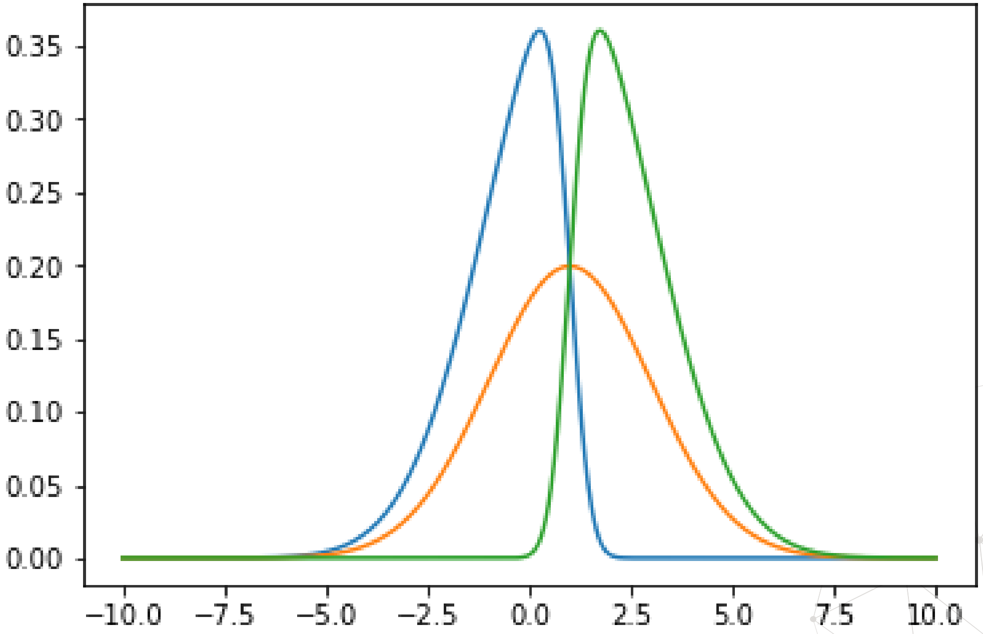


- In Normal distributions, the mean, median and mode are the same
- For skewed distributions, the mean is influenced by the tail

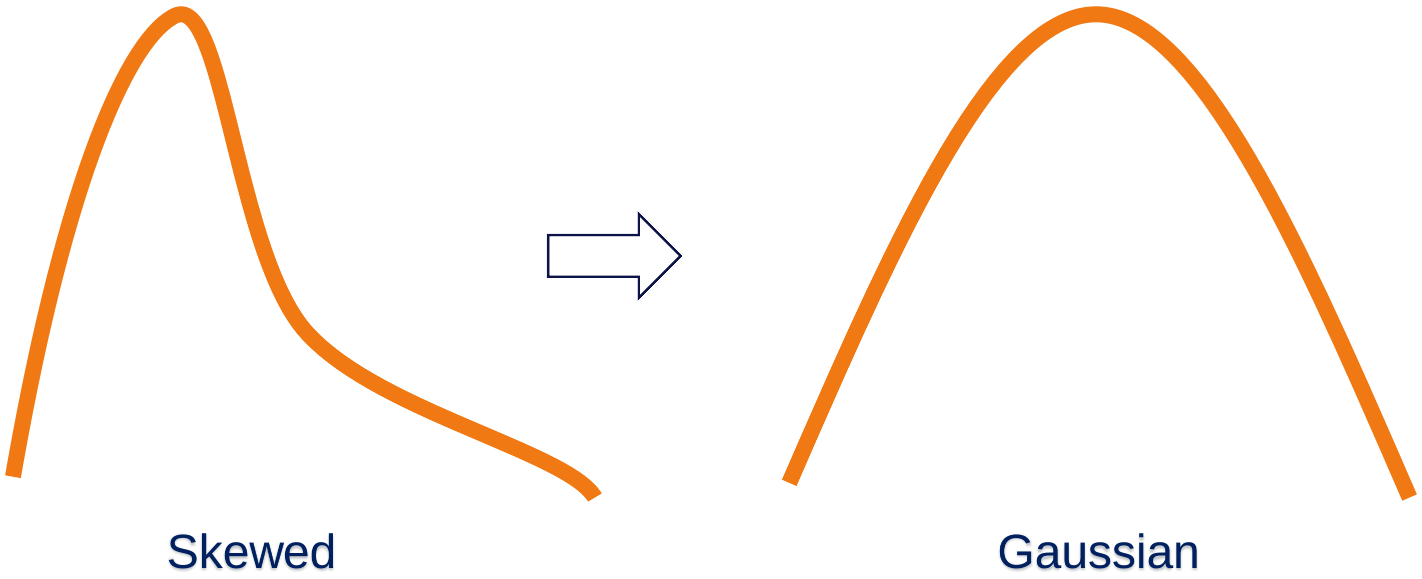
Distributions and model performance

- Linear Models assume that the residuals are normally distributed.

- Other models make no assumption in the distribution of the variables, however a better spread of the values may improve their performance



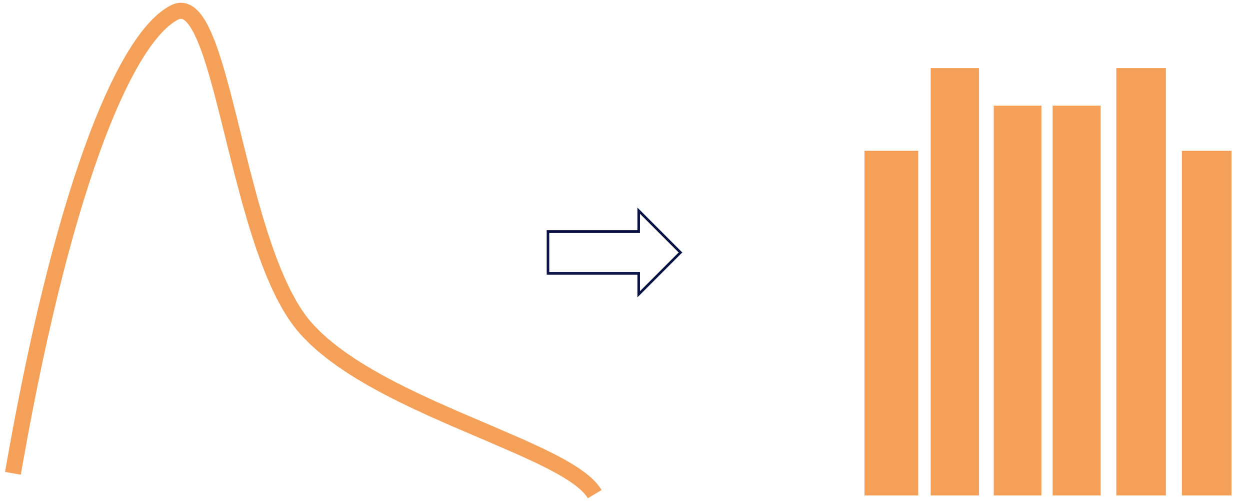
Variable transformation



Variable transformation

- Logarithmic ➔ ln(x)
- Exponential ➔ x Exp (any power)
- Reciprocal ➔ (1 / x)
- Box-Cox ➔ (x Exp (λ) – 1) / λ
  - λ varies from -5 to 5

Variable discretisation



Skewed

Improved value spread

Discretisation

- Equal width bins
  - Bins ➔ (max – min) / n bins
  - Generally does not improve the spread
- Equal frequency bins
  - Bins determined by quantiles
  - Equal number of observations per bin
  - Generally improves spread

Train In Data

THANK YOU

www.trainindata.com