McKinsey&Company

Talking about analytics: Concepts you need to know

McKinsey Analytics

2018

Confidential and proprietary: Any use of this material without specific permission of McKinsey & Company is strictly prohibited

Our focus today

Covered today

1 Identify and validate business need

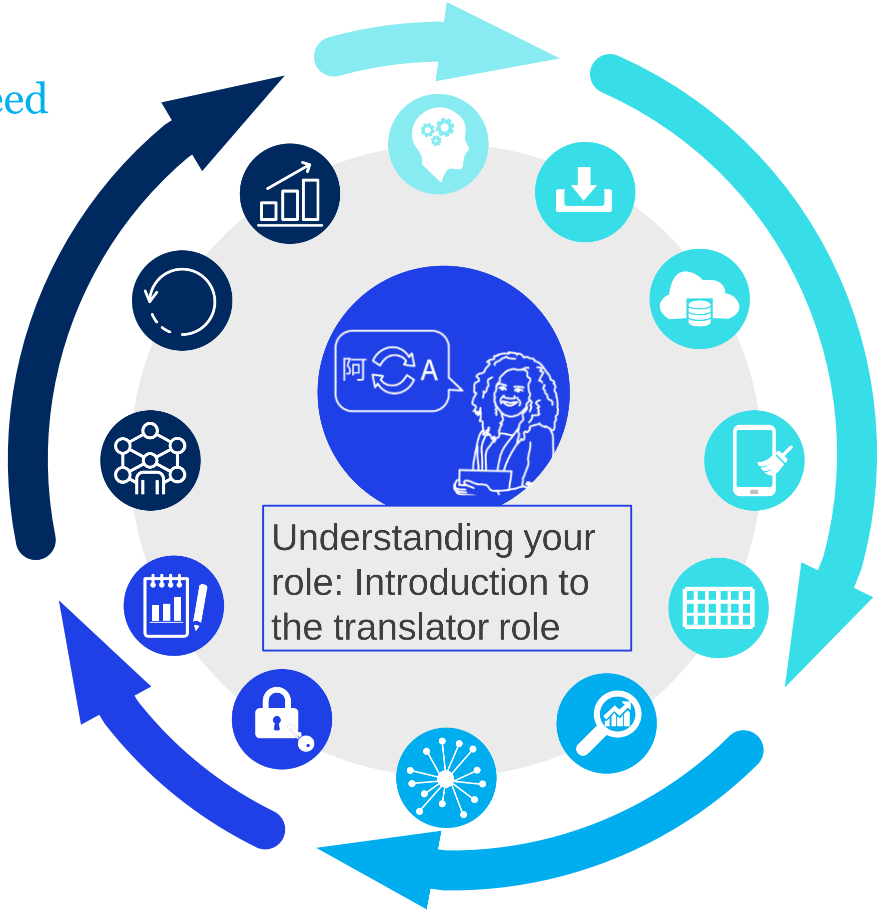
Validating the opportunity: Evaluating use cases

5 Implement and maintain

Bringing it to life: Implementation and sustaining change

4 Validate and derive business implications

Choosing a model: Model performance and business impact



2 Collect, and prepare data

Gathering the right information: Extracting value from data

3 Build the analytical engine

Setting up for success: What to do until you see model output

About our training

What this training is:

- Broad introduction into common analytical concepts
- Provides intuition for basic analytics methods and algorithms

What this training is not:

- In-depth introduction into technical concepts of analytics

- **Does not replace an in-depth conversation with your data scientist** about which methods to use, their pros and cons
- For a **more in-depth introduction, we recommend**

  a reading list at the end of this module

Analytics has many applications

1 http://www.forbes.com/sites/kashmirhill/2012/02/16/how-target-figured-out-a-teen-girl-was-pregnant-before-her-father-did/#12cf258034c6
2 http://observer.com/2016/01/can-we-use-big-data-to-create-hit-tv-shows-as-addictive-as-breaking-bad/
3 https://www.si.com/vault/2014/06/30/106479598/astromatic-baseball-houstons-grand-experiment#

Our focus today

Covered today

1 Identify and validate business need

Validating the opportunity:
Evaluating use cases

2 Collect, and prepare data

Gathering the right information:
Extracting value from data

3 Build the analytical engine

Setting up for success: What to do until you see model output

4 Validate and derive business implications

Choosing a model: Model performance and business impact

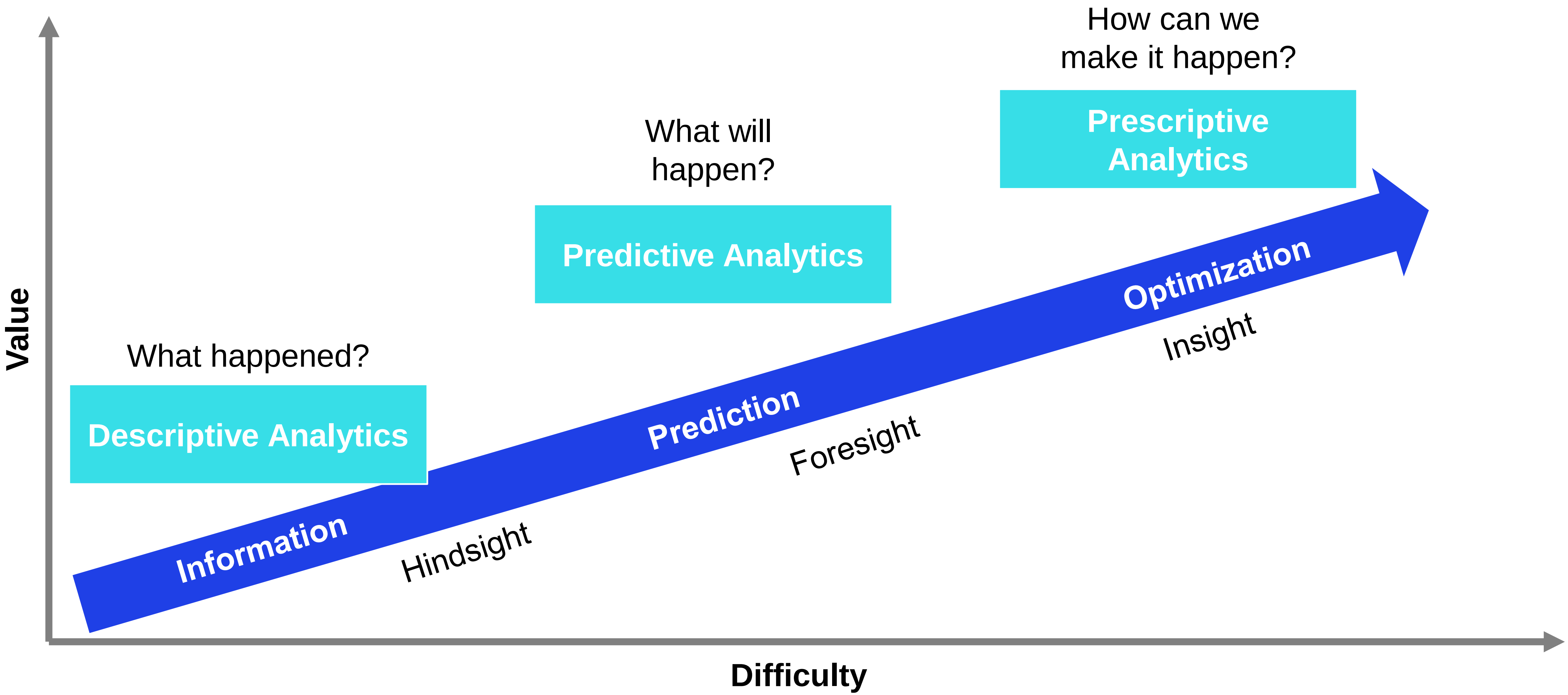
5 Implement and maintain

Bringing it to life: Implementation and sustaining change

Analytics problems and their classification

Analytics problems can be roughly categorized into three problems of varying complexity

Analytics value ladder



Three types of problems being solved with analytics

	Describe	Predict	Prescribe
	What happened in the past and why did it happen?	*What is likely to happen in the future?*	*What should be done to influence the future?*
	Gain insight from historical data	Make predictions about future events	Make decisions to gain an advantage
Type of problem	▪ Concerned with **describing what happened** ▪ Employed heavily across all industries	▪ Extrapolating data to **anticipate behavior** and occurrences (inherently probabilistic) ▪ Used in data-driven organizations as a key source of insight	▪ Principally concerned with **what to do** to achieve goals ▪ Real-time decision making or actionable recommendations and feedback mechanisms
Business questions analytics can answer at this stage	▪ How much did we sell last year? ▪ What is the average spend per customer? ▪ Which supplier is more cost-effective? ▪ Which product has the best profit margins?	▪ What will the oil price be in the next quarter? ▪ Which subscribers are most likely to churn? ▪ Which product will this customer most likely buy?	▪ What is the best price to sell our product in India? ▪ How should we position content on our landing page for this particular customer? ▪ How should we allocate nurses on this shift given admitted patients?

SOURCE: Evans, James R. and Lindner, Carl H. (2012). "Business Analytics: The Next Frontier for Decision Sciences". Decision Line 43 (2)

Many use cases might combine various types of problems to form a solution

	Describe		Predict		Prescribe
Fraud detection	Customer segmentation	+	Classification into high risk vs. low risk of fraud		
Churn detection	Customer segmentation	+	Classification into high vs. low churn risk	+	Optimize churn campaign with respect to budget
Logistics optimization	Describe cost of route segments (incl. wear & tear)		+		Optimize transportation network
Hospital re-admissions			Classify patients into high vs. re-admission risk	+	Identify follow-up inter-ventions in line with budget

Each problem type can be described in more detail…

Descriptive

General exploration — General data analysis including data cuts / pivoting / cross-tabs, time trends, geographic plots, word clouds, etc.

Pattern recognition/ Data mining — Discovering patterns in the data, e.g. similarity among records (clustering) or interdependencies among variables

Inferential statistics — Drawing conclusions about populations based on observed sample

Predictive

Categories (classification) — Predicting binary or ordinal outcomes, e.g. product choice, churn, fraud, purchase events, etc.

Values (continuous) — Predicting continuous outcomes, e.g. revenues, time, growth rates, etc.

Prescriptive

Optimization — Identify set of parameters that optimize system performance under given constraints (often supported by **simulation**)

Recommen-dation — Recommend next best action based on historical pattern and similar customers or items

SOURCE: McKinsey Analytics

… and can be mapped to certain types of analytical techniques that are typically being used

Descriptive

General exploration
- Sample min/max, mean, variance, percentiles
- Simple text analytics
- Network analysis

Pattern recognition/ Data mining
- Clustering (k-means)
- Dimension reduction
- Association Rules

Inferential statistics
- Experimental design
- Hypothesis testing
- Anomaly detection

Predictive

Classification
- Regression (e.g., logistic regression)
- Decision trees
- Random forest
- Support Vector Machines (SVM)
- K-nearest neighbors
- Neural networks

Continuous/ regression
- Regression (linear, ridge, lasso,...)
- Time series analysis
- Support Vector Regression
- Neural networks
- Ensembles

Prescriptive

Optimization
- Continuous: Linear programming
- Discrete: Integer optimization
- Simulation methods:
  - Monte Carlo
  - Agent based modeling

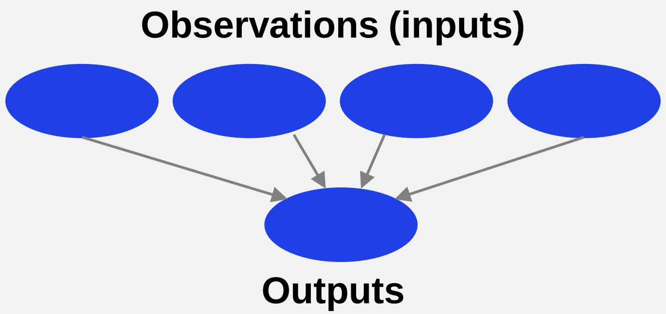
Recommen-dation
- Collaborative Filtering
- Association Rules / Frequent Pattern Mining

NOT MECE – algorithms overlap. And often our business problems are solved by combining different types of algorithms (e.g., first create clusters of clients, then for each cluster predict churn)

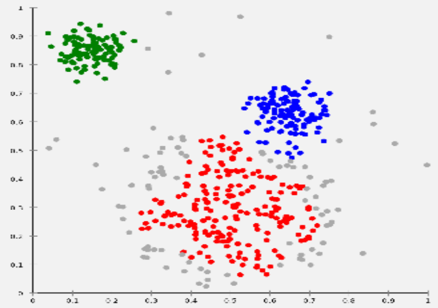
SOURCE: McKinsey Analytics

Two broad categories of techniques: Supervised vs. unsupervised algorithms

Supervised



Unsupervised



When to use it

Supervised: **You know how to classify the data**, but you want the machine to do it for you

Unsupervised: You **do not know how to classify the data**, and you want the machine to find the classifier for you

Business example

Supervised: Determining if a job applicant will be successful based on their application

Unsupervised: Cluster related news articles together for users

How it works

Supervised:

- Person **labels input data** with desired output
- Machine gets **labeled training data** and infers how to label data
- Machine **predicts outputs for new data** based on inferred model

Unsupervised:

- Machine gets **unlabeled** input data
- Machine infers hidden structure from data
- Machine **returns the structure** and classifies data according to that structure

SOURCE: Team analysis

Techniques to tackle analytics problems

Contents

Each problem can be mapped to certain types of analytical techniques that are typically being used

Descriptive

General exploration
- Sample min/max, mean, variance, percentiles
- Simple text analytics
- Network analysis

Pattern recognition/ Data mining
- Clustering (k-means)
- Dimension reduction
- Association Rules

Inferential statistics
- Experimental design
- Hypothesis testing
- Anomaly detection

Predictive

Classification
- Regression (e.g., logistic regression)
- Decision trees
- Random forest
- Support Vector Machines (SVM)
- K-nearest neighbors
- Neural networks

Continuous/ regression
- Regression (linear, ridge, lasso,...)
- Time series analysis
- Support Vector Regression
- Neural networks
- Ensembles

Prescriptive

Optimization
- Continuous: Linear programming
- Discrete: Integer optimization
- Simulation methods:
  - Monte Carlo
  - Agent based modeling

Recommen-dation
- Collaborative Filtering
- Association Rules / Frequent Pattern Mining

SOURCE: McKinsey Analytics

Clustering using k-means

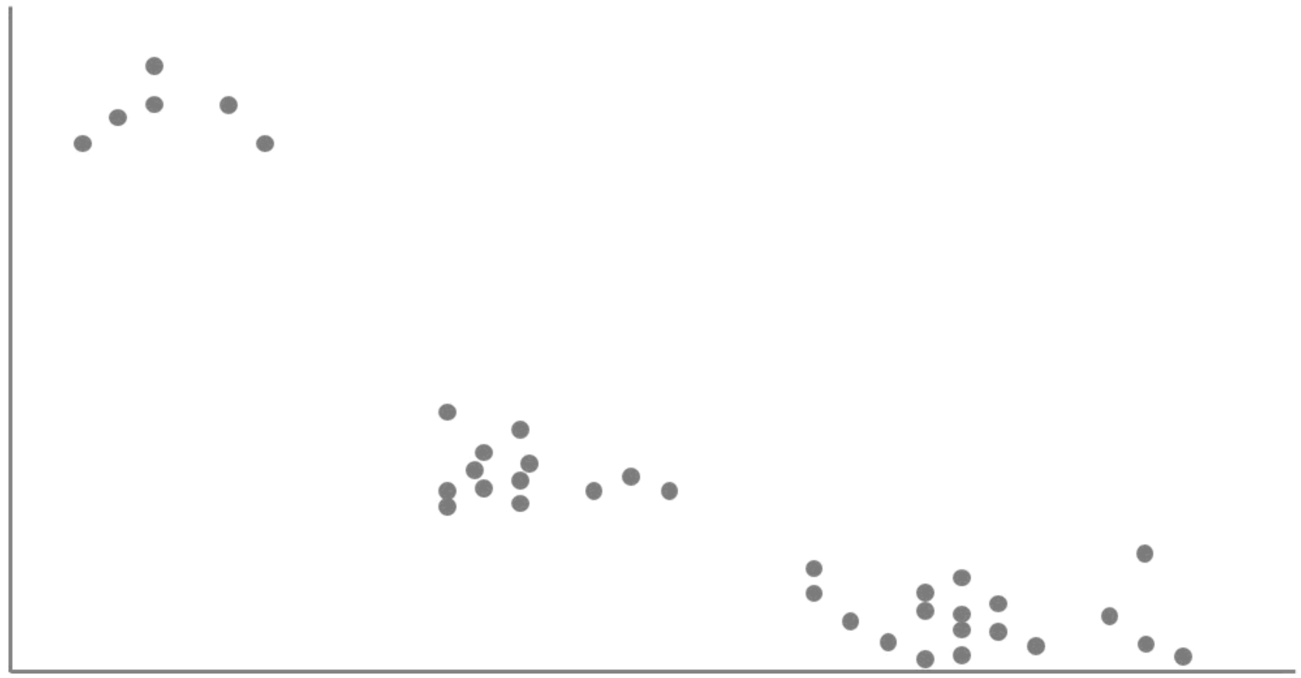
1 Decide on number of clusters (3 clusters in example)

2 Assign centers (centroid) of each cluster randomly ✚✚✚

3 Assign each point to the closest

4 Calculate the new center of the clusters

5 Repeat steps 3 and 4 until centroids do not change



Contents

Each problem can be mapped to certain types of analytical techniques that are typically being used

Descriptive

General exploration
- Sample min/max, mean, variance, percentiles
- Simple text analytics
- Network analysis

Pattern recognition/ Data mining
- Clustering (k-means)
- Dimension reduction
- Association Rules

Inferential statistics
- Experimental design
- Hypothesis testing
- Anomaly detection

Predictive

Classification
- Regression (e.g., logistic regression)
- Decision trees
- Random forest
- Support Vector Machines (SVM)
- K-nearest neighbors
- Neural networks

Continuous/ regression
- Regression (linear, ridge, lasso,...)
- Time series analysis
- Support Vector Regression
- Neural networks
- Ensembles

Prescriptive

Optimization
- Continuous: Linear programming
- Discrete: Integer optimization
- Simulation methods:
  - Monte Carlo
  - Agent based modeling

Recommen-dation
- Collaborative Filtering
- Association Rules / Frequent Pattern Mining

SOURCE: McKinsey Analytics

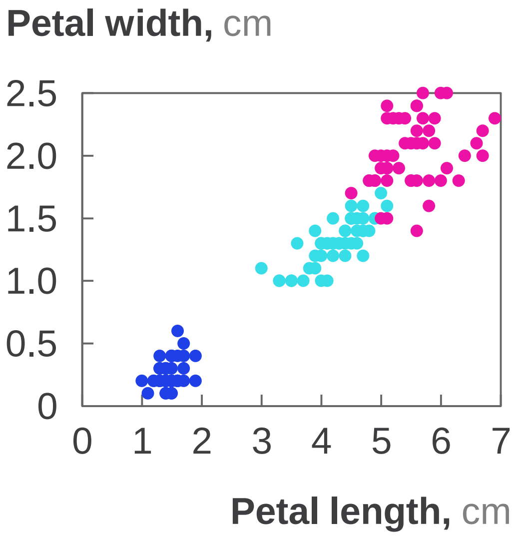
Let's work through a couple of examples using two types of data

Classification example: Iris data set

- Data set consisting of 50 samples from each of three species of Iris flowers

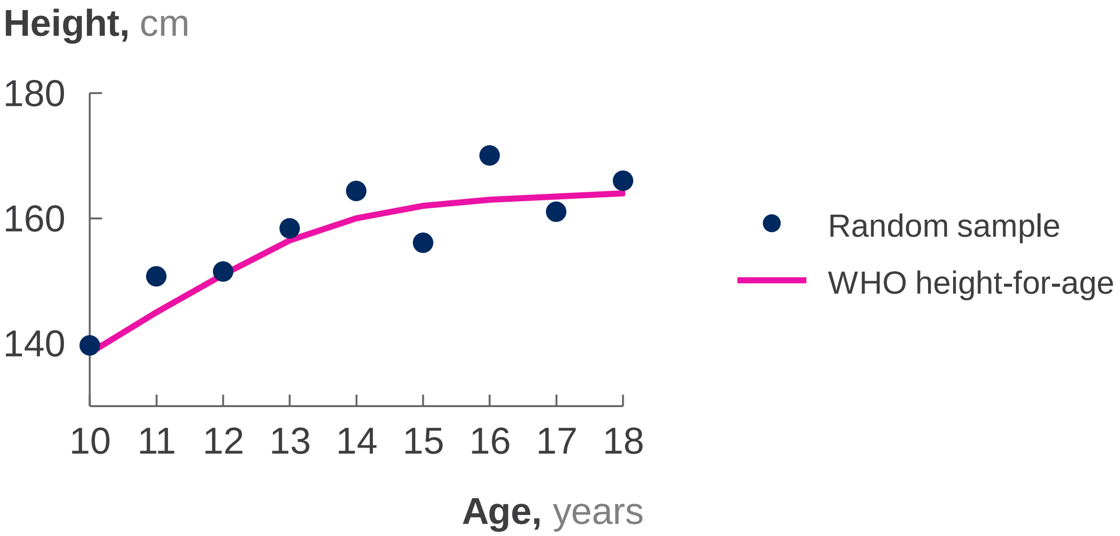
Iris setosa — Iris versicolor — Iris virginica

- Four feature measured for each sample
  - Petal length
  - Petal width
  - Sepal length
  - Sepal width



Continuous example: Height / Age of girls

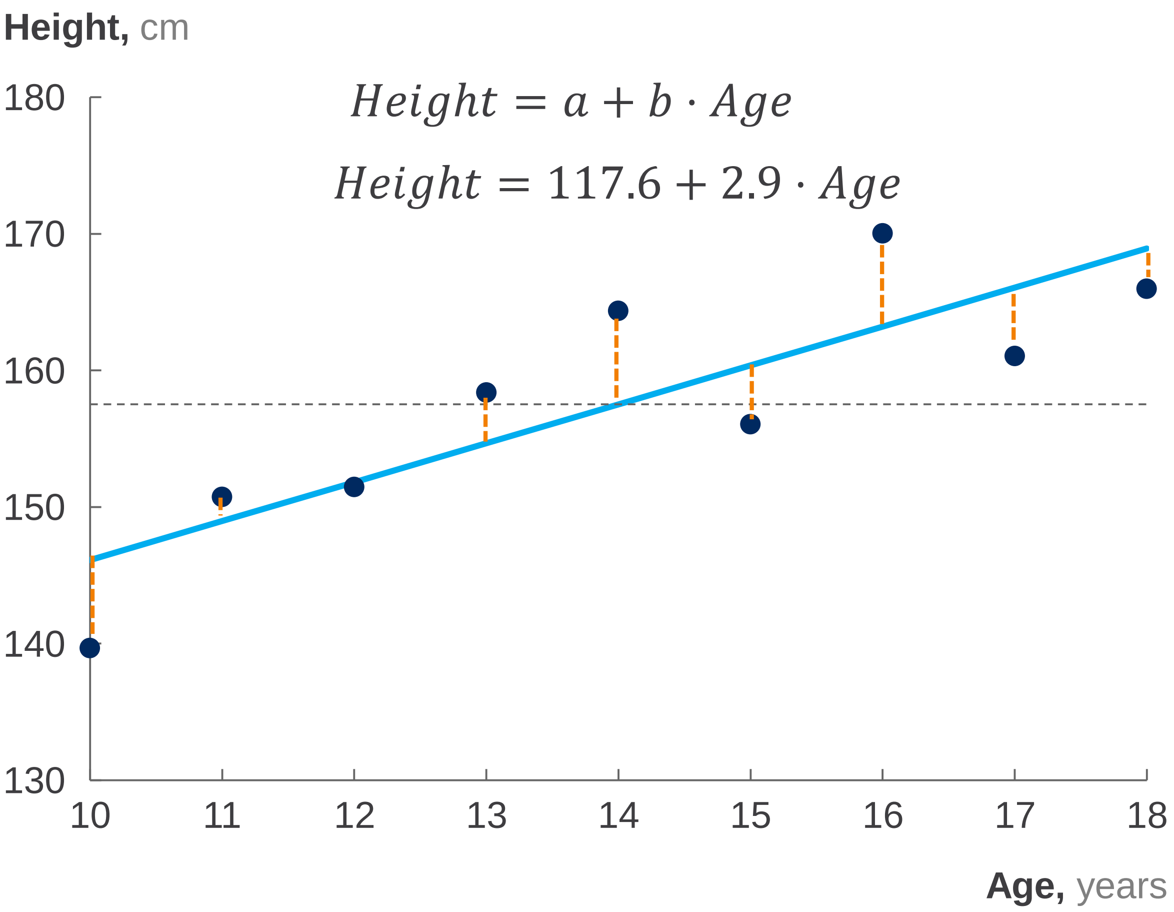
- WHO publishes height-for-age scores for girls between the ages of 5 and 18
- We created a random sample for girls between 10 and 18 with one observation for each age group



Note: For illustration purposes, our examples are shown using only 1 or 2 input variables, in reality a lot more are being used

Linear regressions (most commonly used to predict continuous outcomes) fit a line (or simple functional form) to the data

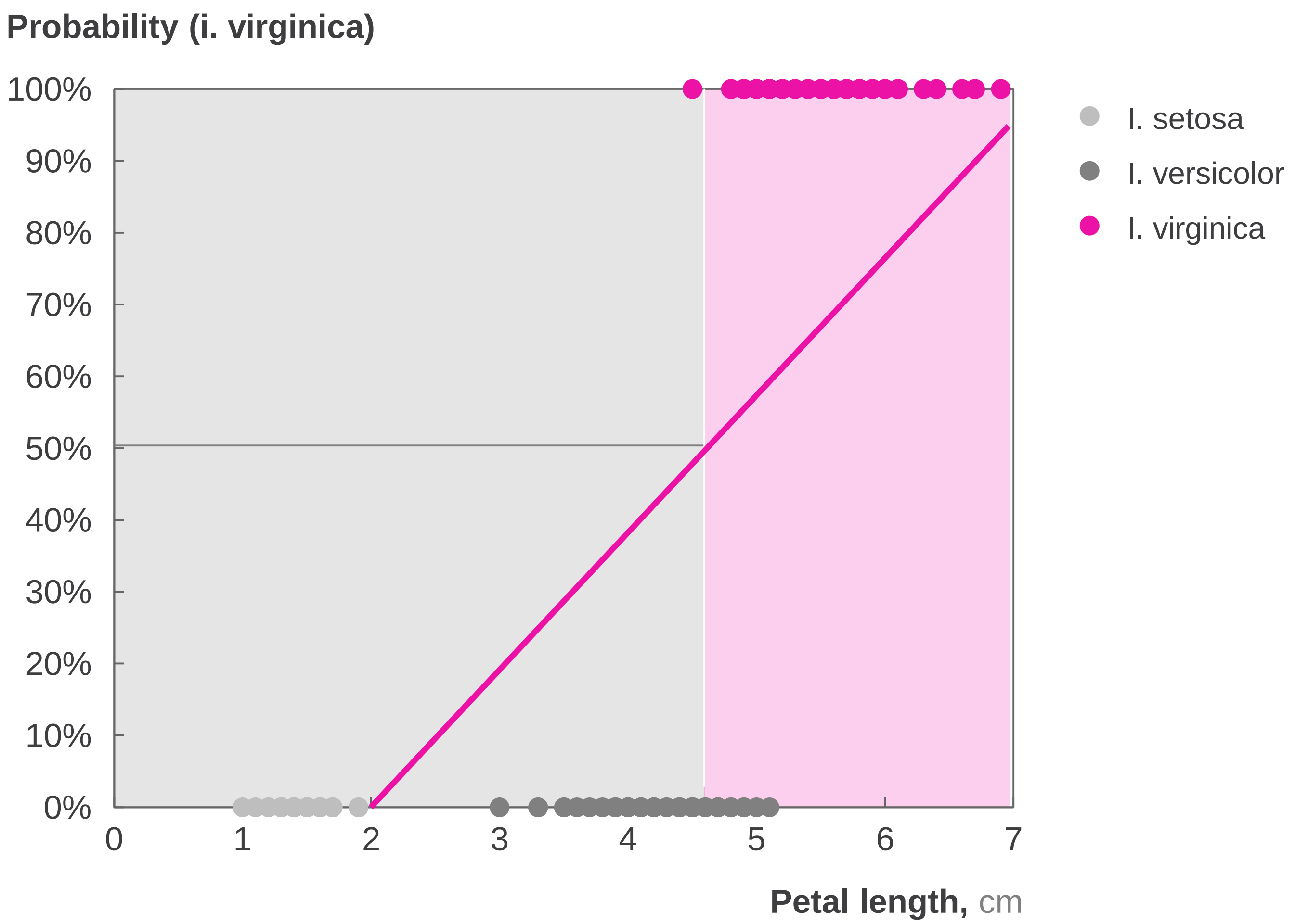
Example: Estimate the functional form of the WHO height-for-age curve for girls (age 10-18)



- Linear regressions **minimize the prediction error** between prediction (regression line) and the observations (dots)
- Coefficients of a linear regression have an **intuitive interpretation**:
  - On average, girls between 10 and 18 years grow about 2.9 cm every year
  - What is the interpretation of the intercept 117.6?
- **More complex functional forms** can be achieved by adding quadratic terms (or higher order polynomials) or interaction terms between variables

When applying a regression technique to a (binary) classification example, we are using the technique to predict 0s and 1s

Example: Estimate the probability of a flower being of “**virginica**” type based on petal length



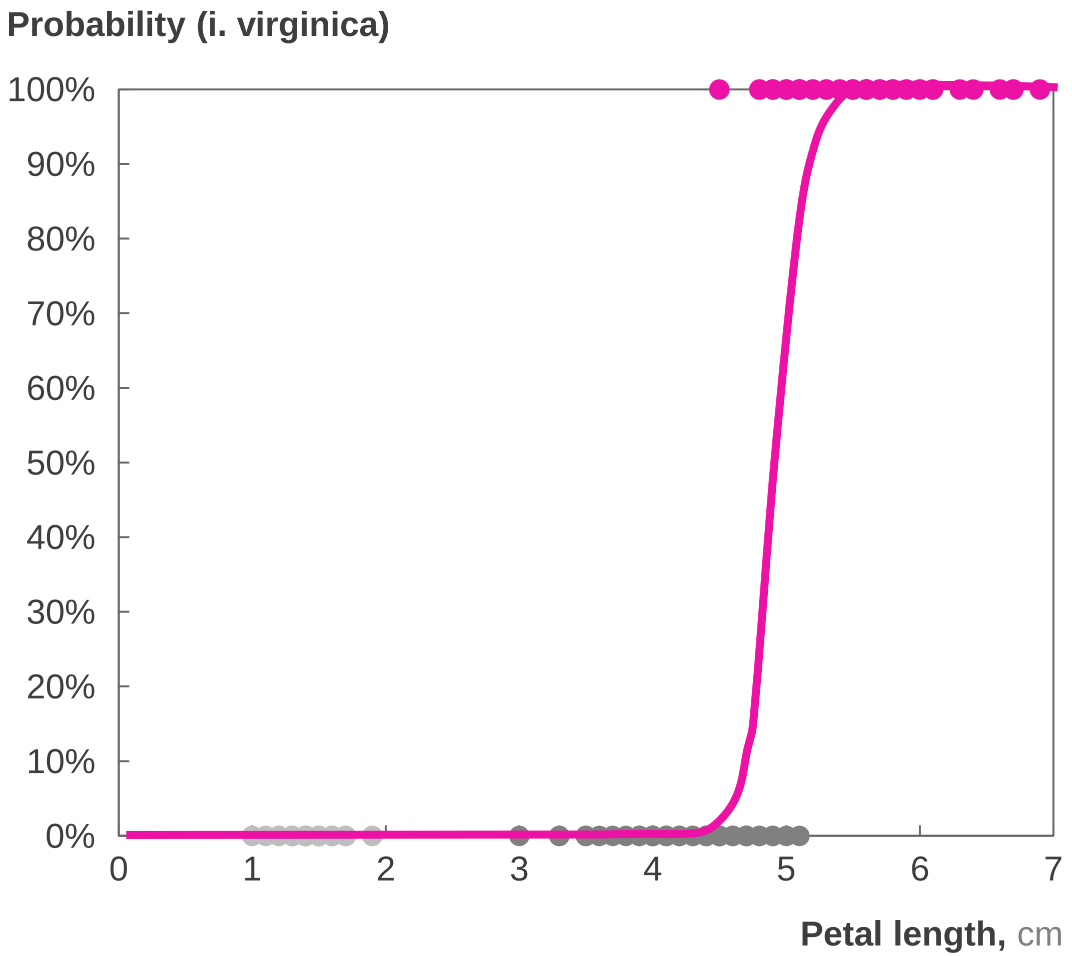
1. Create an outcome variable: probability of being “virginica” (1 for virginica, 0 else)
2. Estimate the linear relationship between petal length (and possibly other variables) and the outcome

$$\Pr(virginica) = -0.39 + 0.19 \cdot PetalLength$$

3. Choose a threshold for when an a prediction is classified as virginica vs. not (typically 50%)
4. When predicted probability is above 50% (here for a petal length >4.7), classify as virginica
5. Else classify as “other”

Logistic regressions use an s-shaped functional form with values between 0 and 1

Example: Estimate the probability of a flower being of "**virginica**" type based on petal length



- The approach to fit a logistic regression is similar to the linear regression, except that the **linear form is transformed into an s-shape using a sigmoid function**

$$\Pr(virginica) = sigmoid(a + b \cdot PetalLength)$$

$$\Pr(virginica) = \frac{1}{1 + e^{-(a+b \cdot PetalLength)}}$$

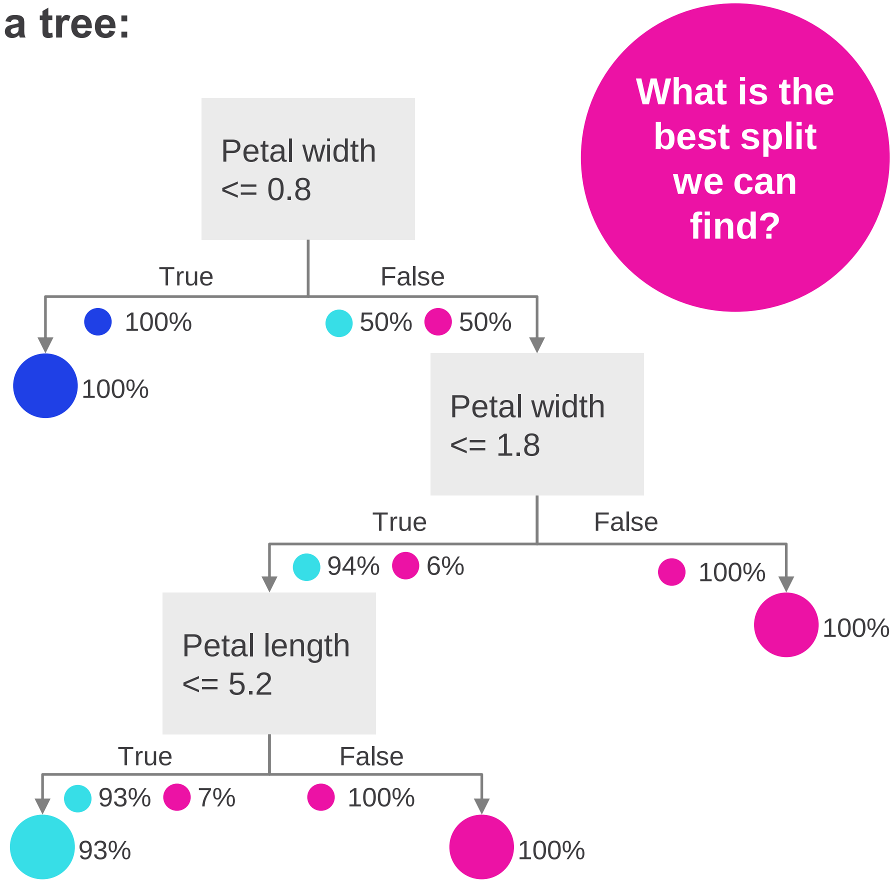
$$\Pr(virginica) = \frac{1}{1 + e^{-(43.8+9.0 \cdot PetalLength)}}$$

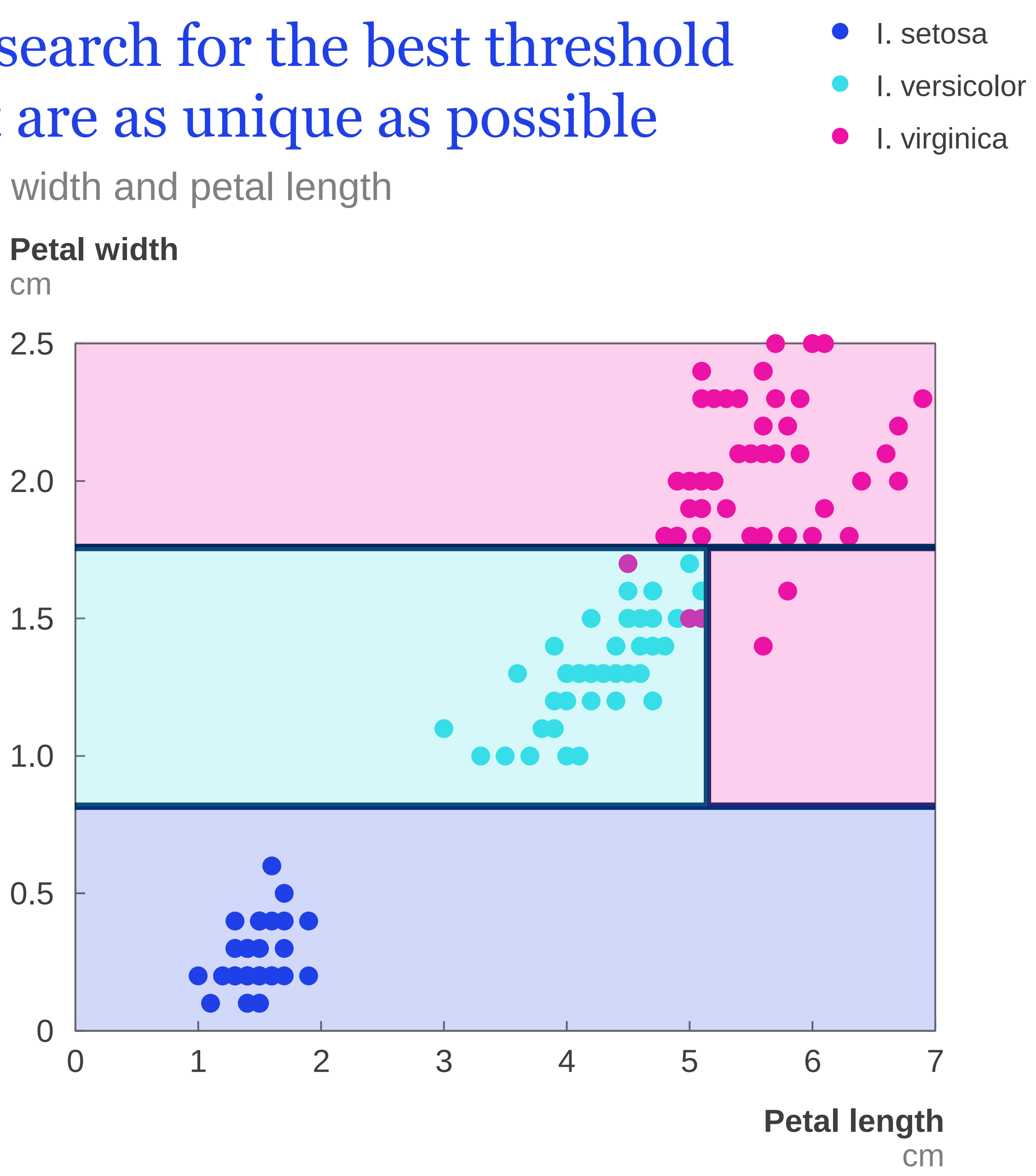
- With an s-shaped form, coefficients have a less intuitive interpretation, e.g.
  - An increase in Petal Length from 1 cm to 2 cm changes probability by close to zero
  - An increase in Petal length from 4cm to 5 cm changes the probability by 0.78

Decision trees: At each node, decision trees search for the best threshold to divide the data in order to get groups that are as unique as possible

Example: Classifying iris flowers based on information about petal width and petal length

Let's build a tree:





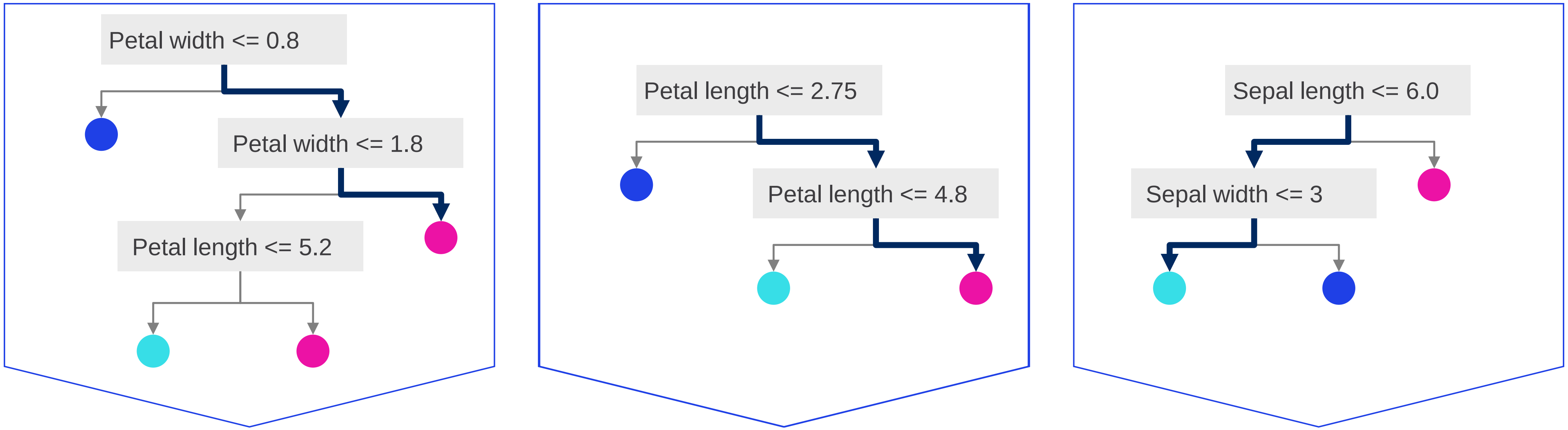
Random forests are a collection of resampled and restricted decision trees that vote on how to classify a single observation

- I. setosa
- I. versicolor
- I. virginica

Example: Let's randomly select rows and features, train a tree for each scenario, then let trees take a vote

	Tree 1	Tree 2	Tree 3
Input rows	**10,000 random rows**	**10,000 random rows**	**10,000 random rows**
Input features	Petal length, Petal width, Sepal length, Sepal width	Petal length, Petal width, Sepal length, Sepal width	Petal length, Petal width, Sepal length, Sepal width

Tree mechanism



Example

Petal length	4.9
Petal width	2.0
Sepal length	5.6
Sepal width	2.8

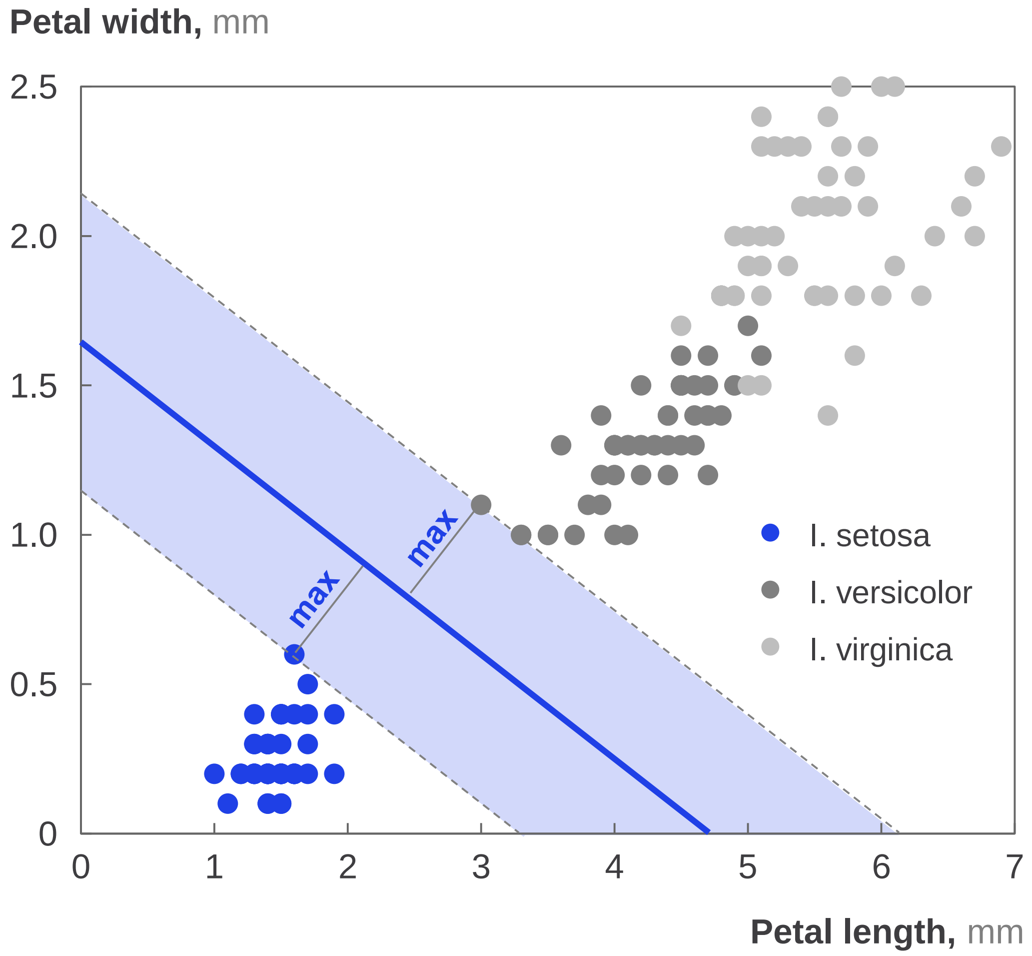
Votes

Tree 1: I. virginica
Tree 2: I. virginica
Tree 3: I. versicolor

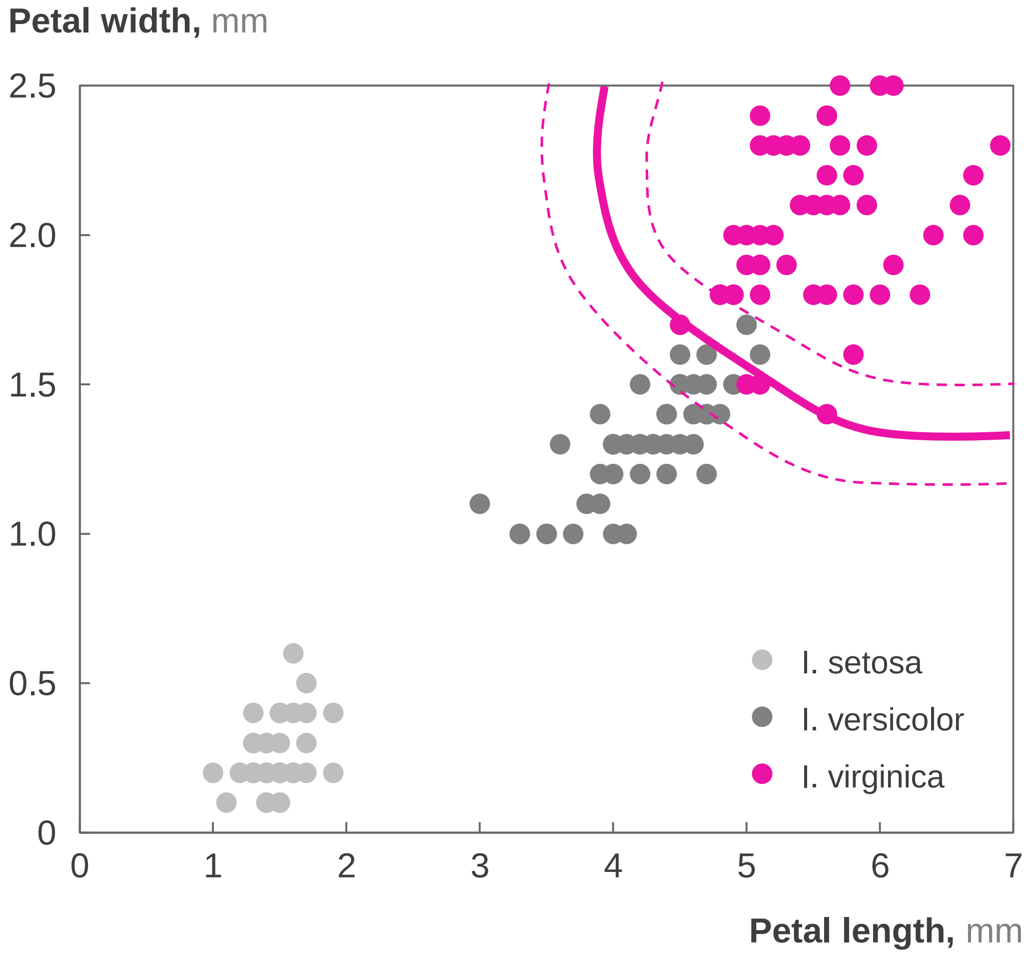
66% probability: I. virginica
33% probability: I. versicolor

Support vector machines separate one class from another by searching for the boundary ("hyperplane") that maximizes the distance to the closest point(s)

For simple classification tasks, a simple line is sufficient to separate classes



For more difficult problems, more complex functional forms can be assumed (e.g., polynomials, radial)

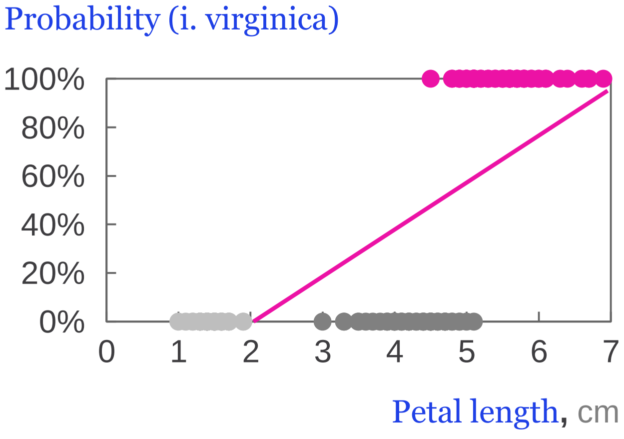


Recap: Different models “see” the world in different ways

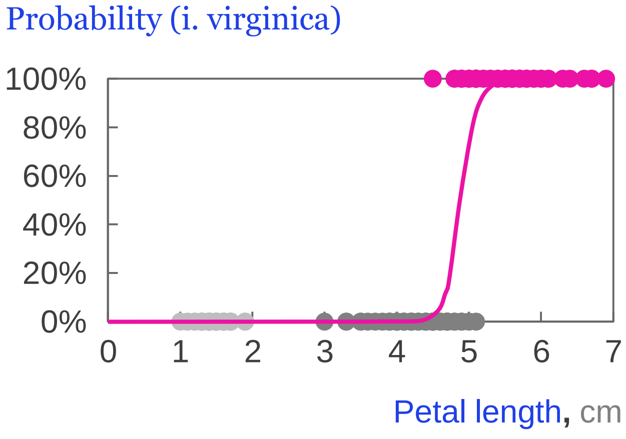
● I. setosa ● I. versicolor ● I. virginica

Example: Distinguishing **i. virginica** from the other two iris types (more details on following pages)

Linear regression

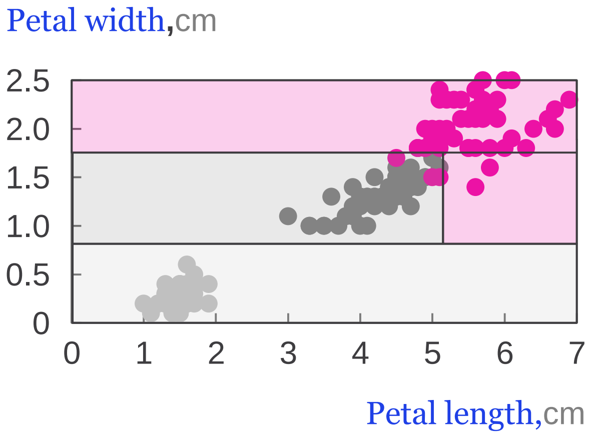


Logistic regression

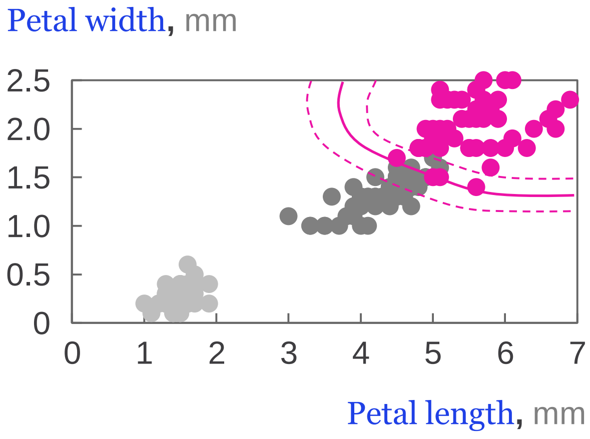


- Focus on **outcome variables**, i.e. finding functional forms by minimizing the prediction error

Decision trees



Support Vector Machines



- Focus on **input features**, i.e. select classes by identifying patterns in the input data

SOURCE: McKinsey Analytics

Be careful with implying causal relationships from regressions (or other machine learning algorithms): Telco churn case

ILLUSTRATIVE

A **linear regression** to estimate the likelihood of churn in TelCo yields the following equation:

$$\Pr(churn) = 0.05 + \boxed{0.20\ campaign\ month} + \boxed{0.13\ WhatsApp} + \boxed{0.05\ Network\ failures}$$

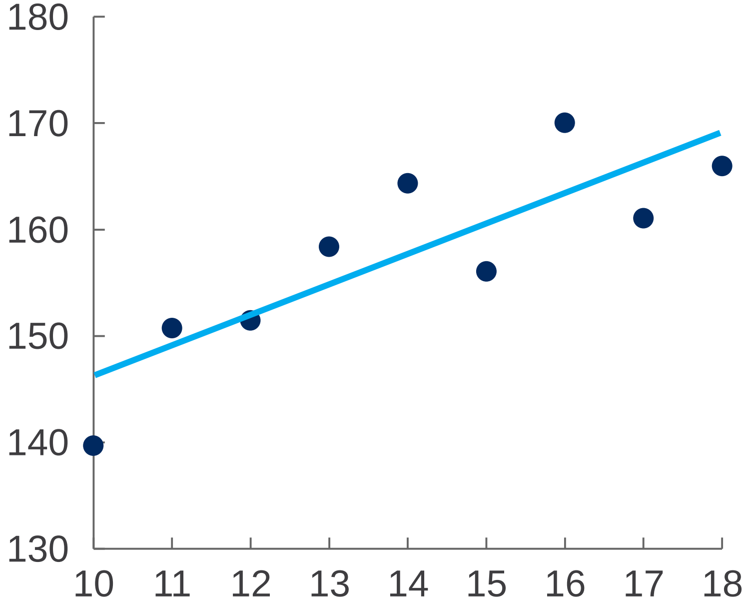
	$0.20\ campaign\ month$	$0.13\ WhatsApp$	$0.05\ Network\ failures$
Technical interpretation	**Running churn campaigns in a given month** is associated with a 20 percentage point *higher* probability of churn	**Increased usage of Whatsapp for calls** is associated with a 13 pp higher churn probability	**Each network failure** in the area of the customer is associated with 0.05 higher probability of churn
Concepts to watch out for when interpreting coefficients (possible alternative explanation)	**Reverse causation:** Higher overall churn leads sales directors to increase the number of churn campaigns	**Common factors or hidden variables** Higher prices for calling increase WhatsApp usage AND churn	**Bidirectional relationship** Higher churn reduces network utilization and thus decreases network failures whereas network failures in turn increase churn

The more complex your model gets, the better it can fit to the data…

Let’s assume we are trying to find the functional form that describes the relationship between Age and Height (cm) for girls between 10 and 18 years of age. We have a sample data set of 9 girls with one observation for each year.

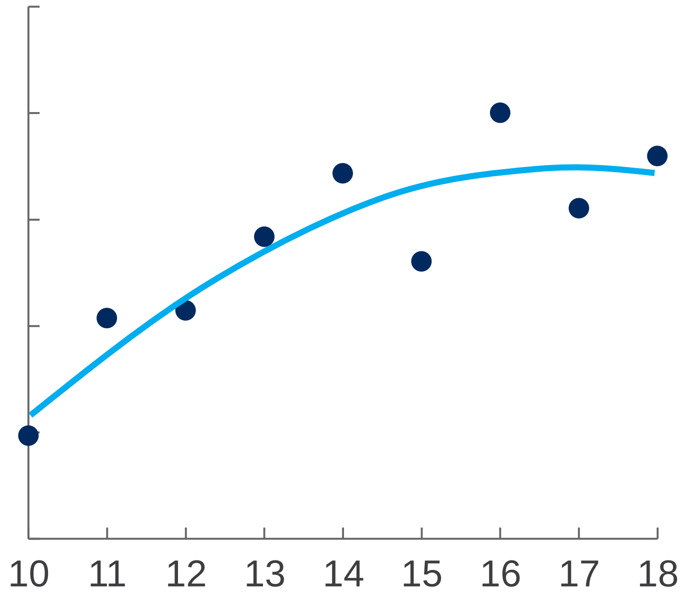
Simple linear regression

$$Height = a + b \cdot Age$$



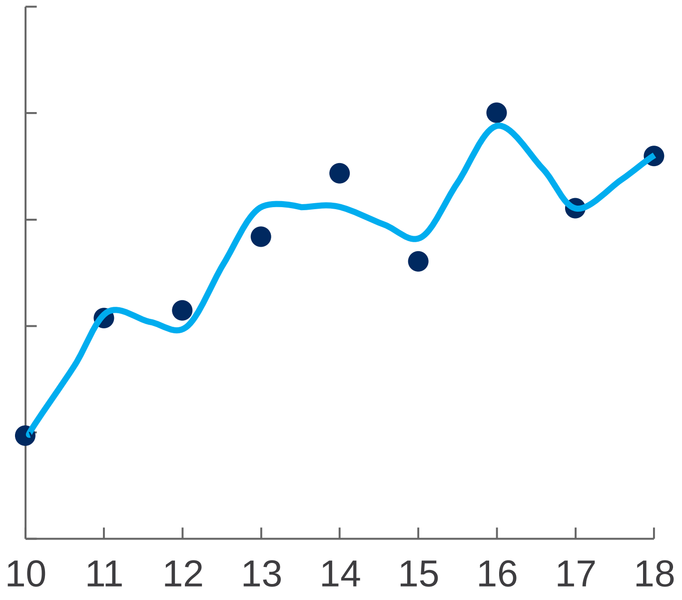
… adding a quadratic term…

$$Height = a + b_1 \cdot Age + b_2 \cdot Age^2$$



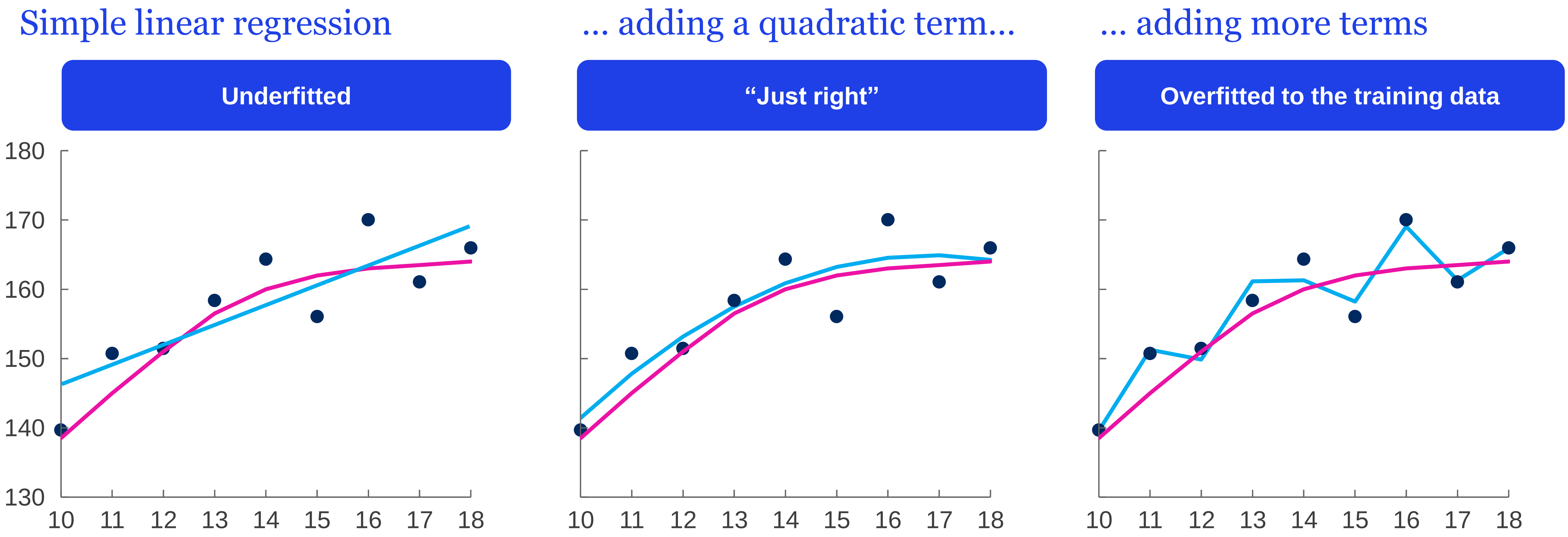
… adding more terms

$$Height \\ = a + b_1 \cdot Age + b_2 \cdot Age^2 + \cdots + b_9 \cdot Age^9$$



… however, the less likely it is to overfit to the sample that you are using to a point where it does not reflect “reality” anymore

Let’s add the **actual functional form** according to WHO size charts…

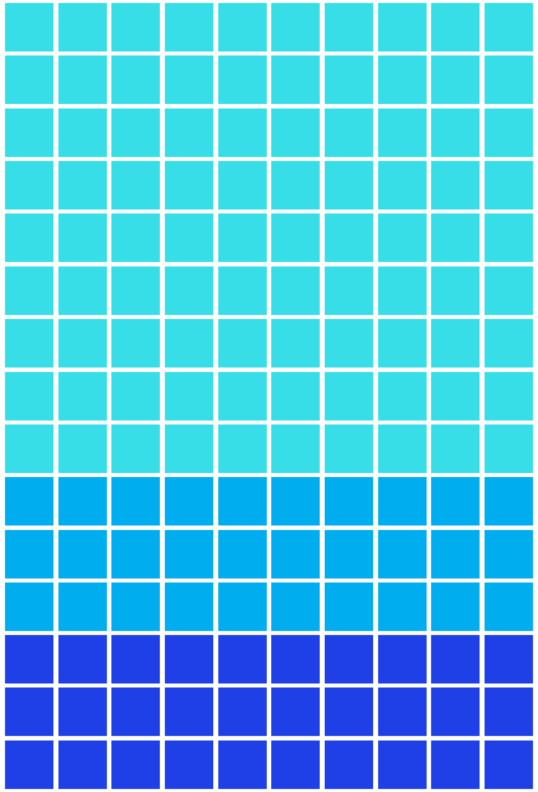


One method to avoid overfitting is to split your data into training sets, test sets and validation sets

Basic idea:

Hold back some of your data to check how the model performs on a new data set.

Thus you split you split your data into three parts (randomly!):



Training set

- Used to estimate the model parameters (e.g. weights, coefficients, decision tree nodes, …)

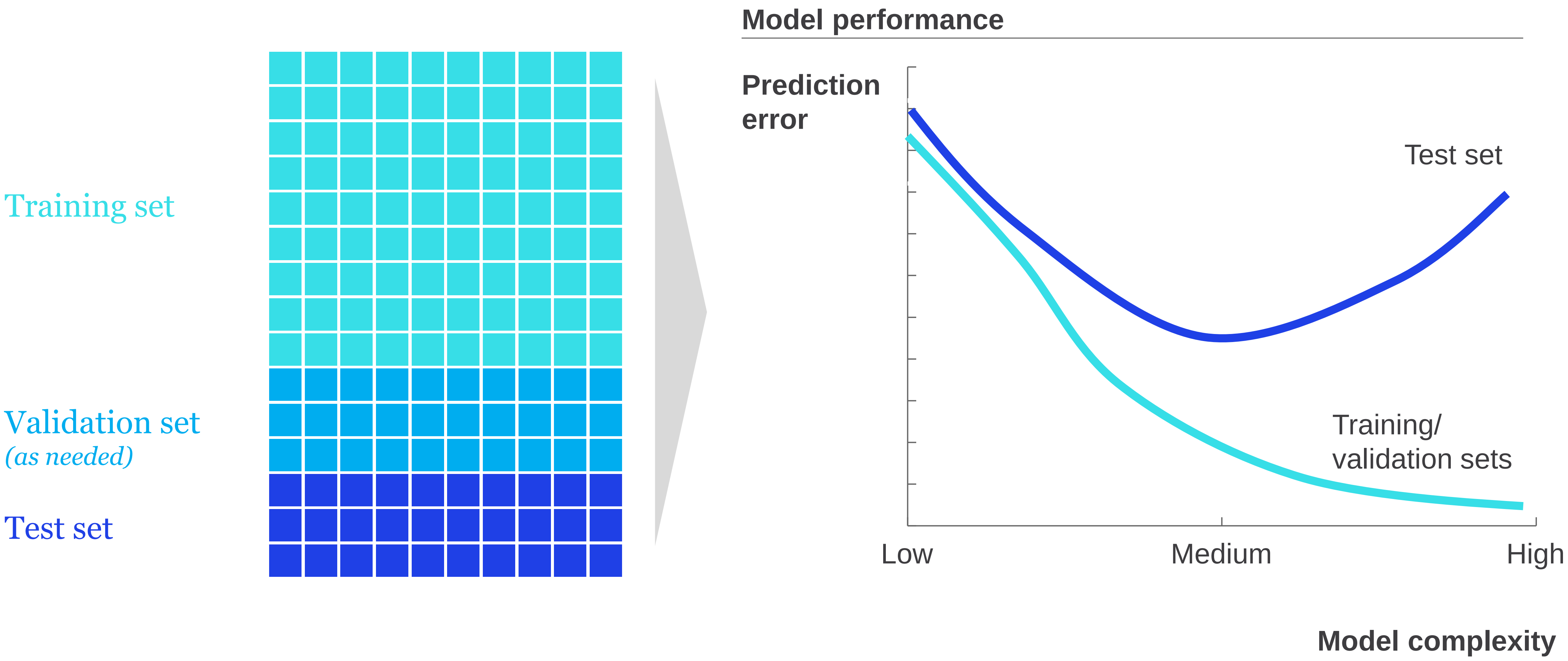
Validation set

- Used to tune hyperparameters (e.g., how many nodes in a decision tree, how many trees in a forest)

Test set

- Used to test how the model performs on a new (out of sample) data set
- If a model performs well on the test set, chances of overfitting are low

The pattern we observed in our previous example is very common: More complex models fit better to the training data, but less well to the test data



Taking the train/test pattern to the next level: Cross-validation

Training set

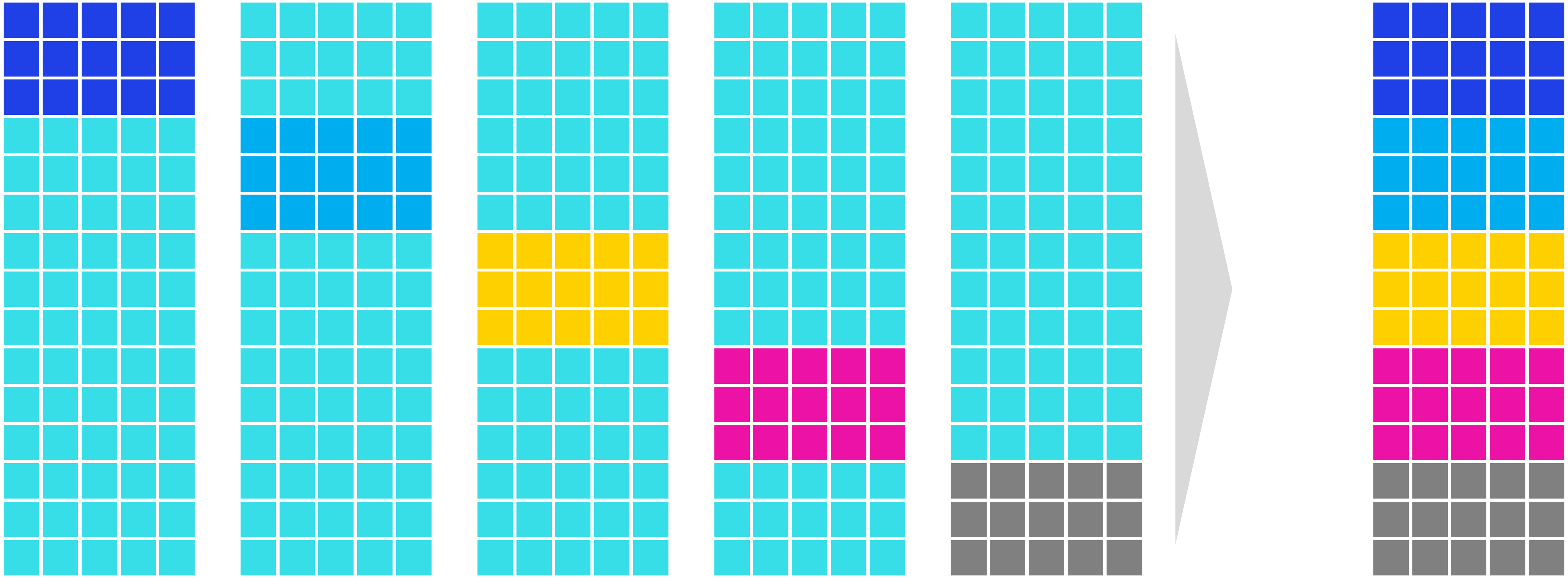
Test set

Basic idea:

- You randomly split all your data 5 times (or more) into train and test sets (“folds”). Each time starting with the full data set
- You then train the model on the 5 different training sets and observe test set performance of each model
- Choose the one with the best performance across all 5 or more runs

Split 1 Split 2 Split 3 Split 4 Split 5

Model performance on each test set



Contents

Each problem can be mapped to certain types of analytical techniques that are typically being used

Descriptive

General exploration
- Sample min/max, mean, variance, percentiles
- Simple text analytics
- Network analysis

Pattern recognition/ Data mining
- Clustering (k-means)
- Dimension reduction
- Association Rules

Inferential statistics
- Experimental design
- Hypothesis testing
- Anomaly detection

Predictive

Classification
- Regression (e.g., logistic regression)
- Decision trees
- Random forest
- Support Vector Machines (SVM)
- K-nearest neighbors
- Neural networks

Continuous/ regression
- Regression (linear, ridge, lasso,...)
- Time series analysis
- Support Vector Regression
- Neural networks
- Ensembles

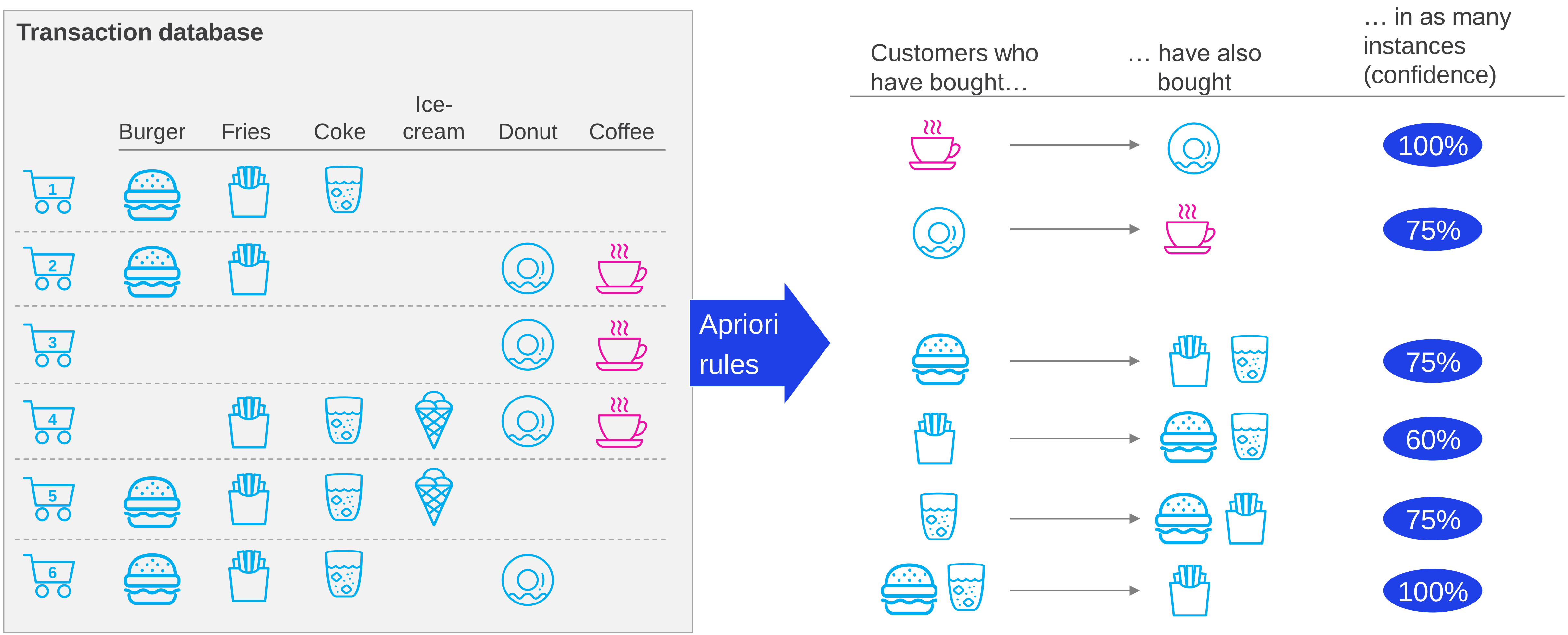
Prescriptive

Optimization
- Continuous: Linear programming
- Discrete: Integer optimization
- Simulation methods:
  - Monte Carlo
  - Agent based modeling

Recommen-dation
- Collaborative Filtering
- Association Rules / Frequent Pattern Mining

SOURCE: McKinsey Analytics

Association rules: the apriori algorithm helps to identify which items often are purchased together



Applications of apriori include product promotions, shelf optimization, next-product-to-buy

Our focus today

Covered today

1 Identify and validate business need

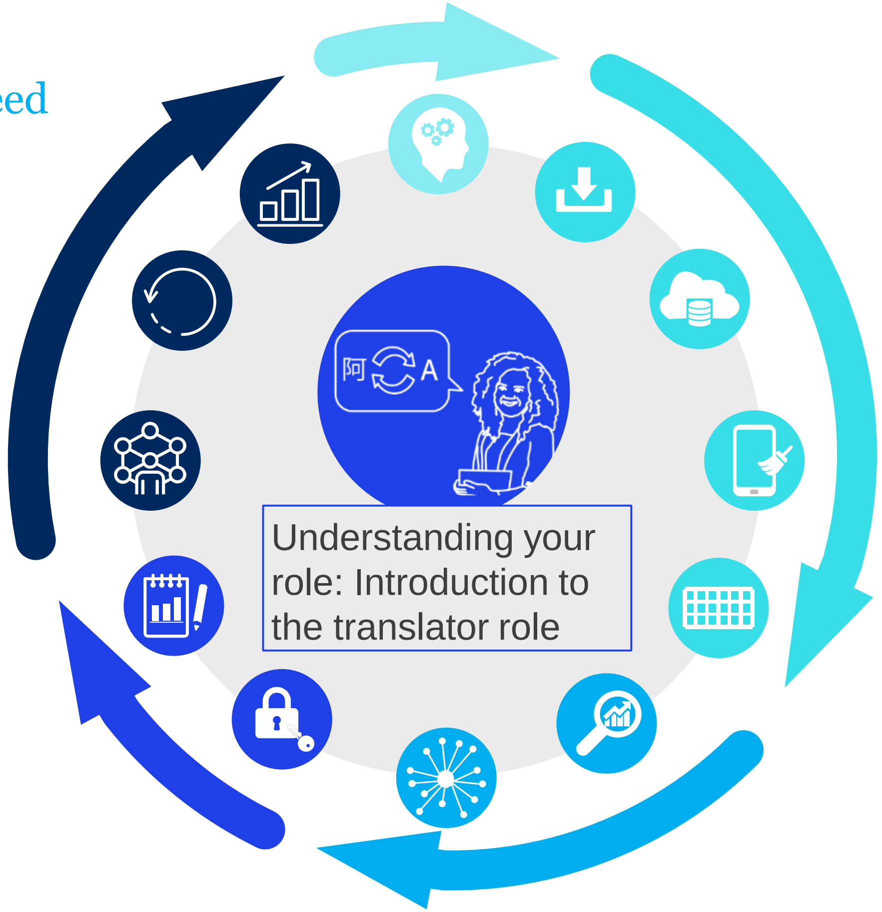
Validating the opportunity: Evaluating use cases

5 Implement and maintain

Bringing it to life: Implementation and sustaining change

4 Validate and derive business implications

Choosing a model: Model performance and business impact



2 Collect, and prepare data

Gathering the right information: Extracting value from data

3 Build the analytical engine

Setting up for success: What to do until you see model output

Choosing the right analytics technique(s) and features requires both business and modelling knowledge

Business Knowledge → **Model Understanding**

Implementation

What will it take to implement across the organization? What mindsets have to shift?

Underlying process

How well does the model characterize the underlying processes that generated the data?

Scalability

Will the model work in practice, e.g. will it be able to deal with exceptions that may occur in a larger data set, different geographies,…?

Business constraints

How do business constraints (e.g., budget) influence model usage?

Interpretability/ Actionability

Is the model easily interpreted? Can clients understand how it works? Is it a 'black-box' solution?

Integration

Can the model easily be integrated into current systems/ workflows or is there a plan to build new systems/ workflows?

Input sensitivity

How well does the model perform on important subsets of the population?

Flexibility

How flexible is the model to change? How does it handle large shifts/ shocks, new information or missing cases?

Model performance metrics

How well does the model predict the target variable/ describe the population / provide recommendations?

Each type of analytics problem comes with a different set of performance metrics that can be used to evaluate the technical performance of a model

Supervised Learning

Classification - binary

Focus today

- Confusion matrix derived measures
  - Accuracy / error rate
  - True/false positive rate
  - True/false negative rate
  - Receiver Operating Characteristic (ROC) curve, Area under the curve (AUC), Gini
- Precision / recall curves, PRAUC (Area under the precision / recall curve)
- Logarithmic Loss
- Brier Score, Hosmer–Lemeshow test (“decile testing”), etc.
- Cross-entropy
- Lift / Gain charts

Classification – multi class

- F1 score
- Matthews Correlation Coefficient
- Cohen Kappa Score
- Macro-*,Micro-*,Weighted-

Continuous / Regression

- Mean squared error (MSE), root mean squared error (RMSE)
- R-Squared, adjusted R-squared
- Mean absolute error (MAE)
- Median absolute error
- Mean Absolute Percentage Error (MAPE)
- Mean Error (ME)

Unsupervised Learning

- Calinski-Harabasz
- Silhouette
- Stability of model output (e.g., for different sub samples)
- Cross-validation likelihood / methods
- Penalty methods (e.g., favoring less complex models)
- A/B Testing on sub-samples
- Simulations based on historical data or re-sampling (“Monte Carlo”)

Not a MECE list, and; please consult with your data scientist for each new model in development

Performance metrics derived from the confusion matrix

In prediction problems, a lot of metrics stem from the confusion matrix – a concept showing correctly vs. incorrectly classified observations

		Prediction	
		Positive	**Negative**
Actual	**Positive**	**True Positive (TP)** Predicted positive, actual positive	**False Negative (FN)** **Type II error** Predicted negative, actual positive
	Negative	**False Positive (FP)** **Type I error** Predicted positive, actual negative	**True Negative (TN)** Predicted negative, actual negative

Important metrics derived from the confusion matrix

		Prediction	
		Positive	**Negative**
Actual	**Positive**	True Positive (TP)	False Negative (FN) Type II error
	Negative	False Positive (FP) Type I error	True Negative (TN)

Accuracy
How often is the model correct?

$$= \frac{\text{TP + TN}}{\text{\# Observations}}$$

Misclassification rate
How often is the model incorrect?

$$= \frac{\text{FP + FN}}{\text{\# Observations}}$$

True positive rate
What share of **positives** has been **correctly** classified?
Also called: Recall, Specificity

$$= \frac{\text{TP}}{\text{Actual positives}}$$

False positive rate
What share of **negatives** has been **incorrectly** classified as positive?

$$= \frac{\text{FP}}{\text{Actual negatives}}$$

Precision
What share of **positive predictions** are actually positive?

$$= \frac{\text{TP}}{\text{Predicted pos.}}$$

Prevalence
What share of our sample is positive (e.g. default rate, churn rate)?

$$= \frac{\text{TP}}{\text{\# Observations}}$$

But which metric should we use?

Misclassification rate

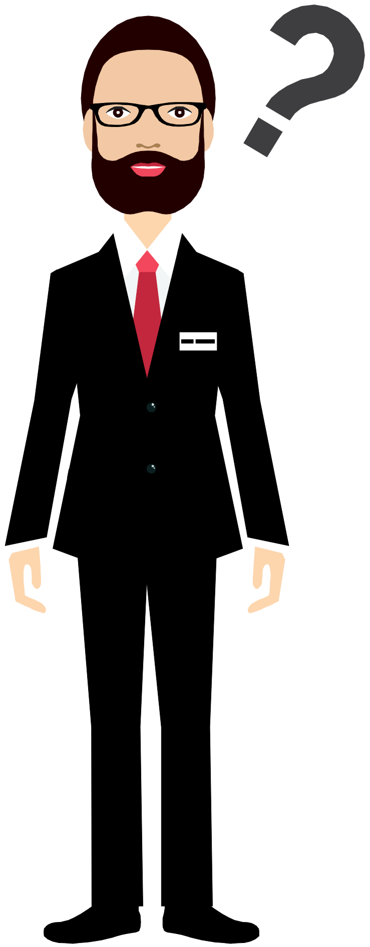
False positive rate

Precision

Accuracy

True positive rate

Prevalence



It depends…

Relevant measures depend on the specific business context

	Pred. Positive	Pred. Negative
Actual Positive	True Positive	False Negative Type 2 error
Actual Negative	False Positive Type 1 error	True Negative

Which error to minimize if we are predicting…

…responses to marketing efforts?

…breast cancer occurrence?

…which bags carry bombs?

How do we interpret the confusion matrix in the context of loan defaults?

Confusion matrix

	Predict default - do not give a loan	**Predict no default** - give out loan
Actual Positive - defaulted loan	**Prevented loss**	**Costly prediction error**
Actual Negative - loan paid back	**Lost business!**	**Good business**

What error would you optimize for if your business strategy was…

- Maximize the return on investment on new loans?
- Maximize total profits?
- Grow the business to capture market share

Most models predict *probability scores* rather than a yes/no decision – to create a confusion matrix a probability threshold needs to be chosen

Many models predict a probability score Pr(Positive)

Observation	Actual	Score
1	Positive	0.99
2	Positive	0.98
3	Negative	0.96
4	Negative	0.9
5	Positive	0.88
6	Negative	0.87
7	Positive	0.85
8	Positive	0.8
9	Negative	0.7
10	Negative	0.6
11	Positive	0.5
12	Negative	0.45
13	Positive	0.43
14	Negative	0.42
15	Negative	0.28
16	Negative	0.27
17	Negative	0.23
18	Negative	0.1
19	Positive	0.05
20	Negative	0.03

Predict positive

Predict negative

The confusion matrix is then built by picking a threshold above which an observation is predicted as positive…

… and counting observations

		Prediction	
		Positive	**Negative**
Actual	**Positive**	6	2
	Negative	5	7

Thus, the same model can yield very different confusion matrices, depending on the chosen threshold

Many models predict a probability score Pr(Positive)

Observation	Actual	Score
1	Positive	0.99
2	Positive	0.98
3	Negative	0.96
4	Negative	0.9
5	Positive	0.88
6	Negative	0.87
7	Positive	0.85
8	Positive	0.8
9	Negative	0.7
10	Negative	0.6
11	Positive	0.5
12	Negative	0.45
13	Positive	0.43
14	Negative	0.42
15	Negative	0.28
16	Negative	0.27
17	Negative	0.23
18	Negative	0.1
19	Positive	0.05
20	Negative	0.03

Threshold: Pr(Positive) ≥ 0.8

5	3
3	9

False positive rate	25%
True positive rate	63%
Accuracy	70%

Threshold: Pr(Positive) ≥ 0.5

6	2
5	7

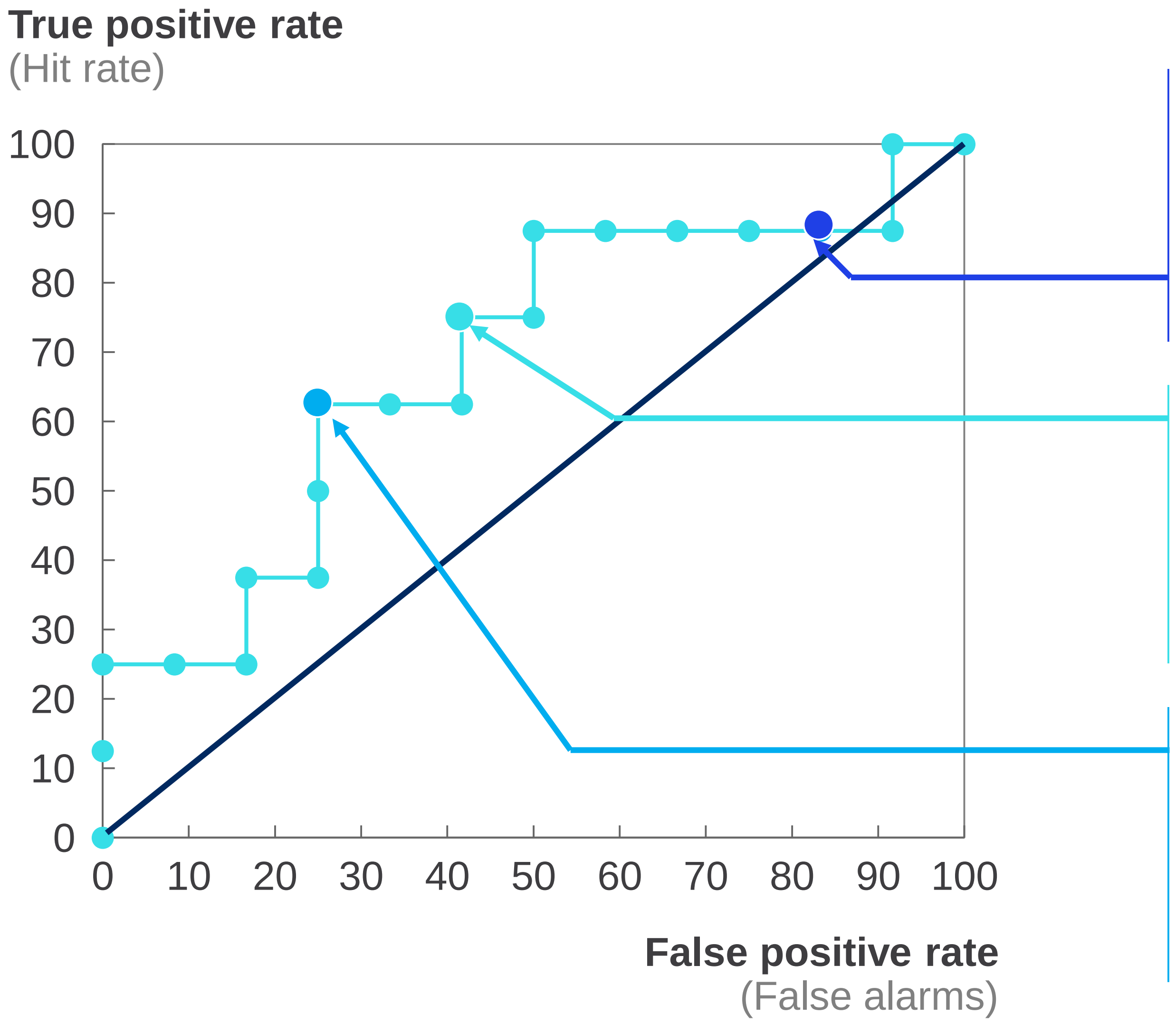
False positive rate	42%
True positive rate	75%
Accuracy	65%

Threshold: Pr(Positive) ≥ 0.2

7	1
10	2

False positive rate	83%
True positive rate	88%
Accuracy	45%

The purpose of a ROC curve is to show performance for all thresholds of a given model



Threshold: Pr(Positive) ≥ 0.2

7	1
10	2

False positive rate	83%
True positive rate	88%
Accuracy	45%

Threshold: Pr(Positive) ≥ 0.5

6	5
2	7

False positive rate	42%
True positive rate	75%
Accuracy	65%

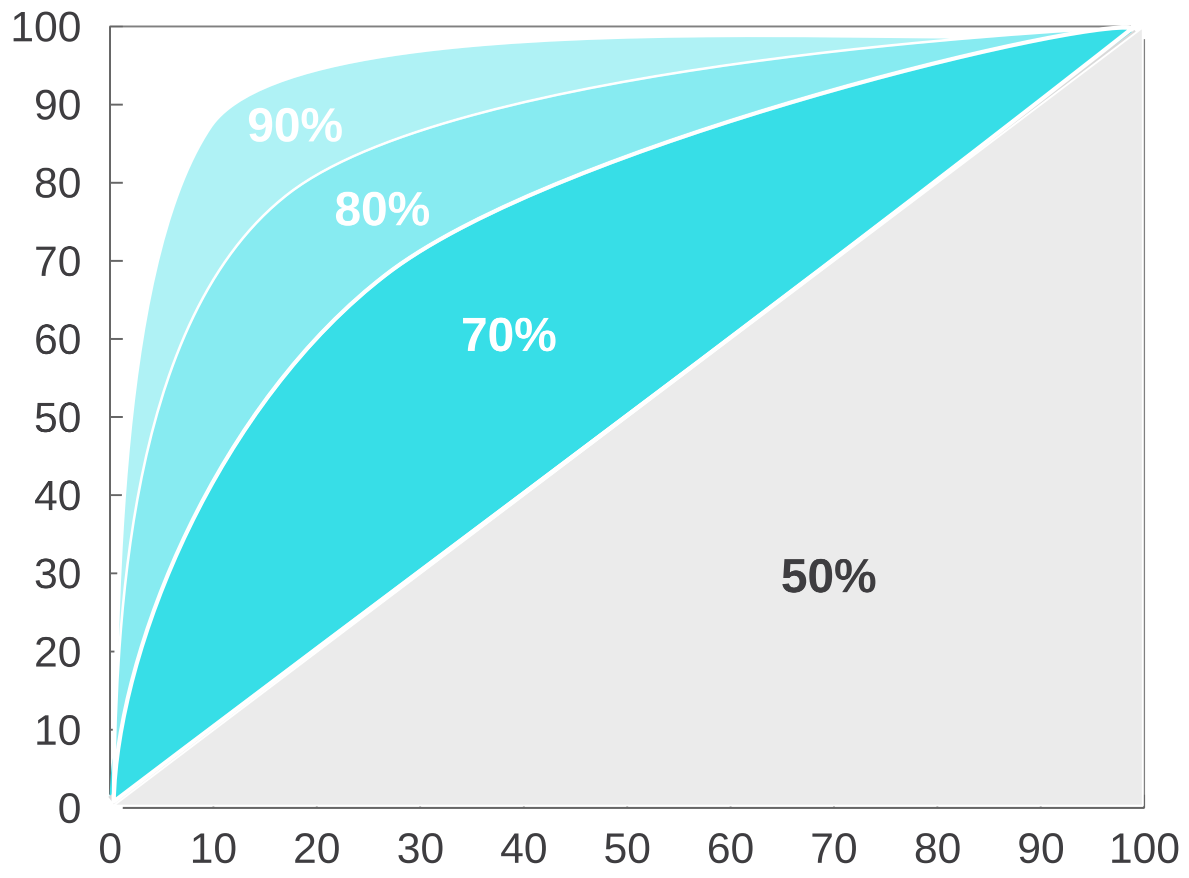
Threshold: Pr(Positive) ≥ 0.8

5	3
3	9

False positive rate	25%
True positive rate	63%
Accuracy	70%

ROC curves and their related metrics AUC/Gini are a common method for data scientists to *compare across models*

True positive rate, (Hit rate)



False positive rate, (False alarms)

ROC curves

- 45 degree line represents random guessing
- Top left corner represents perfect information

Area under the curve (AUC)

- Measure of the explanatory power of the model (generally, higher is better)
- If AUC = 50%, random guessing would have fared as well as the model

Gini coefficient

- GINI = 2*AUC – 1
- If GINI = 0, random guessing would have fared as well as the model

SOURCE: Center for Disease Control and Prevention

Communicating models

While ROC curves, AUC and Gini are great measures to discuss with data scientists, communication needs to be tailored for business audiences

Model performance

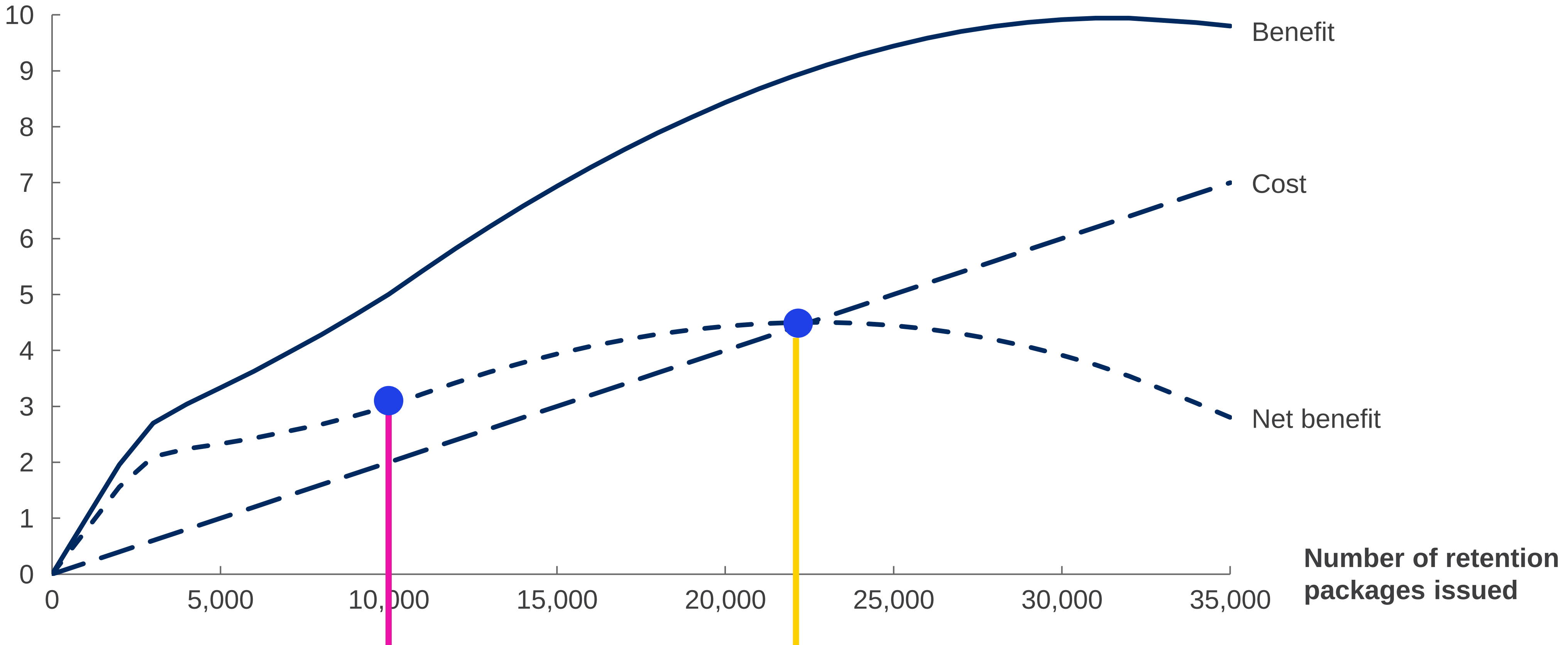
- Expected profit curves
- Profit-probability trade-offs (e.g., \$10m at 2% vs. \$1m at 20%)
- Improvements over baseline (e.g., lift)
- Scenarios
- Sample profiles (e.g., performance for selected observations)
- Performance on historical (out of sample) data
- …

Model functioning

- Step-wise explanations
- Model visualization
- Walk-through for example(s) / historical data
- …

Telco churn example: we might look at expected profit curves, depending on how many customers we offer a retention package

Expected profit and cost curves
USD million



$ 2 million investment for churn campaign not profit maximizing
- 10k packages issued
- 5k churners (worth $5m in revenues)

Profit maximizing campaign size at 22.5k retention packages, $4.5 million budget
- 22.5k packages issued
- 9k churners (worth $4.5m in revenues)

Banking churn example: Explaining model inputs and their usage in the model

x — Proxy of percentage of total prediction power

Most important variables detailed next

More than **700 variables available** and **100 synthetic variables** created:

~450 variables with higher relevance are **used in the model. Redundant and non-significant variables** have been removed and **not considered** in the model

Transactions and Activity — 52%

- ✓ **Months since last ATM transaction**
- ✓ **Days since last Teller transaction**
- ✓ **Debit Transactions on Basic Banking** (across month 1-6)
- ✓ Number of monthly customer transactions
- ✓ Teller Transactions
- ✓ Number & Volume of Mobile Payment
- ✓ Number of E-banking logins
- ✓ Credit transactions on savings & basic banking accounts

Product ownership – Basic & Saving Banking — 26%

- ✓ **Inflow on Basic Banking Account**
- ✓ **End-of-month basic banking balance**
- ✓ Inflow/Outflow on saving banking accounts
- ✓ Volume on saving accounts
- ✓ Credit transactions on savings accounts

Product ownership – Other products balance — 12%

- ✓ **CSI**
- ✓ Pension end-of-month balance
- ✓ Loan home end-of-month balance
- ✓ Fixed deposit end-of-month balance
- ✓ Maximum balance on credit cards

Demographics - Customer data — 8%

- ✓ **Customer Age**
- ✓ **Churn rate in Zip Code**
- ✓ Segment
- ✓ Tenure with client
- ✓ Credit class
- ✓ Municipality change

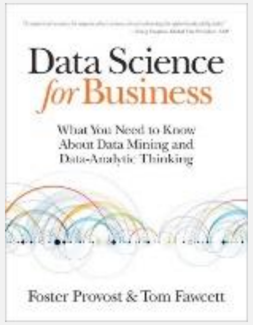
Profitability Variables — 2%

- ✓ Monthly profit from credit interest rate,
- ✓ Monthly profit from fees

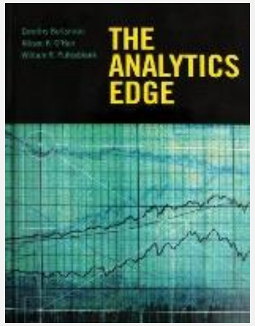
SOURCE: Team analysis

Further reading

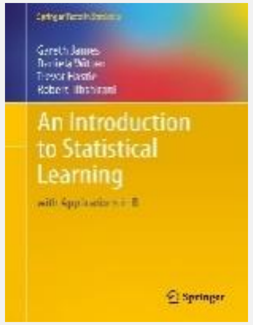
Textbooks



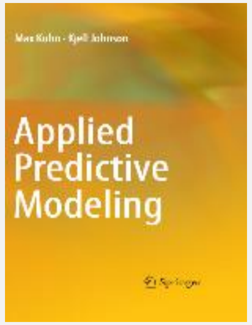
Light math for business: A business first / math light textbook is, *Data Science for Business: What you need to know about data mining and data-analytic thinking*



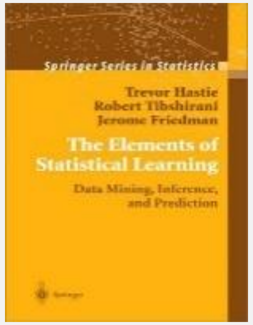
Analytics use cases: For analytics use cases that read like a novel check out *The Analytics Edge,* authored by MIT professor and the leader of our internal Analytics Bootcamp Program



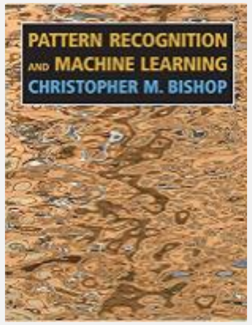
Mathematics behind algorithms (beginners): To learn the mathematics behind algorithms at a beginner level consider *An Introduction to Statistical Learning*



Predictive models: For a deep exploration into Predictive Models, there is *Applied Predictive Modeling*

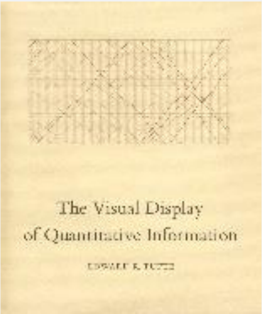


Advanced algorithms: A more advanced explanation of algorithms can be found in, *The Elements of Statistical Learning*

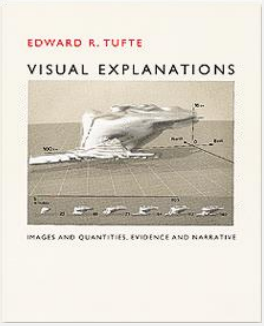


Pattern Recognition and Machine Learning: *Pattern Recognition and Machine Learning (Information Science and Statistics)*

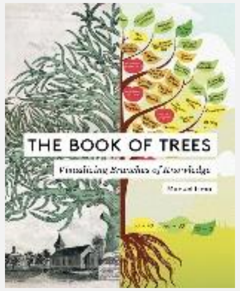
Visualization



The Visual Display of Quantitative Information, Edward Tufte



Visual Explanations, Edward Tufte



The Book of Trees, Manuel Lima

Visualization

FLOWINGDATA

https://flowingdata.com/

Other

- Andrew Ng Coursera
- DataCamp
- Sentdex Youtube

McKinsey&Company