Big data projects: stages, data preparation, exploration and judging fit

Quantitative Methods

1 A story: Ines gets her brief

- A project is a chain of separate jobs
- Menu, shop, wash and cut, taste, cook
- Each job has its own purpose
- Ask: what is this activity for?

IN THE LESSON

- Ines's boss walks over and says, can we build something that warns us about bond defaults?
- Ines does not open a spreadsheet yet.
- Think of cooking a big dinner.
- You do not start by chopping onions.
- First you decide the menu and who is coming.
- Then you shop.
- Then you wash and cut.
- Then you taste as you go.
- Only then do you cook the final dish.
- A data project is the same.
- It is a chain of separate jobs, each with its own purpose, and exam questions give you an activity and ask which job it belongs to.
- The trick is to ask what the activity is for, not what software was used.
- Keep the dinner in mind.

2 Five stages, with a way back

Structured data project

Conceptualise the task → Collect the data → Prepare and wrangle

Explore the data → Train the model → Back to any earlier stage

IN THE LESSON

- Here are the five stages of a project that uses structured data, meaning tidy rows and columns.
- Follow the boxes in the order they are drawn.
- First, conceptualise the modelling task: decide exactly what the model must output and how people will use it.
- For Ines, the output is the chance a bond defaults within twelve months, and the use is a weekly watch list.
- Second, collect the data.
- Third, prepare and wrangle it: fix its errors, then reshape it for the model.
- Fourth, explore it: look at it, pick the useful inputs and build new ones.
- Fifth, train the model: choose a method, score it on data it has not seen, and tune it.
- The last box matters.
- Results often send the team back.

3 Text projects: same chain, new names

Same five purposes; the fifth, model training, keeps its name

Structured step	Text step
Conceptualise	Formulation
Collect	Curation
Prepare and wrangle	Text prep
Explore	Text explore

IN THE LESSON

- A project built on text, such as news stories or filings, follows a parallel path with different names.
- The first stage is called text problem formulation.
- Ines decides she also wants a score for the tone of company news.
- The second stage is data or text curation: gathering the text, for example scraping a website or buying a licensed feed.
- The third is text preparation and wrangling.
- The fourth is text exploration, and the fifth is model training.
- Read the table row by row and match each structured stage on the left with its text twin on the right.
- The purposes are identical.
- Only the material is different, and text needs more careful cleaning.

4 Example 1: which stage is it?

Match each activity to its purpose

Decide: yes or no flag, weekly	output and its use	**Conceptualise**
Buy a licensed bond data feed	get the raw material	**Collect**
Remove duplicate issuers	fix an error	**Prepare**
Plot leverage by rating	look, change nothing	**Explore**
Tune on held back data	score and adjust	**Train**

IN THE LESSON

- First example, a matching exercise on the structured path.
- Ines and her team do five things during the project.
- Deciding that the output will be a yes or no default flag, used on a weekly watch list: that is conceptualising, because it is about the output and its use.
- Buying a licensed bond data feed: that is collecting.
- Removing duplicate issuers: that is preparation, because it fixes an error.
- Plotting leverage against credit rating to see the pattern: that is exploration, because it only looks.
- Tuning the model on data held back from training: that is training.
- Notice that we matched each activity to its purpose, never to a tool.

5 Example 1, part two: the text twin

Text path: decide, gather, fix, look, learn

Output is a tone score	decide what to build	**Formulation**
Scrape headlines from a site	gather the text	**Curation**
Strip web tags and spaces	fix the material	**Text prep**
Draw a word cloud	look at frequencies	**Text explore**

IN THE LESSON

- Now the same exercise for the text path.
- Ines wants a tone score from company news.
- Scraping headlines from a news website is curation, because it gathers the text.
- Stripping out web page tags and extra spaces is text preparation, because it fixes the material.
- Counting which words appear most often and drawing a word cloud is text exploration.
- Deciding that the output is a tone score from negative to positive is problem formulation, and it comes first.
- The exam angle is that a question describes an activity in plain words, and you must name the stage.
- Ask yourself whether the activity decides, gathers, fixes, looks or learns.

6 The loop back, and the trap

- Iterative: later results can send you back
- Poor model? Collect more, build features, reframe
- Do not name a stage by its tool
- Name it by its purpose

IN THE LESSON

- These stages are not a one way street.
- Suppose Ines trains her model and the results are poor.
- The cause may sit far upstream.
- Maybe the data are too thin, so she goes back and collects more.
- Maybe exploring reveals a useful ratio she never built, so she returns to build it.
- Maybe the output was badly framed and she returns to the first stage.
- That is why the process is called iterative.
- The classic trap is naming a stage by its tool.
- A spreadsheet can be used in preparation, in exploration and in training, so the tool tells you nothing.
- The purpose tells you everything.
- And remember, preparation comes before exploration, which comes before training.

7 Washing versus chopping

- Preparation (cleansing): fix errors
- Wrangling: reshape data that are correct
- Bad inputs give confident, wrong answers
- Six errors, five reshapings, outliers, scaling

IN THE LESSON

- Back to the kitchen.
- Washing the vegetables and chopping them are two different jobs.
- Data work splits the same way, and the curriculum treats the two as separate.
- Data preparation, also called cleansing, fixes what is wrong with the data.
- Wrangling then transforms data that are already correct into the shape the model needs.
- Why does this matter to an investor?
- Because a model is only as good as its inputs.
- If a leverage figure is missing or a price is negative, the model will quietly learn from garbage and give confident, wrong warnings.
- Here is the plan.
- Six kinds of error to fix.
- Five kinds of reshaping.
- Then what to do with extreme values, and how to put features on a common scale.

8 Cleansing: the first three errors

Errors one to three (examples from Ines's file)

Error and example	Fix
Incompleteness: blank leverage	Fill or drop
Invalidity: revenue below zero	Correct or drop
Inaccuracy: 450 typed, 504 true	Check source

IN THE LESSON

- Here are the first three errors.
- Incompleteness means a value is missing, such as a blank cell where an issuer's leverage should be.
- The fix is to look it up, estimate it sensibly, or drop the row.
- Invalidity means a value lies outside a meaningful range, such as a negative revenue figure or a credit rating of Z.
- The fix is to correct it or remove it.
- Inaccuracy means a value is allowed but is not a true measure, for example a debt figure typed as four hundred and fifty when the filing says five hundred and four.
- A person has to check it against a source.
- The common feature is that each fails to say what it should.

9 Cleansing: the other three errors

Errors four to six

Error and example	Fix
Inconsistency: maturity before issue	Reconcile
Non uniformity: two date styles	One format
Duplication: issuer listed twice	Remove copies

IN THE LESSON

- Three more.
- Inconsistency means one record conflicts with another, such as a bond with a maturity date before its issue date, or a firm labelled a utility in one column and a retailer in another.
- Non uniformity means the same fact is written in different formats, for example dates as day month year in one source and month day year in another, or amounts in thousands in one file and in millions in another.
- These must be put in one format.
- Duplication means the same record appears more than once, so one firm counts twice and distorts the model's picture.
- The fix is to remove the extra copies.
- Learn the six names as a list: incompleteness, invalidity, inaccuracy, inconsistency, non uniformity and duplication.

10 Wrangling: five ways to reshape

Wrangling transforms correct data

Extraction	year from a date	**New field**
Aggregation	daily to monthly	**Fewer rows**
Filtration	keep investment grade	**Fewer rows**
Selection	drop the record number	**Fewer columns**
Conversion	text rating to number	**New type**

IN THE LESSON

- Once the data are clean, wrangling transforms them.
- There are five operations.
- Extraction pulls a new field out of an existing one, such as the year from a full date.
- Aggregation combines several records into one, such as turning daily prices into a monthly average.
- Filtration removes rows that are not needed, such as keeping only investment grade issuers if that is the brief.
- Selection removes columns that are not needed, such as an internal record number.
- And conversion changes the data type, such as turning a rating written as text into a number, or text dates into real dates.
- Notice that none of these fixes an error.
- They all shape correct data.

11 Example 2: cleansing Ines's file

Vendor file: cleanse before anything else

Rows received		**520**
Less duplicates	520 − 20	**500 issuers**
Missing leverage	incompleteness	**12**
Negative revenue	invalidity	**4**
Issuers needing a fix	16 of 500	**3.2%**

IN THE LESSON

- Second example, with counts.
- The vendor sends Ines a file of five hundred and twenty rows.
- Looking closely, twenty rows are duplicates of other rows.
- Remove them and five hundred unique issuers remain.
- Of those, twelve have a blank leverage figure, which is incompleteness.
- Four show a negative revenue, which is invalidity.
- Assume no issuer has both problems.
- So sixteen issuers need a fix, out of five hundred.
- That is three point two per cent.
- It sounds small, but if those sixteen are the riskiest names, ignoring them would hide exactly the cases Ines cares about.
- Counting the errors before fixing them also tells her how much to trust the vendor.

12 Outliers: trim or winsorise

- Outlier: a value far from the rest
- Trimming: remove extreme observations
- Winsorising: cap them at a percentile
- Winsorising keeps every observation

IN THE LESSON

- Next, outliers: values far from the rest.
- Imagine a class where four students are about the same height and one is a basketball player.
- Ines has to decide what to do with such a value.
- Sometimes it is an error and should be corrected.
- But sometimes it is real, such as one firm with enormous debt.
- There are two common treatments.
- Trimming, also called truncation, removes the extreme observations altogether, so the firm is lost from the data.
- Winsorising replaces the extreme values with the values at chosen percentiles, so the firm stays but its value is capped.
- Winsorising keeps every observation.
- Trimming throws some away.
- Both are wrangling choices, not the correction of errors.

13 Scaling one: normalisation

Normalised X = (X − min) ÷ (max − min)

X: the value	The value to scale
Min: smallest value	Becomes 0
Max: largest value	Becomes 1

IN THE LESSON

- Now scaling.
- Imagine comparing a runner's time in seconds with their age in years.
- The big numbers would dominate.
- Many models, especially those that measure distances, are thrown off when one feature is simply measured in bigger units.
- Scaling puts features on comparable footing.
- The first method is normalisation, sometimes called min max scaling.
- Subtract the smallest value from the value you want, and divide by the range, which is the largest value minus the smallest.
- The result always lies between zero and one: the smallest value becomes zero and the largest becomes one.
- It is easy to understand.
- Its weakness is that a single extreme value sets the largest number and so squeezes everything else.

14 Scaling two: standardisation

Standardised X = (X − mean) ÷ standard deviation

X: the value	The value to scale
Mean of the feature	Subtracted: centres on 0
Standard deviation	Divided by: sets the spread

IN THE LESSON

- The second method is standardisation.
- Subtract the mean of the feature from the value, and divide by the standard deviation, which measures how spread out the values typically are.
- The result says how many standard deviations the value lies from the average.
- The feature is centred on zero, with positive numbers above average and negative numbers below.
- Unlike normalisation, there is no fixed floor or ceiling, and the result is less sensitive to outliers, although a very extreme value still pulls the mean and the standard deviation.
- To avoid mixing them up, remember what each produces.
- Normalisation uses the smallest and largest values and gives zero to one.
- Standardisation uses the mean and the standard deviation, and centres on zero.

15 Example 3: one outlier, two scalings

Feature values 2, 4, 6, 8, 30 (mean 10, SD 11.40)

Normalise 6	(6 − 2) ÷ (30 − 2)	**0.14**
Standardise 6	(6 − 10) ÷ 11.40	**−0.35**
Trim the 30	2, 4, 6, 8	**mean 5**
Winsorise the 30	2, 4, 6, 8, 8	**mean 5.6**
Normalise 6 again	(6 − 2) ÷ (8 − 2)	**0.67**

IN THE LESSON

- Third example.
- One feature, measured across five firms, takes the values two, four, six, eight and thirty.
- Its mean is ten and its sample standard deviation is eleven point four.
- Normalise the value six: six minus two, over thirty minus two, is point one four.
- Standardise it: six minus ten, over eleven point four, is minus point three five.
- Now handle the outlier.
- Trimming the thirty leaves two, four, six and eight, with a mean of five.
- Winsorising caps the thirty at eight, giving two, four, six, eight, eight, with a mean of five point six, and no firm is lost.
- Normalise six again: six minus two, over eight minus two, is point six seven.
- Much healthier.

16 Normalised values, squeezed by one outlier

- A 0.00
- B 0.07
- C 0.14
- D 0.21
- E 1.00

IN THE LESSON

- Let us see the squeeze on a line.
- This is the zero to one scale, with zero for the smallest value on the left and one for the largest on the right.
- Firms A to E have the raw values two, four, six, eight and thirty.
- After normalising, firm A sits at zero.
- Firm B is at point zero seven.
- Firm C is at point one four.
- Firm D is at point two one.
- And firm E, the outlier, sits all the way out at one.
- Four of the five firms are bunched in the bottom fifth of the scale, so the model can barely tell them apart.
- That is why one extreme value can ruin normalisation, and why winsorising first spreads the others out again.

17 Looking, choosing, building

- Explore: summarise and visualise
- Select: keep features that explain the target
- Engineer: create new features from old
- Aim: best inputs before training

IN THE LESSON

- Ines now has clean, scaled data.
- Before training, she explores.
- Think of a chef who tastes the ingredients, picks the best ones, and sometimes blends a new sauce.
- Data exploration has three parts.
- First, exploratory data analysis, summarising and visualising the data to understand it.
- Second, feature selection: keeping only the features that help explain the target, and dropping the rest.
- A feature is simply an input column, and the target is what the model predicts.
- Third, feature engineering: creating new features from the existing ones.
- The aim is to understand the data and give the model the best possible inputs.
- The common confusion is between the second and third.
- Selection chooses among features that already exist.
- Engineering makes new ones.

18 Exploratory data analysis: the toolkit

Match the tool to what it shows

Tool	Shows
Histogram, bar chart, box plot	One feature
Scatterplot, correlation matrix	Pairs of features
Word cloud	Frequent tokens

IN THE LESSON

- Here is the toolkit for the first part.
- For a single feature, Ines uses a histogram, which shows how often values fall in each band, a bar chart for categories, and a box plot, which shows the middle half of the values as a box and marks the extremes.
- For a pair of features, she uses a scatterplot, which puts one feature on each axis, and a correlation matrix, which is a grid of correlations between every pair of features.
- For text, the usual picture is a word cloud, in which the most frequent tokens are drawn largest.
- A token is one word or piece of text.
- Remember that this part only describes.
- It changes nothing in the data.

19 Example 4: explore, select, engineer

Debt 450, EBITDA 150 for one issuer

Box plot of leverage by rating	exploratory analysis	**Look**
Debt and liabilities: 0.97	keep only one	**Selection**
New: debt ÷ EBITDA	450 ÷ 150	**3.0×**
Sector: bank, utility, retailer	three 0/1 columns	**One hot**

IN THE LESSON

- Fourth example.
- Ines predicts default from total debt, total liabilities, E B I T D A and sector.
- E B I T D A is a measure of a firm's operating earnings.
- First, a box plot of leverage by credit rating: that is exploratory analysis.
- Second, the correlation matrix shows debt and liabilities move together, with a correlation of point nine seven.
- Keeping both adds noise, not information, so she drops liabilities.
- That is selection.
- Third, for one issuer, debt is four hundred and fifty and E B I T D A is one hundred and fifty.
- Debt over E B I T D A is three point zero times.
- That is a new feature, so it is engineering.
- Fourth, sector becomes three columns of zeros and ones.
- That is one hot encoding, also engineering.

20 Selection versus engineering

Did the feature exist before you touched it?

Action	Type
Drop liabilities, keep debt	Selection
Debt ÷ EBITDA ratio	Engineering
Log of a skewed feature	Engineering
Sector to 0/1 columns	Engineering

IN THE LESSON

- Here is a quick test to separate the two.
- Ask whether the feature existed before Ines touched it.
- Dropping liabilities removes a column that existed.
- That is selection.
- A ratio of debt to E B I T D A did not exist until she built it.
- That is engineering.
- Taking the logarithm of a skewed feature creates a new, better behaved column.
- That is engineering.
- Keeping the ten most useful of fifty features is selection.
- And turning a category into indicator columns is engineering, called one hot encoding.
- Both reduce the burden on the model.
- Selection cuts noise and the risk of overfitting, which is a model learning the quirks of its data instead of the pattern.
- Engineering gives it clues it could not see alone.

21 Four mistakes to avoid

- 1. Name a stage by purpose, not tool
- 2. Cleansing fixes; wrangling reshapes
- 3. Normalise: min and max. Standardise: mean and SD
- 4. Selection chooses; engineering creates

IN THE LESSON

- Before moving on, four mistakes that cost marks.
- One: naming a stage by its tool instead of its purpose.
- Plotting is exploration whatever software draws it.
- Two: mixing up cleansing and wrangling.
- Cleansing fixes errors, such as missing values and duplicates.
- Wrangling reshapes correct data, by extraction, aggregation, filtration, selection and conversion.
- Three: swapping the two scaling formulas.
- Normalisation uses the smallest and largest values and gives zero to one, so one outlier squeezes it.
- Standardisation uses the mean and standard deviation.
- Four: confusing feature selection, which chooses among existing features, with feature engineering, which creates new ones.
- In the next section there is a fifth: trusting accuracy alone.

22 Ninety-five per cent right, and useless

- Classification: sort cases into categories
- Lazy model, no defaults ever: 95% accurate
- Accuracy misleads when classes are unbalanced
- Better tools start from one small table

IN THE LESSON

- Ines trains her model.
- It scores.
- How good is it?
- Here is a puzzle.
- Suppose of one thousand bonds, only fifty will default.
- A lazy model that simply says, no default, for every bond, is right nine hundred and fifty times.
- That is ninety five per cent accurate, and it has caught no defaults at all.
- So raw accuracy can badly mislead when one outcome is rare.
- The curriculum gives us better tools for a classification model, one that sorts cases into categories.
- Every one of them comes from a small table that counts four kinds of result: hits, false alarms, misses and correct all clears.
- Learn the table first, and the formulas are just arithmetic on it.

23 The confusion matrix

1,000 loans, 50 defaults, 60 flagged

Outcome	Count
Flagged, defaulted (TP)	40
Flagged, repaid (FP)	20
Missed default (FN)	10
Cleared, repaid (TN)	930

IN THE LESSON

- Here is the table, called a confusion matrix.
- Ines tests her model on one thousand loans, of which fifty really default.
- The model raises a flag on sixty loans.
- Of those, forty did default.
- These are true positives, the hits.
- Twenty flagged loans were repaid.
- These are false positives, false alarms.
- The model missed ten defaults, which it did not flag.
- These are false negatives, the misses.
- And nine hundred and thirty loans were not flagged and were repaid.
- These are true negatives, the correct all clears.
- Check the totals: forty plus ten is fifty defaults, and twenty plus nine hundred and thirty is nine hundred and fifty repaid.
- The word positive means the model said default.

24 Precision and recall

Precision = TP ÷ (TP + FP)
Recall = TP ÷ (TP + FN)

TP: true positives	Flagged and defaulted
FP: false positives	Flagged but repaid
FN: false negatives	Defaulted but missed

IN THE LESSON

- Two measures ask different questions.
- Precision asks: of the loans the model flagged, how many really defaulted?
- That is true positives divided by true positives plus false positives.
- It matters when false alarms are costly, for example if a flag freezes a customer's card.
- Recall, also called sensitivity, asks: of the loans that really defaulted, how many did the model catch?
- That is true positives divided by true positives plus false negatives.
- It matters when a miss is costly, as with a lender who must not overlook a default.
- A memory aid: precision looks across the flagged row, recall looks down the defaulted column.
- The bottom of precision has false positives.
- The bottom of recall has false negatives.

25 Accuracy and the F one score

Accuracy = (TP + TN) ÷ all
F1 = 2 × P × R ÷ (P + R)

Correct calls	Hits plus correct clears
All observations	Every loan tested
Precision	TP ÷ (TP + FP)
Recall	TP ÷ (TP + FN)

IN THE LESSON

- Two more.
- Accuracy counts all correct calls, true positives plus true negatives, over all observations.
- It is simple but flattered by an easy, huge true negative cell.
- The F one score blends precision and recall into one number: two times precision times recall, divided by precision plus recall.
- It is a harmonic mean, which means it is dragged toward the weaker of the two, so a model cannot hide a poor recall behind a good precision.
- That is why the F one score suits unbalanced data.
- Notice what is missing from precision, recall and F one: the true negatives.
- The huge easy cell cannot inflate them.
- That is exactly the point.

26 Example 5: scoring Ines's model

TP 40, FP 20, FN 10, TN 930

Precision	40 ÷ (40 + 20)	**0.67**
Recall	40 ÷ (40 + 10)	**0.80**
Accuracy	(40 + 930) ÷ 1,000	**0.97**
F1 score	2 × 0.67 × 0.80 ÷ 1.47	**0.73**
Flags nothing: accuracy	950 ÷ 1,000, recall 0	**0.95**

IN THE LESSON

- Fifth example, using the table.
- Precision is forty over forty plus twenty, so forty over sixty, which is point six seven: two in three flags are right.
- Recall is forty over forty plus ten, so forty over fifty, point eight zero: four in five defaults caught.
- Accuracy is forty plus nine hundred and thirty, over one thousand, point nine seven.
- It looks superb, but it is flattered by the easy clears.
- The F one score is two times point six seven times point eight zero, over point six seven plus point eight zero, which is point seven three.
- Finally the lazy model that flags nothing has accuracy of point nine five and recall of zero.
- Accuracy barely separates the two.
- Recall and F one do.

27 Four scores for one model

Score (per cent)

Precision	**67**
Recall	**80**
F1 score	**73**
Accuracy	**97**

Same model: accuracy flatters it

IN THE LESSON

- Here are the four scores for the same model as bars, in per cent.
- Follow them from left to right.
- Precision is sixty seven.
- Recall is eighty.
- The F one score is seventy three, sitting between the two, a little nearer the weaker one.
- And accuracy towers over them at ninety seven.
- Same model, same loans, and yet accuracy tells a far rosier story, because nine hundred and thirty of the thousand loans were easy correct clears.
- This is the picture to remember when an exam question praises a model for high accuracy on data where one class is rare.
- Ask for precision, recall or F one before you agree it is any good.

28 The ROC curve

Chart: True positive rate vs False positive rate

- Illustrative model
- Random guess
- Bigger area, better

IN THE LESSON

- Models usually give a score, and the user picks a threshold above which a loan is flagged.
- A lower threshold flags more loans.
- The receiver operating characteristic curve, or R O C curve, shows every threshold at once.
- Look at the axes.
- Along the bottom is the false positive rate: the share of repaid loans wrongly flagged.
- Up the side is the true positive rate, which is recall: the share of defaults caught.
- The dashed diagonal is a model that guesses at random.
- The solid curve is an illustrative model like Ines's, bowing up towards the top left corner, which means many defaults are caught for few false alarms.
- The area under the curve, or A U C, is about point eight eight here.
- Bigger is better, and a half is a coin toss.

29 Which error costs more?

- Missing a positive costly: protect recall
- False alarms costly: protect precision
- Both matter, unbalanced classes: F1
- Numeric target: root mean squared error

IN THE LESSON

- Which matters more, precision or recall?
- That is a business question, not a statistical one.
- For Ines's lender, a missed default is the expensive mistake, so recall is the number to protect, and twenty good borrowers wrongly flagged is a price worth paying.
- For a fraud alert that freezes a customer's account, false alarms carry the cost, so precision comes first.
- When both matter and the classes are unbalanced, use the F one score.
- Finally, not every target is a category.
- When the model predicts a number, such as a bond's spread over government bonds, the usual measure of fit is the root mean squared error.
- It asks, typically, how far off are the predictions?
- Smaller is better.

30 Root mean squared error

RMSE = √[Σ(Predicted − Actual)² ÷ n]
Errors 10, −5, 10, −10: squares sum to 325
325 ÷ 4 = 81.25; root ≈ 9 bp

Error	Predicted minus actual
n: observations	How many predictions
Square root	Back to original units

IN THE LESSON

- Here is the formula for a numeric target.
- For each observation, take the error, which is the prediction minus the actual value.
- Square each error so that misses above and below cannot cancel, and so big misses count extra.
- Add the squares, divide by the number of observations to get their average, and then take the square root to return to the original units.
- Let us try four bond spreads, in basis points.
- Actual values are one hundred, one hundred and twenty, one hundred and fifty and two hundred.
- Predictions are one hundred and ten, one hundred and fifteen, one hundred and sixty and one hundred and ninety.
- The errors are ten, minus five, ten and minus ten.
- The squares sum to three hundred and twenty five.
- Divided by four that is eighty one point two five, and the square root is about nine basis points.

31 Recap

- Five stages, iterative; name by purpose
- Cleanse errors; wrangle; trim or winsorise; scale
- Explore, select, engineer
- Precision, recall, F1, AUC; RMSE for numbers

IN THE LESSON

- Let us pull it together.
- A structured project runs through conceptualise, collect, prepare and wrangle, explore and train, and the team loops back when results are poor.
- A text project has the same chain: problem formulation, curation, text preparation, text exploration, training.
- Name a stage by its purpose.
- Cleansing fixes six kinds of error.
- Wrangling reshapes correct data.
- Outliers are trimmed or winsorised.
- Normalisation uses the smallest and largest values; standardisation uses the mean and standard deviation.
- Exploration is analysis, selection and engineering.
- Finally, judge a classifier from the confusion matrix: precision, recall, accuracy, F one and the area under the curve, and root mean squared error for numbers.

32 Try three quick questions

1. Scraped news stories have their web page tags removed. Which stage?
   Answer: Text preparation and wrangling: it fixes the material.
2. Feature values 2, 4, 6, 8, 30. What is the normalised value of 6?
   Answer: (6 − 2) ÷ (30 − 2) = 0.14.
3. A model catches 40 of 50 defaults with 20 false alarms. Precision and recall?
   Answer: Precision 40 ÷ 60 = 0.67; recall 40 ÷ 50 = 0.80.

IN THE LESSON

- Now try three quick questions yourself.
- Pause after each one.
- Question one.
- A team removes web page tags from scraped news stories.
- Which stage is that?
- The answer is text preparation and wrangling, because it fixes the raw material.
- Question two.
- A feature has values two, four, six, eight and thirty.
- What is the normalised value of six?
- It is six minus two, over thirty minus two, which is point one four.
- Question three.
- A model finds forty of fifty defaults and raises twenty false alarms.
- What are its precision and recall?
- Precision is forty over sixty, point six seven, and recall is forty over fifty, point eight zero.