CHARTA LEARNING · LESSON NOTES

Fintech and big data

CFA Level I · Quantitative Methods

You will learn: Say which parts of fintech matter to an investment analyst, explain big data, artificial intelligence and machine learning, and describe how they are used in investment management and why a pattern is not yet a strategy.

Running example: Examples use made-up figures, such as a £20,000 account charged a 1% fee and a backtested pattern that returns 10% a year.

1. Fintech for an analyst: four strands

- Fintech is short for financial technology: technology used to change how financial services are delivered. Only a narrow slice matters to an analyst, not payments apps in general.
- Strand 1, large datasets: raw material such as card payments, searches, satellite photos and posts, often with no neat rows and columns. Strand 2, analytical techniques: the methods that make such data usable.
- Strand 3, automated trading: software carries out orders with nobody pressing buttons, for example slicing a huge order into small pieces. Strand 4, automated advice (robo advice): a questionnaire in, a portfolio kept balanced over time out.
- Strands 1 and 2 change what you can know: an edge is possible only if rivals lack the same data or method. Strands 3 and 4 change what it costs to act and are sold to everyone, so they rarely give an edge. Shared ledgers belong to digital assets.

Fintech for analysts = large datasets + techniques + automated trading + automated advice

WHAT DOES EACH PROPOSAL BUY?

Satellite car park photos	a new dataset	**Edge possible**
Earnings call reader	a technique	**Edge possible**
Order slicing program	automated trading	**Lower cost**
Robo advice front end	automated advice	**Cheaper clients**

Watch out: Thinking fintech means payments apps.

2. Robo advice changes cost, not returns

- Profit equals the fee earned minus the cost of serving the account. Robo advice makes small accounts worth serving because the cost of serving falls.
- Nothing about the market or the client's return changes. Only the cost of serving changes.

Profit = fee income − cost of serving the account

ACCOUNT OF £20,000, FEE 1% A YEAR

Fee income	1% × £20,000	**£200**
Human adviser	200 − 600 cost	**−£400**
Robo adviser	200 − 50 cost	**£150**

Watch out: Believing robo advice raises returns. It lowers the cost of serving the account.

3. Big data

- Big data is described by the Vs. Volume: how much data. Velocity: how fast it arrives and must be handled. Variety: how many shapes (tables, text, photos, audio, video). Veracity: whether the data is true, accurate and free of fakes.
- The first three describe the pile; only veracity tells you whether it can be trusted.
- Sources: traditional records (accounts, prices), data people create (posts, reviews, searches), business processes (card payments) and sensors (phones, satellites, machines).
- Much big data is unstructured: not in neat rows and columns, such as text, pictures and audio. This is the genuinely new part.

> **Big data = volume + velocity + variety (+ veracity)**

THIRTY MILLION SOCIAL MEDIA POSTS FOR SALE

Volume	30 million posts	**Big**
Velocity	50,000 new posts an hour	**Fast**
Variety	text and photos	**Mixed**
Veracity	not stated: fake accounts?	**Unknown**

Watch out: Judging big data by the first three Vs and forgetting veracity.

4. AI, machine learning and overfitting

- Artificial intelligence is machines doing tasks that need human judgement. Machine learning, inside it, works the rule out from data. Deep learning, inside that, stacks many layers to capture very complex relationships. It can be used for either kind of learning.
- Supervised learning learns from labelled examples where the right answer is known, then predicts. Unsupervised learning finds structure with no answers given. Test: were the right answers supplied?
- Overfitting is learning the noise of the training data as if it were signal. A more complex model usually fits its training data better, so judge a model on data it has not seen.
- As complexity rises, training error keeps falling, but unseen-data error falls, reaches a low point, then climbs. Choose the model at that low point.

WHICH MODEL DO WE TRUST? ERROR RATES

Model A (simple)	training / unseen	**20% / 22%**
Model B (moderate)	training / unseen	**12% / 14%**
Model C (very complex)	training / unseen	**2% / 25%**
Best choice	lowest unseen error	**Model B**
Label posts positive or negative	answers supplied	**Supervised**
Group posts by similar topic	no answers given	**Unsupervised**
Predict rise or fall of a price	answers supplied	**Supervised**

Watch out: Judging a model by how well it fits the data it trained on.

5. Uses in investment management

- Text analytics and natural language processing turn words (filings, transcripts, news) into numbers. A sentiment score measures positive or negative tone and feeds a model like any other variable.
- Robo advisers handle allocation and rebalancing cheaply. Risk analytics pulls together every exposure across a firm far faster than by hand. Algorithmic trading works a large order in small pieces to limit price moves against you.
- None of this works unless data is captured, cleaned, stored in a searchable way and moved to the model. That pipeline is most of the real work; the modelling is the small final step.

Sentiment score = (positive words − negative words) ÷ total words

ONE EARNINGS CALL, 1,000 WORDS, SIMPLE WORD COUNTS

Positive / negative words	on the two lists	**50 / 20**
Net count	50 − 20	**30**
Sentiment score	30 ÷ 1,000	**0.03**
Next call: 40 positive, 60 negative	net −20; −20 ÷ 1,000	**−0.02**

6. A pattern is not yet a strategy

- With a huge dataset you can find patterns that are statistically real but not tradable. A backtest tests a strategy on past data; its return is the headline on paper.
- Three things sit between pattern and profit. Transaction costs: fees and price impact. Capacity: how much money the strategy can absorb before it stops working. Decay: others find the same pattern and compete it away.
- Transaction costs can be estimated before risking money, so they are the cheapest test, and the one most often skipped. A better statistical test, such as a lower p value, solves none of these problems.

What is left = backtest return − transaction costs − capacity effects − decay

A PATTERN THAT RETURNS 10% A YEAR IN A BACKTEST (ILLUSTRATIVE)

Backtest return	on paper	**10.0%**
− Transaction costs	10.0 − 3.0	**7.0%**
− Capacity effects	7.0 − 2.0	**5.0%**
− Decay	5.0 − 3.5	**1.5%**

Watch out: Treating a statistically real pattern as a tradable strategy, and ignoring costs, capacity and decay.

Key takeaways

- For an analyst, fintech is four strands: large datasets, techniques, automated trading and automated advice.
- The first two change what you can know; the last two change what it costs to act.
- Big data is volume, velocity, variety and veracity, and much of it is unstructured.
- AI contains machine learning, which contains deep learning. Supervised means answers were supplied.
- Overfitting is learning the noise, so test on unseen data.
- A pattern is not a strategy until it survives costs, capacity and decay.

Quick check

1. A firm buys car park photos to estimate sales before they are reported. Which aspect of fintech is this?
   Answer: Analysis of large, alternative datasets. New data turned into a number.
2. A model is excellent on its training data but poor on new data. What has gone wrong?
   Answer: Overfitting: it learned the noise as if it were signal.
3. A backtest shows a high return. Give two reasons it may not be tradable.
   Answer: Any two of: transaction costs, capacity limits, decay.