Post–Graduate Course:
Big Data & Statistical Learning
(FinTech & AI Engineering Studio Edition)

George Tsomidis, PhD

June 2025

1 Course Vision

Welcome to a **11–week FinTech startup simulation**, where students will **design, build and ship** a Python-based analytics product that converts raw economic, financial or even crypto data into concise investment intelligence for an imaginary board.

All theory is delivered **just–in–time** inside weekly "sprints". Every sprint contains:

- a *1 h Practical–Engineering Power Hour* (theory / live coding / mini–datathon), followed by
- *10 min individual showcases* where each student (analyst henceforth) demos progress to the class (henceforth team, consisting of squads) and your lecturer (who will be playing the "Team Leader" role).

> **Inclusive Project Scoping.** Each project is calibrated to the student's background. For instance, a non-technical person (e.g. lawyer) may focus on regulatory–risk analytics or SQL schema design, while a Computer Science graduate may deep–dive into GPU pipelines. Diversity of skills inside every team is deliberate and assessed.

2 Learning Outcomes

Domain	**Competence**
AI Prompt–Engineering Mastery	Craft, iterate and chain prompts to produce accurate insights and code; benchmark large language models (LLMs) for quality and latency.

Tech Stack	Ingest, clean and store large, multi–frequency data sets; implement classical and modern ML in Python; deploy notebooks to lightweight APIs; version control with GitHub.
Visualization & Storytelling	Build dashboards and publication–quality plots; articulate insights visually and narratively in one–page board briefs.
Process	Operate in Agile Scrum using JIRA: user stories, velocity, retrospectives, code review.
Domain Knowledge	Translate statistical output into financial insight; optionally include crypto–asset analytics (on–chain metrics, DeFi yields, or whatever else).
Communication	Produce a decision–ready one–page report and defend it verbally.

3 Tool Stack

- **Python $\geq$ 3.11**: pandas / polars, statsmodels, scikit–learn, PyTorch–Forecasting, Prophet, MLflow, Plotly.
- **AI & Prompt Toolkit**: ChatGPT, Claude, Gemini, local LLMs via LangChain, etc.; Week 5 includes a prompt benchmark harness.
- **Collaboration**: GitHub Classroom, Slack, **JIRA** (our main project management tool), GitHub Actions CI, Colab, any recommendations are welcome.
- **Data Sources**: Yahoo Finance, Bloomberg, Coinbase, etc.
- **Compute**: Jupyter Notebook, any cloud service, or simply on your PCs.

4 Weekly Sprint Themes and Milestones

Wk	Sprint Focus	Concept Burst (15 min)	Hands-On Objectives
1	Startup Onboarding	Agile Scrum fundamentals; Git/JIRA etiquette	1-page project charter; form diverse squads; initialise repo and JIRA board
1	Data Sourcing	Relational databases; SQL joins; API authentication	Harvest one macro and one market dataset; commit raw data
2	Data Wrangling	Data types; missing-value imputation; tidy-data rules	Clean data tables; create two exploratory plots; quality checklist

3	Classical Econometrics	OLS assumptions; hypothesis testing	Fit baseline regression; examine residuals
4	Time-Series Forecasting	Stationarity tests; ACF/PACF; ARIMA models	Build first forecast; compare against naive benchmark
5	ML Regularization	Ridge/Lasso; K-fold cross-validation	Tune penalized model; log prompt-assisted patterns
6	Tree Ensembles	Random Forest; Gradient Boosting; feature importance	Train ensemble; interpret with SHAP
7	Dimensionality Reduction	PCA; factor models; k-means clustering	Derive latent factors; integrate into pipeline
8	Deep Learning for Sequences	RNN/LSTM basics; sequence modeling	Prototype small LSTM; report performance lift
9	Portfolio Analytics	Mean–variance theory; risk metrics; back-testing	Draft trading rule; run walk-forward test
10	Model Audit	Cross-sample validation; stress tests	Create audit notebook; highlight model limits
11	Productization	Reproducibility; FastAPI; CI/CD workflow	Package API/dashboard; pass Confidence Intervals tests
11	Board Presentation	Storytelling; slidecraft best practice	Deliver 10-min pitch and one-page brief; final repo

5 Assessment

Component	**Weight**	**Evidence**
Sprint Professionalism	20%	JIRA velocity; stand–ups; pull request etiquette.
Individual Technical Depth	20%	Notebook contributions; code quality; peer review.
Visualisation & Storytelling	15%	Clarity and impact of plots/dashboards across sprints.

Squad Product	25%	Live demo plus repository health (tests, CI, docs).
Final Board Report	20%	One–page PDF (insight density, actionability, narrative flow).

- LLM usage **must be declared** in commits; reproducibility failures can cost up to 40% of a component.
- **NO** final exam.
- **Laptops are mandatory in lectures.**
- GR / EN are both acceptable, with a preference though towards EN especially when it comes to reporting.

6 Teaching and Support

- **Weekly Rhythm**: 1 h Practical Power Hour → 10 min showcases → squad retrospective.
- **One–to–One "Team-Leader Hours"**: fortnightly, 15 min slots for code or concept deep dives.
- Slack Ask-Me-Anything (AMA) and bookable calendars for on–demand help.
- Resources: JIRA crash course video; template notebooks; prompt pattern library.
- Provided slides, own notes and discussion in class.
- Free Access First, where possible (e.g., Hilpisch PDF, GitHub repos, arXiv papers).

7 Extra Reading and Resources

- *An Introduction to Statistical Learning* (Python ISLP) — core chapters.
- Hyndman & Athanasopoulos, *Forecasting: Principles and Practice* (Python port).
- Glasserman, *Monte Carlo Methods in Financial Engineering* (selected sections).
- Roncalli, *Machine Learning for Asset Management*, Chapter 8.
- QuantEcon lectures; GitHub `awesome-llm-ops` collection.

References

Barberis, Janos, Douglas W. Arner, and Ross P. Buckley (2019). *The RegTech Book: The Financial Technology Handbook for Investors, Entrepreneurs and Visionaries*. Module 7 – RegTech, Risk & Compliance. Chichester, UK: John Wiley & Sons.

Harvey, Campbell R., Ashwin Ramachandran, and Joey Santoro (2021). *DeFi and the Future of Finance*. Module 5 – Blockchain, DeFi & Web3. Hoboken, NJ: John Wiley & Sons.

Hilpisch, Yves (2019). *Python for Finance: Mastering Data-Driven Finance*. 2nd ed. Module 1 – Python Foundations. Sebastopol, CA: O'Reilly Media.

James, J. Sergios Makram, Paul A. Spindt, and Georgios Sermpinis (2021). *Machine Learning in Finance: From Theory to Practice*. Module 8 – Capstone & Advanced Topics. Cham, Switzerland: Springer.

Jansen, Stefan (2021). *Machine Learning for Algorithmic Trading*. 2nd ed. Module 3 – ML & Algo Trading. Birmingham, UK: Packt Publishing.

Lopez de Prado, Marcos (2018). *Advances in Financial Machine Learning*. Module 2 – Time-Series & Alt-Data. Hoboken, NJ: John Wiley & Sons.

Sivan, Eyal (2024). *Open Banking: A Guide to Banking in the Platform Economy*. Module 6 – Open Banking & API-First Finance. Shelter Island, NY: Manning Publications.

Thomas, Lyn C., David J. Edelman, and Jonathan N. Crook (2017). *Credit Scoring and Its Applications*. 2nd ed. Module 4 – Credit & Lending Risk. Philadelphia, PA: SIAM.