POCKET ATLAS · MINI

₹49 EDITION

Data Science, Minus the Math Panic

The 20-minute map every beginner should read *before* picking a course, an algorithm, or a career lane. No equations. No hype. Just what's true.

WHAT'S INSIDE

Everything that matters, in pointers

This mini gives you the honest map of data science in 20 minutes: which of the five jobs you actually want, the statistics that carry 80% of the work, the algorithms worth knowing, how to pick a chart, and the real path in. Read it once and you'll know more than most applicants.

THE ONE IDEA TO HOLD THE WHOLE WAY

The field is growing fast **and** most of the daily job is prepping data, not modeling. Both are true at once. Learn the boring 80% and the algorithms become the easy part.

CHAPTER 01

The field map — one field, five jobs

- **It's not one job:** data analyst, data scientist, data engineer, analytics engineer and ML engineer each answer a different question with different tools.
- **Data Analyst:** "what happened, and why?" — SQL, Excel, Power BI/Tableau. Dashboards and reports.
- **Data Scientist:** "what will happen? what should we do?" — Python, statistics, experiments and predictions.
- **Data Engineer / Analytics Engineer:** move raw data at scale, then turn it into clean, tested models — the dbt framing: engineers build the infrastructure, analytics engineers build models on it, analysts consume those models.
- **ML Engineer:** puts models into production and keeps them running — Python, PyTorch, Docker, MLOps.

INDIA PAY BANDS (LPA, MID-2026, SOURCED RANGES)

Analyst ₹4–6 fresher → ₹15–20 senior · Scientist ₹6–9 → ₹20–35 (₹60+ AI-specialized) · ML Engineer ₹10–18 → ₹1–2 Cr at top GCCs. Metro adds +15–30%; GenAI skills add a further +60–120%.

CHAPTER 01 · CONTINUED

Growth is real — the entry door is narrower

+34%
BLS-projected data scientist growth 2024–34, 4th fastest US occupation

+113%
WEF-projected growth for big-data specialists by 2030 — the #1 fastest role

#1
AI engineer on LinkedIn's fastest-growing jobs for new grads, 2 years running

The catch: in 2026 the title "Data Scientist" skews mid-senior. Junior postings have shrunk while expectations rose. The growth is real — it just doesn't enter through the door marked "junior data scientist."

- **The realistic ladder:** Data Analyst (yrs 0–2) → Senior Analyst/Analytics Engineer (yrs 2–3) → Data Scientist or ML Engineer (yrs 3+).

WATCH OUT

"Apply to entry-level data scientist roles until one says yes" competes for the scarcest door in the building. Enter through analyst work instead.

CHAPTER 02

The statistics survival kit

- **Mean vs median:** a billionaire walks into a bar — the average patron is now a millionaire, the median patron hasn't moved. Report the median for skewed data like income or house prices.
- **Standard deviation:** the average distance of data from the mean. Two products can both average 4.0 stars — one is reliably fine, the other is polarizing. SD is the number that tells them apart.
- **The 68/95/99.7 rule:** for bell-shaped data, ~68% of values sit within 1 SD, ~95% within 2, ~99.7% within 3. A value 3 SDs out is a once-in-370 event, worth a second look.
- **The Central Limit Theorem:** sample means from almost any distribution pile up into a normal bell curve — the reason A/B tests work even on wildly skewed metrics like revenue per user.

FAT-TAILS WARNING

Stock returns, viral content, incomes and city sizes are fat-tailed — extreme events happen far more than a bell curve predicts. Always plot the distribution before trusting it.

CHAPTER 02 · CONTINUED

Correlation, p-values & the data you don't have

- **Correlation ≠ causation:** ice cream sales and drownings rise together — both follow summer. Every correlation has 3 explanations: A causes B, B causes A, or a hidden C causes both.
- **P-values, honestly:** a p-value is the probability of data this extreme if nothing real were happening — not "the probability you're right." The ASA formally declared p-values "commonly misused" in 2016.
- **Confidence intervals beat verdicts:** "lift between +0.2% and +3.8%" is information; "significant: yes/no" is just a verdict.
- **Sampling bias:** WWII analysts nearly armored the wrong spots on bombers — the planes that came home weren't a random sample. Ask: who never made it into this table?

HUMBLING DATA POINT

At Google and Bing, only 10–20% of experiments produce positive results. If your ideas keep "winning," suspect your testing before you celebrate. — Ron Kohavi, HBR 2017

CHAPTER 03

The algorithm atlas I — supervised

- **Linear regression:** the best straight line through your data — the slope is the story ("every extra ₹1,000 of ad spend adds ~40 sales"). Used by 80%+ of working data scientists per multi-year Kaggle surveys.
- **Logistic regression:** a classifier, despite the name — squashes a score into a probability. Regulators love it because it's explainable: still the default in credit scoring.
- **Decision trees:** a flowchart of if-then questions learned from data. Perfectly readable — but a single tree overfits, memorizing the training data.
- **Random forest:** hundreds of trees, each on a random slice of rows and columns, then they vote. The strong tabular default with almost no tuning.

LEARN THESE FIRST

The boring algorithms pay the bills. Run linear/logistic regression first — every fancier model must beat it to earn its complexity.

Boosting, KNN & naive Bayes

- **Gradient boosting (XGBoost):** trees built in sequence, each fixing the last one's mistakes. The competition-winning, industry-default choice for structured tabular data — fraud, churn, ranking, pricing.
- **KNN:** no learning at all — judge a new point by copying the majority answer of its k most similar past examples. Good for small data and quick baselines.
- **Naive Bayes:** multiplies little independent clues (e.g. "free," "winner," "urgent" in spam) — the independence assumption is wrong, and it still wins surprisingly often on text.

SOURCED VERDICT

Grinsztajn, Oyallon & Varoquaux (2022) found tree-based models like XGBoost still beat deep neural networks on typical tabular benchmarks. On business tables, trees rule.

CHAPTER 04

The algorithm atlas II — unsupervised

- **K-means clustering:** finds k groups hiding in data with no labels — an e-commerce team can discover bargain hunters, loyal spenders and one-time buyers nobody defined in advance.
- **Hierarchical clustering:** builds a dendrogram — a family tree of merges — so you never have to choose k upfront. The tree itself is the insight.
- **PCA:** rotates 50 correlated columns into a handful of "super-columns" carrying most of the information — great for compression, bad for explaining "what a feature means" to a regulator.
- **Picking k:** the elbow method (plot error vs k, find the bend) and the silhouette score are guides, not oracles — always sanity-check clusters against business meaning.

EXAMPLE

30 correlated athlete body measurements collapse via PCA into two components meaning roughly "overall size" and "build type." Two numbers now describe what thirty did.

SVM, neural nets & the 2026 verdict

- **SVM:** finds the boundary with the widest possible margin between classes — the "widest street." Lost mindshare to boosting and deep learning, but still appears in interviews and text/bio niches.
- **Neural networks:** layers of simple units stack up to learn complex patterns — edges → ears → cat, with no manual feature engineering. Nothing else comes close on images, audio and text.
- **When classical ML still wins:** tabular data, under ~100K rows, interpretability/audits required, or limited compute — reach for boosting and trees, not deep learning.

<40%	74%	80%
LLM accuracy on 293 real data-science coding tasks, direct	With an iterative AI agent — better, still not autonomous	Of researchers' solutions used LLM code assistance

THE 2026 CONTEXT

LLMs write the scikit-learn code now. Someone still has to know if it's right — the syntax gap is closing, so judgment moves up the stack.

CHAPTER 05

Which chart, when — four questions

- **Comparison:** "which is bigger / how did it change?" → bar chart for items (must start at zero), line chart for change over time.
- **Composition:** "what are the parts of the whole?" → stacked bar or treemap. Pie only with ≤5 slices and one obvious dominant slice.
- **Distribution:** "what shape does one variable have?" → histogram for shape, box plot to compare shapes across groups.
- **Relationship:** "do these variables move together?" → scatter plot; bubble size for a third variable; heatmap for a whole correlation matrix.
- **Zero-chart option:** need exact values? use a table. Only one number matters? print it big. A chart you don't need just slows the reader down.

CHART CRIME #1

Truncating a bar's axis at 90 instead of 0 turns a real 2% gap into what looks like a 3× rout. Bars always start at zero — only line charts may zoom, with labeled axes.

CHAPTER 06

The real workflow

- **Where the hours go:** ~60% cleaning + ~19% collecting data (CrowdFlower 2016); Anaconda surveys find ~45% on prep. Modeling is typically only 10–20% of the job.
- **Train/test splits:** never grade a student on memorized homework — hold back 20–30% of data the model never sees. For time series, always split by time, never randomly.
- **Overfitting:** 99% training accuracy vs 70% test accuracy is the classic memorization signature. Cures: more data, simpler models, regularization, early stopping.
- **Metrics encode costs:** precision when false alarms are costly (spam filters); recall when misses are costly (cancer screening, fraud). Accuracy alone is a trap on imbalanced data — 99% "accuracy" while catching zero fraud.

THE ONE-LINE LEAKAGE RULE

"If a feature won't exist at prediction time, it can't be in the model." Data leakage produces no error message — just a beautiful score that collapses in production.

CHAPTER 07

The toolkit 2026

- **Python won:** 57.9% of all developers now use it — a +7-point single-year jump in 2025, the largest in Python's history. Learn pandas, NumPy, scikit-learn, matplotlib/seaborn — the core four.
- **SQL is non-negotiable:** the one skill all five data roles share. It appears in more data-role postings than any single Python library — and interviews test it first.
- **Notebooks vs scripts:** Jupyter is the lab bench for exploration, not the factory for production — hidden state and out-of-order cells are the trap. Explore in notebooks, ship as scripts.
- **BI tools:** Power BI for the biggest job pool (especially strong in India), Tableau for visualization craft, Looker for Google-stack shops.

WHERE LLMS GENUINELY HELP

Boilerplate pandas/SQL, explaining errors, drafting first-pass EDA. Not good at: autonomous analysis, knowing when it's wrong, the judgment calls.

CHAPTER 08

Breaking in

- **Delete the anti-list:** Titanic, Iris, MNIST, Boston housing — recruiters read all of them the same way: "can follow a tutorial." ~78% of applicant portfolios are exactly these.
- **The 3–5 project formula:** one strong EDA piece · two ML projects of different types · one deployed, clickable end-to-end project · business-impact READMEs on all of them.
- **Certifications, ranked honestly:** certs get interviews, portfolios get offers. One comprehensive cert (Google Data Analytics) + one tool cert (Power BI/dbt), maximum.
- **The SQL screen:** 3–6 questions in 30–60 minutes; window functions (ROW_NUMBER, RANK, LAG, LEAD) appear in ~40% of the hard ones. The classic trap: you can't filter a window function in WHERE.

THE REALISTIC TIMELINE

6–12 focused months to analyst-ready. The "Data Scientist" title usually takes 2–3 years — it's earned through a role, not a course.

YOU'VE GOT THE MAP. NOW GET THE TERRAIN.

This was the 20-minute version. The full atlas is the whole journey.

You now know more than most people learning data science by watching random videos. The complete **Data Science Pocket Atlas** takes every chapter above and gives you the diagrams, the full sourced tables, and the exact frameworks.

- **The full 8 chapters** — every algorithm, stat and workflow step, illustrated and expanded.
- **12 algorithms explained visually** with strengths, weaknesses and a one-line summary each.
- **The complete chart-choice decision guide** — all 24 rules, plus the six chart crimes illustrated.
- **The India + US salary tables** in full, with the growth data sourced across BLS, WEF and LinkedIn.
- **The FAQ & 20-term glossary** — the questions every beginner actually asks, answered straight.

UPGRADE TO THE FULL EDITION

~~₹199~~ **₹99**

Complete illustrated atlas · instant download

SCAN / TAP
to upgrade

GO FURTHER, PAY LESS

Two ways to keep going

THE AI & TECH STACK BUNDLE

Data Science + AI Engineer Roadmap + Prompt Engineering + Tech Career

Everything you need to understand and grow inside AI & tech, in one download.

~~₹396~~ **₹199**

THE COMPLETE LIBRARY

All 30 Skill Series books

Spirituality, money, career, health, tech — the entire collection, one price.

₹499

GET EVERY NEW BOOK FREE

Follow **@weknowwhatyourelookingfor** and join our WhatsApp list — buyers get first access and reader-only pricing on everything we publish next.

weknowwhatyourelookingfor.com

NOTICE

This guide is for **general educational purposes only**. Individual results vary with effort and circumstances, and no specific outcome is warranted.

Published by weknowwhatyourelookingfor.com · For queries, corrections, or licensing, visit weknowwhatyourelookingfor.com.