Data Science Roadmap (start here)

I have a coding/computer science background

I have a math/stats or science background

I am totally new to this and just want to get started!

Where to focus

bold - essential to get a job in data science

\+ Total newbs start with these

Add the math and you are all set!

Focus on building core coding skills

Ok, you can do this! start with python and then everything with a +

Code skills

This is the foundation, just google/youtube each of these to get started!!

- hello world+
- **data types+**
- **arrays+**
- **If statements**
- **for loops+**
- **indexing+**
- Regular expressions

Programming languages

If you can only learn one, focus on python. It's free and amazingly powerful and well supported

- **python+**
- **SQL**
- R*
- Matlab*

Python Tools

These are the primary data science packages in python.

- **pandas+**
- **numpy+**
- scipy
- keras
- sqlite
- geopy
- **matplotlib+**
- seaborn
- bokeh
- **scikit learn**
- NLTK

Scientific method

Knowing when you "have something" vs just randomness

- **A/B testing**
- **correlation vs causation**
- Confounding variables
- **over-fitting**
- **test/train split**

Programming tools

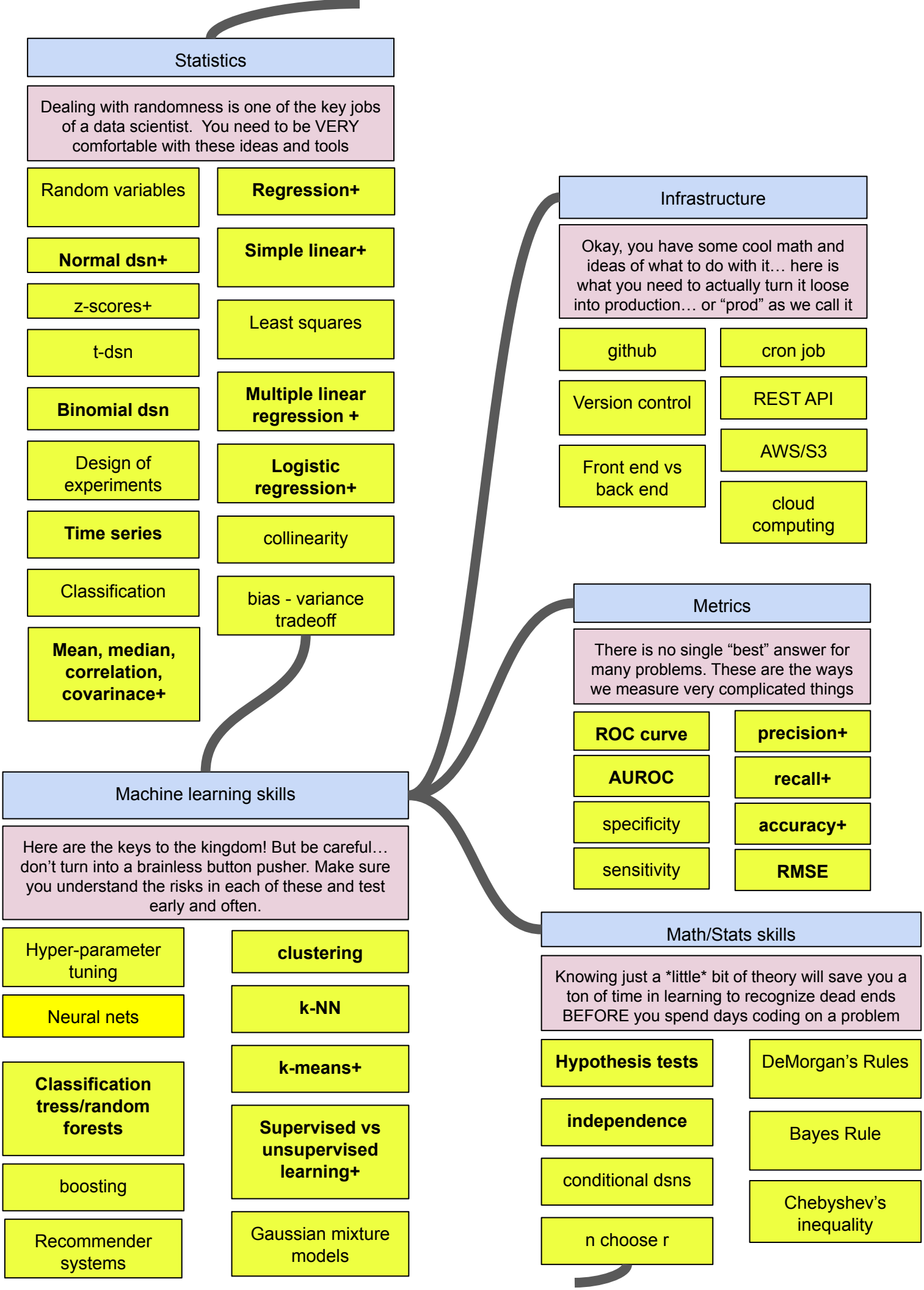
You don't need to be good at all of these, but pick a couple

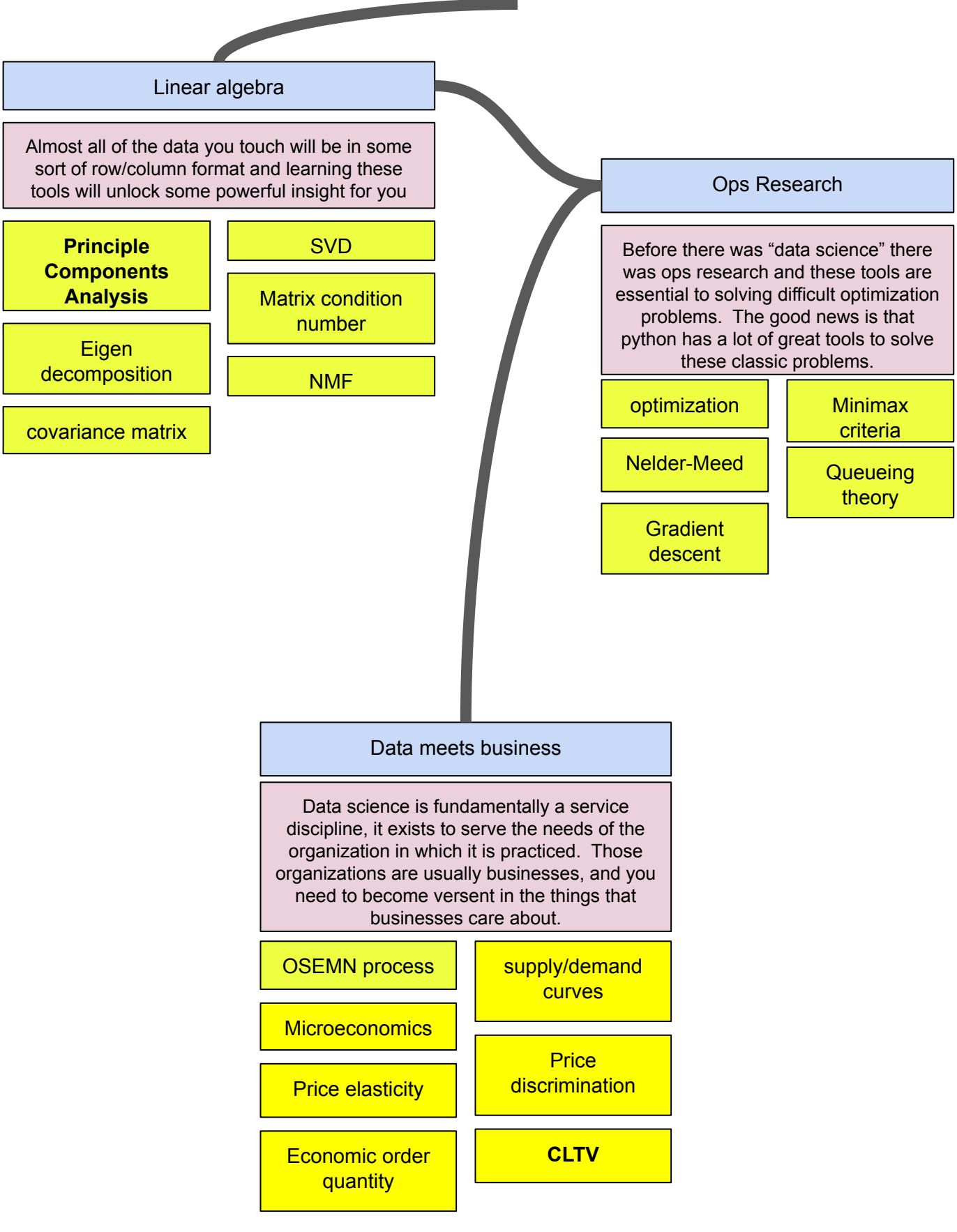
- IDEs
- Anaconda+
- spyder
- Jupyter notebooks+

Data wrangling

Data is often "messy" - weed it, like a garden

- **Pivot Tables**
- **pandas I/O tools**
- groupby
- melt
- datetime
- Anomaly detection
- cleaning/ outlier removal
- **Data cleaning**
- **Boolean indexing**


Statistics
Dealing with randomness is one of the key jobs of a data scientist. You need to be VERY comfortable with these ideas and tools
Random variables
Normal dsn+
z-scores+
t-dsn
Binomial dsn
Design of experiments
Time series
Classification
Mean, median, correlation, covarinace+
Regression+
Simple linear+
Least squares
Multiple linear regression +
Logistic regression+
collinearity
bias - variance tradeoff
Infrastructure
Okay, you have some cool math and ideas of what to do with it… here is what you need to actually turn it loose into production… or "prod" as we call it
github
Version control
Front end vs back end
cron job
REST API
AWS/S3
cloud computing
Metrics
There is no single "best" answer for many problems. These are the ways we measure very complicated things
ROC curve
AUROC
specificity
sensitivity
precision+
recall+
accuracy+
RMSE
Machine learning skills
Here are the keys to the kingdom! But be careful… don't turn into a brainless button pusher. Make sure you understand the risks in each of these and test early and often.
Hyper-parameter tuning
Neural nets
Classification tress/random forests
boosting
Recommender systems
clustering
k-NN
k-means+
Supervised vs unsupervised learning+
Gaussian mixture models
Math/Stats skills
Knowing just a *little* bit of theory will save you a ton of time in learning to recognize dead ends BEFORE you spend days coding on a problem
Hypothesis tests
independence
conditional dsns
n choose r
DeMorgan's Rules
Bayes Rule
Chebyshev's inequality


Linear algebra
Almost all of the data you touch will be in some sort of row/column format and learning these tools will unlock some powerful insight for you
Principle Components Analysis
SVD
Matrix condition number
Eigen decomposition
NMF
covariance matrix
Ops Research
Before there was “data science” there was ops research and these tools are essential to solving difficult optimization problems. The good news is that python has a lot of great tools to solve these classic problems.
optimization
Minimax criteria
Nelder-Meed
Queueing theory
Gradient descent
Data meets business
Data science is fundamentally a service discipline, it exists to serve the needs of the organization in which it is practiced. Those organizations are usually businesses, and you need to become versent in the things that businesses care about.
OSEMN process
supply/demand curves
Microeconomics
Price discrimination
Price elasticity
Economic order quantity
CLTV