**Data Types, Sources and Cleaning**

* Calendar conventions
  + Days: 365 or 360
  + Trading days: 252 or 250
  + Months: 30 days per month, calendar days per month
  + Weeks: 5 day week, calendar week
* Risk free rates
  + US treasury bills - 1, 3, 12 months
  + US treasury bonds - 5, 10, 20, 30 years
  + Libor
* Bonds
* Simple rate of return
* Log rate of return
* Sources
  + Mathematica
  + Yahoo finance
  + Fidelity
  + US Federal Reserve
  + Treasury.gov
  + Etfdb.com
* Missing, incomplete, censored data
  + Drop incomplete observations
  + Use unconditional mean of nearest neighbors
  + Use conditional mean of nearest neighbors (or all data). Better for correlated variables than unconditional mean. Multiple regression. This can be difficult depending on the distribution of missing values in the observations and variables. But this estimates a mean, ignoring the variance and covariance that would be added by the actual observation.
  + There is a process to add the missing variance and covariance using the error residuals from the regression
  + EM algorithm - not for estimating the missing values but for estimating parameters (e.g. mean, covariance) when using data that has missing values)

**Prediction - Statistical, Fundamental, Technical**

* Equities
  + Dividend discount model - constant, 1 stage, 2 stage, 3+ stages
  + Free cash flow discount model
* Market portfolio
  + Theoretical construct proportionally holding all risky assets
  + Proxies (indexes, benchmarks) depend on the universe of assets of interest
  + There are usually many proxies available
  + US equities: S&P 500 Index
  + World equities: MSCI World Index
* Capital asset pricing model (CAPM)
  + Decomposes asset risk into systematic and idiosyncratic components
  + Provides a theoretical rate of return for an asset based on idiosyncratic risk (can't be diversified away, no premium provided for diversifiable risk)
  + Regression of asset return to market return
  + Beta is the regression coefficient. There is also an analytic solution.
  + Alpha is the intercept (constant) term of the regression
  + Can be extended to more predictors besides market return (Fama-French, Carhartt, etc.)
* Factor models
  + Performance analysis - determine alpha of a portfolio
  + Style analysis - determine the contribution of each factor to a portfolio's return
* Derivatives
  + Futures
  + Options
    - Put-Call Parity
    - Implied volatility
    - Greeks - delta, gamma, theta, rho, vega, charm
    - Trading strategies - straddles, strangles, vertical spreads
    - Synthetic positions - using options to create a position that would otherwise not be allowed
  + Stochastic Calculus and Processes
  + Pricing with Monte Carlo simulation and risk neutral measure
  + Pricing with binomial lattice
  + Impact of dividends
* Currencies
* Commodities
* Interest rates
* Stochastic volatility
* Mixture distributions
  + Each component of a portfolio has its own distribution
  + Portfolio return distribution is weighted average of component distributions

**Optimization**

* Markowitz's Modern Portfolio Theory - no risk free asset
  + Returns modeled by multivariate normal distribution
  + Portfolio risk measured by portfolio variance (or standard deviation)
  + Portfolio reward measured by portfolio expected return
  + Minimize risk or maximize return
  + Subject to target return or target risk
  + Analytic solution for long/short portfolio
  + Quadratic programming for long only portfolio
* Markowitz's Modern Portfolio Theory - with risk free asset
  + Allow for a leveraged (borrowing against) (-inf, 1) position in the risk free asset
  + Tangent portfolio: maximizes Sharpe ratio
  + Increase/decrease risk/reward payoff by increasing/decreasing borrowing in risk free asset. This is a line through the risk free portfolio and the tangent portfolio.
  + Analytic solution for long/short tangent portfolio
  + Quadratic programming for long only portfolio
* Markowitz's Modern Portfolio Theory - CAPM optimization
  + From CAPM, covariance matrix decomposed into systematic and idiosyncratic components
  + Idiosyncratic matrix D is a diagonal matrix composed of each asset's variance
  + Systematic matrix is the outer product of the vector of betas with itself
  + Massively decreases number of parameters which is good if not enough data is available to estimate sample covariance matrix
  + Analytic solution exists
* Portfolio optimization with missing data and factor models
  + Estimate missing data and variance/covariance corrections
  + Estimate mean and covariance using EM algorithm
  + Fit factor model using estimated mean and covariance and EM algorithm
  + Denoising the correlation matrix - improves the condition number (thus numerical inversion stability). Replace all eigenvalues below the Marchenko-Pastur distribution parameter range upper bound with their average and reconstruct the covariance matrix
* Improving out of sample performance - caused by over fitting or sample error
  + Enforcing diversification: sampling errors can push optimization towards overweighting and underweighting certain positions. Can fix by adding penalty term to objective function that can be tuned using cross validation. Can also add upper bounds to the positions.
  + Shrinkage estimation: the estimated mean or covariance is a convex combination of the sample estimate and some other estimate. A common choice is the James-Stein shrinkage estimate
  + Combining estimates: for example, a history estimate and factor model estimate weighted by their individual sampling variances
* Numerical issues
  + Inverting
  + Large matrices
* Speed concerns
* Kelly criterion and volatility pumping

**Risk Management - Statistical, Numerical, Qualitative**

* Sensitivity analysis
  + Stress testing
* Risk measures
  + Coherent properties: translation invariance, subadditivity, positive homogeneity, montonocity
  + VaR
  + CVaR
  + Leverage
  + Drawdown
    - Length of drawdown - Lomax distribution?
    - Depth of drawdown - Lomax distribution?
  + CDaR
  + Drawdown Beta
* Types of risk
  + Event risk
* Hedging
  + Options
    - Hold long position and buy puts
    - Hold short position and buy calls
    - Straddles
    - Strangles
  + VIX

**Trading Execution**

* Explicit transaction costs
  + Fixed
  + Variable
* Implicit transaction costs
  + Is it worth it to make the trade? Would portfolio be worth more if you hold than if you rebalance and pay explicit costs?
* Trade impact
* Trade timing
  + Different volatilities and volumes during different times of day, days of week, months, seasons
  + When to open position (buy or short)
  + When to close position (sell or cover)
  + Derivative considerations - I'm right about my position but will it become in the money before expiration? Will it stay in the money or go out of the money?
* Slippage
* Order book

**Statistics, Calculus, Linear Algebra**

* Distributions
  + Normal - univariate, multivariate
  + Student T - univariate, multivariate
  + Non central Student T
  + Moments and central moments - mean, variance, skewness, kurtosis
  + Moment generating functions, characteristic functions
  + Log normal
  + Gamma
  + Copula
    - Copula function maps marginal distributions to a multivariate distribution
    - This is a way to introduce dependence to marginal distributions
  + Mixture distributions - can fit with EM algorithm
    - Number and type of component distributions known
    - Weight and parameters of each distribution unknown
    - Given observations and parameter estimates
    - Membership matrix telling which component distribution each observation was drawn from is unknown
    - E-step: Use observations and current parameter estimates to estimate membership matrix
    - M-step: Use observations and membership matrix estimate to produce updated parameter estimates
    - Repeat previous 2 steps until convergence (e.g. maximum iterations, log likelihood levels off for some number of iterations)
    - Sub-linear convergence and global maximum not guaranteed. May need to try from different starting parameter estimates
    - Issues with numerical stability
    - Can use penalized log likelihood to determine the best fit for different numbers and types of distributions. Different penalization approaches may lead to different results though
  + Elliptical distributions
  + Fitting distribution to data
  + Lomax - gamma/exponential mixture
* Empirical approach over estimating distribution
  + Must have enough data
  + Power law for extreme events
  + Empirical distribution
  + Kernel density estimation
* Measures of central tendency
  + Mean
  + Median
  + Mode
* Measures of dispersion
  + Variance
  + Standard deviation
  + Covariance
  + Mean absolute deviation
  + Median deviation
* Measures of shape
  + Skewness - distribution asymmetry
  + Kurtosis - distribution tailedness
* Measures of dependence
  + Linear Correlation
  + Kendall's Tau
  + Spearman's Rho
* Monte Carlo simulation
* Finite difference method
* Factor models
  + Multiple regression
  + EM algorithm
  + PCA - could result in portfolio with 0 variance
  + Scree plot - plot of eigenvalues in decreasing order, keep eigenvalues up to the elbow
* Matrices
  + Condition number - ratio of largest to smallest eigenvalues, high number is bad for inversion stability
  + Decompositions - cholesky, eigen decomposition, singular value decomposition
* Stochastic Calculus and Processes
  + Jumps
  + Stochastic volatility
* Issues with estimation
  + Mean blur - decreasing the standard error of the mean estimate requires 2n more data due to division by square root of observations. However, increasing the sampling frequency will not work since the mean decreases faster than the standard deviation
  + Variance non-blur - increasing frequency does reduce the standard error of the variance estimate

**Other**

* Arbitrage
* Risk free rate
* Present/Future value
* IRR: interest rate consistent with present value of 0 for a given set of cash flows
* Modelling
  + What does the model do?
  + What does the model leave out or misrepresent?
  + How can the model be used?
  + Know the domain of the problem
  + Errors aren't just noise and contain information
  + Use simple models to validate more complex models
  + Too many parameters with too little data leads to overfitting