1. Hypotheses & Metrics - **What to Test**: pages, funnel, business model, backend functionality, algorithm, new product/feature

- **Hypotheses**: $H_0$ = control & treatment equal, $H_a$ = different
- **Metrics**: Primary = biz objective, e.g. order conversion,  Guardrail = critical biz metrics, e.g. page load time,  Secondary = why primary changed
- **Tests**: 1. robust to extraneous vars (run A/A test)  2. proportions → use Z-test, e.g. CTR  3. avg/user → use T-test, e.g. avg rev/usr, avg posts/usr
- 4. multiple indep vars → use multivariate test (allows testing interaction between indep vars, but requires larger $n$)

2. Design A/B Test - **Population**: unit (usr id, events, cookies),  segment to sample from to reduce dev hrs (geography, device type, browser)

- **Cohort**: if funnel exclude inprogress usrs from test,  if test insignificant test by segment (e.g. behavior {visited 10+ times}, outcome {bought})
- **Sample Size**: determines test length, set ahead, don't stop early, $\pi$ analysis (α=.05, 1-β=.8, min detectable effect (MDE) that justifies feat cost)
- **Lehr's Formula**: approx samp size per group for 2-samp T-test /w equal $\sigma^2$, $n \approx \frac{16\sigma^2}{\delta^2}$ where $\delta$ is MDE =$\mu_0$ - $\mu_1$, approx $\sigma^2$ by $s^2$ using past data
- **Sampling Strategy**: probability sampling (random, stratified, clustered), when to run (season, weekday/weekend)

3. Analyze Results - **Change Aversion**: smaller initial effect    **Novelty**: larger initial effect    **Sol**: compare new usrs to veteran usrs in treatment group

- **Network Eff**: social networks & 2-side markets, treatment interferes /w control violates independ, sol: split groups by location, time, netwrk clusts
- **Simpson's**: trend when groups of data, no trend when groups aggregated → uncontrolled confounder, sol: segment data to see if trend persists
- **Multiple Testing**: rerunning test compounds FPR, sol: use ANOVA for multi-group tests, use Bonferroni Correction (divide $\alpha$ by num of tests)

4. Make Decision - **Independence**: Check usrs were randomized, compare distributions across groups (demographics, device, browser), use $\chi^2$-test

- **Significance Test**: check 1. if statistically significant (primary & guardrail),  and 2. if practically significant (a. lift = $\frac{\text{treatment rate} - \text{control rate}}{\text{control rate}}$,
- b. check effect size > MDE, is full confidence interval for $p_1 - p_2$ > MDE, if only partially rerun with larger $n$ to increase $\pi$)
- **Conflicting Results**: positive primary & negative guardrail → quantify both effects, determine net effect
- **Launch**: ramp up feature release to ensure change aversion or novelty effect aren't permanent