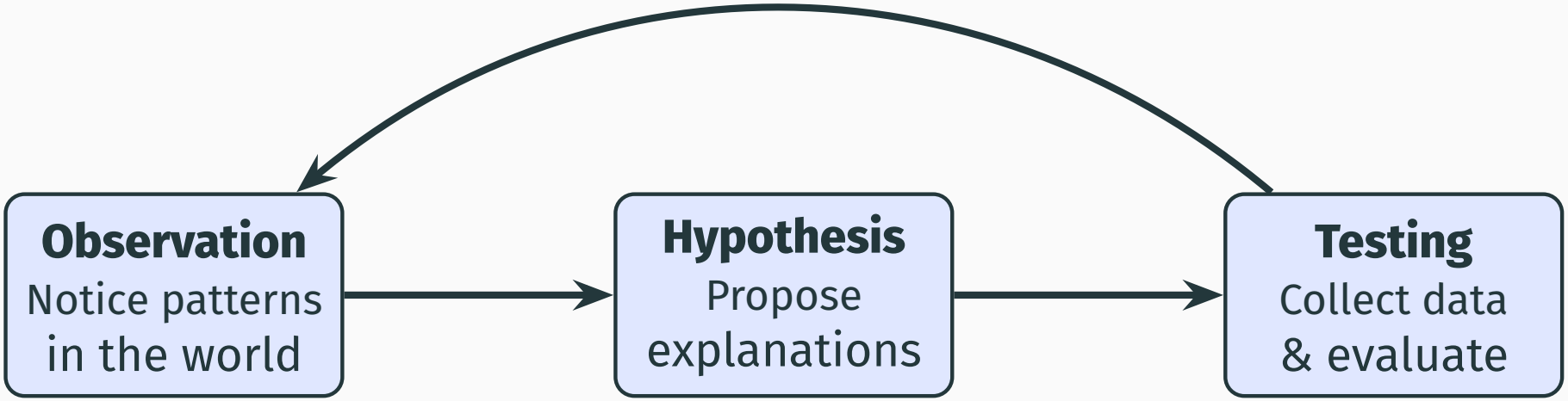
Part I: The Nature of Scientific Inquiry

Introduction to Statistics and Data Analysis

The Scientific Method as a Cycle



Science is iterative, not linear

Each cycle refines our understanding and generates new questions

Example: Discovery of *Helicobacter pylori*

Observation (1980s)
Most stomach ulcers occur in similar patterns; some patients don't respond to stress-reduction treatments

Hypothesis
Ulcers might be caused by bacterial infection, not just stress or spicy food

Testing
Collected stomach tissue samples, cultured bacteria, tested antibiotics on patients

Result
H. pylori confirmed as cause → paradigm shift in treatment → Nobel Prize 2005

The Role of Data as Evidence

Data serves as the **empirical bridge** between our ideas about the world and the world itself

Without Data

- Speculation and opinion
- Anecdote and intuition
- Unverified assumptions
- Competing narratives

With Data

- Evidence-based claims
- Systematic observation
- Testable predictions
- Reproducible findings

"In God we trust. All others must bring data."
— W. Edwards Deming

Building and Refining Scientific Knowledge

- **Incremental Progress**
  Each study adds a piece to the puzzle, building on previous work
- **Self-Correction**
  Replication and peer review help identify and correct errors
- **Increasing Precision**
  Better methods and more data → more accurate understanding
- **Paradigm Shifts**
  Occasionally, accumulated evidence forces revolutionary rethinking

Descriptive vs. Inferential Science

Descriptive Science

Goal: Characterize what is observed

Questions:

- What patterns exist?
- How much/many?
- What are the characteristics?

Example: Measuring the average height of students in this class

Inferential Science

Goal: Generalize beyond observations

Questions:

- Does this apply broadly?
- Is this effect real?
- What can we predict?

Example: Using class data to estimate average height of all university students

Most data analysis involves both: describe your sample, then infer about the population

Example: From Description to Inference

Descriptive Phase

In our trial of 500 patients, the treatment group (n=250) showed an average reduction of 15 mmHg, while the control group (n=250) showed 3 mmHg reduction

Example: From Description to Inference

Descriptive Phase

In our trial of 500 patients, the treatment group (n=250) showed an average reduction of 15 mmHg, while the control group (n=250) showed 3 mmHg reduction

Inferential Phase

Question: Is this 12 mmHg difference likely to be real for all similar patients, or just a fluke in our sample?

Statistical inference: Calculate probability that difference this large could occur by chance (p-value), estimate range for true effect (confidence interval)

Conclusion

If $p < 0.05$ and confidence interval doesn't include zero → we infer the drug likely works for broader population beyond just our 500 patients

Statistics Alone Cannot Determine Causes

Key Principle

Correlation does not imply causation

Classic Example: Ice Cream and Drowning

Observation: Ice cream sales and drowning deaths are strongly correlated

Statistical finding: High correlation ($r \approx 0.9$)

Statistics Alone Cannot Determine Causes

Key Principle

Correlation does not imply causation

Classic Example: Ice Cream and Drowning

Observation: Ice cream sales and drowning deaths are strongly correlated

Statistical finding: High correlation ($r \approx 0.9$) **Wrong conclusion:** Ice cream causes drowning (or vice versa)

Statistics Alone Cannot Determine Causes

Key Principle

Correlation does not imply causation

Classic Example: Ice Cream and Drowning

Observation: Ice cream sales and drowning deaths are strongly correlated

Statistical finding: High correlation ($r \approx 0.9$) **Wrong conclusion:** Ice cream causes drowning (or vice versa) **Actual mechanism:** Both increase during summer; temperature is the confounding variable

Why statistics alone isn't enough:

- Can identify associations but not causal direction
- Cannot distinguish direct from indirect effects
- Cannot reveal confounding variables without assumptions

Understanding Different Types of Relationships

Direct Causation

A causes B

Reverse Causation

B causes A

Understanding Different Types of Relationships

Direct Causation

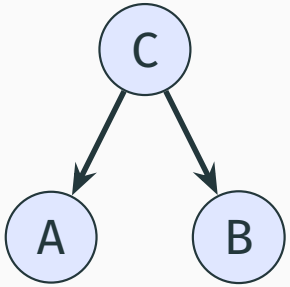
A causes B

Reverse Causation

B causes A

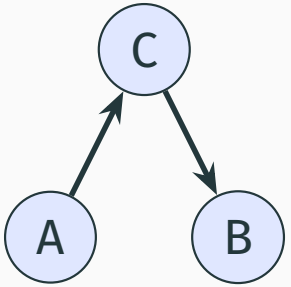
Basic Confounds

The fork



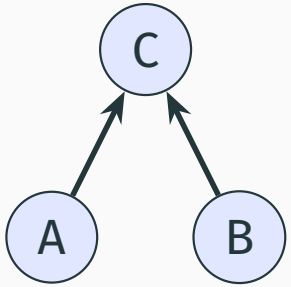
C is common cause to A and B

The pipe



C is a *mediator* of the effect of A on B

The collider



A and B cause C

Statistics identifies associations; models explain mechanisms

The Power of Mechanistic Models

What are Mechanistic Models?

Models that explicitly represent the **underlying processes** and **causal mechanisms** that generate observed patterns

Statistical Models:

- Describe patterns
- Make predictions
- Quantify uncertainty
- ”What” happens

Mechanistic Models:

- Explain processes
- Test mechanisms
- Guide interventions
- ”How” and ”why” it happens

Example: Epidemic Spread

Statistical: Cases increasing exponentially with rate r

Mechanistic: SIR model—susceptible individuals become infected through contact, then recover with immunity. This explains *why* the pattern changes over time.

Why Mechanistic Models Matter

- **Intervention Design**
  Understanding mechanisms tells us *where* to intervene
  Example: Knowing malaria transmission requires mosquitoes → target mosquito populations

Why Mechanistic Models Matter

- **Intervention Design**
  Understanding mechanisms tells us *where* to intervene
  Example: Knowing malaria transmission requires mosquitoes → target mosquito populations

- **Extrapolation**
  Mechanistic models generalize better to new situations
  Example: Physics equations work on Earth and Mars; purely statistical models trained on Earth data might not

Why Mechanistic Models Matter

- **Intervention Design**
  Understanding mechanisms tells us *where* to intervene
  Example: Knowing malaria transmission requires mosquitoes → target mosquito populations

- **Extrapolation**
  Mechanistic models generalize better to new situations
  Example: Physics equations work on Earth and Mars; purely statistical models trained on Earth data might not

- **Counterfactual Reasoning**
  Can answer ”what if” questions about things we haven’t observed
  Example: What would happen if we removed this gene? Changed this policy?

Why Mechanistic Models Matter

- **Intervention Design**
  Understanding mechanisms tells us *where* to intervene
  Example: Knowing malaria transmission requires mosquitoes → target mosquito populations

- **Extrapolation**
  Mechanistic models generalize better to new situations
  Example: Physics equations work on Earth and Mars; purely statistical models trained on Earth data might not

- **Counterfactual Reasoning**
  Can answer ”what if” questions about things we haven’t observed
  Example: What would happen if we removed this gene? Changed this policy?

- **Insight and Understanding**
  Reveals the fundamental principles governing a system
  Example: Newton’s laws explain both falling apples and planetary orbits

Integrating Statistics and Mechanism

Modern data analysis combines both approaches

The Workflow

1. Use **statistical methods** to identify patterns and associations
2. Develop **mechanistic hypotheses** to explain those patterns
3. Use **statistics** to test mechanistic predictions
4. Refine the mechanism based on statistical evidence
5. Iterate!

Example: Drug Development

- Statistics: Drug A correlates with improved outcomes
- Mechanism: Drug A inhibits enzyme X, which blocks pathway Y
- Prediction: Other enzyme X inhibitors should also work
- Statistical test: Clinical trials confirm mechanism-based prediction

Key Takeaways

1. Science is a **cyclical process** of observation, hypothesis, and testing—not a one-way street
2. Data provides the **empirical evidence** that grounds our scientific claims in reality
3. Scientific knowledge is **cumulative and self-correcting**, building over time through many studies
4. We use **descriptive** methods to characterize what we observe and **inferential** methods to generalize beyond our data
5. Statistics alone cannot determine causation—we need models that represent mechanisms
6. **Mechanistic models** explain how and why phenomena occur, enabling better predictions and interventions

Data analysis is the engine that powers this entire scientific process

The bridge between raw observations and scientific understanding

What is Data Analysis?

Definition

Data analysis is the process of **transforming raw data** into **meaningful insights** through systematic examination, cleaning, modeling, and interpretation

Data analysis is:

- The bridge between observation and understanding
- Both a **technical skill** and an **art of reasoning**
- Essential to making **evidence-based decisions**

Two Modes of Data Analysis

Exploratory Analysis

Goal: Discover patterns and generate hypotheses

Approach:

- Open-ended investigation
- Visualization-heavy
- Flexible methods
- Pattern recognition

Mindset:
"What's in this data?"

Example: Plotting gene expression data to identify clusters of co-regulated genes

Confirmatory Analysis

Goal: Test specific hypotheses rigorously

Approach:

- Pre-specified questions
- Formal statistical tests
- Control for error rates
- Hypothesis testing

Mindset:
"Is this effect real?"

Example: Testing whether Gene A expression differs significantly between treatment and control groups

Don’t Confuse Exploration with Confirmation

The Problem

Using the same data to both generate and test hypotheses leads to **false discoveries** and **inflated confidence**

Example: The Garden of Forking Paths

1. Explore data, notice that variable X correlates with outcome Y
2. Test this correlation on the same data → $p < 0.05$!
3. Conclude X causes Y with high confidence
4. Problem: You found the correlation by exploring many variables; the test is invalid

Best Practices:

- Use separate datasets for exploration and confirmation
- Pre-register hypotheses before confirmatory analysis
- Be transparent about which analyses were planned vs. exploratory
- Adjust for multiple comparisons when testing many hypotheses

Data Analysis as Technical Skill

The Technical Toolkit:

Statistical Methods

- Descriptive statistics
- Probability distributions
- Hypothesis testing
- Regression modeling
- Machine learning
- Bayesian inference

Practical Skills

- Data cleaning & wrangling
- Programming (R, Python)
- Visualization
- Database queries
- Reproducible workflows
- Version control

These are **learnable skills** that improve with practice
(We'll cover many of these throughout the course)

Data Analysis as Art of Reasoning

Beyond the formulas, data analysis requires:

- **Judgment** — Which method is appropriate for this question?
- **Creativity** — How can I visualize this pattern effectively?
- **Critical thinking** — Does this result make sense? What could go wrong?
- **Domain knowledge** — What do I know about the subject matter that informs analysis?
- **Communication** — How do I explain these findings clearly to others?
- **Skepticism** — Am I seeing a real pattern or being fooled by randomness?

Good data analysis combines **technical proficiency** with **thoughtful reasoning**

Example: When Technical Skills Meet Reasoning

Scenario: Income Data

You're analyzing household income in a neighborhood:
$35k, $42k, $38k, $45k, $40k, $2.5M, $37k, $41k

Technical calculation:

- Mean = $348,625
- Median = $40,500

Which is "correct"?

- Technical skill: Both are calculated correctly
- Judgment: Median better represents typical household (mean distorted by outlier)
- Critical thinking: Why is there one very high value? Data error or billionaire resident?
- Communication: Report both, explain why median is more meaningful here

Data Analysis and Uncertainty

Core Principle

All data analysis involves **uncertainty**—our goal is to quantify and communicate it honestly

Sources of Uncertainty:

- **Sampling variability** — We observe a sample, not the entire population
- **Measurement error** — Our instruments and methods aren't perfect
- **Model uncertainty** — Our models are simplifications of reality
- **Unknown confounders** — Variables we didn't measure might matter

"The only certainty is that nothing is certain" — Pliny the Elder

Good analysts embrace uncertainty rather than hide from it

Tools for Quantifying Uncertainty

How we express uncertainty in data analysis:

Confidence Intervals

Range of plausible values for a parameter
Example: "Average height is 170cm (95% CI: 168-172cm)"

P-values

Probability of observing data this extreme if null hypothesis is true
Example: "$p = 0.03$ means 3% chance of this result by chance alone"

Prediction Intervals

Range where future observations are likely to fall
Example: "Next patient's blood pressure: 120 mmHg (90% PI: 100-140)"

These tools help us distinguish signal from noise

Data Analysis Never Happens in Vacuum

Every analysis serves a purpose

Consider the context:

- **Who** will use these results?
- **What** decision needs to be made?
- **Why** is this question important?
- **When** do results need to be ready?
- **How** will findings be communicated?

Example: Clinical Trial Analysis

- **Who:** Doctors, patients, regulators
- **What:** Approve drug or not?
- **Why:** Patient lives at stake
- **Result:** Need higher standards of evidence, clearer communication of risks

Common Pitfalls in Data Analysis

1. **P-hacking**
   Trying many analyses until finding $p < 0.05$

2. **HARKing** (Hypothesizing After Results are Known)
   Pretending you predicted what you actually discovered

3. **Cherry-picking**
   Reporting only results that support your hypothesis

4. **Confusing correlation with causation**
   Assuming association implies causal relationship

5. **Ignoring assumptions**
   Applying methods without checking if assumptions are met

6. **Overfitting**
   Creating models that fit your data perfectly but predict poorly

7. **Survivorship bias**
   Analyzing only successful cases while ignoring failures

Best Practices for Rigorous Data Analysis

Before Analysis

- Clearly define your research question
- Pre-register your analysis plan (when possible)
- Understand your data's origin and limitations

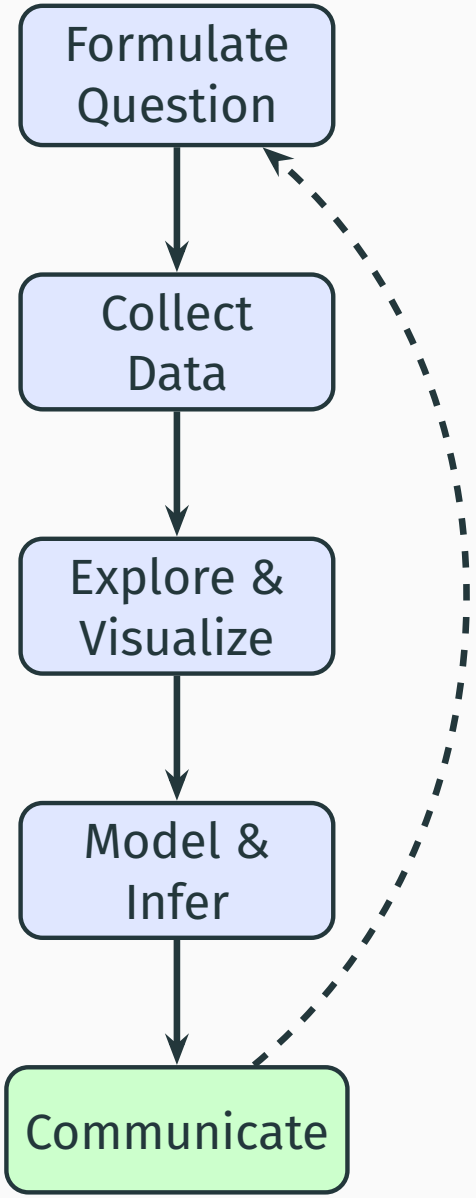
During Analysis

- Document every step (reproducibility!)
- Check assumptions of your methods
- Visualize data before modeling
- Consider alternative explanations

After Analysis

- Report all analyses performed, not just significant ones
- Be transparent about limitations
- Distinguish between exploratory and confirmatory findings
- Make data and code available (when appropriate)

Preview: The Data Analysis Pipeline



Key Takeaways

1. Data analysis is the **bridge** between raw observations and scientific understanding
2. **Exploratory** analysis discovers patterns; **confirmatory** analysis tests hypotheses—don't confuse them!
3. Data analysis requires both **technical proficiency** and **thoughtful reasoning**
4. All analysis involves uncertainty—our job is to quantify and communicate it honestly
5. Beware common pitfalls: p-hacking, cherry-picking, confusing correlation with causation
6. Good analysis is **transparent**, **reproducible**, and considers the broader context

Data analysis is both a science and an art —master both!

The deep questions underlying statistical inference

Why Philosophy Matters for Data Analysis

The Core Question

How do we gain reliable knowledge from limited, uncertain observations?

Philosophical questions shape practical decisions:

- How much evidence is enough to conclude something is real?
- What does ”probability” actually mean?
- Can we ever prove causation from observational data?
- How should prior knowledge influence our conclusions?
- What makes one explanation better than another?

These aren’t just abstract questions—they affect how we design studies, analyze data, and interpret results in biology every day

A. Epistemology: How Do We Know What We Know?

Epistemology
The branch of philosophy concerned with the nature, sources, and limits of knowledge

Key Questions for Scientists:

1. What counts as evidence?
2. How do we justify believing our conclusions?
3. What are the limits of what we can know from data?
4. How certain do we need to be before acting?

In Biology
When we sequence a genome, how do we know the sequence is correct?
When we observe a correlation, how do we know it's real and not chance?

The Problem of Induction

Inductive Reasoning
Drawing general conclusions from specific observations

The Logic:

1. Observed: Swan 1 is white
2. Observed: Swan 2 is white
3. Observed: Swan 3 is white
4. ... (1000 swans)
5. Conclude: All swans are white

The Problem:

- We can't observe *all* swans
- Next swan could be black
- No matter how many white swans we see, we can't be *certain*
- Yet science relies on this!

(And indeed, black swans exist in Australia—discovered 1697)

The Problem of Induction in Biology

Example: Drug Testing

Observation: New drug reduces tumor size in 100 mice

Inductive inference: Drug will reduce tumors in humans

The problem:

- We only tested 100 mice, not all mice
- We tested mice, not humans
- We tested under specific conditions
- Future trials might fail

How do we proceed despite uncertainty?

- Use statistical inference to quantify uncertainty
- Require replication across studies
- Test in multiple model systems before humans
- Accept that science gives us provisional knowledge, not certainty

Uncertainty is Inherent in Empirical Knowledge

There is no escape from uncertainty in science

Why statistical thinking is essential:

- We can never observe everything (all cells, all organisms, all conditions)
- Biological systems are inherently variable
- Measurements contain error
- Chance events occur

The Solution

Statistics provides tools to:

- **Quantify** how uncertain we should be
- **Distinguish** real patterns from random noise
- **Make decisions** with explicit error rates

Statistics is the grammar of science in the face of uncertainty

“All models are wrong,
but some are useful”

— George E.P. Box

What This Means

- Every model (statistical or mechanistic) is a simplification of reality
- No model captures every detail—nor should it
- The question is not ”Is this model true?” but ”Is it useful?”
- Good models capture essential features while ignoring irrelevant complexity

Like a map: it’s not the actual terrain, but it helps you navigate

Statistical Models in Biology

Example: Linear Regression for Gene Expression

Model: Expression = β_0 + $\beta_1 \times$ Temperature + error

What the model assumes:

- Linear relationship
- Constant variance
- Independent observations
- Normally distributed errors

Reality:

- Relationship might be curved at extremes
- Variance might change with temperature
- Gene networks create dependencies
- Distributions might be skewed

Is the model wrong? Yes. **Is it useful?** Often!
It captures the main trend and lets us make predictions

Choosing Between Models

How do we decide which model to use?

1. **Purpose** — What question are we trying to answer?
2. **Assumptions** — Which model's assumptions are most reasonable?
3. **Fit vs. Simplicity** — Balance between accuracy and parsimony
4. **Interpretability** — Can we understand what the model tells us?
5. **Generalizability** — Will it work on new data?

Occam's Razor

"Entities should not be multiplied beyond necessity"

When two models explain the data equally well, prefer the simpler one

Example: The Complexity Trade-off

Simple Model:
Growth = a + $b \times$ Nitrogen

Pros:

- Easy to interpret
- Few parameters
- Stable predictions

Cons:

- Misses saturation effect
- Ignores other nutrients

Complex Model:
Growth = f(N, P, K, pH, temp, light, water, ...) with interactions

Pros:

- More realistic
- Better fit to training data
- Captures interactions

Cons:

- Hard to interpret
- Many parameters
- Might overfit

The art: Finding the sweet spot between simplicity and realism

C. Objectivity and Subjectivity in Analysis

The Paradox

Science aims for **objectivity**, but data analysis involves countless **subjective decisions**

Subjective choices analysts make:

- Which variables to measure
- How to define/categorize variables
- Which data points to exclude (outliers?)
- Which statistical test to use
- How to transform data
- Significance threshold ($\alpha = 0.05$?)
- How to visualize results

These are called "researcher degrees of freedom"

Example: Analyzing Cell Morphology

Scenario: Do cells change shape in response to drug?

Choice 1: How to measure shape?

- Area? Perimeter? Aspect ratio? Circularity? All of them?

Choice 2: Which cells to include?

- Only complete cells? What about cells touching image border?
- How to handle dividing cells?

Choice 3: How to handle outliers?

- Exclude cells $>$ 3 SD from mean? Or keep all?

Choice 4: Which statistical test?

- t-test? Wilcoxon test? Linear model with covariates?

Each choice is defensible—but different choices can lead to different conclusions!

The Replication Crisis in Science

The Problem

Many published findings fail to replicate when other labs try to reproduce them

Contributing factors:

- **P-hacking:** Trying many analyses until finding $p < 0.05$
- **HARKing:** Hypothesizing After Results are Known
- **Publication bias:** Journals prefer positive results
- **Flexibility in analysis:** Using researcher degrees of freedom to get desired result
- **Low statistical power:** Small samples lead to unreliable estimates

Example from Psychology

Study found that listening to ”When I’m Sixty-Four” made people younger
(Result from p-hacking and selective reporting—obviously false)

Solution 1: Pre-registration and Transparency

Pre-registration

Publicly specify your hypotheses, methods, and analysis plan before collecting/analyzing data

What to pre-register:

- Research questions and hypotheses
- Sample size and stopping rules
- Which variables will be analyzed
- Primary vs. secondary outcomes
- Statistical tests and significance thresholds
- Criteria for excluding data

Benefits:

- Prevents p-hacking and HARKing
- Distinguishes confirmatory from exploratory analyses
- Increases trust in results

Platforms: OSF (osf.io), AsPredicted, ClinicalTrials.gov

Solution 2: Open Science Practices

Make research transparent and reproducible:

Open Data
Share raw data (when ethically possible) so others can verify analyses

Open Code
Share analysis scripts so methods are completely transparent

Open Materials
Describe methods in enough detail for others to replicate

Example in Biology
Publishing sequencing data to GEO/SRA, sharing microscopy images, providing detailed protocols, making analysis code available on GitHub

Transparency is the best antidote to bias

Solution 3: Better Reporting Standards

What journals and reviewers increasingly expect:

- Report all outcome measures, not just significant ones
- Distinguish pre-planned from exploratory analyses
- Report effect sizes with confidence intervals, not just p-values
- Show data distributions, not just summary statistics
- Acknowledge limitations honestly
- Include negative results
- Provide enough detail for replication

Guidelines

- CONSORT (clinical trials)
- ARRIVE (animal research)
- STROBE (observational studies)
- PRISMA (systematic reviews)

Balancing Objectivity and Subjectivity

We cannot eliminate subjectivity—but we can manage it

Best practices:

1. **Acknowledge** that subjective choices exist
2. **Make choices explicit** through transparent reporting
3. **Pre-commit** to choices when possible (pre-registration)
4. **Test robustness** by trying alternative analysis approaches
5. **Seek independent replication** as ultimate test

The goal is not to be perfectly objective (impossible)
but to be transparent and honest about our choices

Three Key Philosophical Tensions

1. **Induction vs. Certainty**
   We must generalize from limited data, accepting uncertainty is inevitable

2. **Simplicity vs. Realism**
   Models must be simple enough to understand yet complex enough to be useful

3. **Objectivity vs. Subjectivity**
   Science aims for objectivity but requires subjective judgment throughout

Good scientists navigate these tensions thoughtfully
rather than pretending they don't exist

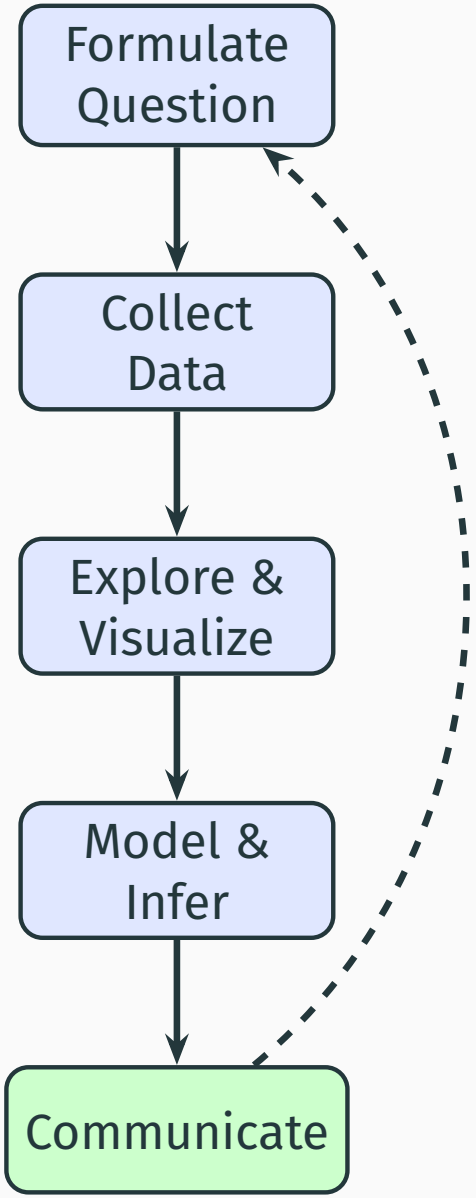
Key Takeaways

1. **Induction problem:** We can never be certain, but statistics helps us quantify uncertainty
2. **Models are simplifications:** The question isn't "Is it true?" but "Is it useful?"
3. **All analysis involves subjective choices**—researcher degrees of freedom are real
4. **Flexibility in analysis** can lead to false discoveries (p-hacking, HARKing)
5. **Solutions:** Pre-registration, transparency, open science, better reporting
6. **Accept uncertainty and subjectivity**, but manage them through rigor and honesty

Understanding these foundations makes you a better, more thoughtful analyst

From question to insight

The Data Analysis Pipeline



A. Formulating Questions

Why this matters

Your research question determines everything that follows:
study design, data collection, statistical methods, interpretation

Characteristics of good research questions:

- **Clear and specific** — Not vague or ambiguous
- **Answerable** — Can be addressed with available methods
- **Relevant** — Matters to science or society
- **Feasible** — Realistic with available resources
- **Well-defined** — Terms and concepts clearly specified

"A problem well stated is a problem half solved" — Charles Kettering

Types of Research Questions

Descriptive:

- What is the average?
- How common is X?
- What patterns exist?

Example: What is the mutation rate in *E. coli* under normal conditions?

Comparative:

- Does A differ from B?
- Is group X different?

Example: Do knockout mice grow slower than wild-type?

Relational:

- Are X and Y associated?
- Does X predict Y?

Example: Is gene expression correlated with metabolite levels?

Causal:

- Does X cause Y?
- What effect does X have?

Example: Does this drug reduce tumor growth?

test

Causal questions require special study designs!

Causal vs. Associational Questions

Scenario: Exercise and Longevity

Association: People who exercise live longer (observational study)

Can we conclude exercise causes longer life?

Possible confounders:

- Health-conscious people both exercise AND eat better
- Wealthy people can afford gym memberships AND better healthcare
- Genetics might influence both activity levels and health

To establish causation, we need:

- Randomized controlled trials (when ethical/feasible)
- Careful control of confounders
- Mechanistic evidence
- Temporal ordering (cause precedes effect)
- Dose-response relationships

No amount of clever analysis can fix a poorly designed study

Key design considerations:

1. **Experimental vs. Observational**
2. **Sample size and power**
3. **Randomization and controls**
4. **Blinding**
5. **Replication**
6. **Measurement quality**

Investment in good design pays enormous dividends later

Experimental vs. Observational Studies

Experimental Design

Researcher manipulates the variable of interest

Strengths:

- Can establish causation
- Control confounders
- Test mechanisms

Weaknesses:

- May be unethical
- Artificial conditions
- Time/cost intensive

Example: Randomly assign mice to drug vs. placebo

Observational Design

Researcher observes without intervention

Strengths:

- Study natural variation
- Often more feasible
- Real-world relevance

Weaknesses:

- Cannot prove causation
- Confounding variables
- Selection bias

Example: Survey wild populations for genetic variants

Sample Size and Statistical Power

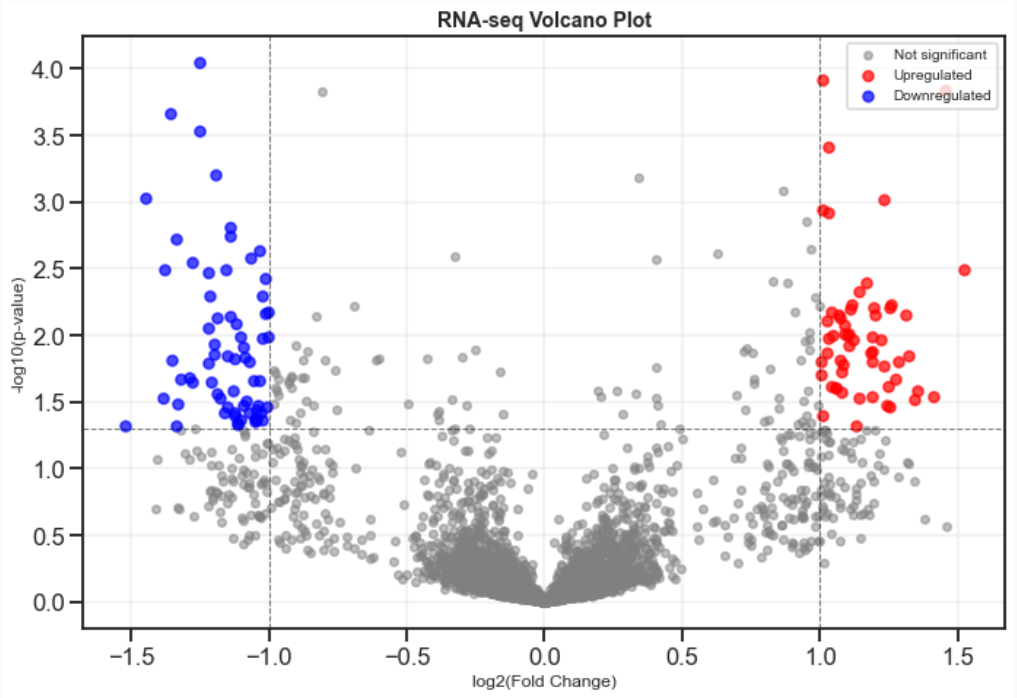
Statistical Power

The probability of detecting a real effect if it exists

Why this matters:

- **Too small:** Miss real effects (false negatives)
- **Too large:** Waste resources, ethical concerns
- **Underpowered studies** plague the literature

- simulate 5000 genes



Randomization and Controls

Randomization

Randomly assign subjects to treatment groups to eliminate systematic bias

Why randomize?

- Balances known AND unknown confounders across groups
- Foundation for causal inference
- Justifies statistical tests

Example: Cell Culture Experiment

Bad: Treat cells in left well with drug, right well as control
(Position effects, temperature gradients, pipetting order...)

Good: Randomly assign wells to treatment/control
(Eliminates systematic spatial effects)

Better: Block by plate, randomize within blocks
(Controls for plate-to-plate variation)

Blinding: Controlling Observer Bias

Blinding (Masking)
Keeping certain individuals unaware of treatment assignments to prevent bias

Types of blinding:

- **Single-blind:** Subjects don't know their group
- **Double-blind:** Neither subjects nor researchers know
- **Triple-blind:** Data analysts also unaware

Example: Scoring Phenotypes
Researcher scoring mouse behavior might unconsciously:

- Give benefit of doubt to treatment group
- Be more attentive to expected outcomes
- Interpret ambiguous cases differently

Solution: Code samples so scorer doesn't know which group they're from

Cardinal Rule

Always visualize your data before formal analysis

Why exploration matters:

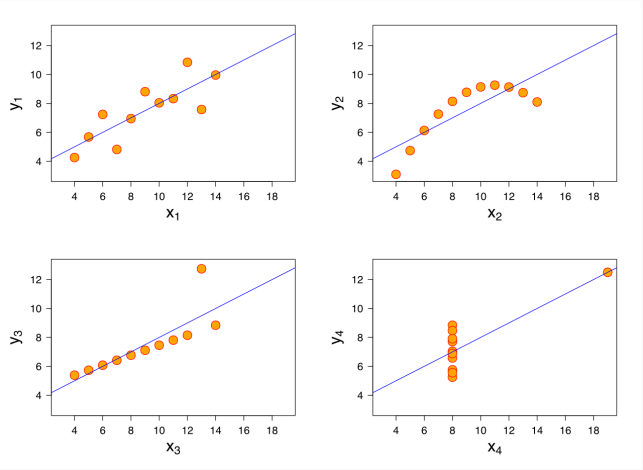
- Detect outliers and errors
- Understand distributions
- Identify patterns and relationships
- Check assumptions
- Generate hypotheses
- Catch problems early

"The simple graph has brought more information to the data analyst's mind than any other device" — John Tukey

The Power of Visualization: Anscombe's Quartet

Four datasets with identical statistics

- Same mean of X and Y
- Same variance of X and Y
- Same correlation ($r = 0.816$)
- Same regression line ($y = 3 + 0.5x$)



Source: Anscombe (1973)

But completely different patterns when plotted:

1. Linear relationship (appropriate for linear regression)
2. Curved relationship (needs nonlinear model)
3. Linear with one outlier (outlier drives the relationship)
4. No relationship except one high-leverage point

Lesson

Summary statistics alone can be deeply misleading—always plot your data!

Effective Visualization in Biology

Common plot types and their uses:

- **Histograms/density plots:** Distribution of continuous variables
  Example: Distribution of cell sizes
- **Box plots / violin plots:** Compare distributions across groups
  Example: Gene expression in treatment vs. control
- **Scatter plots:** Relationship between two variables
  Example: Body size vs. metabolic rate
- **Heatmaps:** High-dimensional data
  Example: Gene expression across samples and conditions
- **Time series plots:** Changes over time
  Example: Population growth curves

Principles of Good Data Visualization

1. **Show the data** — Not just summaries (avoid bar plots of means!)
2. **Respect visual perception** — Use appropriate scales, colors
3. **Avoid chartjunk** — Remove unnecessary elements
4. **Label clearly** — Axes, units, legends
5. **Be honest** — Don’t distort to make patterns look stronger
6. **Consider color blindness** — Use accessible color palettes

Common Mistake in Biology

Showing bar plots with error bars instead of showing the actual data points.
This hides the distribution, outliers, and sample size!

D. Modeling and Inference

This stage involves:

Statistical Modeling

Fitting mathematical models to data to:

- Describe relationships
- Estimate parameters
- Make predictions
- Test hypotheses

Statistical Inference

Drawing conclusions about populations from samples:

- Hypothesis testing (p-values)
- Confidence intervals
- Effect size estimation
- Model comparison

Hypothesis Testing Framework

The Logic

1. Assume the null hypothesis (H_0: no effect) is true
2. Calculate: How likely is our observed data under H_0?
3. If very unlikely ($p < \alpha$), reject H_0
4. Otherwise, fail to reject H_0

Example: Testing a New Antibiotic

H_0: New antibiotic has same efficacy as standard
H_A: New antibiotic is more effective

Collect data, calculate test statistic
$p = 0.02$ (only 2% chance of this result if H_0 true)
Reject H_0 → Evidence that new antibiotic is better

Important: Failing to reject $\neq$ proving null is true!

Understanding P-values

What a p-value IS

The probability of observing data as extreme as yours (or more extreme) if the null hypothesis were true

What a p-value is NOT

- NOT the probability the null hypothesis is true
- NOT the probability your result is due to chance
- NOT a measure of effect size or importance
- NOT the probability you made a mistake

The $\alpha = 0.05$ threshold:

- Conventional, but arbitrary
- Doesn't make $p = 0.049$ meaningful and $p = 0.051$ meaningless
- Consider p-values as continuous measures of evidence

Statistical vs. Biological Significance

Critical Distinction

Statistical significance $\neq$ Biological importance

Example: Gene Expression

Scenario: Large RNA-seq study (10,000 samples per group)

Gene A: 1.01-fold change, $p < 0.0001$ (highly significant!)
Gene B: 5-fold change, $p = 0.03$ (barely significant)

Which matters more biologically? Gene B!

With huge samples, tiny meaningless effects become ”significant”
Always report **effect sizes** (fold-change, Cohen’s d, etc.) with confidence intervals

Ask: *“Is this difference large enough to care about?”*

The Multiple Testing Problem

The Problem

Test many hypotheses → inflated false positive rate

Example: Testing 20 Genes

Using $\alpha = 0.05$, expect 1 false positive even if none are real
Test 1000 genes → expect 50 false positives!
Test 20,000 genes (typical RNA-seq) → expect 1,000 false positives!

Solutions:

- **Bonferroni correction:** Divide α by number of tests
  (Very conservative, low power)
- **False Discovery Rate (FDR):** Control proportion of false positives
  (More powerful, common in genomics: Benjamini-Hochberg)
- **Pre-registration:** Limit number of primary hypotheses

Checking Model Assumptions

Critical Step

All statistical models make assumptions—check them!

Common assumptions to check:

- **Normality:** Are residuals normally distributed?
  Check: Q-Q plots, histograms, Shapiro-Wilk test
- **Homoscedasticity:** Is variance constant across groups?
  Check: Residual plots, Levene's test
- **Independence:** Are observations independent?
  Consider: Repeated measures, spatial/temporal correlations
- **Linearity:** Is relationship actually linear?
  Check: Scatter plots, residual plots

If assumptions violated: Transform data, use different test, or robust methods

Your analysis is worthless if you can't communicate it

Good statistical communication:

- **Know your audience** — Adjust technical level appropriately
- **Tell a story** — Guide reader through logic
- **Visualize effectively** — One good plot worth thousands of words
- **Be precise but accessible** — Avoid jargon when possible
- **Quantify uncertainty** — Show confidence intervals, not just point estimates
- **Be honest about limitations** — Build trust through transparency
- **Distinguish strength of evidence** — Avoid over-confident claims

What to Report in Results

Essential elements:

1. **Sample sizes** — For all groups
2. **Effect sizes** — With confidence intervals
   Not just ”significant” but ”5-fold increase (95% CI: 3.2-7.8)”
3. **Statistical tests used** — Name the test, report test statistic
4. **P-values** — Exact values when possible, not just $p < 0.05$
5. **Visualization** — Show the data, not just summaries
6. **Assumptions checked** — Note transformations, exclusions
7. **Multiple testing corrections** — If applicable

Provide enough detail for readers to evaluate your conclusions

Translating Statistical Findings to Decisions

Example: Drug Development Decision

Statistical finding: New drug reduces tumor size by 30% (95% CI: 15-45%, $p = 0.001$, n=100 mice)

Decision considerations:

- Effect size clinically meaningful? (30% → yes, promising)
- Uncertainty acceptable? (CI doesn't include zero → reliable)
- Side effects tolerable?
- Cost-benefit analysis?
- What's next step? (Larger trial? Different dose?)

Decision: Proceed to Phase II trials with refined protocol

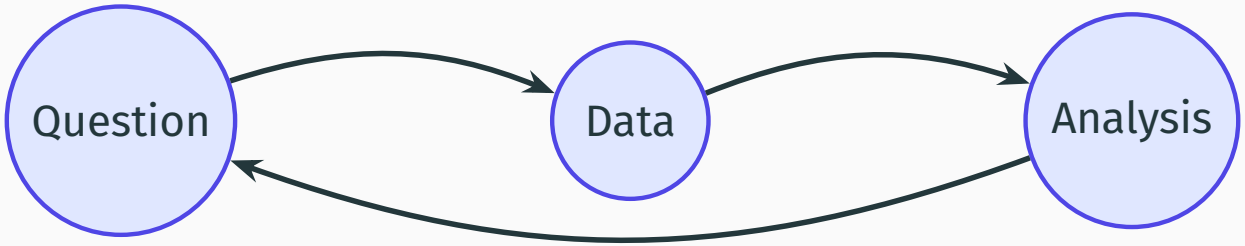
Statistics informs decisions, but doesn't make them—
context and values matter

The Iterative Nature of Data Analysis

Real data analysis is rarely linear—it's iterative!

Common cycles:

- Exploration reveals need to collect more data
- Modeling shows violations of assumptions → return to exploration
- Results raise new questions → formulate new hypotheses
- Peer review identifies issues → reanalyze
- Replication attempts → refine methods



Embrace iteration—it's a sign of thorough, careful science

Key Takeaways

1. **Start with clear questions**—they determine everything else
2. **Good design matters**—randomization, adequate power, controls
3. **Always visualize first**—avoid Anscombe's Quartet mistakes
4. **Check assumptions**—models only work when assumptions hold
5. **Report effect sizes**, not just p-values—statistical $\neq$ biological significance
6. **Correct for multiple testing**—especially in high-throughput biology
7. **Communicate clearly**—your science is only as good as your explanation
8. **Iterate**—real analysis is messy and cyclical

Navigating the fundamental trade-offs in data analysis

Three Major Philosophical Tensions

Real data analysis involves navigating fundamental trade-offs:

A. Frequentist vs. Bayesian Approaches
What does probability mean? How should prior knowledge influence analysis?

B. Prediction vs. Explanation
Machine learning's predictive power vs. traditional statistics' interpretability

C. Complexity vs. Simplicity
The bias-variance tradeoff and Occam's Razor

Understanding these tensions makes you a more sophisticated analyst

A. Frequentist vs. Bayesian Approaches

The Core Disagreement

What does “probability” actually mean?

Frequentist View

Probability = **long-run frequency**

“If we repeated this experiment infinitely many times, what proportion would show this result?”

Parameters are fixed but unknown

Bayesian View

Probability = **degree of belief**

“Given the evidence, how confident should we be that this is true?”

Parameters are random variables with distributions

Both are mathematically rigorous—they differ in philosophy, not correctness

The Frequentist Framework

Key concepts:

- **P-values:** Probability of data given null hypothesis
- **Confidence intervals:** If we repeated sampling many times, 95% of CIs would contain true parameter
- **Hypothesis testing:** Control long-run error rates (α, β)
- **No prior knowledge:** Analysis based only on current data

Example: Testing Gene Expression

H_0: No difference between treatment and control

Frequentist: "If there truly is no difference, we'd see data this extreme only 3% of the time ($p = 0.03$). So we reject H_0."

Note: This does NOT tell us probability that H_0 is true!

Strengths: Objective, well-established, controls error rates
Weaknesses: Often misinterpreted, can't incorporate prior knowledge

The Bayesian Framework

Key concepts:

- **Prior distribution:** What we believe before seeing data
- **Likelihood:** Probability of data given parameters
- **Posterior distribution:** Updated beliefs after seeing data
- **Bayes' Theorem:** Posterior $\propto$ Prior $\times$ Likelihood

Example: Testing Gene Expression

Prior: Based on literature, most genes don't respond (skeptical prior)

Data: Observe 2-fold change

Posterior: After updating with data, 85% probability of real effect

Bayesian: "Given the data and prior knowledge, there's 85% chance this gene responds to treatment."

Strengths: Intuitive interpretation, incorporates prior knowledge
Weaknesses: Subjective priors, computationally intensive

Example: Rare Disease Testing

Scenario: Genetic Disease Test

Disease prevalence: 1 in 10,000 people
Test accuracy: 99% sensitive, 99% specific
You test positive. What's the probability you have the disease?

Example: Rare Disease Testing

Scenario: Genetic Disease Test

Disease prevalence: 1 in 10,000 people
Test accuracy: 99% sensitive, 99% specific
You test positive. What's the probability you have the disease?
Intuitive answer: 99% (wrong!)

Correct answer: Only about 1%!

Why? In 10,000 people:

- 1 truly has disease → tests positive (99% chance)
- 9,999 don't have disease → 99 test positive (1% false positive rate)
- So 100 positive tests, but only 1 is truly diseased
- Probability = 1/100 ≈ 1%

Lesson: Base rates (priors) matter! Bayesian thinking helps here.

Frequentist vs. Bayesian: When to Use Which?

Use Frequentist When:

- You want to control long-run error rates (clinical trials)
- No strong prior information exists
- Regulatory requirements (FDA accepts frequentist)
- Simple hypothesis testing suffices

Use Bayesian When:

- You have strong prior knowledge to incorporate
- You want direct probability statements about parameters
- Small sample sizes (priors help stabilize estimates)
- Complex hierarchical models
- Sequential updating as data arrives

In biology: **Frequentist still dominates**, but Bayesian approaches growing
(especially in genomics, phylogenetics, systems biology)

B. Prediction vs. Explanation

The Tension

Do we want to **predict** outcomes accurately or **understand** the underlying mechanisms?

Prediction Focus

Goal: Maximize accuracy

"Black box" models OK if they work

Examples:

- Deep neural networks
- Random forests
- Gradient boosting

Explanation Focus

Goal: Understand relationships

Interpretability crucial

Examples:

- Linear regression
- ANOVA
- Generalized linear models

Often a trade-off: most interpretable models less accurate, most accurate models less interpretable

Example: Predicting Disease Risk

Scenario: Predicting Cancer from Gene Expression

Machine Learning Approach:

- Train deep neural network on 10,000 genes
- Achieves 95% accuracy on test set
- But: Can't explain *why* it predicts cancer
- Can't tell which genes matter most

Statistical Modeling Approach:

- Logistic regression with 5 key genes (from prior knowledge)
- Achieves 88% accuracy
- Can interpret: Each unit increase in Gene A multiplies odds by 2.5
- Can validate biological mechanism

Which is better? Depends on your goal!

When to Prioritize Prediction

Prediction is the primary goal when:

- **Clinical decision support** — Need accurate diagnosis/prognosis
  Example: Predicting patient response to immunotherapy
- **Screening/classification** — Identify candidates for further study
  Example: Identifying potential drug compounds from chemical structure
- **Mechanism partially known** — Just need to predict outcome
  Example: Predicting protein structure from sequence (we know physics, but it's complex)
- **Validation available** — Can test predictions experimentally
  Example: ML predicts gene function → validate with knockout

If the model works and improves decisions, interpretability is secondary

When to Prioritize Explanation

Explanation is the primary goal when:

- **Basic science** — Want to understand biological mechanisms
  Example: Which genes regulate cell cycle progression?
- **Hypothesis generation** — Need to guide future experiments
  Example: What pathways are perturbed in disease?
- **Intervention design** — Need to know what to manipulate
  Example: Which metabolic enzyme to target with drug?
- **Trust and transparency** — Stakeholders need to understand
  Example: Explaining treatment decisions to patients
- **Scientific publishing** — Reviewers expect mechanistic insight
  Example: Most biological journals want causal stories

If you can't explain it, you don't really understand it

Finding the Middle Ground

Strategies to balance prediction and explanation:

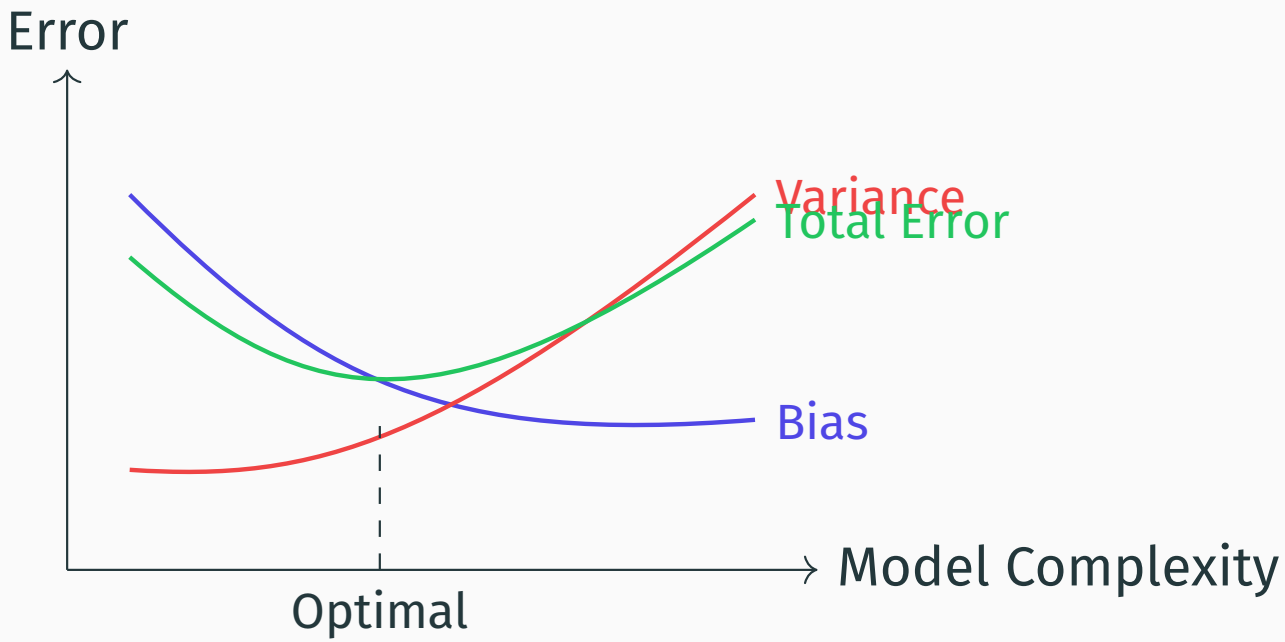
1. **Interpretable ML models**
   Use methods like LASSO, elastic net, decision trees that offer some interpretability
2. **Post-hoc interpretation**
   Apply SHAP values, feature importance, or partial dependence plots to black boxes
3. **Two-stage approach**
   Use ML for prediction, then simpler models to understand top features
4. **Mechanistic ML**
   Incorporate biological constraints into neural networks
5. **Ensemble thinking**
   Use both approaches and compare insights

The best approach depends on your scientific question and available data

C. Complexity vs. Simplicity

The Fundamental Tradeoff

- **Simple models:** May miss important patterns (high bias)
- **Complex models:** May fit noise as if it were signal (high variance)



Goal: Find the sweet spot that minimizes total error

Underfitting vs. Overfitting

Underfitting

Problem: Model too simple

Symptoms:

- Poor fit to data
- High error on training set
- High error on test set

Example: Linear model for curved relationship

Just Right

Balance: Captures signal, ignores noise

Symptoms:

- Good fit to data
- Low training error
- Low test error

Example: Polynomial degree matches true curve

Overfitting

Problem: Model too complex

Symptoms:

- Perfect fit to data
- Very low training error
- High test error

Example: High-degree polynomial fits noise

Training error always decreases with complexity, but test error may increase!

Example: Modeling Bacterial Growth

Scenario: Fitting Growth Curve to 10 Time Points

Model 1: Linear (too simple)

- Misses exponential phase
- High error
- Underfits

Model 2: Logistic curve (just right)

- Captures lag, exponential, stationary phases
- Biologically motivated
- Generalizes well

Model 3: 9th-degree polynomial (too complex)

- Passes through every point perfectly
- Wiggles unrealistically between points
- Terrible predictions for new data
- Overfits

Occam's Razor in Statistics

Principle of Parsimony

"Among competing hypotheses that explain the data equally well, the simplest one is most likely to be correct"

Why prefer simpler models?

- **Interpretability** — Easier to understand and communicate
- **Generalizability** — Less likely to overfit
- **Stability** — Small data changes don't drastically alter model
- **Falsifiability** — Simpler hypotheses easier to test
- **Biological plausibility** — Nature often operates by simple principles

But: Don't oversimplify! Use as complex a model as needed, but no more

Tools for Balancing Complexity

How to choose model complexity:

Cross-Validation

Split data into training/test sets; evaluate performance on held-out data
Most reliable method for assessing generalization

Information Criteria

AIC (Akaike) / **BIC** (Bayesian): Balance fit and complexity
Lower values = better; BIC penalizes complexity more heavily

Regularization

Add penalty for model complexity (LASSO, Ridge, Elastic Net)
Automatically shrinks unimportant coefficients toward zero

Example in RNA-seq

You have 20,000 genes but only 50 samples → severe overfitting risk
Solution: Use LASSO to select most important genes automatically

Example: Predicting Protein Expression

Scenario: Predict protein levels from mRNA

Available predictors:

- mRNA abundance (obvious choice)
- mRNA half-life
- Codon usage
- 5’ UTR structure
- Ribosome binding site strength
- … 50 other features

Approach 1: Include all 50+ features
Problem: Perfect fit to training data, terrible predictions (overfitting)

Approach 2: Use only mRNA abundance
Problem: Misses important biology (underfitting)

Approach 3: Use LASSO or cross-validation to select 5-10 key features
Result: Good predictions, interpretable, captures main biology

The Role of Domain Knowledge

Key Insight

In biology, domain knowledge should guide the bias-variance tradeoff

Use biological knowledge to:

- **Choose functional forms**
  Use mechanistic equations (Michaelis-Menten, Hill, logistic) rather than arbitrary polynomials
- **Select variables**
  Include biologically relevant predictors, exclude implausible ones
- **Set constraints**
  Force parameters to be positive, bounded, etc. based on biology
- **Interpret results**
  If model suggests implausible biology, probably overfitting

Don't let the algorithm alone decide—inject biological reasoning!

When Complex Models Are Justified

Sometimes complexity is necessary:

- **Biological systems are complex**
  Gene networks, metabolic pathways, ecosystems—reality has many interactions
- **Large datasets support it**
  With millions of data points (genomics), can fit complex models reliably
- **Prediction is the goal**
  If you just need accurate forecasts, complexity OK if validated
- **Simple models demonstrably fail**
  If linear model gives terrible fit and residuals show clear patterns

The principle is not "always use simple models" but
"use the simplest model that adequately captures the biology"

Navigating All Three Tensions Together

Real analysis involves all three tensions simultaneously:

Example: Analyzing Clinical Trial Data

Frequentist vs. Bayesian:

- Should we incorporate prior trial results? (Bayesian)
- Or analyze this trial independently? (Frequentist)

Prediction vs. Explanation:

- Do we need to predict individual response? (ML)
- Or understand which factors matter? (Statistical modeling)

Complexity vs. Simplicity:

- Include all patient characteristics? (Complex)
- Or just treatment group? (Simple)

Decision depends on: Study goals, sample size, regulatory requirements, stakeholder needs

Practical Guidance for Biological Research

General recommendations:

1. **Start simple, add complexity only if needed**
   Fit basic model first; add terms only if significantly improve fit
2. **Use biological knowledge to constrain choices**
   Don't treat analysis as pure math—inject domain expertise
3. **Validate, validate, validate**
   Use held-out data, cross-validation, independent replication
4. **Match method to question**
   Prediction task? ML is fine. Causal inference? Need careful design
5. **Be transparent about choices**
   Report why you chose Bayesian vs. frequentist, complex vs. simple
6. **Consider multiple approaches**
   Try both frequentist and Bayesian; compare simple and complex models

There Is No Perfect Answer

All of these tensions involve genuine trade-offs

There is no universally “correct” choice

What matters is:

- **Awareness** — Recognize these tensions exist
- **Thoughtfulness** — Make deliberate choices based on context
- **Transparency** — Explain and justify your decisions
- **Humility** — Acknowledge limitations of your approach
- **Iteration** — Be willing to try alternative approaches

The mark of a sophisticated analyst:
Understanding the trade-offs and navigating them wisely

Key Takeaways

1. **Frequentist vs. Bayesian:** Different philosophies of probability; choose based on goals and whether prior knowledge exists

2. **Prediction vs. Explanation:** ML excels at prediction; statistical models provide understanding. Match method to question

3. **Complexity vs. Simplicity:** Balance bias and variance. Use cross-validation and information criteria to find sweet spot

4. **Occam's Razor:** Prefer simpler models when they explain data equally well, but don't oversimplify complex biology

5. **Domain knowledge:** Use biological understanding to guide statistical choices

6. **No perfect answer:** These are genuine trade-offs; thoughtful navigation matters more than finding the "right" approach

Conclusion: Becoming a Thoughtful Analyst

Understanding these philosophical tensions transforms you from a button-pusher into a thoughtful scientist

Key principles for practice:

- Question your assumptions
- Understand your tools deeply
- Match methods to questions
- Validate rigorously
- Communicate transparently
- Stay humble about uncertainty

Good data analysis is both a science and an art—master the philosophy as well as the techniques

Syllabus

- Introduction
- Probability and descriptive statistics
- Statistical Inference
- Linear Models
- Basics of Experimental Design