Hypothesis testing vs. exploratory data analysis

- Hypothesis testing \( \Rightarrow \) controlled experiments
  - pros:
    - limit the number of independent variables.
    - sharpens thinking
    - encourage precise measurement.
  - cons:
    - far from reality and therefore may ignore important variables that effect outcomes.
    - biased toward the initial hypothesis
    - overlook interesting phenomena not related to original hypothesis.

Exploratory data analysis

- Pros
  - based on voluminous datasets and search for interesting patterns, no more hypotheses.

- Cons
  - generalizability
  - causality

Reductionist

- The dominant methodology of science has been analytic and reductionist in nature.
- This was clearly articulated by Descartes in 1628: In studying any phenomenon, simplify it to its essential components, dissecting away everything else.
- Reductionism is motivated by the belief that complicated systems are most fruitfully investigated at the lowest possible level, piece by piece.
- By reducing explanations to the smallest possible entities, the hope is that we will find entities that are simple enough to fully analyze and explain.
- The spectacular success of this methodology in modern science is undeniable.
Synthetic Approach

- Reductionist has not led to a corresponding understanding of how large collections of such components operate as systems.
- And, indeed, across the sciences, it is becoming increasingly clear that reduction must be complemented by synthesis.
- While reduction emphasizes decomposition and localization, synthesis is focused on understanding how the organization of the interactions between these components gives rise to the original phenomenon of interest.

The purpose of visualization

- Temporal behavior of data elements
- When was something greatest/least?
- Is there a pattern?
- Are two series similar?
- Do any of the series match a pattern?
- Provide simpler, faster access to the series

More tasks

- Does data element exist at time $t$?
- When does a data element exist?
- How long does a data element exist?
- How often does a data element occur?
- How fast are data elements changing?
- In what order do data elements appear?
- Do data elements exist together?

Patterns?
Statistical measures generate results in a compact form (instead of the full dataset).
However, they hide interesting features, such as the distributions of the data, and whether they are outliers.
Time Series Data

- Real-Valued function over time
- Goal: find patterns
  - “Starts Low, Ends High”
  - Outliers
  - Periodic Patterns
  - Laggards and Leaders
- Hypothesis generation
- Visualization alone is not enough

Basic Design

- Is the visualization time-dependent?
- Static: shows history, multiple perspectives.
- Dynamic: give feel for process & changes over time

ThemeRiver

- Visualize themes over time in large document collection
- Suitable for presenting multiple attributes over time
- Relying on basic perception rules

River Metaphor

River metaphor: Each attribute is mapped to a “current” in the “river”, flowing along the timeline
Visual cues

- Current width ~ strength of theme
- River width ~ global strength
- Color mapping (similar themes – same color family)
- Time line
- External events

Extended exploration

Presenting other data types

Critique

Strong points:
- Intuitive exploration of temporal changes and relations
- Applicable to general attributes

Weak points:
- Limited number of themes / attributes
- No ability to reorder currents
Ben Shneiderman

**Overview**
first, zoom and filter, then
details-on-demand

---

**Timeboxes**

- Rectangular query regions
- Value must be in range for all time points in region
- Combine multiple timeboxes for conjunctive query

---

**TimeSearcher - Timeboxes**

- Visual query operator for time series
  - [e.g. 1500 stocks]
- Rectangular region drawn on the
timeline display
- X-axis of the box = *time period*
- Y-axis of the box = *constraint on the\n  values*
- Multiple timeboxes = *conjunctive\n  queries*
Leader & Laggards

• Offset by one time period

Query inversion

Multiple time-varying attributes

Variable-time box

At least R consecutive time periods between X1 and X2 items must have values in the range.
Angular Queries

Query

• Perception -> query -> quantify
• Hypothesis -> query -> confirm/reject

More query functions?

• Input: pattern
• Output: events (timestamps)
• Repeated single patterns within a stream
• Repeated sequential patterns within a stream or across multiple streams
• Event-based query: before/after/during a certain event, repeated patterns of another stream.

Pattern

• Timebox-like
• Specify a subsequence
Pattern Comparison

- Dynamic time warping