# Table of Contents

|  |
| --- |
| Research | Current Trends | Use-cases | Project Cycle | Presentation | Best Practices |
| Situation Understanding |
| Data quality | Acquisition | Data Mining/Sql | Profiling/Descriptive Statistics |
| Cleaning | Transformation | Prelim Exploration/Slice & Dicing | Exploration/Visualization | Big Data |
| Methods Selection | ∇ Feature Engineering | ∇ Pre-processing |
| * Techniques & Methods | Application Notes |
| Insight | Action | Results |
| * Data Architecture & Engineering | Databases |
| Bash | Surveys | Open Data | Scaling | ML/Cloud Tools | Software Eng. | Q&A |

# Techniques & Methods

* Computational Fundamentals
* Algorithms
* Statistics
* Probability
* Inferencial Statistics
* A/B Testing
* Bayesian Analysis
* Markov Models
* Forecasting Analysis
* Fourier Analysis
* Simulation and Random Walks
* Rule-based Learning
* Unsupervised Learning
* Ensemble Models
* Re-inforcement Learning
* On-line Learning
* Deep Learning
* Artificial Intelligence
* Tensors
* Optimization & Multiple Goals
* Stochastic Optimization
* Scheduling and multi-period planning
* Sensitivity Analysis
* Operations Research & Analytics : Queing,

To get started, the Python sections are linked at the left -- [Python Set Up](https://developers.google.com/edu/python/set-up) to get Python installed on your machine, [Python Introduction](https://developers.google.com/edu/python/introduction) for an introduction to the language, and then [Python Strings](https://developers.google.com/edu/python/strings) starts the coding material, leading to the first exercise. The end of each written section includes a link to the code exercise for that section's material. The lecture videos parallel the written materials, introducing Python, then strings, then first exercises, and so on. At Google, all this material makes up an intensive 2-day class, so the videos are organized as the day-1 and day-2 sections.

This material was created by [Nick Parlante](http://www-cs-faculty.stanford.edu/%7enick/) working in the engEDU group at Google. Special thanks for the help from my Google colleagues John Cox, Steve Glassman, Piotr Kaminksi, and Antoine Picard. And finally thanks to Google and my director Maggie Johnson for the enlightened generosity to put these materials out on the internet for free under the [Creative Commons Attribution 2.5](http://creativecommons.org/licenses/by/2.5/) license -- share and enjoy!

## findall and Groups

|  |
| --- |
| Step > Goal/ Value |
| Assumptions: |
| Additional Resources: |

The parenthesis ( ) group mechan

  # Open file  
  f = open('test.txt', 'r')  
  # Feed the file text into findall(); it returns a list of all the found strings  
  strings = re.findall(r'some pattern', f.read())

# Statistics

## p-values

American Statistical Association (ASA) – Statement on Statistical Significance and P-values- March 7, 2016

<http://www.amstat.org/newsroom/pressreleases/P-ValueStatement.pdf>

<http://www.nature.com/news/scientific-method-statistical-errors-1.14700>

<http://www.nature.com/news/statisticians-issue-warning-over-misuse-of-p-values-1.19503?WT.mc_id=FBK_NatureNews>

“In its statement, the ASA advises researchers to avoid drawing scientific conclusions or making policy decisions based on *P* values alone. “

Misuse of the *P* value — a common test for judging the strength of scientific evidence — is contributing to the number of research findings that [cannot be reproduced](http://www.nature.com/news/reproducibility-1.17552),

the *P* value was [being misapplied](http://www.nature.com/news/scientific-method-statistical-errors-1.14700) in ways that cast doubt on statistics generally,

Researchers should describe not only the data analyses that produced statistically significant results, the society says, but all statistical tests and choices made in calculations. Otherwise, results may seem falsely robust.

## Industry

### Ten Simple Rules for Effective Statistical Practice

<http://journals.plos.org/ploscompbiol/article?id=10.1371%2Fjournal.pcbi.1004961#sec002>

#### Rule 1: Statistical Methods Should Enable Data to Answer Scientific Questions

start with the underlying question, such as, “Where are the differentiated genes?” and, from there, would consider multiple ways the data might provide answers.

This shift in perspective from statistical technique to scientific question may change the way one approaches data collection and analysis.

After learning about the questions, statistical experts discuss with their scientific collaborators the ways that data might answer these questions and, thus, what kinds of studies might be most useful.

## Rule 2: Signals Always Come with Noise

In some cases variability is good, because we need variability in predictors to explain variability in outcomes.

For example, Google Flu Trends debuted to great excitement in 2008, but turned out to overestimate the prevalence of influenza by nearly 50%, largely due to bias caused by the way the data were collected; see Harford [[8](http://journals.plos.org/ploscompbiol/article?id=10.1371%2Fjournal.pcbi.1004961#pcbi.1004961.ref008)], for example.

<http://www.ft.com/cms/s/2/21a6e7d8-b479-11e3-a09a-00144feabdc0.html>

## Rule 3: Plan Ahead, Really Ahead

Asking questions at the design stage can save headaches at the analysis stage: careful data collection can greatly simplify analysis and make it more rigorous.

## Rule 4: Worry about Data Quality

Units of measurement should be understood and recorded consistently. It is important that missing data values can be recognized as such by relevant software. For example, 999 may signify the number 999, or it could be code for “we have no clue.” There should be a defensible rule for handling situations such as “non-detects,” and data should be scanned for anomalies such as variable 27 having half its values equal to 0.00027. Try to understand as much as you can how these data arrived at your desk or disk. Why are some data missing or incomplete? Did they get lost through some substantively relevant mechanism? Understanding such mechanisms can help to avoid some seriously misleading results. For example, in a developmental imaging study of attention deficit hyperactivity disorder, might some data have been lost from children with the most severe hyperactivity because they could not sit still in the MR scanner?

## Rule 5: Statistical Analysis Is More Than a Set of Computations

A reader will likely want to consider the fundamental issue of whether the analytic technique is appropriately linked to the substantive questions being answered.

## Rule 6: Keep it Simple

This rule has been rediscovered and enshrined in operating procedures across many domains and variously described as “Occam’s razor,” “KISS,” “less is more,” and “simplicity is the ultimate sophistication.”

Having said this, scientific data have detailed structure, and simple models can’t always accommodate important intricacies. The common assumption of independence is often incorrect and nearly always needs careful examination. See [Rule 8](http://journals.plos.org/ploscompbiol/article?id=10.1371%2Fjournal.pcbi.1004961#sec009). Large numbers of measurements, interactions among explanatory variables, nonlinear mechanisms of action, missing data, confounding, sampling biases, and so on, can all require an increase in model complexity.

## Rule 7: Provide Assessments of Variability

A basic purpose of statistical analysis is to help assess uncertainty,

## Rule 8: Check Your Assumptions

In addition to nonlinearity and statistical dependence, missing data, systematic biases in measurements, and a variety of other factors can cause violations of statistical modeling assumptions, even in the best experiments.

## Rule 9: When Possible, Replicate!

http://journals.plos.org/ploscompbiol/article?id=10.1371%2Fjournal.pcbi.1004961#pcbi.1004961.ref002

## Rule 10: Make Your Analysis Reproducible

http://journals.plos.org/ploscompbiol/article?id=10.1371%2Fjournal.pcbi.1004961#pcbi.1004961.ref002