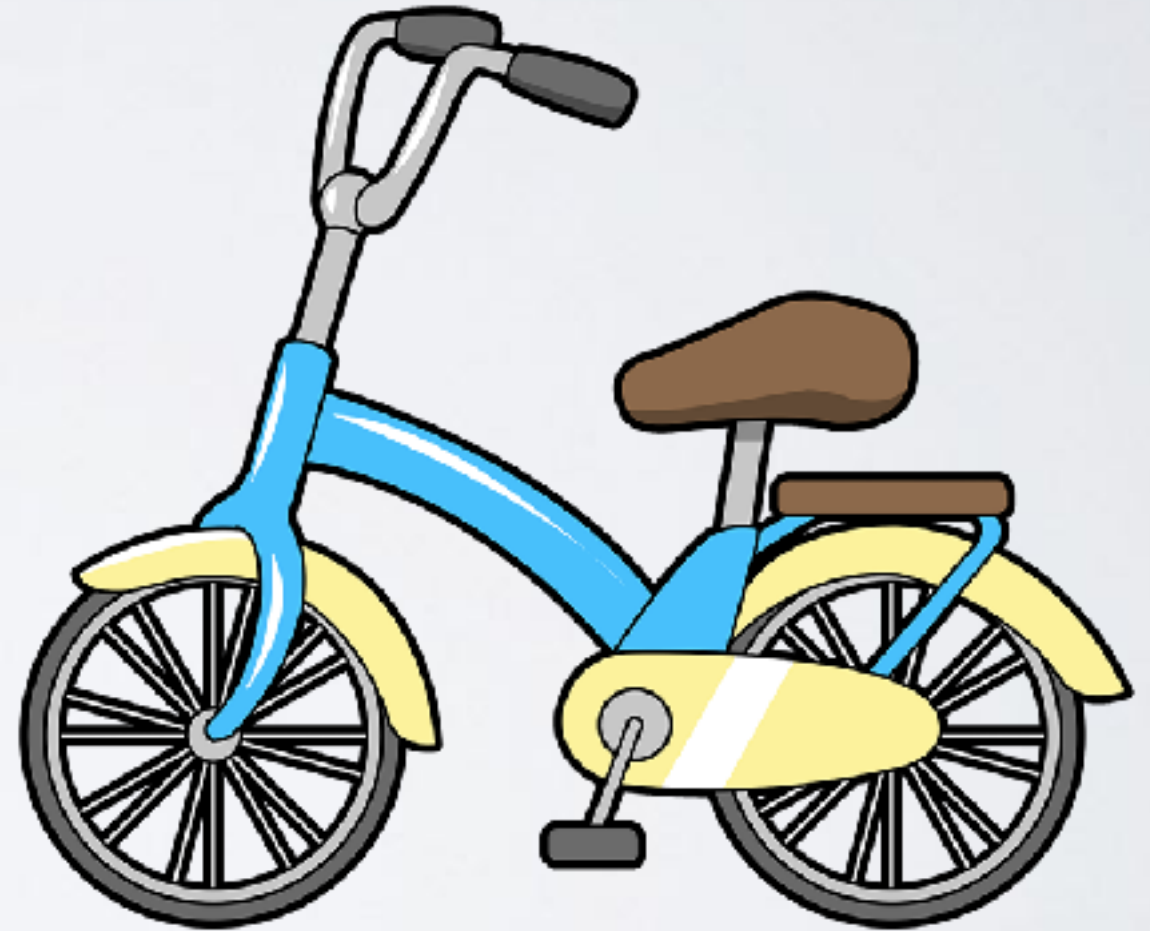


# Turning Data Into Knowledge



Tips and tricks that will empower you to become an Excel Master Jedi Black Belt Guru, and a more powerful Python White Belt Novice

# TRICYCLE VS BICYCLE



# AUTO VS MANUAL TRANSMISSION





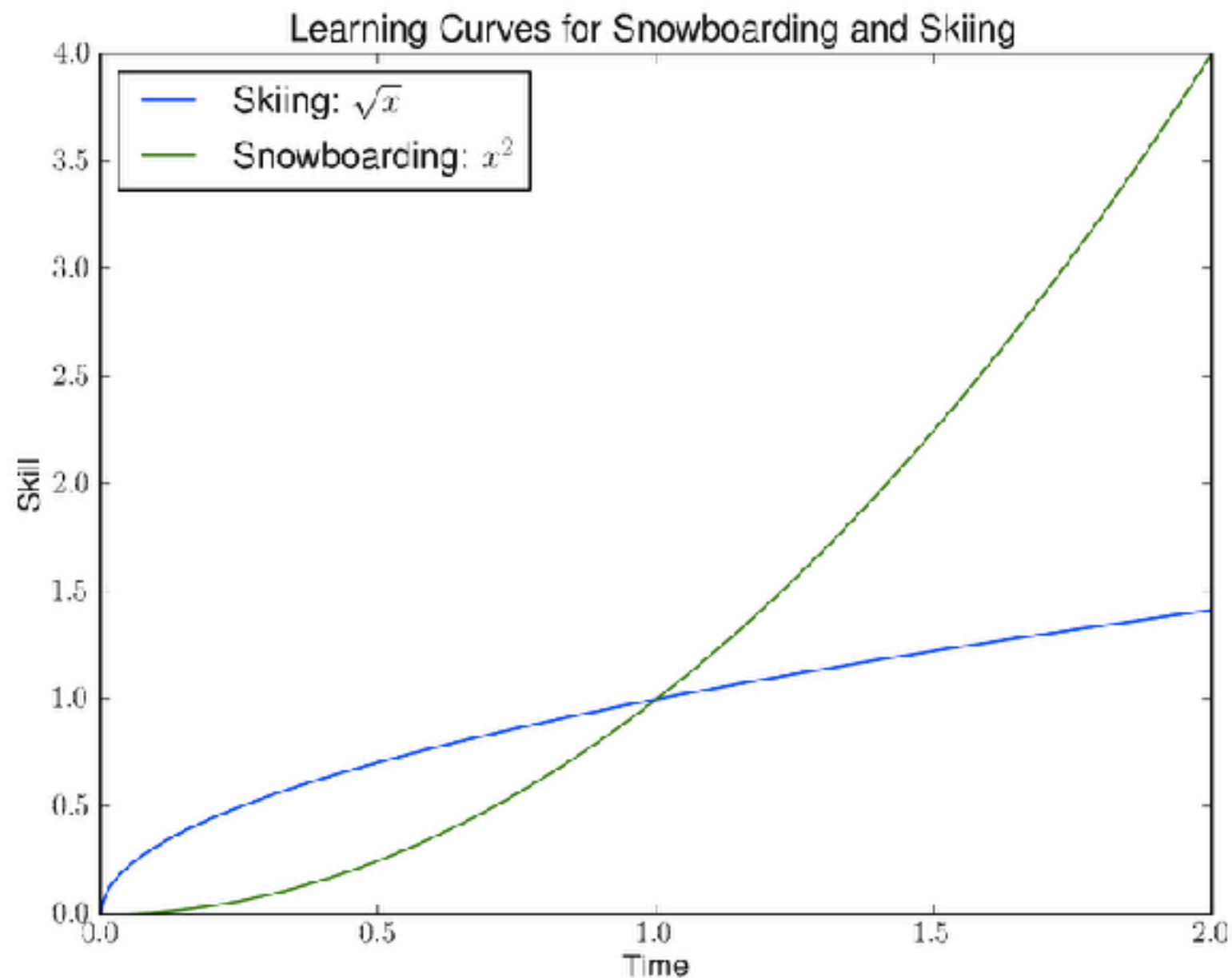
# SKIING VS SNOWBOARDING



# EXCEL VS PYTHON



# WHAT DO THESE HAVE IN COMMON?



# EXPLORE THE UNKNOWNNS

*There are known knowns; there are things we know that we know.*

*There are known unknowns; that is to say, there are things that we now know we don't know.*

*But there are also unknown unknowns – there are things we do not know we don't know.*

-Donald Rumsfeld



- Browse Excel's functionality from the Ribbon
- Google your Excel problems, someone probably has asked them online

# THE KC™ DATA ANALYSIS LEARNING CURVE

- Use more formatting, the **Excel White Belt**
- Use more shortcuts
- Use Pivot Tables
- Use Macros (but not really)
- Use custom functions, the **Excel Black Belt**
- Use Python, the **eternal White Belt**



# XR EXAMPLE: FORMATTING

- Use Table Formatting
  - Filters get automatically applied
  - Easier to read with alternating row colours
- Conditional Formatting also helps you process information
- Freeze panes helps you read datasets with a lot of rows and columns
- Group columns to hide columns (avoid Hide column feature)
- Tab colours

# KEYBOARD SHORTCUTS

- F2 to edit cells
- Arrow keys to navigate spreadsheet, or TAB
- CTRL-Home to move to top-left most cell
- CTRL End to move to bottom-right most cell
- CTRL-A to select
- CTRL-PgUp and CTRL-PgDn to access other tabs
- CTRL-TAB to access other workbooks

# KEYBOARD SHORTCUTS

- Access keyboard shortcuts via ALT
  - pre and post Office 2007 shortcuts work
- Example: Name Manager
  - ALT I+N+D (Office 2003 and earlier)
  - ALT M+N (Office 2007 and later)
  - Directly Name box

# OTHER DATA MANIPULATION

- Sort and Filter
- Text to columns
- Remove duplicates
- Arrays (requires CTRL-SHIFT-ENTER)
  - Example use find to count *Scheffersomyces* entries



# PREFERRED METHODS FOR DATA ANALYSIS

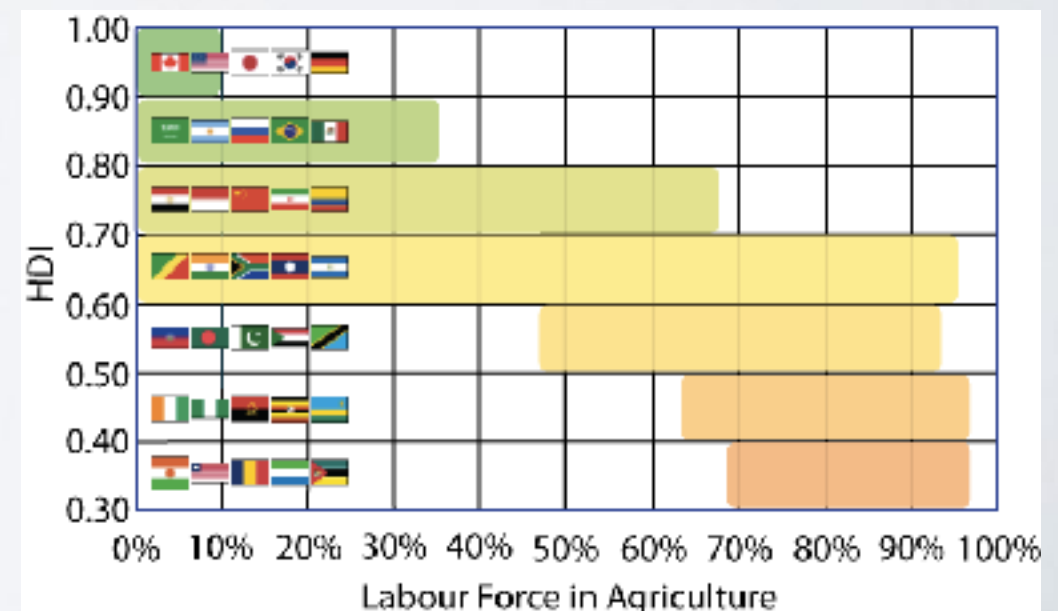
- Keep raw data, meta data, and summary of data on separate sheets
  - Only add meta or additional calculations to raw data
- Defined names are your friends
- I prefer advanced functions over Pivot tables

# ANALYZE 2008 US ELECTION

- Download data from the web within Excel
  - Can use other sources such as XML, databases, text files
  - Can refresh data
- Apply 'advanced functions' to further analyze data
- Paste special features
- Pivot table

# HDI EXAMPLE

- `sumifs(sum range, criteria range, criteria 1, criteria range criteria 2`
  - can only be used with equality
- `sum(if( (condition 1)*(condition 2),...))`
  - can use equality or inequalities



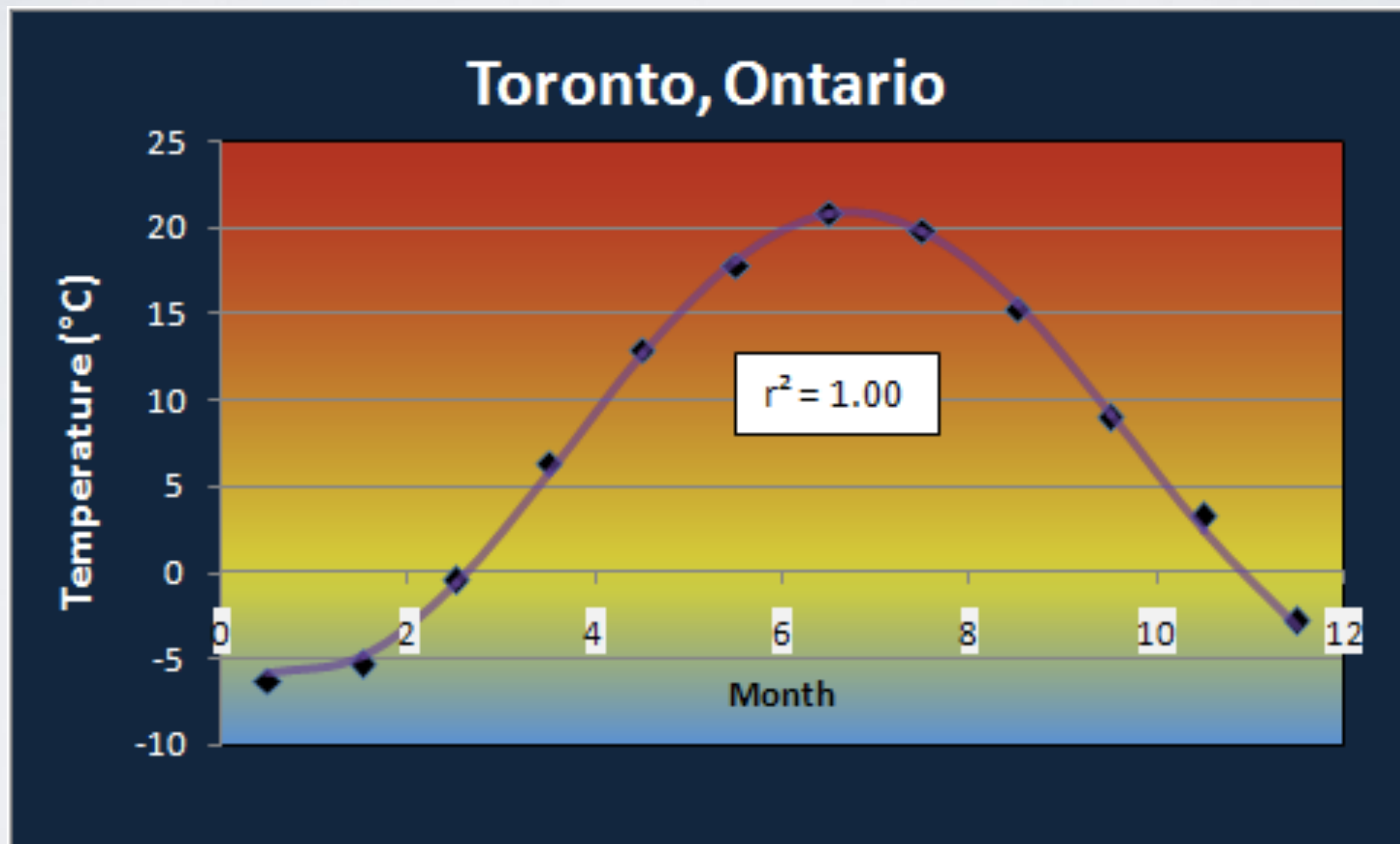
# MOVIE ANALYSIS

- Use of advanced functions to analyze data from different perspectives
- all movies vs blockbusters vs revenue







# TEMPERATURE AS A SINUSOIDAL FUNCTION










# PYTHON EXAMPLES

- pandas (think Excel in Python)
- scrapping with BeautifulSoup
- UniProt
- Entrez / NCBI

# AUTOMATED EMAIL SCRIPTS

**Daily scripts**  **Inbox** 

 **kevin.correia@gmail.com** Jan 19 (7 days ago) ★    
to me 

 Categorize this message as: **Personal**  [Never show this again](#) 

430363 RNA-SEQ analysis of a *C. albicans* rim101-/- disrupted strain, a strain overexpressing RIM101 and a reference strain at both alkaline (7.6) and acidic (4) pH Transcriptome or Gene expression The aim of the study was to compare the RNA-SEQ profiles of a wild type *C. albicans* strain (SC5314), a rim101-/- mutant strain (DAY25) and a strain overexpressing RIM101 (CGY1) at both alkaline (7.6) and acidic (4) pH, in order to identify Rim-dependent genes involved in tolerance to antifungals Overall design: Samples were collected after growth at pH 4 (Rim pathway inactivated) and pH 7.6 (Rim pathway activated), in triplicates. RNAs were extracted, prepared, and sequenced using standard RNA-Seq methods and Illumina technology. Transcriptome Sequencing 5476.0 35441 *Candida albicans* nan PRJNA430363 22221 CHU Grenoble Alpes

370154 *Yarrowia lipolytica* YB392 Resequencing genome sequencing Genome sequencing Genome sequencing of *Yarrowia Lipolytica* strain Genome Sequencing 4952.0 270187 *Yarrowia lipolytica* YB392 PRJNA370154 78114 DOE Joint Genome Institute

370155 *Yarrowia lipolytica* YB419 Resequencing genome sequencing Genome sequencing Genome sequencing of *Yarrowia Lipolytica* strain Genome Sequencing 4952.0 270186 *Yarrowia lipolytica* YB419 PRJNA370155 78113 DOE Joint Genome Institute

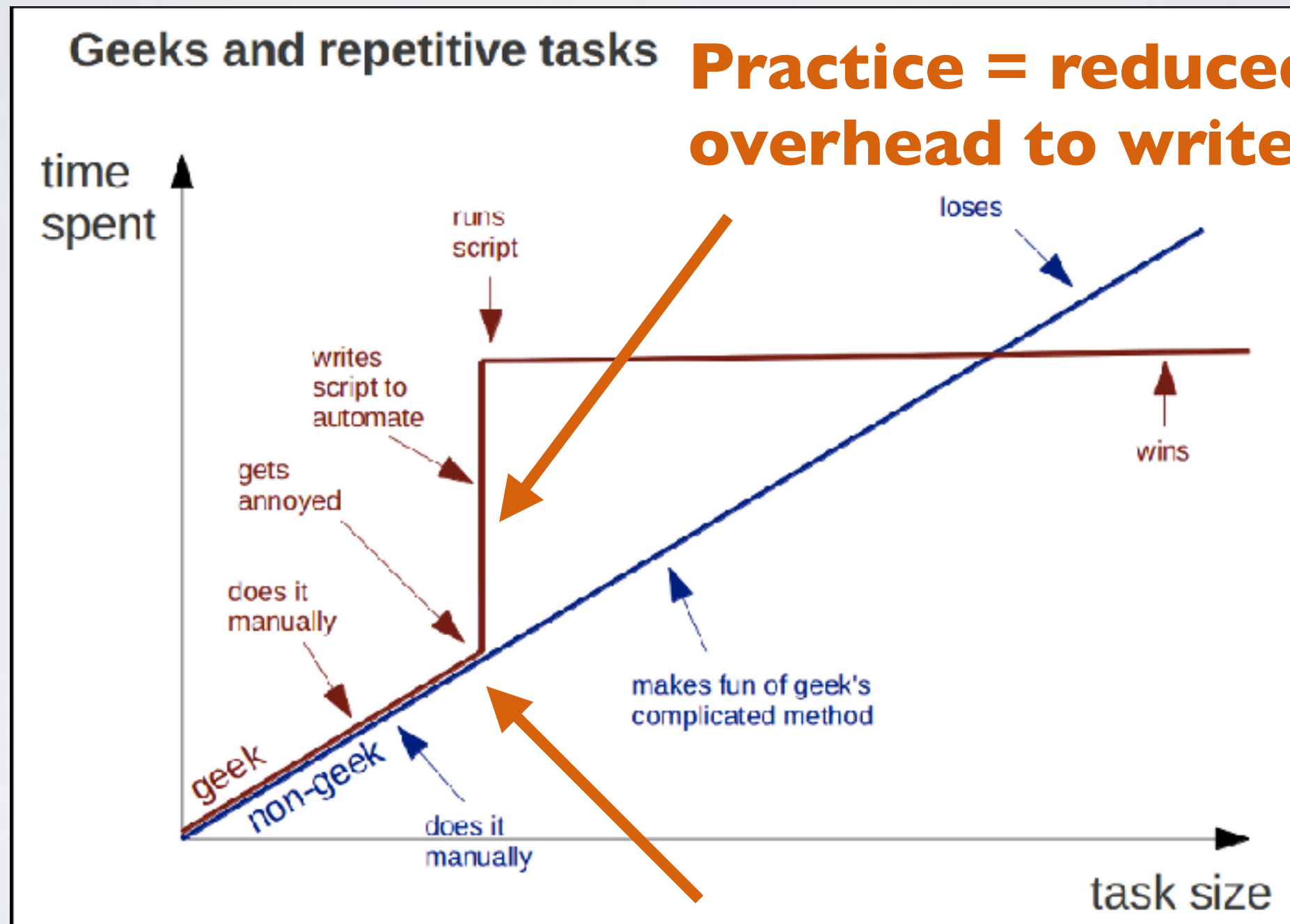
370156 *Yarrowia lipolytica* YB420 Resequencing genome sequencing Genome sequencing Genome sequencing of *Yarrowia Lipolytica* strain Genome Sequencing 4952.0 270185 *Yarrowia lipolytica* YB420 PRJNA370156 78112 DOE Joint Genome Institute

370157 *Yarrowia lipolytica* YB566 Resequencing genome sequencing Genome sequencing Genome sequencing of *Yarrowia Lipolytica* strain Genome Sequencing 4952.0 270184 *Yarrowia lipolytica* YB566 PRJNA370157 78111 DOE Joint Genome Institute

370158 *Yarrowia lipolytica* YB567 Resequencing genome sequencing Genome sequencing Genome sequencing of *Yarrowia Lipolytica* strain Genome Sequencing 4952.0 270183 *Yarrowia lipolytica* YB567 PRJNA370158 78110 DOE Joint Genome Institute

370159 *Yarrowia lipolytica* YIAK001 Resequencing genome sequencing Genome sequencing Genome sequencing of *Yarrowia Lipolytica* strain Genome Sequencing 4952.0 270182 *Yarrowia lipolytica* YIAK001 PRJNA370159 78109 DOE Joint Genome Institute

# YOU ALWAYS WIN WITH PROGRAMMING



**Practice = reduced overhead to write scripts**

**Practice = reduced threshold to write scripts**