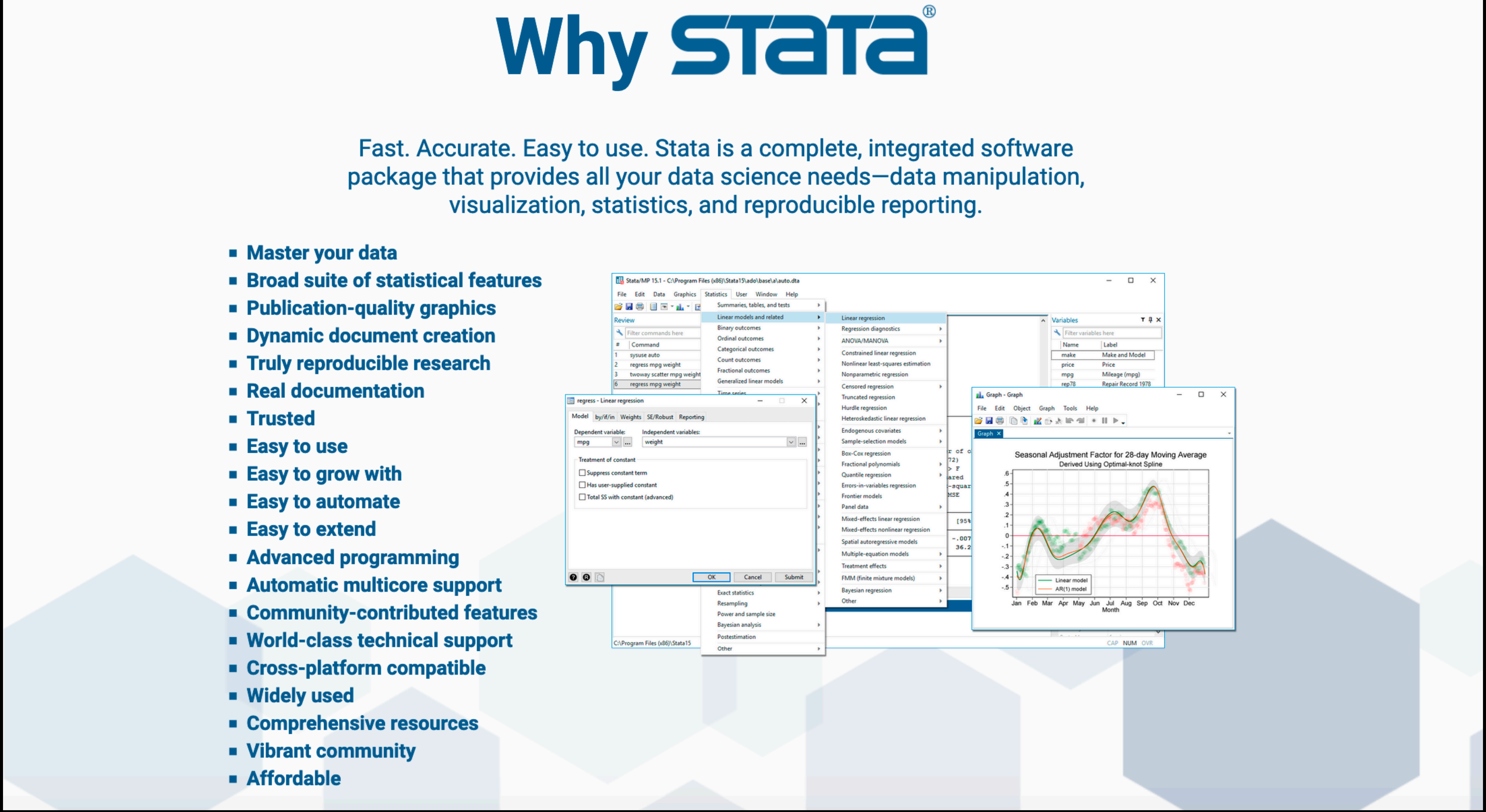
# Machine Learning and AI in Stata

Kiel.AI

# What is stata?

- Statistical Program like SPSS or Matlab



## Why STATA®

Fast. Accurate. Easy to use. Stata is a complete, integrated software package that provides all your data science needs—data manipulation, visualization, statistics, and reproducible reporting.

- Master your data
- Broad suite of statistical features
- Publication-quality graphics
- Dynamic document creation
- Truly reproducible research
- Real documentation
- Trusted
- Easy to use
- Easy to grow with
- Easy to automate
- Easy to extend
- Advanced programming
- Automatic multicore support
- Community-contributed features
- World-class technical support
- Cross-platform compatible
- Widely used
- Comprehensive resources
- Vibrant community
- Affordable

Stata/MP 15.1 - C:\Program Files (x86)\Stata15\ado\base\a\auto.dta

| # | Command |
|---|---|
| 1 | sysuse auto |
| 2 | regress mpg weight |
| 3 | twoway scatter mpg weight |
| 6 | regress mpg weight |

regress - Linear regression

Dependent variable: mpg

Independent variables: weight

Treatment of constant

□ Suppress constant term

□ Has user-supplied constant

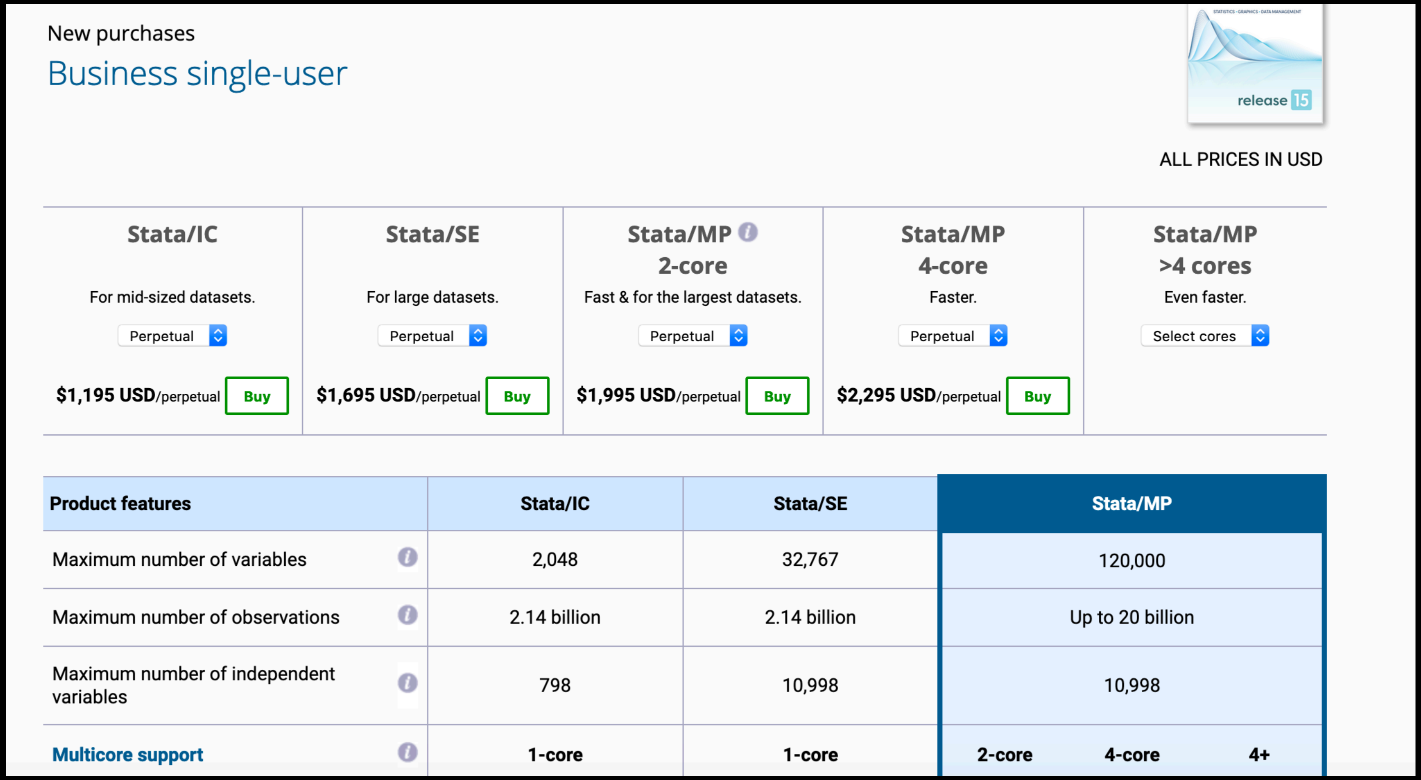
□ Total SS with constant (advanced)

| Name | Label |
|---|---|
| make | Make and Model |
| price | Price |
| mpg | Mileage (mpg) |
| rep78 | Repair Record 1978 |

C:\Program Files (x86)\Stata15

# Where to get it

- stata.com
- Not for free



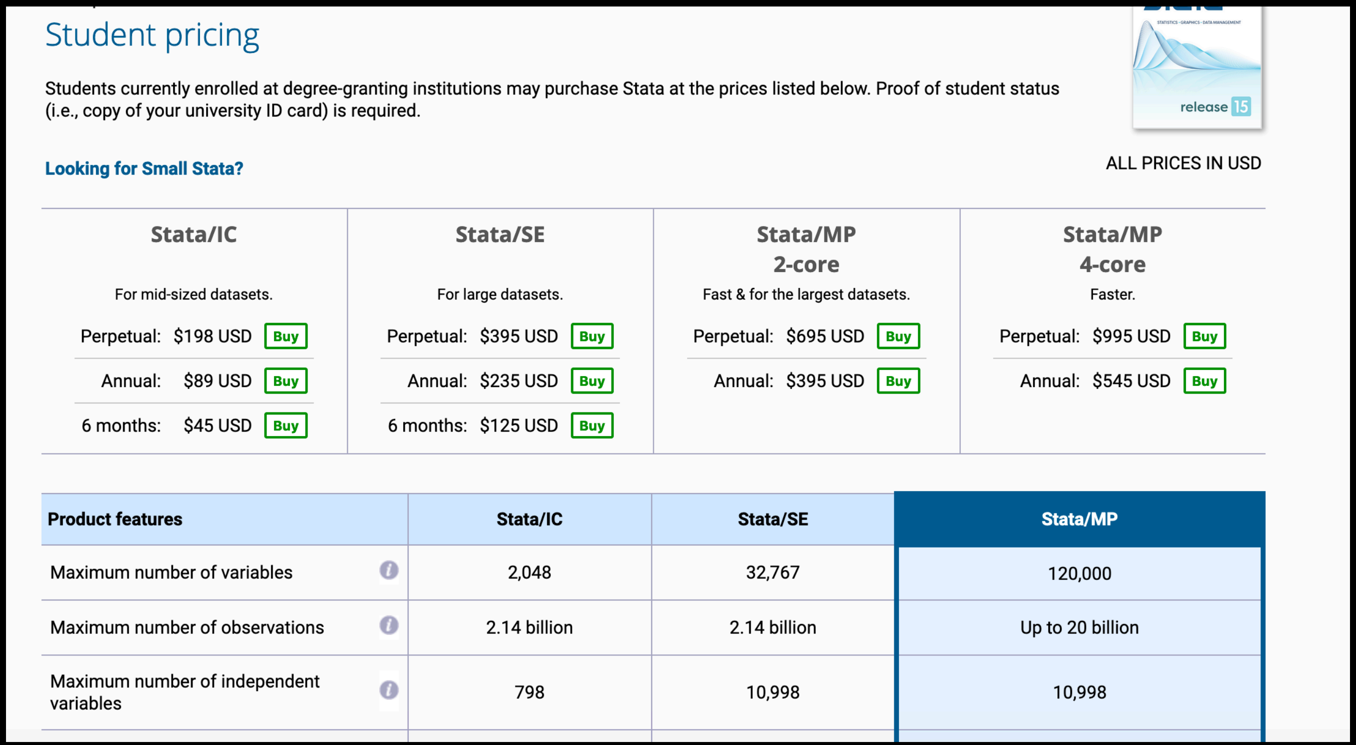
New purchases

Business single-user

ALL PRICES IN USD

| Stata/IC | Stata/SE | Stata/MP 2-core | Stata/MP 4-core | Stata/MP >4 cores |
|---|---|---|---|---|
| For mid-sized datasets. | For large datasets. | Fast & for the largest datasets. | Faster. | Even faster. |
| Perpetual | Perpetual | Perpetual | Perpetual | Select cores |
| $1,195 USD/perpetual Buy | $1,695 USD/perpetual Buy | $1,995 USD/perpetual Buy | $2,295 USD/perpetual Buy | |

| Product features | Stata/IC | Stata/SE | Stata/MP |
|---|---|---|---|
| Maximum number of variables | 2,048 | 32,767 | 120,000 |
| Maximum number of observations | 2.14 billion | 2.14 billion | Up to 20 billion |
| Maximum number of independent variables | 798 | 10,998 | 10,998 |
| Multicore support | 1-core | 1-core | 2-core 4-core 4+ |



Student pricing

Students currently enrolled at degree-granting institutions may purchase Stata at the prices listed below. Proof of student status (i.e., copy of your university ID card) is required.

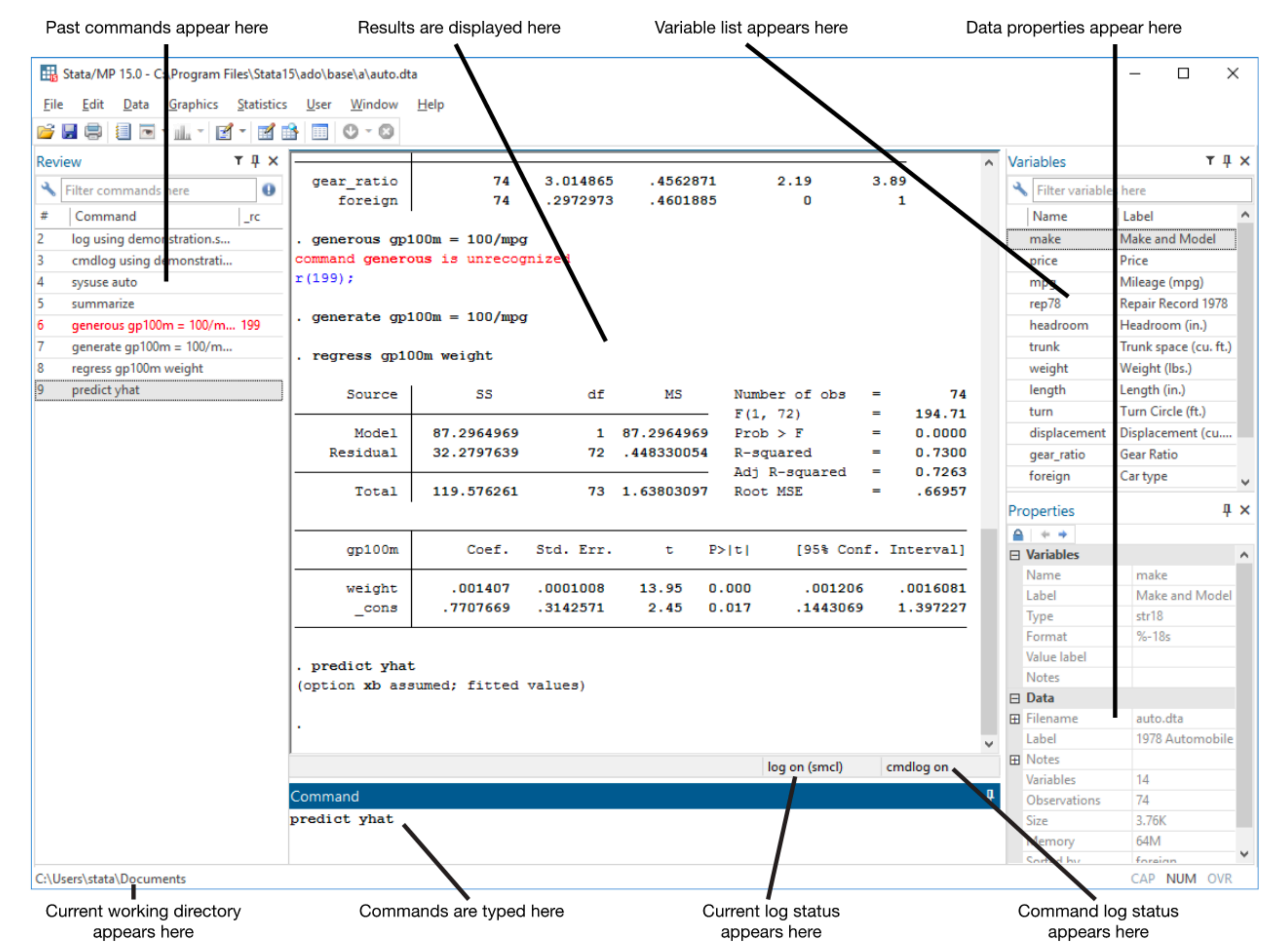
Looking for Small Stata?

ALL PRICES IN USD

| Stata/IC | Stata/SE | Stata/MP 2-core | Stata/MP 4-core |
|---|---|---|---|
| For mid-sized datasets. | For large datasets. | Fast & for the largest datasets. | Faster. |
| Perpetual: $198 USD Buy | Perpetual: $395 USD Buy | Perpetual: $695 USD Buy | Perpetual: $995 USD Buy |
| Annual: $89 USD Buy | Annual: $235 USD Buy | Annual: $395 USD Buy | Annual: $545 USD Buy |
| 6 months: $45 USD Buy | 6 months: $125 USD Buy | | |

| Product features | Stata/IC | Stata/SE | Stata/MP |
|---|---|---|---|
| Maximum number of variables | 2,048 | 32,767 | 120,000 |
| Maximum number of observations | 2.14 billion | 2.14 billion | Up to 20 billion |
| Maximum number of independent variables | 798 | 10,998 | 10,998 |

# Interface



Past commands appear here

Results are displayed here

Variable list appears here

Data properties appear here

Current working directory appears here

Commands are typed here

Current log status appears here

Command log status appears here

| # | Command | _rc |
|---|---|---|
| 2 | log using demonstration.s... | |
| 3 | cmdlog using demonstrati... | |
| 4 | sysuse auto | |
| 5 | summarize | |
| 6 | generous gp100m = 100/m... | 199 |
| 7 | generate gp100m = 100/m... | |
| 8 | regress gp100m weight | |
| 9 | predict yhat | |

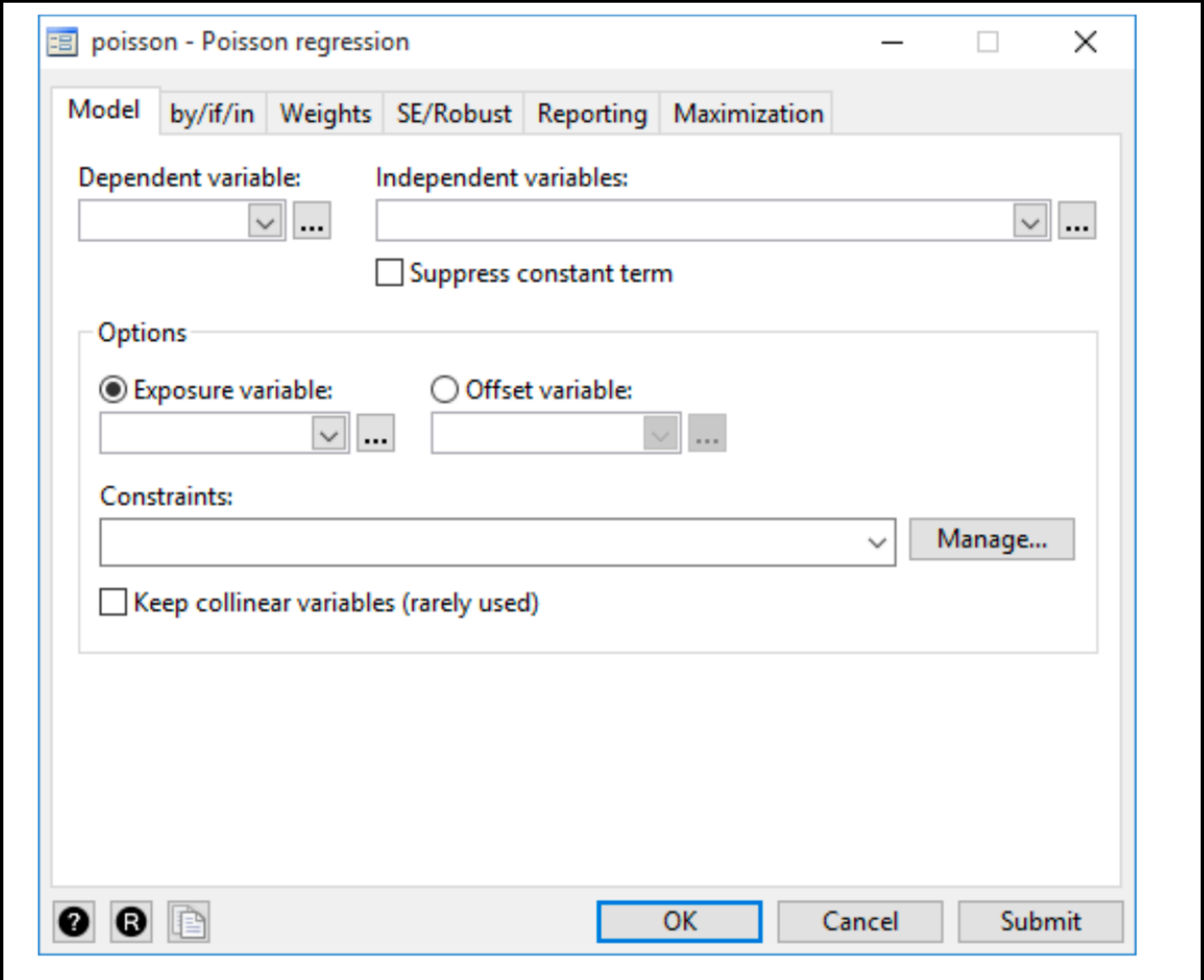
```
  gear_ratio |         74    3.014865    .4562871       2.19       3.89
     foreign |         74    .2972973    .4601885          0          1

. generous gp100m = 100/mpg
command generous is unrecognized
r(199);

. generate gp100m = 100/mpg

. regress gp100m weight

      Source |       SS           df       MS      Number of obs   =        74
-------------+----------------------------------   F(1, 72)        =    194.71
       Model |  87.2964969         1  87.2964969   Prob > F        =    0.0000
    Residual |  32.2797639        72  .448330054   R-squared       =    0.7300
-------------+----------------------------------   Adj R-squared   =    0.7263
       Total |  119.576261        73  1.63803097   Root MSE        =    .66957

------------------------------------------------------------------------------
      gp100m |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      weight |    .001407   .0001008    13.95   0.000      .001206    .0016081
       _cons |   .7707669   .3142571     2.45   0.017     .1443069    1.397227
------------------------------------------------------------------------------

. predict yhat
(option xb assumed; fitted values)

.
```

```
predict yhat
```

| Name | Label |
|---|---|
| make | Make and Model |
| price | Price |
| mpg | Mileage (mpg) |
| rep78 | Repair Record 1978 |
| headroom | Headroom (in.) |
| trunk | Trunk space (cu. ft.) |
| weight | Weight (lbs.) |
| length | Length (in.) |
| turn | Turn Circle (ft.) |
| displacement | Displacement (cu.... |
| gear_ratio | Gear Ratio |
| foreign | Car type |

| | |
|---|---|
| ⊟ **Variables** | |
| Name | make |
| Label | Make and Model |
| Type | str18 |
| Format | %-18s |
| Value label | |
| Notes | |
| ⊟ **Data** | |
| ⊞ Filename | auto.dta |
| Label | 1978 Automobile |
| ⊞ Notes | |
| Variables | 14 |
| Observations | 74 |
| Size | 3.76K |
| Memory | 64M |
| Sorted by | foreign |

# Interface



poisson - Poisson regression

Dependent variable:

Independent variables:

□ Suppress constant term

Options

● Exposure variable:

○ Offset variable:

Constraints:

□ Keep collinear variables (rarely used)

# Differences

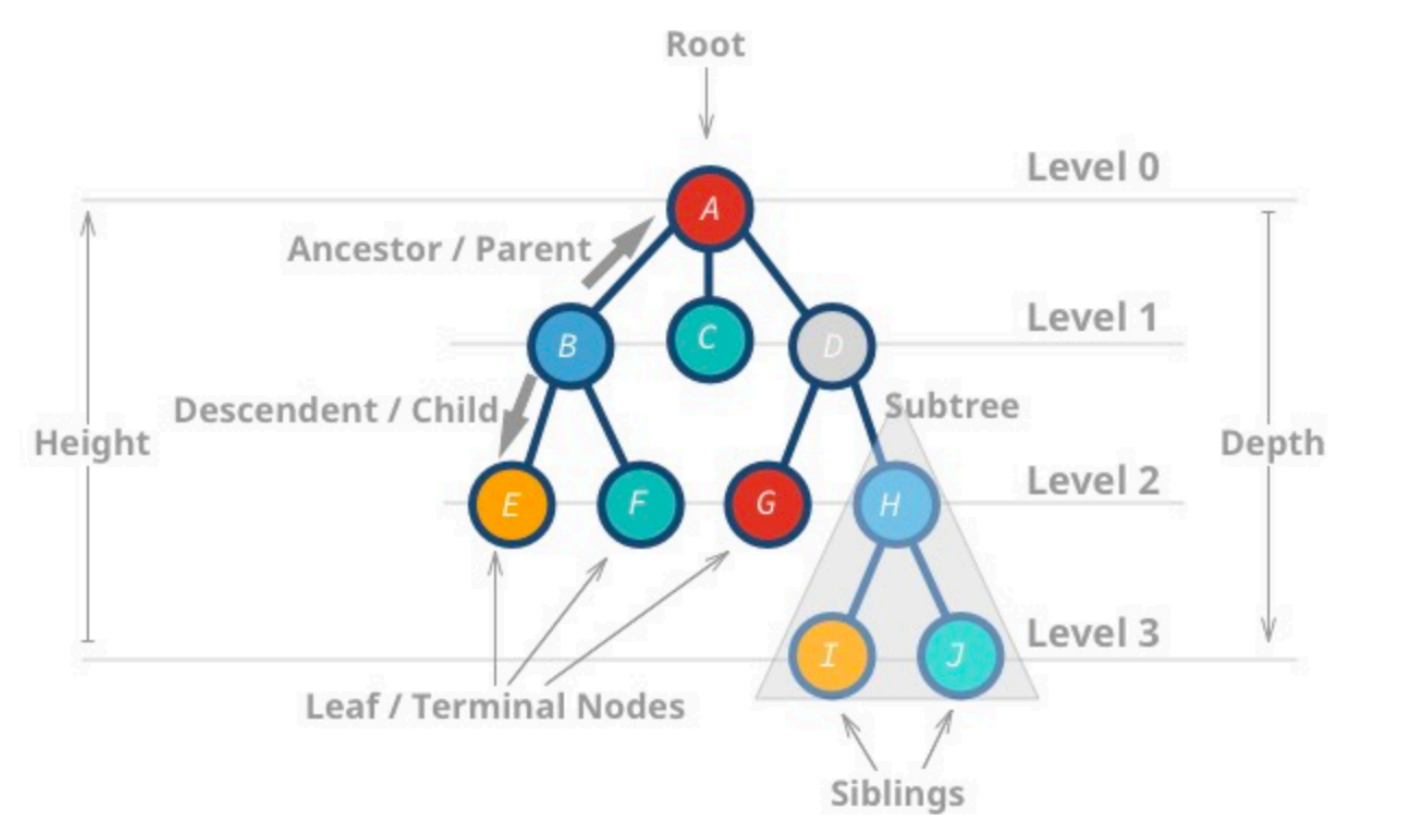
- Hybrid between Syntax and Click based (Do-Files)
- Syntax based (R, Python, Matlab)
- Click based (SPSS)

# AI and ML in Stata

- i.e. module CART (Classification and regression trees):
- Install by typing "ssc install cart”

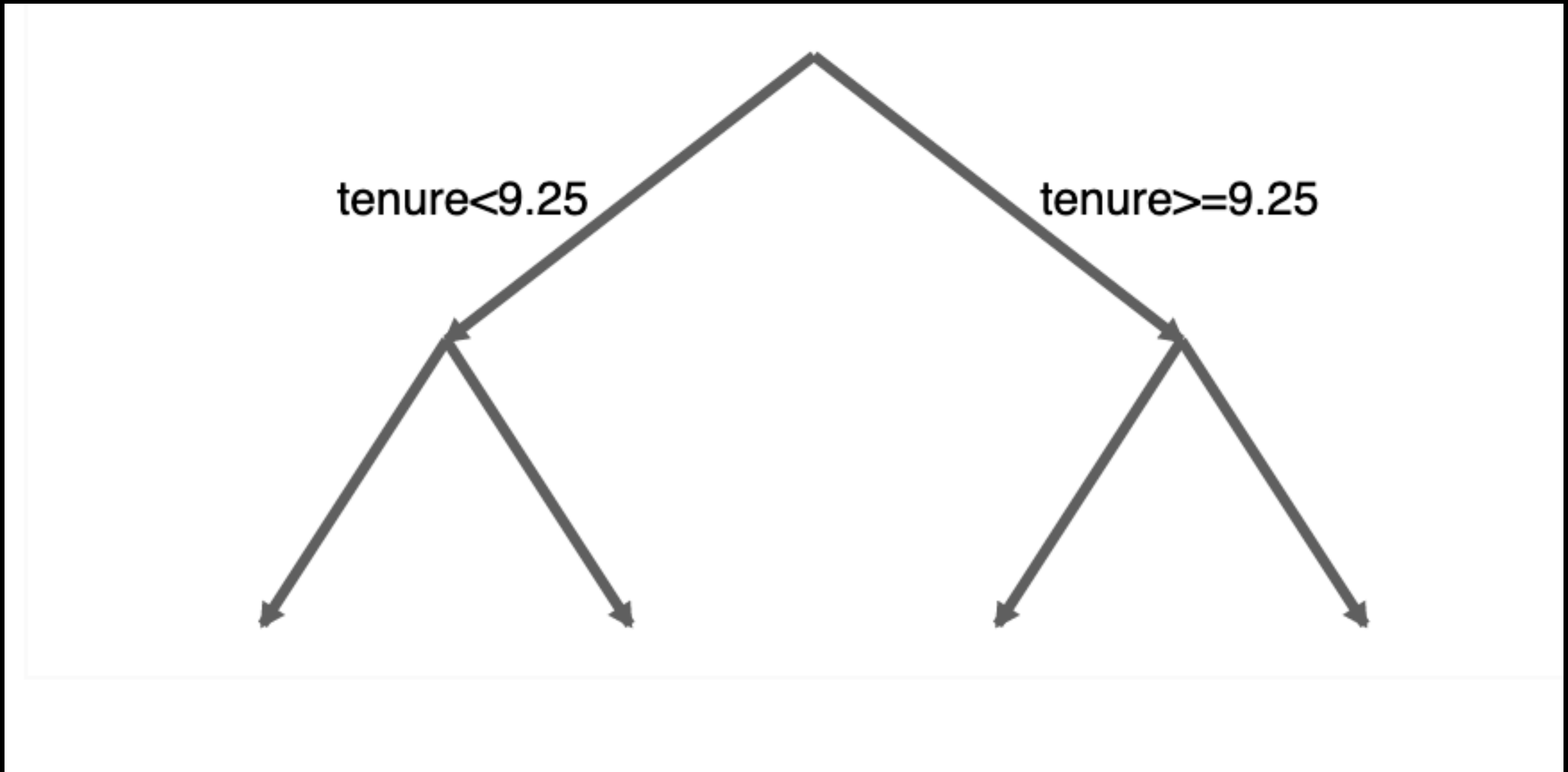
# Trees in Stata

- Root (top node), Nodes, Leafs (terminal nodes), …



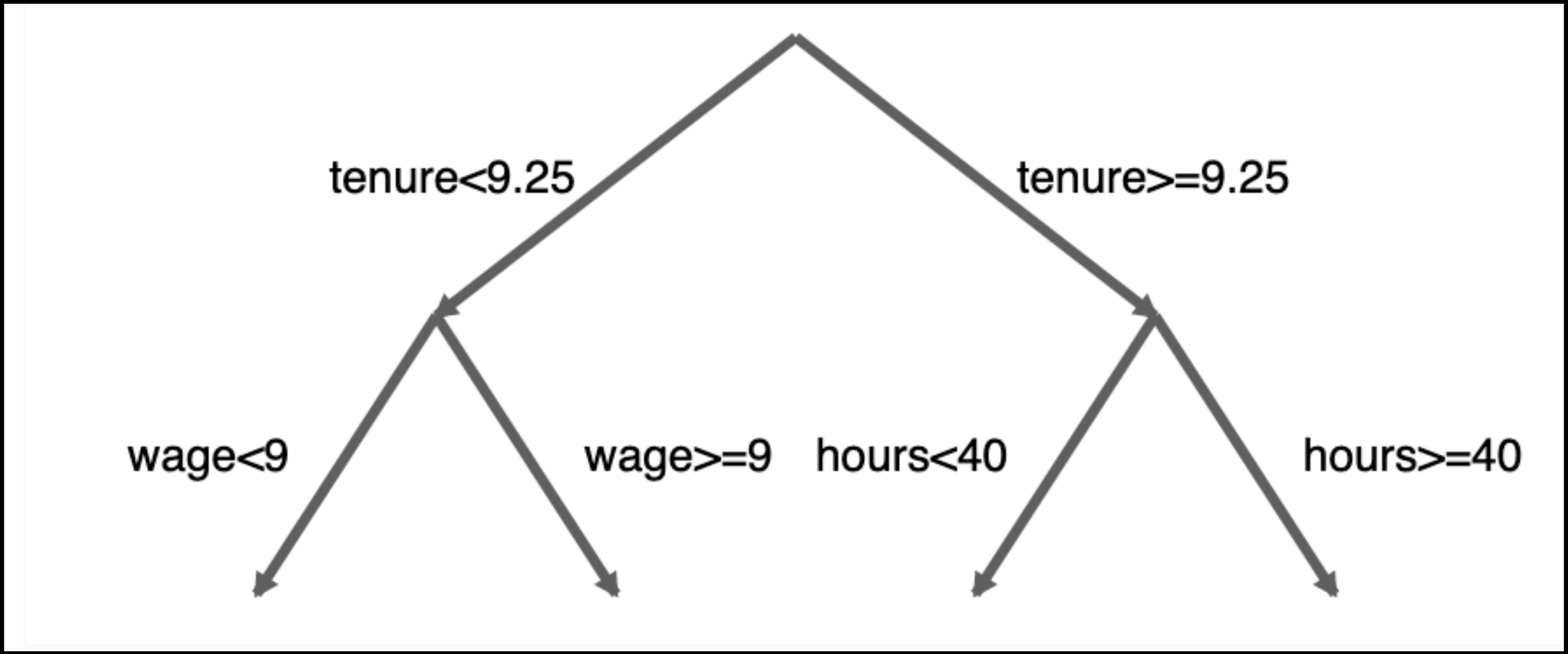
# Trees in Stata

- Classification in two different groups at each node



# Trees in Stata

- Gets complex very fast: 10 levels of binary splits give you 2^10 = 1024 terminal nodes (leaves)



# Trees in Stata

- Splits at the nodes are by default based on SSR, for binary outcomes it is the number of misclassified observations

# Random Forest

- Collection of Decision Trees (default: 500) with a conclusion or final result.

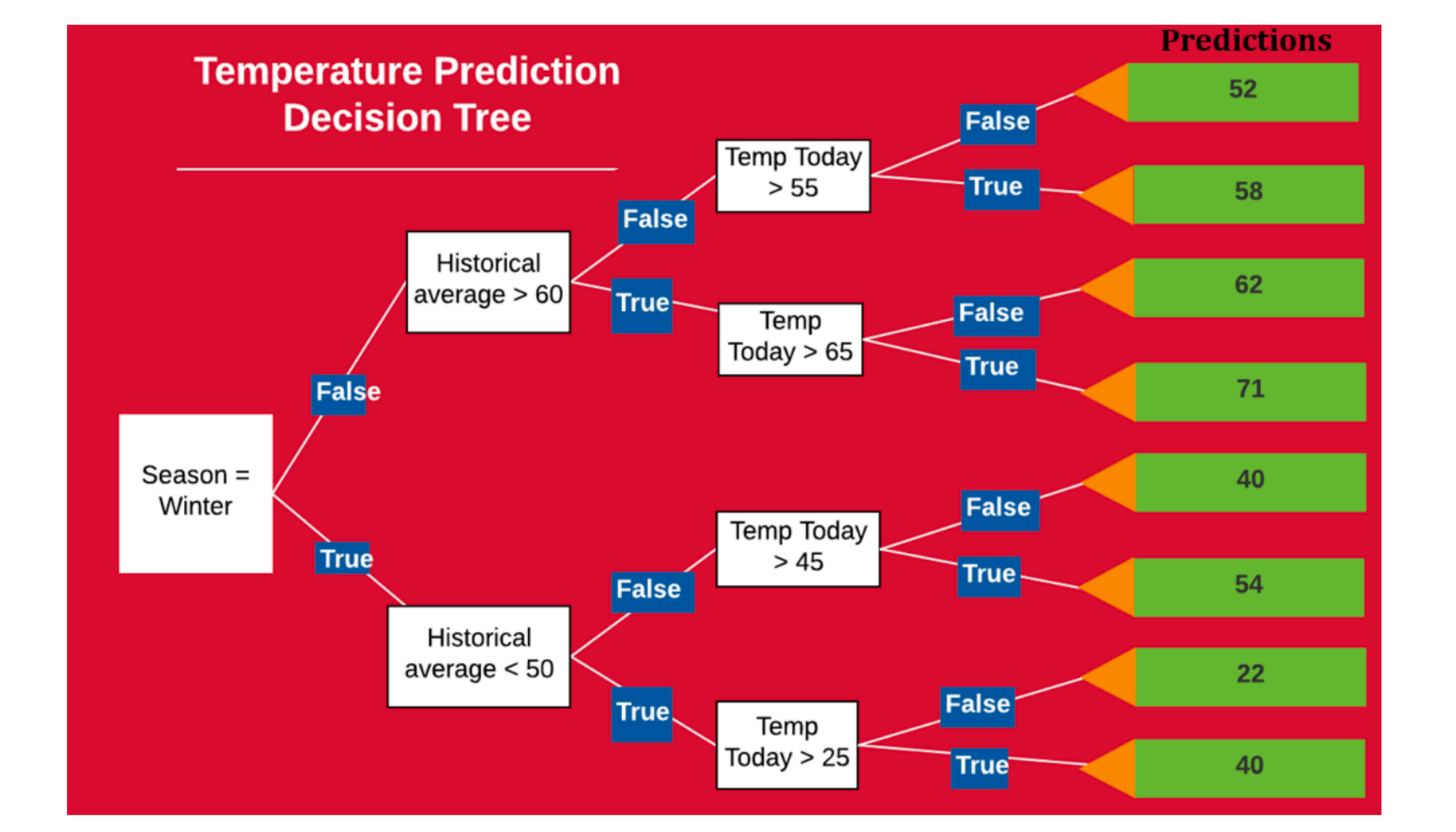
- Random Forest has a very small MSE (Mean Squared Error) in general

- ML has to be able to react well to new data

- 1. Random sampling of training data

- 2. Random subset of features when splitting nodes

# Temperature prediction



# Conclusion

- Hybrid between click and syntax (intermediate level)
- R or Python are better for more advanced tasks
- But easy tasks like subsets are difficult or not possible