



*Source of image: <http://www.collectifbam.fr/thomas-thibault-au-fabshop/>*

**“A proposal for a new Stata licensing scheme based on blockchain, cloud computing, and grid computing”**

**Alexander Zlotnik, PhD**

Technical University of Madrid (Universidad Politécnica de Madrid)

**David Arroyo Manzano, MsSc**

Why?

# Everyone will be using...

“Big data”

+

Complex algorithms

=

Lots of computational resources

# Examples

- (very) big data & simple operations (such as *sort*)
- big data & regression analysis
- big data & multiple imputation
- (just) data & bayesian analysis

# Current Stata solutions

- **Custom programming in C++**
- Stata / MP
- Stata distributed processing  
(several computers)  
...example: Stata PARALLEL

# Custom programming in C++

- Example:

2015 UK Stata Users Group meeting



## **Big Data in Stata**

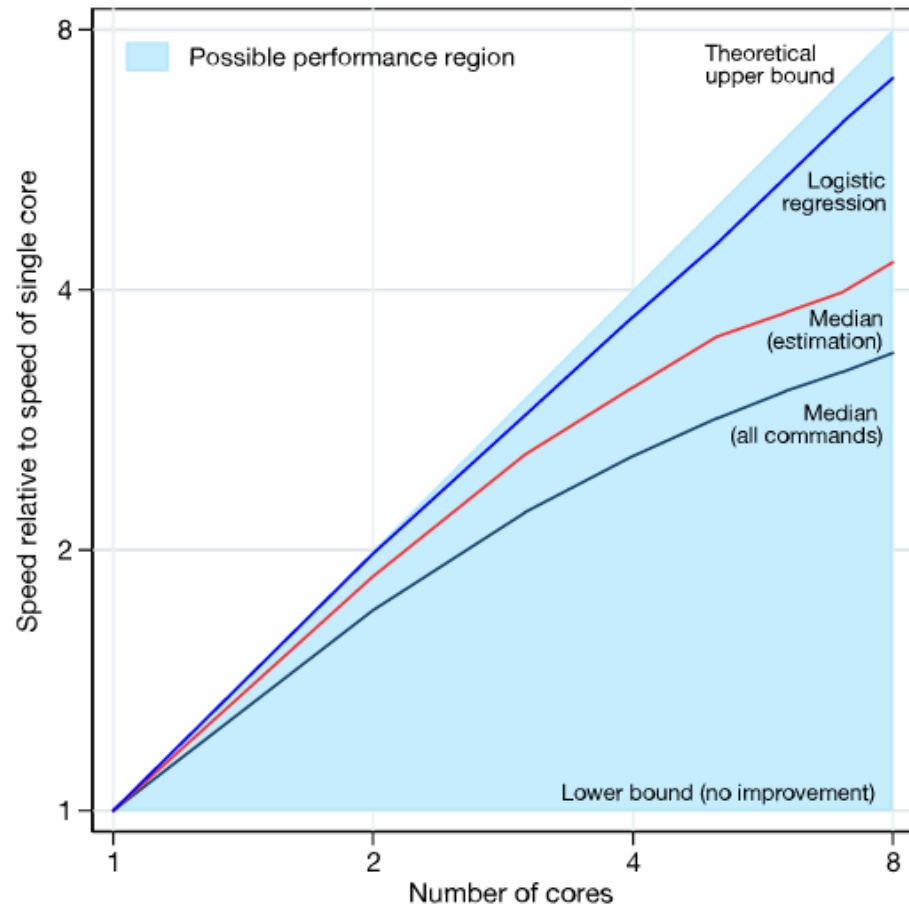
Andrew Maurer

*Quantitative Risk Management*

# Current Stata solutions

- Custom programming in C++
- **Stata / MP**
- Stata distributed processing  
(several computers)  
...example: Stata PARALLEL

# Stata / MP



Source: <https://www.stata.com/statamp/>



# Current Stata solutions

- Custom programming in C++
- Stata / MP
- **Stata distributed processing**  
(several computers)  
...example: **Stata PARALELL**

# Stata PARELLEL

GitHub, Inc. (US) | <https://github.com/gvegayon/parallel>

Features Business Explore Marketplace Pricing

gvegayon / parallel

Code Issues 7 Pull requests 0 Projects 0 Wiki Insights

## Authors

George G. Vega [aut,cre] g.vegayon %at% gmail

Brian Quistorff [aut] Brian.Quistorff %at% microsoft

PARALLEL: Stata module for parallel computing

stata parallelization bootstrap simulation hpc parallel

325 commits 3 branches 4 releases 3 contributors MIT

Branch: master New pull request

Find file Clone or download





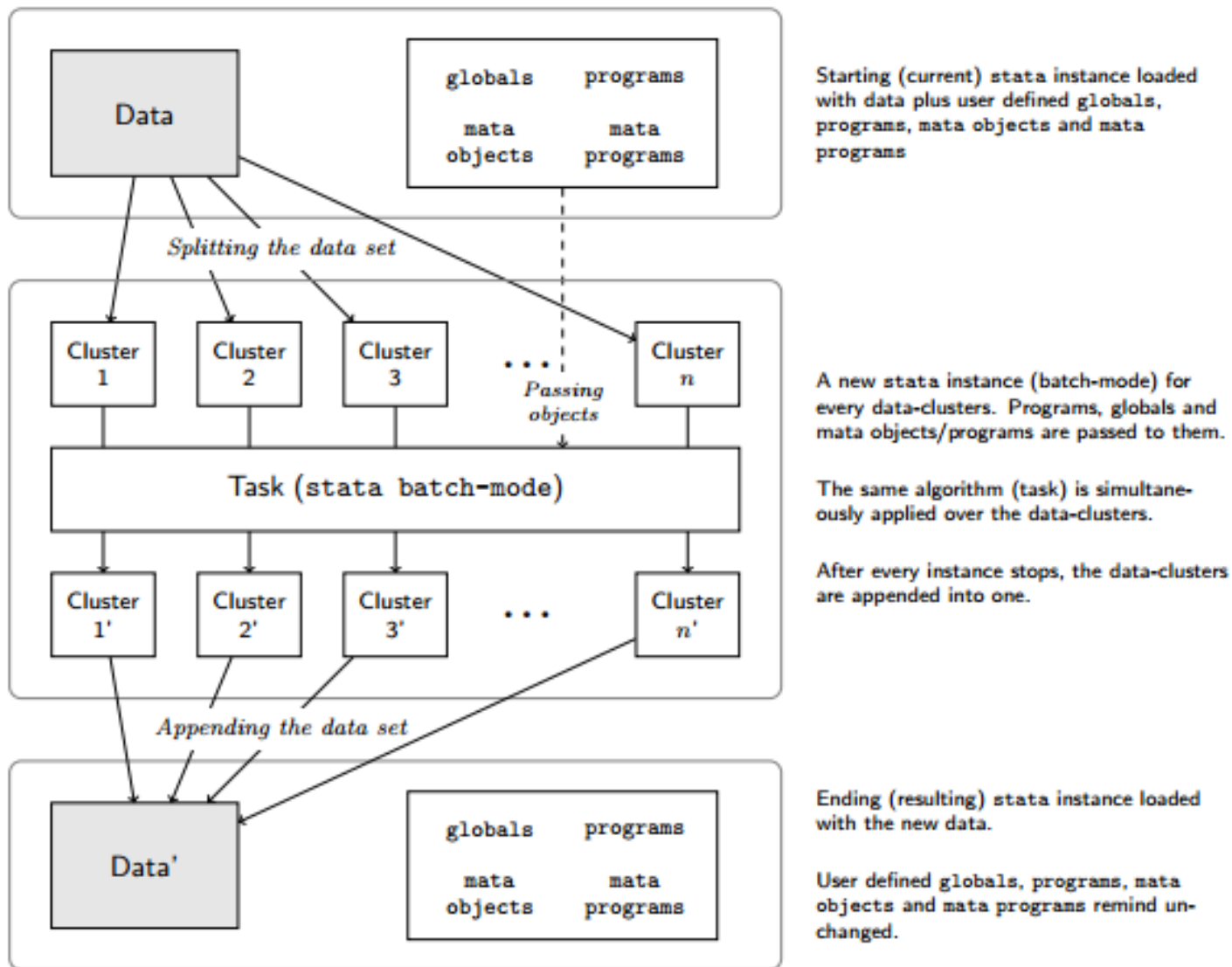
 <b>bquistorff</b> update HTML from sthlp update	Latest commit 72ee46F on 29
 .github	Update ISSUE_TEMPLATE.md 3 months
 ado	update HTML from sthlp update 2 months
 man	Normalize all the line endings 3 years

Figure 1: How parallel works



Starting (current) stata instance loaded with data plus user defined globals, programs, mata objects and mata programs

A new stata instance (batch-mode) for every data-clusters. Programs, globals and mata objects/programs are passed to them.

The same algorithm (task) is simultaneously applied over the data-clusters.

After every instance stops, the data-clusters are appended into one.

Ending (resulting) stata instance loaded with the new data.

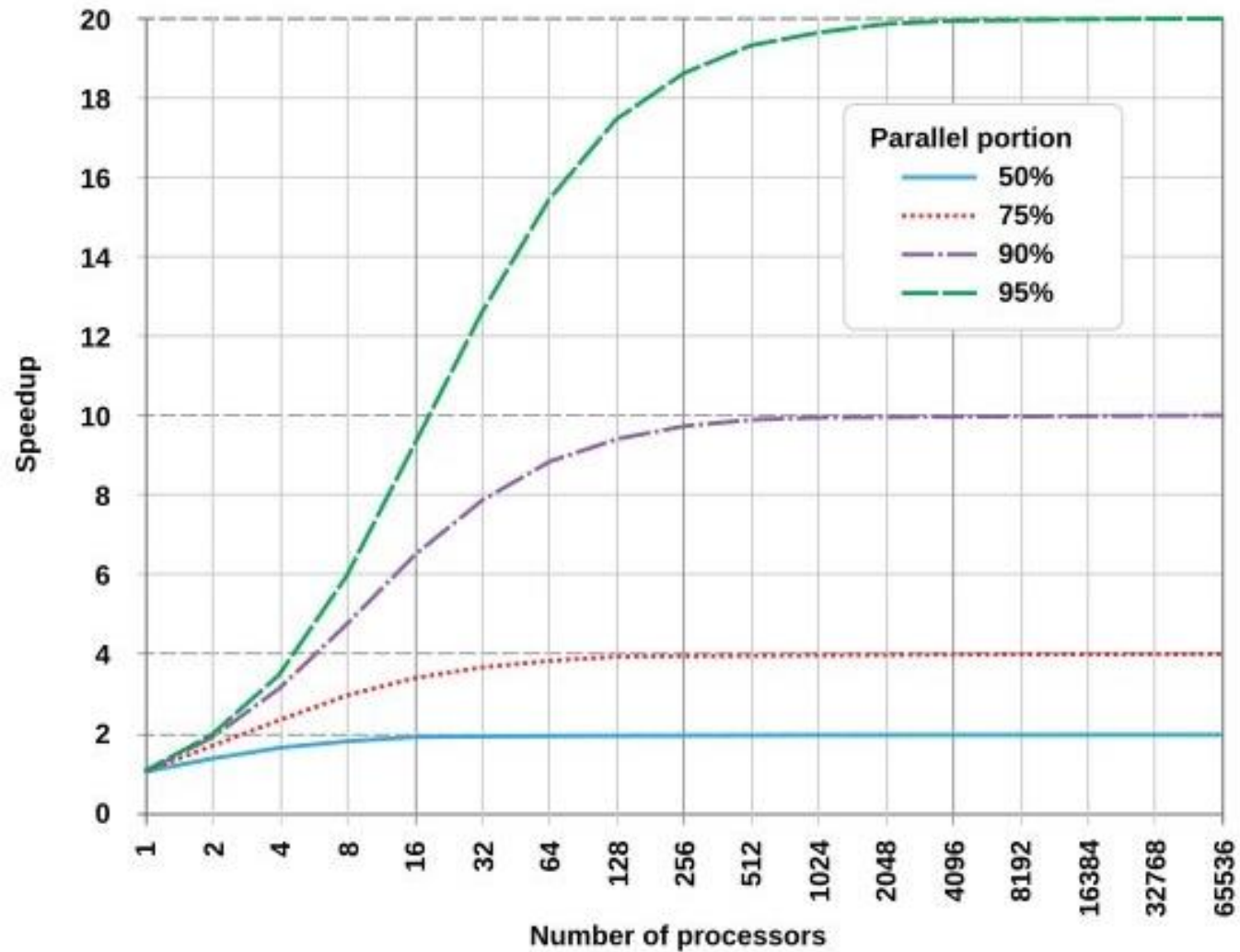
User defined globals, programs, mata objects and mata programs remind unchanged.

# Distributed processing

- **Important concepts**
  - **Algorithmic complexity**
  - **Ahmdal's law**
- **Decision criteria**



# Ahmdal's law



# Distributed processing

- Important concepts
  - Algorithmic complexity
  - Ahmdal's law
- **Decision criteria**

# Decision criteria

- **High n ?**
- **$O(n)$  = algorithmic complexity ?**
- **Parallelizable code (Ahmdal's law) ?**



# Example: Multiple Imputation

- **High n ?**
  - Many experiments => High n => Yes
- **$O(n)$  = algorithmic complexity ?**
  - $O(n) \approx n$  (regressions)
- **Parallelizable code (Ahmdal's law) ?**
  - Many independent experiments => Yes

# Ideas for future Stata versions

# Ideas for future Stata versions

- **Stata private cloud**
- Stata public cloud (grid computing)  
... with blockchain licensing

# Chessbase private cloud

The screenshot displays the Chessbase software interface. The main window title is "Anand,Viswanathan (2797) - Carlsen,Magnus (2865), Grenke Chess Classic 3rd 2015 - A90 (Roiz,M), 0-1". The interface includes a menu bar (File, Home, Insert, Board, Training, Analysis, Engine, View, Help) and a toolbar with options like Board Sounds, Coordinates, Always Promote To Queen, and various board views (Square, Pieces, Table, 3D Board, Clocks). The central area shows a chessboard with pieces in their starting positions. The right-hand pane is titled "Notation + Openings Book" and contains the following text:

**Anand,Viswanathan 2797 - Carlsen,Magnus 2865 0-1**  
**A90 Grenke Chess Classic 3rd Baden-Baden (4) 06.02.2015 (Roiz,M)**

**28...Qf7! 29.Re6!** Vishy manages to find the best way to develop his counterplay.  
[ 29.Rd6 was much worse: Rfe8 30.Rde6 Ng4 31.Bxg4 hxg4 32.Qb3 Rxe6 33.fxe6 Qf3 34.Qc4 Qd5 35.Qc2 Kg7 36.e7 Re8 37.Re3 Bd4 38.Rxa3 Rxe7-+ ]

**29...Ng4?** This natural move is not the best from an objective point of view, though White's task is becoming extremely tough.  
[ After the correct 29...Rfe8! 30.b5 cxb5 31.Bg2 b4 32.Bf1 Kh8 33.Bc4 Rxe6 34.fxe6 Qe7 ♠ White still would have some counter-chances, but Black's advantage is indisputable there. ]

Below the text, there is a "Komodo 9.02 64-bit" engine control panel with buttons for "Stop", "3 CPUs", and "Cloud". It also shows engine statistics: "♙ ♚ (-0.52)", "Depth=27", "29...Kh8 (1/39)", and "5049 kN/s". A list of moves is displayed:

1. ♚ (-0.52): 29...Kh8 30.Rde1 Rae8 31.Qc4 Qg7  
2. ♚ (-0.35): 29...Rfe8 30.b5 Rxe6 31.fxe6 Qe7

A smaller chessboard is visible in the bottom right corner of the interface.

# Stata private cloud proposal

- Do some processing locally.
- Remove all identifying information (variable names, variable encoding, non-numerical values, et cetera).
- Send complex optimization problems to Stata cloud.
- Get results in local instance of Stata.

# Ideas for future Stata versions

- Stata private cloud
- **Stata public cloud (grid computing)  
... with blockchain licensing**

# Stata public cloud (grid)

- Many computers ...  
... in different geographical locations  
... working on the same problem
  
- Example: SETI@home

## Data analysis

Chirping data

Doppler drift rate -16.8770 Hz/sec Resolution 0.075 Hz

Best Gaussian: power 2.31, fit 0.480, score 4.814



Overall 86.364% done

CPU time: 3 hr 4 min 27.06 sec

## Data info

From: 15 hr 2' 40" RA, +14 deg 23' 34" Dec

Recorded on: Tue Mar 16 08:32:04 2004

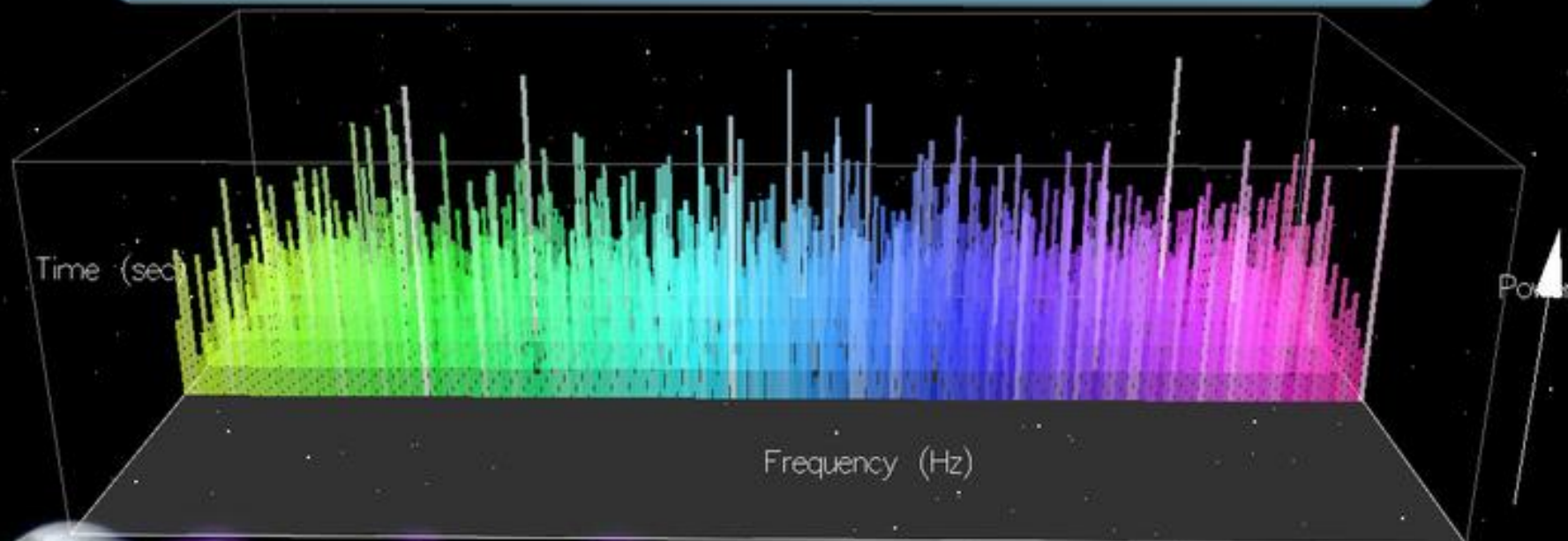
Base frequency: 1.421220703 GHz

## User info

Name: nnn

Team:

Total credit: 3166.43



# SETI@home

The Search for Extraterrestrial Intelligence



# Stata public cloud (grid)

- The same approach could be used with Stata.
- But... how could Stata users be incentivized to provide their instances of Stata for distributed processing?
  - With **blockchain** licensing !

# What is a blockchain?



# Blockchain applications

- blockchain = distributed database (distributed ledger) with transactional integrity guarantees not controlled by a single entity based on many processing nodes (anonymous or publicly known).
- It is very hard (almost impossible, given certain conditions) to falsify an entry in the blockchain.

# Blockchain applications

- Civil registries.
- Land ownership registries.
- Notary registries.

# Blockchain applications

- Cryptocurrencies (Bitcoin, “ether”, etc) which are not controlled by a central bank (or any kind of central entity).
- International financial transactions (alternatives to the SWIFT system).

# Blockchain applications

- Smart electricity grids (intelligent electricity production, distribution and billing).
- Distributed organizations (such as cooperatives with no managers).
- e-Administration / Open Government (Malta, Russia, Ukraine, Estonia, ...)

# Stata public cloud + **Blockchain licensing**

- “Free” Stata license which is paid for by computational time for Stata Corp.
- Computational time given to Stata Corp is logged in a blockchain thus guaranteeing transparency and irrevocability.

Thank you !