Architektura dużych projektów bioinformatycznych

Pakiety do obliczeń: naukowych, Inżynierskich i statystycznych Przegląd i porównanie

Bartek Wilczyński

10.4.2019

Plan na przyszły tydzień: quiz

- Kto używał debugerów: pdb, pycharm, idle
- Kto używał profilerów: profile,cProfile, kernprofiler, runsnakerun
- Kto używał narzędzi do benchmarkowania: timeit, pystone, etc.
- Kto używał modułu inspect
- Kto używał systemów do logowania/raportowania błędów (logging, warnings)?

Plan na dziś

- Pakiety do obliczeń: przegląd zastosowań
- różnice w zapotrzebowaniu: naukowcy, inżynierowie, statystycy/medycy
- Matlab/octave/scipy
- S-Plus/SPSS/projekt R
- Mathematica/Maxima/Sage
- Pakiety komercyjne vs. Open Source
- Excel?

Typowi użytkownicy pakietów obliczeniowych

- Inżynierowie i projektanci (budownictwo, lotnictwo, motoryzacja, itp.)
- Naukowcy doświadczalni (fizycy, chemicy, materiałoznawcy, itp.)
- Statystycy (zastosowania w medycynie, ekonomii, biologii molekularnej, psychologii, socjologii, ubezpieczeniach, itp.)
- Matematycy (przede wszystkim matematyka stosowana)

Obliczenia naukowe

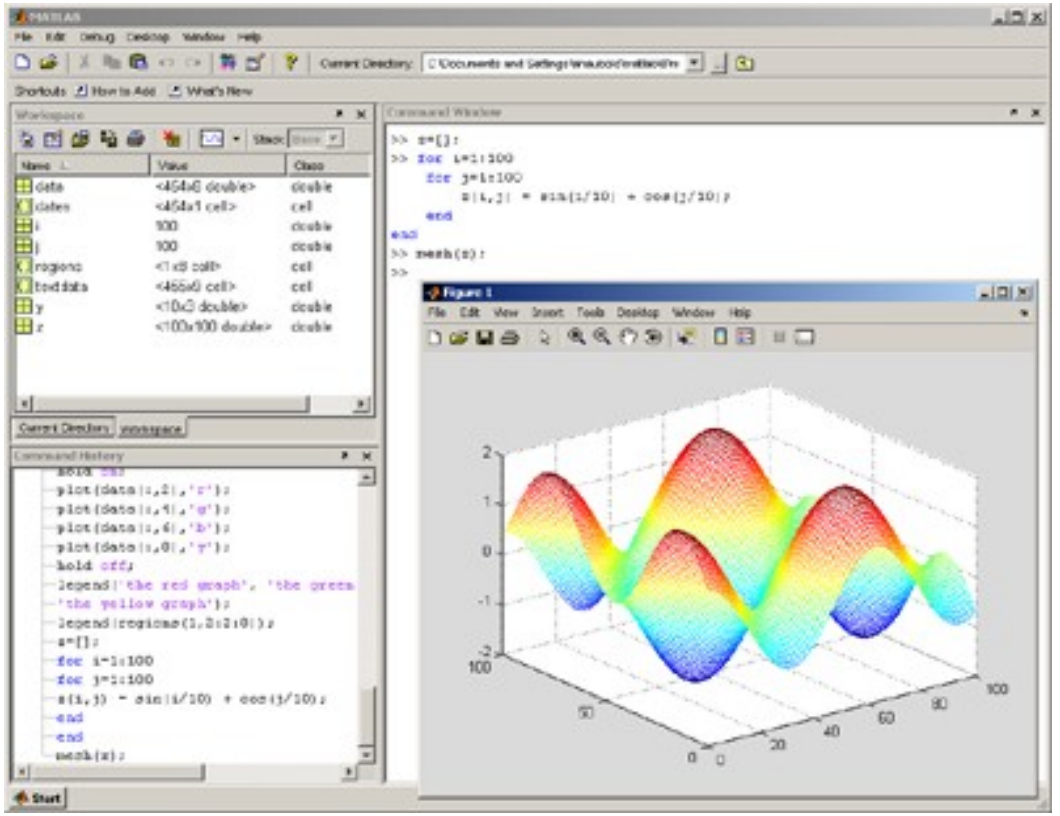
- Komputer jako “potężniejszy kalkulator”
- W zasadzie wszystko można zaprogramować samemu, ale każdemu mogą się przydać:
  - Interfejs użytkownika łatwiejszy niż typowego kompilatora
  - Możliwość zaawansowanej grafiki
  - Dobrze przetestowane standardowe procedury
  - Interfejsy do urządzeń
  - Wsparcie fachowców

Matlab i pakiety “inżynierskie”

- Rozwijany w latach 70'tych przeze Clive Moler'a jako narzędzie dla studentów informatyki, aby nie musieli używać zaawansowanych bibliotek fortranu
- Firma mathworks powstaje w 1984 i wydaje pierwszą wersję Matlab'a
- Najpopularniejszy wśród inżynierów, dobre całki numeryczne, rozwiązywanie równań i wykresy (również 3d)
- Bardzo popularny także do przetwarzania sygnałów i symulacji (simulink)
- Licencja komercyjna – niedrogi dla studentów, droższy dla uczelni, bardzo drogi dla przemysłu

Toolbox'y Matlab'a

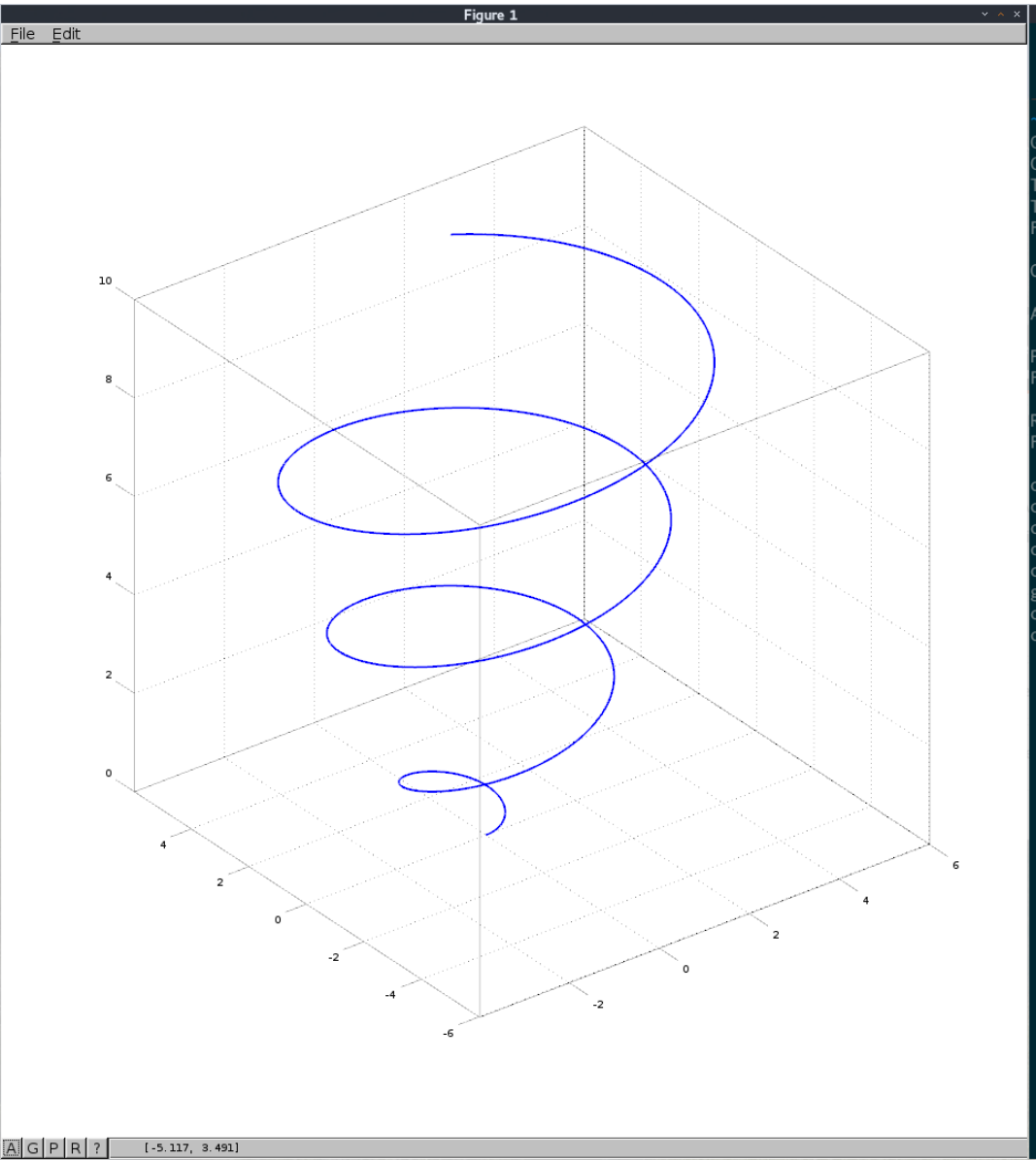
- Wiele dodatkowych (płatnych) bibliotek dla specjalistów
  - Symbolic math
  - Image processing
  - Financial toolbox
  - Bioinformatics
  - Optimization
  - SimBiology

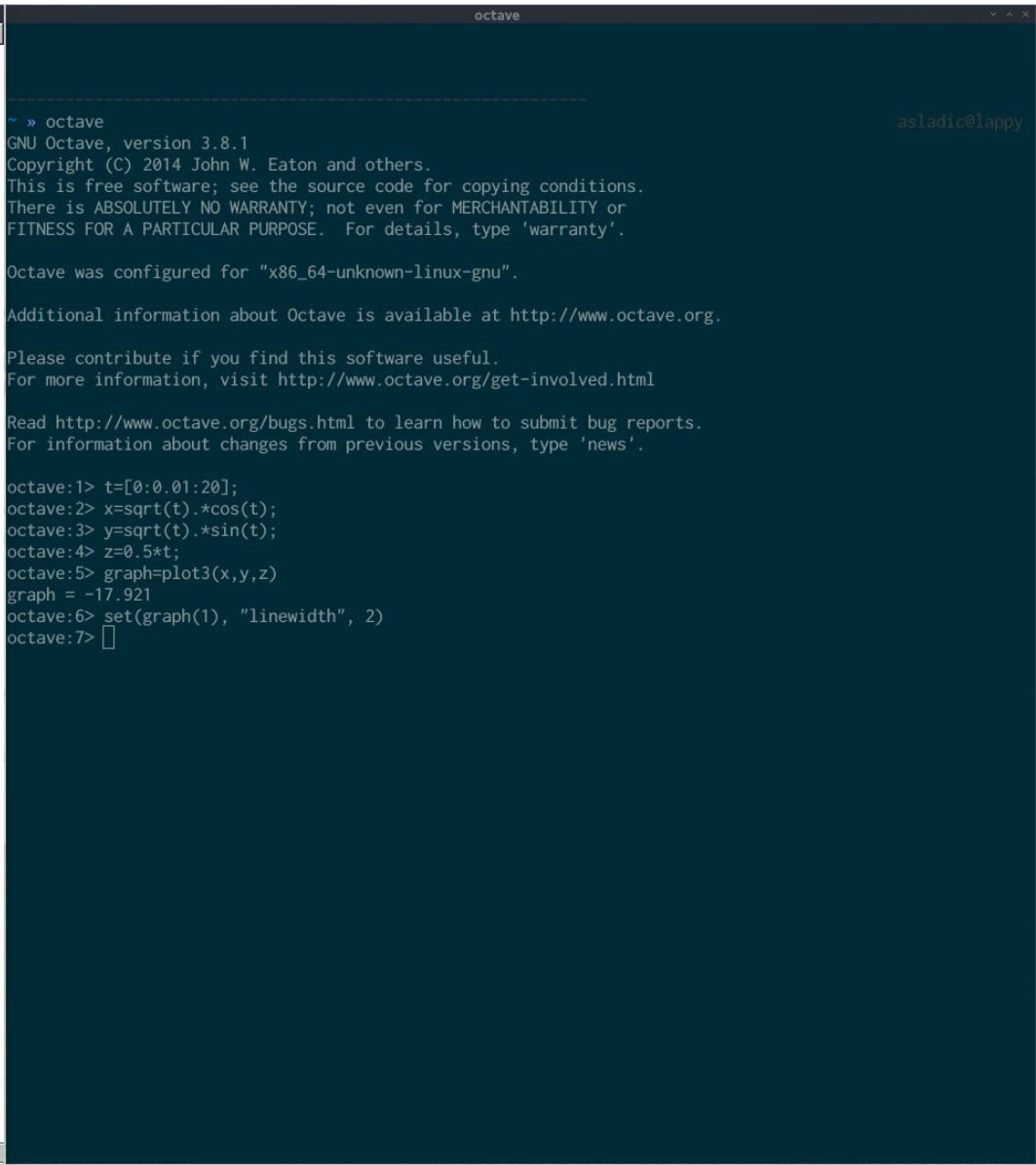


Alternatywy openSource

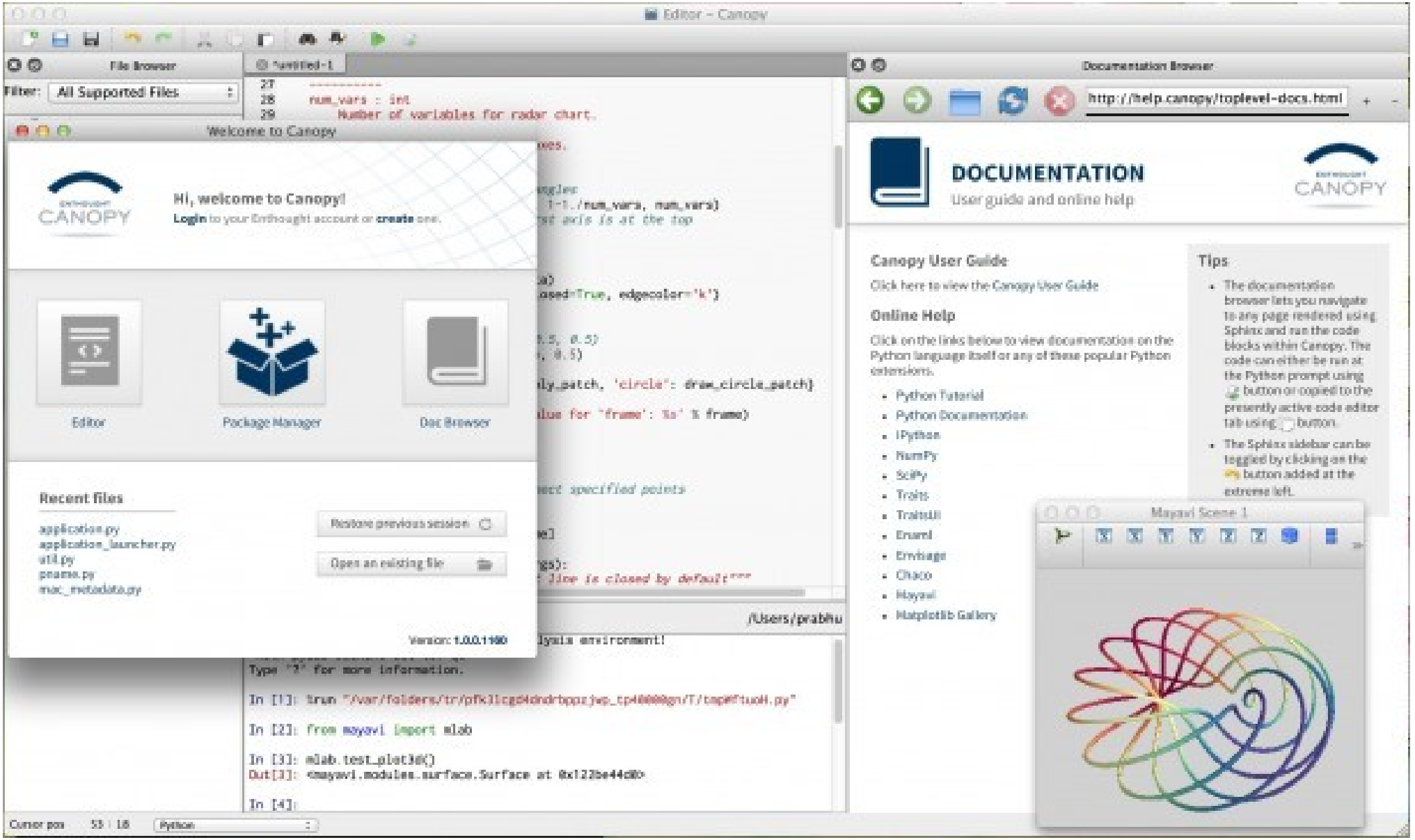
- GNU Octave (rozpoczęty w 1988, wydania od 1992, rozwijany przez John'a W. Eatona, chemika z University of Wisconsin-Madison)
  - W zasadzie kompatybilny z Matlab'em
  - John W. Eaton Inc. - consulting, firma, która czerpie dochody z konsultingu Octave’a
- SciLab – dawniej psilab, rozwijany od lat 90tych w INRIA we Francji, przekształcony w spółkę Scilab enterprises, przejęty w 2017 przez ESI Group, dostępny na licencji GNU
- Scipy stack – zestaw bibliotek python'a do obliczeń naukowych
  - Wiele bibliotek, rozwijanych przez niezależne grupy
  - System pakietów, edytor i dystrybucja organizowana przez firmę Enthought, również komercyjne dystrybucje i konsulting
  - Wiele konferencji tematycznych dla naukowców I pracowników przemysłu - także źródło dochodu

Interfejs Octave





Interfejs Enthought Canopy



S-Plus dla statystyków

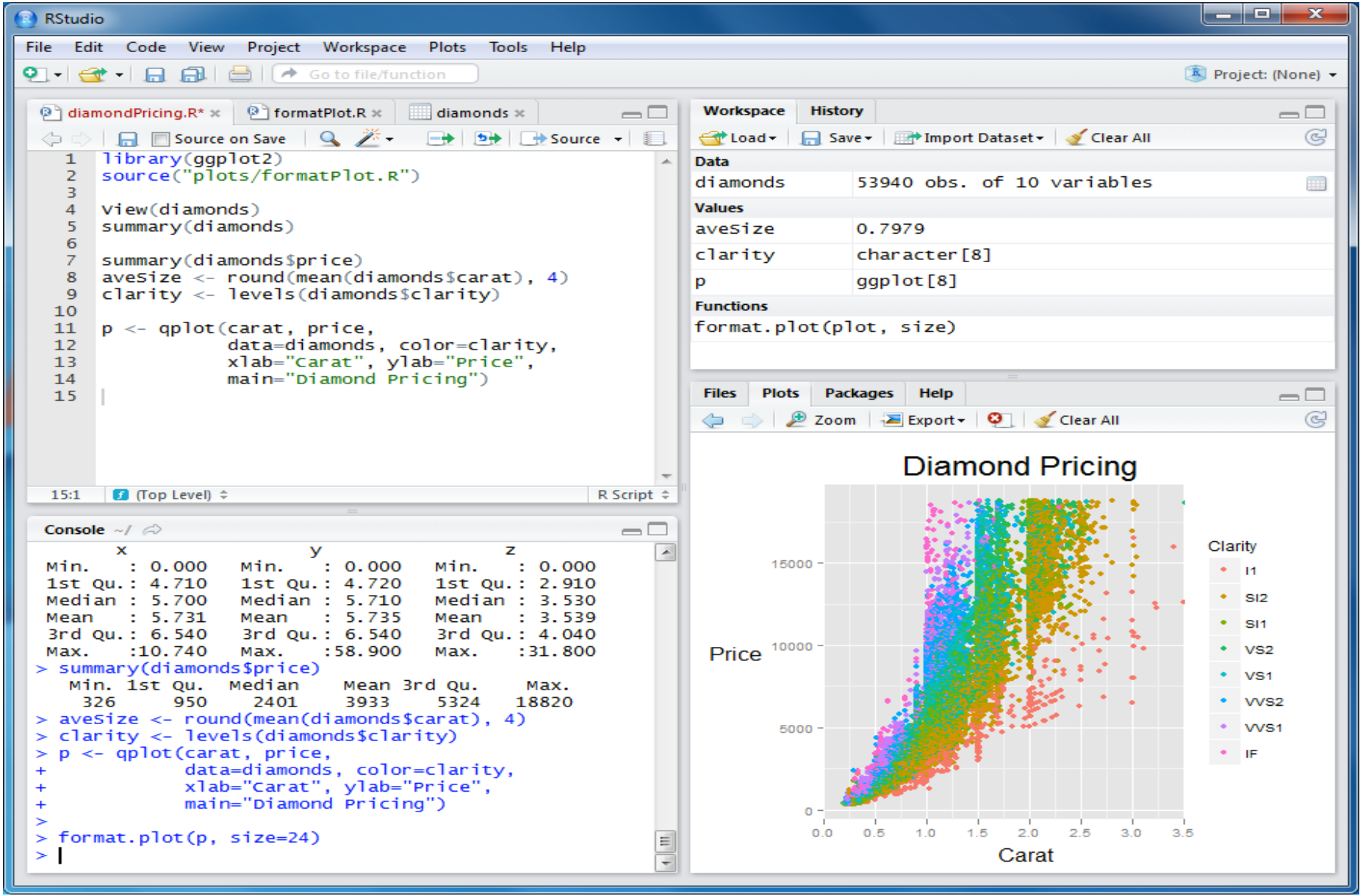
- Język S zaprojektowany w laboratoriach Bell Labs przez Johna Chambers'a
- Implementacja przez R. Douglas'a Martina, profesora statystyki w Seattle
- Wydany komercyjnie w 1988 jako S-Plus, potem kolejno “przejmowany” przez różne korporacje aż do 2008, kiedy przejęła go firma TIBCO
- Adresowany do statystyków akademickich I przemysłowych
- Ogólny, bez specjalizacji w jakiejś dziedzinie zastosowań

Projekt R – Implementacja OpenSource języka S

- Rozpoczęty ok. 1995 projekt stworzenia darmowej implementacji języka S
- Ross Ithaka (obecnie Genentech) and Robert Gentelman (obecnie Univ. Auckland)
- W tej chwili zarządzany przez “R foundation”
- Wiele firm “wspierających” R
- Zachęcam do obejrzenia slajdów Chambers'a:

  www.r-project.org/useR-2006/Slides/Chambers.pdf

Rstudio – interfejs do R'a



System pakietów w R

- Istotna jest możliwość rozwijania własnych “pakietów” w R (coś na kształt “toolbox'ów” Matlab'a)
- Jest to proces całkowicie demokratyczny, każdy może wysłać pakiet i umieścić go w repozytorium CRAN (Comprehensive R Archive Network)
- Możliwość automatycznej instalacji pakietów z CRAN
- Pewne standardy dokumentacji (winiety) zgodne z “literate programming”

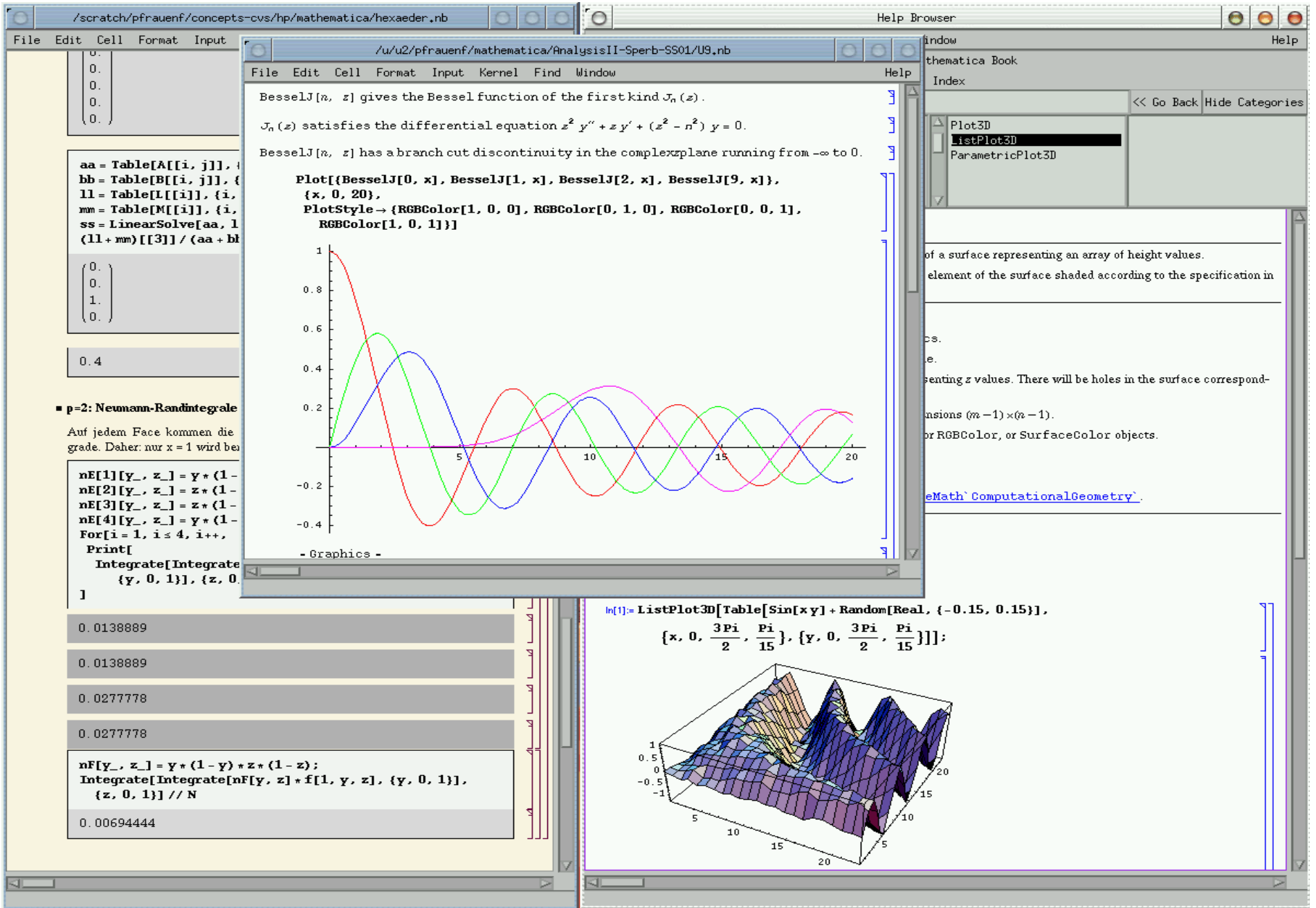
Bioconductor

- Dystrybucja wybranych pakietów R do analizy danych biomedycznych
- Nacisk na łatwiejszą instalację i lepszą dokumentację pakietów
- Stabilny cykl wydań (2 razy do roku)
- Szkolenia adresowane do biologów I medyków
- Kompletnie not-for-profit
- Finansowany z grantów (ok 7-8 osób core team)

Obliczenia symboliczne - Mathematica

- Opracowana w latach 1980'tych przez Stephen'a Wolframa
- Jeden z pierwszych w historii pakietów umożliwiających obliczenia symboliczne
- Bardzo popularna wśród studentów amerykańskich, którzy muszą “zaliczyć” rachunek różniczkowy
- Obecnie także w wersji online: Wolfram Alpha
- Konkurencyjne pakiety: Maple, Mathcad

Mathematica - interfejs



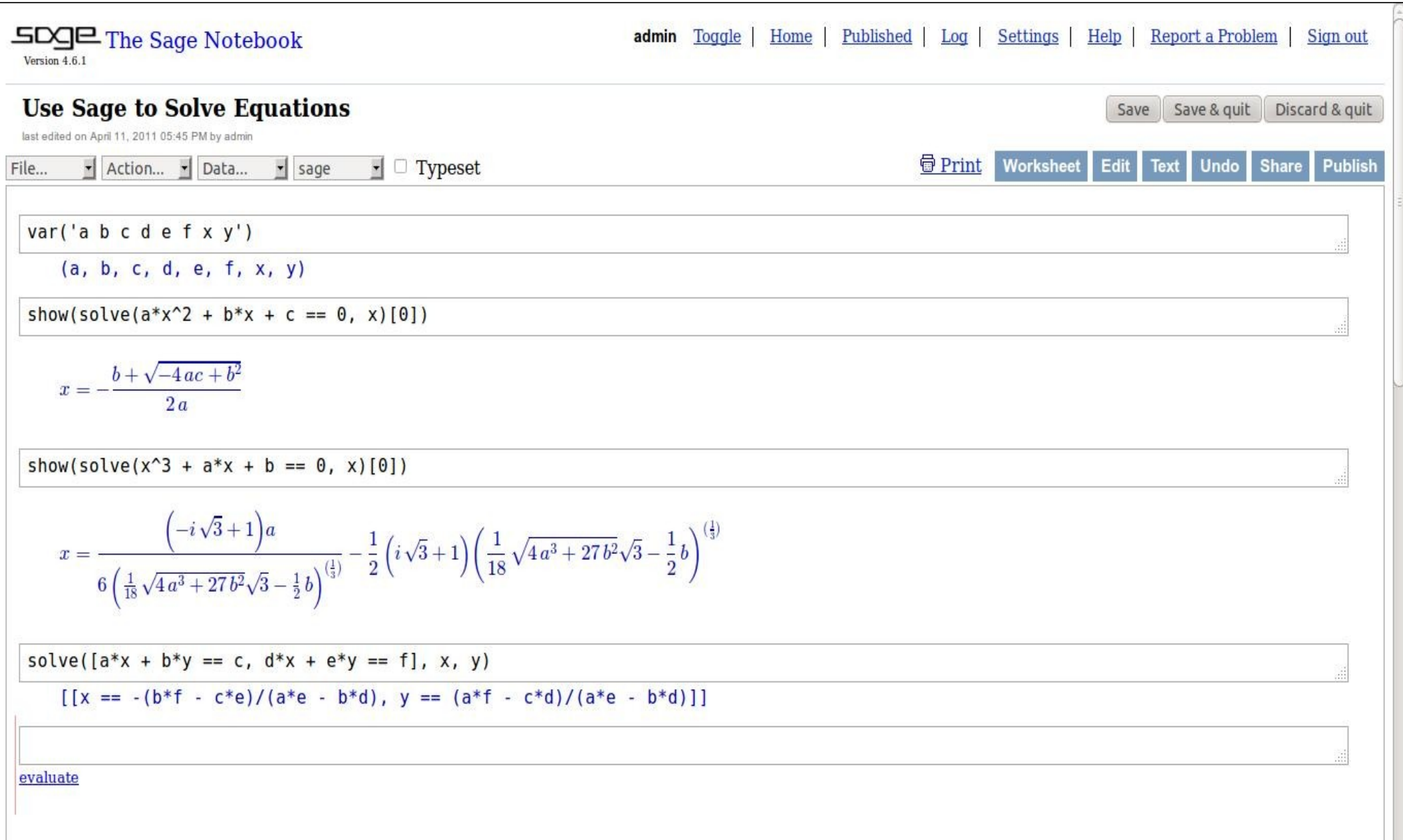
Obliczenia symboliczne Open Source

- Maxima (1992-), a wcześniej Macsyma (1968-1982)
- Wydana w 1998 na licencji GPL
- Napisana w języku lisp
- Wiele konkurencyjnych interfejsów (WXMaxima, Gmaxima itp)
- Maxima skupiona na obliczeniach symbolicznych, bez większej funkcjonalności w numeryce

SAGE math cloud
(dawniej sage notebook)

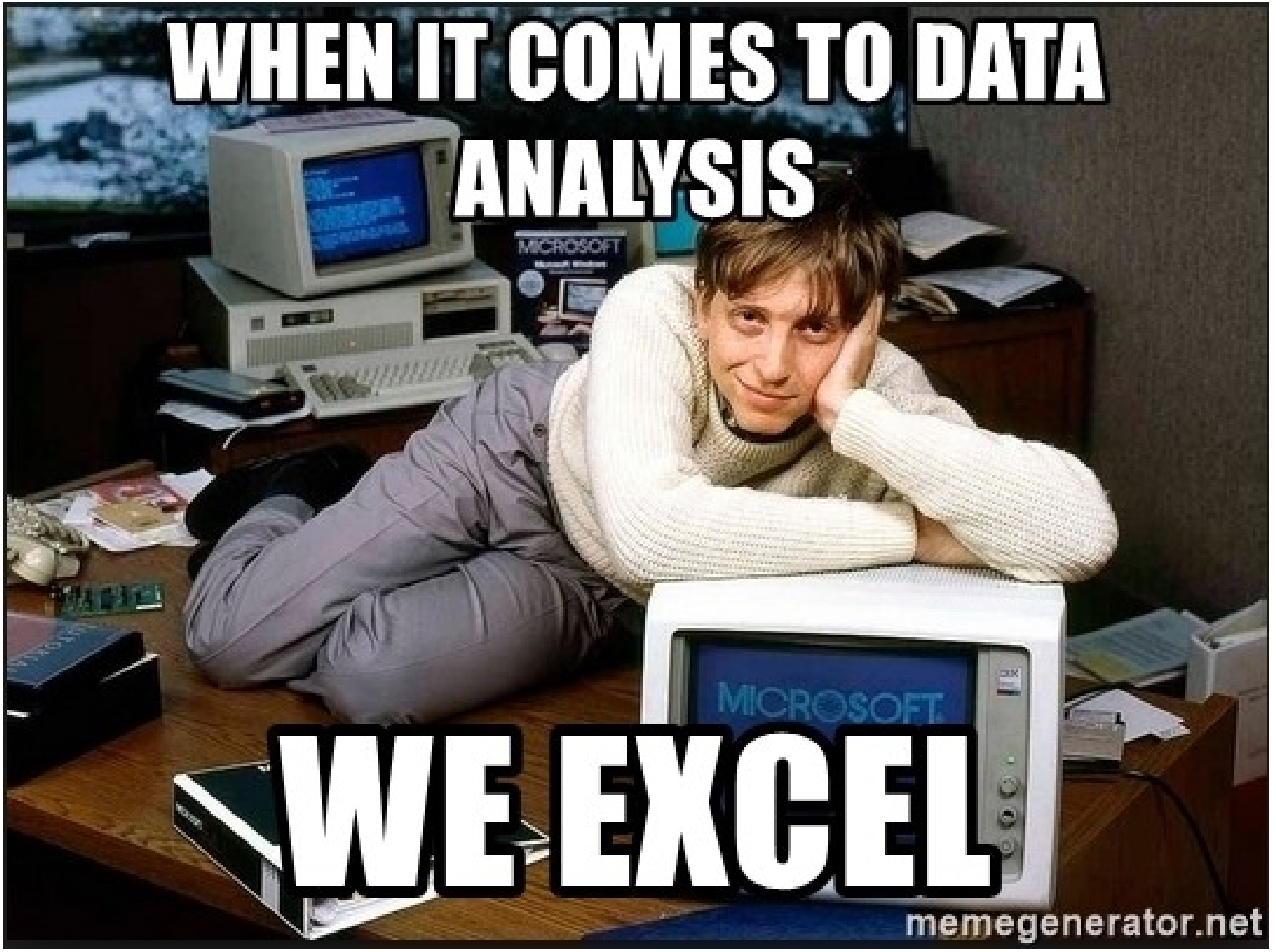
- Stosunkowo nowy projekt (inny niż sage synapse)
- Połączenie wielu środowisk obliczeniowych
  - Python (Numpy, Scipy, Sympy, matplotlib, Networkx)
  - Maxima
  - R
  - GAP, FLINT, GD, JMOL, PALP, Singular
- Środowisko w przeglądarce, sesja na serwerze lub “w chmurze”

Interfejs SAGE



Excel?

- Najpopularniejszy pakiet do obliczeń
- Bardzo prosty interfejs
- Często stosowany również w bio-informatyce
- Ma spore ograniczenia (np. Maksymalna liczba linii w arkuszu), które utrudniają rozwój projektów prowadzonych w arkuszu
- Brak możliwości efektywnego testowania,
- brak debuggerów


WHEN IT COMES TO DATA ANALYSIS
MICROSOFT
MICROSOFT
WE EXCEL
memegenerator.net

Ziemann et al. Genome Biology (2016) 17:177
DOI 10.1186/s13059-016-1044-7

Genome Biology

COMMENT Open Access

Gene name errors are widespread in the scientific literature

Mark Ziemann[1], Yotam Eren[1,2] and Assam El-Osta[1,3*]

Abstract

The spreadsheet software Microsoft Excel, when used with default settings, is known to convert gene names to dates and floating-point numbers. A programmatic scan of leading genomics journals reveals that approximately one-fifth of papers with supplementary Excel gene lists contain erroneous gene name conversions.

Keywords: Microsoft Excel, Gene symbol, Supplementary data

Abbreviations: GEO, Gene Expression Omnibus; JIF, journal impact factor

The problem of Excel software (Microsoft Corp., Redmond, WA, USA) inadvertently converting gene symbols to dates and floating-point numbers was originally described in 2004 [1]. For example, gene symbols such as *SEPT2* (Septin 2) and *MARCH1* [Membrane-Associated Ring Finger (C3HC4) 1, E3 Ubiquitin Protein Ligase] are converted by default to '2-Sep' and '1-Mar', respectively. Furthermore, RIKEN identifiers were described to be automatically converted to floating point numbers (i.e. from accession '2310009E13' to '2.31E+13'). Since that report, we have uncovered further instances where frequently reused. Our aim here is to raise awareness of the problem.

We downloaded and screened supplementary files from 18 journals published between 2005 and 2015 using a suite of shell scripts. Excel files (.xls and.xlsx suffixes) were converted to tabular separated files (tsv) with ssconvert (v1.12.9). Each sheet within the Excel file was converted to a separate tsv file. Each column of data in the tsv file was screened for the presence of gene symbols. If the first 20 rows of a column contained five or more gene symbols, then it was suspected to be a list of gene symbols, and then a regular expression (regex) search of the entire column was applied to identify gene symbol errors. Official gene symbols from Ensembl version 82, accessed November 2015, were obtained for *Arabidopsis thaliana*, *Caenorhabditis elegans*, *Drosophila melanogaster*, *Danio rerio*, *Escherichia coli*, *Gallus gallus*, *Homo sapiens*, *Mus musculus*, *Oryza sativa* and *Saccharomyces cerevisiae* [2]. The regex search used was similar to that described previously by Zeeberg and colleagues [1], with the added screen for dates in other formats (e.g. DD/MM/YY and MM-DD-YY). To expedite analysis of supplementary files from multi-disciplinary journals, we limited the articles screened to those that have the keyword 'genome' in the title or abstract (*Science*, *Nature* and *PLoS One*). Excel files (.xls and.xlsx) deposited in NCBI Gene Expression Omnibus (GEO) [3] were also

Table 1 Results of the systematic screen of supplementary Excel files for gene name conversion errors

Journal[a]	Number of Excel files screened	Number of gene lists found	Number of papers with gene lists	Number of supplementary files affected	Number of papers affected	Number of gene names converted
PLoS One	7783	2202	994	220	170	4240
BMC Genomics	11464	1650	801	218	158	4932
Genome Res	2607	580	251	114	68	3180
Nucleic Acids Res	2117	540	315	88	67	1661
Genome Biol	2678	664	257	97	63	1878
Genes Dev	932	395	190	75	55	1593
Hum Mol Genet	980	372	168	48	27	1724
Nature	482	150	74	27	23	1375
BMC Bioinformatics	1790	235	152	26	21	534
RNA	569	127	77	20	15	1341
Nat Genet	264	70	37	12	9	178
Bioinformatics	731	112	67	11	6	339
PLoS Comput Biol	177	79	32	6	6	46
PLoS Biol	143	54	29	7	5	206
Mol Biol Evol	995	112	79	7	4	56
Science	172	36	19	7	3	451
Genome Biol Evol	490	32	25	2	2	121
DNA Res	801	57	30	2	2	6
Total	35175	7467	3597	987	704	23861

[a]**The 18 journals investigated are ordered by the number of papers affected by gene name conversion errors**