Reproducible research with SAS

- SAS ODS, Macro and Graphics

Zhengming Chen

05/09/2017

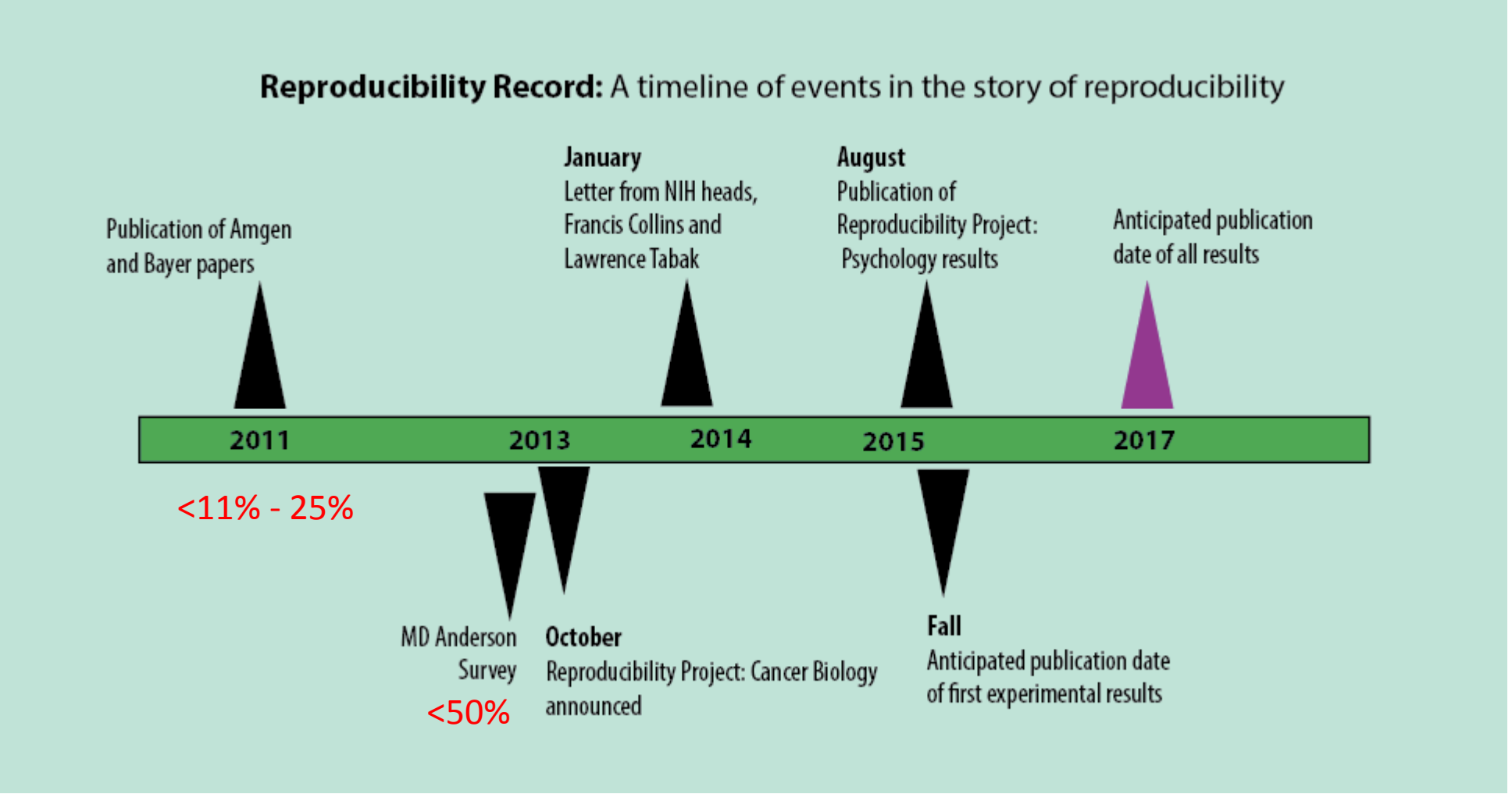
Outline

1. Reproducible research
2. How to run SAS on Mac computer?
3. How to generate high quality report and do it efficiently with SAS:

    Two SAS Macros with ODS:

    - **Macro 1**: Report dataset information and summary of all of its variables;
    - **Macro 2**: Correlation and figures;

1. Reproducible research

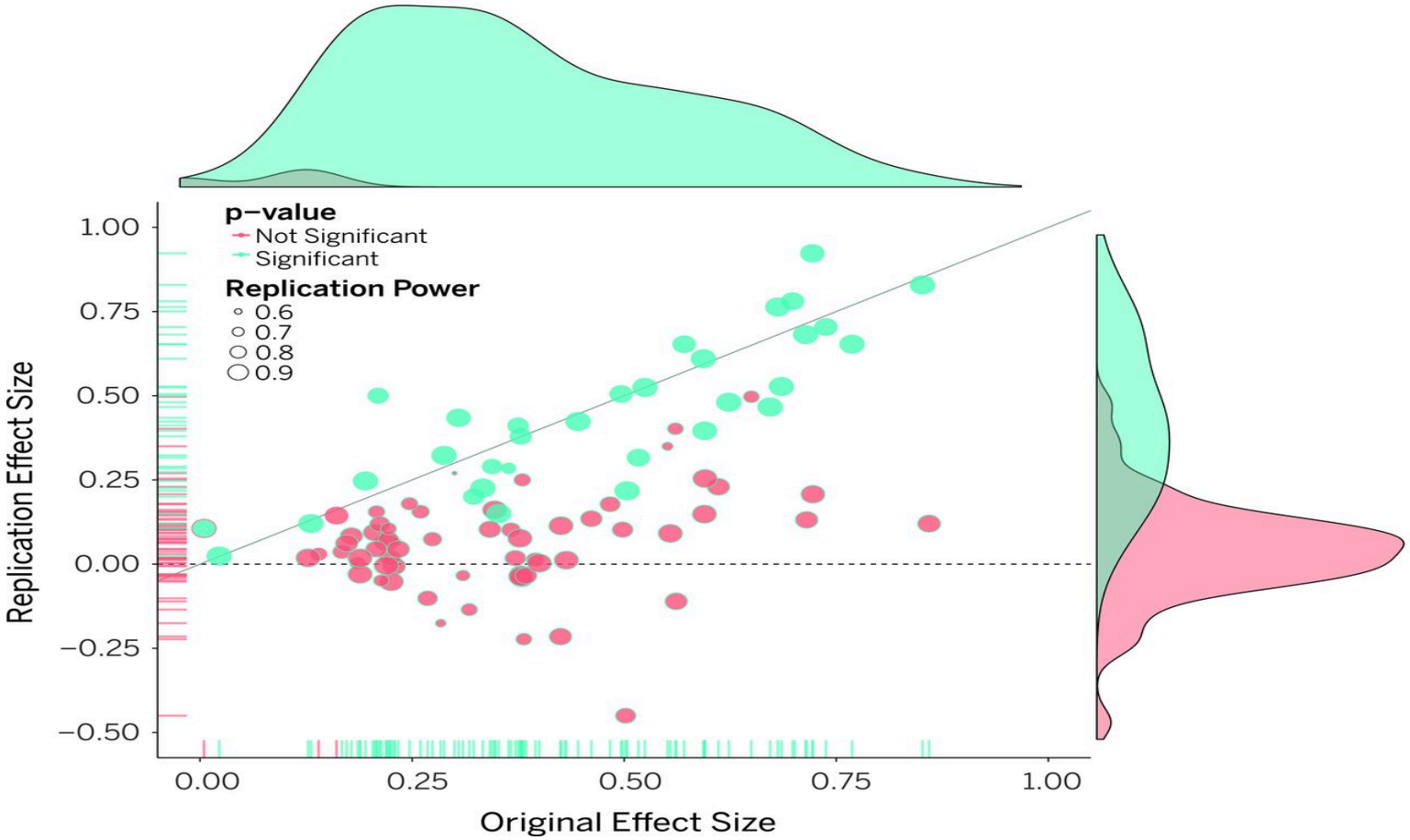


http://sitn.hms.harvard.edu/flash/2015/reproduce-or-bust-bringing-reproducibility-back-to-center-stage/

By Steph Guerra
Figure by Krissy Lyon

The Reproducibility Project: Psychology

- describe the replication of 100 experiments reported in papers published in 2008 in three high-ranking psychology journals



Significance rate:
36% vs 97%

Average effect size:
<50% of original results

Science
AAAS

Open Science Collaboration *Science* 2015;349:aac4716

The Reproducibility Project: Cancer Biology

- To scrutinize 50 cancer papers published in *Nature, Science, Cell* and other high-impact journals.

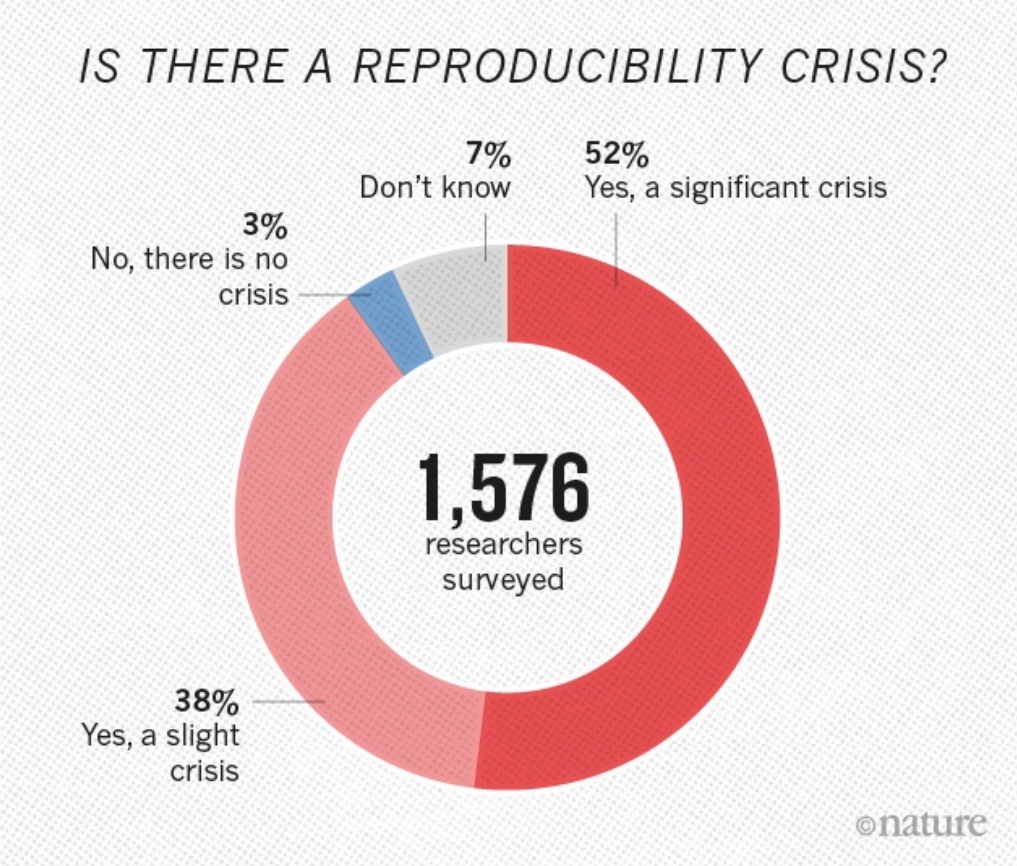
Early results:

"seven of the replication studies are now complete, and *eLife* is publishing five fully analysed efforts on 19 January 2017…

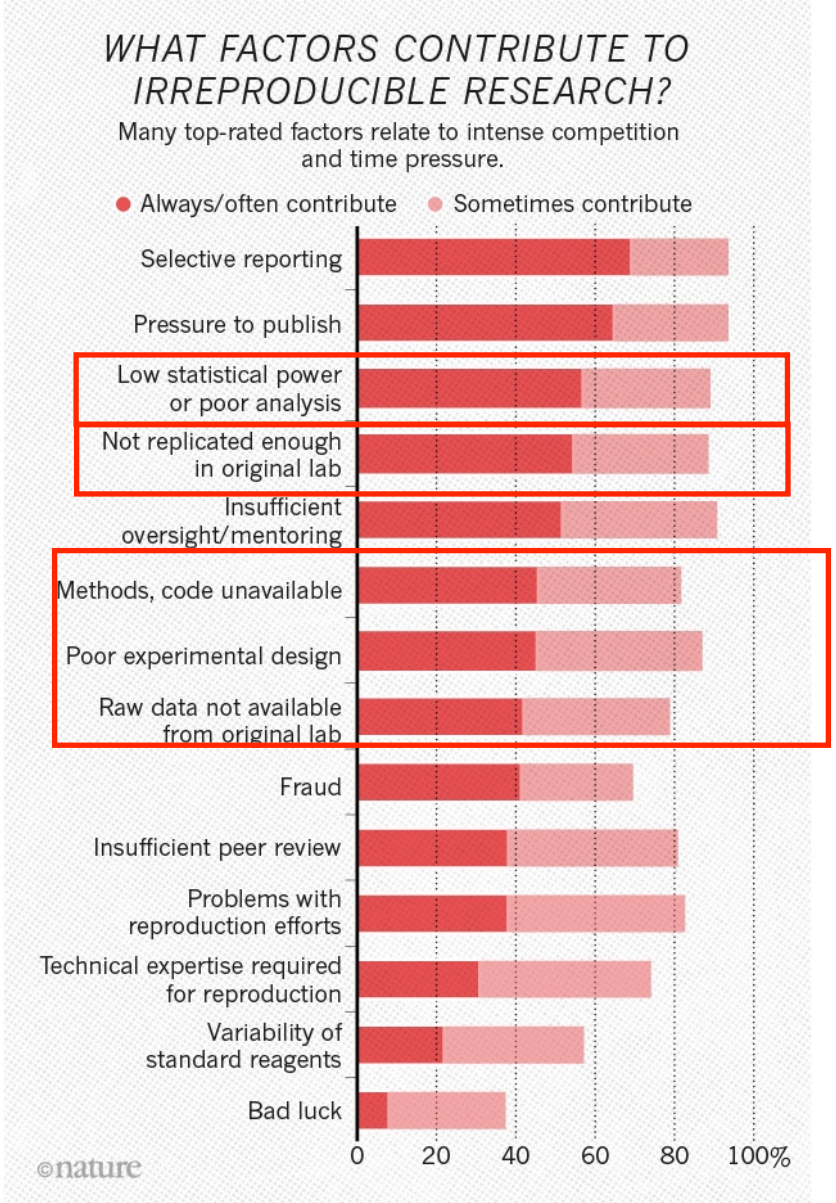
"the overall take-home message was that two studies generated findings similar to the original, one did not replicate the original, and two others were inconclusive."

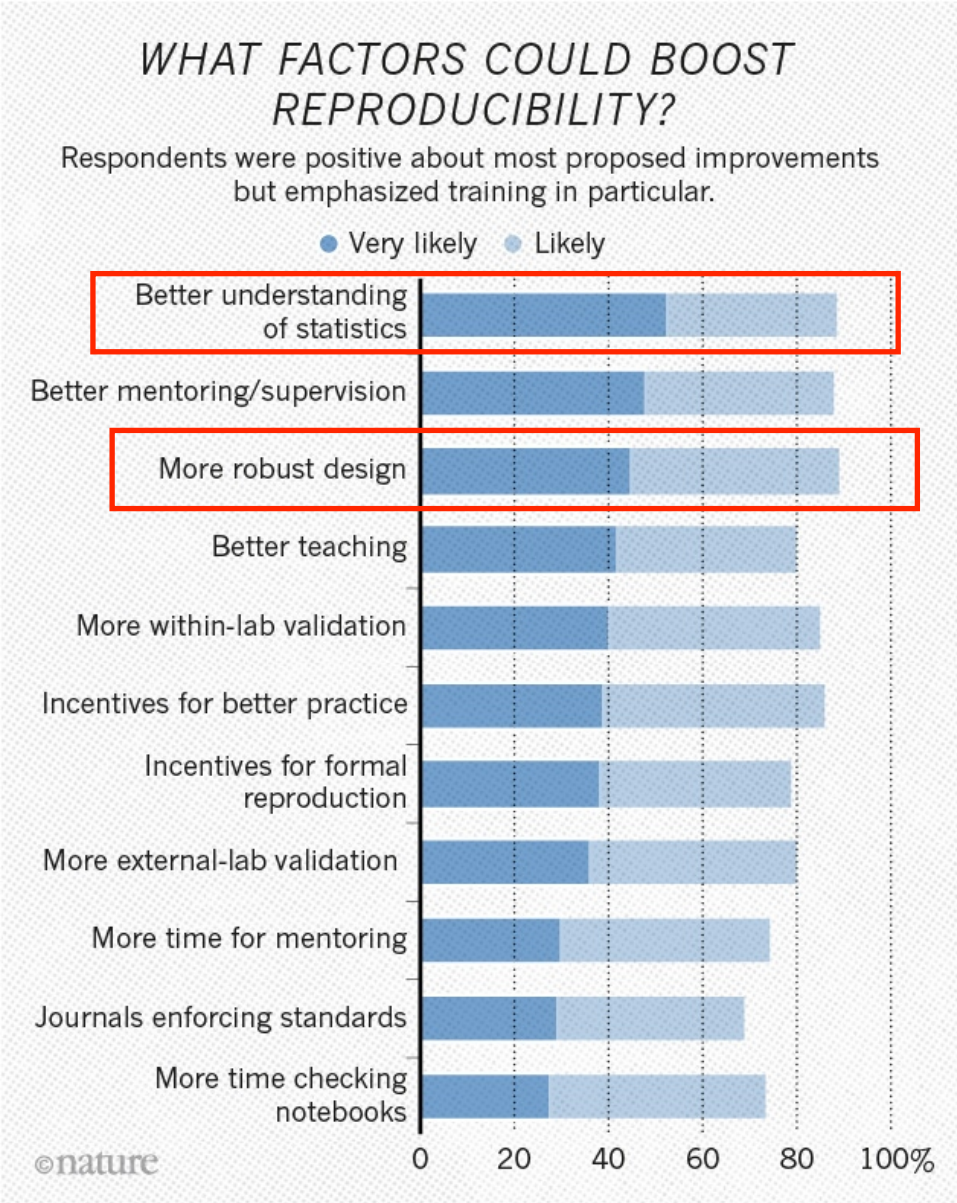
Nature | News
Cancer reproducibility project releases first results
An open-science effort to replicate dozens of cancer-biology studies is off to a confusing start.
Monya Baker & Elie Dolgin
18 January 2017

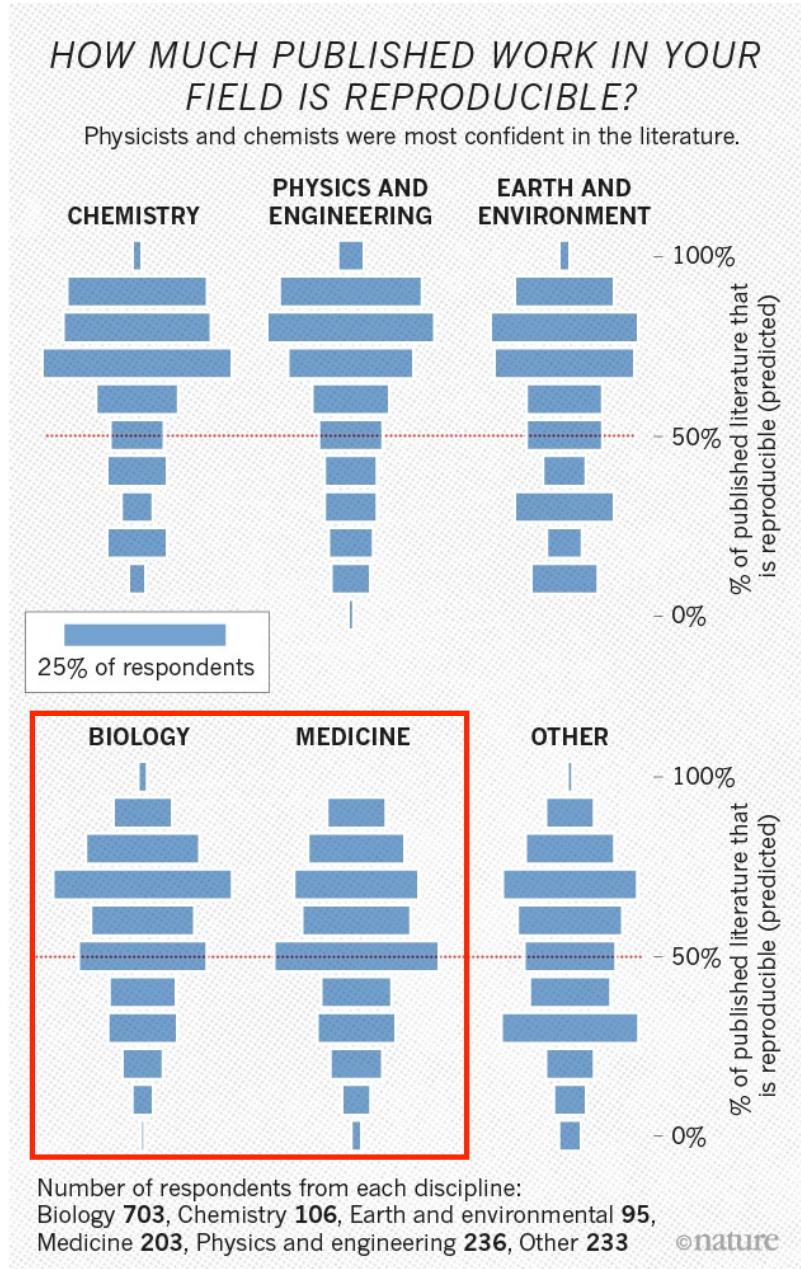
A survey conducted by *Nature* in 2016



1,500 scientists lift the lid on reproducibility
Survey sheds light on the 'crisis' rocking research.
Monya Baker[1]
Nature, 25 May 2016 Corrected: 28 July 2016


WHAT FACTORS CONTRIBUTE TO IRREPRODUCIBLE RESEARCH?
Many top-rated factors relate to intense competition and time pressure.
Always/often contribute
Sometimes contribute
Selective reporting
Pressure to publish
Low statistical power or poor analysis
Not replicated enough in original lab
Insufficient oversight/mentoring
Methods, code unavailable
Poor experimental design
Raw data not available from original lab
Fraud
Insufficient peer review
Problems with reproduction efforts
Technical expertise required for reproduction
Variability of standard reagents
Bad luck
0
20
40
60
80
100%
©nature


WHAT FACTORS COULD BOOST REPRODUCIBILITY?
Respondents were positive about most proposed improvements but emphasized training in particular.
Very likely
Likely
Better understanding of statistics
Better mentoring/supervision
More robust design
Better teaching
More within-lab validation
Incentives for better practice
Incentives for formal reproduction
More external-lab validation
More time for mentoring
Journals enforcing standards
More time checking notebooks
0
20
40
60
80
100%
©nature


HOW MUCH PUBLISHED WORK IN YOUR FIELD IS REPRODUCIBLE?
Physicists and chemists were most confident in the literature.
CHEMISTRY
PHYSICS AND ENGINEERING
EARTH AND ENVIRONMENT
100%
50%
0%
% of published literature that is reproducible (predicted)
25% of respondents
BIOLOGY
MEDICINE
OTHER
100%
50%
0%
% of published literature that is reproducible (predicted)
Number of respondents from each discipline:
Biology 703, Chemistry 106, Earth and environmental 95,
Medicine 203, Physics and engineering 236, Other 233
©nature

Best Jobs in America

Fastest-growing jobs

1. Biostatistician

Demand for biostatisticians are projected to grow a solid 34% over 10 years.
- 01/05/2017, CNN Money

Correlation/Association?

http://money.cnn.com/gallery/pf/2017/01/05/fastest-growing-jobs-2017/

2. How to run SAS on Mac computer

Option 1: Install dual operation systems

- **Use Virtual Machine to install Win system (Win 7, 8 or 10 professional);**
- **Install Windows SAS under the Win system;**

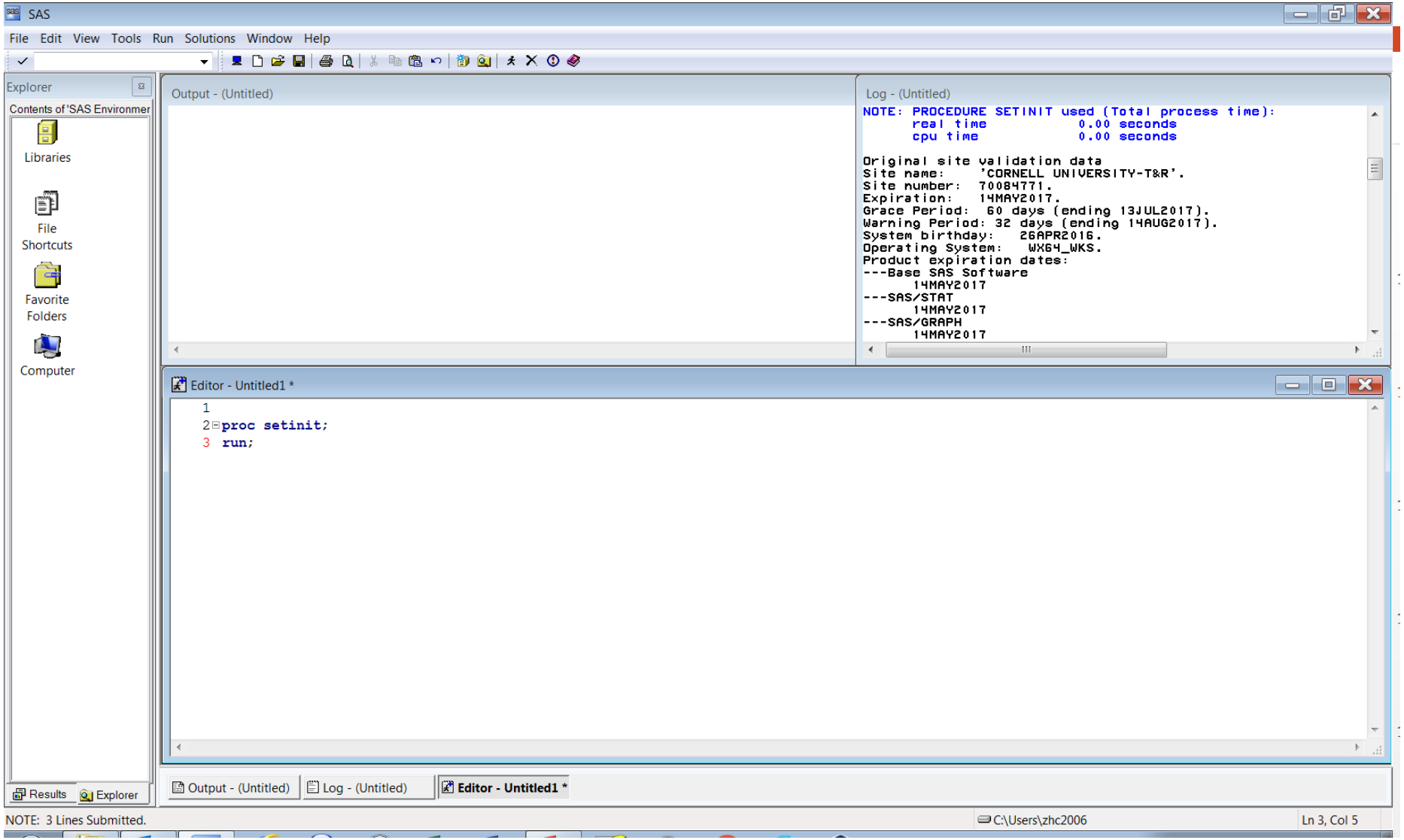
Pros:

- Can have full version with full capacity;
- No need internet to run;

Cons:

- Need to buy Win system, and pay SAS license fee.
- Each time, needs to run Virtual Machine, and then Windows;
- SAS full version alone takes hours to install (>35 GB in size for full version);
- If need to upgrade, need install again;

Classic SAS Windowing interface



Option 2: Install SAS University Edition

- **Use Virtual Machine to Install and run SAS University Edition**

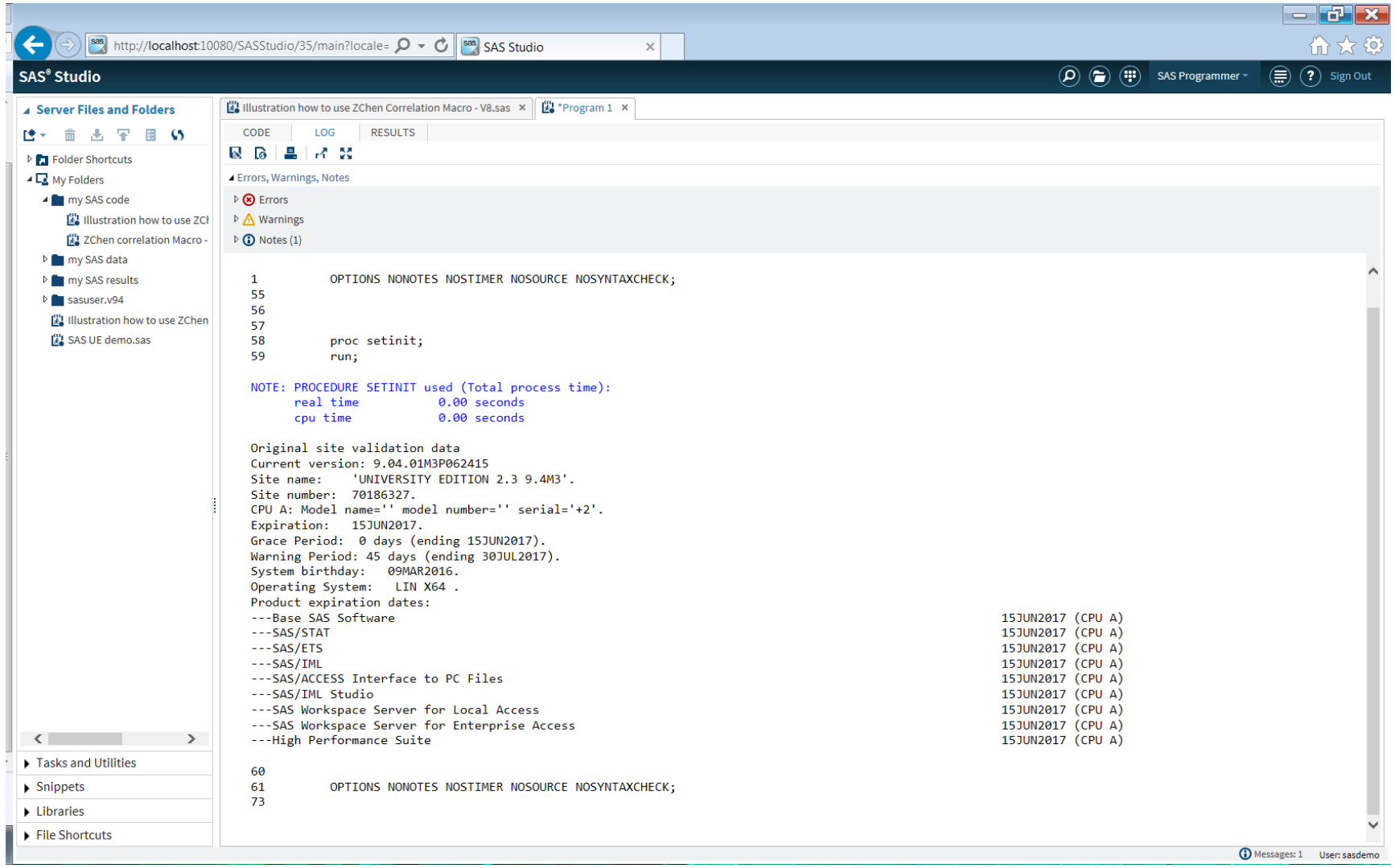
Pros:

- Free for academics;
- No need internet to use; (but you can run on Amazon Web Services (AWS) Marketplace);
- Smaller and easy to install;
- Runs with SAS studio interface in a browser so it can run on any type of computers, including Mac;

Cons:

- Needs to run Virtual Machine;
- Not full version (11GB in size, 8 modules);
- Can be slow;

SAS studio interface for SAS University Edition



https://www.sas.com/en_us/software/university-edition.html?keyword=sas%252520university%252520edition&matchtype=p&publisher=google&gclid=CI_5jN-wz9MCFYtXDQodlSoPiQ

Option 3: Use SAS OnDemond for Academics

- **Run SAS full version in cloud;**

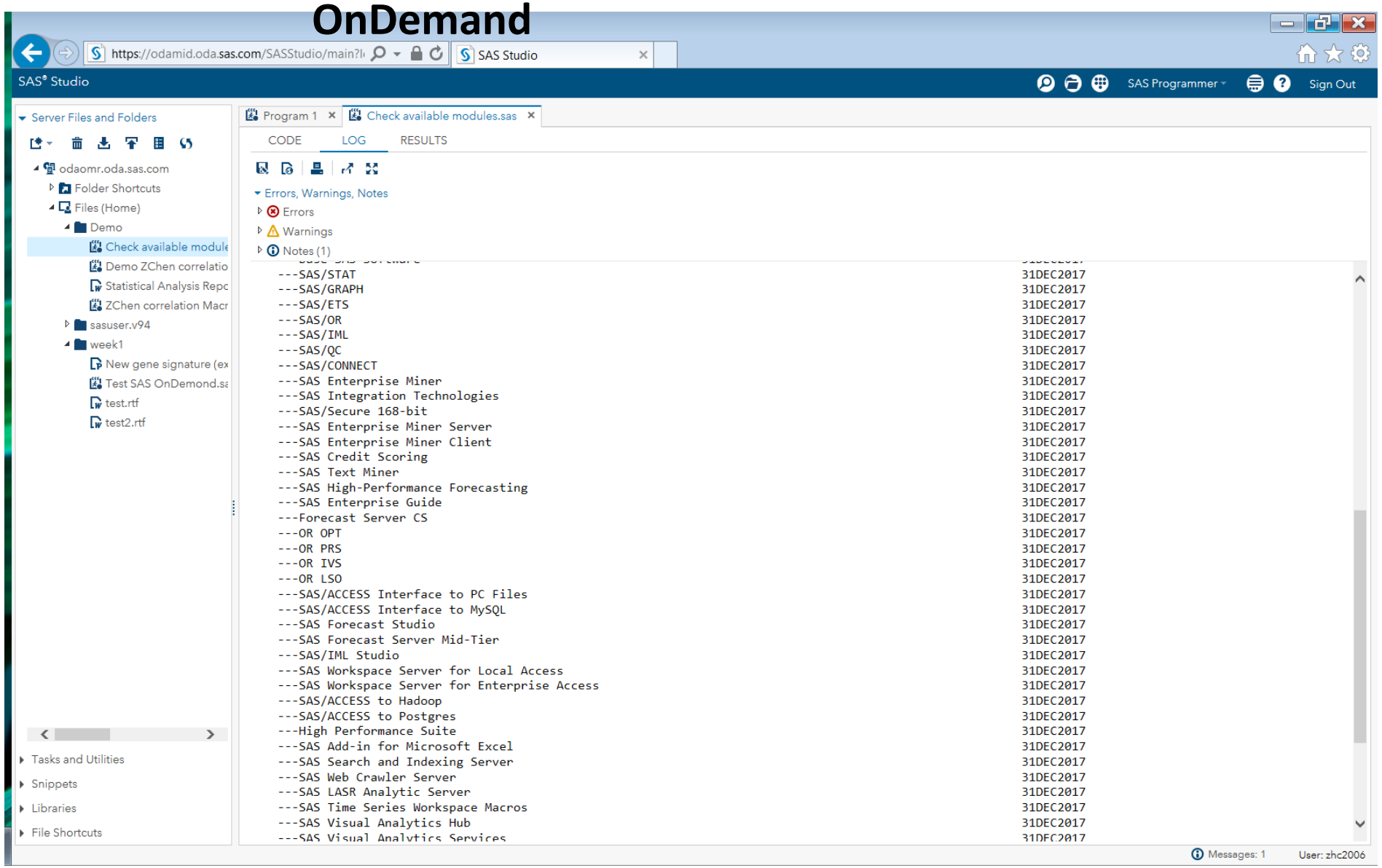
Pros:

- In cloud, no need to install;
- Most up-to-date full version;
- Free (free to use, free 5GB online space);
- Runs on SAS studio interface in browser;
- Runs from any device with internet accessibility;

Cons:

- Need internet to use;
- Needs to login and out;
- Need upload and download files;
- Sometimes ‘scheduled maintenance’ and you cannot use;

SAS studio interface for SAS OnDemand



https://www.sas.com/en_us/software/on-demand-for-academics.html

3. Two SAS Macro - demonstration

- **Macro 1**: Report dataset information and summary of all of its variables;

  %***DataInfo*** (Dataset=sashelp.class);

Features:

- Only one parameter – SAS dataset name;
- Run one line of SAS code, you get these information about any SAS dataset in one file:

  - Attributes of the dataset;
  - Descriptive statistics of all numerical variables;
  - Descriptive statistics of all categorical variables;

- **Macro 2**: Correlation and figures;

%***CorrTest*** (dataset=sashelp.class, V1=Age, V2=Weight, type="spearman“, H0=0.5);

Features:

- More parameters, with default values;
- Run one line of SAS code, you get these information about the correlation between pairs of numerical variables in one file:

  - Descriptive statistical of each variables;
  - Correlation (different types) coefficient, with 95% CI;
  - P values testing different null hypothesis;
  - Publishable correlation plots;

Thank you!