Data Science

Vladimir Svetnik
Round table discussion
at NISS-SAMSI Affiliates Meeting
Sunday, March 6
Austin, TX
vladimir_svetnik@merck.com

Outline

- John W. Tukey 100th Birthday Celebration
- D.Donoho's views on Data Science
- Merck&Co and Data Science
  - John W. Tukey
  - Leo Breiman
  - Jerome H. Friedman

D.Donoho's views on Data Science (cont.)

  - Six Divisions of Data Science
  - Common Task Framework (CTF)
- Merck&Co. Data Science Examples
  - Deep Learning Networks in QSAR
  - Flow Cytometry Data Analysis

tukey-100th-hi-res.jpeg

genomics.princeton.edu/storeylab/tukey/tukey-100th-hi-res.jpeg

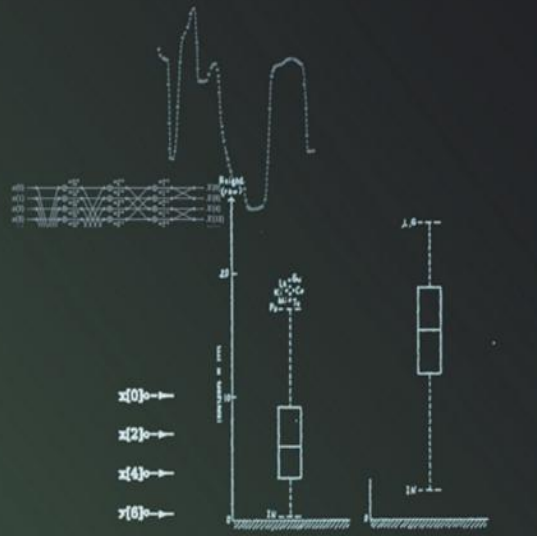
Apps www.music.mahidol....

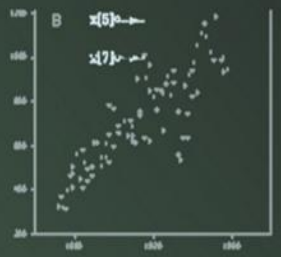
John W. Tukey 100th Birthday Celebration at Princeton University

Friday, September 18, 2015 | 9:00am-6:00pm
McDonnell Hall A02, Princeton University
Full details available at:
csml.princeton.edu/tukey

Speakers:

Yoav Benjamini
Persi Diaconis
David Donoho
Jianqing Fan
Luisa Fernholz
Jerome Friedman
Rafael Irizarry
Karen Kafadar
Stephan Morganthaler
Scott Zeger





CENTER FOR STATISTICS AND MACHINE LEARNING

PRINCETON UNIVERSITY

David Donoho: 50 years of Data Science

https://dl.dropboxusercontent.com/u/23421017/50YearsDataScience.pdf

"Ten to twenty years ago, JohnChambers, Bill Cleveland and Leo Breiman independently once again urged academic statistics to expand its boundaries beyond the classical domain of theoretical statistics; Chambers called for more emphasis on data preparation and presentation rather than statistical modeling; and Breiman called for emphasis on prediction rather than inference. Cleveland even suggested the catchy name "Data Science" for his envisioned field"

"A recent and growing phenomenon is the emergence of "Data Science" programs at major universities, including UC Berkeley, NYU, MIT, and most recently the Univ. of Michigan, which on September 8, 2015 announced a $100M "Data Science Initiative" that will hire 35 new faculty. Teaching in these new programs has significant overlap in curricular subject matter with traditional statistics courses; in general, though, the new initiatives steer away from close involvement with academic statistics departments"

"Drawing on work by Tukey, Cleveland, Chambers and Breiman, I present a vision of data science based on the activities of people who are 'learning from data', and I describe an academic field dedicated to improving that activity in an evidence-based manner. This new field is a better academic enlargement of statistics and machine learning than today's Data Science Initiatives, while being able to accommodate the same short-term goals."

Merck&Co. and Data Science

- John W Tukey
- Leo Breiman
- Jerome H Friedman

The Six Divisions of Data Science

Data Exploration and Preparation

Sanity check of data basic properties, anomalies and artifact remediation, and also grouping, smoothing subsetting , etc.

Data Representation and Transformation

Combining data from different databases with different formats and mathematically represent special type of data using Fourier/Wavelet and other transformations

Computing with Data

R (tidyr, dplyr, ggplot2), Python (Pandas,Scikit-learn, NumPy, matplotlib); Java, C++, Git/GitHub, Hadoop and MapReduce for parallel computing; HPC clusters

The Six Divisions of Data Science (cont.)

Data Modeling

Generative modeling: "This roughly speaking coincides with traditional Academic statistics"

Predictive modeling: "…in which one constructs methods which predict well over a given data universe. This roughly coincides with modern Machine Learning"

Data Visualization and Presentation

Science about Data Science

Uncovering emergent phenomena in data analysis, for example new patterns arising in data analysis workflows, or disturbing artifacts in published analysis results.

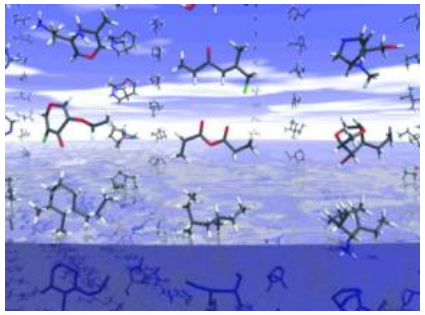
Foundational work to make future such science possible – such as encoding documentation of individual analyses and conclusions in a standard digital format for future harvesting and meta analysis.

The Common Task Framework (CTF)

Predictive Modeling culture together with CTF is the 'secret sauce' of machine learning

a) A publicly available training dataset involving, for each observation, a list of (possibly many) feature measurements, and a class label for that observation.
b) A set of enrolled competitors whose common task is to infer a class prediction rule from the training data.
c) A scoring referee, to which competitors can submit their prediction rule. The referee runs the prediction rule against a testing dataset which is sequestered behind a Chinese wall. The referee objectively and automatically reports the score (prediction accuracy) achieved by the submitted rule

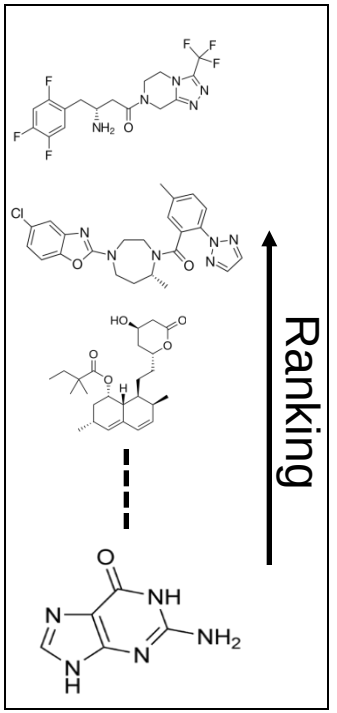
QSAR: Quantitative Structure & Activity Relationship



- Potency
- Some AEs & drug ADME properties

Computer Predictive QSAR Models

- Potency
- **~33 AEs & drug ADME properties**



Correlation(Lab, Computer) : 0.30 — 0.91

12 Years of Domination in QSAR Area

Pre-Random-Forest Era

Many different methods, e.g. PLS

2003

Svetnik, V.; Liaw, A.; Tong, C.; Culberson, J.C.; Sheridan, R.P.; Feuston, B.P. *Random Forest: a classification and regression tool for compound classification and QSAR modeling*. *J. Chem. Inf. Comput. Sci.* **2003**

Random-Forest Era

Random Forest (RF)

Deep-Learning : Disruptive Technology

- Revolutionizing many research areas (e.g. computer vision, speech recognition)
- Pursued by almost all innovation-drive companies (e.g. Google, IBM, Microsoft, Facebook, etc.)
- Attracted our attention by winning the 1st prize in Merck's QSAR competition, and beating RF by ~16%.

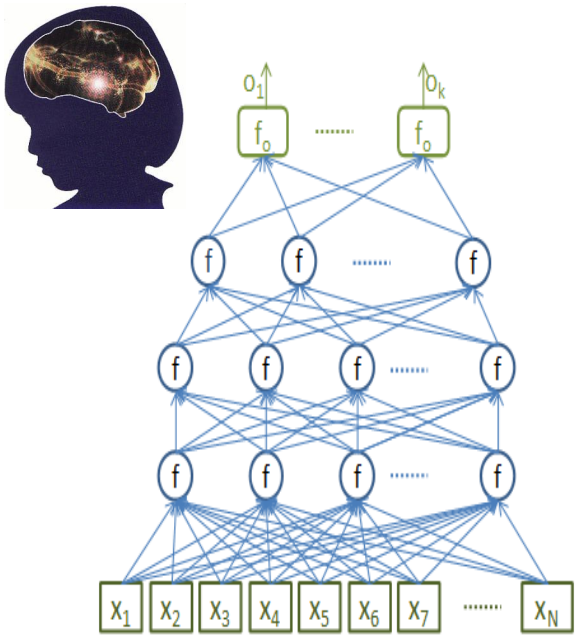
The New York Times

Scientists See Promise in Deep-Learning Programs

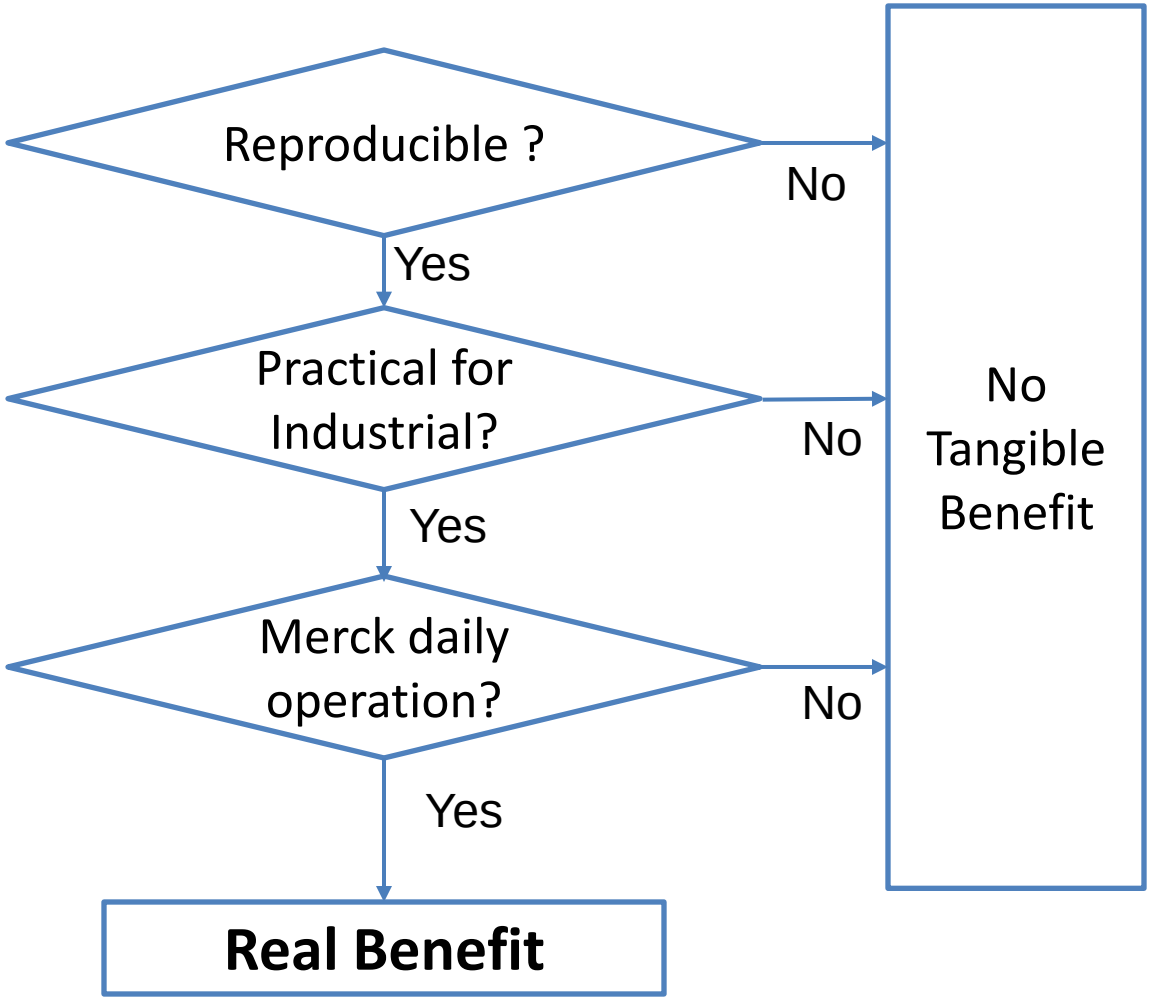
By JOHN MARKOFF

Published: November 23, 2012

Is Deep-Learning Truly Beneficial to Drug Discovery?

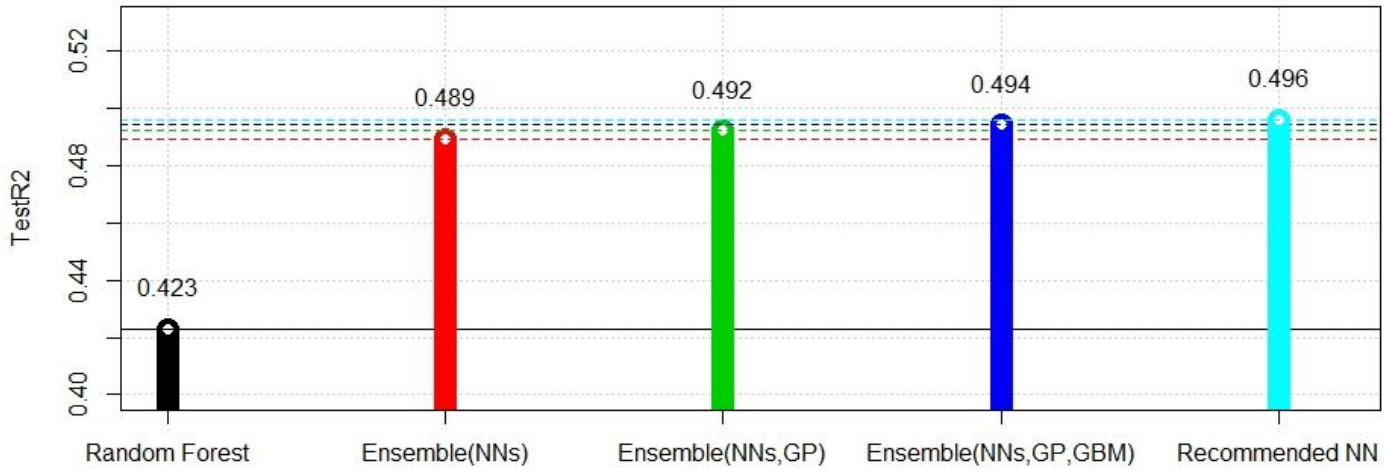


Dataset:
a 165,000 x 10,000 matrix
Model:
a network of 51,008,000 parameters
Usage:
Many options



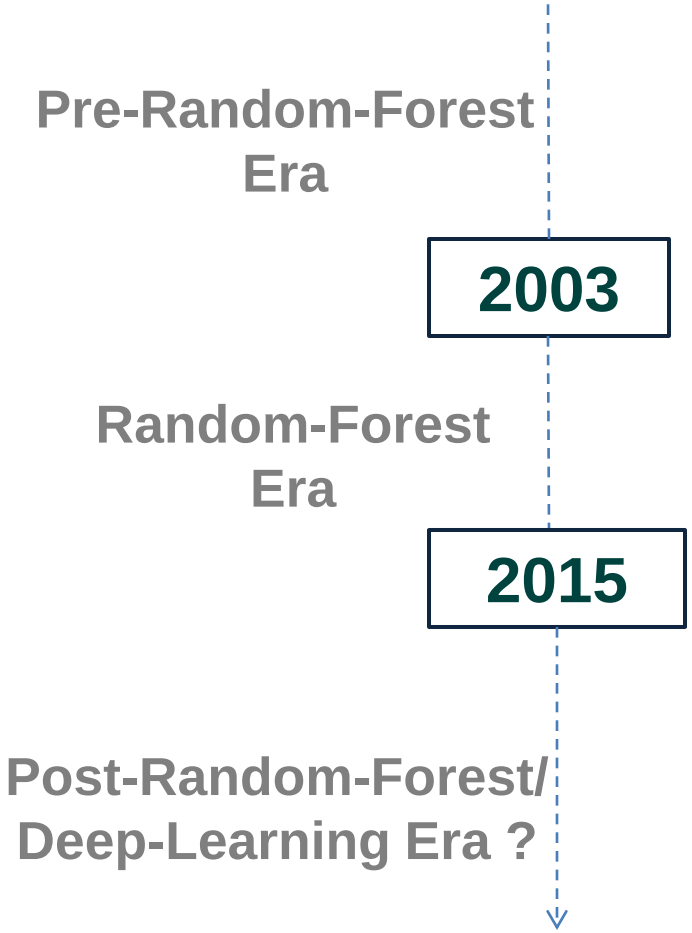
Bringing Deep-Learning

- Created state-of-the-art parallel computing infrastructure
- Identified key factors for deep-learning's success in QSAR
- Developed and deployed software Merck QSAR operations



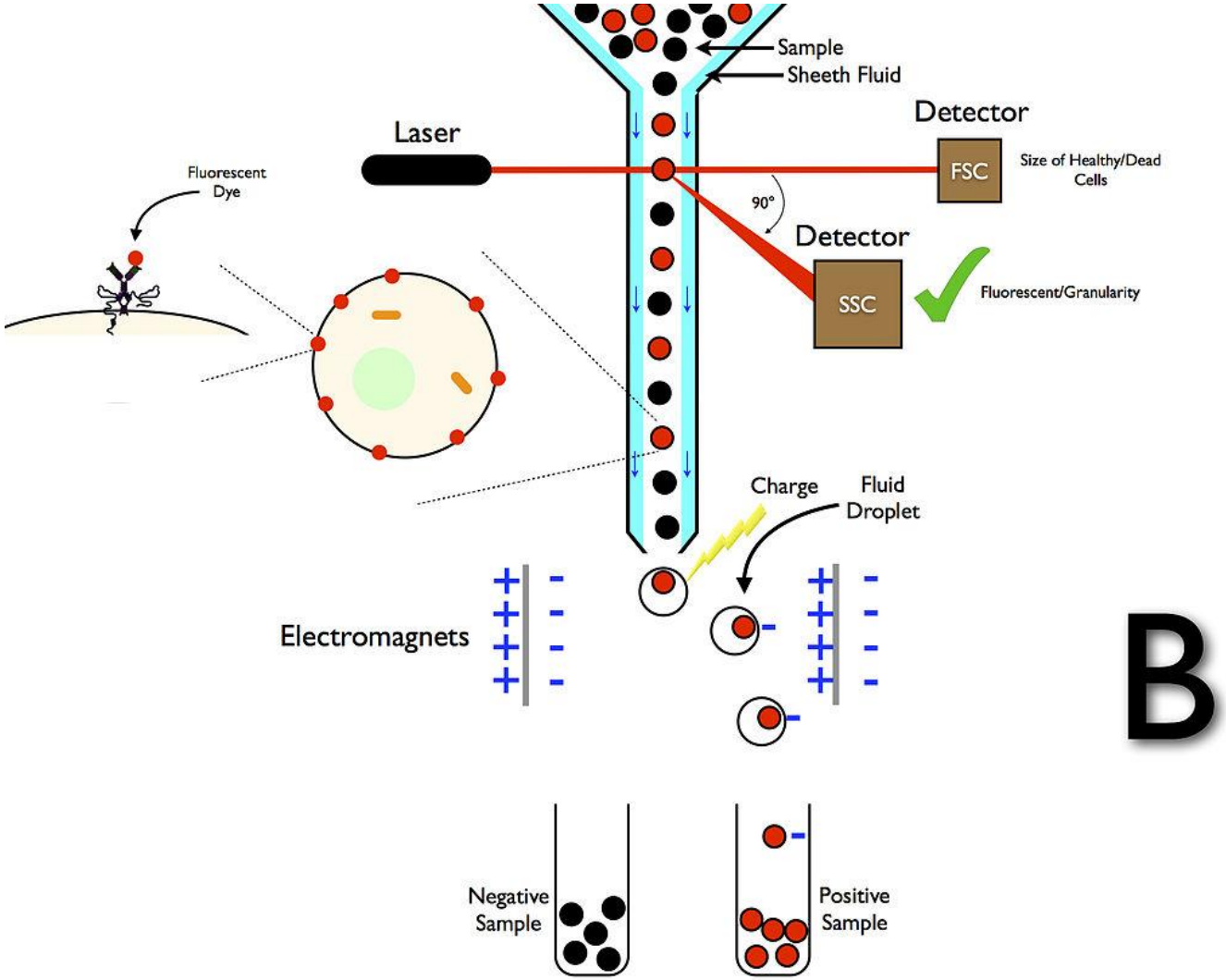
17.1% improvement over RF

Deep-learning era?



Ma, J.; Sheridan R.P.; Liaw, A.; Dahl, G.E.; Svetnik, V. Deep Neural Nets as a Method for Quantitative Structure-Activity Relationships. *J. Chem. Inf. Comput. Sci., 2015*.

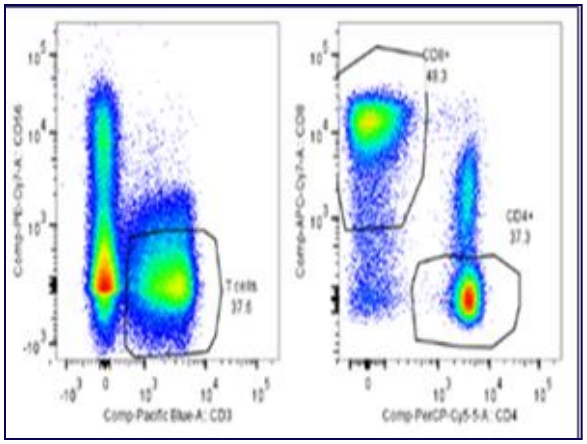
Flow Cytometry



https://commons.wikimedia.org/w/index.php?curid=18139883

Cytometry

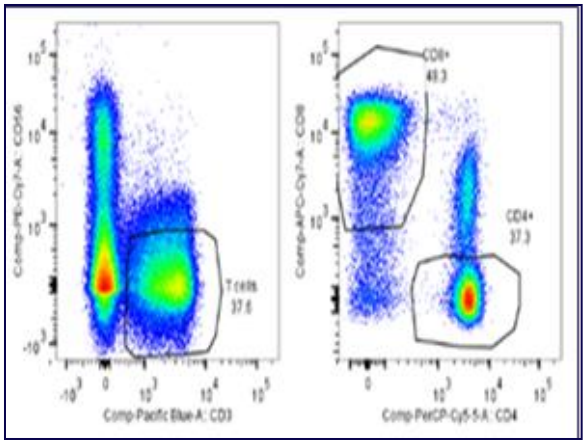
- Critical Importance
  - Ubiquitously used in the industry



- Unmet needs in cytometry data analysis
  - Feedback from scientists: “*time-consuming*”, “*bottleneck*”
  - Our observations:
    - Too much subjective judgment (frequently un-blinded)
    - Different approaches/platforms used by different teams
    - Reproducibility and transferability issues

Cytometry

- Critical Importance
  - Ubiquitously used in the industry



- Unmet needs in cytometry data analysis
  - Feedback from scientists: “*time-consuming*”, “*bottleneck*”
  - Our observations:
    - Too much subjective judgment (frequently un-blinded)
    - Different approaches/platforms used by different teams
    - Reproducibility and transferability issues

Sources of Variability

Sample Collection

Sample Preparation

Data Acquisition

Data Analysis

Biological Variability

Procedure and Assay Variability

Equipment-related Variability

Analyst-induced Variability

Quality-Control (QC) Feedbacks

Minimizing

FlowCAP - Flow Cytometry: Critical Assessment of Population Identification Methods, 2010-2014

- “In response to this need, we are pleased to announce the **FlowCAP**project.
- The goal of FlowCAP is to advance the development of computational methods for the identification of cell populations of interest in flow cytometry data. FlowCAP will provide the means to objectively test these methods, first by comparison to manual analysis by experts using common datasets, and second by prediction of a clinical/biological outcome. ”

Solution: **CASK-Cyto**

Customized **A**dvanced but **S**imple **K**its for **Cyto**metry

- A high-throughput and objective cytometry data analysis **platform**
- 2500+ lines of **R codes** built using Bioconductor
- Statistical **algorithms** specially developed for this platform to ensure high-throughput

"CASK-Cyto to Accelerate Multiplex Immuno-Cytometry Assay Development" Shubing Wang: Poster session 8-10pm on Sunday (March 6th).