Hillman Cancer Center

Pitt Public Health

Global research productivity in the era of big data: Linking Statistics and Epidemiology

Faina Linkov, PhD, MPH
Associate Professor
Magee Womens Research Institute
University of Pittsburgh School of Medicine

February, 2018

Magee-Womens Research Institute

MAGEE-WOMENS RESEARCH INSTITUTE

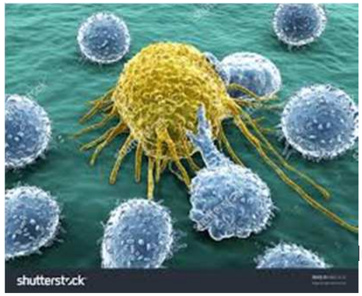
Faina Linkov

Associate Professor, Department of Obstetrics, Gynecology, and Reproductive Sciences and Epidemiology

UNIVERSITY OF PITTSBURGH

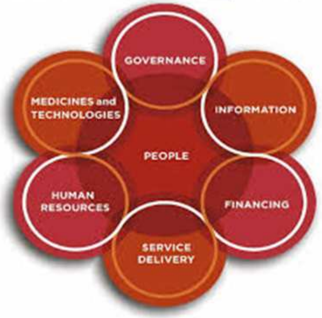
~100 publications, H-Factor 19 (Google Scholar)

Areas of Research



u/

1. Molecular / patho epidemiology, Cancer Prevention



2. Health Systems Research

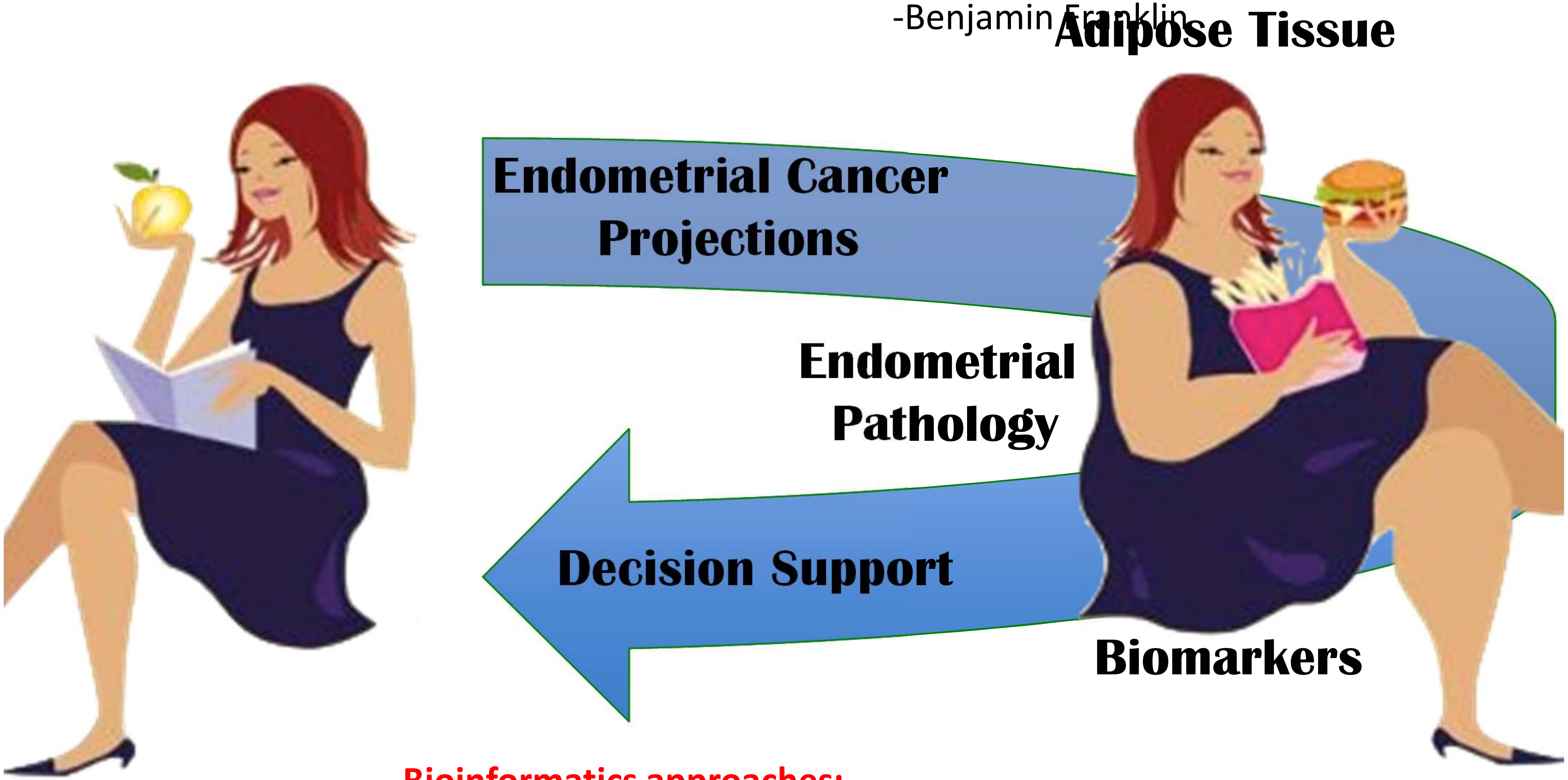
http://cajgh.pitt.edu

3. Global health and research productivity

Unifying theme: Data on Prevention?

“An ounce of prevention is worth a pound of cure”

-Benjamin Franklin



Bioinformatics approaches:
Improved data capturing?
Building models?

Key challenges in statistics today

1. Explosion of big data and data sources
2. Analysis and interpretation of existing data: integration of multiple data types
3. Statistical knowledge is not equally distributed
4. Research productivity is statistics depended

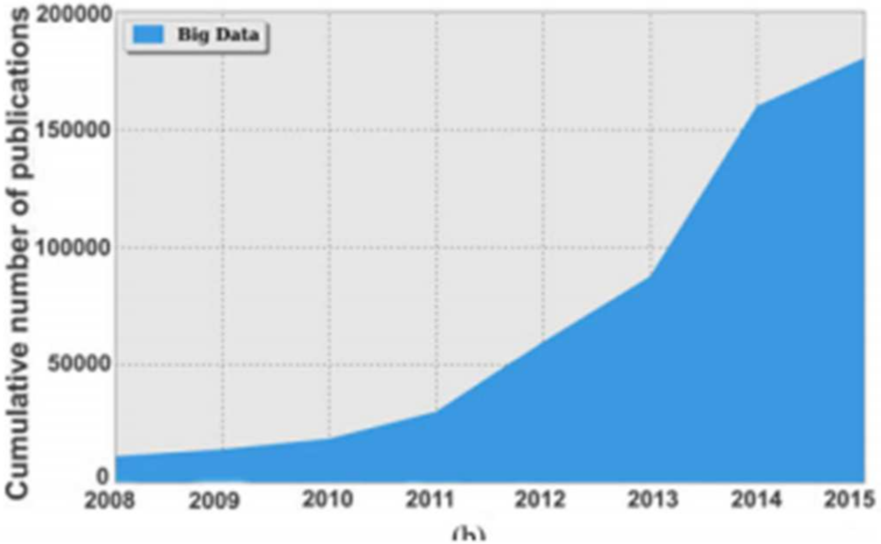
Solutions?

Statistics education

Improvement of research Productivity: Central Asian Journal of Global Health

Integration of statistical knowledge into current public health challenges: obesity?

Challenge 1
Explosion of big data



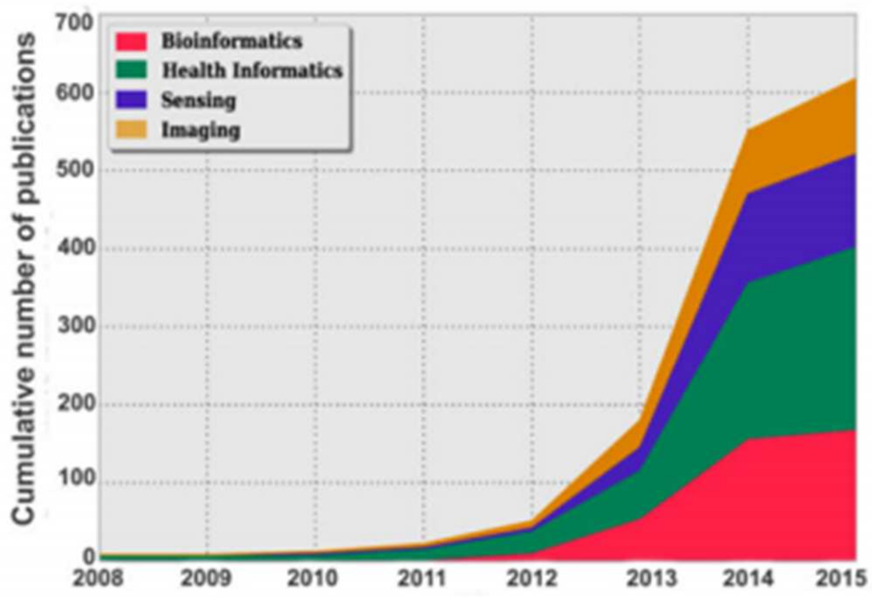
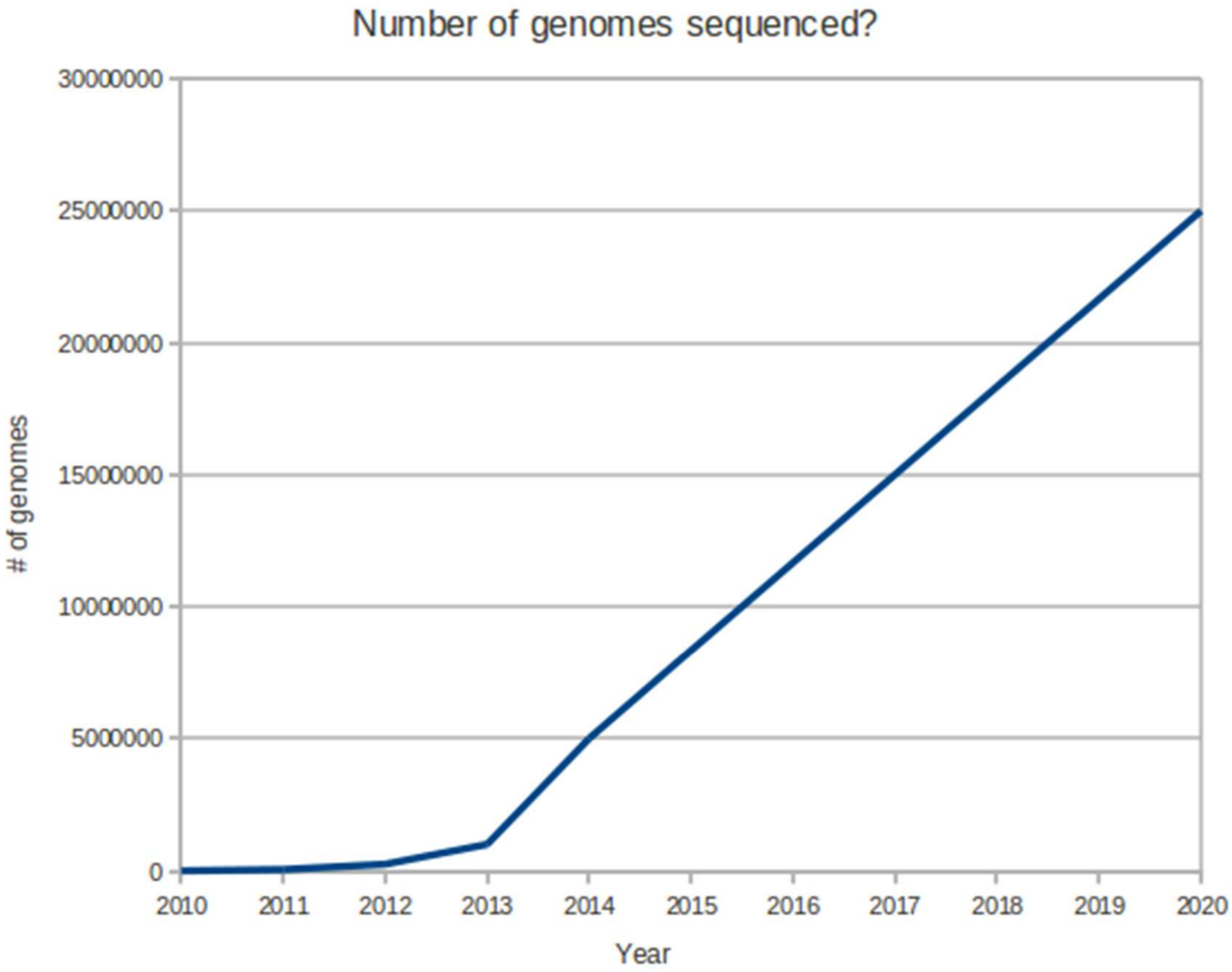
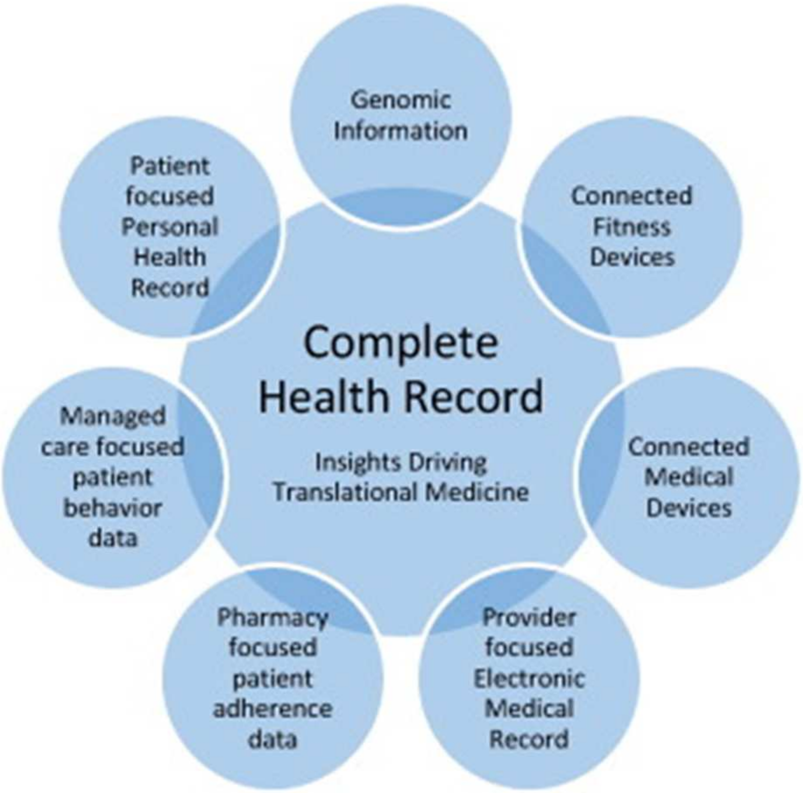


Fig. 1. (a) Cumulative number of publications referring to "big data" indexed by Google Scholar. (b) Cumulative number of publications per health research area referring to "big data," as indexed in IEEE Xplore, ACM Digital library, PubMed (National Library of Medicine, Bethesda, MD), Web of Science, and Scopus.


Number of genomes sequenced?
# of genomes
30000000
25000000
20000000
15000000
10000000
5000000
0
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
Year

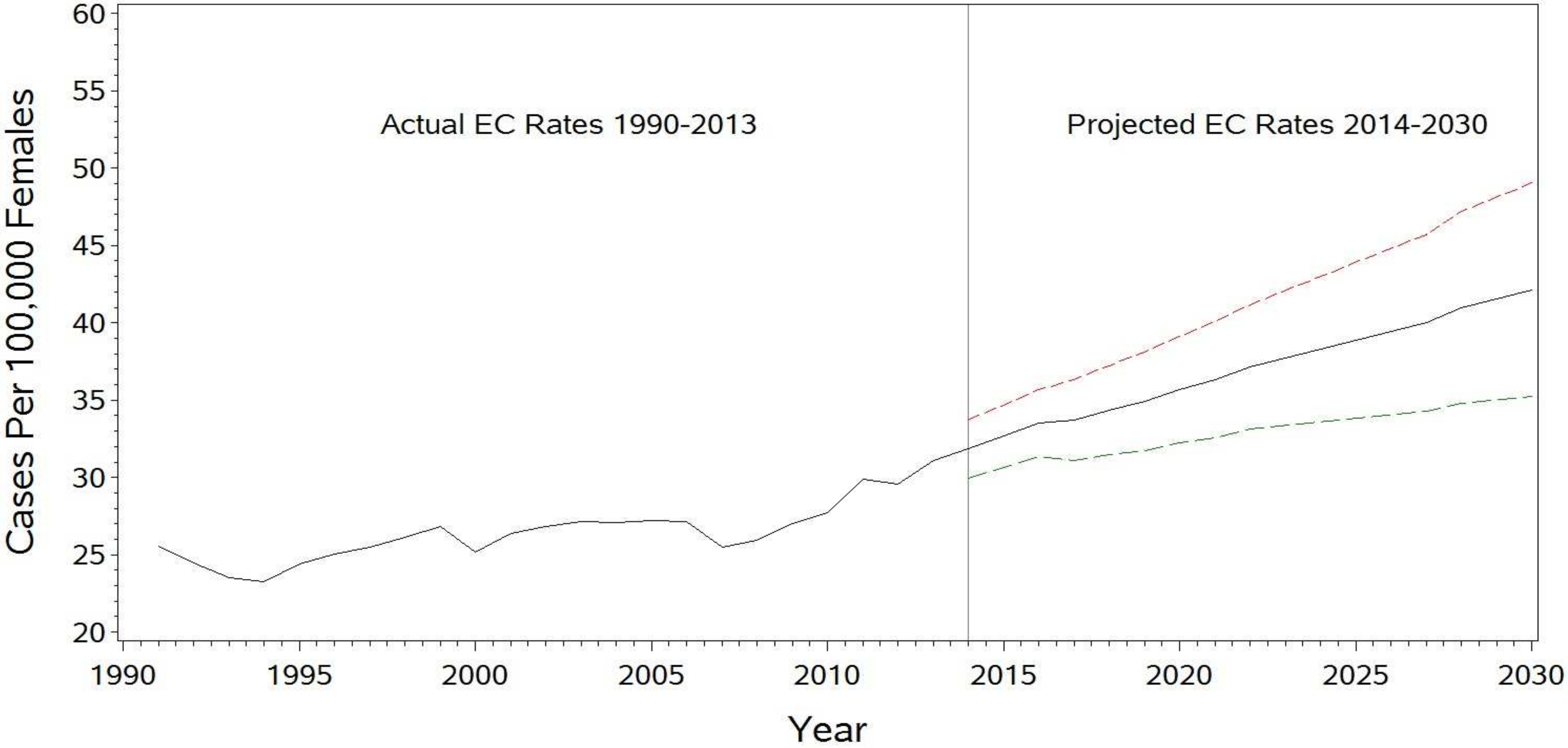
Health Records Data (EMR)-Mostly incomplete


Genomic Information
Connected Fitness Devices
Connected Medical Devices
Provider focused Electronic Medical Record
Pharmacy focused patient adherence data
Managed care focused patient behavior data
Patient focused Personal Health Record
Complete Health Record
Insights Driving Translational Medicine

Many essential items are missing in the current systems!!!

The Idea of Complete Health Record - Includes all the health information about a patient, not just what is in the EMR.

Analysis of future EC trends: Harness publically available data instead of collecting new data?



The best-fitting model based on multivariate regression projected an increase to 42.13 EC cases per 100,000 by the year 2030, a 50% increase over 2010 EC rates (Future oncology, 2014)

Challenge 2: Integration of multiple types of data into one study

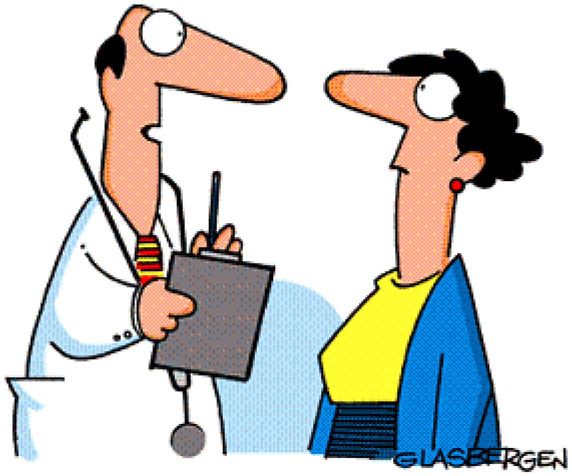
What strategy do you use to combine genetics data, protein data, health history data, psychological data, tissue data, and protein marker data?

Development of Bariatric Surgery Cohort:

Building bridge between obesity, inflammation, biobehavioral factors, and cancer and ultimately design interventions targeting multiple mechanisms that cause malignancies

Endometrial cancer risk reduction in the context of weight loss through bariatric surgery

FunnyandJokes.com

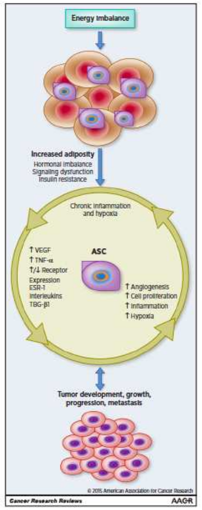


"You'll lose weight on any strict diet, but it's mostly water...from crying."

Information collected:

- Blood (protein and genetics markers)
- Urine
- Health forms (including General Health form, reproductive health form, SF-36 (quality of life), CES-D (depression), MAQ (physical activity), Sleep scale)
- Endometrial tissue

Adipose derived stem cells



Genetic

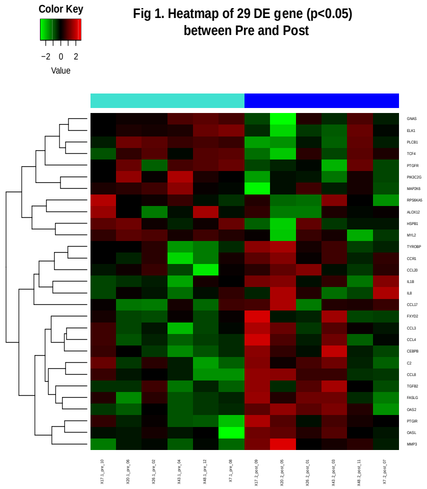
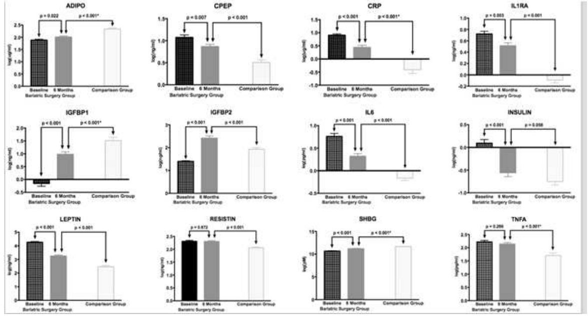


Fig 1. Heatmap of 29 DE gene (p<0.05) between Pre and Post

Protein



Lessons learned and future directions?

Problem:

New technologies have enabled new scientific questions to be addressed and old questions addressed in new ways.

There is problem of integration of multiple data types to get actionable knowledge

Solution?

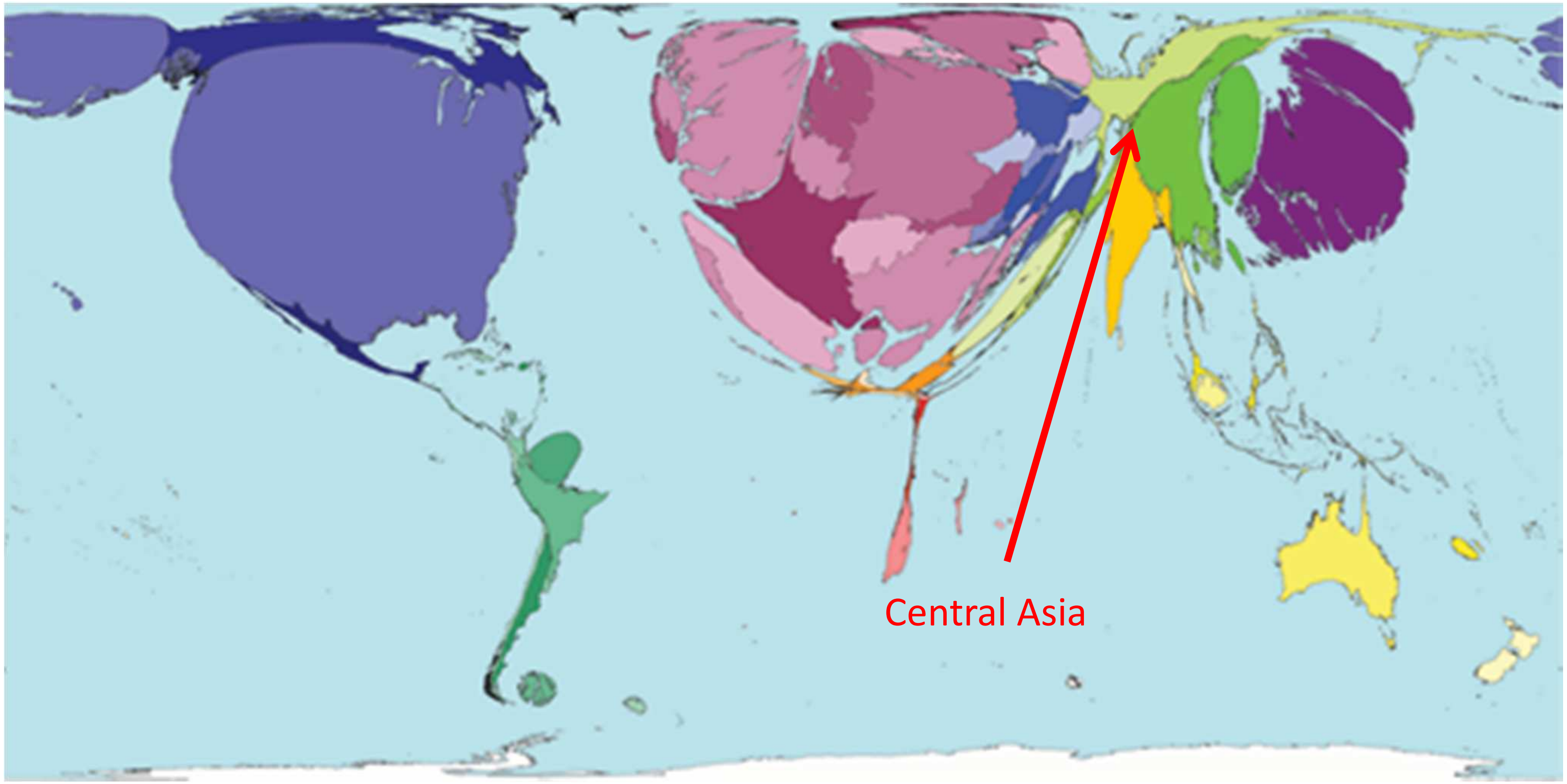
Cross disciplinary collaboration to come up with robust statistical and informatics approaches for combining data.

Develop innovative bioinformatics systems which permit biological scientists to directly perform detailed analyses of high dimensional data without the knowledge of computer programming?

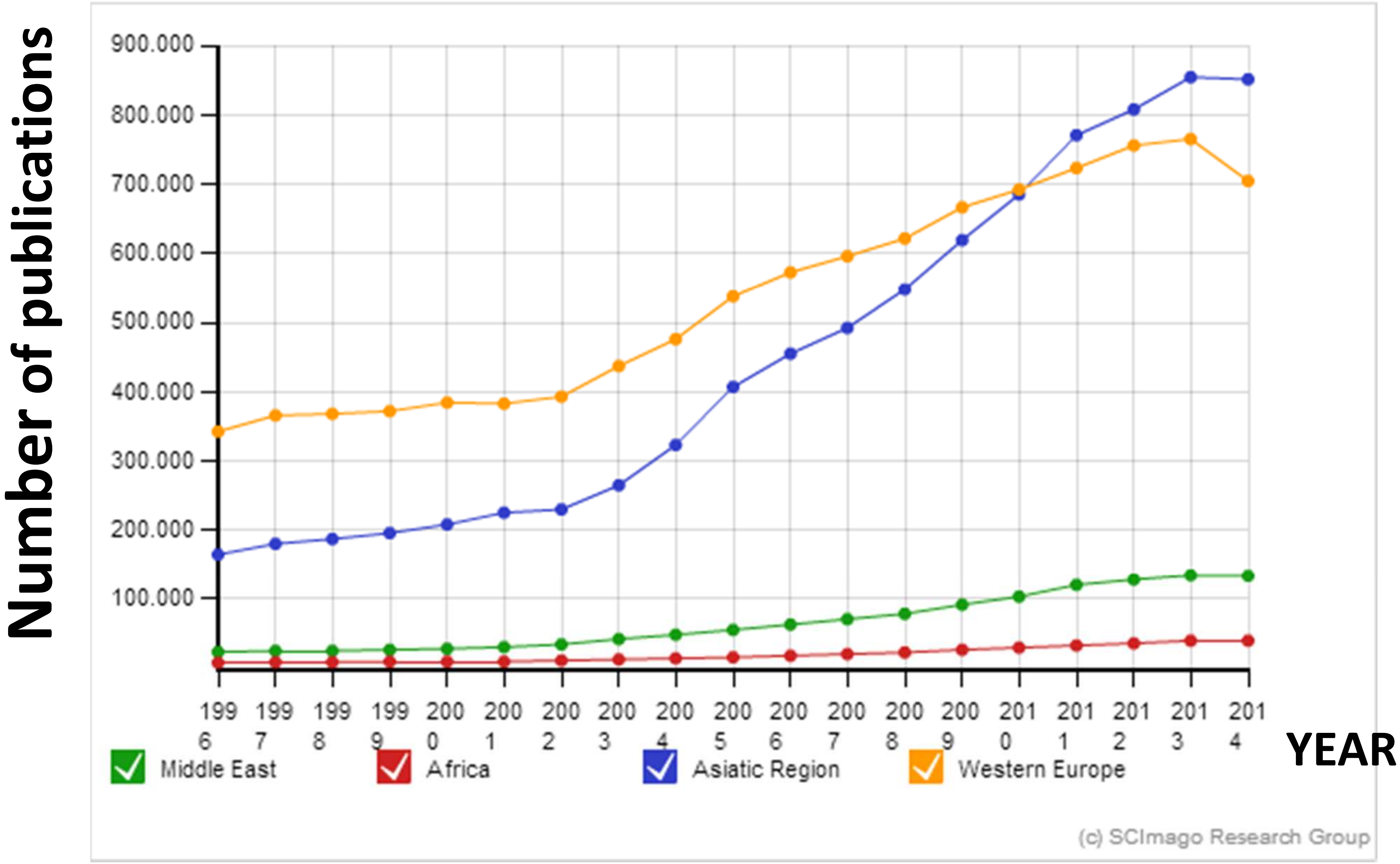
Many “big data” problems that biostatisticians encounter involve genomic data and involve biological discovery or prediction rather than hypothesis testing. Time to revisit traditional hypothesis testing?

Challenge 3: Statistical knowledge is not equally distributed

Publication productivity in Central Asia and Africa are low, mostly due to lack of statistical training among local scientists

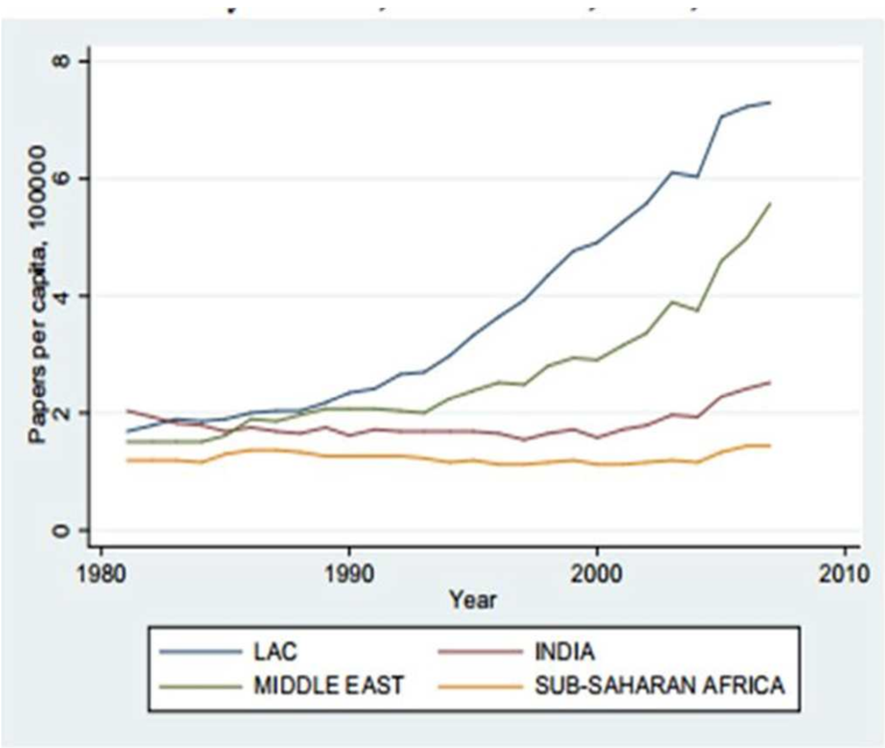


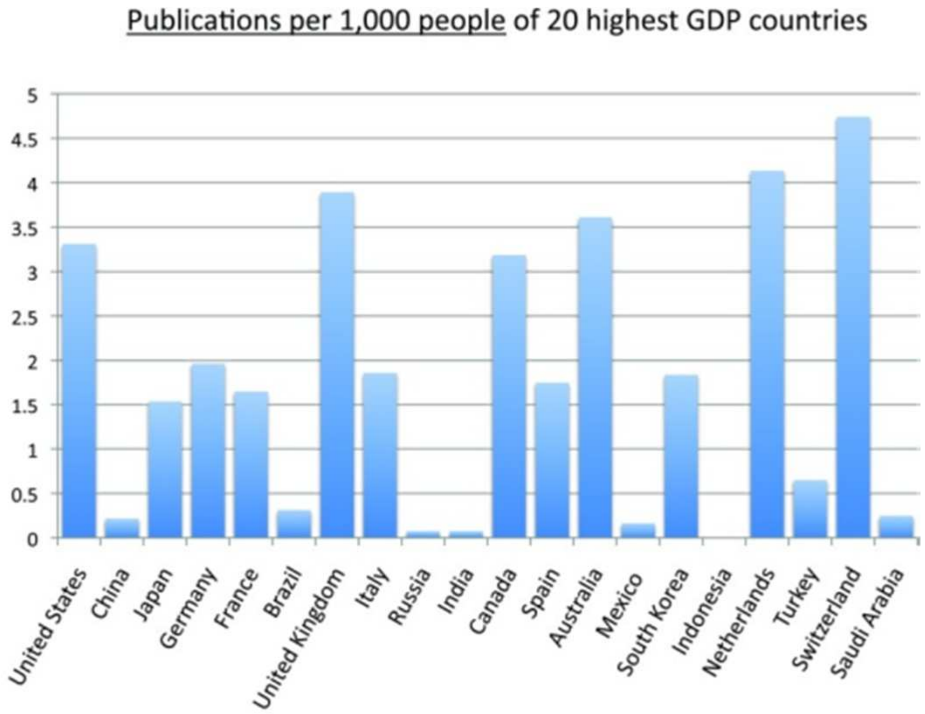
Productivity gap between Africa/Middle East and Asia/Western Europe



(c) SCImago Research Group

Per capita publications




Publications per 1,000 people of 20 highest GDP countries

Xu, Pubmed, 2014

Concept

- We need to **provide access to publishing** in quality peer reviewed journal to scientists who have limited publishing opportunities
- We are establishing a sustainable business model for the Central Asian Journal of Global Health (CAJGH), a peer-reviewed open access mentored journal with specific focus on developing countries.

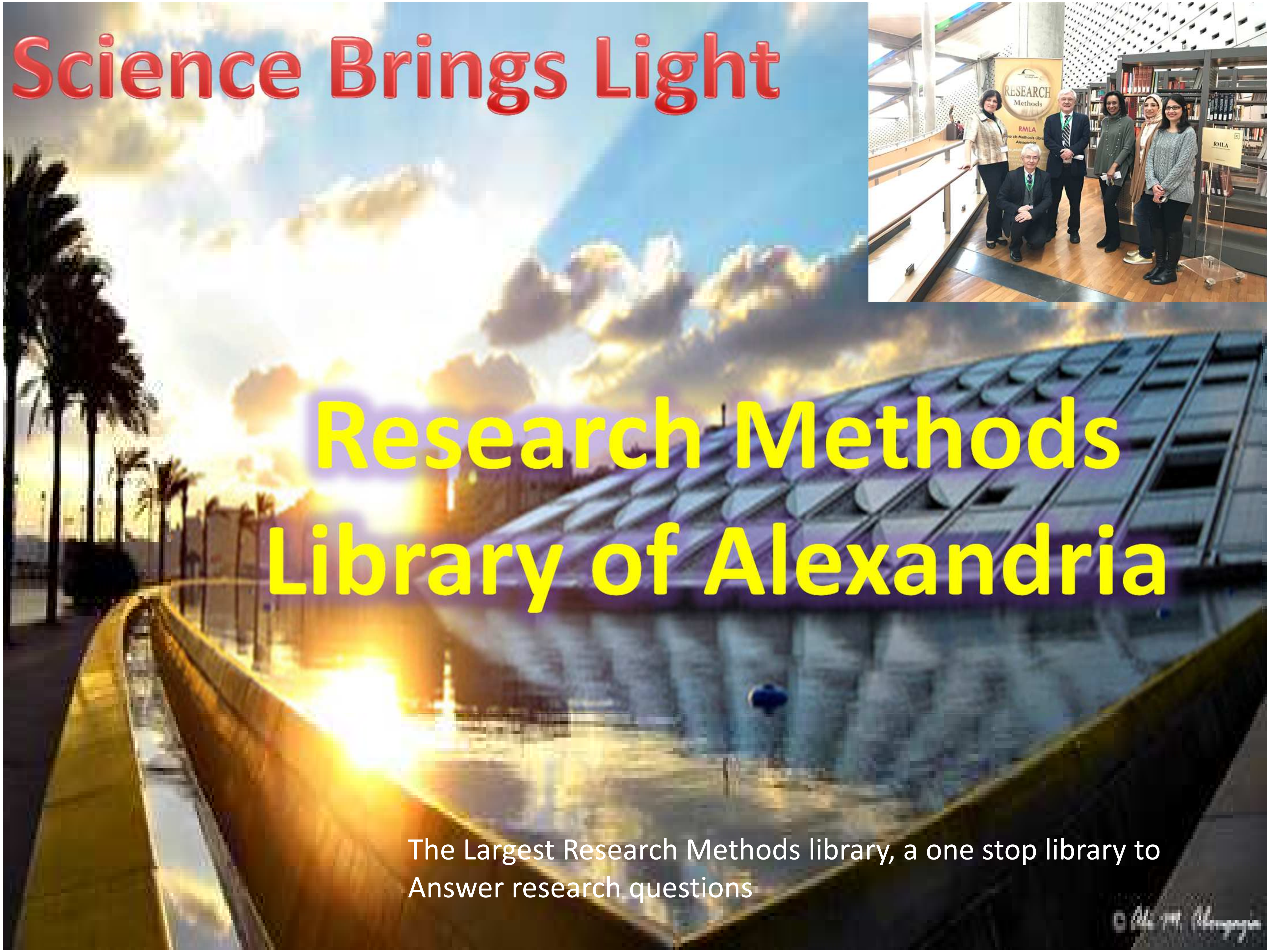
Challenge 4: Research productivity is statistics depended

"There are lies, damn lies, and statistics. We're looking for someone who can make all three of these work for us."

Why low publication rates in developing countries?

- Difficulty in publishing academic work for junior investigators and those in the developed world
- Traditional journals are difficult to get into, open access journals charge high publication fees, while their quality is in question
- Limited publishing skills “Publish or Perish” concept
- Large percentage of authors from Africa and Central Asia publish in predatory/low quality journals

What are the solutions?

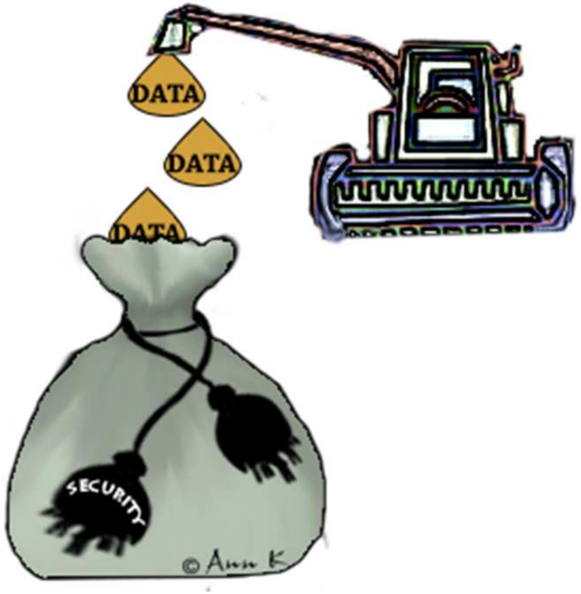


Science Brings Light

Research Methods Library of Alexandria

The Largest Research Methods library, a one stop library to Answer research questions

Establishing methodologies for effectively analyzing mass data harvests



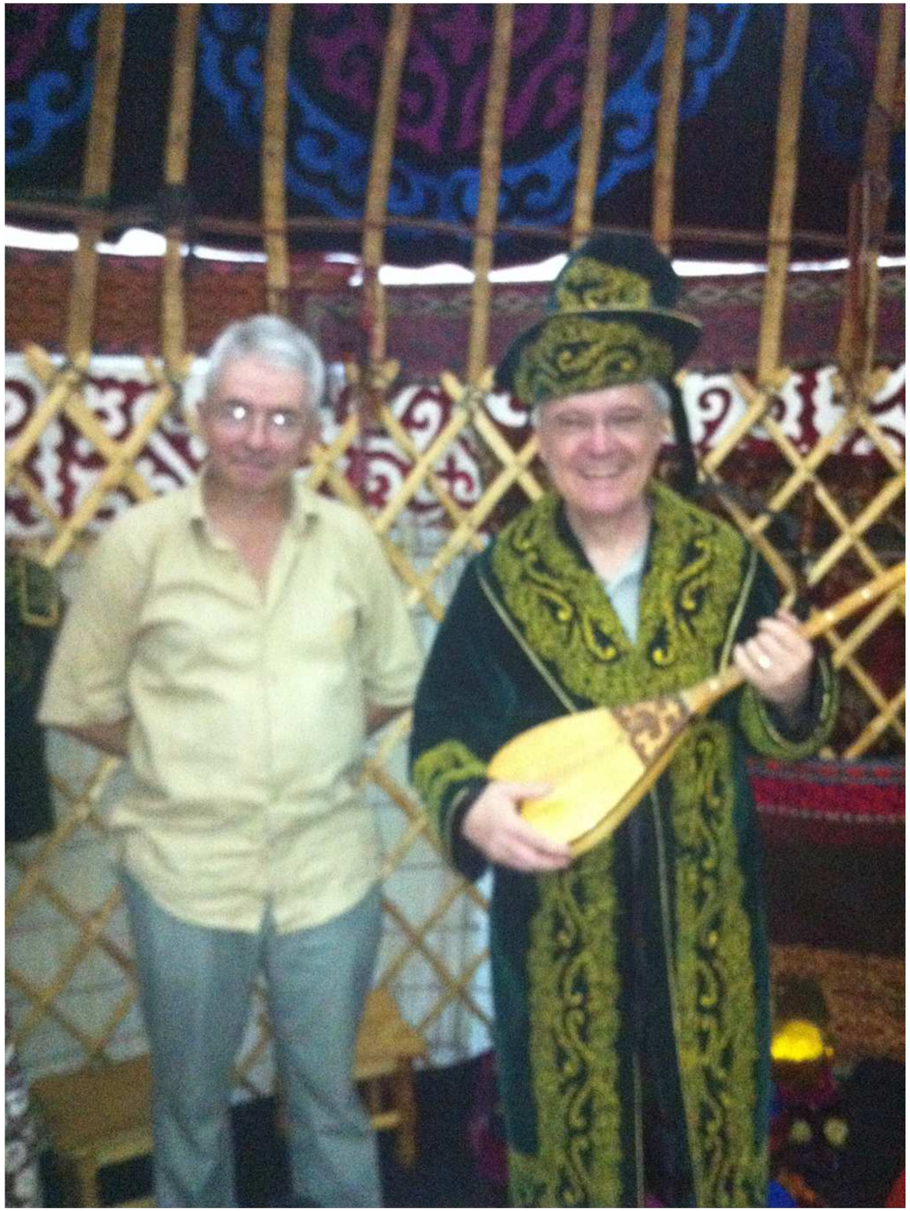
Rows of servers inside Facebook Data Center in Oregon

Building a Mentoring Journal

CAJGH achievements

- 10 issues published
- The only PubMed referenced journal in CA
- Application to SCOPUS has been submitted
- Initial financial support from Nazarbayev University/Kazakhstan government/USAID
- Network of over 2000 scientists
- Collaboration with ministries and universities
- Recognized as leader in academic publishing in Kazakhstan

Research Methods Training course in Astana, Kazakhstan, 2012
Should we establish more training courses?



Research Methods Library of Alexandria

RESEARCH
Methods
RMLA
Research Methods Library of Alexandria
Serageldin-Euclid Library
A Library of Classical Statistics, Research Methodology and On-line Resources
RMLA

American Association for Advancement in Science, 2016

Vinton Cerf, Ismail Serageldin, Ronald LaPorte, Faina Linkov

Conclusions

- The number of data sources is exploding in global health
- Our tools for handling these data are limited
- Cross disciplinary collaboration is needed to harness the data and improve research productivity
- Library of Alexandria can be the center for data exploration on global health

Questions?
Collaborations?

Please contact Faina Linkov at linkfy@mail.magee.edu