LOS ANGELES EAST

R

USER GROUP

meetup

Introduction to the Biostatistics Division

October 8, 2018

Paul Marjoram
(pmarjora@usc.edu)

The Division

- **Dept. of Preventive Medicine: ~160 faculty**
- **Formed in 1984**
- **Division: 23 faculty; 52 PhD students; 36 Masters students**
- **Roles in education and service…**

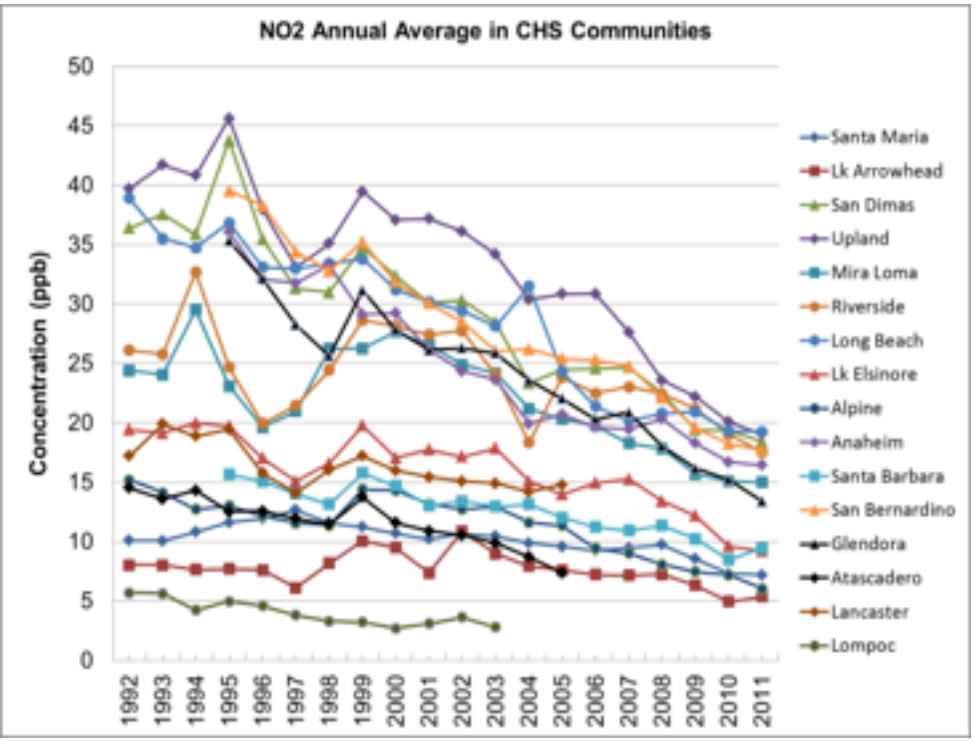
- **Focal research areas:**
  - Statistical genetics and genetic epidemiology
  - Environmental statistics and epidemiology
  - Clinical trials
  - Study design and analysis
  - Big data and Data science...**(emerging theme)**
  - **Typically funded by the National Institutes of Health (National Cancer Institute)**

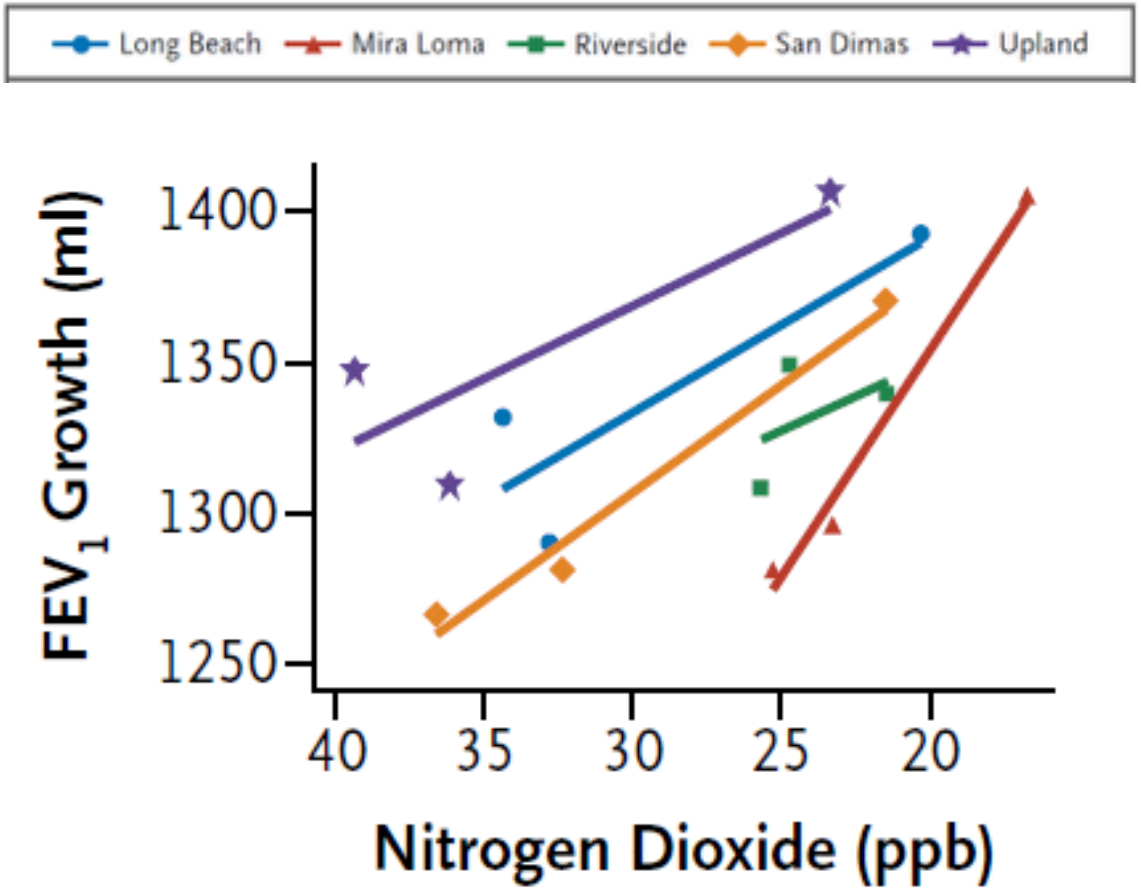
Research: Environmental Statistics

Example: The Children's Health Study [CHS]

Improved air quality → better health

NO_2 trends 1992-2011



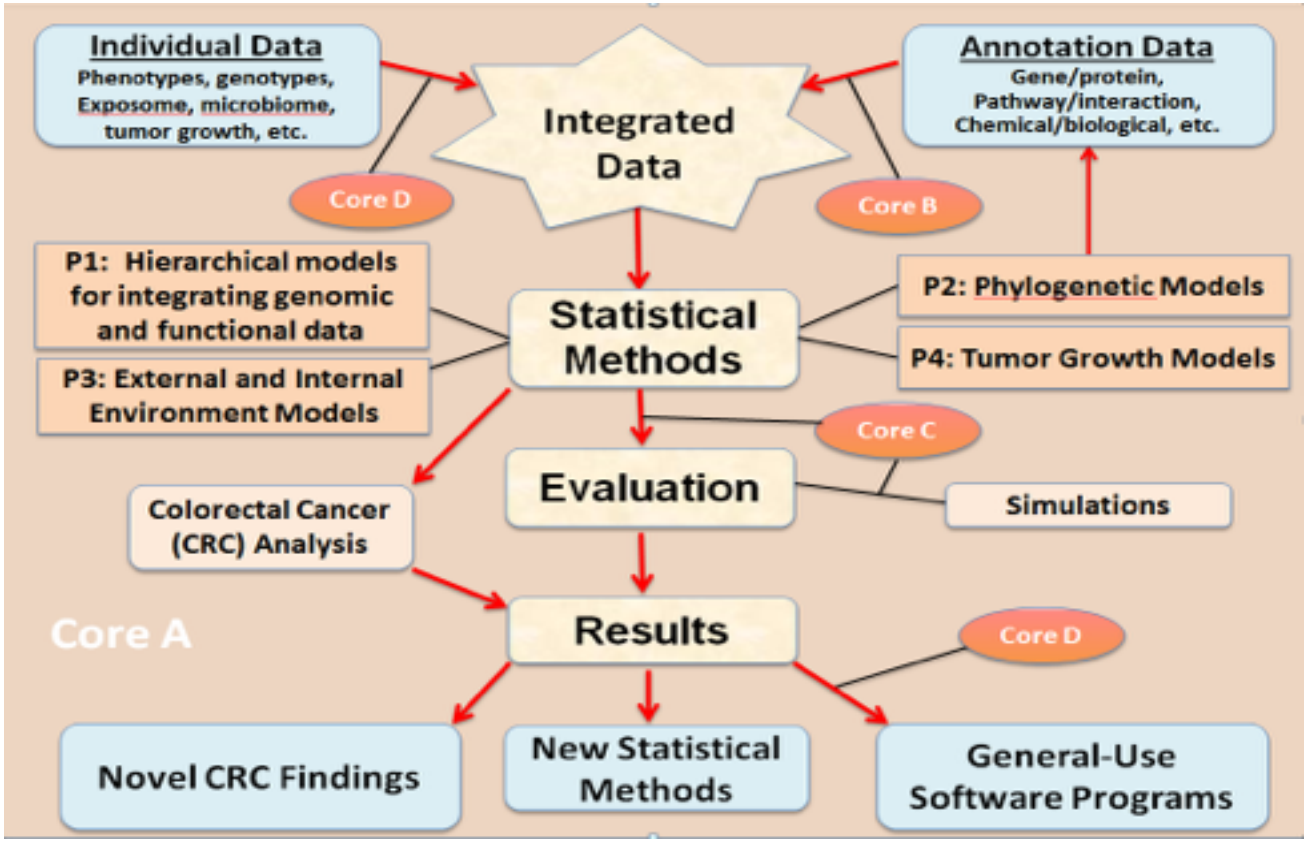


Gauderman et al., NEJM, 2015
Berhane et al., JAMA 2016

Research: Genomics

Statistical Methods for Integrative Genomics in Cancer [P01]

Goal: Develop new methods and software to integrate diverse "-omics" data in models for disease risk.



FIGI: Functionally Informed GxE

- 50,000+ cases, 50,000+ controls (D)
  - "Genetic" data (G):
    - HRC imputation: >39 million Single Nucleotide Polymorphisms
  - "Epi" data (E):
    - Harmonized individual-level environmental exposure data

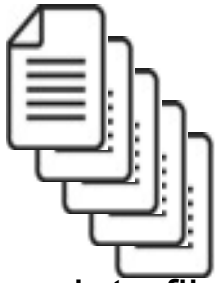






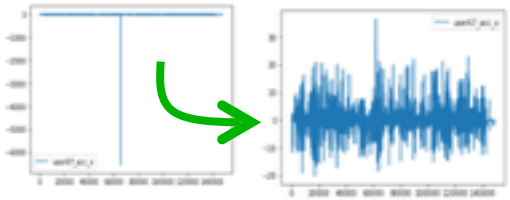
Data Analysis Pipeline

1. Data processing



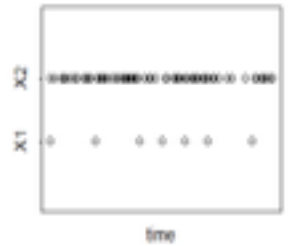
Raw data files

Data manipulation



Noise filters

Data analysis



Temporal alignment

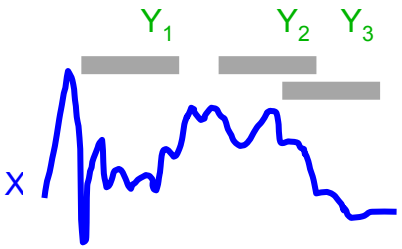
Subject matter expertise

$$x \rightarrow \begin{matrix} f_1(x) \\ \vdots \\ f_k(x) \end{matrix}$$

Transformation

Statistical theory

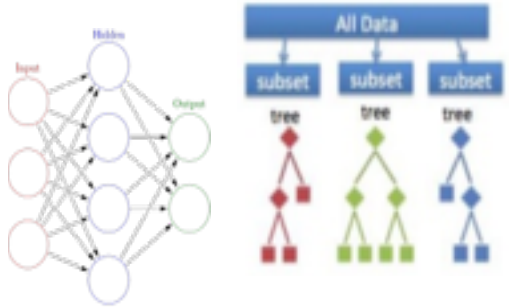
2. Feature engineering



Summarize within windows

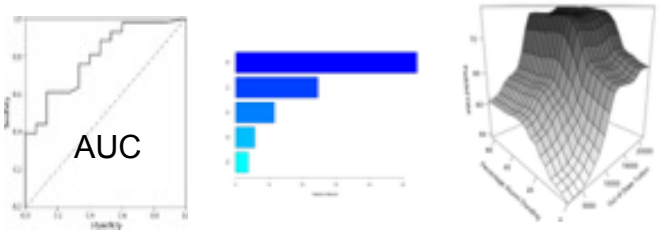
Advanced data analysis, Computational statistics

3. Modeling



Prediction model: Y ~ **X**

Machine learning



Model performance, interpretation, application

Statistical concepts, contextual knowledge, "salesmanship"

USC Center for High-Performance Computing [HPC]

- 2,700 compute nodes - available to all at USC
- Our cluster:
  - 45 nodes
    - 760 cores

George G. Vega Yon
@gvegayon

Follow

It's so nice to be able to request 150 cores for #hpc without having to ask for permission :) #phdlife #rstats #phylogenetics

JOBID	USER	ACCOUNT	PARTITION	NAME	TASKS	CPUS_PER_TASK	MIN_MEMORY	START_TIME	TIME	TIME_LIMIT	STATE	NODELIST(REASON)
1731189	vegayon	lc_pdt	thomas	01-gold-standard-wrong-prior	10	1	500M	2018-09-27T00:33:23	8:15	2:00:00	RUNNING	hpc0953
1731190	vegayon	lc_pdt	thomas	01-gold-standard-wrong-prior	10	1	500M	2018-09-27T00:33:23	8:15	2:00:00	RUNNING	hpc0954
1731191	vegayon	lc_pdt	thomas	01-gold-standard-wrong-prior	10	1	500M	2018-09-27T00:33:23	8:15	2:00:00	RUNNING	hpc0955
1731192	vegayon	lc_pdt	thomas	01-gold-standard-wrong-prior	10	1	500M	2018-09-27T00:33:23	8:15	2:00:00	RUNNING	hpc0956
1731193	vegayon	lc_pdt	thomas	01-gold-standard-wrong-prior	10	1	500M	2018-09-27T00:33:23	8:15	2:00:00	RUNNING	hpc0957
1731194	vegayon	lc_pdt	thomas	01-gold-standard-wrong-prior	10	1	500M	2018-09-27T00:33:23	8:15	2:00:00	RUNNING	hpc0958
1731195	vegayon	lc_pdt	thomas	01-gold-standard-wrong-prior	10	1	500M	2018-09-27T00:33:23	8:15	2:00:00	RUNNING	hpc0959
1731196	vegayon	lc_pdt	thomas	01-gold-standard-wrong-prior	10	1	500M	2018-09-27T00:33:23	8:15	2:00:00	RUNNING	hpc0960
1731197	vegayon	lc_pdt	thomas	01-gold-standard-wrong-prior	10	1	500M	2018-09-27T00:33:23	8:15	2:00:00	RUNNING	hpc0961
1731198	vegayon	lc_pdt	thomas	01-gold-standard-wrong-prior	10	1	500M	2018-09-27T00:33:23	8:15	2:00:00	RUNNING	hpc0962
1731199	vegayon	lc_pdt	thomas	01-gold-standard-wrong-prior	10	1	500M	2018-09-27T00:33:23	8:15	2:00:00	RUNNING	hpc0963
1731200	vegayon	lc_pdt	thomas	01-gold-standard-wrong-prior	10	1	500M	2018-09-27T00:33:23	8:15	2:00:00	RUNNING	hpc0964
1731201	vegayon	lc_pdt	thomas	01-gold-standard-wrong-prior	10	1	500M	2018-09-27T00:33:23	8:15	2:00:00	RUNNING	hpc[0954-0958]
1731202	vegayon	lc_pdt	thomas	01-gold-standard-wrong-prior	10	1	500M	2018-09-27T00:33:23	8:15	2:00:00	RUNNING	hpc[0959-0963]
1731185	vegayon	lc_pdt	thomas	01-gold-standard-wrong-prior	10	1	500M	2018-09-27T00:33:23	8:15	2:00:00	RUNNING	hpc3327
1731171	vegayon	lc_pdt	thomas	01-gold-standard-wrong-prior	1	1	8G	2018-09-27T00:33:03	8:35	12:00:00	RUNNING	hpc0953

12:47 AM - 27 Sep 2018

There is a dude from Viterbi owns 384 nodes with 5,124 cores.

"Happy Scientist" Seminar Series

- Good R habits
  - Github for sharing and version control
  - Producing R packages
  - Parallel processing
  - R "Tidyverse"

USCIMAGE
Integrative Methods of Analysis for Genetic Epidemiology

The Happy Scientist Workshop, 2018: #2

Introduction to the Tidyverse, pt 1: Data Wrangling with dplyr

We are pleased to announce the latest of the educational workshops sponsored by the IMAGE Program of USC's Biostatistics Division. This series, the Happy Scientist workshop series, is aimed at providing educational material for researchers, both students and faculty, about a variety of tools and methods that might prove useful to them. If you have any suggestions for subjects that you would like to learn about in future, please send email to (pmarjora@usc.edu). Our agenda will be driven by your specific interests as far as is possible.

Description of this seminar: Developments in the last few years by the R community have revolutionized

dplyr
www.rstudio.com

R Training

(Current Ph.D. students: George Vega Yon, Malcolm Barrett)

- "R Bootcamp": 1 week R training and Hackathon (Aug)
  - 80 participants.

Looking to hire part-time/consultant programmer with expertise in R and high-performance computing (contact me at pmarjora@usc.edu

Software currently on Biostat GitHub

Program	Function
aphylo	Statistical inference of genetic functions in phylogenetic trees
sluRm	Running jobs on HPCC
BinaryDosage	Converts VCF files to a binary format
LUCid	Latent or unobserved clustering with integrated data
GxEScanR	Genomewide scan of GxE, standard and 2-step methods
rbootcamp	Materials for the recent r bootcamp workshop
hierr	An R package for hierarchical regularized regression
bvs	Bayesian variable selection
hpc-with-r	materials for introduction to R for HPC users workshop
partition	Network analysis
rslurm	Submit R calculations to a SLURM cluster
amcmc	Adaptive and other MCMC methods
polygons	Flexible functions for computing polygons coordinates in R
CASI	Canonical Analysis of Set Interactions
software-dev	Coding standards for the USC biostats group
fdrci	Permutation-Based FDR Analysis
TumorModeling	Tumor modeling methods for IMAGE PO1
rphyloxml	Read and write phyloXML files in R
jsPhyloSVG	htmlwidgets for the jsPhyloSVG javascript library
FIGI_analysis	Scripts for analyzing CRC data
qr_regularized_reg	Regularized regression with quantile(Q1) penalty
admixture_bma	Simulations for Admixture Project
multiethn_finemap_meth	Implementation of MJAM

New Masters in Health Data Science

- Quantitative training in biostats and computational skills needed to manage, analyze, and model big data.
- Prepare students to learn from data to address important questions in public health and biomedical sciences.
- Will begin in Fall 2019.

Thanks to…

- Emil Hvitfeld, Szilard Pafka
- Malcolm Barrett, George Vega Yon, Zhi Yang
- Department of Prev. Med -> pizzas!

END