Statistics 215A: Applied Statistics
Fall, 2015
T/Th: 11-12:30 330 Evans

Instructor: Bin Yu binyu@stat.berkeley.edu, 409 Evans Hall
GSI: Karl Kumbier, kkumbier@berkeley.edu

Self introduction (statistics.berkeley.edu/~binyu)

- 28th year on campus

- Current research outlook: understand and solve data problems through critical thinking, domain knowledge, fast algorithms, insights from theory, and good communication and interpersonal skills

- Current research interests: statistical machine learning, high dimensional inference, interdisciplinary research (neuroscience, genomics, remote sensing, text documents,…)

Introduction among students

- Introduce yourself to your neighbors with your affiliation on campus

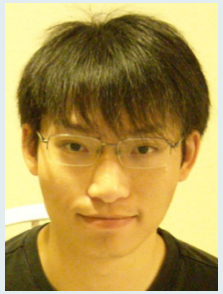
Data collection: heights

- Ask students to write down their estimates
- Pass around a paper for them to enter the numbers
- Two studies

Why are we here?

Anchor project for this class from systems biology

Statistics, UCB



Siqi Wu

Antony Joseph

Karl Kumbier

Erwin Frise

Ann Hammonds

Susan Celniker

Tsinghua Univ, IIIS

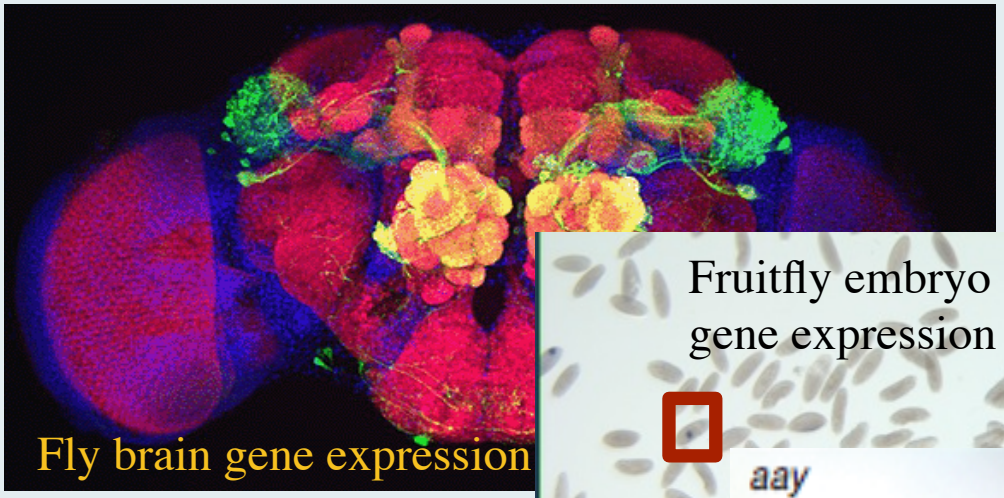
Wei Xu

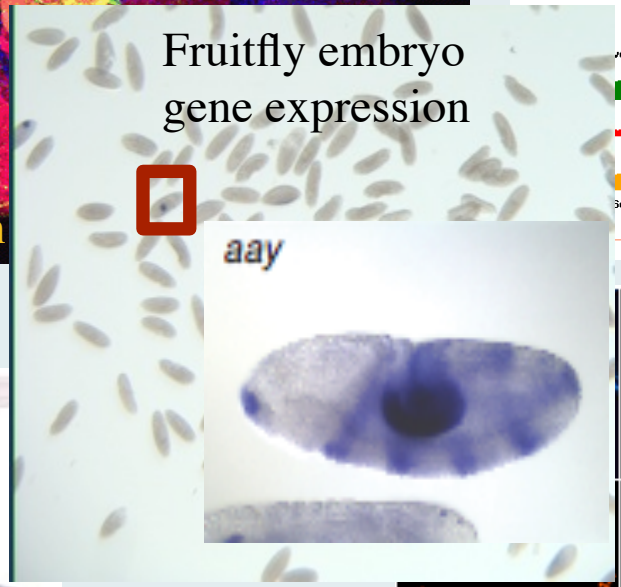
Yang Zhang

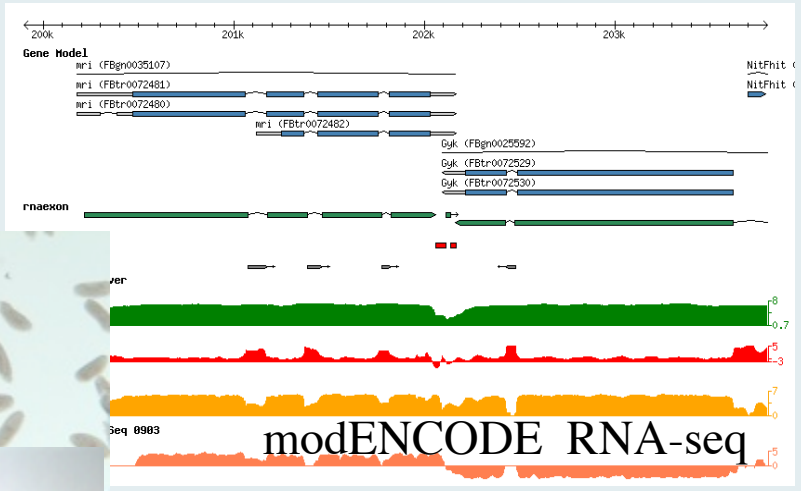
Fangzhou Xu

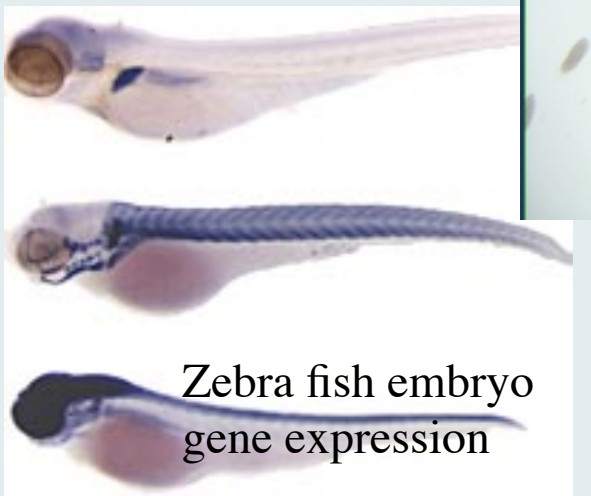
Andy Yao

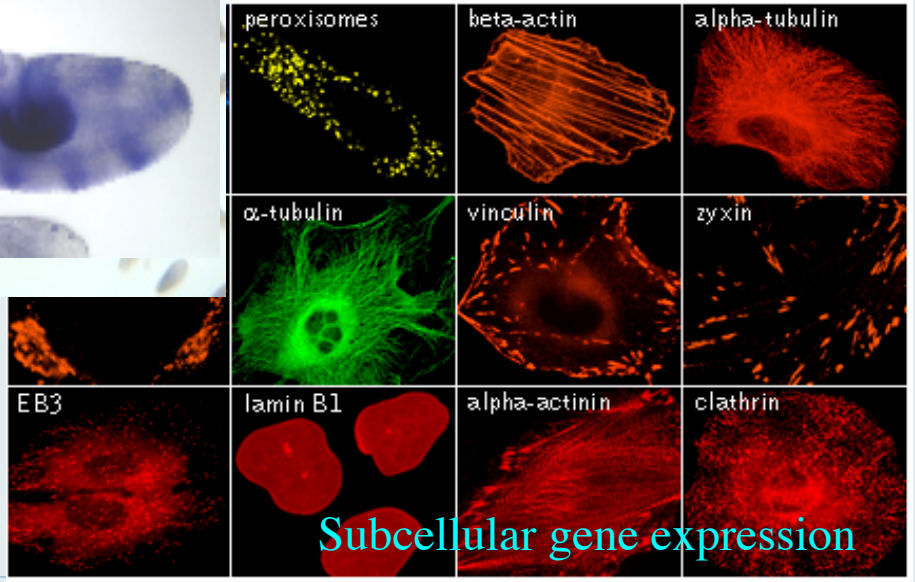
Abundance of systems biology data aims to answer:
how do genes interact (where and when)?





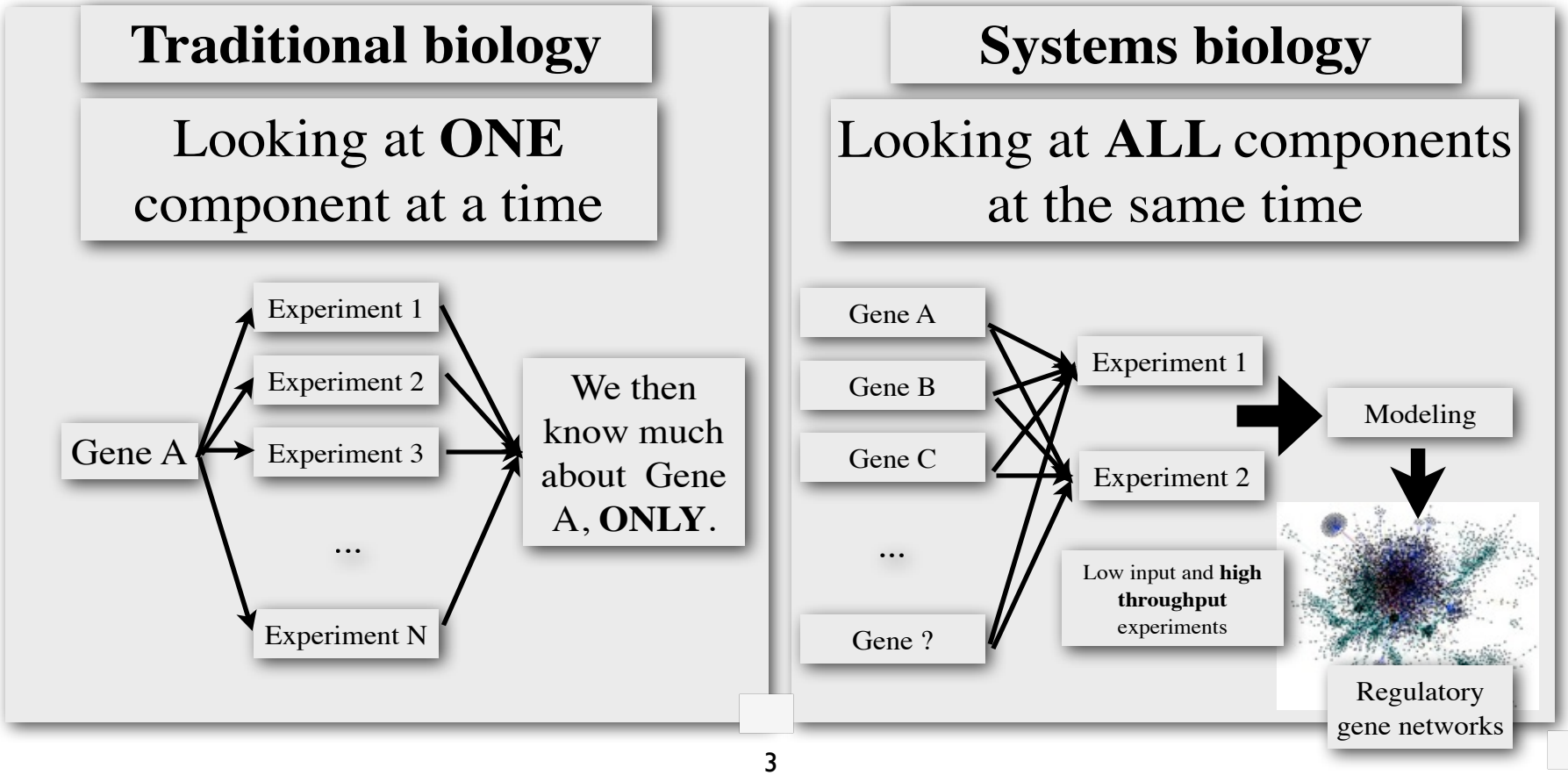






Big picture: Systems Biology

Systems Biology is the study of an organism as an **integrated and interacting network** of genes, proteins and biochemical reactions.



3

Guiding principles for data-intensive science

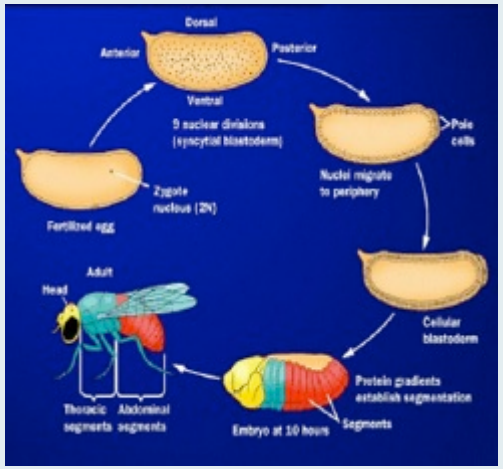
Seed scientific problem(s)

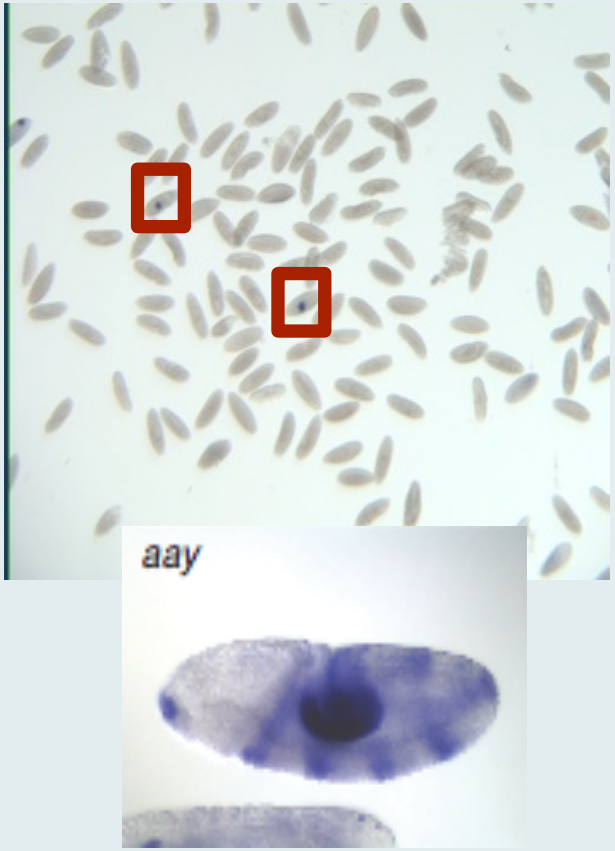
Generalization

"Embedded" students/postdocs work on site,
in the wet lab

The Berkeley Drosophila Genome Project (BDGP)

(my biology collaborator Frise is in the BDGP Celniker Lab)



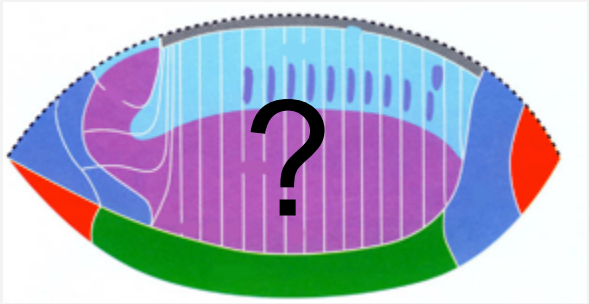


Genes interact in space:
spatial (image) gene expression data

The Berkeley Drosophila Genome Project (BDGP) (cont)

7K+ genes examined – about 1 TB data

We seek answers to the following questions:



Drosophila (fruitfly) embryo

How many functional regions are there?

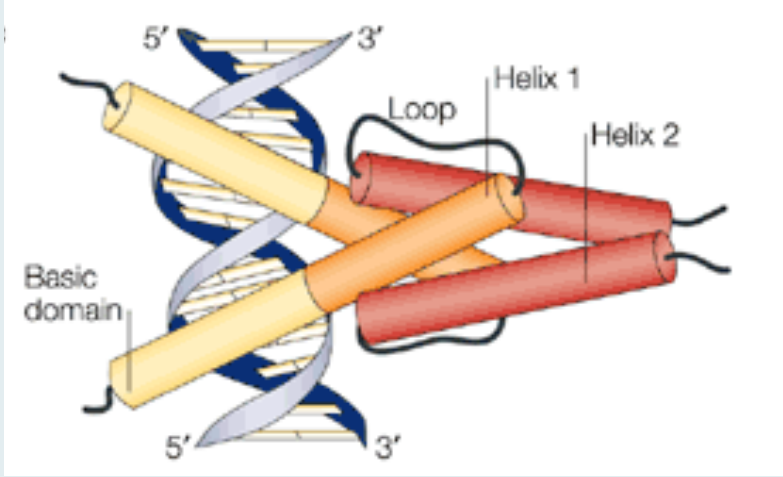
Or can we re-produce the fate map with our data?

Gene A Gene B

What are the new gene functions and gene-gene interactions?

Transcription Factors (TFs):
trigger molecules, corresponding to genes

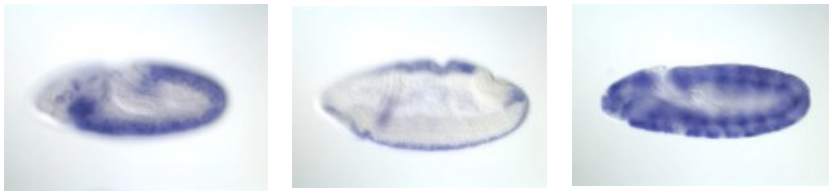
1. What are Transcription Factors(TFs)? DNA binding molecules. On-off switch – triggering other TFs/genes to express.



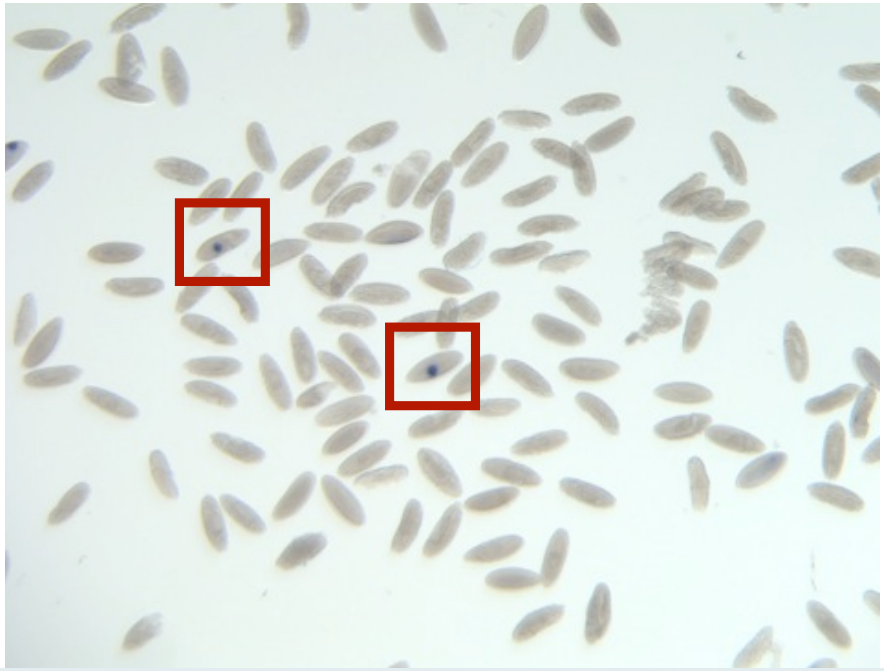
2. Why do we study TFs? They are drivers of gene function cascades and are believed to be almost all of the genes at work in early stages.

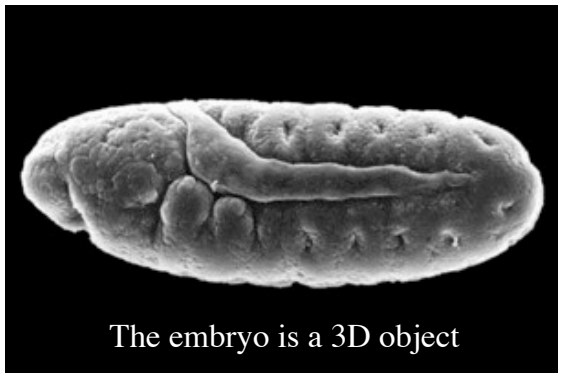
Data collection procedure (1/2)

* Using dye chemistry to visualize spatial gene patterns

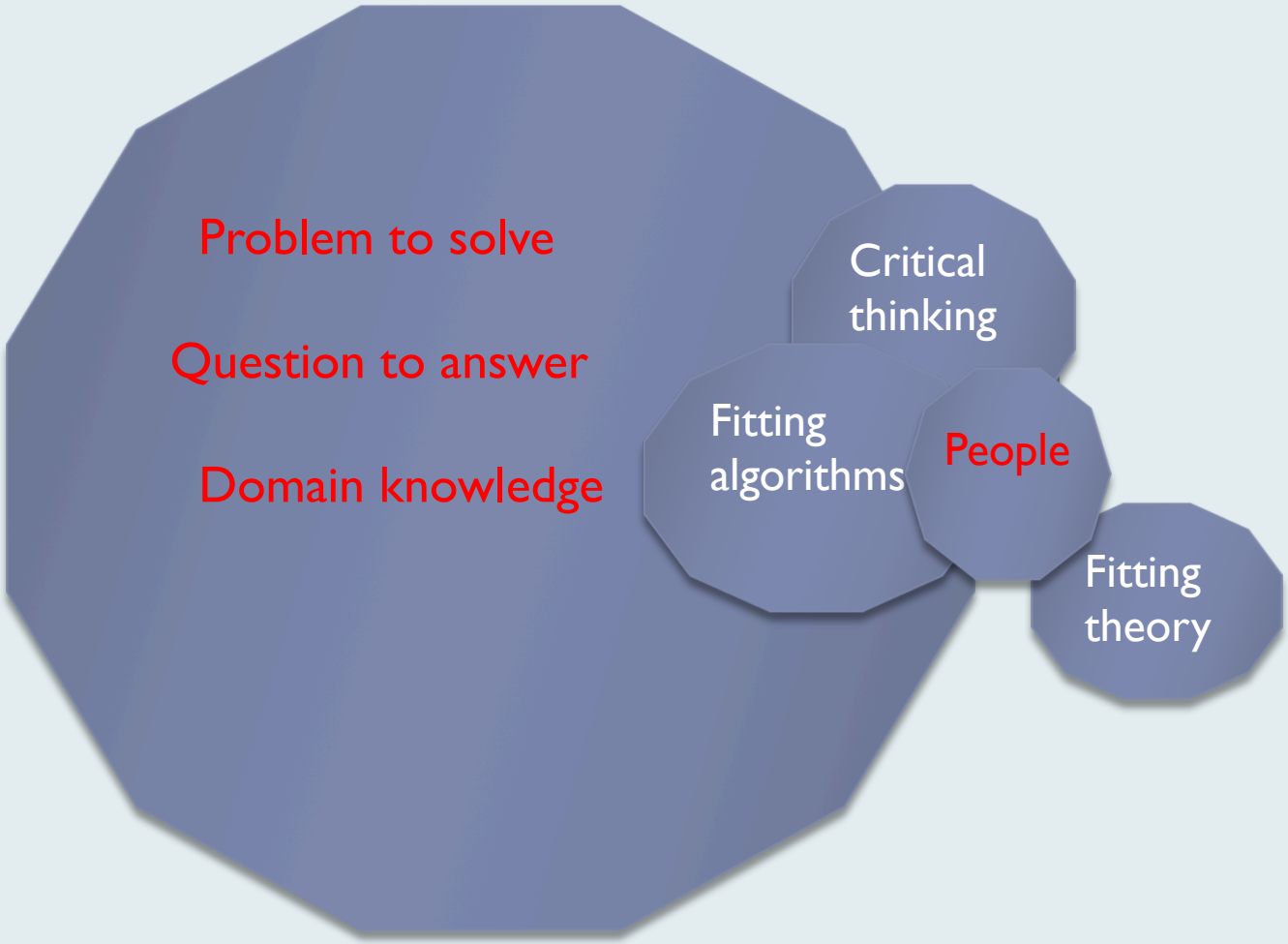


* Imaging under a microscope





Our goal: how to solve real world problems



Our goal dictates the approach we take

- Problem-centered. Multiple skills are musts:

- People skills/Communication skills
- Good judgement calls, common sense, critical thinking skills

- Fitting or appropriate methods/algorithms

- Fitting or appropriate theory

Our goal dictates the approach we take

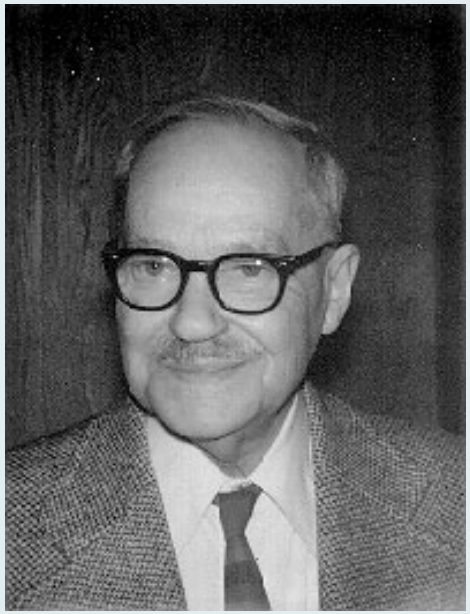
Role of mathematics/methods in this class:

- Neyman's argument to start establish a separate statistics department at UC Berkeley in 1954: statistics is different from mathematics since the mathematics in statistics has to be relevant to real problems.

Jerzy Neyman

Founder of Berkeley Statistics

(April 16, 1894 – August 5, 1981)



"Life is complicated, but not uninteresting."

Neyman-Pearson Lemma
Neyman causal model

Role of math in this class:

Good theorems are always appreciated by me, but not all are relevant to this class.

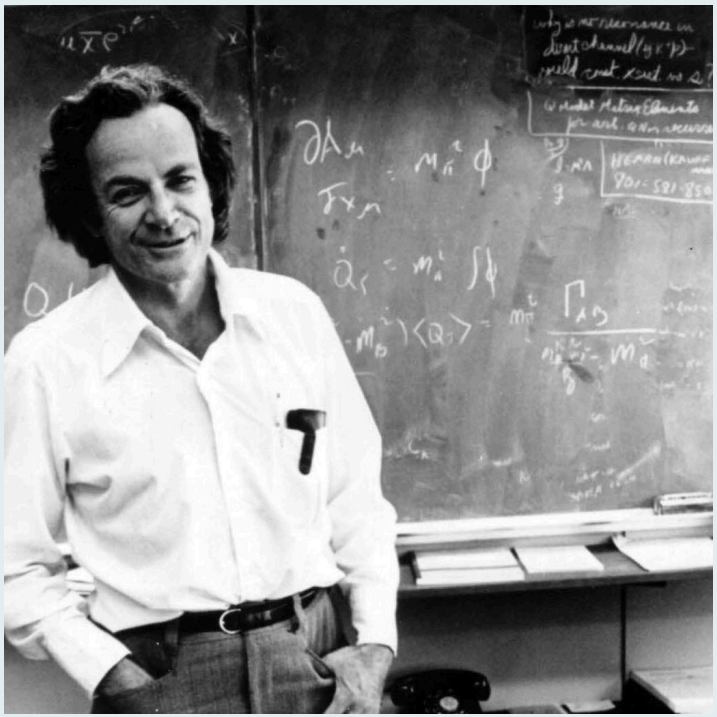
"A good theorem is a good joke" – Tom Cover, information theorist and statistician

(August 7, 1938 – March 26, 2012)

Cover, T. and Hart, P. (1967). Nearest neighbor pattern Classification. IEEE Transactions on Information Theory.

Role of math and ego in this class:

"It doesn't matter how beautiful your theory is, it doesn't matter how smart you are. If it doesn't agree with experiment, it's wrong."



Richard Feynma

(May 11, 1918 – February 15, 1988)

My views on the class (statistics.berkeley.edu/~binyu)

- I do not have all the answers.

- My goals are

to cultivate an effective learning environment that is interactive, respectful, and with straightforward communications in class and office hours and lab sessions, and

to learn together by working through past work (case studies and methodologies/ theories) to illustrate a first-principle approach to solving data problems. This approach can be used for new problems and new methodology developments.

Pedagogy aim for this class: "learn how to learn in context"

- Symbols have meanings in context just like in art

"Untitled" (1989-1990)

Doris Salcedo (born 1958)
Colombian-born sculptor who lives and works in Bogotá.

.

http://www.nytimes.com/2015/02/15/arts/design/doris-salcedo-whose-art-honors-lives-lost-gets-a-retrospective-in-chicago.html?_r=0

- "The Colombian sculptor's work is about mourning. Influenced by the horrific violence she observed throughout the world, but especially in her native Bogotá, Colombia, Salcedo wanted to find a way to bring humanity to the losses. She worries that society has become hardened to violence and that victims have become mere statistics or headlines."

- "These eleven "Untitled" sculptures (1989-90), composed of white cotton shirts in plaster and impaled by steel rebar, are Salcedo's response to two 1988 massacres that took place on banana plantations in La Negra and La Honduras. The shirts represent the standard dress of the plantation workers while alluding to the absence of their bodies as well as the funerary dress for the dead"

http://www.chicagonow.com/show-me-chicago/2015/02/who-is-doris-salcedo/

Our statistical work impacts real world

- Responsibility

there are consequences that could lift the world and

those that could harm

- Professional ethics

honesty, reproducibility

Christine H. Fox ASA President's Invited Speaker at JSM 2015

"Fox, who as acting deputy secretary of defense was the highest-ranking civilian woman ever to serve in the U.S. Department of Defense, will present a talk titled 'The role of analysis in supporting strategic decisions.' For her presentation, she will draw from her extensive experience in senior leadership positions at the Center for Naval Analysis (CNA), Department of Defense and AP"

http://www.eurekalert.org/pub_releases/2015-06/asa-chf060815.php

"Fox, who left the Pentagon earlier this year for a job as a senior adviser at Johns Hopkins Applied Physics Laboratory, is "a brilliant defense thinker and proven manager," Defense Secretary Chuck Hagel said"

http://news.yahoo.com/top-gun-fox-mcgillis-pentagon-144004637.html

Christine H. Fox: B.A. in Math, M.A. in Applied Math

- She spoke much about how to convey statistical results to three different secretaries of defense with different styles.

http://news.yahoo.com/top-gun-fox-mcgillis-pentagon-144004637.html

Real world is messy, symbols are clean

- This class bridges between the messy and the clean – the goal is to help you impact the world

- It differs in many ways from a traditional math/methods class

1. It is NON-linear which might be perceived incorrectly as "disorganized"

2. Students are active partners

Our plan for this class to help you impact the world

- Lecture class time:

  a combination of traditional lecture and group discussions

Discussion time: computing and data labs

Re-branding best of applied statistics

BIG DATA AND ANALYTICAL DATA PLATFORMS - ARTICLES / EXPERT ARTICLES

Data Wisdom for Data Science

Bin Yu, *Departments of Statistics and EECS, University of California at Berkeley*

data wisdom is the ability to combine domain, mathematical, and methodological knowledge with experience, understanding, common sense, insight, and good judgment in order to think critically about data and to make decisions based on data.

http://www.odbms.org/2015/04/data-wisdom-for-data-science/

Active participation is key for this class

- Speak one's mind
- Be an active listener

Goal for improved communication:

Speakers get better at listening; listeners get better at speaking.

Speak or not to speak

- Pros

Practice communication skills – for your work to be appreciated.

To be corrected

Helps to think, clarify your own thoughts

Get people to know you and you to know others

Concentrate better

- Cons:

Embarrassment

Self-deception, confusing people

Grading policy and reading homework

- Handout

- Reading assignments for Tuesday:

  Spence et al (2012)

  Box (1986)

  Freedman's Ch. 1

More reading on Bin's career and views

- 1. "A Conversation with Professor Bin Yu": Full interview at: http://www.icsa.org/bulletin/issues/ICSABulletin13Jul.pdf

- 2. IMS Presidential Address by Bin in 2014

Slides:: http://www.stat.berkeley.edu/~binyu/ps/papers2014/IMS-pres-address14-yu.pdf

Write-up in IMS Bulletin: http://bulletin.imstat.org/2014/10/ims-presidential-address-let-us-own-data-science/

video of a re-run on youtube at IMS new researcher conference at Harvard in Aug. 2014:: https://www.youtube.com/watch?v=92OjsYQJC1U