

# DON'T GET CAUGHT WITH YOUR PANTS DOWN

(QUESTIONABLE DATA AND  
QUESTIONABLE DATA HANDLING!)

Data After Dark  
January 2016

# You will encounter\* ethical dilemmas



(\*Or may have already encountered)

# Core ethical principles in research and data science

- **Moral principles - Belmont Report**
  - Respect for persons
  - Beneficence
  - Justice
- **Regulations – e.g., HIPAA**
- **Practices – Make it easy to do the right thing**

# Three primary categories of ethical problems in health informatics:

- **Healthcare** – how it is performed, its successes and failures
- **Information/data** – management of information, EHRs, data exchange, confidentiality
- **Software** – the tools we develop and use to manage information, diagnostics, analysis

# Health Insurance Portability & Accountability Act (1996) - HIPAA

The dreaded law .... What does it mean for your research?

- Protection for the privacy of Protected Health Information
- Protection for the security of Protected Health Information
- Standardization of electronic data interchange in health care transactions

<http://www.hhs.gov/ocr/privacy/hipaa/understanding/>

[https://en.wikipedia.org/wiki/Health\\_Insurance\\_Portability\\_and\\_Accountability\\_Act](https://en.wikipedia.org/wiki/Health_Insurance_Portability_and_Accountability_Act)

<http://www.gpo.gov/fdsys/pkg/PLAW-104publ191/html/PLAW-104publ191.htm>

# Information: confidentiality versus privacy

## Privacy applies to people

- How participants are identified
- The setting that participants interact with the research team
- Methods used to collect information about the participants
- Type of information
- Access to the minimum information necessary

## Confidentiality applies to data

- Pertains to identifiable data
- Agreement about access and maintenance
- Procedures to ensure authorized access
- Limitations to confidentiality procedures
- HIPPA protection from disclosure of PHI (personal health Information) data

What if informatics analysis allows re-identification?

# Privacy is NOT the same as security

- Need to define authorized access:
  - Individual Patient?
  - Family Member / Caregiver?
  - Personal Physician
    - Nurses? Other physicians? Medical Assistants?
  - Payer? Health Plan? Government?
  - Employer?
- You can have privacy breaches with secure technology

# Follow the data

- Hospital
- Outpatient Clinics
- Patients' Homes
- Pharmacy
- Outsourced Services
- Home Health

What are the risks?



# Public access to data

Are there cases when it is important to have public access to personal health data?

- Public health – surveillance, epidemiological investigations, population-based interventions
- Research
- Quality assurance / monitoring fraud / abuse

# Evaluating an informatics software tool

1. Does it work as designed?
2. Is it used by whom it was designed for?
3. Does it produce the desired results?
4. Does it work better than the procedures it replaced?
5. To what extent do effects depend on practice setting?
6. Is it cost effective?
7. What training is available in its use and how effective is this training?
8. What are the long-term effects on the delivery of medical care?
9. How does the tool impact the organizations in which it is implemented?

# Ethics is a team sport

- Codes of ethics
- Case studies
- Ethics committees and personnel
- Informal discussion

# OHSU Ethics Resources

OHSU Center for Ethics

<http://www.ohsu.edu/xd/education/continuing-education/center-for-ethics/>

OHSU Ethics programs:

<http://www.ohsu.edu/xd/education/continuing-education/center-for-ethics/ethics-programs/>

- Ethics Consult Service (ECS). Health care professionals address challenging ethical issues that confront patients, families and their care team through education, policy development, and consultation.
- Institutional Ethics Committee (OIEC). Faculty address organizational ethics issues that have significant effect on clinical care, research, and system administration.
- Interprofessional Ethics Fellowship. 2-year certificate program.

NIH ethics Program: <http://ethics.od.nih.gov/>

# Code of Ethics -Resources

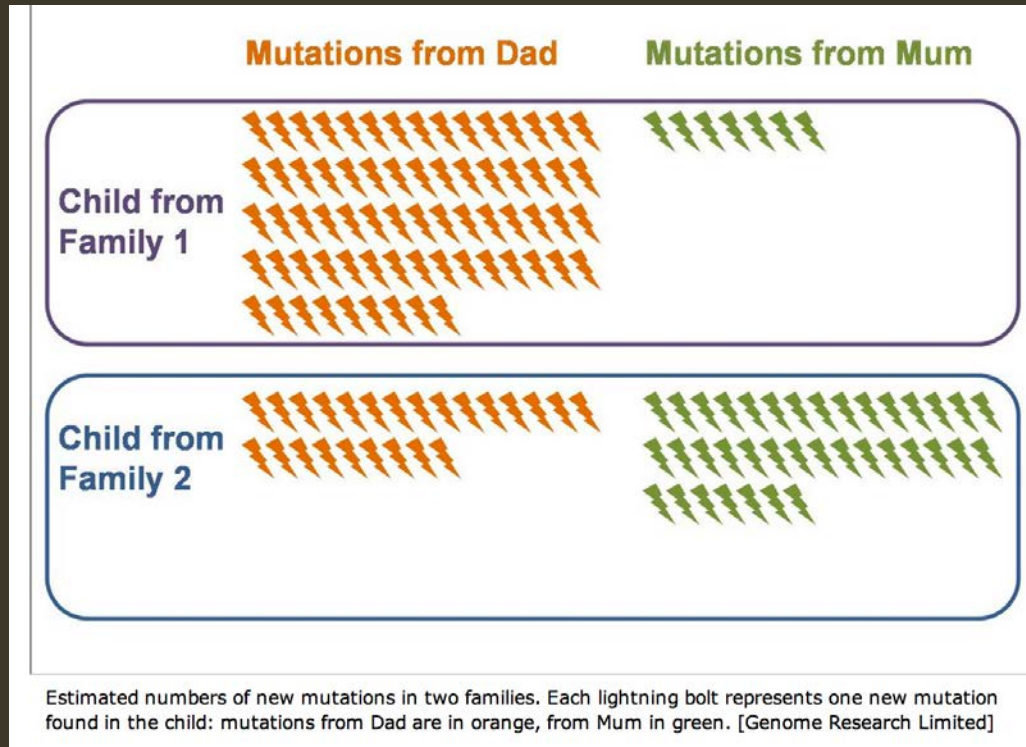
- World Health Organization (WHO)  
<http://www.who.int/ethics/en/>
- International Medical Informatics Association (IMIA)  
[http://www.imia-medinfo.org/new2/pubdocs/Ethics\\_Eng.pdf](http://www.imia-medinfo.org/new2/pubdocs/Ethics_Eng.pdf)
- British Computer Society (BCS)  
<http://www.bcs.org/category/6030>
- American Health Information Management Association (AHIMA)  
[http://library.ahima.org/xpedio/groups/public/documents/ahima/bok1\\_024277.hcsp?dDocName=bok1\\_024277](http://library.ahima.org/xpedio/groups/public/documents/ahima/bok1_024277.hcsp?dDocName=bok1_024277)
- American Medical Informatics Association (AMIA)  
<http://jamia.bmj.com/content/20/1/141.full.pdf+html?sid=63e076ee-a2a1-4e19-842f-4f58bbe044c0>

**SWITCHING  
GEARS  
ABIT...**



Are you a “mutant”?

# We are all “mutants”



We each get approximately 60 new “mutations” in our genome from our parents

<http://www.sanger.ac.uk/about/press/2011/110612.html>

Variation in genome-wide mutation rates within and between human families.  
1000 Genomes Project. [Nature genetics](#) 2011;43;7;712-4



# You are identifiable by your DNA

It has been estimated that only about 100 single nucleotide polymorphisms (SNP) are required to distinguish an individual's DNA record

Lin et al., 2006;

<http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3621020/#R42>

# Principles to determine identifiability

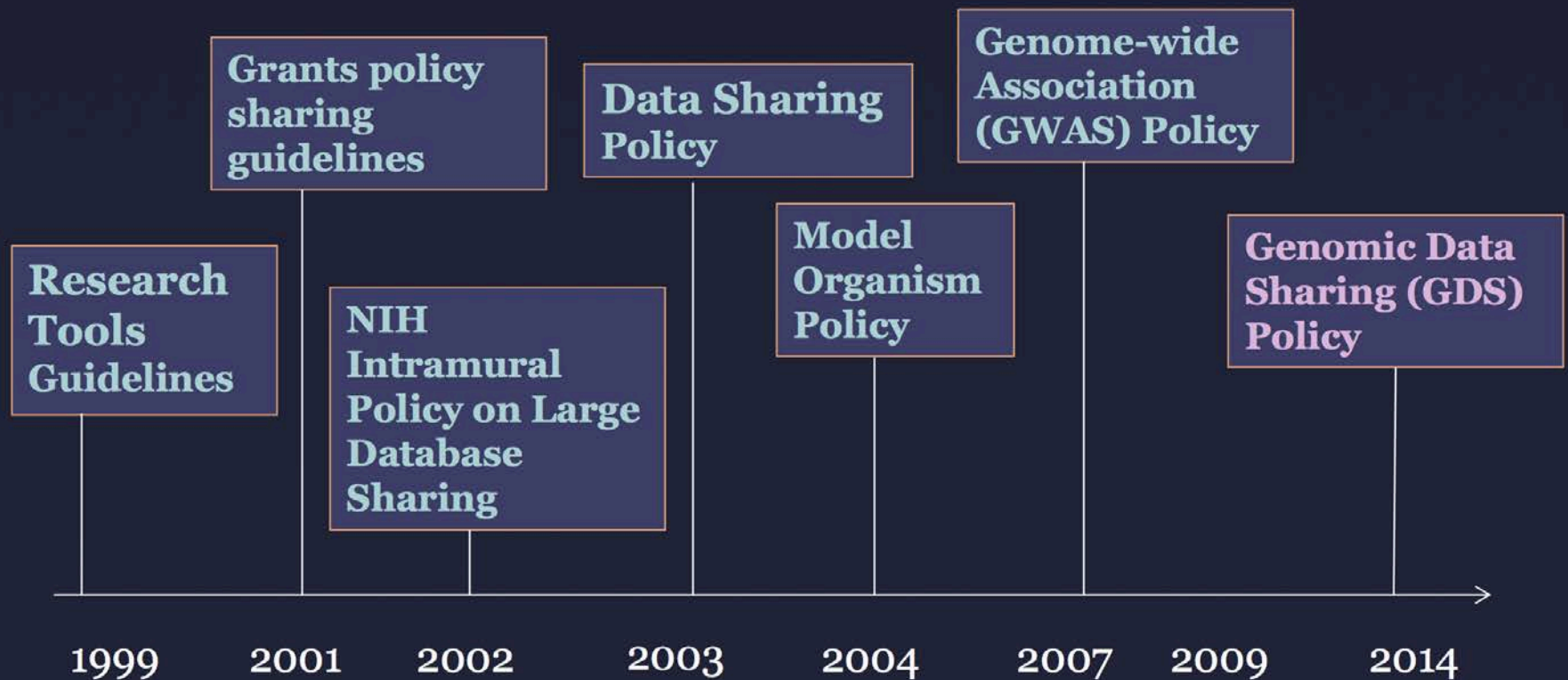
Principle	Description	Examples
Replication	Prioritize health information features into levels of risk according to the chance it will consistently occur in relation to the individual	Low: results of a patient's blood glucose level test will vary  High: Demographics of a patient (e.g. birthdate) are relatively static
Resource availability	Determine which external resources contain the patients' identifiers and the replicable features in the health information, as well as who is permitted access to these resources	Low: The results of laboratory reports are not often disclosed with identity beyond healthcare environments  High: Patient identity and demographics are often in public resources, such as vital records—birth, death, and marriage registries.
Distinguishability	Determine the extent to which the subject's data can be distinguished if health data is disseminated	Low: It has been estimated that the combination of <i>Year of Birth</i> , <i>Gender</i> , and <i>3-Digit ZIP Code</i> is unique for approximately 0.04% of residents in the United States (Sweeney 2007). This means that very few residents could be indentified through this combination of data alone  High: It has been estimated that the combination of a patient's <i>Date of Birth</i> , <i>Gender</i> , and <i>5-Digit ZIP CODE</i> is unique for over 50% of residents in the United States (Golle, 2006, Sweeney 2002a, b). This means that over half of US residents could be uniquely described just with these three data elements

# dbGAP – The NCBI database of Genotypes and Phenotypes

- Had put online aggregate case–control information for each SNP in a study (i.e., the likelihood a person from the case group had a SNP variant, and similarly for the control group).
- Even though aggregated, one could determine if a given person was in the case group, control group, or neither group if you had their DNA
- In 2008, NIH and Wellcome Trust removed these summaries from the public section of databanks, including dbGaP

Look at the dbGAP: <http://www.ncbi.nlm.nih.gov/gap>

# Evolution of Data Sharing at NIH



# Informed Consent & GDS

Studies should ask participant's consent for genomic and phenotypic data to be used for future research purposes and to be **shared broadly**\*

- Explicit explanation regarding sharing via unrestricted- or controlled-access repositories
- If participant does not consent to broad sharing of data, still can be enrolled in the study, but the data may not be shared

For Data Submission, requirement for Institutional Certification  
with Assurance from the IRB

\*As of 1/2015

How can we possibly learn how the genome works without sharing all the data?

And how can we possibly share data if all of it is identifiable?

# Ambiguation of the very personal genome

- NIH dictates that clinical data should be disseminated in a manner that is devoid of identifiers. What to do when the data is itself an identifier?
- Distinguishing records, whether genomic or clinical, is not the same as the ability to identify from whom they came
- Difference between describing path for re-identification and likelihood that path would be leveraged by an adversary
- HIPAA Privacy Rule states that health information designated as de-identified must account for the context of the anticipated recipients, not that the data can never be re-identified

⇒ **Risk Based Framework**

# What are the ethical considerations of genetic screening?

Privacy, inaccuracy, discrimination, eugenics, resource allocation

Watch this GATTACA trailer:

[https://www.youtube.com/watch?v=ZppWok6SX88&list=PLN-SnMXbRhkMkyOCE1-wYq0V8Q6I\\_YMT\\_](https://www.youtube.com/watch?v=ZppWok6SX88&list=PLN-SnMXbRhkMkyOCE1-wYq0V8Q6I_YMT_)



# The Burlington Northern Santa Fe Railroad (BNSF)

- Obtained blood samples from employees who were seeking disability compensation as a result of carpal tunnel syndrome
- Employees were not told the purpose of the tests (and therefore did not consent), which was to perform genetic testing for a mutation on Chromosome 17 that had been associated with hereditary neuropathy with liability to pressure palsies
- Workers were threatened with discharge if they did not provide the sample
- Lewin T. Commission sues railroad to end genetic testing in work injury cases. [New York Times](#). February 10, 2001:A7. => violation of the Americans With Disabilities Act
- Girion L. Railroad Settles Suit Over Genetic Testing. [LA Times](#). May 9, 2002. => Workers paid between \$5,900 to \$75,000, depending on whether they were tested

# Federal law against discrimination

- The presence of certain gene variations could be used against someone in their employment, as we have seen
- 2008 federal law signed by George Bush: The Genetic Information Nondiscrimination Act (GINA)
  - Bill passed Senate unanimously and House by vote of 414 to 1 (who was that person, anyway?)
- GINA bans health insurance companies and employers from requesting or requiring genetic testing; using it for decisions regarding coverage, rates, or preexisting conditions; hiring, firing, or promotion or terms of employment
- States also have genetic discrimination laws, some are weaker and some stronger
- The law doesn't apply to life insurance or long-term care insurance, or to employers with fewer than 15 employees.
- Does not prohibit health insurers or health plan administrators from obtaining and using genetic test results in making health insurance payment determinations.

# Scientific communication and data sharing

# In Paper We Trust

- The peer-reviewed article is the chief means of communicating new knowledge.....and **unfortunately data**
- Scientific publication is a systematic process: a “touchstone” of the scientific method
- Readers and scientific community assume standards have been met

[http://en.wikipedia.org/wiki/Peer\\_review](http://en.wikipedia.org/wiki/Peer_review)

# What are those standards?

- Work is original
- Contributions are accurately acknowledged
- Findings are reproducible, data is available
- Ideas, experimental design, and data have been objectively and independently evaluated.

Like any system, there are  
breaks.

# Both Dramatic....



**This article has been retracted**

[< Prev](#) | [Table of Contents](#) | [Next >](#)

Published Online May 19 2005

*Science* 17 June 2005:

Vol. 308 no. 5729 pp. 1777-1783

DOI: 10.1126/science.1112286

REPORT

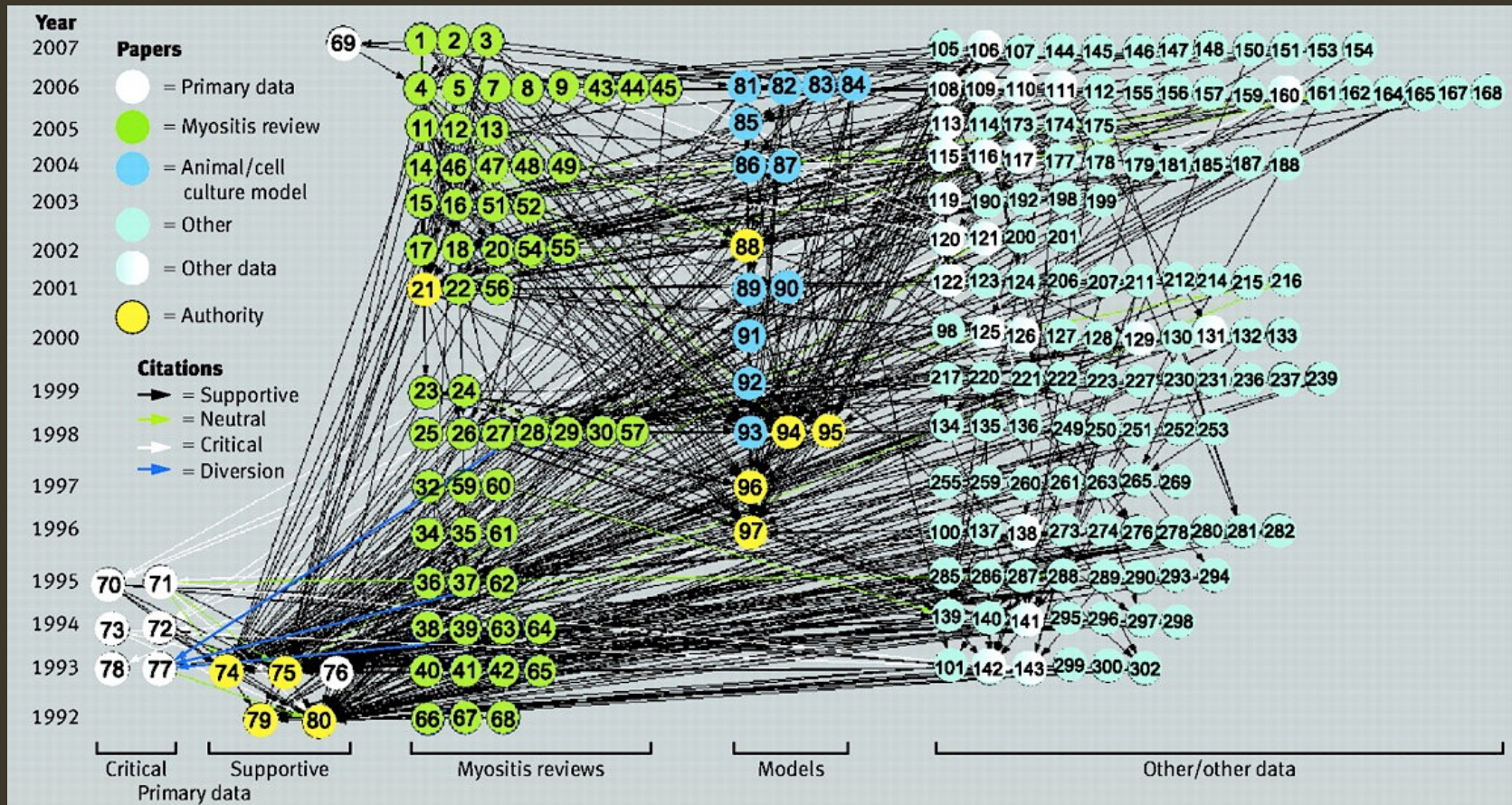
## **Patient-Specific Embryonic Stem Cells Derived from Human SCNT Blastocysts**

Woo Suk Hwang<sup>1,2,\*</sup>, Sung Il Roh<sup>3</sup>, Byeong Chun Lee<sup>1</sup>, Sung Keun Kang<sup>1</sup>, Dae Kee Kwon<sup>1</sup>, Sue Kim<sup>1</sup>, Sun Jong Kim<sup>3</sup>, Sun Woo Park<sup>1</sup>, Hee Sun Kwon<sup>1</sup>, Chang Kyu Lee<sup>2</sup>, Jung Bok Lee<sup>3</sup>, Jin Mee Kim<sup>3</sup>, Curie Ahn<sup>4</sup>, Sun Ha Paek<sup>4</sup>, Sang Sik Chang<sup>5</sup>, Jung Jin Koo<sup>5</sup>, Hyun Soo Yoon<sup>6</sup>, Jung Hye Hwang<sup>6</sup>, Youn Young Hwang<sup>6</sup>, Ye Soo Park<sup>6</sup>, Sun Kyung Oh<sup>4</sup>, Hee Sun Kim<sup>4</sup>, Jong Hyuk Park<sup>7</sup>, Shin Yong Moon<sup>4</sup>, Gerald Schatten<sup>7,\*</sup>

<http://www.sciencemag.org/site/feature/misc/webfeat/hwang2005/>



# And Insidious....



[Greenberg, BMJ 2009;339:b2680](#)



Unreliable research

# Trouble at the lab

The  
Economist

Scientists like to think of science as self-correcting. To an alarming degree, it is not



<http://www.economist.com/news/briefing/21588057-scientists-think-science-self-correcting-alarming-degree-it-not-trouble>

Should Science be Reproducible?

Reproducibility is dependent *at a minimum*, on using the same resources. But...



Bashir

@Bashir\_Course9

method isn't described here b/c this High Impact Report is 200 words. see Supplement Appendix L for vague description  
[#overlyhonestmethods](#)



2 MONTHS AGO



REPLY



RETWEET



FAVORITE



- A well-known journal

Journal guidelines for methods are often poor and space is limited

# What does it mean to be reproducible?

- What is primary conclusion being tested?
- Which experiments need to be reproduced?
- Does the data support the primary conclusion?
- Compare study results statistically
  - Is there an *experimental effect*?
  - A *lab effect*?
  - A synergy between the two?