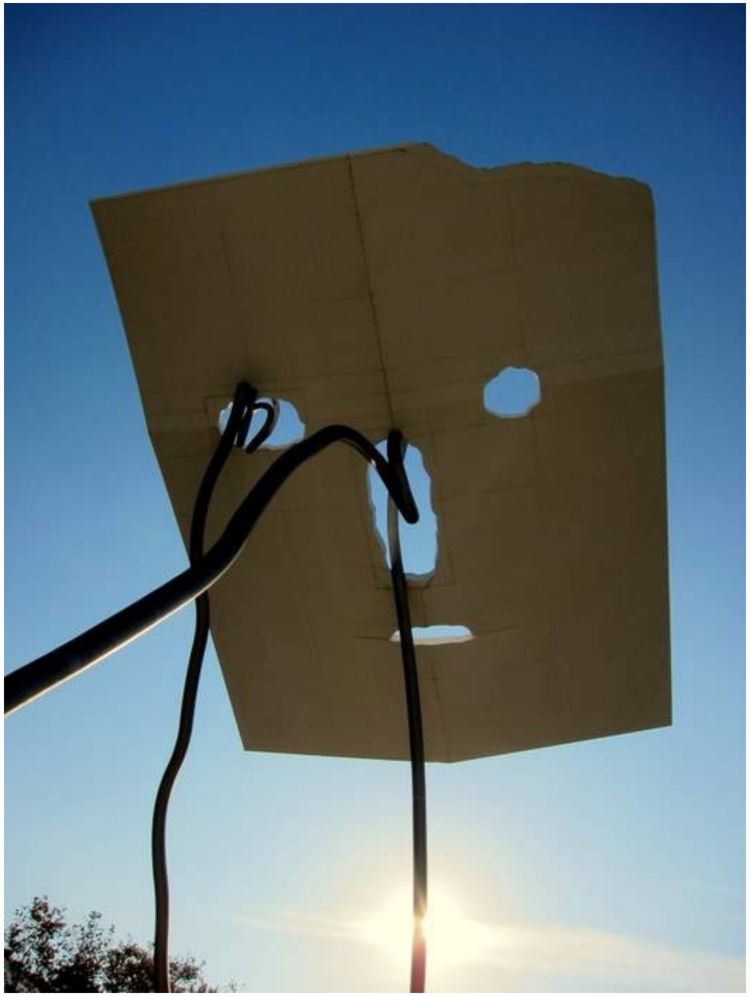
Genomic Privacy:

Proposed Social & Technological Solutions to Issues of Data Privacy in Personal Genomics

(Licensure, Secondary Datasets, Enclaves)

Mark Gerstein
Yale



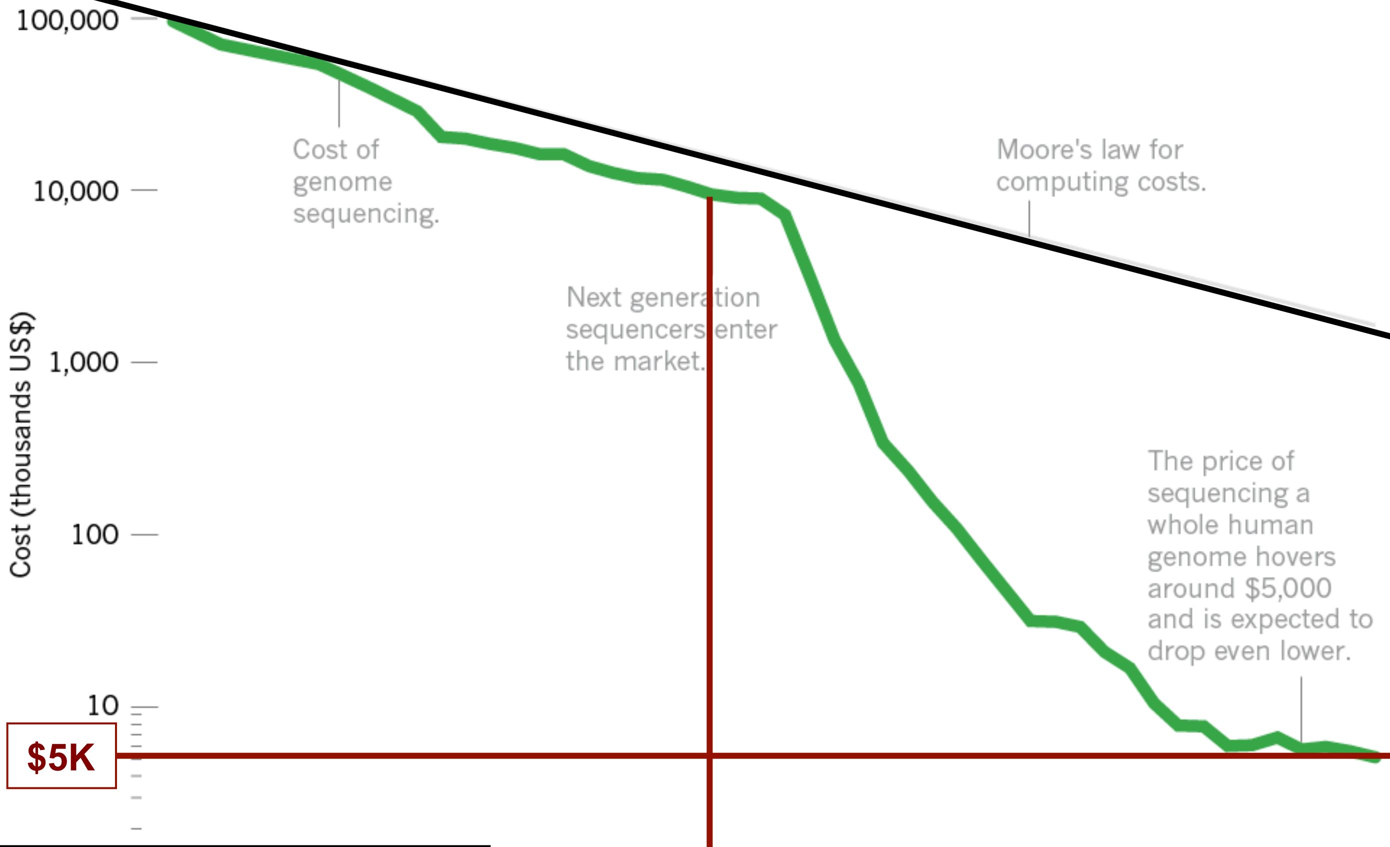
Slides freely downloadable from
Lectures.GersteinLab.org
& “tweetable” (via @markgerstein).
See last slide
for references & more info.

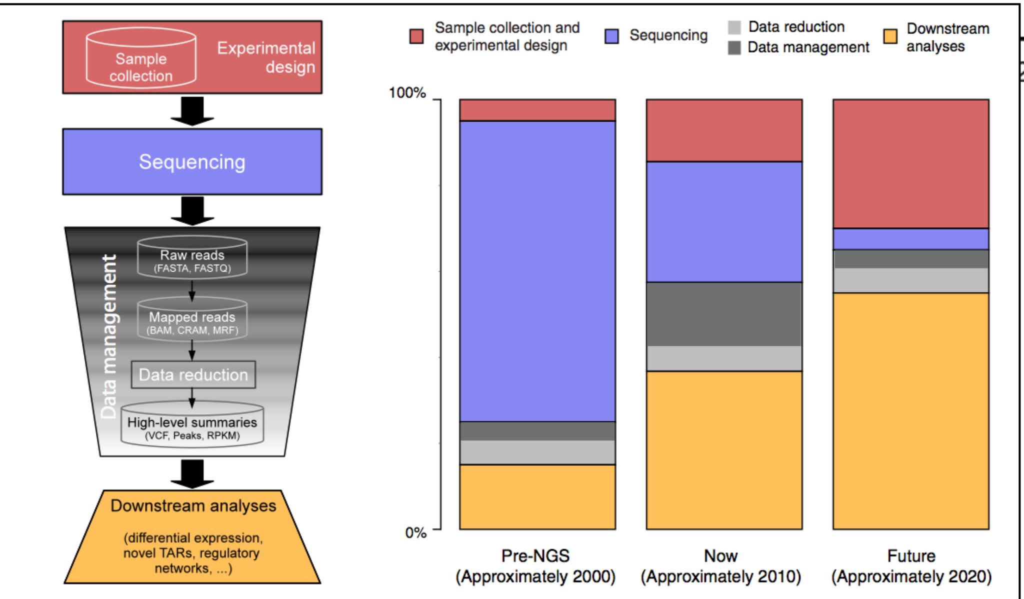
Setting the Stage: the Advent of Personal Genomics

- Human Genome sequence in 2000 for ~$3 billion
- A Human Genome can be sequenced today for ~$1000
- Thousands of SNPs can be interrogated for ~$99

IEEE
SPECTRUM
FOR THE TECHNOLOGY INSIDER | 03.12
BOEING'S BATTERY NIGHTMARE
Why there's no alternative to lithium ion
P. 11
THE NEW CYBORG
Steve Mann on living in augmented reality
P. 42
TRACKING DOWN STUXNET
How the super-worm infected Iran's centrifuges
P. 48
APPLE'S 1987 VISION OF SIRI
Watch the video that foretold the future
SPECTRUM.IEEE.ORG
My Genome
A QUICK AND AFFORDABLE TEST TOLD ME LOTS OF THINGS I WANTED TO KNOW ABOUT MY HEREDITY—
AND ONE THING I DIDN'T
IEEE

The Explosion of Data in Genomics: the Numbers





From ‘00 to ~’ 20, cost of DNA sequencing expt. shifts from the actual seq. to sample collection & analysis

[Nature 507, 294; Sboner et al. (‘11) GenomeBiology]

DTC Genomics

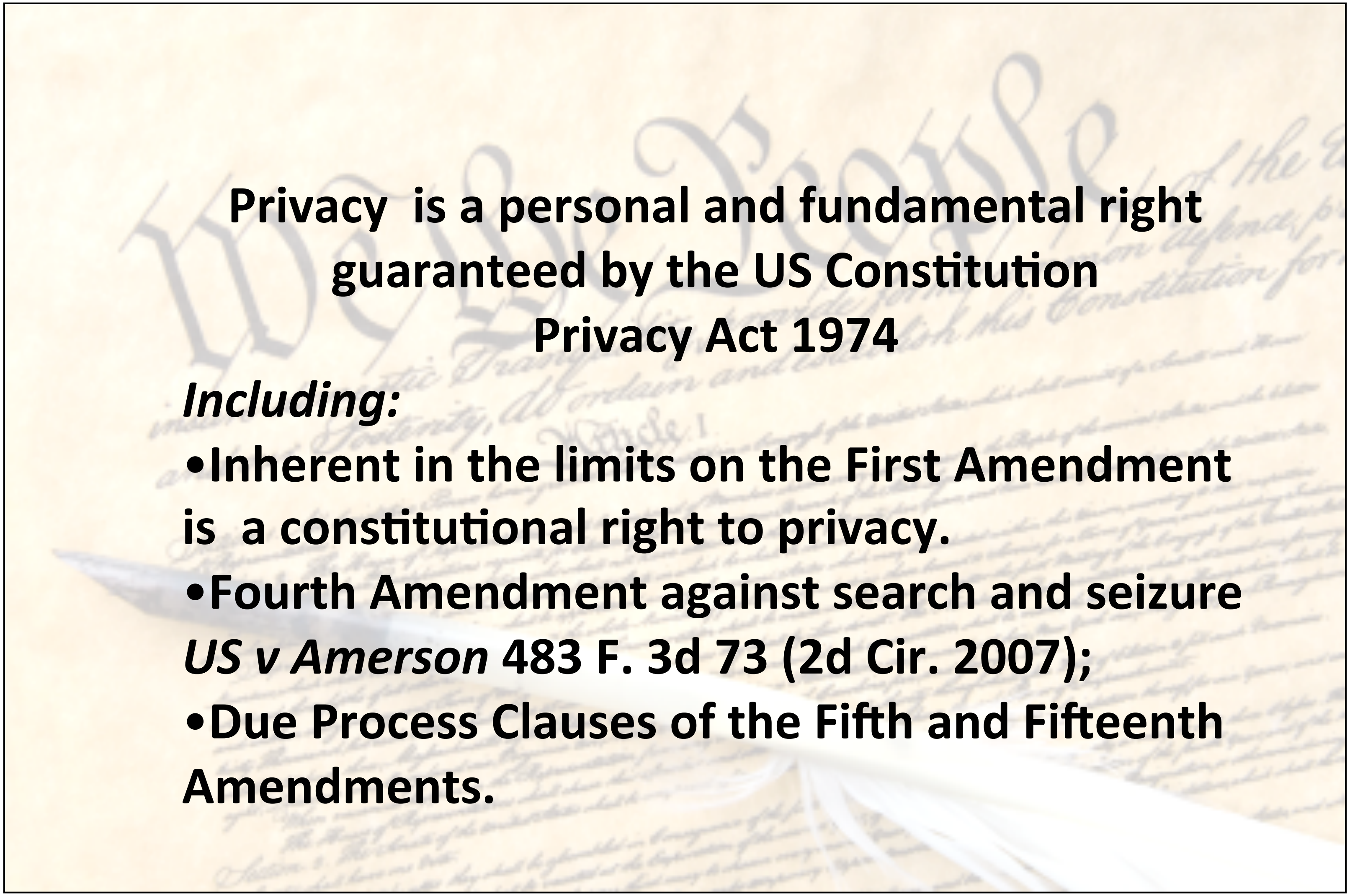
- Industry spurred by falling prices of sequencing and computation
- Major players were Navigenics, DeCode and 23andMe.
- 23andMe
  - has 600,000 Customers
  - $99 per analyiss
  - Promotes sharing of Data
  - Currently in trouble with the FDA and limited to only recreational (e.g., ancestry related) analysis
  - Millions in VC funding

Proposed Social & Technological Solutions to Issues of Data Privacy in Personal Genomics

(Licensure, Secondary Datasets, Enclaves)

- Setting the Stage: the Advent of Personal Genomics
- The of **Conundrum Genomic Privacy**
  - Fundamental, inherited info v need for data-sharing to enable research
- Current Social & Tech Approaches
  - GINA, Consents & **"Secure" use of dbGAP**
  - Ways the solutions have been **"hacked"**
- **Strawman Hybrid Soc-Tech Approach**
  - Soc: Licensure for genomic researchers
  - Enclaves in the cloud
  - Intelligent **data formats mostly splitting private & public info** (ex of MRF)

Privacy



Privacy is a personal and fundamental right guaranteed by the US Constitution

Privacy Act 1974

Including:

- Inherent in the limits on the First Amendment is a constitutional right to privacy.
- Fourth Amendment against search and seizure *US v Amerson* 483 F. 3d 73 (2d Cir. 2007);
- Due Process Clauses of the Fifth and Fifteenth Amendments.

The Conundrum of Genomic Privacy: Is it a Problem?

Yes

Genetic Exceptionalism

Not yet sure of the relevance of the data (but the internet doesn't forget)

Testing discloses both yours' and your family's fundamental data

No

Shifting societal foci

No one really cares about *your* genes

You might not care

Cost Benefit Analysis: how helpful is identifiable data in genomic research?

[Klitzman & Sweeney ('11), J Genet Couns 20:98l; Greenbaum & Gerstein ('09), New Sci. (Sep 23)]

Tricky Privacy Considerations in Personal Genomics

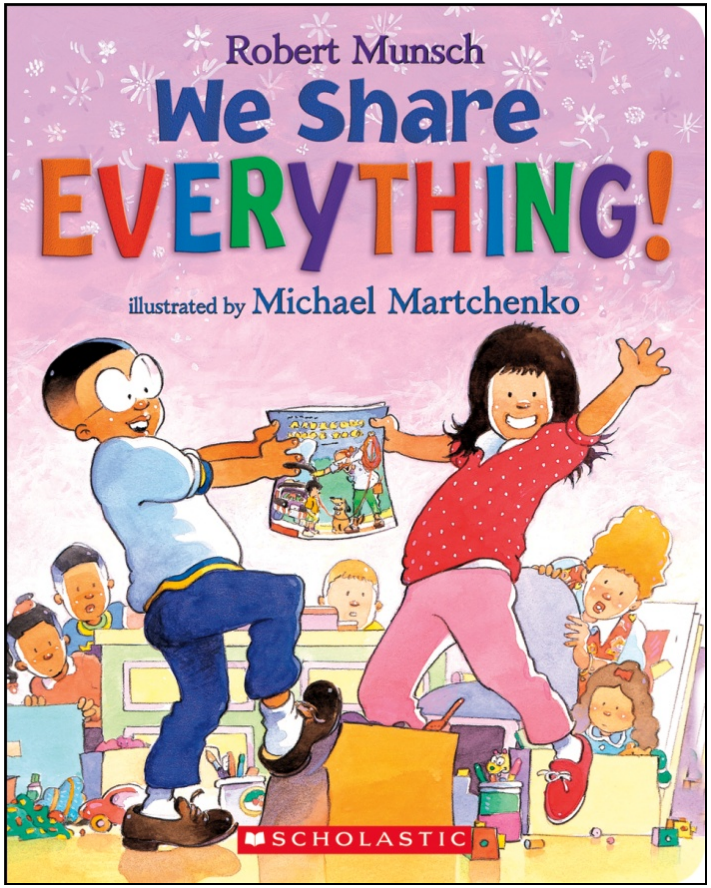
- Personal Genomic info. essentially meaningless currently but will it be in 20 yrs? 50 yrs?
  - Genomic sequence very revealing about one's children
  - Once put on the web it can't be taken back
- Not clear whether they can be treated with "open data" ethos of circa 2000
- Ownership of the data & what consent means (Hela)
  - Could your genetic data give rise to a product line?

- Large discussion of Identification Risk but what about Characterization Risk
  - Finding you were in study X vs identifying that you have trait Y from studying your identified genome

[D Greenbaum & M Gerstein ('08). Am J. Bioethics; D Greenbaum & M Gerstein, Hartford Courant, 10 Jul. '08 ; SF Chronicle, 2 Nov. '08; Greenbaum et al. *PLOS CB* ('11) ; Greenbaum & Gerstein ('13), The Scientist; Photo from NY Times]

The Other Side of the Coin: Why we should share

- Sharing helps speed research
  - Large-scale mining of this information is important for medical research
  - Privacy is cumbersome, particularly for big data
  - Sharing is important for reproducible research
  - Sharing is useful for education
- The individual (harmed?) v the collective (benefits)
  - But do sick patients care about their privacy?
- What is acceptable risk? What is acceptable data leakage?
- Maybe a we need a few "test pilots" (ala PGP)?
  - Sports stars & celebrities?



[Yale Law Roundtable ('10). Comp. in Sci. & Eng. 12:8; D Greenbaum & M Gerstein ('09). Am. J. Bioethics; D Greenbaum & M Gerstein ('10). SF Chronicle, May 2, Page E-4; Greenbaum et al. *PLOS CB* ('11)]

Genomics has similar "Big Data" Issues to the Rest of Society

- Sharing & "peer-production" is central to success of many new ventures, with the same risks as in genomics
- We confront privacy risks every day we access the internet
- (...or is the genome more exceptional & fundamental?)

[Seringhaus & Gerstein ('09), Hart. Courant (Jun 5); Greenbaum & Gerstein ('11), Times (6 Oct)]

Proposed Social & Technological Solutions to Issues of Data Privacy in Personal Genomics

(Licensure, Secondary Datasets, Enclaves)

- Setting the Stage: the Advent of Personal Genomics
- The of **Conundrum Genomic Privacy**
  - Fundamental, inherited info v need for data-sharing to enable research
- Current Social & Tech Approaches
  - GINA, Consents & **"Secure" use of dbGAP**
  - Ways the solutions have been **"hacked"**
- **Strawman Hybrid Soc-Tech Approach**
  - Soc: Licensure for genomic researchers
  - Enclaves in the cloud
  - Intelligent **data formats mostly splitting private & public info** (ex of MRF)

HIPAA

Health Insurance Portability and Accountability Act of 1996

- protects individuals from being charged higher premiums based on Genetics
- does not protect groups from being charged higher premiums
- **Treats Genetic information like all other health information:**
  - **to be protected it must meet the definition of protected health information (PHI):**
    - **it must be individually identifiable**
    - **and maintained by a covered health care provider, health plan, or health care clearinghouse.**
    - **See 45 C.F.R 160.103 and 164.501**
  - **a use or disclosure of genetic information in violation of the HIPAA Privacy Rule could result in a fine of $100 to $50,000 or more for each violation.**
- HOWEVER
- the regulation does not address the type of information that is protected but, rather, who holds it
- many facilities that perform direct-to-consumer genetic testing and analysis are exempt
- **HIPAA** Doesn't prohibit
  - Requiring or requesting genetic tests
  - Disclosure of genetic data without permission
  - Excluding coverage for a condition
- Anonymized biological material is not considered PHI

GINA

GENETIC INFORMATION NONDISCRIMINATION ACT

Genetic Information Nondisclosure Act of 2008

- Title I relating to Health Insurance
- Title II relating to Employment Discrimination
- GINA Prohibits:
  - group and individual health insurers from using genetic data for determining eligibility or premiums
  - insurers from requesting that the insured undergo genetic testing
  - employers from using genetic data to may employment decisions
  - Employers from requesting genetic data about an employee or their family

Current Social & Technical Solutions

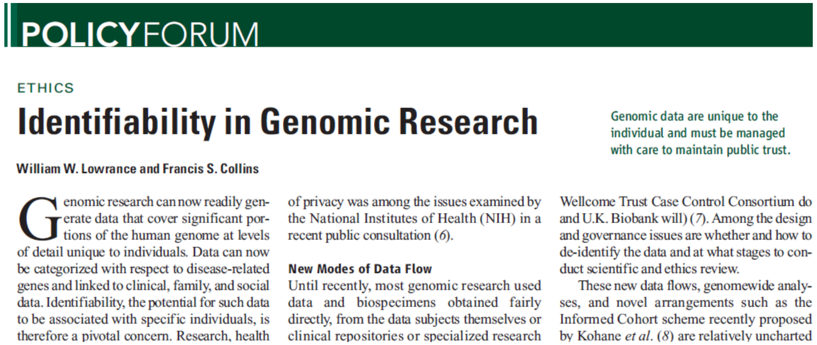
- Consents
- dbGAP distribution of data
- Local computes on secure computer
- Issues
  - Non-uniformity of consents & paperwork
  - Encryption & computer security creates burdensome requirements on data sharing & large scale analysis
    - Security increasingly becoming harder & harder
  - Different international norms

[Smith et al ('05), Genome Biol.; Greenbuam et al ('04), Nat. Biotech; Greenbaum & Gerstein ('13), The Scientist]

Proposed Social & Technological Solutions to Issues of Data Privacy in Personal Genomics

(Licensure, Secondary Datasets, Enclaves)

- Setting the Stage: the Advent of Personal Genomics
- The of **Conundrum Genomic Privacy**
  - Fundamental, inherited info v need for data-sharing to enable research
- Current Social & Tech Approaches
  - GINA, Consents & **"Secure" use of dbGAP**
  - Ways the solutions have been **"hacked"**
- **Strawman Hybrid Soc-Tech Approach**
  - Soc: Licensure for genomic researchers
  - Enclaves in the cloud
  - Intelligent **data formats mostly splitting private & public info** (ex of MRF)



POLICYFORUM

ETHICS

Identifiability in Genomic Research

Genomic data are unique to the individual and must be managed with care to maintain public trust.

William W. Lowrance and Francis S. Collins

Genomic research can now readily generate data that cover significant portions of the human genome at levels of detail unique to individuals. Data can now be categorized with respect to disease-related genes and linked to clinical, family, and social data. Identifiability, the potential for such data to be associated with specific individuals, is therefore a pivotal concern. Research, health

of privacy was among the issues examined by the National Institutes of Health (NIH) in a recent public consultation (*6*).

New Modes of Data Flow

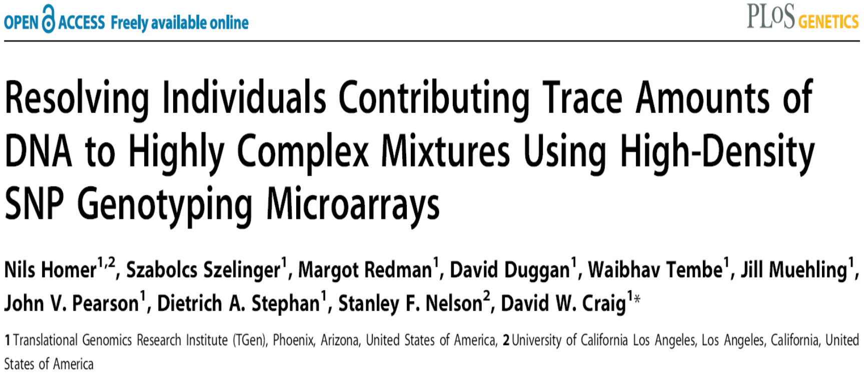
Until recently, most genomic research used data and biospecimens obtained fairly directly, from the data subjects themselves or clinical repositories or specialized research

Wellcome Trust Case Control Consortium do and U.K. Biobank will) (*7*). Among the design and governance issues are whether and how to de-identify the data and at what stages to conduct scientific and ethics review.

These new data flows, genomewide analyses, and novel arrangements such as the Informed Cohort scheme recently proposed by Kohane *et al*. (*8*) are relatively uncharted

***Matching against reference genotype*. The number of DNA markers such as single-nucleotide polymorphisms (SNPs) that are needed to uniquely identify a single person is small; Lin *et al.* estimate that only 30 to 80 SNPs could be sufficient**

***Linking to nongenetic databases*. A second route to identifying genotyped subjects is deduction by linking and then matching geno-type- plus-associated data (such as gender, age, or disease being studied) with data in health-care, administrative, criminal, disaster response, or other databases … If the nongenetic data are overtly identified, the task is straightforward**



OPEN ACCESS Freely available online　PLoS GENETICS

Resolving Individuals Contributing Trace Amounts of DNA to Highly Complex Mixtures Using High-Density SNP Genotyping Microarrays

Nils Homer[1,2], Szabolcs Szelinger[1], Margot Redman[1], David Duggan[1], Waibhav Tembe[1], Jill Muehling[1], John V. Pearson[1], Dietrich A. Stephan[1], Stanley F. Nelson[2], David W. Craig[1]*

1 Translational Genomics Research Institute (TGen), Phoenix, Arizona, United States of America, 2 University of California Los Angeles, Los Angeles, California, United States of America

- a framework for accurately and robustly resolving whether individuals are in a complex genomic DNA mixture using high-density single nucleotide polymorphism (SNP) genotyping microarrays.
- We demonstrate an approach for rapidly and sensitively determining whether a trace amount (<1%) of genomic DNA from an individual is present within a complex DNA mixture
- "Identifying Personal Genomes by Surname Inference,"Gymrek, McGuire, Golan, Halperin, Erlich ('13). Science.
  - Identifying anonymized 1000G individuals through DB xref

About the James Watson Genotype Viewer

On May 31, 2007, Nobel Laureate James Watson received his personal genome sequence in a ceremony at the Baylor College of Medicine. This genome sequence describes the six billion base pairs of DNA that James Watson received from his two parents, the unique combination of which are responsible for James Watson's genetic individuality. Dr. Watson is making his genome sequence available to the public in the hope that it will encourage the development of an era of "personalized medicine" when the information contained in our genomes is used to identify and prevent diseases to which we are genetically prone before they appear, and to create personalized medical therapies that have the maximum benefit and the minimum risk. This simple browser allows you to view the places where Watson's sequence is different from the "reference" human genome sequence, as well as to view the genes and some of the common diseases associated with them.

What the Watson Sequence is

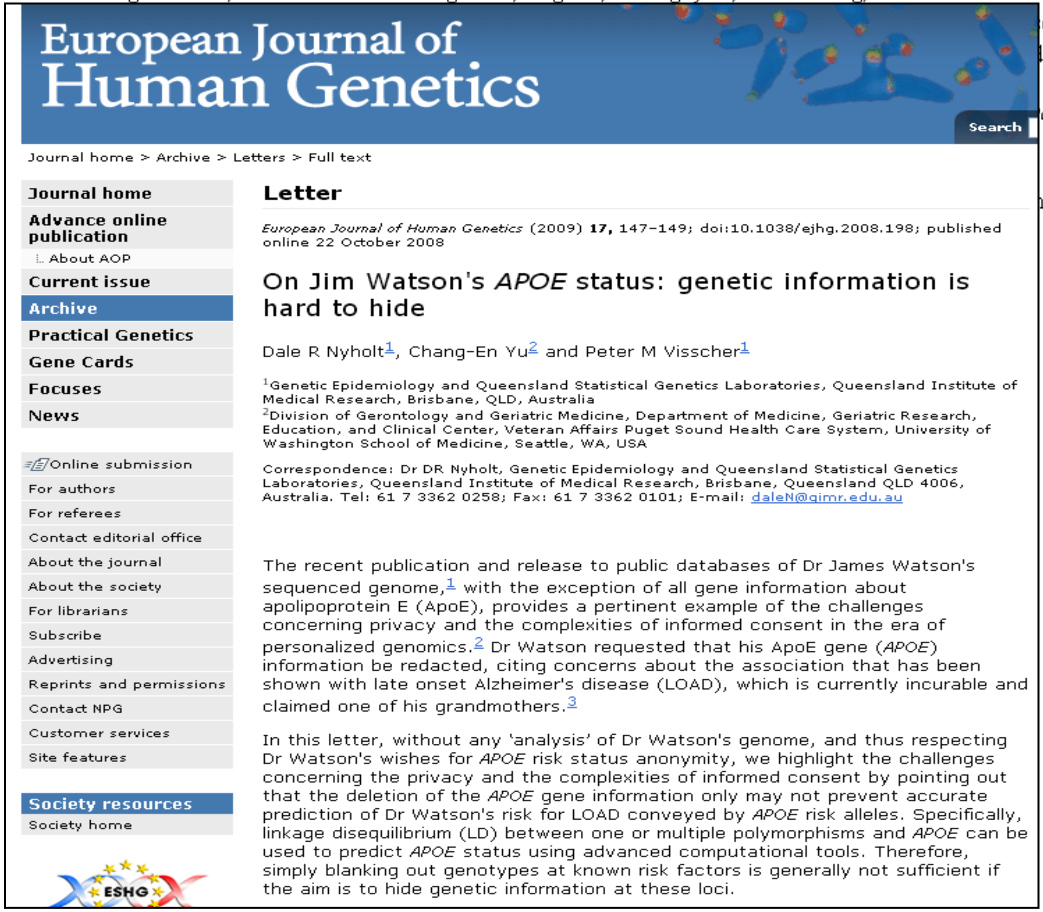
Dr. Watson's genome was sequenced at 6x coverage using 454 Life Sciences Technology. This means that each position on the genome was sequenced roughly six times. However, because of the probabilistic nature of the technology, some positions have been seen more than six times, and some less or not at all. The 454 technology produces short stretches of sequences called "reads" that are roughly 100 bases long. However, the functional units of the genome, the genes, are roughly 50,000 bases long, or 500 times the size of a 454 sequence. To interpret the Watson sequence, it was matched to the

equence, with the exception of the ApoE gene, variants of which are associated

l will be available from many other web sites in the future.

ariants or *polymorphisms*. Because each of these differences involves only a

n as *alleles*. Each SNP has two possible alleles.



Letter

European Journal of Human Genetics (2009) **17**, 147–149; doi:10.1038/ejhg.2008.198; published online 22 October 2008

On Jim Watson's *APOE* status: genetic information is hard to hide

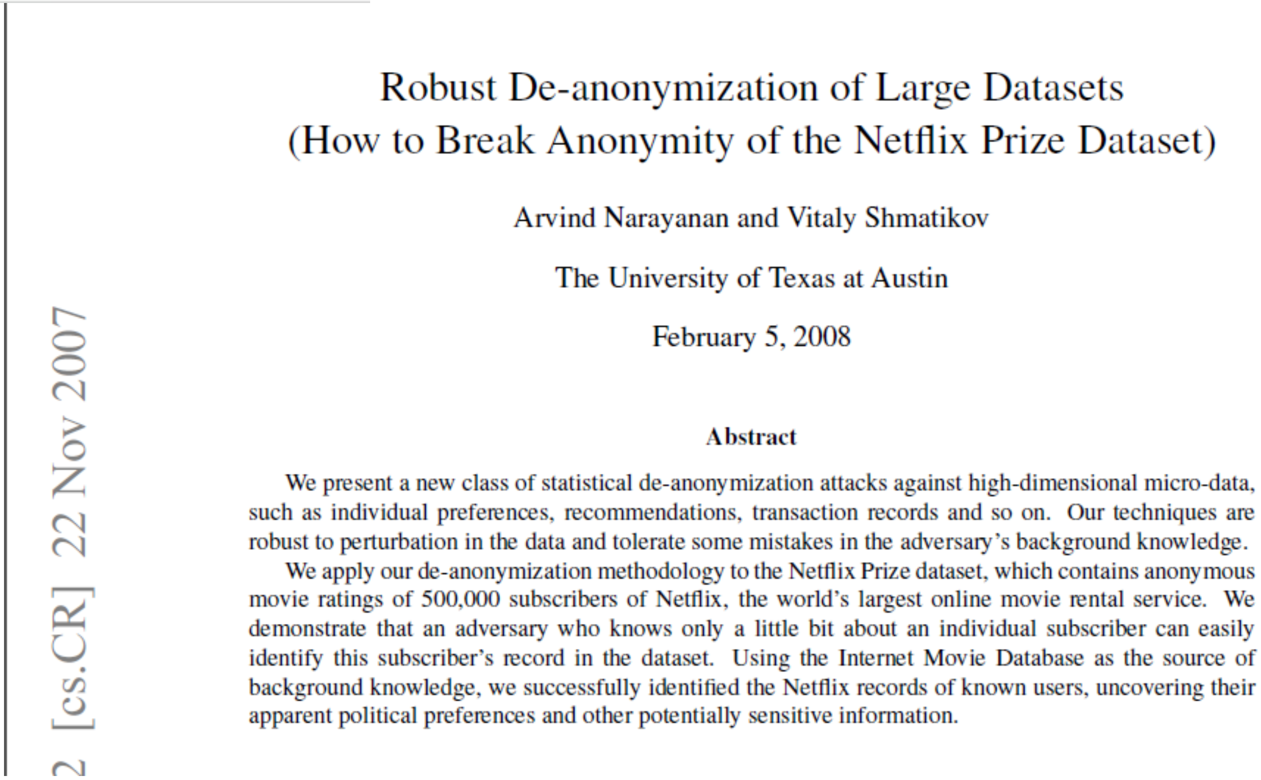
Dale R Nyholt[1], Chang-En Yu[2] and Peter M Visscher[1]

[1]Genetic Epidemiology and Queensland Statistical Genetics Laboratories, Queensland Institute of Medical Research, Brisbane, QLD, Australia
[2]Division of Gerontology and Geriatric Medicine, Department of Medicine, Geriatric Research, Education, and Clinical Center, Veteran Affairs Puget Sound Health Care System, University of Washington School of Medicine, Seattle, WA, USA

Correspondence: Dr DR Nyholt, Genetic Epidemiology and Queensland Statistical Genetics Laboratories, Queensland Institute of Medical Research, Brisbane, Queensland QLD 4006, Australia. Tel: 61 7 3362 0258; Fax: 61 7 3362 0101; E-mail: daleN@qimr.edu.au

The recent publication and release to public databases of Dr James Watson's sequenced genome,[1] with the exception of all gene information about apolipoprotein E (ApoE), provides a pertinent example of the challenges concerning privacy and the complexities of informed consent in the era of personalized genomics.[2] Dr Watson requested that his ApoE gene (*APOE*) information be redacted, citing concerns about the association that has been shown with late onset Alzheimer's disease (LOAD), which is currently incurable and claimed one of his grandmothers.[3]

In this letter, without any 'analysis' of Dr Watson's genome, and thus respecting Dr Watson's wishes for *APOE* risk status anonymity, we highlight the challenges concerning the privacy and the complexities of informed consent by pointing out that the deletion of the *APOE* gene information only may not prevent accurate prediction of Dr Watson's risk for LOAD conveyed by *APOE* risk alleles. Specifically, linkage disequilibrium (LD) between one or multiple polymorphisms and *APOE* can be used to predict *APOE* status using advanced computational tools. Therefore, simply blanking out genotypes at known risk factors is generally not sufficient if the aim is to hide genetic information at these loci.



Robust De-anonymization of Large Datasets (How to Break Anonymity of the Netflix Prize Dataset)

Arvind Narayanan and Vitaly Shmatikov

The University of Texas at Austin

February 5, 2008

Abstract

We present a new class of statistical de-anonymization attacks against high-dimensional micro-data, such as individual preferences, recommendations, transaction records and so on. Our techniques are robust to perturbation in the data and tolerate some mistakes in the adversary's background knowledge.

We apply our de-anonymization methodology to the Netflix Prize dataset, which contains anonymous movie ratings of 500,000 subscribers of Netflix, the world's largest online movie rental service. We demonstrate that an adversary who knows only a little bit about an individual subscriber can easily identify this subscriber's record in the dataset. Using the Internet Movie Database as the source of background knowledge, we successfully identified the Netflix records of known users, uncovering their apparent political preferences and other potentially sensitive information.

Cross correlated small set of identifiable IMDB movie database rating records with large set of "anonymized" Netflix customer ratings

Proposed Social & Technological Solutions to Issues of Data Privacy in Personal Genomics

(Licensure, Secondary Datasets, Enclaves)

- Setting the Stage: the Advent of Personal Genomics
- The of **Conundrum Genomic Privacy**
  - Fundamental, inherited info v need for data-sharing to enable research
- Current Social & Tech Approaches
  - GINA, Consents & **"Secure" use of dbGAP**
  - Ways the solutions have been **"hacked"**
- **Strawman Hybrid Soc-Tech Approach**
  - Soc: Licensure for genomic researchers
  - Enclaves in the cloud
  - Intelligent **data formats mostly splitting private & public info** (ex of MRF)

Strawman Hybrid Social & Tech Solution?

- Fundamentally, researchers have to keep genetic secrets
  - Genetic Licensure
- Technology to make it easier
  - Cloud computing & enclaves
  - Selection of stubb & "test pilot" datasets for benchmarking
  - Careful separation of private & public data
    - Lightweight, freely accessible secondary datasets
- Technological barriers shouldn't create a social incentive for “hacking”

[D Greenbaum, M Gerstein (‘11). Am J Bioeth 11:39. Greenbaum & Gerstein, The Scientist ('13)]

Example Format for Functional Genomics (Gene Expression Levels)

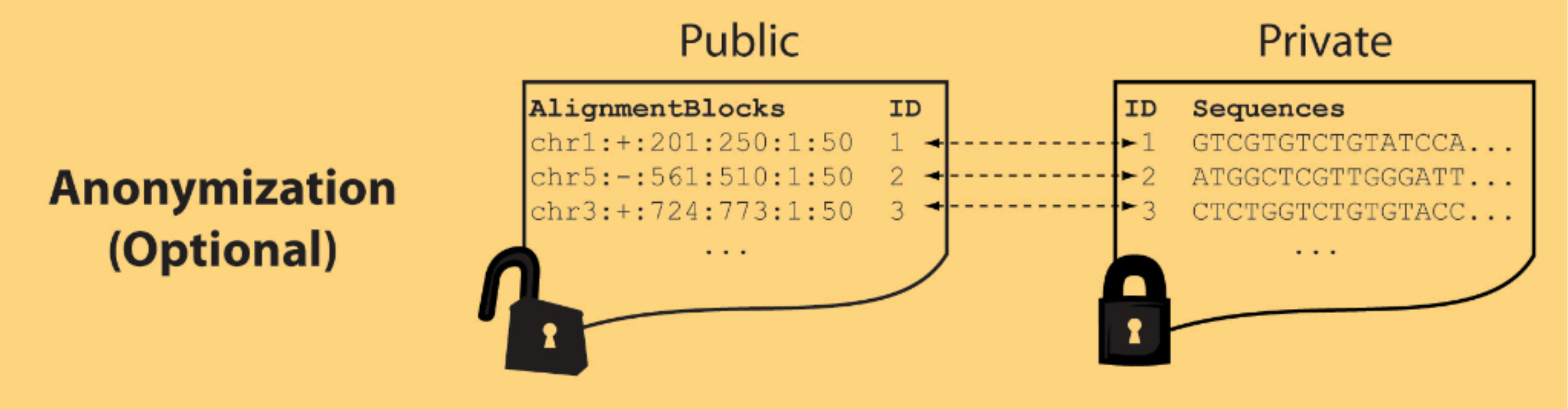
- Human genome reseq. all about variants vs. reference
- Situation diff. for func. genomics
  - Often variant info. Is determined incidentally

- On one hand: **Reads have variant information** in most functional regions (deep RNA-seq expt. essentially exome seq.)
- On other hand: **high-level summaries and signal tracks mostly what is used (80%) and do not involve variant info.** Helpful to make this freely available and easy to use

[Greenbaum et al. *PLOS CB* ('11)]

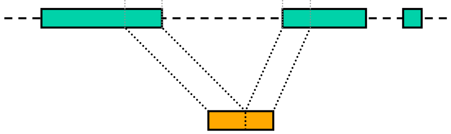
Light-weight formats

- Some lightweight format clearly separate public & private info., aiding exchange
- Files become much smaller
- Distinction between formats to compute on and those to archive with – become sharper with big data



AlignmentBlocks	ID
chr1:+:201:250:1:50	1
chr5:-:561:510:1:50	2
chr3:+:724:773:1:50	3
...	

ID	Sequences
1	GTCGTGTCTGTATCCA...
2	ATGGCTCGTTGGGATT...
3	CTCTGGTCTGTGTACC...
	...



Mapping coordinates without variants (MRF)

Reads (linked via ID, 10X larger than mapping coord.)

[*Bioinformatics* 27: 281]

MRF Examples

10X Compression Ex.

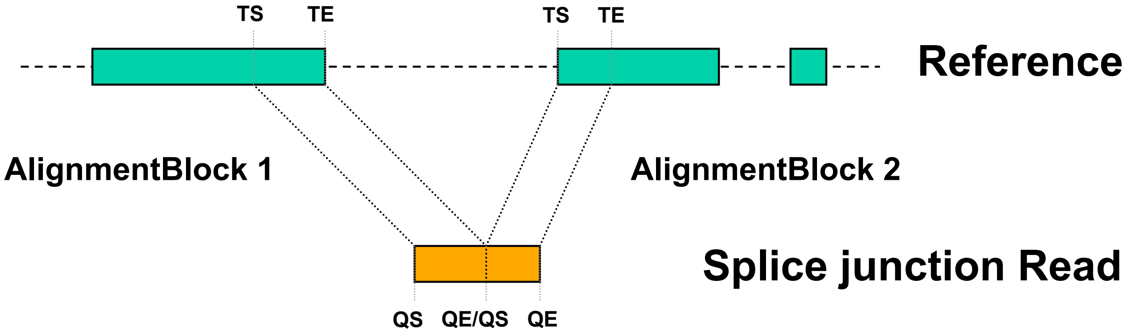
Raw ELAND export file has uncompressed file size: ~4 GB; total number of reads: ~20 million; number of mapped reads: ~12 million .

MRF file is significantly smaller (~400 MB uncompressed, ~130 MB compressed with gzip).

BAM file
has a size of ~1.2 GB.

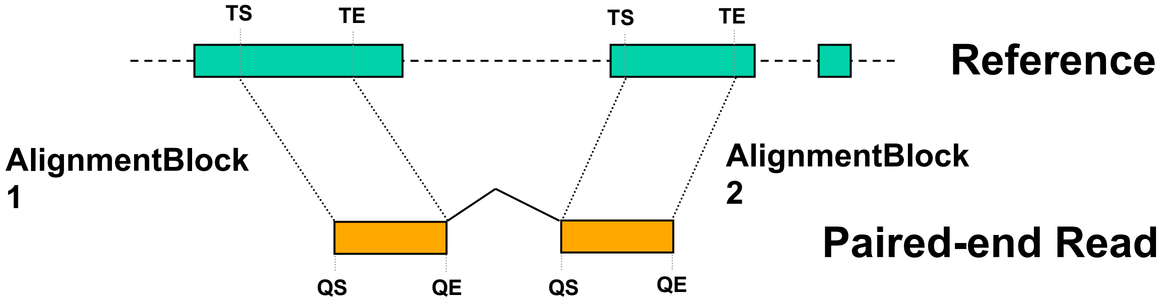
Reference based compression (ie CRAM) is similar but it stores actual variant beyond just position of alignment block

`chr2:+:601:630:1:30,chr2:+:921:940:31:50`



Legend: TS = TargetStart, TE = TargetEnd, QS = QueryStart, QE = QueryEnd

`chr9:+:431:480:1:50|chr9:+:945:994:1:50`



Legend: TS = TargetStart, TE = TargetEnd, QS = QueryStart, QE = QueryEnd

[Habegger et al., Bioinformatics ('11)]

Proposed Social & Technological Solutions to Issues of Data Privacy in Personal Genomics

(Licensure, Secondary Datasets, Enclaves)

- Setting the Stage: the Advent of Personal Genomics
- The of **Conundrum Genomic Privacy**
  - Fundamental, inherited info v need for data-sharing to enable research
- Current Social & Tech Approaches
  - GINA, Consents & **"Secure" use of dbGAP**
  - Ways the solutions have been **"hacked"**
- **Strawman Hybrid Soc-Tech Approach**
  - Soc: Licensure for genomic researchers
  - Enclaves in the cloud
  - Intelligent **data formats mostly splitting private & public info** (ex of MRF)

Acknowledgements

papers.gersteinlab.org/subject/privacy

D Greenbaum

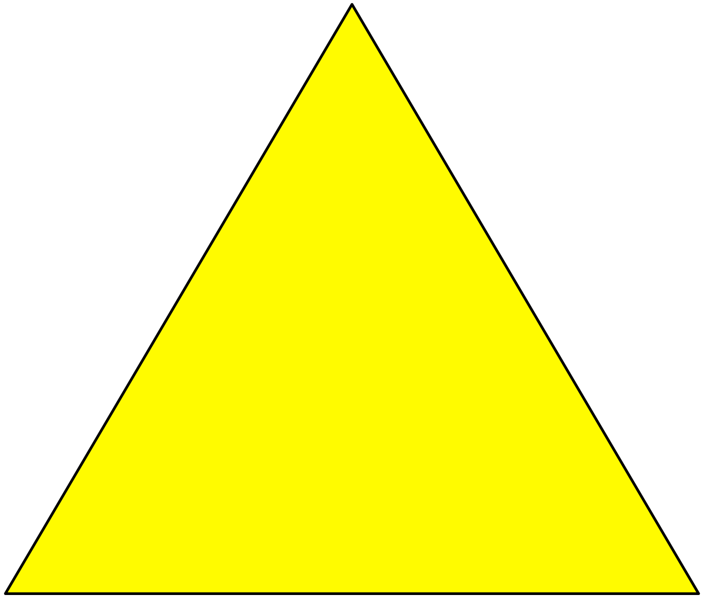
A Harmanci

A Sboner, XJ Mu

L Habegger, A Sboner, TA Gianoulis,
J Rozowsky, A Agarwal, M Snyder

Default Theme

- Default Outline Level 1
  - Level 2



More Information on this Talk

SUBJECT: Networks

DESCRIPTION:

NOTES:
This PPT should work on mac & PC. Paper references in the talk were mostly from Papers.GersteinLab.org.

PHOTOS & IMAGES. For thoughts on the source and permissions of many of the photos and clipped images in this presentation see http://streams.gerstein.info . In particular, many of the images have particular EXIF tags, such as kwpotppt , that can be easily queried from flickr, viz: http://www.flickr.com/photos/mbgmbg/tags/kwpotppt .