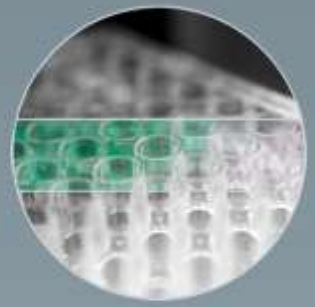
Security in a Distributed Network Environment

Kenneth R. Evans, PhD

CEO, OCBN;

Associate Professor, Dept of Pathology and Laboratory Medicine, Queen's University

Ontario Cancer Biomarker Network



High Dimensional Medical Databases

- The opportunities are enormous…

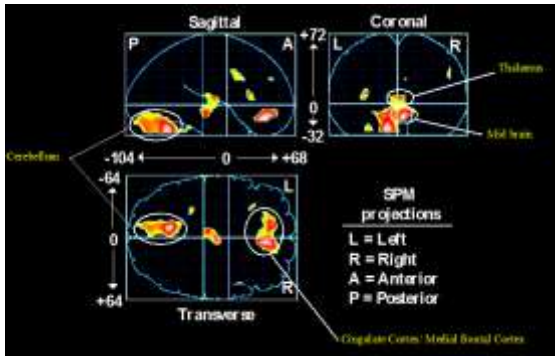
"Depressive Symptoms" Example

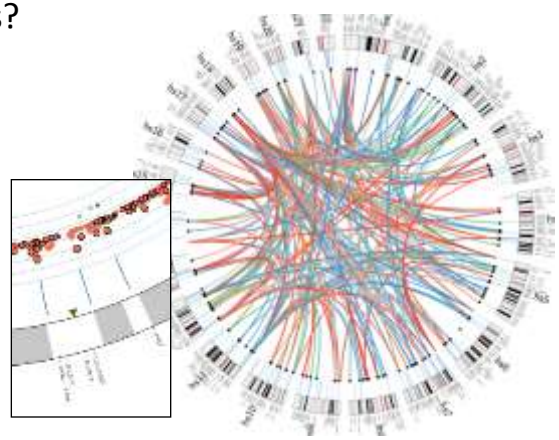
Autism | Depression | Neuro-degeneration | Cerebral Palsy | Addictions | Epilepsy | Traumatic Brain Injury

Common Symptoms?

"Worthlessness, Sadness, Helplessness"

Common mechanisms?





Personalized medicine

- Changing what we mean by “disease”
- The opportunities are enormous…
- So, too are the challenges

Security and Privacy

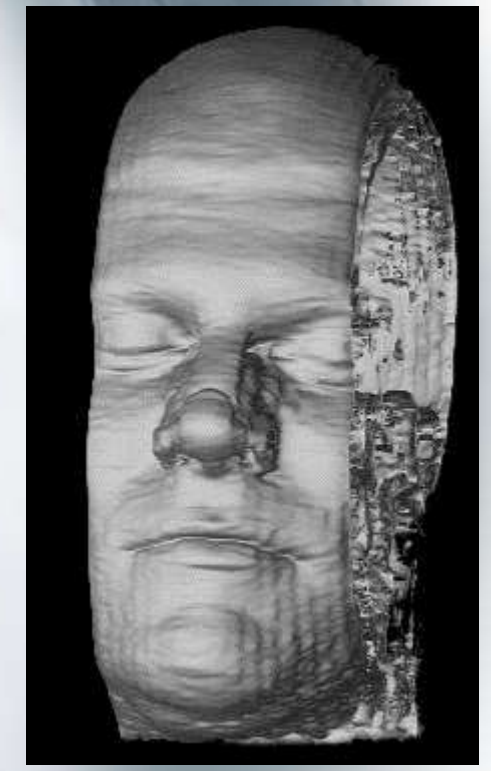
- Privacy is an enormous concern in medicine
  - But the full extent of the issue is only just beginning to be understood.
- Personal Health Information:
  - “Information in any format that identifies the individual, including demographic information collected from an individual that can reasonably be used to identify the individual.
  - Additionally, PHI is information created or received by a health care provider, health plan, employer, or health care clearinghouse; and relates to the past, present, or future physical or mental health or condition of an individual.”

HIPAA Guidance

US Health Insurance Portability and Accountability Act (HIPAA): 18 “identifiers”

1. Name
2. Telephone numbers
3. Fax numbers
4. Electronic mail addresses
5. Social security numbers
6. Medical record numbers
7. Health plan beneficiary numbers
8. Account numbers
9. Certificate / license numbers
10. Vehicle identifiers and serial numbers, including license plate numbers
11. Device identifiers and serial numbers
12. Web Universal Resource Locators (URLs)
13. Internet Protocol (IP) address numbers
14. Location; all geographic subdivisions smaller than a state, including street address, city, county, precinct, zip code, and their equivalent geocodes.
15. Dates (all dates related to the subject of the information, e.g. birth dates, admission dates, discharge dates, encounter dates, surgery dates, etc.)
16. Biometric identifiers, including finger and voice prints
17. Full face photographic images and any comparable images
18. Any other unique identifying number, characteristic, or code

Brain Image Data

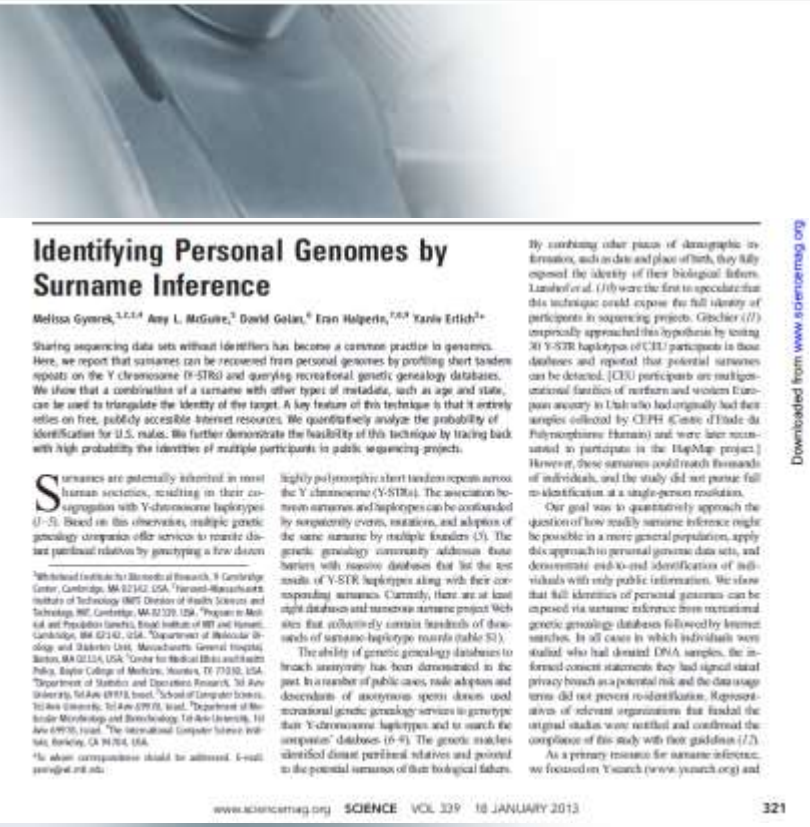


- It is possible to reassemble an image of an individual from an MRI scan
- Studies have shown 30 – 70% success rates in identifying the individual from photographs
- Full face photographic images and ***any comparable images*** are PHI

...and evermore shall be so?

- The concept of PHI is evolving

Erlich and Gymrek, 2013



Identifying Personal Genomes by Surname Inference

Melissa Gymrek,[1,2,3,4] Amy L. McGuire,[5] David Golan,[6] Eran Halperin,[7,8,9] Yaniv Erlich[2]*

Sharing sequencing data sets without identifiers has become a common practice in genomics. Here, we report that surnames can be recovered from personal genomes by profiling short tandem repeats on the Y chromosome (Y-STRs) and querying recreational genetic genealogy databases. We show that a combination of a surname with other types of metadata, such as age and state, can be used to triangulate the identity of the target. A key feature of this technique is that it entirely relies on free, publicly accessible Internet resources. We quantitatively analyze the probability of identification for U.S. males. We further demonstrate the feasibility of this technique by tracing back with high probability the identities of multiple participants in public sequencing projects.

Surnames are paternally inherited in most human societies, resulting in their co-segregation with Y-chromosome haplotypes (1–5). Based on this observation, multiple genetic genealogy companies offer services to reunite distant patrilineal relatives by genotyping a few dozen highly polymorphic short tandem repeats across the Y chromosome (Y-STRs). The association between surnames and haplotypes can be confounded by nonpaternity events, mutations, and adoption of the same surname by multiple founders (3). The genetic genealogy community addresses these barriers with massive databases that list the test results of Y-STR haplotypes along with their corresponding surnames. Currently, there are at least eight databases and numerous surname project Web sites that collectively contain hundreds of thousands of surname-haplotype records (table S1).

The ability of genetic genealogy databases to breach anonymity has been demonstrated in the past. In a number of public cases, male adoptees and descendants of anonymous sperm donors used recreational genetic genealogy services to genotype their Y-chromosome haplotypes and to search the companies' databases (6–9). The genetic matches identified distant patrilineal relatives and pointed to the potential surnames of their biological fathers. By combining other pieces of demographic information, such as date and place of birth, they fully exposed the identity of their biological fathers. Lunshof *et al.* (10) were the first to speculate that this technique could expose the full identity of participants in sequencing projects. Gitschier (11) empirically approached this hypothesis by testing 30 Y-STR haplotypes of CEU participants in those databases and reported that potential surnames can be detected. [CEU participants are multigenerational families of northern and western European ancestry in Utah who had originally had their samples collected by CEPH (Centre d'Etude du Polymorphisme Humain) and were later reconsented to participate in the HapMap project.] However, these surnames could match thousands of individuals, and the study did not pursue full re-identification at a single-person resolution.

Our goal was to quantitatively approach the question of how readily surname inference might be possible in a more general population, apply this approach to personal genome data sets, and demonstrate end-to-end identification of individuals with only public information. We show that full identities of personal genomes can be exposed via surname inference from recreational genetic genealogy databases followed by Internet searches. In all cases in which individuals were studied who had donated DNA samples, the informed consent statements they had signed stated privacy breach as a potential risk and the data usage terms did not prevent re-identification. Representatives of relevant organizations that funded the original studies were notified and confirmed the compliance of this study with their guidelines (12).

As a primary resource for surname inference, we focused on Ysearch (www.ysearch.org) and

[1]Whitehead Institute for Biomedical Research, 9 Cambridge Center, Cambridge, MA 02142, USA. [2]Harvard-Massachusetts Institute of Technology (MIT) Division of Health Sciences and Technology, MIT, Cambridge, MA 02139, USA. [3]Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA. [4]Department of Molecular Biology and Diabetes Unit, Massachusetts General Hospital, Boston, MA 02114, USA. [5]Center for Medical Ethics and Health Policy, Baylor College of Medicine, Houston, TX 77030, USA. [6]Department of Statistics and Operations Research, Tel Aviv University, Tel Aviv 69978, Israel. [7]School of Computer Science, Tel Aviv University, Tel Aviv 69978, Israel. [8]Department of Molecular Microbiology and Biotechnology, Tel Aviv University, Tel Aviv 69978, Israel. [9]The International Computer Science Institute, Berkeley, CA 94704, USA.

*To whom correspondence should be addressed. E-mail: yaniv@wi.mit.edu

Downloaded from www.sciencemag.org

- Accessed publicly available information from the 1000 Genomes project, as well as geneology sites
- Successfully identified 50 individuals
- This is the reality: if we allow even “de-identified” data (ie without personal identifiers) into an open access environment, it will be possible to re-identify them.

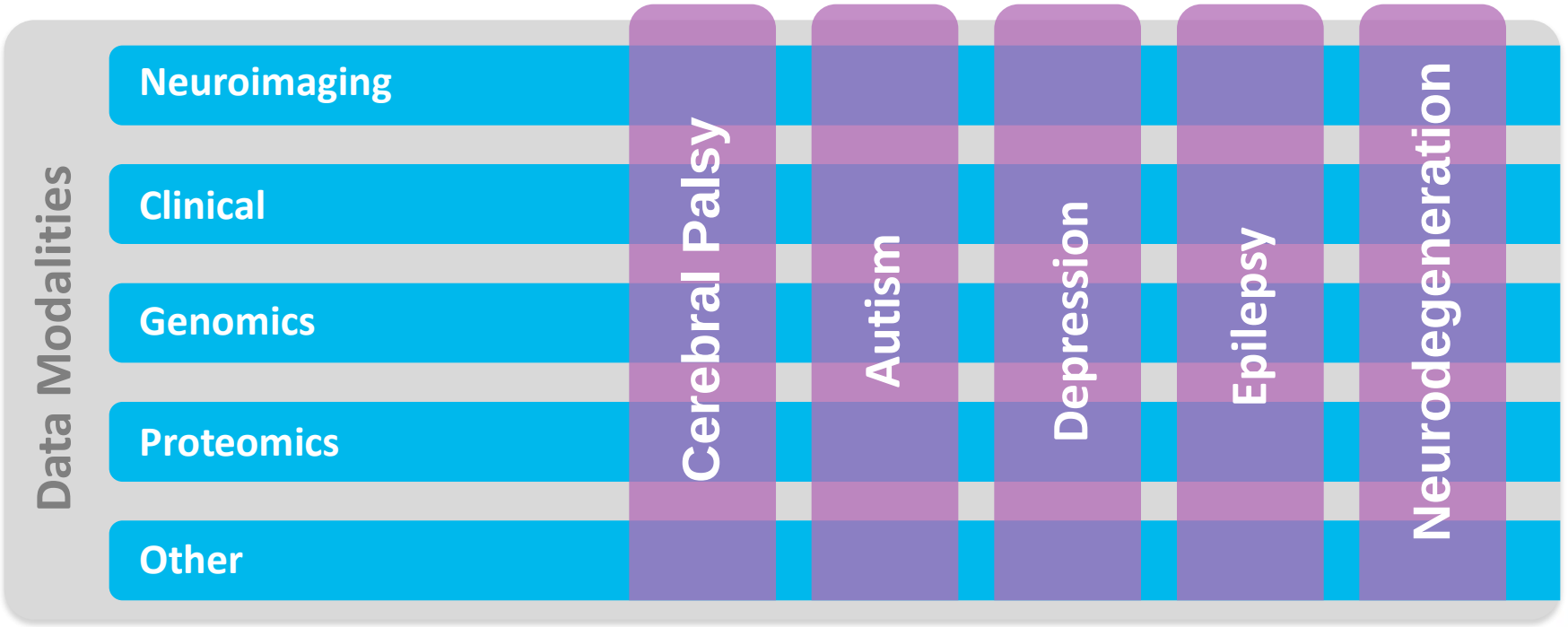
Building a medical database

- Even with consent, is open access medical data a good idea?
  - … or even ethical?
- Should we build a database under the assumption that the rules will remain as they are now?
- When OBI undertook the building of Brain-CODE these questions were foremost in all of our minds

ONTARIO BRAIN INSTITUTE

OBI's Big Data Needs

Integrated Discovery System



» Disease-themed, multidisciplinary collaborations collecting multiple modalities of data

What Brain-CODE means:

- To researchers
  - Collect, share and analyze ‘big’ data from anywhere, across multiple technologies, data sources, and diseases
  - Augment value of patient’s research data through integration with accrued health information (e.g., ICES, OHS)
- To patients
  - Engagement in the research process
  - Education about their diseases, treatment options, and ongoing research
- To the economy
  - Accelerated development of novel treatments and diagnostics
  - Amplified government investment in CNS R&D

Brain-CODE

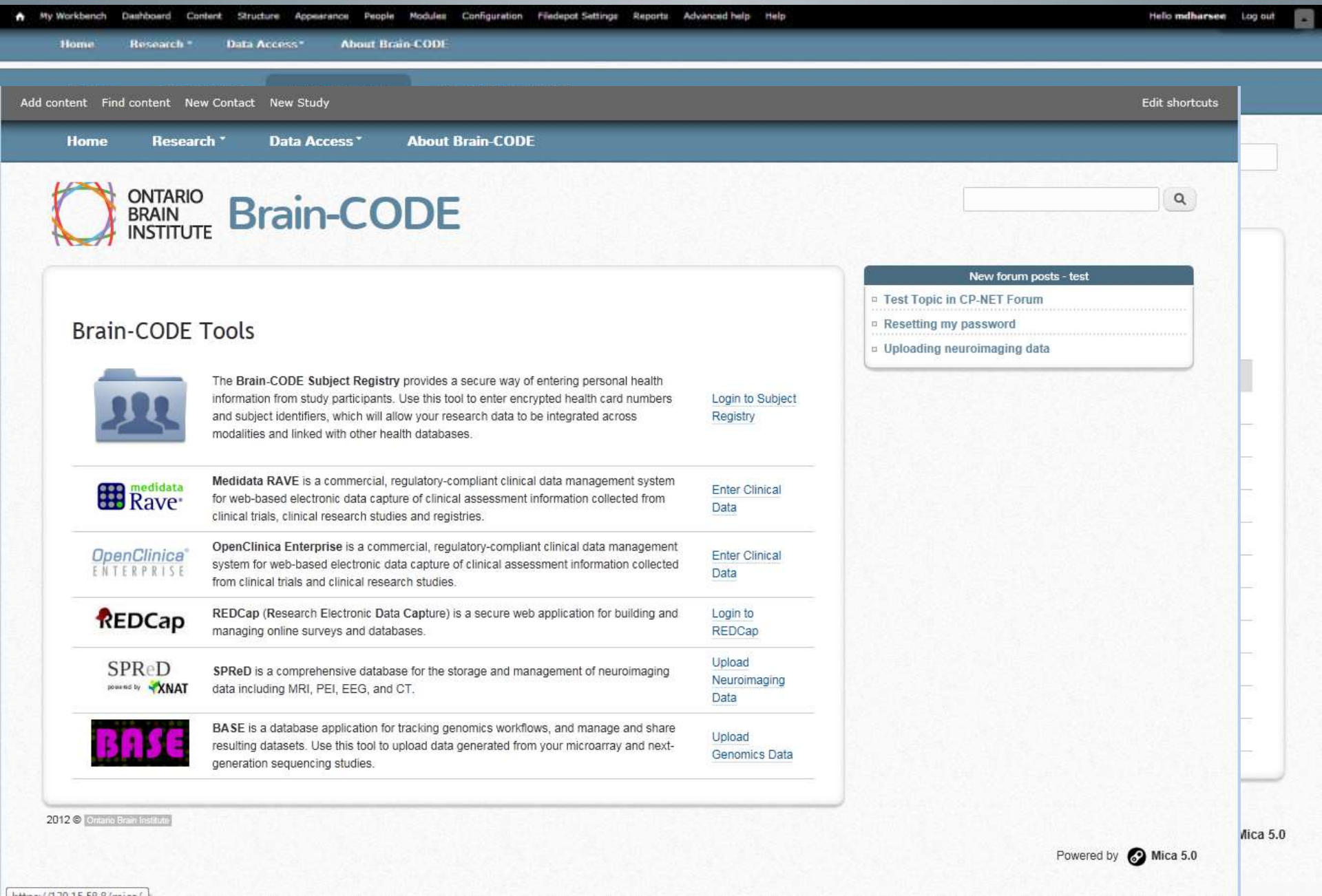
- *Leverages existing infrastructure and expertise wherever possible*
- "InDOC" Consortium:☐

Group	Lead	Expertise
Ontario Cancer Biomarker Network (OCBN)	Dr. Kenneth Evans (Exec Dir)	Molecular biomarker and clinical specimen data management; Overall leadership of BrainCODE
Rotman Research Institute	Dr. Stephen Strother	Brain imaging data management
Applied Health Research Centre (AHRC) at St. Michael's Hospital	Dr. Muhammad Mamdani	Clinical trials infrastructure
High Performance Computing Virtual Laboratory (HPCVL) in Kingston	Dr. Ken Edgecombe	Secure, regulatory-compliant computing and large scale data storage
E-Health Information Laboratory at the Children's Hospital of Eastern Ontario (CHEO)	Dr. Khaled El-Emam	Tools for de-identification, encryption, secure linkage, secure computation

Privacy and Security

- The OBI and Brain-CODE *adhere to the highest levels of data privacy and security*:
  - **Security:** HPCVL, controlled access, security of data transfer and storage higher than bank-level
  - **Encryption**: Sophisticated algorithms for record linkage
  - **Privacy systems**: Full PIA completed in collaboration with Ontario Privacy Commissioner: designated "*Privacy by Design Ambassador*" by same
  - **Governance**: Extensive policy framework and governance apparatus ensures ongoing compliance with regulatory and REB requirements

Brain-CODE Portal



Designed to seamlessly link with autonomous databases

Patient enrols in study

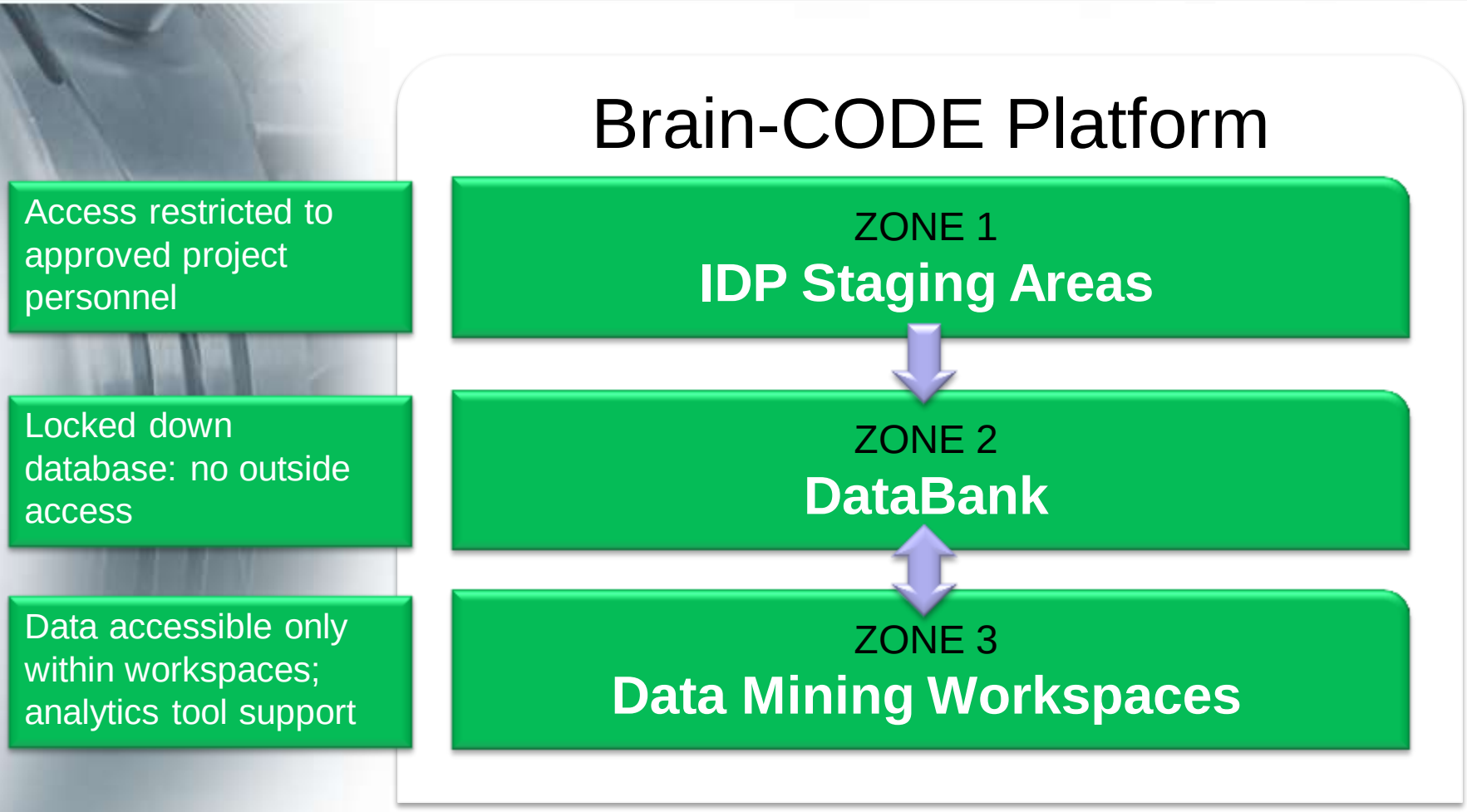
Patient contributes data to other study

Mathematical properties of encrypted numbers allow them to be identified as coming from same person, without ever de-encrypting the health card number itself

Encryption issues

- Some efforts to encrypt even highly complex data (e.g. Imaging, sequencing)
- Perhaps a solution to security issues, but what happens when you need to analyze the data?
- For many large data sets you have another issue: they're too large for widespread downloading to be practical
- One solution:
  - Allow access to data for mining, but within a secure environment
  - Only allow results to be extracted, not data
  - Brain-CODE approach, but also being proposed by others (e.g. NCI)

Secure, Comprehensive, High Performance Data Management Platform



Brain-CODE Team

Executive Management

Ken Evans
Muhammad Mamdani
Stephen Strother
Ken Edgecombe
Khaled El-Emam

Project Management

Anthony Vaccarino
Moyez Dharsee
Barbara McQuaid

High Performance Computing

Chris MacPhee
Costa Dafnas
Michael Hanlan

Data Management

Rino La Grassa
Natascha Kozlowski
Anthony Vaccarino
Stephen Strother
Tom Gee
Paulo Nuin

Software Development

Tom Gee
Nima Nourhaghighi
Fan Dong
Fletcher Johnson
Paulo Nuin
Rashad Badrawi
Rino La Grassa
Chris Ducharme
Ben Eze
Aleks Essex

Common Data Elements

Anthony Vaccarino
Stephen Strother
Tom Gee
Natascha Kozlowski
Moyez Dharsee