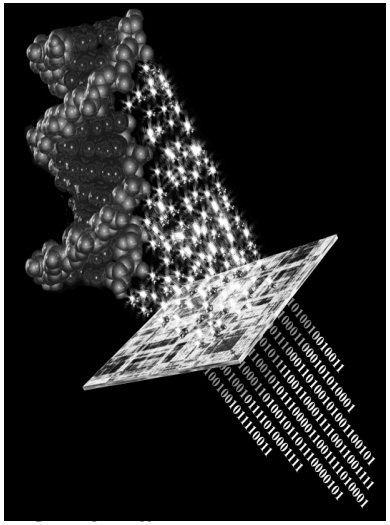
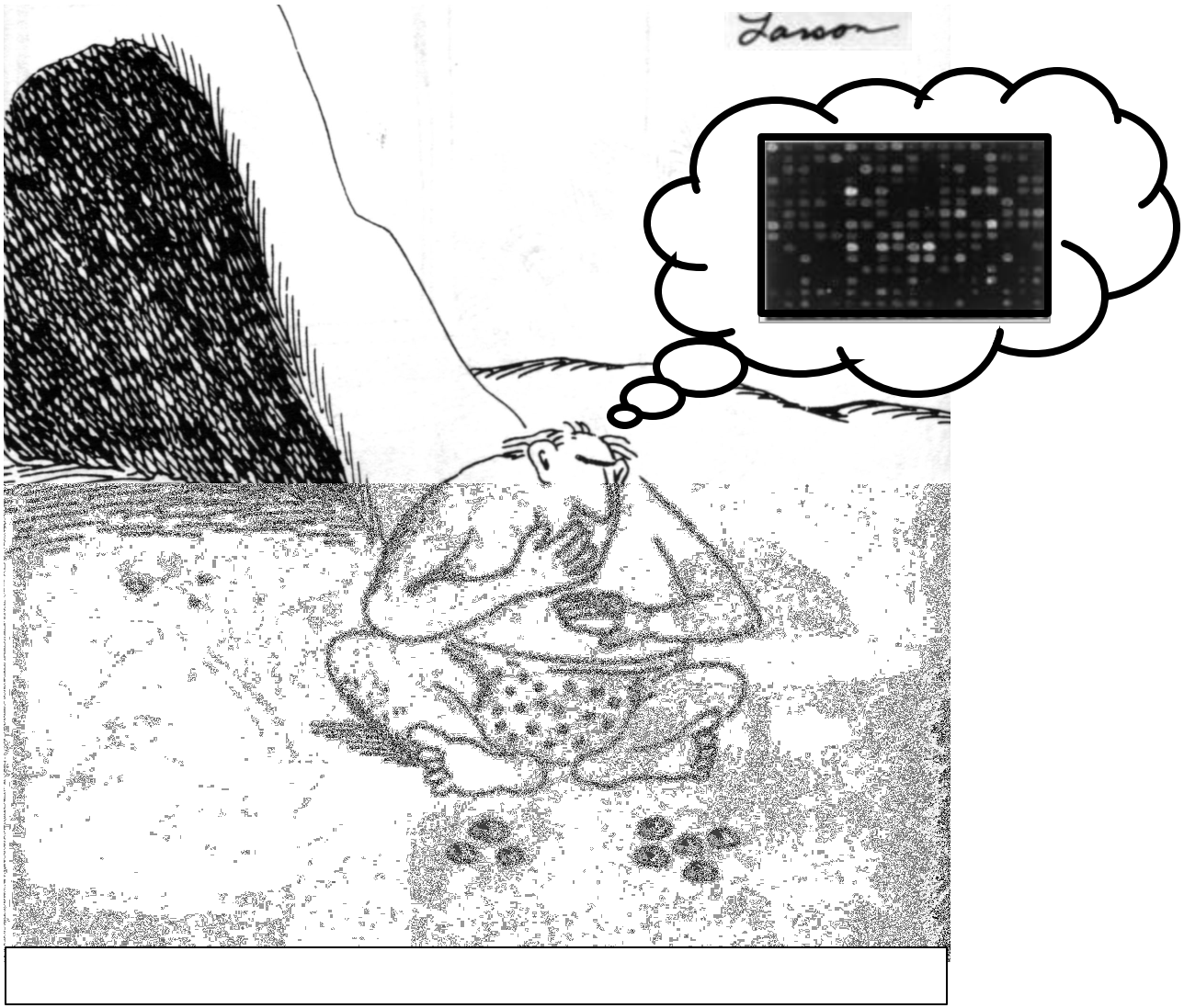
The NCI-60 Pilot and caBIG

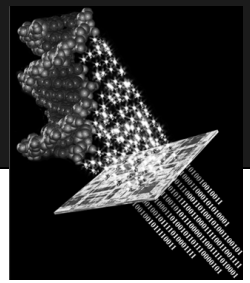


John N. Weinstein, NCI
Ari Kahn, SRA International
David Kane, SRA International

The Genomics and Bioinformatics Group
LMP, CCR, NCI

Analyzing and Integrating Omic Data



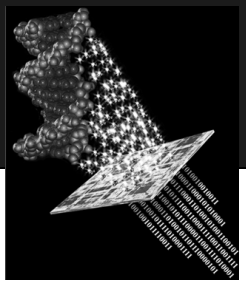


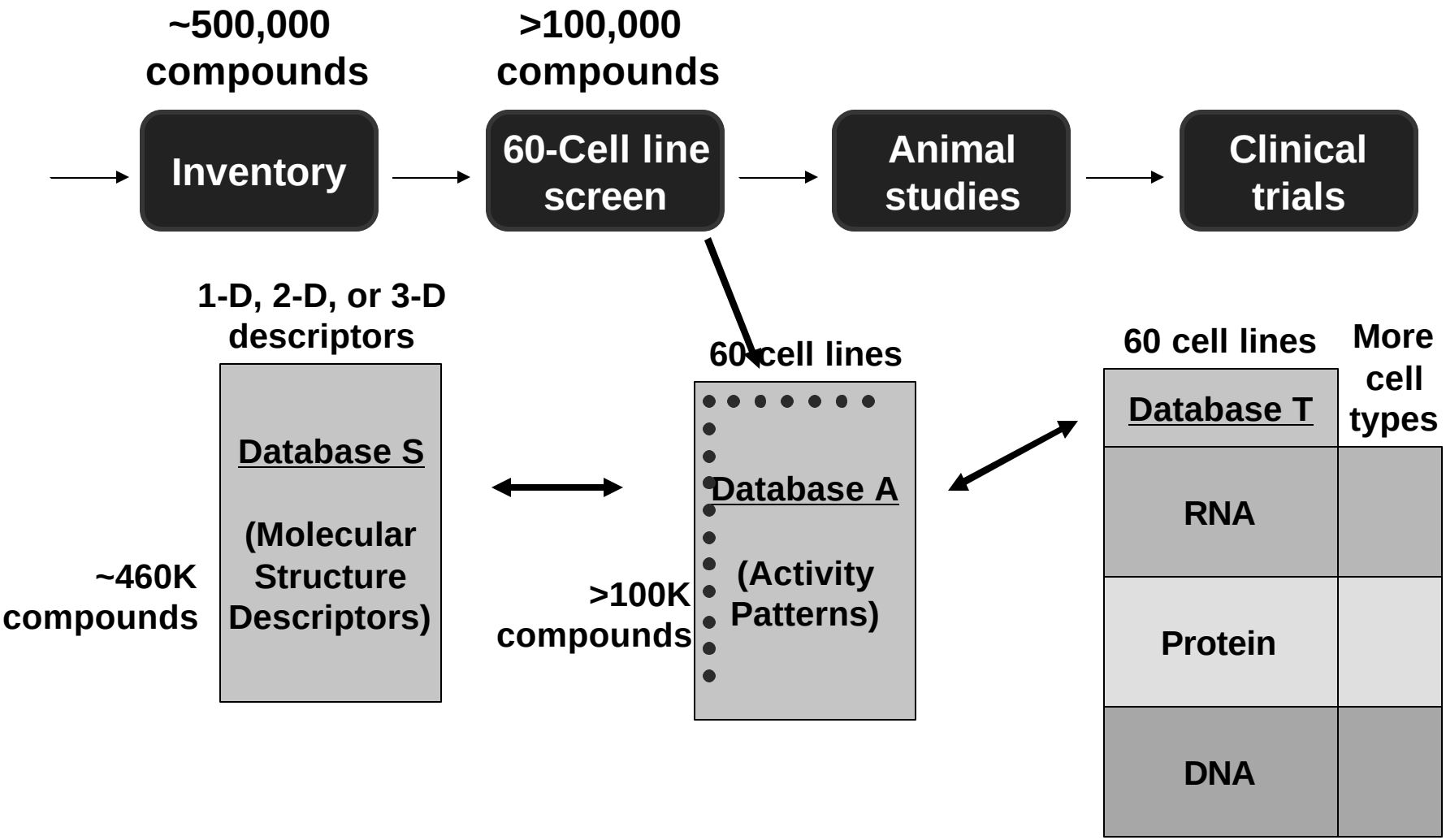
Flow in one portion of NCI Drug Discovery Program

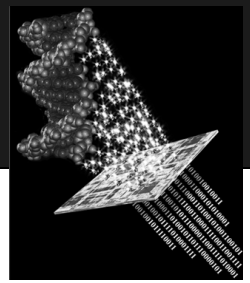
~500,000 compounds → **Inventory** → ~100,000 compounds → **60-Cell line screen** → **Animal studies** → **Clinical trials**

- Breast
- Prostate
- Melanoma
- Leukemia
- Lung (NSCL)
- Colon
- Ovarian
- Renal
- CNS

Flow in one portion of the NCI-DTP Drug Discovery Program







Molecular Databases: The NCI-60 Cancer Cells

Pharmacological
- >100,000 compounds

Misc. "Targets"
- Oncogenes
- Tumor Suppressors
- Enzymes
- Transporters
- P53-related

Tissue Arrays (with Raffeld, Hewitt, Liotta)

mRNA

- cDNA Arrays (9K) (with Brown/Botstein)
- Affy Chips (6.8K) (with Lander/Golub)
- Affy Chips (U95 and U133)(with Scherf)
- Transporters (with Gottesman, Sadee)

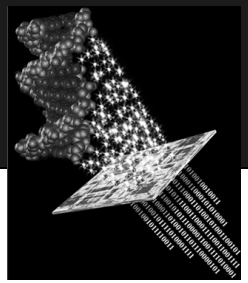
Protein

- 2-D Gels (with Anderson)
  - -Mass Spec. I.D.
- Lysate Arrays (700) (with Liotta/Petricoin)
- Cell Arrays (with Raffeld, Hewitt, Liotta)

DNA

- SNPs (with Buetow)
- SNPs(with Sanger Center)
- SKY (with Kirsch)
- CGH (with Kirsch)
- aCGH (with Gray/Pinkel)
- CpG Methylation (with Feinberg, Munroe)

Flow in one portion of NCI Drug Discovery Program



~500,000 compounds

>100,000 compounds

Inventory → 60-Cell line screen → Animal studies → Clinical trials

1-D, 2-D, or 3-D descriptors

Cells

Cells

Database S

(Molecular Structure Descriptors)

~460K compounds

270,000,000 structural elements

>100K compounds

Database A

(Activity Patterns)

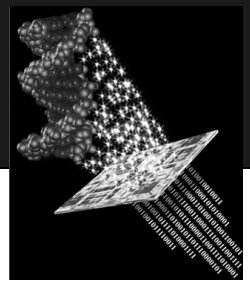
3,600,000 potencies

genes, proteins

Database T

(Molecular Targets)

6,000 target data points



Freely Available Bioinformatic Resources from the Genomics & Bioinformatics Group, LMP, CCR, NCI

(Head: John N. Weinstein, M.D., Ph.D.; jw4i@nih.gov)

http://discover.nci.nih.gov

MatchMiner: **translates among gene identifier types for lists of hundreds or thousands of genes. Includes: GenBank accession numbers, IMAGE clone IDs, common gene names, HUGO names, gene symbols, UniGene clusters, FISH-mapped BAC clones, Affymetrix identifiers, and chromosome locations. (Bussey, et al., Genome Biology 2003; 4:R27)**

GoMiner: **addresses the question, "Now that I've done the gene expression experiment and identified a set of 'interesting' genes, what do those genes mean biologically?" GoMiner batch-processes and organizes lists of thousands or tens of thousands of genes and provides two fluent, robust visualizations of the genes in the framework of the Gene Ontology hierarchy. (Zeeberg, et al., Genome Biology 2003; 4:R28)**

CIMMaker: **generates color-coded Clustered Image Maps (CIMs) ("heat maps") to represent "high-dimensional" data sets such as gene expression profiles. We introduced CIMs for data on drug activity, target expression, gene expression, and proteomics. Clustering axes brings like together with like to create patterns of color. (Weinstein, et al., Science 1997; 275:343-349).**

MedMiner: **searches and organizes the biomedical literature on genes, gene-gene relationships, and gene-drug relationships. It uses GeneCards, PubMed, syntactic analysis, truncated-keyword filtering of relationals, and user-controlled sculpting of Boolean queries to generate key sentences from pertinent abstracts. Abstracts selected can be automatically entered into EndNote. MedMiner speeds up 5-fold the capture and organization of literature from PubMed searches. (Tanabe, et al., BioTechniques 1999; 27: 1210.)**

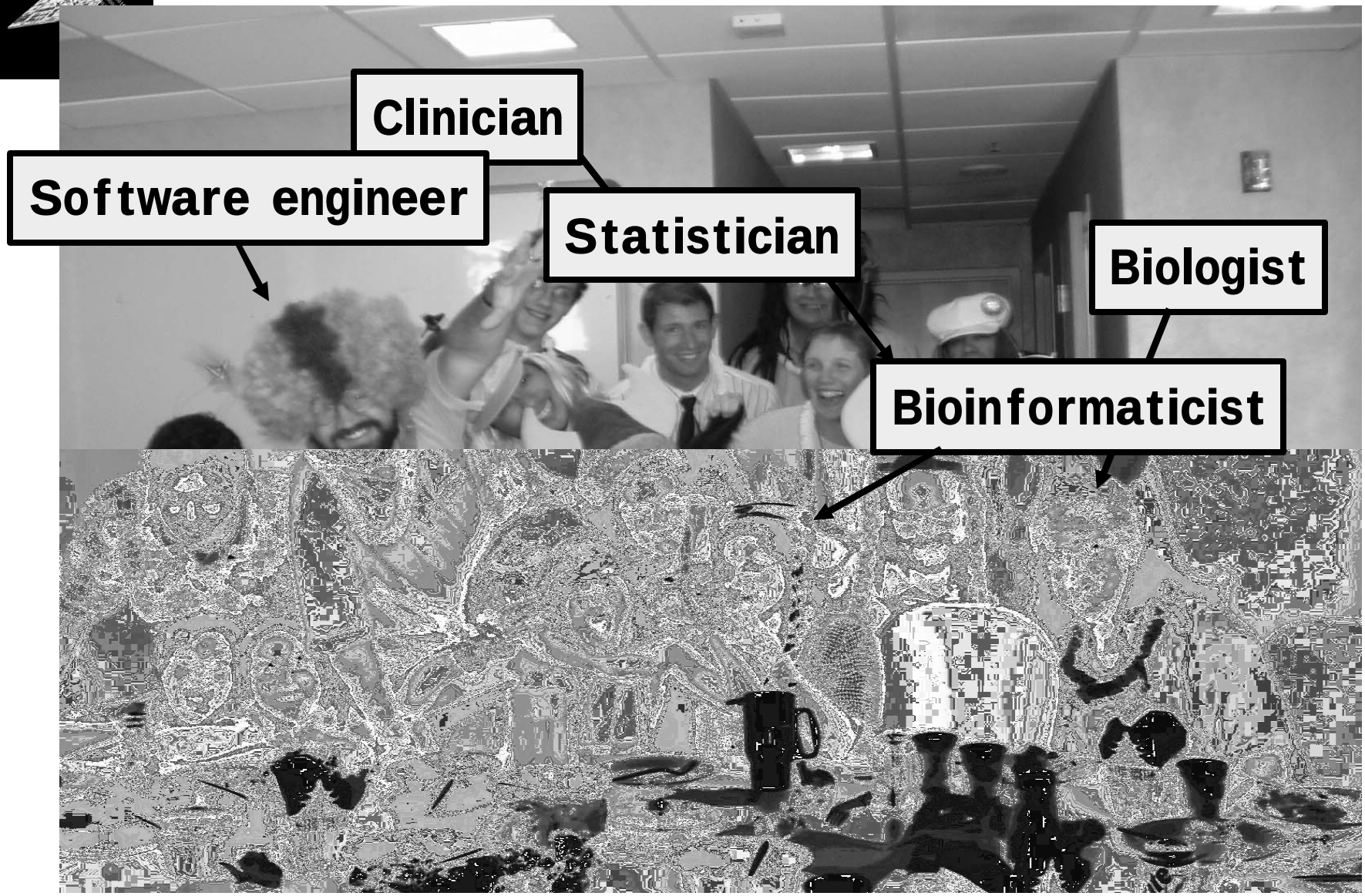
AbMiner: **Relational database of commercially available antibodies and quality-control information.**

LeadMiner: **links molecular markers and the drug discovery process. It links gene expression profiles the NCI-60 (or other screening cell panel) to a set of 27,000 chemical substructure descriptors of compounds tested against the cells. It can be used to identify substructure classes correlated with activity against cell types that express large amounts of a particular gene. (Blower, et al., Nature's Pharmacogenomics J. 2002; 2:259) Available in full-capability version for the NCI-60.**

NCI-60 cancer cell databases: **includes gene expression data from 9,700-gene cDNA array & 6,800-gene Affymetrix oligonucleotide array.**

Software development spearheaded by D. Kane, et al., SRA Intl.

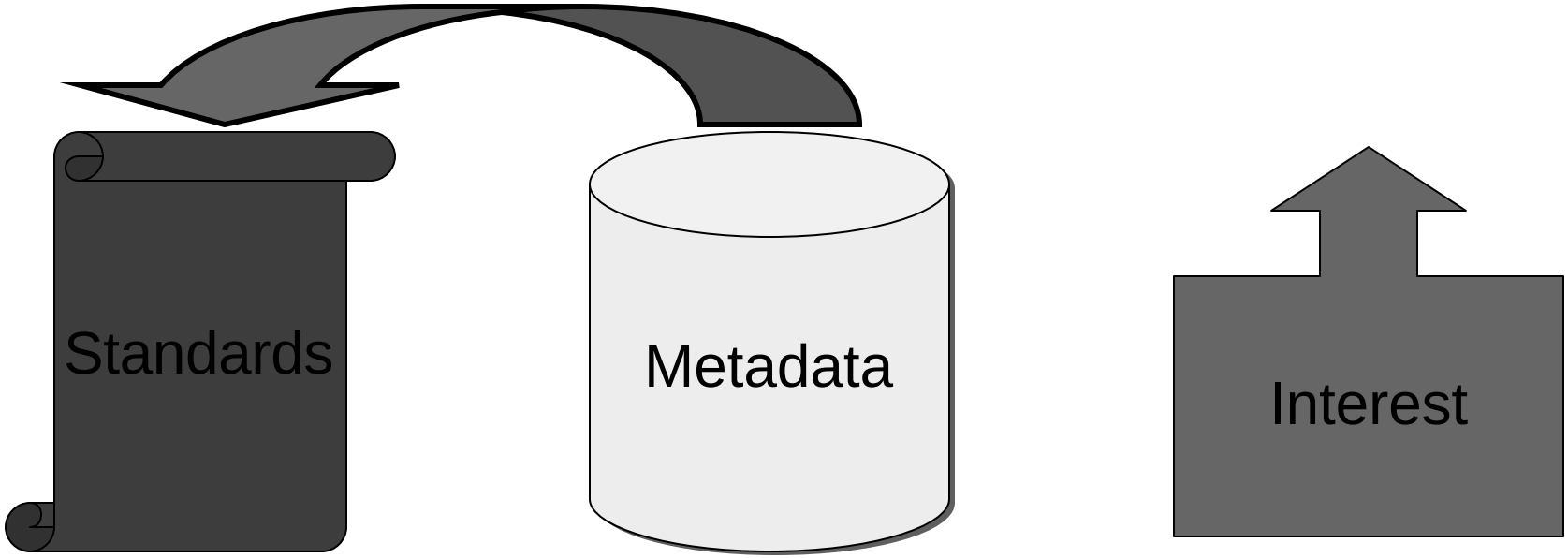
Genomics & Bioinformatics Group Lab Meeting



ICR Goals

- **Goals of caBIG Integrative Cancer Research Workspace**
  - **"Create a common, extensible informatics platform that integrates diverse data types and supports interoperable analytic tools"**
- **NCI-60 Pilot and the These Goals**
  - **Validation of approach**
  - **Validate the existing tools and standards**
    - **Identify gaps in each**
  - **Data available to researchers**
- **NCI-60 Adopters**
  - **Memorial Sloan-Kettering Cancer Center**
  - **Alex Lash**
  - **http://www.mskcc.org**

NCI-60 Pilot Selection Criteria



- MIAME
- MIAMEtox
- SKYWEB

Data Standards and Hosts

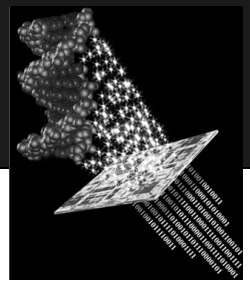
Data Source	Target Standards	Target caBIG Host	Other Target Host
U95 and U133	MIAME (MAGE-ML)	caArray	mAdb, MIAMExpress
DTP Drug Response	MIAME-tox (MAGEtox-ML)	Local	NCT-CEBS
cDNA	MIAME (MAGE-ML)	caArray	mAdb, MIAMExpress
SKY	SkyWeb	Local	SKY DB
BAC array CGH	MIAME (MAGE-ML)	caArray	Local, SKY DB

U133, U95, cDNA

- **MIAME and MAGE-ML**
  - **Annotation**
    - **Archeology**
    - **EVS**
  - **Create MAGE-ML**
    - **caArray**
    - **MAGEstk (MGED)**
    - **GCOS (Affymetrix)**
  - **Validation**
    - **Submit to Repository**
    - **Feedback from Adopter**

DTP Drug Response Data

- **MIAMEtox**
  - **Standards Development**
    - **CEBS**
    - **Controlled vocabularies (caDSR/EVS)**
  - **Create MAGEtox-ML**
    - **Evolving**
  - **Validation**
    - **CEBS, caArray?**
    - **Adopter Feedback**
  - **Coordinate with DTP**



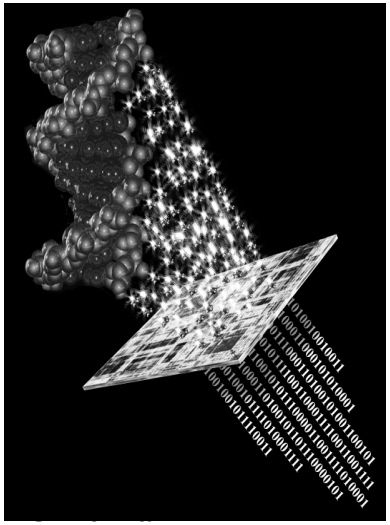
SKY and aCGH

- **SKYWEB**
  - **No model available**
  - **Extend MIAME and MAGE-OM**
  - **Local DB**
  - **Validation**
    - **Adopter feedback**

Conclusion

- **Why we see this pilot as valuable to caBIG**
  - **Cross platform**
  - **Extremely good use-case**
  - **Helps to identify areas of future development**
  - **Helps to develop standards and tools**
    - **to produce data (us)**
    - **to host (repository)**
    - **to analyze (adopter)**

The Genomics and Bioinformatics Group
LMP, CCR, NCI



John N. Weinstein, NCI
Ari Kahn⊙, SRA International
David Kane°⊙, SRA International
Robert Sfeir°, SRA International

Uma Shankavaram, NCI
William Reinhold, NCI
Barry Zeeberg, NCI

⊙caBIG NCI-60
°caBIG GOMiner

Uwe Scherf
Larry H. Smith
Lorrie Tanabe
Preetha Umamaheswaran
Y. Zhou (Joe)

Ajay
Jes Alexande
Sayed Daoud
Hosein Kouros-Mehr
Alika Maunakea
Darren Andrews
Yi Fan
Fuad Gwadry
Jae K. LeeNishizuka
Samuel Richman

Kimberly Bussey
Hong Cao
Sohyoung Kim
Joseph Ludwig
Sylvia Major
Daisaku Morita
Sudarshan Narasimhan
Satoshi Nishizuka
Mark Reimers
Margot Sunshine