


# **The Data Deluge and what to do with it**

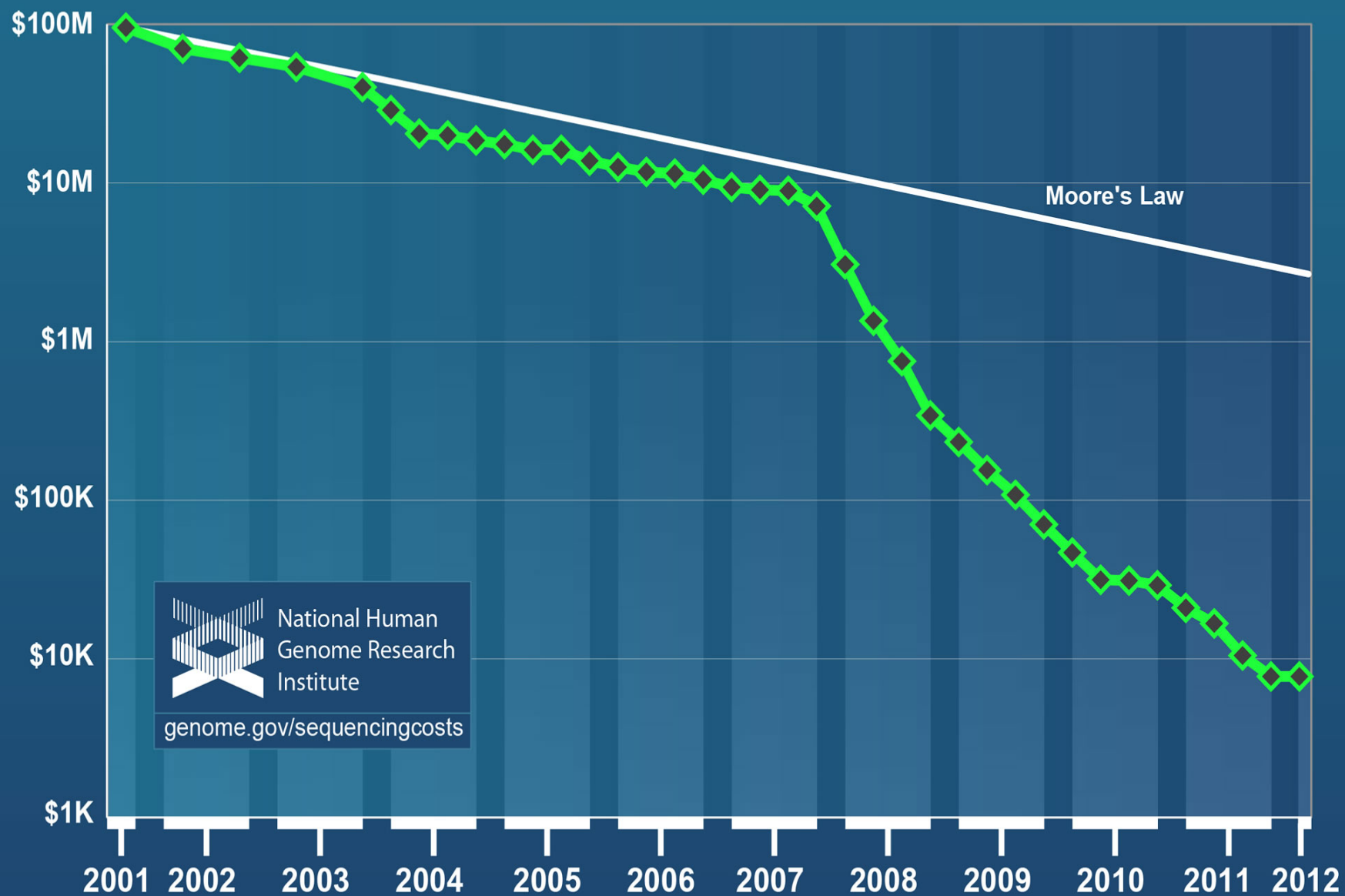
**Richard S. Savage**

**(Warwick Systems Biology/Warwick Medical  
School)**



**Biomedicine has become a  
data-rich subject**

## Cost per Genome



# Molecular Cancer Data

---

'n' samples

- Genomic
- Transcriptomics
- Epigenomics
- Proteomics
- Metabolomics

'p'  
features

(now)

$n \sim 10^3$

$p \sim 10^4$  (gene expression)  
 $10^6$  (SNPs)

(soon)

$n \sim 10^4$  (larger studies)

$10^7$  (UK population scale)

$p \sim 10^9$  (genome seq.)

# Clinical covariates

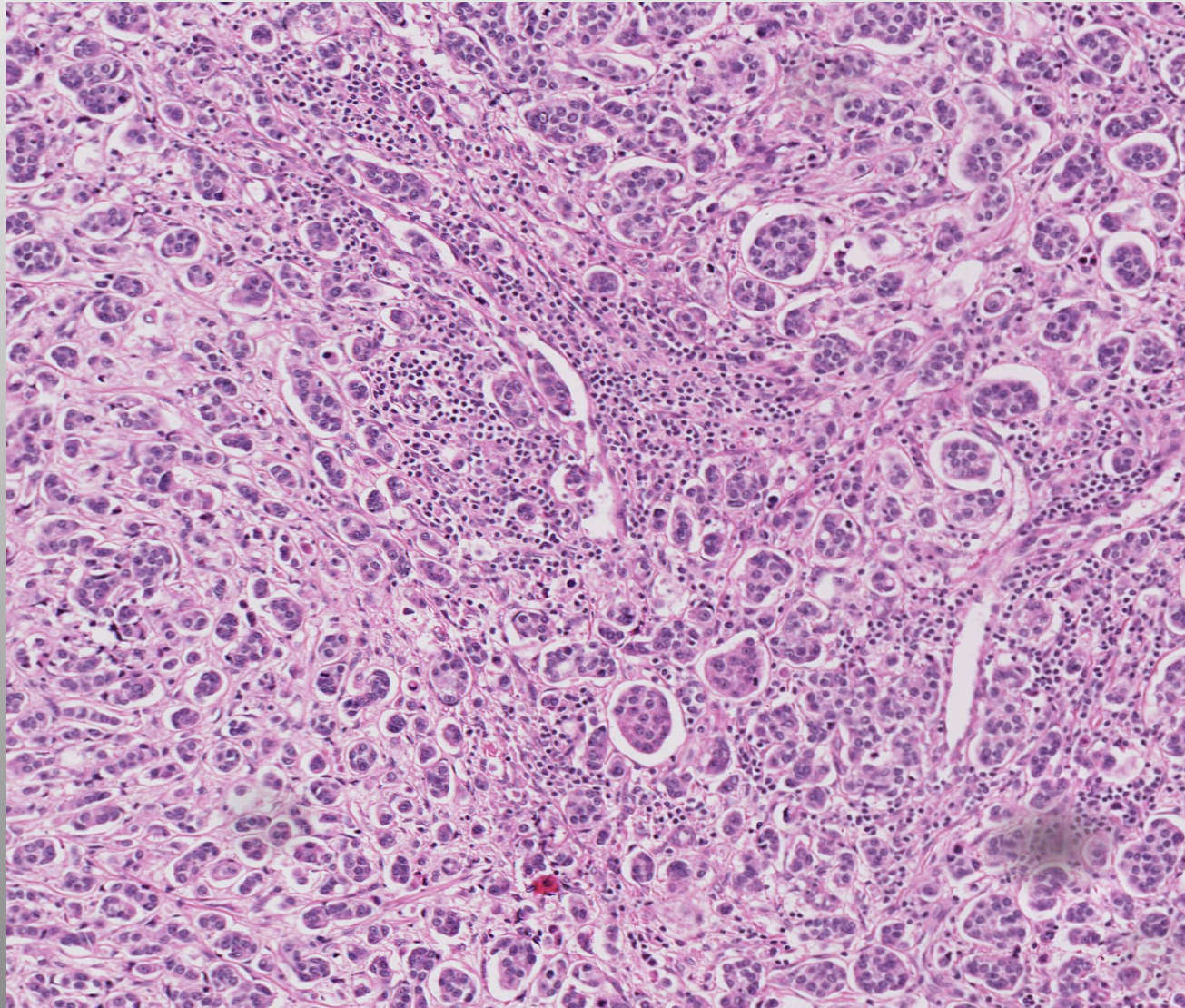
---

Clinical Feature
Age at diagnosis
Treatment group
Disease grade
Tumour size
No. of positive lymph nodes
Histological type
ER status
Tumour cellularity
PAM50 subtype
Treatment received
Batch identifier



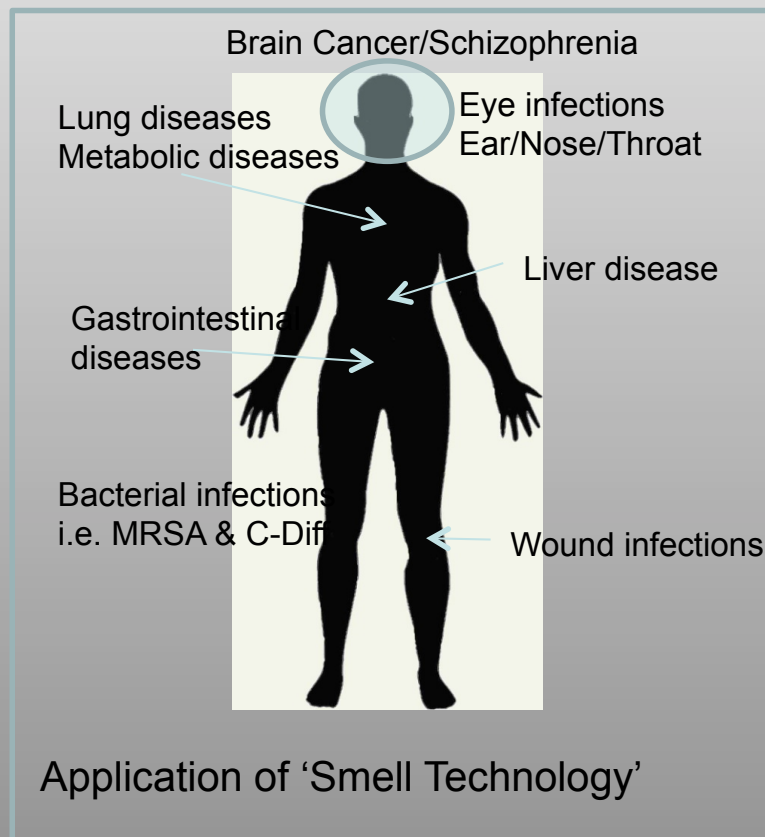
# Digital Pathology

---

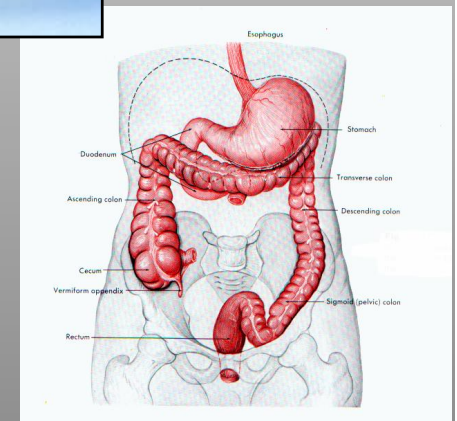


(image courtesy of  
Yinyin Yuan/METABRIC)

# Gas-phase biomarkers



- May not be specific marker
- Change in total profile
- For disease management/ healthy living



**A quick case study....**



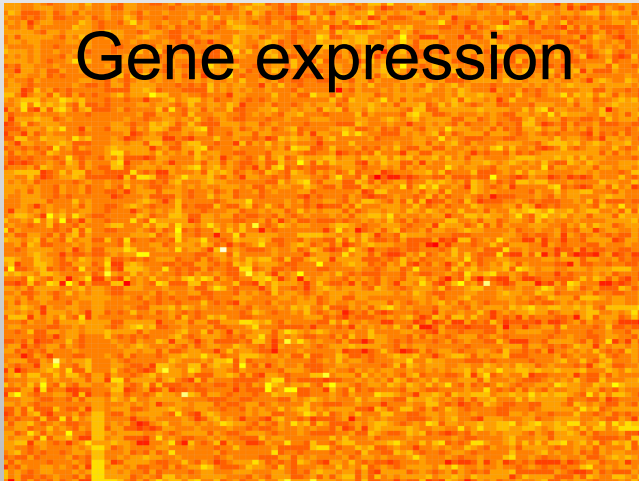
# **The Sage Bionetworks-DREAM Breast Cancer Prognosis Challenge**

**Aim:** predict breast cancer  
survival times

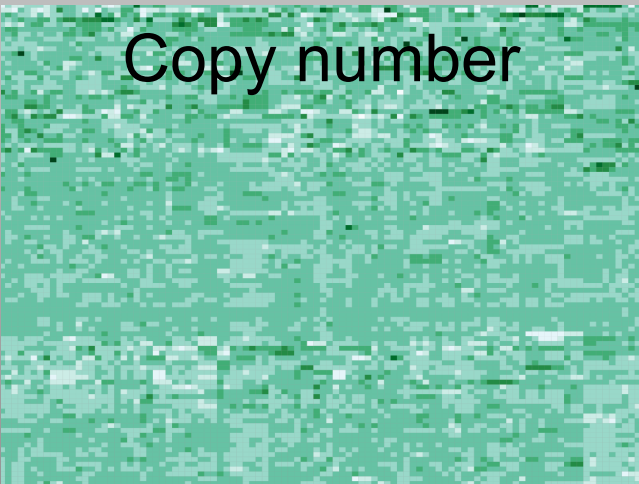
# METABRIC (~2000 patients)

---

Gene expression



Copy number



## Clinical Feature

Age at diagnosis

Treatment group

Disease grade

Tumour size

No. of positive lymph nodes

Histological type

ER status

Tumour cellularity

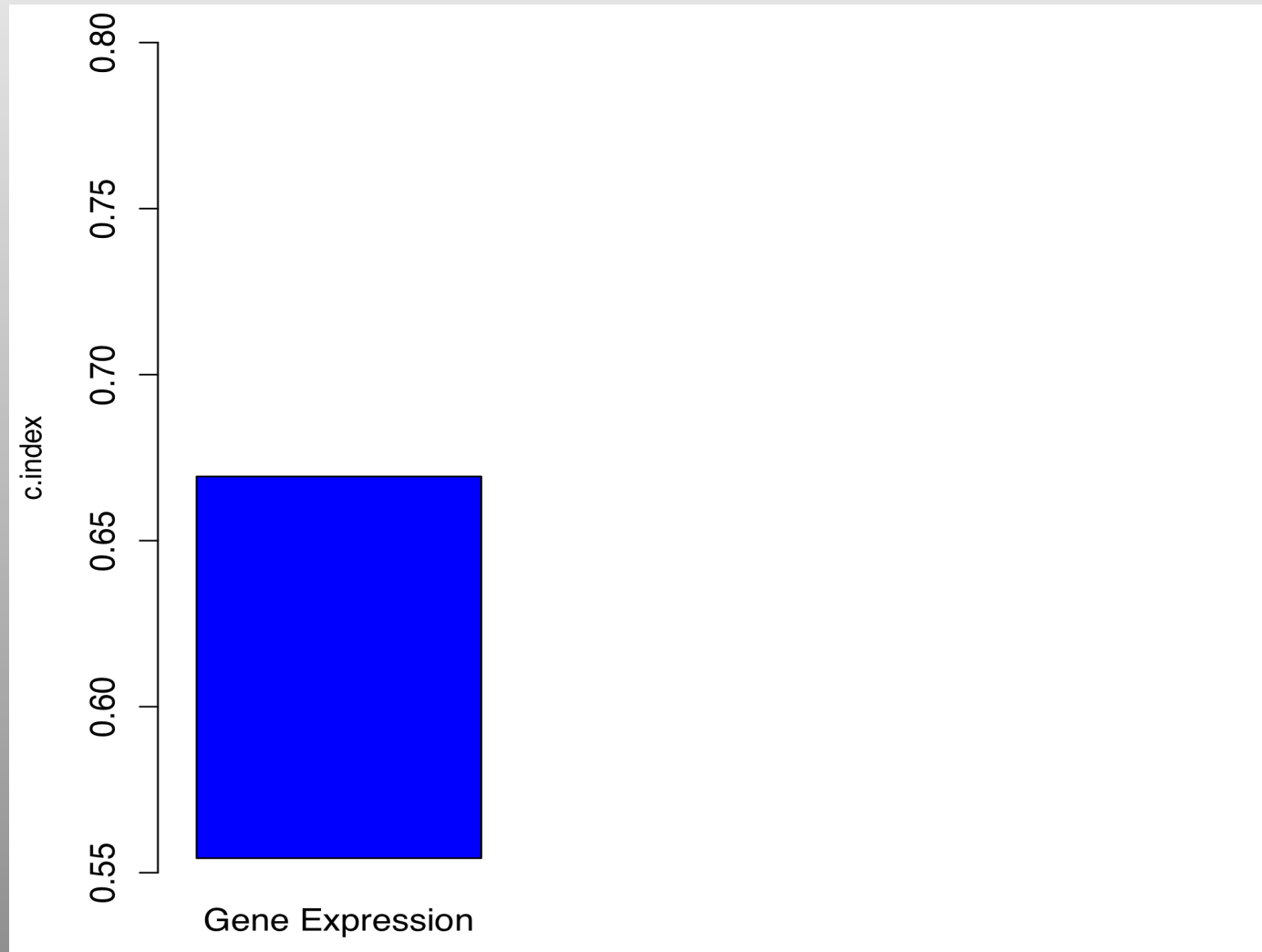
PAM50 subtype

Treatment received

Batch identifier

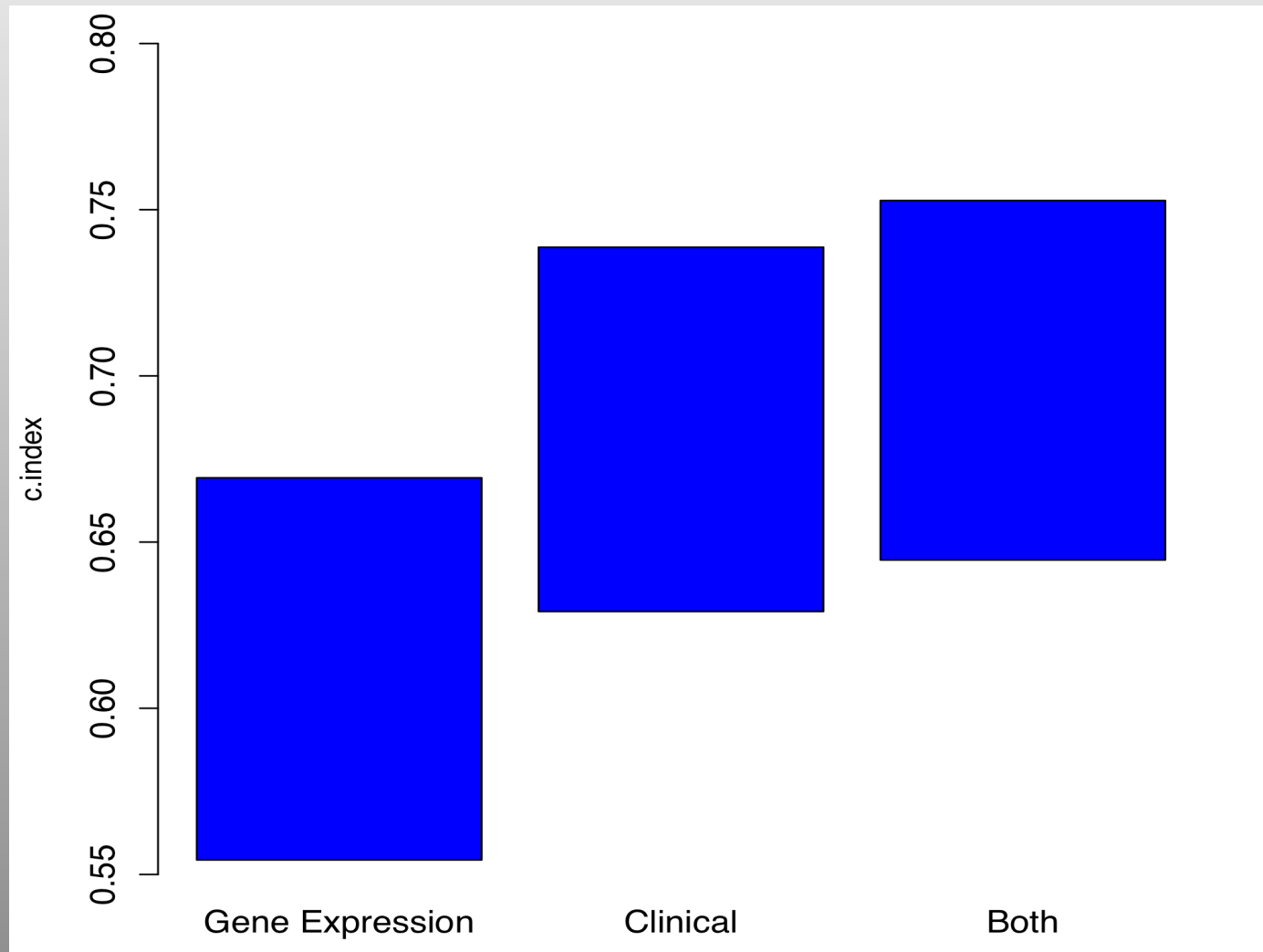
# Gene Expression

---



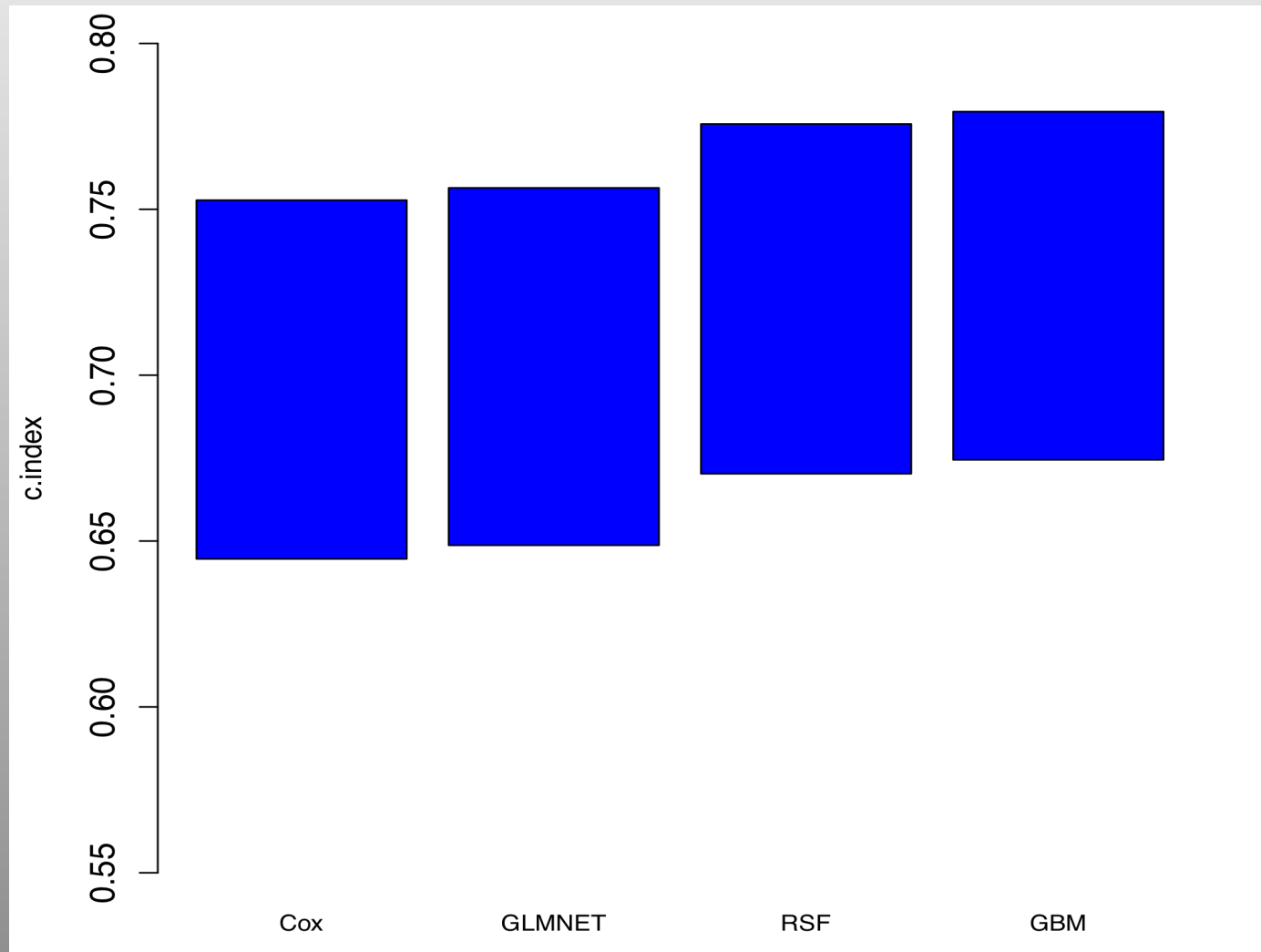
# Adding in Clinical Covariates

---



# Range of Models

---





# Further Directions

---

- Better feature learning models
  - Better survival prediction models
  - Data fusion methods
- 
- (samples from more patients)
  - (more data types)



**Big models?**

(image courtesy of Mark Morgan)

**It depends...**

# Big Models? - Pros

---

- Might be better at capturing complex underlying structure
- Better able to integrate different data types?

# Big Models? - Cons

---

- Might **not** be better at capturing complex underlying structure
- Computationally more intensive
- Can be harder to develop



# Big Models? – a Thought

---

- Big Data needs models that scale well, in CPU/storage terms
- **Perhaps we also need models that scale their complexity, in response to the underlying structure and/or computing budget?**

**Questions?**

# Acknowledgements

---

- Sage Bionetworks
- The DREAM Challenge organisers
- The other teams from the Challenge
- Yinyin Yuan (METABRIC)
- James Covington (gas-phase biomarkers)

**Questions?**





# Why Cancer?

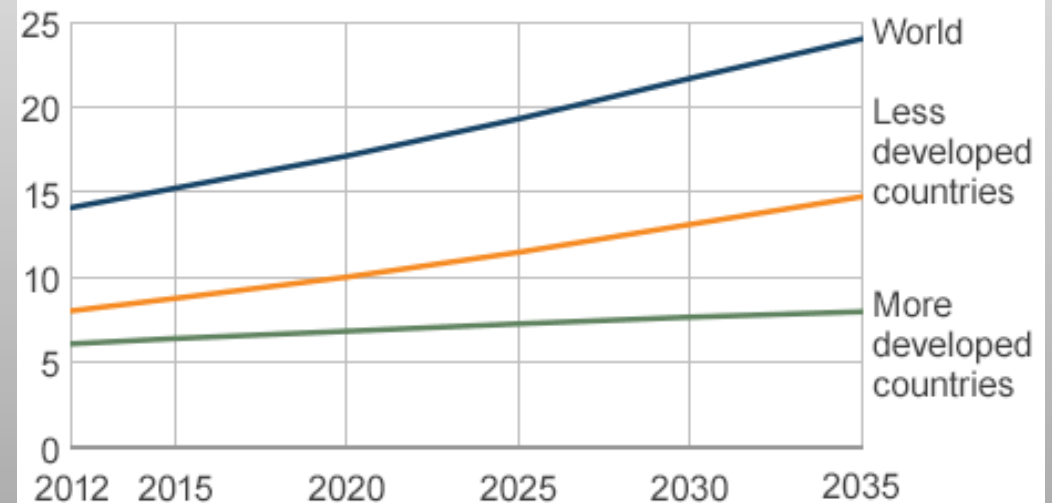
# Why Cancer?

---

- Per year...
  - 14 million cases
  - 8.2 million deaths
  - 1 in 7 deaths worldwide
  - (WHO Globocan 2012 report)

Predicted global cancer cases

Cases (millions)



Source: WHO GloboCan



# Concordance Index

---

If actual survival time of patient A  $>$  patient B

Concordance Index (CI) is the probability that our model correctly predicts this

For example:

CI = 0.5  $\rightarrow$  coin-flip (useless)

CI = 1  $\rightarrow$  perfect prediction

# Final Leaderboard

Model ID	Model Name	Team ID	Train Score	Initial Test Score	Final Test Score
syn1444444	Attractor Metagenes Model 101515	317809	0.7519	0.7236	0.7225
syn1444370	Attractor Metagenes Model 101509	317809	0.7463	0.7269	0.7210
syn1444400	Attractor Metagenes Model 101511	317809	0.7537	0.7245	0.7199
syn1444472	Attractor Metagenes Model 101516	317809	0.7447	0.7288	0.7192
syn1444424	Attractor Metagenes Model 101514	317809	0.7475	0.7273	0.7177
syn1443133	PittAttractomeHyb.2	422262	0.7538	0.7207	0.7175
syn1435273	Attractor Metagenes Model 101099	323618	0.7294	0.7243	0.7170
syn1426931	NCIS_S01E06	323618	0.7302	0.7240	0.7169
syn1443594	WarwickSystemsBiology (Mon Oct 15 21.36.33 2012)	362302	0.7493	0.7255	0.7168
syn1443155	PittAttractomeHyb.20Features	422262	0.7588	0.7229	0.7162
syn1426690	ENSEMBLE 100902	962237	0.7536	0.7254	0.7136
syn1436850	ENSEMBLE 101001	962237	0.7404	0.7246	0.7131
syn1444362	Attractor Metagenes Model 101507 OSDS	323618	0.7520	0.7242	0.7130
syn1444367	Attractor Metagenes Model 101508 OSDS	323618	0.7555	0.7235	0.7124
syn1443598	WarwickSystemsBiology (Mon Oct 15 21.46.00 2012)	362302	0.7494	0.7258	0.7122