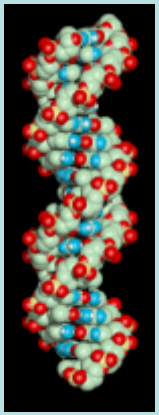
Gupta Lab



10110101

# IT495 DNA Storage and Security
# Lecture 1:Admin Details and Background


**Prof. Manish K. Gupta**

**Laboratory of Natural Information Processing**

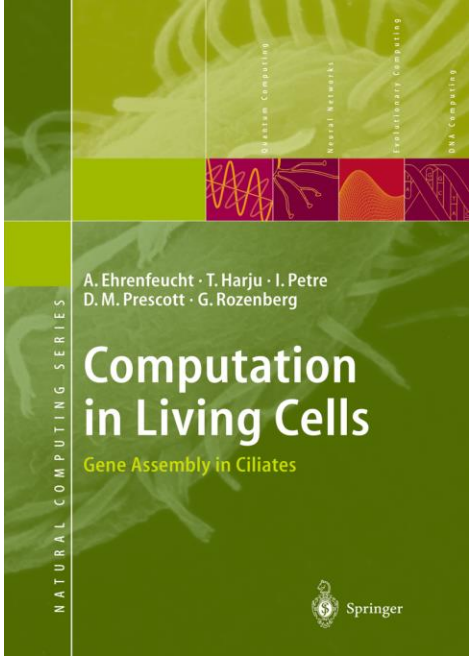
# Outline of Presentation

- **Admin Details**
- **Overview and Subject Classification**
- **Historical Introduction & Motivation**


http://en.wikipedia.org/wiki/File:UT_HelloWorld.jpg

2004 @ iGEM: Students of Boston University, Caltech, MIT, Princeton and U of Texas

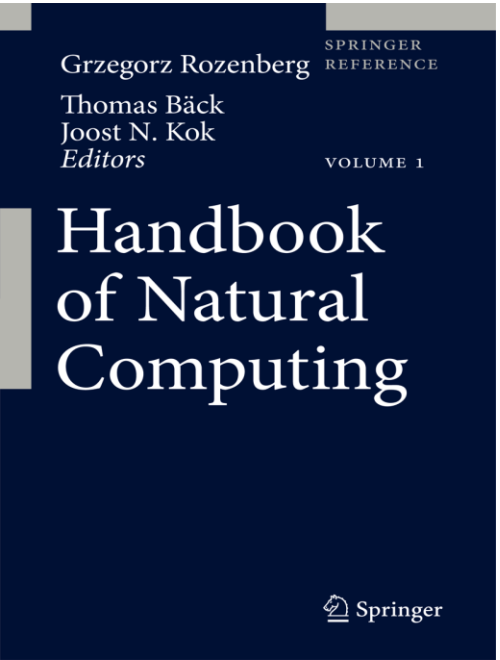
http://partsregistry.org/cgi/htdocs/SBC04/index.cgi

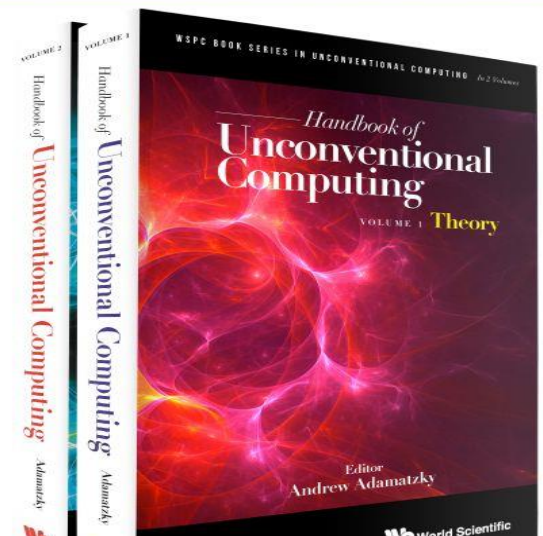

# Administrative Details IT495

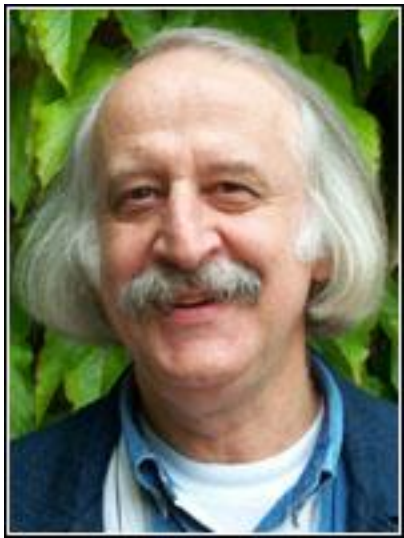
Natural computing builds a bridge between computer science and natural sciences.

**Lila Kari and Grzegorz Rozenberg**




**Forthcoming**



# Instructor

- **Manish K Gupta (www.mankg.com)**
- **Office: Room 2209 Faculty Block 2**
- **Email: mankg@guptalab.org**
- **Phone: 91-79-68261549**

YouTube Channel: https://www.youtube.com/c/ManishGuptamankg

Twitter: https://twitter.com/mankg

Lab: http://www.guptalab.org

Biography (http://www.guptalab.org/mankg/public_html/WWW/shortbio.html )

Gupta Lab

http://www.guptalab.org

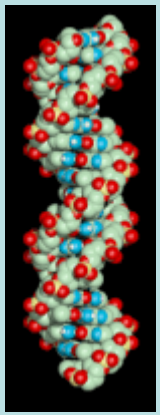
# Marks Distribution (Tentative) / Grading Policy

| | |
|---|---|
| **Mid Term Test 1** | **30%** |
| **Scribe Notes** | **20%** |
| **Projects** | **50%** |

# Opportunities Available

- **Theoretical**
- **Software Development**
- **Experimental**


# How to Store Elephants?



10110101


**Prof. Manish K. Gupta**

**Laboratory of Natural Information Processing**

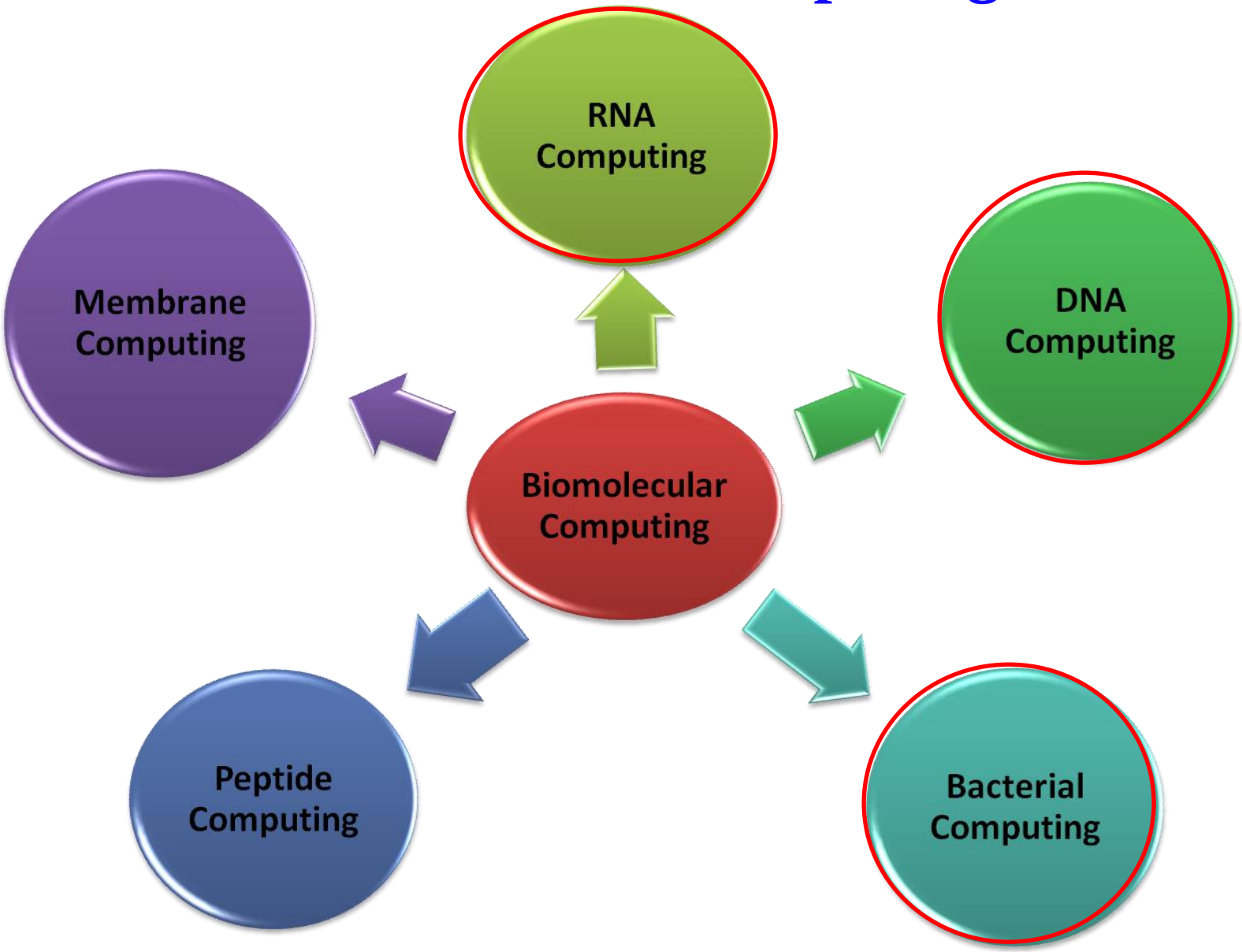
http://en.wikipedia.org/wiki/Elephant#mediaviewer/File:Elephant_Walking_animated.gif

# Motivation

- "**Biology** and **computer science**
  – **life** and **computation** – are related.
  I am confident that at their interface great discoveries await those who seek them"

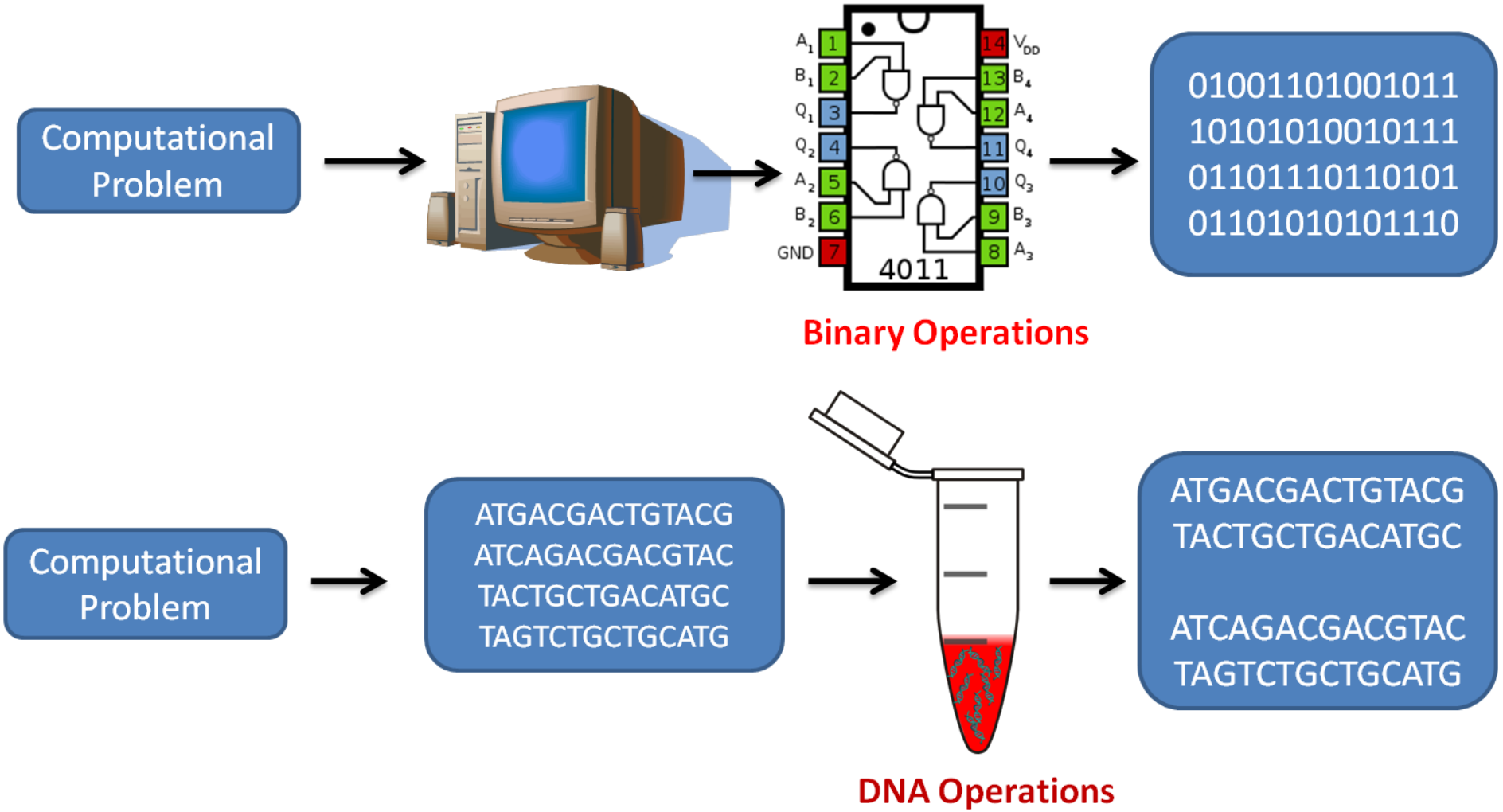
[Leonard Adleman, Scientific American, August 1998]

# Biomolecular Computing

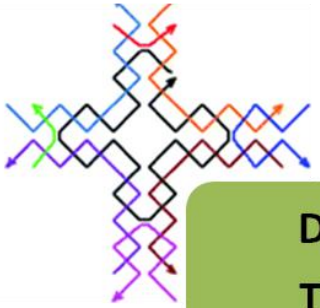


# DNA Computing

Computing using DNA strands. DNA hybridization is main tool for DNA Computing

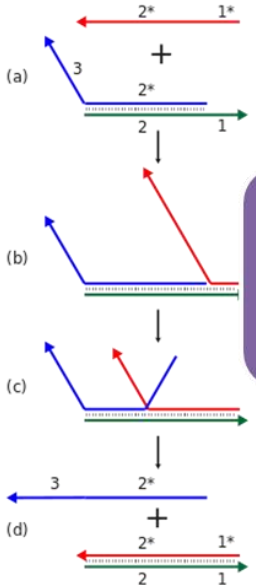


# DNA Computing Applications



DNA Tiles

DNA Nanostructures

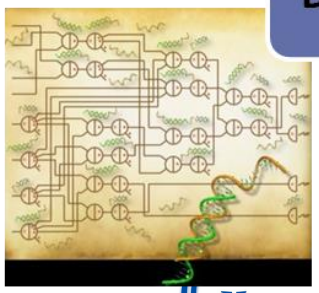

DNA Strand Displacement

DNA Computing

DNA Cryptography

DNA Circuits

DNA Data Storage Systems



**Storage is the fundamental need for both life and computation**

**Every life needs storage…..**

**Storage of resources such as food…**

**Storage is also a basic computing primitive…**

**What is life?**

**Life = Storage + Information Processing (Computing)**

**Modern Humans Stores data…**

# Some Examples of Natural Storage…

http://weirdimals.files.wordpress.com/2010/11/honeypotant.jpg

# History

Cave Paintings

40000 BC

Tally Sticks

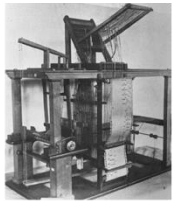
5000 BC

Papyrus

2000 BC

Paper

100 BC



Punch Card

1725 AD

Hard Disk

1956

DRAM

1966

8" Floppy disk

1971

5.25" floppy disk

1976

1877

Phonograph

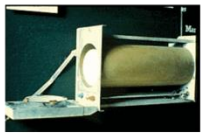
1898

Telegraphone

1928

Magnetic Tape

1948



Williams Tube

1980

Other floppy disks

1982

Compact Disc

1988

CD-ROM

1990

Magneto Optical Disc (MOD)

Compact Flash

1994

DVD

1995

Memory Stick

1998

2001

USB Flash Drive

2003

Blu-Ray

2004


HD-DVD

today


Cloud Data Storage

# Cloud Data Center

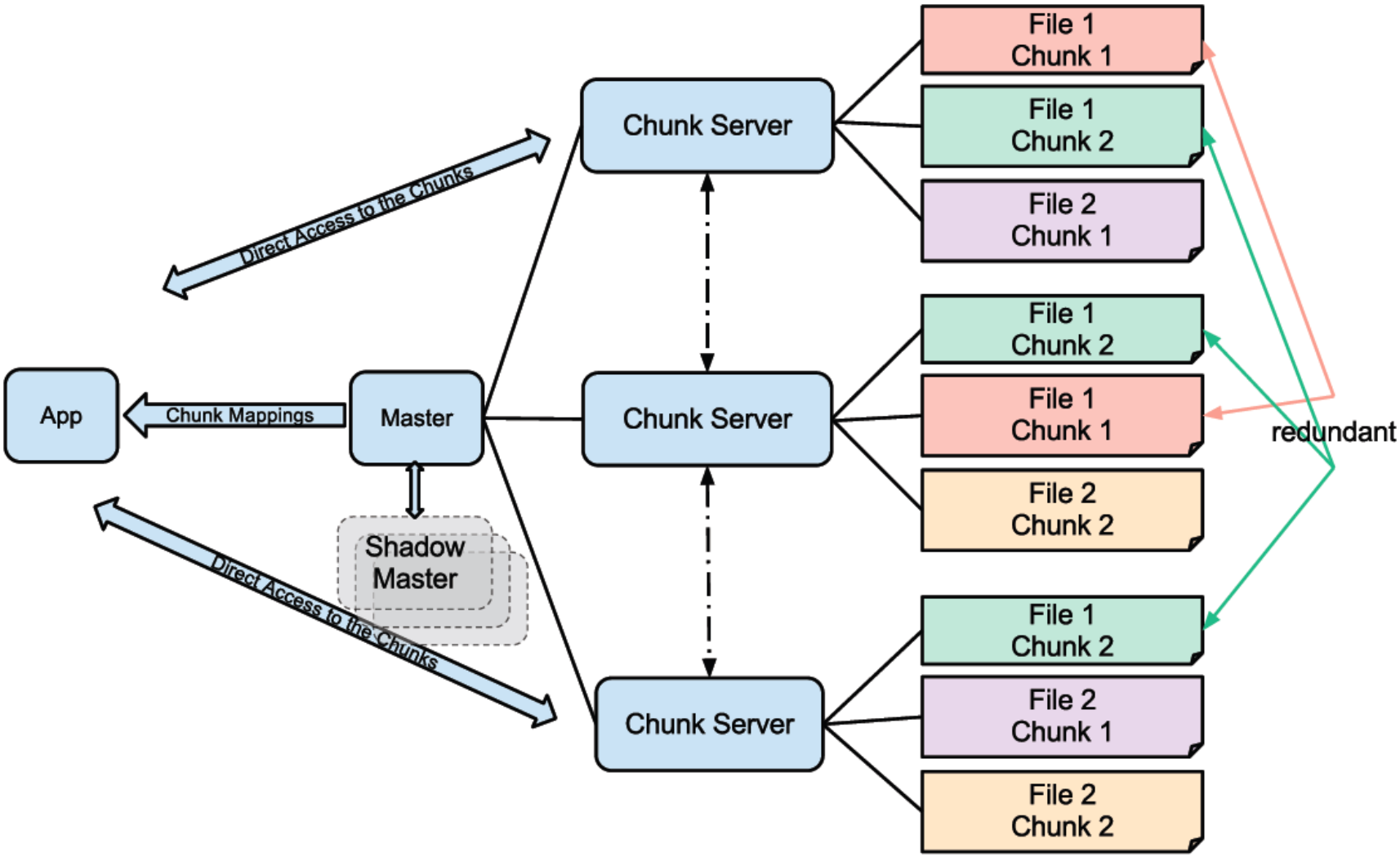
http://www.highperformancedatacenternews.com/images/Ibm_Data_Center_Auckland_NZ_2011_432.png

IBM Cloud Data Center, Auckland, New Zealand


http://www.koreaittimes.com/images/imagecache/large/cheonan%20cloud%20center.JPG

Cheonan Cloud Data Center, South Korea

# File Storage Systems: Google File System



# What about Elephants?


http://www.informationweek.com/big-data/big-data-analytics/5-big-wishes-for-big-data-deployments/d/d-id/1109606?

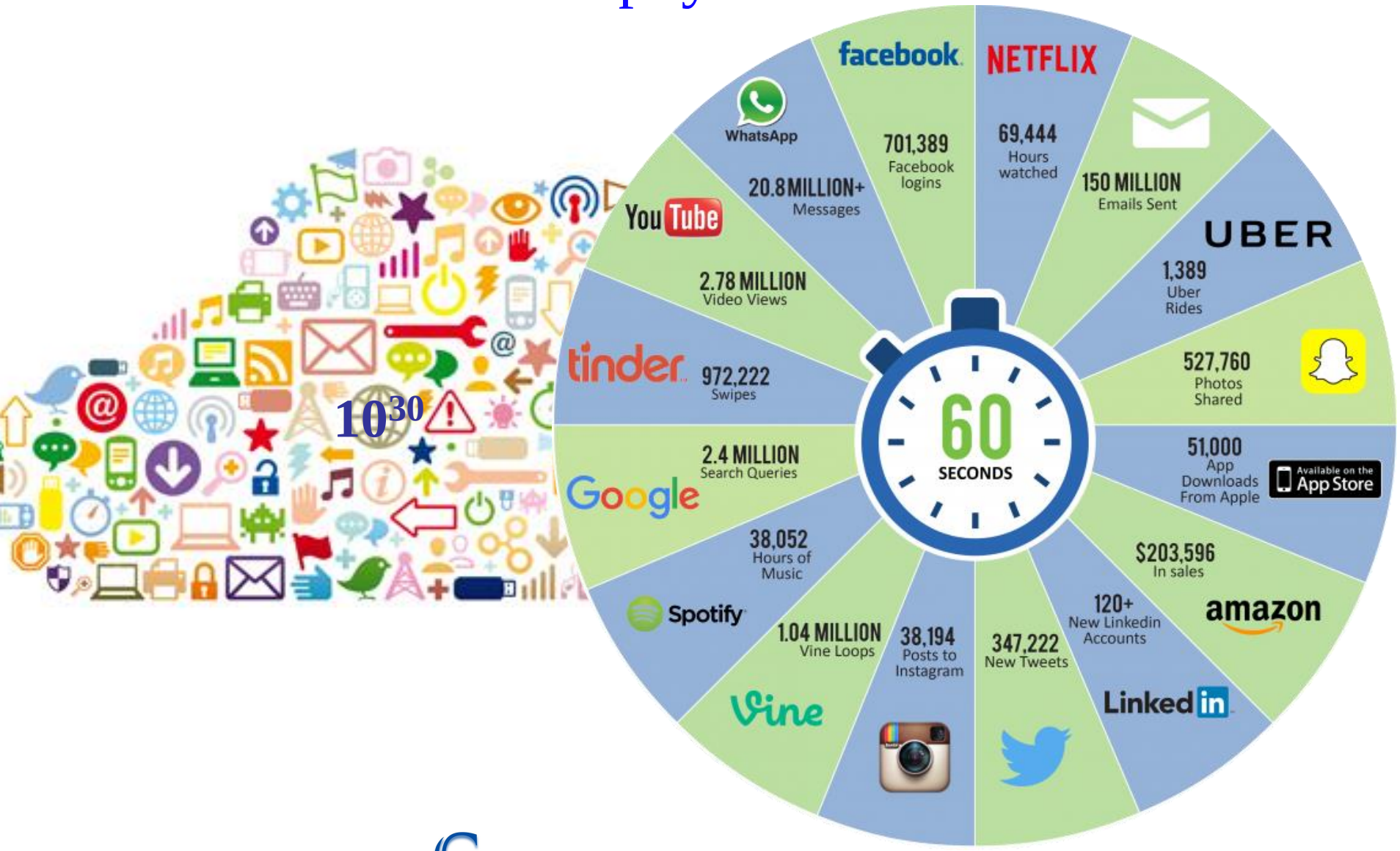
# How many elephants are we generating now?

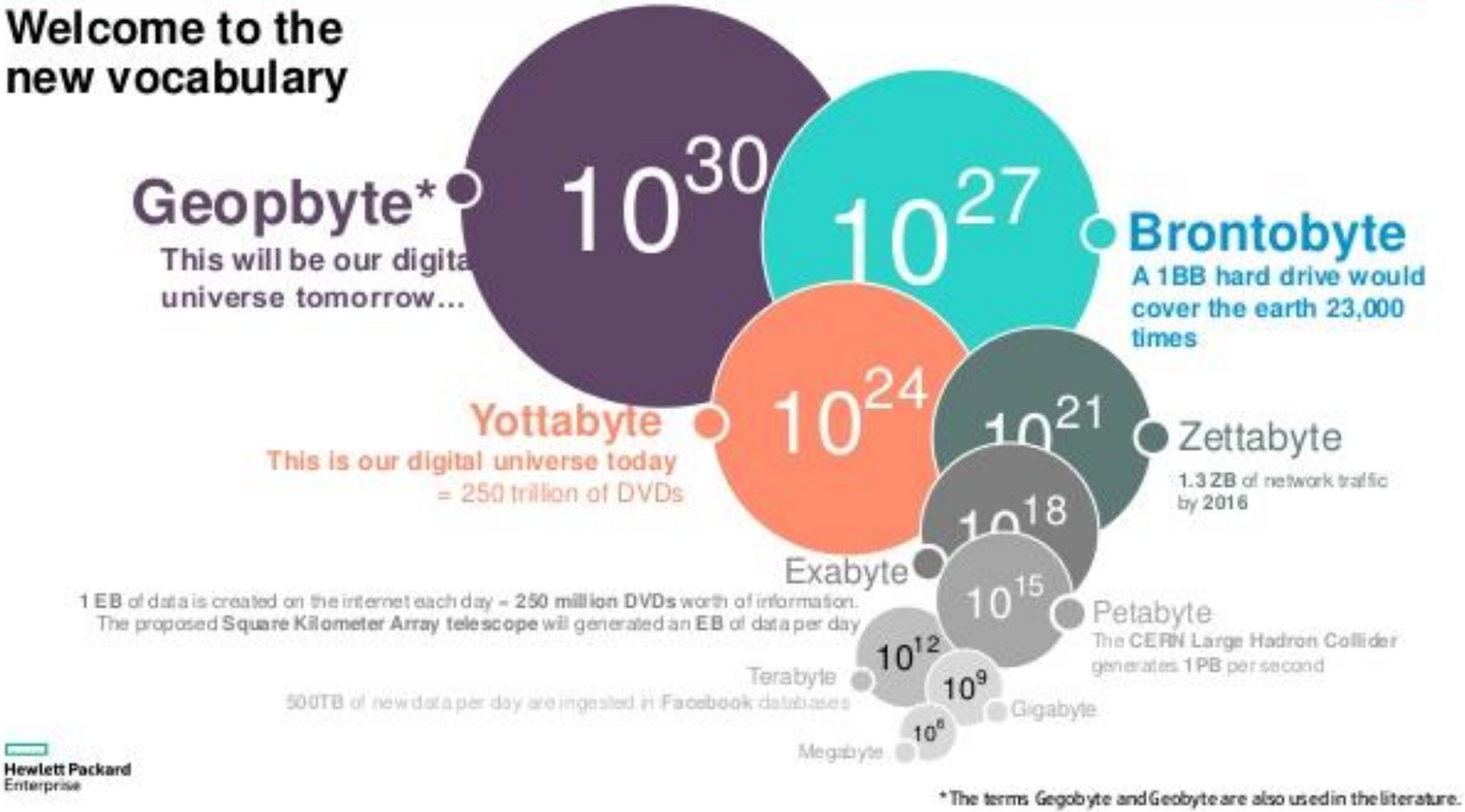
***"From the dawn of civilization until 2003, humankind generated five exabytes (1 exabytes = 1 billion gigabytes) of data. Now we produce five exabytes every two days and the pace is accelerating.***

***Eric Schmidt, Executive Chairman, Google, August 4, 2010.***

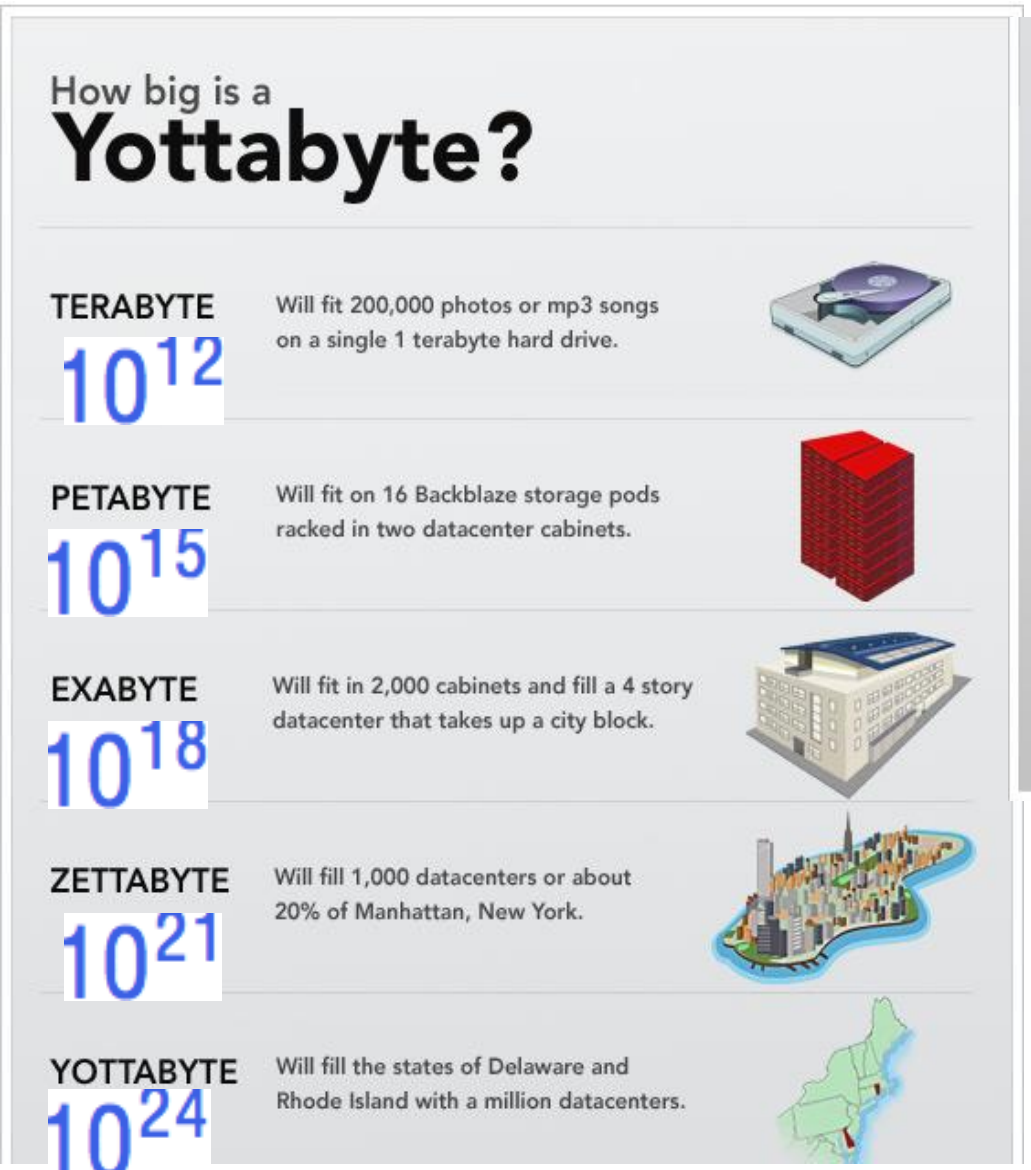
http://upload.wikimedia.org/wikipedia/commons/e/e5/Eric_Schmidt_at_the_37th_G8_Summit_in_Deauville_037.jpg

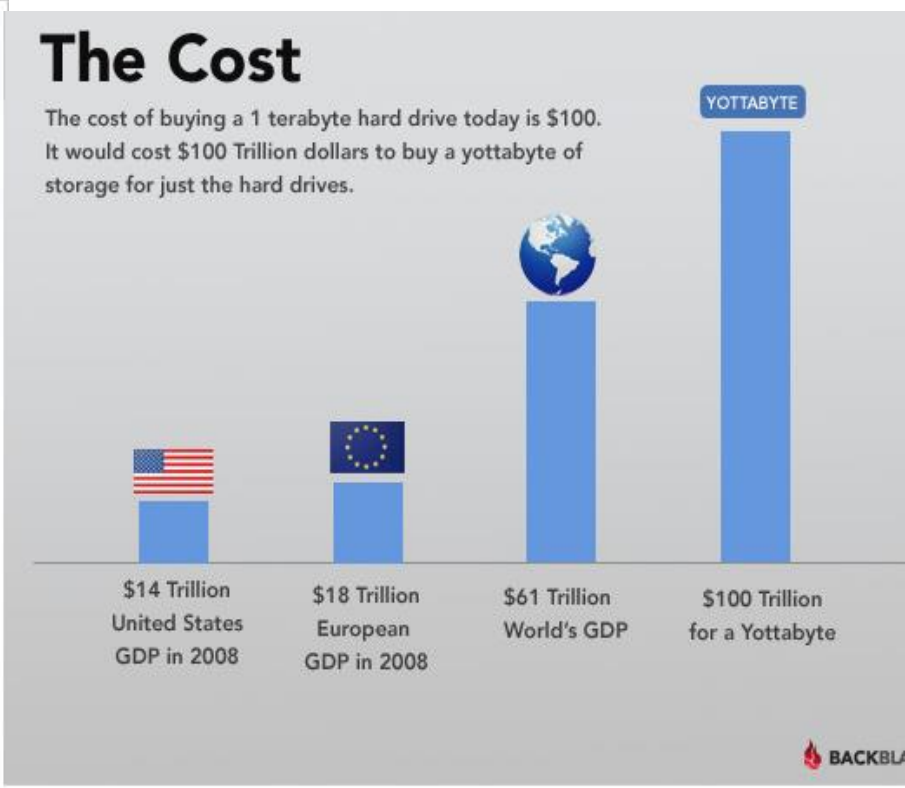
# Cloud of Internet of Things (IoT) will produce Geopbyte: $10^{30}$




Welcome to the new vocabulary
Geopbyte*
This will be our digita
universe tomorrow...
10^30
10^27
Brontobyte
A 1BB hard drive would cover the earth 23,000 times
Yottabyte
This is our digital universe today
= 250 trillion of DVDs
10^24
10^21
Zettabyte
1.3 ZB of network traffic by 2016
10^18
Exabyte
1 EB of data is created on the internet each day = 250 million DVDs worth of information.
The proposed Square Kilometer Array telescope will generated an EB of data per day
10^15
Petabyte
The CERN Large Hadron Collider generates 1 PB per second
10^12
Terabyte
500TB of new data per day are ingested in Facebook databases
10^9
Gigabyte
10^6
Megabyte
Hewlett Packard Enterprise
*The terms Gegobyte and Geobyte are also used in the literature.

# How Big is the Elephant?




What is the cost for storing the Elephants?

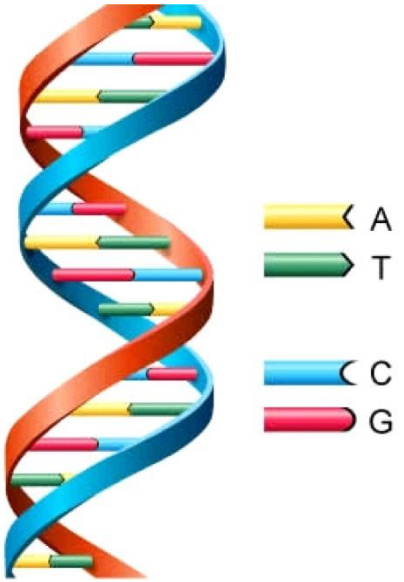
# What is the ultimate storage device?

# Next Generation Storage Device - DNA

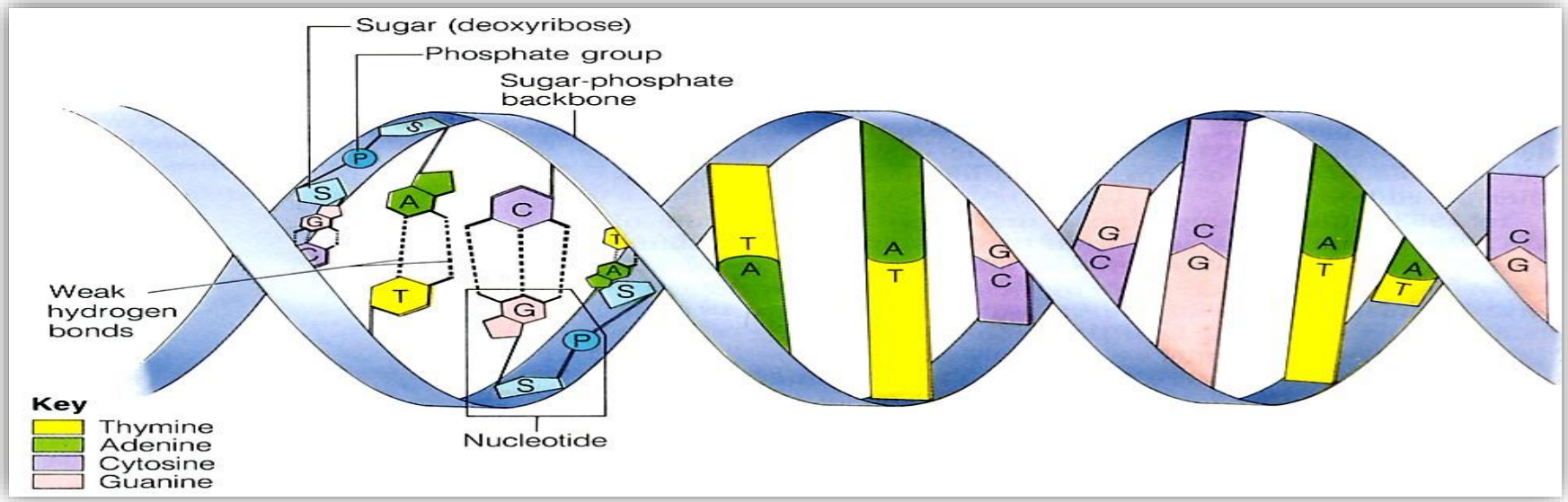



Image Credit: http://media.economist.com/sites/default/files/imagecache/full-width/images/print-

# DNA: The Blueprint of Life

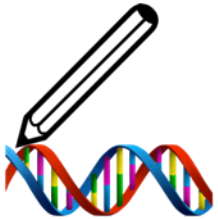



- Four letter alphabet (nucleotides, bases):
  - A (adenine)
  - C (cytosine)
  - G (guanine)
  - T (thymine)
- Complementary base pairs CG, AT
- Hybridization via base pairing

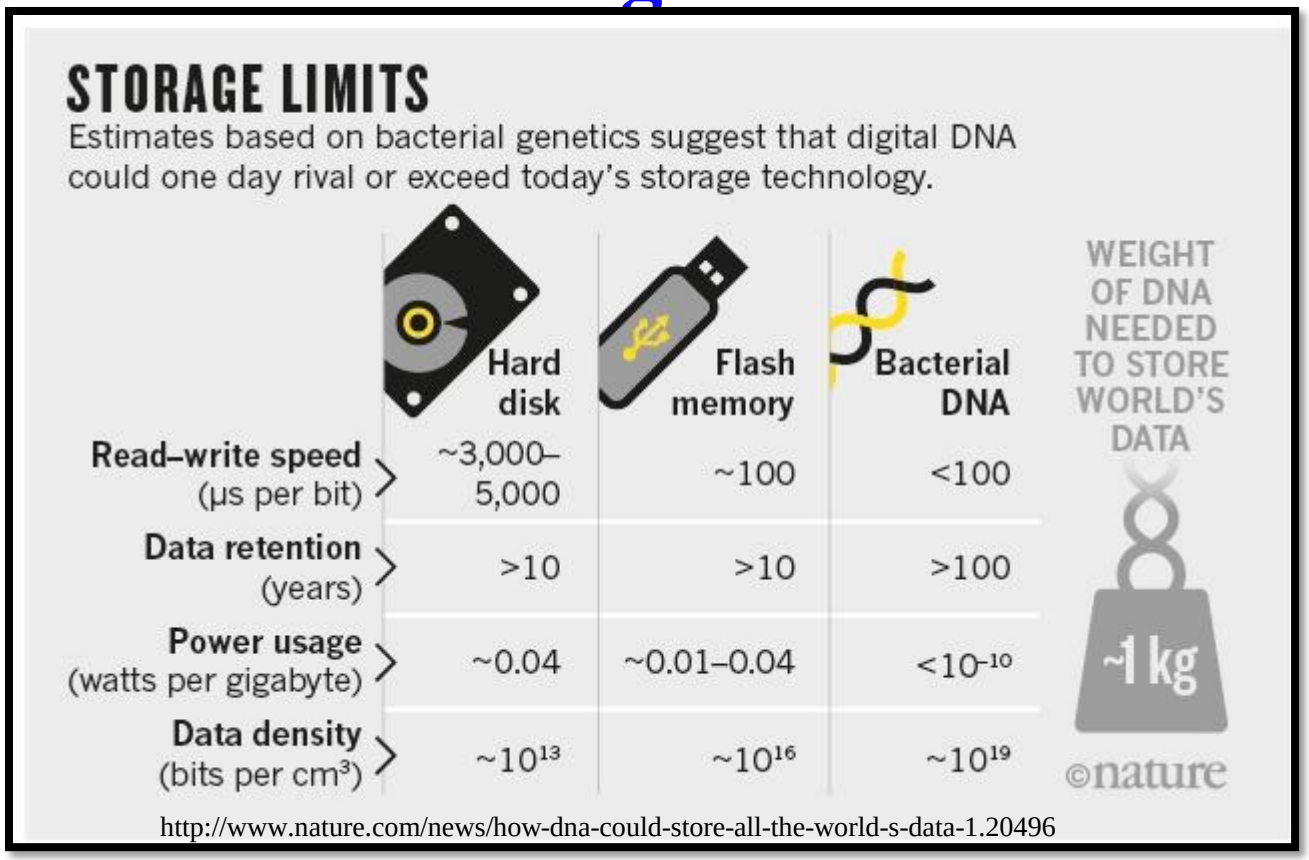


# DNA Reading and Writing

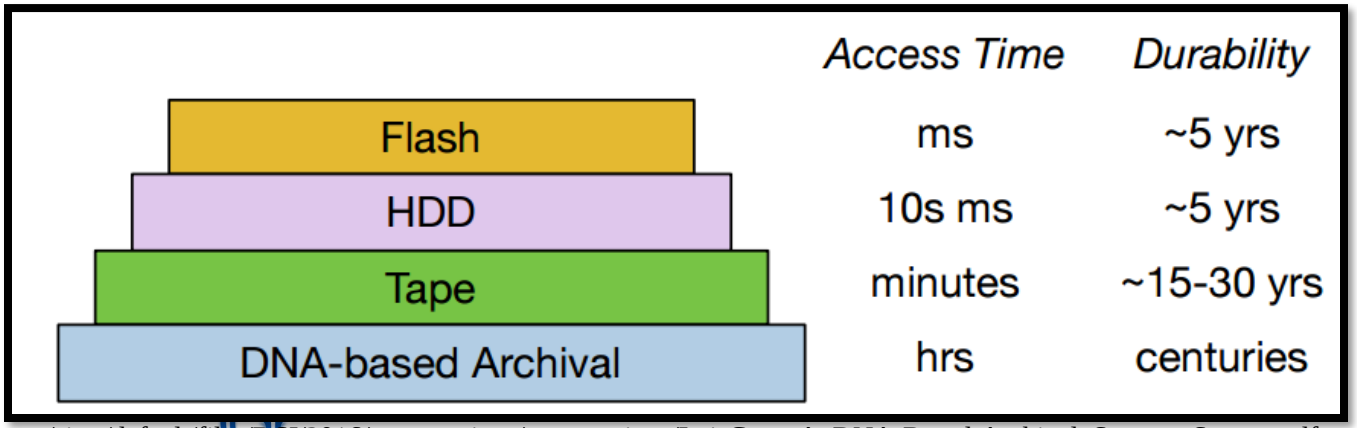
- **Central to a DNA manipulation: DNA Synthesis and Sequencing.**
- **DNA synthesis = write process.**
- **DNA sequencing = read process.**




# DNA as Storage Medium

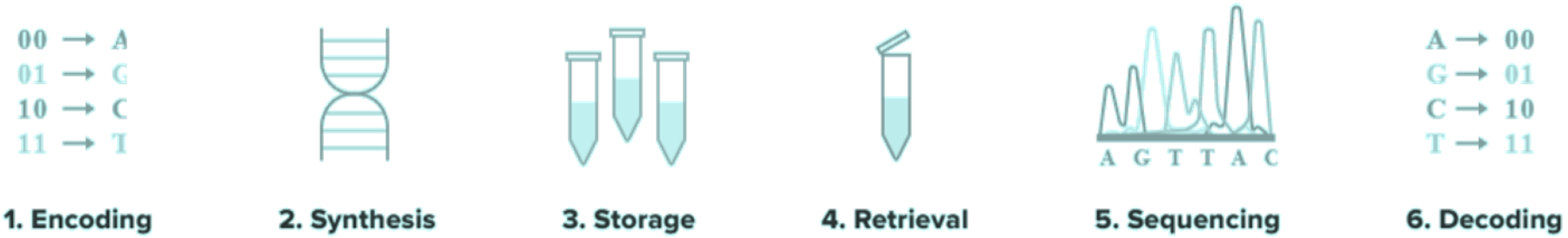


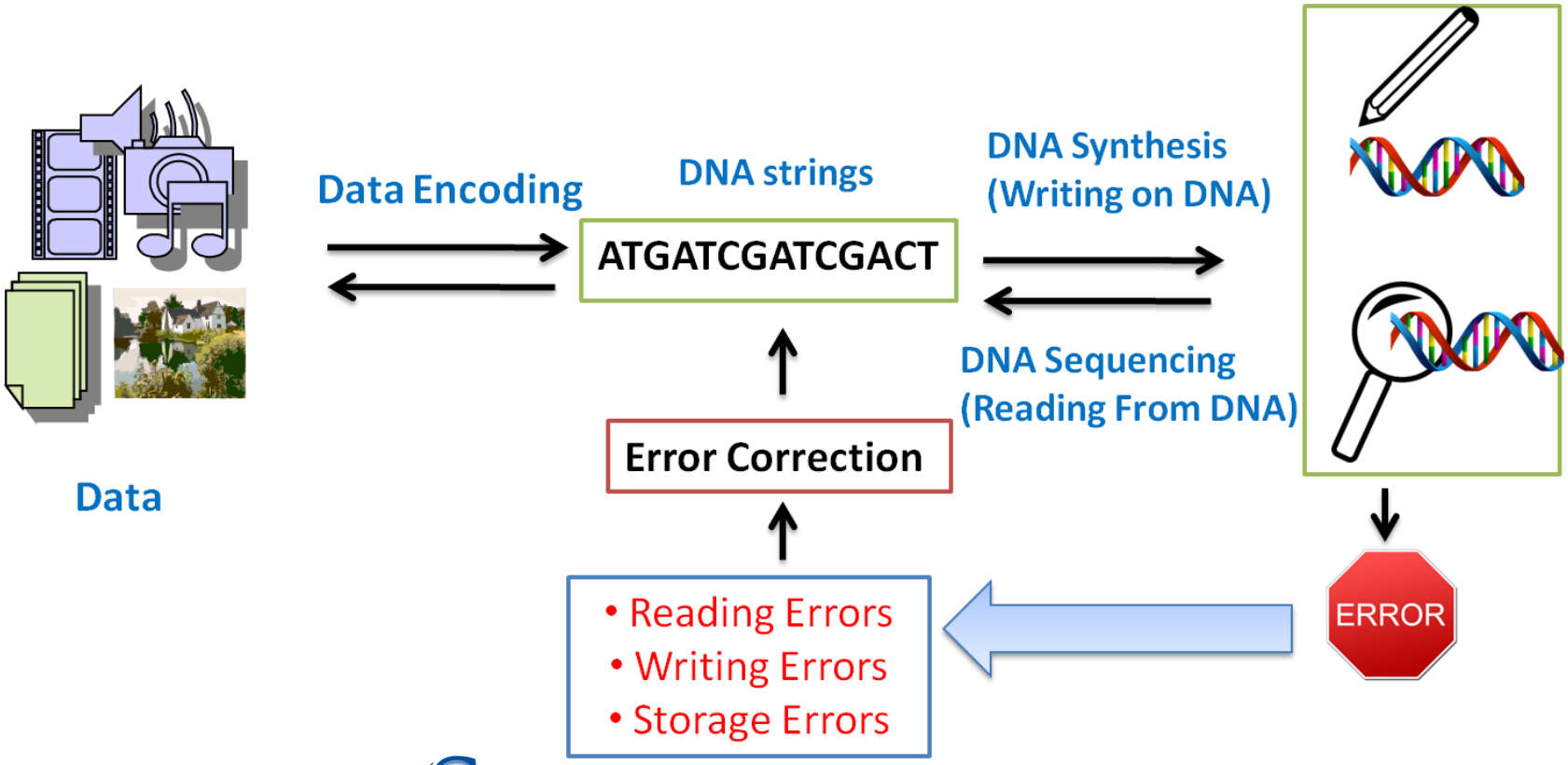

http://www.nature.com/news/how-dna-could-store-all-the-world-s-data-1.20496



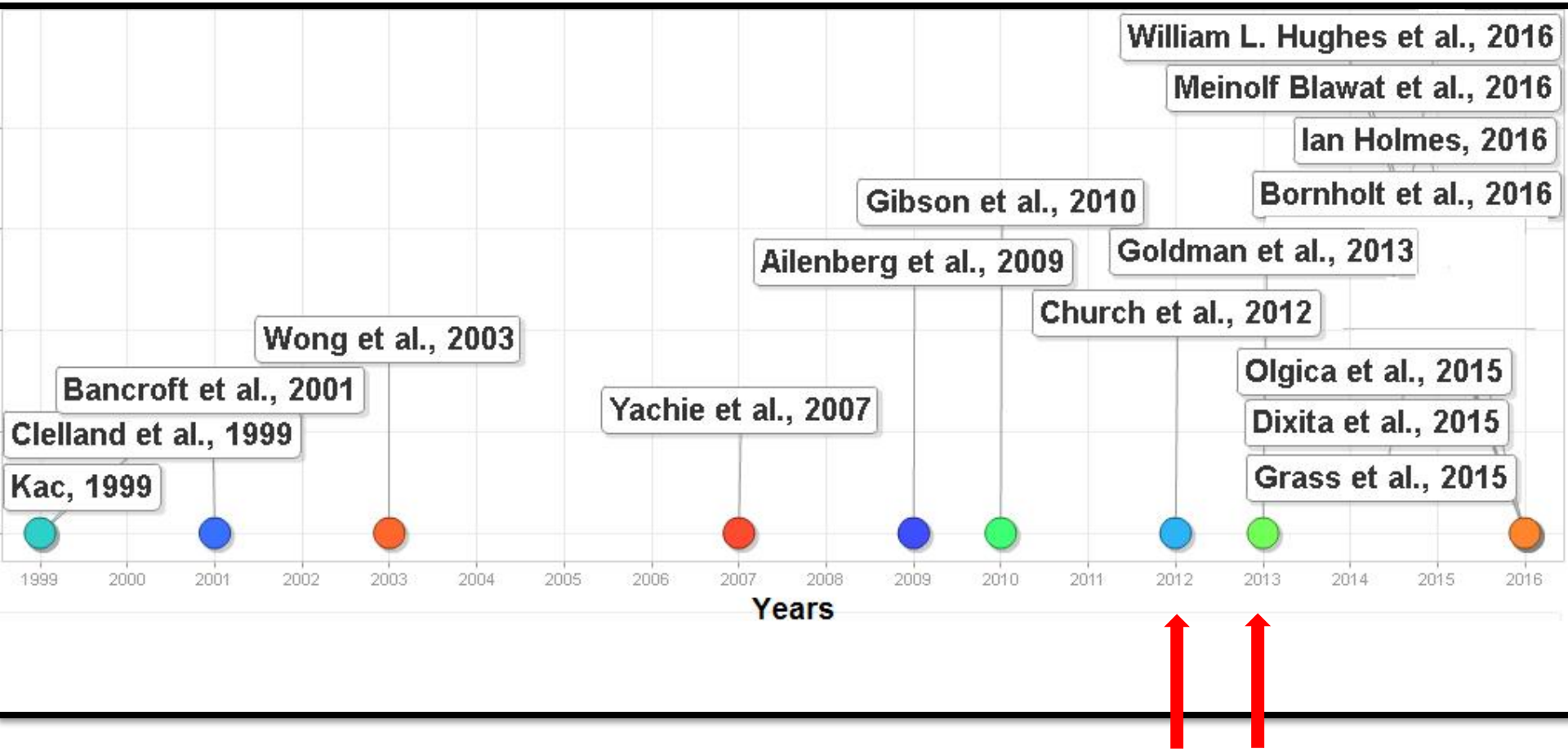

http://www.snia.org/sites/default/files/DSI/2016/presentations/gen_sessions/LuisCeze_A_DNA-Based-Archival_Storage_System.pdf

# How DNA Data Storage System Works?



# Timeline for DNA based Data Storage




DNA Cloud, 2013

# DNA based Data Storage Systems


•Church, George M., Yuan Gao, and Sriram Kosuri. "Next-generation digital information storage in DNA." Science 337.6102 (2012): 1628-1628.

•Goldman, N., Bertone, P., Chen, S., Dessimoz, C., LeProust, E. M., Sipos, B., & Birney, E. (2013). Towards practical, high-capacity, low-maintenance information storage in synthesized DNA. Nature, 494(7435), 77-80.

•Shalin Shah, Dixita Limbachiya and Manish K. Gupta, "DNACloud: A Potential Tool for storing Big Data on DNA" Foundations of Nanoscience: Self-Assembled Architectures and Devices (FNANO14), SnowBird, Utah, USA, 2014.

•Grass, R. N., Heckel, R., Puddu, M., Paunescu, D., & Stark, W. J. (2015). Robust Chemical Preservation of Digital Information on DNA in Silica with Error-Correcting Codes. Angewandte Chemie International Edition, 54(8), 2552-2555.

•Kiah, Han Mao, Gregory J. Puleo, and Olgica Milenkovic. Codes for DNA storage channels." Information Theory Workshop (ITW), 2015 IEEE. IEEE, 2015.

•Gabrys, Ryan, Han Mao Kiah, and Olgica Milenkovic. Asymmetric Lee distance codes for DNA-based storage." 2015 IEEE International Symposium on Information Theory (ISIT). IEEE, 2015.

•Gabrys, Ryan, Han Mao Kiah, and Olgica Milenkovic. "Asymmetric Lee distance codes: New bounds and constructions." Information Theory Workshop (ITW), 2015 IEEE. IEEE, 2015.

•Yazdi, SM Hossein Tabatabaei, et al. "A rewritable, random-access DNA-based storage system." Scientific reports 5 (2015).

Holmes, Ian. "Modular non-repeating codes for DNA storage.“ arXiv preprint arXiv:1606.01799 (2016).

•Dixita Limbachiya, Vijay Dhameliya, Madhav Khakhar, Manish K Gupta, "On Optimal Family of Codes for Archival DNA Storage” Proceedings of IWSDA’15 (2015): 123-127

•Hunt, F. H., Perkins, S., & Smith, D. H. (2015). Channel models and error correction codes for DNA information storage. International Journal of Information and Coding Theory, 3(2), 120-136.

•Gabrys, R., Yaakobi, E., & Milenkovic, O. (2016). Codes in the Damerau Distance for DNA Storage. arXiv preprint arXiv:1601.06885.

•Jain, S., Farnoud, F., Schwartz, M., & Bruck, J. (2016). Duplication-Correcting Codes for Data Storage in the DNA of Living Organisms.

Dixita Limbachiya, Manish K Gupta and Vaneet Aggarwal, “ Family of Constrained Codes for Archival DNA Data Storage” IEEE Communcition Letters, pp 1972 - 1975 Volume: 22 , Issue 10 , Oct. 2018

•Zhirnov, V., Zadegan, R. M., Sandhu, G. S., Church, G. M., & Hughes, W. L. (2016). Nucleic acid memory. Nature Materials, 15(4), 366-370.

•Bornholt, J., Lopez, R., Carmean, D. M., Ceze, L., Seelig, G., & Strauss, K. (2016). A DNA-Based Archival Storage System. Microsoft Research

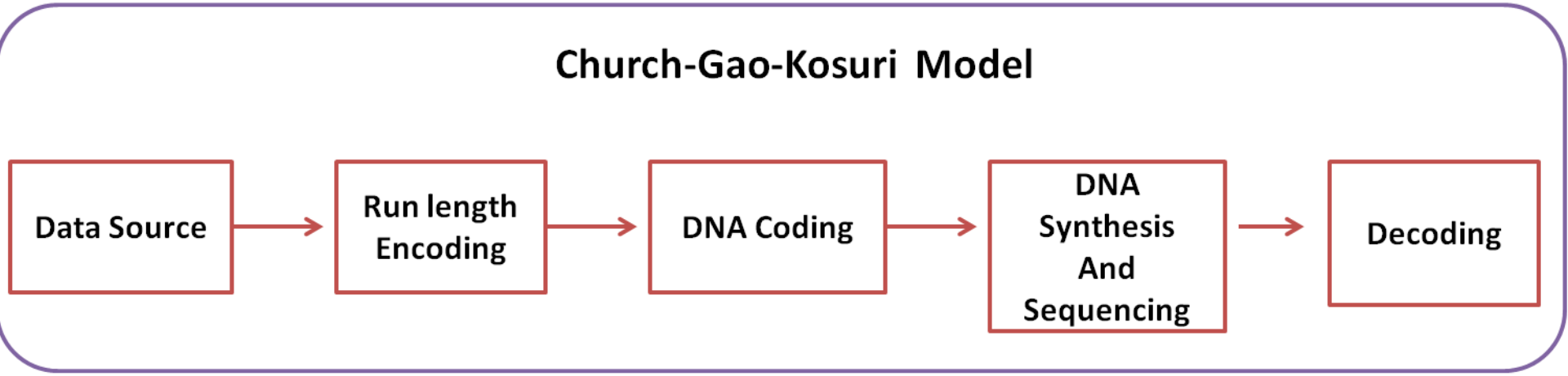
•Mayer, C., McInroy, G. R., Murat, P., Van Delft, P., & Balasubramanian, S. (2016). An Epigenetics-Inspired DNA-Based Data Storage System.Angewandte Chemie International Edition.

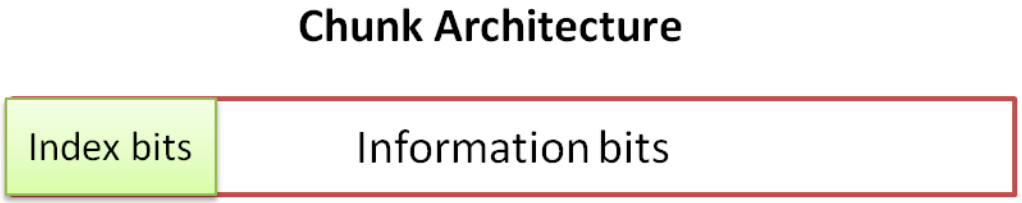
Blawat, Meinolf, et al. "Forward Error Correction for DNA Data Storage."Procedia Computer Science 80 (2016): 1011-1022.

Erlich, Yaniv, and Dina Zielinski. "Capacity-approaching DNA storage." bioRxiv (2016): 074237.

Laure, Chloé, et al. "Coding in 2D: Using Intentional Dispersity to Enhance the Information Capacity of Sequence-Coded Polymer Barcodes." Angewandte Chemie 128.36 (2016): 10880-10883.

# Archival DNA Storage Model



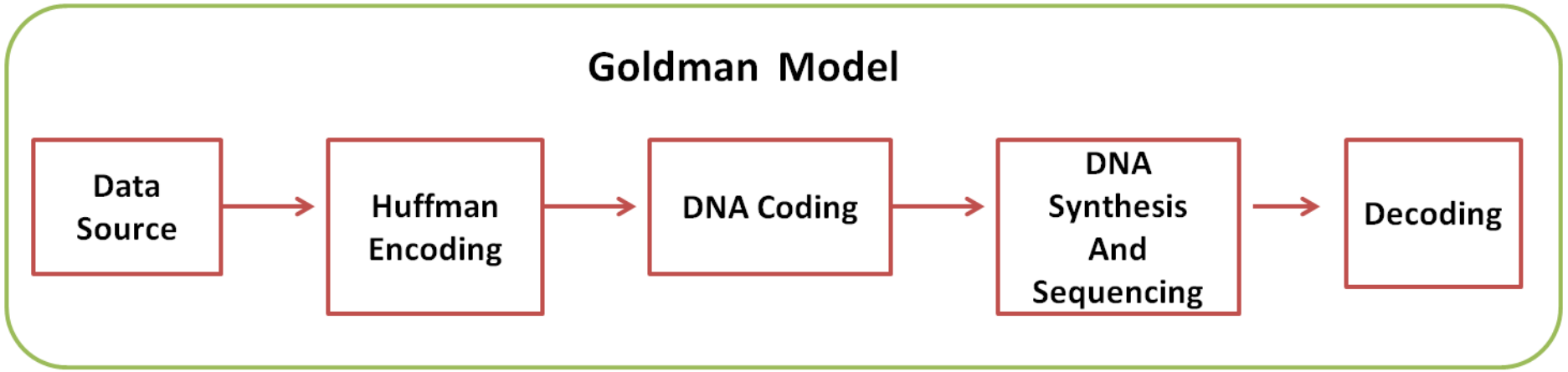



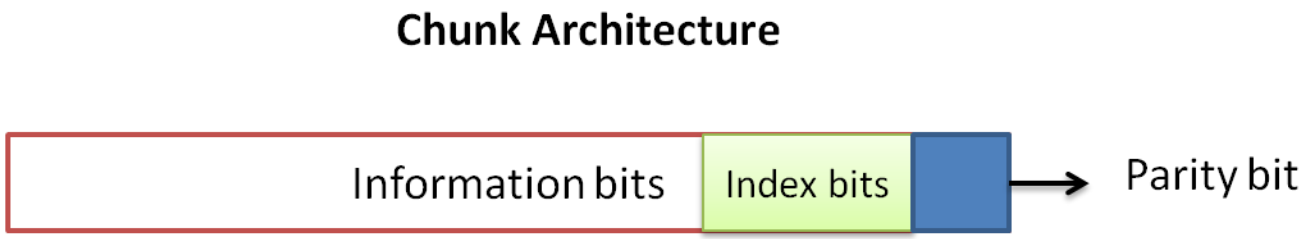

- Length of chunk = 115
- Number of chunks = 54898
- No error correction

stored a 5.27 MB book

Church, George M., Yuan Gao, and Sriram Kosuri. "Next-generation digital information storage in DNA." Science 337.6102 (2012): 1628-1628.

# Archival DNA Storage Model




- Length of chunk = 117
- Number of chunks = 153335
- Four-fold redundancy added for error correction

Goldman, Nick, et al. "Towards practical, high-capacity, low-maintenance information storage in synthesized DNA." Nature 494.7435 (2013): 77-80.

# DNA Cloud 1.0:
# A Tool for Storing BIG Data on DNA

http://www.guptalab.org/dnacloud

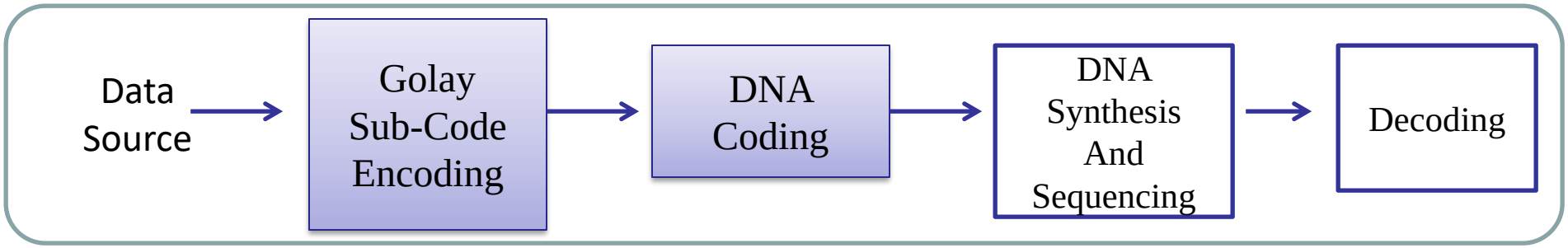
Shalin Shah

Dixita Limbachiya

Shalin Shah, Dixita Limbachiya, and Manish K. Gupta, DNA Cloud: A Potential Tool for Storing Big Data on DNA, In Proceedings of 11th Annual Conference on Foundations of Nanoscience: Self-Assembled Architectures and Devices 2014 (FNANO 14) pp. 204-205 .

http://www.geoengineeringwatch.org/wp-content/uploads/2013/01/helix-cloud-contrail-spotted-near-moscow-russia-december-24-2012-2.jpg

# Our DNA Golay sub-code Model




- Seven families of non linear ternary codes with the (n, M, d) parameters are

$(9, 256, 3)_3$, $(11, 256, 5)_3$ , $(15, 256, 7)_3$ , $(18, 256, 9)_3$ , $(21, 256, 11)_3$ , $(24, 256, 13)_3$ and $(26, 256, 15)_3$ respectively.

- A subcode $(11, 256, 5)_3$ of ternary Golay code is used to encode the data to DNA with 2 bits flips error correction in DNA code.

- DNA data storage capacity achieved is 115 EB (Exabytes) per gram of DNA.

- Decrease in the length of the DNA that minimize the cost of the DNA storage.

- Variable length chunk architecture.

# Codewords of DNA Golay Subcode $(11, 256, 5)_3$

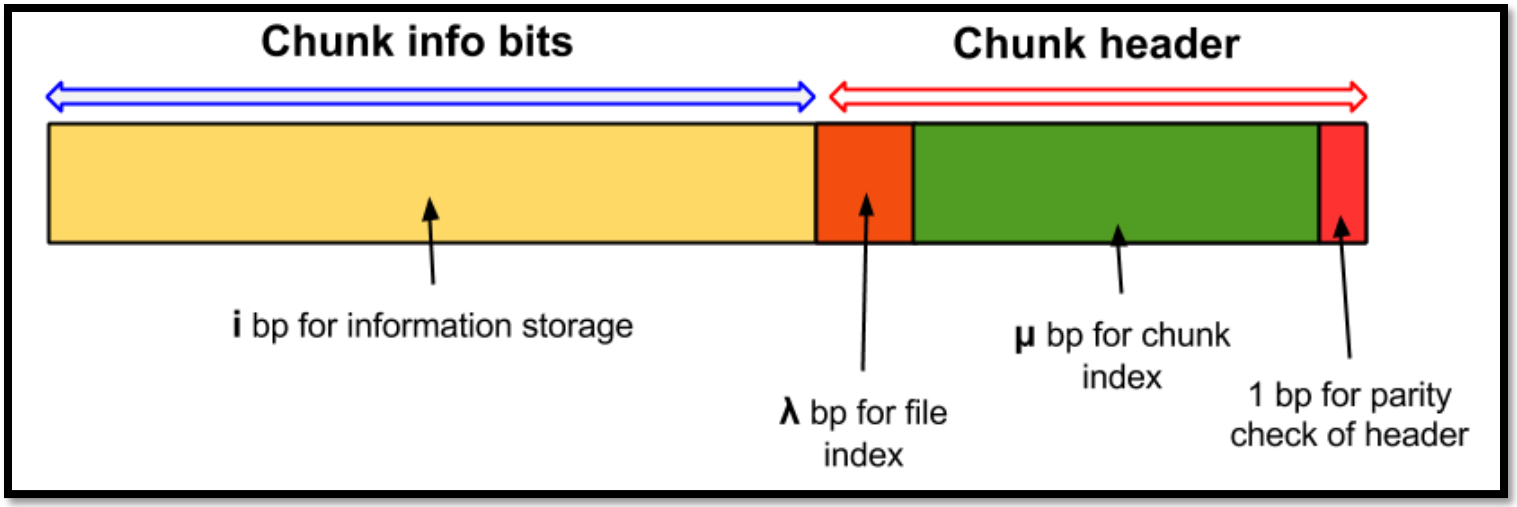
Table 1

Table 2: Codewords from subcode of Ternary Golay Code i.e. $[11, 6, 5]_3$ assigned to 256 ASCII values is given in the table.

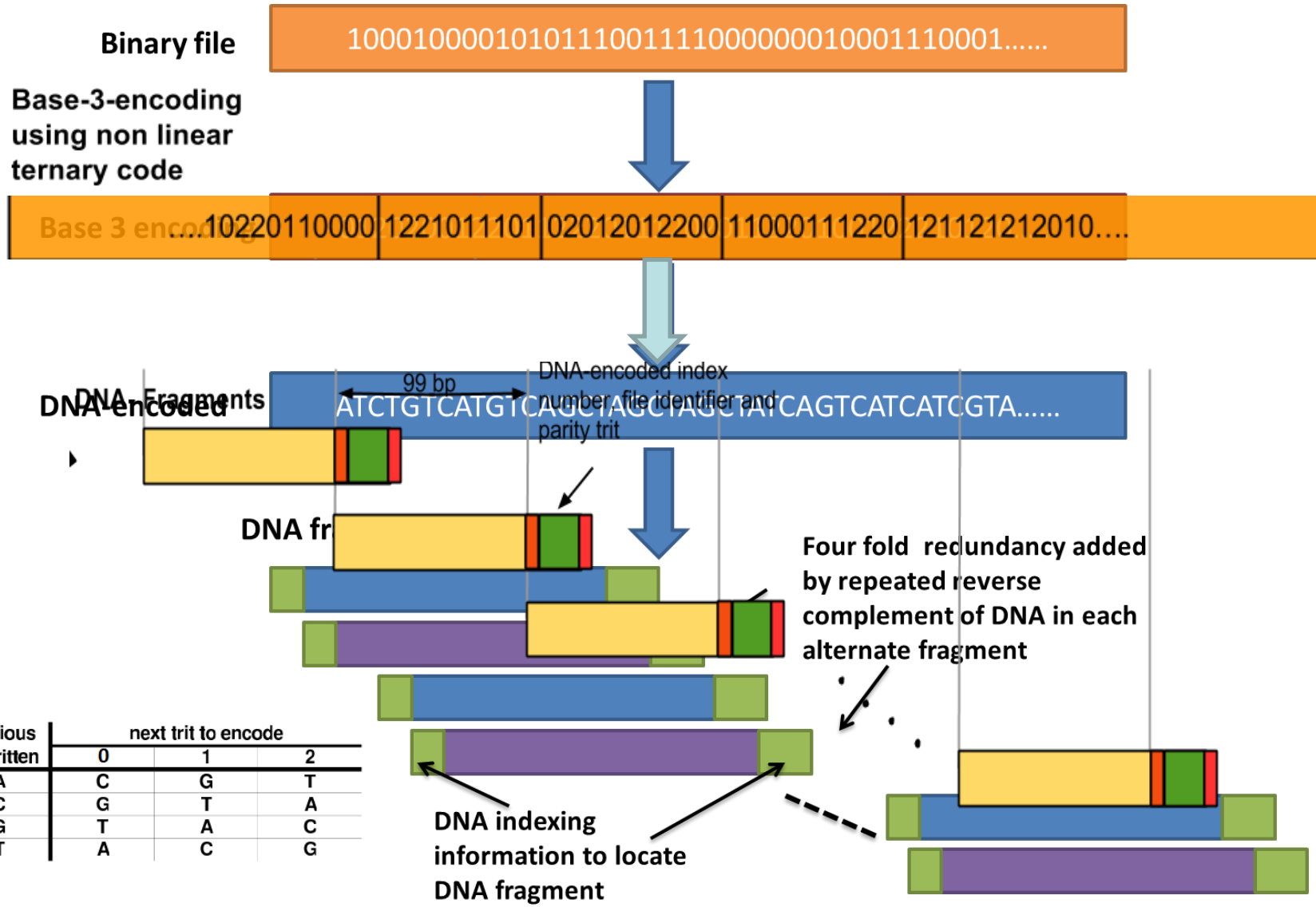
| ASCII Values | Golay codes | Weight | ASCII values | Golay codes | Weight | ASCII values | Golay codes | Weight | ASCII values | Golay codes | Weight |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 86 | 00002111202 | 6 | 170 | 00001222101 | 6 | 127 | 00020220222 | 6 | 253 | 00022001121 | 6 |
| 52 | 00021112020 | 6 | 138 | 00010110111 | 6 | 41 | 00012221010 | 6 | 86 | 00011002212 | 6 |
| 42 | 00201010122 | 6 | 100 | 00200121021 | 6 | 44 | 00202202220 | 6 | 250 | 00221200011 | 6 |
| 132 | 00220011210 | 6 | 161 | 00222122112 | 9 | 98 | 00211120200 | 6 | 8 | 00210201102 | 6 |
| 34 | 00212012001 | 6 | 10 | 00102020211 | 6 | 149 | 00101101110 | 6 | 87 | 00100212012 | 6 |
| 21 | 00122210100 | 6 | 74 | 00121021002 | 6 | 36 | 00120102201 | 6 | 69 | 00112100022 | 6 |
| 177 | 00111211221 | 9 | 20 | 00110022120 | 6 | 213 | 02012212122 | 9 | 163 | 02011020021 | 6 |
| 229 | 02010101220 | 6 | 255 | 02002102011 | 6 | 197 | 02001210210 | 6 | 133 | 02000021112 | 6 |
| 252 | 02022022200 | 6 | 26 | 02021100102 | 6 | 173 | 02020211001 | 6 | 151 | 02210222211 | 9 |
| 82 | 02212000110 | 6 | 75 | 02211111012 | 9 | 37 | 02200112100 | 6 | 166 | 02202220002 | 6 |
| 191 | 02201001201 | 6 | 88 | 02220002022 | 6 | 63 | 02222110221 | 9 | 68 | 02221221120 | 9 |
| 150 | 02111202000 | 6 | 76 | 02110010202 | 5 | 4 | 02112121101 | 9 | 154 | 02101122222 | 9 |
| 234 | 02100200121 | 6 | 22 | 02102011020 | 6 | 162 | 02121012111 | 9 | 105 | 02120120010 | 6 |
| 102 | 02122201212 | 9 | 171 | 01021121211 | 9 | 104 | 01020202110 | 6 | 169 | 01022010012 | 6 |
| 196 | 01011011100 | 6 | 208 | 01010122002 | 6 | 84 | 01012200201 | 6 | 130 | 01001201022 | 6 |
| 146 | 01000012221 | 6 | 72 | 01002120120 | 6 | 16 | 01222101000 | 6 | 66 | 01221212202 | 9 |
| 24 | 01220020101 | 6 | 106 | 01212021222 | 9 | 223 | 01211102121 | 9 | 58 | 01210210020 | 6 |
| 137 | 01202211111 | 9 | 73 | 01201022010 | 6 | 101 | 01200100212 | 6 | 168 | 01120111122 | 9 |
| 181 | 01122222021 | 9 | 175 | 01121000220 | 6 | 251 | 01110001011 | 6 | 40 | 01112112210 | 9 |
| 140 | 01111220112 | 9 | 17 | 01100221200 | 6 | 83 | 01102002102 | 6 | 254 | 01101110001 | 6 |
| 240 | 20121202122 | 9 | 214 | 20120010021 | 6 | 53 | 20122121220 | 9 | 202 | 20111222011 | 9 |
| 25 | 20110200210 | 6 | 18 | 20112011112 | 9 | 247 | 20101012200 | 6 | 174 | 20100120102 | 6 |
| 112 | 20102201001 | 6 | 89 | 20022212211 | 9 | 210 | 20021020110 | 6 | 217 | 20020101012 | 6 |
| 248 | 20012102100 | 6 | 194 | 20011210002 | 6 | 182 | 20010021201 | 6 | 80 | 20002022022 | 6 |
| 79 | 20001100221 | 6 | 195 | 20000211120 | 6 | 12 | 20220222000 | 6 | 209 | 20222000202 | 6 |
| 165 | 20221111101 | 9 | 245 | 20210112222 | 9 | 2 | 20212220121 | 9 | 81 | 20211001020 | 6 |
| 38 | 20200002111 | 6 | 141 | 20202110010 | 6 | 211 | 20201221212 | 9 | 239 | 22100111211 | 9 |
| 95 | 22102222110 | 9 | 43 | 22101000012 | 6 | 224 | 22120001100 | 6 | 203 | 22122112002 | 9 |
| 145 | 22121220201 | 9 | 147 | 22110221022 | 9 | 19 | 22112002221 | 9 | 50 | 22111110120 | 9 |
| 136 | 22001121000 | 6 | 107 | 22000202202 | 6 | 134 | 22002010101 | 6 | 109 | 22021011222 | 9 |
| 153 | 22020122121 | 9 | 148 | 22022200020 | 6 | 205 | 22011201111 | 9 | 212 | 22010012010 | 6 |
| 54 | 22012120212 | 9 | 241 | 22202101122 | 9 | 156 | 22201212021 | 9 | 115 | 22200020220 | 6 |
| 116 | 22222021011 | 9 | 78 | 22221102210 | 9 | 67 | 22220210112 | 9 | 70 | 22212211200 | 9 |
| 178 | 22211022102 | 9 | 159 | 22210100001 | 6 | 142 | 21112020000 | 6 | 92 | 21111101202 | 9 |
| 48 | 21110212101 | 9 | 90 | 21102210222 | 9 | 218 | 21101021121 | 9 | 126 | 21100102020 | 6 |
| 39 | 21122100111 | 9 | 219 | 21121211010 | 9 | 167 | 21120022212 | 9 | 114 | 21010000122 | 6 |
| 172 | 21012111021 | 9 | 14 | 21011222220 | 9 | 120 | 21000220011 | 6 | 139 | 21002001210 | 6 |
| 160 | 21001112112 | 9 | 33 | 21020110200 | 6 | 179 | 21022221102 | 9 | 117 | 21021002001 | 6 |
| 225 | 21211010211 | 9 | 129 | 21210121110 | 9 | 183 | 21212202012 | 9 | 230 | 21201200100 | 6 |
| 35 | 21200011002 | 6 | 93 | 21202122201 | 9 | 6 | 21221120022 | 9 | 32 | 21220201221 | 9 |
| 56 | 21222012120 | 9 | 158 | 10212101211 | 9 | 185 | 10211212110 | 9 | 47 | 10210020012 | 6 |
| 143 | 10202021100 | 6 | 123 | 10201102002 | 6 | 204 | 10200210201 | 6 | 242 | 10222211022 | 9 |
| 111 | 10221022221 | 9 | 103 | 10220100120 | 6 | 108 | 10110111000 | 6 | 9 | 10112222202 | 9 |
| 65 | 10111000101 | 6 | 249 | 10100001222 | 6 | 13 | 10102112121 | 9 | 180 | 10101220020 | 6 |
| 226 | 10120221111 | 9 | 144 | 10122002010 | 6 | 15 | 10121110212 | 9 | 57 | 10011121122 | 9 |
| 128 | 10010202021 | 6 | 135 | 10012010220 | 6 | 243 | 10001011011 | 6 | 190 | 10000122210 | 6 |
| 207 | 10002200112 | 6 | 77 | 10021201200 | 6 | 45 | 10020012102 | 6 | 91 | 10022120001 | 6 |
| 192 | 12221010000 | 6 | 186 | 12220121202 | 9 | 216 | 12222202101 | 9 | 97 | 12211200222 | 9 |
| 118 | 12210011121 | 9 | 246 | 12212122020 | 9 | 215 | 12201120111 | 9 | 51 | 12200201010 | 6 |
| 206 | 12202012212 | 9 | 184 | 12122020122 | 9 | 227 | 12121101021 | 9 | 233 | 12120212220 | 9 |
| 237 | 12112210011 | 9 | 188 | 12111021210 | 9 | 113 | 12110102112 | 9 | 49 | 12102100200 | 6 |
| 201 | 12101211102 | 9 | 155 | 12100022001 | 6 | 222 | 12020000211 | 6 | 231 | 12022111110 | 9 |
| 5 | 12021222012 | 9 | 27 | 12010220100 | 6 | 131 | 12012001002 | 6 | 164 | 12011112201 | 9 |
| 3 | 12000110022 | 6 | 46 | 12002221221 | 9 | 119 | 12001002120 | 6 | 28 | 11200222122 | 9 |
| 176 | 11202000021 | 6 | 23 | 11201111220 | 9 | 64 | 11220112011 | 9 | 157 | 11222220210 | 9 |
| 187 | 11221001112 | 9 | 244 | 11210002200 | 6 | 238 | 11212110102 | 9 | 96 | 11211221001 | 9 |
| 235 | 11101202211 | 9 | 60 | 11100010110 | 6 | 1 | 11102121012 | 9 | 110 | 11121122100 | 9 |
| 200 | 11120200002 | 6 | 221 | 11122011201 | 9 | 99 | 11111012022 | 9 | 31 | 11110120221 | 9 |
| 198 | 11112201120 | 9 | 193 | 11002212000 | 6 | 125 | 11001020202 | 6 | 124 | 11000101101 | 6 |
| 152 | 11022102222 | 9 | 122 | 11021210121 | 9 | 71 | 11020021020 | 6 | 94 | 11012022111 | 9 |
| 220 | 11011100010 | 6 | 29 | 11010211212 | 9 | 199 | 00000201211 | 5 | 61 | 00000102122 | 5 |
| 11 | 00002012110 | 5 | 228 | 00002210021 | 5 | 62 | 00001021220 | 5 | 55 | 00001120012 | 5 |
| 121 | 00020121100 | 5 | 7 | 00020022011 | 5 | 30 | 00022100210 | 5 | 232 | 00022202002 | 5 |
| 189 | 00021010201 | 5 | 59 | 00010212200 | 5 | 236 | 00010011022 | 5 | 0 | 00000000000 | 0 |

# DNA Golay Subcode Model

## Chunk Architecture

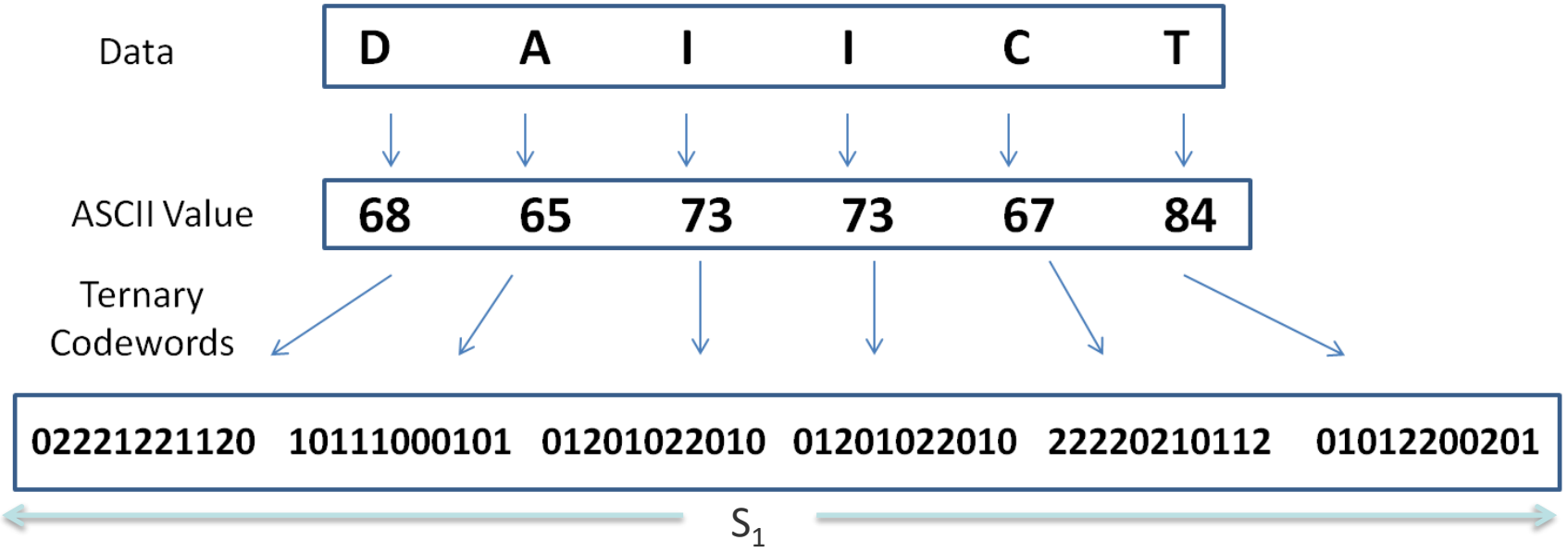


# DGold Binary Code Encoding




| previous nt written | next trit to encode | | |
|---|---|---|---|
| | 0 | 1 | 2 |
| A | C | G | T |
| C | G | T | A |
| G | T | A | C |
| T | A | C | G |

# DNA Golay Subcode Example




- $S_2$ = 00202202220 (Ternary Golay Subcode for separator , )
- $S_4$ = 01210210020 (Ternary Golay Subcode for separator : )
- $S_3$ = 22222021011 21000220011 22222021011 (File Extension )
  t x t

- $S_5$ = 22012120212 22012120212 (File size N = |S1| ( here = 66 ))

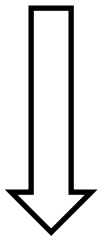
- n = |S1|+|S2|+|S3|+|S4|+|S5|
  = 66+11+33+11+22 = 143
- Padding zeros $S_6$ such that
  (n + S6) mod 99 = 0 (here $S_6$ =55)

# DNA Golay Subcode Example Cont.

Final String $S_7 = S_1.S_2.S_3.S_4.S_6.S_5$

02221221120 10111000101 01201022010 01201022010 22220210112 01012200201

00202202220 22222021011 21000220011 22222021011 01210210020 00000000000
00000000000 00000000000 00000000000 00000000000 22012120212 22012120212



| previous nt written | next trit to encode | | |
|---|---|---|---|
| | 0 | 1 | 2 |
| A | C | G | T |
| C | G | T | A |
| G | T | A | C |
| T | A | C | G |

Final DNA String $S_8$

CATGATGAGCG ACTCTACGACT AGCGACATAGT AGCGACATAGT GCATATCGAGC GACTGCACGCT
ACACATATGCG CATGTGACTCT GACGTGCGTCT GCATGTGACTC GATCGCTACAC GTACGTACGTA
CGTACGTACGT ACGTACGTACG TACGTACGTAC GTACGTACGTA TGTCAGCGCTG TGTCAGCGCTG

# DNA Golay Subcode Example Cont.

Final DNA String $S_8$

CATGATGAGCG ACTCTACGACT AGCGACATAGT AGCGACATAGT GCATATCGAGC GACTGCACGCT
ACACATATGCG CATGTGACTCT GACGTGCGTCT GCATGTGACTC GATCGCTACAC GTACGTACGTA
CGTACGTACGT ACGTACGTACG TACGTACGTAC GTACGTACGTA TGTCAGCGCTG TGTCAGCGCTG

DNA Chunks of length 99

$L_1$ CATGATGAGCG ACTCTACGACT AGCGACATAGT AGCGACATAGT GCATATCGAGC
GACTGCACGCT ACACATATGCG CATGTGACTCT GACGTGCGTCT

$L_2$ GCATGTGACTC GATCGCTACAC GTACGTACGTA CGTACGTACGT ACGTACGTACG
TACGTACGTAC GTACGTACGTA TGTCAGCGCTG TGTCAGCGCTG

# DNA Golay Subcode Example Cont.

Indexing the Chunk

- Let $i_1$ and $i_2$ be chunk index for L1 and L2 respectively. Assume $i_1=0$ and $i_2=1$
- Let Identifier of the file $ID_1 = 00$.
- Let Parity bit P= Σ odd positions in chunk index I and ID. Here $P_1 = 0+0 = 0$ and $P_2 = 1+0=1$

Chunk Index = ID.i.P
Chunk index 1 = 0000 and Chunk index 2 = 0011
Golay sub code for Chunk 1 = CGTA and Chunk 2 = CGAG

$DNA\ Chunk_1$ CATGATGAGCG ACTCTACGACT AGCGACATAGT AGCGACATAGT GCATATCGAGC
GACTGCACGCT ACACATATGCG CATGTGACTCT GACGTGCGTCT CGTA

$DNA\ Chunk_2$ GCATGTGACTC GATCGCTACAC GTACGTACGTA CGTACGTACGT ACGTACGTACG
TACGTACGTAC GTACGTACGTA TGTCAGCGCTG TGTCAGCGCTG CGAG

# DNA Cloud 2.0:
# A Tool for Storing Elephants on DNA

http://www.guptalab.org/dnacloud

Vijay Dhameliya

Dixita Limbachiya

Madhav Khakhar

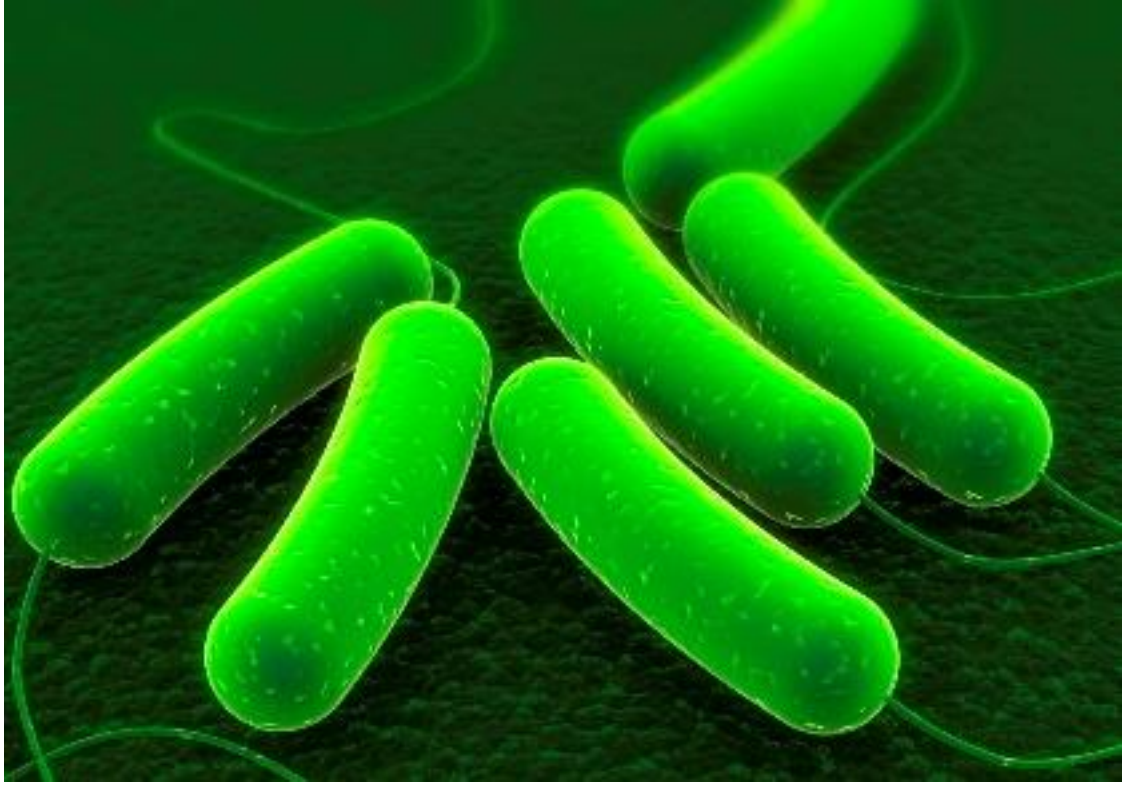
Dixita Limbachiya, Vijay Dhameliya, Madhav Khakhar, Manish K Gupta,
"On Optimal Family of Codes for Archival DNA Storage"
IEEE Proceedings of IWSDA'15 (2015): pp. 123-127

http://www.geoengineeringwatch.org/wp-content/uploads/2013/01/helix-cloud-contrail-spotted-near-moscow-russia-december-24-2012-2.jpg

# Comparison of files encoded by Goldman Encoding & DNA Golay Subcodes by using DNA Cloud

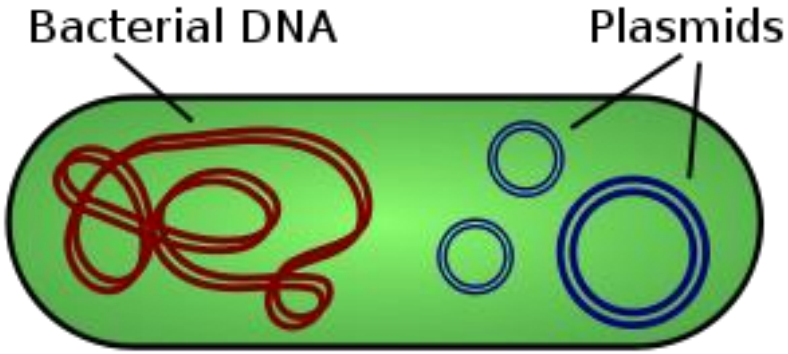
| | | | Goldman's Result | | | DNA Golay subcode Results | | |
|---|---|---|---|---|---|---|---|---|
| **Original Filename** | **File Size** | **Bytes** | **Chunk Size** | **No. of Chunks** | **No. of nucleotides** | **Chunk Size** | **No. of Chunks** | **No. of nucleotides** |
| EBI.jp2 | 179.9 KB | 184264 | 117 | 37423 | 4378491 | 112 | 20476 | 2293312 |
| MLK_excerpt_VBR_45-85.mp3 | 164.6 KB | 168539 | 117 | 34164 | 3997188 | 111 | 18728 | 2078808 |
| View_huff3.cd.new | 15.3 KB | 15646 | 117 | 3163 | 370071 | 109 | 1740 | 189660 |
| watsoncrick.pdf | 274.3 KB | 280864 | 117 | 56911 | 6658587 | 112 | 31209 | 3495408 |
| wssnt10.txt | 105.2 KB | 107738 | 117 | 21650 | 2533050 | 111 | 11973 | 1329003 |
| **Total** | **739.3 KB** | **757051** | | 153335 | 17937387 | | 84126 | 9386191 |

E. Coli



A Gut bacteria

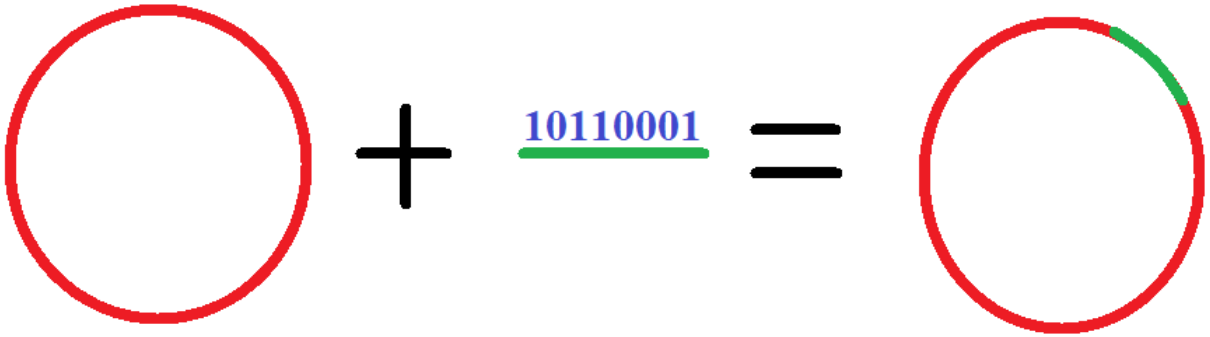
# Bacteria Based Storage Systems

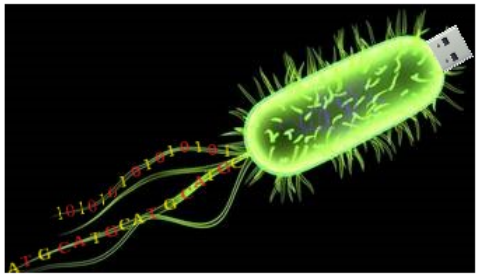


# Store all your data in your stomach!

## Bacteria Data Storage

# BacSoft

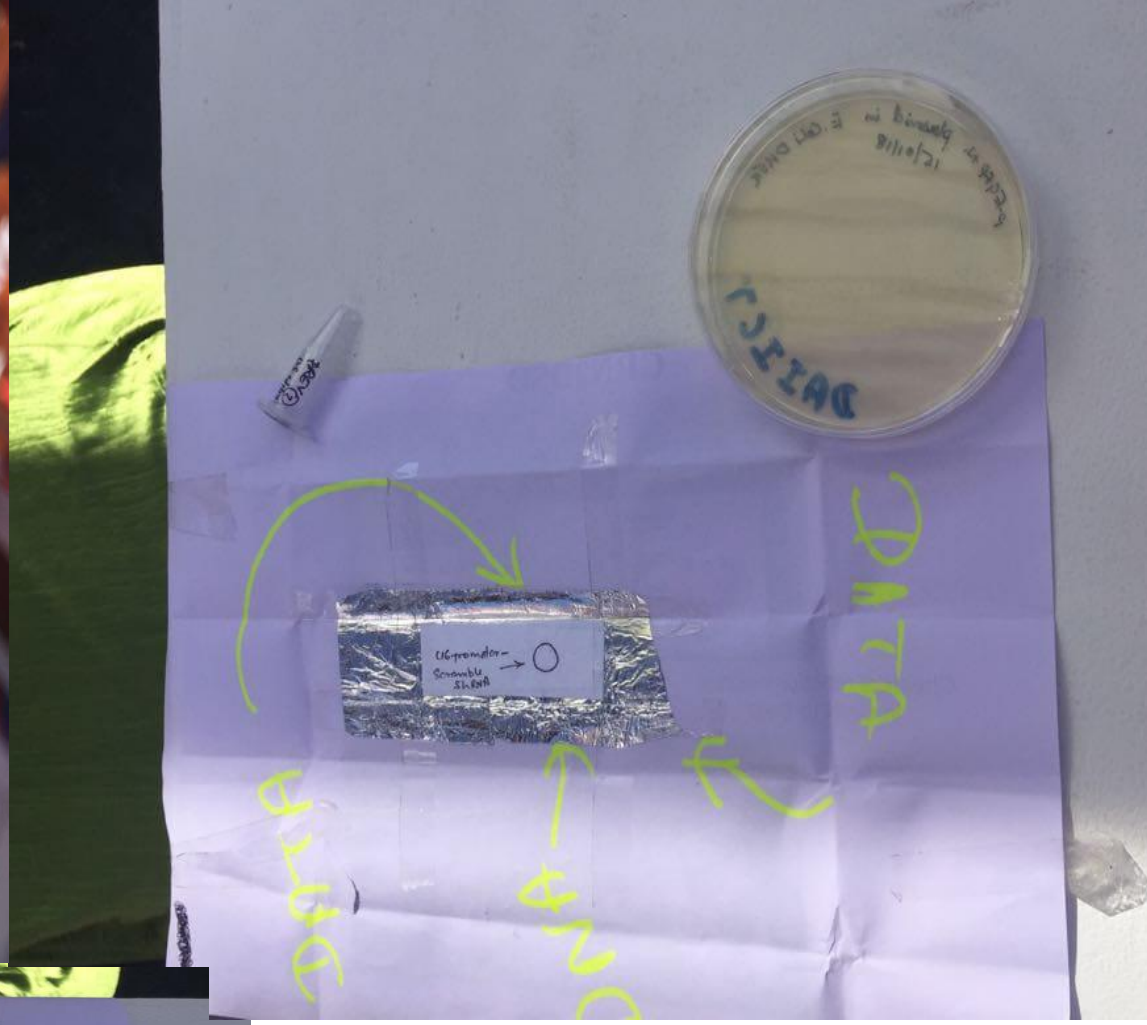



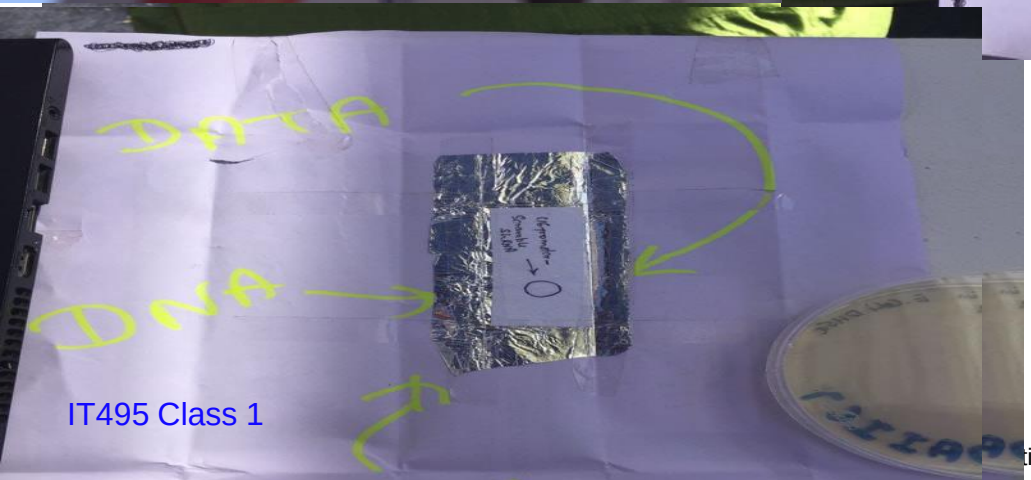



Bacteria Cloud : A Tool For Archival Data Storage

http://www.guptalab.org/bacsoft

First Software for Bacterial Data Storage




## Our Experiment

## Demo: Prime Minister India and Prime Minister Israel January 2018

# I2 T2: Indian talent times Israeli technology

# Text encoded in DNA

**I2 T2: Israel-India's ties for tomorrow**

# DNA Sequences corresponding to the above text data

AGAGTGATCGATACAGATGACGAGATCTGCTACTCTGAGCAGTAGAGCTCATCGTCTCAGATGACGAGATCTGCTACTCTGAGC
AGTAGAGCTCATCGTCTCAGATGACGAGATCTGCTACTCTAGTCATGCATATCGAGCTACGCACGAGCATGCGCTAGACTACGC
ACAGCGATACTATGAGAGACTGTGCATGTGACTCATCTCTCGATACTAGTCTCTGTCGCACTATGAGAGACTGTGATGTGTCAT
CGACTGCGTGTCATCTCTCGATACTGACAGTATATCATATAGCAGTCACTATGTCGCATACGCGCACGACACGTAGTCGTACTG
CTGCTCACGCATAGCGTCGTGATCGAGTCTCGTAGTATACTGACACTGTCGCACTAGAGCTCATCGTACACGATCGCTATCACT
GTAGTAGACTGACATCTGTATGCAGATGACGAGATGCGTATATGTGATGTGTCATCATGCACAGTCTATCACTGTAGTAGCGTC
GTGATGCACGTGTGCGCTGCGCGATGACTCTCGACAGCATAGCTGACATCGCATGATCGACGTAGCATCATATAGCAGCATGCG
CTAGAGTGCTATGCAGCACAGTCTGCAGTGCTATGCAGCTAGTACGATGTCTGAGACTAGTACGATGACATCGCATGAGCGTCG
TGATACTGTGC

For comments and feedback, contact us at mankg@guptalab.org or visit us at: Gupta Lab

*Tot

# Image



http://guptalab.org/pmdemo/image.html

# Audio

http://guptalab.org/pmdemo/audio.html

# Video


http://guptalab.org/pmdemo/video.h

# Times of India: 25 March 2018

# DATA STORAGE ON DNA
# FROM FICTION TO REALITY

*Researchers At DA-IICT Achieve Breakthrough In Technology, Hopeful Of Its Commercial Viability In Near Future*

Parth.Shastri@timesgroup.com

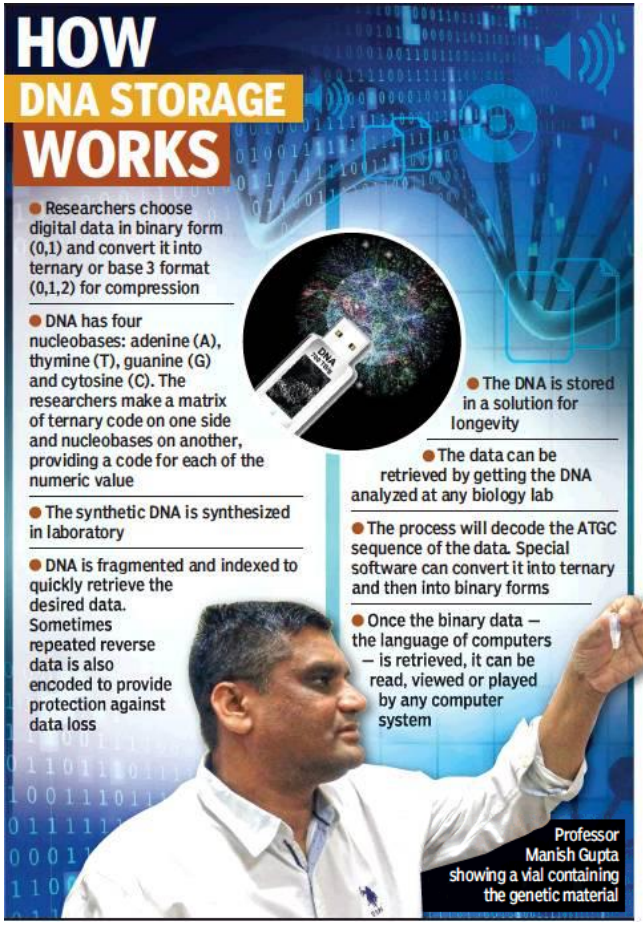
When the Indian and Israeli prime ministers inaugurated iCreate campus near Bavla in January this year, they witnessed a demonstration of how their joint message of I2T2 will be saved for at least a thousand years — on a strand of DNA!

The tiny speck of DNA in a vial not only contained their images and the text of the speech but also an audio and a video file. The encoding of data on synthetic DNA was done by Gandhinagar-based Dhirubhai Ambani Institute of Information and Communication Technology (DA-IICT). The institute is one of the few working in the field of DNA computing in India.

## Era of bio-computing

Professor Manish Gupta, spearheading the project, said that data storage on DNA has come out of the realm of science fiction and has become a reality. "Earlier, computer systems had data storage capacity of megabytes (MBs)," he said. "Now your smartphone has more capacity. With phenomenal growth of internet and social media, we are generating a very large quantity of data on a daily basis which highlights the need for a more efficient data storage system."

And the answer was found in DNA. While the concept has been worked on for decades, the major breakthrough was achieved in 2012 when a team of researchers from




## Masters of DNA software

Prof Gupta's team has made rapid strides in the field of developing open-source software for bio-computing, ranging from developing DNA Cloud, convert-

a range of DNA computing applications. Dixita Limbachiya, one of his doctoral students, and others have in a research paper said that with their method they can theoretically store 115 exabytes (EB) — equivalent to 1 million terabytes (TB) — of data on 1 gram of DNA.

### THE NEXT FRONTIER?

DA-IICT researchers have already set their eyes on the next step — storage of digital data on E. Coli bacteria. Amay Agrawal, a BTech student, is working on the project along with other students. Agarwal is currently in Germany, working on a DNA sculpting project. The living organisms with 4,290 protein coding genes provide an opportunity for coders to take the data to next generation of progeny with high production rate. One can literally save their data in their gut! The team applied for a patent for the process in 2016. The researchers want to gradually move upwards towards more complex organisms including plants.

pers in the field; today there are 30-35. "We have filed a patent for data storage with error correction mechanism," he said. "Companies like Microsoft have taken deep interest in the technology as they want to come up with a single device for the entire process. We are hopeful that the technology will be available commercially in a decade." Gupta said his team had been invited to Microsoft Research Faculty Summit in 2009 for collaborations in the area of DNA computing.

## Making it feasible

What's the hindrance at the moment? Gupta said that software is available to convert media data (any file) into DNA sequences and vice versa. "One can send DNA sequences to any biotech company, which can convert the DNA strands to physical DNA using DNA synthesizer machines and get a courier with data in a test tube," he said. "The same process can be reversed. Better and better algorithms are being developed for error-correction in the process."

The challenge now is to develop a printer-like device, or a device like

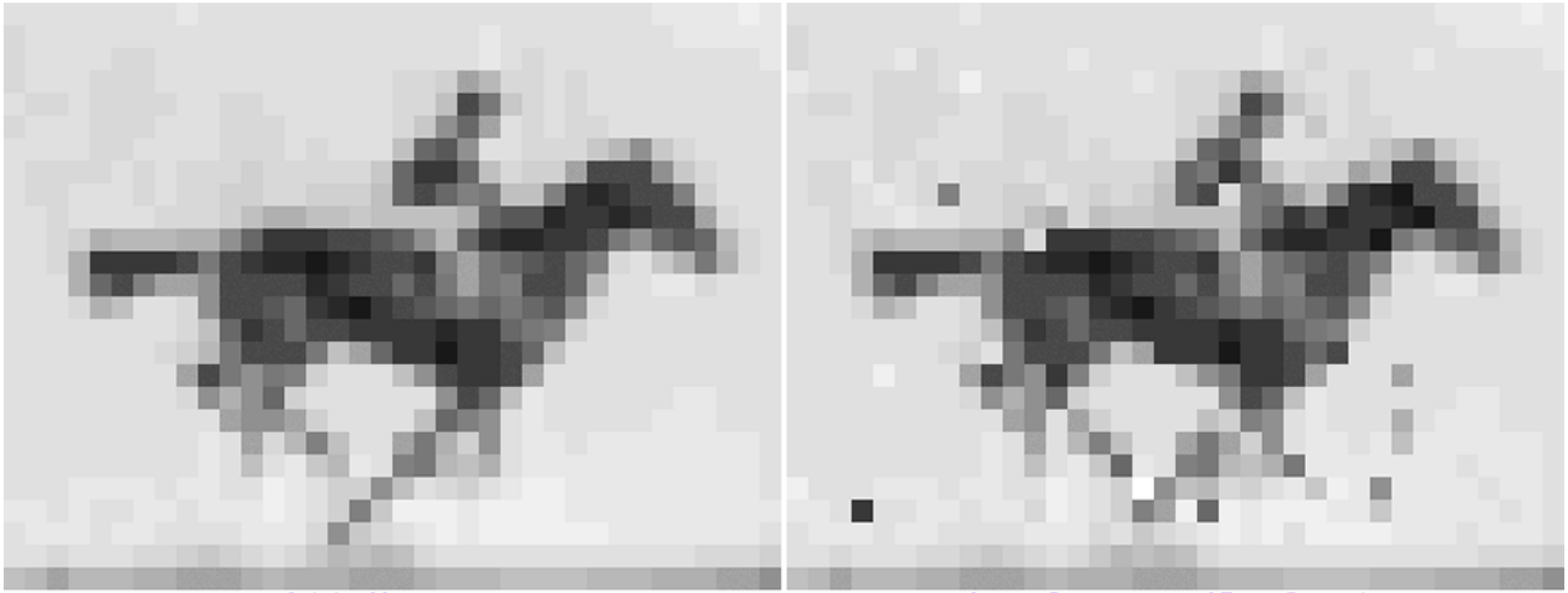
# Times of India: 7 December 2021

storage capacity of conventional magnetic data storage devices used for the past five decades. But in academic circles, the concept has remained limited to labs across the globe for the past decade due to two factors —uniformity and cost-effectiveness.


https bad/
missi .cms

# Some Recent News Headlines….

# CRISPR–Cas encoding of a digital movie into the genomes of a population of living bacteria



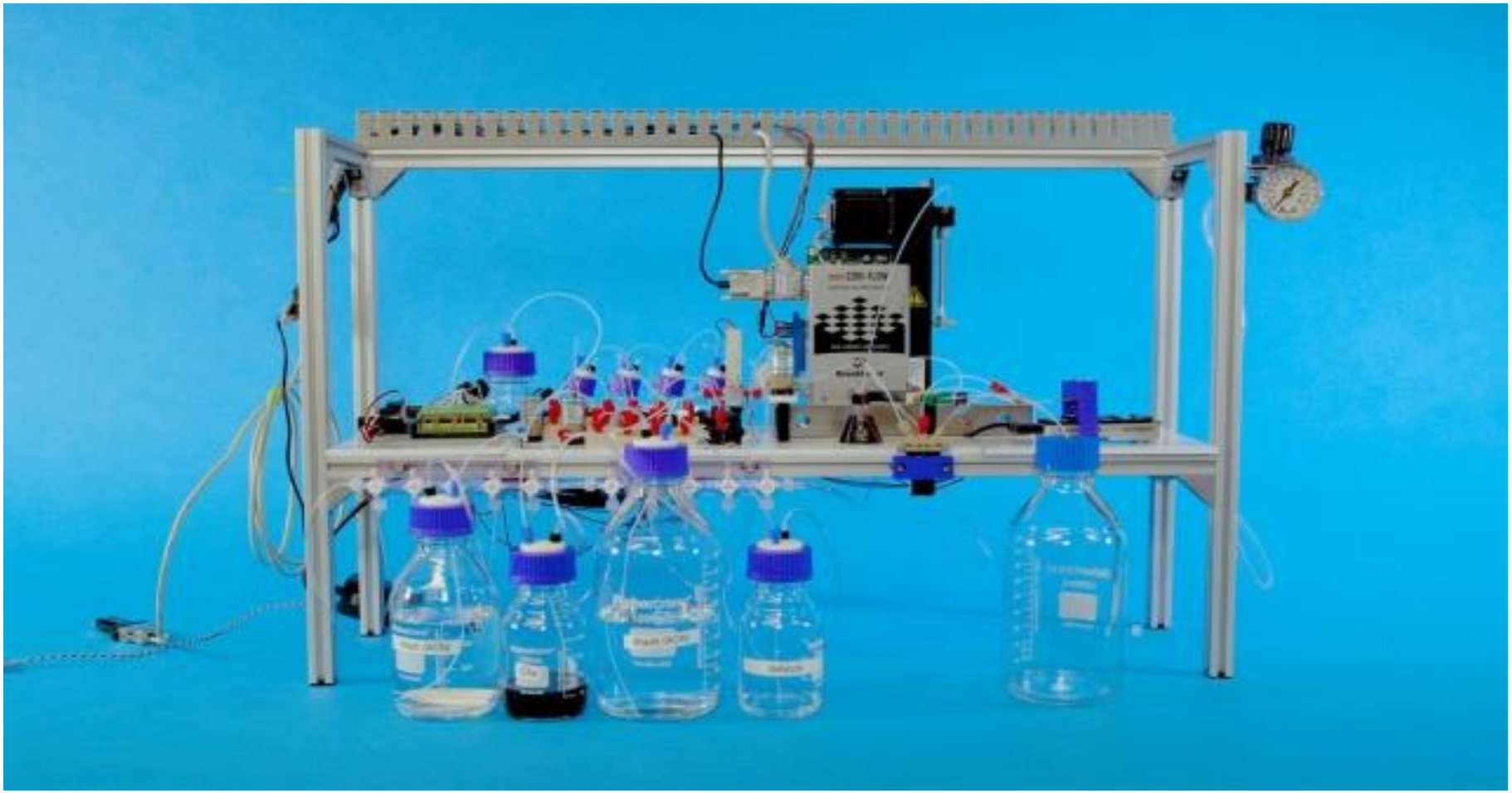
Original Image

Image Reconstructed From Bacteria

https://www.nature.com/articles/nature23017

https://www.nytimes.com/2017/07/12/science/film-clip-stored-in-dna.html

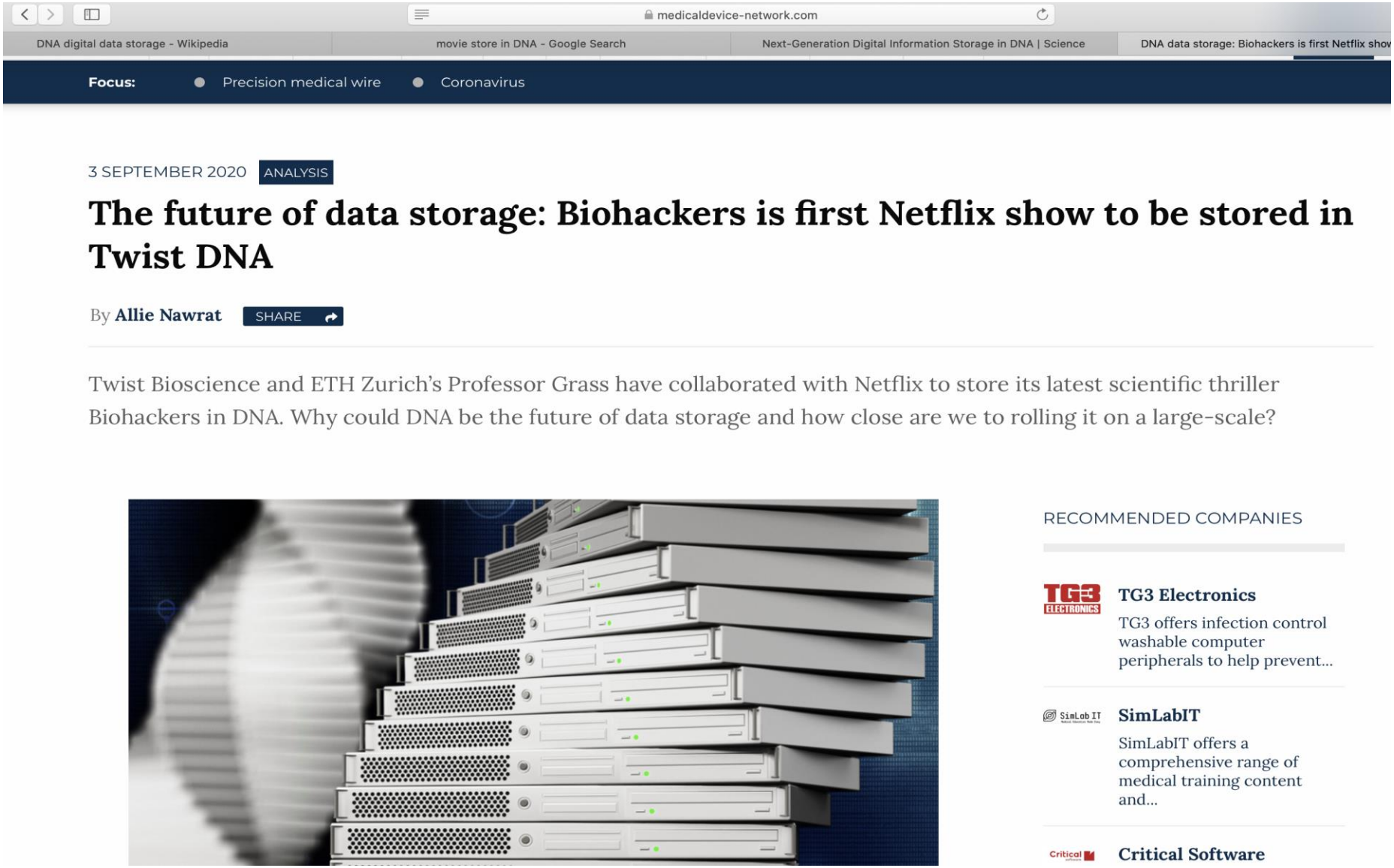
# Microsoft US$ 10,000 Prototype



HELLO (01001000 01000101 01001100 01001100 01001111 in bits) yielded approximately 1 mg of DNA – write and read took 21 hrs. (March 2019)

https://www.engadget.com/2019-03-21-microsoft-dna-storage-device.html

# 1 September 2020




https://www.medicaldevice-network.com/features/dna-data-storage/

# Headlines.. Total-RAD:Rewritable DNA Storage

MAY 21, 2012

## Totally RAD: Bioengineers create rewritable digital data storage in DNA

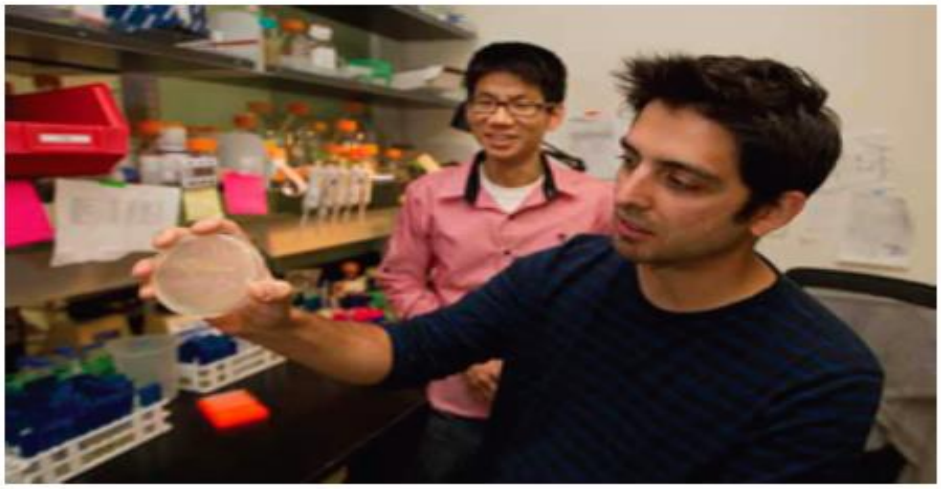
BY ANDREW MYERS

Sometimes, remembering and forgetting are hard to do.

"It took us three years and 750 tries to make it work, but we finally did it," said Jerome Bonnet, PhD, of his latest research, a method for repeatedly encoding, storing and erasing digital data within the DNA of living cells.

Bonnet, a postdoctoral scholar at Stanford University, worked with graduate student Pakpoom Subsoontorn and assistant professor Drew Endy, PhD, to reapply natural enzymes adapted from bacteria to flip specific sequences of DNA back and forth at will. All three scientists work in the Department of Bioengineering, a joint effort of the School of Engineering and the School of Medicine.

Norbert von der Groeben



Pakpoom Subsoontorn (left) and Jerome Bonnet show off a petri dish with two arrows formed of cells modified using their Recombinase Addressable Data storage device.

In practical terms, they have devised the genetic equivalent of a binary digit — a "bit" in data parlance. "Essentially, if the DNA section points in one direction, it's a zero. If it points the other way, it's a one," Subsoontorn explained.

# Start-ups to watch…










https://www.nanalyze.com/2018/09/dna-data-storage-companies/

# DNA DATA STORAGE ALLIANCES

## FOUNDERS

Illumina | Microsoft | Twist Bioscience | Western Digital

25 Member Organizations

https://dnastoragealliance.org

"DNA Data Storage Alliance, a SNIA Technology Affiliate."

Gupta Lab

https://www.snia.org/technology-focus/snia-dna-technology-affiliate

~432 exabytes per one gram of DNA



Article | Open Access | Published: 25 April 2022

# Towards practical and robust DNA-based data archiving using the yin–yang codec system

Zhi Ping, Shihong Chen, Guangyu Zhou, Xiaoluo Huang, Sha Joe Zhu, Haoling Zhang, Henry H. Lee, Zhaojun Lan, Jie Cui, Tai Chen, Wenwei Zhang, Huanming Yang, Xun Xu ✉, George M. Church ✉ & Yue Shen ✉

*Nature Computational Science* **2**, 234–242 (2022) | Cite this article

**3689** Accesses | **2** Citations | **16** Altmetric | Metrics

A preprint version of the article is available at bioRxiv.

## Abstract

DNA is a promising data storage medium due to its remarkable durability and space-efficient storage. Early bit-to-base transcoding schemes have primarily pursued information density, at the expense of introducing biocompatibility challenges or decoding failure. Here we propose a robust transcoding algorithm named the yin–yang codec, using two rules to

Download PDF

## Associated Content

**The yin–yang codec for archival DNA storage**

Manish K. Gupta

Nature Computational Science | **News & Views**

25 Apr 2022

Check for updates | news & views

DNA COMPUTING

### The yin–yang codec for archival DNA storage

A robust and reliable codec is the backbone for any digital DNA storage. A recent work introduces a codec based on ancient Chinese philosophy, yin–yang, that outperforms other codecs in terms of reliability and physical information density.

Manish K. Gupta

In the modern world, whenever we use our mobile phones to click on a picture and post it on social media accounts, we do not worry about the availability of storage, as we assume that companies such as social media platforms have enough storage space. However, data storage is a burning issue, not only for these companies, but also for our society at large, as we produce a substantial amount of data daily. More specifically, we produce data in the range of a couple of exabytes (1 exabyte = 1 billion gigabytes) every day[1], and the pace is accelerating: the Internet of Things (IoT) will potentially produce data in the order of geopbytes (1 geopbyte = 1 trillion exabytes). Ultimately, this poses immediate challenges related to cost and space: current data storage technologies will require a substantial amount of space to store our data, which can be fairly expensive. A promising avenue for improving our data storage capabilities is using DNA as a storage medium. For instance, one gram of DNA can store up to 455 exabytes of data[2], meaning that 2 grams of DNA can store the entire Internet (which takes approximately 700 exabytes), and 1 kg of DNA can store all

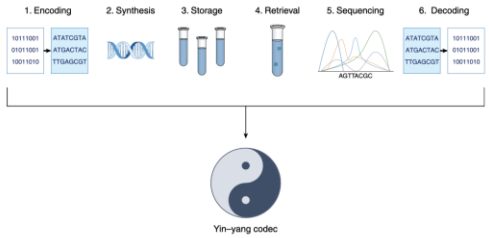



Fig. 1 | **The six steps of the DNA storage process.** Steps 1 (encoding) and 6 (decoding) are performed by the codec. Machines carry out DNA writing (synthesis) and DNA reading (sequencing).

2, the DNA strings are synthesized (written) using a machine, and then stored in test tubes (step 3). To access the data, one would need to retrieve the test tube (step 4) and use a DNA sequencer machine to read the data in the form of DNA strings (step 5). These DNA strings would then need to be further

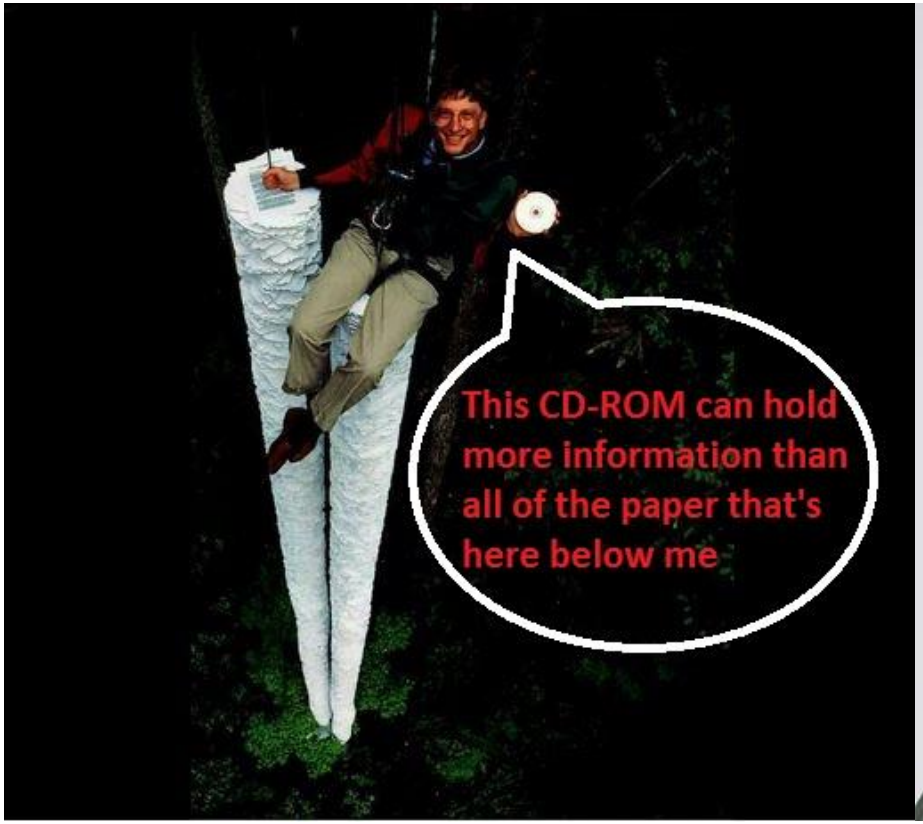
The yin–yang codec[3] is motivated by Goldman's rotating encoding strategy (where binary strings are converted to ternary strings using Huffman coding and then using a specific table[4] to get homopolymer-free DNA strings) and the DNA fountain coding strategy (erasure codes involving

# Challenges in DNA based Storage Systems

- Storing Elephants is still costly
- Developing new error correction techniques
- Rewritable DNA storage
- Exploring other living organism for storage such as Bacteria and Plants or even Humans !

# This is how you store Elephants!

**Early**




**Now**


http://cs618427.vk.me/v618427903/15d6a/_JQnRCNg80c.jpg

# Natural Questions…

# Storing life on life…

- Can we store data on synthetic DNA?
- Can we store data on plants?
- Can we store data on proteins?
- Can we store data on bacteria?
- Can we store data on Humans?

# Tentative Course Content

Historical Introduction and Motivation of DNA storage, Basic review of elementary number theory and algebra, Overview of biology, coding theory and cryptography. Overview of digital storage. Green storge (carbon footprints of storage technologies), Basics of storage systems and their parameters. Overview of DNA synthesis (DNA writing) and DNA sequencing (DNA reading) techniques. Enzymatic DNA synthesis and Nanopore sequencing, NGS methods. DNA printers. DNA semiconductor interfaces, DNA storage architecture. DNA file systems and Operating Systems, DNA storage models and their capacities. Types of Errors in DNA storage. Archival DNA storage techniques. Rewritable DNA storage techniques. Different coding schemes of DNA storage. DNA media standards, Limitations of DNA storage. DNA codes and their properties. DNA cryptography. Basic definitions; Different schemes and approaches, Computational complexity of DNA Cryptography, Limitations of DNA Cryptography; Security of DNA Storage Systems. DNA storage of classified data. Software implementations of DNA storage systems. Overview of natural data storage mediums such as bacteria, protein and plants. Future directions and commercialization efforts of the DNA storage. Applications and use cases. Energy consumption of DNA storage, ethics.


Researchers from Microsoft and the University of Washington have demonstrated the first fully automated system to store and retrieve data in manufactured DNA — a key step in moving the technology out of the research lab and into commercial data centers.
https://aka.ms/AA4japg Read more on Microsoft Research: https://aka.ms/dna-storage/

https://youtu.be/60Gi5lqL-dA

# Collaborators on DNA Storage

Taslimarif Saiyed
Director
C-CAMP
Centre for Cellular and Molecular Platforms
NCBS, Bangalore

David M. Smith
Head
Fraunhofer IZI
Fraunhofer Institute for Cell Therapy and Immunology IZI

Vaneet Aggrawal
Professor
PURDUE UNIVERSITY

Martin Sajfutdinow
PhD scholar
Fraunhofer IZI
Fraunhofer Institute for Cell Therapy and Immunology IZI

# Support

Microsoft Research

विज्ञान एवं प्रौद्योगिकी विभाग
DEPARTMENT OF SCIENCE & TECHNOLOGY

DAAD
Deutscher Akademischer Austauschdienst
German Academic Exchange Service

kuliza

# Team at Gupta Lab

## Current Graduate(PhD) student

Sourav Deb

## Past Graduate (PhD) students

Dr. Dixita Limbachiya
LGC, Biosearch Technologies
United Kingdom

Dr. Krishna Gopal Benerjee
IIT Kanpur

# Lab Alumni

*Shikhar Kumar Gupta*

TECHNISCHE UNIVERSITÄT DRESDEN

*Foram Joshi*

TECHNISCHE UNIVERSITÄT DRESDEN

*Arnav Goyal*

*Bansari Rao*

UF UNIVERSITY of FLORIDA

*Shalin Shah*

Duke UNIVERSITY

*Muskan Kukreja*

*Naman Turakhia*

NVIDIA

*Dhaval Trivedi*

*Sudhanshu Dwivedi*

TITAN COMPANY

*Nilay Chheda*

*Priyanka Shukla*

ZS

*Sonam Jain*

Infosys

*Anshul Chaurasia*

IESEG SCHOOL OF MANAGEMENT

*Sandeep Vasani*

dgit

*Prateek Jain*

mdfirst

*Vijay Dhameliya*

ISKCON

Shreyansh Prajapati

QUALCOMM

Nikhil Agrawal

Anurag Nigam

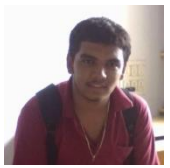
Akshay Soni

YAHOO!

Tanvi Sharma

NVIDIA

Pallav Vvas



Dhruv Raval

Parit Bansal

SIB Swiss Institute of Bioinformatics

# Lab Alumni

Sahil Sikka

Lavish Mantri

Vaibhav Devpura

Akshita Sahai

Shivam Khare

Anupam Agrawal

Srijan Anil

Kavin Parekh

Rahul Bhaskar

Akanksha Modi

Srajan Paliwal

Sauhil kansal

Ashwin Jain

Jinay Kothari

Mit Sheth

Hritik Bhardwaj

Parul Sharma

Archit Jain

Arun Agrawal

Richa Misra

Vikram Jaglan

**In science if you know what you are doing you should not be doing it.**

**In engineering if you do not know what you are doing you should not be doing it**

**Richard W. Hamming**

Email: mankg@guptalab.org

## Any Questions ?

Subscribe, like and share YouTube: https://youtube.com/c/ManishGuptamankg