

# Data organization: normalization and modeling

Brad Chapman  
Ginkgo Bioworks

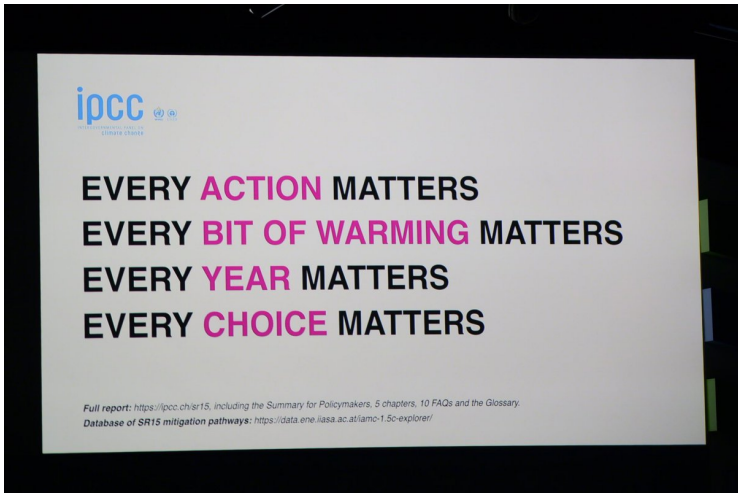
28 March 2019



**GINKGO** BIOWORKS™  
THE ORGANISM COMPANY

<https://www.ginkgobioworks.com/>

# Motivation



The image shows the cover page of the IPCC Summary for Policymakers (SR15) report. It features the IPCC logo at the top left, which includes the text 'ipcc' and 'Intergovernmental Panel on Climate Change'. Below the logo, the text 'EVERY ACTION MATTERS', 'EVERY BIT OF WARMING MATTERS', 'EVERY YEAR MATTERS', and 'EVERY CHOICE MATTERS' is displayed in large, bold, black capital letters, with the words 'ACTION', 'BIT OF WARMING', 'YEAR', and 'CHOICE' highlighted in pink. At the bottom, there is a line of text providing the full report URL: 'Full report: <https://ipcc.ch/sr15>, including the Summary for Policymakers, 5 chapters, 10 FAQs and the Glossary.' and a line for the database of SR15 mitigation pathways: 'Database of SR15 mitigation pathways: <https://data.ene.iiasa.ac.at/iamc-1.5c-explorer/>'.

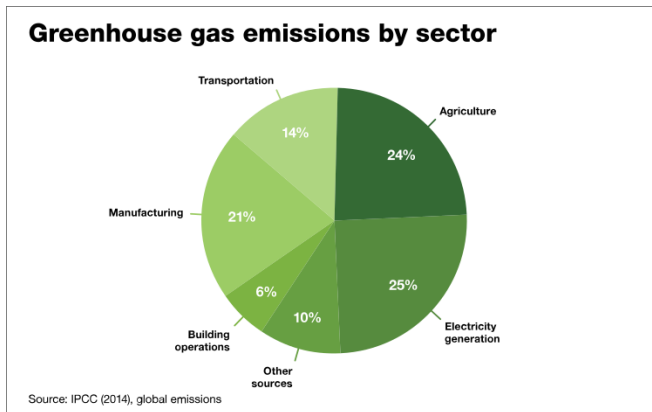
ipcc  
Intergovernmental Panel on Climate Change

**EVERY ACTION MATTERS**  
**EVERY BIT OF WARMING MATTERS**  
**EVERY YEAR MATTERS**  
**EVERY CHOICE MATTERS**

Full report: <https://ipcc.ch/sr15>, including the Summary for Policymakers, 5 chapters, 10 FAQs and the Glossary.  
Database of SR15 mitigation pathways: <https://data.ene.iiasa.ac.at/iamc-1.5c-explorer/>

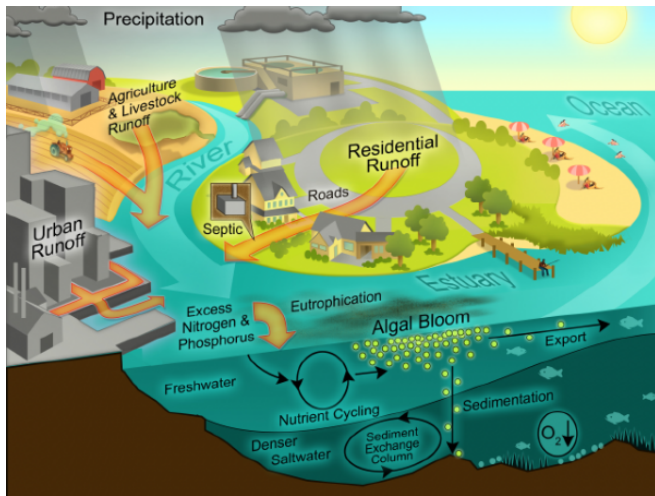
<https://www.theguardian.com/environment/2018/oct/08/global-warming-must-not-exceed-15c-warns-landmark-un-report>

# Agriculture



<https://www.gatesnotes.com/Energy/We-should-discuss-soil-as-much-as-coal>

# Agricultural fertilizer usage





We **design microbes** for more sustainable agriculture.



By **engineering the bacteria** that live in soil, we aim to nurture healthy and sustainable communities—from the ground up.

<https://joynbio.com/>

# Animal proteins

## ENVIRONMENTAL IMPACT

- 18% of global anthropogenic greenhouse gas emissions come from livestock farming. By contrast, global transportation accounts for 13%.<sup>1</sup>
- 26% Earth's ice-free surface is used for livestock farming. This represents 70% of all agricultural land.<sup>2</sup>
- 27-29% of humanity's freshwater footprint is used for the production of animal products.<sup>3</sup>
- Livestock farming is a top contributor to deforestation, land degradation, water pollution and desertification.<sup>4</sup>

<https://www.new-harvest.org/about>

# Motif Ingredients

**EVERYONE SEES A DIFFERENT FUTURE FOR FOOD.**

WE'RE MAKING INGREDIENTS  
FOR **EVERYONE.**



- Motif makes ingredients for the next generation of
- plant-based and healthy foods. We use fermentation to brew vital proteins and nutrients that power your body and please your palate. We collaborate with chefs, health experts, and food visionaries to create new building blocks for tomorrow's food revolution. Motif combines tradition, innovation, and biotechnology to help build a more sustainable future, for everyone.

<https://www.motifingredients.com/>



# Synthetic Biology as a solution



<https://grist.org/series/panic-free-gmos/>

# What does Ginkgo do?

- Make an organism that produces some product
- Multiple hosts: bacteria, yeasts, mammalian cells
- Multiple goals: find activity, improve activity, make protein...

# Design: what to make

- Source input enzymes and pathways
  - Existing biological knowledge
  - Permutations of known enzymes
- Adjust to be possible to make
  - Codon optimization
  - Restriction site/design issues
- Transcriptional machinery

# Build: how to make it

- Order parts:  
<https://www.twistbioscience.com/>
- Synthesize difficult parts: internal BioFab (<- Gen9 <- Codon)
- Combine parts into pathways: internal assembly approaches

# Test: how well did we make it

- Develop high throughput assays for compound of interest
- Scaling
  - Samples (plate-based)
  - Size (fermentation)
- Analysis: combining results from multiple inputs

# Data challenges

- Multiple levels of highly specialized work
- Custom designs, builds, assays required
- Lots of simultaneous projects with different needs

# Data goals

- Capture intent and process in a lightweight way
- Data interoperability
- Enable permissionless data analysis
- Inform decision making

# Data reality



**Nick Heltzman**

@NickDoesData



Follow

## Roles and Responsibilities:

- Automate horrible business practices
- Write ad hoc SQL as needed

## REQUIRED EXPERIENCE:

- 15 years exp deep learning in Python
- PhD thesis on Bayesian modeling
- NLP experience in 7 languages
- 10 years of creating Hadoop clusters from scratch

9:18 PM - 11 Feb 2019

<https://twitter.com/NickDoesData/status/1095160141207531520>



# Do science, not cleaning



**Vicki Boykis**

@vboykis

Follow



Just a personal anecdote, but, in the past 2 years, % of any given project:

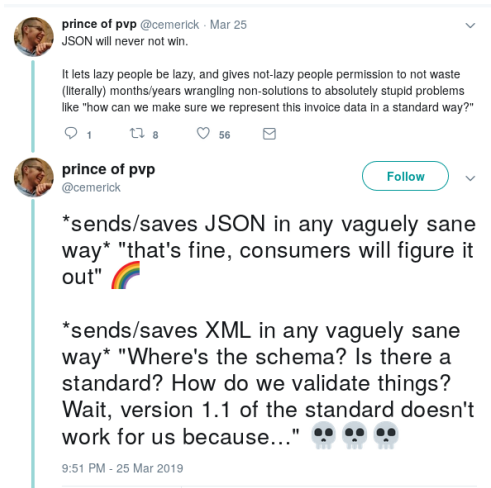
+ that involves ML: 15%

+ that involves moving, monitoring, and counting data to feed ML: 85%

8:34 AM - 15 Jan 2019

<https://twitter.com/vboykis/status/1085183529527508992>

# Interoperable structured data is hard



<https://twitter.com/cemerick/status/1110373707183210501>

# Analogous to workflows

```
#!/usr/bin/env cwl-runner
cwlVersion: v1.0
class: Workflow
doc: BWA-mapping-PE is a mapping workflow using BWA for Paired-end reads.

inputs:
  fastq1_url:
    type: string
    label: Download link of FastQ file from next generation sequencer
  fastq2_url:
    type: string
    label: Download link of FastQ file from next generation sequencer
```

```
module load bwa/0.7.15
```

```
bwa mem -t 16 -M /clusterfs/rosalind/users/makman/GATK/bwa_mem/HanXRQr1.0-20151230.fa
Anzac_Pueblo_R1_trimmed.fastq.gz Anzac_Pueblo_R2_trimmed.fastq.gz > Anzac_Pueblo.sam
# bwa mem -t 16 -M /clusterfs/rosalind/users/makman/GATK/bwa_mem
/HanXRQr1.0-20151230.fa Arikara_R1_trimmed.fastq.gz Arikara_R2_trimmed.fastq.gz >
Arikara.sam
```

# Common issue everywhere

- Most knowledge in documents, presentations, people's brains
- A lot of work to represent specialized knowledge in a structured way
- Need a lot of context in new biological areas

# Data approach

Without imposing too much extra work:

- Capture scientific intent
- Improve naming with ontologies




# Scientific intent: challenges

- High level data structures to organize projects
- Multiple ways of doing design and test
- Ad-hoc capture of intent: Jupyter, Excel, Slack







# Scientific intent: approach

- Examine characteristic analyses
- Model uncaptured data
- Provide data structures

# Example: lab workflow

 Step 31: **Reaction plates**  

*LiquidTransfer*

Containers and Data

abs340\_raw,  
Timepoint

Containers and Data

Timepoint,  
abs340\_raw:340

Containers and Data

Experiment 4061  
(Plate Reader Assay)

Containers and Data

Experiment 4062  
(Plate Reader Assay)

Containers and Data

Experiment 4064  
(Plate Reader Assay)

Containers and Data

Experiment 4066  
(Plate Reader Assay)

Containers and Data

Experiment 4067  
(Plate Reader Assay)

Containers and Data

Experiment 4070  
(Plate Reader Assay)



# Example: analysis setup

```
# Container IDs  
uninduced_plate = range(97852, 97852+4)+[97849]  
uninduced_od_plate = range(98235, 98235+5)  
induced_plate = range(98286, 98286+5)  
induced_od_plate = range(98401, 98401+5)  
reaction_plate = range(99665, 99665+5)  
glycerol_plate = range(95062, 95062+5)
```

# Example: analysis output

## 11 Files Out

In [459]: *# Writes dataframes into multi-sheet excel files*

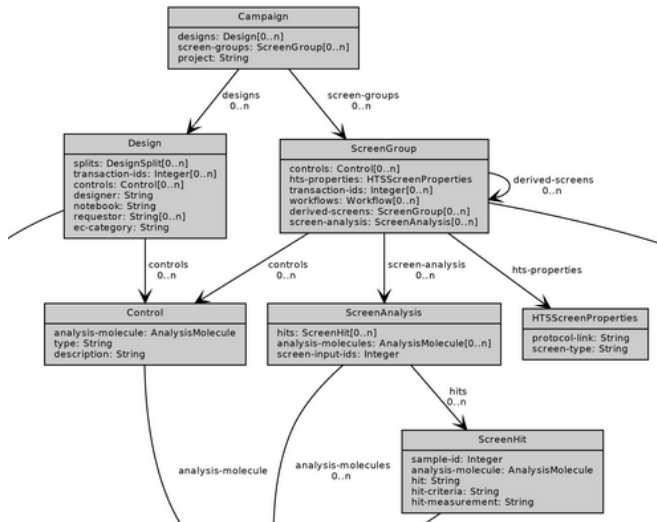
```
def save_xls(list_dfs, sheet_names, xls_path):  
    writer = pd.ExcelWriter(xls_path)  
    for i, df in enumerate(list_dfs):  
        df.to_excel(writer, sheet_names[i])  
    writer.save()  
  
write_directory = '/ginkgo/bitome/data/users/ashepard/PythonData/'
```

In [460]: `list_dfs = [df, posdf, negdf, condf, cont_stat, df10, df11, df12, df13, dfall]`  
`sheet_names = ['All Data', 'Positive Control', 'Negative Control', 'WT Control',`

# Example: missing connections

- Experimental intent
- Design to assay connection
- Assay to analysis connection
- Analysis output hits to reporting

# Data model view



# Flexible data model

```
^{:graphviz/color "gray80"}
ScreenGroup
[^{:type Control :cardinality [0 n]} controls
  ^HTSScreenProperties hts-properties
  ^{:type Integer :cardinality [0 n]} transaction-ids
  ^{:type Workflow :cardinality [0 n]} workflows
  ^{:type ScreenGroup :cardinality [0 n]} derived-screens
  ^{:type ScreenAnalysis :cardinality [0 n]} screen-analysis
]

^{:graphviz/color "gray80"}
ScreenHit
[Integer sample-id
  ^AnalysisMolecule analysis-molecule
  ^String hit
  ^String hit-criteria
  ^String hit-measurement
]
```

<https://github.com/luchiniatwork/hodur-engine>

# Translating into database storage



## Flexible

Datomic's provides the power of schema without requiring that you define everything up-front. Add attributes dynamically at any time, without worrying about fixed tables or disruptive migrations. [Learn More »](#)



## Hierarchical

A universal relation lets you handle row-oriented, column-oriented, graph, and hierarchical data in a single system without impedence. [Learn More »](#)

<https://www.datomic.com/>

# Ontologies: extend key/values

- Need
  - Standard naming
  - Flexibility
- Adopt existing ontologies
  - Avoid work of inventing
  - Contribute and extend community standards

# Adoption: descriptions and assays

- Feature descriptions: Sequence Ontology
  - <http://www.sequenceontology.org/>
- HTS: BioAssay Ontology
  - <http://bioassayontology.org/>
- Mapping to Gene Ontology
  - <http://geneontology.org/docs/download-mappings/>



# Example key names: OD600

600

OD600

OD600:600

abs600\_raw

# BioAssay Ontology

OLS > BioAssay Ontology BAO > BAO:0040015

## cell density determination

Search BAO



[http://www.bioassayontology.org/bao#BAO\\_0040015](http://www.bioassayontology.org/bao#BAO_0040015)



Cell density, typically measured by absorbance (often indicated as A600 or OD600), is often used for bacterial or fungal cultures as an indication of cell number, viability or proliferation.

Tree view

Term history

assay method component

assay method

assay design method

viability measurement method

cell density determination

Graph view

Reset tree

Show all siblings

Term Info

definition source

PubChem AID: 775

editor note

Date from BAE: 6/15/2016

<https://www.ebi.ac.uk/ols/ontologies/bao>

# Microbial Conditions Ontology

OLS > Microbial Conditions Ontology MCO > MCO:0000059

## OD600

Search MCO

[http://purl.obolibrary.org/obo/MCO\\_0000059](http://purl.obolibrary.org/obo/MCO_0000059)

an optical density that specifies the amount of light of 600 nm of wavelength the bearer is able to transmit

Tree view

Term history

entity

continuant

specifically dependent continua

quality

physical object quality

physical quality

radiation quality

electromagnetic (EM) radiation quality

optical quality

optical density

OD600

Graph view

Reset tree

Show all siblings

Term Info

[database cross reference](#)

- colombos:OD600

Term relations

**Subclass of:**

- [optical density](#)

<https://www.ebi.ac.uk/ols/ontologies/mco>

# Other issues

- Where to put units
- Which mean the same thing?

Time

Timepoint

Timepoint (second)

Timestamp

# Initial steps: data mine from existing

- Need to make practical
- Extract examples
- SciGraph
- Semi-automatically map to ontologies

[https://github.com/ginkgobioworks/  
ontology-clean](https://github.com/ginkgobioworks/ontology-clean)

# SciGraph

vocabulary Vocabulary services	
GET	<code>/vocabulary/autocomplete/{term}</code> Find a concept by its prefix
GET	<code>/vocabulary/categories</code> Get all categories
GET	<code>/vocabulary/id/{id}</code> Find a concept by its ID
GET	<code>/vocabulary/prefixes</code> Get all CURIE prefixes
GET	<code>/vocabulary/search/{term}</code> Find a concept from a term fragment

<https://github.com/SciGraph/SciGraph>

# Search example

```
$ curl 'http://localhost:9000/scigraph/vocabulary/search/time' \  
| json_pp  
{  
  "definitions" : [  
    "A unit which is a standard measure of the dimension "  
    "in which events occur in sequence."],  
    "labels" : ["time unit"],  
    "iri" : "http://purl.obolibrary.org/obo/UO_0000003"  
  ],  
  {  
    "definitions" : [  
      "A quality in which events occur in sequence."],  
      "labels" : ["time"],  
      "iri" : "http://purl.obolibrary.org/obo/PATO_0000165"  
    }  
  }  
}
```

# RDF like modeling

Ontologies map naturally to flexible  
entity - attribute - value

assay1 - sample - sample1

assay1 - sample - sample2

sample1 - time - 30

sample1 - time unit - seconds

sample1 - OD600 - 0.986

sample2 - time - 30

sample2 - time unit - seconds

sample2 - OD600 - 1.13



# Table like modeling

Many tools work with tabular formats, but have to transform arbitrary columns into final analysis tables

assay	sample	time	time unit	OD600
assay1	sample1	30	seconds	0.986
assay1	sample2	30	seconds	1.13

# Elements and Principles of Data Science



Stephanie Hicks  
@stephaniehicks

Following



So, @rdpeng and I have had a lot fun recently thinking about data analyses and the field of #datascience as a whole. We wrote down some ideas in a #preprint called Elements and Principles of Data Analysis. feedback is welcomed #statistics #AcademicTwitter [arxiv.org/abs/1903.07639](https://arxiv.org/abs/1903.07639)

3:18 PM - 20 Mar 2019

<https://arxiv.org/abs/1903.07639>

<https://twitter.com/stephaniehicks/status/1108462768099856384>

# Framework for organizing analyses

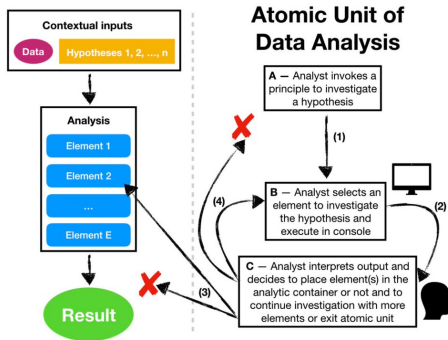


Figure 8: **The atomic unit of data analysis.** Given the contextual inputs, the data analysis is built upon atomic units of data analyses, which occurs both on the computer and in the analyst's head. In each atomic unit, the data analyst first chooses a principle to investigate a hypothesis or scientific question (Stage A). Then, the analyst alternates between Stages B and C until the analyst exits the atomic unit, either by choosing to end the line of investigation or choosing to invoke a new principle. The actions in the atomic unit are denoted by (1) the analyst selects an element to investigate a hypothesis or scientific question, which is executed in the computer or console, (2) the analyst interprets the output in his or her head, (3) the analyst decides whether or not to place the element from (1) in the data container or data product or not, which means the element is never recorded in the data container or product and the audience does not see it as part of the analysis, and (4) the analyst decides to whether to continue in the atomic unit by selecting another element to investigate the hypothesis or question or to exit the atomic unit entirely and end this line of investigation.

# In conclusion: machine learning

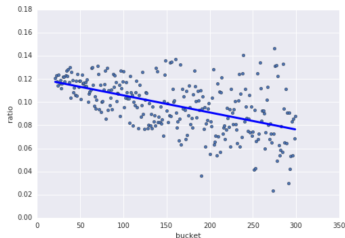
```
from sklearn import linear_model

first300 = joined[joined['bucket'] < 300]
x300 = first300['bucket'].values.reshape(-1,1)
y300 = first300['ratio'].values.reshape(-1,1)

regr = linear_model.LinearRegression()
regr.fit(x300, y300)

first300.plot(x='bucket',y='ratio',kind='scatter')
plot.plot([first300['bucket'], regr.predict(x300), color='blue', linewidth=3])

[<matplotlib.lines.Line2D at 0x7f6371a23cf8>]
```



# Real conclusion

- Ginkgo: synthetic biology for agricultural sustainability
- General data challenges
- Capture scientific intent
- Improve naming with ontologies