HEALTHY BIRTH,
GROWTH & DEVELOPMENT

knowledge integration

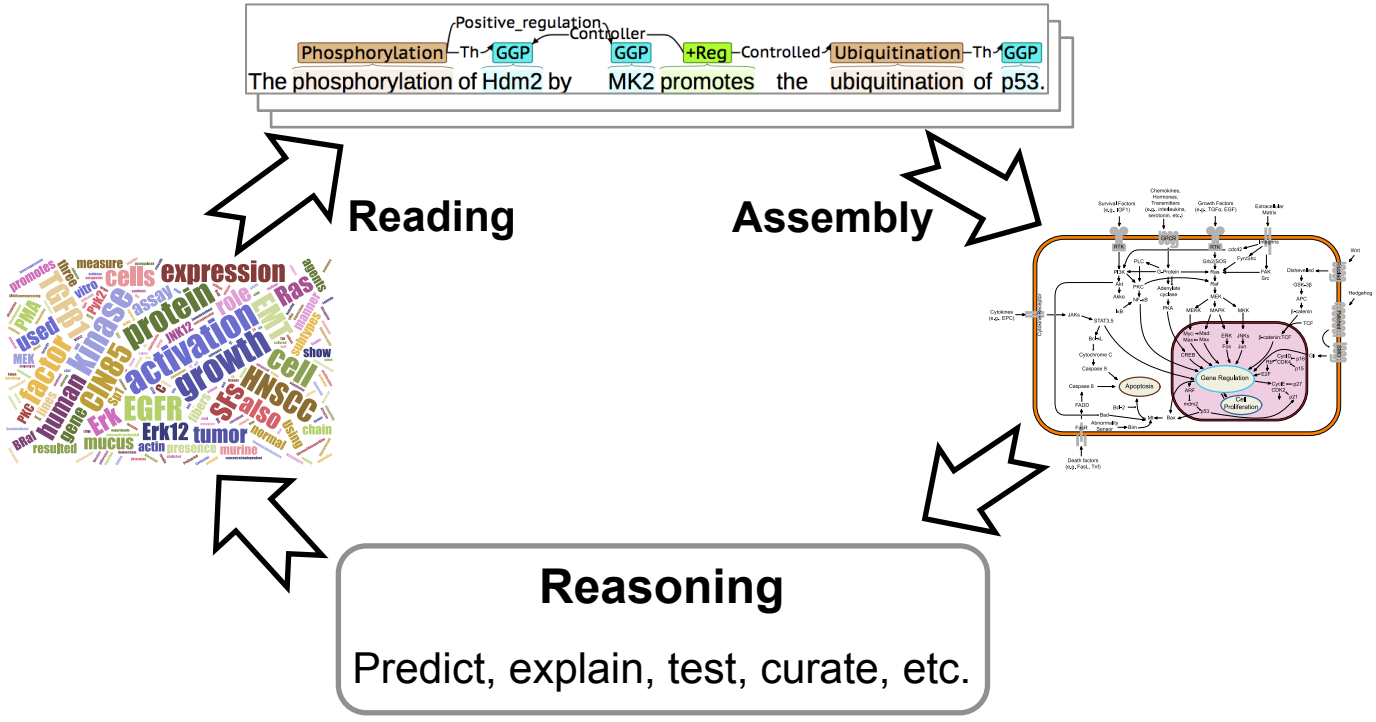
Leveraging Machines for Causal Modeling

MIHAI SURDEANU

October 25, 2016

PROBLEM STATEMENT AND DATA ANALYSIS METHODOLOGY

Goal: Build machines that help humans to model and analyze very complicated systems by reading fragmented literatures, and assembling and reasoning with models.



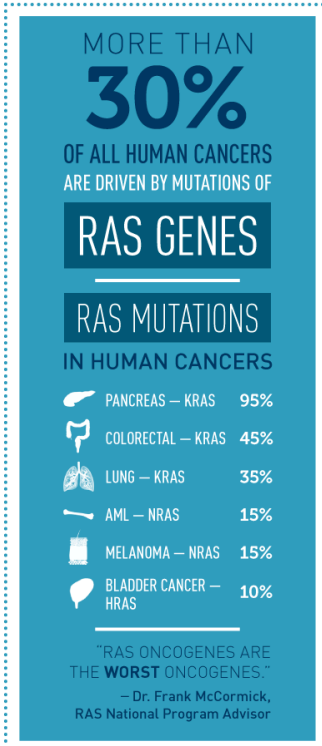
© Bill & Melinda Gates Foundation

PROBLEM STATEMENT AND DATA ANALYSIS METHODOLOGY

Goal: Build machines that help humans to model and analyze very complicated systems by reading fragmented literatures, and assembling and reasoning with models.

- **Use case 1**: modeling cancer pathways, the solution to which involves reasoning over models of these systems, which have been assembled, usually from text sources, by automatic or semi-automatic methods.

- **Use case 2**: understanding biological and socio-economic factors that impact children's health, through automated reading of literature.



© Bill & Melinda Gates Foundation

KEY FINDINGS AND INSIGHTS

- High-quality machine reading at scale is possible
  - Read all PubMed Open Access literature at 5 seconds/paper
  - Accuracy comparable to human cancer experts, at much higher throughput

- Use case 1: discovery of novel cancer driving mechanisms
  - Works for 11 different cancer types
  - New hypotheses for precision medicine for cancer treatment

- Use case 2: discovery of causal mechanisms in children's health
  - Reduced model development time from ~1 month to 2 days

© Bill & Melinda Gates Foundation

HOW WE STARTED, AND WHY

This effort is initiated by DARPA's Big Mechanism program.

"Technology to understand complicated systems"

© Bill & Melinda Gates Foundation

TECHNOLOGY TO UNDERSTAND COMPLICATED SYSTEMS

- Our reach exceeds our grasp, which can be dangerous
  - For many problems and many reasons – methodology, cognitive limitations, specialization – we have component-level, not system-level understanding.
  - Understanding means cause ➔ effect. Big data associations are not enough.

- **If humans can't understand complicated systems, then machines must help!**

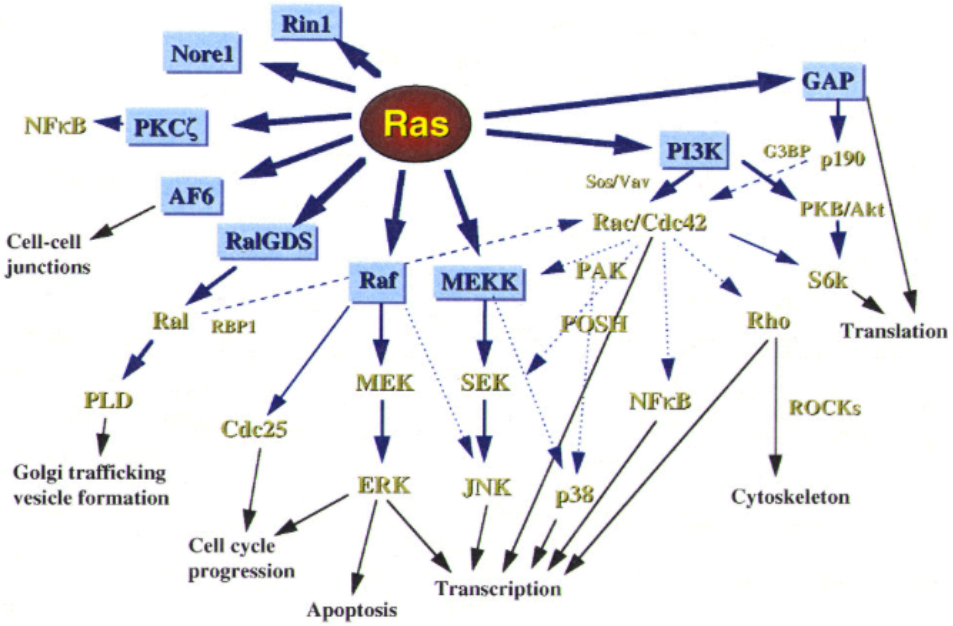
Flash crash of US t-bills; Oct. 24, 2014

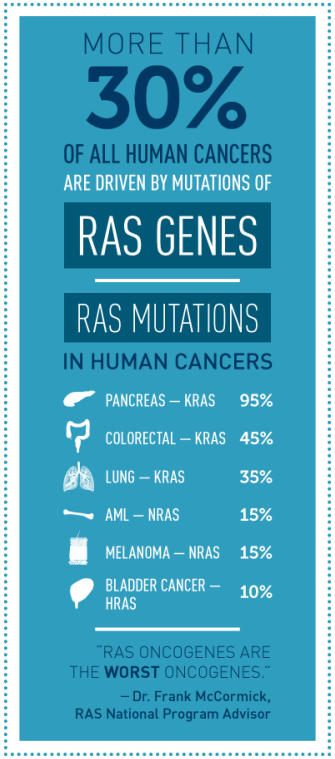
Slide courtesy of: Paul Cohen, DARPA

© Bill & Melinda Gates Foundation

WHICH COMPLICATED SYSTEM?

Big Mechanism is not cancer biology program, but molecular signaling in cancer, specifically Ras-driven cancers, is hugely complicated.





Slide courtesy of: NIH Ras Initiative

© Bill & Melinda Gates Foundation

WHICH COMPLICATED SYSTEM?

Thousands of molecular interactions within two “hops” of Ras in Pathway Commons. Tiny fragment of a model built by computers reading the literature:

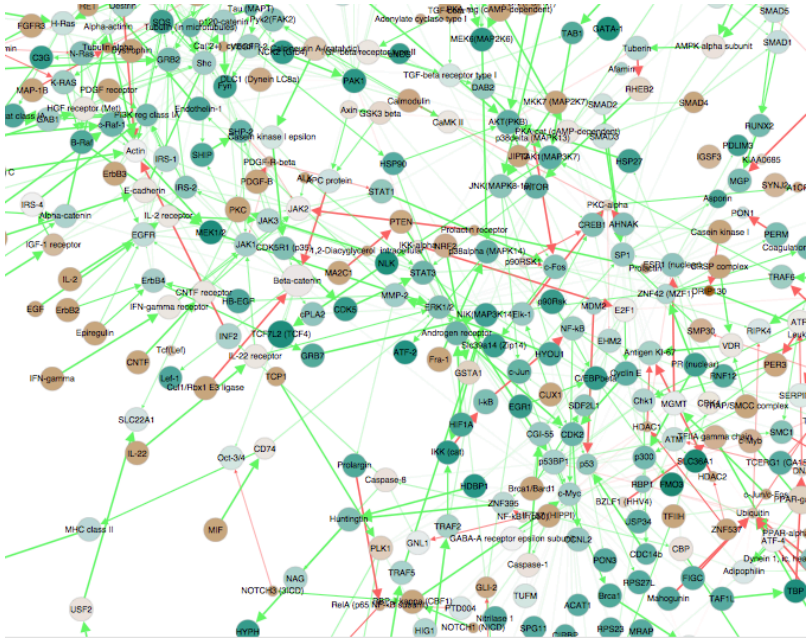


Image courtesy of: Andrey Rzhetsky, University of Chicago

© Bill & Melinda Gates Foundation

OVERVIEW

- Machine reading system
- Use case 1: discovery of novel cancer drivers
- Use case 2: modeling children's health

© Bill & Melinda Gates Foundation

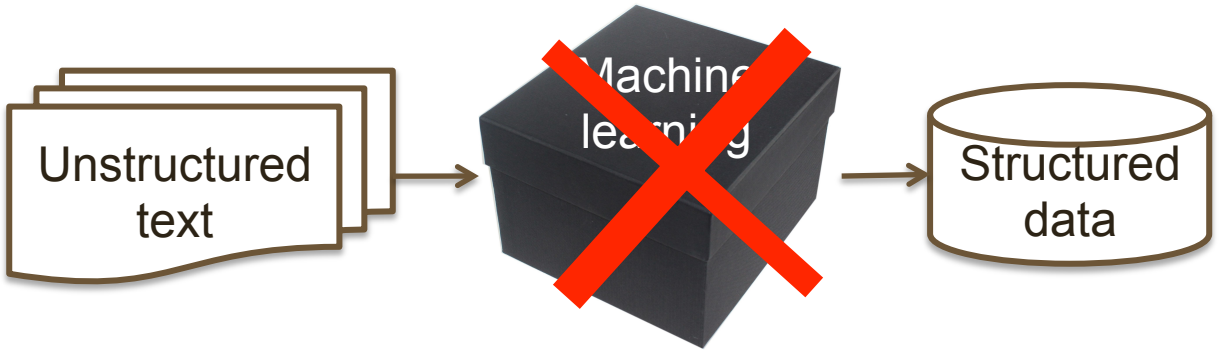
OVERVIEW

- **Machine reading system**
- Use case 1: discovery of novel cancer drivers
- Use case 2: modeling children's health

© Bill & Melinda Gates Foundation

MACHINE LEARNING DOES NOT APPLY HERE…

- "Black box" approaches to reading using machine learning not a great choice, especially in inter-disciplinary projects!



- There was no training data to match the Big Mechanism vision

© Bill & Melinda Gates Foundation

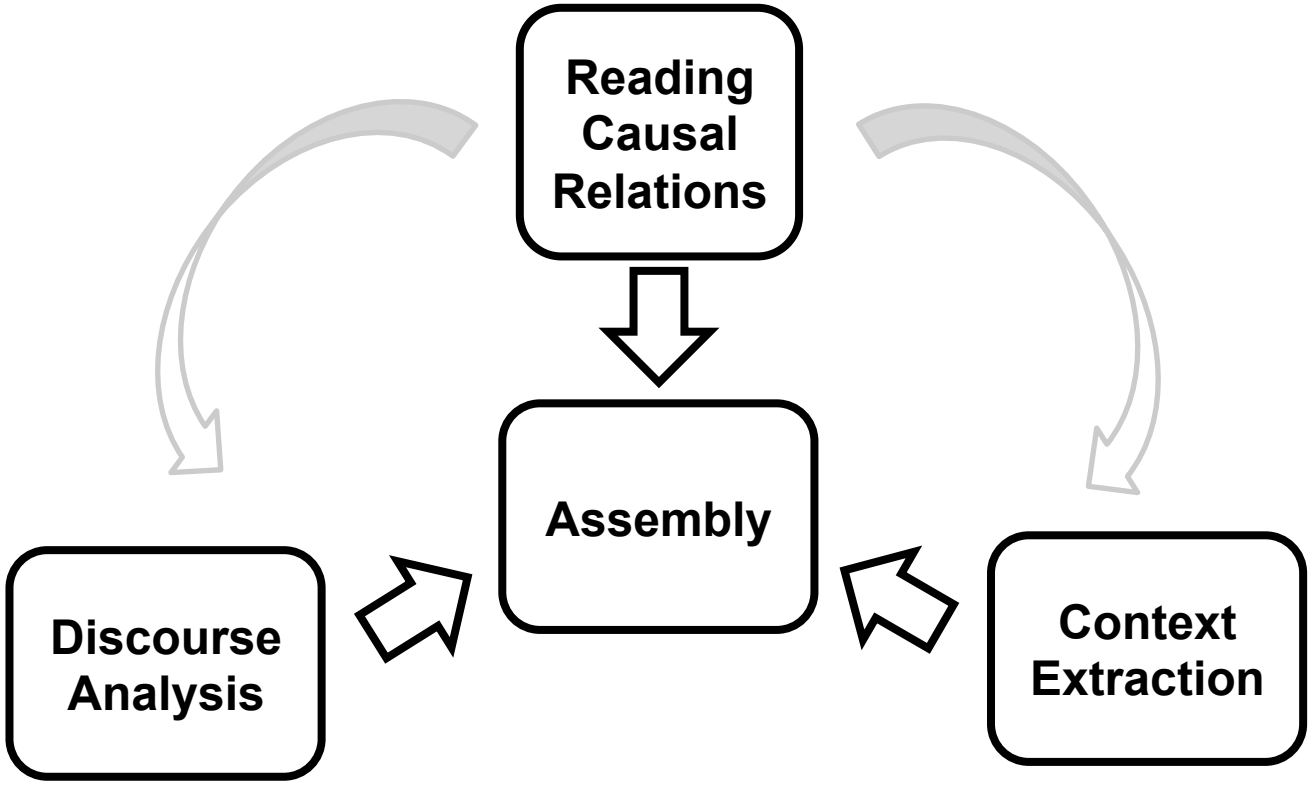
NEW PLATFORM

- So we ended up designing a new platform for machine reading: information extraction language + runtime environment
  - *"Mark the object of the verb 'phosphorylate' as a protein being phosphorylated, and the subject as the controller of the interaction"*

- **Expressive**: patterns over syntax or over plain text
- Captures **complex** language structures such as nested interactions
- Handles language **ambiguity**
- **Fast** runtime

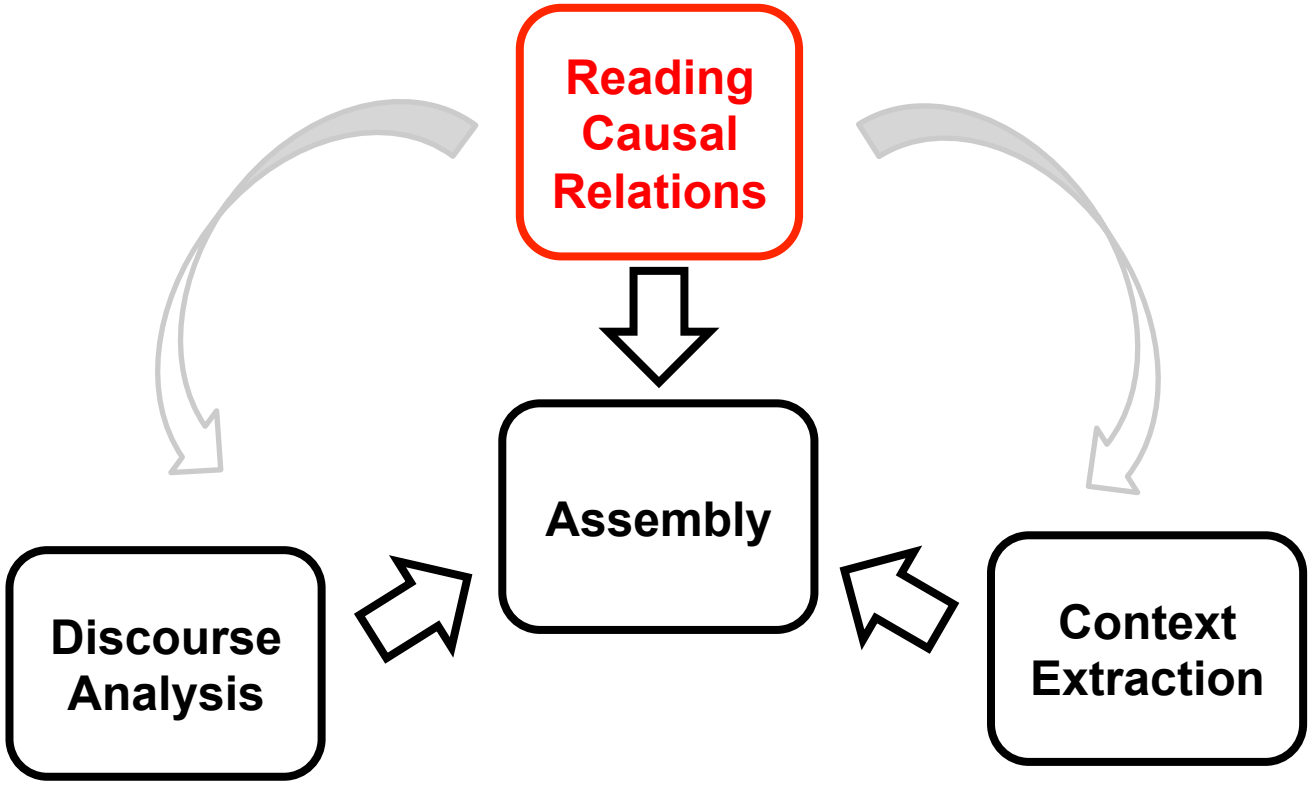
© Bill & Melinda Gates Foundation

HIGH-LEVEL ARCHITECTURE



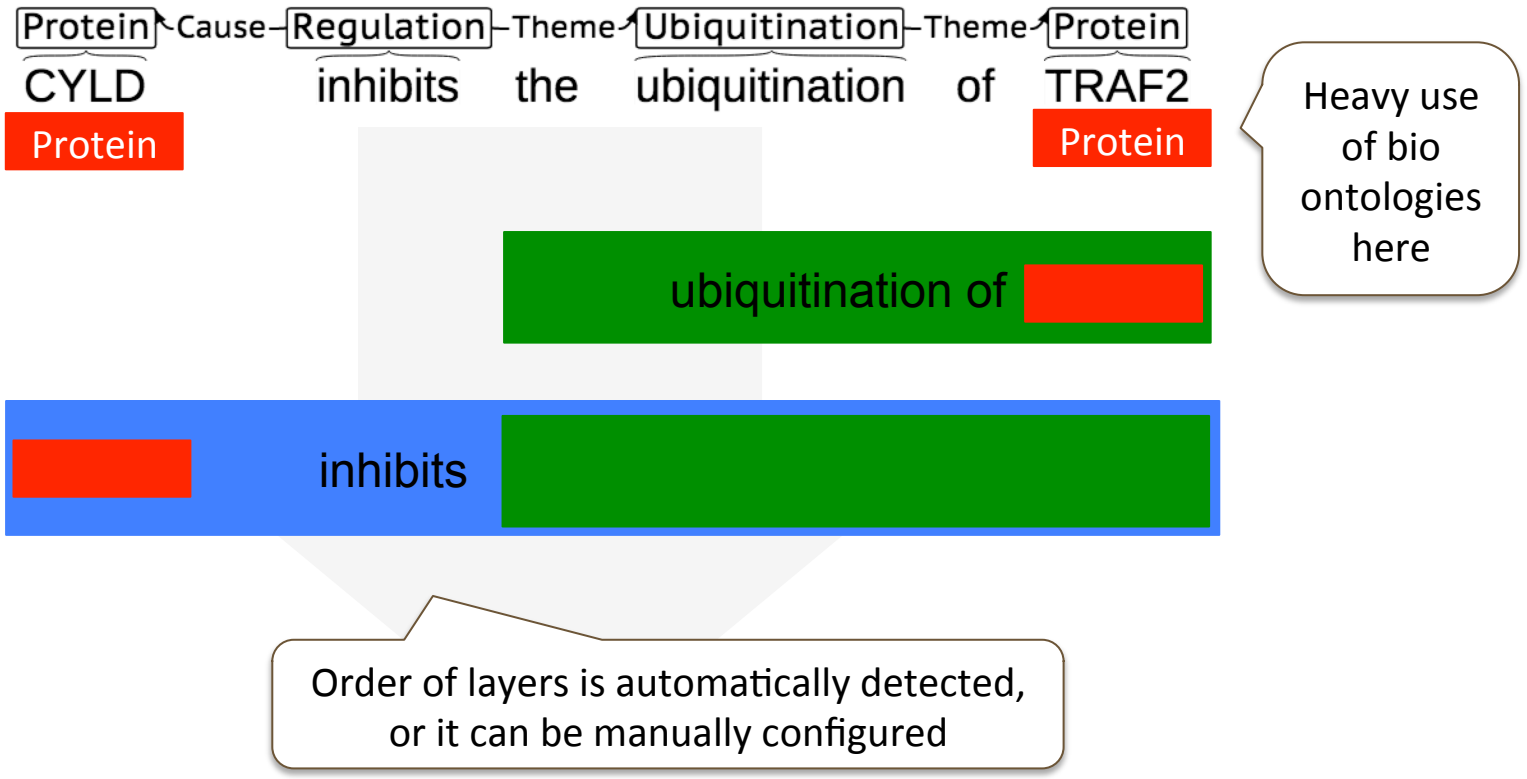
© Bill & Melinda Gates Foundation

HIGH-LEVEL ARCHITECTURE



© Bill & Melinda Gates Foundation

WALKTHROUGH EXAMPLE FOR READING OF CAUSAL RELATIONS



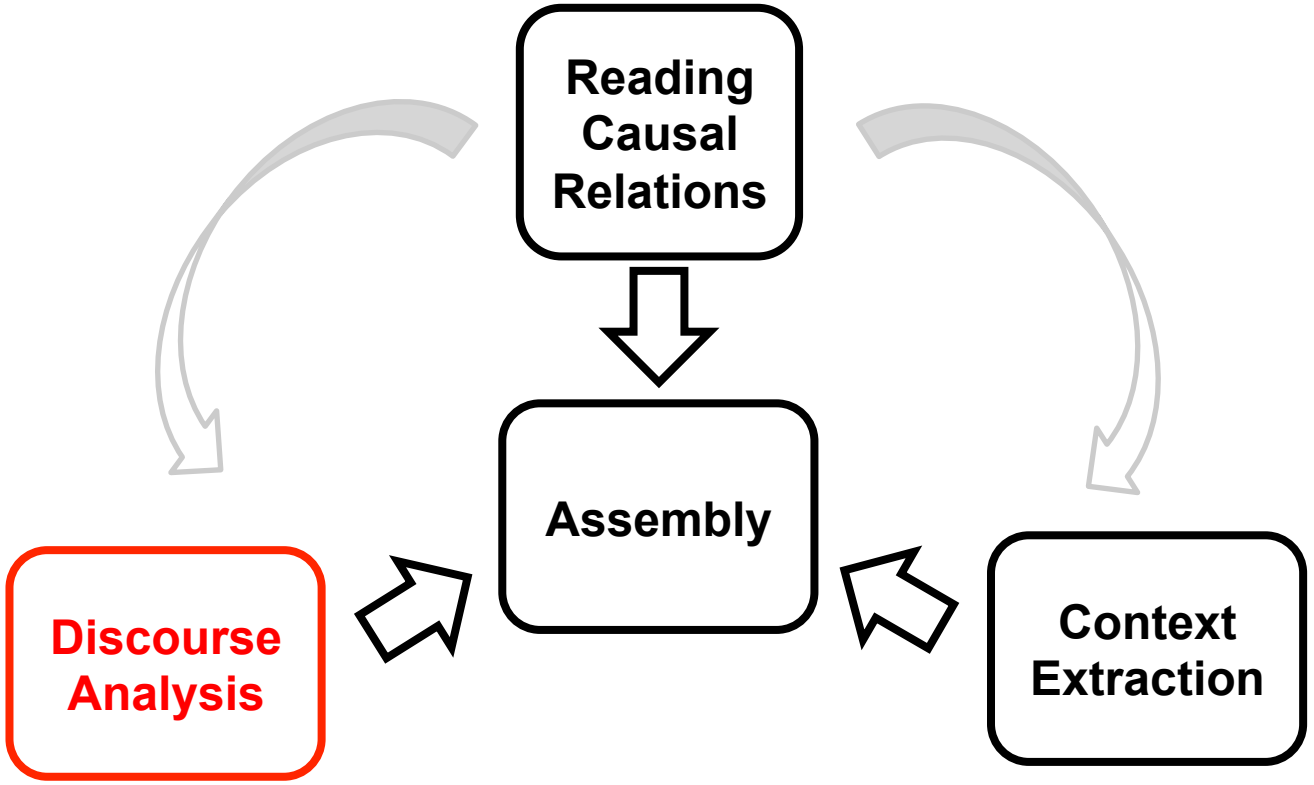
© Bill & Melinda Gates Foundation

FEW GRAMMAR RULES

Type	Syntax	Surface	Total
Entities	0	15	15
Generic entities	0	2	2
Modifications	0	6	6
Mutants	0	9	9
Total entities	**0**	**32**	**32**
Simple events	15	11	26
Binding	30	7	37
Hydrolysis	8	2	10
Translocation	12	0	12
Positive regulation	16	4	20
Negative regulation	14	3	17
Total events	**95**	**27**	**122**
Total	**95**	**59**	**154**

© Bill & Melinda Gates Foundation

HIGH-LEVEL ARCHITECTURE



© Bill & Melinda Gates Foundation

COREFERENCE RESOLUTION

"Cells were transfected with *N540K, G380R, R248C, Y373C, K650M and K650E-FGFR3 mutants* and analyzed for activatory STAT1(Y701) phosphorylation 48 hours later. [...] In 293T and RCS cells, **all six FGFR3 mutants** induced activatory ERK(T202/Y204) phosphorylation (Fig. 2)."

Unspecified mutations, leading to 6 x 2 events

"LL-37 forms a complex together with the IGF-1R ... and **this binding** results in IGF-1R activation …"

Common noun phrase refers to a specific interaction

© Bill & Melinda Gates Foundation

HEDGING

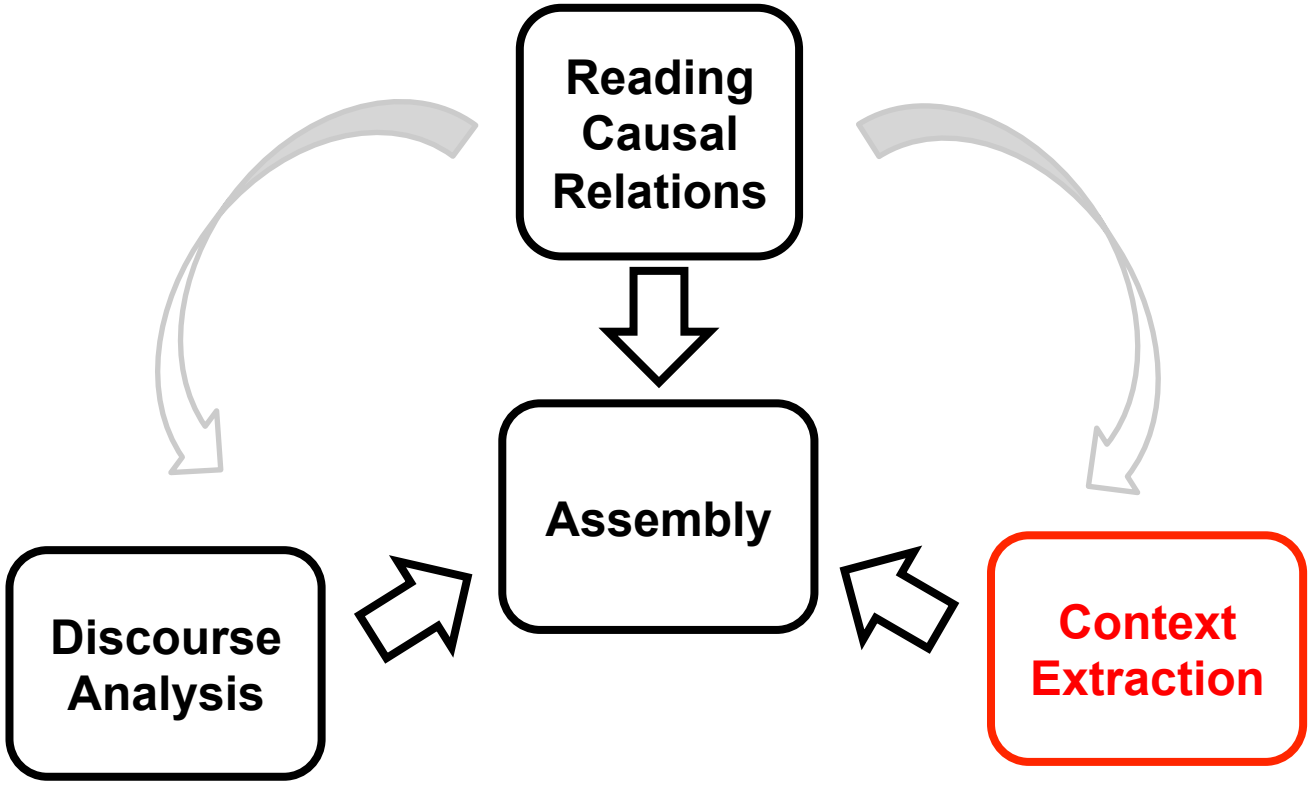
- "We hypothesize that O_2 inhibits the Cdc25-mediated wt Ras GDP dissociation"
- "The result suggests that both the intrinsic and the Cdc25-mediated wt Ras GDP dissociation are insensitive to H2O2."

Degree of substantiation	Precision
None	43%
Partial	60%
Full	80%

Analysis courtesy of: Dayne Freitag, SRI

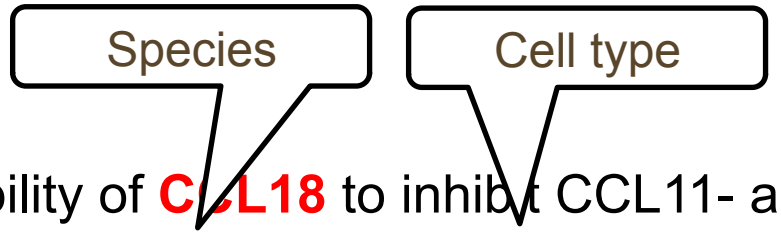
© Bill & Melinda Gates Foundation

HIGH-LEVEL ARCHITECTURE



© Bill & Melinda Gates Foundation

CONTEXTUALIZING THE EXTRACTIONS IS CRUCIAL



"The ability of **CCL18** to inhibit CCL11- and CCL13-induced chemotactic responses of **human eosinophils** has previously been reported [27]. We show here that **it is able to inhibit the chemotactic responses of other CCR3 agonists** (Figure 1A)."

© Bill & Melinda Gates Foundation

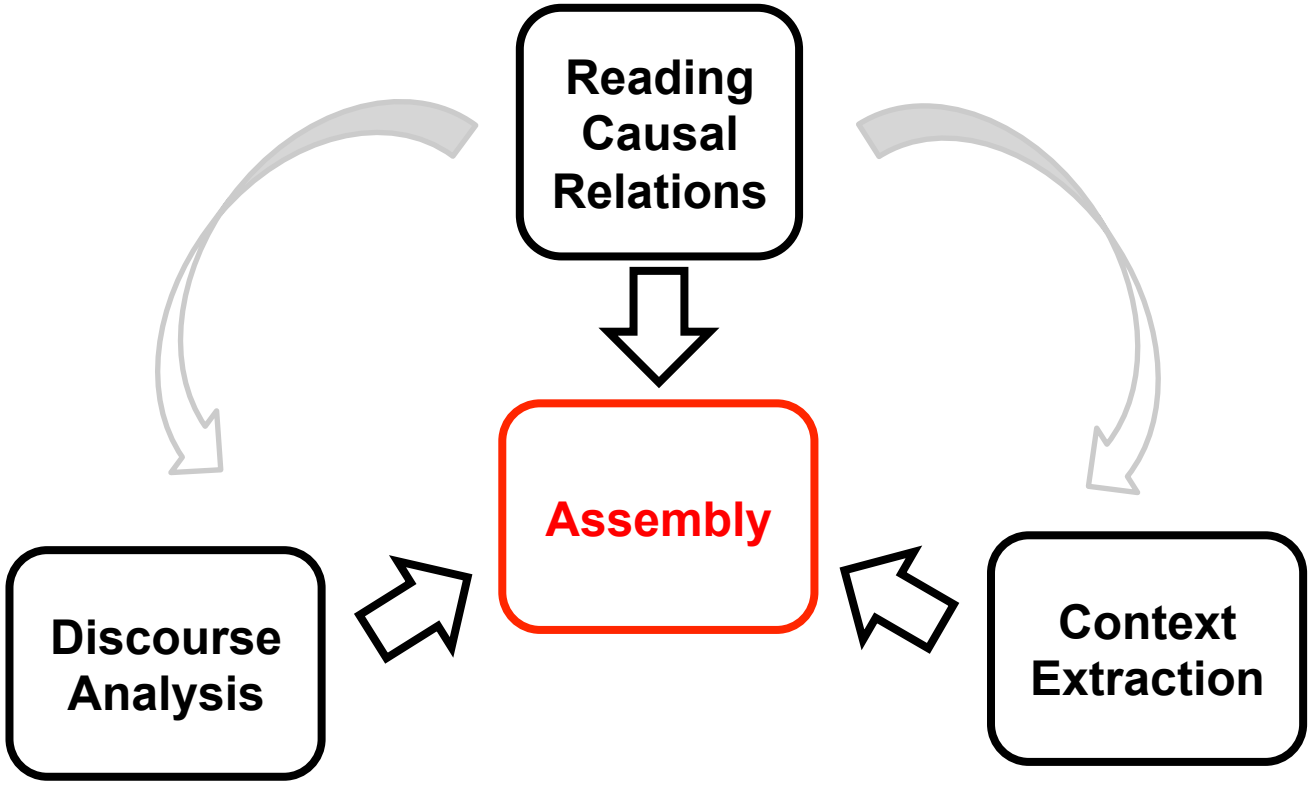
BIOLOGICAL CONTEXT

- Biological context serves as an index to the type of biological system: which mechanisms are (possibly) present / active.

- Context as specification of biological containers:
  - **Species**: human, yeast
  - **Organ**: liver, lung
  - **Tissue type**: lymphoid, embryo
  - **Cell type**: t-cell, endothelial
  - **Cell line**: PCS-100-020 (human, artery, endothelial)
  - **Subcellular locations:** endosome, nucleus

© Bill & Melinda Gates Foundation

HIGH-LEVEL ARCHITECTURE



© Bill & Melinda Gates Foundation

ASSEMBLY DRIVEN BY CAUSAL PRECEDENCE

Causal precedence = Interaction A precedes B if and only if the output of A is **necessary** for the successful execution of B

© Bill & Melinda Gates Foundation

WHY CAUSAL PRECEDENCE IS IMPORTANT

The SH2 domain of c-SRC *binds* VEGFR2,
when VEGFR2 is *phosphorylated* at Y1057

precedes

© Bill & Melinda Gates Foundation

WHY CAUSAL PRECEDENCE IS IMPORTANT

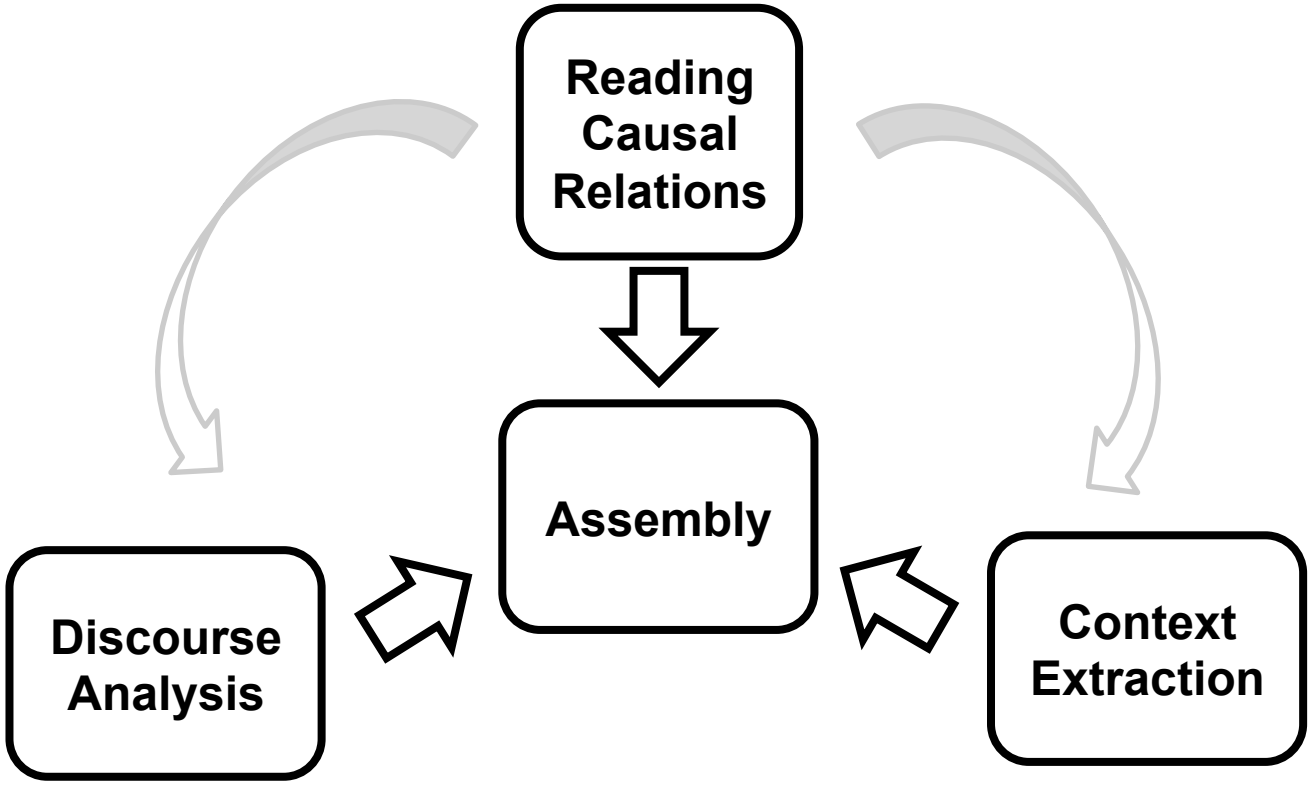
The SH2 domain of c-SRC *binds* VEGFR2,
before VEGFR2 is *phosphorylated* at Y1057

follows

(skipping the technical details in the interest of time)

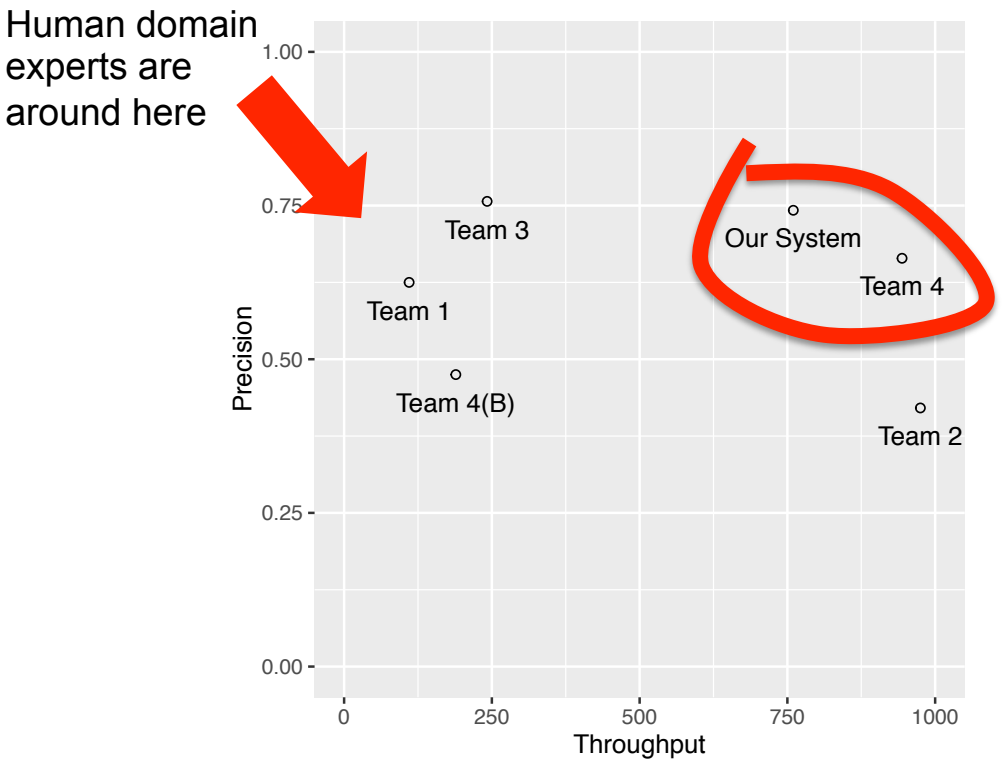
© Bill & Melinda Gates Foundation

A COMPLETE MACHINE READING SYSTEM



© Bill & Melinda Gates Foundation

HOW WELL DOES MACHINE READING WORK?



© Bill & Melinda Gates Foundation

OVERVIEW

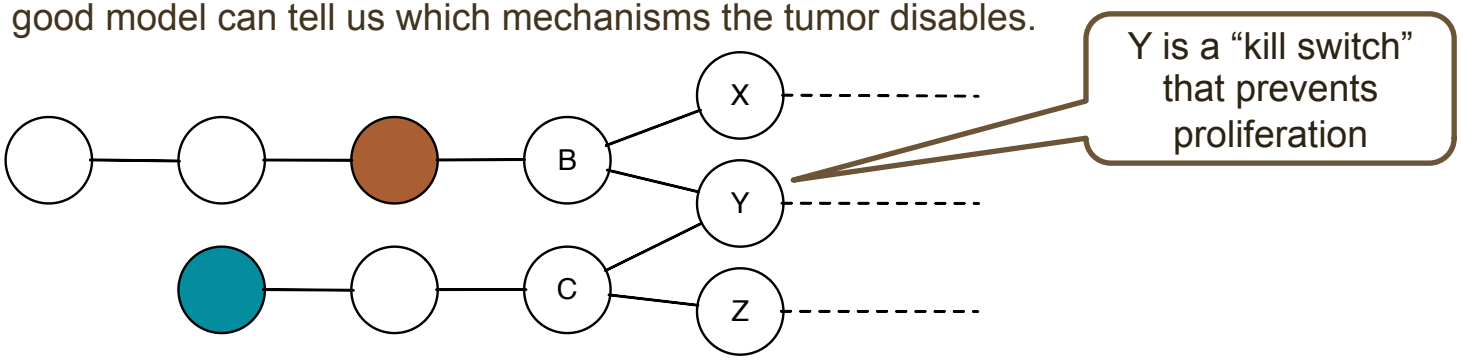
- Machine reading system
- **Use case 1: discovery of novel cancer drivers**
- Use case 2: modeling children's health

© Bill & Melinda Gates Foundation

THE MUTEX ALGORITHM

We have huge amounts of data relating gene expression levels to cancers. We don't know the underlying mechanisms.

Mutex insight: If a tumor "wants" to disable a mechanism, it will mutate something upstream, but it generally won't "pay" for two mutations that do the same thing. So mutually exclusive mutations plus a good model can tell us which mechanisms the tumor disables.

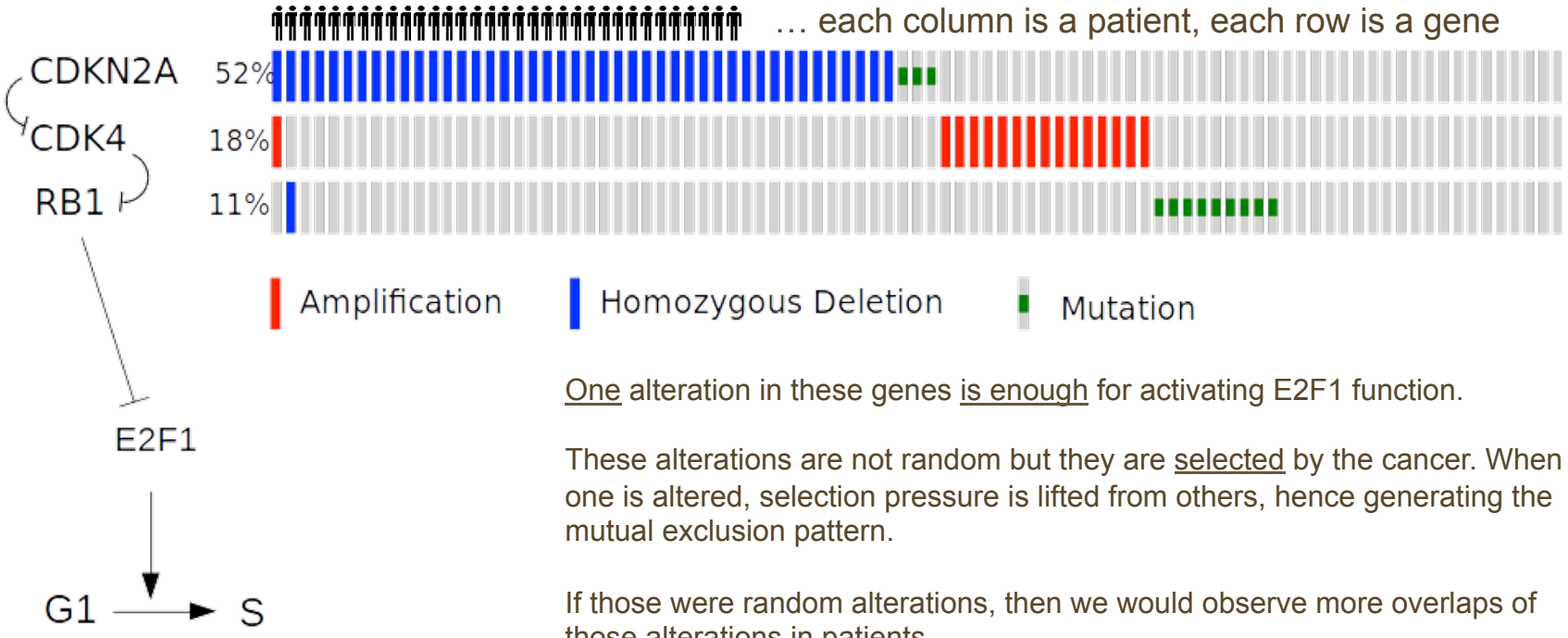


Slide courtesy of: Emek Demir, OHSU

© Bill & Melinda Gates Foundation

OBSERVATION OF MUTUAL EXCLUSIVITY OF GENOMIC ALTERATIONS IN CANCER

For instance, in Glioblastoma:



One alteration in these genes is enough for activating E2F1 function.

These alterations are not random but they are selected by the cancer. When one is altered, selection pressure is lifted from others, hence generating the mutual exclusion pattern.

If those were random alterations, then we would observe more overlaps of those alterations in patients.

Slide courtesy of: Emek Demir, OHSU

© Bill & Melinda Gates Foundation

WHY IT MATTERS

- Why it matters:
  - Mutations often disable “kill switches” that prevent proliferation.
  - Fixing the downstream effects of one mutation is not enough: The cell will mutate again to find another way to have the same downstream effects.
  - Therapy should disable *all* upstream mutations, or the driver itself.

- Knowing more is better.

© Bill & Melinda Gates Foundation

HOW MUTEX WORKS

- Mutex does a graph search on the signaling network to find subgraphs of genes that
  - Are altered in mutually exclusive manner, and
  - Have a common downstream signaling target.

- The important issue is what data is used to search for downstream targets.

© Bill & Melinda Gates Foundation

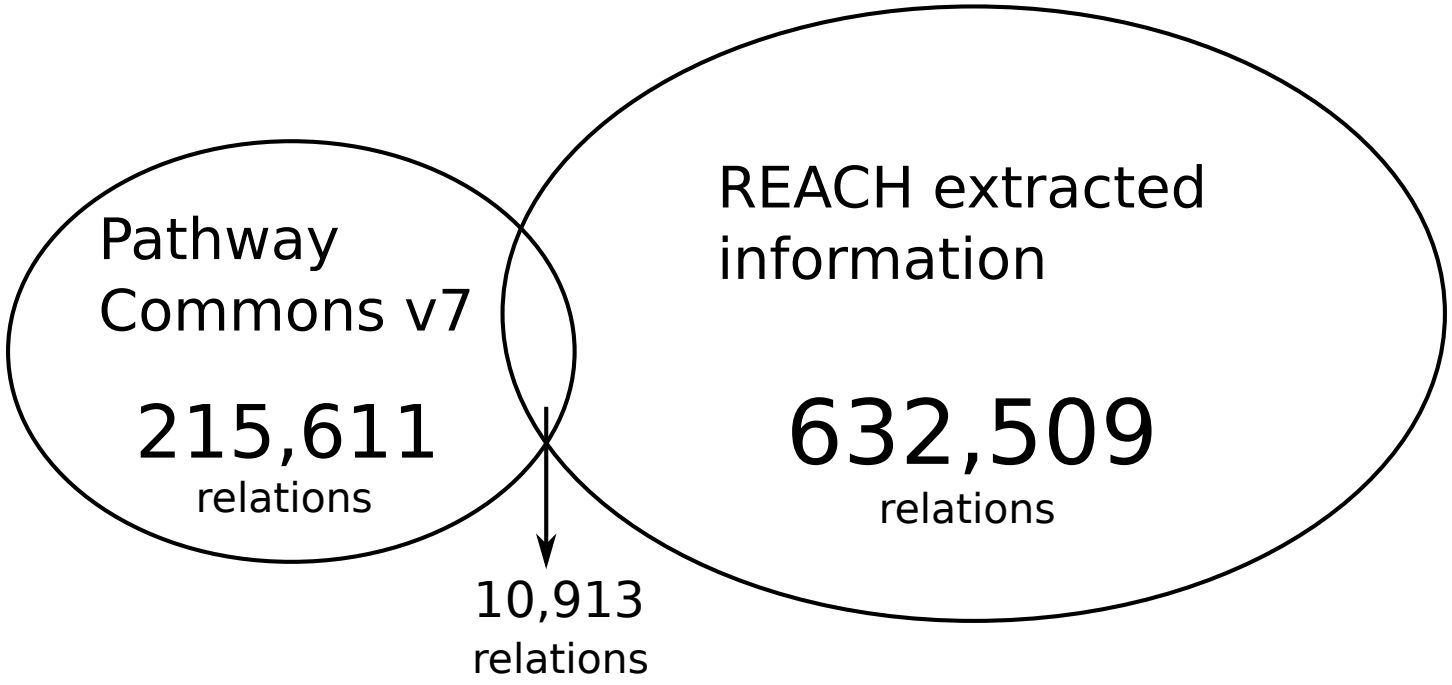
SEARCHING FOR DOWNSTREAM PROTEINS

- Previously Demir et al. were searching over Pathway Commons (PC), a manually-curated database of biochemical interactions.
- PC is estimated to cover **1%** of the literature.

- We expanded the dataset automatically by reading all PubMed Open Access papers using the system introduced before.

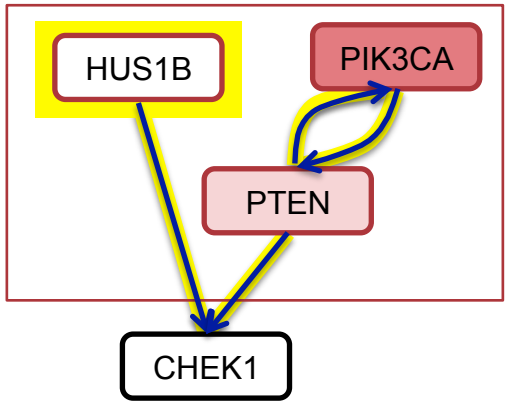
© Bill & Melinda Gates Foundation

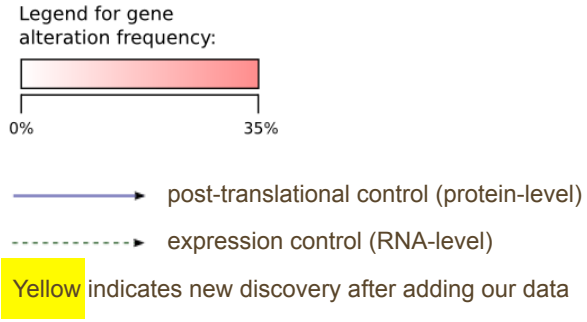
DATA COMPARISON



© Bill & Melinda Gates Foundation

POTENTIAL CANCER DRIVING MECHANISMS DISCOVERED





HUS1B → CHEK1: “Hus1 loss results in abnormal gammaH2AX localization and increased CHK1 phosphorylation.”

PTEN → CHEK1: “PTEN also induces phosphorylation and monoubiquitination of DNA damage checkpoint kinase, Chk1”

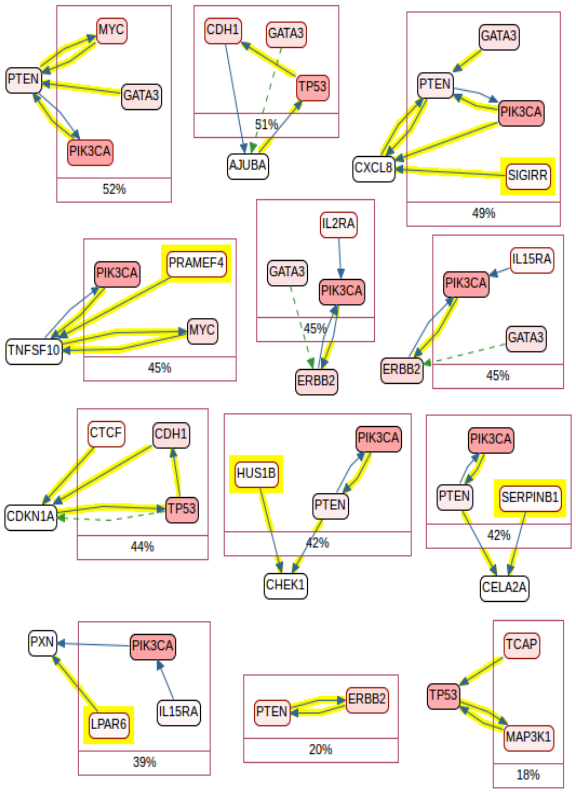
PIK3CA → PTEN: “p110alpha inactivation can inhibit the impact of PTEN loss”

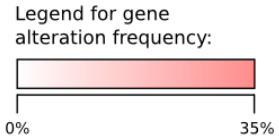
Members of a mutex group are shown in a compound node, where its label indicates sample-coverage of the alterations.

Highlighted edges do not exist in Pathway Commons (PC), and highlighted nodes cannot be detected using PC only.

© Bill & Melinda Gates Foundation

POTENTIAL CANCER DRIVING MECHANISMS FOR TCGA BREAST CANCER DATASET (BRCA)





post-translational control (protein-level)

expression control (RNA-level)

Yellow indicates new discovery after adding our data

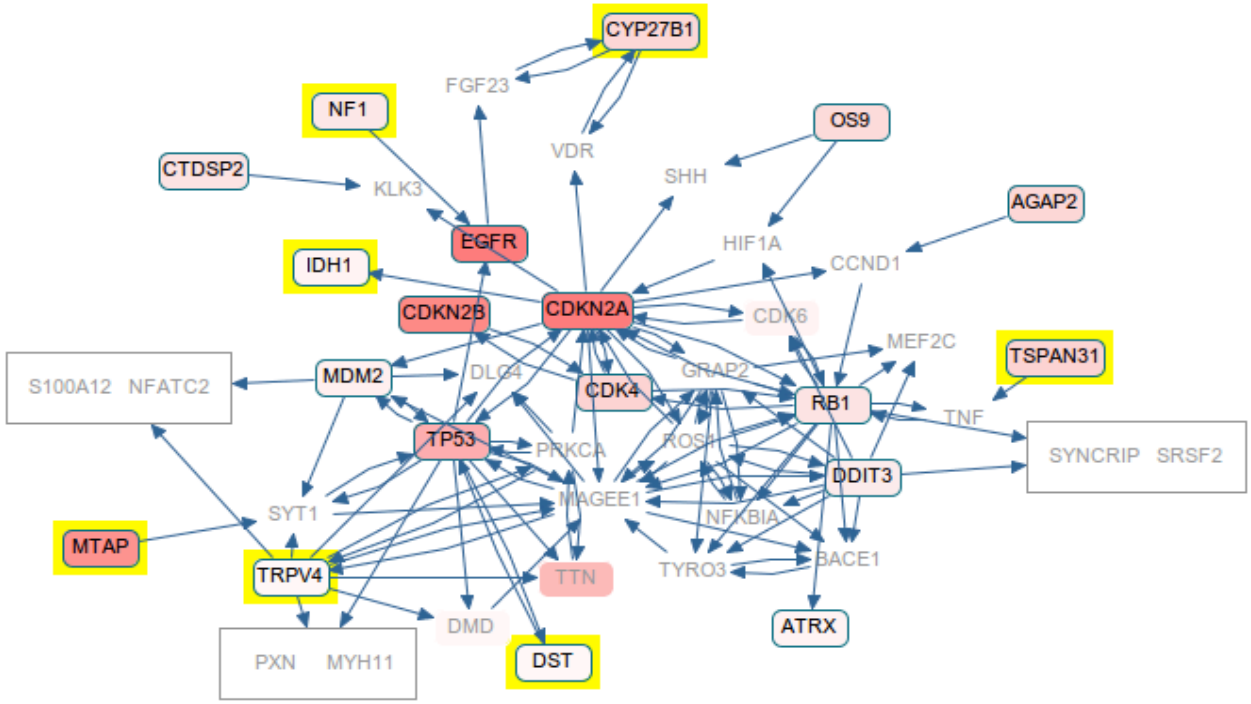
Members of a mutex group are shown in a compound node, where its label indicates sample-coverage of the alterations.

Highlighted edges do not exist in Pathway Commons (PC), and highlighted nodes cannot be detected using PC only.

Slide courtesy of: Emek Demir, OHSU

© Bill & Melinda Gates Foundation

GBM MUTEX RESULTS AFTER ADDING THE NEW DATA

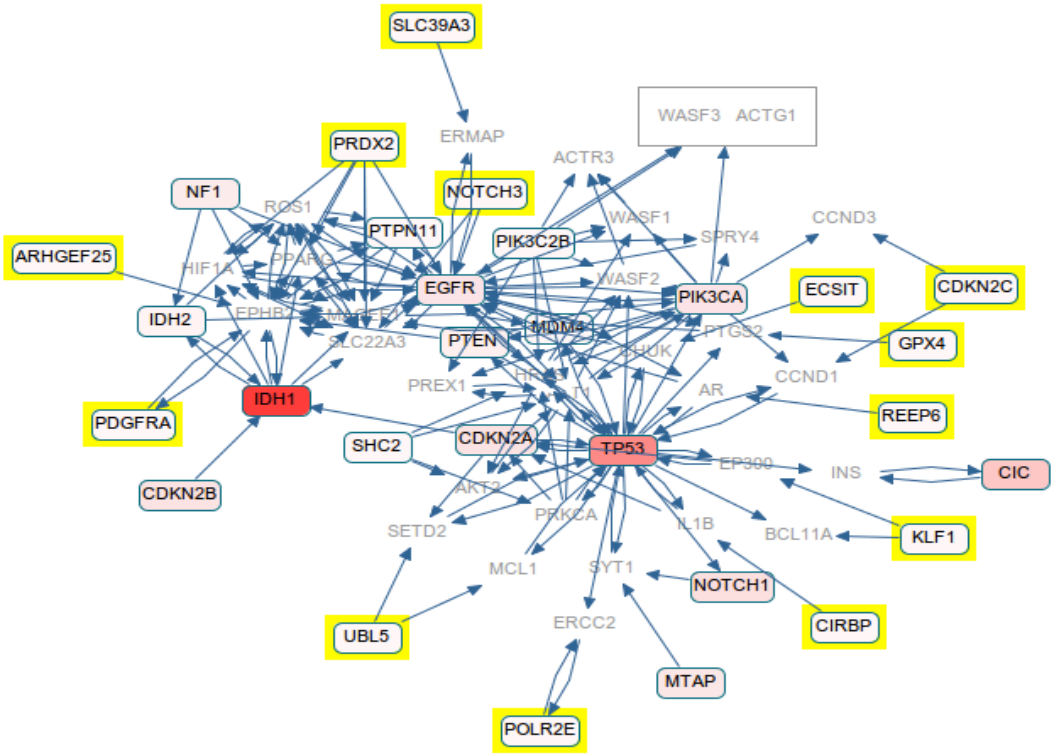


Highlighted genes can be detected after adding our data

Image courtesy of: Emek Demir, OHSU

© Bill & Melinda Gates Foundation

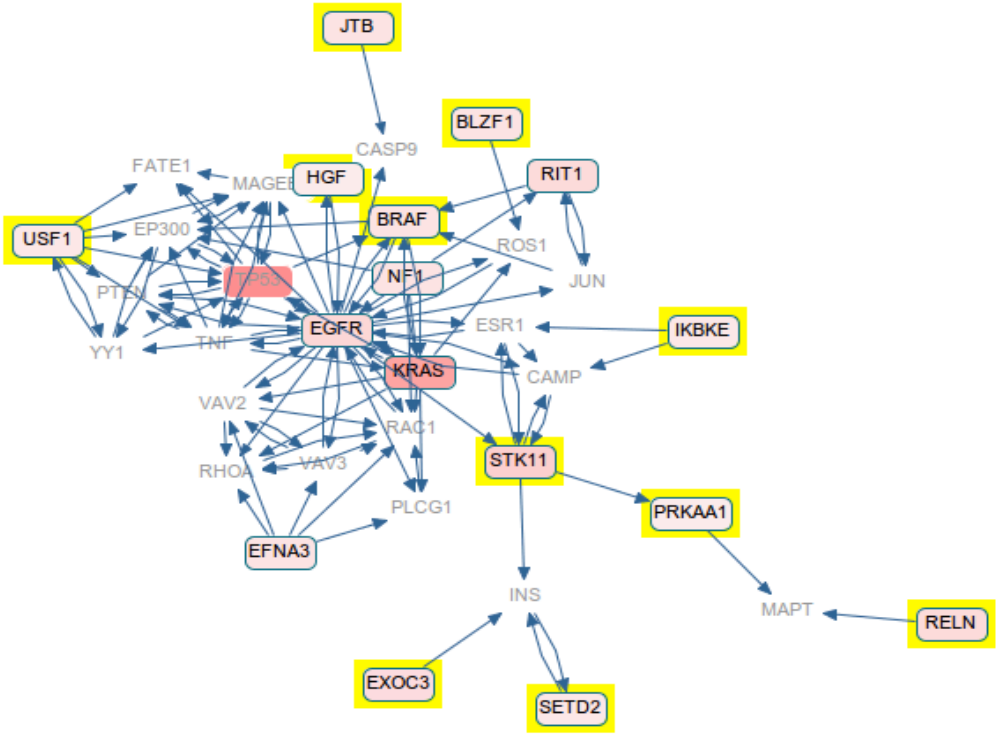
LGG MUTEX RESULTS AFTER ADDING THE NEW DATA



Highlighted genes can be detected after adding our data

Image courtesy of: Emek Demir, OHSU

LUAD MUTEX RESULTS AFTER ADDING THE NEW DATA



Highlighted genes can be detected after adding our data

Image courtesy of: Emek Demir, OHSU

© Bill & Melinda Gates Foundation

HBGD*ki*

OVERVIEW

- Machine reading system
- Use case 1: discovery of novel cancer drivers
- **Use case 2: modeling children's health**

© Bill & Melinda Gates Foundation

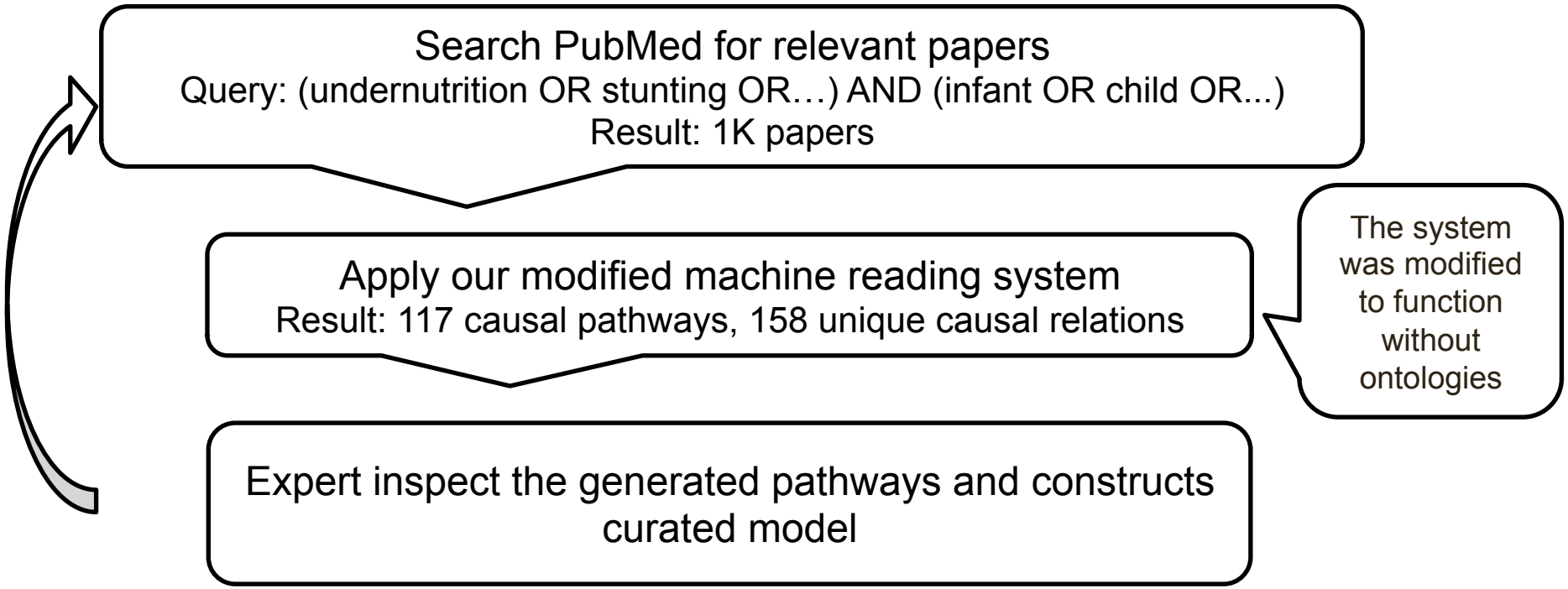
NEW USE CASE: CHILDREN'S HEALTH

- Do exactly the same, but model children's health

- Collaboration with Lyn Powell – HBGDki-qPM team
- Very preliminary work: we had < 2 weeks to adapt our system
- Turns out biology was the "easy" use case
  - Many factors impact children's health, from biology to socio-economic
  - No ontologies

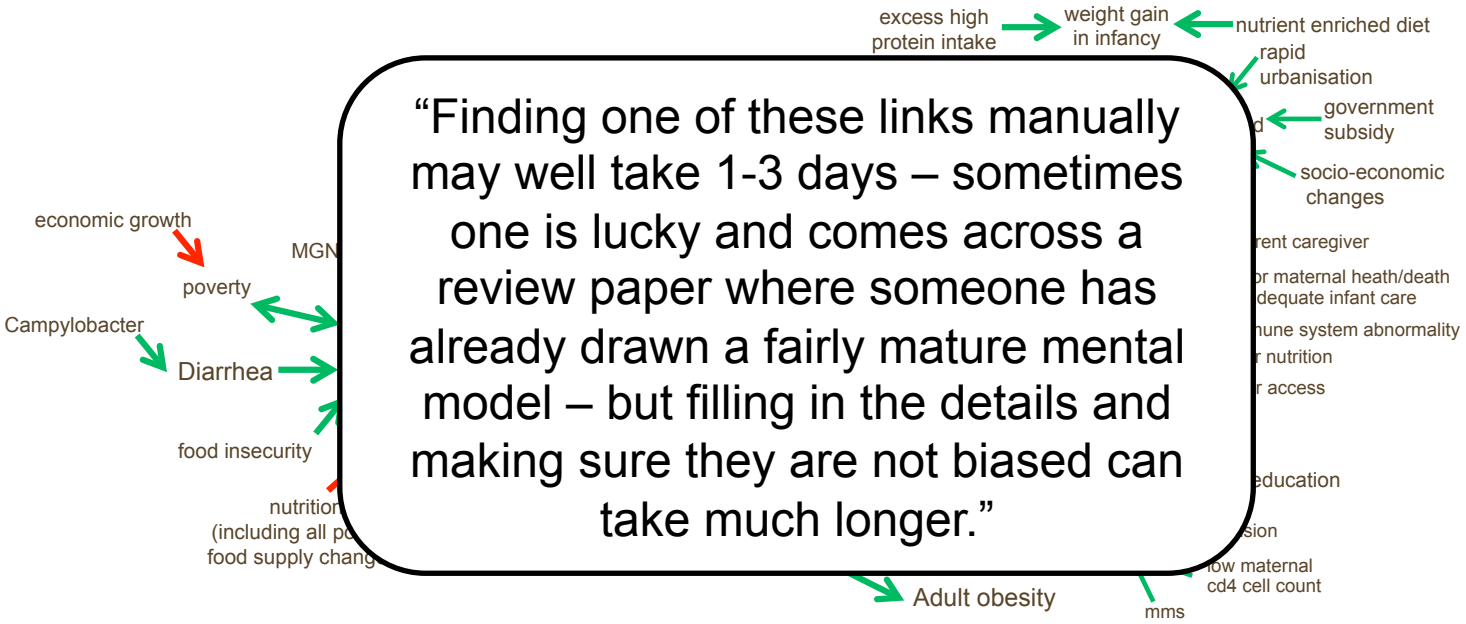
© Bill & Melinda Gates Foundation

PROCESS



© Bill & Melinda Gates Foundation

RESULTING MODEL



"Finding one of these links manually may well take 1-3 days – sometimes one is lucky and comes across a review paper where someone has already drawn a fairly mature mental model – but filling in the details and making sure they are not biased can take much longer."

Constructed in 2 days. Normally, it takes 1 month.

Model courtesy of: Lyn Powell, HBGDki-qPM team

© Bill & Melinda Gates Foundation

KEY FINDINGS AND INSIGHTS

- High-quality machine reading at scale is possible
  - Read all PubMed Open Access literature at 5 seconds/paper
  - Accuracy comparable to human cancer experts, at much higher throughput

- Use case 1: discovery of novel cancer driving mechanisms
  - Works for 11 different cancer types
  - New hypotheses for precision medicine for cancer treatment

- Use case 2: discovery of causal mechanisms in children's health
  - Reduced model development time from ~1 month to 2 days

© Bill & Melinda Gates Foundation

THANK YOU!

- Acknowledgments:
  - Vision: **Paul Cohen**, DARPA Big Mechanism program manager
  - Actually implementing it: the University of Arizona team**: Marco A. Valenzuela-Escarcega, Gus Hahn-Powell, Dane Bell, Thomas Hicks, Enrique Noriega-Atala, Clayton Morrison**
  - Use case 1: **Emek Demir, Özgün Babur** – Oregon Health & Science University
  - Use case 2: **Lyn Powell** – HBGDki-qPM team

© Bill & Melinda Gates Foundation