INTERVENTIONS MODELING ON CAUSAL GRAPHS: FROM CANCER TREATMENT TO SOLVING HUNGER

STEFAN ANDJELKOVIC

OUBES TALK, NOV 18TH, 2020

THE BEGINNING OF MY JOURNEY INTO COMP BIO

- MASt in Physics at “The Other Place” University 
- Scholarship condition: return to home country (🇷🇸) and work for 2 years
- Quantitative skills + coding + data visualization ⇒ Data Science (ML)
- Entrepreneurship course – business strategy for a bioinformatics startup
- Bioinformatics Engineering ≈ Data Science for Biology

MY FIRST FULL TIME POSITION: BIOINFORMATICS ENGINEER

- Seven Bridges Genomics – yes, named after 7 bridges of Königsberg
- Platform for running genomics analyses on cloud
- Company found in 2009, following development of Next-Generation Sequencing (NGS)
- At the time I joined in 2016 it had 250+ employees
- Offices in Cambridge (MA), SF, London & Belgrade

SevenBridges

NGS TECH AND THE ROLE OF BIOINFORMATICS

- DNA is read in fragments (~100 bp), and then reassembled
- Input files ~100 GB
- Alignment: reads aligned to a reference genome - extremely computationally intensive
- Variant calling: SNVs, indels, CNVs
- Manual exploration

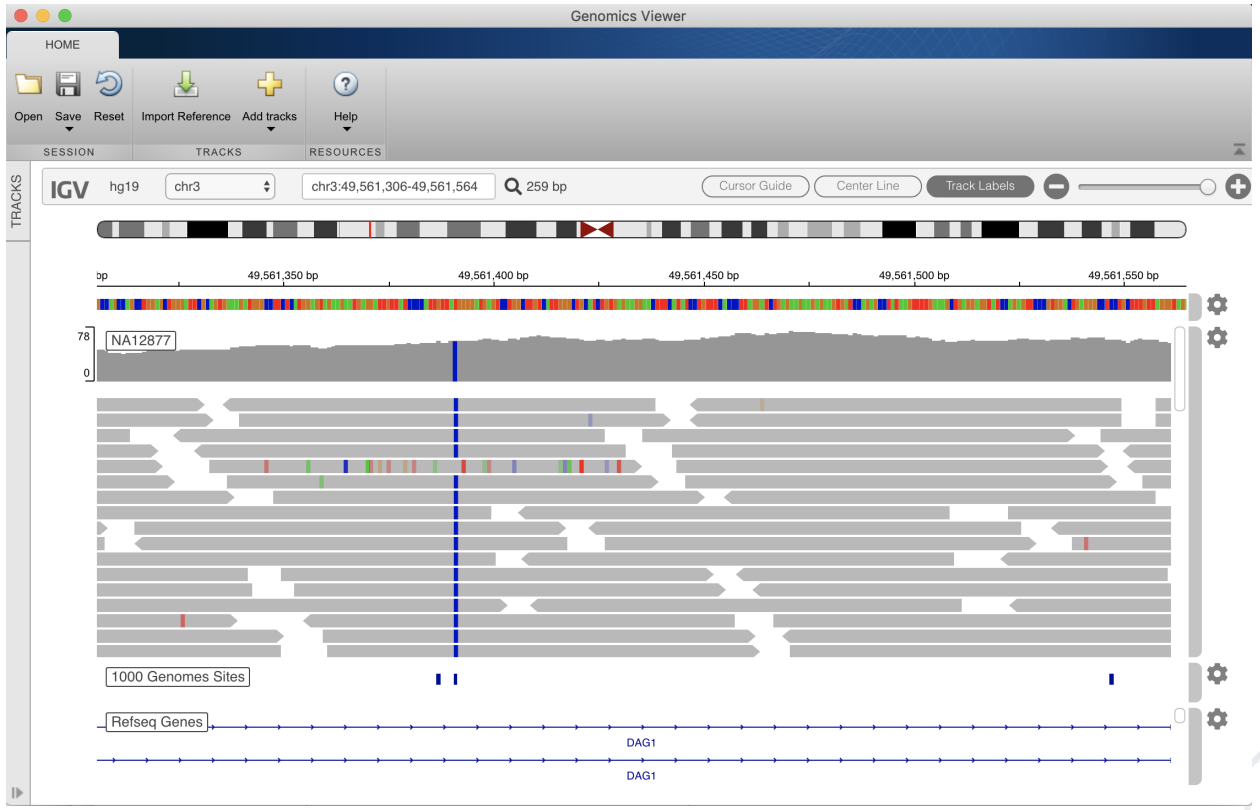


Figure adapted from https://www.mathworks.com/help/bioinfo/ref/genomicsviewer-app.html

THE ROLE OF THE BIOINFORMATICS ENGINEER

- Bioinformatics tool wrapping and porting on cloud
- Pipeline assembly
- Running, testing, curating
- Data analysis and curation
- Literature exploration

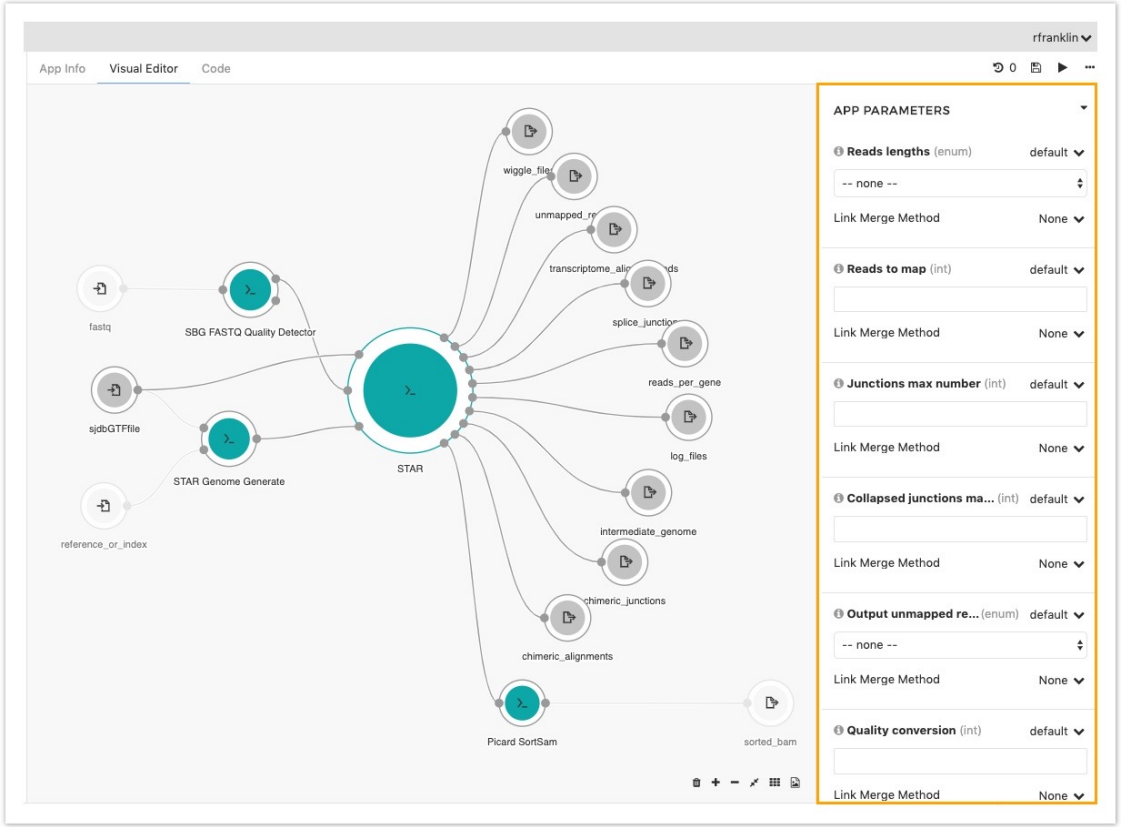


Figure adapted from https://docs.sevenbridges.com/docs/build-a-workflow-tutorial

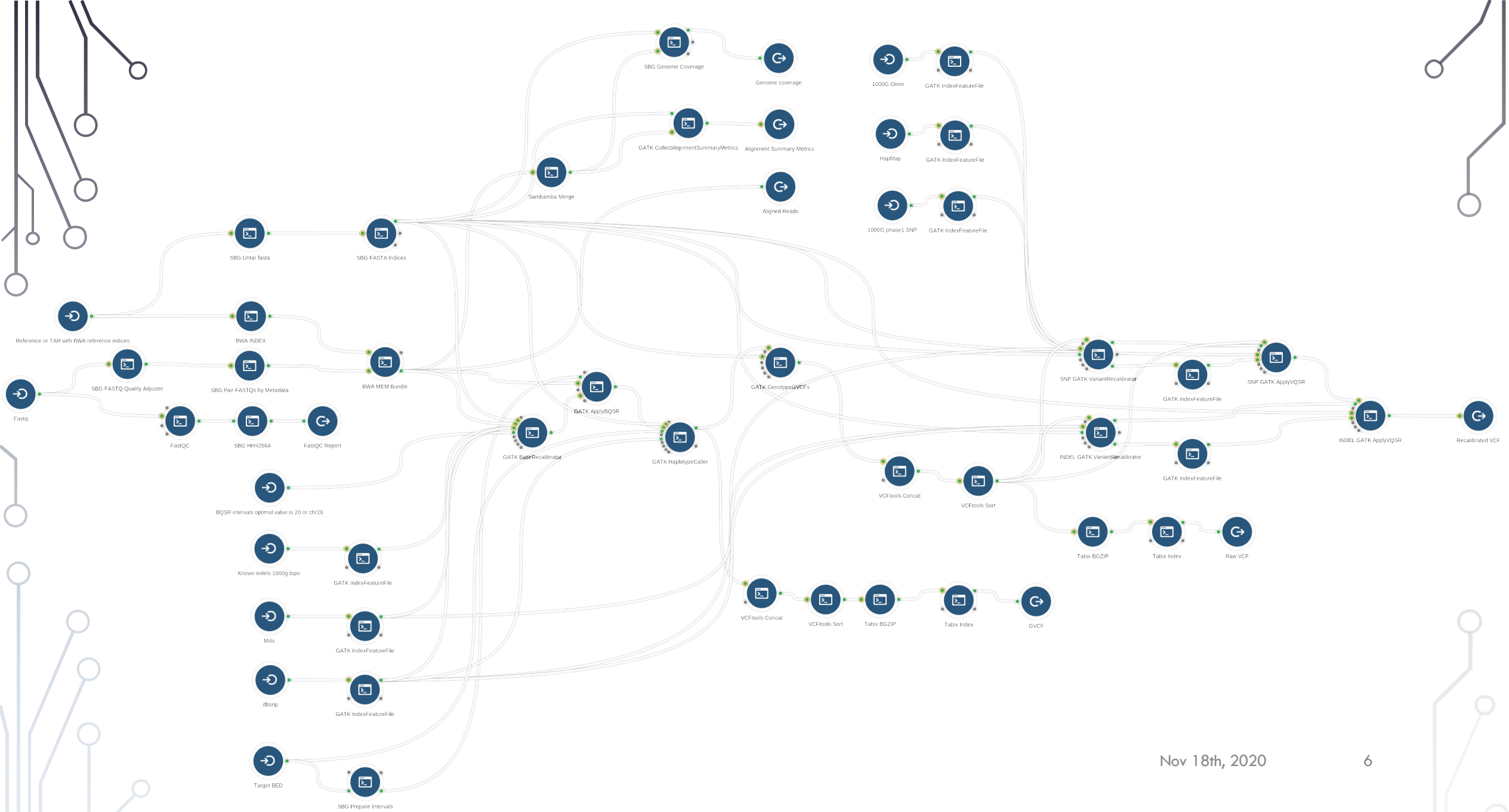


Figure adapted from https://www.sevenbridges.com/first-look-gatk4/

MY MAIN PROJECT – CTDNA ANALYSIS PIPELINE

- Circulating tumo(u)r DNA (ctDNA) – tumo(u)r DNA fragments floating in bloodstream
- Non-invasive diagnostics (liquid biopsy)
- Similar methods used for non-invasive prenatal testing (NIPT)

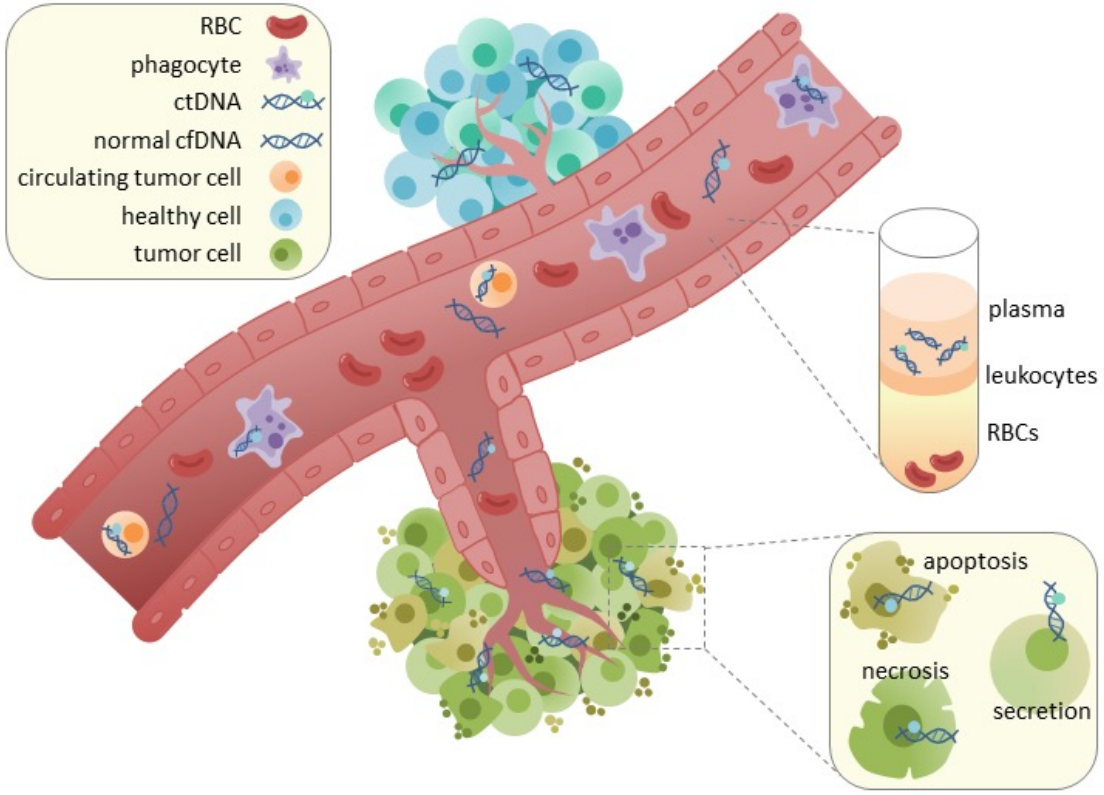


Figure adapted from https://commons.wikimedia.org/wiki/File:CtDNA_in_circulation.png

LESSONS LEARNT #1

- Job positions of the future are converging on “jack of all trades, master of none” types, but “learning on the go” is a double-edged sword
- Reading papers outside of your area of expertise is hard and time consuming
- You can’t segregate data from its production

Piece of advice:

- Make sure you’re aware of what’s expected of you, and deflect responsibilities by asking for clarifications

NEXT STOP → PERSIDA

- New startup offering specialized bioinformatics services to bigger companies
- Small and compact team – less diversification of roles, better learning opportunity
- Focus on delivering solutions to diagnostics companies for more direct impact on global healthcare

ADVANTAGES

- Structured data – software was tailored to the customers
- Close coordination with customers allowed for iterative analyses
- Exciting new projects in dynamic environment (whole startup was R&D department)
- Flat hierarchy – fewer management layers allow for more transparency, and optimal training ground for future executive positions

LESSONS LEARNT #2

- Smaller team means more mundane work for you, but also more growth space
- Less hierarchy means more dynamic decision making
- The ability to estimate how much time it will take you to obtain a new skill is as important as developing this skill

Piece of advice:

- If you're working on cutting edge technology development, always account for overhead for learning and exploring in your estimates

SO, WHAT ABOUT LEARNING?

- There are different aspects:
  - Basic knowledge (needed for reasoning and understanding results) – requires reading
  - Skills (needed for high performance in work) – requires practicing
  - Expertise (needed for competitive edge) – requires staying up to date
  - Broad knowledge (needed for interdisciplinary work) – requires education
- Rosalind Info – web platform for learning bioinformatics
- CMU-Pitt Comp Bio PhD Program – a junction of cutting edge AI and Medicine

CPCB PHD PROGRAM

- Jointly offered by Carnegie Mellon University School of Computer Science & University of Pittsburgh School of Medicine
- 4 specializations: Computational Genomics, Structural Biology, Bioimage Informatics, and Cell & Systems Modeling
- 5 core courses: Cellular & Systems Modeling, Intro to Structural Bio, Computational Genomics, Intro to ML, and Lab Methods for Comp Biologists
- 3 electives: 1 life sciences, 1 specialization and 1 open elective

MELODY LAB

- **Me**chanisms and **Lo**gic of **Dy**namics
- Specializing in mechanistic modeling of biological networks dynamics.
- Network models represented as digital logic circuits.
- Data sources: databases and machine reading outputs
- The overarching goal is **automation** of building **interpretable** network models for **hypothesis generation and evaluation**

BIOLOGICAL NETWORKS

- Gene regulatory networks
- Cell signaling pathways
- Other causal networks

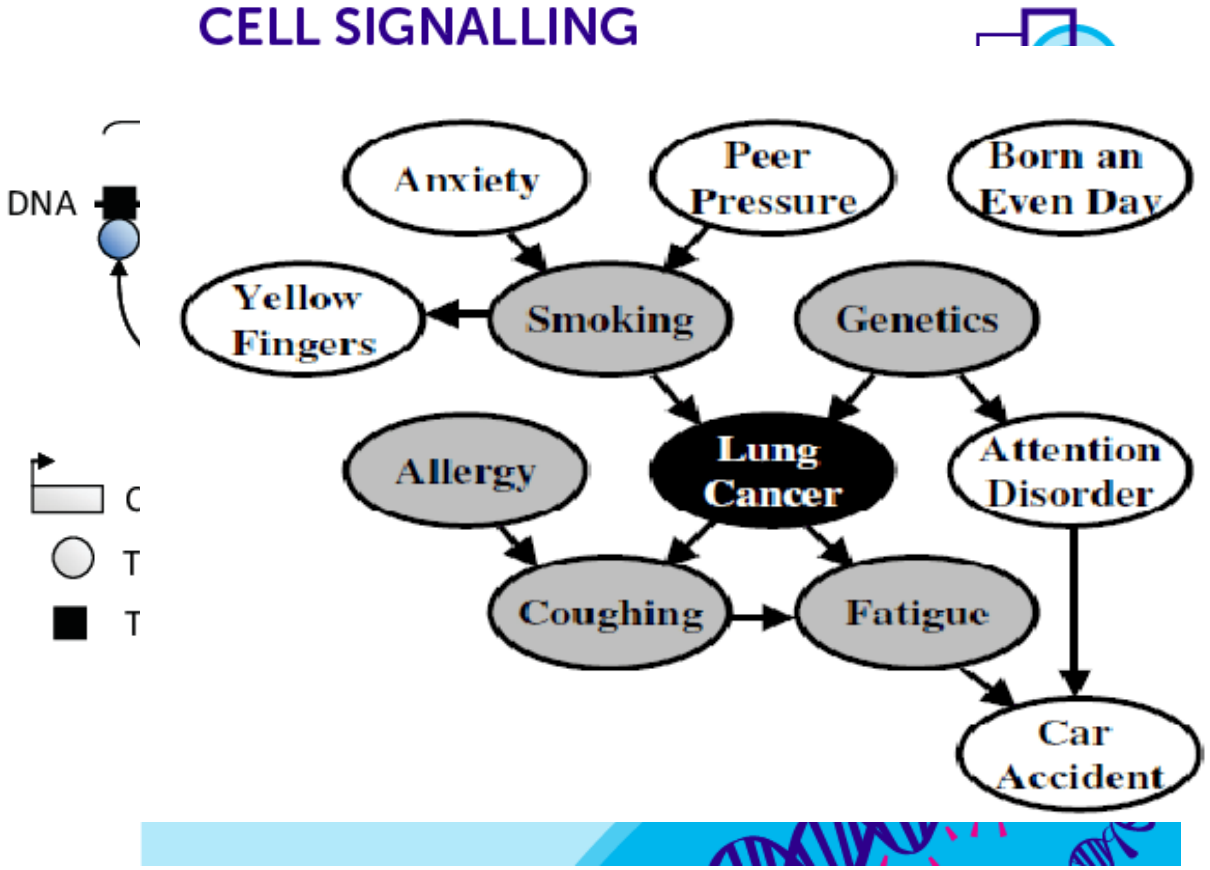


Figure adapted from Reuter et al (2010)

Fig. adapted from https://scienceresearchuk.org/2015/04/29/our-milestones-the-discovery-of-the-nras-signalling-gene

HOW DO WE MODEL SIGNAL PROPAGATION?

- ODE-based approach
  - Ex. 1: A + 2B → C
  - Updates: $\frac{d[C]}{dt} = k[A][B]^2$, $\frac{d[A]}{dt} = -k[A][B]^2$, $\frac{d[B]}{dt} = -2k[A][B]^2$
- Boolean networks
  - Ex. 1: A + 2B → C (note: in Boolean representation this is equivalent to A + B → C)
  - Updates: C = A OR B
- Discrete networks (our method)
  - Ex. 1: A + 2B → C
  - Updates: C = min(A+2B, max_C)

APPLICATIONS TO CANCER

- Ras signaling – most common oncogenes across cancer types
- Network structure plays crucial role (feedback loops)
- Multiple treatment targets, but which ones are most effective?

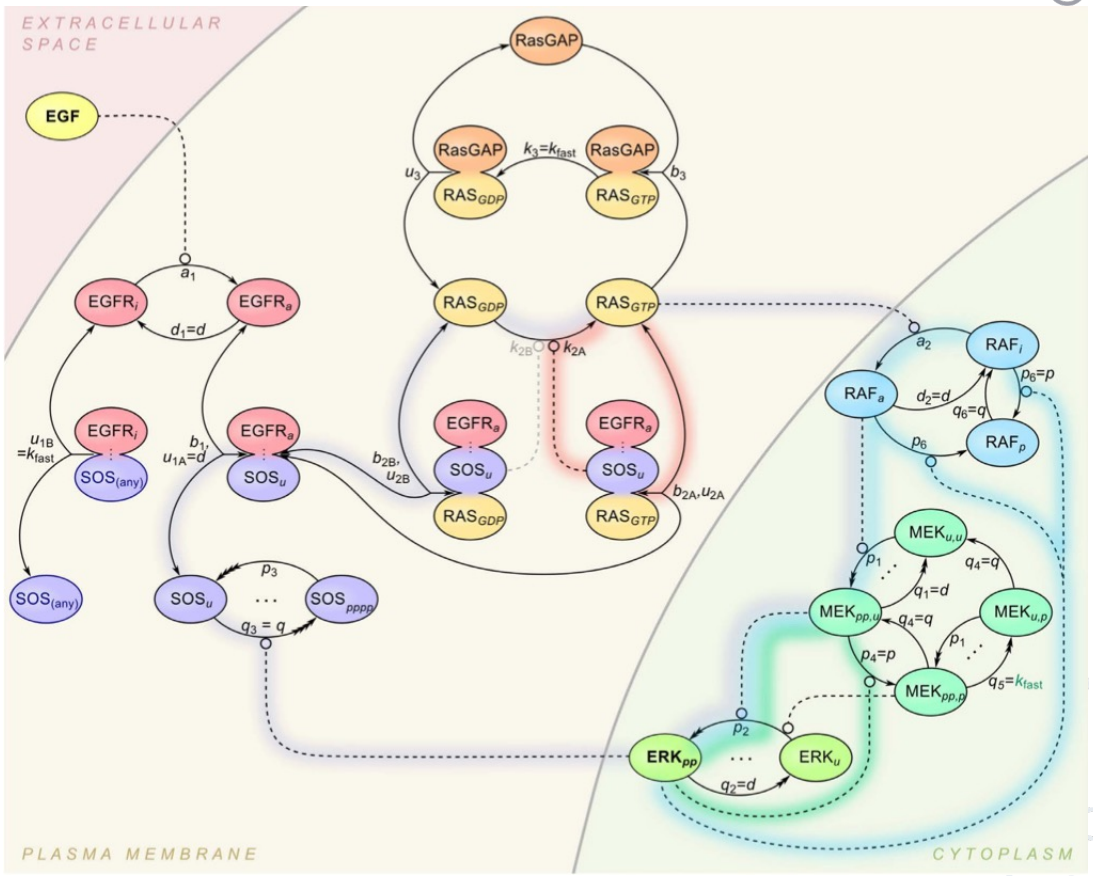
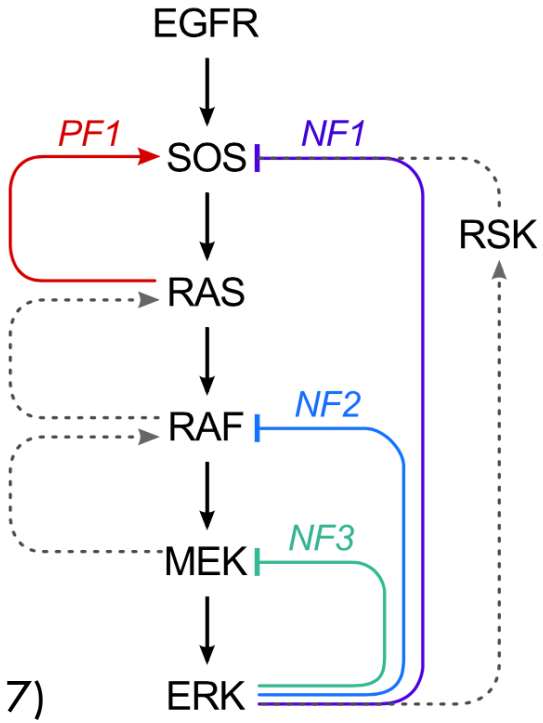
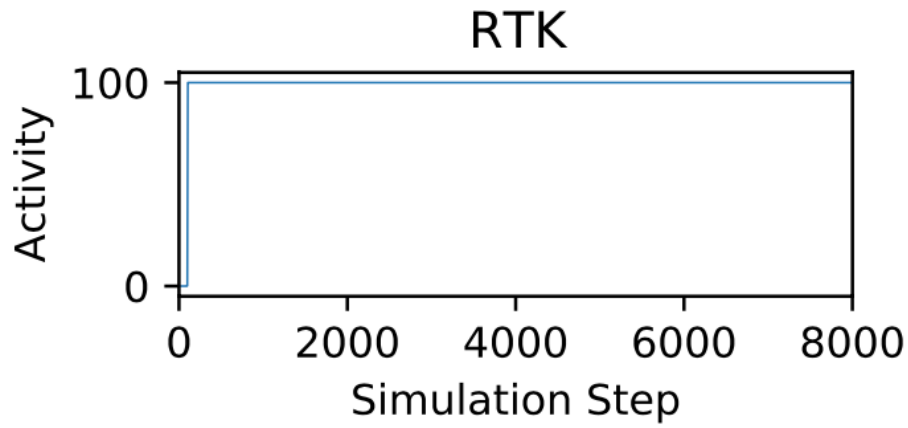


Figure adapted from Kochańzyk et al (2017)

ERK SIGNAL

- Stimulates cell proliferation
- Negative feedback causes oscillations
- On the right hand side, simulated signal





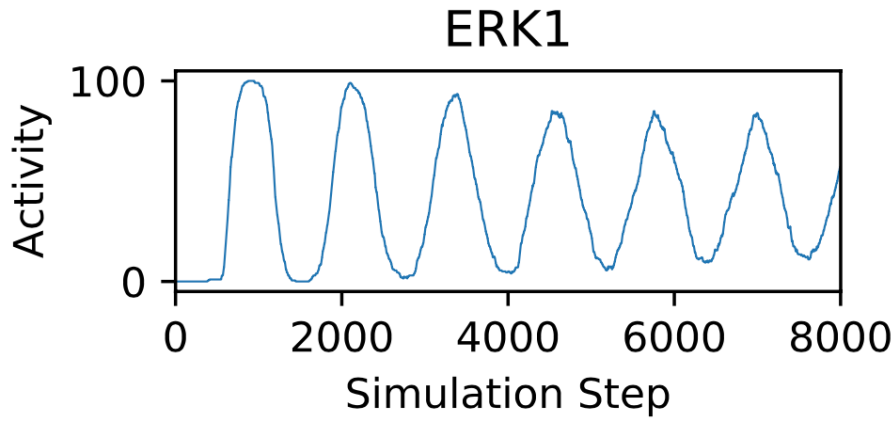


Figure adapted from Kochańzyk et al (2017)

HOW CANCER DRUGS WORK?

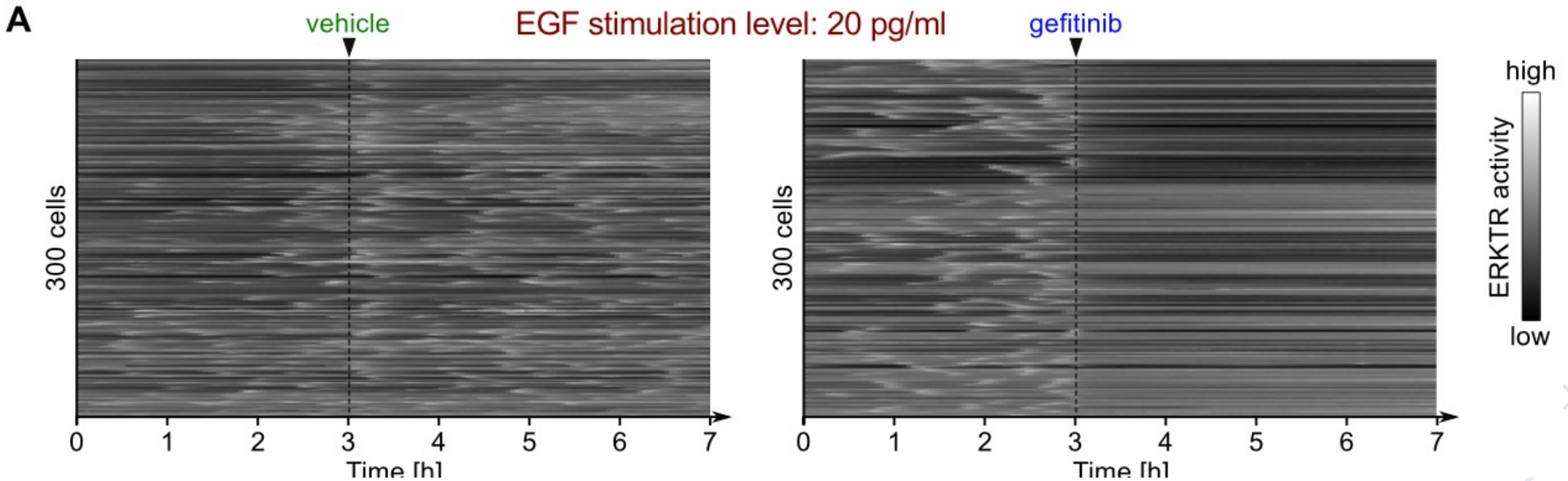
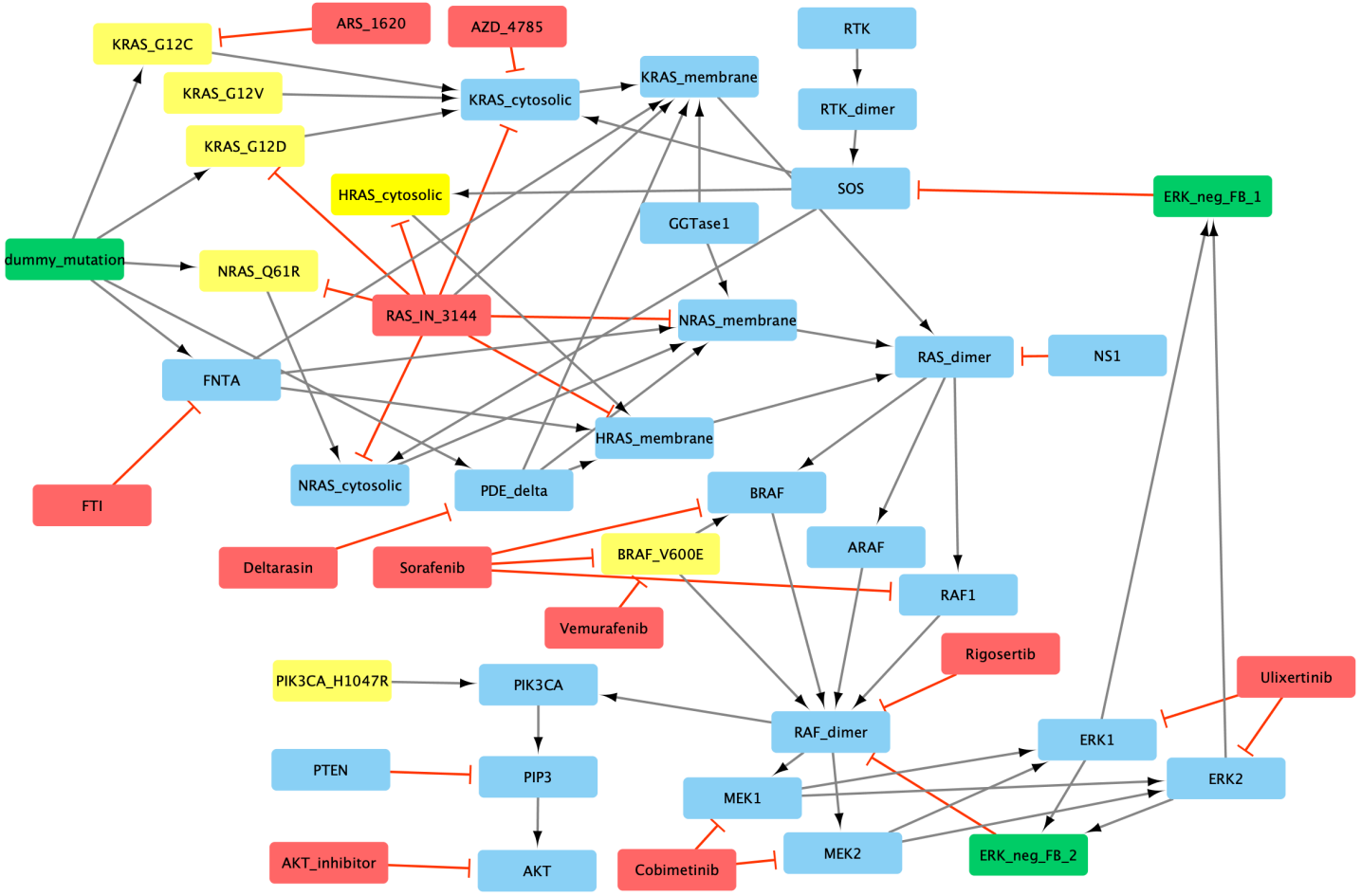


Figure adapted from Kochańzyk et al (2017)

MOST COMPREHENSIVE RAS PATHWAYS CANCER TREATMENT MODEL



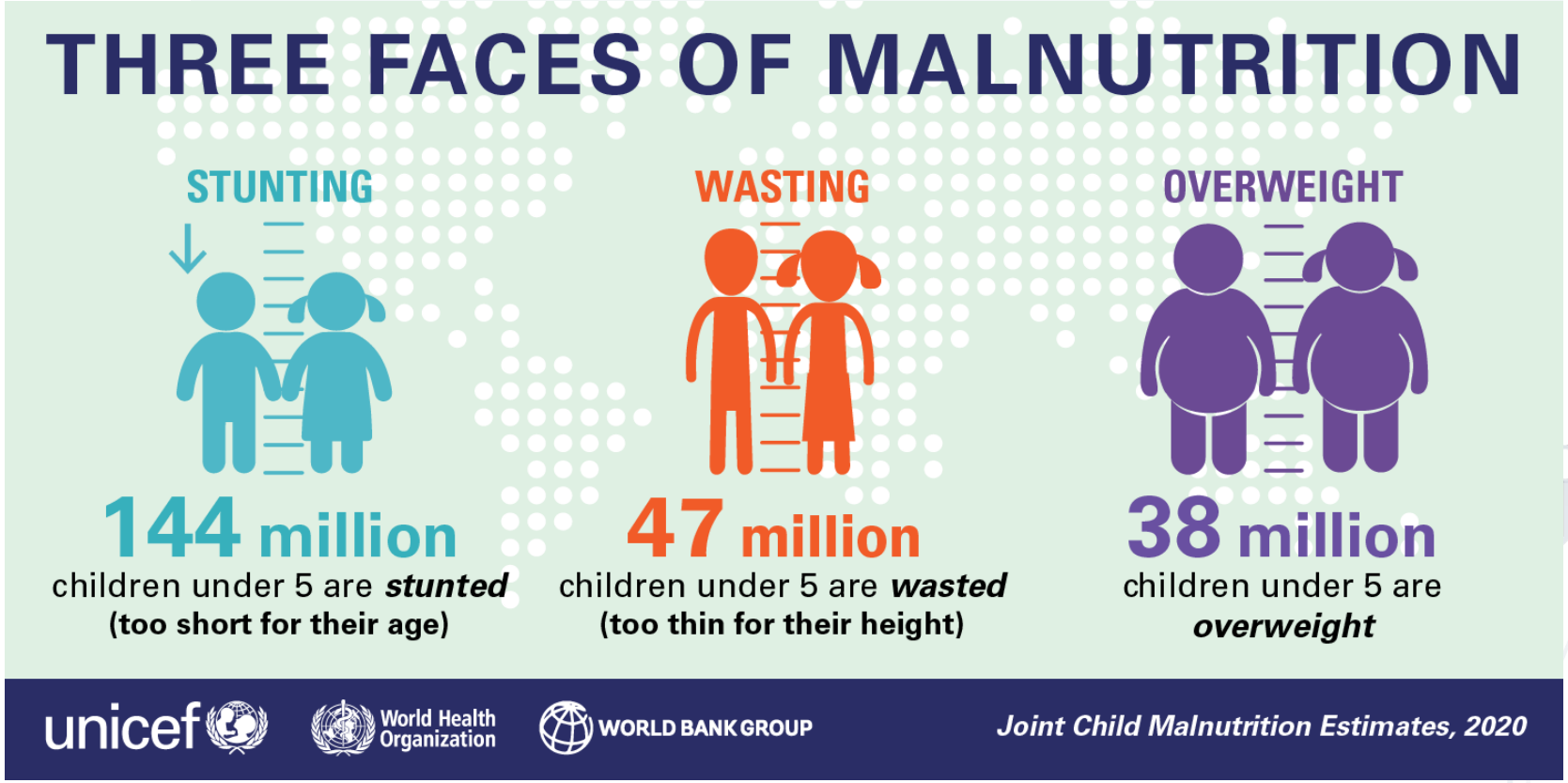
- Protein
- Inhibitor
- Mutation
- Auxilliary variable

FUTURE STEPS

- Model parameter tuning
- Validation on experimental data

CHILDREN MALNUTRITION

- Metrics (children under 5):
  - Stunting
  - Wasting
  - Overweight



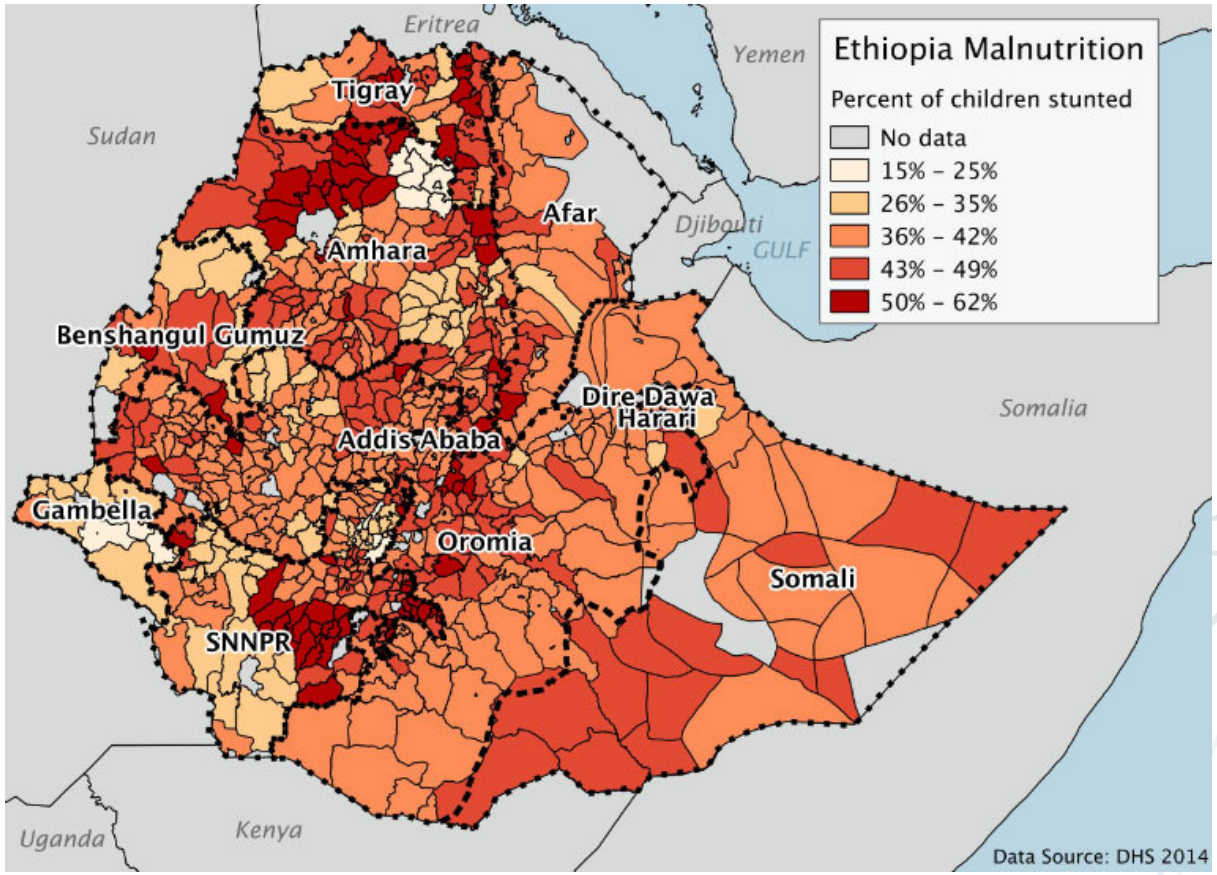
https://data.unicef.org/topic/nutrition/malnutrition/

CHILDREN MALNUTRITION IN ETHIOPIA

- Current statistics:

	Ethiopia	World
Severe wasting	1.2	2.1
Wasting	7.2	6.9
Overweight	2.1	5.6
Stunting	36.8	21.3
Underweight	21.1	13.0

Joint UNICEF, WHO, and World bank estimates in July 2020



World bank visualization of DHS 2014 data

FOOD INSECURITY IN ETHIOPIA

- Previous challenges:
  - 1983-1985 famine
  - 2011-present immigration from South Soudan
  - 2016 floods
- Anticipated shocks
  - COVID-19
  - Locust swarms

Photo: Locusta migratoria, by Gilles San Martin Wikimedia commons

Cases

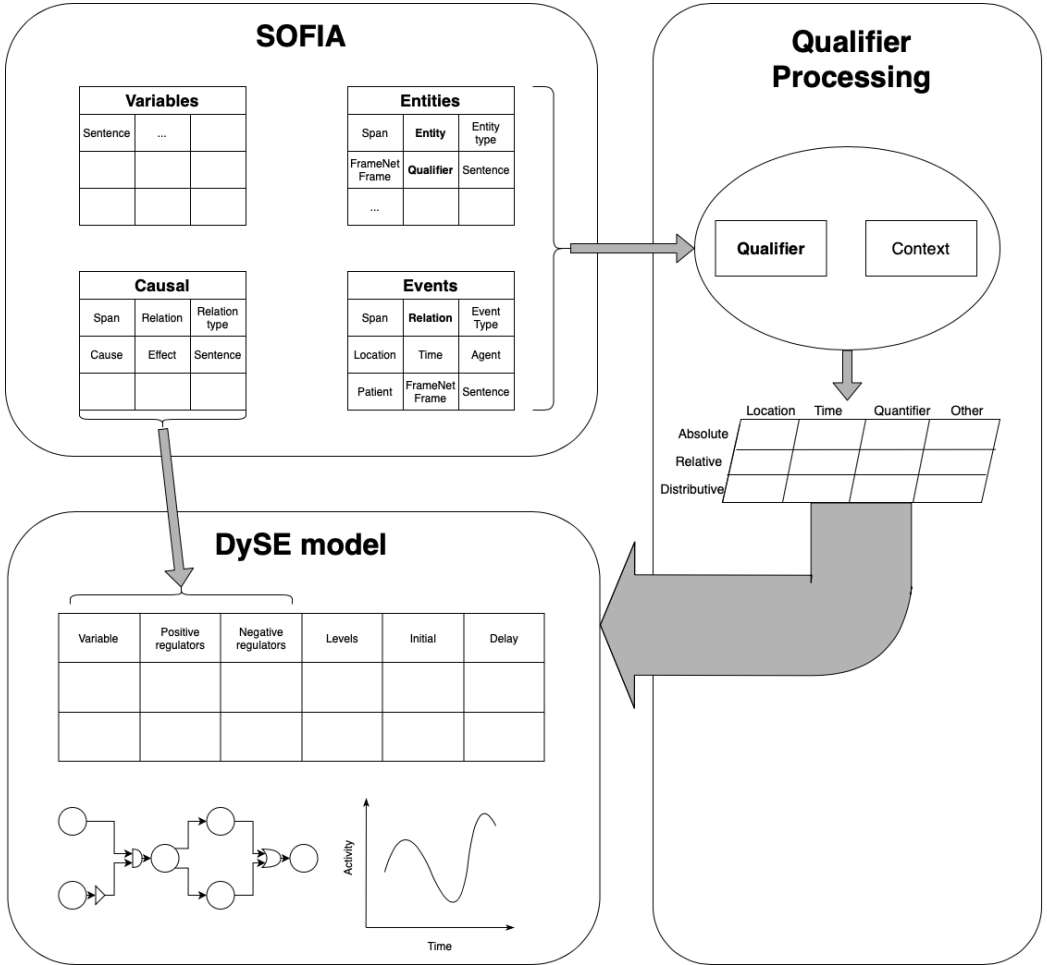
Photo: Wikimedia Commons https://www.theguardian.com/world/2011/sep/20/ethiopia-famine/

HOW DO WE OPTIMIZE AID TO ETHIOPIA WHILE MINIMIZING COSTS?

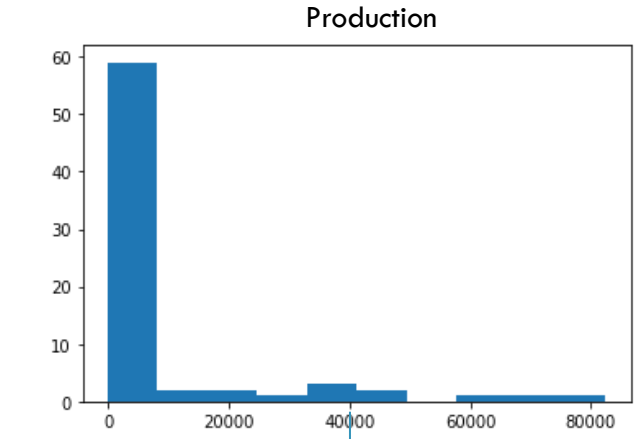
- Build a causal analysis graph (CAG)
  - Identify the relevant factors (e.g. crop yield, food price, available medicine, physical security, climate, etc.)
  - Identify interactions between these factors and their directionality
- Integrate data
  - Define indicators (features) to represent the factors (nodes)
  - Obtain indicator datasets
- Define the dynamic model
  - Infer dynamic model parameters
  - Perform dynamic simulations for forecasting
- Conduct goal optimization

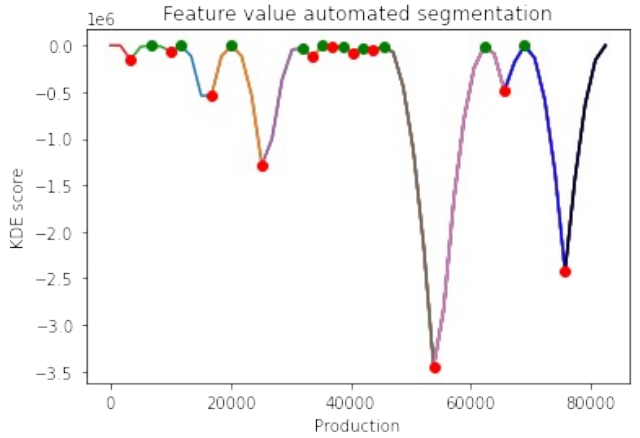
AVAILABLE MODELING RESOURCES

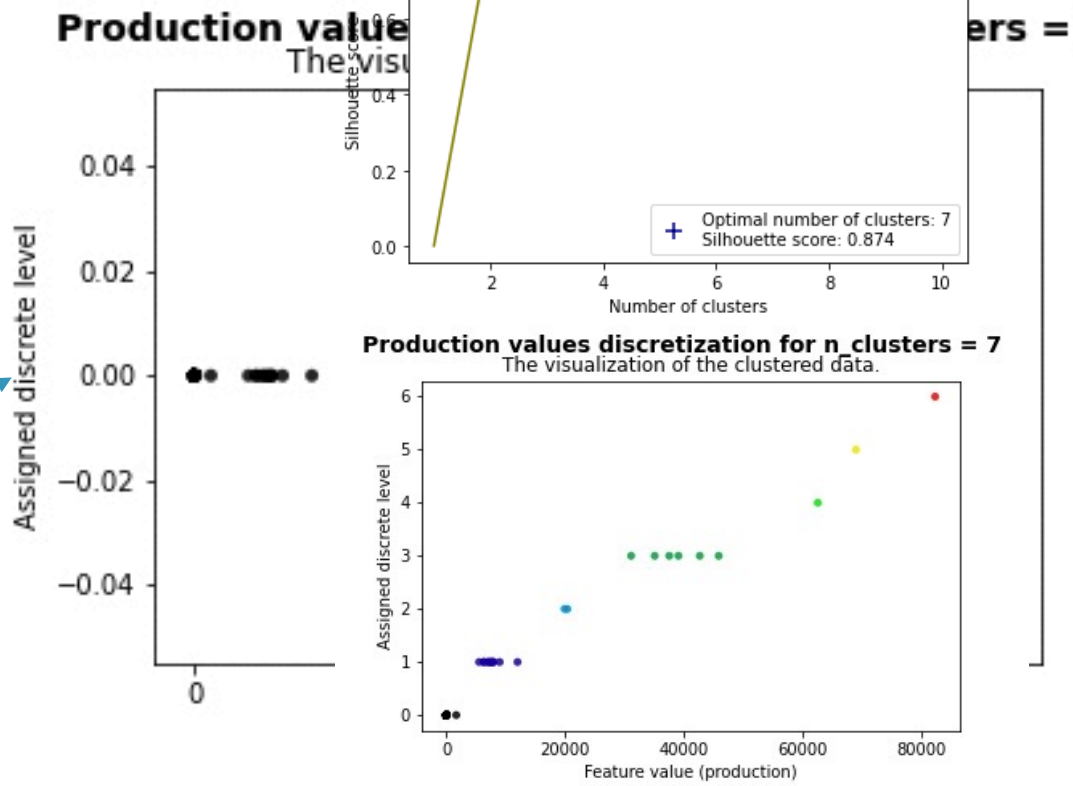
- Domain expert simulated datasets
- Historical spatiotemporal data
- Machine readers (SOFIA)
- **Dynamic Systems Explanation (DySE) framework**
  - Discrete extension of Boolean networks
  - Requires:
    - list of nodes
    - positive and negative regulation
    - value discretization and calibration



DISCRETIZATION AND CALIBRATION

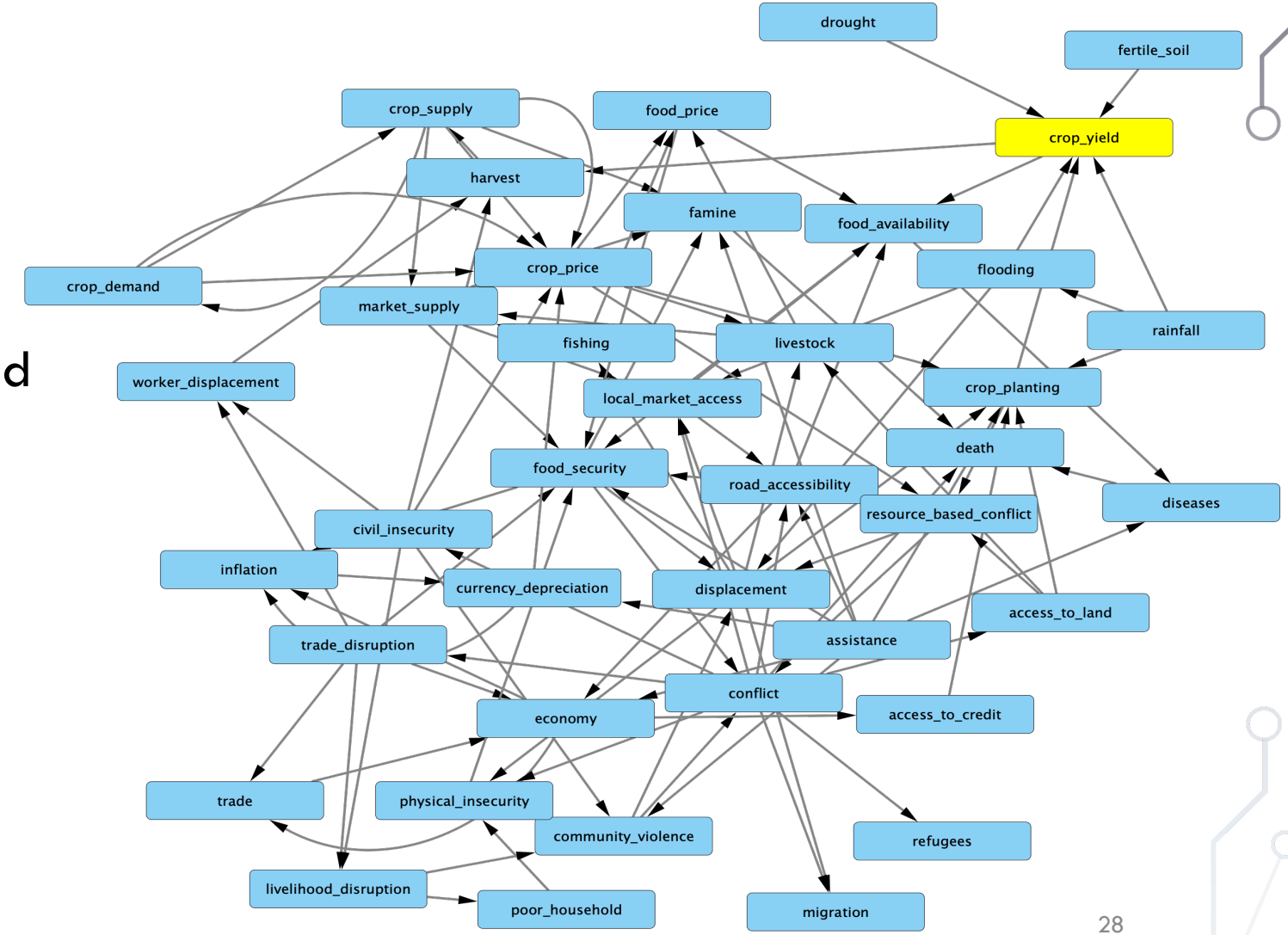






CAG

- Factors can globally be grouped into:
  - Climate
  - Agricultural production
  - Public health
  - Economy
  - Physical security
  - **Aid/Intervention**



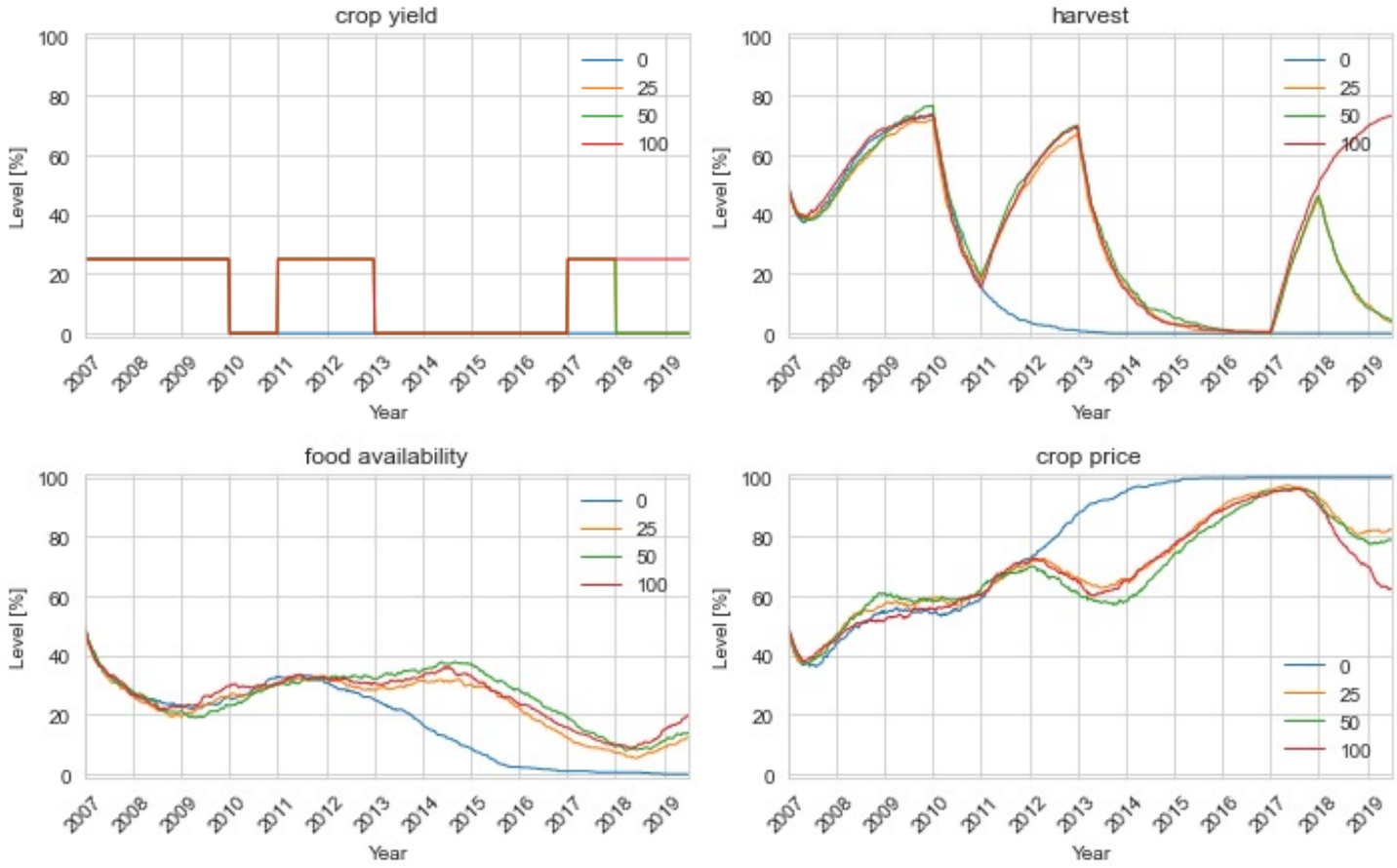
AGRICULTURAL PRODUCTION MODELING

- DSSAT model (Decision Support System for Agrotechnology Transfer, November 2019)
- Selected a single location to study:
  - Oromia, Misraq Sheva,
    8.625° N, 38.792° E
- There are 4 modeled crop types: maize, teff, sorghum, and wheat
- 3 parameters variation comparisons:
  - Fertilizer
  - Management practice
  - Planting window shift

FERTILIZER COMPARISON

Parameter	Value
Fertilizer*	**0,25,50,100**
Planting window shift	-30
Management practice	rf_lowN
Rainfall	0.50 (low)
Crop	Wheat
Season	Meher

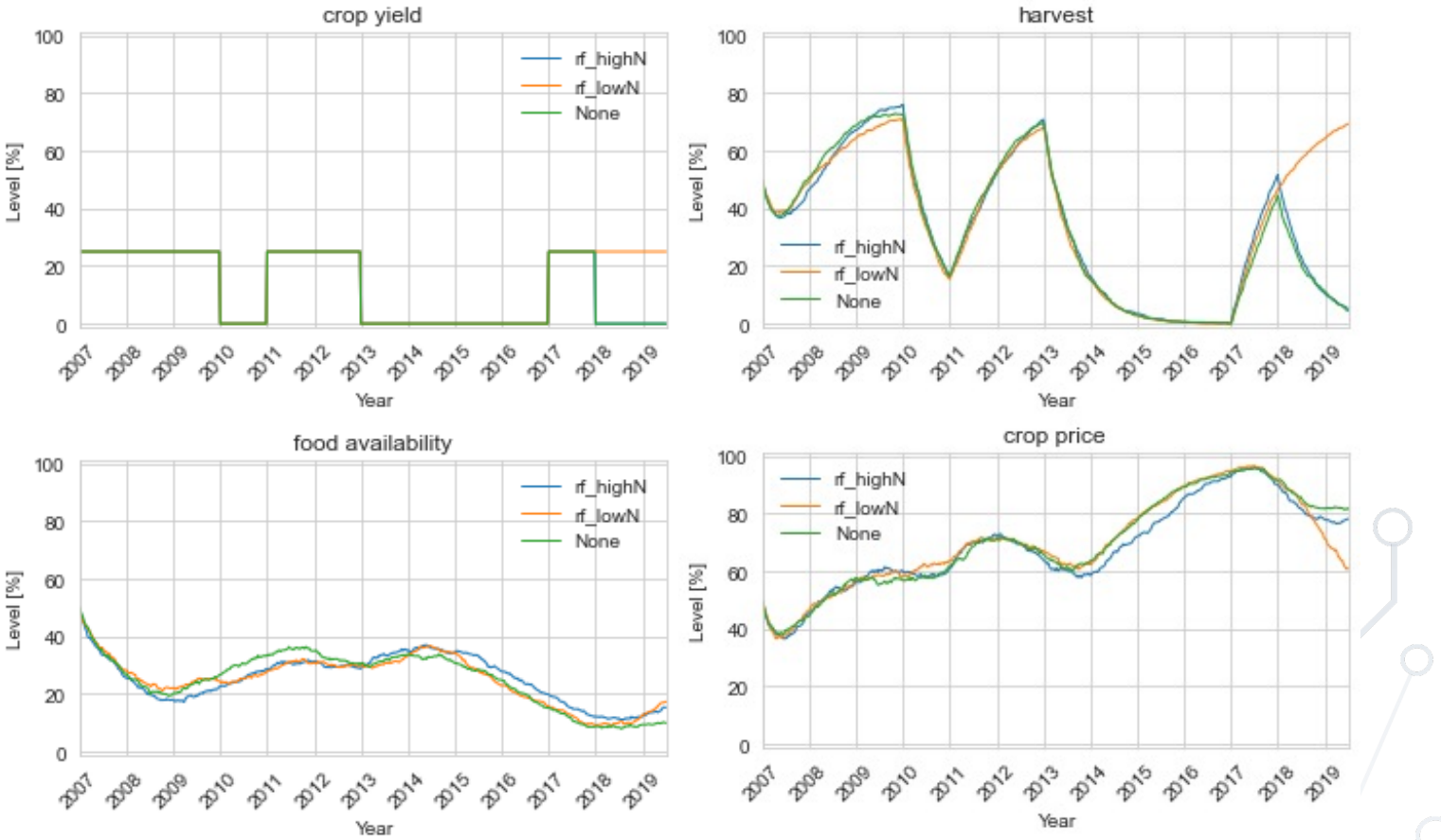
*varied parameter



MANAGEMENT PRACTICE COMPARISON

Parameter	Value
Fertilizer	100
Planting window shift	-30
Management practice*	**rf_highN, rf_lowN, None**
Rainfall	0.50 (low)
Crop	Wheat
Season	Meher

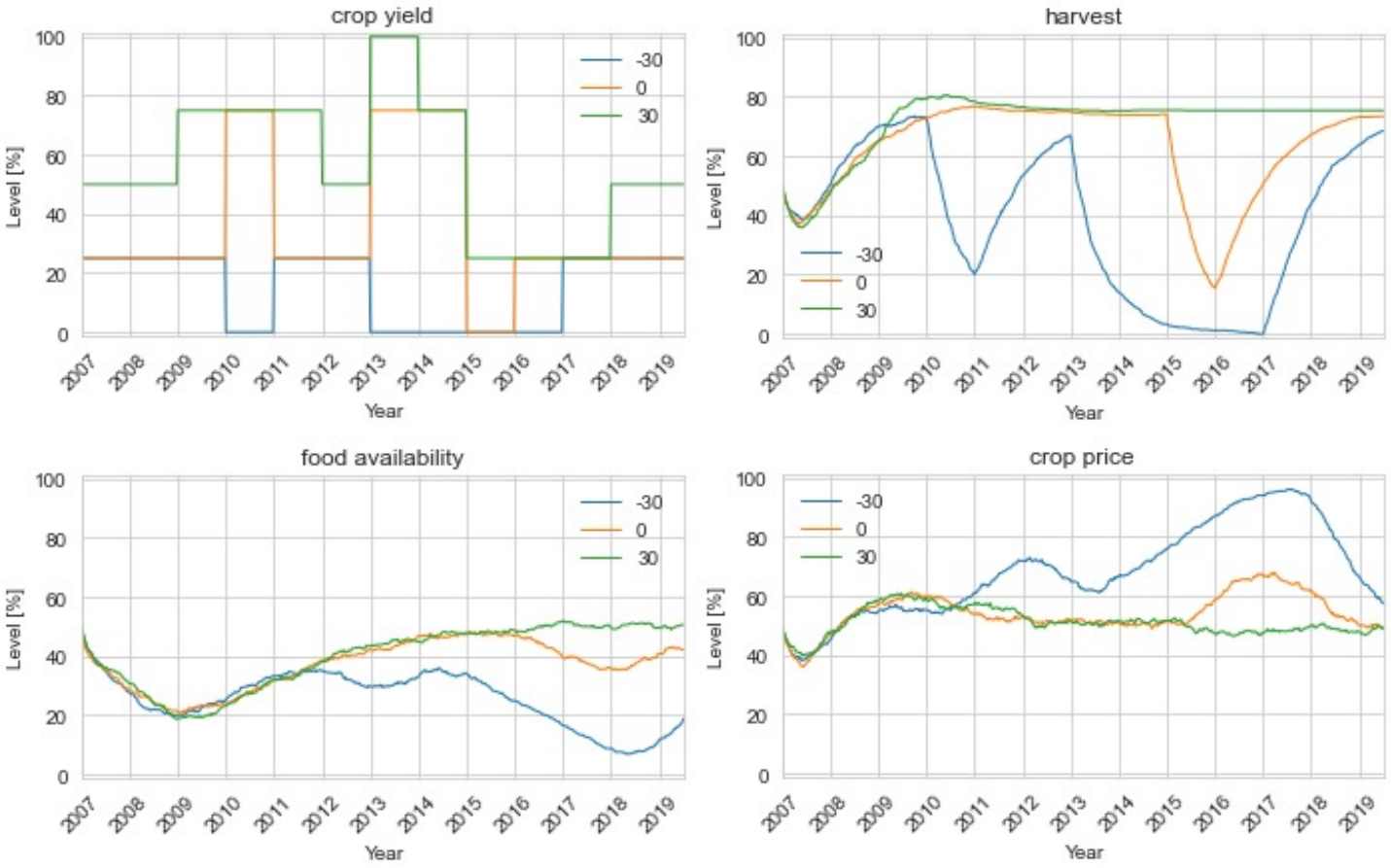
*varied parameter



PLANTING WINDOW SHIFT COMPARISON

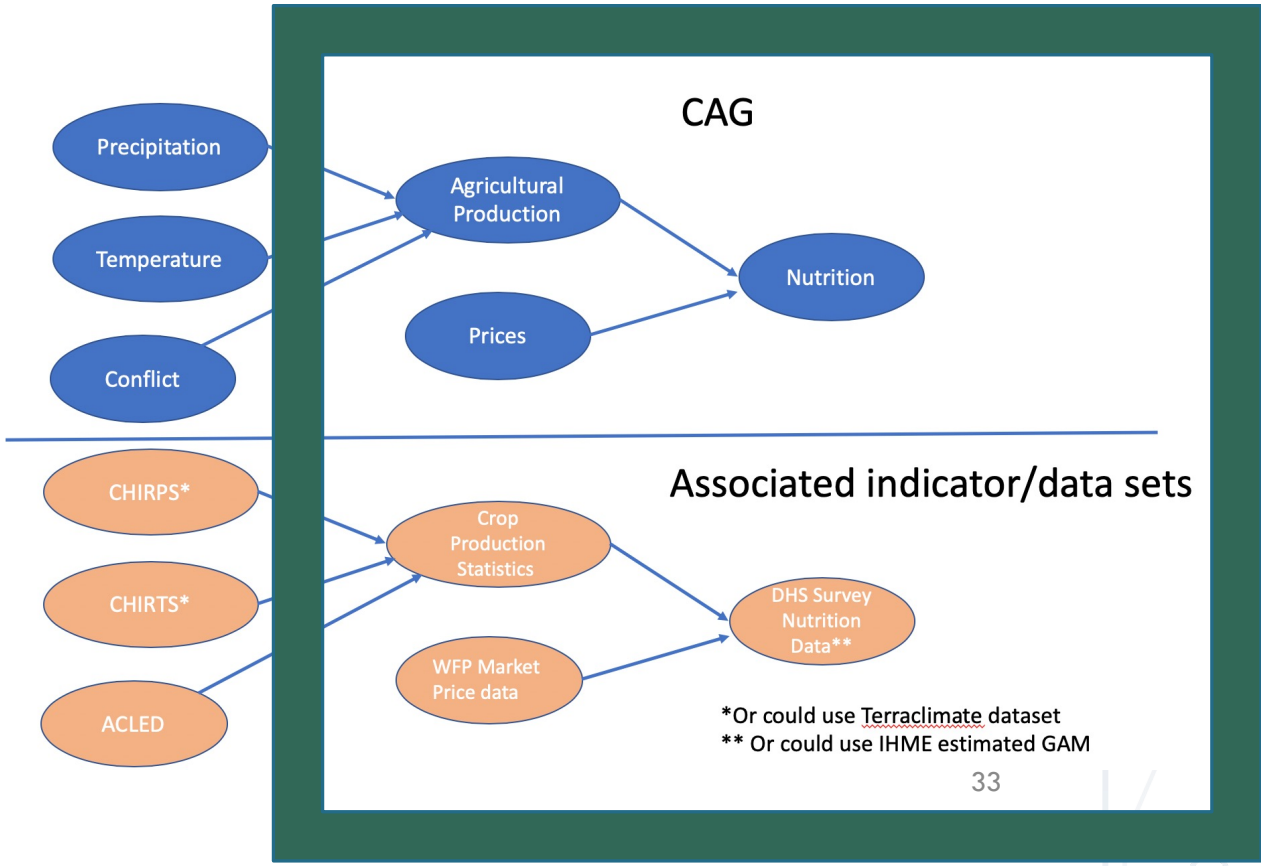
Parameter	Value
Fertilizer	100
Planting window shift*	**-30,0,30**
Management practice	rf_lowN
Rainfall	0.50 (low)
Crop	Wheat
Season	Meher

*varied parameter

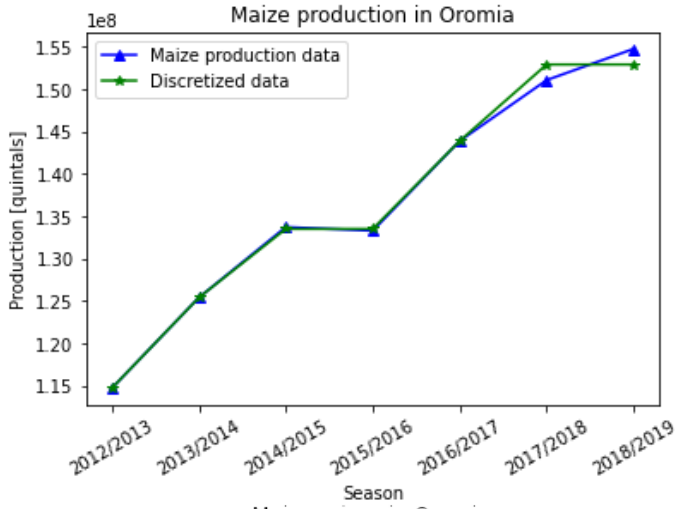


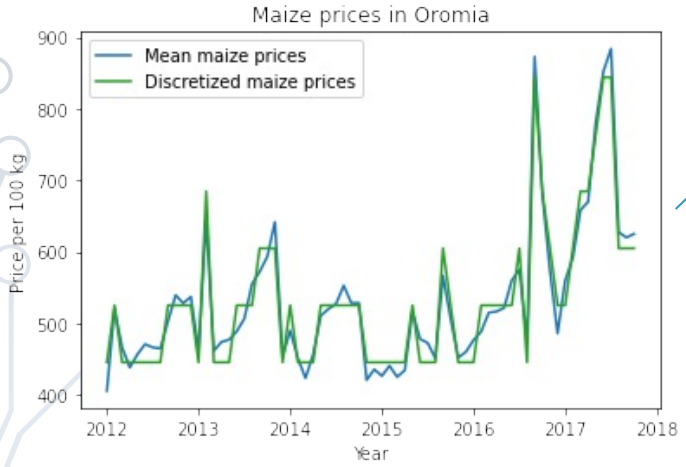
MINI USE-CASE: FROM AGRICULTURAL PRODUCTION TO MALNUTRITION

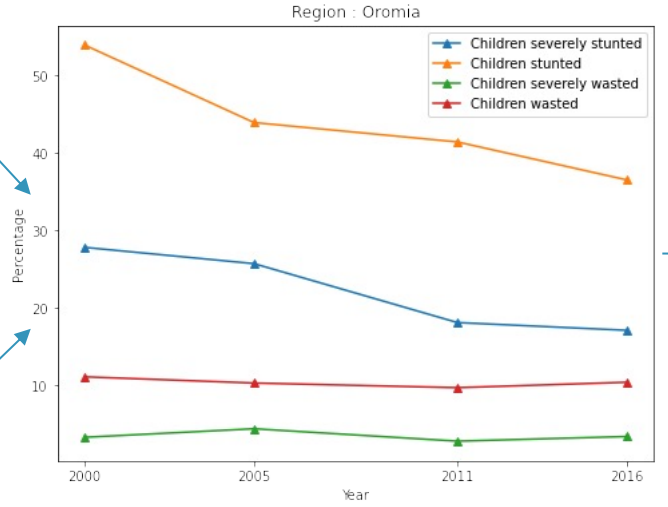
- We built a mini CAG
- Agricultural production historical spatiotemporal data from Food and Agriculture Organization (FAO) of the United Nations
- Food prices historical spatiotemporal data from World Food Program (WFP, Nobel Laureate for Peace 2020)
- 2 datasets for malnutrition:
  - Simulated (from IHME)
  - Historical survey estimates from (DHS): from 2000, 2005, 2011, and 2016

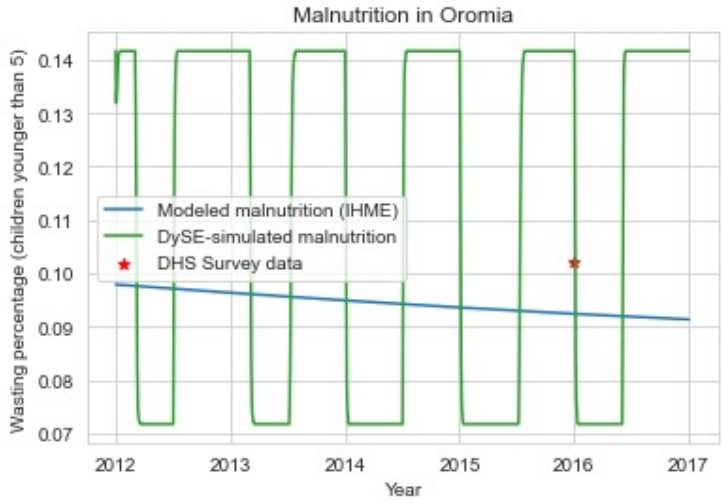


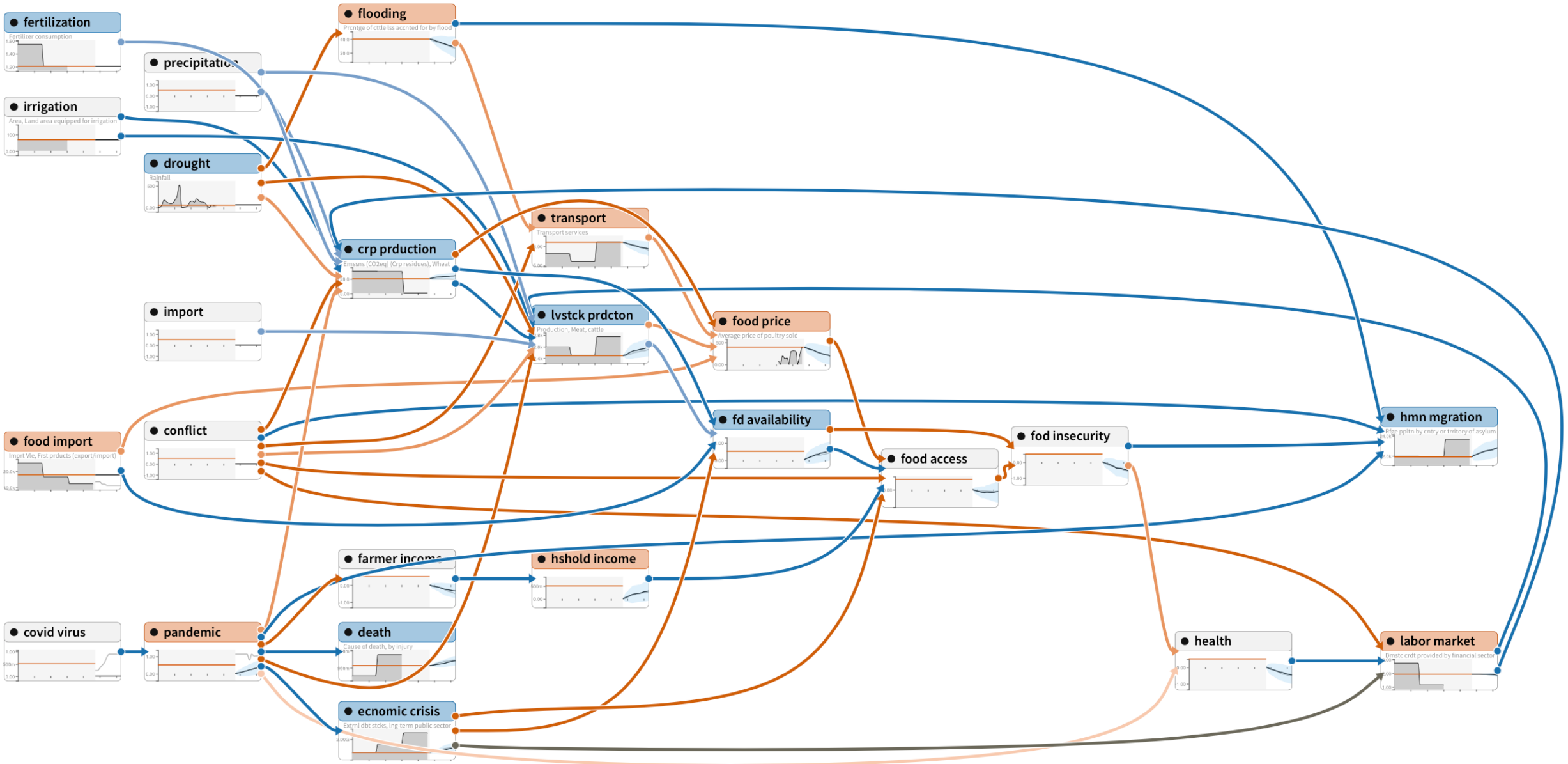
PRELIMINARY RESULTS










fertilization
Fertilizer consumption
flooding
Prcntge of cttle lss accnted for by flood
precipitation
irrigation
Area, Land area equipped for irrigation
drought
Rainfall
transport
Transport services
crp prduction
Emssns (CO2eq) (Crp residues), Wheat
import
lvstck prdcton
Production, Meat, cattle
food price
Average price of poultry sold
food import
Imprt Vle, Frst prducts (export/import)
conflict
fd availability
fod insecurity
hmn mgration
Rfge ppltn by cntry or trritory of asylum
food access
farmer income
hshold income
covid virus
pandemic
death
Cause of death, by injury
health
labor market
Dmstc crdt provided by financial sector
ecnomic crisis
Extrnl dbt stcks, lng-term public sector

FUTURE WORK

- Sensitivity analysis
- Goal optimization
- Exploring different shock scenarios

LESSONS LEARNT #3

- Academia can look a lot like industry, depending on your funding source and management structure
- As opposed to industry where learning and personal development is side effect, in academia it's central and present in all activities
- Academia will teach you modesty, understanding pitfalls of your methods, while industry is less perfectionist, but more aware of commercial value of your work

Piece of advice:

- If you decide on doing a PhD, try to plan and stay organized, spare effort on time management, because these will be the key factors to your PhD duration.

ACKNOWLEDGEMENTS

- Thanks to OUBES for the invitation!
- Thanks to MeLoDy lab for the exciting research opportunities
- Thanks to CPCB PhD program for overwhelming me with ML and biomedical knowledge
- Thanks to DARPA for funding our projects