Models and Simulation

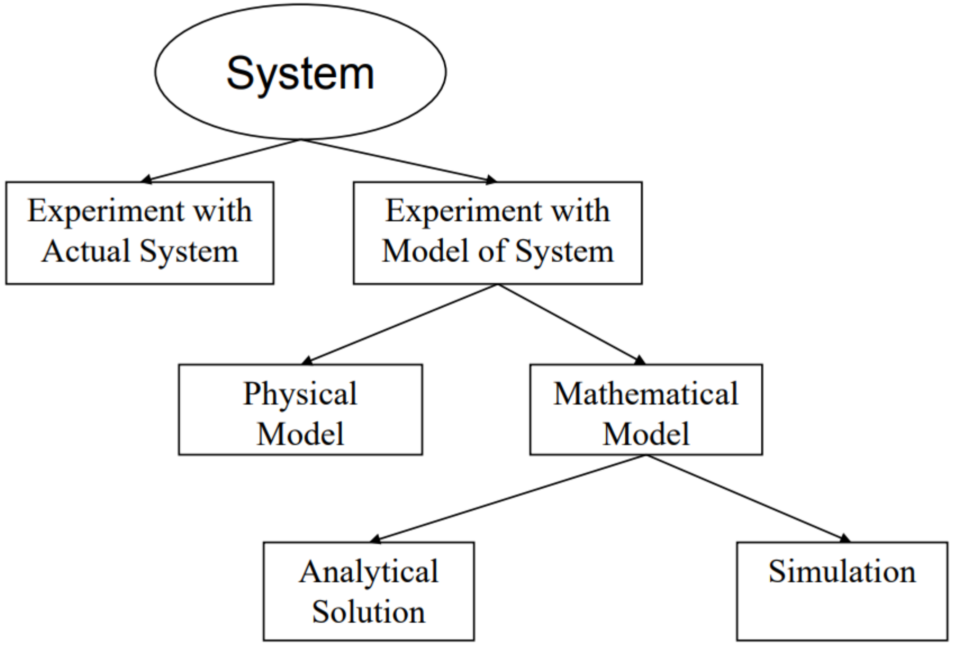
Note:

Switched class order so that there would be no Homework associated with today's lecture. Enjoy your Thanksgiving!

Outline

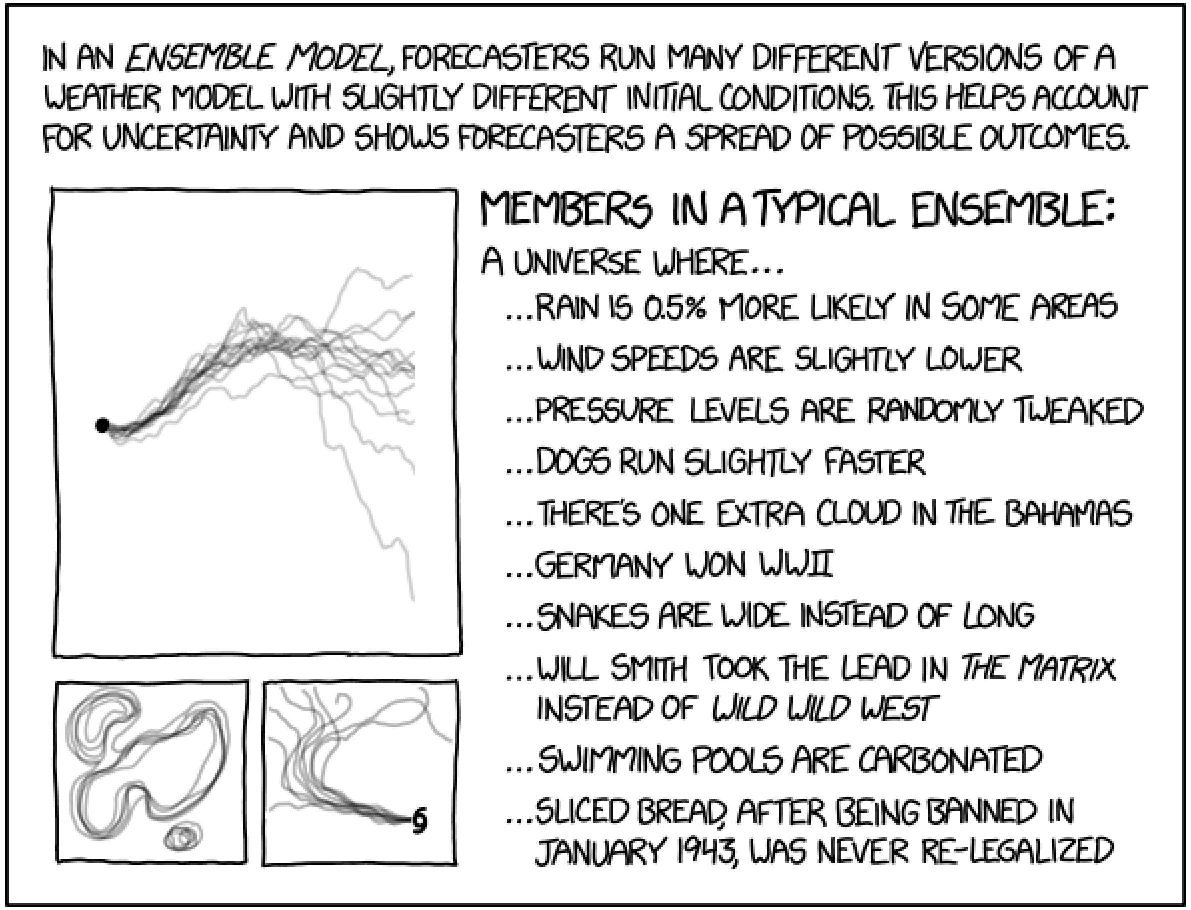
- Models
- Best Practices
- Discrete Event Simulation
- Agent Based Modeling

Models and Simulations



Law & Kelton (2000), Simulation Modeling and Analysis 3rd ed., McGraw-Hill, Inc.

- A **model** is a mathematical description of a system that is used to approximate the behavior under a defined set of operating conditions and simplifying assumptions.
- A **simulation** is the running of a model for a particular set of inputs.


IN AN *ENSEMBLE MODEL*, FORECASTERS RUN MANY DIFFERENT VERSIONS OF A WEATHER MODEL WITH SLIGHTLY DIFFERENT INITIAL CONDITIONS. THIS HELPS ACCOUNT FOR UNCERTAINTY AND SHOWS FORECASTERS A SPREAD OF POSSIBLE OUTCOMES.
MEMBERS IN A TYPICAL ENSEMBLE:
A UNIVERSE WHERE...
...RAIN IS 0.5% MORE LIKELY IN SOME AREAS
...WIND SPEEDS ARE SLIGHTLY LOWER
...PRESSURE LEVELS ARE RANDOMLY TWEAKED
...DOGS RUN SLIGHTLY FASTER
...THERE'S ONE EXTRA CLOUD IN THE BAHAMAS
...GERMANY WON WWII
...SNAKES ARE WIDE INSTEAD OF LONG
...WILL SMITH TOOK THE LEAD IN *THE MATRIX* INSTEAD OF *WILD WILD WEST*
...SWIMMING POOLS ARE CARBONATED
...SLICED BREAD, AFTER BEING BANNED IN JANUARY 1943, WAS NEVER RE-LEGALIZED

Example: Boarding Airplanes (Actual System)

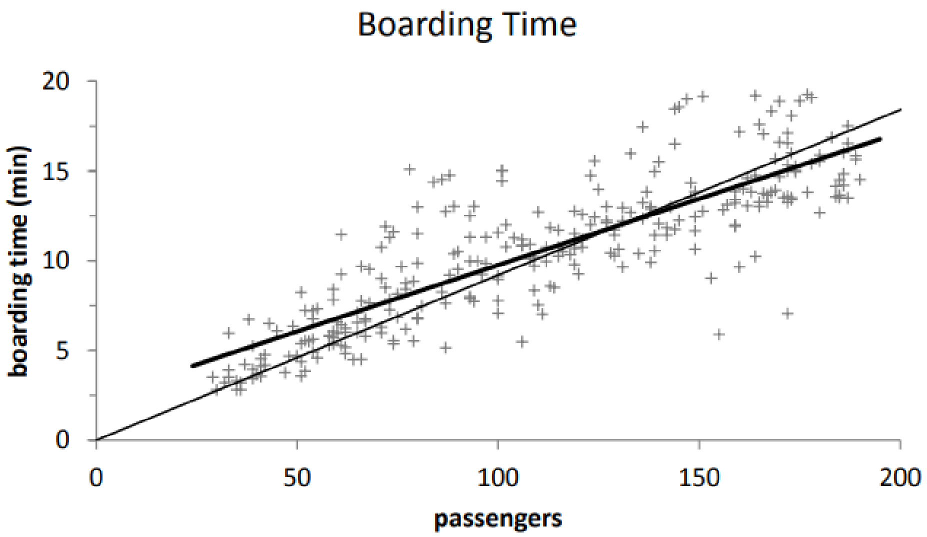


Figure 3. Boarding times (282 measured flights)

- Actual boarding times from airlines with different boarding policies can be measured or a single airline can do A/B Testing to compare boarding times
- From "Aircraft Boarding – Data, Validation, Analysis" M Schultz, ATM2017

Example: Boarding Airplanes (Actual System)

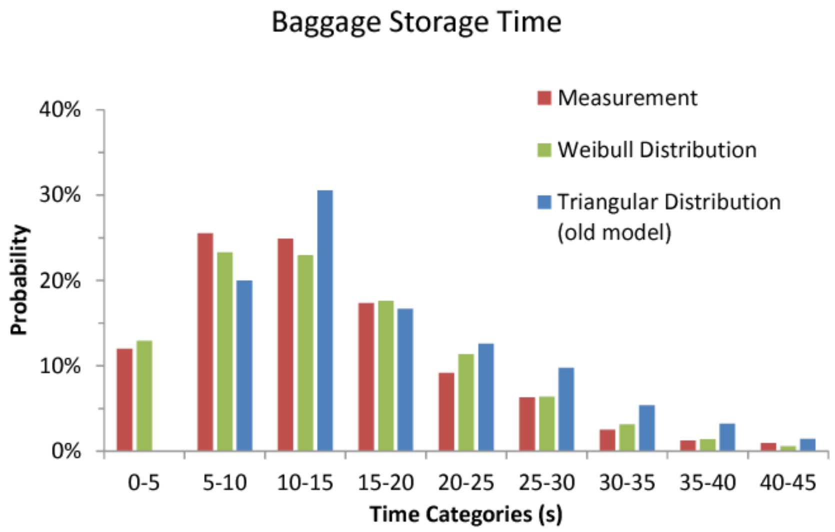


TABLE III. COMPARISON OF BOARDING PROGRESS USING REAL DATA FOR CALIBRATION

Boarding Strategies	Boarding time (%)						
	data	*sim.*	*diff.*	*Q.10*	*Q.25*	*Q.75*	*Q.90*
random	101.4	100.0	*1.4*	-8.6	-4.6	4.9	9.5
airline - S1	93.7	104.5	*-10.8*	-9.3	-5.1	5.2	10.2
airline - S2	87.0	83.8	*3.2*	-7.4	-4.0	4.4	8.4
**airline - S2*		**80.5**					
	Seat Load Factor 76% ± 5%						
random	102.6	100.0	2.6	-10.6	-5.7	6.2	11.8
airline - S1	94.8	98.7	-3.9	-11.5	-6.3	6.6	12.7
airline - S2	88.0	83.4	4.6	-8.9	-4.8	5.2	10.2
**airline - S2*		**80.8**					

- Validating model parameters with real world measurements is costly but often important for model accuracy
- From “Aircraft Boarding – Data, Validation, Analysis” M Schultz, ATM2017

Example: Boarding Airplanes (Physical Model)

Front						
24						18
12						6
23						17
11						5
22						16
10					28	4
21						15
9					27	3
20						14
8					26	2
19						13
7					25	1

Figure 4: Steffen method seating order.

Table 1: Elapsed time for each boarding method in minutes and seconds.

Method	Official Time	Extended Time
Back–Front	6:11	6:16
Blocks	6:54	6:56
Wilma	4:13	4:21
Steffen	3:36	3:40
Random	4:44	4:48

- Using a mock 757 fuselage on a soundstage with 12 rows of six seats and 72 passengers of various ages
- Taken from“Experimental test of airplane boarding methods” JH Steffen and J Hotchkiss, Fermi Lab Publications

Potential Shortcomings of Physical Model

- Types and amount of luggage, coats, food, etc.
- Groups of people traveling together (with children)
- Distribution of mobility or health issues
- Experience level of travelers (plus repeated experiments has potential for learning)
- Seat preference for open seating arrangements
- Other Factors: Limited overhead, multi-leg flights

Example: Boarding Airplanes (Analytic Model)

- D – A delay distribution. The delay values d_i will be sampled from the distribution D.
- h – The number of passengers per row.
- l – Distance between rows (leg room).

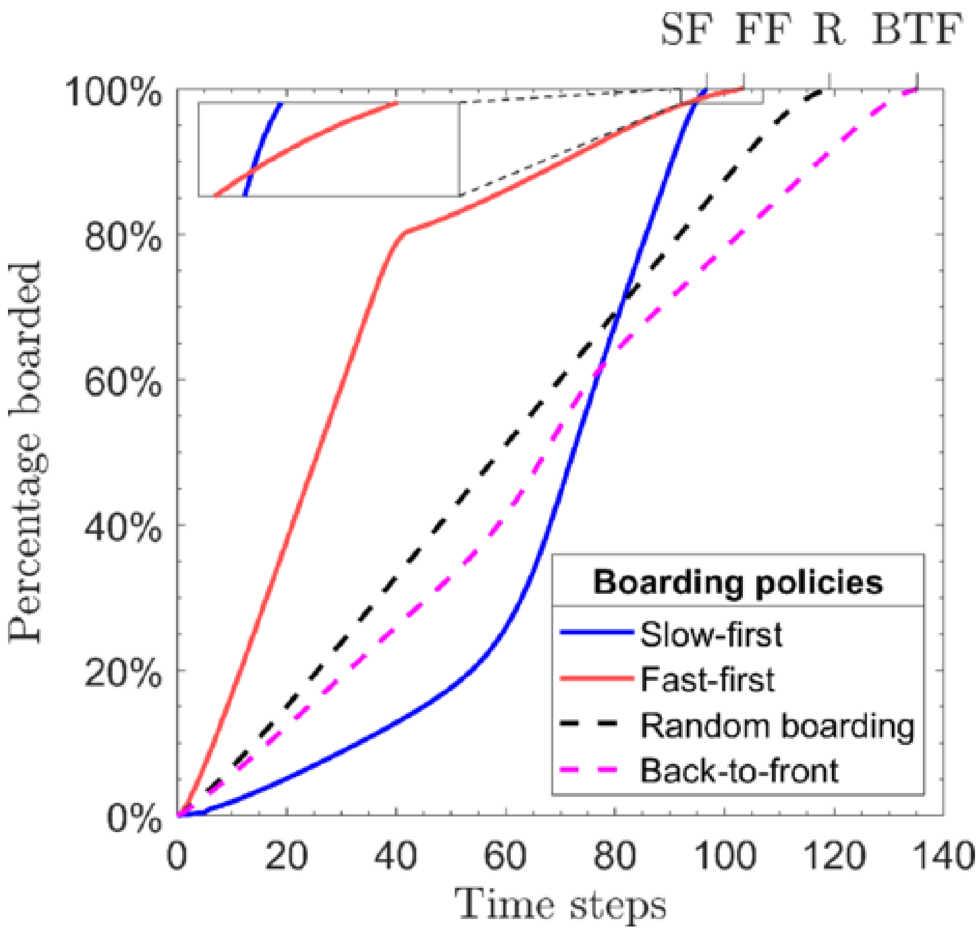
For simplicity of presentation we will assume that D, h, l are all constant throughout the paper.

- W – A width distribution. The values w_i will be sampled from W.
- F – An airline boarding policy. We represent an airline boarding policy by a function F. $F(r)$ indicates the first time at which passengers from row r are allowed to join the boarding queue. If passengers follow the boarding policy, their (q, r)-coordinates satisfy $q \geq F(r)$.
- Ω – A passenger's reaction model. We need to make an assumption as to the nature of the reaction of passengers to the airline policy. The reaction determines the effect of the boarding policy on the boarding process. For instance, in the extreme case in which passengers do not pay any attention to the airline's announcements, the boarding policy is irrelevant. In the other extreme, if passengers join the queue immediately after being allowed, the airline can fully control the queuing order. In this paper we will use the following parameterized reaction model:

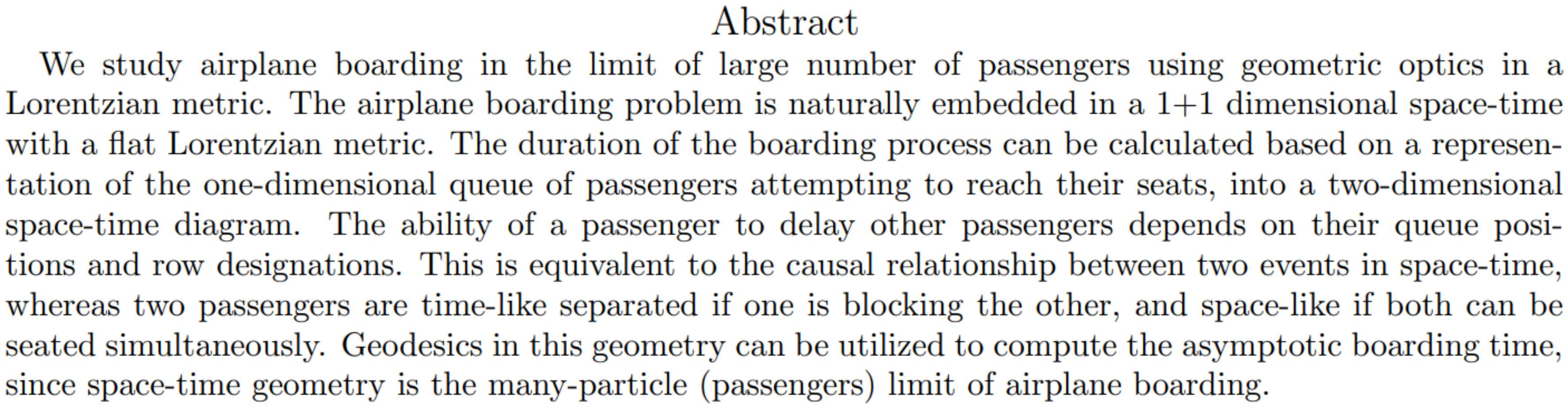
- The attentive reaction model with parameter T. In this model, passengers join the queue at uniformly distributed times, within T time units of being allowed to board.

- Analytic models typically rely on a more idealized version of the system, but can generate general rules and make certain optimality statements that more detailed systems cannot
- "Analysis of Airplane Boarding Times" E Bachmat et al.

Who is the Audience?



- Erland, Sveinung, et al. "Lorentzian-geometry-based analysis of airplane boarding policies highlights “slow passengers first” as better." *Physical Review E* 100.6 (2019): 062313.


Abstract

We study airplane boarding in the limit of large number of passengers using geometric optics in a Lorentzian metric. The airplane boarding problem is naturally embedded in a 1+1 dimensional space-time with a flat Lorentzian metric. The duration of the boarding process can be calculated based on a representation of the one-dimensional queue of passengers attempting to reach their seats, into a two-dimensional space-time diagram. The ability of a passenger to delay other passengers depends on their queue positions and row designations. This is equivalent to the causal relationship between two events in space-time, whereas two passengers are time-like separated if one is blocking the other, and space-like if both can be seated simultaneously. Geodesics in this geometry can be utilized to compute the asymptotic boarding time, since space-time geometry is the many-particle (passengers) limit of airplane boarding.

Example: Boarding Airplanes (Simulation)

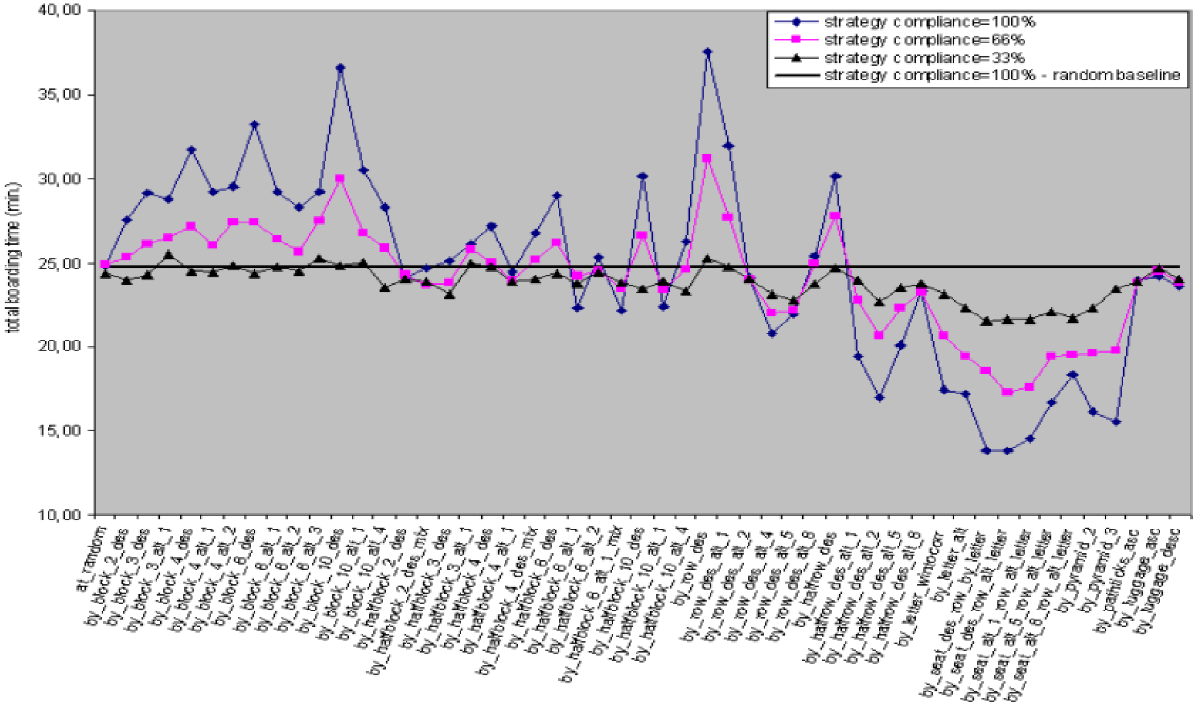


Figure 3: Effect when passengers have a strategy compliance degree of respectively 100% , 66% and 33%

Parameter name	Parameter Value
Step multiplier	1
Path and seat ticks {min; modus; max}	{ 1.8 ; 2.4 ; 3}
Install ticks {min; modus; max}	{ 6; 9; 30}
Queue-delay ticks	5
Strategy compliance	100 %
air plane occupancy	100 %
Hand luggage distribution (1;2;3 pieces)	{60 % ; 30 % ;10 %}
Group Size {min; modus; max}	{ 1 ; 1 ; 1}

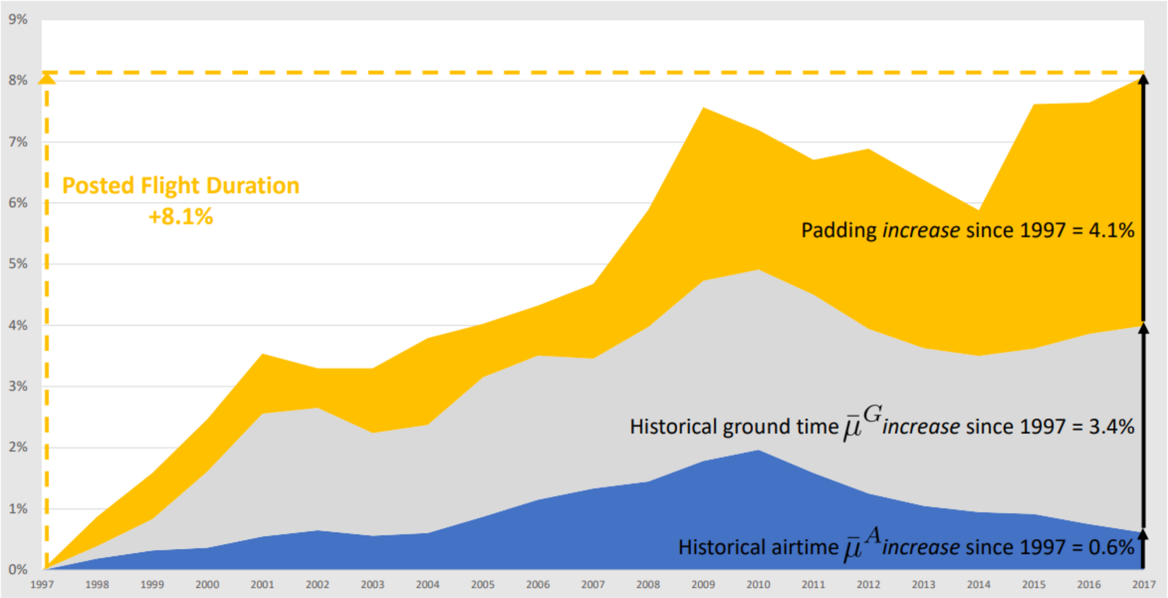
Table 1: Standard simulation model parameters; ticks and luggage distribution values originates from [5]

- Simulations can account for a greater variety of scenarios but can be dependent on the fidelity of the model
- Audenaert, Jan, Katja Verbeeck, and Greet Vanden Berghe. "Multi-agent based simulation for boarding." *The 21st Belgian-Netherlands Conference on Artificial Intelligence*. 2009.

Final Thoughts on Boarding

- What is(are) the true objective(s)? (Hint: $$$)
- What is the value of speeding up the boarding process? When is it a blocker to takeoff?
- Are there potential unintended consequences to changing the boarding process?
- Are there other pieces of the system that could be changed?

Consequences



- There were 965M domestic passengers in 2017, theoretically using the block boarding would take 12,730yrs vs an optimal 6,732yrs (a difference greater than recorded history)
- In practice software to reduce taxi times in a single airport has saved 275k gallons of fuel and reduce delays by 916 hrs in a year
- *Where did the time go? On the increase in airline schedule padding over 21 years*. Zhang et al. 2018

Why simulate?

- Cost effective
- Avoids possible ethical issues
- As a proof of concept before embarking on a more expensive trial run
- Can test a variety of “what if” scenarios

Some Types of Simulations

- Static Model
  - Contains no history of previous input values, internal variables, or output values (i.e. it has no sense of evolution with time)

- Dynamic Model
  - Allow for time-dependent behavior of systems, maintaining an internal state that tracks some combination of prior inputs, internal variables, or outputs

Some Types of Simulations

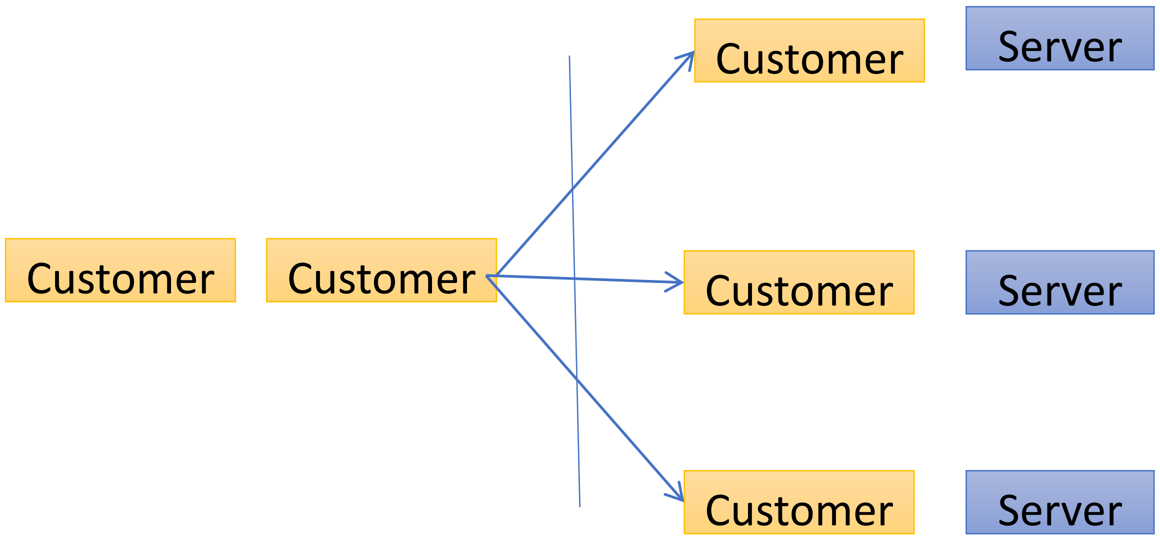
- Deterministic Model
  - The program returns the exact same output given the same input (does not use random number generator)

- Stochastic Model
  - The model contains random elements (though should be reproducible with the use of a seeded random number generator)

Some Types of Simulations

- Discrete
  - Variables such as time and location are discretized to ease the progression of the model
- Continuous
  - Variables are continuous and evolve as such

Some Popular Classes of Simulation

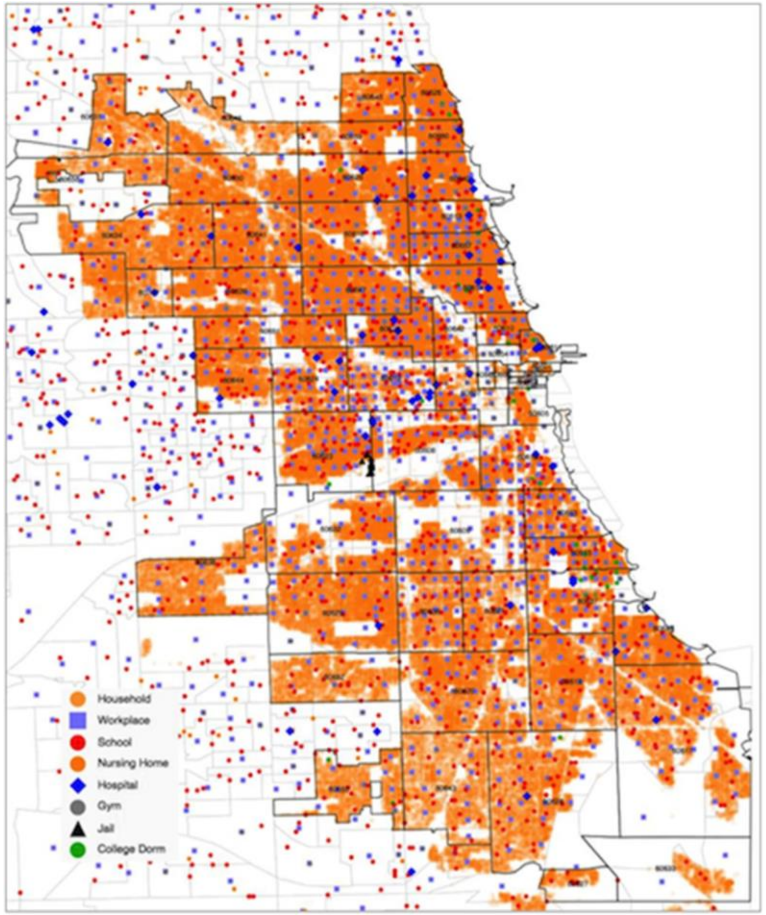


- Discrete-Event Simulation (DES)
  - Is a *stochastic* model that can simulate system *dynamics* with significant changes happening at *discrete* time instances

Example:

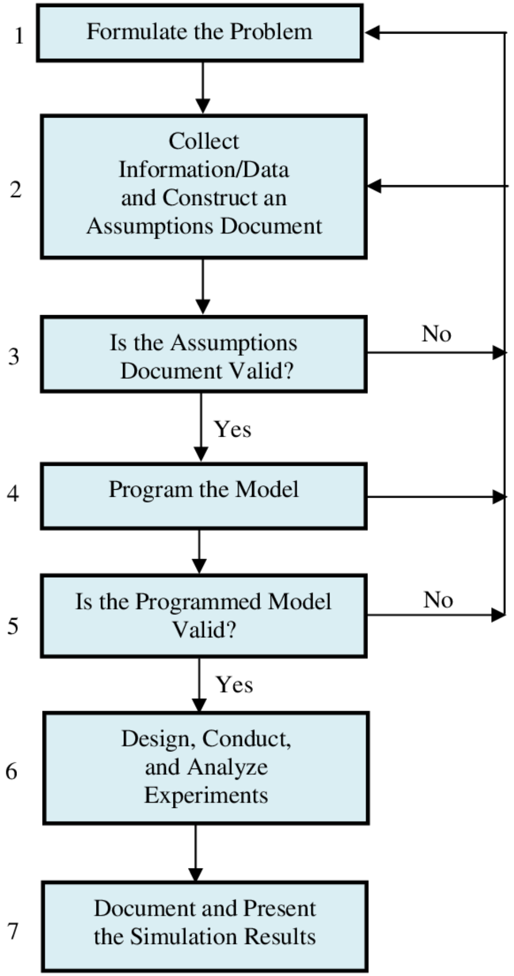
A queueing system with arrival times occurring according to a random distribution and changes in the system parameters (i.e. queue length, etc) occurring at discrete times

Some Popular Classes of Simulation



- Agent-Based Models are a class of models for simulating the actions and interactions of autonomous agents with a view of creating emergent system wide behavior from individual agents (stochastic, dynamic, discrete)

Some Best Practices



- Validation is the process of determining whether a simulation model is an accurate representation of the system, *for the particular objectives of the study*
- Law, Averill M. "How to build valid and credible simulation models." *2019 Winter Simulation Conference (WSC)*. IEEE, 2019.

Credibility

- The decision-maker’s understanding of and agreement with the model’s assumptions
- Demonstration of model validation
- Decision-maker’s ownership of and involvement with the project

For simulations and models to have an impact they need to inform the decisions of those with authority over they system. To do this the model needs to be seen as credible by decision-makers.

Step 1: Formulate the Problem

- Stating the problem of interest to the decision-maker
- Including:
  - Overall objectives for the study
  - Specific questions to be answered by the study
  - Performance measures that will be used to evaluate the system
  - Model scope
  - System configurations and time frame

Step 2: Collect Information/Data/Assumptions

- Collect information on system layout and operating procedures (and if possible, verify that the actual system matches)
- Collect data for necessary model parameters and distributions\
- Document model assumptions

Step 3: Verify the Assumptions are Valid

- Walk through the assumptions with Subject Matter Experts (SMEs), try and include multiple experts as often people are viewing the system from different perspectives
- Repeat steps 1-2 as necessary before moving on

Step 4: Program the Model

- Choose a model type and implement (relatively easy)

Step 5: Is the Programmed Model Valid?

- Compare to existing models/ system outputs
- Review the outputs with a SME to verify results are reasonable
- Perform sensitivity analysis to gauge how much the results are a function of small changes in initial conditions, system parameters, or random numbers

Step 6: Design, Conduct, and Analyze Experiments

- For each system configuration of interest, decide on tactical issues such as run length, warmup period, and the number of independent model replications
- Analyze the results and decide if additional experiments are required

Step 7: Document and Present the Results

- Documentation for the model should clearly state all assumptions, detail the computer program, and the results
- Presentation of the model should discuss the model building and validation process to promote model credibility

Examples: Epidemiology

$$\frac{dS}{dt} = -\frac{\beta I S}{N}$$
$$\frac{dE}{dt} = \frac{\beta I S}{N} - \alpha D$$
$$\frac{dI}{dt} = \alpha D - \gamma I$$
$$\frac{dR}{dt} = \gamma I$$

- Classic mathematical models such as the SIR model give us concepts such as $R_0 = \frac{\beta}{\gamma}$
- These are used to provide high level dynamics, but have not had great success in modeling actual infection levels with human responses

Examples: Epidemiology

- Brookings has developed the TRACE model (Testing Responses through Agent-based Computational Epidemiology) to answer questions about:
  - How effective can a test-and-trace policy really be in containing future waves of infection? Is such an approach feasible in the United States?
  - How much testing capacity is needed for effective containment? How much capacity to trace contacts will be needed?
  - How accurate must tests be?
  - What is the most efficient way to use limited testing capacity?
  - How might success depend on still-uncertain assumptions about the spread of the disease itself?
  - What social distancing measures might still be needed to enhance containment?

Patient Scheduling

Motivation

- Given the large capital investment and scarcity of proton therapy treatment centers, it is important that each center treats as many patients as possible while minimizing patient wait times
- Schedules need to be constructed with awareness of the limiting constraints of the facility
- Schedules need to account for uncertainty in treatment time and should account for problems faced in an outpatient setting, such as patient tardiness and absenteeism

What information do you need for modeling?

- What is the basic workflow to see a patient?
  - Check-in → Imaging → Alignment → Radiation → Rotation → …
- How long does each step take?
  - What do the distributions for each step look like? What affects these? Can steps be done in parallel? Are times correlated?
- What additional steps might be needed?
  - Immobilization, IV, Breathing Correction, Cleanup, …
- What do the treatment plans of patients look like?
  - i.e. how many angles, how much exposure, …
- What information is known and when?

Proton Therapy

- A course of treatment typically has five treatments per week for three to five weeks
- A patient is immobilized and then transported through the treatment process on a motorized patient carrier that docks with the imaging and gantry equipment
- Before each treatment, a patient is imaged to ensure precise delivery of radiation
- A patient receives radiation from multiple beam angles to mitigate any incident radiation to healthy tissue

Patient Flow Through the System

Waiting Room

- Arrive for treatment

Imaging Room

- Loaded onto transporter
- Immobilized
- Imaged by CT scanner

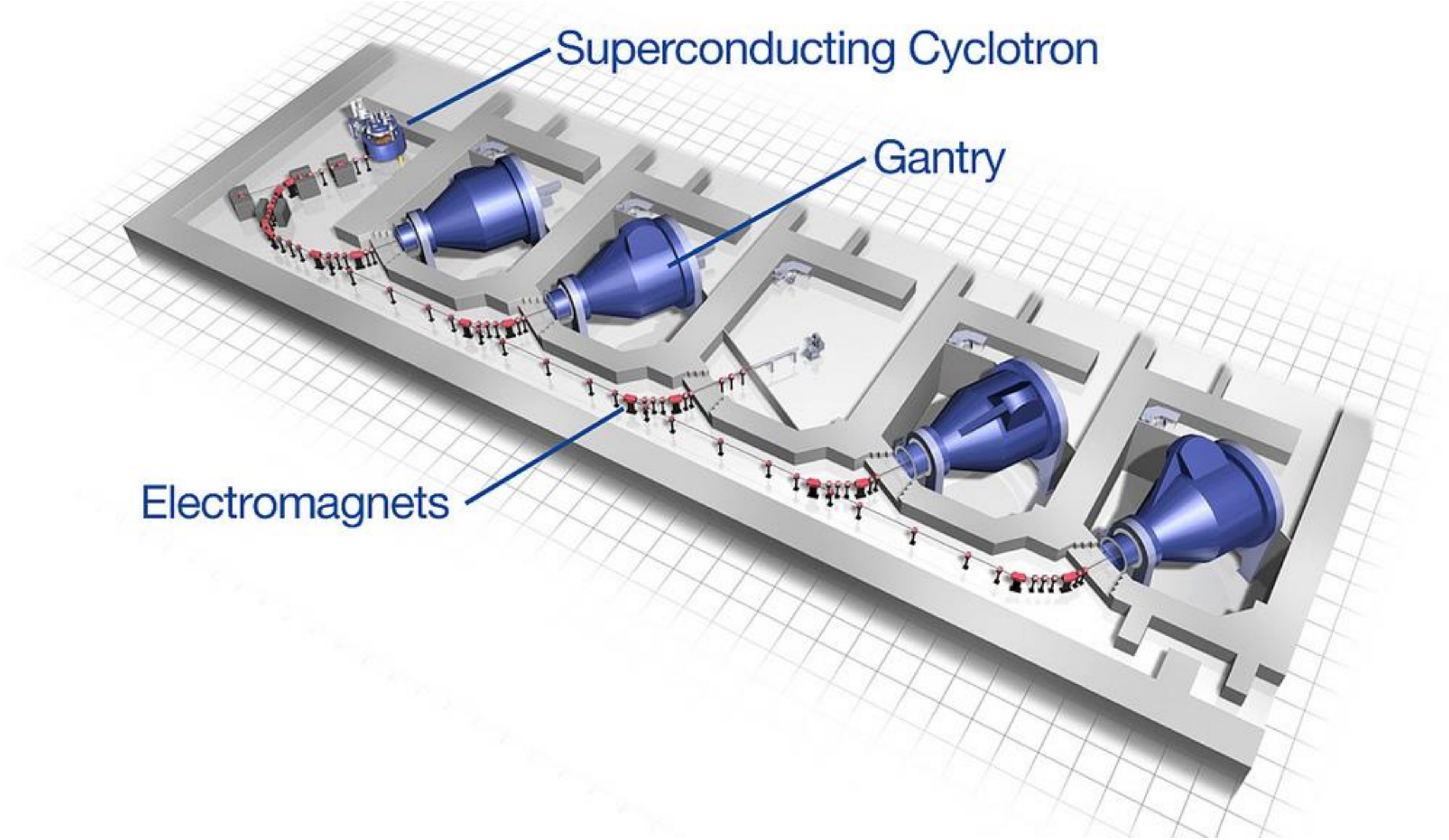
Gantry Room

- Dock transporter with gantry
- Receive treatment
- Dismount from transporter and discharged

X-ray Computed Tomography

- Scan is taken prior to every treatment
- Patient is immobilized in imaging room, then scanned
- Landmarks are placed to aid in the correct delivery of radiation

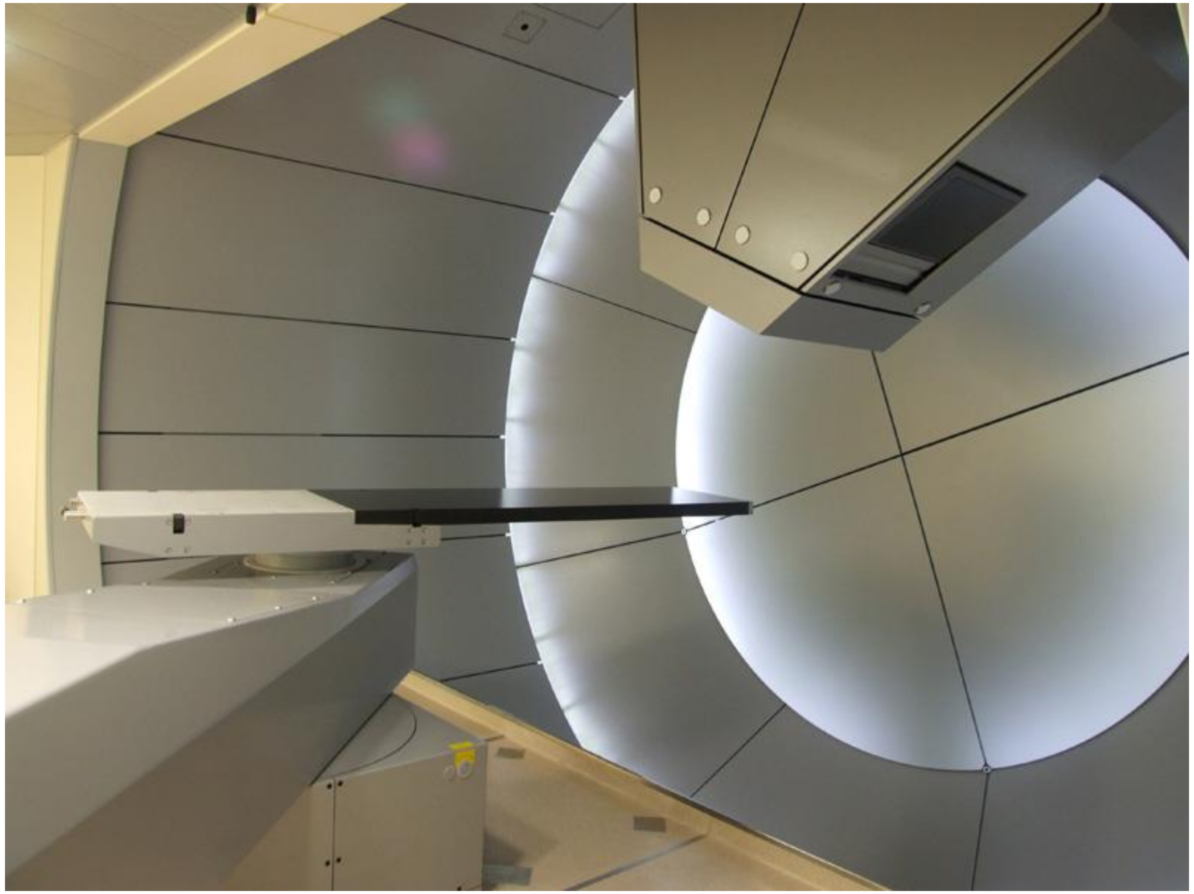
Sample Facility Layout



Cyclotron and Room Switching

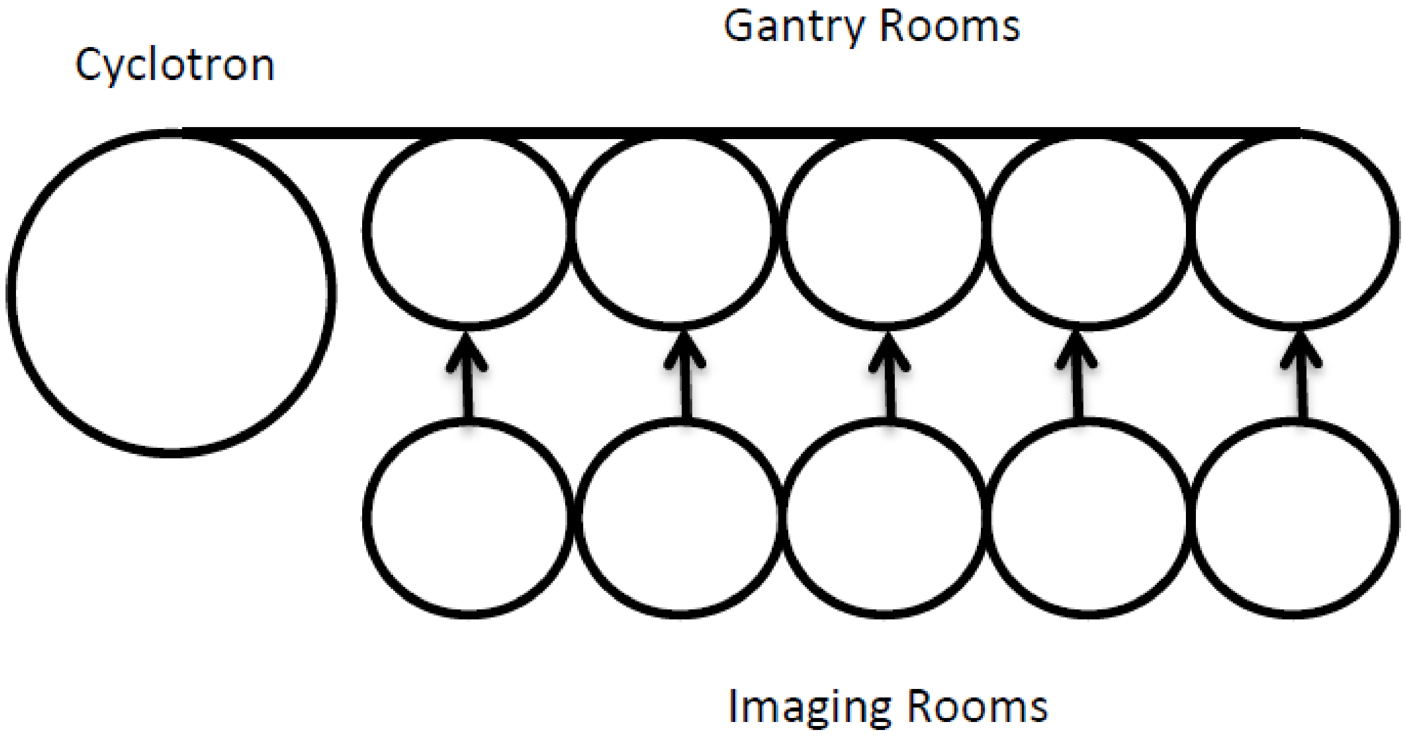
- Cyclotron is able to deliver protons to a single gantry room at any point in time
- There is a single cyclotron, making it the limiting resource of the system
- Switching delivery of protons from one gantry room to another incurs a delay of 45 seconds

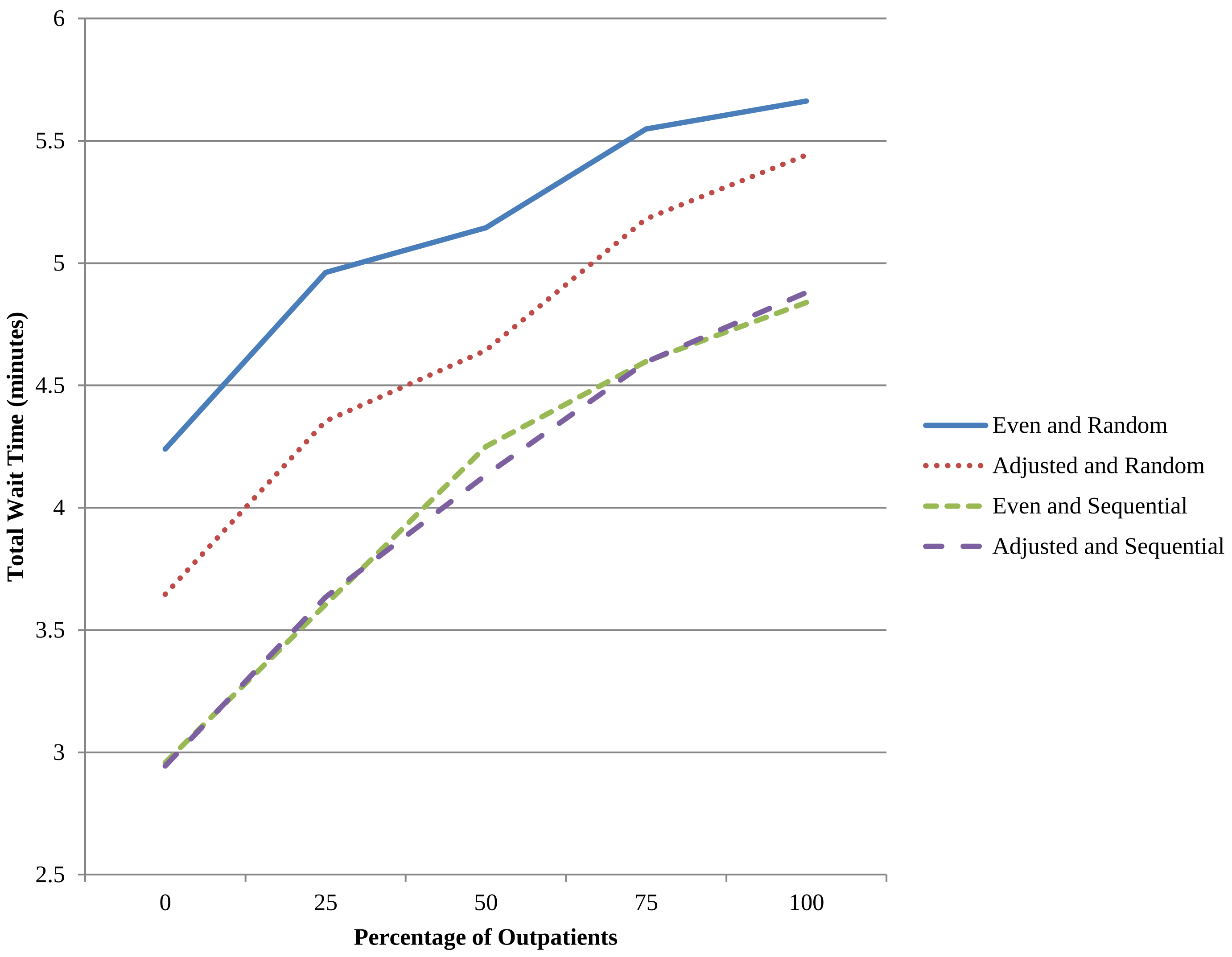
The Gantry Room



- Gantry must be rotated between each beam angle, a process that takes approximately 90 seconds

The Complete System



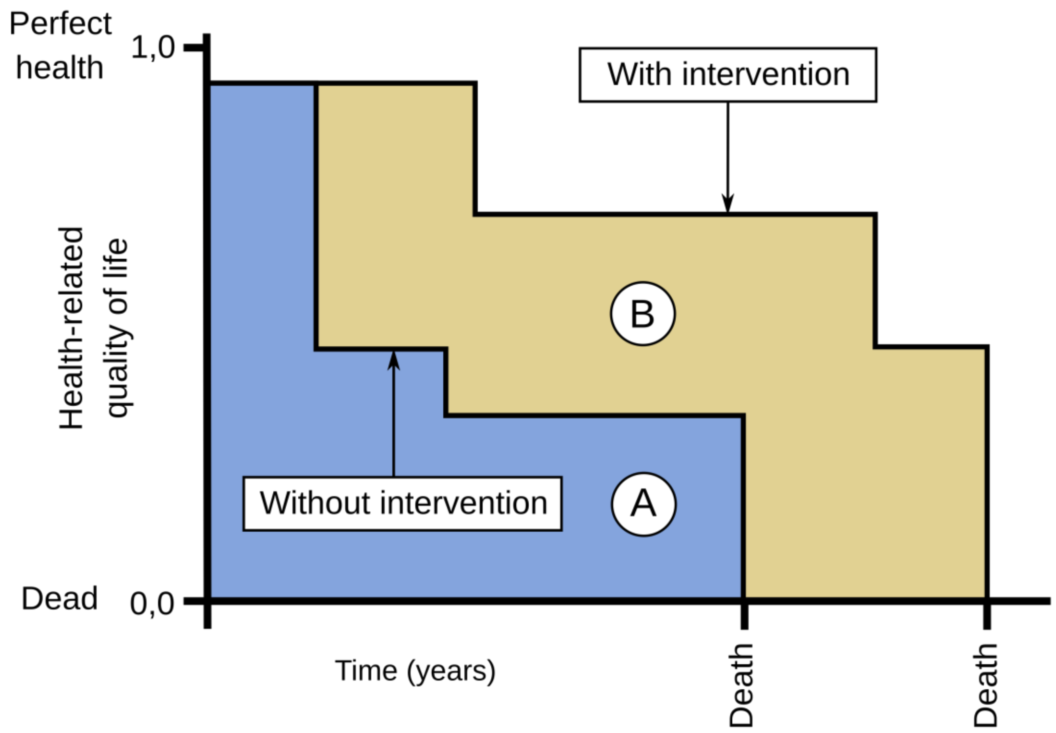


- Simulation allows for estimates of the effects of underlying system parameters, granting understanding of how the system might behave in a range of situations

Other Important Factors

- Patient Absenteeism
  - Sick patients have a harder time of making it to their appointments
  - Missed appointments lead not only to uncertainty, but potentially worse outcomes
  - Cost of rides or reminders is small compared to treatment, but not often considered within scope of treatment
- Patient Priority
  - Due to high cost, access to the treatment is still somewhat limited
  - Marginal improvement compared to next best treatment could be used in making decisions

Quality Adjusted Life Years (QALYs)



Wikimedia commons Author: Jmarchn

- Designed to capture tradeoffs between health and longevity that might occur with treatment
- Imperfect measure as health-related quality of life is highly subjective, and potentially unfairly devalues some states

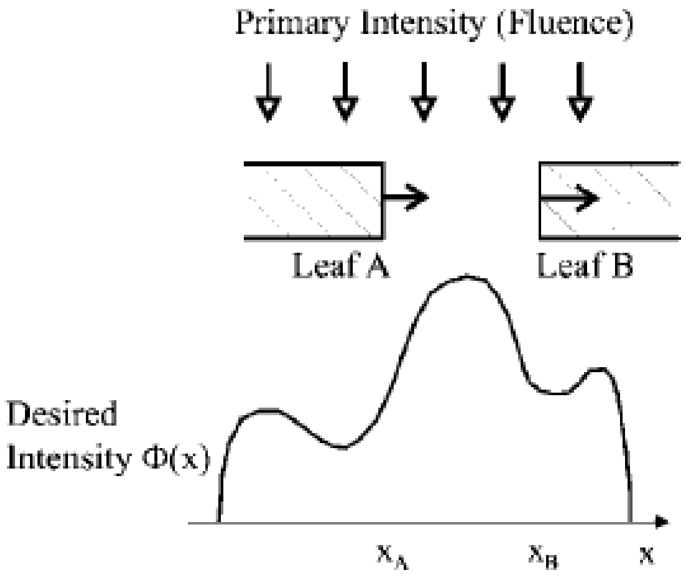
Treatment Optimization

Motivation

- Cancer is the second leading cause of death in the United States, with nearly 1 of every 4 deaths attributable to cancer (2019)
- Approximately 60% of all U.S. patients with cancer are treated with radiation therapy, most of them with external beam radiation therapy
- Intensity Modulated Radiation Therapy (IMRT) is the most common form of external beam radiation therapy

- IMRT uses a gantry arm to allow radiation to be delivered from multiple angles

- A multi-leaf collimator is adjusted for each beam angle to shape the radiation



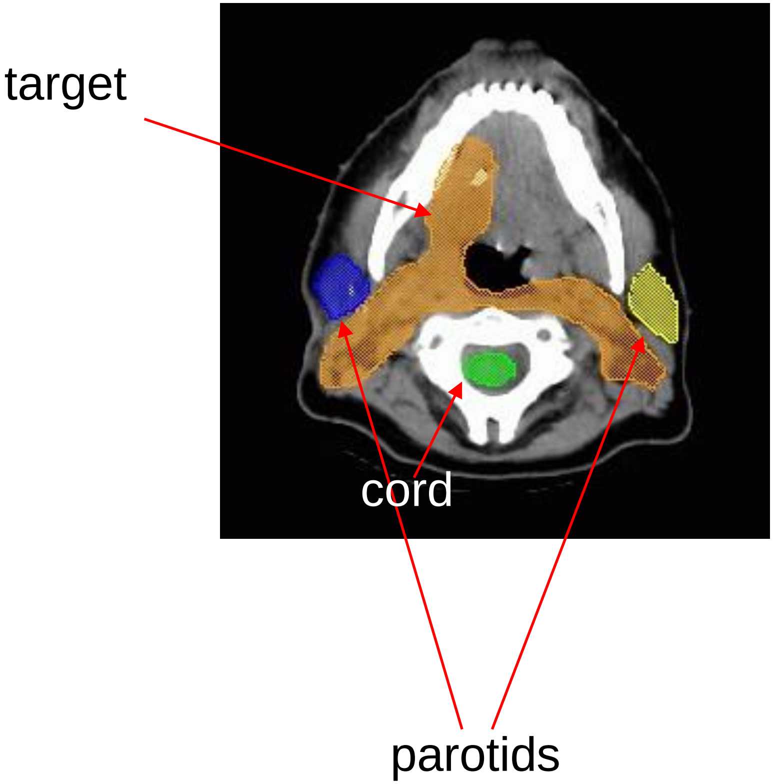
- IMRT can adjust the amount of radiation received at each pixel offering far greater control than previous treatments such as 3D conformal mapping radiation therapy (3DCRT)
- The multi-leaf collimator is dynamically adjusted controlling the intensity to each pixel

- Images from Webb 2003 ©2003 by The British Institute of Radiology

IMRT Planning Problem

1. Identify the tumor and organs at risk (OAR)
   - The objective function is defined in this step
2. Select a set of beam angles to be used for treatment plan
   - Automated set selection not currently integrated into commercial software
3. Calculate the intensity profiles for each angle
   - Currently automated by commercially available software, but can take up to 30 minutes

Step 1: Identify Tumor and OAR



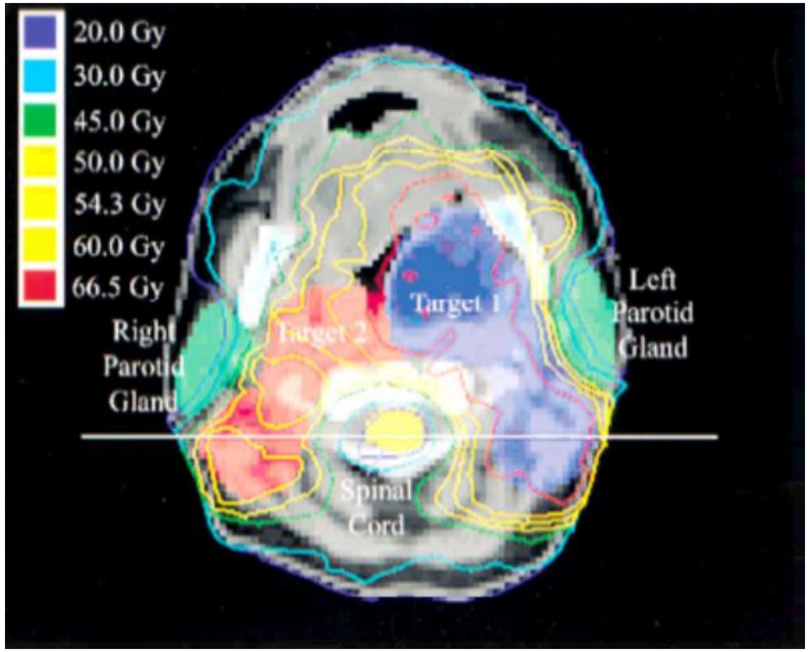
- A number of sensitive structures including the parotids (salivary glands), spinal cord, and lymph nodes
- Different tissues have different tolerance to radiation
- The tumor and immediate adjacent tissue form the planning treatment volume (PTV) for which target radiation levels are set

Penalty Score

$$S = \sum_{i \in OAR \cup PTV} \beta_i \max(A_i - d_i, 0) + \sum_{i \in PTV} \beta_i \max(d_i - A_i, 0)$$

- β_i penalty weight
- d_i desired dose level, in Greys (Gy)
- A_i actual dose level
- OAR set of organs at risk
- PTV set of tissue in planned treatment volume

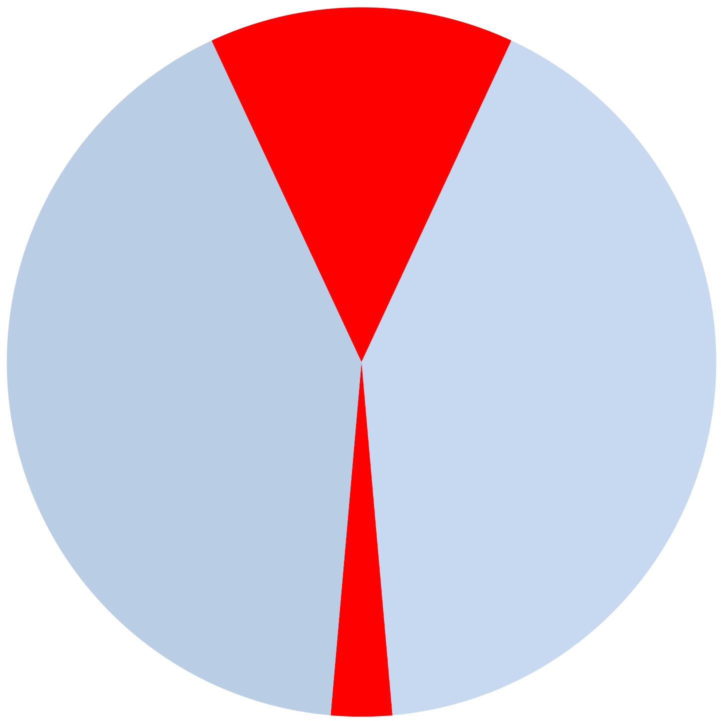
Penalty Score



Constraint	Desired level	Weight
Less than 66% of the left parotid receiving	26 Gy	**3**
Less than 33% of the left parotid receiving	32 Gy	**3**
Less than 66% of the right parotid receiving	26 Gy	**3**
Less than 33% of the right parotid receiving	32 Gy	**3**
Less than 90% of the oral mucosa receiving	30 Gy	**8**
Less than 30% of the oral mucosa receiving	40 Gy	**8**
Maximum spinal cord dose	45 Gy	**15**
Maximum brain stem dose	54 Gy	**15**
More than 95% of the low-risk PTV receiving	54 Gy	**6**
Less than 5% of the low-risk PTV receiving	59.4 Gy	**6**
More than 95% of the high-risk PTV receiving	59.4 Gy	**6**
Less than 5% of the high-risk PTV receiving	70 Gy	**6**

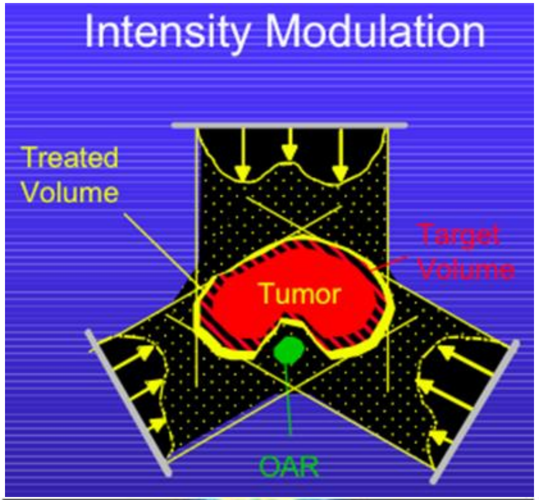
Step 2: Selecting Beam Angles

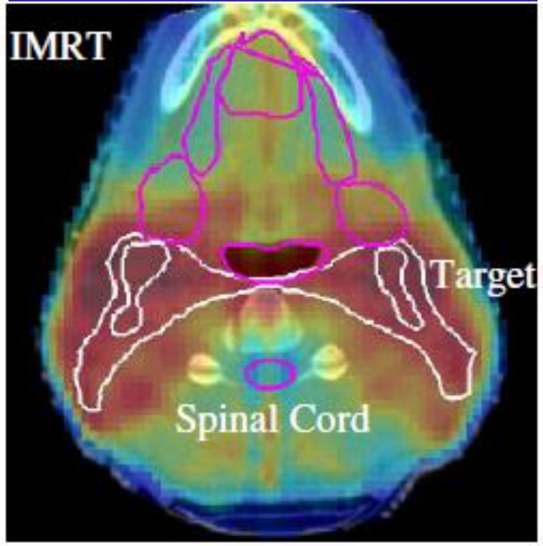
- There are diminishing returns for using more beam angles; our plans were constructed using 7 beam angles
- Plans were constructed such that beam angles were spaced at least 30 degrees apart
- Since some radiation passes through the tumor, beam angles were not placed between 170 and 190 degrees of another angle



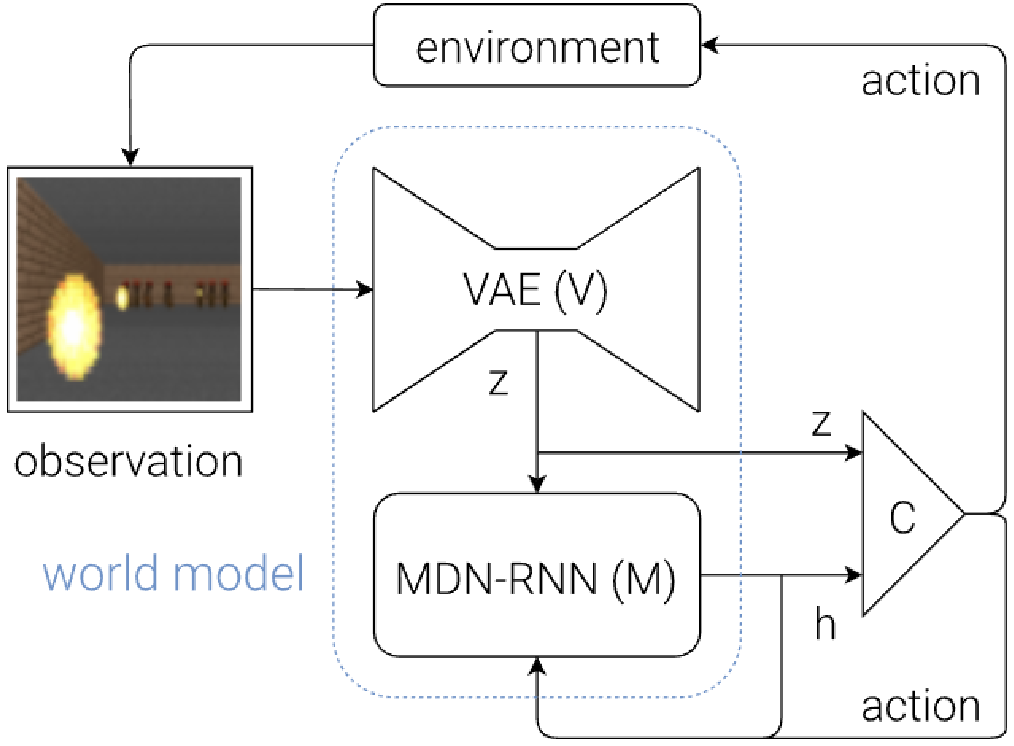
Step 3: Calculation of Intensity Profiles

- Each beam angle has about 200 pixels (also called beamlets or bixels) each of which can receive a different amount of radiation
- Beam intensities are optimized by P^3 IMRT which uses the CT scan to help ensure precise calculations
- Current software and computers take several minutes to calculate and simulate all beam intensities





Teaching Computers to make their own Models



World Models

Can agents learn inside of their own dreams?

DAVID HA
Google Brain
Tokyo, Japan

JÜRGEN SCHMIDHUBER
NNAISENSE
Swiss AI Lab, IDSIA (USI & SUPSI)

March 27
2018

NIPS 2018
Paper

YouTube
Talk

Download
PDF

Abstract

We explore building generative neural network models of popular reinforcement learning environments. Our *world model* can be trained quickly in an unsupervised manner to learn a compressed spatial and temporal representation of the environment. By using features extracted from the world model as inputs to an agent, we can train a very compact and simple policy that can solve the required task. We can even train our agent entirely inside of its own dream environment generated by its world model, and transfer this policy back into the actual environment.