

# THE CLOUD HUNTER'S PROBLEM

## AN AUTOMATED DECISION ALGORITHM TO IMPROVE THE PRODUCTIVITY OF SCIENTIFIC DATA COLLECTION IN STOCHASTIC ENVIRONMENTS

Arthur Small

based on work with JASON B. STEFIK, JOHANNES VERLINDE, NATHANIEL C.  
JOHNSON

Decision Analysis



# BEGINNINGS: PROBABILISTIC FORECASTING

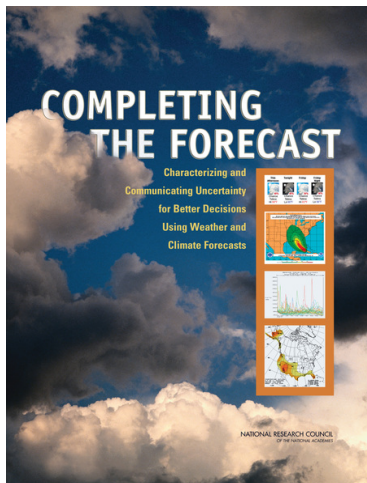




FIGURE 4.2 Headline from *The Forum* newspaper, April 24, 1997. SOURCE: Forum Communications Company.

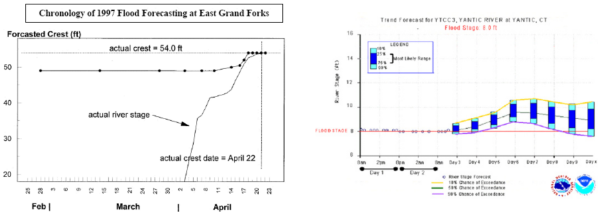


FIGURE 4.3 Left: Deterministic forecasts issued by NWS prior to the Red River flood of 1997. Right: Probabilistic river stage forecast from AHPS. SOURCE: NWS.

## FIGURE 1: Communication of forecast information during the Red River Flood of 1997 in Grand Forks, North Dakota

# THE PROMISE: DATA-DRIVEN DECISION MAKING

Data science hype: “Better information → Better decisions!”

*Claim:* That's wrong.

Data-driven decision-making also requires complementary analytics and products:

- *Prediction* : calibration of error, → *probabilistic* forecasts
- *Optimization* : application-specific models and tools, attention to user objectives
- *Visualization/communication* : intelligence *actionable* in terms of user's decision problem

# REALIZING THE PROMISE OF DATA-DRIVEN DECISION-MAKING

On prediction, data science gallops along: machine learning, classification, prediction, inference. . .

*Claim:* The promise of data science to empower data-driven decision-making lags its potential, because of insufficient focus on *complementary* analytics:

- Rigorous quantification of *uncertainty* and *error* in outputs of predictive models
- Accurate formal representation of decision problems – including decision-maker's objectives and tolerance for risk
- Appropriate use of optimization techniques
- Visualizations or other products customized to decision contexts

*Goal:* Identify good *actions*, conditional on predictions made with error

# THE CLOUD HUNTER'S PROBLEM

# MEET THE CLOUD HUNTER

The Cloud Hunter is an atmospheric scientist.

The Cloud Hunter wants to collect data from inside *liquid boundary layer clouds*.



That takes aircraft. Which are expensive.



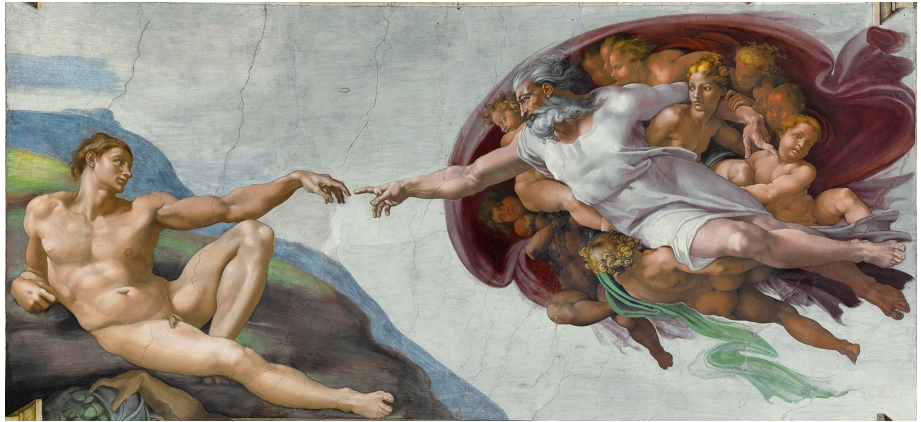
# THE CLOUD HUNTER'S DECISION PROBLEM

The Cloud Hunter's Problem concerns how to allocate a fixed budget of flight hours between dates over the course of a field season.

Fly/No-fly decisions must be made 1 day ahead, based on imperfect day-ahead forecasts of whether conditions are good or bad for collecting required data.



## PROPOSED DECISION-MAKING PROCESS:



# FORMAL MODEL OF THE CLOUD HUNTER'S DECISION PROBLEM

# MODEL SETUP: CONDITIONS FOR DATA COLLECTION OVER THE COURSE OF THE FIELD SEASON

$D$  : length of the field season in days

$d = D, \dots, 1$  : index of dates

$X_d$  : quality of conditions for data collection:

- $X_d = 1$  if conditions on date  $d$  are good, 0 otherwise
- Each  $X_d$  a binary random state variable, i.e., a Bernoulli trial

A field season is a particular realization  $x_D, \dots, x_1$  of the stochastic process  $X_D, \dots, X_1$ .

Will assume the  $X_d$  are independent and identically distributed (i.i.d.).

- Assumption not actually required, but keeps things simpler and clearer.

Vector notation:  $\mathbf{X} = \langle X_D, \dots, X_1 \rangle$  denotes the stochastic process;  
 $\mathbf{x} = \langle x_D, \dots, x_1 \rangle$  denotes a particular realization.

# MODEL SETUP: DECISIONS, RESOURCE CONSTRAINTS

$F \leq D$  : number of flights in the Cloud Hunter's budget.

$f = F, \dots, 1$  : index of flights remaining in budget

$a_d$  : binary control variables ("actions")

- $a_d = 1$  iff they opt to fly on date  $d$ , 0 otherwise

$\mathbf{a} = \langle a_D, \dots, a_1 \rangle$  : actions chosen on dates  $d = D, \dots, 1$

Resource constraint:  $\sum_d a_d \leq F$ .

## PAYOFFS AND OBJECTIVES

*Payoffs:* For a given sequence of choices  $\mathbf{a}$  and realizations  $\mathbf{x}$ , the realized amount of data collected  $U$  is given by

$$U = \mathbf{a} \cdot \mathbf{x} = \sum_d \mathbf{a}_d \mathbf{x}_d$$

Assume: Flights left over at the end of the season have no residual value.

*Decision-maker's objective:* Choose a fly/no-fly decision rule to maximize data collected in expectation, subject to the resource constraint on total allowable flights:

Choose  $\mathbf{a}$  to  $\max_{\mathbf{a}} E[\mathbf{a} \cdot \mathbf{X}]$ , subject to  $\sum_d a_d \leq F$ .

**Important:** This is a *substantive assumption* about the decision-maker's goals.

- Models a decision-maker with a high tolerance for *risk*.

# FORECASTS

Decision taken on basis of a day-ahead forecast.

Before taking each decision, decision-maker receives a forecast signal  $s_d \in \mathbb{S}$ .

Calibration: map this signal to a probability of good conditions:

$$p(s) = \Pr\{X_d = 1 | s_d = s\}$$

.

(Will assume stationarity.)



# DISTRIBUTION OF FORECAST SIGNALS

More than one day ahead, don't know which forecast signals  $s \in \mathbb{S}$  will be received.

But, *do* know the the likelihood of receiving different signals.

$\pi(s)$  : probability that forecasting system will generate signal  $s$ .

$\pi(\cdot)$  defines a probability distribution over the set  $\mathbb{S}$  of possible forecast signals.

# THE TASK OF THE DECISION ANALYST

Given this set-up, the job of the decision analyst is to devise an *optimal decision rule*  $a^* = a(d, f|p)$  that delivers a recommended action—fly or no-fly—as a function of

- $d$  the number of days left in the field season,
- $f$  the number of flights left in the budget, and
- $p(s)$  the forecast probability that a flight today would be successful.

## DERIVING A SOLUTION

# INTERTEMPORAL OPTIMIZATION VIA DYNAMIC PROGRAMMING

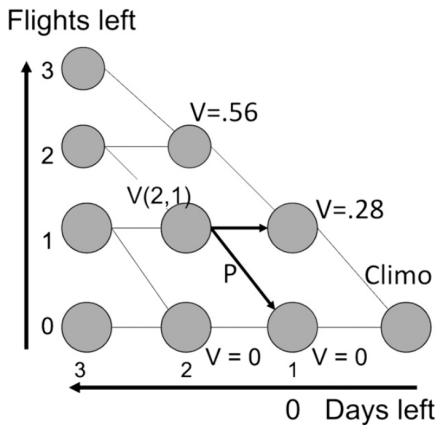


FIGURE 2: Graphical representation of the decision algorithm.

# PROBABILISTIC FORECASTING OF FAVORABLE CONDITIONS USING SELF-ORGANIZING MAPS

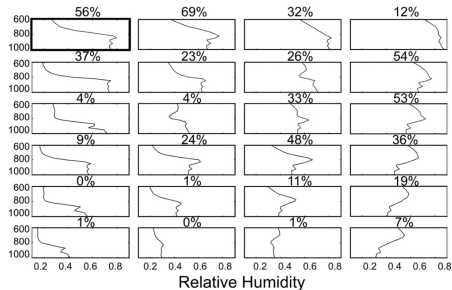


FIGURE 3: 6 X 4 SOM grid for relative humidity profiles

13.3%	20.6%	4.4%	3.7%
1.5%	8.1%	1.5%	4.4%
0%	3.7%	4.4%	11.8%
2.9%	3.7%	7.4%	6.6%
0%	0%	0.7%	1.5%
0%	0%	0%	0%

FIGURE 4: Conditional probability distribution of SOM state realizations following a forecast of SOM state 1

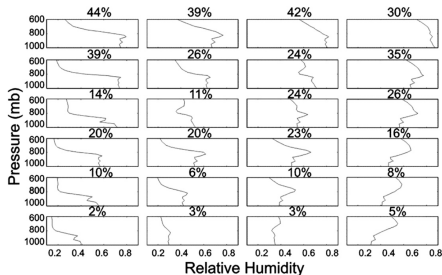


FIGURE 5: Estimated probabilities of good conditions for data collection, as a function of day-ahead SOM forecast



# RESULTS

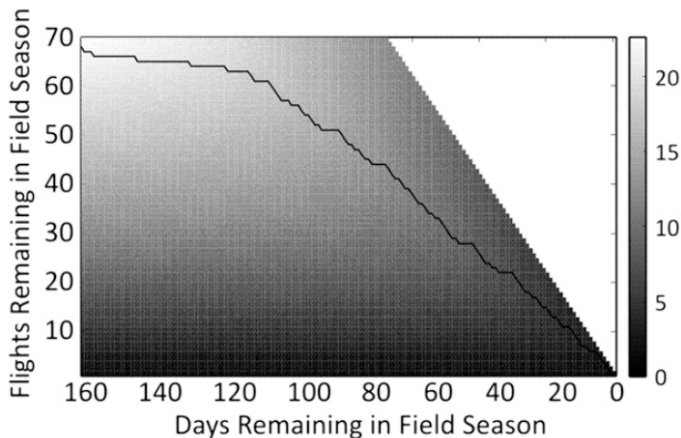


FIGURE 6: Computed values for the value function  $V(d, f)$

Dark line : simulated sequence of flight decisions

- Diagonal movements = fly dates, horizontal movements = no-fly dates

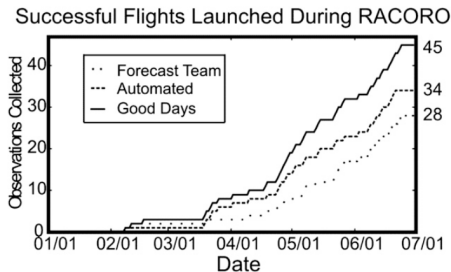


FIGURE 7: Results: the algorithm's simulated performance during 2009 field season, compared with realized performance of heuristic decision procedure.

The algorithm achieves a 21% increase in the number of successful flights.

	Heuristic procedure	Automated algorithm
Flights launched	56	66
Successes	28	34
Type I errors	28	32
Type II errors	17	11

FIGURE 8: Summary of outcomes

Successes are flights launched on days with good conditions.

Type I errors are decisions to fly only to find no clouds.

Type II errors are decisions to stand down only to find that the desired conditions existed.