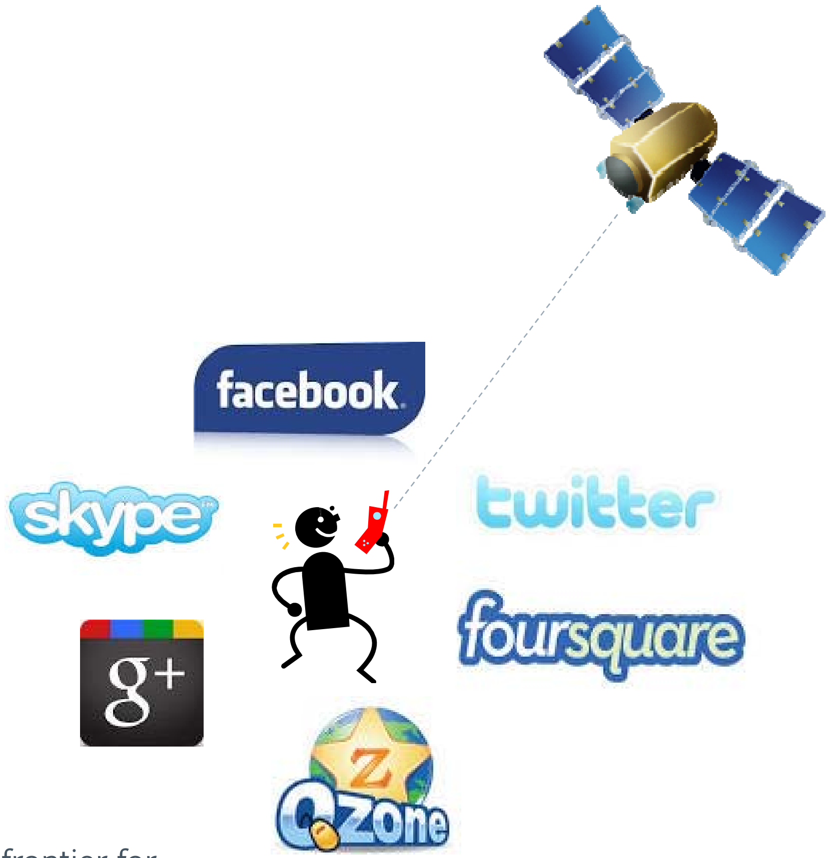
# “Veracity”
# Managing Uncertain Data

Skript zur Vorlesung Datenbanksystem II
Dr. Andreas Züfle

# Geo-Spatial Data

- Huge flood of geo-spatial data
  - Modern technology
  - New user mentality

- Great research potential
  - New applications
  - Innovative research
  - Economic Boost
    - "$600 billion potential annual consumer surplus from using personal location data" [1]



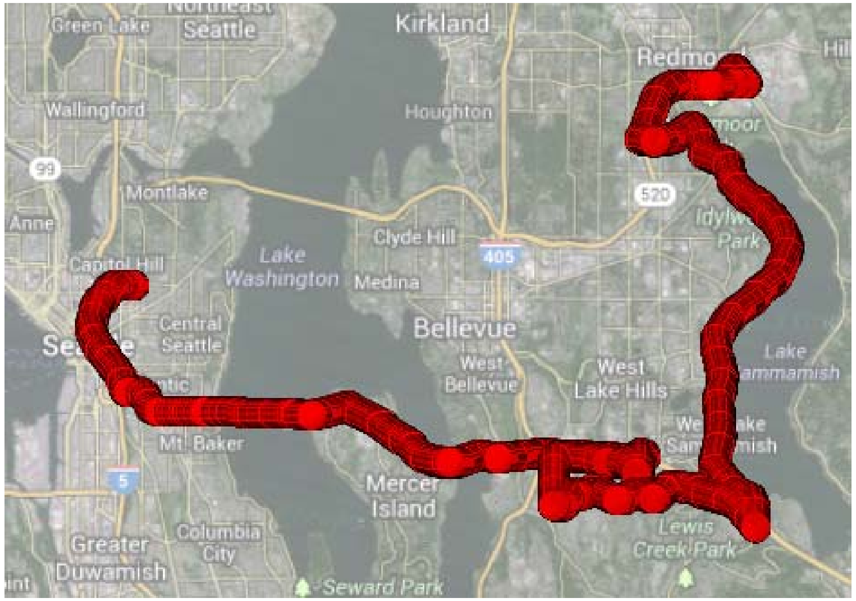

[1] McKinsey Global Institute. Big data: The next frontier for innovation, competition, and productivity. June 2011.

# Spatio-Temporal Data

- (object, location, time) triples

- Queries:
  - “Find friends that attended the same concert last saturday”

- Best case: Continuous function
  $time \rightarrow space$

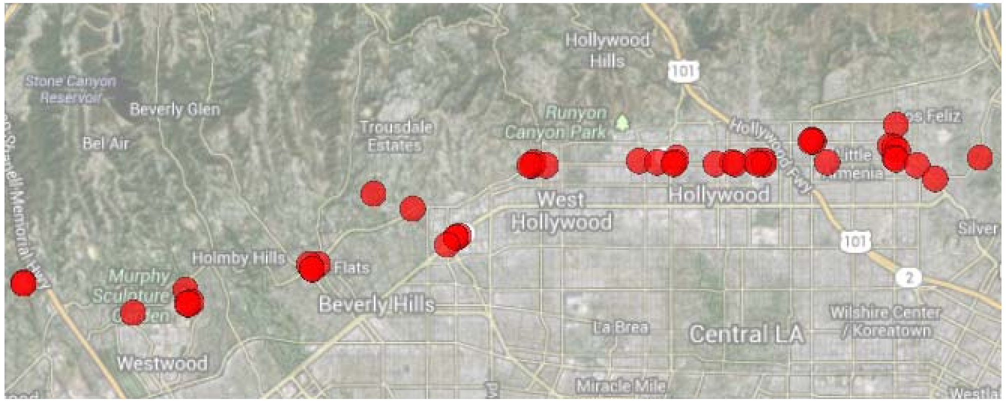



GPS log taken from a thirty minute drive through Seattle

Dataset provided by: P. Newson and J. Krumm. Hidden Markov Map Matching Through Noise and Sparseness. ACMGIS 2009.

# Sources of Uncertainty

- Missing Observations
  - Missing GPS signal
  - RFID sensors available in discrete locations only
  - Wireless sensor nodes sending infrequently to preserve energy
  - Infrequent check-ins of users of geo-social networks

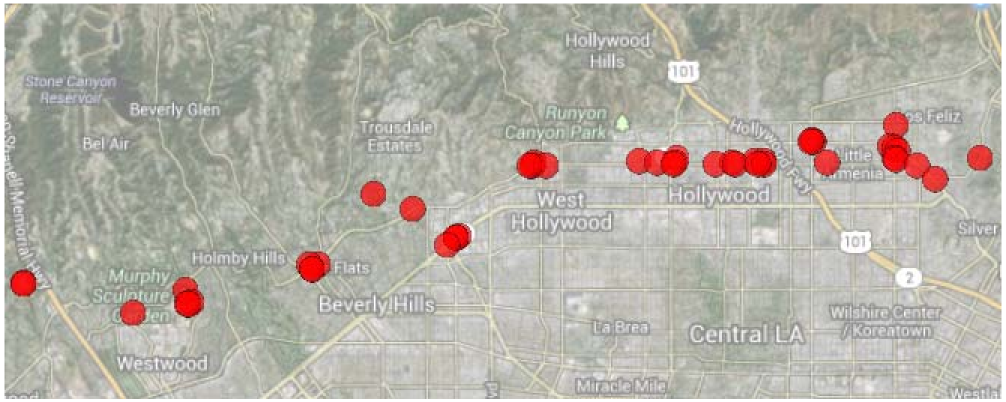



Dataset provided by: E. Cho, S. A. Myers and J. Leskovek. Friendship and Mobility: User Movement in Location-Based Social Networks. SIGKDD 2011.

# Sources of Uncertainty

- Uncertain Observations
  - Imprecise sensor measurements (e.g. radio triangulation, Wi-Fi positioning)
  - Inconsistent information (e.g. contradictive sensor data)
  - Human errors (e.g. in crowd-sourcing applications)

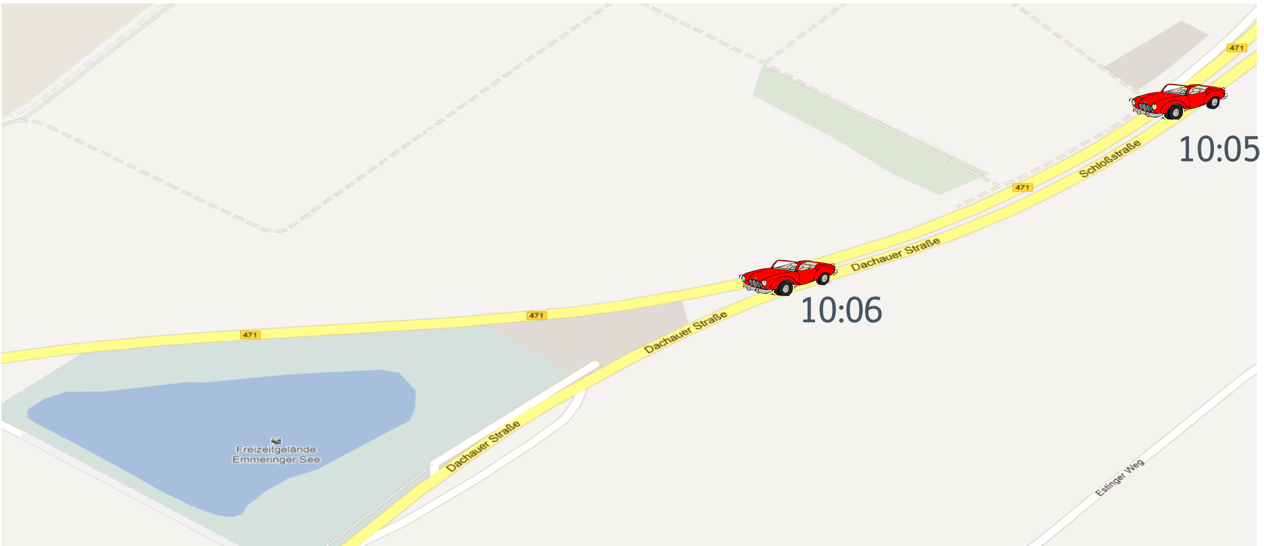
➢ From database perspective, the position of a mobile object is uncertain




Dataset provided by: E. Cho, S. A. Myers and J. Leskovek. Friendship and Mobility: User Movement in Location-Based Social Networks. SIGKDD 2011.

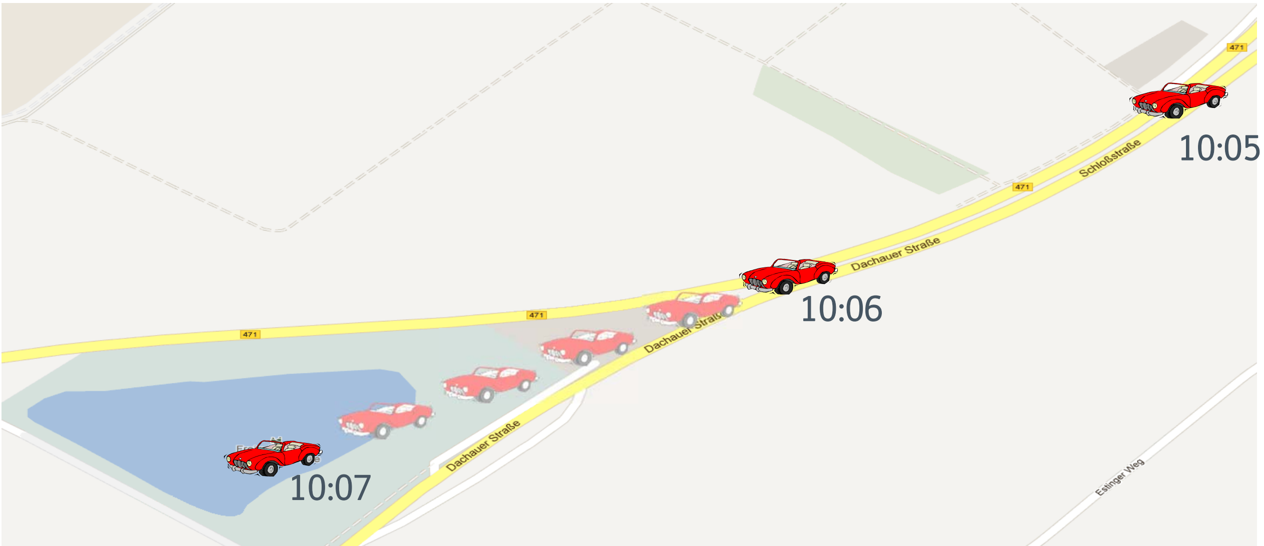
# Uncertainty in Spatial Data

- At time 10:07: Where is an object having past observations at times 10:05am and 10:06am?



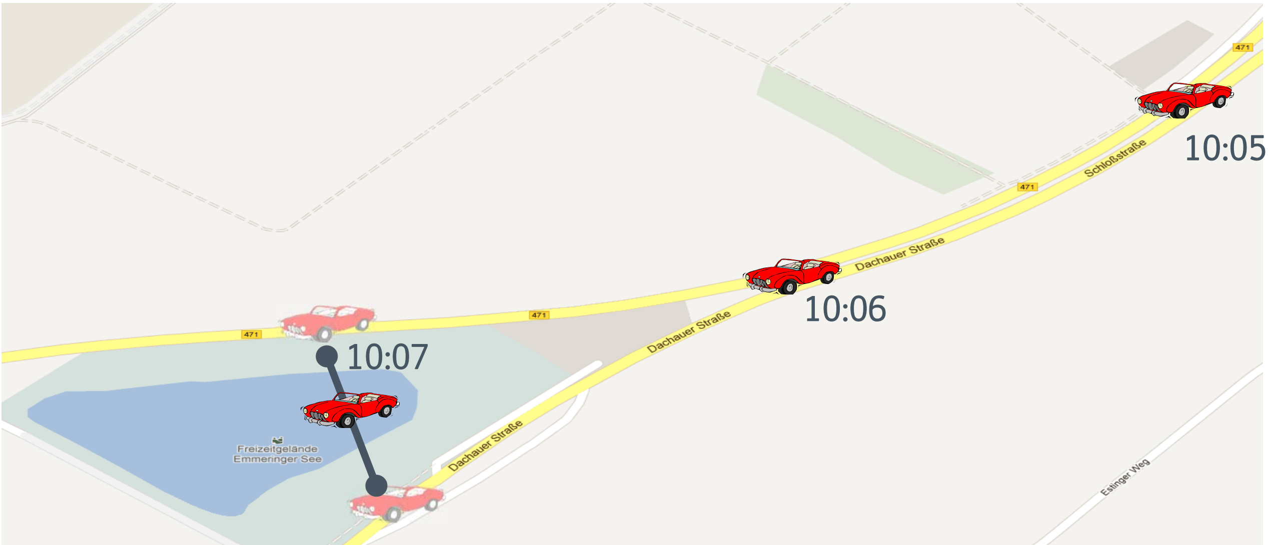
# Previous Solution: Extrapolation

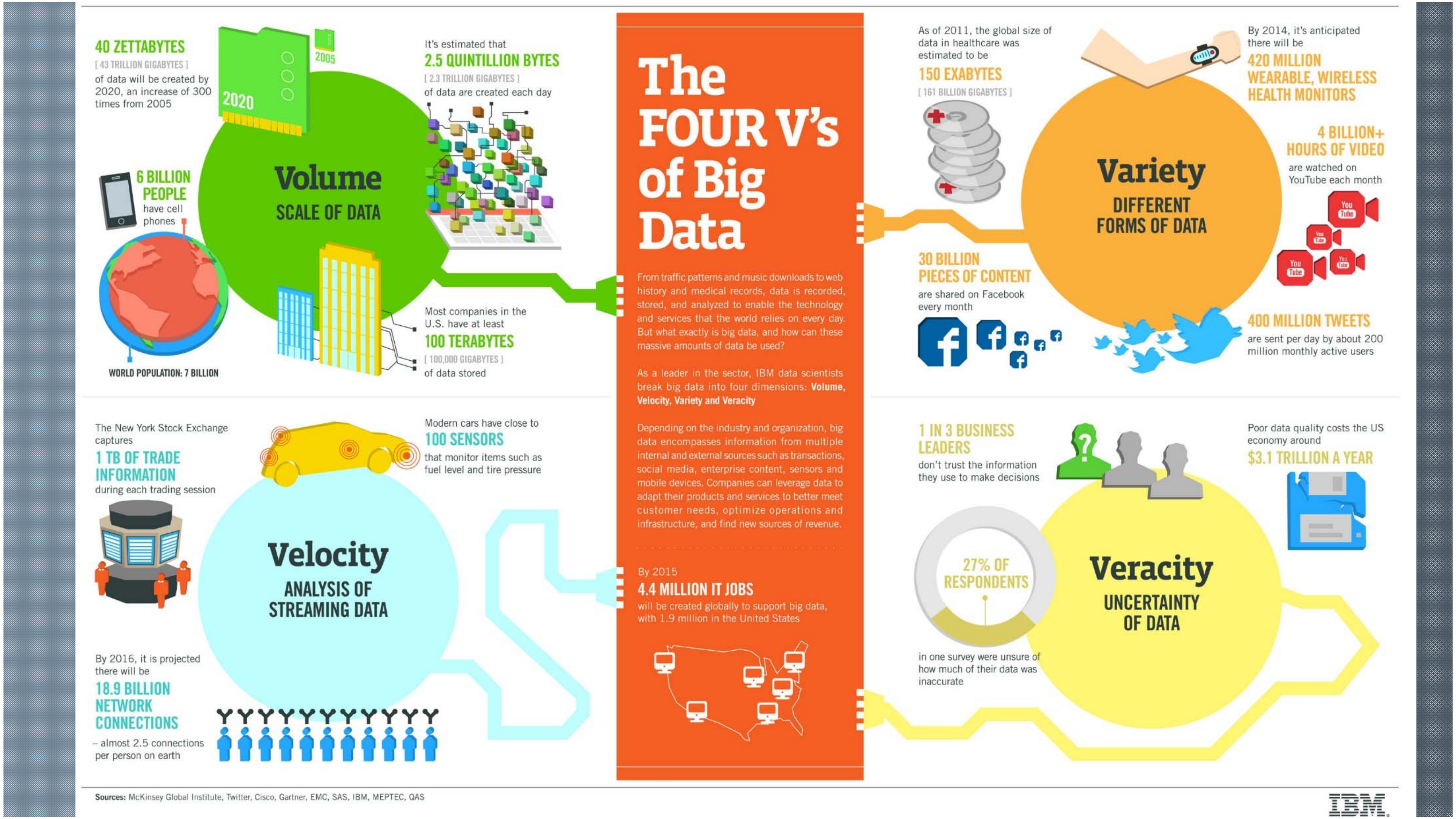
- Unknown positions are estimated using past observations
- No semantic information (road network, driver behaviour etc.)



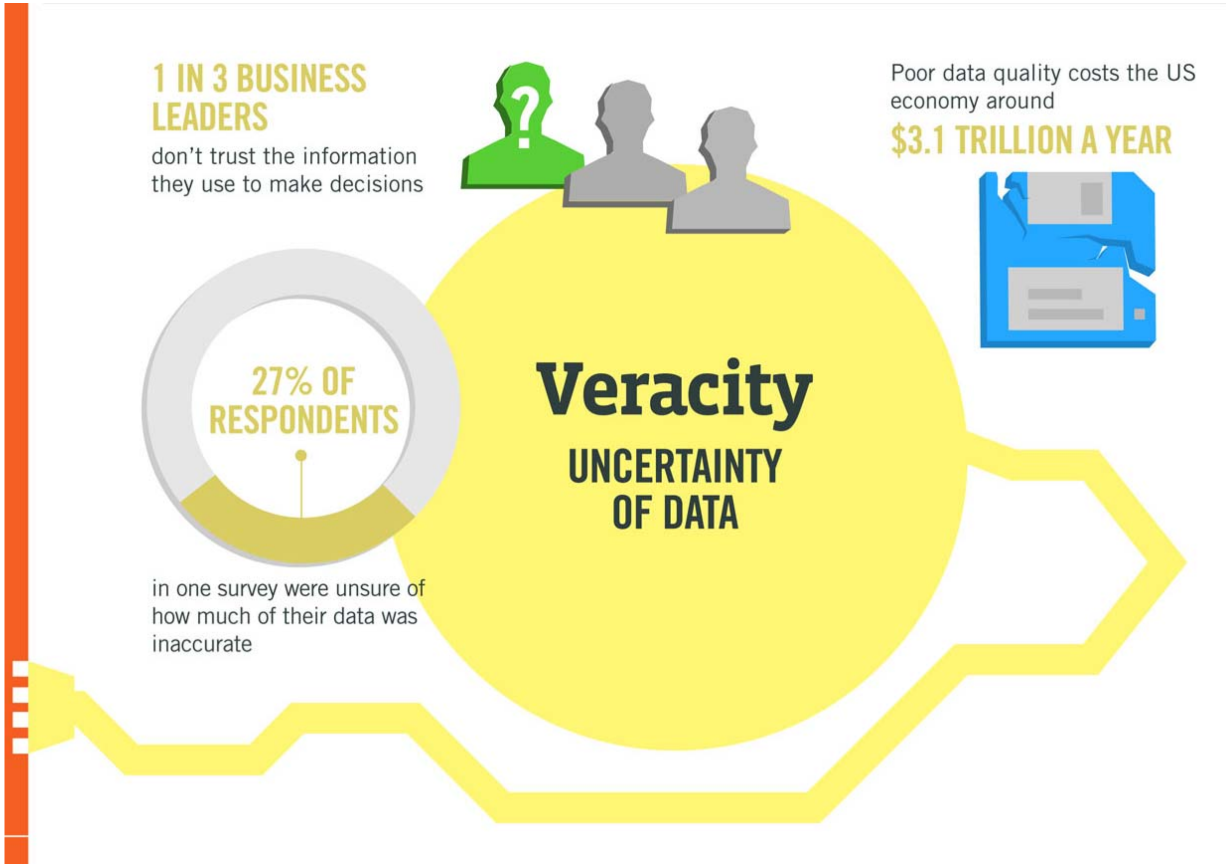
# Previous Solution: Aggregation

- Exploit semantic knowledge to obtain possible positions of an object
- Aggregate possible positions (expected position, most-likely position)





The FOUR V's of Big Data
From traffic patterns and music downloads to web history and medical records, data is recorded, stored, and analyzed to enable the technology and services that the world relies on every day. But what exactly is big data, and how can these massive amounts of data be used?
As a leader in the sector, IBM data scientists break big data into four dimensions: Volume, Velocity, Variety and Veracity
Depending on the industry and organization, big data encompasses information from multiple internal and external sources such as transactions, social media, enterprise content, sensors and mobile devices. Companies can leverage data to adapt their products and services to better meet customer needs, optimize operations and infrastructure, and find new sources of revenue.
By 2015
4.4 MILLION IT JOBS
will be created globally to support big data, with 1.9 million in the United States
Volume
SCALE OF DATA
40 ZETTABYTES
[ 43 TRILLION GIGABYTES ]
of data will be created by 2020, an increase of 300 times from 2005
2020
2005
It's estimated that
2.5 QUINTILLION BYTES
[ 2.3 TRILLION GIGABYTES ]
of data are created each day
6 BILLION PEOPLE
have cell phones
WORLD POPULATION: 7 BILLION
Most companies in the U.S. have at least
100 TERABYTES
[ 100,000 GIGABYTES ]
of data stored
Velocity
ANALYSIS OF STREAMING DATA
The New York Stock Exchange captures
1 TB OF TRADE INFORMATION
during each trading session
Modern cars have close to
100 SENSORS
that monitor items such as fuel level and tire pressure
By 2016, it is projected there will be
18.9 BILLION NETWORK CONNECTIONS
– almost 2.5 connections per person on earth
Variety
DIFFERENT FORMS OF DATA
As of 2011, the global size of data in healthcare was estimated to be
150 EXABYTES
[ 161 BILLION GIGABYTES ]
By 2014, it's anticipated there will be
420 MILLION WEARABLE, WIRELESS HEALTH MONITORS
4 BILLION+ HOURS OF VIDEO
are watched on YouTube each month
30 BILLION PIECES OF CONTENT
are shared on Facebook every month
400 MILLION TWEETS
are sent per day by about 200 million monthly active users
Veracity
UNCERTAINTY OF DATA
1 IN 3 BUSINESS LEADERS
don't trust the information they use to make decisions
Poor data quality costs the US economy around
$3.1 TRILLION A YEAR
27% OF RESPONDENTS
in one survey were unsure of how much of their data was inaccurate
Sources: McKinsey Global Institute, Twitter, Cisco, Gartner, EMC, SAS, IBM, MEPTEC, QAS


1 IN 3 BUSINESS LEADERS
don't trust the information they use to make decisions
27% OF RESPONDENTS
in one survey were unsure of how much of their data was inaccurate
Veracity
UNCERTAINTY OF DATA
Poor data quality costs the US economy around
$3.1 TRILLION A YEAR

# Research Challenge

Include the uncertainty directly in the querying and mining process.

# Research Challenge

Include the uncertainty directly in the querying and mining process.

↓

Assess the reliability of similarity search and data mining results

# Research Challenge

Include the uncertainty directly in the querying and mining process.

↓

Assess the reliability of similarity search and data mining results

↓

Enhance the underlying decision-making process.

# Overview

1. Introduction to Probability Theory
2. Case Study: Probabilistic Count Queries

# Overview

1. Introduction to Probability Theory
2. Case Study: Probabilistic Count Queries

# Probability Theory: Random Variables

A random variable $X$ is a variable whose value is subject to variations due to chance.

The set of possible outcomes of $X$ is denoted as $\Omega$.

# Probability Theory: Random Variables

A random variable $X$ is a variable whose value is subject to variations due to chance.

The set of possible outcomes of $X$ is denoted as $\Omega$.

Example 1: Coin toss

$\Omega = \{$  ,  $\}$

# Probability Theory: Random Variables

A random variable $X$ is a variable whose value is subject to variations due to chance.

The set of possible outcomes of $X$ is denoted as $\Omega$.

Example 1: Coin toss

$\Omega = \{$ ,  $\}$

Example 2: Dice throw

$$\Omega = \{1,2,3,4,5,6\}$$

# Probability Theory: Random Events

Any $\omega \subseteq \Omega$ is called a random event.

# Probability Theory: Random Events

Any $\omega \subseteq \Omega$ is called a random event.

Example 3: Dice throw $\Omega = \{1,2,3,4,5,6\}$
Event A := “An even number is thrown” = $\{2,4,6\} \subseteq \Omega$

# Probability Theory: Random Events

Any $\omega \subseteq \Omega$ is called a random event.

Example 3: Dice throw $\Omega = \{1,2,3,4,5,6\}$
Event A := "An even number is thrown" = $\{2,4,6\} \subseteq \Omega$

Example 4: Throw of two dice. $\Omega = \{1,2,3,4,5,6\}^2 = \{(1,1),(1,2),\ldots,(6,6)\}$
Event B := "The sum of points thrown equals 4" = $\{(1,3),(2,2),(3,1)\} \subseteq \Omega$

# Probability Theory: Random Events

Any $\omega \subseteq \Omega$ is called a random event.

Example 3: Dice throw $\Omega = \{1,2,3,4,5,6\}$
Event A := "An even number is thrown" = $\{2,4,6\} \subseteq \Omega$

Example 4: Throw of two dice. $\Omega = \{1,2,3,4,5,6\}^2 = \{(1,1),(1,2),\ldots,(6,6)\}$
Event B := "The sum of points thrown equals 4" = $\{(1,3),(2,2),(3,1)\} \subseteq \Omega$

Let $X$ be a random variable and let $\omega$ be a random event. Then $P(X = \omega)$ denotes the probability that random variable $X$ takes a value in $\omega$.

# Probability Theory: Probability Mass Function

Let $\Omega$ be finite or countably infinite.
A function

$$p: \Omega \rightarrow [0,1]$$

such that

$$\sum_{\omega \in \Omega} p(w) = 1$$

is called probability mass function (pmf).

# Probability Theory: Probability Mass Function

Let $\Omega$ be finite or countably infinite.
A function

$$p: \Omega \rightarrow [0,1]$$

such that

$$\sum_{\omega \in \Omega} p(w) = 1$$

is called probability mass function (pmf).

A pmf $p_X$ is called pmf of a random variable X if for any $\omega \in \Omega$:

$$P(X = w) = p_X(w)$$

# Probability Theory: Probability Mass Function

Let $\Omega$ be finite or countably infinite.
A function

$$p: \Omega \to [0,1]$$

such that

$$\sum_{\omega \in \Omega} p(w) = 1$$

is called probability mass function (pmf).

A pmf $p_X$ is called pmf of a random variable X if for any $\omega \in \Omega$:

$$P(X = w) = p_X(w)$$

Example 5: Dice throw $\Omega = \{1,2,3,4,5,6\}$

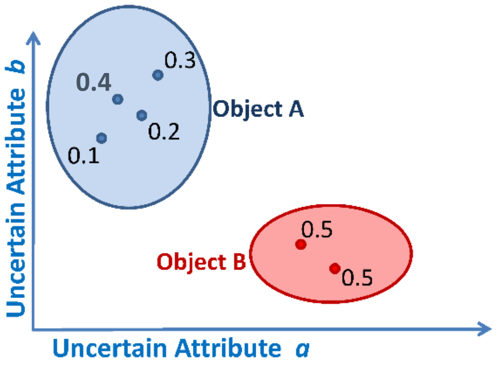
$$P(X = 1) = p_X(1) = \frac{1}{6}$$

# Uncertain Data

- In an uncertain database $DB = \{o_1, \ldots, o_N\}$, each object $o \in DB$ is a random variable.

# Uncertain Data

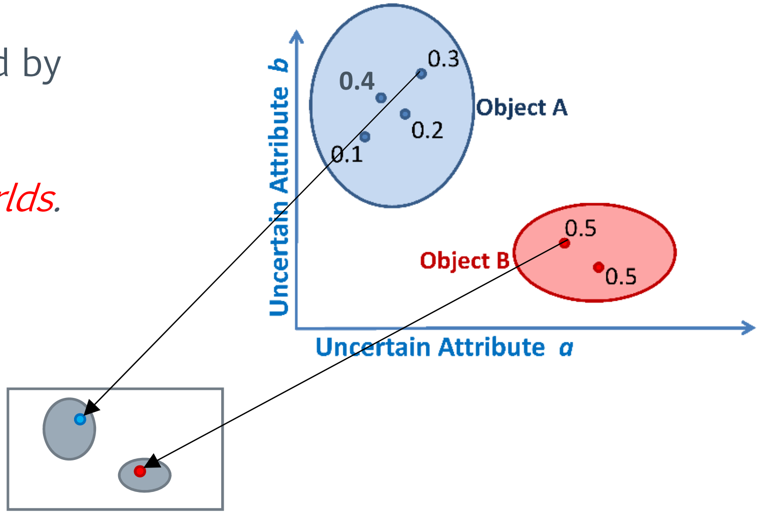
- In an uncertain database $DB = \{o_1, \ldots, o_N\}$, each object $o \in DB$ is a random variable.



# Possible World Semantics

The sample space $\Omega_{DB}$ is defined by $\Omega_1 \times \cdots \times \Omega_N$

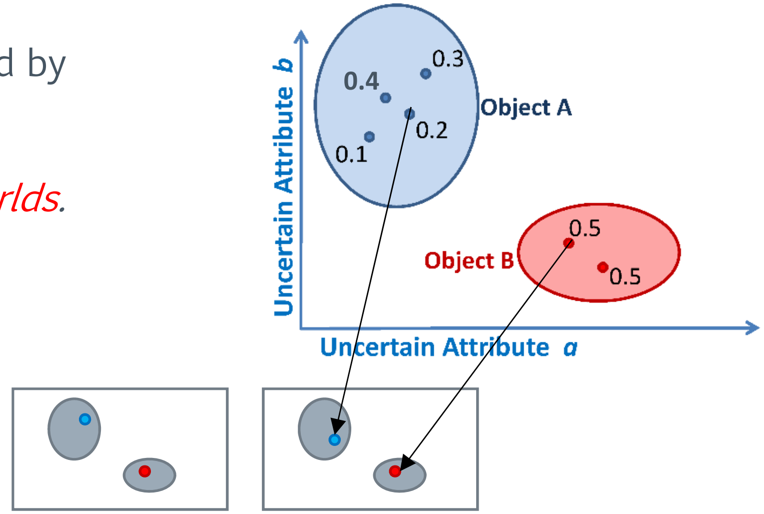
Samples are called *Possible Worlds*.



# Possible World Semantics

The sample space $\Omega_{DB}$ is defined by $\Omega_1 \times \dots \times \Omega_N$

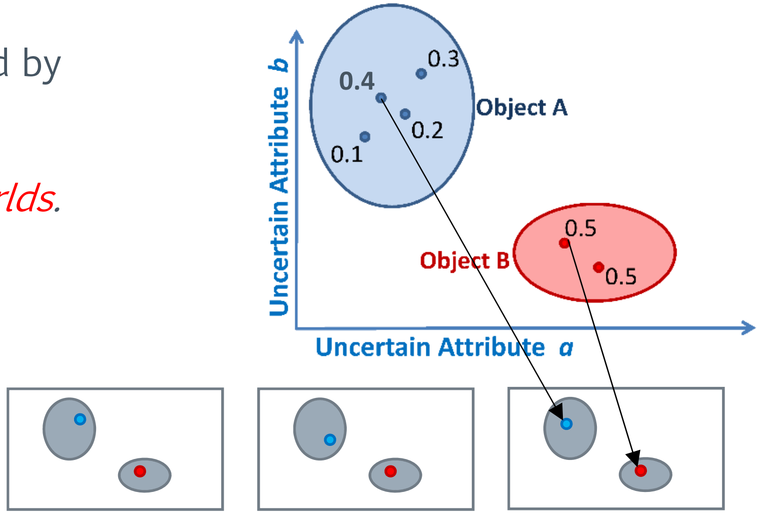
Samples are called *Possible Worlds*.



# Possible World Semantics

The sample space $\Omega_{DB}$ is defined by $\Omega_1 \times \cdots \times \Omega_N$

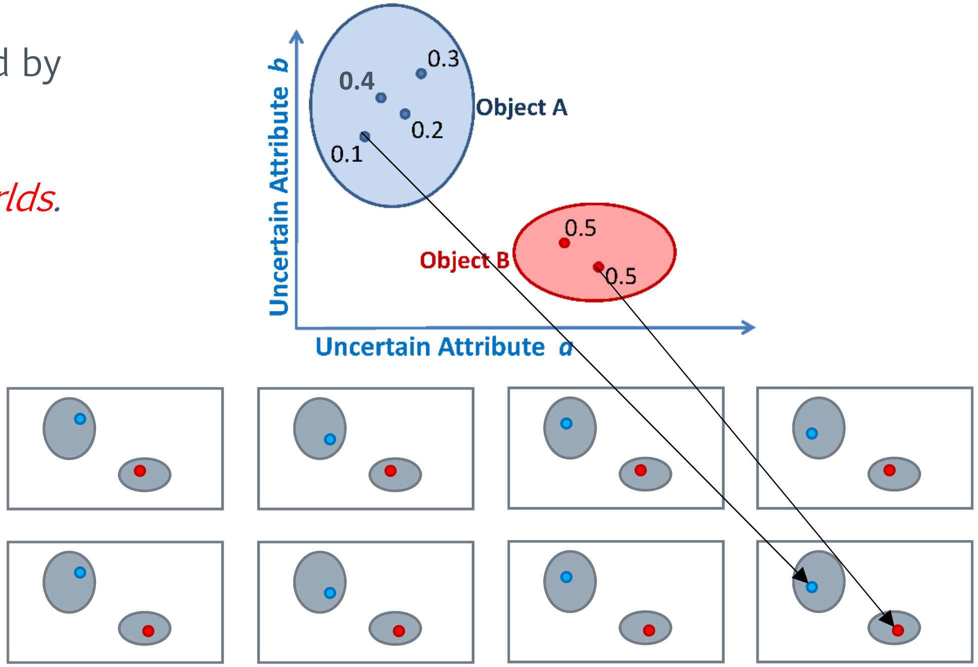
Samples are called *Possible Worlds*.



# Possible World Semantics

The sample space $\Omega_{DB}$ is defined by $\Omega_1 \times \cdots \times \Omega_N$

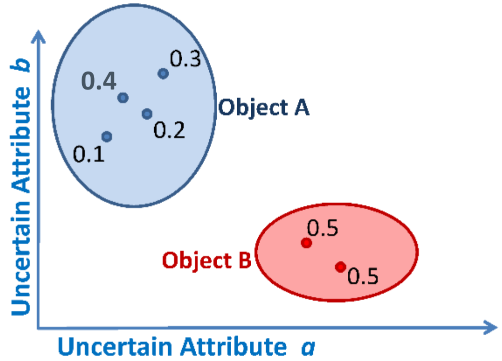
Samples are called *Possible Worlds*.



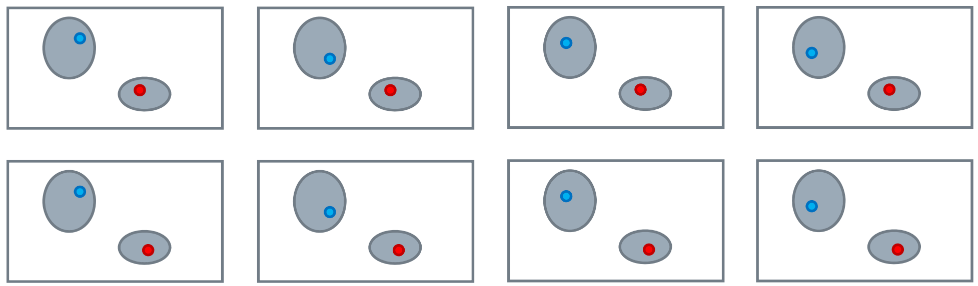
# Possible World Semantics

The sample space $\Omega_{DB}$ is defined by $\Omega_1 \times \cdots \times \Omega_N$

Samples are called *Possible Worlds*.



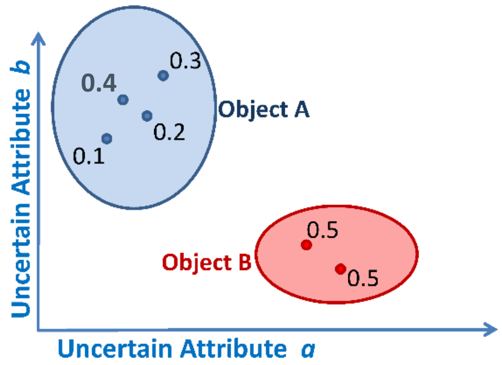

$$\Omega_{DB} = \{ \quad \}$$



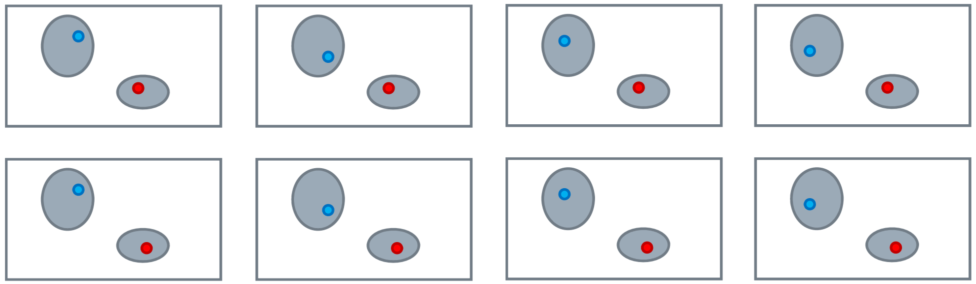
# Possible World Semantics

The sample space $\Omega_{DB}$ is defined by $\Omega_1 \times \cdots \times \Omega_N$

Samples are called *Possible Worlds*.




$\Omega_{DB} =$ {



}

Assumption: $p_{DB}: \Omega \rightarrow [0,1]$ can be computed efficiently.

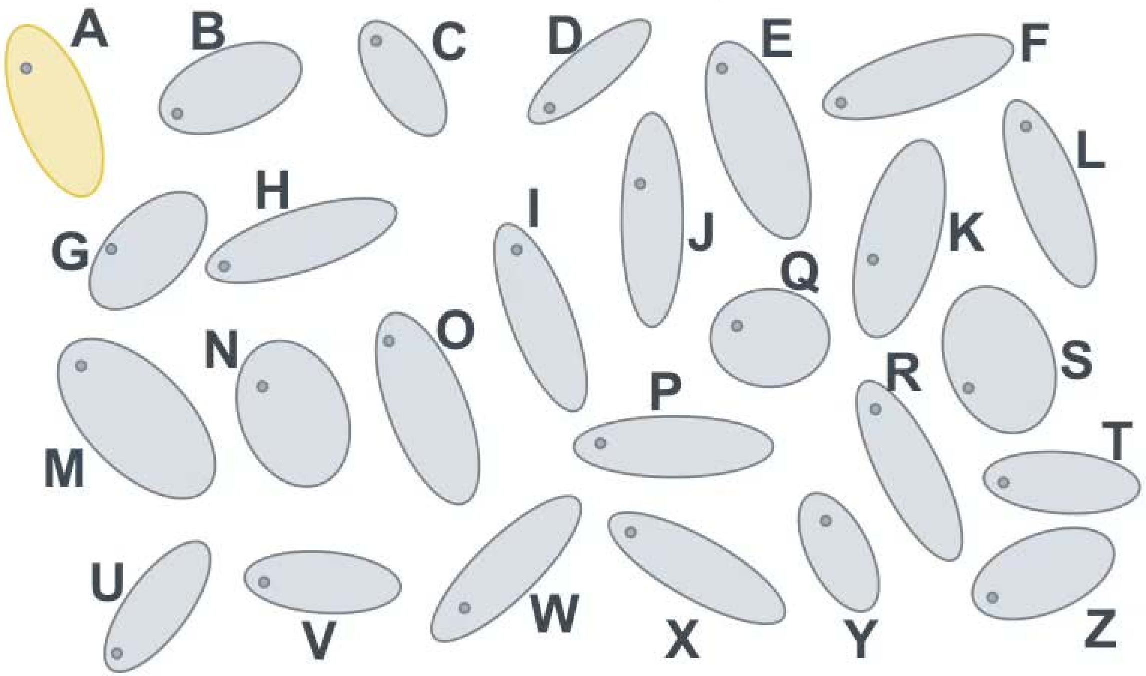
# Answering Queries using PWS

Let $\varphi$ be a query predicate and let $I(\varphi, w \in \Omega_{DB})$ be an indicator function returning one if predicate $\varphi$ holds in world $w$ and zero otherwise.

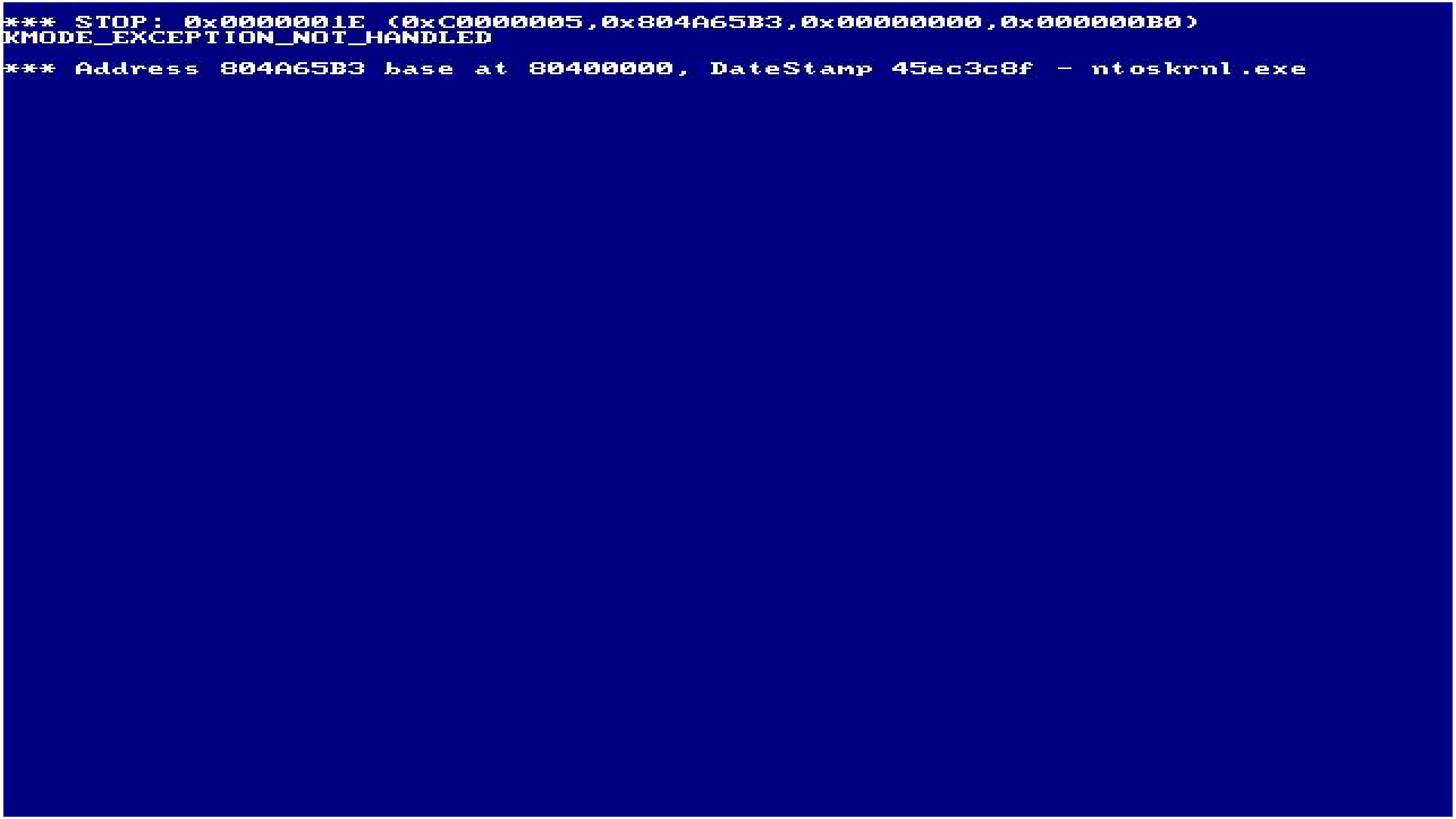
The probability $P(\varphi, D)$ of the event that a query predicate $\varphi$ holds on an uncertain database $DB$ is defined as

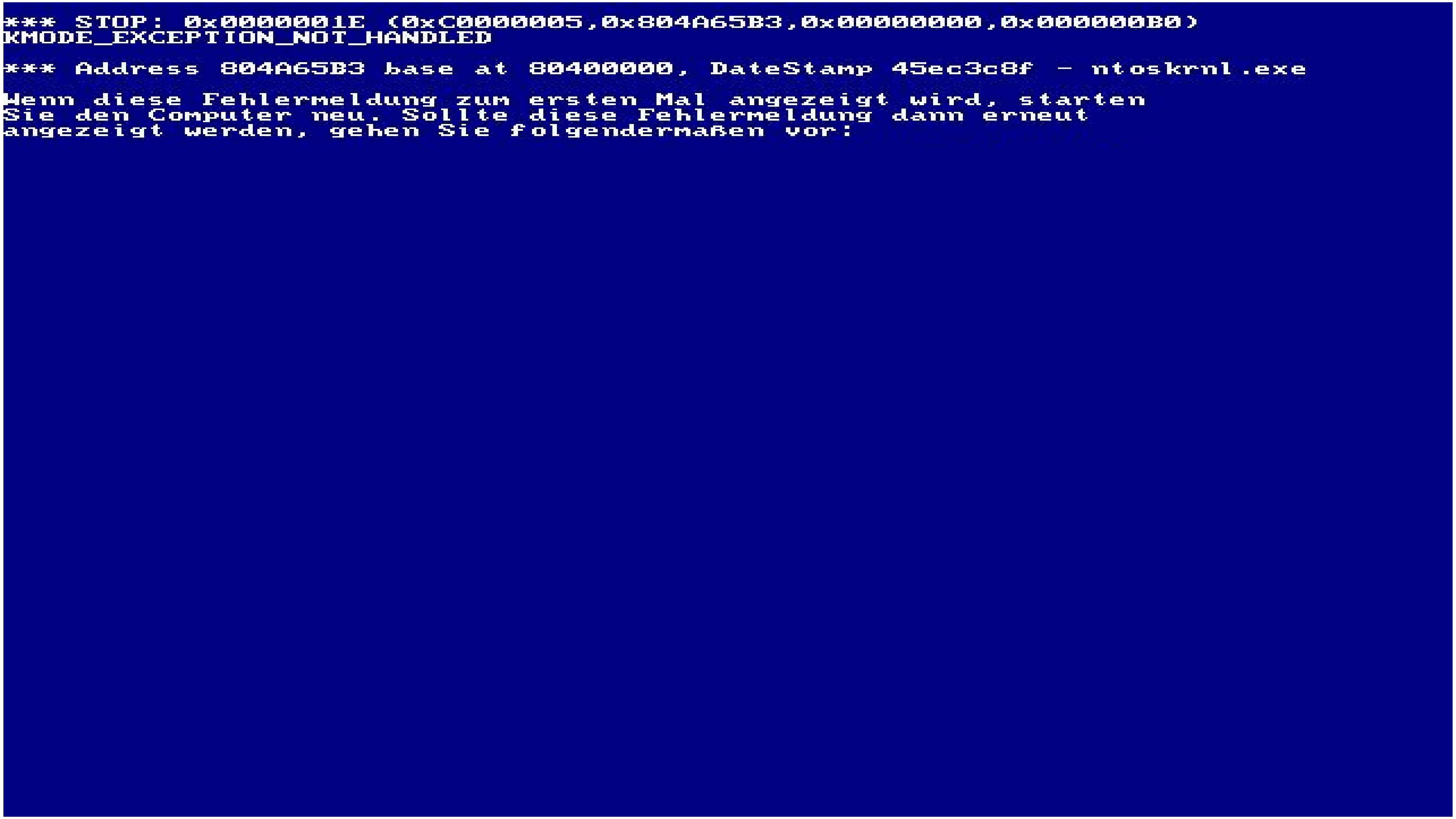
$$P(\varphi, D) = \sum_{w \in \Omega_{DB}} I(\varphi, w) P(w)$$

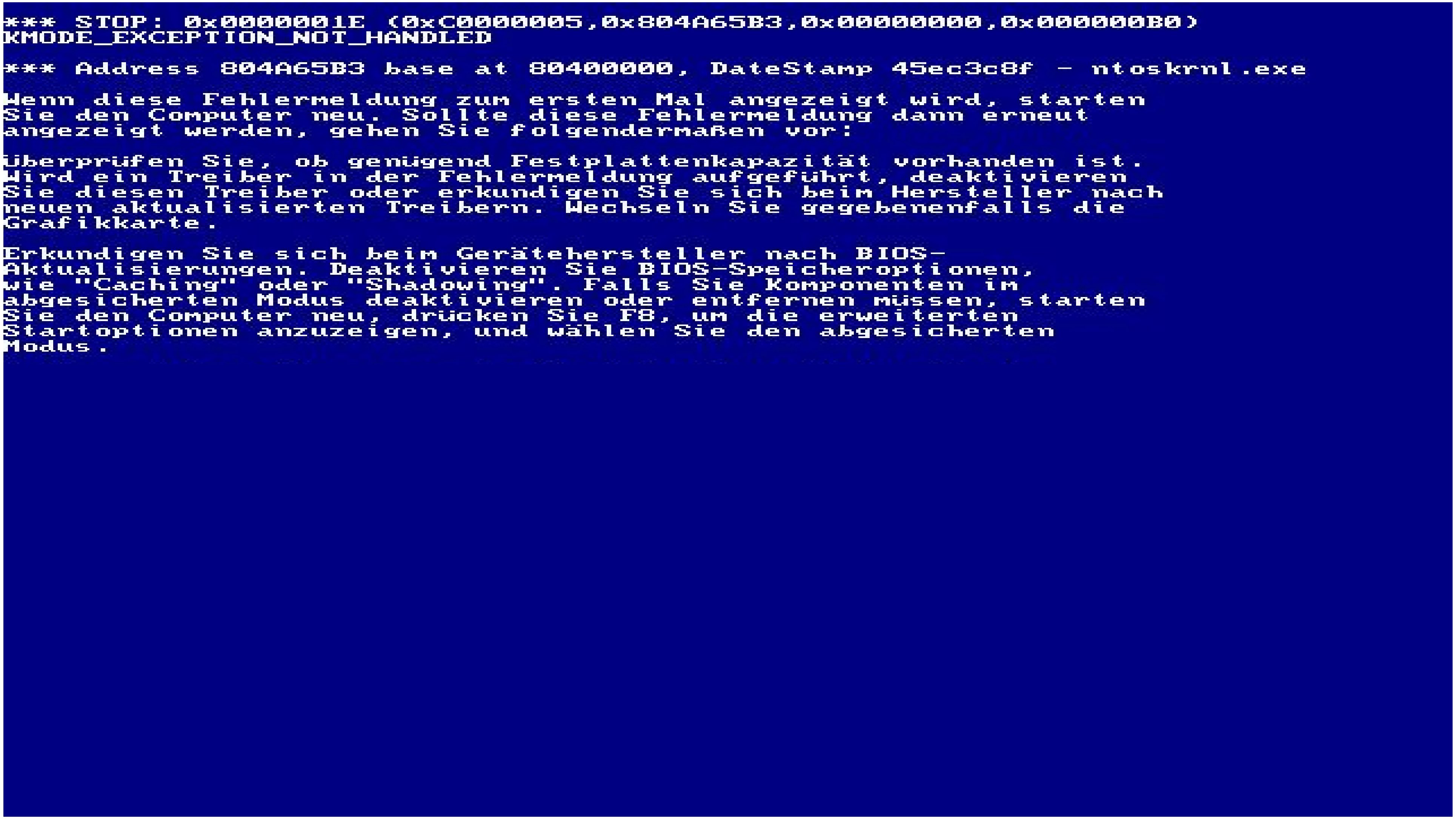
# Possible Worlds: Example II

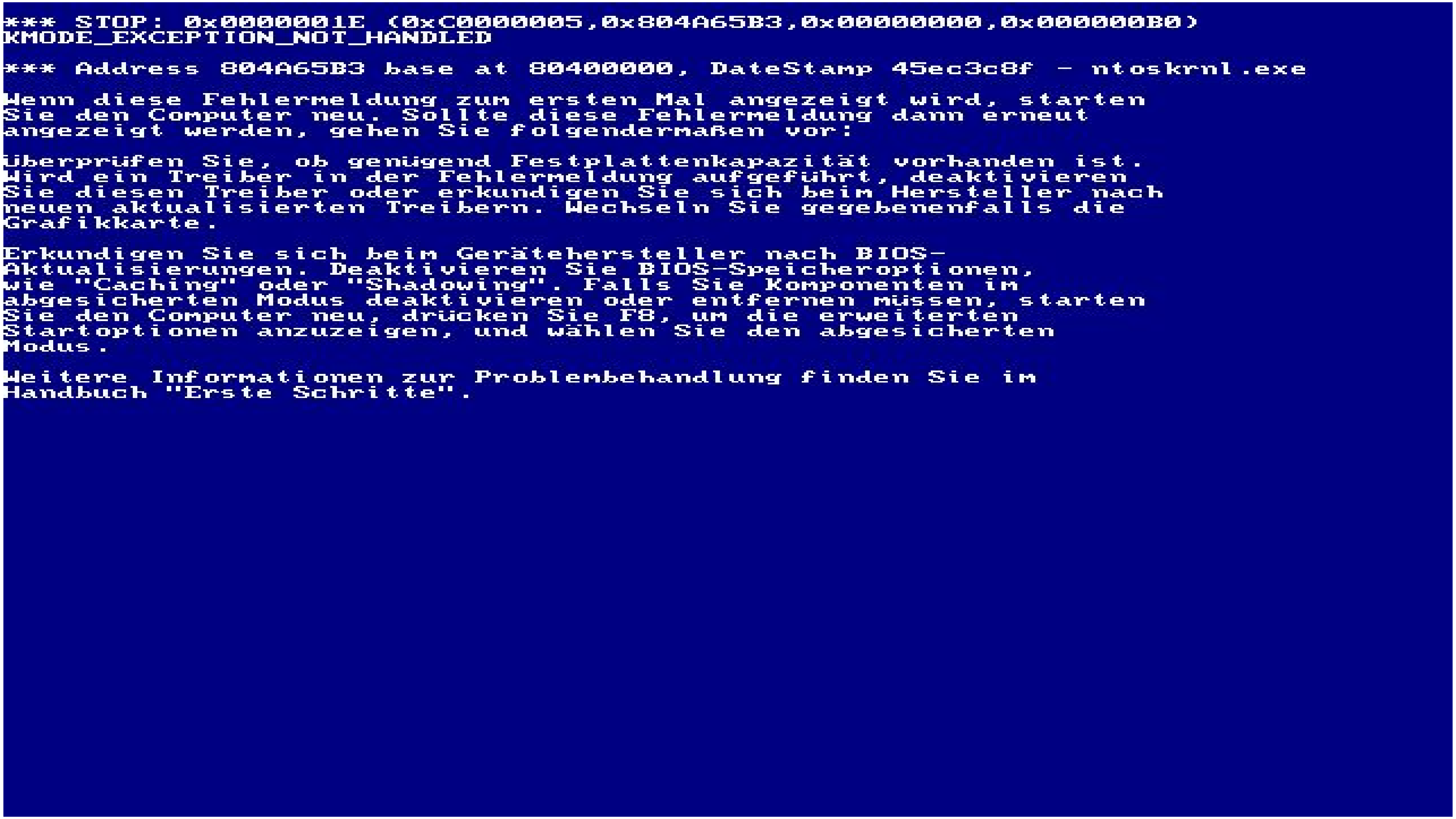


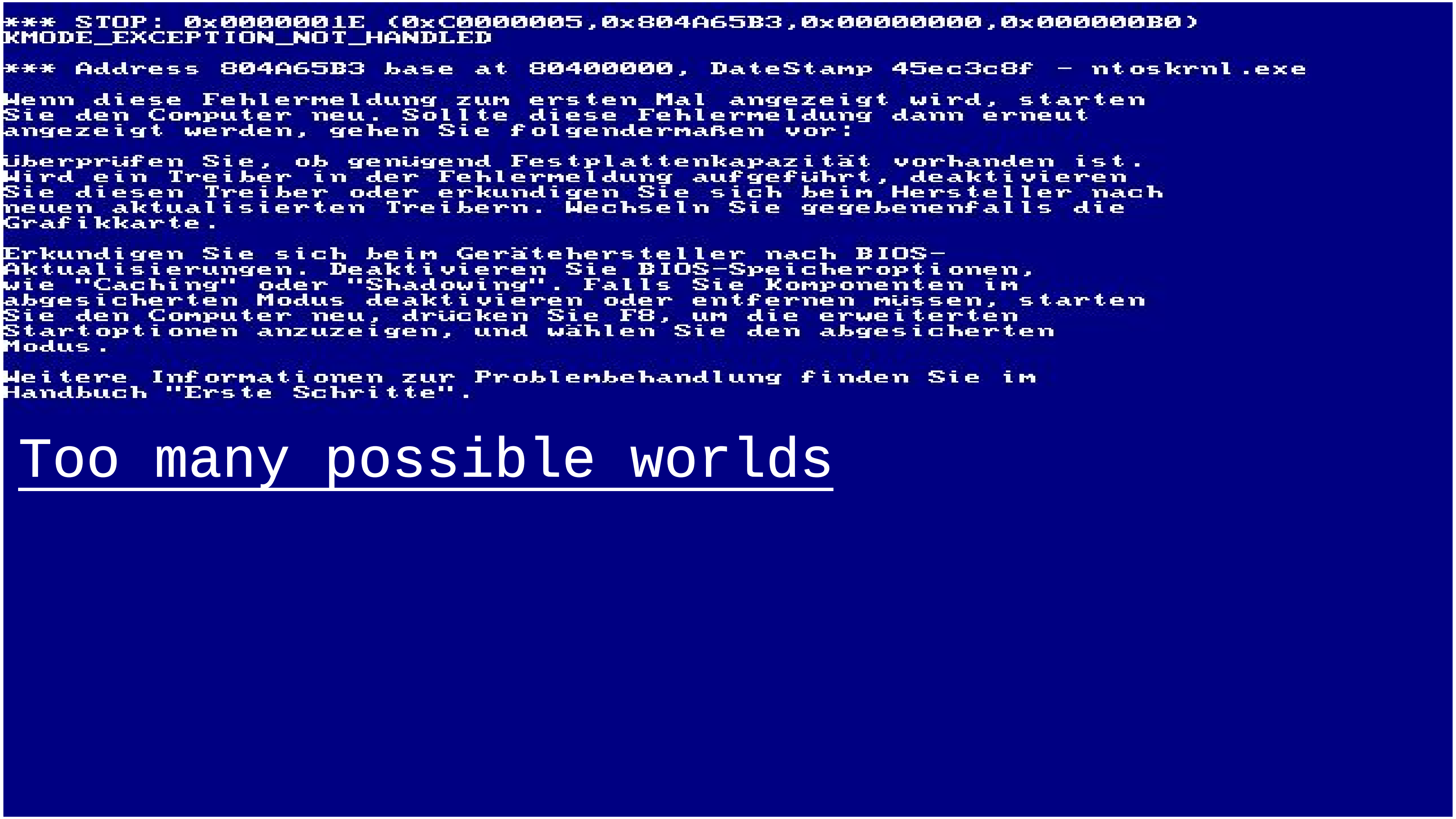

*** STOP: 0x0000001E (0xC0000005,0x804A65B3,0x00000000,0x000000B0)
KMODE_EXCEPTION_NOT_HANDLED


*** STOP: 0x0000001E (0xC0000005,0x804A65B3,0x00000000,0x000000B0)
KMODE_EXCEPTION_NOT_HANDLED
*** Address 804A65B3 base at 80400000, DateStamp 45ec3c8f - ntoskrnl.exe


*** STOP: 0x0000001E (0xC0000005,0x804A65B3,0x00000000,0x000000B0)
KMODE_EXCEPTION_NOT_HANDLED
*** Address 804A65B3 base at 80400000, DateStamp 45ec3c8f - ntoskrnl.exe
Wenn diese Fehlermeldung zum ersten Mal angezeigt wird, starten
Sie den Computer neu. Sollte diese Fehlermeldung dann erneut
angezeigt werden, gehen Sie folgendermaßen vor:


*** STOP: 0x0000001E (0xC0000005,0x804A65B3,0x00000000,0x000000B0)
KMODE_EXCEPTION_NOT_HANDLED
*** Address 804A65B3 base at 80400000, DateStamp 45ec3c8f - ntoskrnl.exe
Wenn diese Fehlermeldung zum ersten Mal angezeigt wird, starten Sie den Computer neu. Sollte diese Fehlermeldung dann erneut angezeigt werden, gehen Sie folgendermaßen vor:
Überprüfen Sie, ob genügend Festplattenkapazität vorhanden ist. Wird ein Treiber in der Fehlermeldung aufgeführt, deaktivieren Sie diesen Treiber oder erkundigen Sie sich beim Hersteller nach neuen aktualisierten Treibern. Wechseln Sie gegebenenfalls die Grafikkarte.
Erkundigen Sie sich beim Gerätehersteller nach BIOS-Aktualisierungen. Deaktivieren Sie BIOS-Speicheroptionen, wie "Caching" oder "Shadowing". Falls Sie Komponenten im abgesicherten Modus deaktivieren oder entfernen müssen, starten Sie den Computer neu, drücken Sie F8, um die erweiterten Startoptionen anzuzeigen, und wählen Sie den abgesicherten Modus.


*** STOP: 0x0000001E (0xC0000005,0x804A65B3,0x00000000,0x000000B0)
KMODE_EXCEPTION_NOT_HANDLED
*** Address 804A65B3 base at 80400000, DateStamp 45ec3c8f - ntoskrnl.exe
Wenn diese Fehlermeldung zum ersten Mal angezeigt wird, starten Sie den Computer neu. Sollte diese Fehlermeldung dann erneut angezeigt werden, gehen Sie folgendermaßen vor:
Überprüfen Sie, ob genügend Festplattenkapazität vorhanden ist. Wird ein Treiber in der Fehlermeldung aufgeführt, deaktivieren Sie diesen Treiber oder erkundigen Sie sich beim Hersteller nach neuen aktualisierten Treibern. Wechseln Sie gegebenenfalls die Grafikkarte.
Erkundigen Sie sich beim Gerätehersteller nach BIOS-Aktualisierungen. Deaktivieren Sie BIOS-Speicheroptionen, wie "Caching" oder "Shadowing". Falls Sie Komponenten im abgesicherten Modus deaktivieren oder entfernen müssen, starten Sie den Computer neu, drücken Sie F8, um die erweiterten Startoptionen anzuzeigen, und wählen Sie den abgesicherten Modus.
Weitere Informationen zur Problembehandlung finden Sie im Handbuch "Erste Schritte".




# Too many possible worlds


# Too many possible worlds

Main challenge:

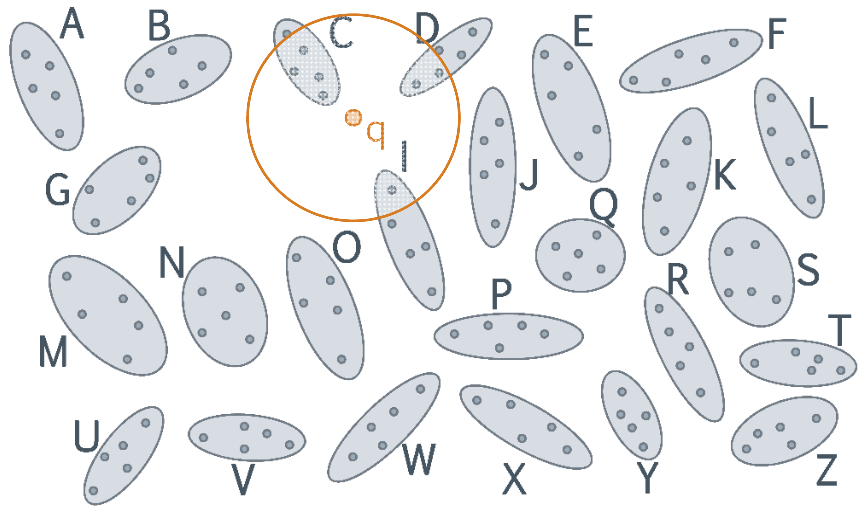
- Answer queries efficiently.
- Despite an exponential number of possible worlds

# Overview

1. Introduction to Probability Theory
2. Case Study: Probabilistic Count Queries

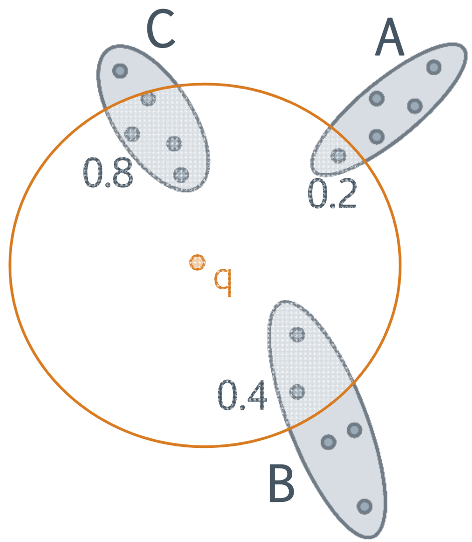
# Count Queries on Uncertain Data

How many objects are located in the depicted circular region centered at query point q?



# Count Queries on Uncertain Data

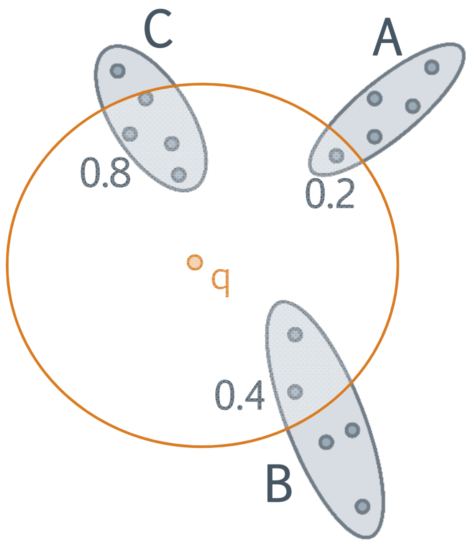
- $2^{|DB|}$ possible worlds
- Main idea: Use polyomial multiplication to enumerate possible results



# Count Queries on Uncertain Data

*Example:*

$\mathcal{F} =$
$(P(A) \cdot x + 1 - P(A)) \cdot$
$(P(B) \cdot x + 1 - P(B)) \cdot$
$(P(C) \cdot x + 1 - P(C))$



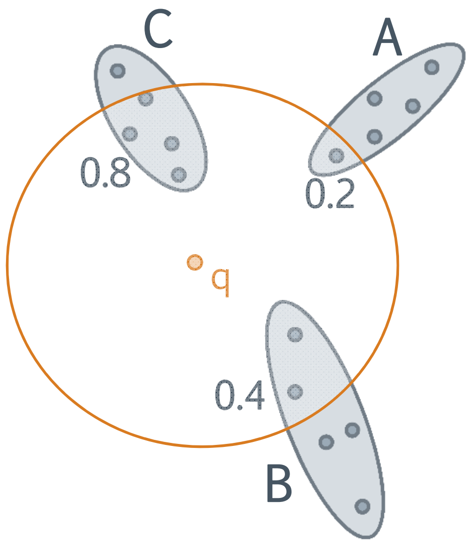
# Count Queries on Uncertain Data

*Example:*

$\mathcal{F} =$
$(P(A) \cdot x + 1 - P(A)) \cdot$
$(P(B) \cdot x + 1 - P(B)) \cdot$
$(P(C) \cdot x + 1 - P(C)) =$

$(0.2x+0.8) \cdot (0.4x+0.6) \cdot (0.8x + 0.2) =$

$(0.08x^2 + 0.12x + 0.32x + 0.48) \cdot (0.8x + 0.2)$



# Count Queries on Uncertain Data

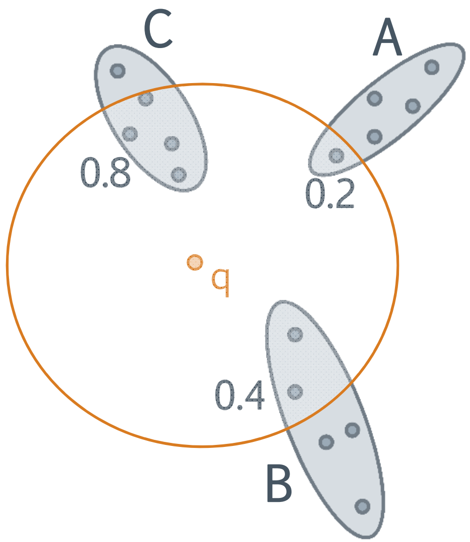
*Example:*

$\mathcal{F} =$
$(P(A) \cdot x + 1 - P(A)) \cdot$
$(P(B) \cdot x + 1 - P(B)) \cdot$
$(P(C) \cdot x + 1 - P(C)) =$

$(0.2x+0.8) \cdot (0.4x+0.6) \cdot (0.8x + 0.2) =$

$(0.08x^2 + 0.12x + 0.32x + 0.48) \cdot (0.8x + 0.2) =$

$(0.08x^2 + 0.44x + 0.48) \cdot (0.8x + 0.2)$



# Count Queries on Uncertain Data

*Example:*

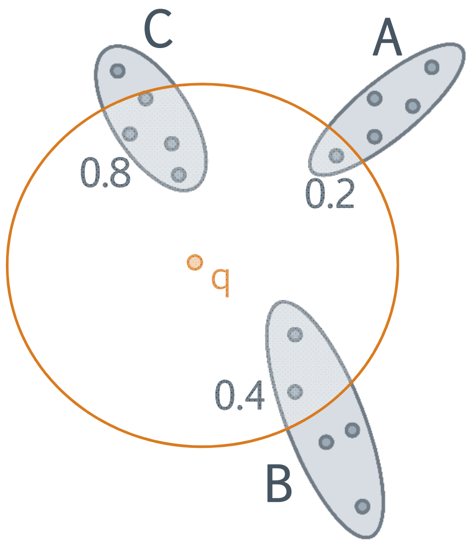
$\mathcal{F} =$
$(P(A) \cdot x + 1 - P(A)) \cdot$
$(P(B) \cdot x + 1 - P(B)) \cdot$
$(P(C) \cdot x + 1 - P(C)) =$

$(0.2x+0.8) \cdot (0.4x+0.6) \cdot (0.8x + 0.2) =$

$(0.08x^2 + 0.12x + 0.32x + 0.48) \cdot (0.8x + 0.2) =$

$(0.08x^2 + 0.44x + 0.48) \cdot (0.8x + 0.2) =$

$(0.032x^3 + 0.224x^2 + 0.456x^1 + 0.288x^0)$



# Count Queries on Uncertain Data

*Example:*

$\mathcal{F} =$
$(P(A) \cdot x + 1 - P(A)) \cdot$
$(P(B) \cdot x + 1 - P(B)) \cdot$
$(P(C) \cdot x + 1 - P(C)) =$

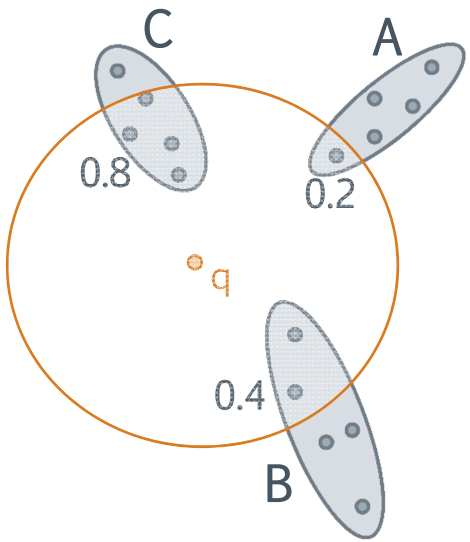
$(0.2x+0.8) \cdot (0.4x+0.6) \cdot (0.8x + 0.2) =$

$(0.08x^2 + 0.12x + 0.32x + 0.48) \cdot (0.8x + 0.2) =$

$(0.08x^2 + 0.44x + 0.48) \cdot (0.8x + 0.2) =$

$(0.032x^3 + 0.224x^2 + 0.456x^1 + 0.288x^0)$

Probability that exactly two objects are inside the query region



# Count Queries on Uncertain Data

*Example:*

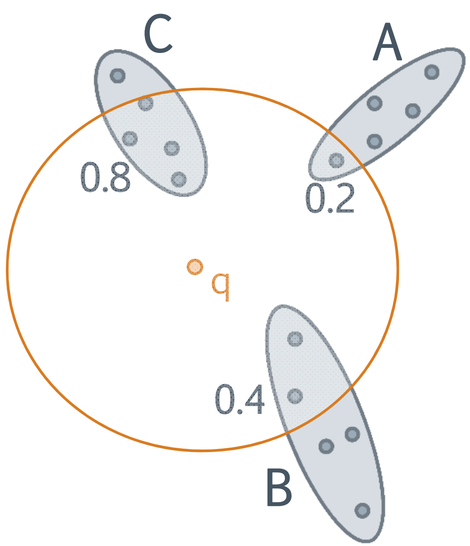
$\mathcal{F} =$
$(P(A) \cdot x + 1 - P(A)) \cdot$
$(P(B) \cdot x + 1 - P(B)) \cdot$
$(P(C) \cdot x + 1 - P(C)) =$

$(0.2x+0.8) \cdot (0.4x+0.6) \cdot (0.8x + 0.2) =$

$(0.08x^2 + 0.12x + 0.32x + 0.48) \cdot (0.8x + 0.2) =$

$(0.08x^2 + 0.44x + 0.48) \cdot (0.8x + 0.2) =$

$(0.032x^3 + 0.224x^2 + 0.456x^1 + 0.288x^0)$




Polynomial time solution: Unify worlds that are equivalent with respect to the query predicate!