Multivariate Pareto Records

James Allen Fill and Daniel Q. Naiman

Department of Applied Mathematics and Statistics
The Johns Hopkins University

AofA at CIRM, Luminy, France
June 25, 2019

(Goal: "executive summary" of 3 papers in ≤ 25 minutes!)

A "null" continuous multivariate model

We have new and basic things to say about **multivariate records** (stay tuned for a definition!) for the very simple "null" model of iid d-dimensional observations $X^{(1)}, X^{(2)}, \ldots$ (copies of X) with iid coordinates.

In this talk, we will take the distribution of each coordinate to be either Uniform$(0,1)$ or (by taking negative logs) Exponential(1).

We will focus on the following questions, especially the second:

1. How can we sample (**g**enerate) multivariate records efficiently? (F & Naiman, 2019**g**)

2. How does the record-setting **f**rontier behave asymptotically? (F & Naiman, 2019**f**)

3. What can we say about the number of records **b**roken when a new record is set? (F, 2019**b**)

All three papers have been submitted, and all are on the arXiv.

Basic definitions: records

- Let $x_+ := \sum_{j=1}^{d} x_j$.
- Write $x \prec y$ to mean that $x_j < y_j$ for $1 \le j \le d$.
- Write $x \le y$ to mean that $x_j \le y_j$ for $1 \le j \le d$.
- We say that $X^{(k)}$ is a (Pareto) record (or that it sets a record at time k) if $X^{(k)} \not\prec X^{(i)}$ for all $1 \le i < k$.
- If $1 \le k \le n$, we say that $X^{(k)}$ is a current record (or remaining record, or maximum) at time n if $X^{(k)} \not\prec X^{(i)}$ for all $1 \le i \le n$.
- If $1 \le k \le n$, we say that $X^{(k)}$ is a broken record at time n if it is a record but not a current record, that is, if $X^{(k)} \not\prec X^{(i)}$ for all $1 \le i < k$ but $X^{(k)} \prec X^{(\ell)}$ for some $k < \ell \le n$; in that case, the observation corresponding to the smallest such ℓ is said to break or kill the record $X^{(k)}$.

Basic definitions: record-setting region and its frontier

- The record-setting region at time n is the (random) closed set of points

$$\mathrm{RS}_n := \{x \in \mathbb{R}^d : 0 \leq x \not\succeq X^{(i)} \text{ for all } 1 \leq i \leq n\}.$$

- We call the (topol.) boundary of RS_n (relative to the closed positive orthant determined by the origin) its frontier and denote it by F_n.

Remark: The terminology for RS_n is natural since the next observation $X^{(n+1)}$ sets a record if and only if it falls in the record-setting region.

Record frontier

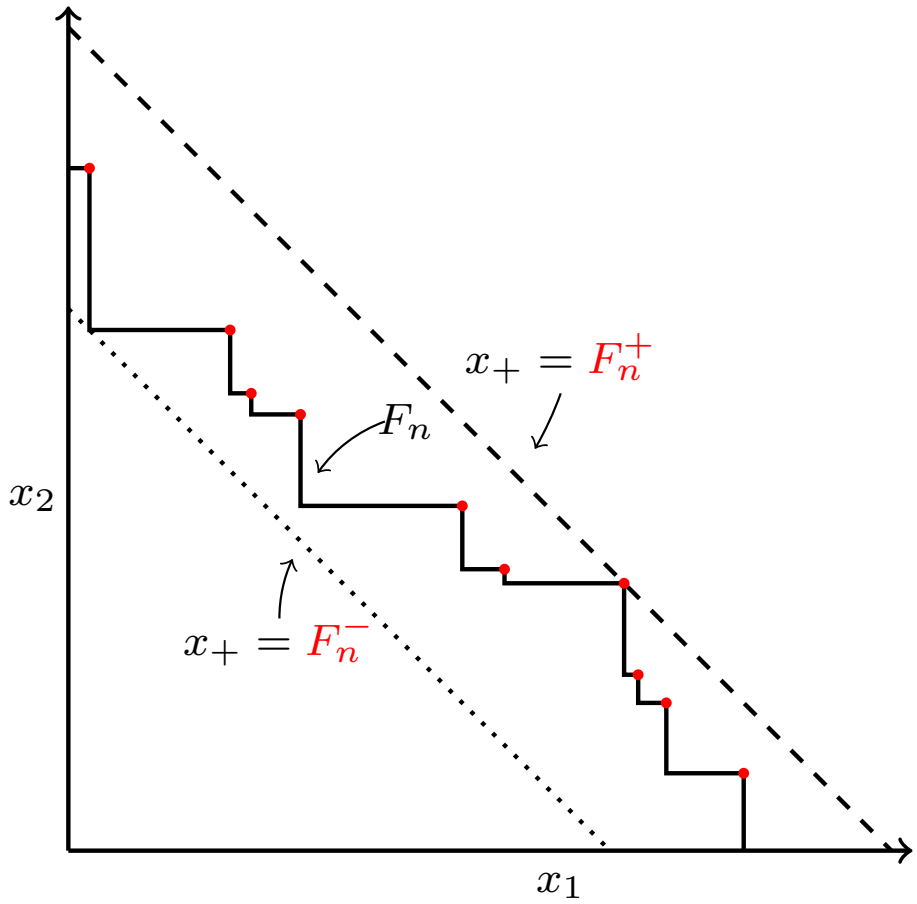


Figure: Record frontier F_n based on n bivariate Exponential(1) observations resulting in 10 current records (shown as solid points). The values $F_n^- = \min\{x_+ : x \in F_n\}$ and $F_n^+ = \max\{x_+ : x \in F_n\}$ determine two hyperplanes $x_+ = F_n^-$ and $x_+ = F_n^+$.
A new observation sets a record if and only if it falls in the region to the upper right of F_n.

1. Efficient sampling of multivariate records

1. How can we sample multivariate records efficiently?

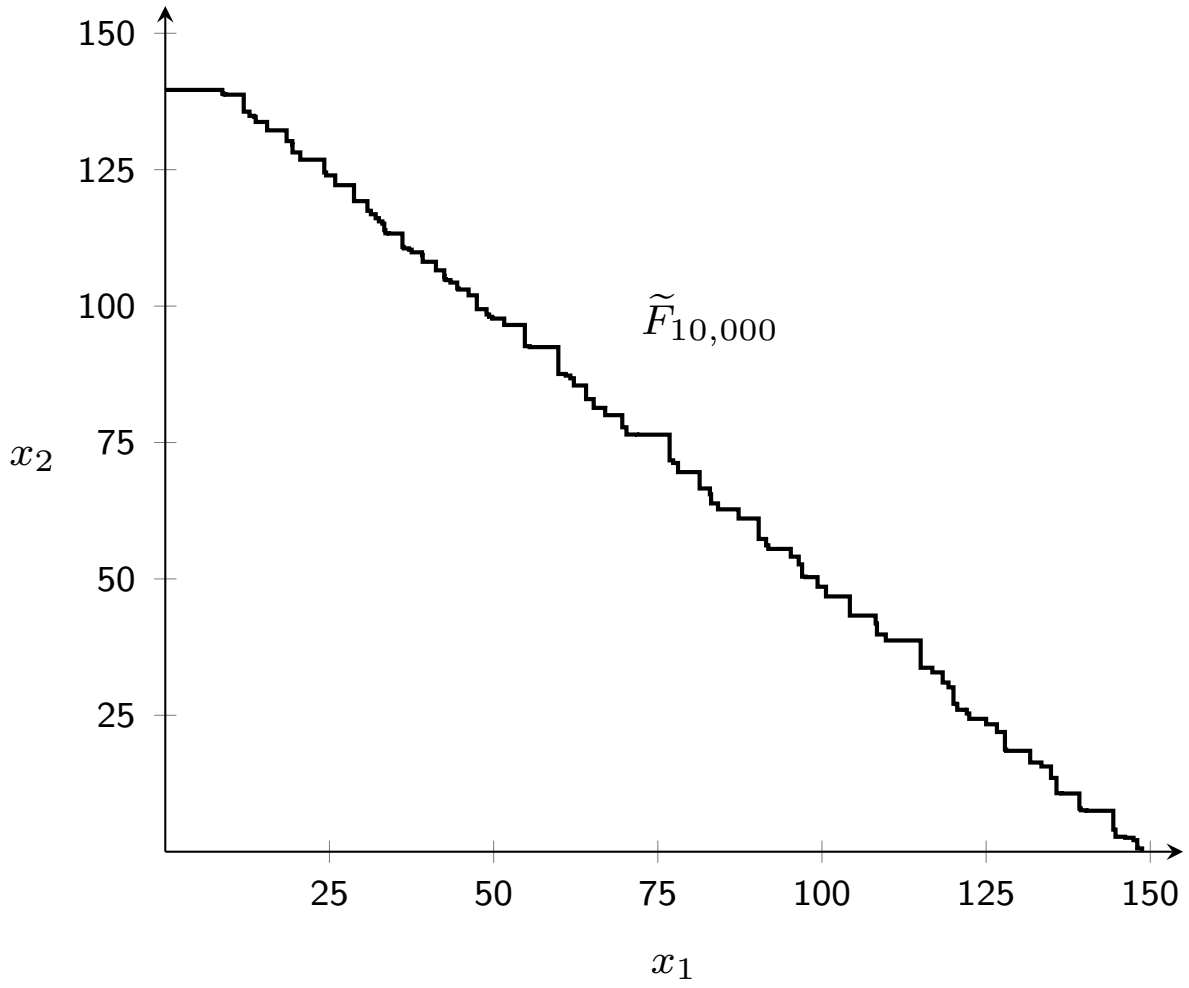


Figure: Record frontier $\widetilde{F}_{10,000}$ after 10,000 records generated using the importance-sampling algorithm described in F & Naiman (2019g).

1. Efficient sampling of multivariate records

How was this figure generated?

Spoiler: not by generating bivariate observations $X^{(1)}, X^{(2)}, \ldots$ and waiting for 10,000 records to be generated!

Let T_m denote the number of observations required for m records to be set. Among other more precise results, F & Naiman (2019f) show that

$$\frac{\mathrm{L}\, T_m}{(d!m)^{1/d}} \xrightarrow{\text{a.s.}} 1. \qquad (\mathrm{L} \equiv \ln,\ \mathrm{L}_2 \equiv \ln\ln, \text{ etc.})$$

With $d = 2$ and $m =$ 10,000 as in the figure, this gives the estimate

$$T_{10{,}000} \approx 10^{61} \quad (!!)$$

Suppose here that coordinates are distributed Uniform$(0,1)$.

The record-setting region RS_n after n observations can be represented as the non-disjoint union

$$\mathsf{RS}_n = \cup_{g \in G_n} O_g^+$$

of the translated positive orthants

$$O_g^+ := \{y \in [0,1)^d : y \geq g\},$$

where G_n is the set of minima of the frontier F_n.

a bivariate example: three slides back

a trivariate example: next slide

Trivariate example of generators

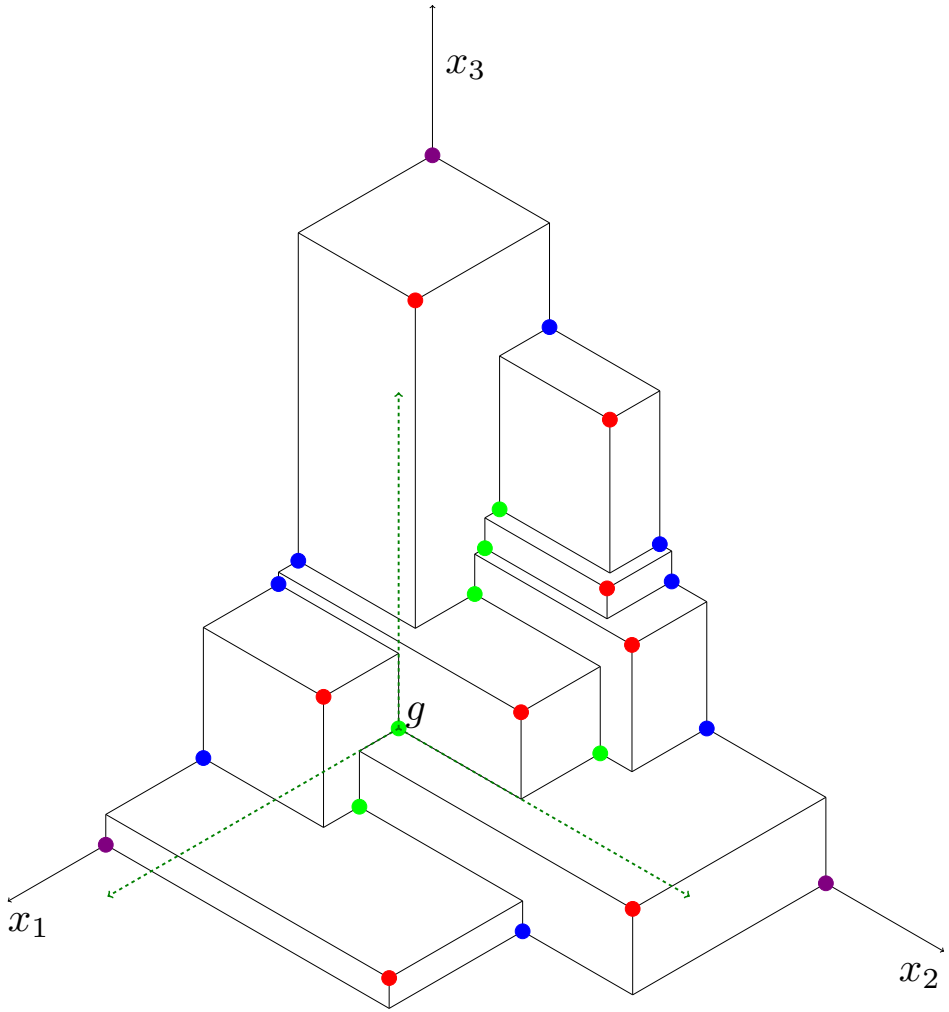


Figure: Example of a record frontier in dimension $d = 3$ with 8 remaining records shown in red and the resulting 17 generators: three one-dimensional generators shown in violet, eight two-dimensional generators shown in blue, and six three-dimensional (interior) generators shown in green. The lower boundary of one of the orthants O_g^+ is shown using green dashed lines.

Generating the next multivariate record

Importance sampling subroutine for generating a new record $\mathbf{R}$ (given the current set G of generators):

1. Sample a generator $\mathbf{g} = g$ from G proportional to the volume of the corresponding orthants O_g^+. (Computing volumes of orthants is very easy!, so this sampling is easy, too.)

2. Sample $\mathbf{R}$ uniformly from the orthant $O_{\mathbf{g}}^+$. (This is easy, too.)

3. Accept $\mathbf{R}$ as a new record with probability equal to the reciprocal of the number of generator-orthants O_g^+ to which it belongs. If $\mathbf{R}$ is rejected, repeat Steps 1–3.

4. Update G to G', as described (& analyzed) in F & Naiman (2019g). (This isn't so easy!)

Remarks:

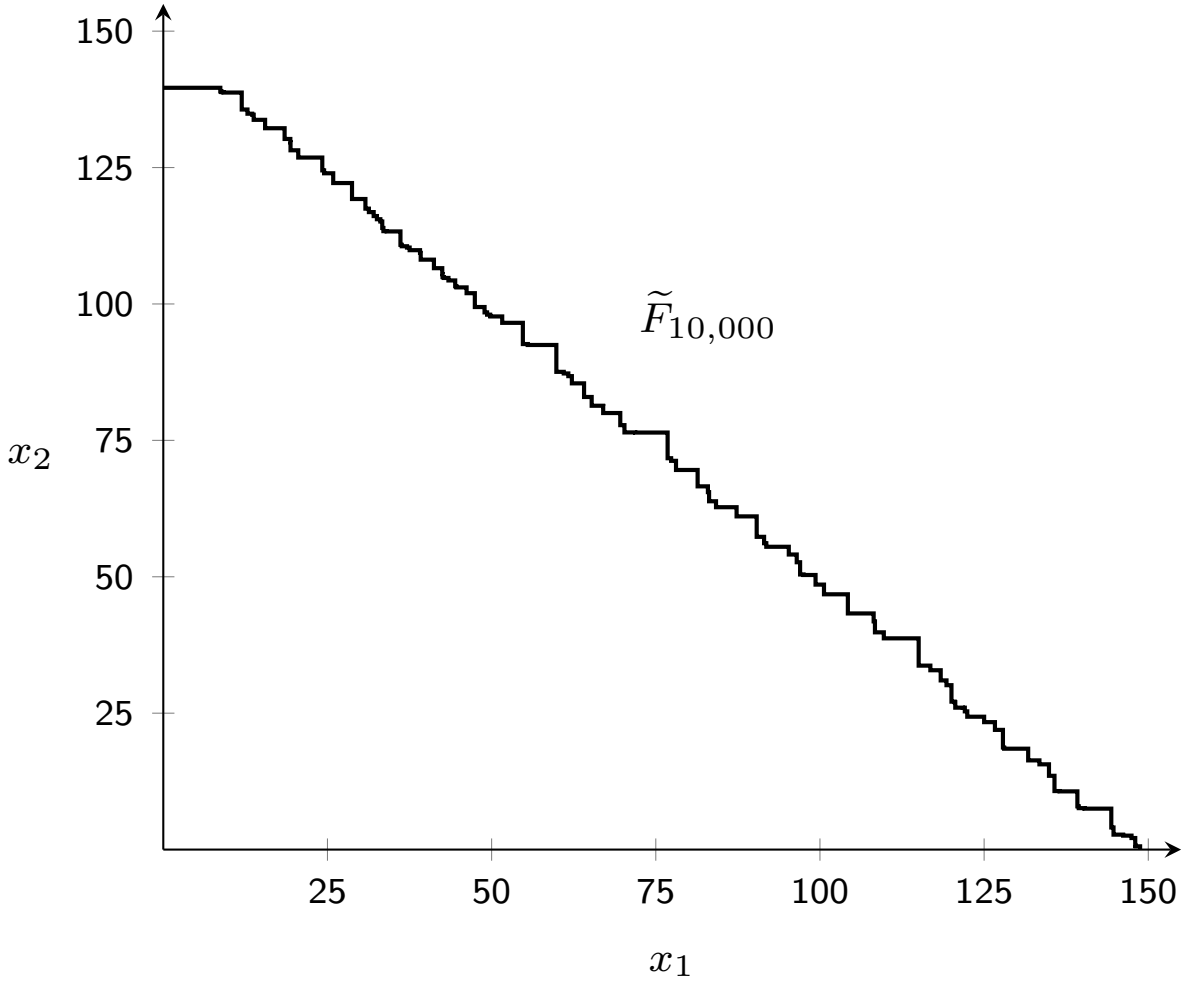
1. Results are easier to discover empirically in "records-time" but easier to prove in "observations-time"!

2. Results in "observations-time" can be transferred to results in "records-time" using time change arguments that exploit information concerning the record counts R_n or (equivalently) the record epochs T_m: See Sections 4–5 of F & Naiman (2019f).

From now on, we take observation coords. to be distributed Exp(1).

The frontier is nearly planar

Recall this nearly planar (i.e., linear) frontier figure:



Questions about the frontier

Questions to ask about the frontier:

(a) Where is the frontier located? I.e., what is an approximating hyperplane?

(b) How thick (i.e., wide) is the frontier?

For sharper answers than in today's talk, see the paper F & Naiman (2019f).

Approximate location of frontier: $x_+ = \mathrm{L}\, n$

Question (a) is easy to answer: Deviations of the sum of coordinates for a generic current record at time n from $\mathrm{L}\, n$ are typically of constant order.

Observe that the conditional distribution of $X_+^{(k)}$ given that $X^{(k)}$ is a current record at time n doesn't depend on $k \in \{1, \ldots, n\}$; in particular, it's the conditional distribution of $X_+^{(n)}$ given that $X^{(n)}$ sets a record. Let Y_n be a random variable with that distribution.

Let G denote a random variable with the standard Gumbel distribution (i.e., distribution function $x \mapsto e^{-e^{-x}}$, $x \in \mathbb{R}$), and write $\xrightarrow{\mathcal{L}}$ for convergence in law (i.e., in distribution).

Theorem

We have

$$Y_n - \mathrm{L}\, n \xrightarrow{\mathcal{L}} G.$$

The proof, which makes use of Scheffé's theorem (so that there is in fact convergence in total variation) is quite elementary.

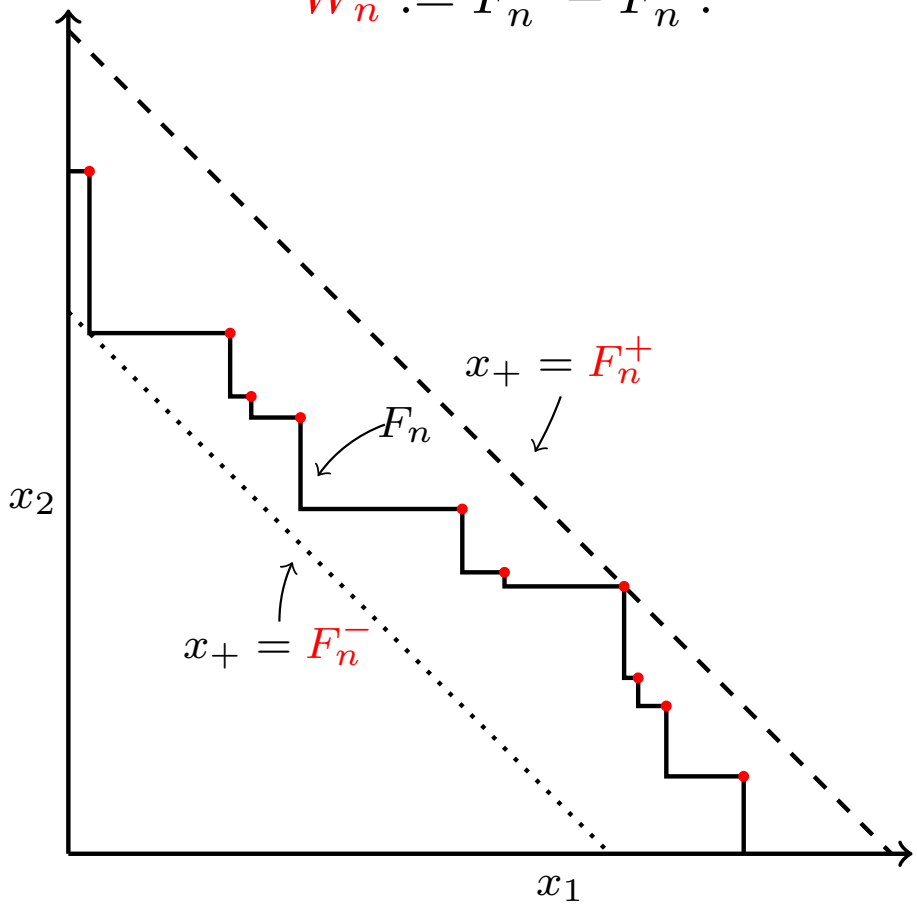
Width of frontier

- Recall that F_n denotes the frontier of RS_n, and let

$$F_n^- := \min\{x_+ : x \in F_n\} \quad \text{and} \quad F_n^+ := \max\{x_+ : x \in F_n\}.$$

- We define the width of F_n as

$$W_n := F_n^+ - F_n^-.$$



The leading edge F_n^+ of the frontier

Fortunately, to study the width $W_n = F_n^+ - F_n^-$, we can get away with studying the leading edge F_n^+ and the trailing edge F_n^- *separately*.

Lemma (characterization of F_n^+)

We have

$$F_n^+ = \max\{X_+^{(k)} : 1 \le k \le n\},$$

which is nondecreasing in n.

The proof is easy.

So the process F^+ is simply the partial-maximum process corresponding to iid Gamma$(d, 1)$ random variables, and classical extreme value theory due to Jack Kiefer (1972) (involving rather sophisticated use of the Borel–Cantelli lemmas) can be brought to bear.

Asymptotic behavior of the leading edge F_n^+

Theorem (derived from Kiefer, Sixth Berkeley Symposium, 1972)

(a) Typical behavior of F^+:

$$\frac{F_n^+ - \mathrm{L}\, n}{\mathrm{L}_2\, n} \xrightarrow{\mathrm{P}} d - 1.$$

(b) Almost sure behavior for F^+:

$$\liminf \frac{F_n^+ - \mathrm{L}\, n}{\mathrm{L}_2\, n} = d - 1 < d = \limsup \frac{F_n^+ - \mathrm{L}\, n}{\mathrm{L}_2\, n} \text{ a.s.}$$

Asymptotic behavior of the trailing edge F_n^-

Theorem

(a) Typical behavior of F^-:

$$\frac{F_n^- - \mathrm{L}\, n}{\mathrm{L}_2\, n} \xrightarrow{\mathrm{P}} 0.$$

(b) Almost sure behavior for F^-: *If $d \geq 2$, then*

$$\lim \frac{F_n^- - \mathrm{L}\, n}{\mathrm{L}_2\, n} = 0 \text{ a.s.}$$

The proof combines routine probabilistic arguments with a novel geometric lemma.

A geometric lemma, illustrated

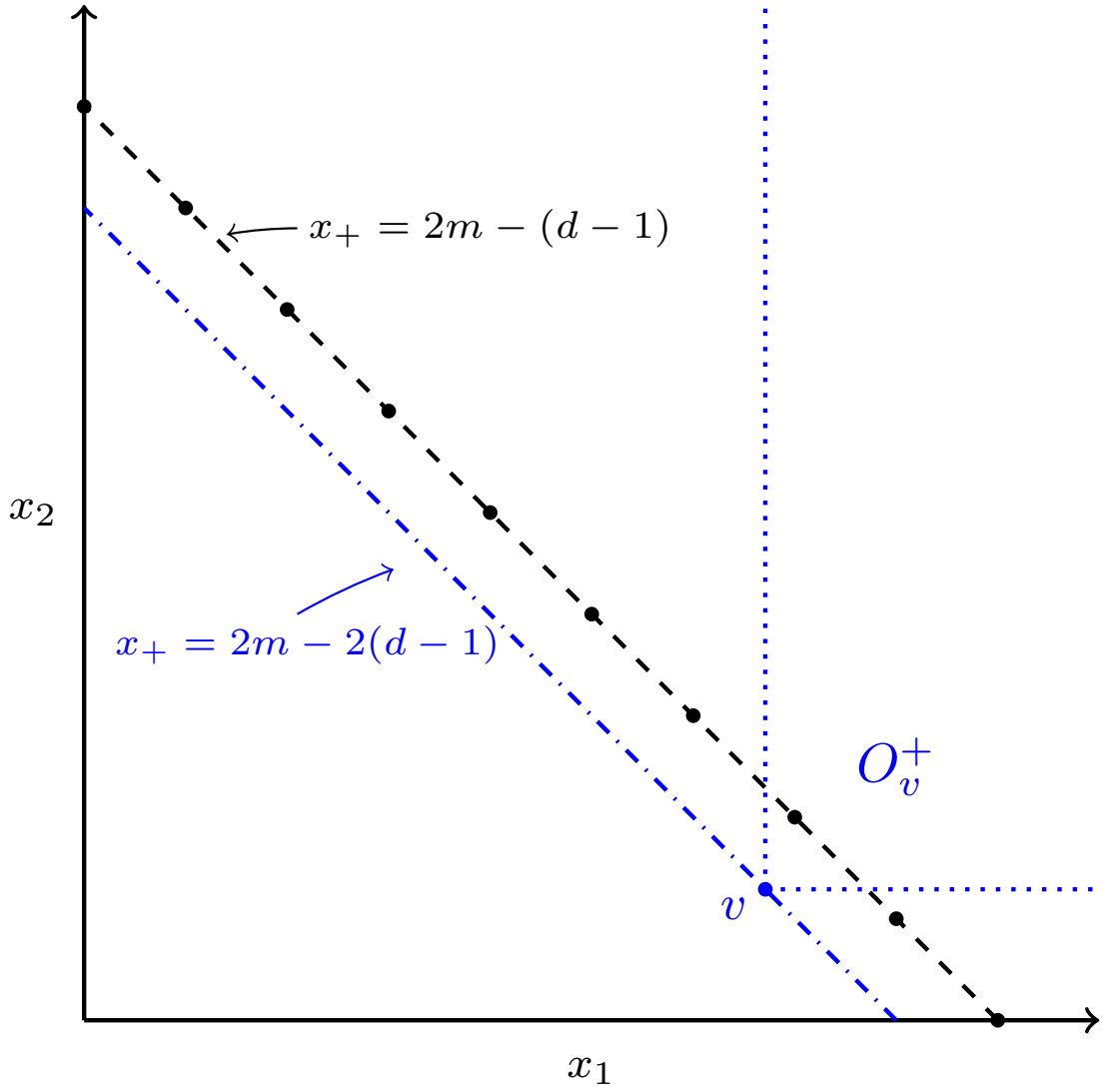


Figure: Geometric lemma illustrated for $d = 2$. Given v with $v_+ = 2m - 2(d-1)$, the orthant O_v^+ determined by v must contain a point i with integer coordinates on the hyperplane $x^+ = 2m - (d-1)$.

Asymptotic behavior of the width W_n

Theorem

(a) Typical behavior of W:

$$\frac{W_n}{\mathrm{L}_2\, n} \xrightarrow{\mathrm{P}} d-1.$$

(b) Almost sure behavior for W: *If $d \geq 2$, then*

$$\liminf \frac{W_n}{\mathrm{L}_2\, n} = d-1 < d = \limsup \frac{W_n}{\mathrm{L}_2\, n} \text{ a.s.},$$

and, in particular,

$$W_n = \Theta(\mathrm{L}_2\, n) \text{ a.s.} \quad \square$$

Remark: When $d = 1$, at each time $n \geq 1$ there is one current record, $F_n^+ = F_n^-$ is the value of that record, RS_n is the interval $[F_n^+, \infty)$, and $W_n = 0$.

3. Broken records: simulation of 100,000 bivariate records

N_k = number of records (out of 100,000) that break k current records.

k	N_k
0	50,334
1	24,667
2	12,507
3	63,35
4	3,040
5	1,571
6	782
7	364
8	202
9	94
10	48
11	24
12	18
13	8
14	4
16	1
17	0
18	1

Geometric$(1/2)$ distribution for bivariate record-breaking

Here is the main theorem of F (2019b). That paper also presents a first-order correction term.

Theorem

Consider our "null" bivariate model for observations. Let $K_n = -1$ *if the* n^{th} *observation is not a new record, and otherwise let* K_n *denote the number of remaining records killed by the* n^{th} *observation. Then* K_n*, conditionally given* $K_n \geq 0$*, converges in distribution to* $G - 1$*, where* $G \sim \text{Geometric}(1/2)$*, as* $n \to \infty$*.*

The paper provides a possible roadmap for the proof of a stronger result:

Conjecture

The fractions $\tilde{p}_{M,k}$ *of the first* M *records that break precisely* k *remaining records satisfy*

$$\sup_{k \geq 0} \left| \tilde{p}_{M,k} - 2^{-(k+1)} \right| \xrightarrow{\text{a.s.}} 0 \text{ as } M \to \infty.$$

Record-breaking in higher dimensions

Similarly, the following conjecture arises from data generated by the importance-sampling algorithm for higher dimensions:

Conjecture

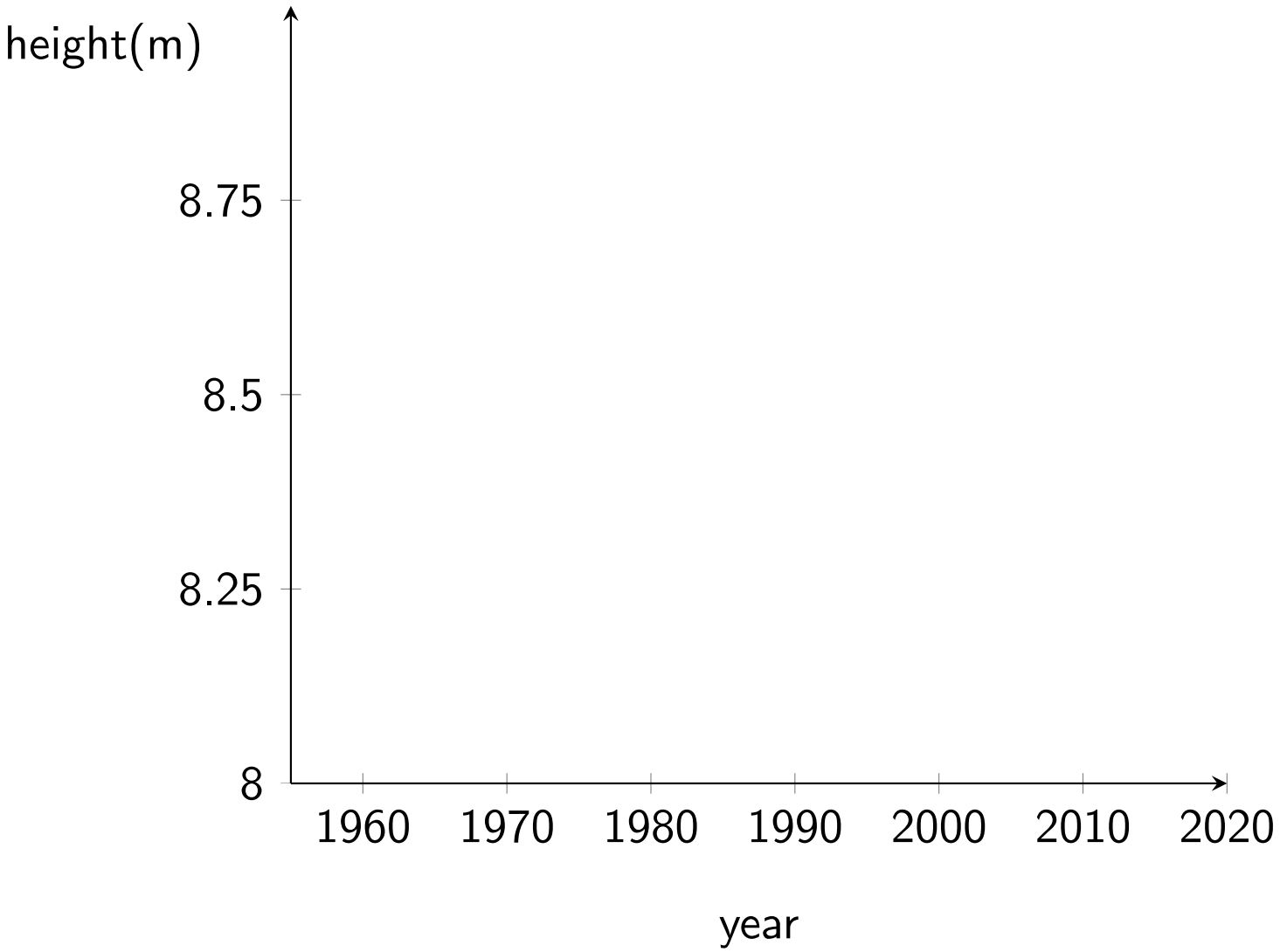
Consider dimension $d \geq 2$. Let $f_{d,m}$ denote the fraction of the first m records set that break 0 records. Then there exist constants $p_d \in (0,1)$ such that, almost surely, $f_{d,m} \to p_d$ as $m \to \infty$. Further, $p_d \to 1$ as $d \to \infty$.

- The data also suggest that perhaps $p_d = 1 - d^{-1}$ for every $d \geq 2$.
- Even for $d = 2$, the conjecture is stronger than what is proved in F (2019b).
- For $d \geq 3$ and $k \geq 1$, we do not know what to conjecture concerning the limiting behavior of the fraction of the first m records set that break k records.

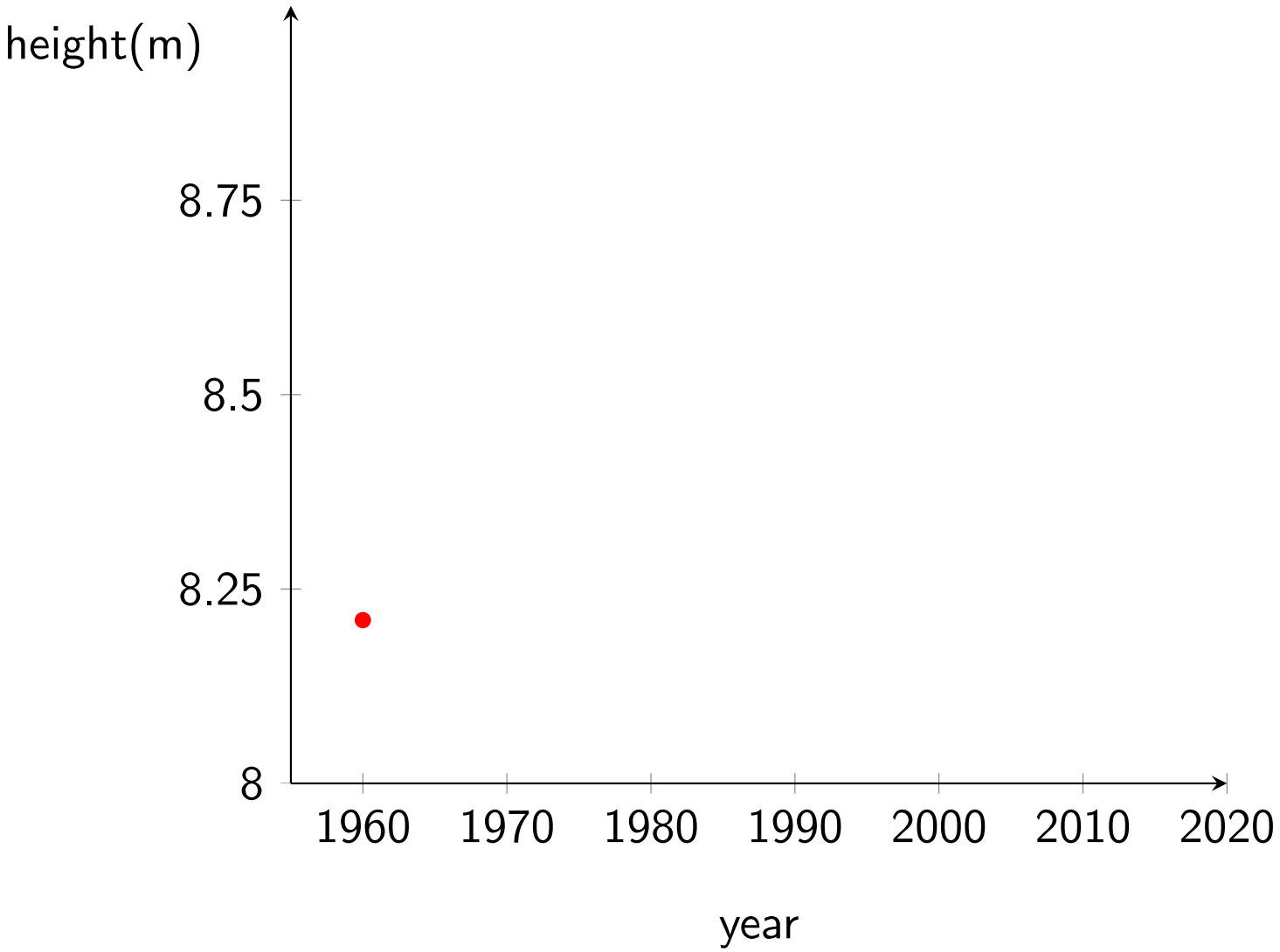
I will present more details in Klagenfurt in 2020.

See you then!

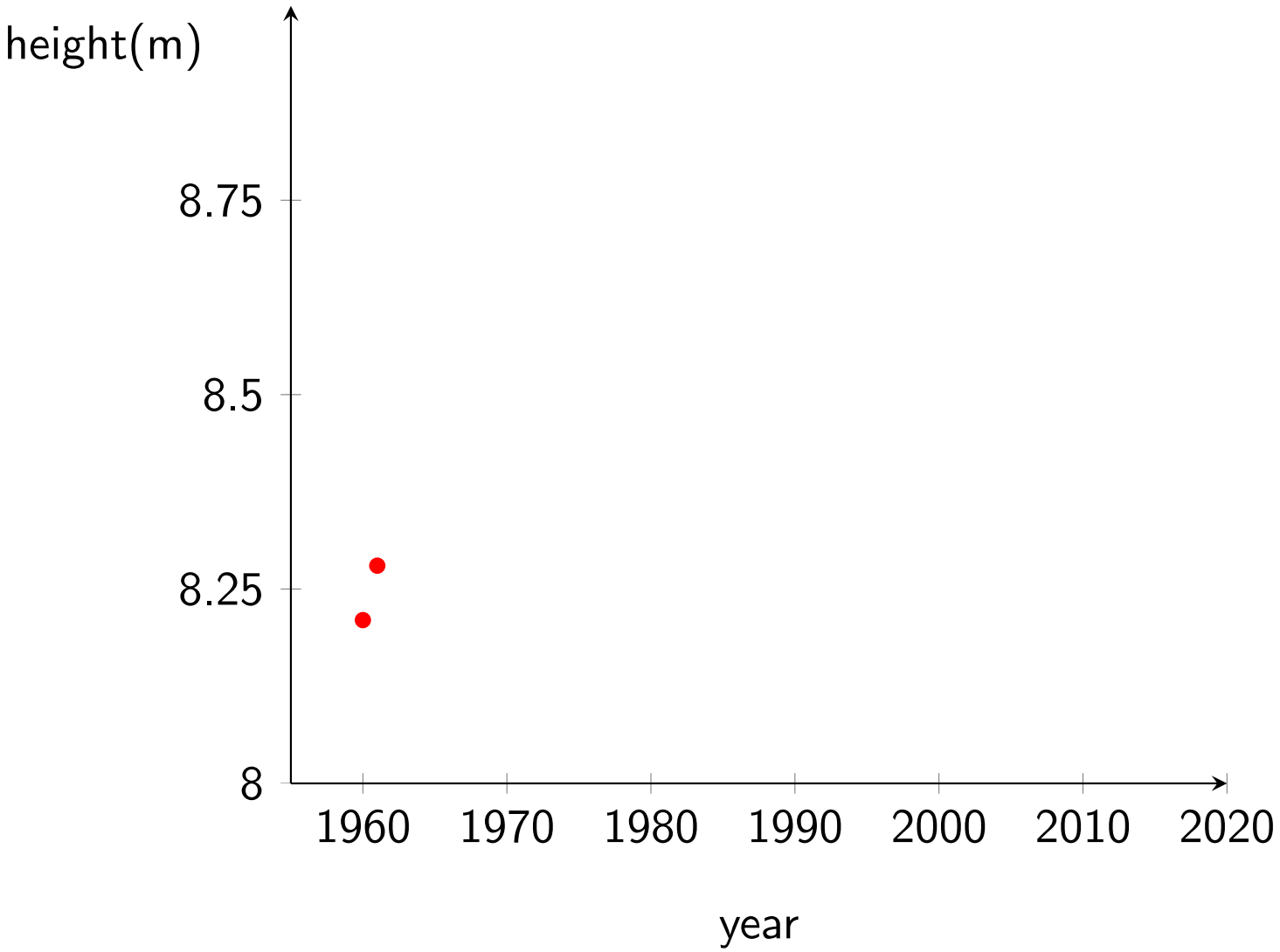
Men's Best Long-Jumps by Year



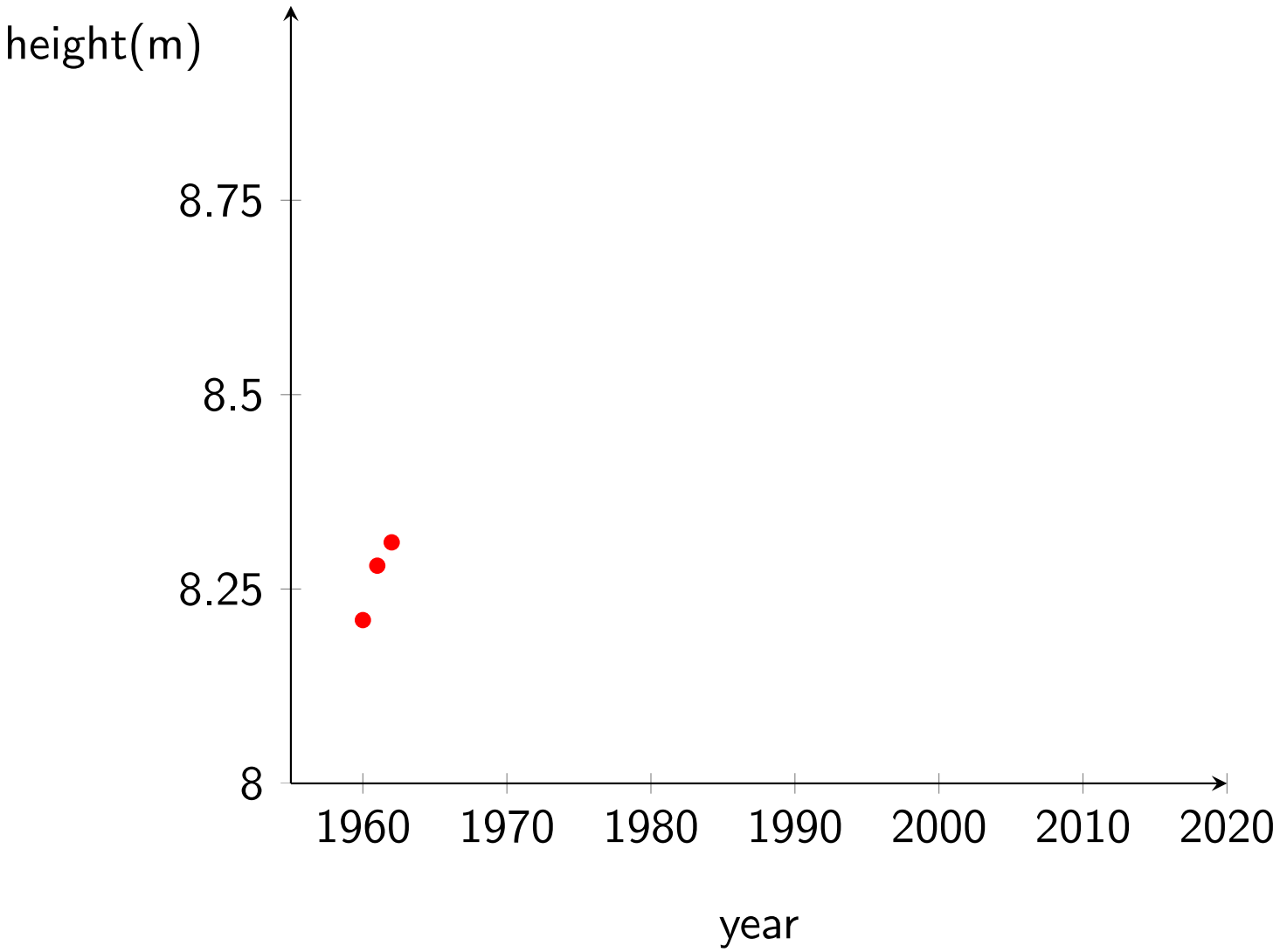
Men's Best Long-Jumps by Year



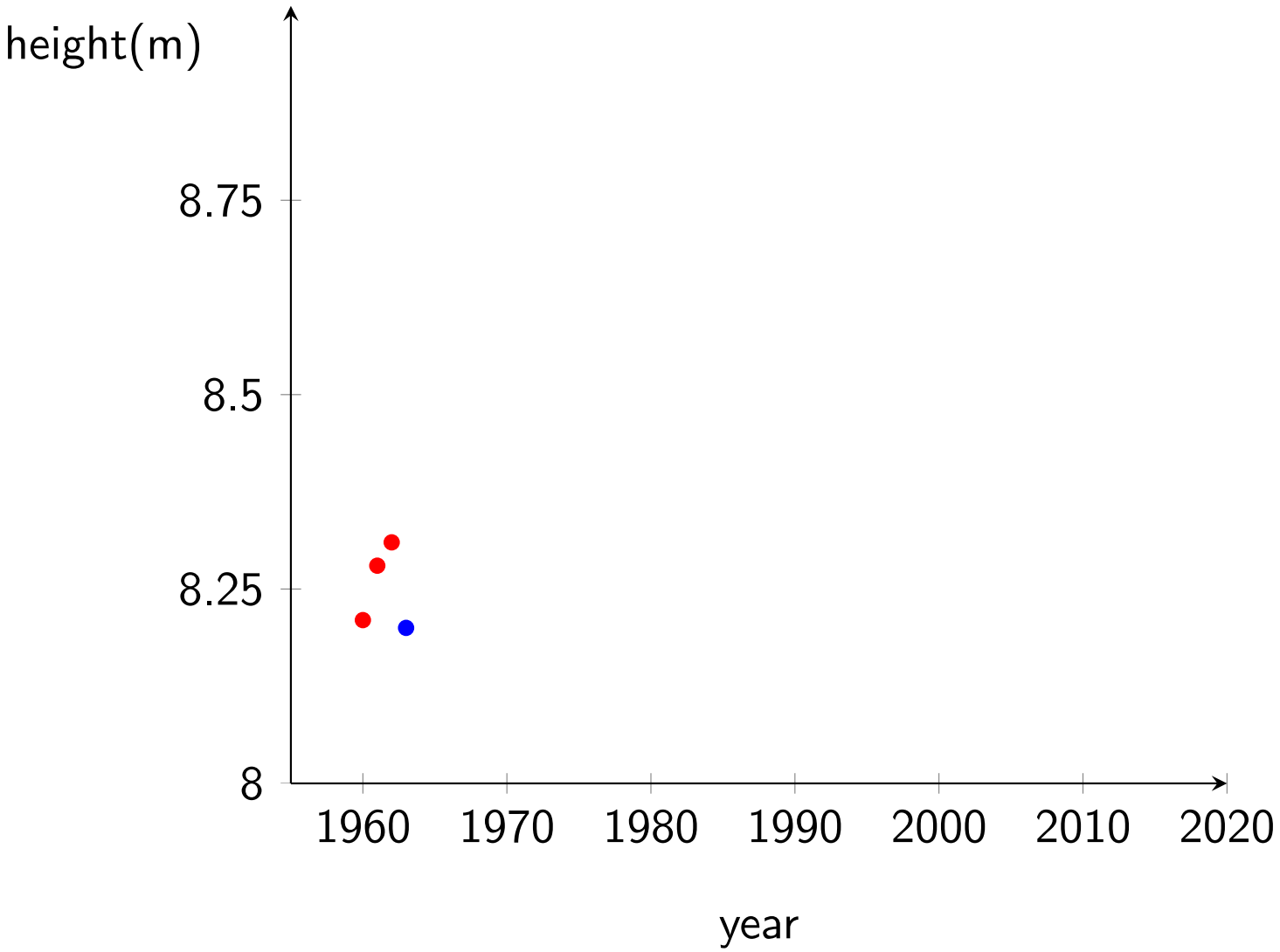
Men's Best Long-Jumps by Year



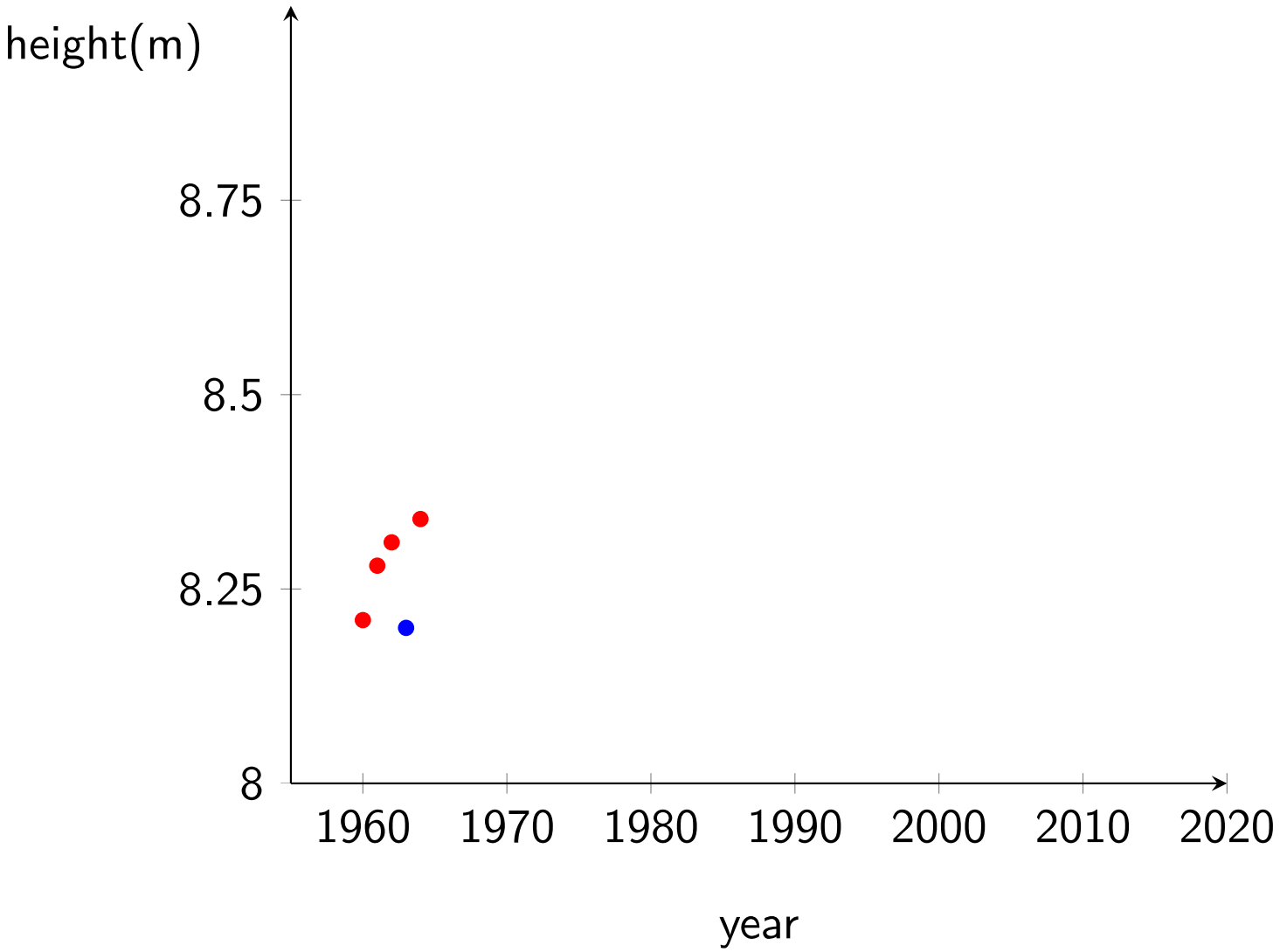
Men's Best Long-Jumps by Year



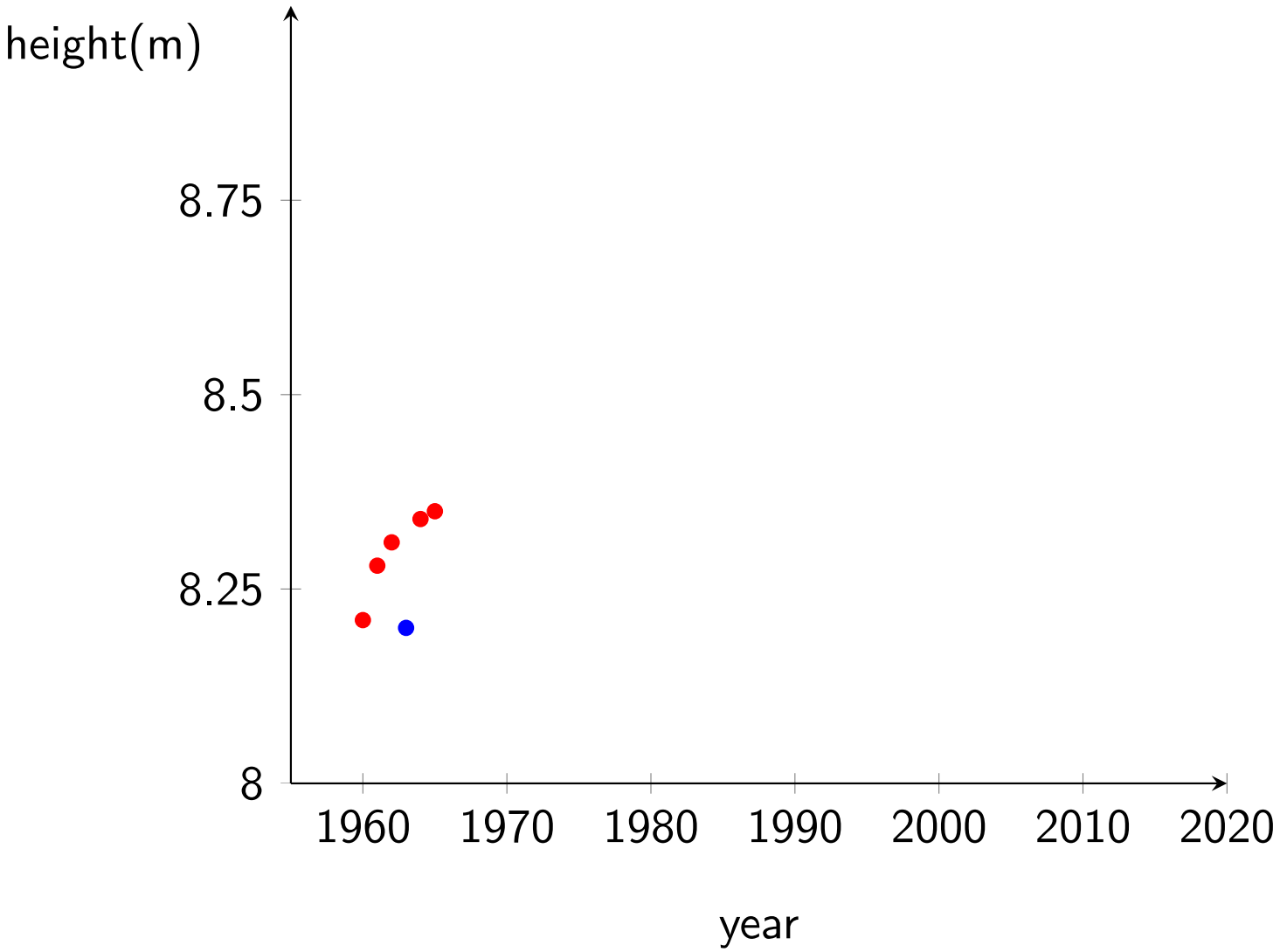
Men's Best Long-Jumps by Year



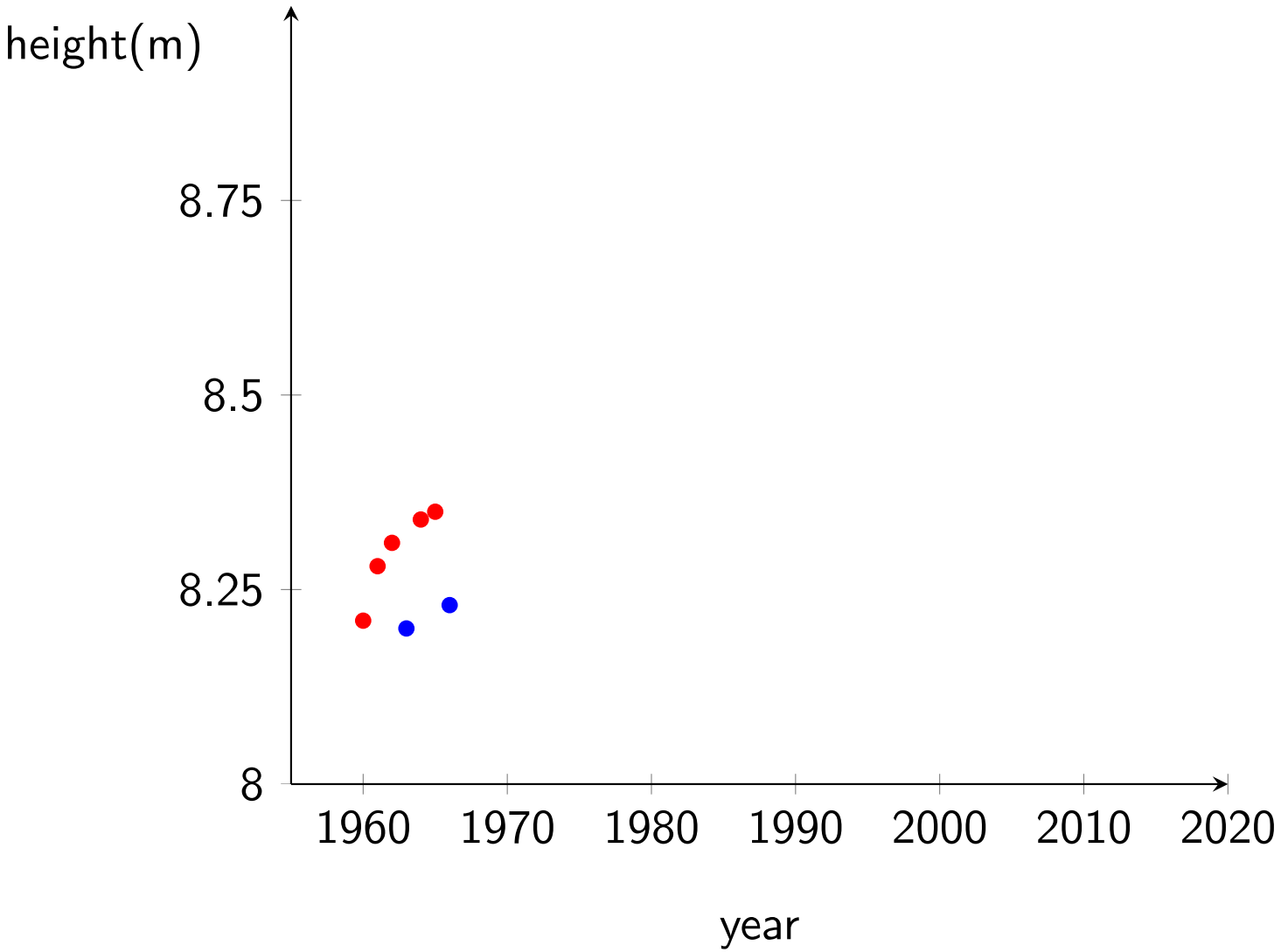
Men's Best Long-Jumps by Year



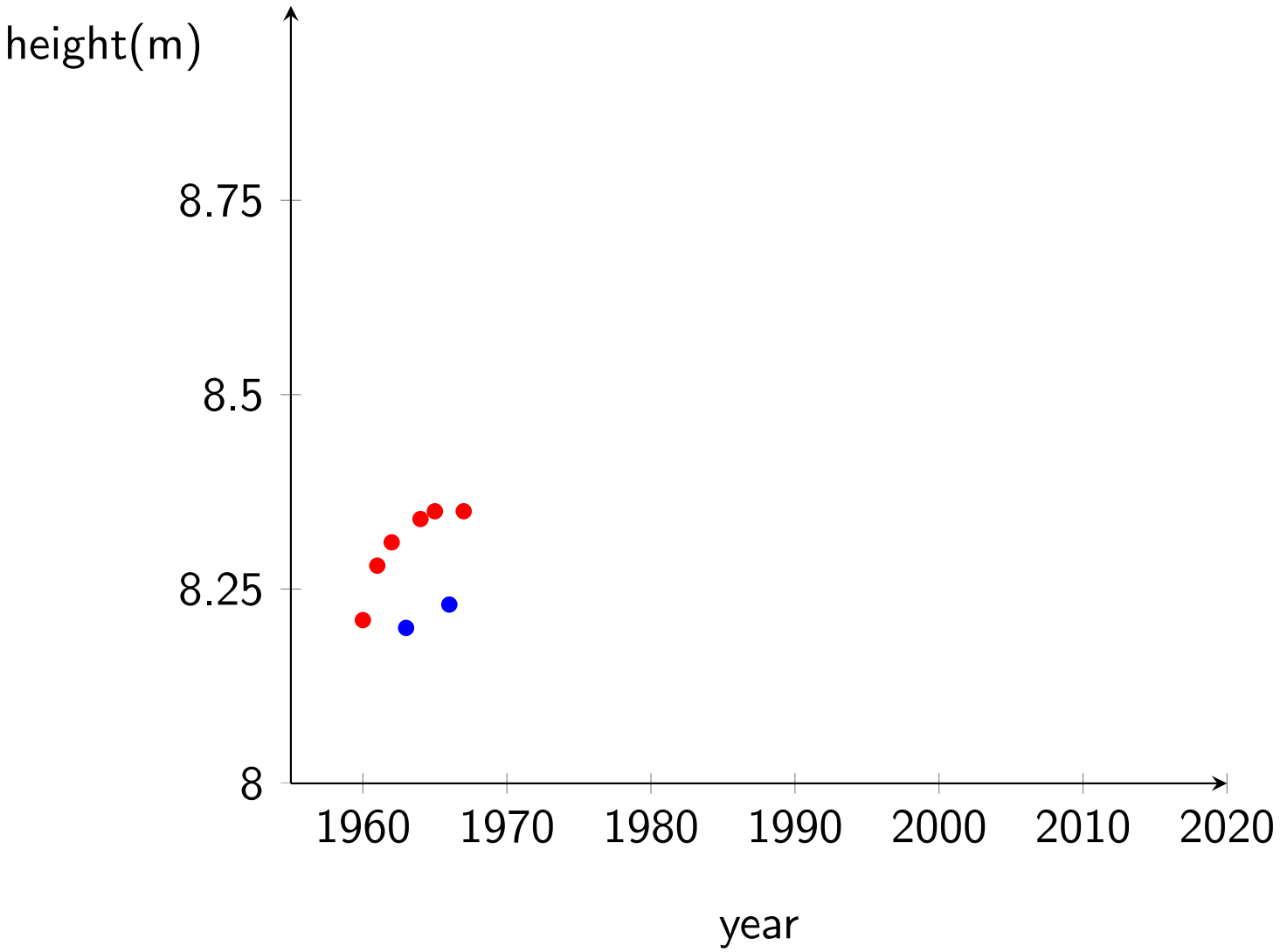
Men's Best Long-Jumps by Year



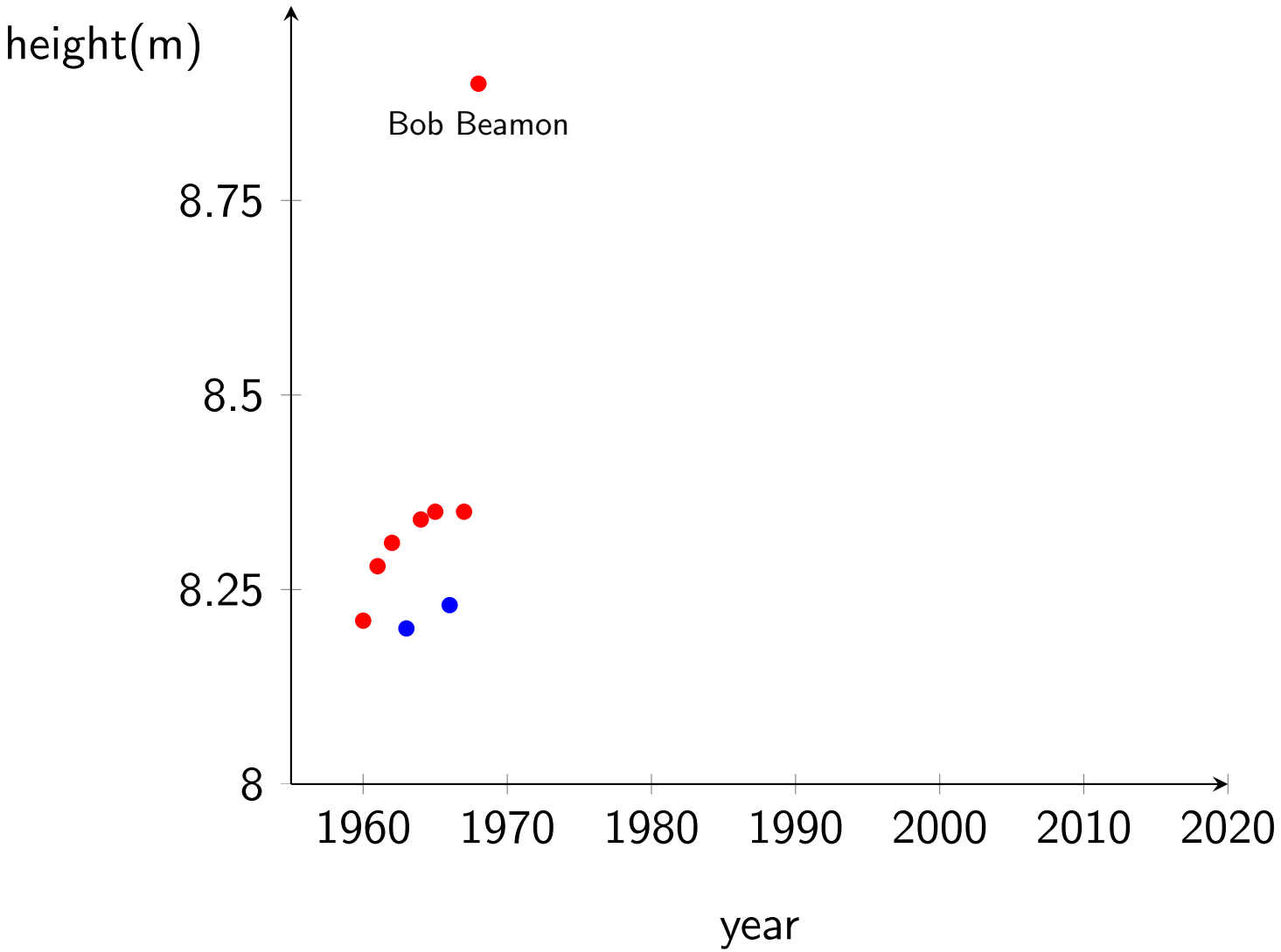
Men's Best Long-Jumps by Year



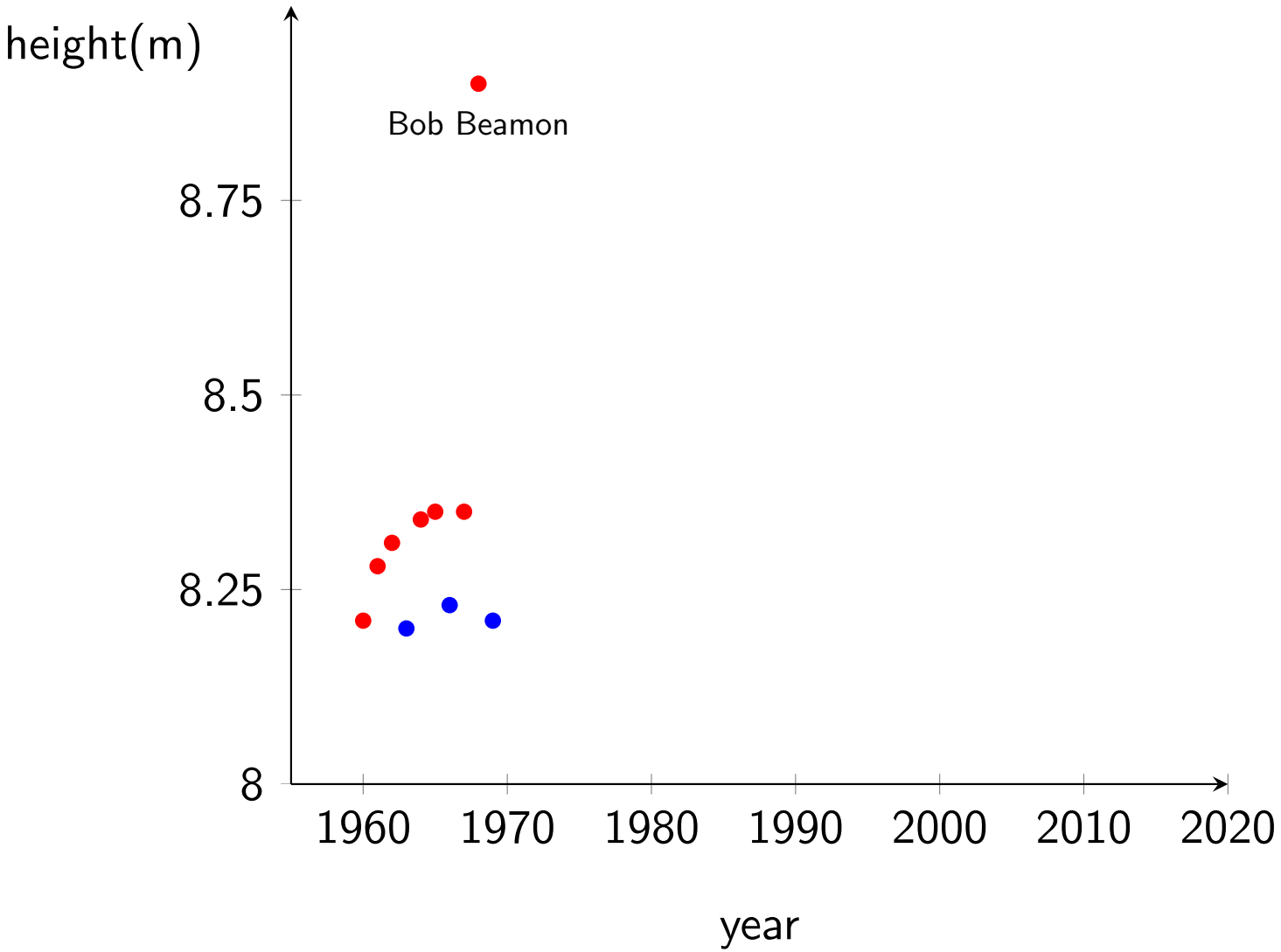
Men's Best Long-Jumps by Year



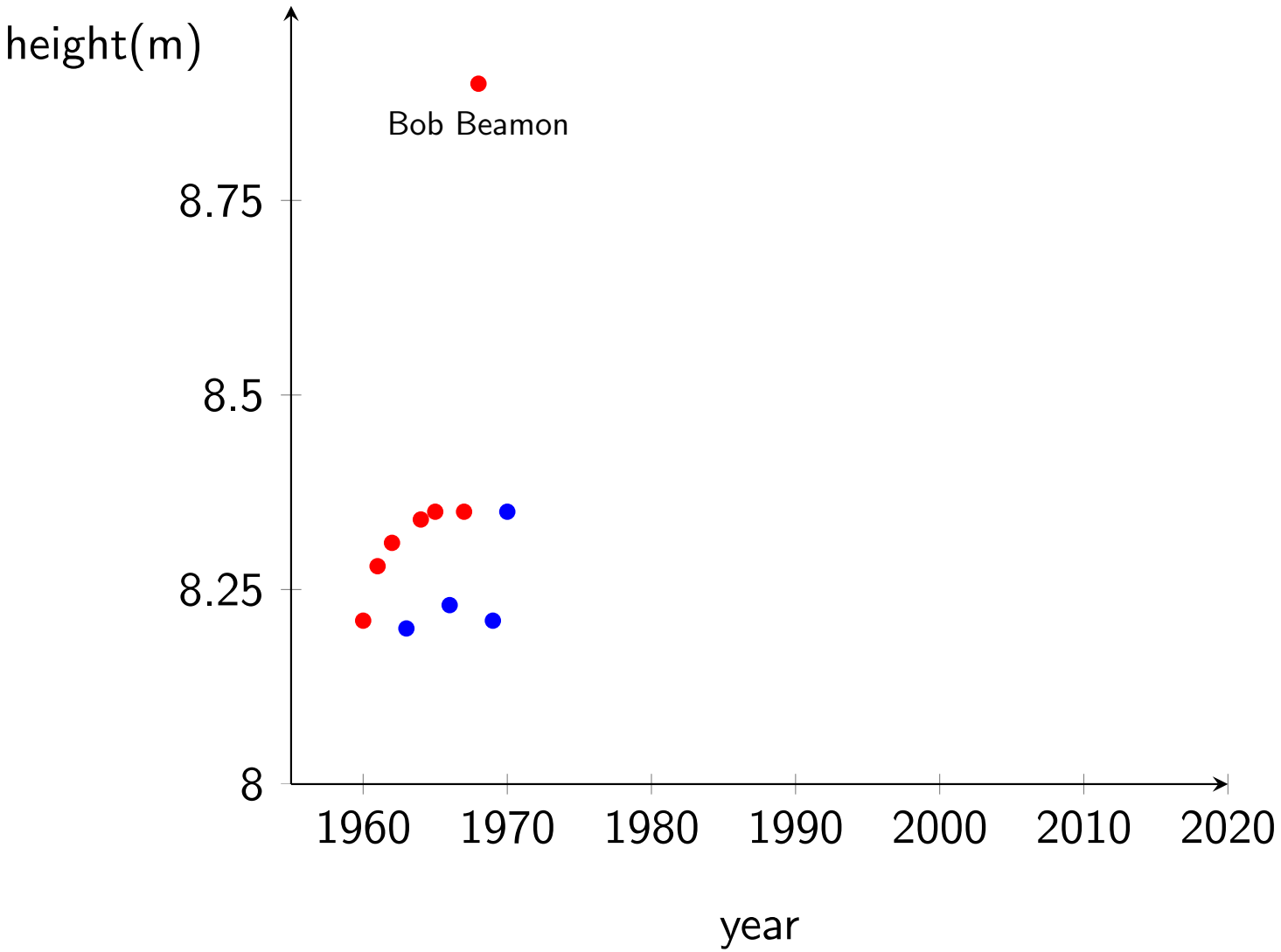
Men's Best Long-Jumps by Year



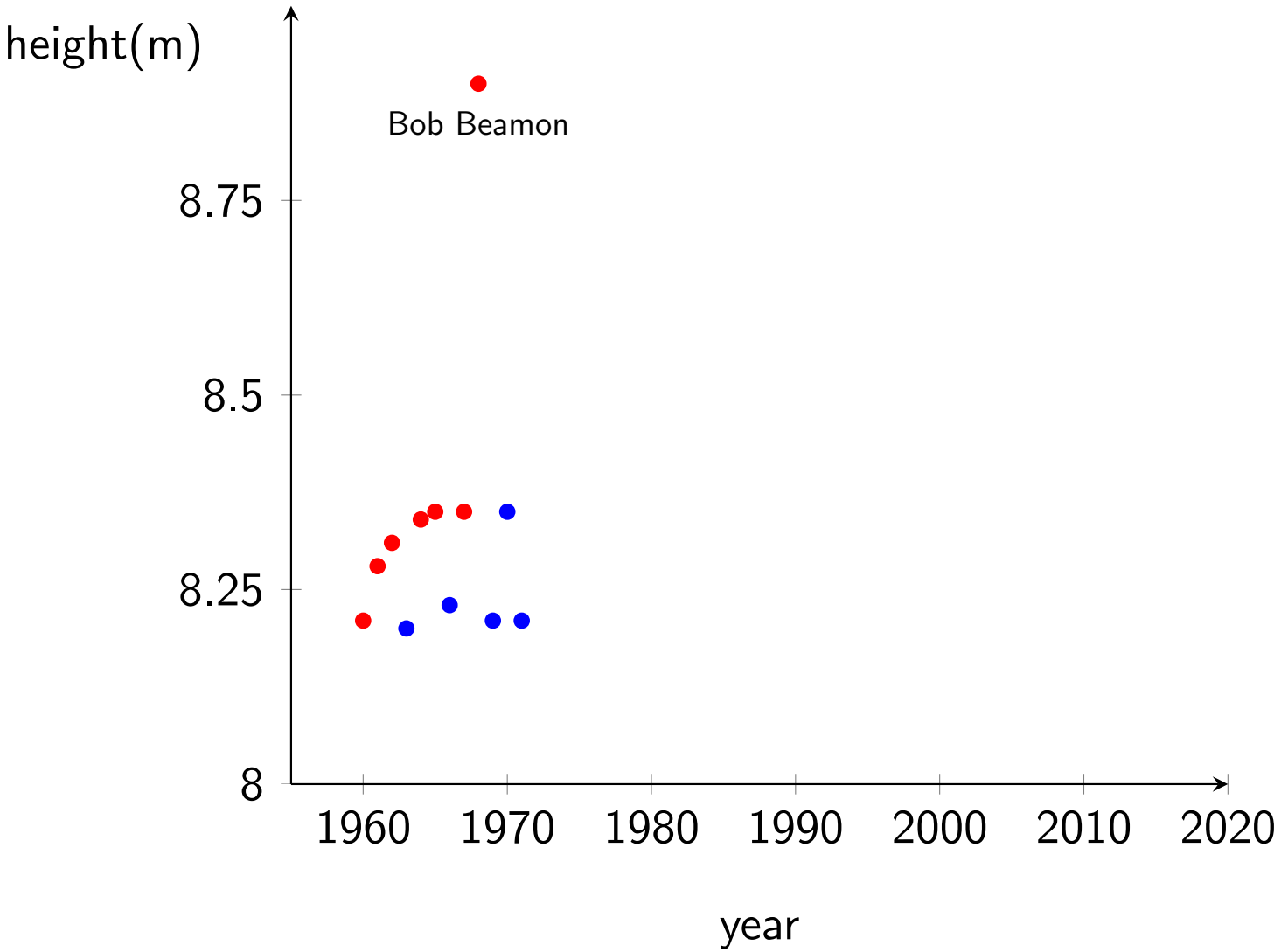
Men's Best Long-Jumps by Year



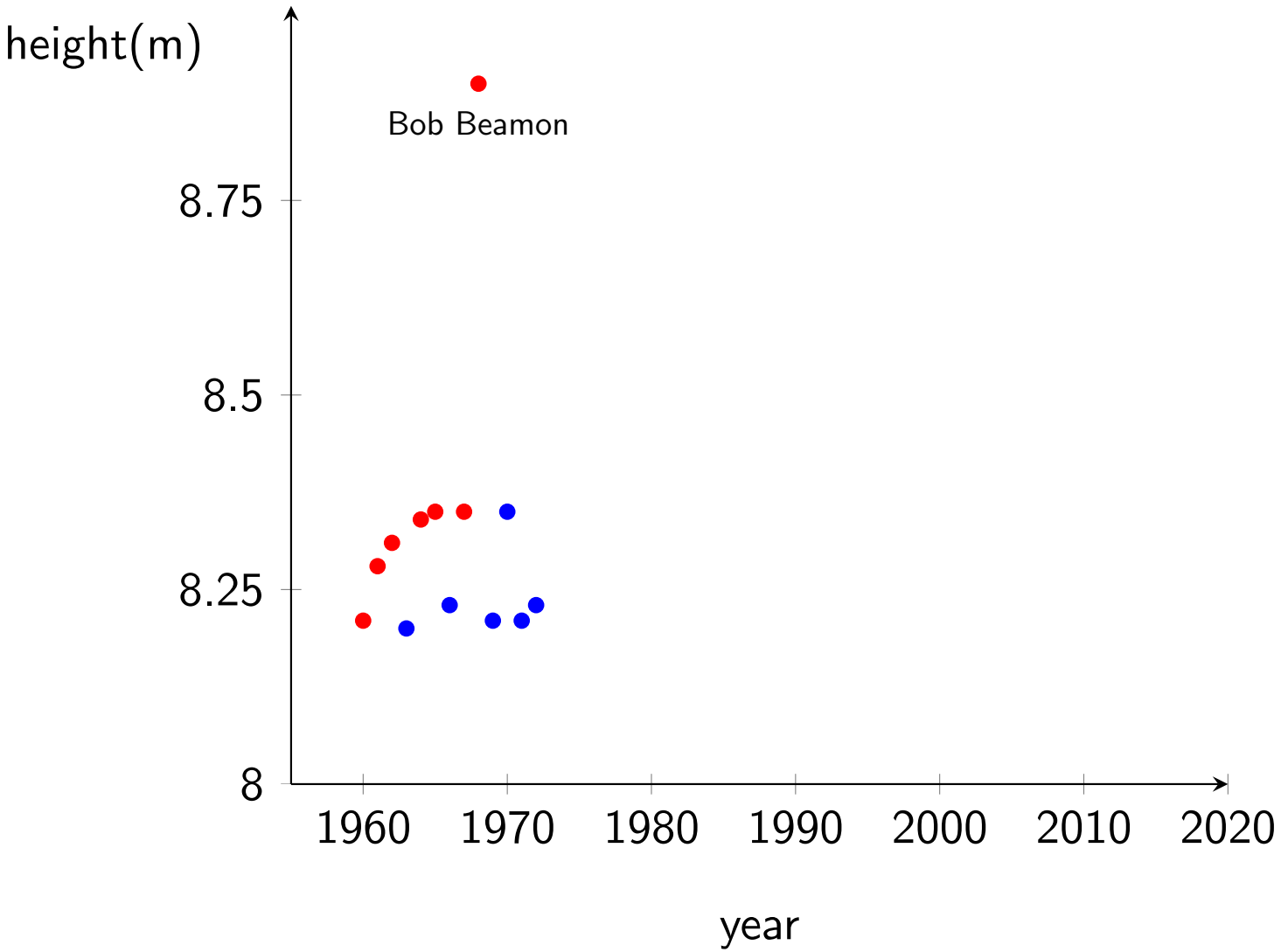
Men's Best Long-Jumps by Year



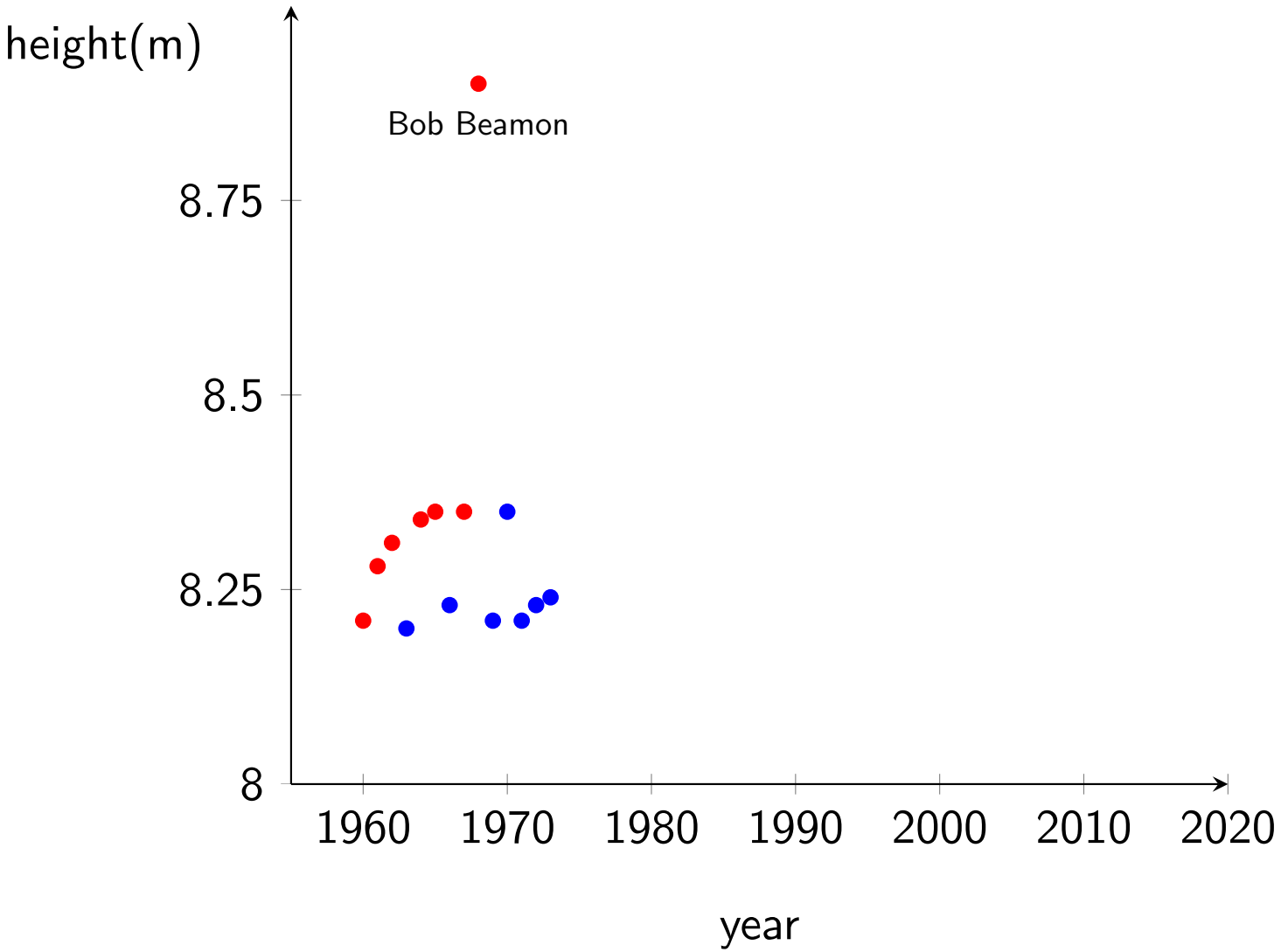
Men's Best Long-Jumps by Year



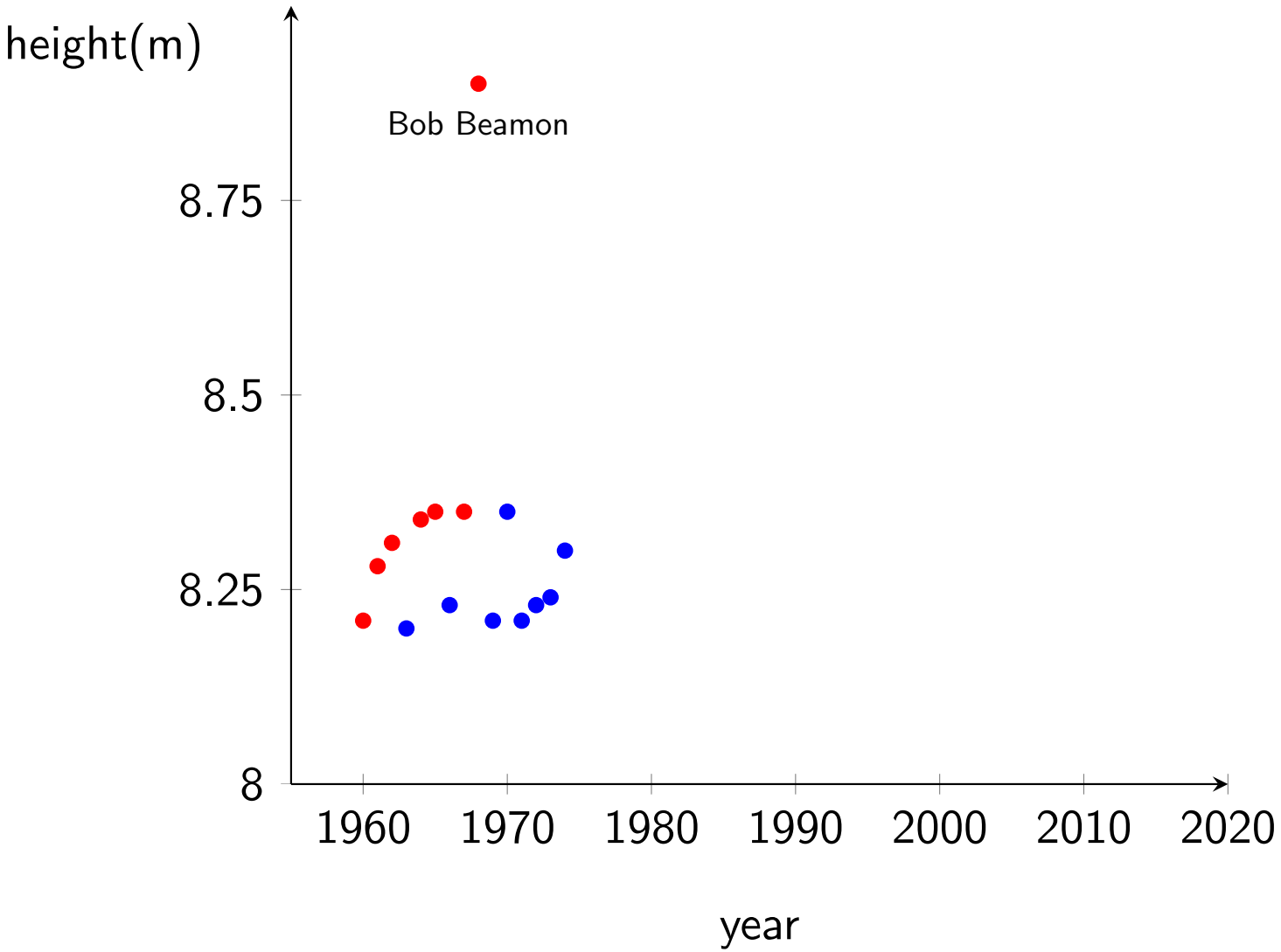
Men's Best Long-Jumps by Year



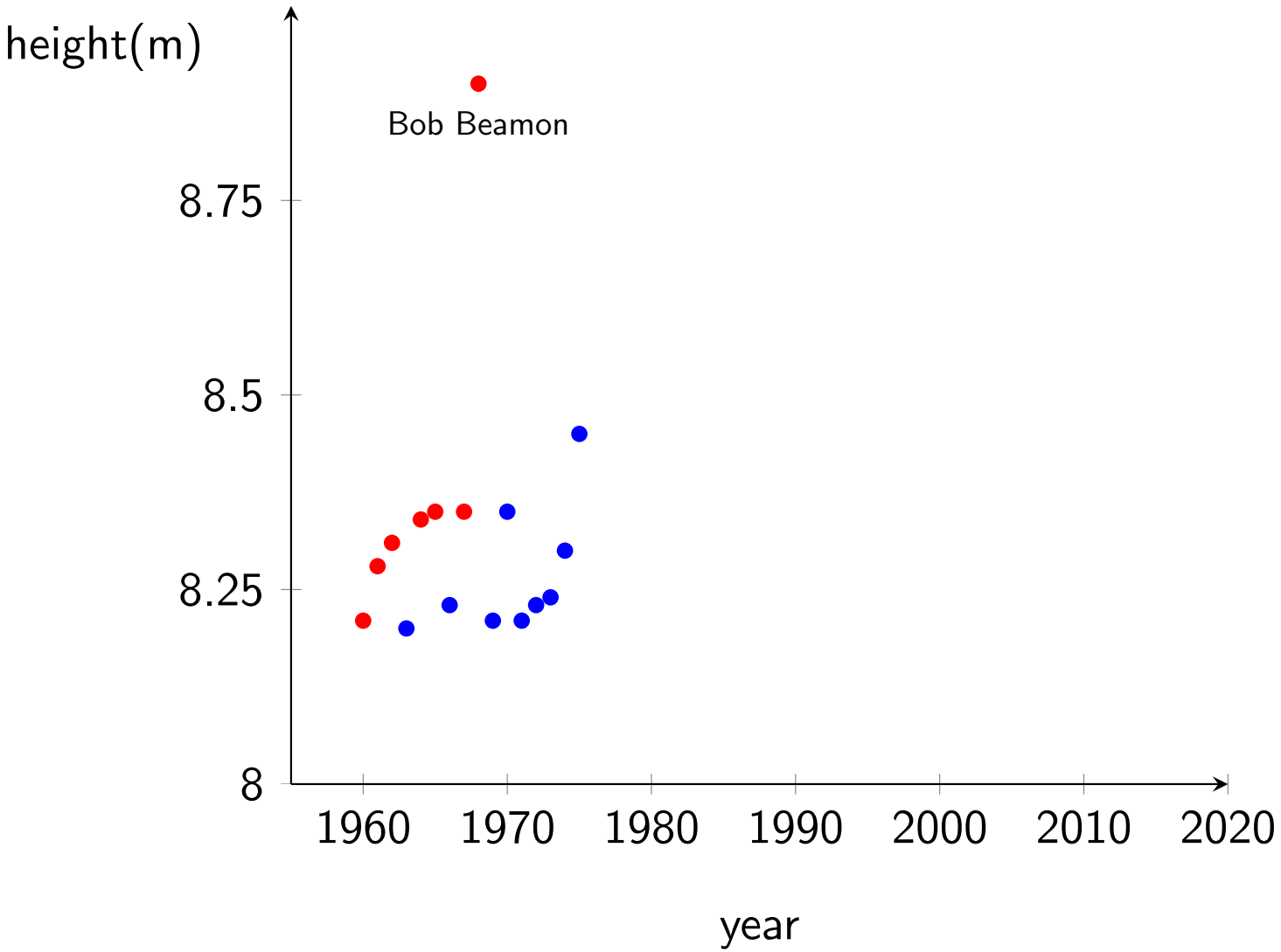
Men's Best Long-Jumps by Year



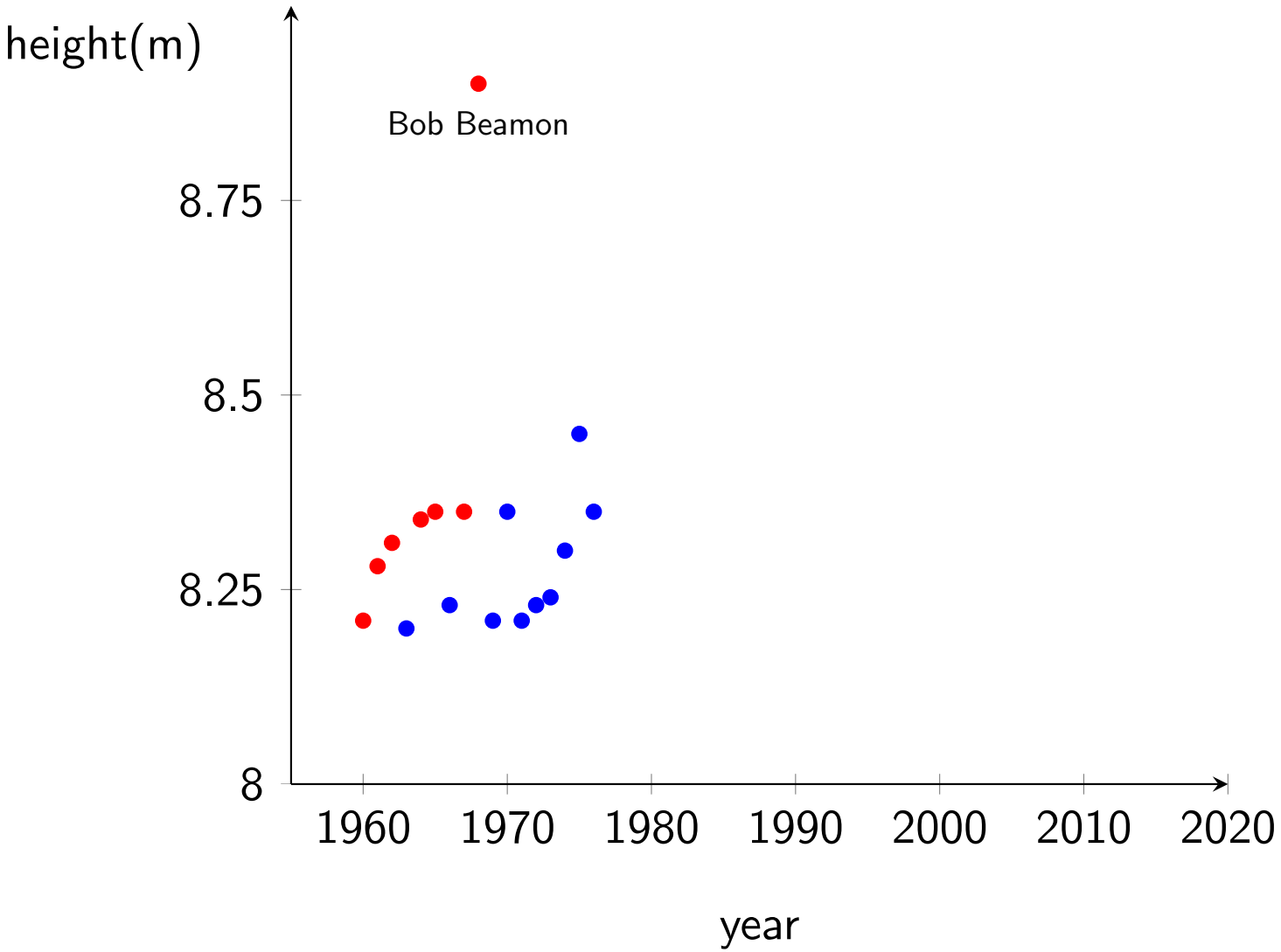
Men's Best Long-Jumps by Year



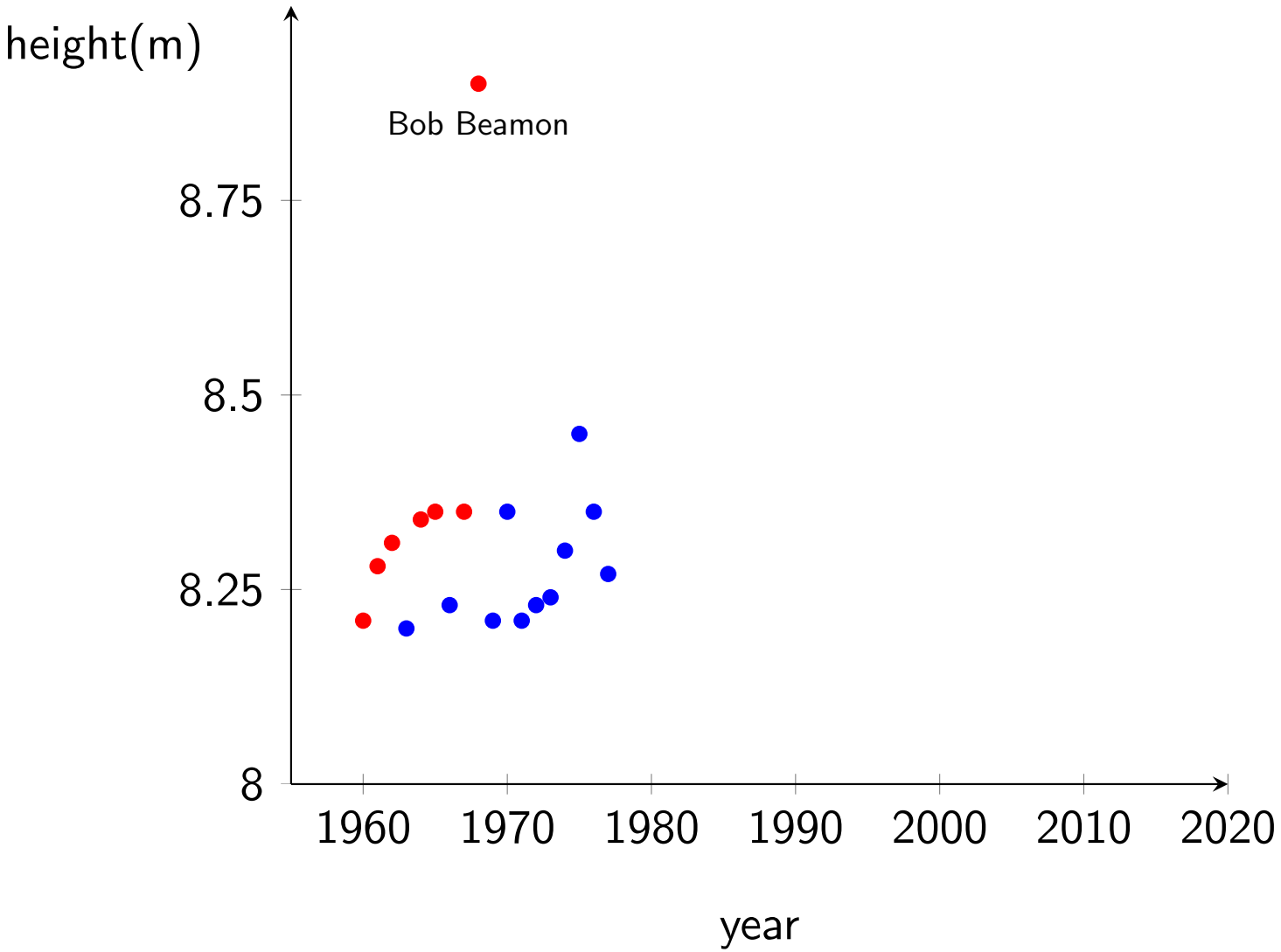
Men's Best Long-Jumps by Year



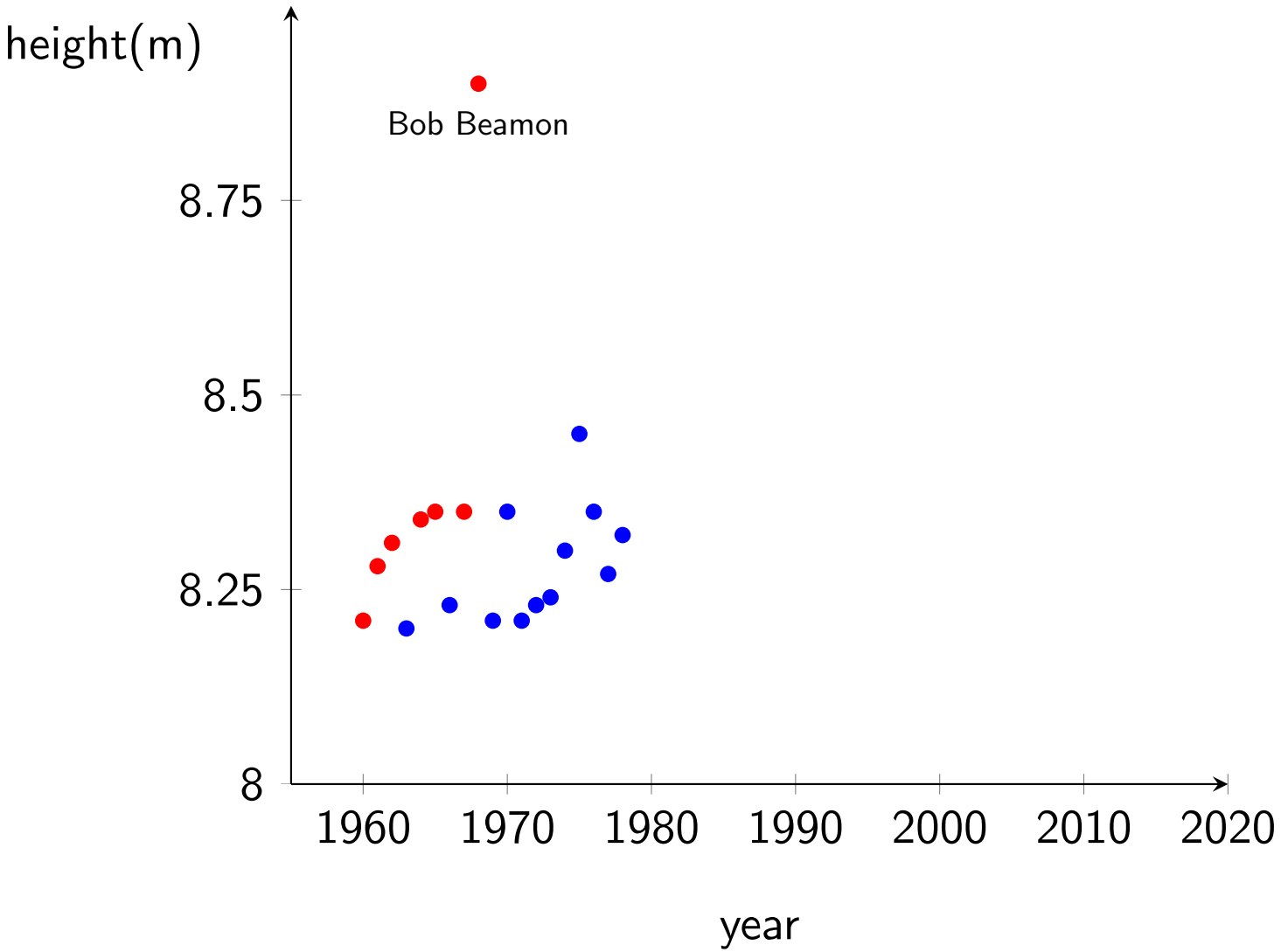
Men's Best Long-Jumps by Year



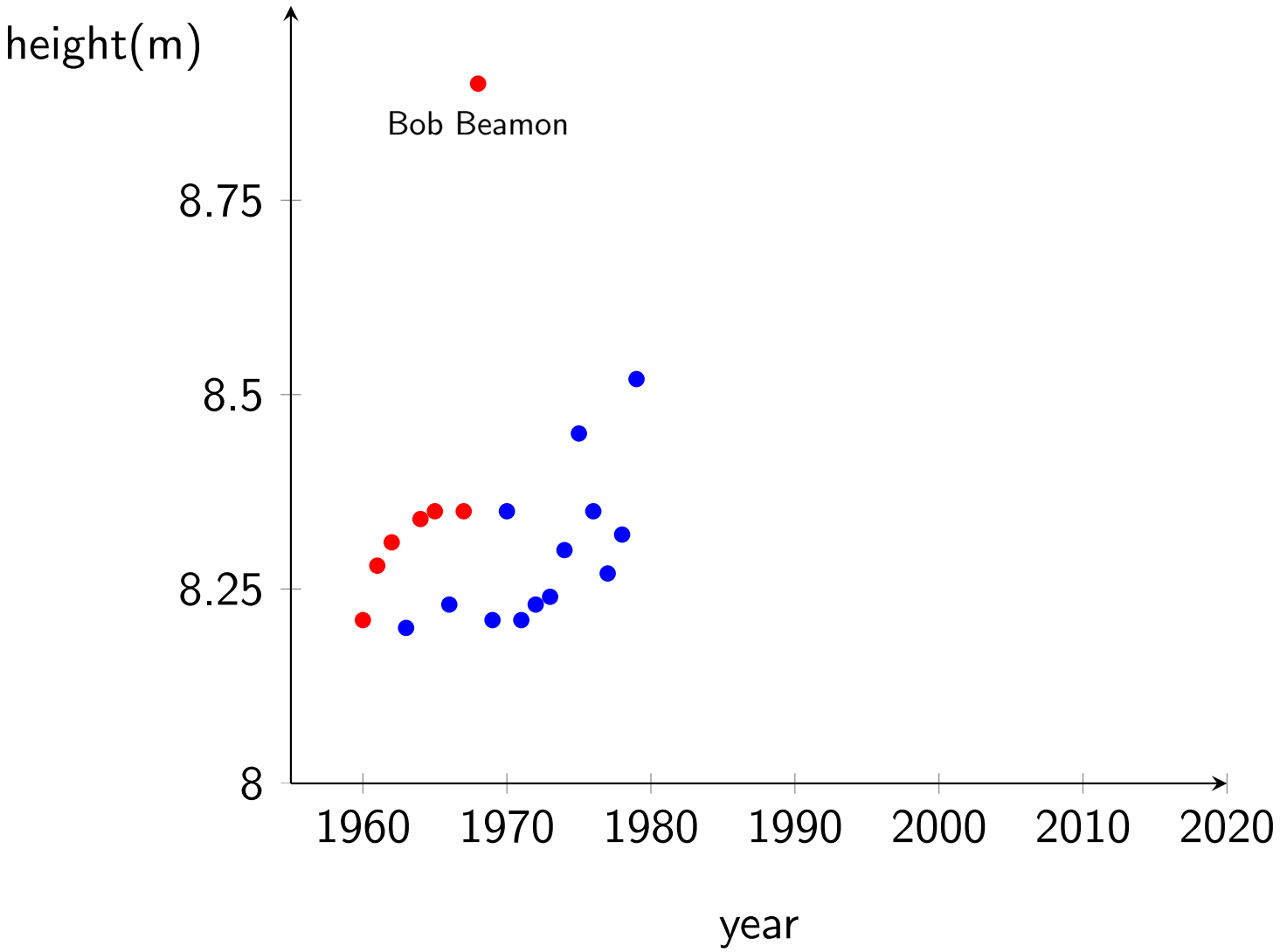
Men's Best Long-Jumps by Year



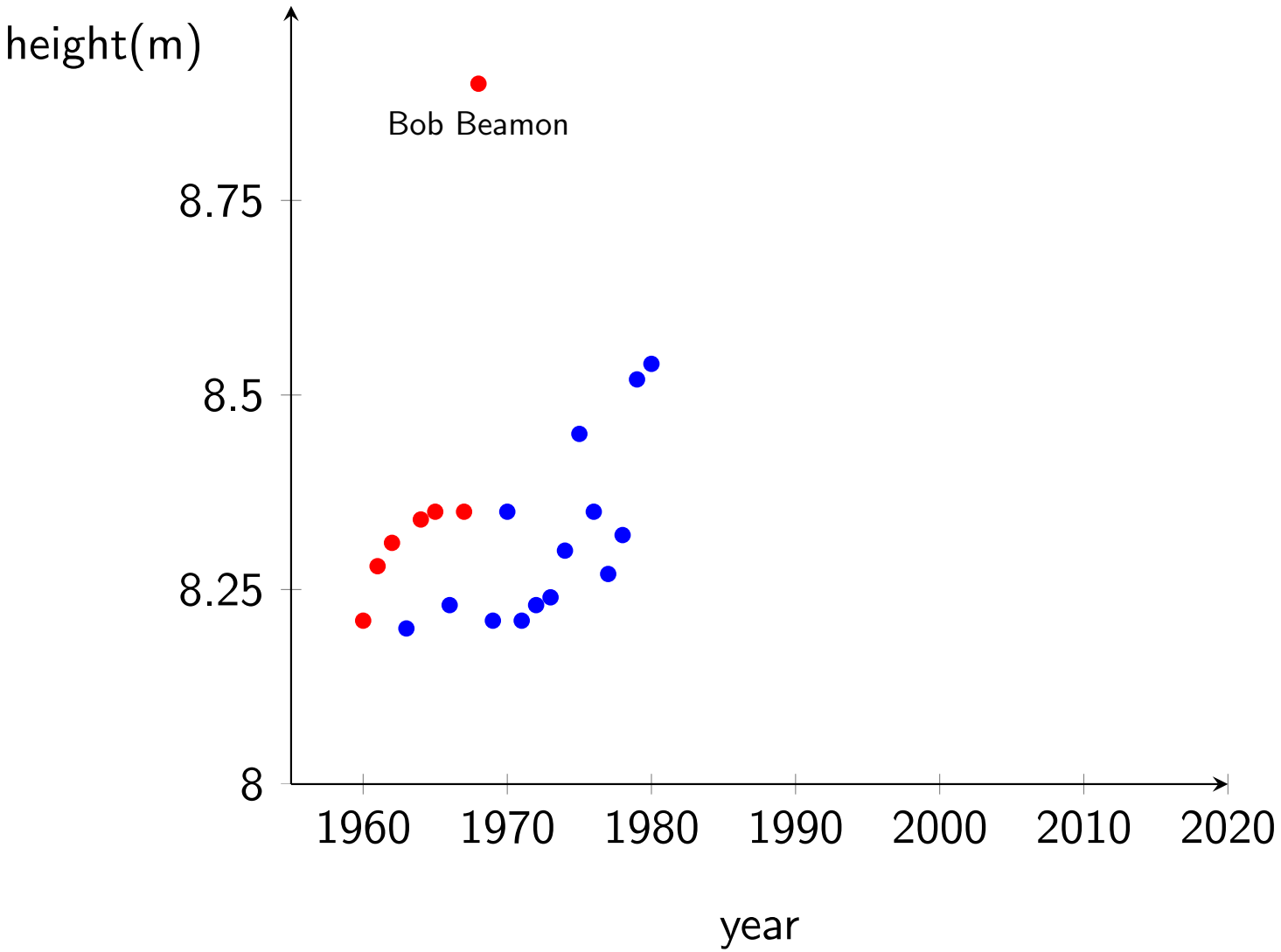
Men's Best Long-Jumps by Year



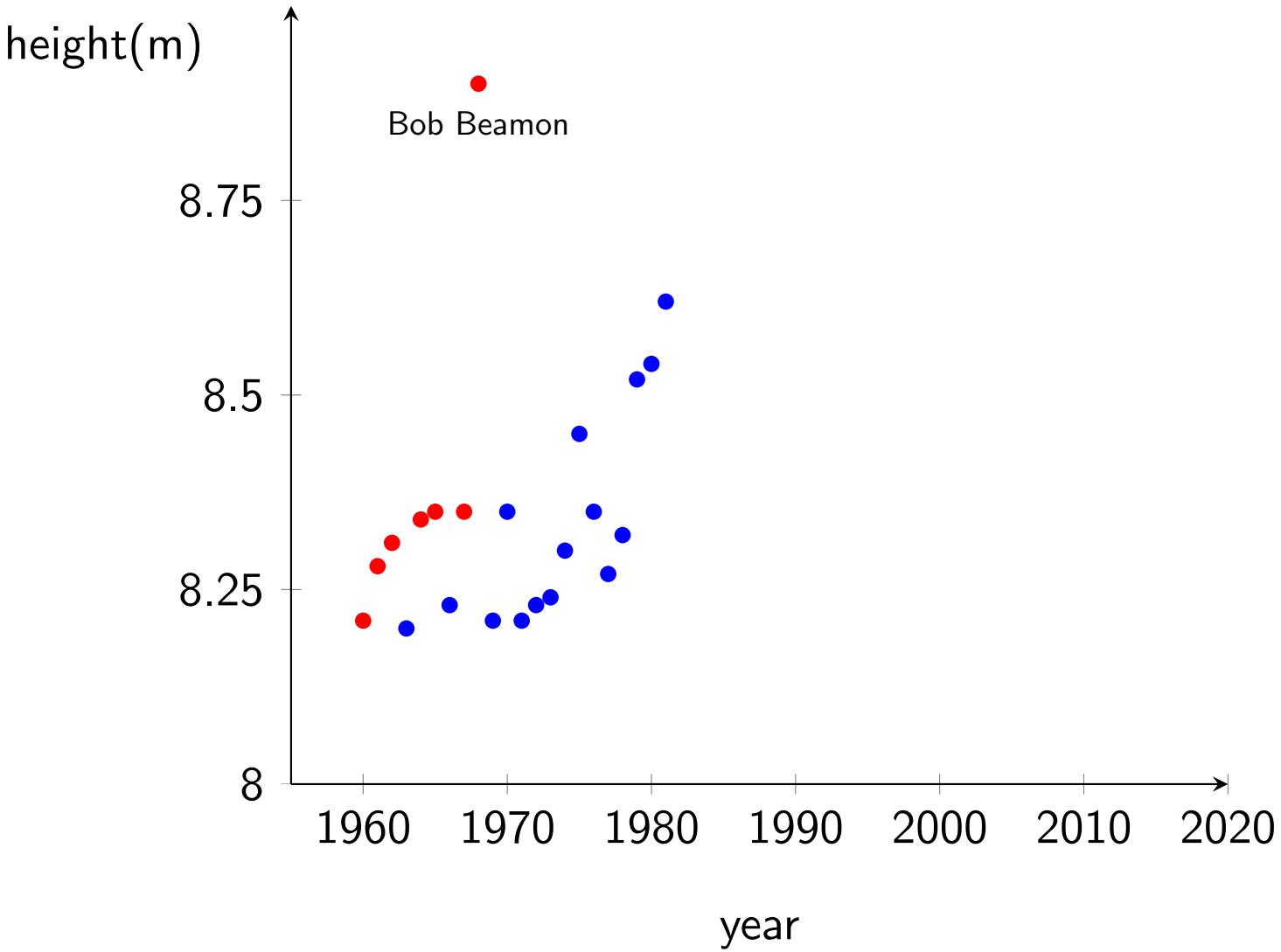
Men's Best Long-Jumps by Year



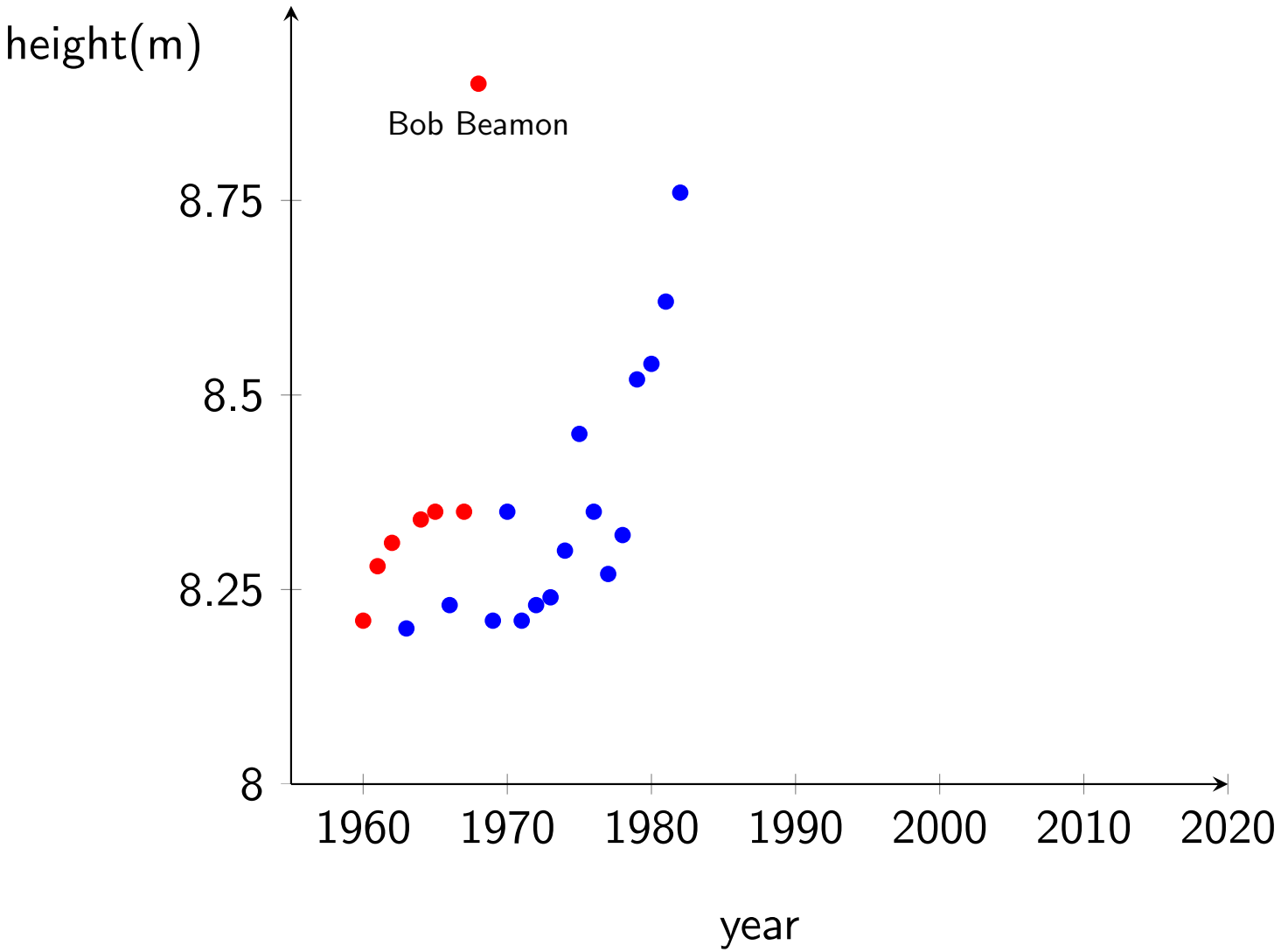
Men's Best Long-Jumps by Year



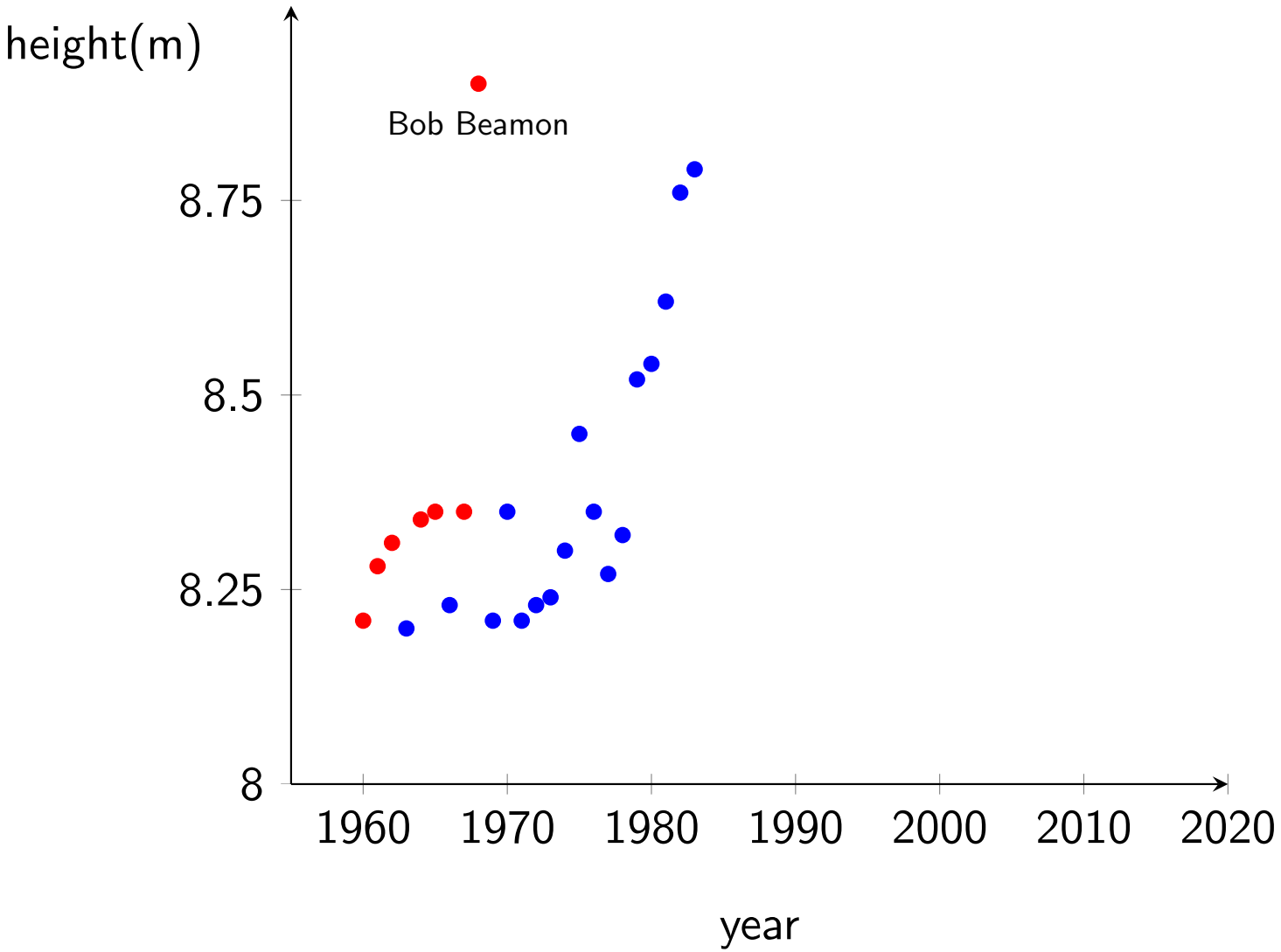
Men's Best Long-Jumps by Year



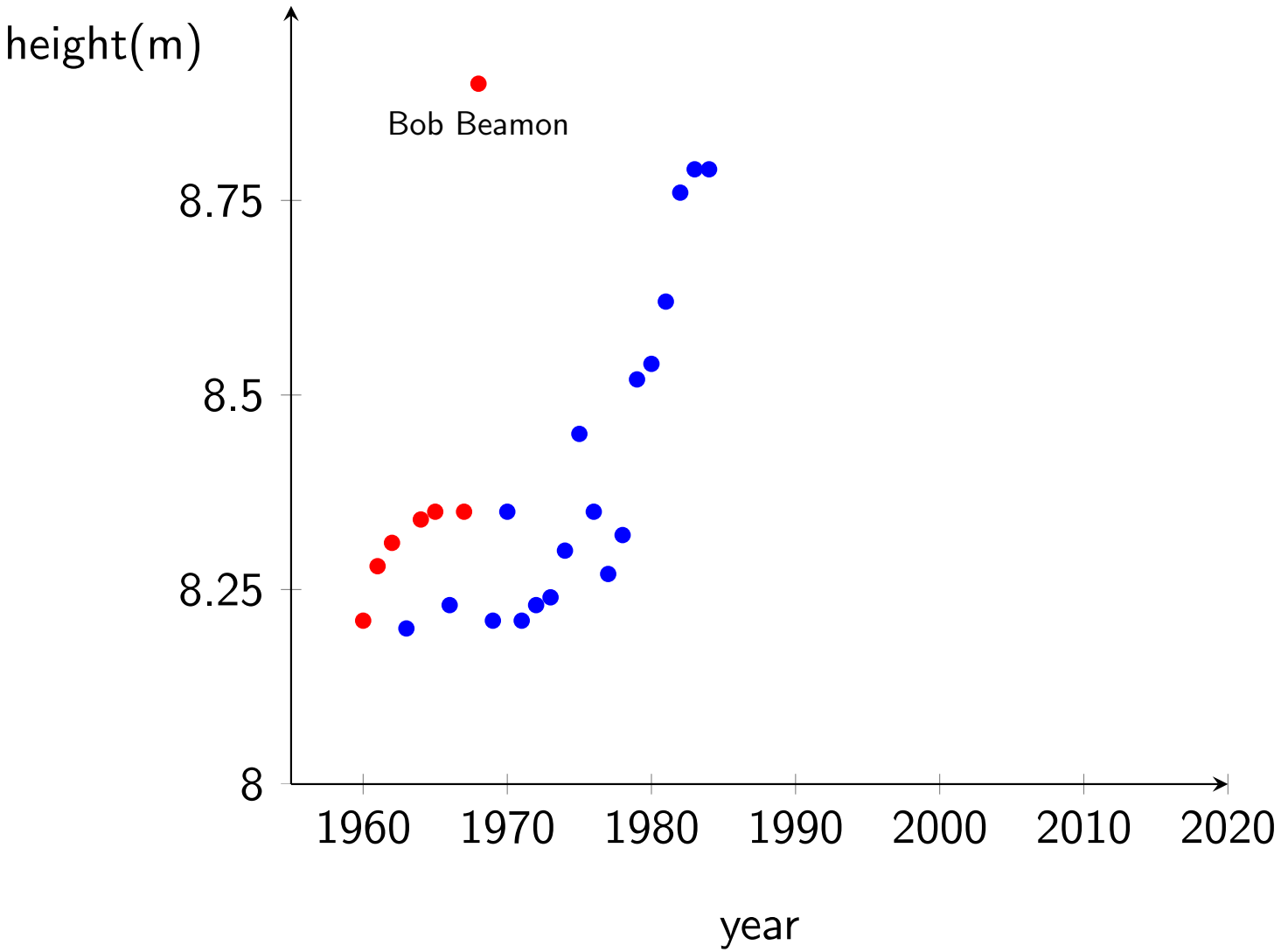
Men's Best Long-Jumps by Year



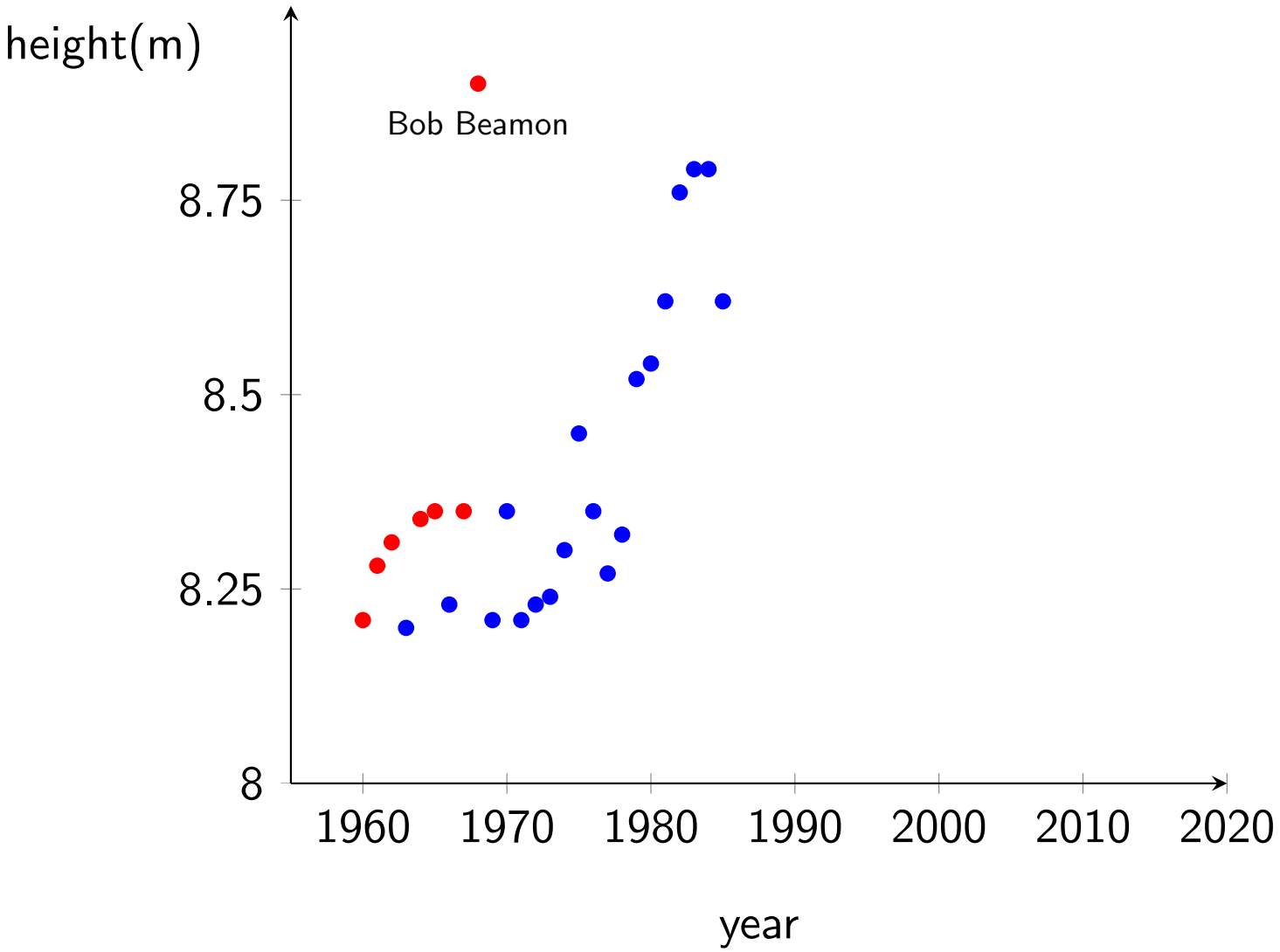
Men's Best Long-Jumps by Year



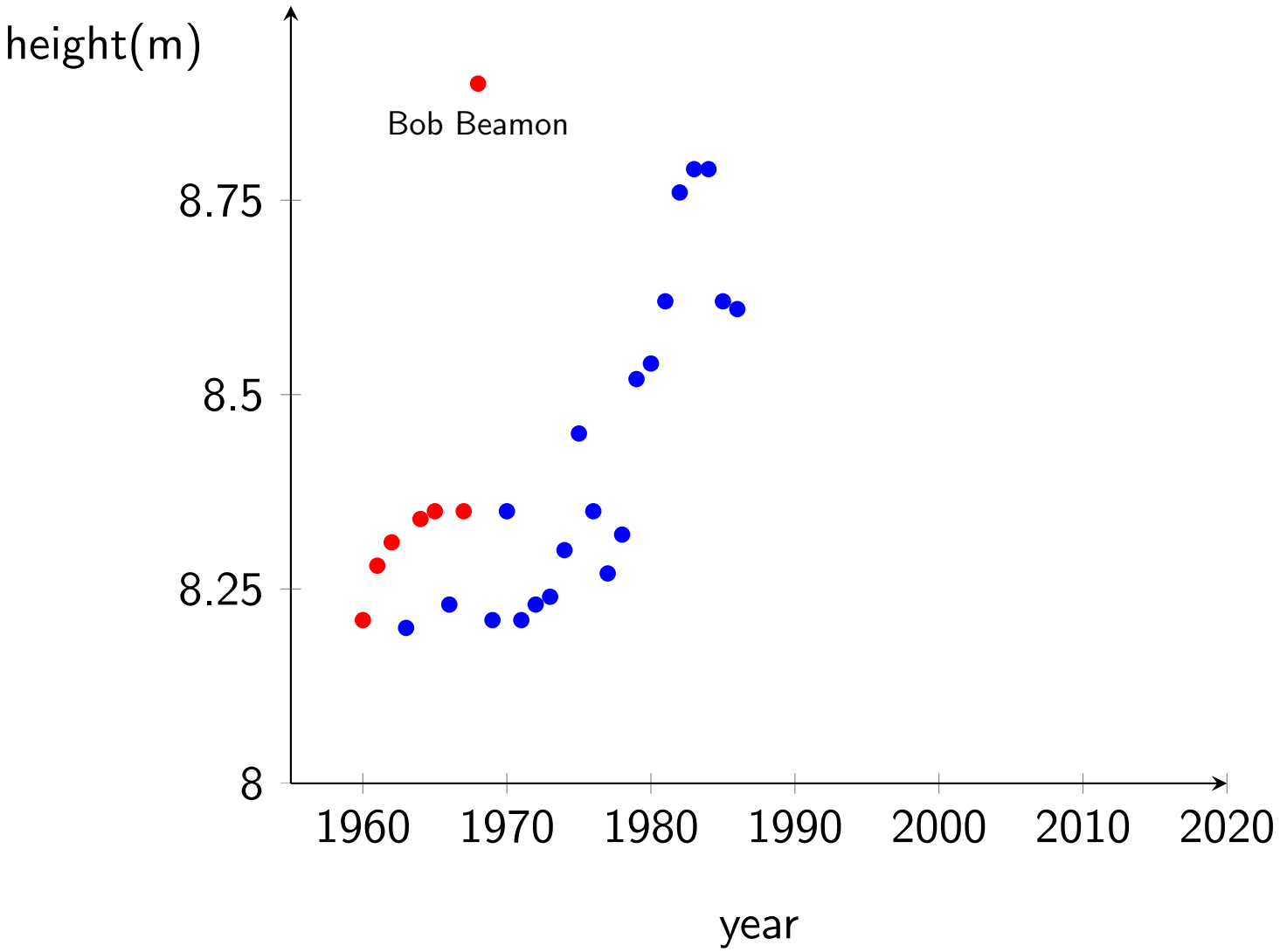
Men's Best Long-Jumps by Year



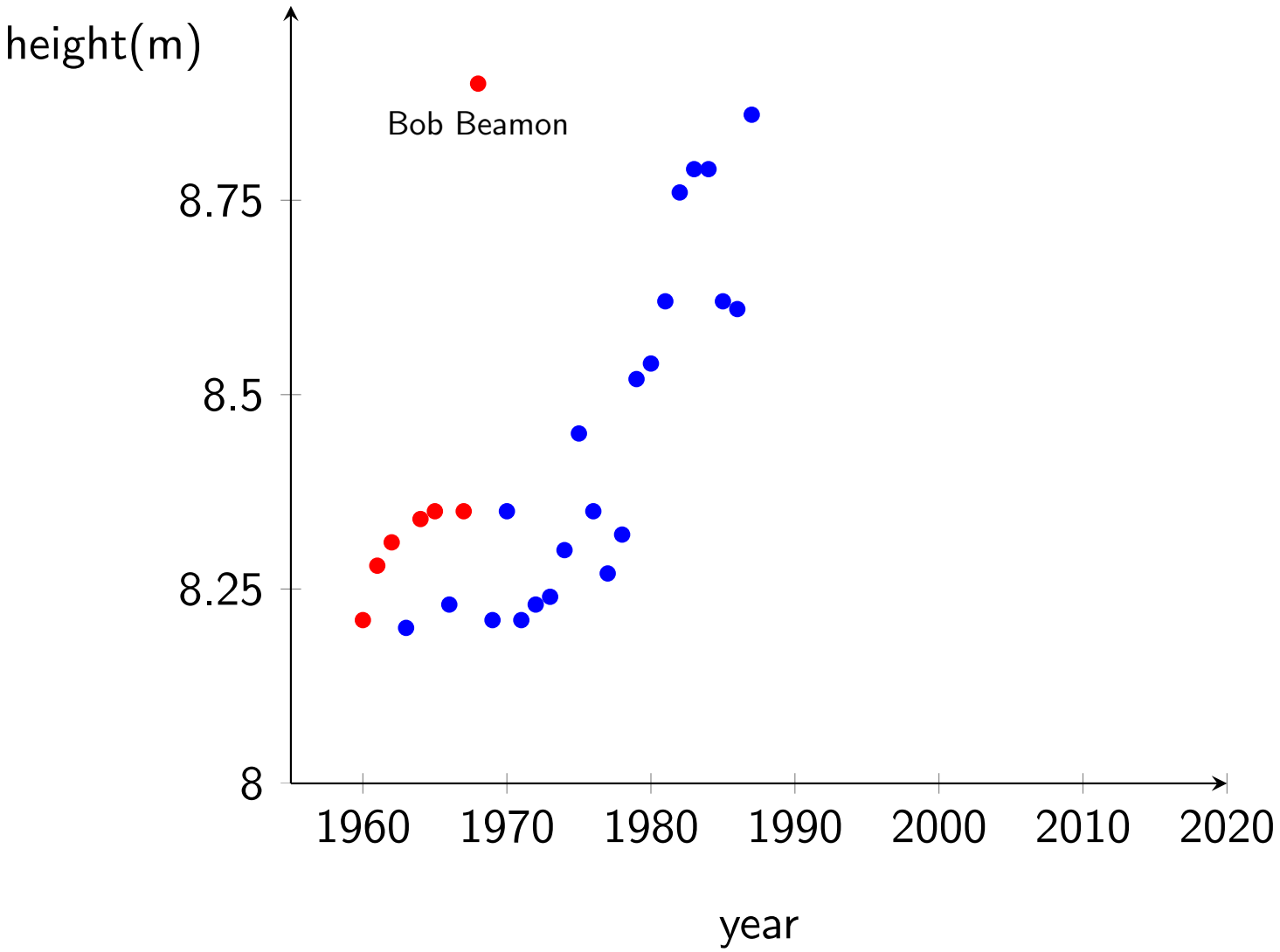
Men's Best Long-Jumps by Year



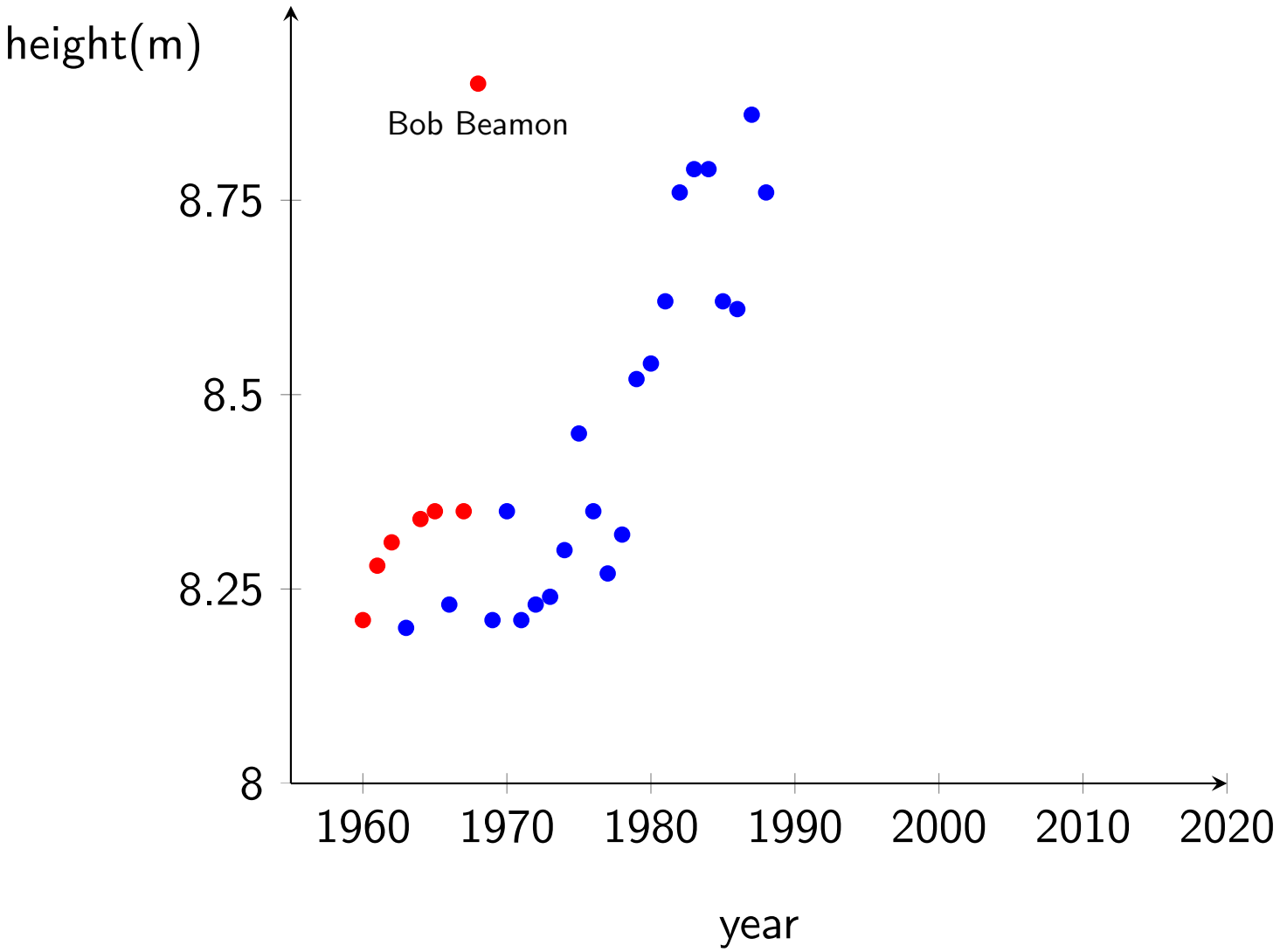
Men's Best Long-Jumps by Year



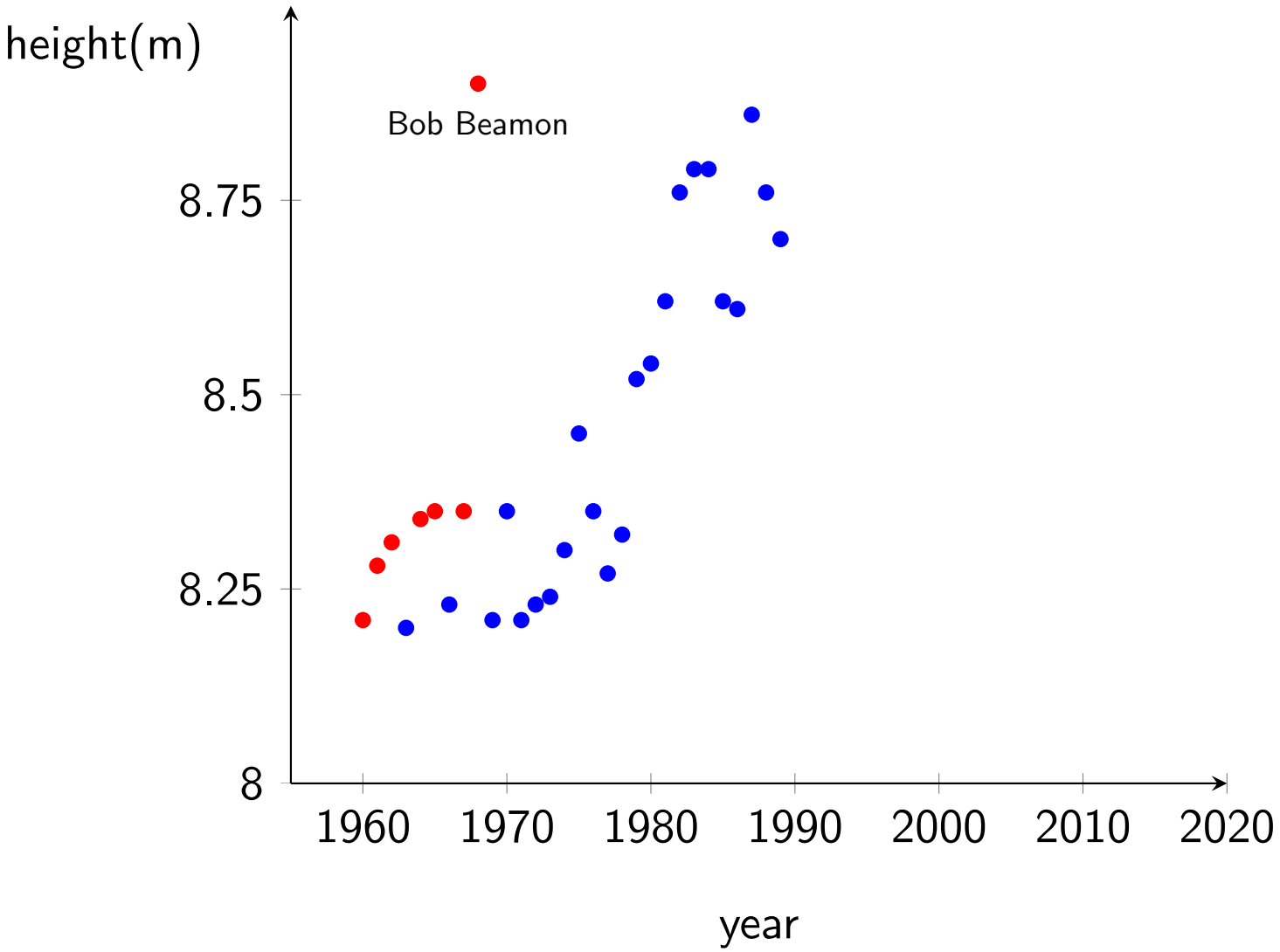
Men's Best Long-Jumps by Year



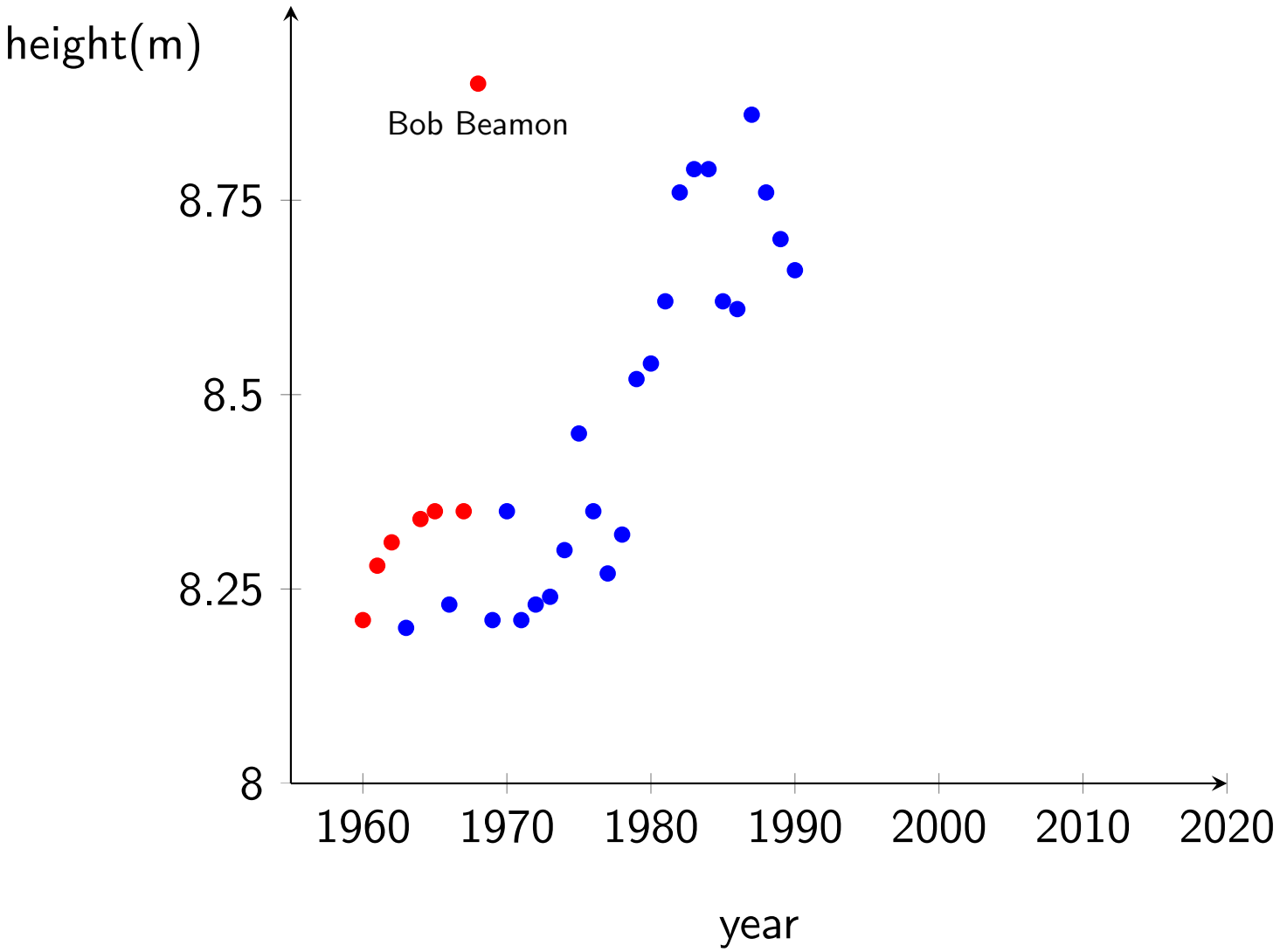
Men's Best Long-Jumps by Year



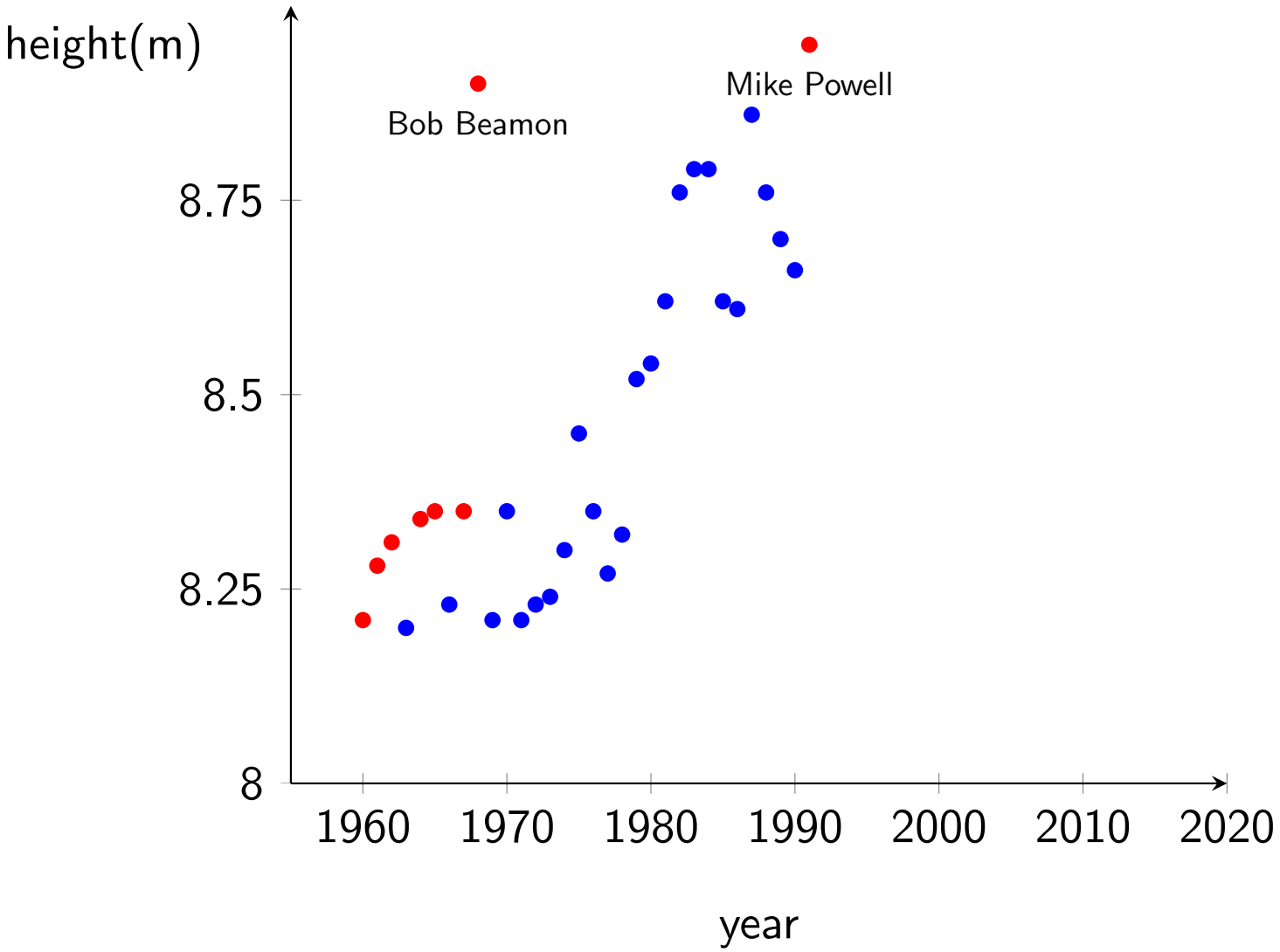
Men's Best Long-Jumps by Year



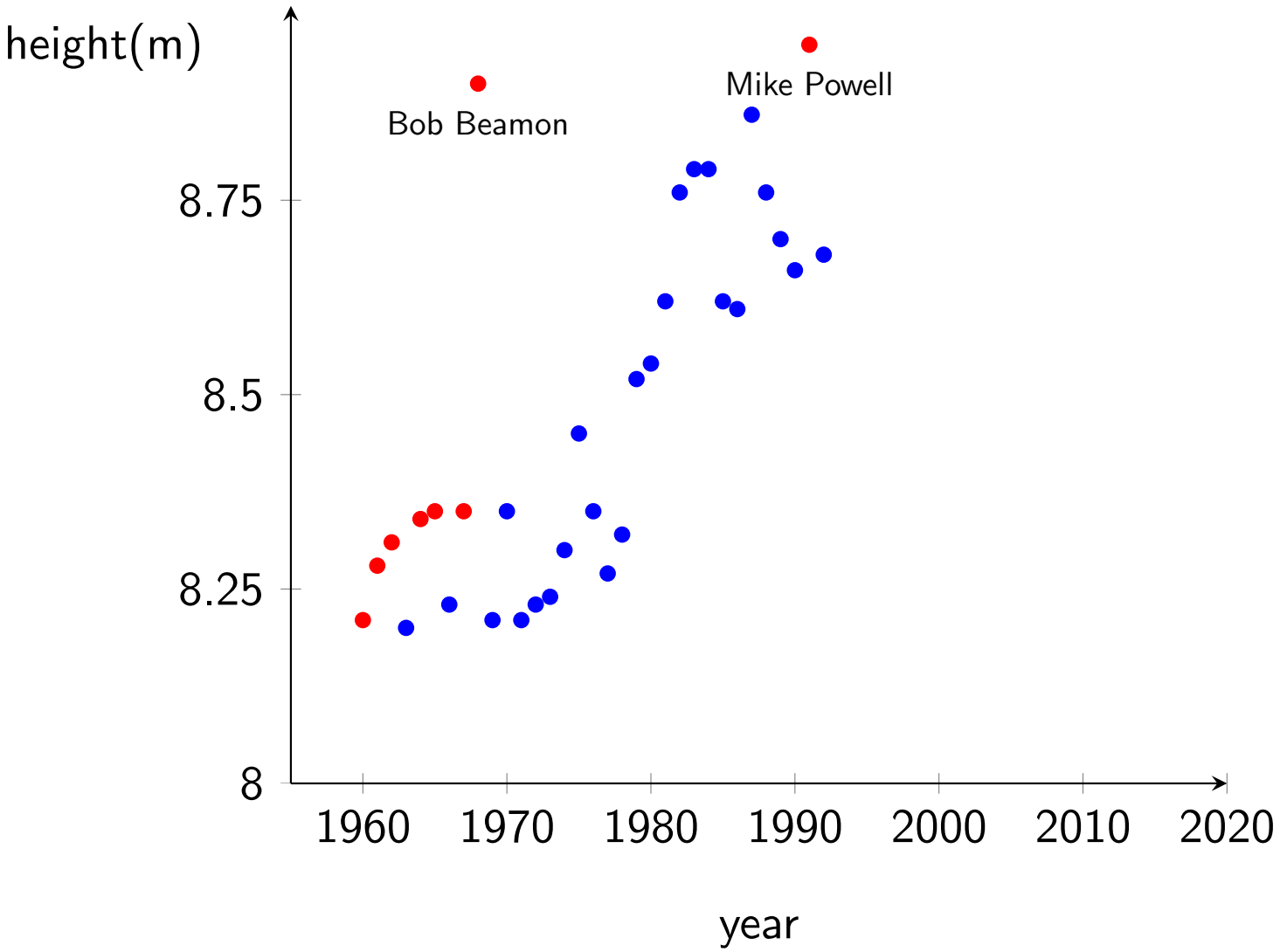
Men's Best Long-Jumps by Year



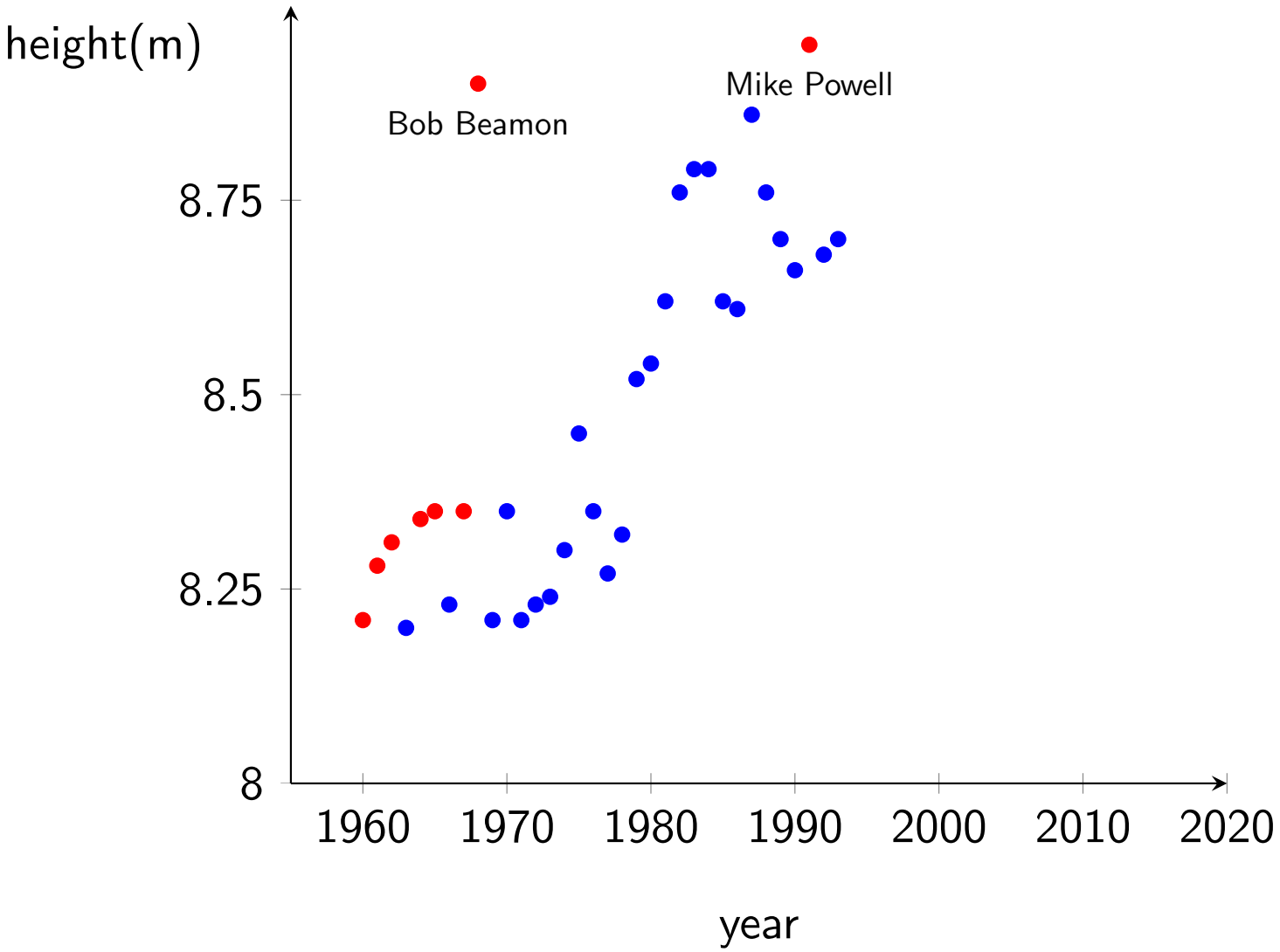
Men's Best Long-Jumps by Year



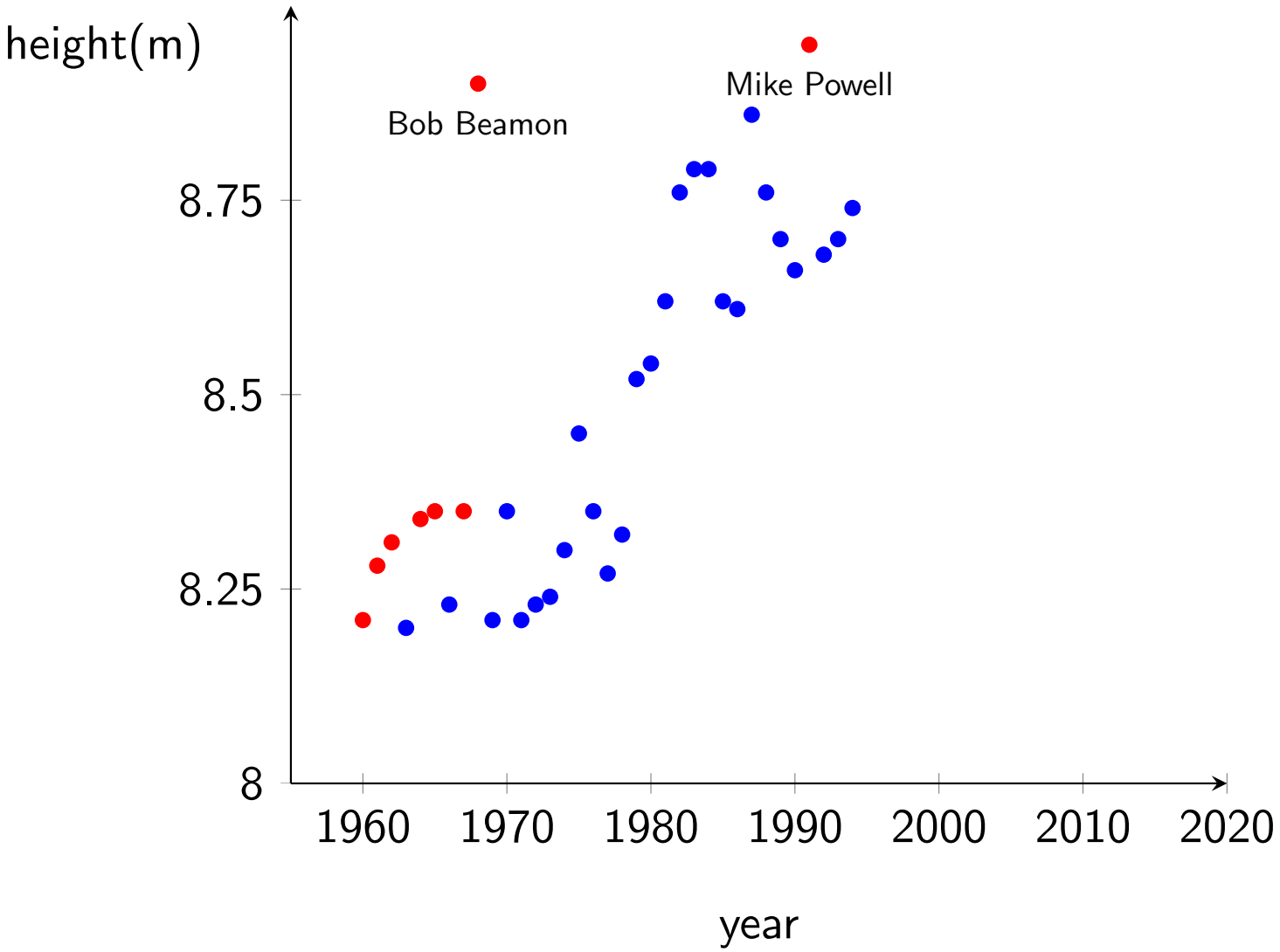
Men's Best Long-Jumps by Year



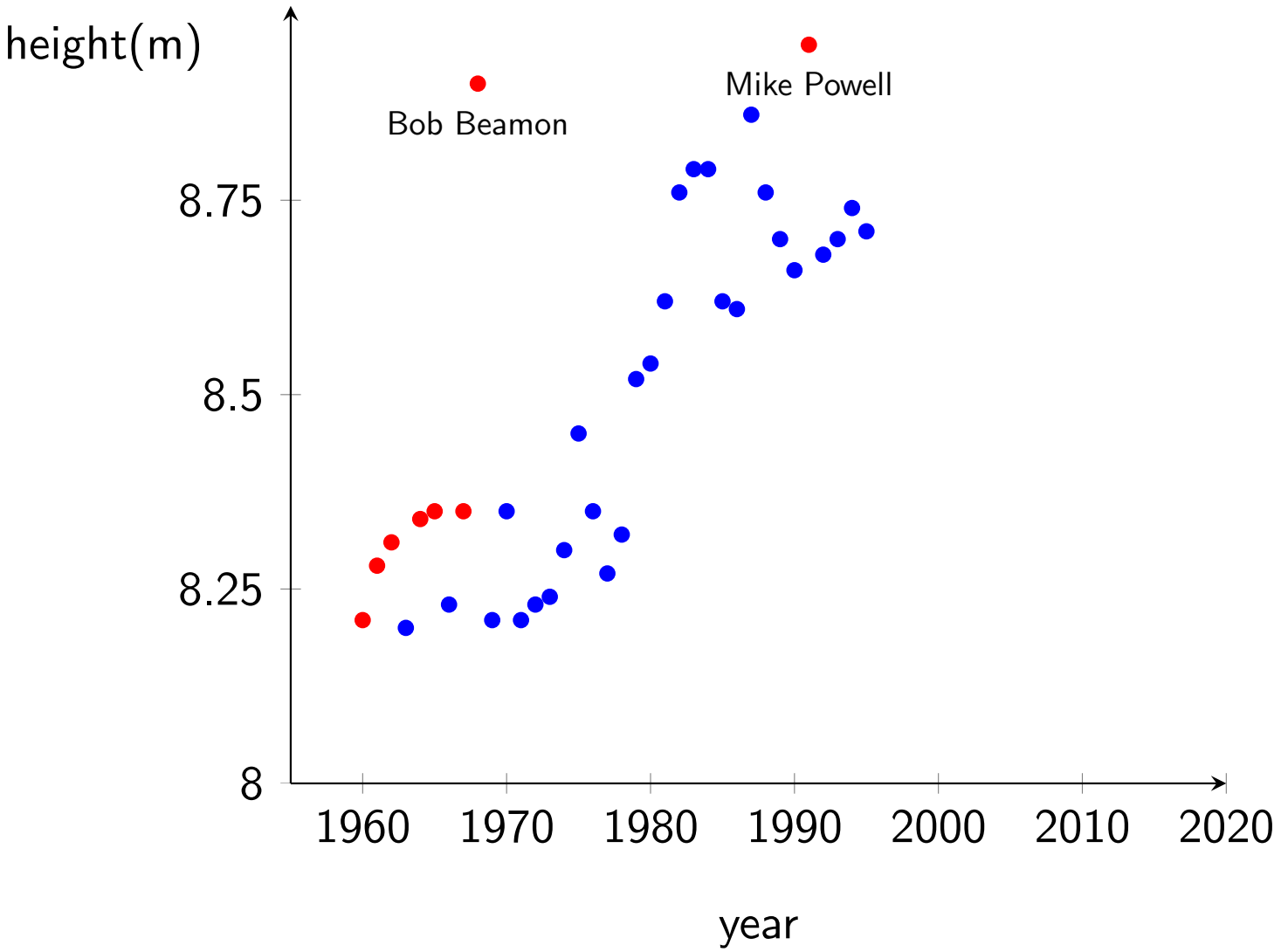
Men's Best Long-Jumps by Year



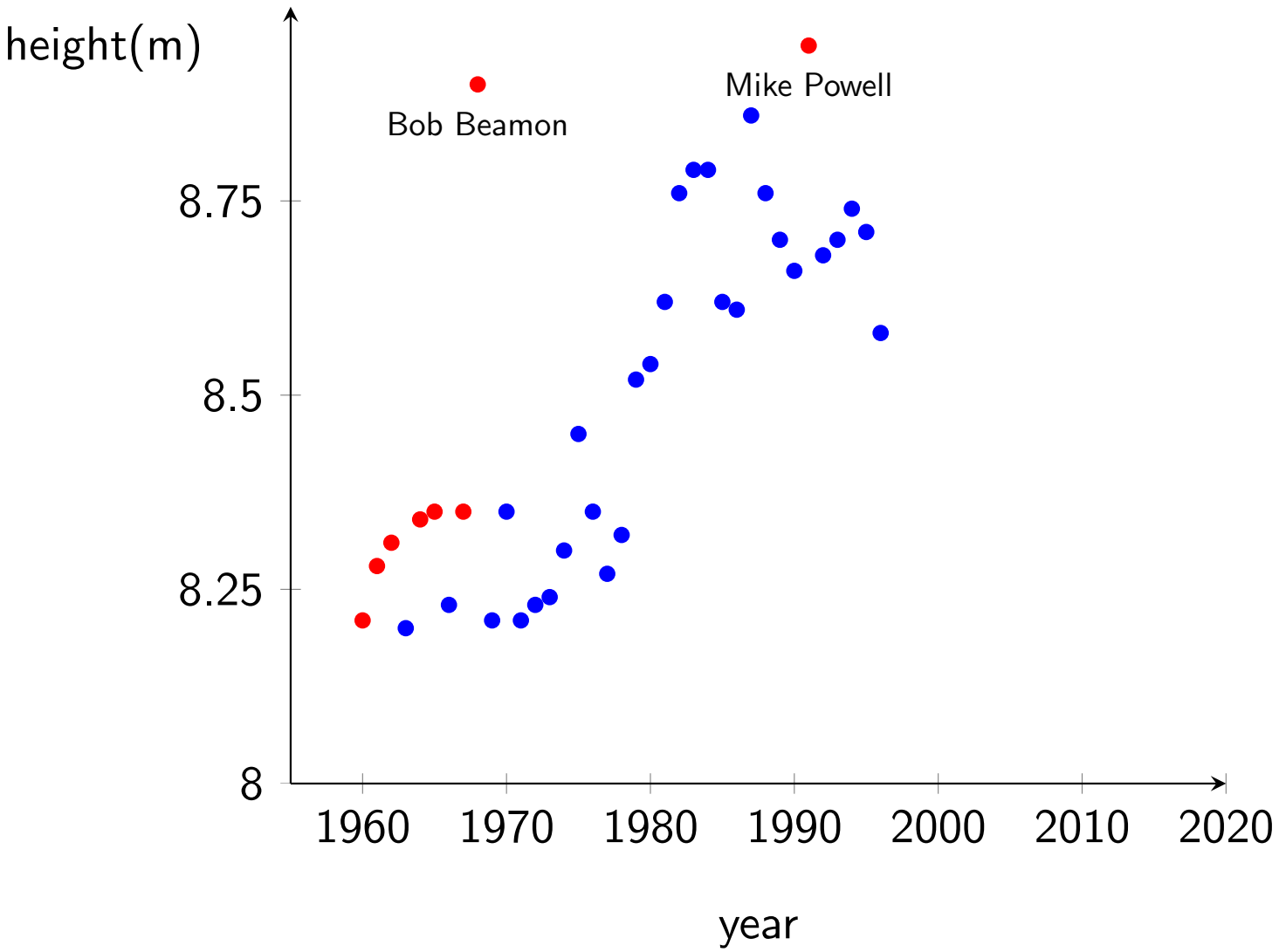
Men's Best Long-Jumps by Year



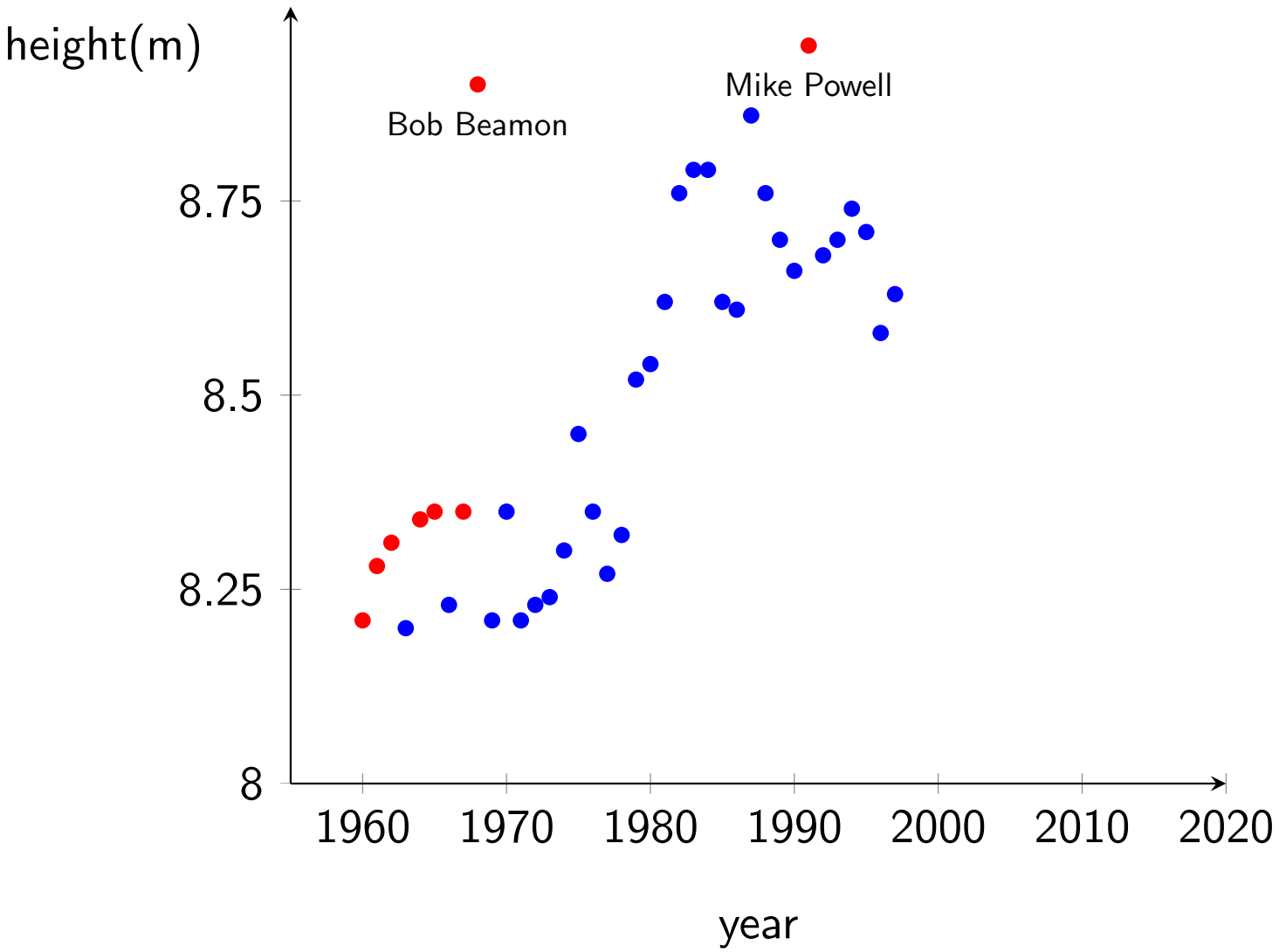
Men's Best Long-Jumps by Year



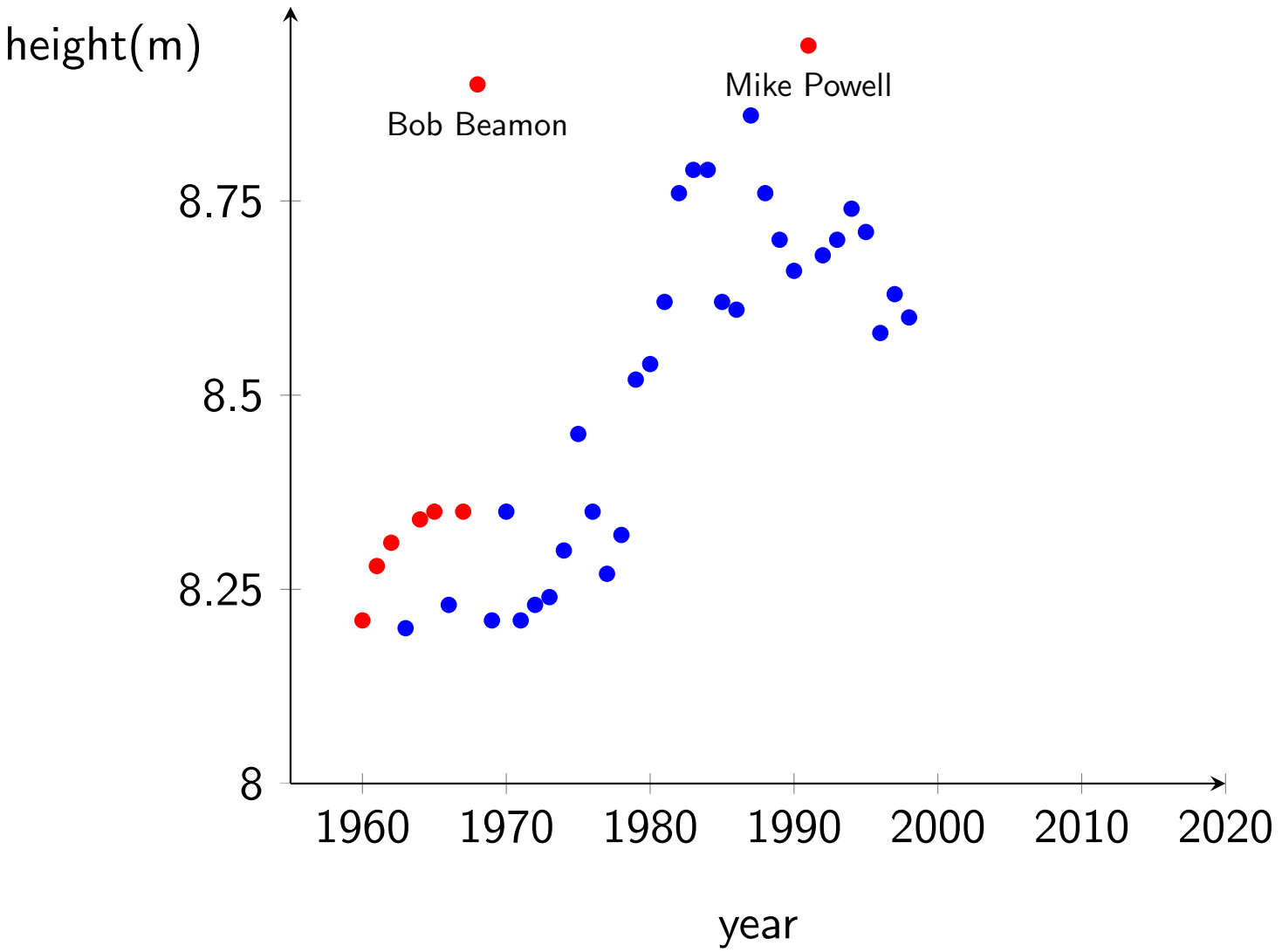
Men's Best Long-Jumps by Year



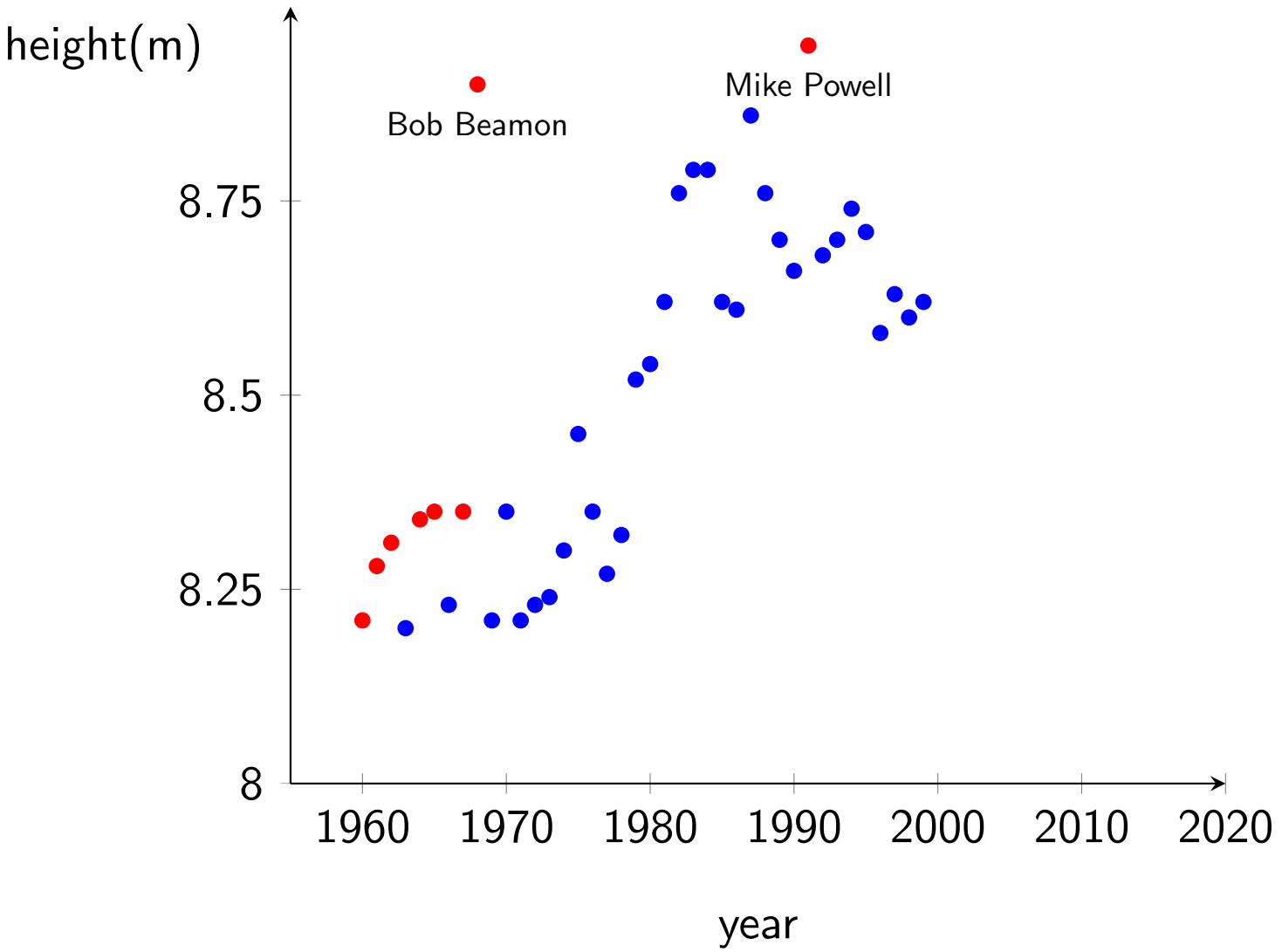
Men's Best Long-Jumps by Year



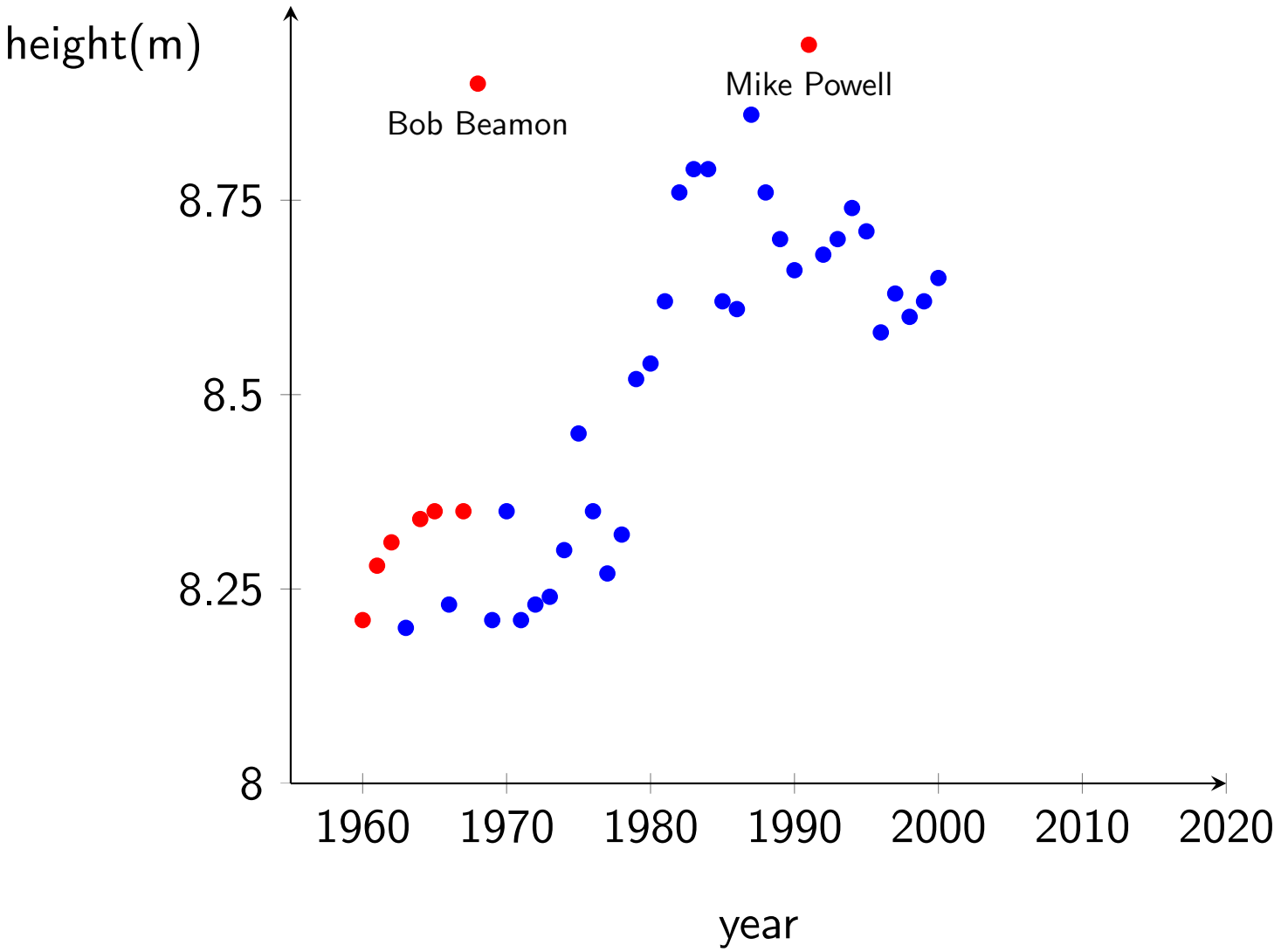
Men's Best Long-Jumps by Year



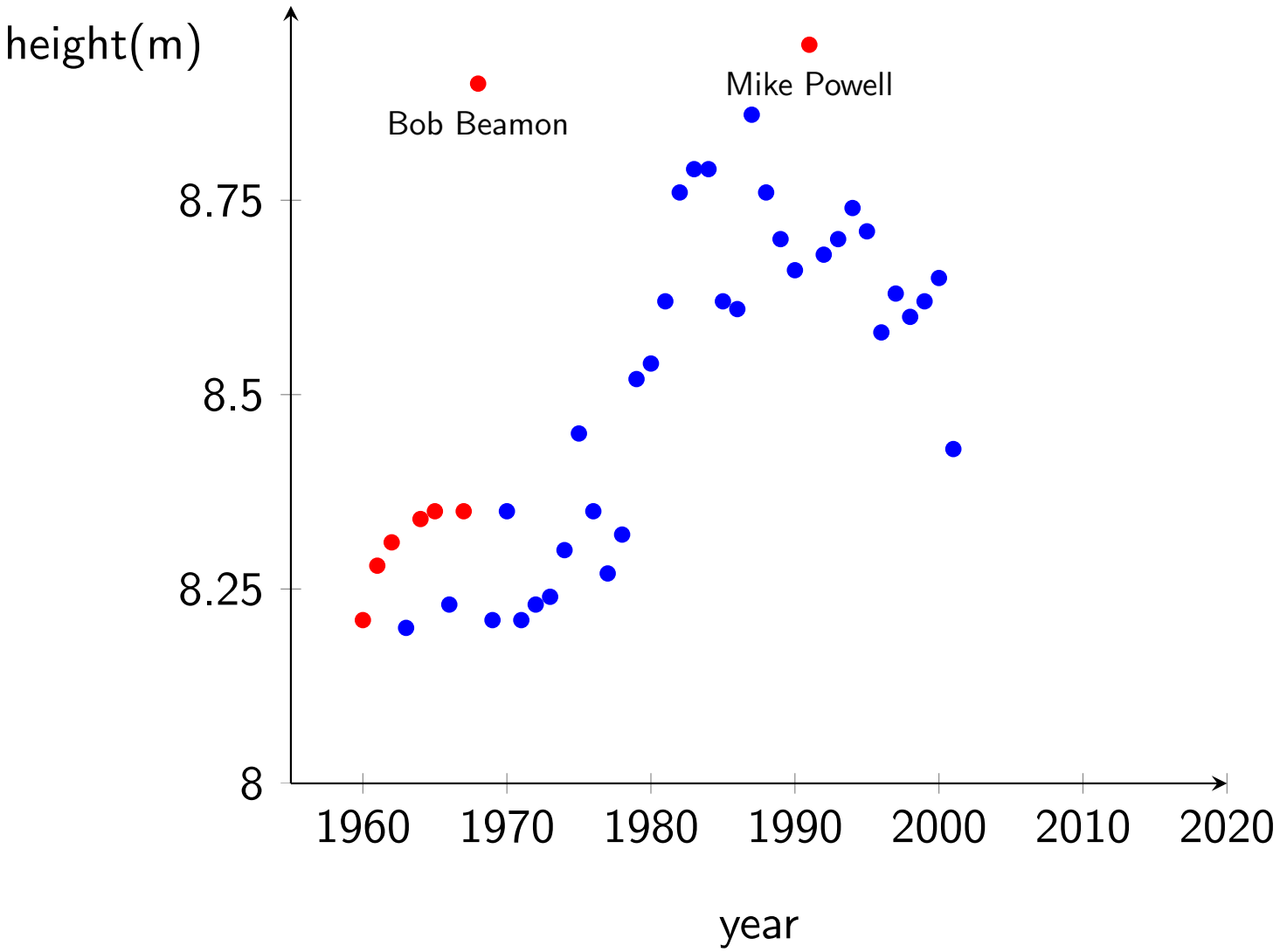
Men's Best Long-Jumps by Year



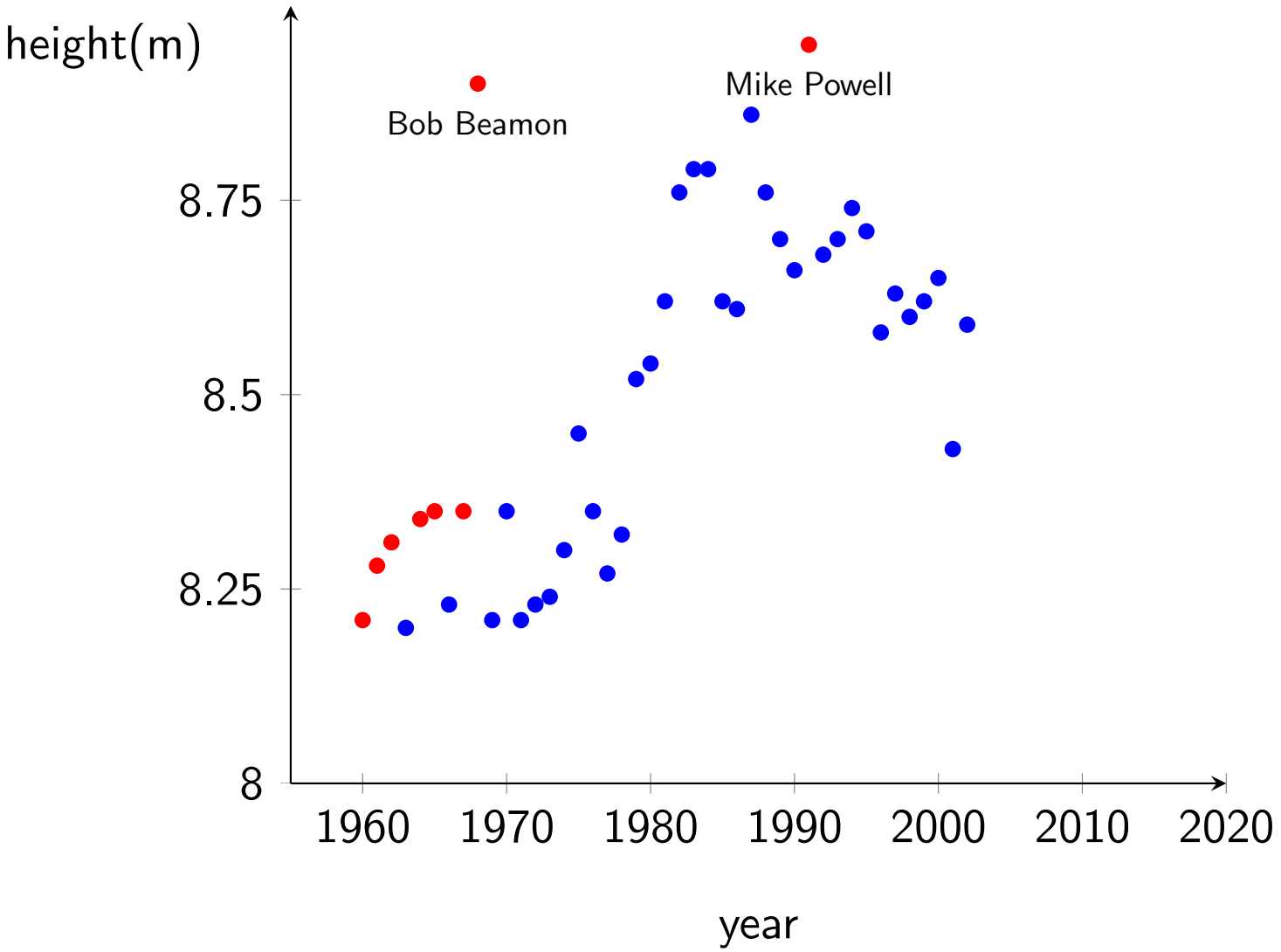
Men's Best Long-Jumps by Year



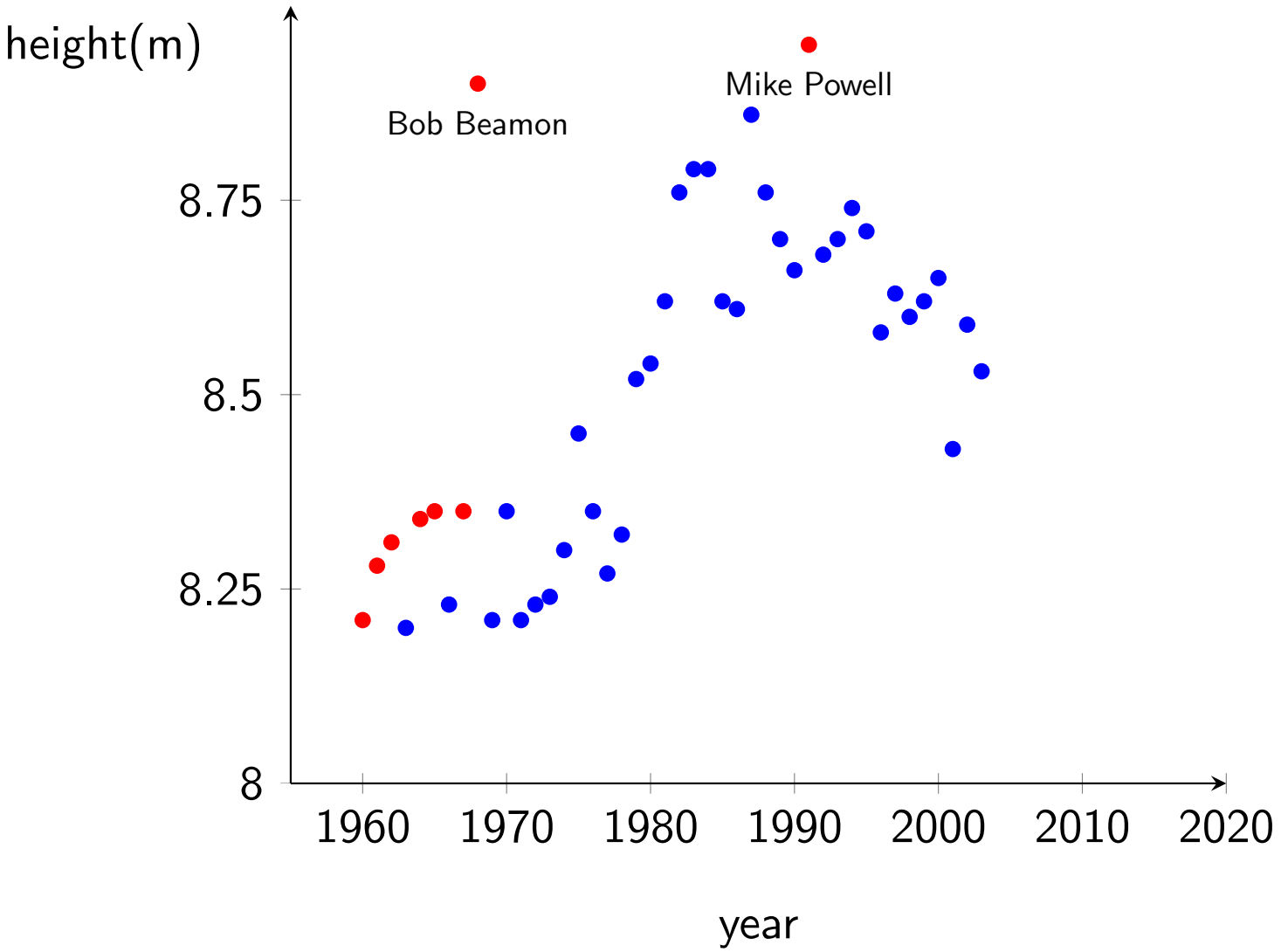
Men's Best Long-Jumps by Year



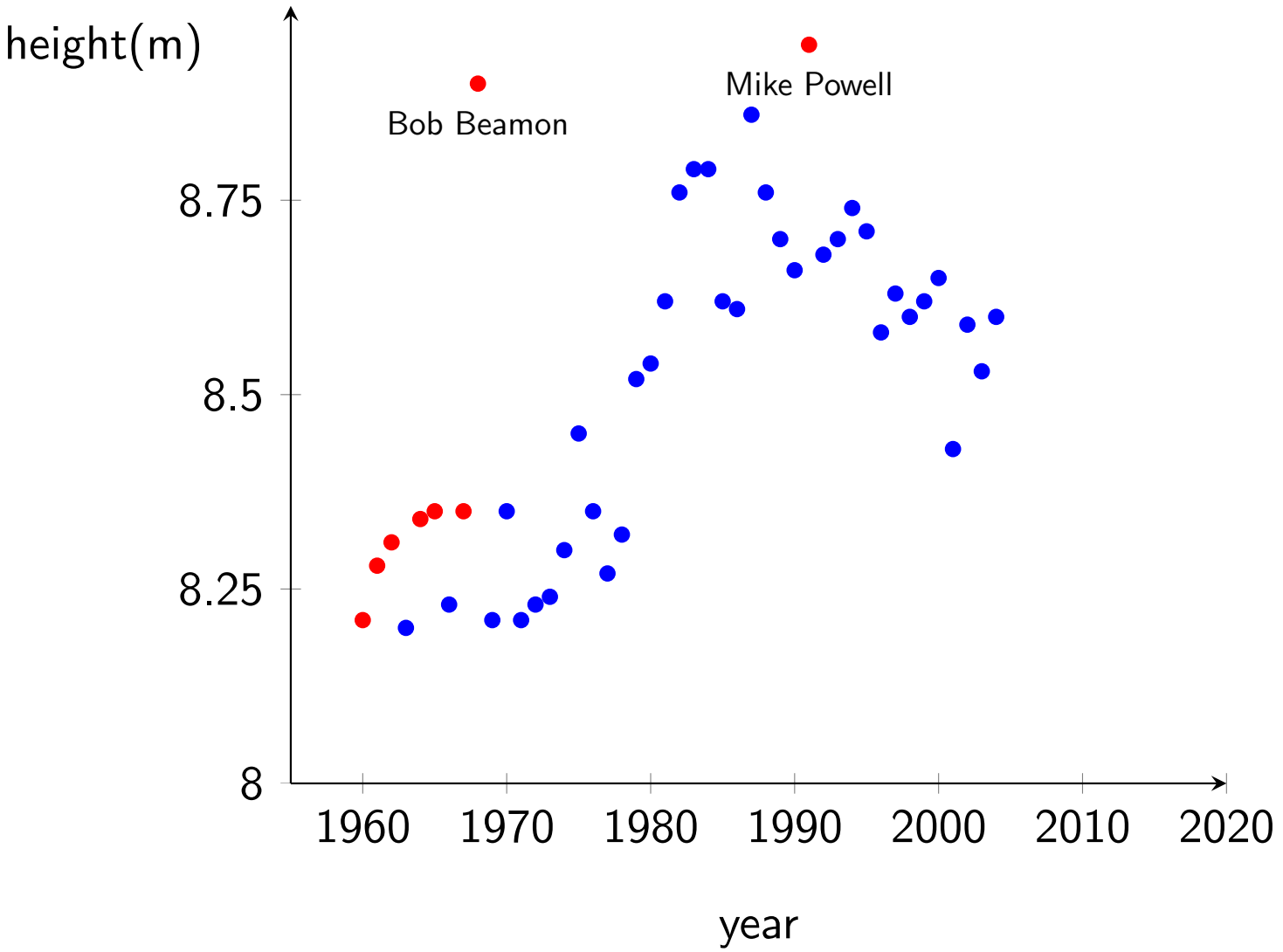
Men's Best Long-Jumps by Year



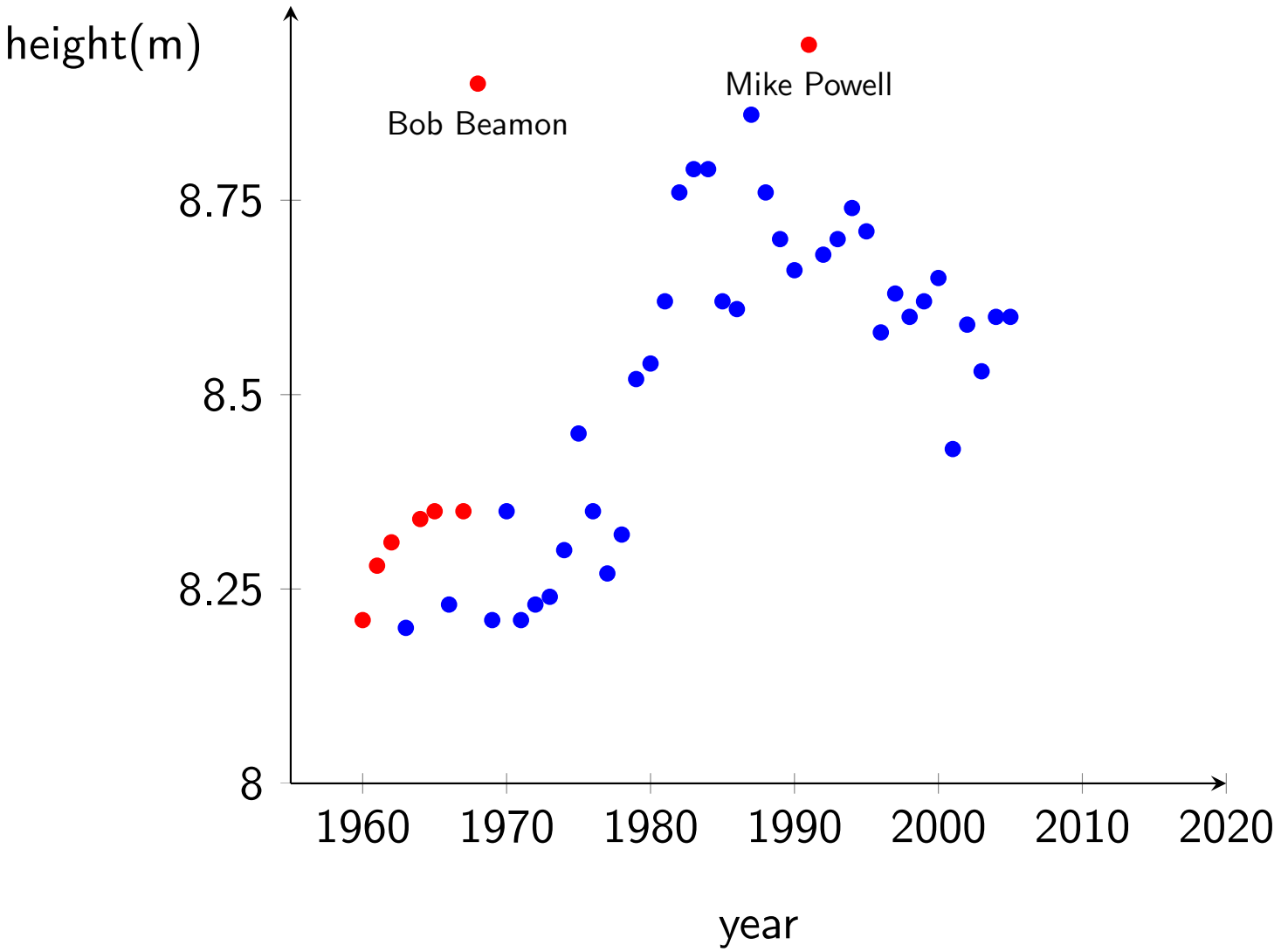
Men's Best Long-Jumps by Year



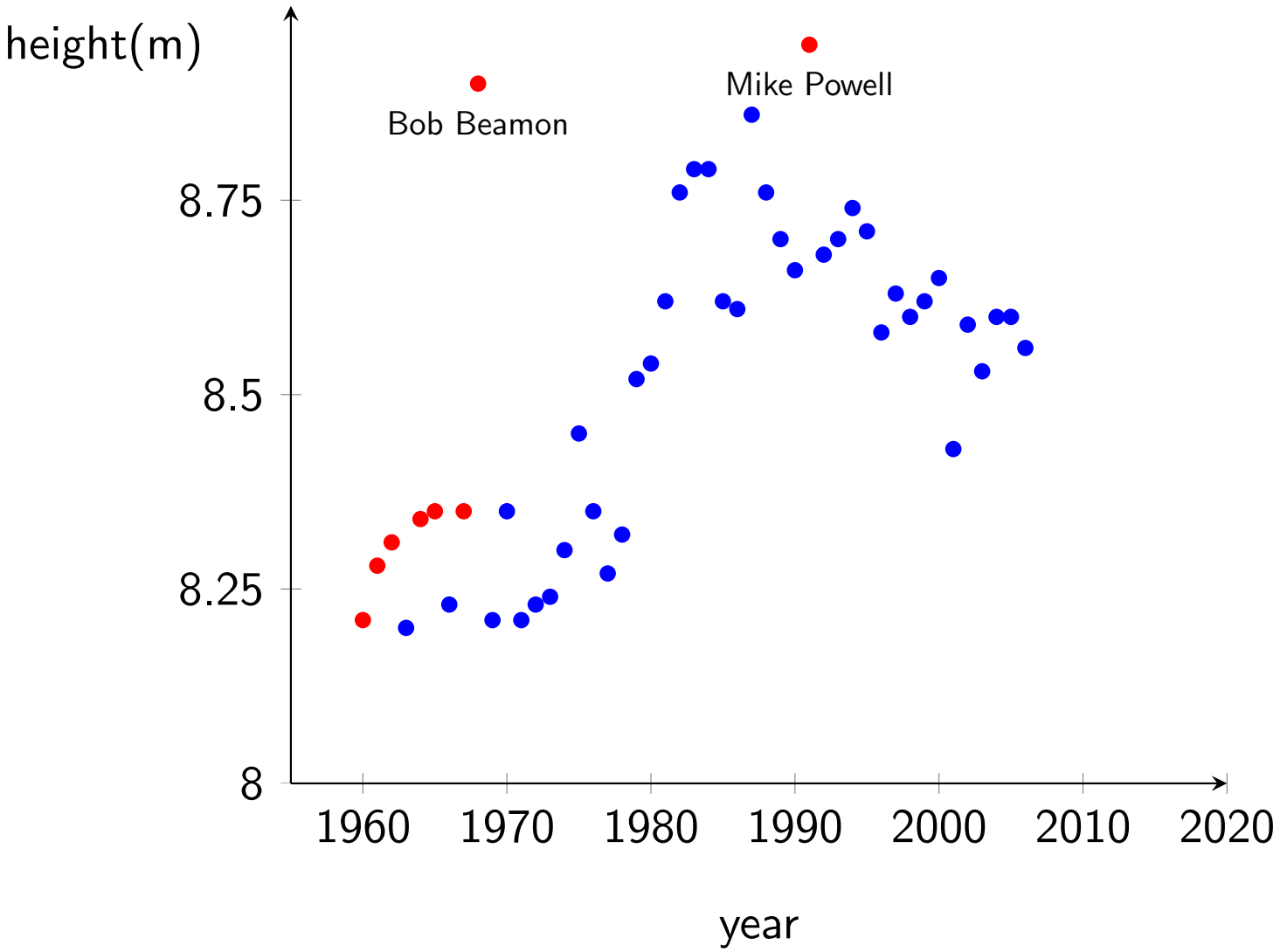
Men's Best Long-Jumps by Year



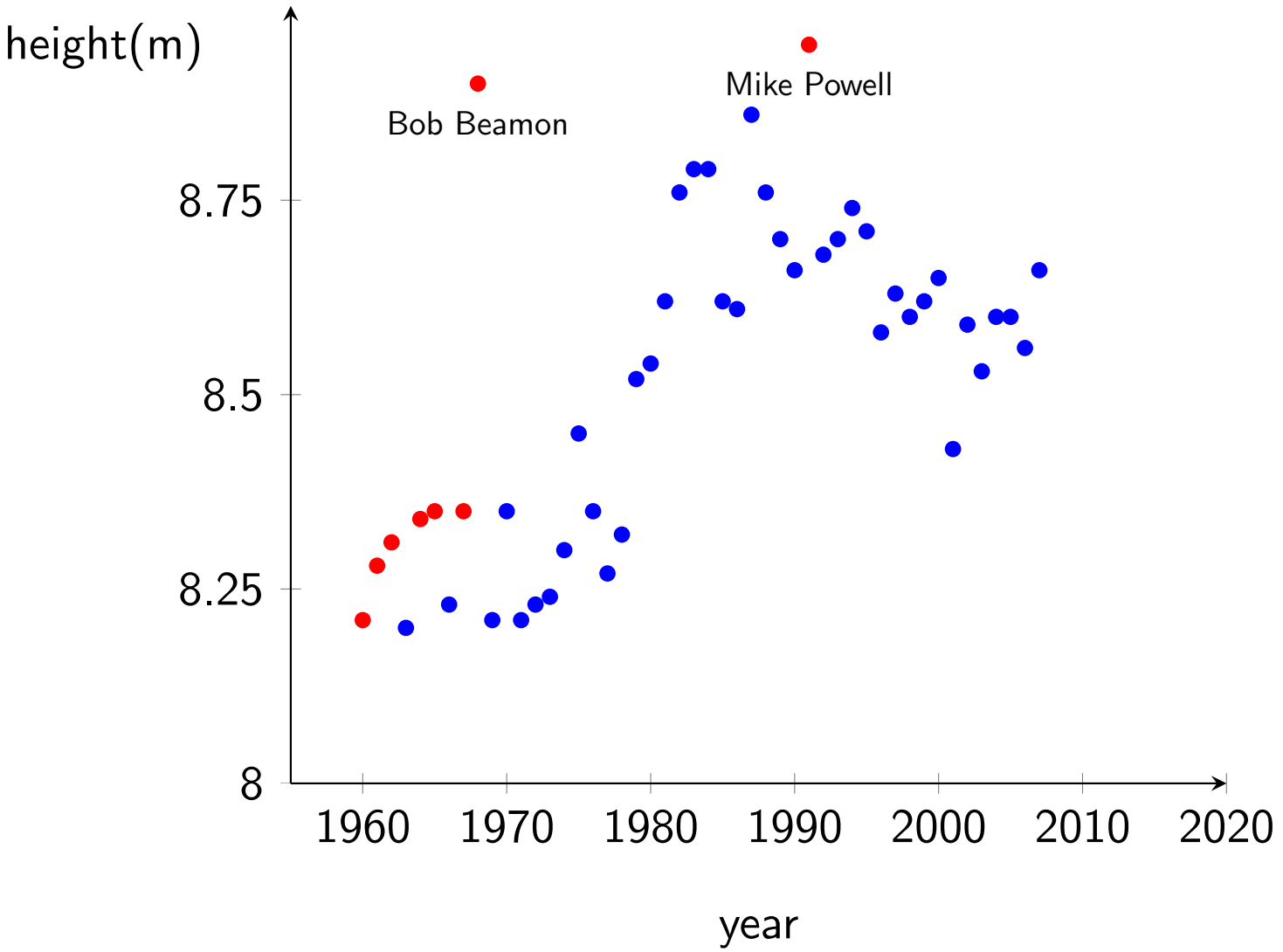
Men's Best Long-Jumps by Year



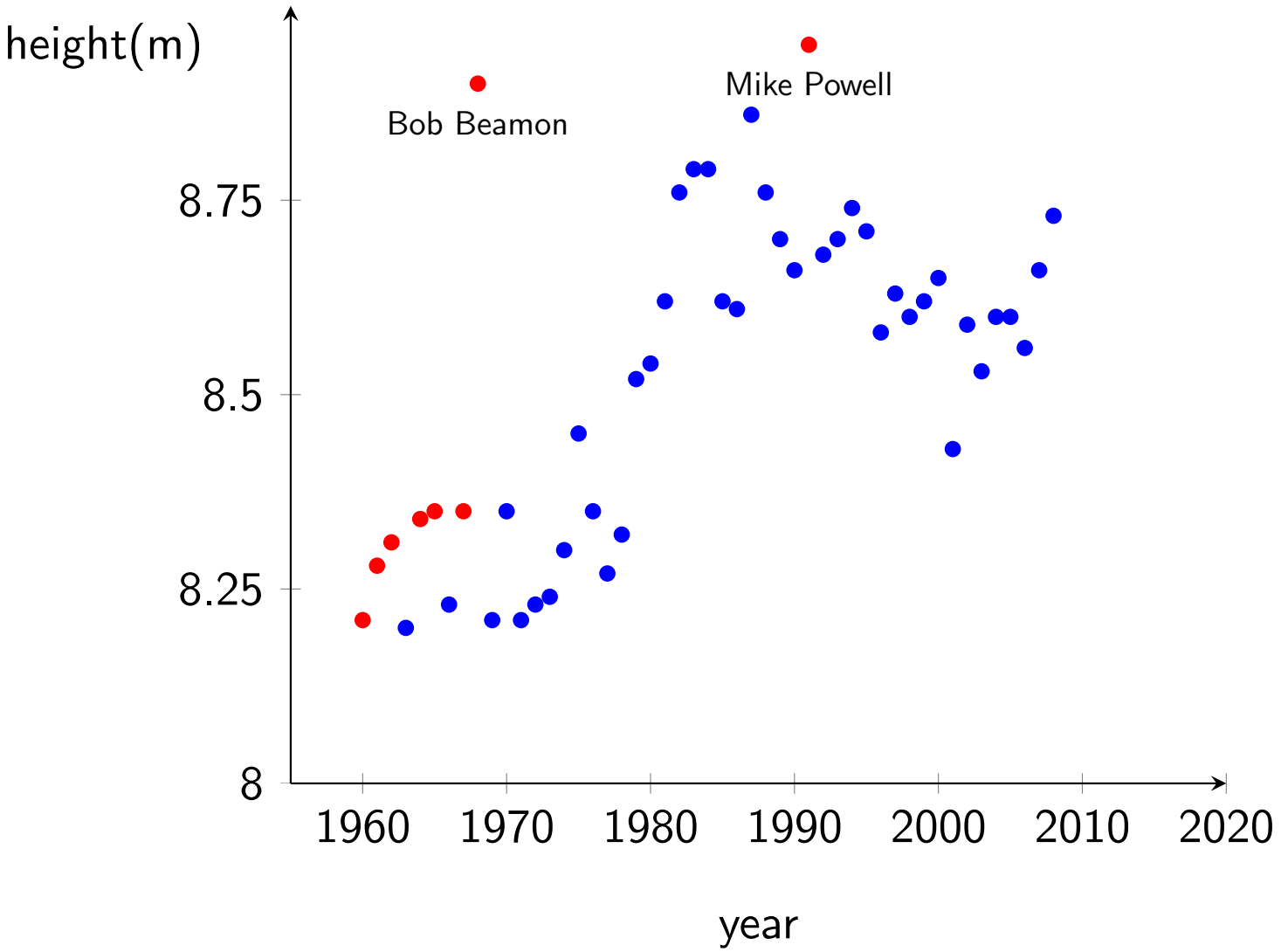
Men's Best Long-Jumps by Year



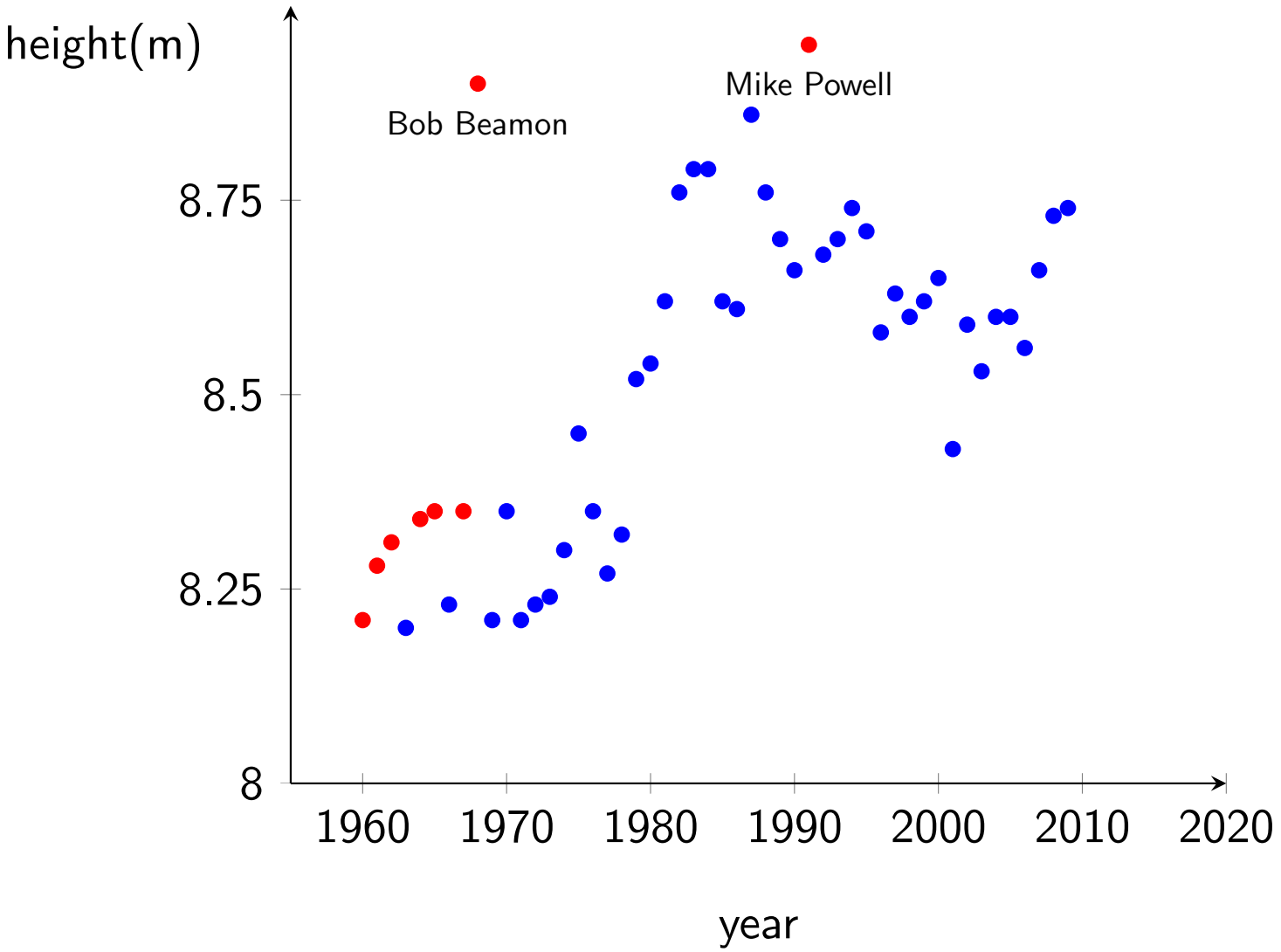
Men's Best Long-Jumps by Year



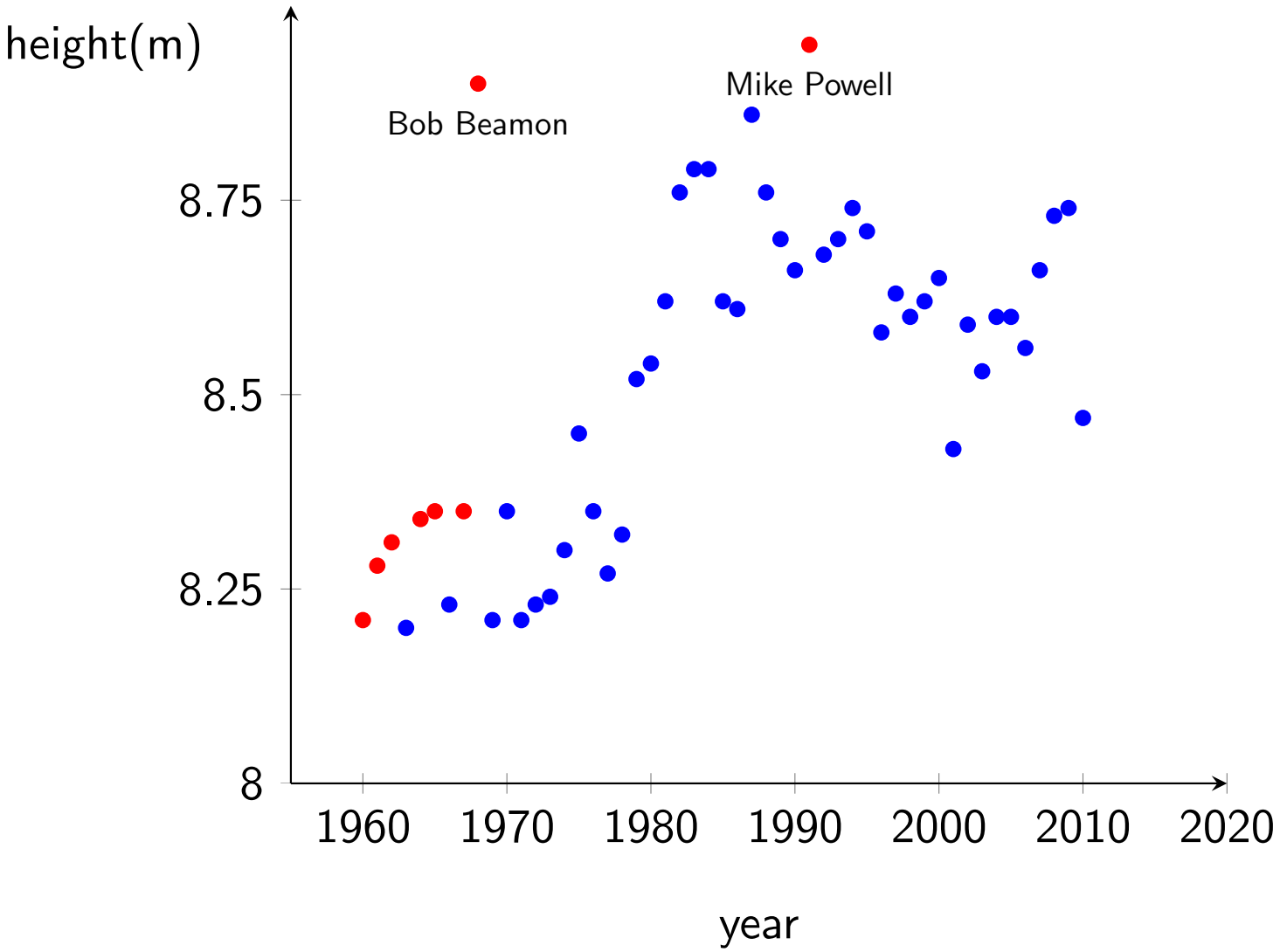
Men's Best Long-Jumps by Year



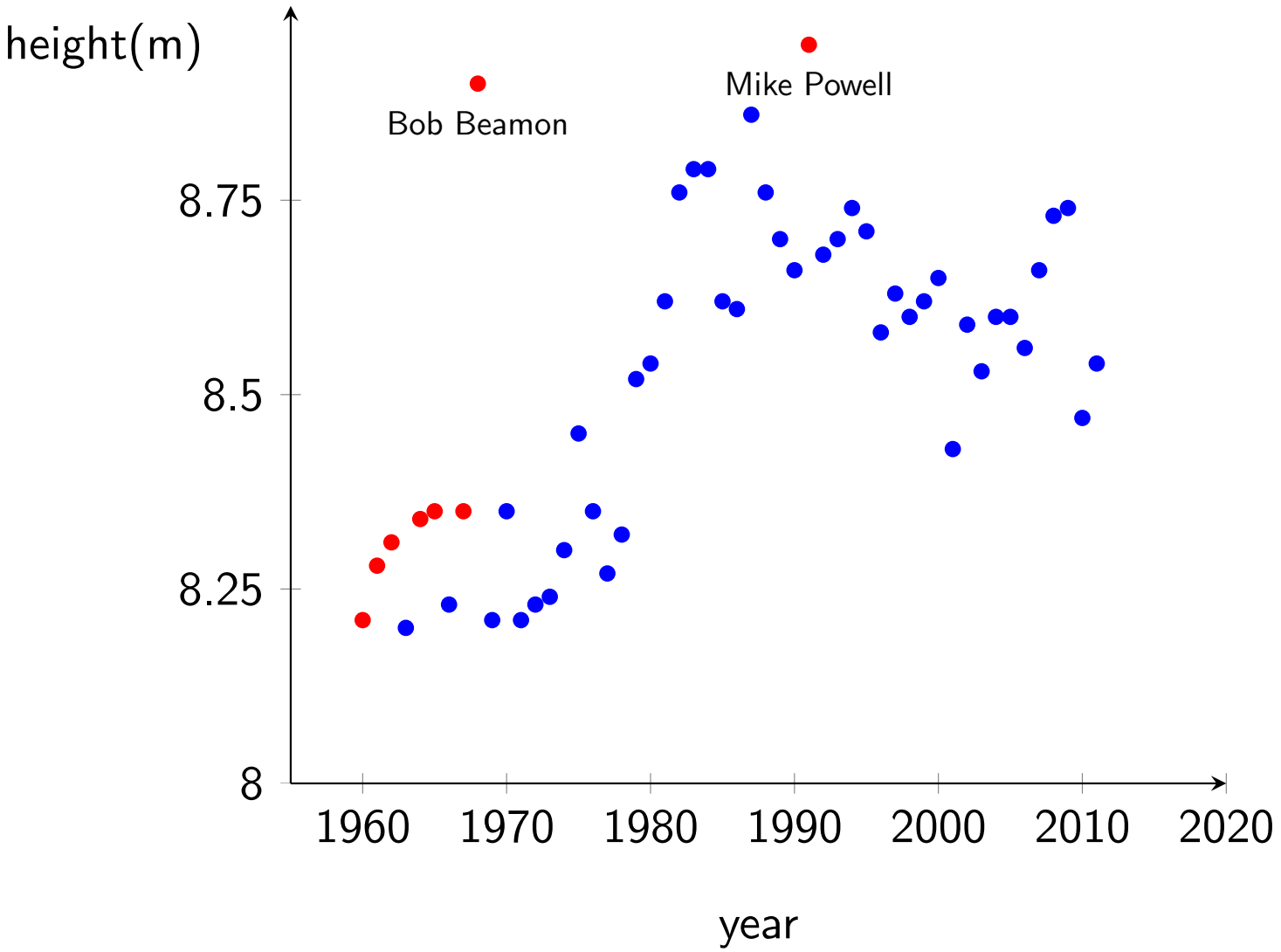
Men's Best Long-Jumps by Year



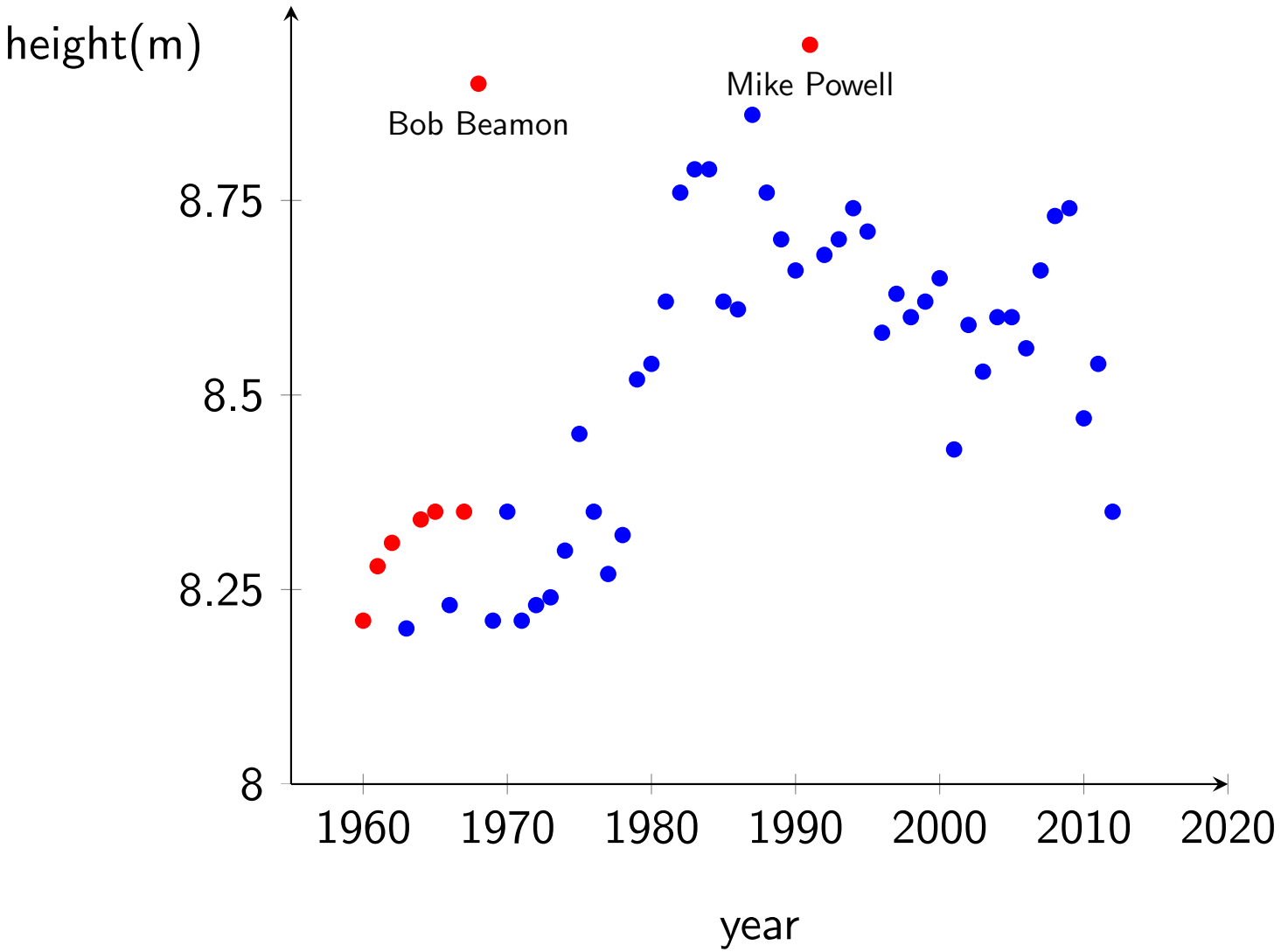
Men's Best Long-Jumps by Year



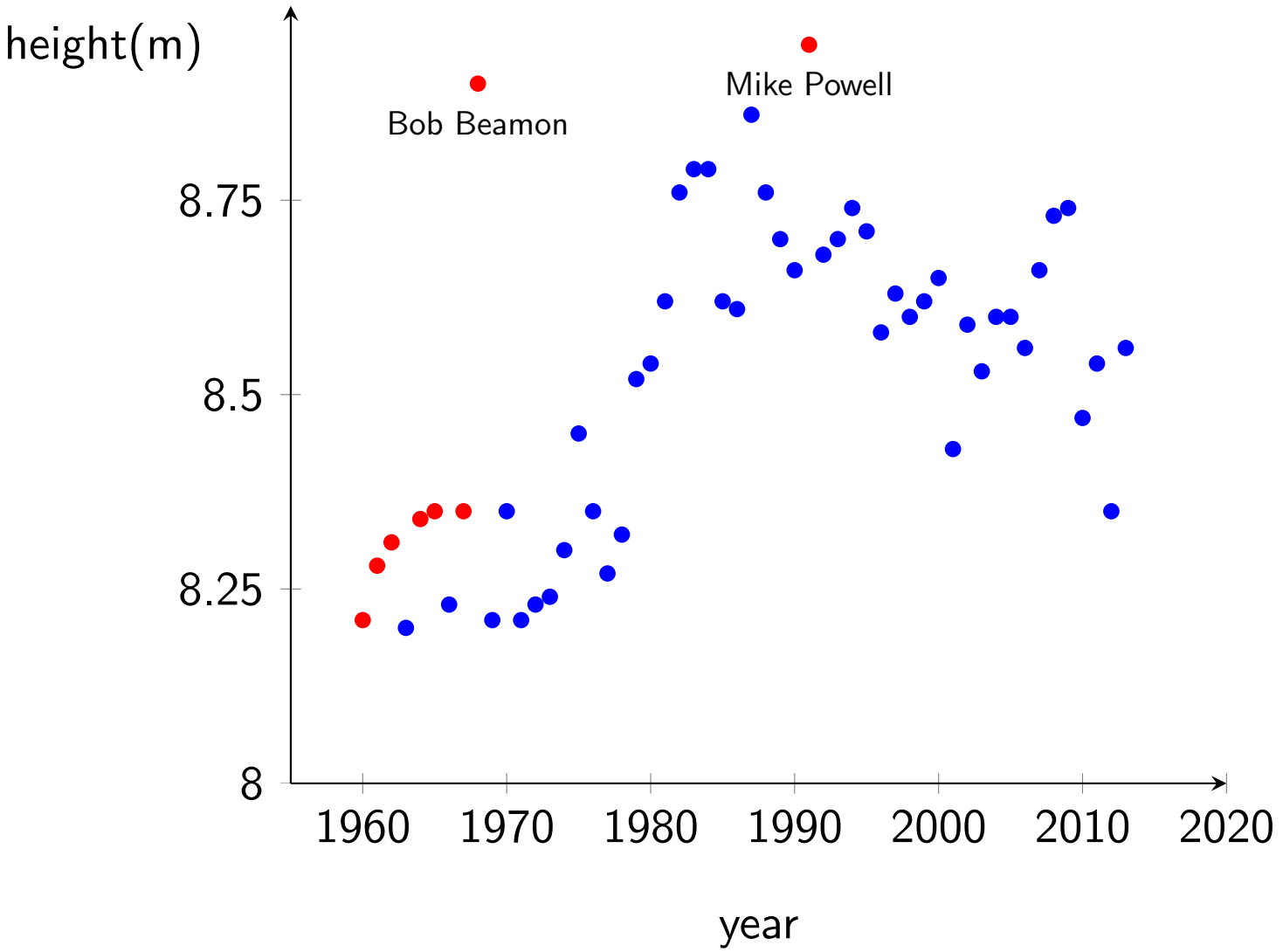
Men's Best Long-Jumps by Year



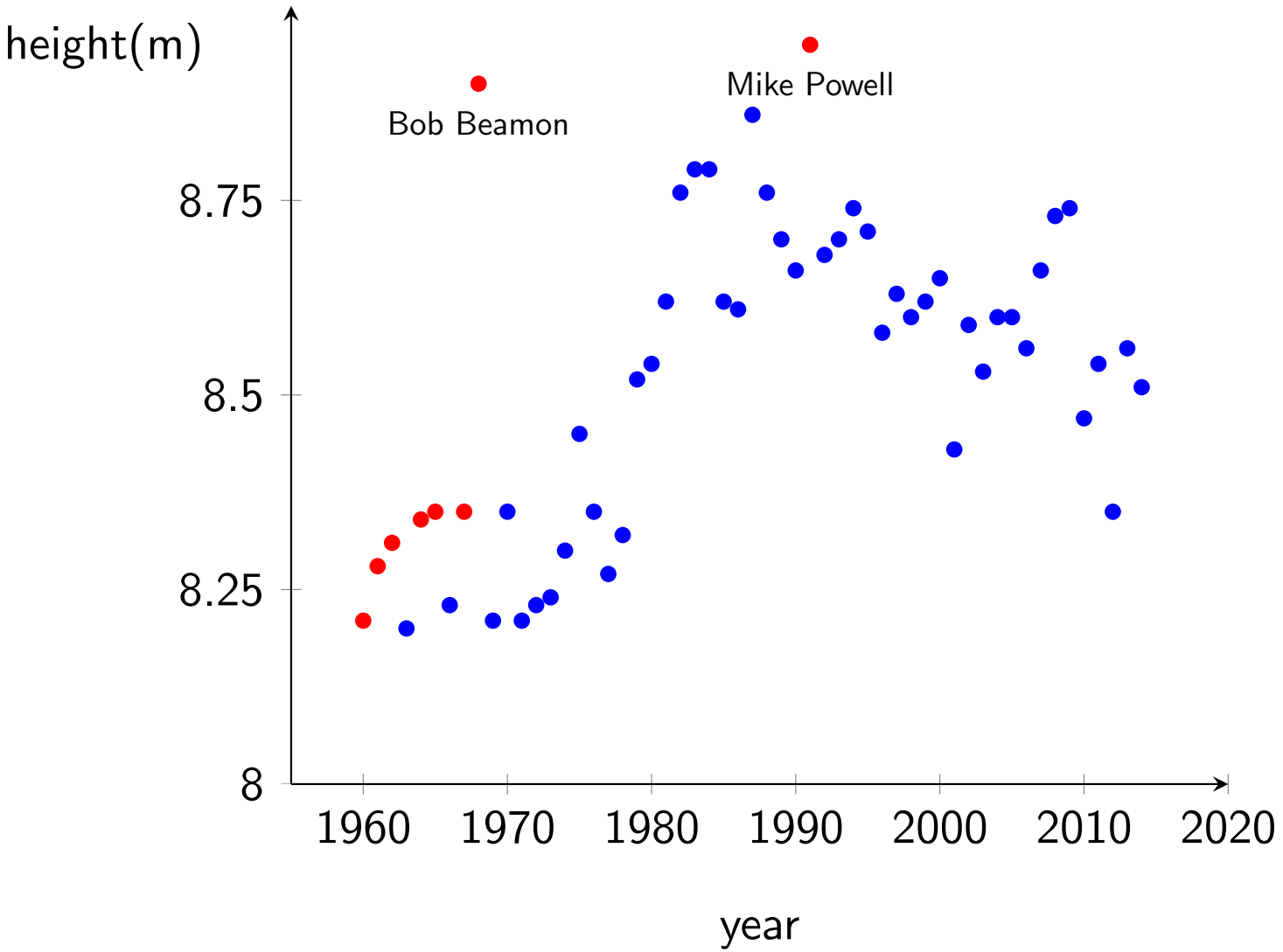
Men's Best Long-Jumps by Year



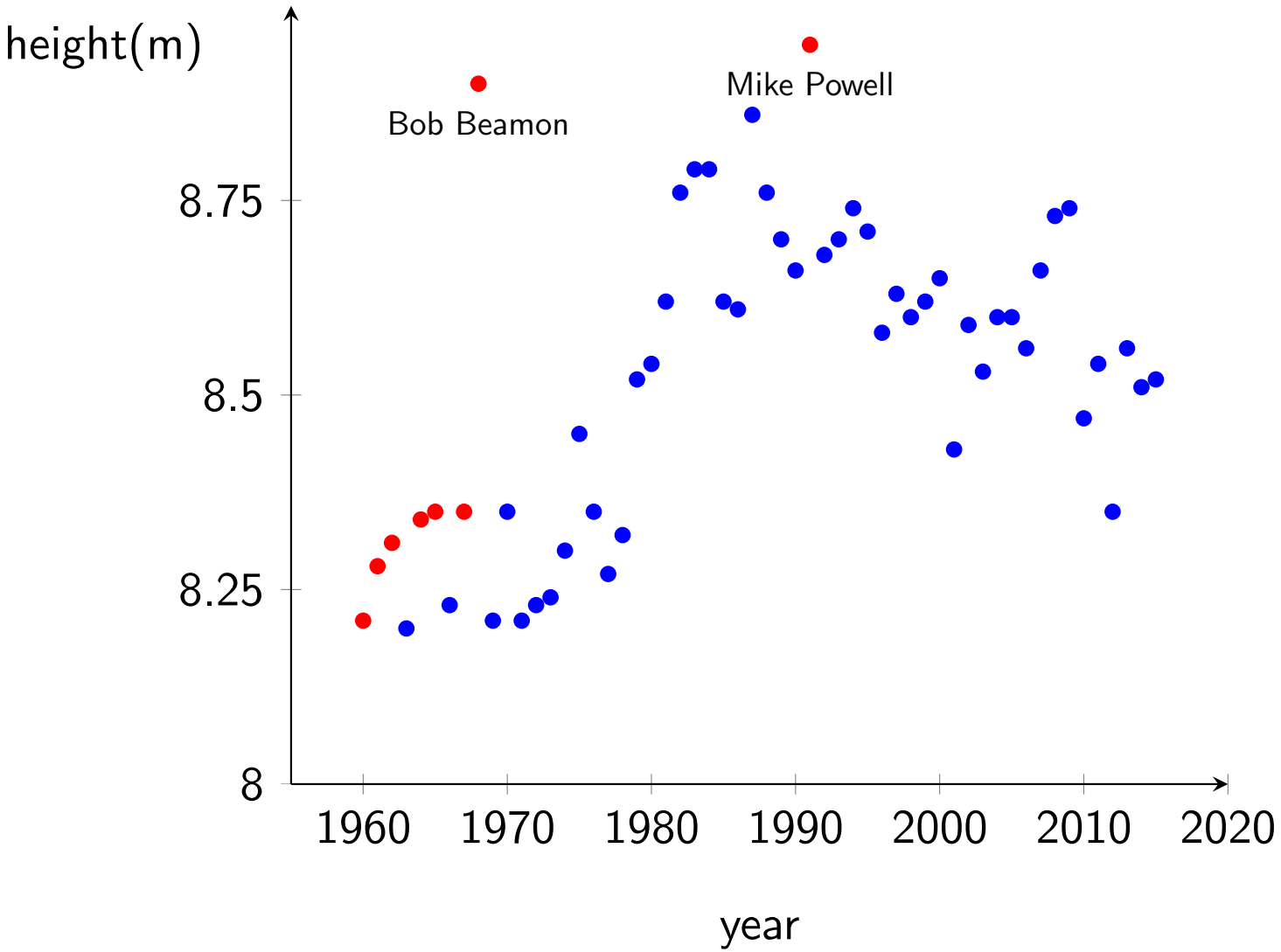
Men's Best Long-Jumps by Year



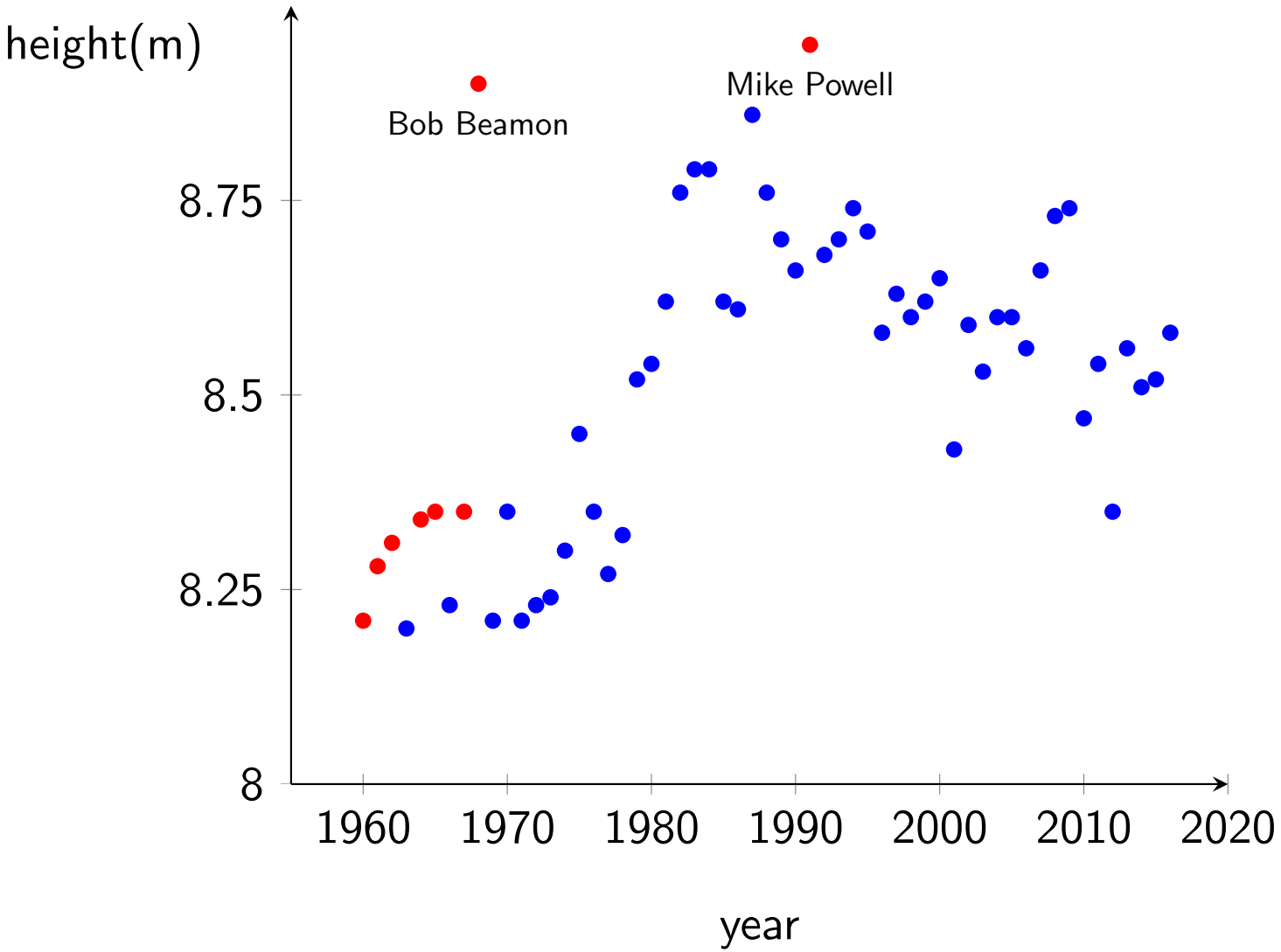
Men's Best Long-Jumps by Year



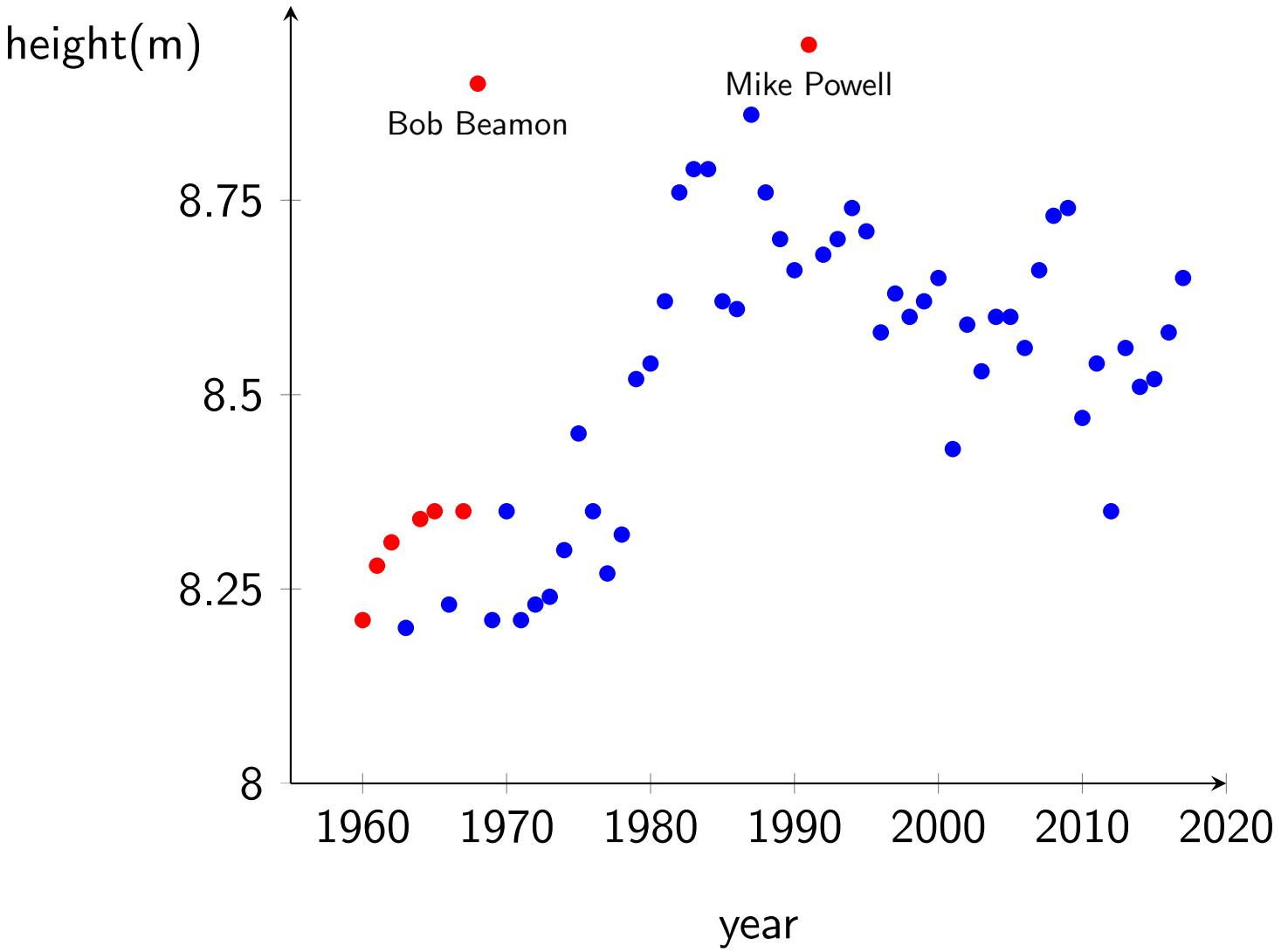
Men's Best Long-Jumps by Year



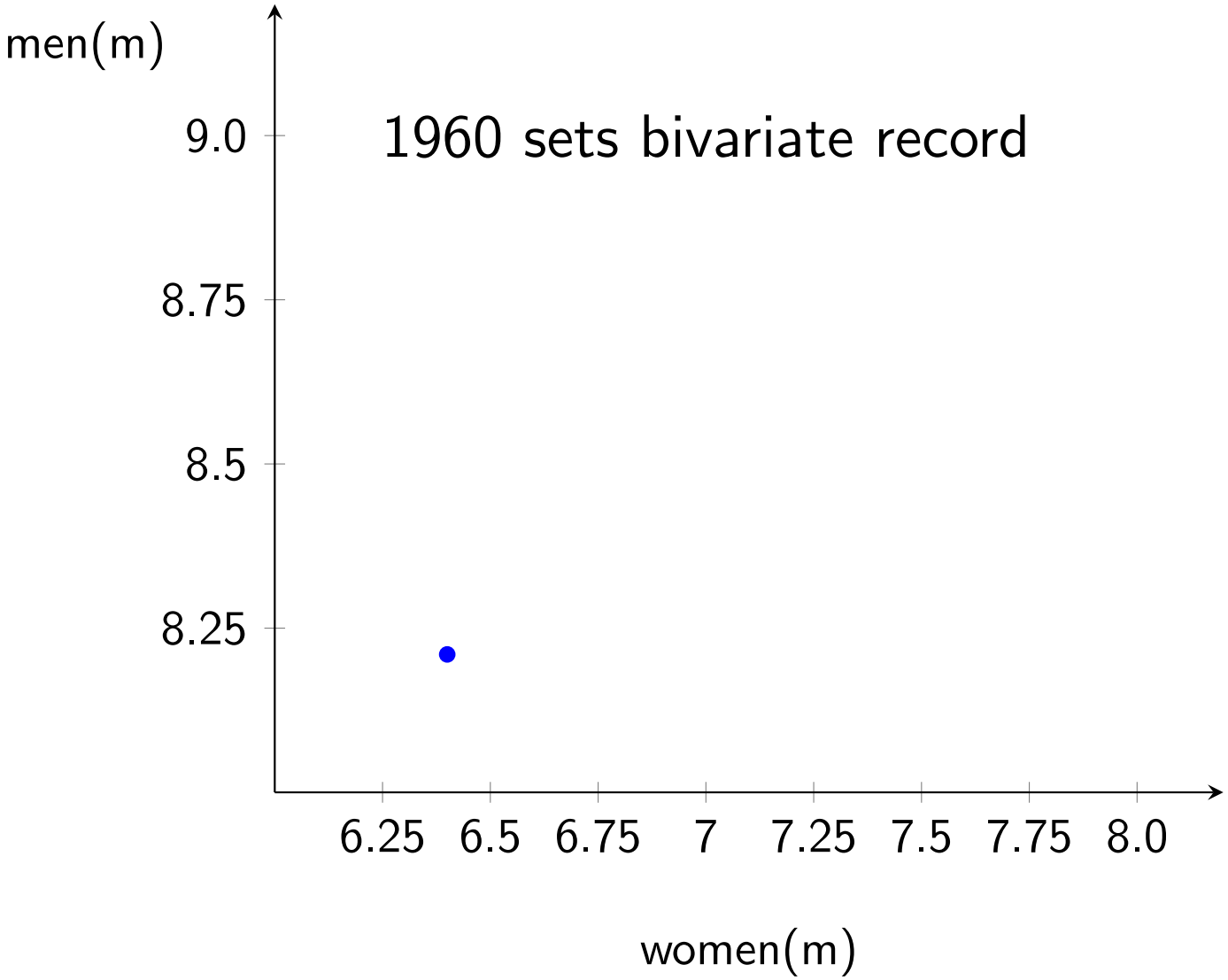
Men's Best Long-Jumps by Year



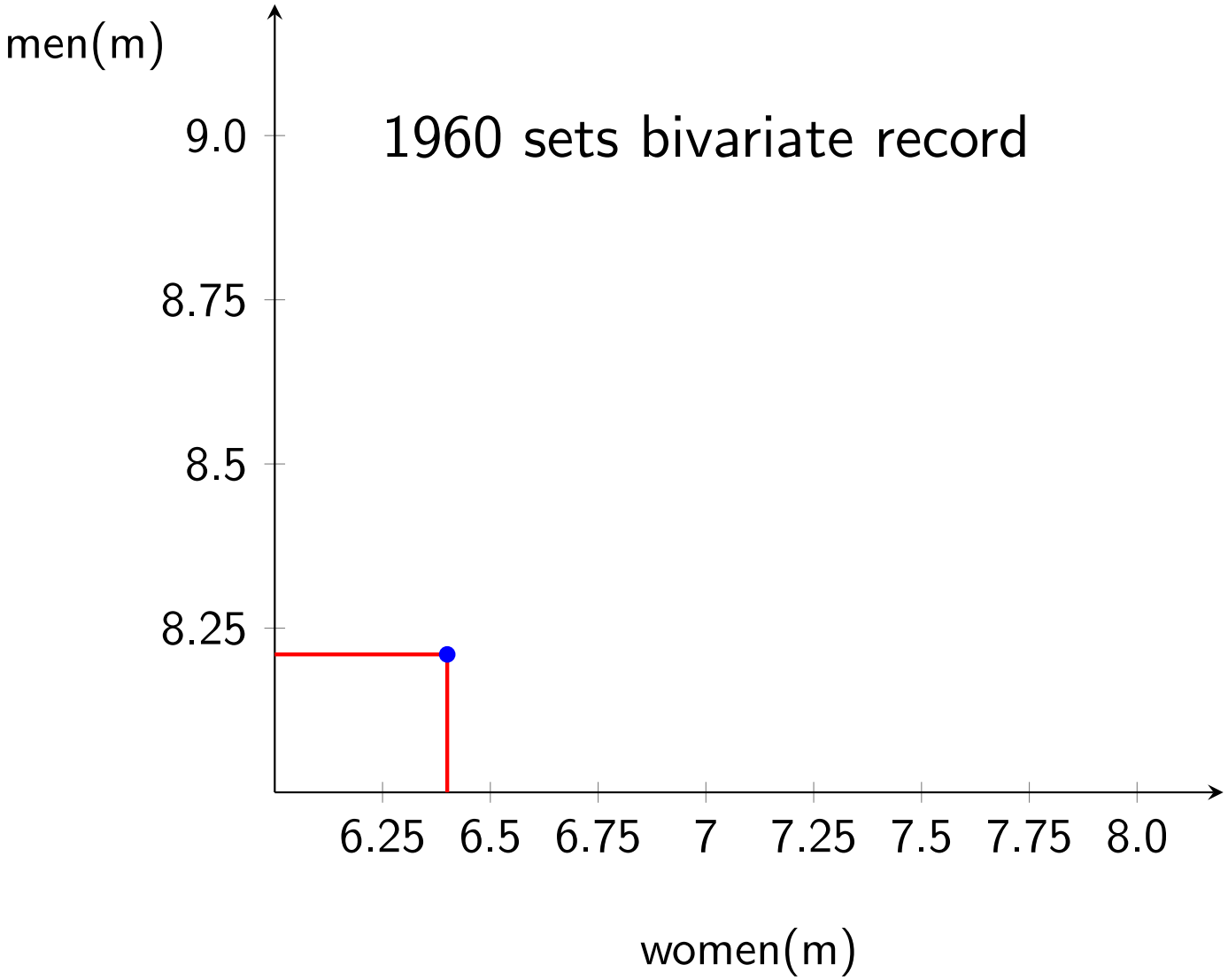
Men's Best Long-Jumps by Year



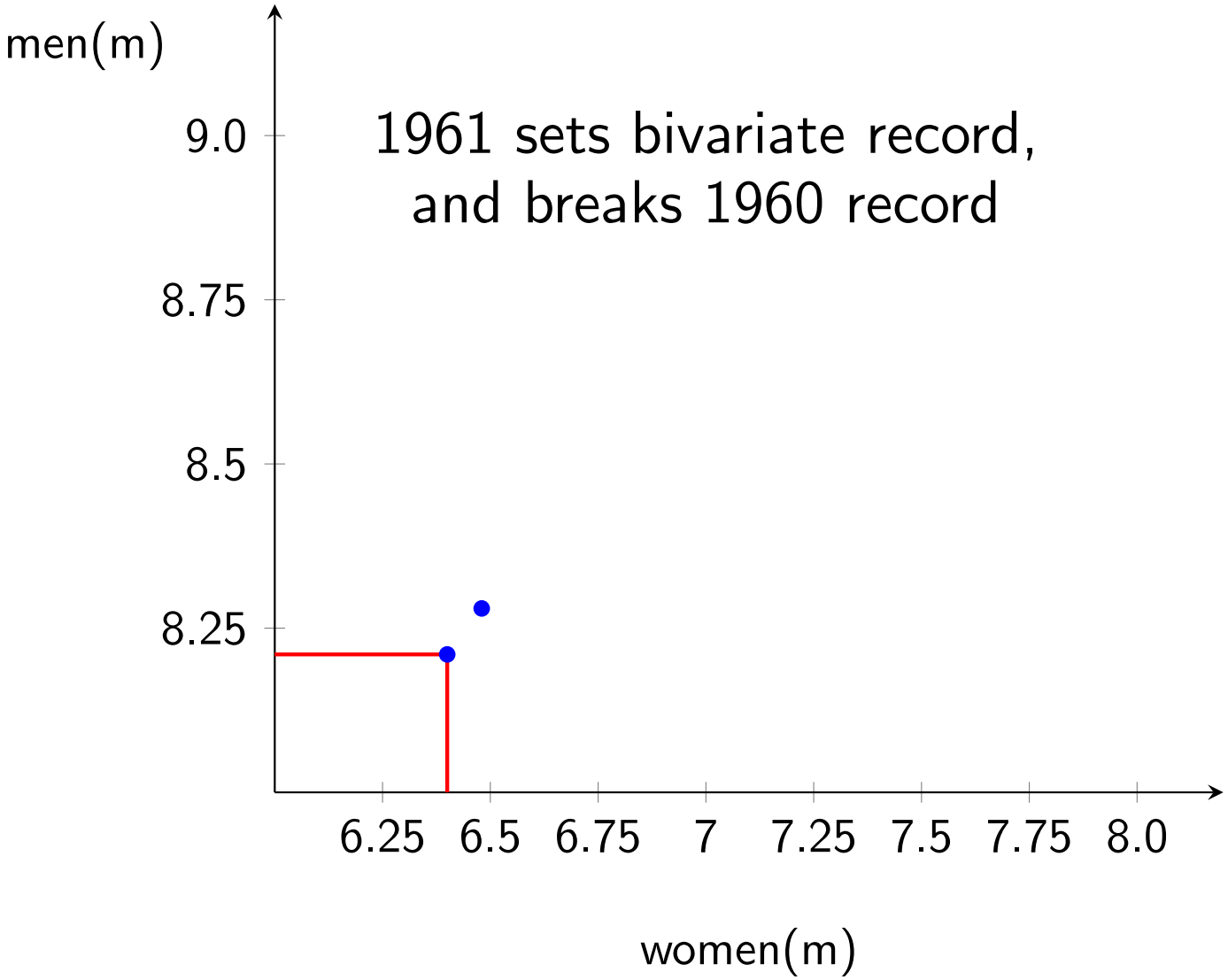
Men and Women's Best Long Jumps by Year



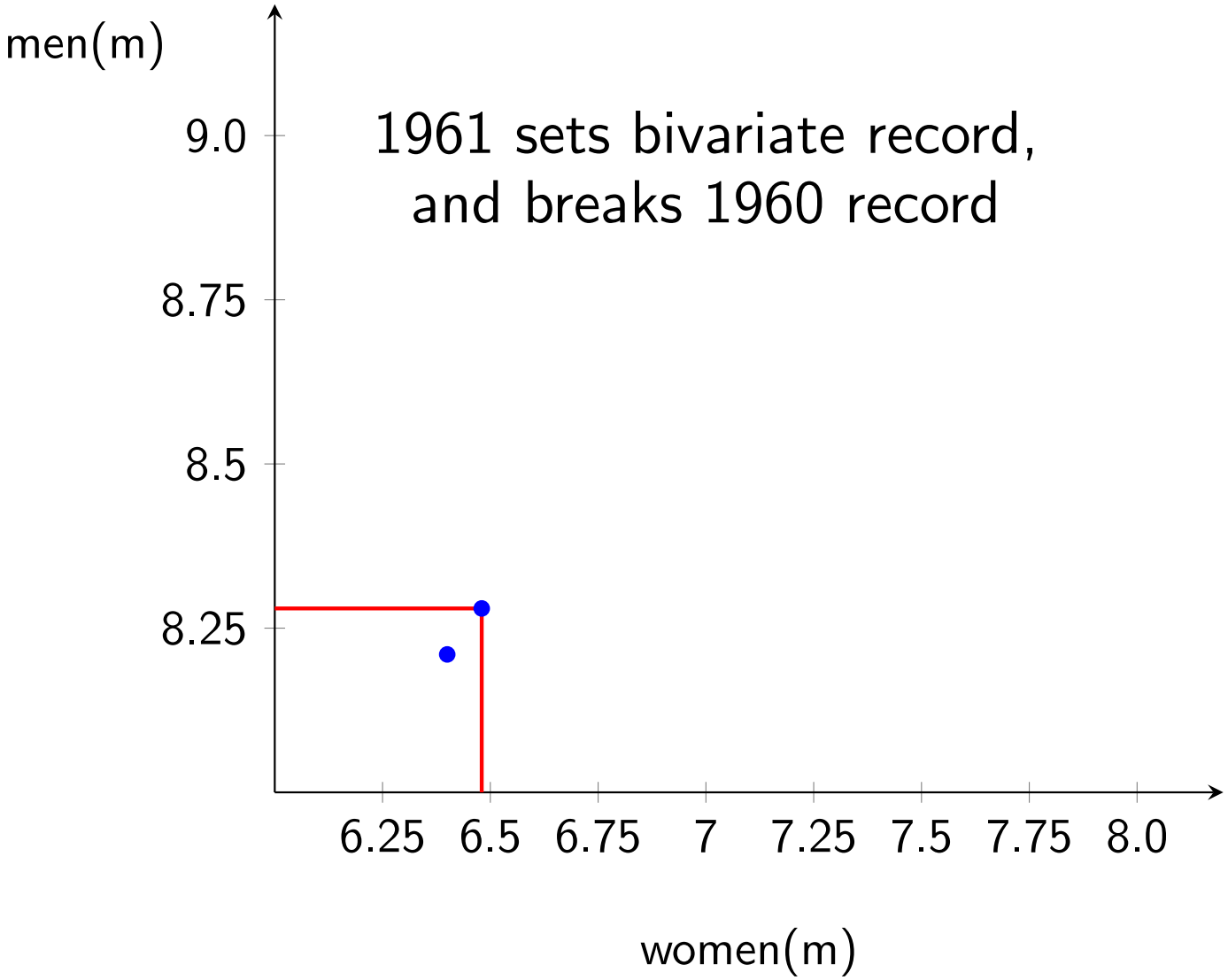
Men and Women's Best Long Jumps by Year



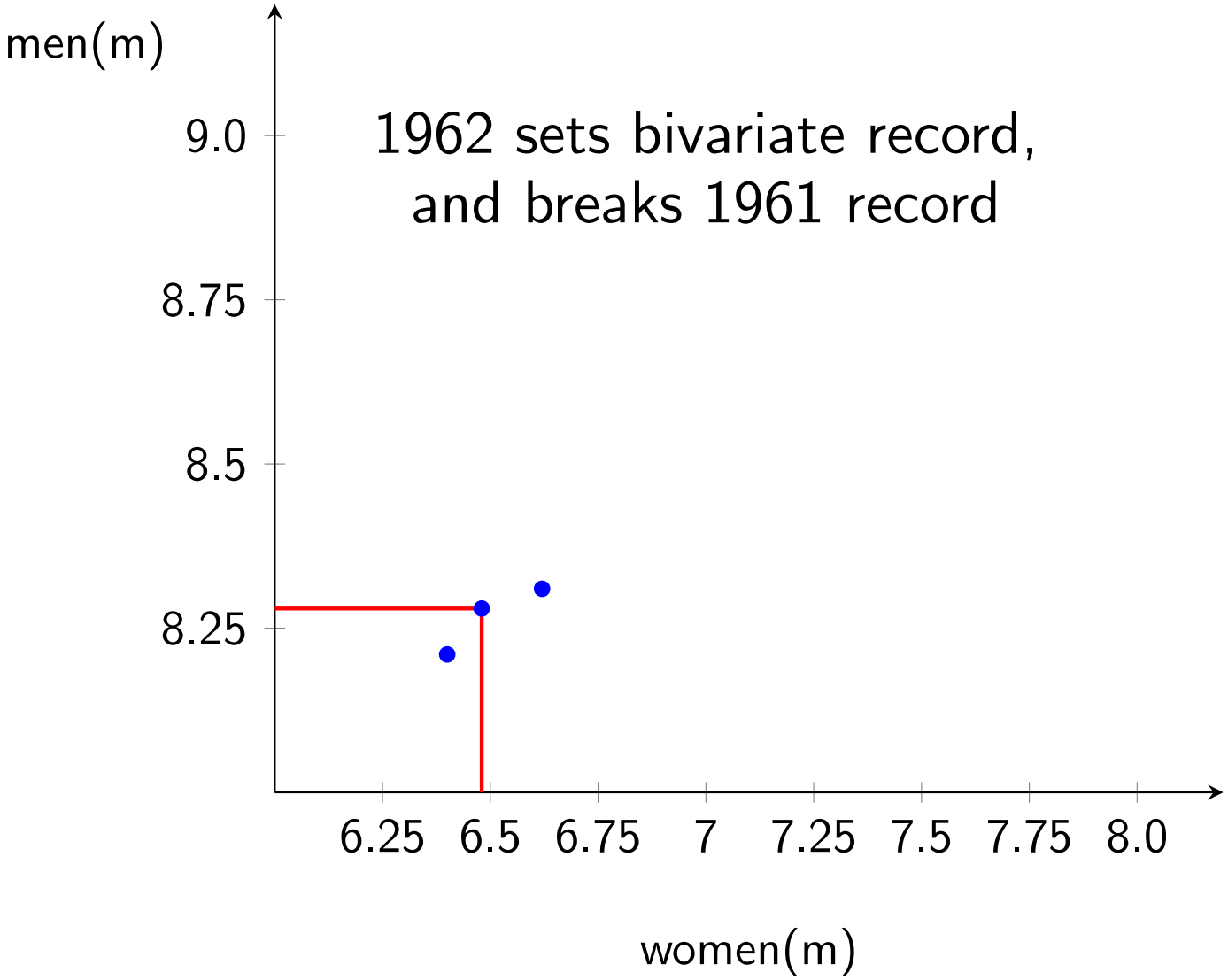
Men and Women's Best Long Jumps by Year



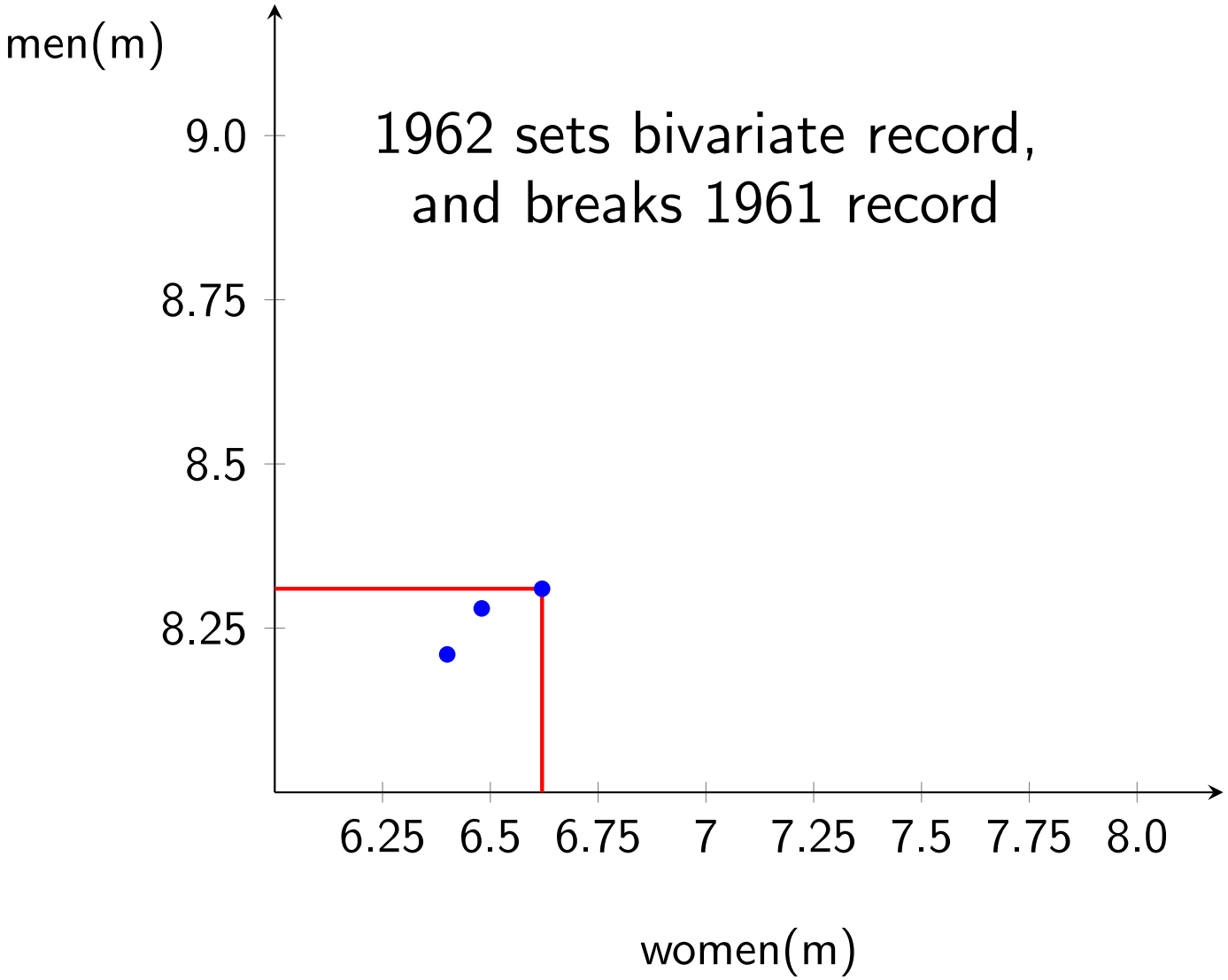
Men and Women's Best Long Jumps by Year



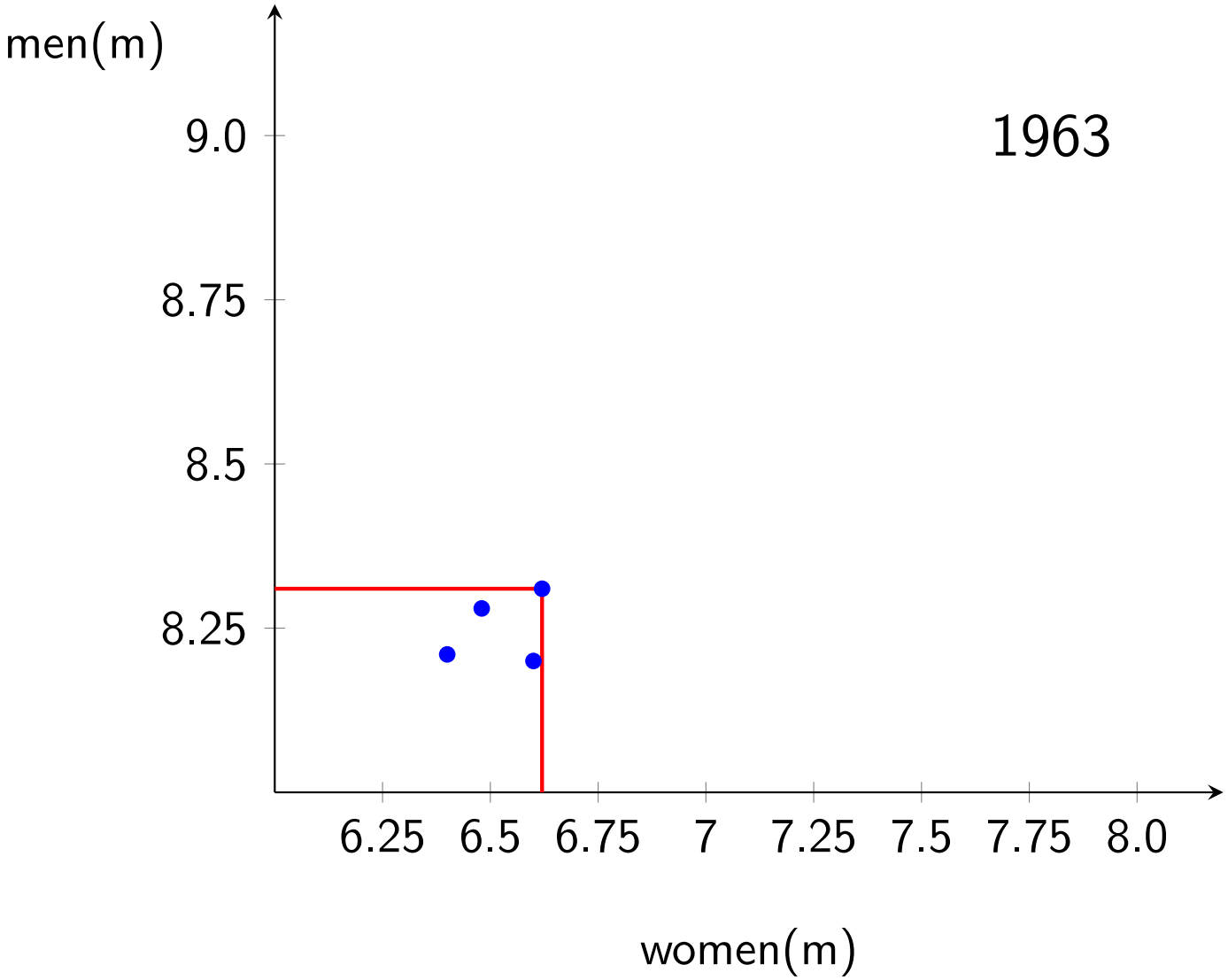
Men and Women's Best Long Jumps by Year



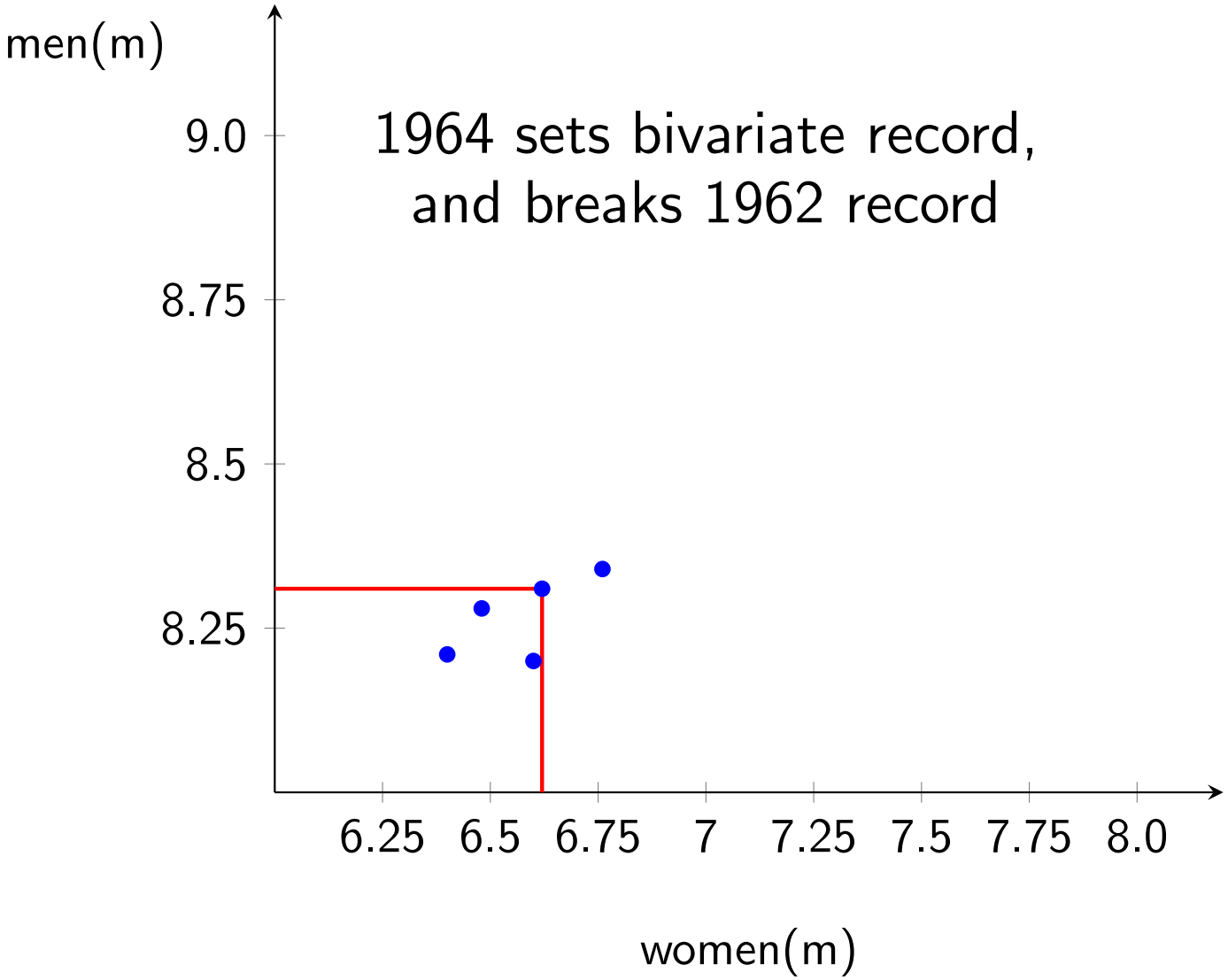
Men and Women's Best Long Jumps by Year



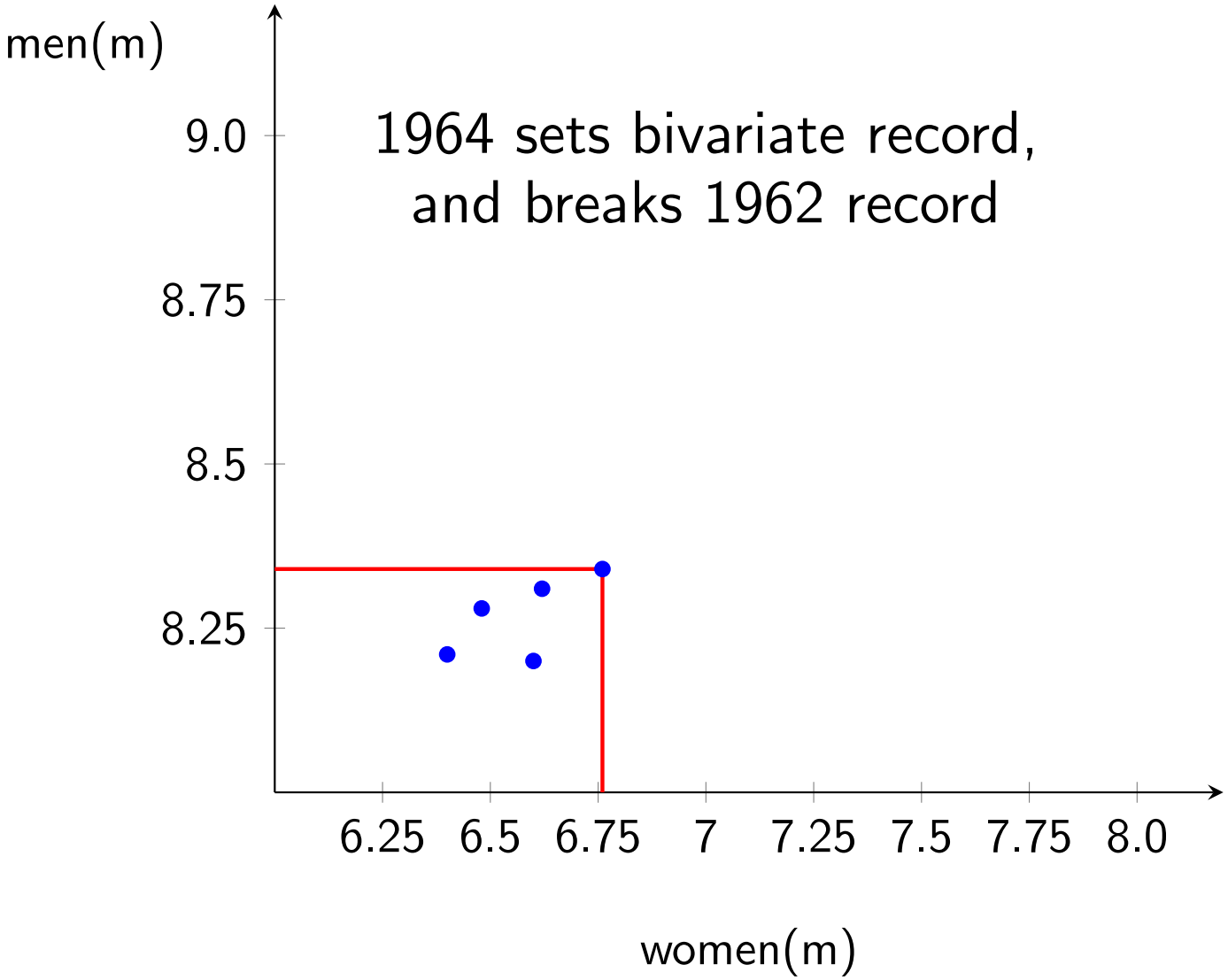
Men and Women's Best Long Jumps by Year



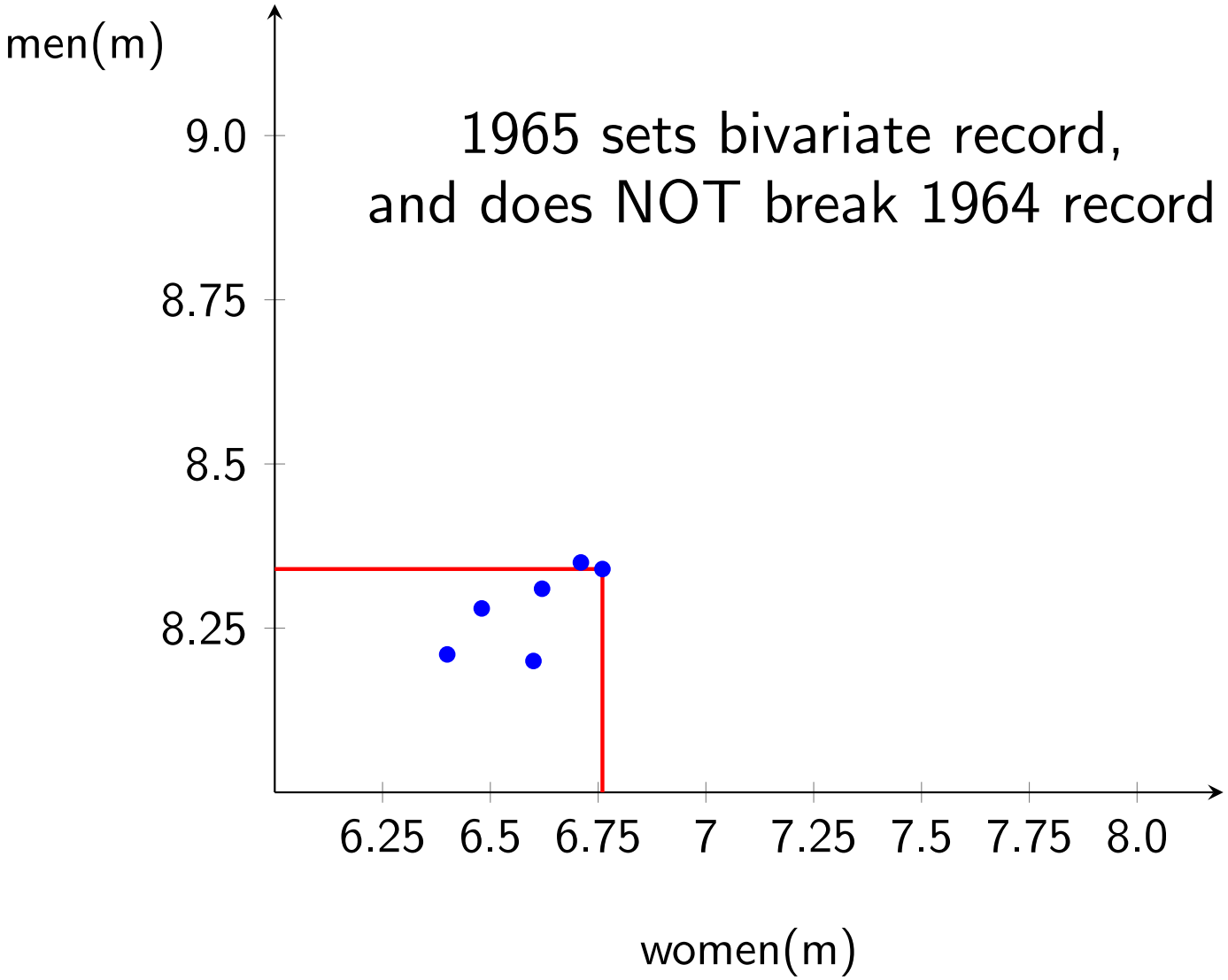
Men and Women's Best Long Jumps by Year



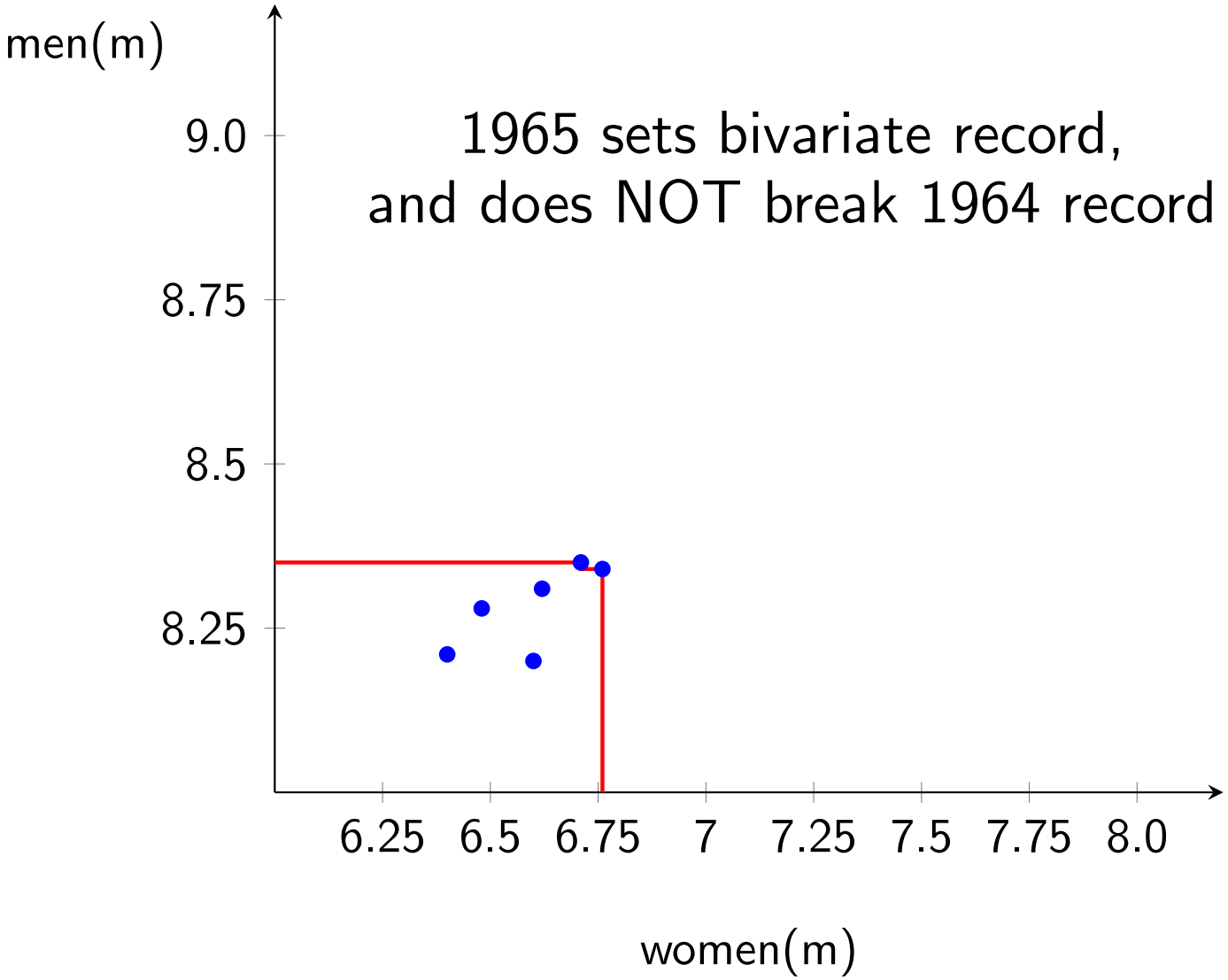
Men and Women's Best Long Jumps by Year



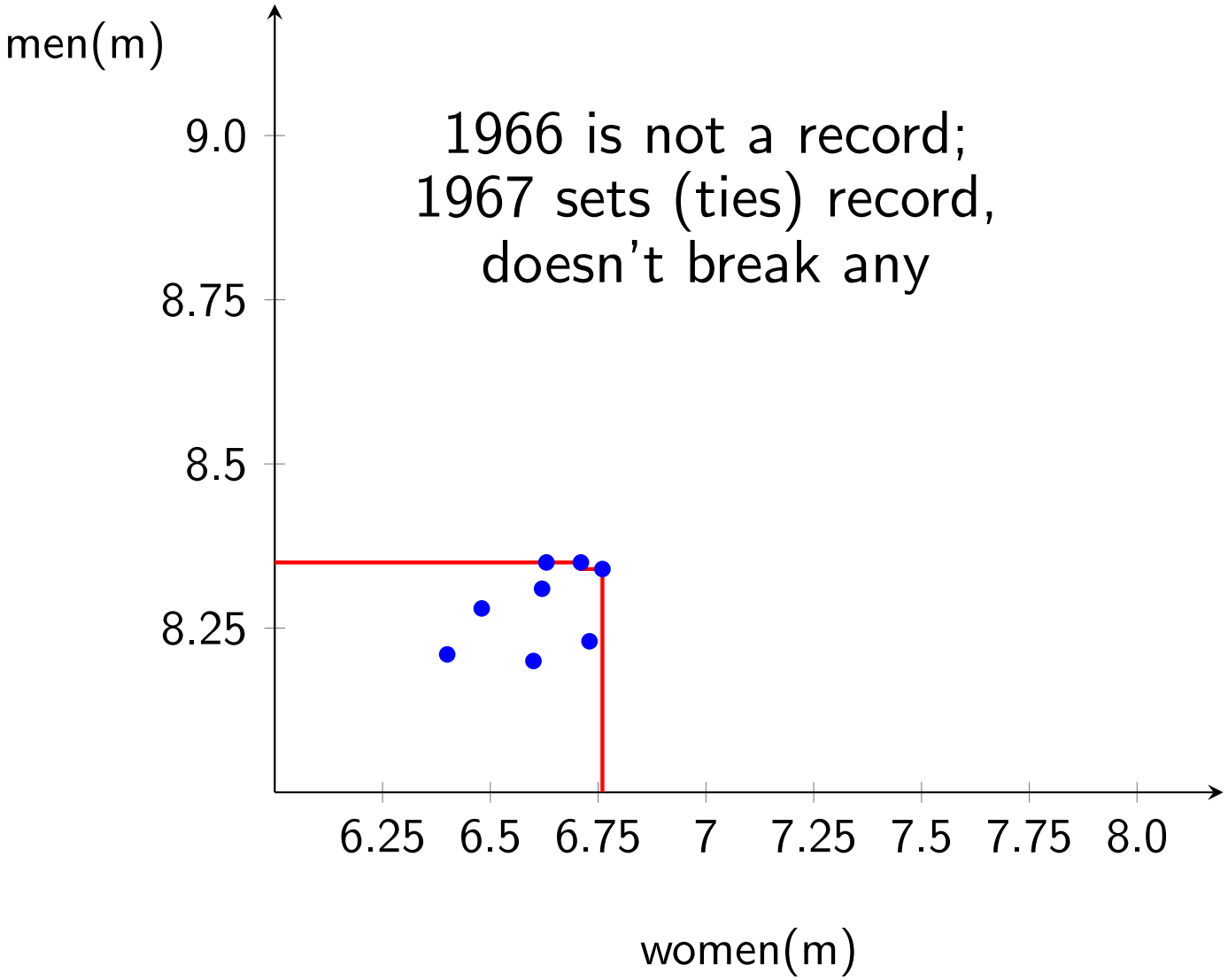
Men and Women's Best Long Jumps by Year



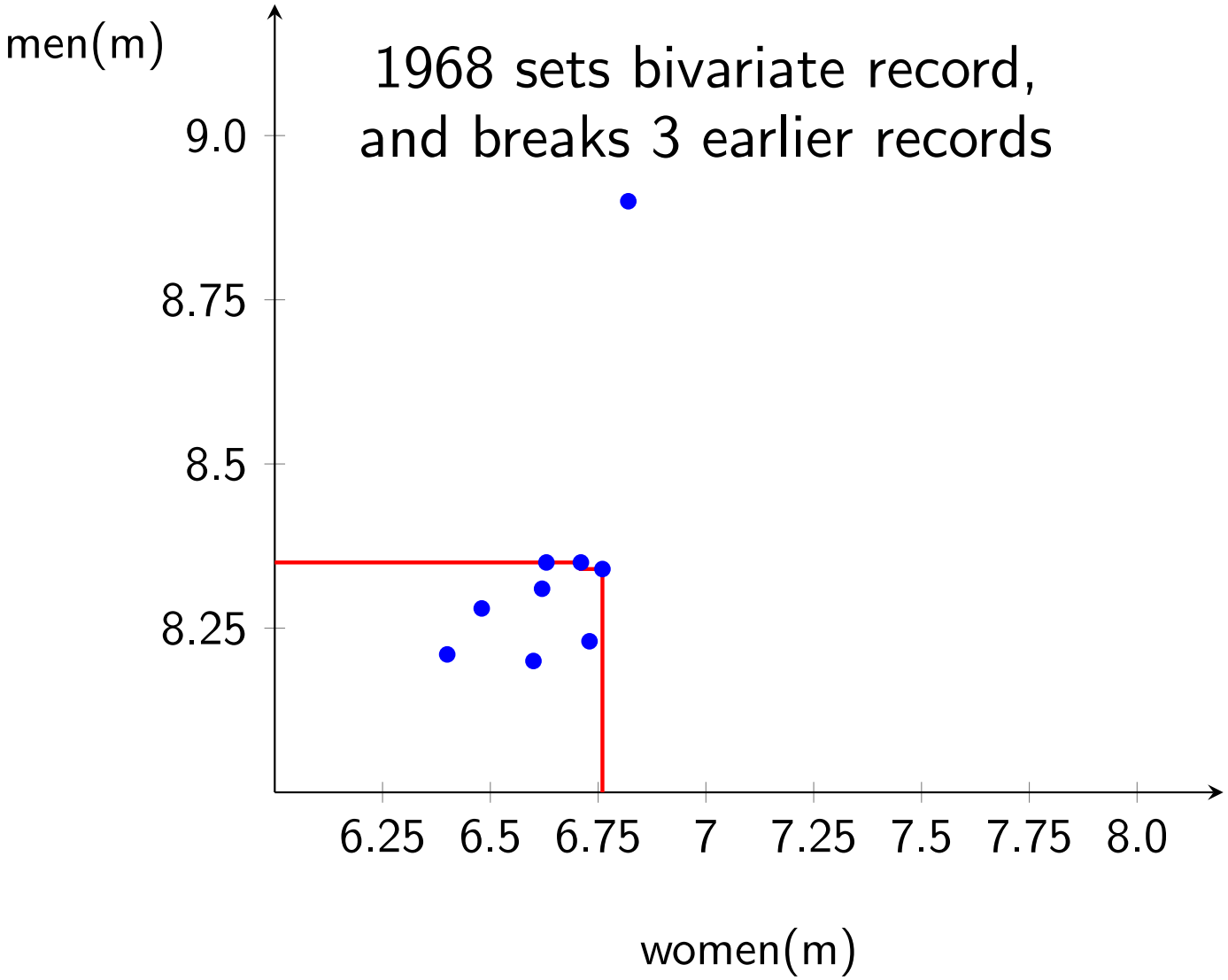
Men and Women's Best Long Jumps by Year



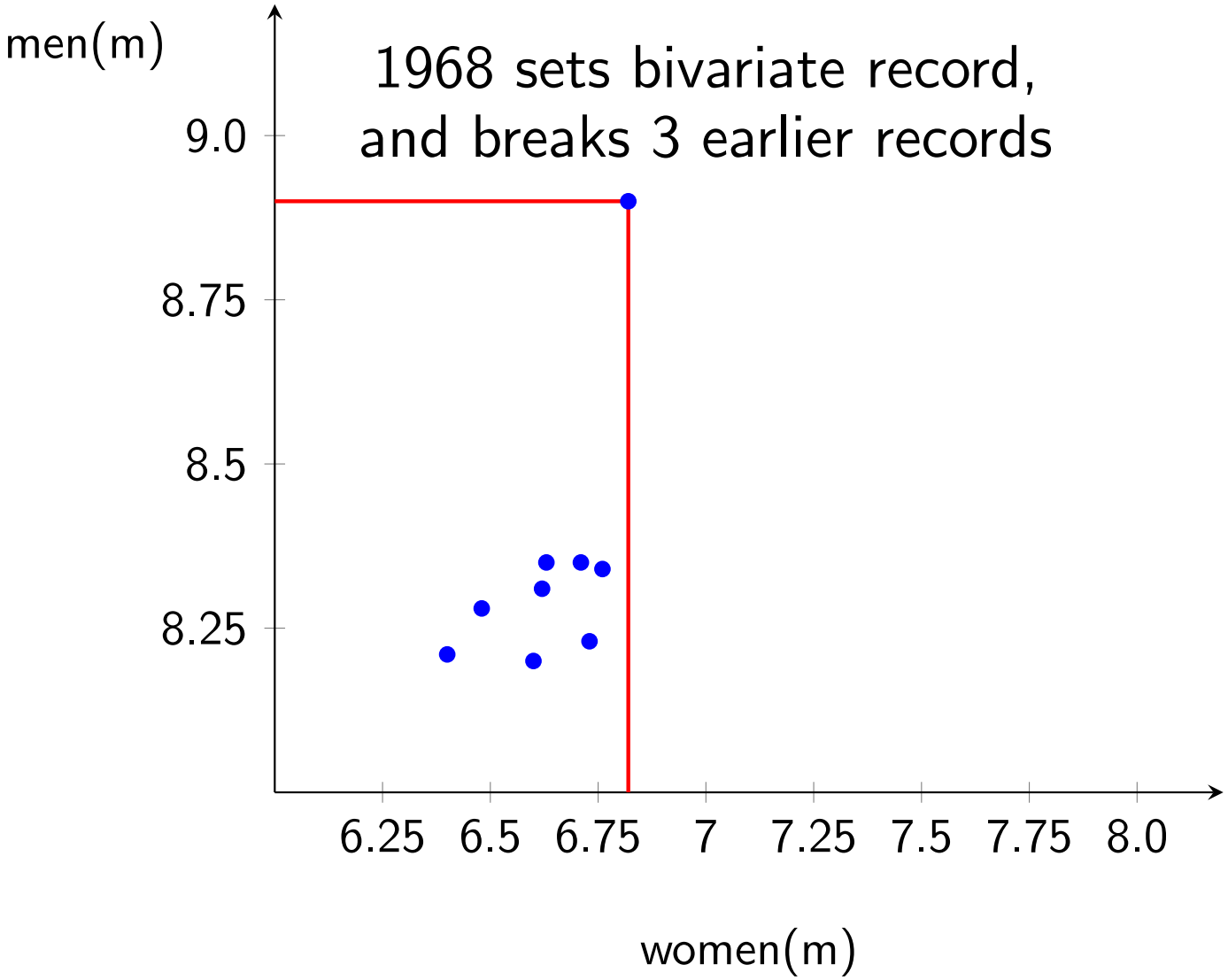
Men and Women's Best Long Jumps by Year



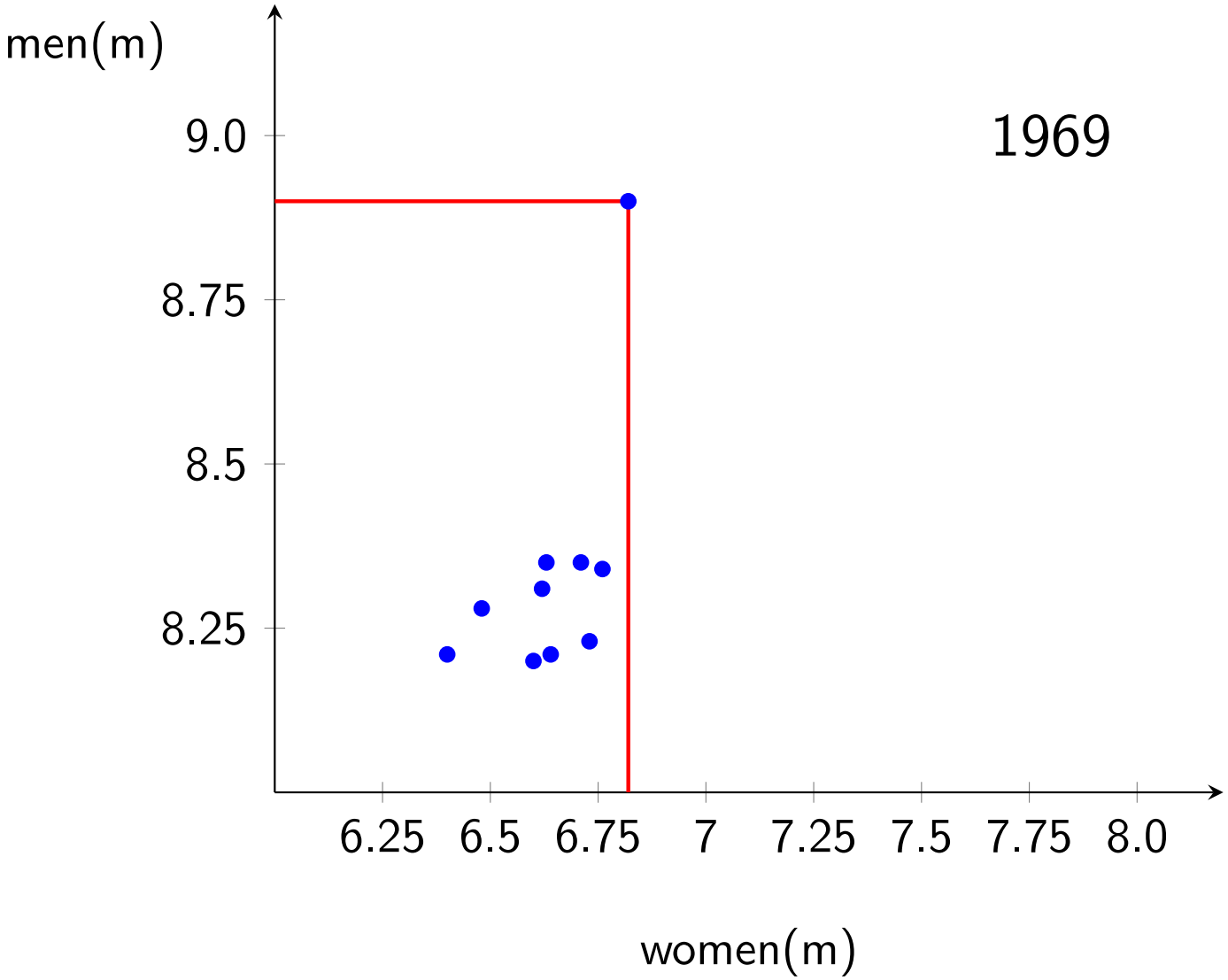
Men and Women's Best Long Jumps by Year



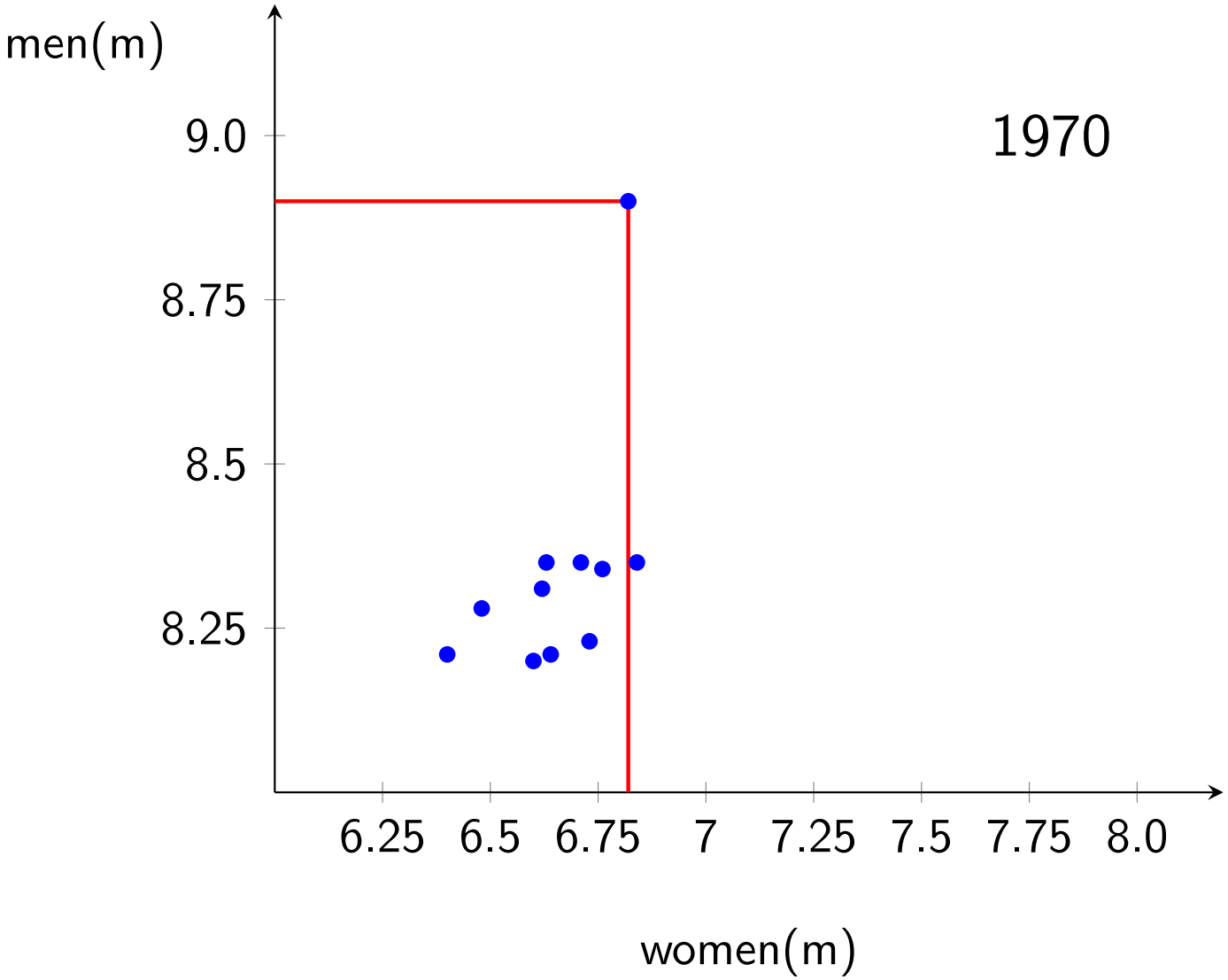
Men and Women's Best Long Jumps by Year



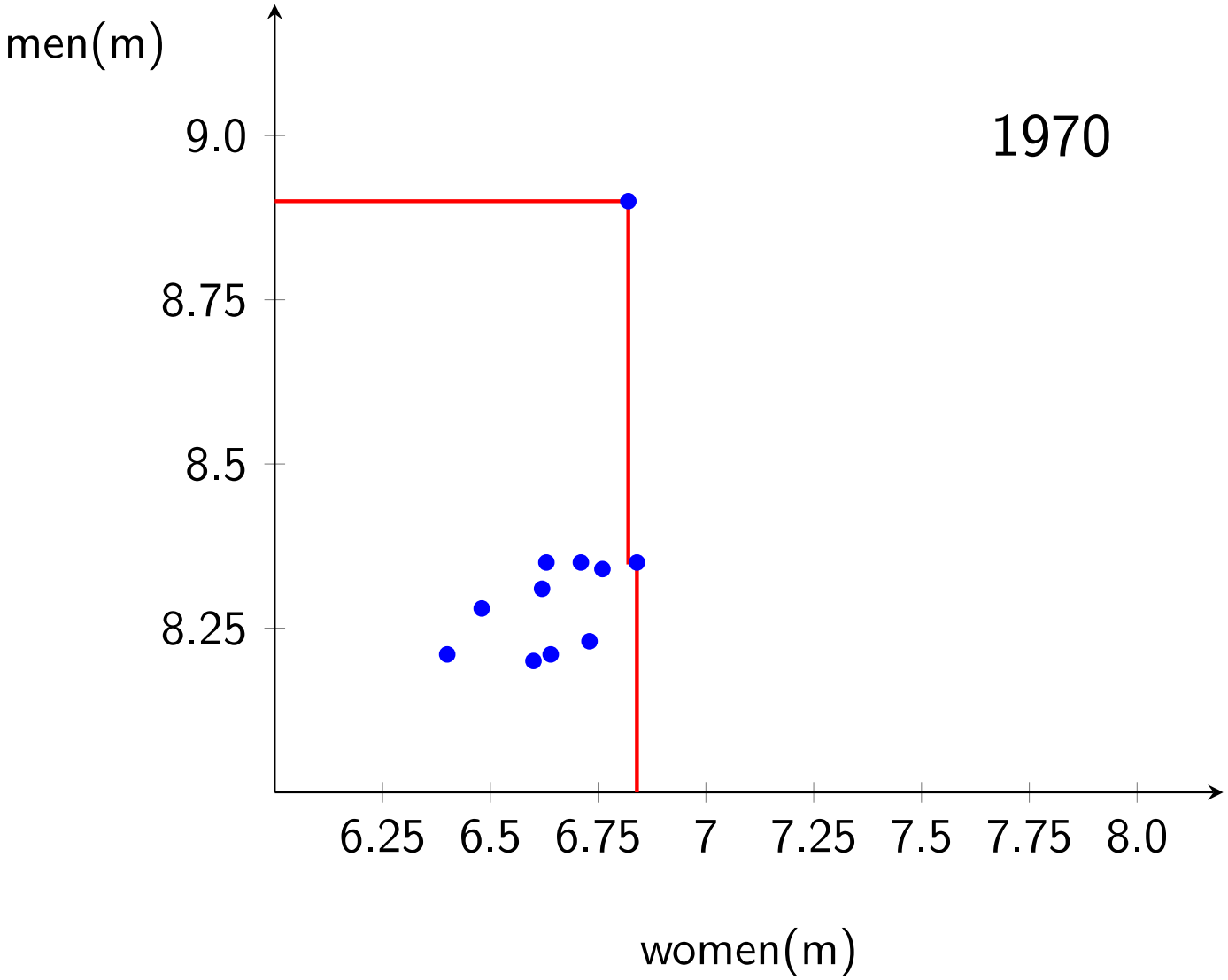
Men and Women's Best Long Jumps by Year



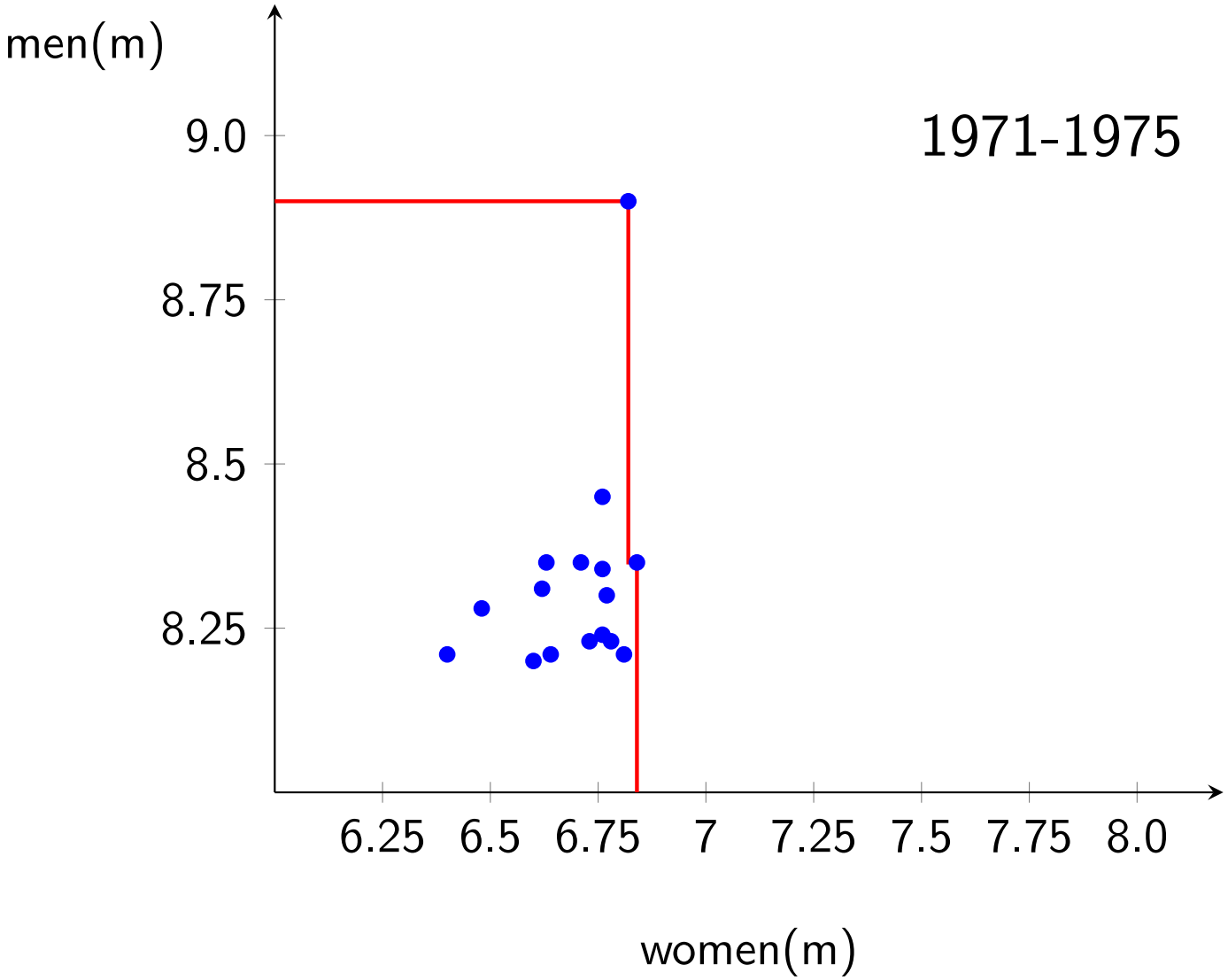
Men and Women's Best Long Jumps by Year



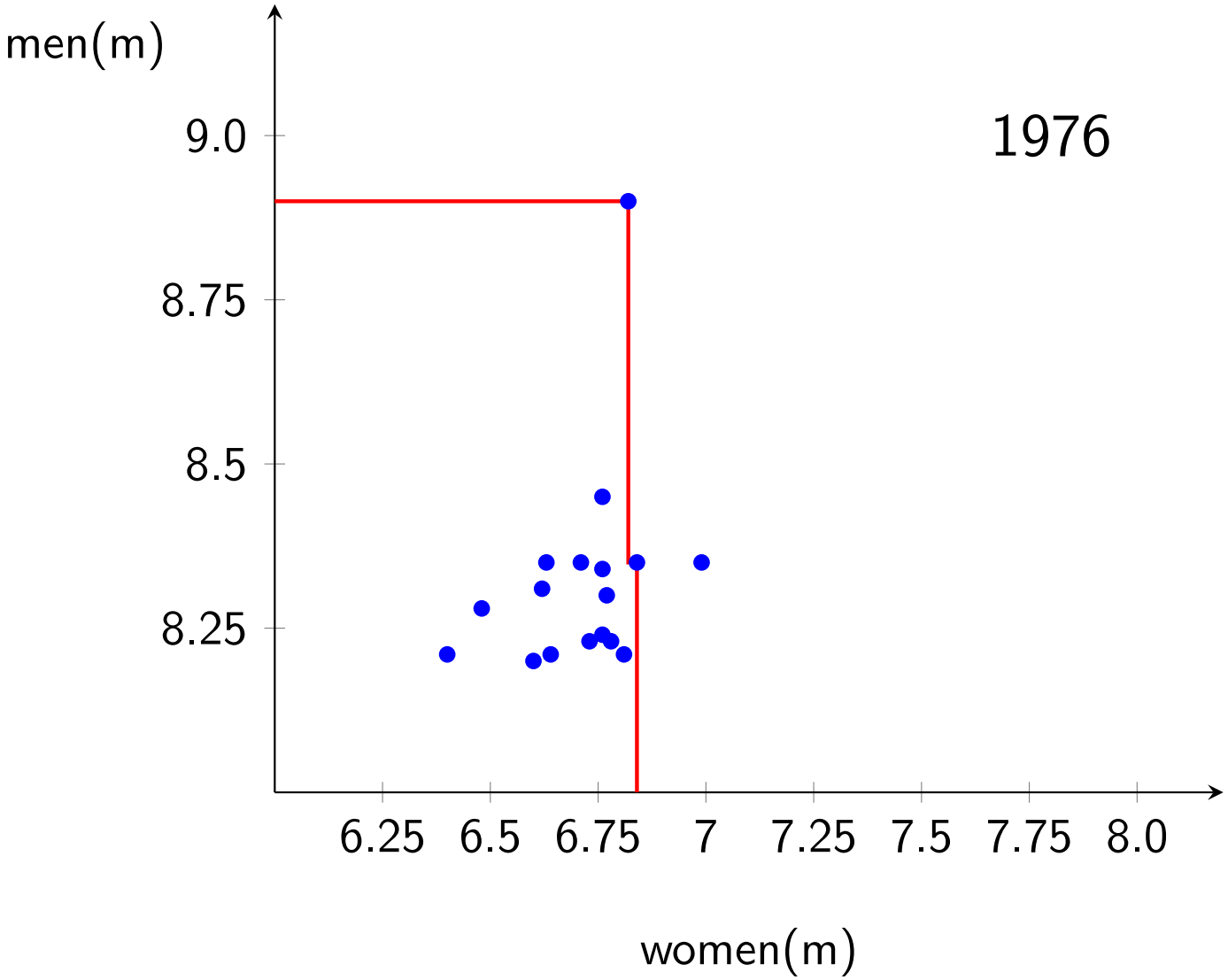
Men and Women's Best Long Jumps by Year



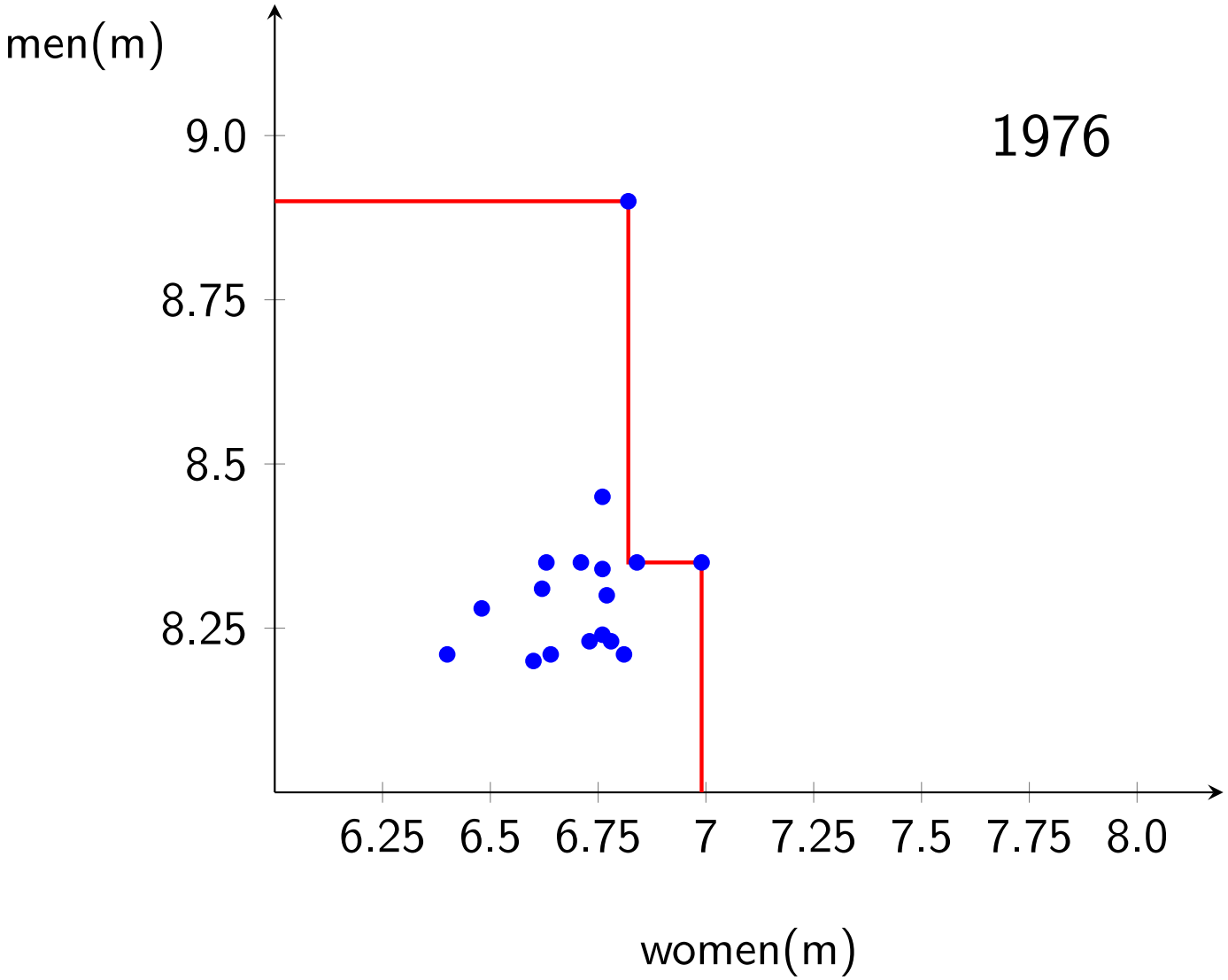
Men and Women's Best Long Jumps by Year



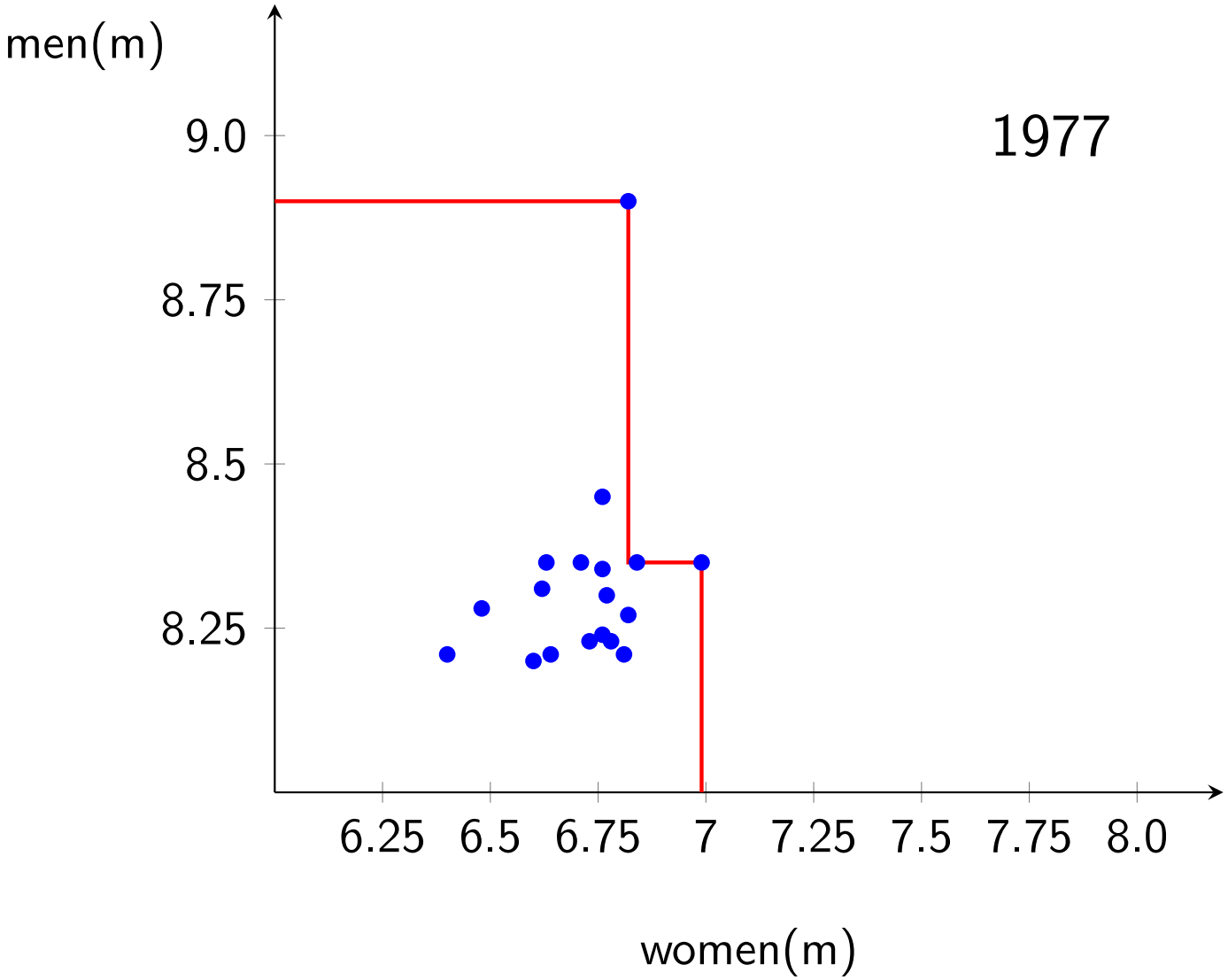
Men and Women's Best Long Jumps by Year



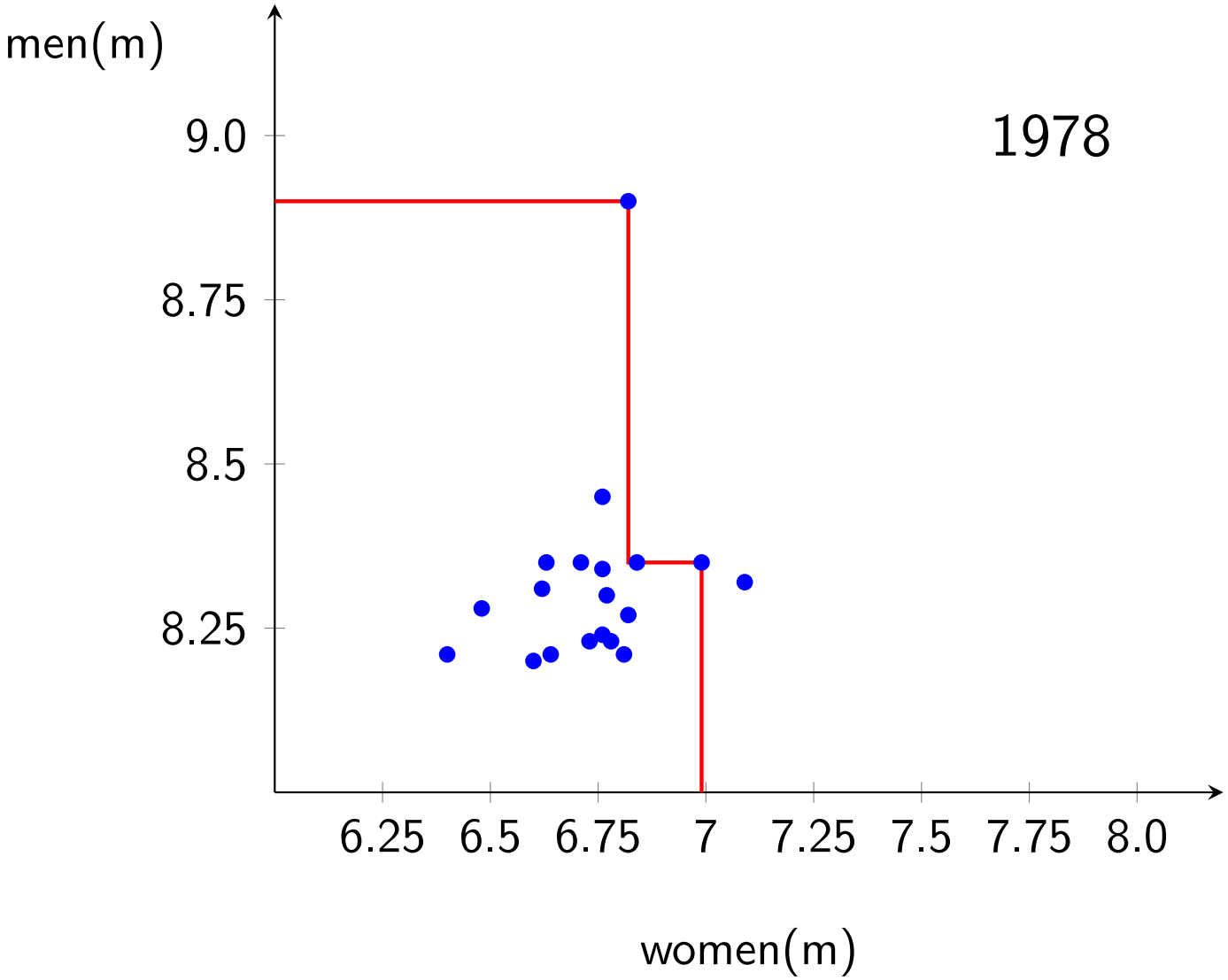
Men and Women's Best Long Jumps by Year



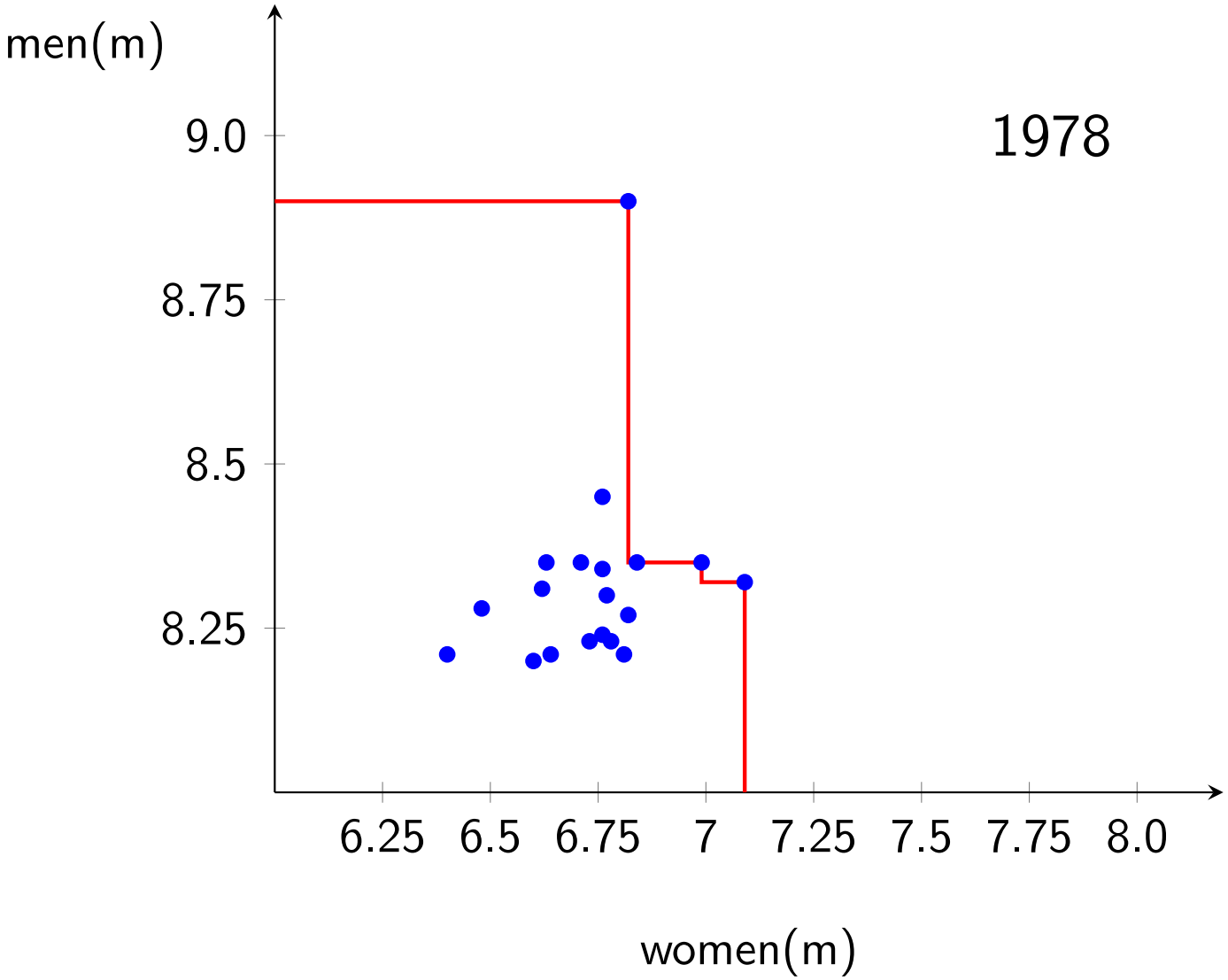
Men and Women's Best Long Jumps by Year



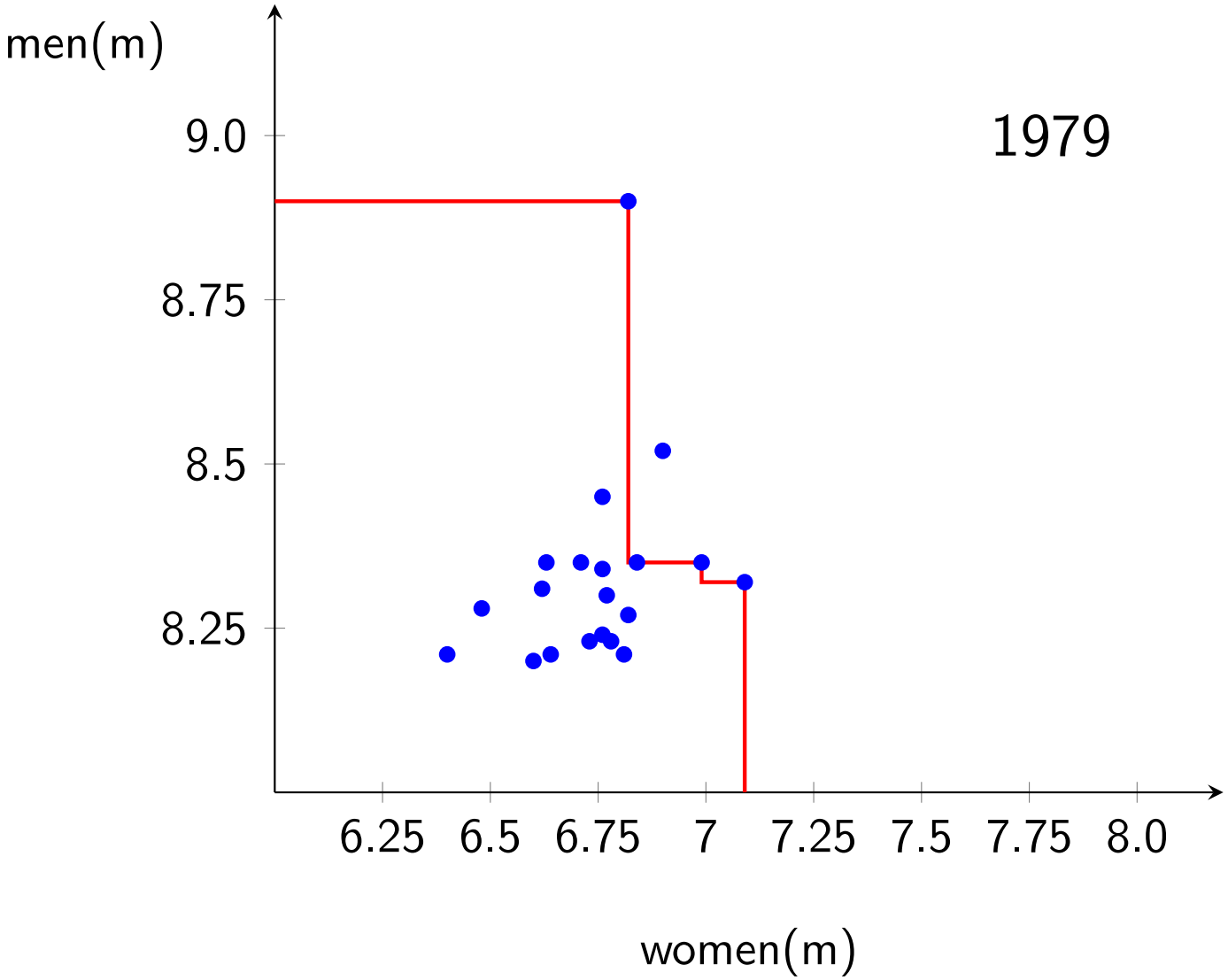
Men and Women's Best Long Jumps by Year



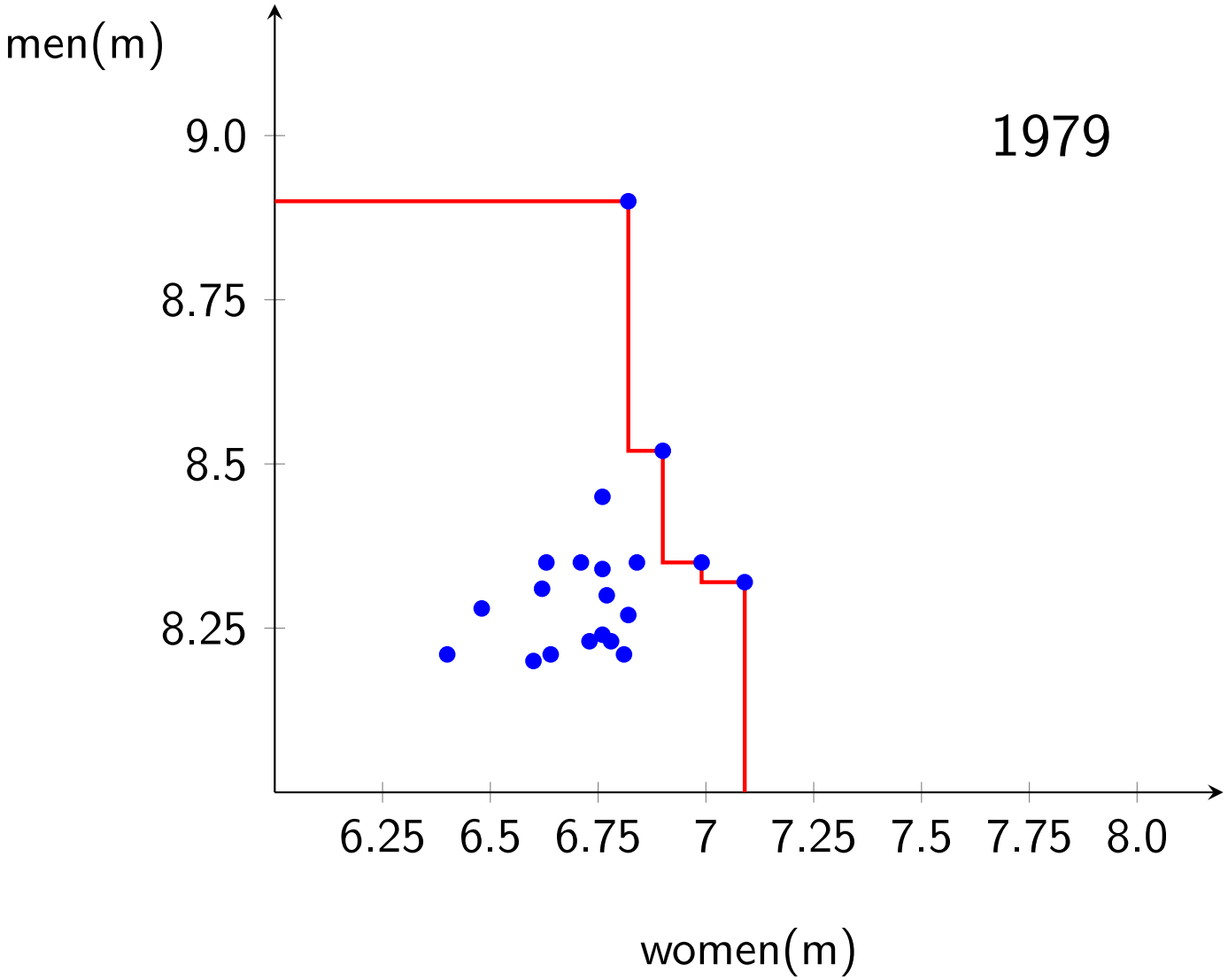
Men and Women's Best Long Jumps by Year



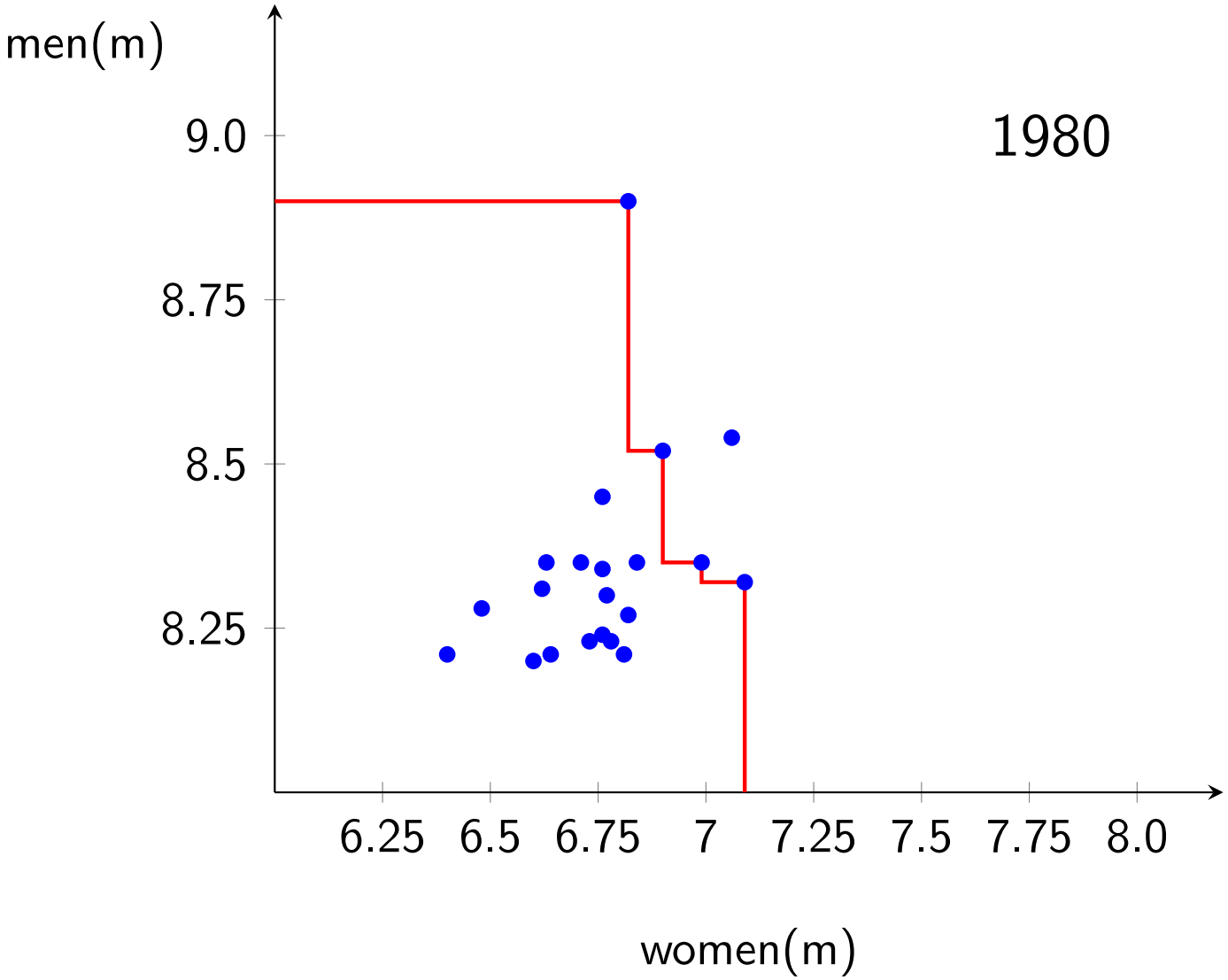
Men and Women's Best Long Jumps by Year



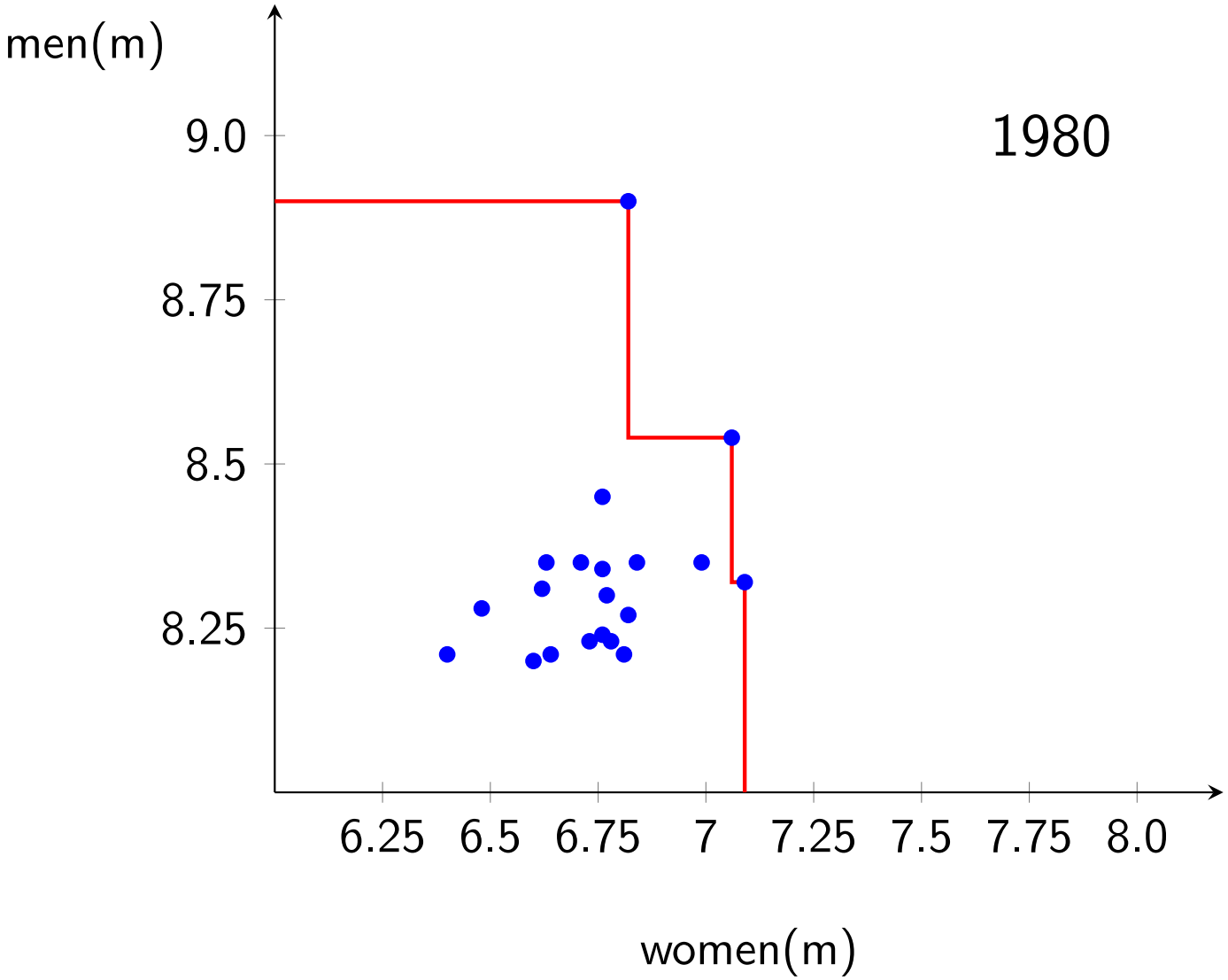
Men and Women's Best Long Jumps by Year



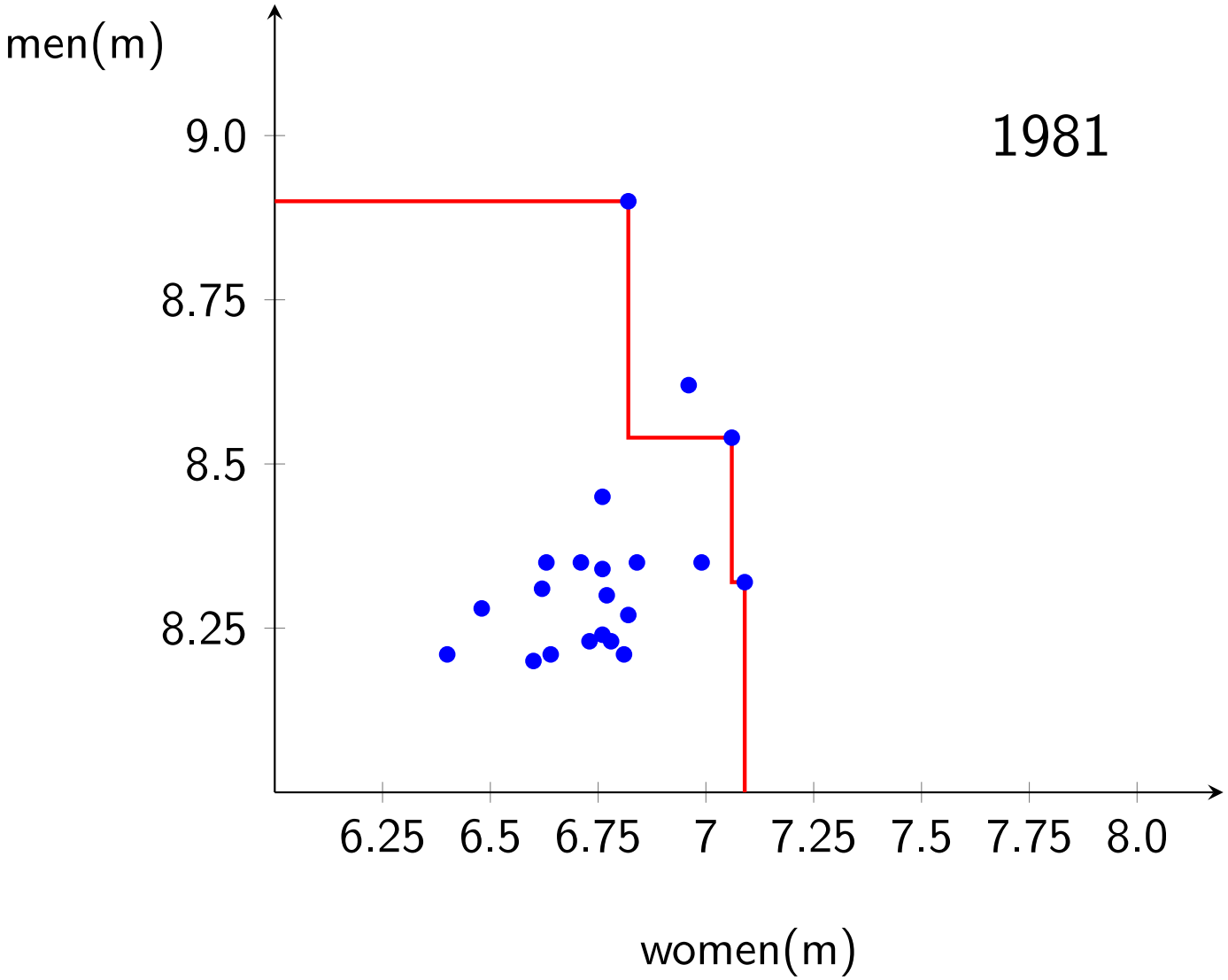
Men and Women's Best Long Jumps by Year



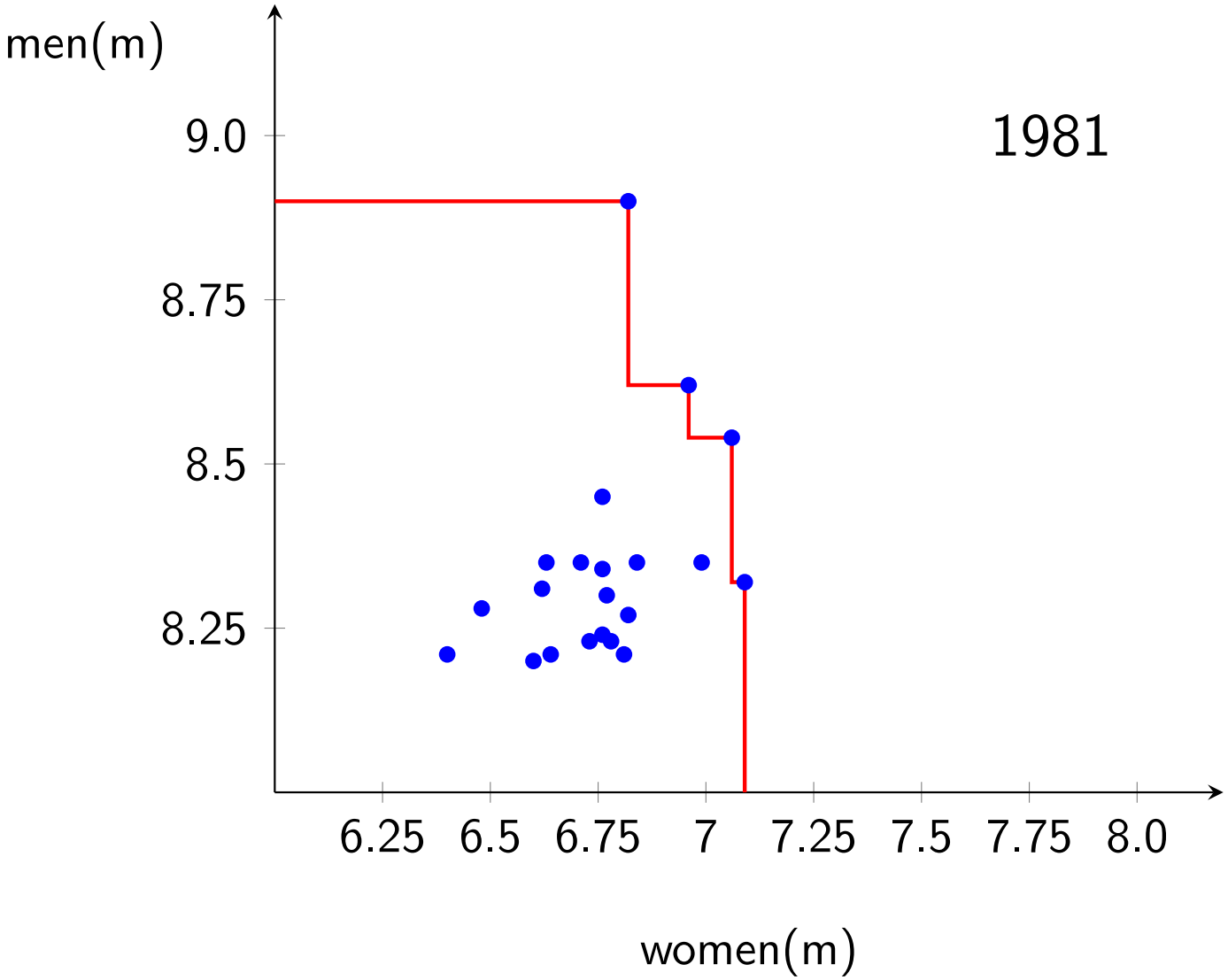
Men and Women's Best Long Jumps by Year



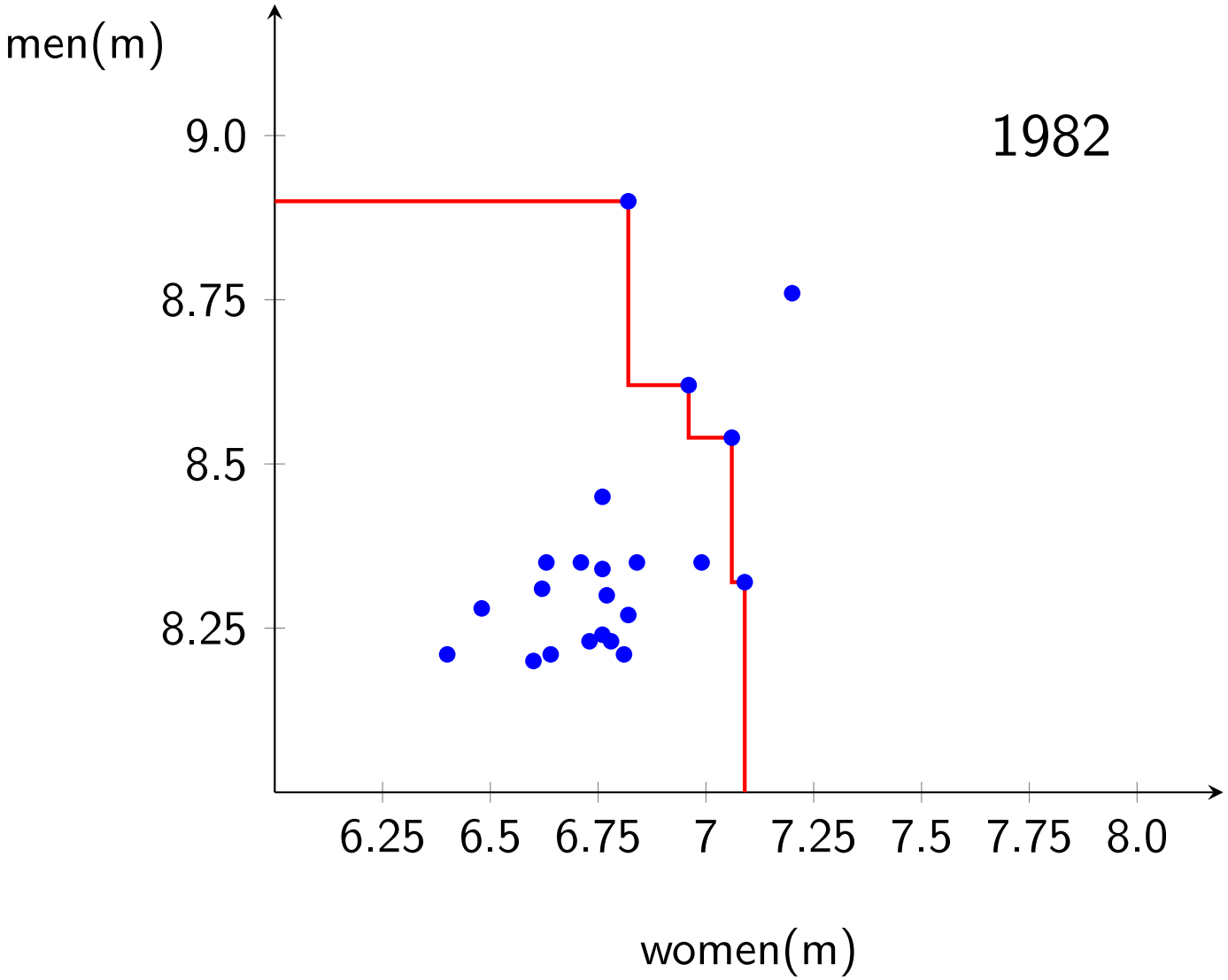
Men and Women's Best Long Jumps by Year



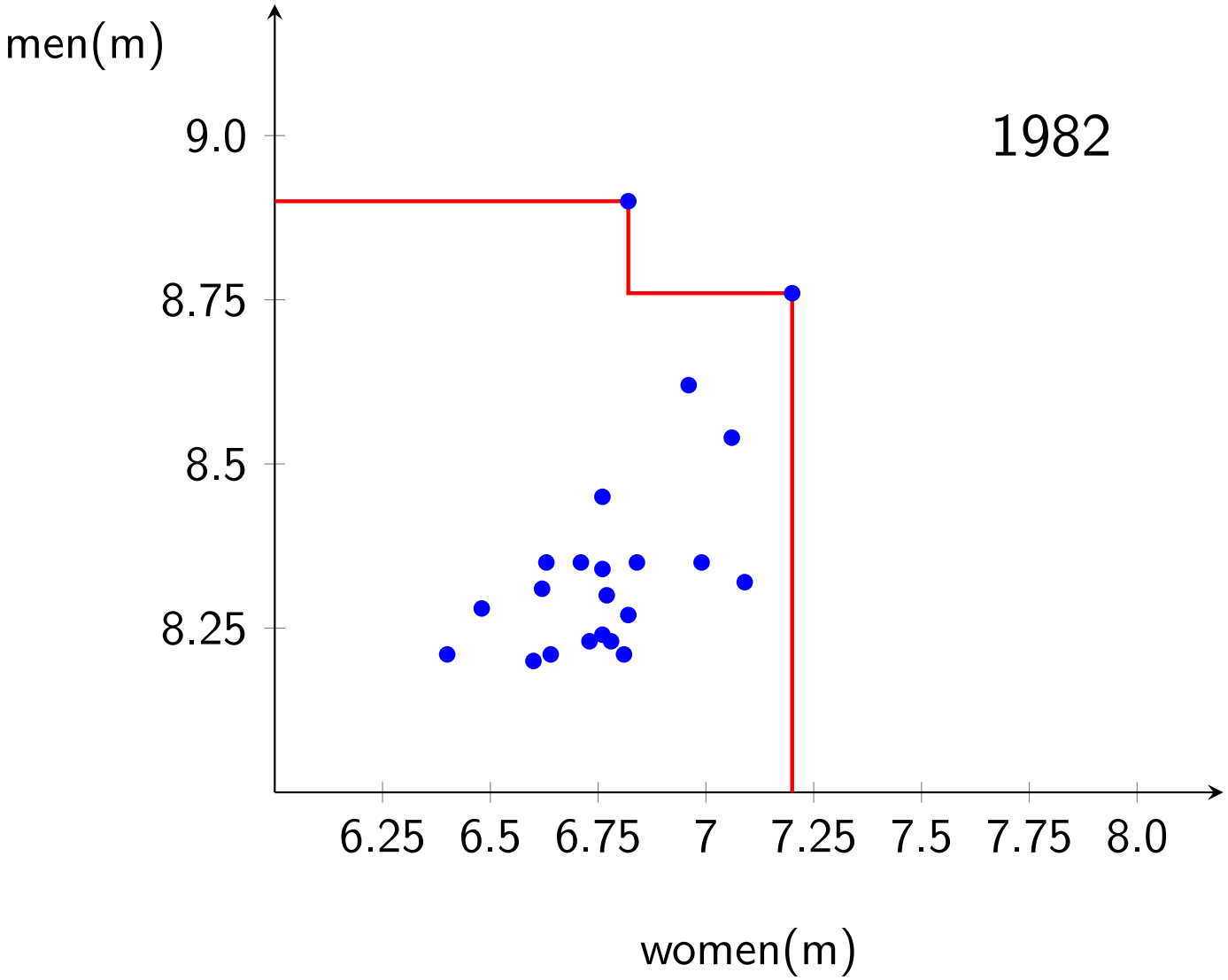
Men and Women's Best Long Jumps by Year



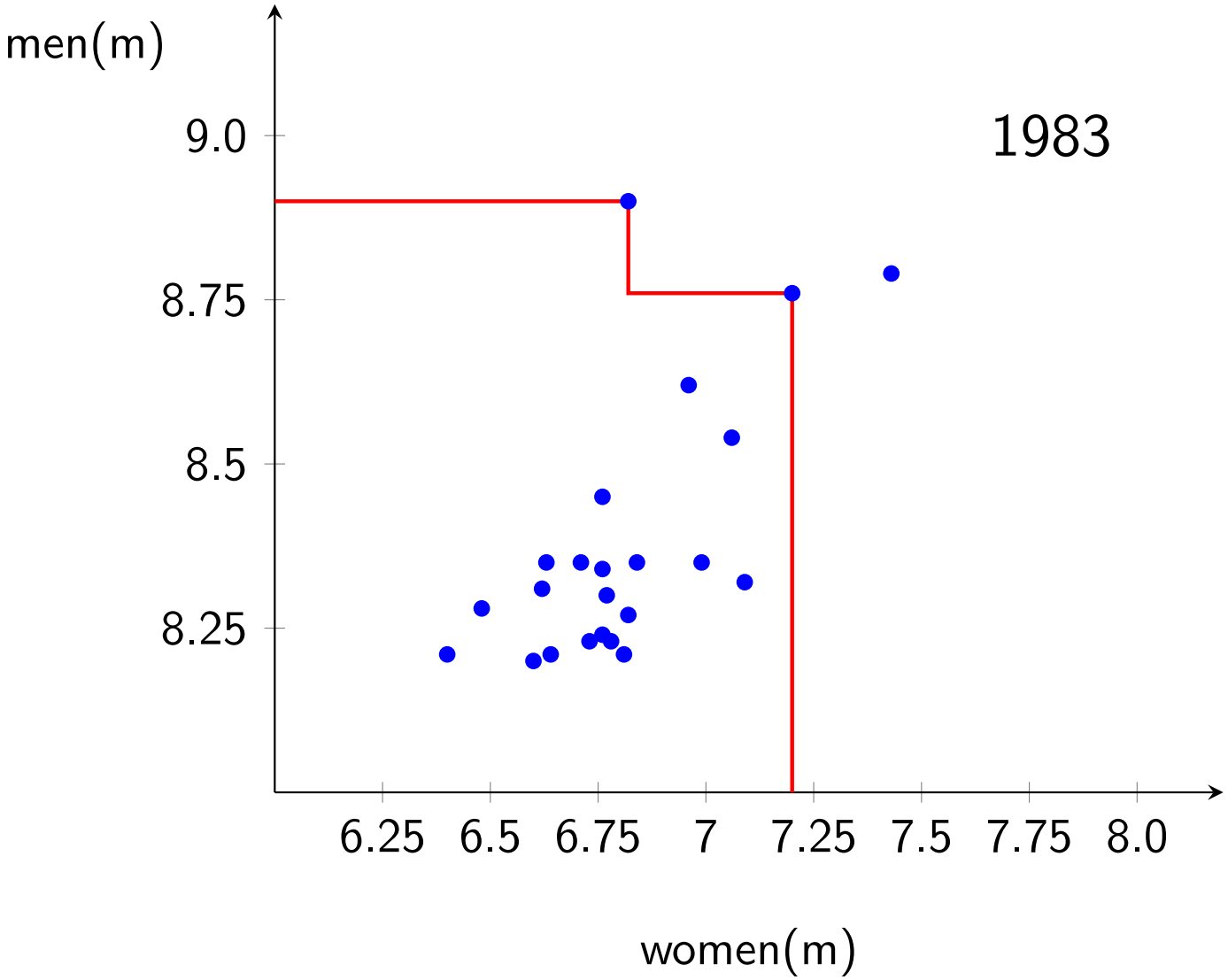
Men and Women's Best Long Jumps by Year



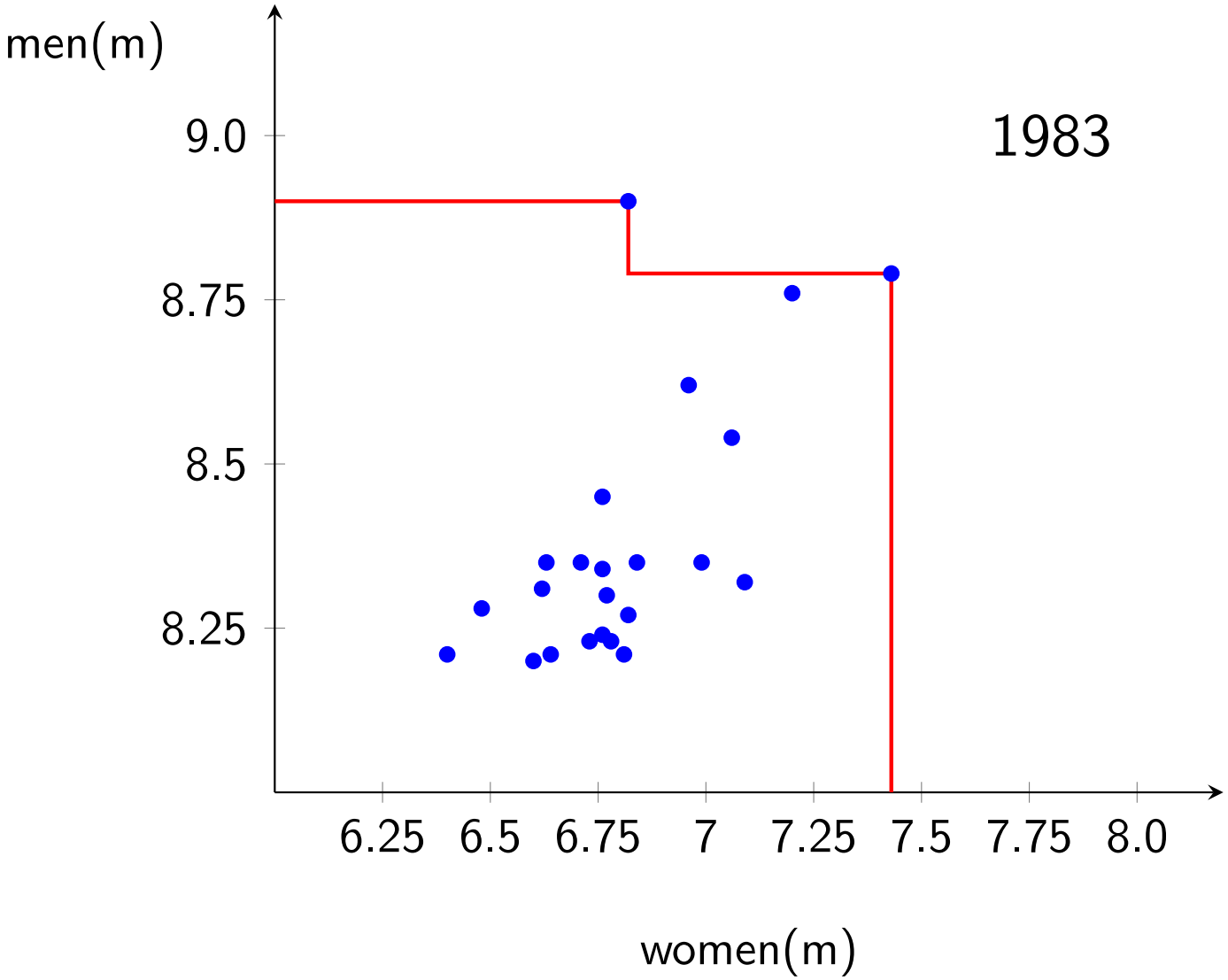
Men and Women's Best Long Jumps by Year



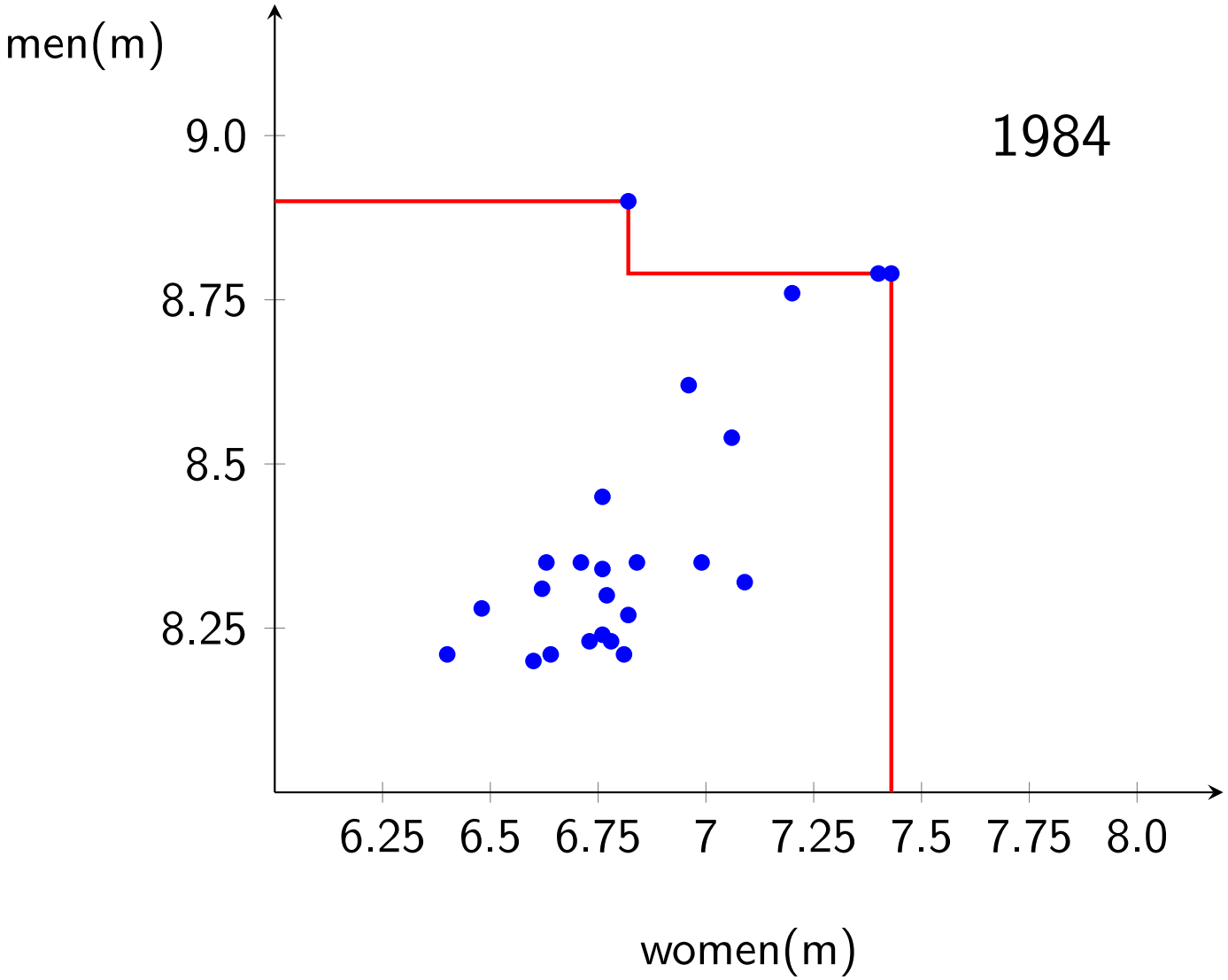
Men and Women's Best Long Jumps by Year



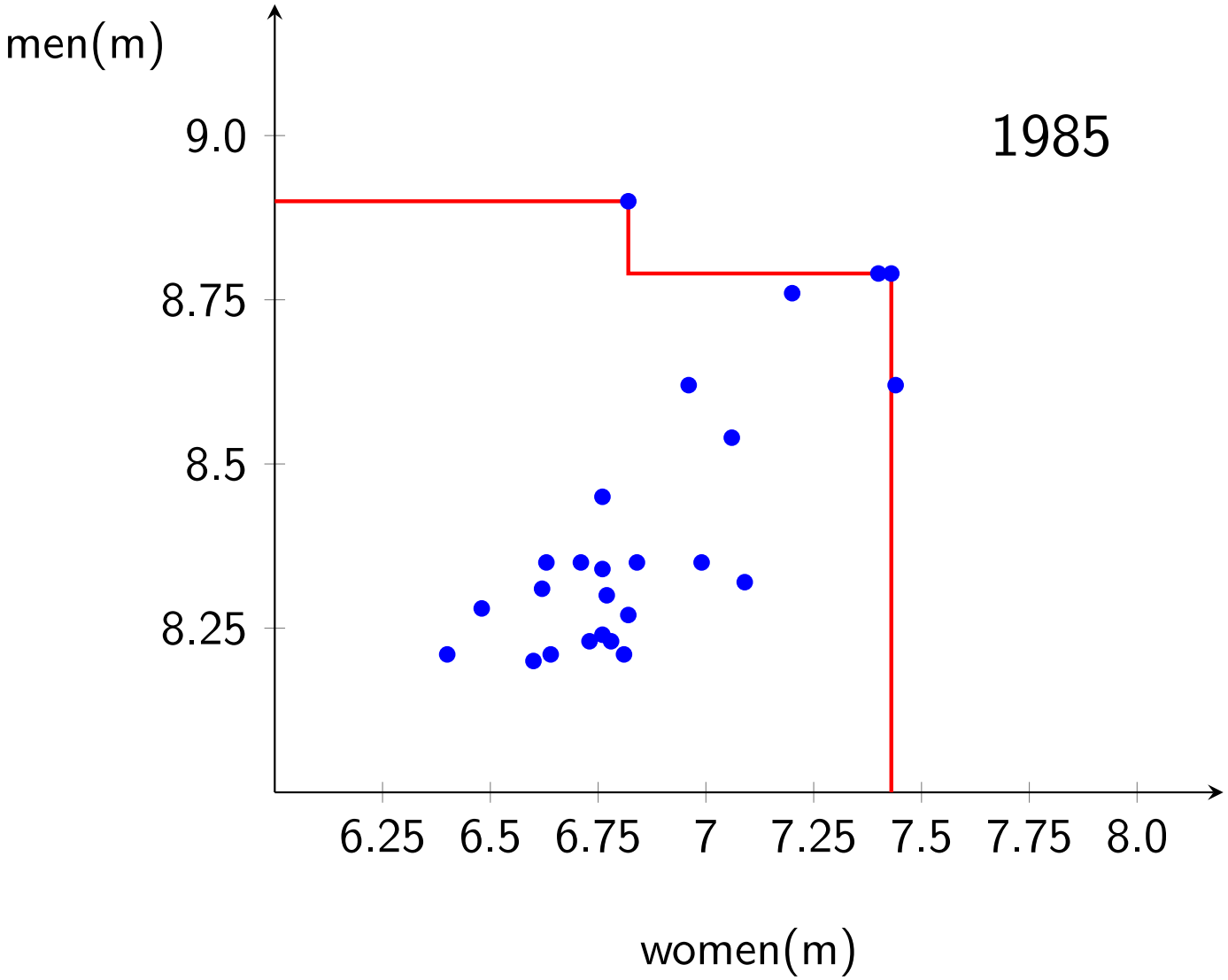
Men and Women's Best Long Jumps by Year



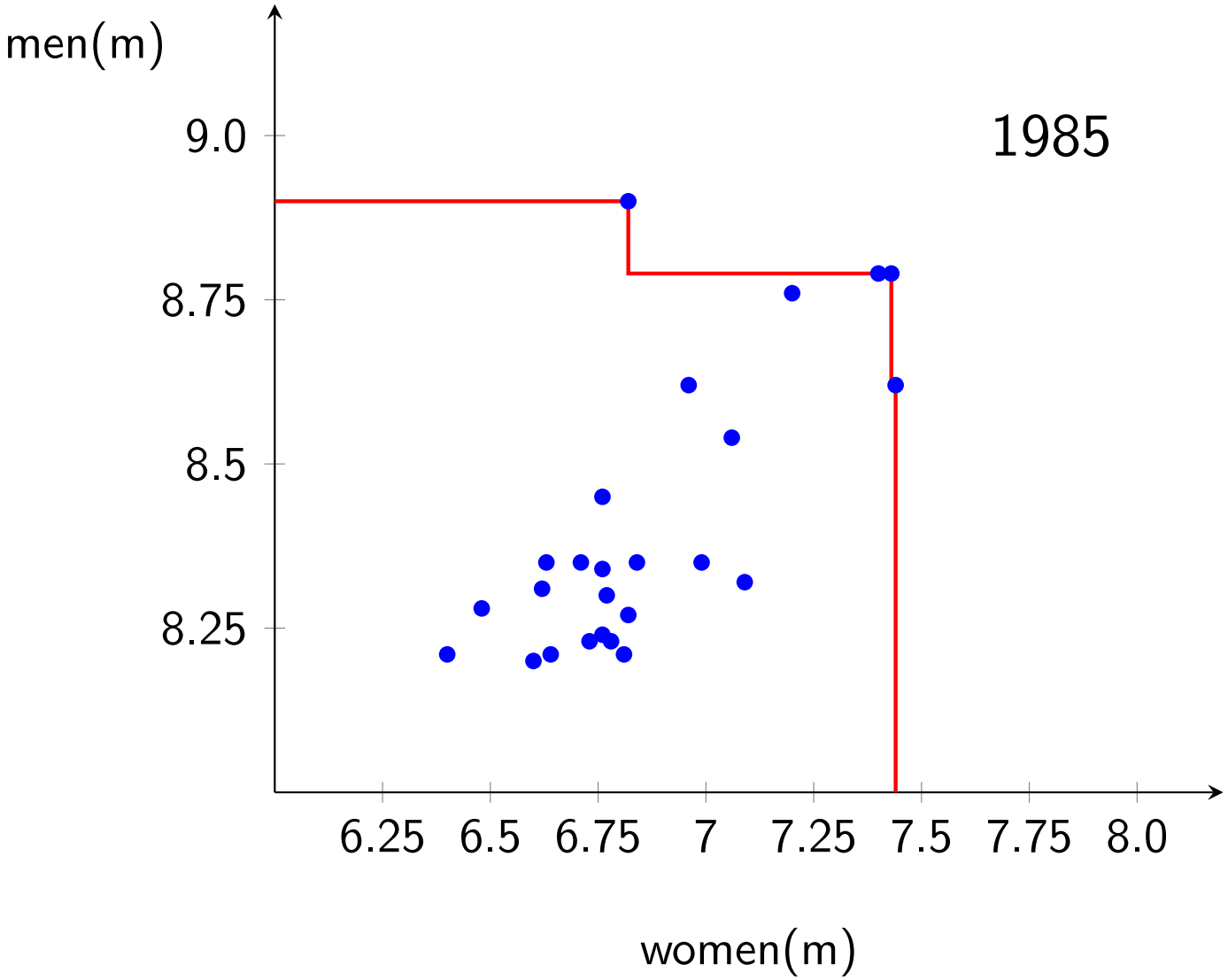
Men and Women's Best Long Jumps by Year



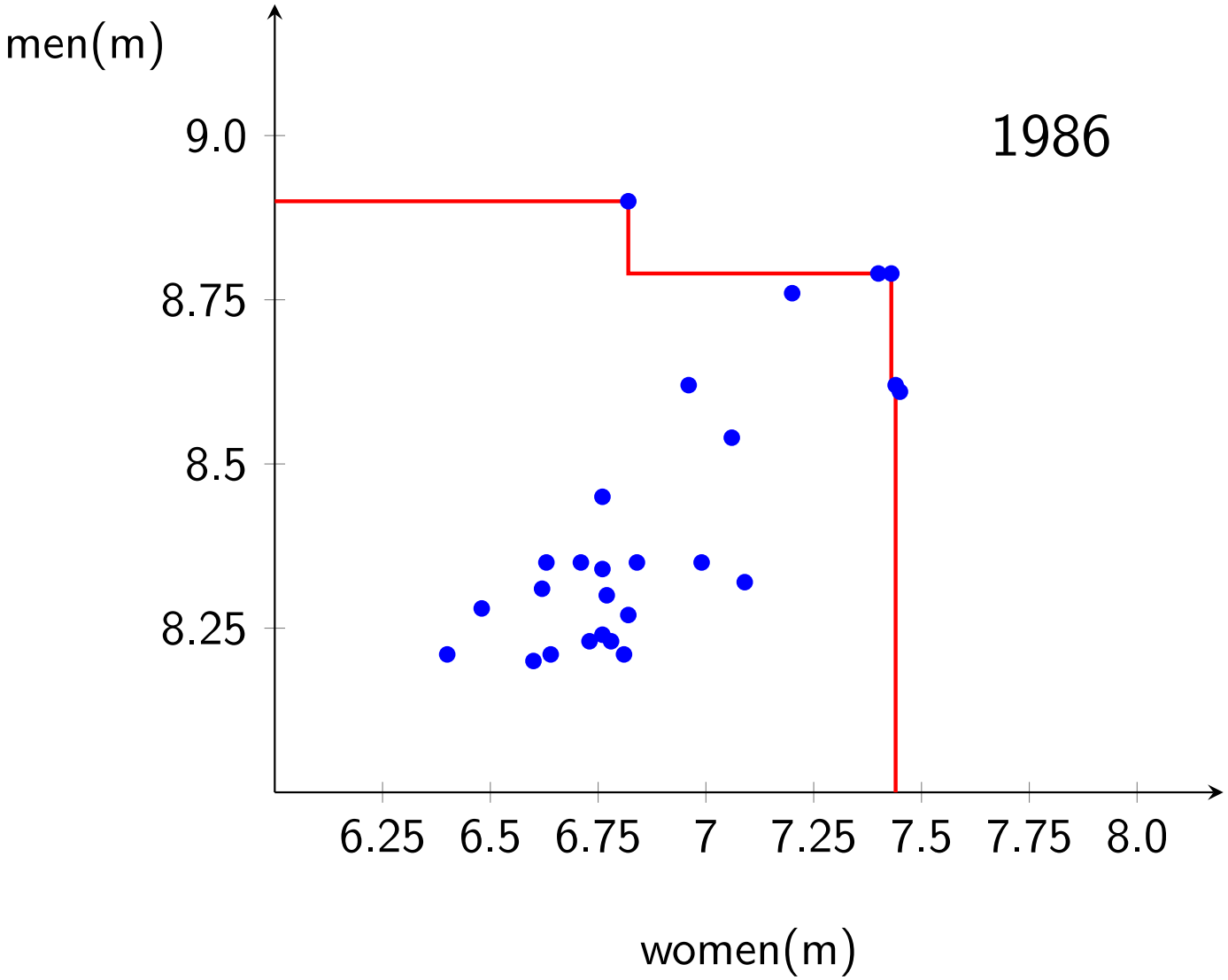
Men and Women's Best Long Jumps by Year



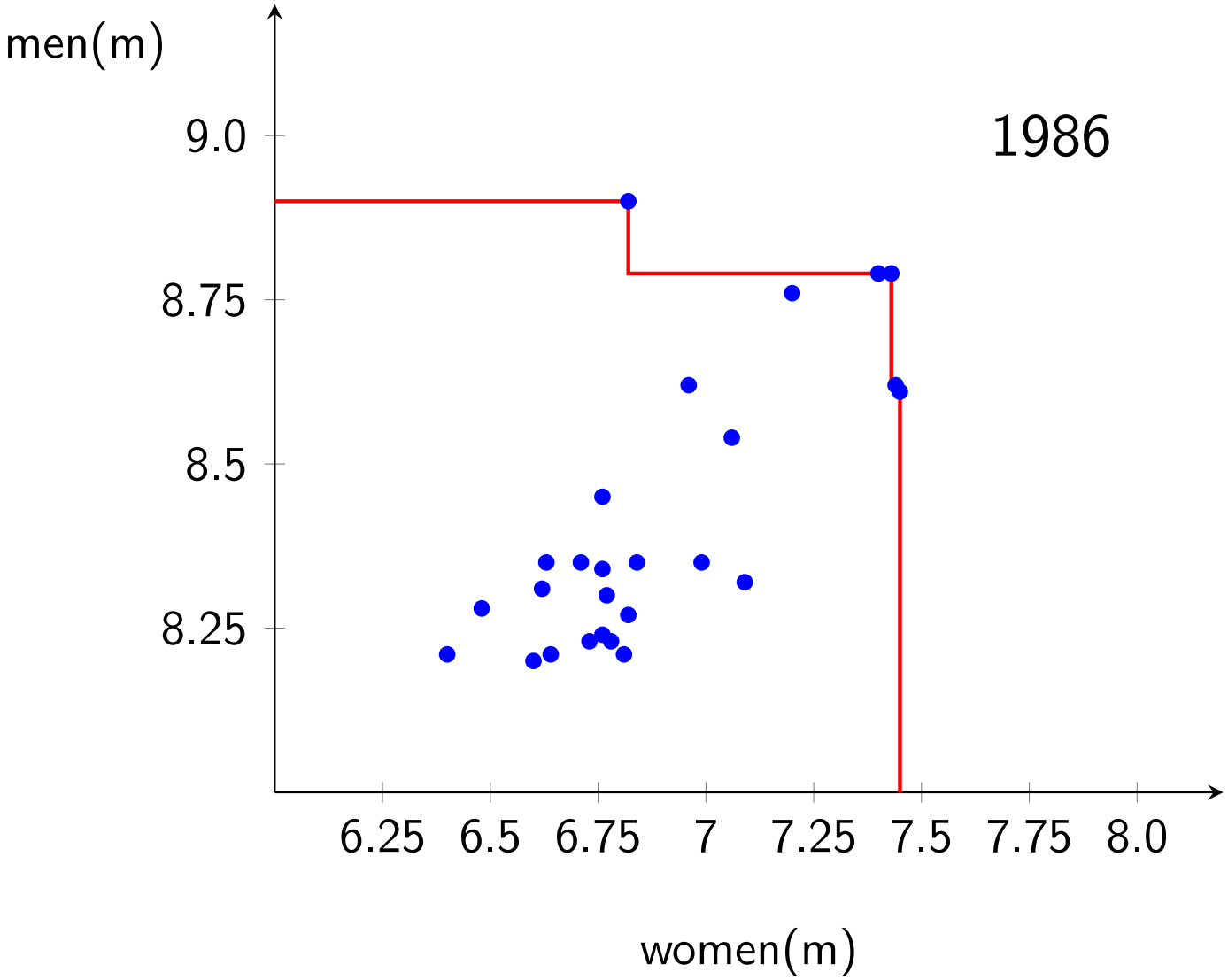
Men and Women's Best Long Jumps by Year



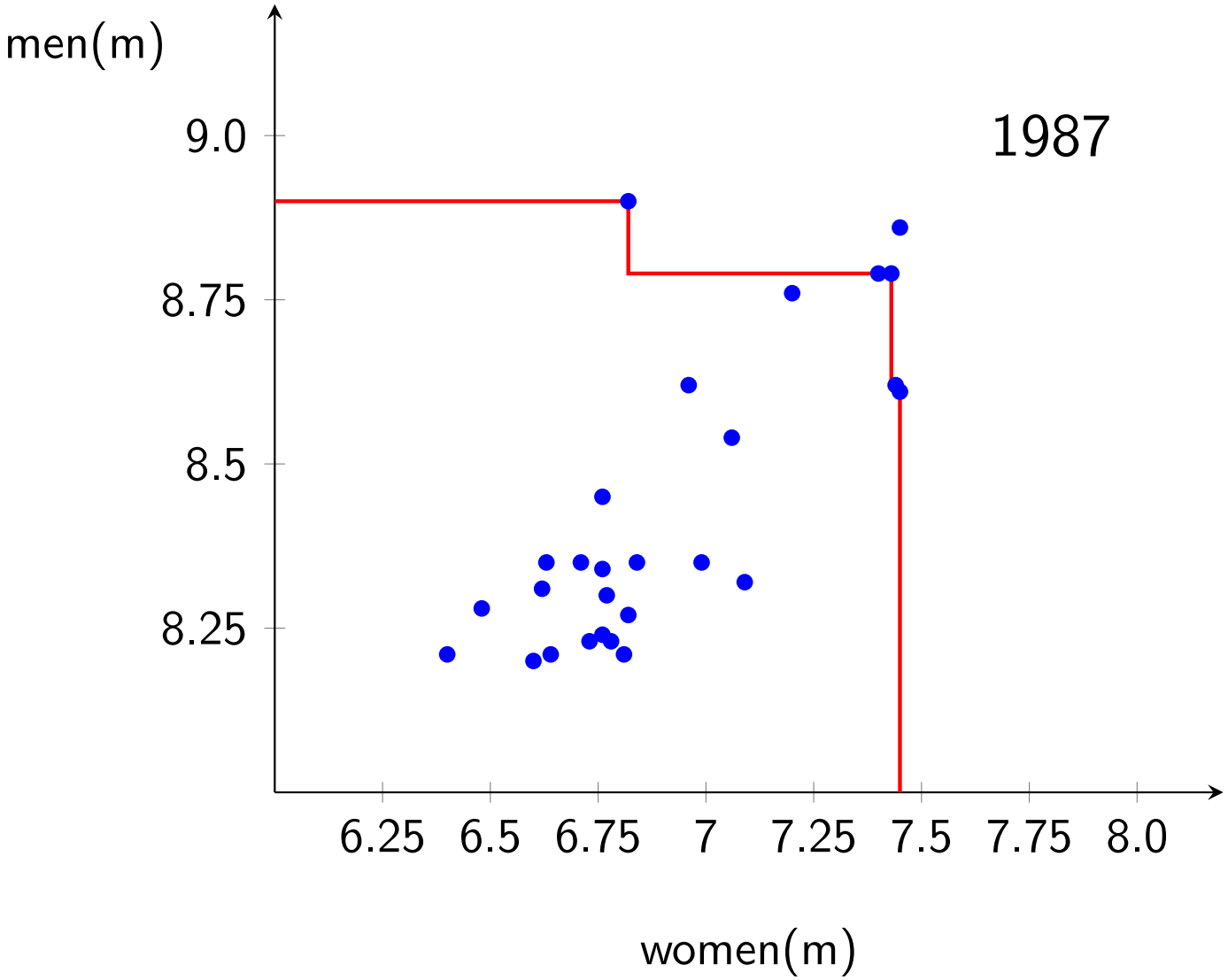
Men and Women's Best Long Jumps by Year



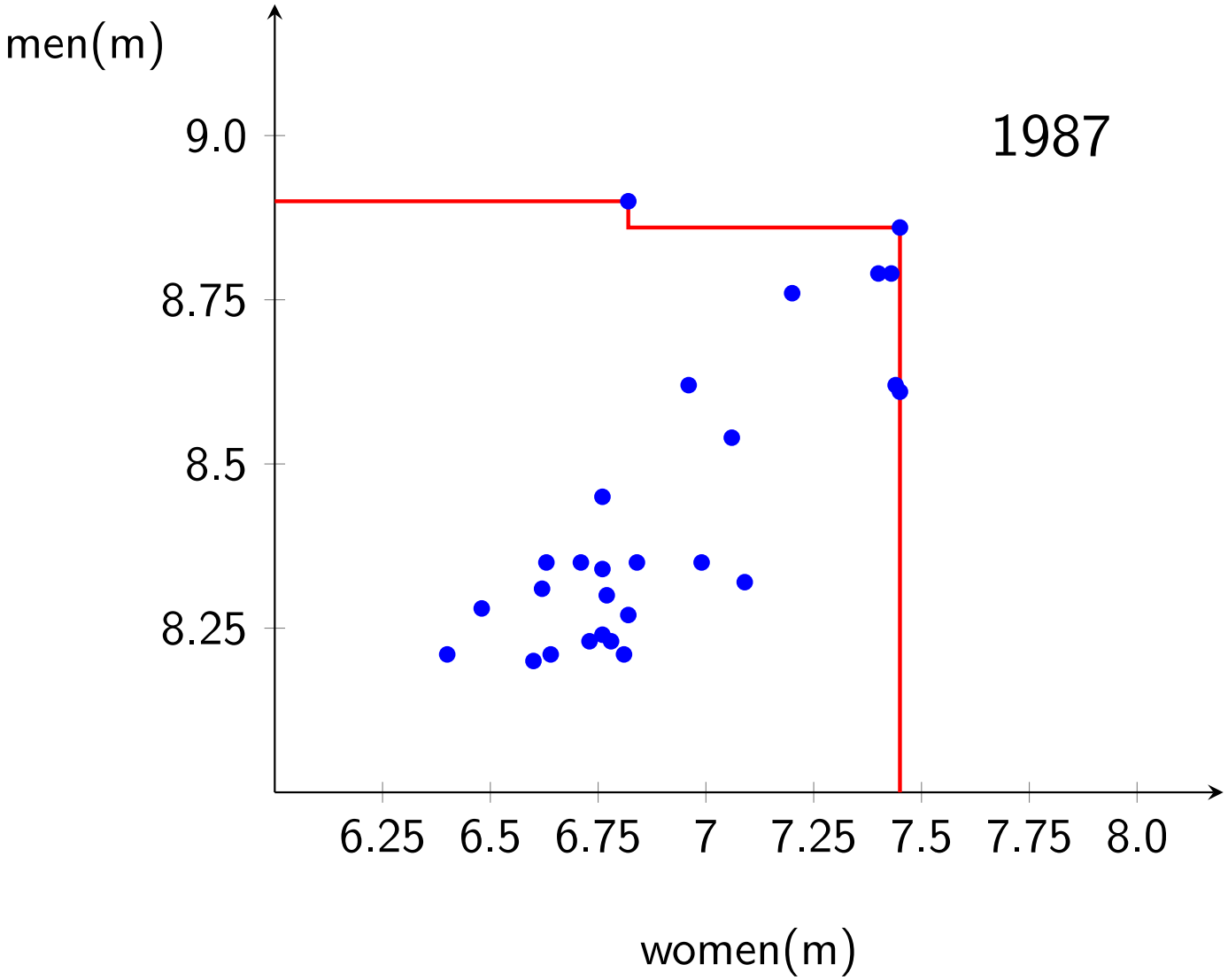
Men and Women's Best Long Jumps by Year



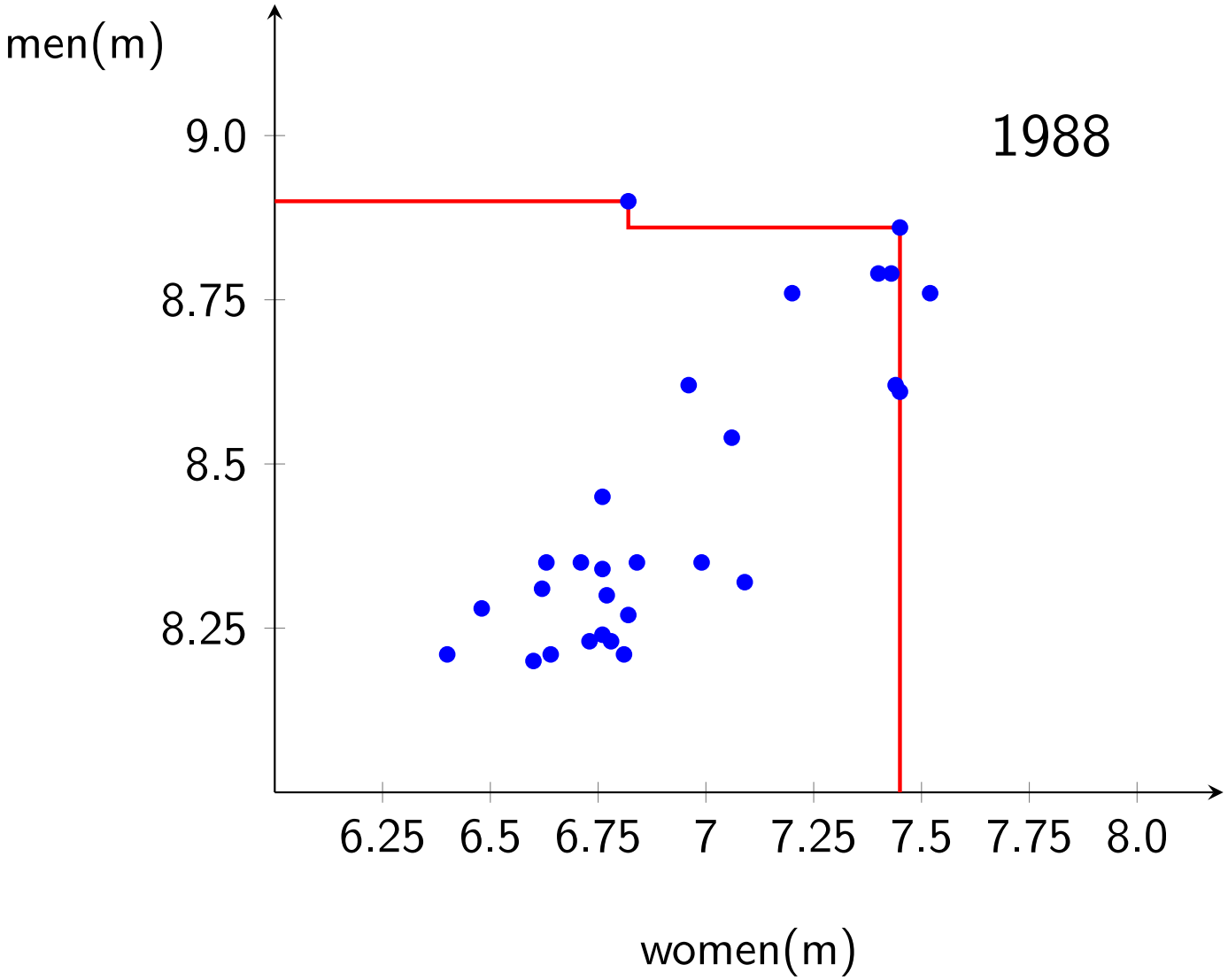
Men and Women's Best Long Jumps by Year



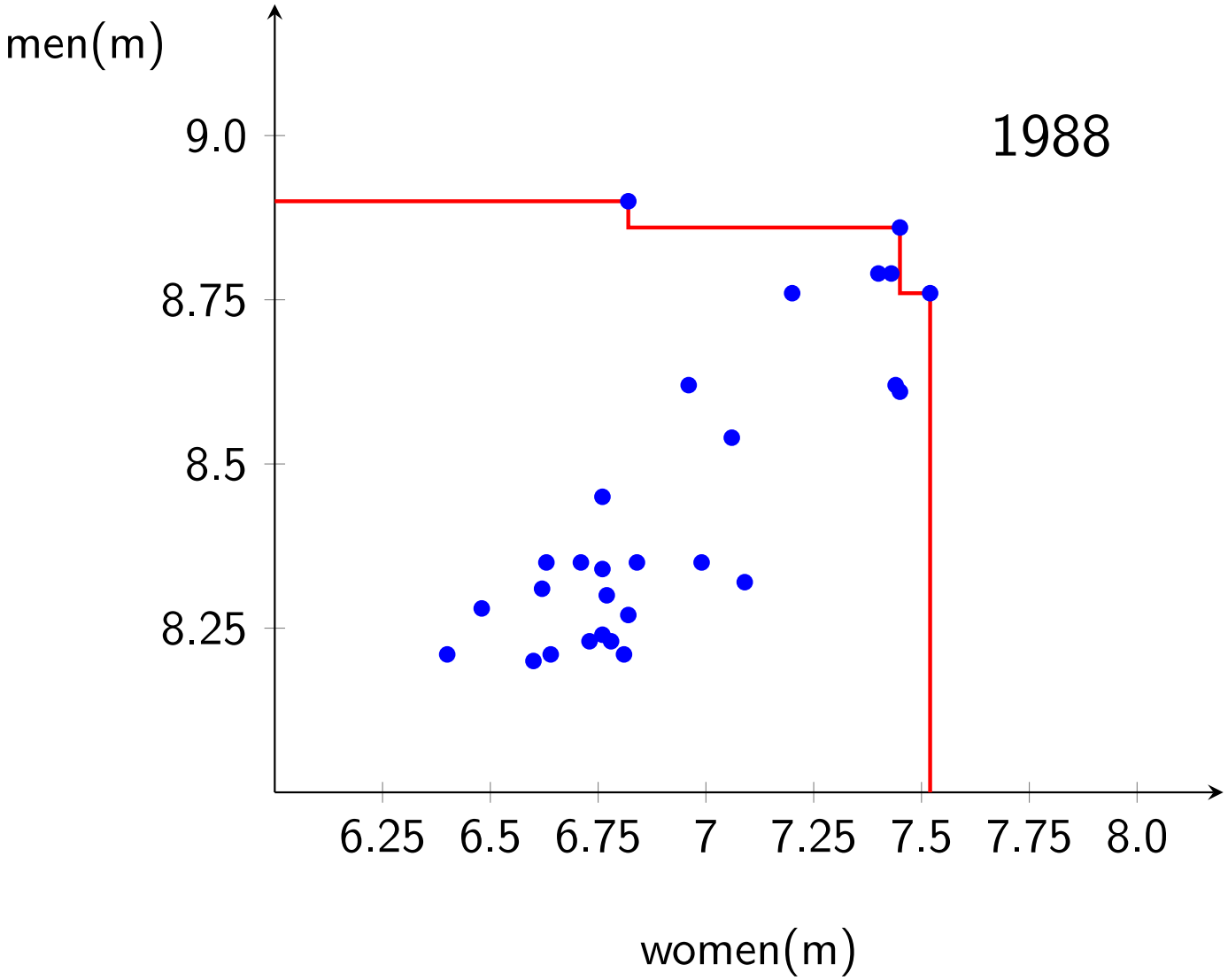
Men and Women's Best Long Jumps by Year



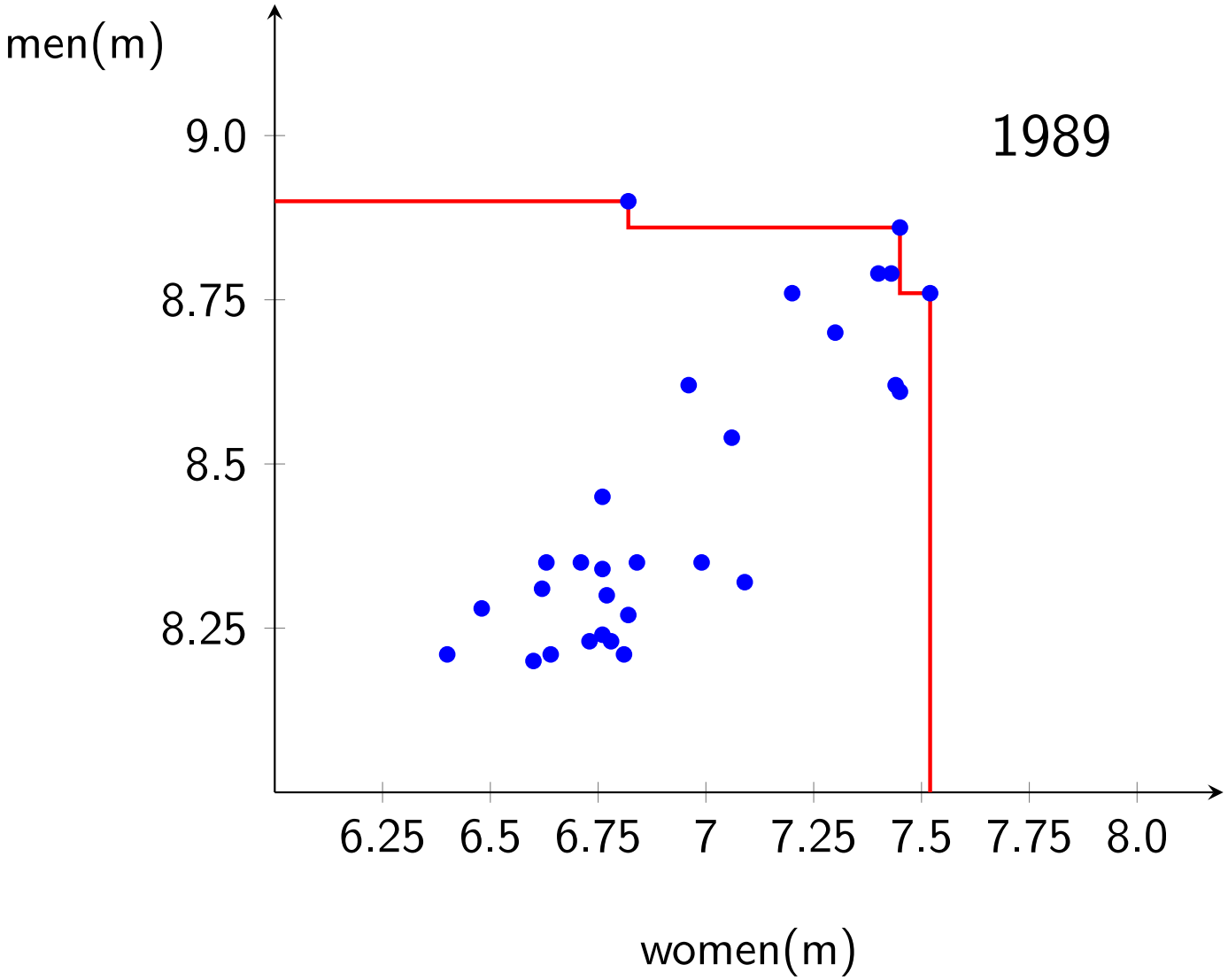
Men and Women's Best Long Jumps by Year



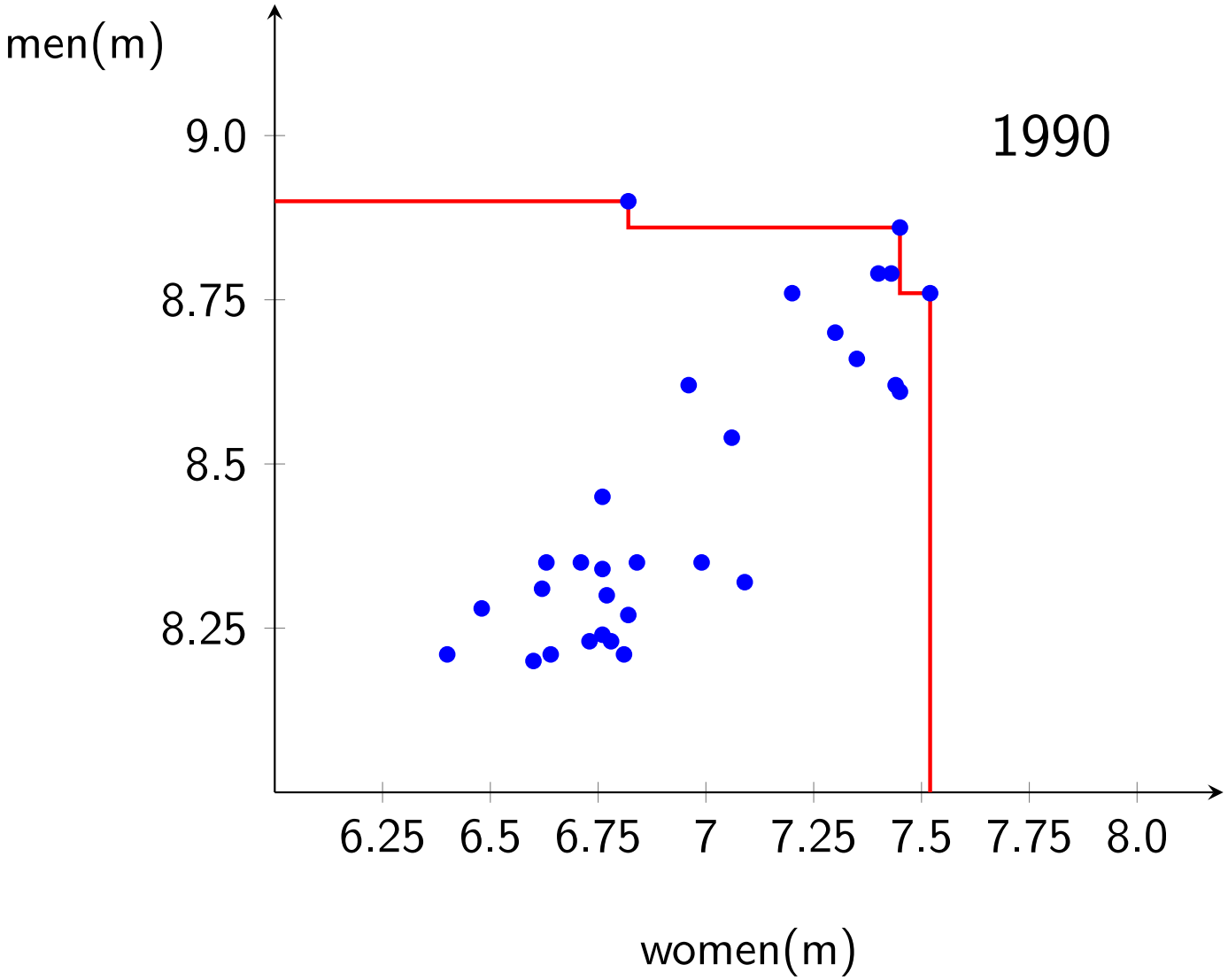
Men and Women's Best Long Jumps by Year



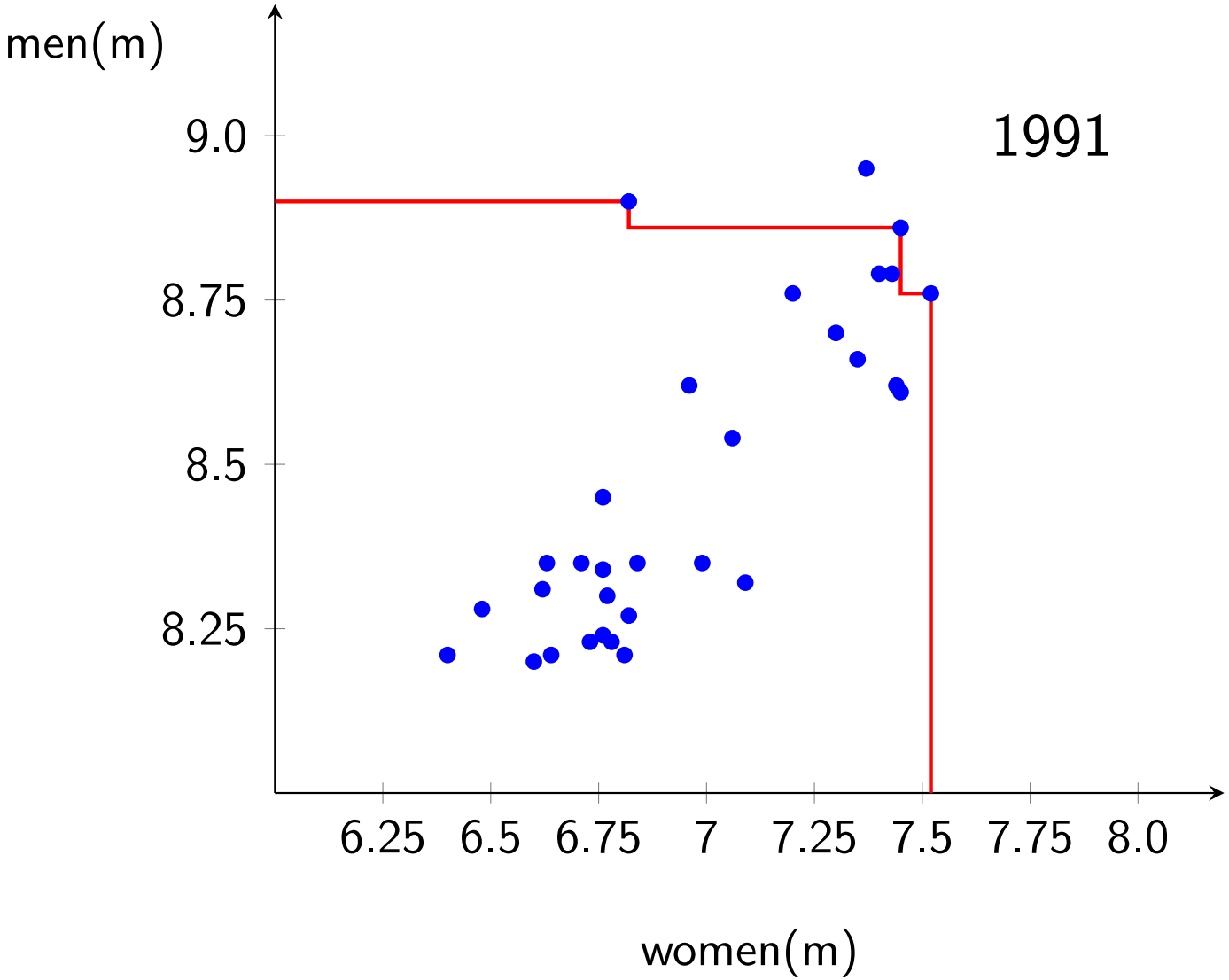
Men and Women's Best Long Jumps by Year



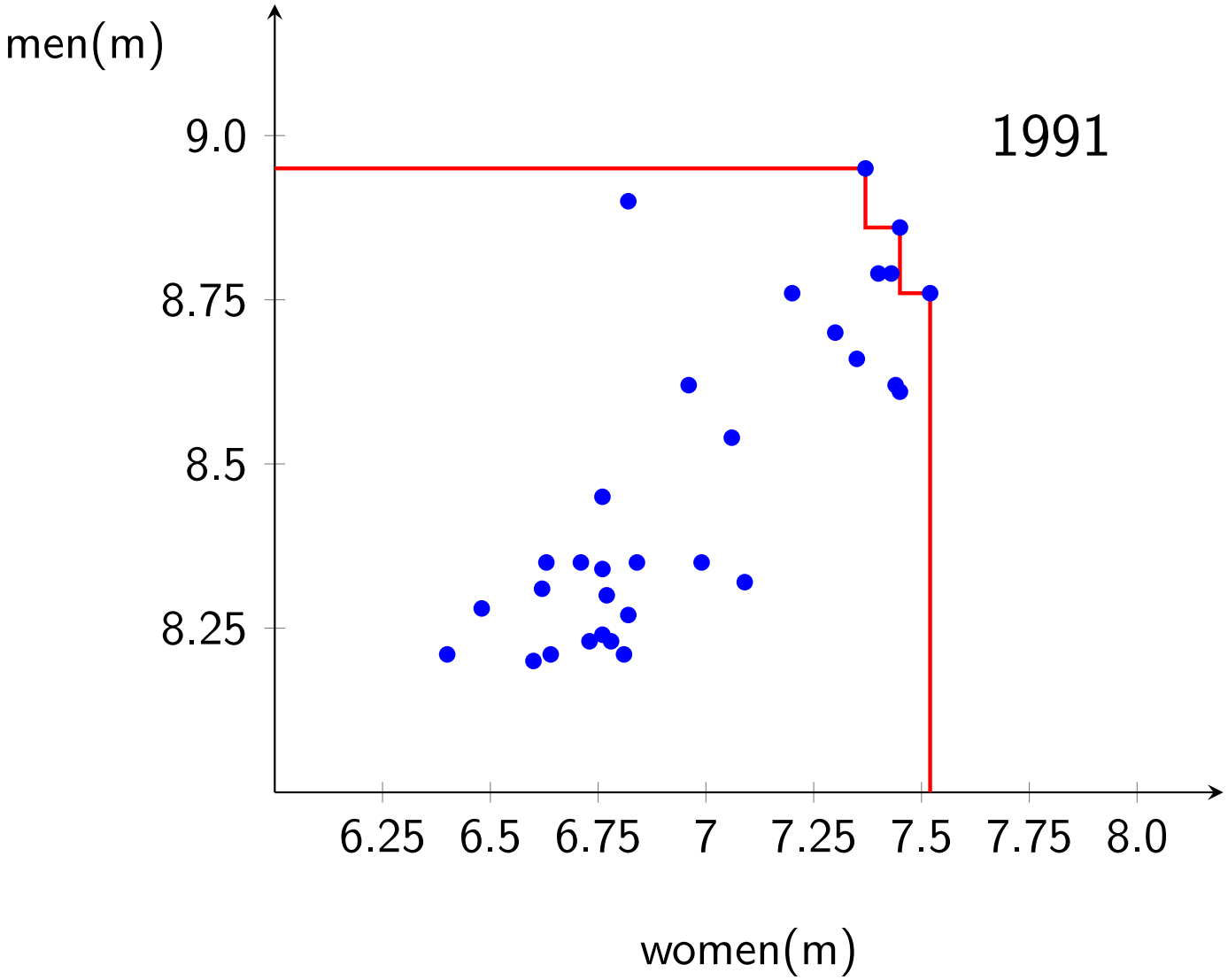
Men and Women's Best Long Jumps by Year



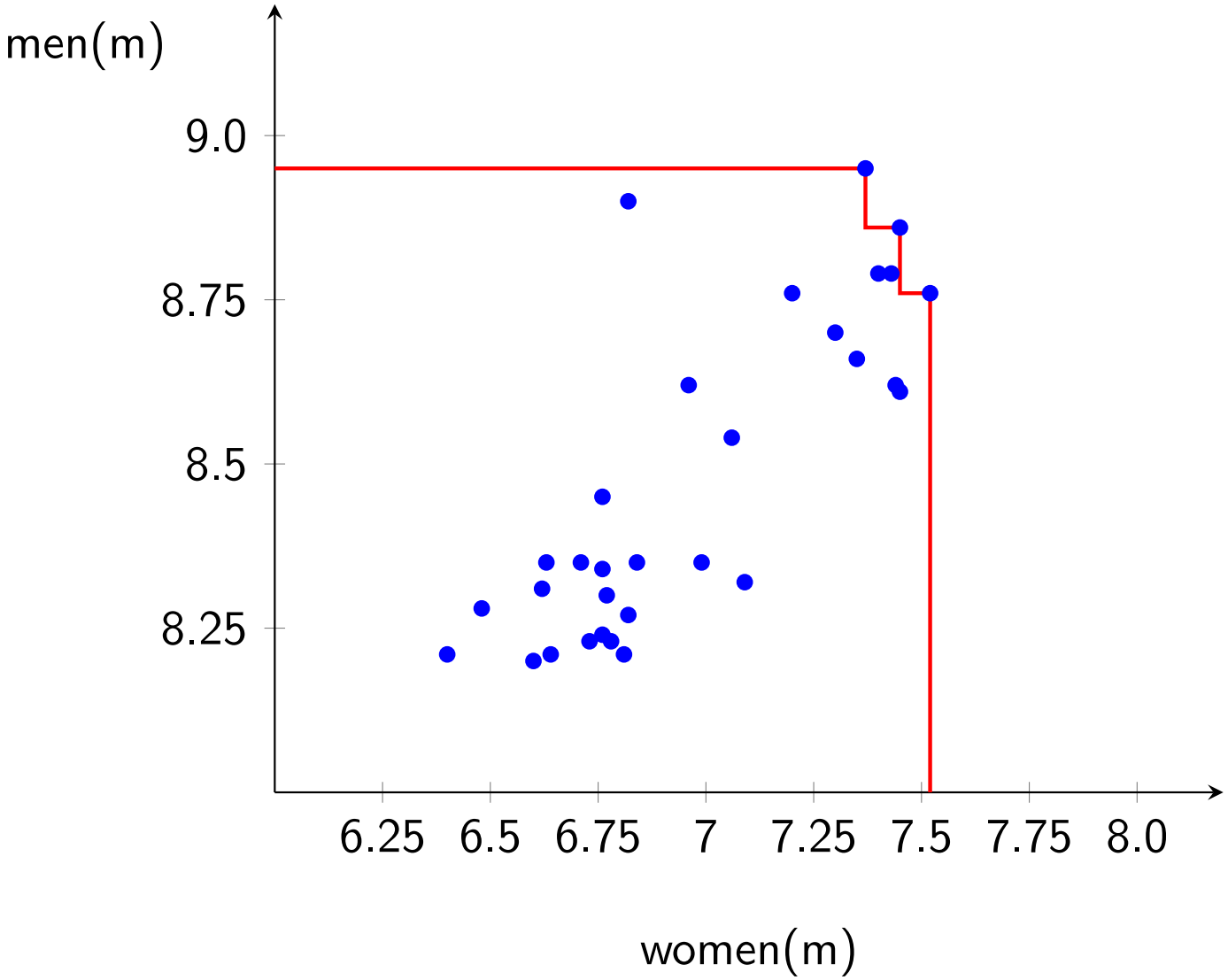
Men and Women's Best Long Jumps by Year



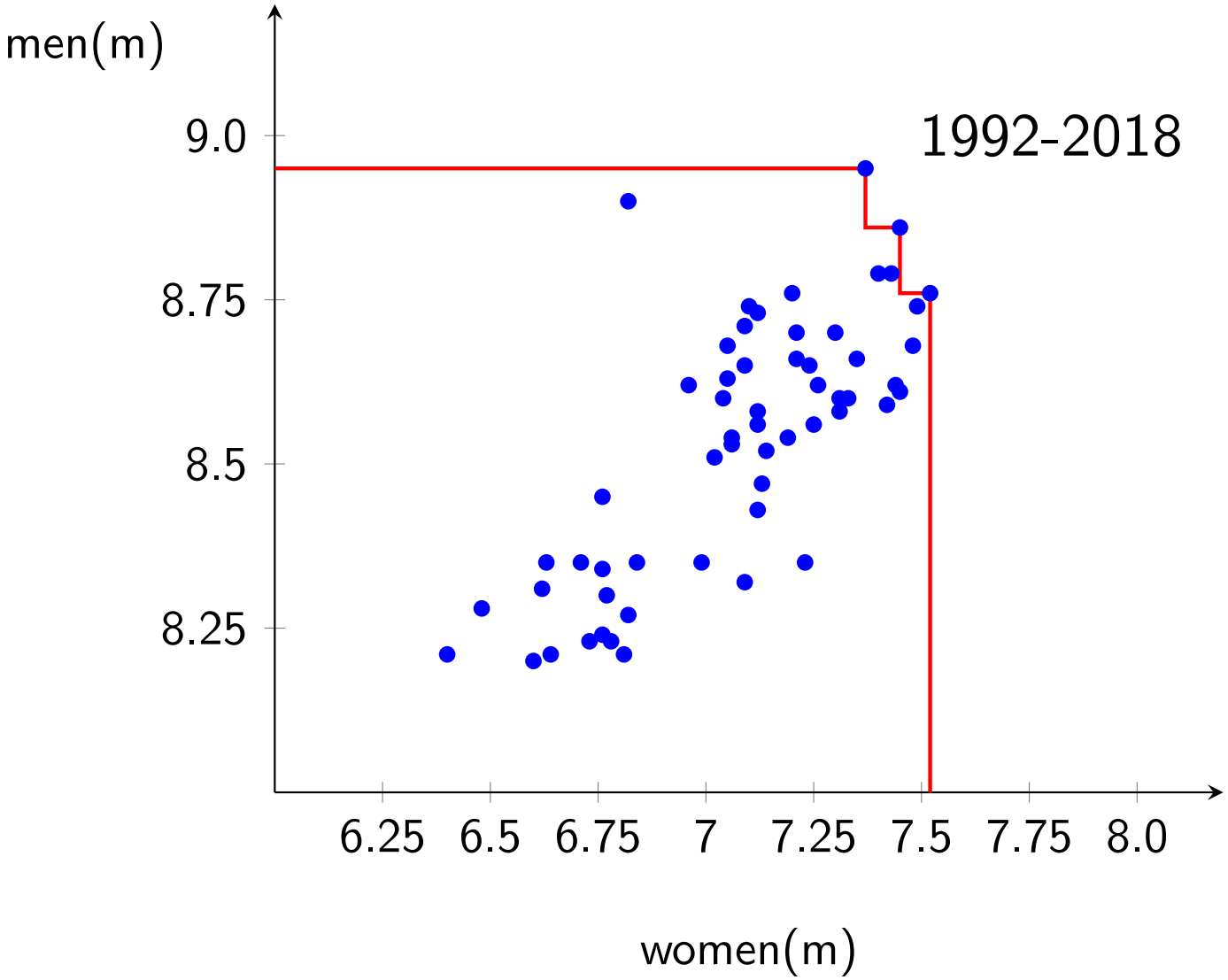
Men and Women's Best Long Jumps by Year



Men and Women's Best Long Jumps by Year



Men and Women's Best Long Jumps by Year



A "null" continuous multivariate model

We will examine the "null" model of iid d-dimensional observations $X^{(1)}, X^{(2)}, \ldots$ (copies of X) with independent coordinates, which WLOG are identically distributed.

In this talk, we will take the distribution of each coordinate to be either Uniform$(0,1)$ or (by taking negative logs) Exponential(1).

We will focus on the following questions, especially the second:

1. How can we sample (**g**enerate) multivariate records efficiently? (F & Naiman, 2019**g**)

2. How does the record-setting **f**rontier behave asymptotically? (F & Naiman, 2019**f**)

3. What can we say about the number of records **b**roken when a new record is set? (F, 2019**b**)

All three papers have been submitted, and all are on the arXiv.

Basic definitions: records

- Let $x_+ := \sum_{j=1}^{d} x_j$.
- Write $x \prec y$ to mean that $x_j < y_j$ for $1 \le j \le d$.
- Write $x \le y$ to mean that $x_j \le y_j$ for $1 \le j \le d$.
- We say that $X^{(k)}$ is a (Pareto) record (or that it sets a record at time k) if $X^{(k)} \not\prec X^{(i)}$ for all $1 \le i < k$.
- If $1 \le k \le n$, we say that $X^{(k)}$ is a current record (or remaining record, or maximum) at time n if $X^{(k)} \not\prec X^{(i)}$ for all $1 \le i \le n$.
- If $1 \le k \le n$, we say that $X^{(k)}$ is a broken record at time n if it is a record but not a current record, that is, if $X^{(k)} \not\prec X^{(i)}$ for all $1 \le i < k$ but $X^{(k)} \prec X^{(\ell)}$ for some $k < \ell \le n$; in that case, the observation corresponding to the smallest such ℓ is said to break or kill the record $X^{(k)}$.

Basic definitions: record counts

For $n \geq 1$ (or $n \geq 0$, with the obvious conventions),

- let R_n denote the number of records $X^{(k)}$ with $1 \leq k \leq n$,
- let r_n denote the number of remaining records at time n, and
- let $\beta_n := R_n - r_n$ denote the number of broken records.

Note that R_n and β_n (but not r_n) are nondecreasing in n.

For dimension $d \geq 2$, by standard consideration of concomitants,

$$r_n(d) \stackrel{\mathcal{L}}{=} R_n(d-1) \text{ for each } n$$

(but *not* as processes!).

Basic definitions: record-setting region and its frontier

- The record-setting region at time n is the (random) closed set of points

$$\mathrm{RS}_n := \{x \in \mathbb{R}^d : 0 \leq x \not\leq X^{(i)} \text{ for all } 1 \leq i \leq n\}.$$

- We call the (topol.) boundary of RS_n (relative to the closed positive orthant determined by the origin) its frontier and denote it by F_n.

Remark: The terminology for RS_n is natural since the next observation $X^{(n+1)}$ sets a record if and only if it falls in the record-setting region.

Note that

$$\mathrm{RS}_n = \{x \in \mathbb{R}^d : 0 \leq x \not\leq X^{(i)} \text{ for all } 1 \leq i \leq n \text{ such that } X^{(i)} \text{ is a current record at time } n\},$$

and that the current records at time n all belong to RS_n but lie on its frontier.

Observe also that F_n is a closed subset of RS_n.

Record frontier

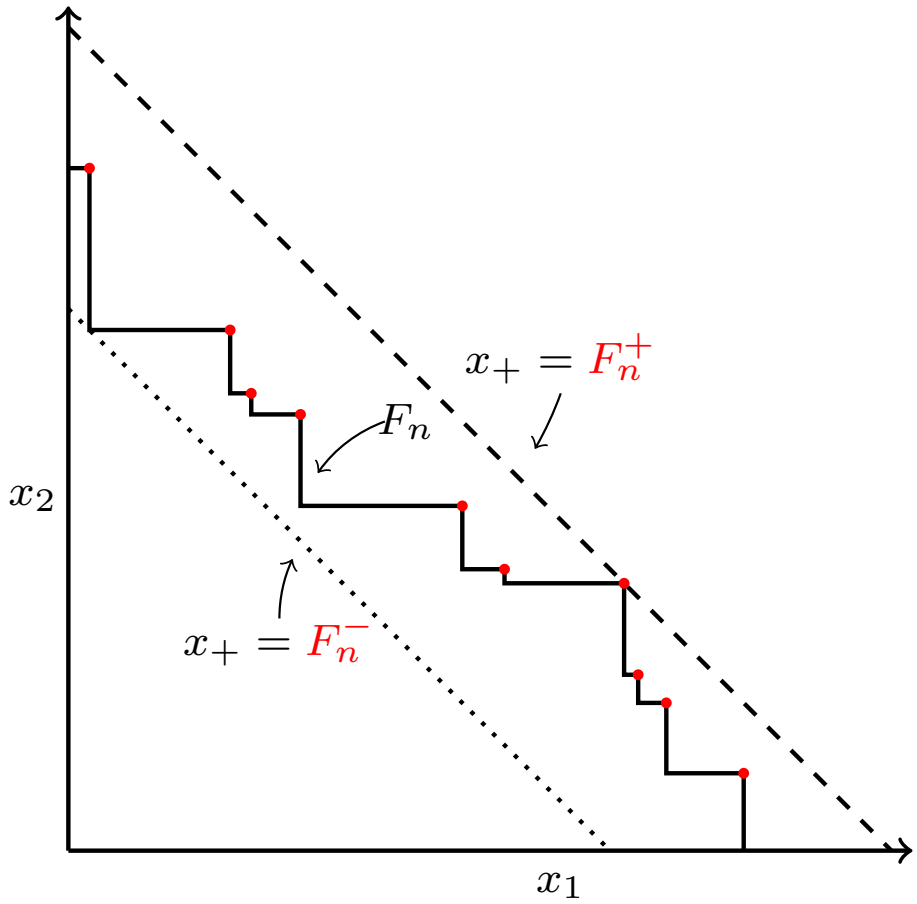


Figure: Record frontier F_n based on n bivariate Exponential(1) observations resulting in 10 current records (shown as solid points). The values $F_n^- = \min\{x_+ : x \in F_n\}$ and $F_n^+ = \max\{x_+ : x \in F_n\}$ determine two hyperplanes $x_+ = F_n^-$ and $x_+ = F_n^+$.
A new observation sets a record if and only if it falls in the region to the upper right of F_n.

1. Efficient sampling of multivariate records

1. How can we sample multivariate records efficiently?

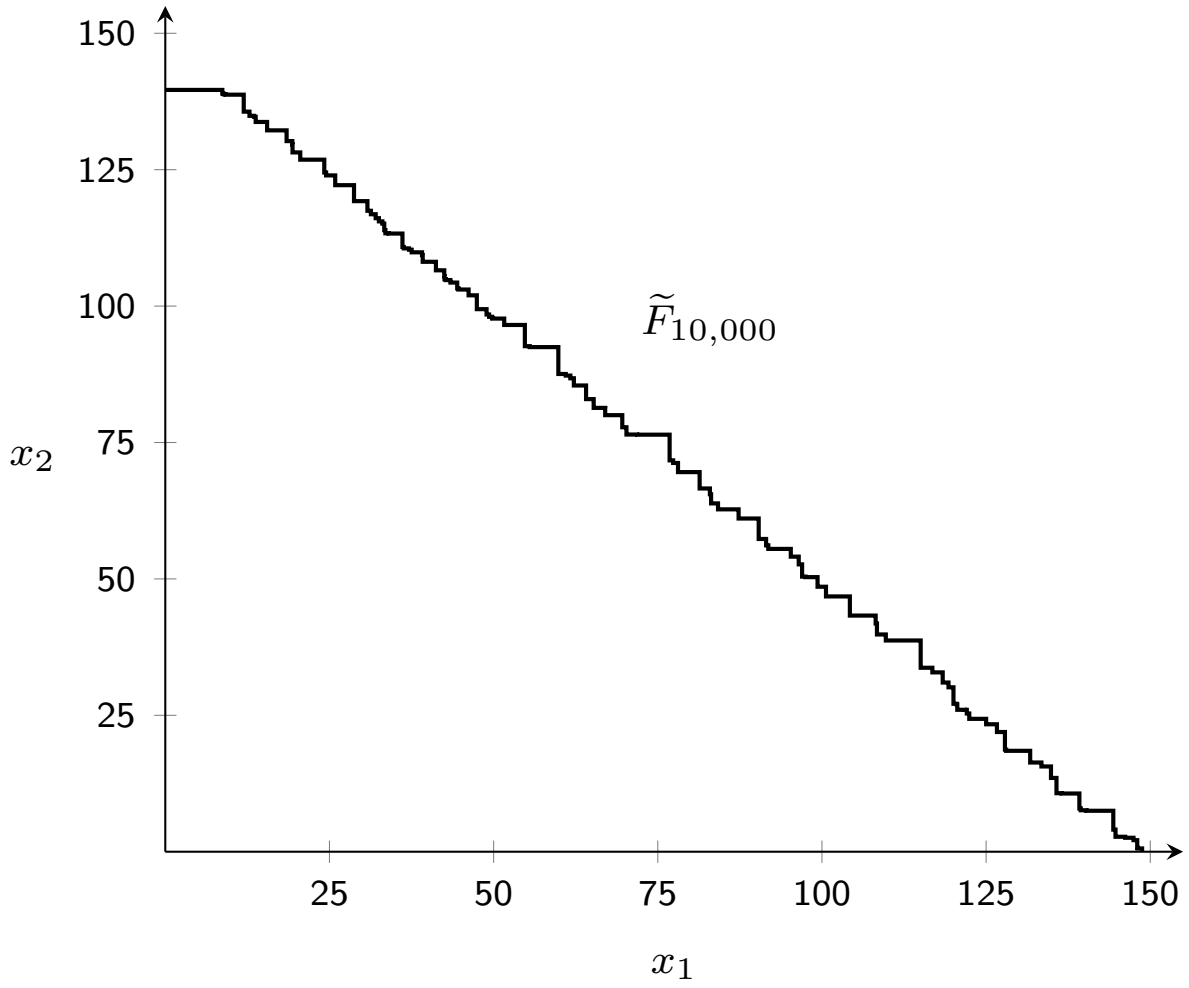


Figure: Record frontier $\widetilde{F}_{10,000}$ after 10,000 records generated using the importance-sampling algorithm described in F & Naiman (2019g).

1. Efficient sampling of multivariate records

How was this figure generated?

Spoiler: not by generating bivariate observations $X^{(1)}, X^{(2)}, \ldots$ and waiting for 10,000 records to be generated!

Let T_m denote the number of observations required for m records to be set. Among other more precise results, F & Naiman (2019f) show that

$$\frac{\mathrm{L}\, T_m}{(d!m)^{1/d}} \xrightarrow{\text{a.s.}} 1. \qquad (\mathrm{L} \equiv \ln)$$

With $d = 2$ and $m =$ 10,000 as in the figure, this gives the estimate

$$T_{10{,}000} \approx 10^{61} \quad (!!)$$

1. Efficient sampling of multivariate records

Suppose here that coordinates are distributed Uniform$(0, 1)$.

The record-setting region RS_n after n observations can be represented as the non-disjoint union

$$\mathrm{RS}_n = \cup_{g \in G_n} O_g^+$$

of the translated positive orthants

$$O_g^+ := \{y \in [0, 1)^d : y \geq g\},$$

where G_n is the set of minima of the frontier F_n.

a bivariate example: three slides back

a trivariate example: next slide

Trivariate example of generators

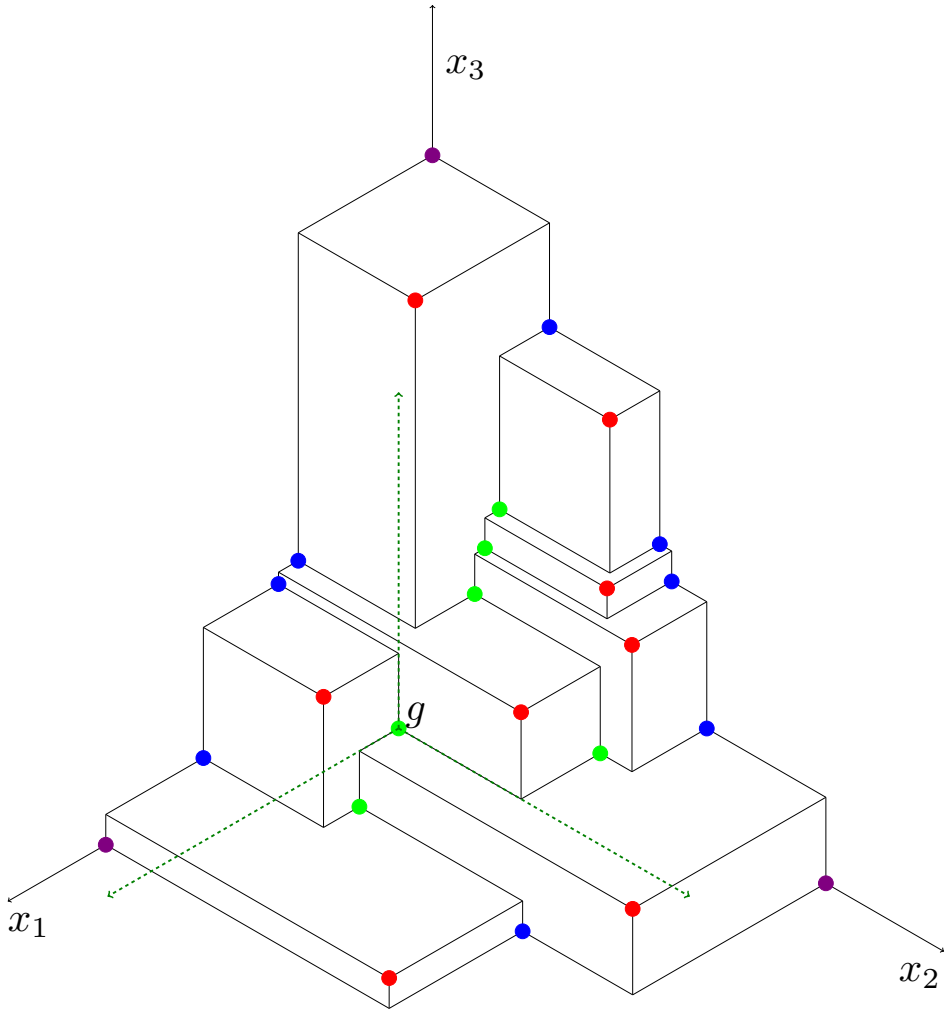


Figure: Example of a record frontier in dimension $d = 3$ with $\rho = 8$ remaining records shown in red and the resulting $\gamma = 17$ generators: three one-dimensional generators shown in violet, eight two-dimensional generators shown in blue, and six three-dimensional (interior) generators shown in green. The lower boundary of one of the orthants O_g^+ is shown using green dashed lines.

Generating the next multivariate record

Importance sampling subroutine for generating a new record R:

1. Sample $\mathbf{g}$ from G according to the distribution

$$\mathbb{P}(\mathbf{g} = g) = \frac{\mathbb{P}(X \in O_g^+)}{\sum_{h \in G} \mathbb{P}(X \in O_h^+)} = \frac{\prod(1 - g_j)}{\sum_{h \in G} \prod(1 - h_j)}.$$

2. Sample a r. vector $\mathbf{U}$ of independent Uniform$[0, 1)$ coords. and set

$$\mathbf{R}_j = \mathbf{g}_j + (1 - \mathbf{g}_j)\mathbf{U}_j, \quad j \in [d].$$

3. If $\mathbf{R} = R$, accept R as a new record with probability

$$1/\#\{g : R \in O_g^+\}.$$

   If R is rejected, repeat Steps 1–3.

4. Update G to G', as described (& analyzed) in F & Naiman (2019g).

Remarks:

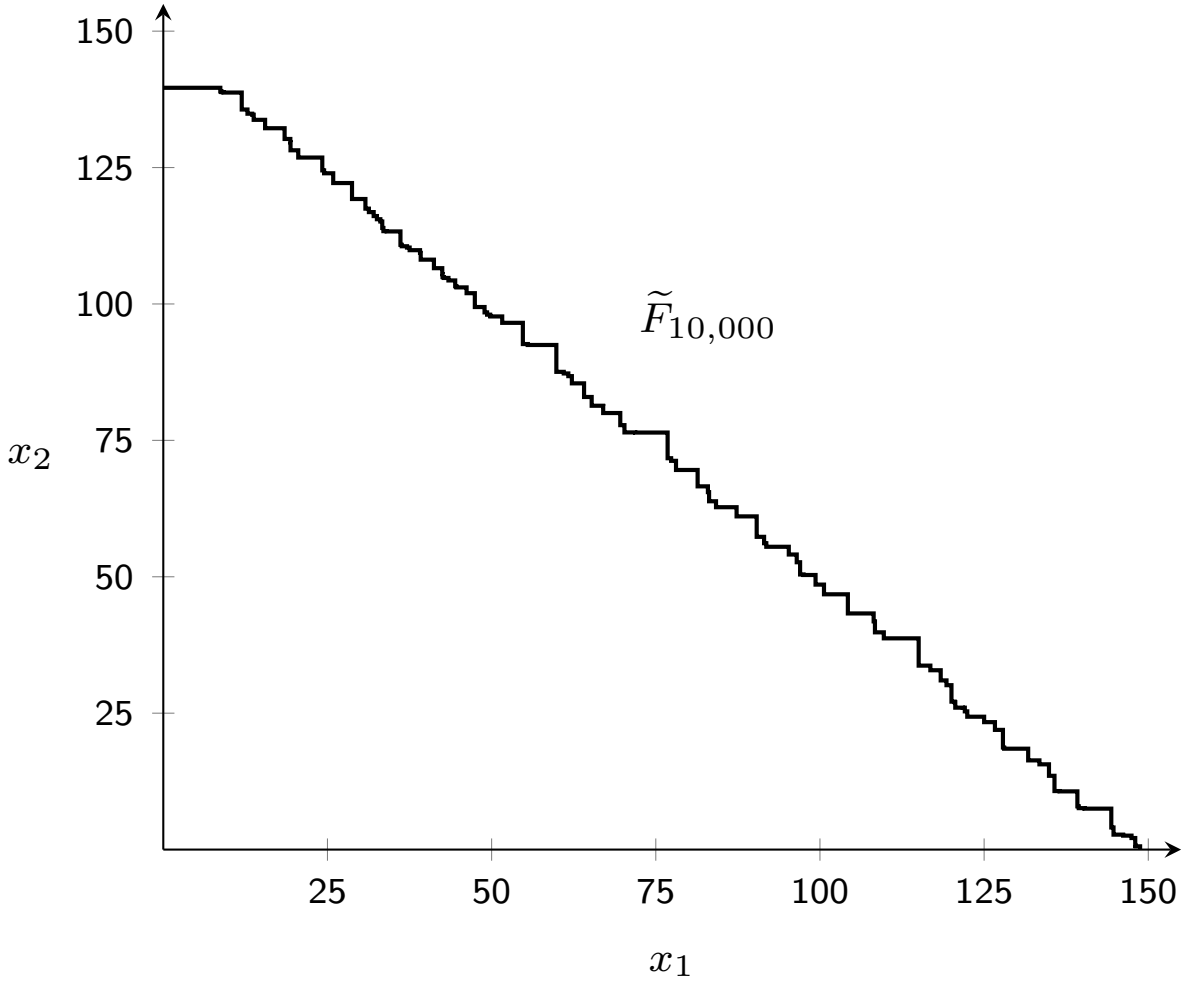
1. Results are easier to discover empirically in "records-time" but easier to prove in "observations-time"!

2. Results in "observations-time" can be transferred to results in "records-time" using time change arguments that exploit information concerning the counts R_n or (equivalently) the record epochs T_m: See Sections 4–5 of F & Naiman (2019f).

From now on, we take observation coords. to be distributed Exp(1).

The frontier is nearly planar

Recall this nearly planar (i.e., linear) frontier figure:



Questions about the frontier

Questions to ask about the frontier:

(a) Where is the frontier located? I.e., what is an approximating plane?

(b) How thick (i.e., wide) is the frontier?

Approximate location of frontier: $x_+ = \mathrm{L}\, n$

Question (a) is easy to answer: Deviations of the sum of coordinates for a generic current record at time n from $\mathrm{L}\, n$ are typically of constant order.

Observe that the conditional distribution of $X_+^{(k)}$ given that $X^{(k)}$ is a current record at time n doesn't depend on $k \in \{1, \ldots, n\}$; in particular, it's the conditional distribution of $X_+^{(n)}$ given that $X^{(n)}$ sets a record. Let Y_n be a random variable with that distribution.

Let G denote a random variable with the standard Gumbel distribution (i.e., distribution function $x \mapsto e^{-e^{-x}}$, $x \in \mathbb{R}$), and write $\xrightarrow{\mathcal{L}}$ for convergence in law (i.e., in distribution).

Theorem

We have

$$Y_n - \mathrm{L}\, n \xrightarrow{\mathcal{L}} G.$$

Approximate location of frontier: $x_+ = \mathrm{L}\, n$

Proof: This is quite elementary. Let p_n denote the probability that $X^{(n)}$ sets a record. Fix $n \geq 2$ for the moment. For $x \succ 0$ we have

$$\begin{aligned}
&\mathbb{P}(X^{(n)} \in dx \,|\, X^{(n)} \not\prec X^{(i)} \text{ for all } 1 \leq i \leq n) \\
&= p_n^{-1}\, \mathbb{P}(X^{(n)} \in dx,\, X^{(n)} \not\prec X^{(i)} \text{ for all } 1 \leq i \leq n) \\
&= p_n^{-1}\, \mathbb{P}(X^{(n)} \in dx,\, x \not\prec X^{(i)} \text{ for all } 1 \leq i \leq n-1) \\
&= p_n^{-1}\, \mathbb{P}(X^{(n)} \in dx)\, \mathbb{P}(x \not\prec X^{(i)} \text{ for all } 1 \leq i \leq n-1) \\
&= p_n^{-1} e^{-x_+} [1 - \mathbb{P}(x \prec X^{(1)})]^{n-1}\, dx = p_n^{-1} e^{-x_+} (1 - e^{-x_+})^{n-1}\, dx,
\end{aligned}$$

and so the conditional density depends on x only through x_+. It follows that the density $f_n(y)$ of Y_n satisfies

$$f_n(y) = p_n^{-1} \frac{y^{d-1}}{(d-1)!} e^{-y} (1 - e^{-y})^{n-1}, \quad y > 0.$$

Using the well-known $p_n \sim n^{-1} (\mathrm{L}\, n)^{d-1}/(d-1)!$ as $n \to \infty$, it is easy to check that, for each fixed $z \in \mathbb{R}$, the density of $Y_n - \mathrm{L}\, n$ at z converges to the standard Gumbel density $e^{-z} e^{-e^{-z}}$ as $n \to \infty$. The claimed result thus follows from Scheffé's theorem, which shows that there is in fact convergence in total variation. □

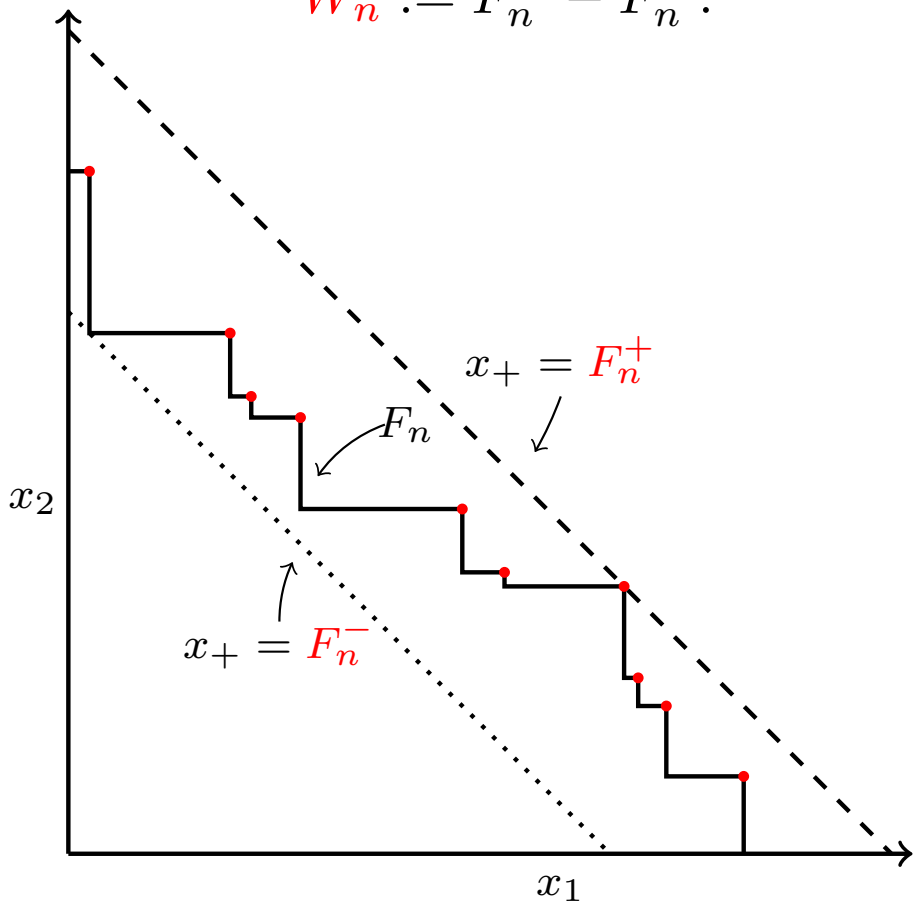
Width of frontier

- Recall that F_n denotes the frontier of RS_n, and let

$$F_n^- := \min\{x_+ : x \in F_n\} \quad \text{and} \quad F_n^+ := \max\{x_+ : x \in F_n\}.$$

- We define the width of F_n as

$$W_n := F_n^+ - F_n^-.$$



Rough statement of main results about frontier

Very rough statement of main results (L_k denotes the kth iterate of L):

Unlike for the generic Y_n,

- deviations of F_n^+ from $\mathrm{L}\, n$ are exactly of order $\mathrm{L}_2\, n$;
- deviations of F_n^- from $\mathrm{L}\, n$ are of smaller order than $\mathrm{L}_2\, n$.
- Therefore, the width W_n of the frontier is exactly of order $\mathrm{L}_2\, n$.

In the following slides I will state the results less roughly—but still not as sharply as in F & Naiman (2019f).

The leading edge F_n^+ of the frontier

Lemma (characterization of F_n^+)

We have

$$F_n^+ = \max\{X_+^{(k)} : 1 \le k \le n\},$$

which is nondecreasing in n.

Proof: The current records at time n all belong to F_n, and broken records and non-records all have coordinate-sums (strictly) smaller than some current record. Thus $F_n^+ \ge \max\{X_+^{(k)} : 1 \le k \le n\}$.

Conversely, if $x \in F_n$, then $x \le X^{(i)}$ for some i; it follows that $F_n^+ \le \max\{X_+^{(k)} : 1 \le k \le n\}$. □

So the process F^+ is simply the partial-maximum process corresponding to iid Gamma$(d, 1)$ random variables, and classical extreme value theory due to Jack Kiefer (1972) (involving rather sophisticated use of the Borel–Cantelli lemmas) can be brought to bear.

Asymptotic behavior of the leading edge F_n^+

Theorem (derived from Kiefer, Sixth Berkeley Symposium, 1972)

(a) Typical behavior of F^+:

$$\frac{F_n^+ - \mathrm{L}\, n}{\mathrm{L}_2\, n} \xrightarrow{\mathrm{P}} d - 1.$$

(b) Almost sure behavior for F^+:

$$\liminf \frac{F_n^+ - \mathrm{L}\, n}{\mathrm{L}_2\, n} = d - 1 < d = \limsup \frac{F_n^+ - \mathrm{L}\, n}{\mathrm{L}_2\, n} \text{ a.s.}$$

Remark: In fact, one can show rather simply from this theorem and the fact that F^+ has nondecreasing sample paths that the set of limit points of the sequence $(F_n^+ - \mathrm{L}\, n)/\mathrm{L}_2\, n$ is almost surely the closed interval $[d-1, d]$.

Asymptotic behavior of the trailing edge F_n^-

Theorem

(a) Typical behavior of F^-:

$$\frac{F_n^- - \mathrm{L}\, n}{\mathrm{L}_2\, n} \xrightarrow{\mathrm{P}} 0.$$

(b) Almost sure behavior for F^-: *If $d \geq 2$, then*

$$\lim \frac{F_n^- - \mathrm{L}\, n}{\mathrm{L}_2\, n} = 0 \text{ a.s.}$$

I will come back to the proof of this theorem as time permits.

Asymptotic behavior of the width W_n

Theorem

(a) Typical behavior of W:

$$\frac{W_n}{\mathrm{L}_2\, n} \xrightarrow{\mathrm{P}} d-1.$$

(b) Almost sure behavior for W: *If $d \geq 2$, then*

$$\liminf \frac{W_n}{\mathrm{L}_2\, n} = d-1 < d = \limsup \frac{W_n}{\mathrm{L}_2\, n} \text{ a.s.},$$

and, in particular,

$$W_n = \Theta(\mathrm{L}_2\, n) \text{ a.s.} \quad \square$$

Remark:

(a) When $d = 1$, at each time $n \geq 1$ there is one curr. rec., $F_n^+ = F_n^-$ is the value of that record, RS_n is the interval $[F_n^+, \infty)$, and $W_n = 0$.

(b) Part (b) can be strengthened to the conclusion that the set of limit points of the sequence $W_n/\,\mathrm{L}_2\, n$ is a.s. the closed interval $[d-1, d]$.

Proof of theorem about trailing edge F_n^-: upper bound

Lemma (upper bound on F_n^-)

(a) *Let $1 \le m \le n$. Define*

$$B_{m,n} := m^{\text{th}}\text{-largest value among } X_+^{(k)} \text{ with } 1 \le k \le n.$$

Then, over the event $\{r_n \ge m\}$ that there are at least m remaining records at time n, we have

$$F_n^- \le B_{m,n}.$$

(b) *The procs. F^- & $B_{m,\cdot}$ (for any m) have nondecreasing sample paths.*

Proof: (a) Over the event $\{r_n \ge m\}$, F_n^- is at most the mth-largest sum of coordinates of remaining records, which is in turn at most $B_{m,n}$.

(b) The asserted monotonicity is clear for the bounding processes. The asserted monotonicity of F^- follows easily from the observation that $F_{n+1} \subseteq \mathsf{RS}_{n+1} \subseteq \mathsf{RS}_n$. □

Proof of theorem about trailing edge F_n^-: upper bound

Lemma

Assume $d \geq 2$. Let r_n denote the number of remaining records at time n. Then

$$\liminf \frac{r_n}{(\mathrm{L}\, n)/(d\, \mathrm{L}_2\, n)} \geq 1 \text{ a.s.}$$

The proof of this lemma is in the paper. □

We can now establish half of the asserted a.s. behavior of F^- in the following "top outer boundary" form: If $d \geq 2$, then

$$\mathbb{P}(F_n^- \geq \mathrm{L}\, n + c\, \mathrm{L}_2\, n \text{ i.o.}) = 0 \text{ if } c > 0.$$

Proof: In light of the last two lemmas, it is sufficient that for each fixed positive integer m we have

$$\mathbb{P}\left(B_{m,n} \geq \mathrm{L}\, n + \frac{a}{m}\, \mathrm{L}_2\, n \text{ i.o.}\right) = 0$$

if $a > 1$. But this is known from Kiefer(1972). □

Proof of theorem about trailing edge F_n^-: lower bound

Towards a stochastic lower bound on F_n^-:

For this we use the definitions of the frontier F_n and the closed record-setting region RS_n to argue as follows. For any set $S \subseteq \mathbb{R}^d$, let $N_n(S)$ denote the number of observations $X^{(i)}$ with $1 \le i \le n$ that fall in S. Then

$$\begin{aligned}
\{F_n^- \le b\} &= \{x_+ \le b \text{ for some } x \in F_n\} = \{x_+ \le b \text{ for some } x \in \mathsf{RS}_n\} \\
&= \{x_+ \le b \text{ for some } x \ge 0 \text{ satisfying } x \not\prec X^{(i)} \text{ for all } 1 \le i \le n\} \\
&= \bigcup_{x \ge 0:\, x_+ \le b} \{N_n(O_x^+) = 0\} \\
&= \bigcup_{x \ge 0:\, x_+ = b} \{N_n(O_x^+) = 0\}.
\end{aligned}$$

The difficulty with upper-bounding the probability of this event is of course that the last union is uncountable.

A geometric lemma, illustrated

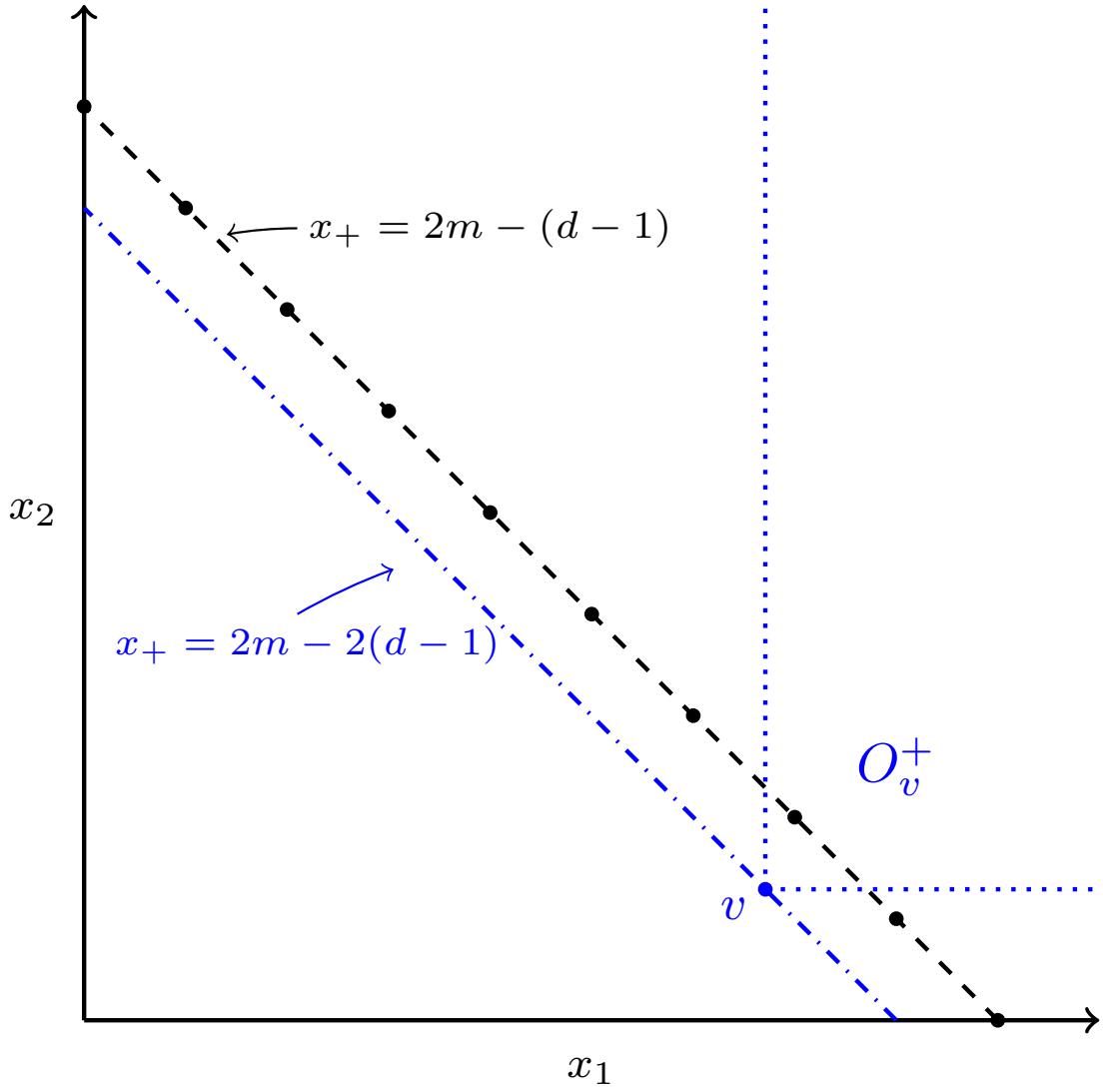


Figure: Geometric lemma illustrated for $d = 2$. Given v with $v_+ = 2m - 2(d-1)$, the orthant O_v^+ determined by v must contain a point i with integer coordinates on the hyperplane $x^+ = 2m - (d-1)$.

A key geometric lemma

Lemma (key geometric lemma)

Given a positive integer $m \geq d-1$, *and* $0 \leq x \in \mathbb{R}^d$ *with*

$$x_+ = 2m - 2(d-1), \tag{1}$$

there exists $0 \leq i \in \mathbb{Z}^d$ *with*

$$i_+ = 2m - (d-1) \tag{2}$$

such that

$$O_i^+ \subseteq O_x^+. \tag{3}$$

Proof: We need to prove the existence of $0 \leq i \in \mathbb{Z}^d$ satisfying (2) and (3) (i.e., $x \leq i$). The frugal choice $0 \leq i' \in \mathbb{Z}^d$ defined by

$$i'_j := \lceil x_j \rceil, \quad j = 1, \ldots, d,$$

satisfies (3) but not necessarily (2). However, using (1) we observe that i'_+ is at least the integer

$$x_+ = 2m - 2(d-1)$$

and strictly less than the integer $2m - 2(d-1) + d = 2m - (d-2)$, i.e., is at most $2m - (d-1)$. Thus we need only (arbitrarily) "sweeten" (i.e., add 1 to) precisely $2m - (d-1) - i'_+ \in \mathbb{Z} \cap [0, d-1]$ of the entries i'_j to obtain i with the desired properties. □

A stochastic lower bound on F_n^-

Now using routine arguments/calculations (including finite subadditivity), we obtain the following result:

Proposition (Stochastic lower bound on F_n^-)

Let $0 \le b_n < \mathrm{L}\, n$ with $b_n = (1 - o(1))\, \mathrm{L}\, n$ and $\mathrm{L}\, n - b_n \to \infty$. Then

$$\mathbb{P}(F_n^- \le b_n) \le (\mathrm{L}\, n)^{d-1} \exp\left[- \exp\left\{ (1 + o(1)) \tfrac{1}{2} (\mathrm{L}\, n - b_n) \right\} \right]. \quad \square$$

We can now establish the other half of the asserted a.s. behavior of F^- in the following "bottom outer boundary" form on the scale of $\mathrm{L}_3\, n$:

$$\mathbb{P}(F_n^- \le \mathrm{L}\, n - 3\, \mathrm{L}_3\, n \text{ i.o.}) = 0.$$

Proof: The process F^- has nondecreasing sample paths. From this it follows that if (b_n) is (ultimately) monotone nondecreasing and (n_j) is any strictly increasing sequence of positive integers, then

$$\{F_n^- \le b_n \text{ i.o.}(n)\} \subseteq \{F_{n_j}^- \le b_{n_{j+1}} \text{ i.o.}(j)\}.$$

To complete the proof, we choose $b_n \equiv \mathrm{L}\, n - 3\, \mathrm{L}_3\, n$ and $n_j \equiv 2^j$, bound $\mathbb{P}(F_{n_j}^- \le b_{n_{j+1}})$ using the proposition, and apply 1BC. $\square$

3. Broken records: simulation of 100,000 bivariate records

N_k = number of records that break k current records.

k	N_k	$\tilde{p}_k$
0	50,334	0.50334
1	24,667	0.24667
2	12,507	0.12507
3	63,35	0.06335
4	3,040	0.03040
5	1,571	0.01571
6	782	0.00782
7	364	0.00364
8	202	0.00202
9	94	0.00094
10	48	0.00048
11	24	0.00024
12	18	0.00018
13	8	0.00008
14	4	0.00004
16	1	0.00001
17	0	0.00000
18	1	0.00001

Geometric$(1/2)$ distribution for bivariate record-breaking

Here is the main theorem of F (2019b). That paper also presents a first-order correction term.

Theorem

Consider our "null" bivariate model for observations. Let $K_n = -1$ *if the* n^{th} *observation is not a new record, and otherwise let* K_n *denote the number of remaining records killed by the* n^{th} *observation. Then* K_n*, conditionally given* $K_n \geq 0$*, converges in distribution to* $G - 1$*, where* $G \sim \text{Geometric}(1/2)$*, as* $n \to \infty$*.*

The paper provides a possible roadmap for the proof of a stronger result:

Conjecture

The fractions $\tilde{p}_{M,k}$ *of the first* M *records that break precisely* k *remaining records satisfy*

$$\sup_{k \geq 0} \left| \tilde{p}_{M,k} - 2^{-(k+1)} \right| \xrightarrow{\text{a.s.}} 0 \text{ as } M \to \infty.$$

Record-breaking in higher dimensions

Similarly, the following conjecture arises from data generated by the importance-sampling algorithm for higher dimensions:

Conjecture

Consider dimension $d \geq 2$. Let $f_{d,m}$ denote the fraction of the first m records set that break 0 records. Then there exist constants $p_d \in (0,1)$ such that, almost surely, $f_{d,m} \to p_d$ as $m \to \infty$. Further, $p_d \to 1$ as $d \to \infty$.

- The data also suggest that perhaps $p_d = 1 - d^{-1}$ for every $d \geq 2$.
- Even for $d = 2$, the conjecture is stronger than what is proved in F (2019b).
- For $d \geq 3$ and $k \geq 1$, we do not know what to conjecture concerning the limiting behavior of the fraction of the first m records set that break k records.